


# Numerical Regularization for Atmospheric Inverse Problems

**Adrian  
Doicu**

**Thomas  
Trautmann**

**Franz  
Schreier**

 Springer

PRAXIS

# Numerical Regularization for Atmospheric Inverse Problems

---



Adrian Doicu, Thomas Trautmann, and  
Franz Schreier

---

# Numerical Regularization for Atmospheric Inverse Problems



Published in association with  
**Praxis Publishing**  
Chichester, UK



Dr Adrian Doicu  
Professor Dr Thomas Trautmann  
Dr Franz Schreier  
Deutsches Zentrum für Luft- und Raumfahrt  
Remote Sensing Technology Institute  
Oberpfaffenhofen  
Germany

---

SPRINGER-PRAXIS BOOKS IN ENVIRONMENTAL SCIENCES  
SUBJECT *ADVISORY EDITOR*: John Mason, M.B.E., B.Sc., M.Sc., Ph.D.

---

ISBN 978-3-642-05438-9 e-ISBN 978-3-642-05439-6  
DOI 10.1007/978-3-642-05439-6

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010920974

© Springer-Verlag Berlin Heidelberg 2010

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law. The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: Jim Wilkie  
Project copy editor: Mike Shardlow  
Author-generated LaTeX, processed by EDV-Beratung Herweg, Germany

Printed on acid-free paper

Springer is part of Springer Science + Business Media ([www.springer.com](http://www.springer.com))

To our families



# Table of Contents

<b>Preface</b>	XI
<b>1 Remote sensing of the atmosphere</b>	1
1.1 The atmosphere – facts and problems . . . . .	1
1.1.1 Greenhouse gases . . . . .	3
1.1.2 Air pollution . . . . .	4
1.1.3 Tropospheric ozone . . . . .	4
1.1.4 Stratospheric ozone . . . . .	4
1.2 Atmospheric remote sensing . . . . .	4
1.3 Radiative transfer . . . . .	8
1.3.1 Definitions . . . . .	9
1.3.2 Equation of radiative transfer . . . . .	9
1.3.3 Radiative transfer in the UV . . . . .	10
1.3.4 Radiative transfer in the IR and microwave . . . . .	14
1.3.5 Instrument aspects . . . . .	17
1.3.6 Derivatives . . . . .	17
1.4 Inverse problems . . . . .	18
<b>2 Ill-posedness of linear problems</b>	23
2.1 An illustrative example . . . . .	23
2.2 Concept of ill-posedness . . . . .	27
2.3 Analysis of linear discrete equations . . . . .	28
2.3.1 Singular value decomposition . . . . .	28
2.3.2 Solvability and ill-posedness . . . . .	29
2.3.3 Numerical example . . . . .	32
<b>3 Tikhonov regularization for linear problems</b>	39
3.1 Formulation . . . . .	39
3.2 Regularization matrices . . . . .	41
3.3 Generalized singular value decomposition and regularized solution . . . .	45
3.4 Iterated Tikhonov regularization . . . . .	49



3.5	Analysis tools . . . . .	50
3.5.1	Filter factors . . . . .	50
3.5.2	Error characterization . . . . .	51
3.5.3	Mean square error matrix . . . . .	56
3.5.4	Resolution matrix and averaging kernels . . . . .	57
3.5.5	Discrete Picard condition . . . . .	58
3.5.6	Graphical tools . . . . .	61
3.6	Regularization parameter choice methods . . . . .	66
3.6.1	A priori parameter choice methods . . . . .	67
3.6.2	A posteriori parameter choice methods . . . . .	68
3.6.3	Error-free parameter choice methods . . . . .	74
3.7	Numerical analysis of regularization parameter choice methods . . . . .	83
3.8	Multi-parameter regularization methods . . . . .	93
3.8.1	Complete multi-parameter regularization methods . . . . .	94
3.8.2	Incomplete multi-parameter regularization methods . . . . .	98
3.9	Mathematical results and further reading . . . . .	103
<b>4</b>	<b>Statistical inversion theory</b>	<b>107</b>
4.1	Bayes theorem and estimators . . . . .	107
4.2	Gaussian densities . . . . .	109
4.2.1	Estimators . . . . .	110
4.2.2	Error characterization . . . . .	112
4.2.3	Degrees of freedom . . . . .	113
4.2.4	Information content . . . . .	118
4.3	Regularization parameter choice methods . . . . .	121
4.3.1	Expected error estimation method . . . . .	121
4.3.2	Discrepancy principle . . . . .	124
4.3.3	Hierarchical models . . . . .	125
4.3.4	Maximum likelihood estimation . . . . .	126
4.3.5	Expectation minimization . . . . .	128
4.3.6	A general regularization parameter choice method . . . . .	130
4.3.7	Noise variance estimators . . . . .	135
4.4	Marginalizing method . . . . .	137
<b>5</b>	<b>Iterative regularization methods for linear problems</b>	<b>141</b>
5.1	Landweber iteration . . . . .	141
5.2	Semi-iterative regularization methods . . . . .	144
5.3	Conjugate gradient method . . . . .	146
5.4	Stopping rules and preconditioning . . . . .	154
5.4.1	Stopping rules . . . . .	155
5.4.2	Preconditioning . . . . .	156
5.5	Numerical analysis . . . . .	160
5.6	Mathematical results and further reading . . . . .	162

<b>6</b>	<b>Tikhonov regularization for nonlinear problems</b>	163
6.1	Four retrieval test problems . . . . .	164
6.1.1	Forward models and degree of nonlinearity . . . . .	164
6.1.2	Sensitivity analysis . . . . .	169
6.1.3	Prewhitening . . . . .	171
6.2	Optimization methods for the Tikhonov function . . . . .	173
6.2.1	Step-length methods . . . . .	174
6.2.2	Trust-region methods . . . . .	178
6.2.3	Termination criteria . . . . .	179
6.2.4	Software packages . . . . .	183
6.3	Practical methods for computing the new iterate . . . . .	183
6.4	Error characterization . . . . .	190
6.4.1	Gauss–Newton method . . . . .	191
6.4.2	Newton method . . . . .	196
6.5	Regularization parameter choice methods . . . . .	199
6.5.1	A priori parameter choice methods . . . . .	200
6.5.2	Selection criteria with variable regularization parameters . . . . .	203
6.5.3	Selection criteria with constant regularization parameters . . . . .	206
6.6	Iterated Tikhonov regularization . . . . .	209
6.7	Constrained Tikhonov regularization . . . . .	212
6.8	Mathematical results and further reading . . . . .	217
<b>7</b>	<b>Iterative regularization methods for nonlinear problems</b>	221
7.1	Nonlinear Landweber iteration . . . . .	222
7.2	Newton-type methods . . . . .	222
7.2.1	Iteratively regularized Gauss–Newton method . . . . .	223
7.2.2	Regularizing Levenberg–Marquardt method . . . . .	232
7.2.3	Newton–CG method . . . . .	237
7.3	Asymptotic regularization . . . . .	239
7.4	Mathematical results and further reading . . . . .	246
<b>8</b>	<b>Total least squares</b>	251
8.1	Formulation . . . . .	252
8.2	Truncated total least squares . . . . .	254
8.3	Regularized total least squares for linear problems . . . . .	258
8.4	Regularized total least squares for nonlinear problems . . . . .	267
<b>9</b>	<b>Two direct regularization methods</b>	271
9.1	Backus–Gilbert method . . . . .	271
9.2	Maximum entropy regularization . . . . .	280
<b>A</b>	<b>Analysis of continuous ill-posed problems</b>	285
A.1	Elements of functional analysis . . . . .	285
A.2	Least squares solution and generalized inverse . . . . .	288
A.3	Singular value expansion of a compact operator . . . . .	290
A.4	Solvability and ill-posedness of the linear equation . . . . .	291

<b>B</b>	<b>Standard-form transformation for rectangular regularization matrices</b>	295
B.1	Explicit transformations . . . . .	295
B.2	Implicit transformations . . . . .	299
<b>C</b>	<b>A general direct regularization method for linear problems</b>	303
C.1	Basic assumptions . . . . .	303
C.2	Source condition . . . . .	305
C.3	Error estimates . . . . .	306
C.4	A priori parameter choice method . . . . .	306
C.5	Discrepancy principle . . . . .	307
C.6	Generalized discrepancy principle . . . . .	310
C.7	Error-free parameter choice methods . . . . .	313
<b>D</b>	<b>Chi-square distribution</b>	319
<b>E</b>	<b>A general iterative regularization method for linear problems</b>	323
E.1	Linear regularization methods . . . . .	323
E.2	Conjugate gradient method . . . . .	327
E.2.1	CG-polynomials . . . . .	328
E.2.2	Discrepancy principle . . . . .	332
<b>F</b>	<b>Residual polynomials of the LSQR method</b>	343
<b>G</b>	<b>A general direct regularization method for nonlinear problems</b>	349
G.1	Error estimates . . . . .	350
G.2	A priori parameter choice method . . . . .	353
G.3	Discrepancy principle . . . . .	354
<b>H</b>	<b>A general iterative regularization method for nonlinear problems</b>	365
H.1	Newton-type methods with a priori information . . . . .	365
H.1.1	Error estimates . . . . .	368
H.1.2	A priori stopping rule . . . . .	368
H.1.3	Discrepancy principle . . . . .	370
H.2	Newton-type methods without a priori information . . . . .	373
<b>I</b>	<b>Filter factors of the truncated total least squares method</b>	385
<b>J</b>	<b>Quadratic programming</b>	391
J.1	Equality constraints . . . . .	391
J.2	Inequality constraints . . . . .	394
	<b>References</b>	407
	<b>Index</b>	423

# Preface

The retrieval problems arising in atmospheric remote sensing belong to the class of the so-called discrete ill-posed problems. These problems are unstable under data perturbations, and can be solved by numerical regularization methods, in which the solution is stabilized by taking additional information into account.

The goal of this research monograph is to present and analyze numerical algorithms for atmospheric retrieval. The book is aimed at physicists and engineers with some background in numerical linear algebra and matrix computations. Although there are many practical details in this book, for a robust and efficient implementation of all numerical algorithms, the reader should consult the literature cited.

The data model adopted in our analysis is semi-stochastic. From a practical point of view, there are no significant differences between a semi-stochastic and a deterministic framework; the differences are relevant from a theoretical point of view, e.g., in the convergence and convergence rates analysis.

After an introductory chapter providing the state of the art in passive atmospheric remote sensing, Chapter 2 introduces the concept of ill-posedness for linear discrete equations. To illustrate the difficulties associated with the solution of discrete ill-posed problems, we consider the temperature retrieval by nadir sounding and analyze the solvability of the discrete equation by using the singular value decomposition of the forward model matrix.

A detailed description of the Tikhonov regularization for linear problems is the subject of Chapter 3. We use this opportunity to introduce a set of mathematical and graphical tools to characterize the regularized solution. These comprise the filter factors, the errors in the state space and the data space, the mean square error matrix, the averaging kernels, and the L-curve. The remaining part of the chapter is devoted to the regularization parameter selection. First, we analyze the parameter choice methods in a semi-stochastic setting by considering a simple synthetic model of a discrete ill-posed problem, and then present the numerical results of an extensive comparison of these methods applied to an ozone retrieval test problem. In addition, we pay attention to multi-parameter regularization, in which the state vector consists of several components with different regularization strengths. When analyzing one- and multi-parameter regularization methods, the focus is on the pragmatic aspects of the selection rules and not on the theoretical aspects

associated with the convergence of the regularized solution as the noise level tends to zero.

At first glance, it may appear that Chapter 4, dealing with statistical inversion theory, is an alien to the main body of the textbook. However, the goal of this chapter is to reveal the similitude between Tikhonov regularization and statistical inversion regarding the regularized solution representation, the error analysis, and the design of regularization parameter choice methods. The marginalizing method, in which the auxiliary parameters of the retrieval are treated as a source of errors, can be regarded as an alternative to the multi-parameter regularization, in which the auxiliary parameters are a part of the retrieval.

Chapter 5 briefly surveys some classical iterative regularization methods such as the Landweber iteration and semi-iterative methods, and then treats the regularizing effect of the conjugate gradient method for normal equations (CGNR). The main emphasis is put on the CGNR and the LSQR implementations with reorthogonalizations. Finally, we analyze stopping rules for the iterative process, and discuss the use of regularization matrices as preconditioners.

The first five chapters set the stage for the remaining chapters dealing with nonlinear ill-posed problems. To illustrate the behavior of the numerical algorithms and tools we introduce four test problems that are used throughout the rest of the book. These deal with the retrieval of  $O_3$  and BrO in the visible spectral region, and of CO and temperature in the infrared spectral domain.

In Chapter 6 we discuss practical aspects of Tikhonov regularization for nonlinear problems. We review step-length and trust-region methods for minimizing the Tikhonov function, and present algorithms for computing the new iterate. These algorithms rely on the singular value decomposition of the standard-form transformed Jacobian matrix, the bidiagonalization of the Jacobian matrix, and on iterative methods with a special class of preconditioners constructed by means of the Lanczos algorithm. After characterizing the solution error, we analyze the numerical performance of Tikhonov regularization with a priori, a posteriori and error-free parameter choice methods.

Chapter 7 presents the relevant iterative regularization methods for nonlinear problems. We first examine an extension of the Landweber iteration to the nonlinear case, and then analyze the efficiency of Newton type methods. The following methods are discussed: the iteratively regularized Gauss–Newton method, the regularizing Levenberg–Marquardt method and the Newton–CG method. These approaches are insensitive to overestimations of the regularization parameter, and depend or do not depend on the a priori information. Finally, we investigate two asymptotic regularization methods: the Runge–Kutta regularization method and the exponential Euler regularization method.

In Chapter 8 we review the truncated and the regularized total least squares method for solving linear ill-posed problems, and put into evidence the likeness with the Tikhonov regularization. These methods are especially attractive when the Jacobian matrix is inexact. We illustrate algorithms for computing the regularized total least squares solution by solving appropriate eigenvalue problems, and present a first attempt to extend the total least squares to nonlinear problems.

Chapter 9 brings the list of nonlinear methods to a close. It describes the Backus–Gilbert method as a representative member of mollifier methods, and finally, it addresses the maximum entropy regularization.

For the sake of completeness and in order to emphasize the mathematical techniques which are used in the classical regularization theory, we present direct and iterative methods for solving linear and nonlinear ill-posed problems in a general framework. The analysis is outlined in the appendices, and is performed in a deterministic and discrete setting. Although discrete problems are not ill-posed in the strict sense, we prefer to argue in this setting because the proofs of convergence rate results are more transparent, and we believe that they are more understandable by physicists and engineers.

Several monographs decisively influenced our research. We learned the mathematical fundamentals of the regularization theory from the books by Engl et al. (2000) and Rieder (2003), the mathematical foundation of iterative regularization methods from the recent book by Kaltenbacher et al. (2008), and the state of the art in numerical regularization from the book by Hansen (1998). Last but not least, the monograph by Vogel (2002) and the book by Kaipio and Somersalo (2005) have provided us with the important topic of regularization parameter selection from a statistical perspective.

This book is the result of the cooperation of more than six years between a mathematically oriented engineer and two atmospheric physicists who are interested in computational methods. Therefore, the focus of our book is on practical aspects of regularization methods in atmospheric remote sensing. Nevertheless, for interested readers some mathematical details are provided in the appendices.

The motivation of our book is based on the need and search for reliable and efficient analysis methods to retrieve atmospheric state parameters, e.g., temperature or constituent concentration, from a variety of atmospheric sounding instruments. In particular, we were, and still are, involved in data processing for the instruments SCIAMACHY and MIPAS on ESA's environmental remote sensing satellite ENVISAT, and more recently for the spectrometer instruments GOME-2 and IASI on EUMETSAT's MetOp operational meteorological satellite. This resulted in the development of the so-called DRACULA (aDvanced Retrieval of the Atmosphere with Constrained and Unconstrained Least squares Algorithms) software package which implements the various methods as discussed in this book. A software package like DRACULA, complemented by appropriate radiative transfer forward models, could not exist without the support we have received from many sides, especially from our colleagues at DLR in Oberpfaffenhofen. To them we wish to address our sincere thanks.

Finally, we would like to point out that a technical book like the present one may still contain some errors we have missed. But we are in the fortunate situation that each author may derive comfort from the thought that any error is due to the other two. In any case, we will be grateful to anyone bringing such errors or typos to our attention.

Oberpfaffenhofen, Germany  
March, 2010

Adrian Doicu  
Thomas Trautmann  
Franz Schreier

# 1

## Remote sensing of the atmosphere

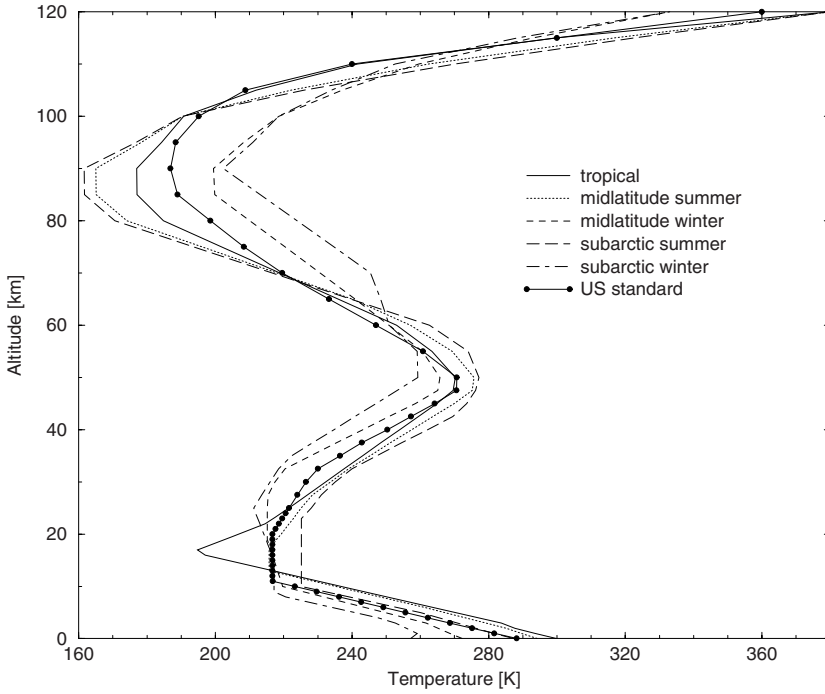
Climate change, stratospheric ozone depletion, tropospheric ozone enhancement, and air pollution have become topics of major concerns and made their way from the scientific community to the general public as well as to policy, finance, and economy (Solomon et al., 2007). In addition to these atmospheric changes related to human activities, natural events such as volcanic eruptions or biomass burning have a significant impact on the atmosphere, while the demands and expectations on weather forecasting are steadily increasing (Chahine et al., 2006). Furthermore, the discovery of extrasolar planets with the possibility of hosting life (Des Marais et al., 2002) has brought a new momentum to the subject of planetary atmospheres.

In view of all these developments, atmospheric science comprising various fields of physics, chemistry, mathematics, and engineering has gained new attraction. Modeling and observing the atmosphere are keys for the advancement of our understanding the environment, and remote sensing is one of the superior tools for observation and characterization of the atmospheric state.

In this chapter a brief introduction to atmospheric remote sensing will be given. After a short survey of the state of the atmosphere and some of its threats, the atmospheric sounding using spectroscopic techniques is discussed. A review of the radiative transfer in (Earth's) atmosphere and a general characterization of atmospheric inverse problems will conclude our presentation.

### 1.1 The atmosphere – facts and problems

The state of planetary atmospheres, i.e., its thermodynamic properties, composition, and radiation field, varies in space and time. For many purposes it is sufficient to concentrate on the vertical coordinate and to ignore its latitude, longitude, and time-dependence. Various altitude regions of the atmosphere are defined according to the temperature structure: troposphere, stratosphere, mesosphere, and thermosphere (Figure 1.1).



**Fig. 1.1.** AFGL (Air Force Geophysics Laboratory) reference-atmospheric models: temperatures (Anderson et al., 1986). The circles attached to the US standard profile indicate the altitude levels.

Pressure  $p$  decreases monotonically with increasing altitude  $z$ ; according to the ideal gas law  $p = nk_B T$  and the hydrostatic equation  $dp = -g\rho dz$  we have

$$p(z) = p_0 \exp \left( - \int_0^z \frac{dz}{\bar{H}} \right).$$

Here,  $n$  is the number density,  $g$  is the gravity acceleration constant,  $k_B$  is the Boltzmann constant,  $\rho = mn$  is mass density, and  $m$  is the mean molecular mass ( $m \approx 29 \text{ amu} = 4.82 \cdot 10^{-23} \text{ g}$  for dry air in Earth's lower and mid atmosphere). Ignoring the altitude-dependence of the factors defining the scale height

$$H(z) = \frac{k_B T(z)}{mg},$$

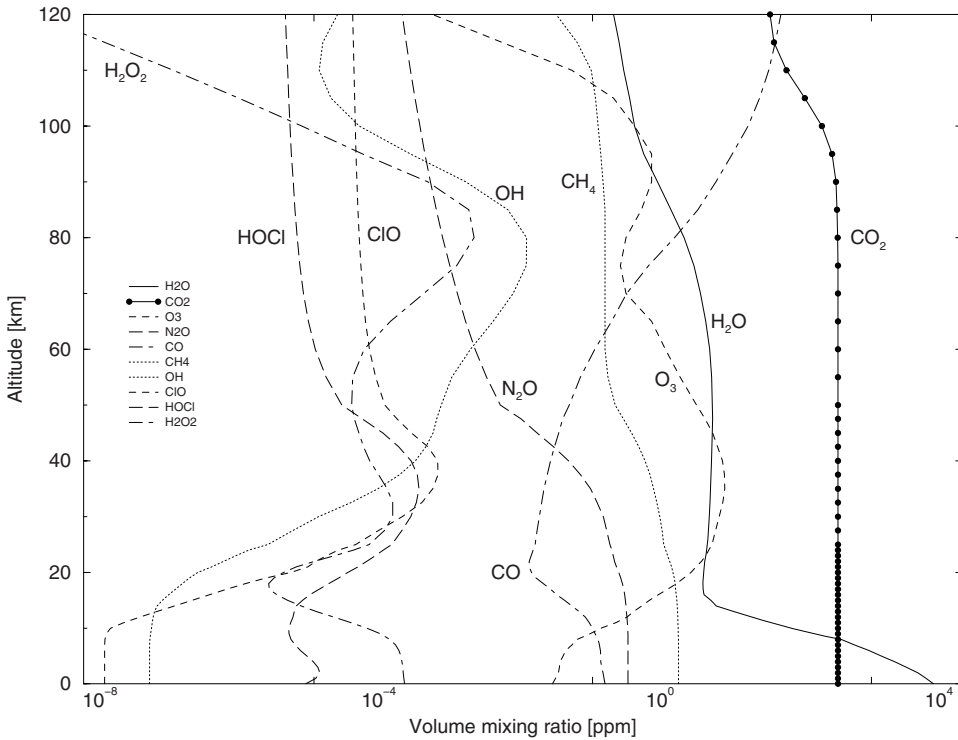
yields

$$p(z) = p_0 \exp \left( - \frac{z}{\bar{H}} \right), \quad (1.1)$$

where  $p_0$  is the surface pressure ( $p_0 = 1 \text{ bar} = 1013.25 \text{ mb}$  for standard STP). Then, assuming a mean atmospheric temperature  $T = 250 \text{ K}$ , gives the scale height  $\bar{H} = 7.3 \text{ km}$ .

The terrestrial atmosphere is composed of a large number of gases and various solid and liquid particles (hydrometeors and aerosols), see Figure 1.2. The water- and aerosol-free atmosphere is made up of nitrogen ( $\text{N}_2$ , 78%) and oxygen ( $\text{O}_2$ , 21%) with almost constant mixing ratios in the lower and middle atmosphere. Water is present in all three phases





**Fig. 1.2.** AFGL reference atmospheric models: volume mixing ratios of selected molecules (Anderson et al., 1986).

(vapor, liquid droplets, and ice crystals) and varies significantly in space and time. The remaining 1% of the atmospheric gases are noble gases (0.95%) and trace gases (0.05%). The trace gases, which are mainly carbon dioxide, methane, nitrous oxide and ozone, have a large effect on Earth's climate and the atmospheric chemistry and physics.

Precise knowledge of the distribution and temporal evolution of trace gases and aerosols is important in view of the many challenges of the atmospheric environment.

### 1.1.1 Greenhouse gases

The greenhouse gases (carbon dioxide  $\text{CO}_2$ , methane  $\text{CH}_4$ , tropospheric ozone  $\text{O}_3$ , chlorofluorocarbons and to a lesser extent water  $\text{H}_2\text{O}$ ) are responsible for Earth's natural greenhouse effect which keeps the planet warmer than it would be without an atmosphere. These gases block thermal radiation from leaving the Earth atmosphere and lead to an increase in surface temperature. In the last century, the concentration of greenhouse gases increased substantially:  $\text{CO}_2$  from its pre-industrial level of about 280 ppm by more than 30% due to combustion of fossil fuels, and  $\text{CH}_4$  by even more than 100%. As a consequence, one expects an average global warming of about  $2^\circ\text{C}$  to  $4^\circ\text{C}$  in the coming century. Hence, substantial changes of the environment can be expected with significant effects for the existing flora and fauna (Solomon et al., 2007).

### 1.1.2 Air pollution

Pollutants from natural processes and human activities like  $\text{NO}_2$  and  $\text{CO}$  are emitted into the troposphere. In the northern hemisphere, the main source of pollutants is fossil fuel combustion coupled with some biomass burning, while in the southern hemisphere, biomass burning is the primary source. Acid rain produces severe damage to forests and aquatic life, especially in regions with a lack of natural alkalinity. This forms when  $\text{SO}_2$  and  $\text{NO}_2$  build up in the atmosphere. Sulfur dioxide and nitrogen dioxide are oxidized by reaction with the hydroxyl radical and generate sulfuric acid and nitric acid, respectively. These acids with a pH normally below 5.6 are then removed from the atmosphere in rain, snow, sleet or hail. It should be pointed out that the release of  $\text{SO}_2$  into the atmosphere by coal and oil burning is at least two times higher than the sum of all natural emissions.

### 1.1.3 Tropospheric ozone

Ozone is a toxic and highly oxidizing agent. Photochemical ozone production in the troposphere, also known as summer smog, produces irritation of the respiratory system and reduces the lung function. The majority of tropospheric ozone formation occurs when nitrogen oxides, carbon monoxide and volatile organic compounds react in the atmosphere in the presence of sunlight. High concentrations of ozone arise when the temperature is high, humidity is low, and air is relatively static, and when there are high concentrations of hydrocarbons.

### 1.1.4 Stratospheric ozone

While ozone behaves like a greenhouse gas in the troposphere, in the stratosphere it helps to filter out the incoming ultraviolet radiation from the Sun, protecting life on Earth from its harmful effects. It is produced from ultraviolet rays reacting with oxygen at altitudes between 20 and 50 km, where it forms the so-called stratospheric ozone layer. In the upper stratosphere, ozone is removed by catalytic cycles involving halogen oxides. In addition, a very substantial depletion of stratospheric ozone over Antarctica and the Arctic has been observed during springtime. The main source of the halogen atoms in the stratosphere is photodissociation of chlorofluorocarbon compounds, commonly called freons, and of bromofluorocarbon compounds known as halons. These compounds are transported into the stratosphere after being emitted at the surface from industrial production. The loss of ozone in the stratosphere is also affected, in a synergistic manner, by the tropospheric emission of greenhouse gases.

## 1.2 Atmospheric remote sensing

Remote sensing means that measurements are performed at a large distance from the object or the medium to be investigated. The interaction of electromagnetic or acoustic waves with the medium is determined by the state of the medium, and the modification of the waves can be used for the retrieval of the medium's properties. The following discussion

concentrates on measurements of the electromagnetic radiation, but the mathematical tools for the solution of the inverse problem can equally well be applied to acoustic measurements, e.g., SONAR (SOund Navigation and Ranging) or SODAR (SOund Detection And Ranging).

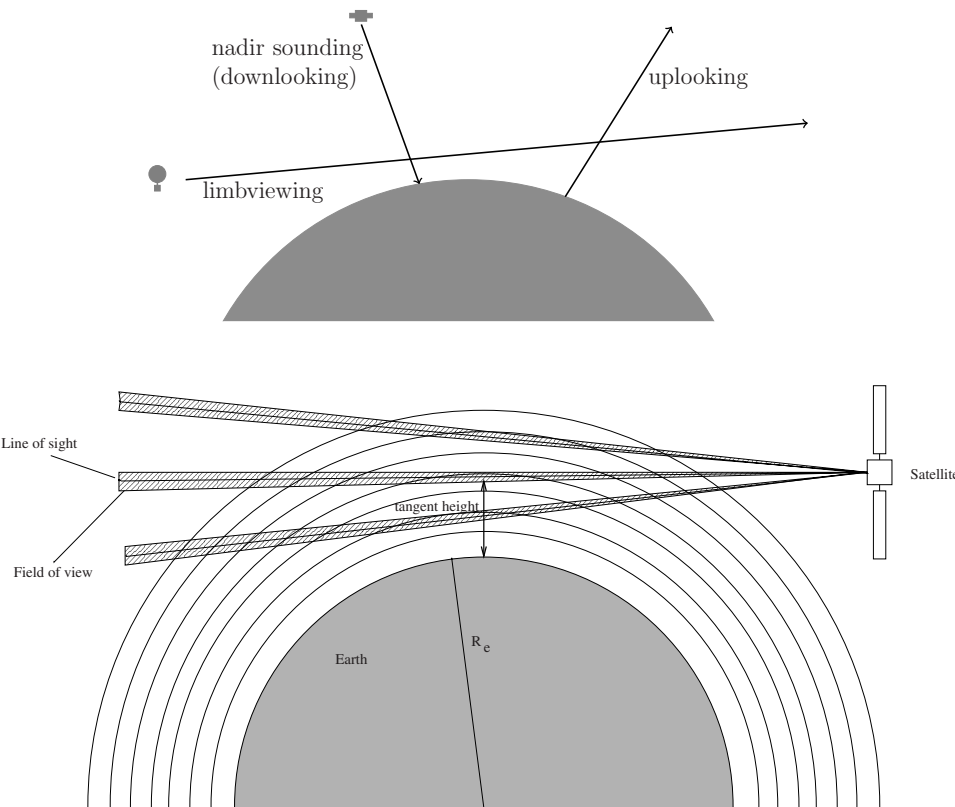
Remote sensing can be passive or active. Active remote sensing utilizes an artificial radiation source such as a laser emitting light pulses; the laser light is scattered by gas molecules and aerosols and it is partially absorbed by the target gas. A portion of the emitted light is collected by a detector telescope, and the analysis of the recorded laser light reveals information about the composition of the atmosphere. In LIDAR (LIght Detection And Ranging) systems, the transmitter and the detector are usually co-located and the technique is based on backscattering. Radar (radio detection and ranging) systems employ a similar technique using microwave-emitting antennas.

In contrast, passive remote sensing utilizes natural radiation sources. The observation of short-wave solar radiation propagating through the atmosphere, interacting with its constituents and partly being reflected by Earth's surface, and the observation of long-wave thermal emission of both atmosphere and surface are the main approaches. Passive remote sensing can be achieved by analyzing absorption or emission spectra as follows:

- (1) Thermal emission. Instruments based upon the emission technique detect the long-wave radiation (infrared or microwave) thermally emitted in the atmosphere along the observer's line-of-sight. The signals from atmospheric constituents can be regarded as thermal 'fingerprints' of the atmosphere, and from the emission line properties, temperature or trace gas concentrations are derived.
- (2) Absorption of solar radiation. The upwelling radiation at the top of the atmosphere from the ultraviolet to the near-infrared comprises the solar radiation that has been scattered by air molecules and aerosols, partially absorbed by the target gas and reflected at the Earth's surface. Information on trace gas concentrations is encapsulated in that part of the incoming solar radiation that has been removed by absorption.
- (3) Absorption of direct radiation. This category includes occultation instruments that measure solar, lunar, and even stellar radiation directly through the limb of the atmosphere during Sun, Moon and star rise and set events. By measuring the amount of absorption of radiation through the atmosphere, occultation instruments can infer the vertical profiles of trace gas constituents.

A further classification of remote sensing systems is based on the sensor location and the observation geometry (Figure 1.3):

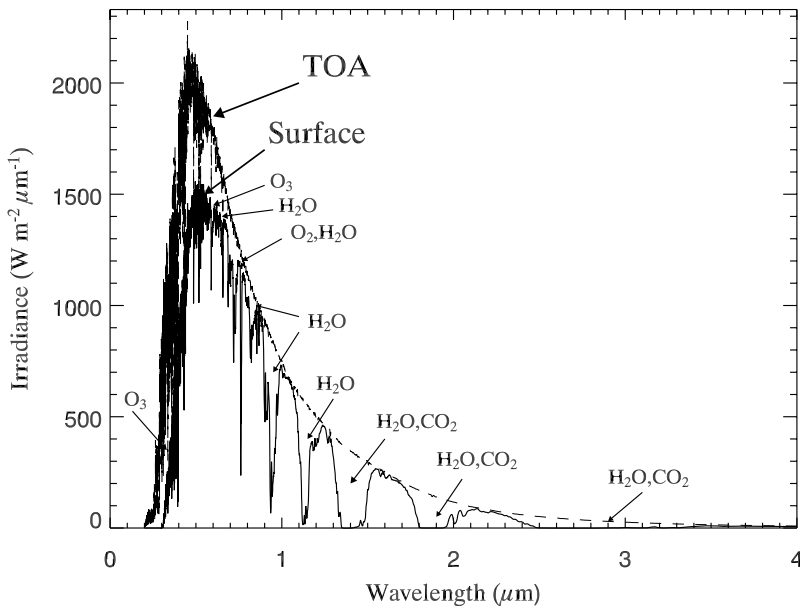
- (1) Ground-based systems deployed in laboratory buildings usually observe the atmosphere in an 'uplooking' geometry. Observatories in mountain regions are frequently used with altitudes up to several kilometers, for example, in the Network for Detection of Atmospheric Composition Change (NDACC).
- (2) Airborne remote sensing systems work with instruments onboard of aircraft or balloons. Whereas conventional aircraft operate in altitudes more or less confined to the troposphere, some aircraft such as the American ER-2 or the Russian Geophysica can reach altitudes of about 20 km, well in the lower stratosphere. Stratospheric balloons can reach altitudes of almost 40 km, hence permitting observation of the atmosphere in 'limb sounding' geometry.



**Fig. 1.3.** Observation geometries for atmospheric remote sensing.

- (3) Spaceborne systems aboard satellites, the Space Shuttle, or the International Space Station (ISS) work in limb viewing or in nadir viewing (downlooking) mode. A large number of sensors for environmental and meteorological studies is mounted on polar orbiting satellites flying at altitudes of about 800 km. Furthermore geostationary satellites with an altitude of about 36 000 km are utilized, especially for meteorological purposes. In contrast, Space Shuttles and the ISS are orbiting at altitudes of about 400 km or less.

Figure 1.4 illustrates the incoming extraterrestrial solar radiation at the top of the atmosphere (TOA) versus wavelength. It is noted that for solar wavelengths beyond  $1.4 \mu\text{m}$  the solar emission curve closely resembles a blackbody radiator having a temperature of about 6000 K. The lower curve depicts a MODTRAN4 (MODerate resolution atmospheric TRANsmission) calculation (Berk et al., 1989) for the downwelling solar flux density reaching the ground. The solar zenith angle has been set to  $60^\circ$ , while for the composition and state of the atmosphere a midlatitude summer case has been adopted. All relevant absorbing atmospheric trace gases, as shown in the figure, were included in the radiative transfer computation which had a moderate spectral resolution of about  $20 \text{ cm}^{-1}$ . Similarly, in Figure 1.5 we show the infrared spectrum of the Earth atmosphere. The results

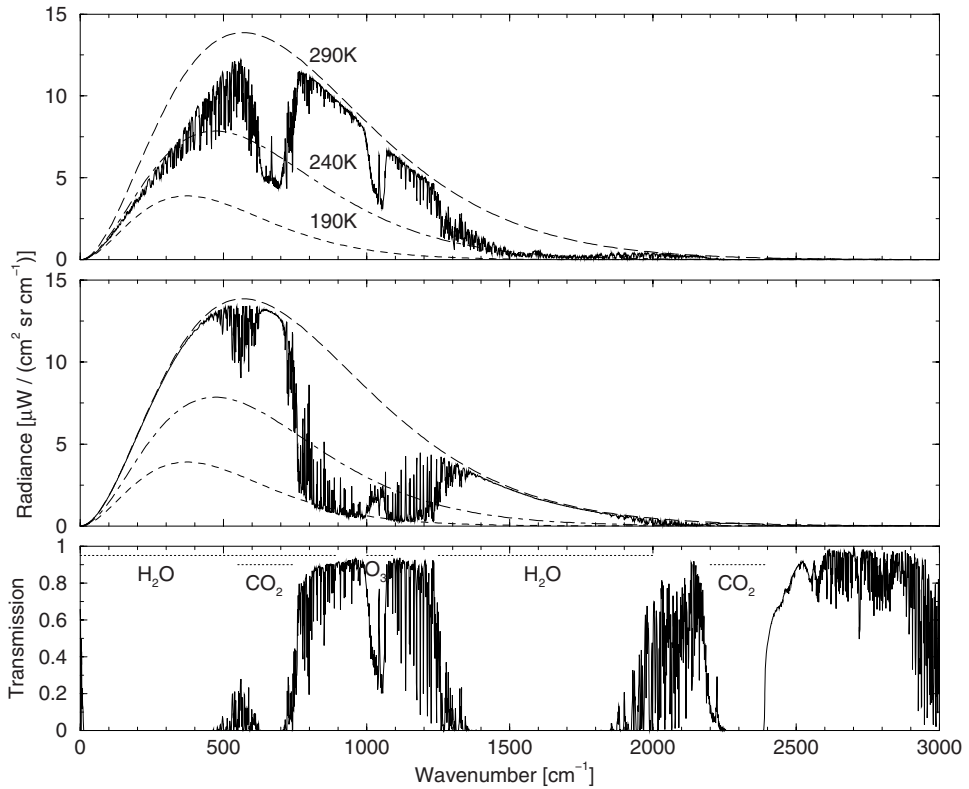


**Fig. 1.4.** Spectral distribution of the incoming solar flux density at the top of the atmosphere (TOA) and at ground level for a clear sky atmosphere and a nonreflecting ground. The solar zenith angle has been set to  $60^\circ$ . (Adapted from Zdunkowski et al. (2007).)

correspond to a clear sky US standard atmosphere and are also computed with the radiative transfer band model MODTRAN4. Figures 1.4 and 1.5 clearly demonstrate that UV and IR spectra of the terrestrial atmosphere contain a wealth of information about its state, and, in particular, signatures of a large number of molecular absorbers can be identified. Two examples will serve to illustrate the basic principles of atmospheric remote sensing.

In the UV wavelength range  $290\text{--}330\text{ }\mu\text{m}$ , not only do spaceborne nadir observations of the radiance enable determination of the total column amount of ozone below the sub-satellite point but scanning from smaller to larger wavelengths also allows us to ‘sound’ the atmosphere as a function of increasing distance from the sensor. Ozone molecules absorb solar radiation strongly at short wavelengths, i.e., photons entering the atmosphere are not able to penetrate the ozone layer in the stratosphere (with maximum concentration around 20 or 25 km). On the other hand, photons with higher wavelengths have a better chance to reach a greater depth (lower altitude) before they are absorbed.

Weather forecasting heavily relies on sounding of the atmospheric temperature profile using satellite observations in the infrared or microwave region following the pioneering work of King and Kaplan. King (1956) showed that the vertical temperature profile can be estimated from satellite radiance scan measurements. Kaplan (1959) demonstrated that intensity measurements in the wing of a  $\text{CO}_2$  spectral band probe the deeper regions of the atmosphere, whereas observations closer to the band center see the upper part of the atmosphere. Analogously, the complex of  $\text{O}_2$  lines in the microwave spectral range can be used. In both cases one utilizes emission from a relatively abundant gas with known and uniform distribution.



**Fig. 1.5.** Infrared spectrum of the Earth atmosphere: upwelling radiation seen by an observer above the atmosphere (top), downwelling radiation seen by an observer at sealevel (middle) and atmospheric transmission for a vertical path (bottom). The blackbody radiation according to Planck's function for three representative values and the main absorption bands are indicated too.

In summary, the spectral absorption or emission characteristics combined with monotonically increasing path length allows a mapping between altitude and wavelength, thus providing a direct link between absorber amount or temperature and observed radiation.

### 1.3 Radiative transfer

In atmospheric remote sensing, the radiation seen by an observer is described by the theory of radiative transfer with an appropriate instrument model. Before discussing radiative transfer models for the UV/vis and IR/mw spectral ranges, we define some quantities of central importance. For a thorough discussion of the material presented in this section we recommend classical textbooks on atmospheric radiation as for example, Goody and Yung (1989), Thomas and Stamnes (1999), Liou (2002), and Zdunkowski et al. (2007).

### 1.3.1 Definitions

Different variables are used to characterize the ‘color’ of electromagnetic waves: wavelength  $\lambda$  with units  $\mu\text{m}$ , nm, or  $\text{\AA}$  are common in the ultraviolet and visible range, wavenumbers  $\nu = 1/\lambda$  in units of  $\text{cm}^{-1}$  are used in the infrared, and frequencies  $\tilde{\nu} = c\nu$  (with  $c$  being the speed of light) are employed in the microwave regime. Numerically one has  $\nu [\text{cm}^{-1}] = 10\,000/\lambda [\mu\text{m}] \approx 30\tilde{\nu} [\text{GHz}]$ .

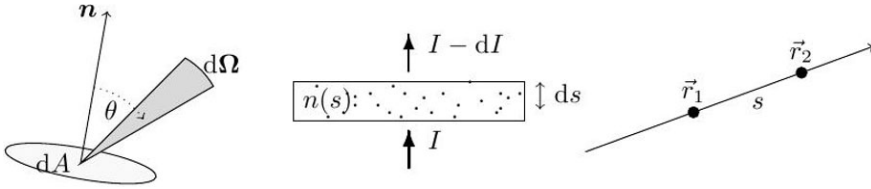
Monochromatic radiance or intensity is defined as the differential amount of energy  $dE_\lambda$  in a given wavelength interval  $(\lambda, \lambda + d\lambda)$  crossing an area  $dA$  into a solid angle  $d\Omega$ , oriented with an angle  $\theta$  relative to the normal  $\mathbf{n}$  of the area, within a time interval  $dt$  (Figure 1.6),

$$I_\lambda = \frac{dE_\lambda}{\cos \theta d\Omega dt dA d\lambda}. \quad (1.2)$$

The definition of the radiance  $I_\nu$  is done in a similar manner.

For a beam of radiation traveling in a certain direction, with distances measured by the path variable  $s = |\mathbf{r}_1 - \mathbf{r}_2|$ , the ratio of the radiances at two different locations defines the transmission

$$\mathcal{T}(\mathbf{r}_1, \mathbf{r}_2) = \frac{I(\mathbf{r}_1)}{I(\mathbf{r}_2)}. \quad (1.3)$$



**Fig. 1.6.** Concepts of radiative transfer. Left: illustration of radiance definition (1.2). Middle: schematics of radiation attenuation  $dI$  traversing a path element  $ds$  with absorber density  $n$ . Right: path  $s = |\mathbf{r}_1 - \mathbf{r}_2|$  relevant for the definition of optical depth and transmission.

### 1.3.2 Equation of radiative transfer

A beam of radiation traversing the atmosphere will be attenuated by interactions with the atmospheric constituents, and the extinction (absorption and scattering) is proportional to the amount of incoming radiation, the path distance  $ds$  in the direction  $\Omega$ , and the density  $n$  of the medium, i.e.,  $dI \propto -In ds$  (Figure 1.6). On the other hand, the thermal emission of the medium and the scattering processes will result in an increase of the radiation energy described by a ‘source function’  $J(\mathbf{r}, \Omega)$ . The total change of radiation is given by the equation of radiative transfer

$$\frac{1}{n(\mathbf{r})C_{\text{ext}}(\mathbf{r})} \frac{dI}{ds}(\mathbf{r}, \Omega) = -I(\mathbf{r}, \Omega) + J(\mathbf{r}, \Omega). \quad (1.4)$$

The quantity  $C_{\text{ext}}$  is called the extinction cross-section, and its product with the number density is the extinction coefficient  $\sigma_{\text{ext}} = nC_{\text{ext}}$ .

In the absence of any sources, the differential equation can be readily solved and we have (Beer–Lambert–Bouguer law)

$$\mathcal{T}(\mathbf{r}_1, \mathbf{r}_2) = \frac{I(\mathbf{r}_1)}{I(\mathbf{r}_2)} = \exp \left( - \int_{|\mathbf{r}_1 - \mathbf{r}_2|} C_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) ds \right), \quad (1.5)$$

where the integral in the exponent is the so-called (extinction) optical depth between the points  $\mathbf{r}_1$  and  $\mathbf{r}_2$ ,

$$\tau_{\text{ext}}(\mathbf{r}_1, \mathbf{r}_2) = \int_{|\mathbf{r}_1 - \mathbf{r}_2|} C_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) ds = \int_{|\mathbf{r}_1 - \mathbf{r}_2|} \sigma_{\text{ext}}(\mathbf{r}) ds.$$

Equation (1.4) is a linear first-order differential equation that can be formally integrated giving

$$I(\mathbf{r}_o, \boldsymbol{\Omega}) = I(\mathbf{r}_s, \boldsymbol{\Omega}) \exp(-\tau_{\text{ext}}(\mathbf{r}_o, \mathbf{r}_s)) + \int_{|\mathbf{r}_o - \mathbf{r}_s|} J(\mathbf{r}, \boldsymbol{\Omega}) \exp(-\tau_{\text{ext}}(\mathbf{r}_o, \mathbf{r})) ds. \quad (1.6)$$

The integral form of the radiative transfer equation (1.6) describes the radiation seen by an observer at  $\mathbf{r}_o$ ; the first term is the source radiation at  $\mathbf{r}_s$  (e.g., Earth's surface in case of a downlooking observer) attenuated according to Beer's law (1.5) and the second term represents the radiation due to emission and scattering at intermediate points along the line of sight.

The atmospheric energy budget is essentially determined by solar insolation (roughly in the UV–vis–IR spectral range 0.2–0.35  $\mu\text{m}$ ) and emission by the Earth and its atmosphere (in the infrared spectral range 3.5–100  $\mu\text{m}$ ). For most practical purposes, these two spectral regions may be treated separately: in the solar spectral range it is justified to neglect the thermal emission of the Earth–atmosphere system, whereas in the infrared the scattering processes are usually important only in the so-called atmospheric window region 8–12.5  $\mu\text{m}$  (Figure 1.5). However, as the clear atmosphere is almost transparent to the infrared radiation in this region, the atmospheric window is of minor importance for remote sensing of trace gases (except for ozone).

### 1.3.3 Radiative transfer in the UV

The radiation field can be split into two components: the direct radiation, which is never scattered in the atmosphere and reflected by the ground surface, and the diffuse radiation, which is scattered or reflected at least once. Neglecting the thermal emission, the source function  $J$  can be decomposed as

$$J(\mathbf{r}, \boldsymbol{\Omega}) = J_{\text{ss}}(\mathbf{r}, \boldsymbol{\Omega}) + J_{\text{ms}}(\mathbf{r}, \boldsymbol{\Omega}), \quad (1.7)$$

where the single and the multiple scattering source functions are given by

$$J_{\text{ss}}(\mathbf{r}, \boldsymbol{\Omega}) = F \frac{\omega(\mathbf{r})}{4\pi} P(\mathbf{r}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_{\text{sun}}) e^{-\tau_{\text{ext}}(\mathbf{r}, \mathbf{r}_{\text{max}})},$$



and

$$J_{\text{ms}}(\mathbf{r}, \Omega) = \frac{\omega(\mathbf{r})}{4\pi} \int_{4\pi} P(\mathbf{r}, \Omega, \Omega') I(\mathbf{r}, \Omega') d\Omega',$$

respectively. In the above relations,  $\omega = \sigma_{\text{scat}}/\sigma_{\text{ext}}$  is the single scattering albedo,  $\sigma_{\text{scat}}$  is scattering coefficient,  $F$  is the incident solar flux,  $P$  is the phase function,  $\Omega_{\text{sun}}$  is the unit vector in the sun direction, and  $\mathbf{r}_{\text{max}}$  is the point at the top of the atmosphere corresponding to  $\mathbf{r}$ , that is,  $\mathbf{r}_{\text{max}} = \mathbf{r} - |\mathbf{r}_{\text{max}} - \mathbf{r}| \Omega_{\text{sun}}$ . It should be pointed out that technically, there is no absolute dividing line between the Earth's atmosphere and space, but for studying the balance of incoming and outgoing energy on the Earth, an altitude at about 100 kilometers above the Earth is usually used as the 'top of the atmosphere'.

An accurate interpretation of the measurements performed by satellite instruments in arbitrary viewing geometries requires the solution of the radiative transfer equation in a three-dimensional inhomogeneous spherical atmosphere. For this type of radiative transfer problems, the Monte Carlo technique (Marchuk et al., 1980) is a possible candidate. In a Monte Carlo simulation the radiance at the top of the atmosphere is determined statistically by simulating a large number of individual photon trajectories through the atmosphere. This method is computationally very expensive in the calculation of the backscattered radiance, because many photons are lost when they leave the atmosphere at other positions and in other directions than the one to the satellite. For atmospheric applications, the so-called backward Monte Carlo method is more efficient. Here, the photons are started from the detector and their path is followed backward to the point where they leave the atmosphere in solar direction. The disadvantages of this method are, however, its poor accuracy for optically thick or weakly absorbing media, and that for each viewing geometry, a new backward calculation has to be performed. Additionally, the required linearization of such Monte Carlo models is a challenging task. Applications of the Monte Carlo method for radiance calculations in a spherical atmosphere can be found in Oikarinen et al. (1999).

### ***Radiative transfer models***

In practice, simplified radiative transfer models are used to simulate the radiances at the observer's position and in the direction of the instrument line-of-sight. These can be categorized depending on the assumptions made for the geometry of the model atmosphere.

*Plane-parallel radiative transfer* calculations have been applied successfully for nadir measurements with solar zenith angles up to  $75^\circ$ . The discrete ordinate method (Stamnes et al., 1988), the doubling-adding method (Hansen, 1971), the finite difference method (Barkstrom, 1975) and the Gauss–Seidel iteration method (Herman and Browning, 1965) have been used to solve the radiative transfer equation in a plane-parallel atmosphere. Further details on the mentioned solution methods can be found in Lenoble (1985).

For nadir viewing geometries with large solar zenith angles and for limb viewing geometries, the so-called *pseudo-spherical approximation* has been developed (Dahlback and Stamnes, 1991). In this approximation, the single scattering radiance is computed in a spherical atmosphere, whereas the multiple scattering radiance is still calculated in a plane-parallel geometry. For limb measurements, the effect of a varying solar zenith angle along the line of sight is accounted for by performing a set of independent pseudo-spherical calculations for different values of the solar zenith angle. This model is equivalent to the *independent pixel approximation* for three-dimensional radiative transfer in clouds, and

can be regarded as a first-order spherical correction to the plane-parallel formulation of the radiative transfer. Solution methods for radiative transfer in a pseudo-spherical atmosphere include the discrete ordinate method (Spurr, 2001, 2002), the finite difference method (Rozanov et al., 2000), and the discrete ordinate method with matrix exponential (Doicu and Trautmann, 2009a).

For a subhorizon Sun as well as for lines of sight with large tangent heights, the independent pixel approximation leads to errors of about 4%. For such problems, the *spherical shell approximation* (Rozanov et al., 2001; Walter et al., 2005; Doicu and Trautmann, 2009e) delivers more accurate results. Here, the atmosphere is approximated by homogeneous spherical shells and no horizontal inhomogeneities in the optical parameters are considered. The radiative transfer equation is solved by means of a Picard iteration with a long or a short characteristic method (Kuo et al., 1996).

Accurate simulations of radiances in ultraviolet and visible spectral regions should take into account that light scattered by the atmosphere is polarized and that approximately 4% of molecular scattering is due to the inelastic rotational Raman component.

### Polarization

The radiation and state of polarization of light can be described by the Stokes vector  $\mathbf{I} = [I, Q, U, V]^T$ , where  $I$  is the radiance,  $Q$  is a measure for the polarization along the  $x$ - and  $y$ -axis of the chosen reference frame,  $U$  is a measure of the polarization along the  $+45^\circ$  and  $-45^\circ$  directions, and  $V$  describes the circular polarization. The vector radiative transfer equation reads as

$$\frac{d\mathbf{I}}{ds}(\mathbf{r}, \boldsymbol{\Omega}) = -\sigma_{\text{ext}}(\mathbf{r}) \mathbf{I}(\mathbf{r}, \boldsymbol{\Omega}) + \sigma_{\text{ext}}(\mathbf{r}) \mathbf{J}(\mathbf{r}, \boldsymbol{\Omega}),$$

where  $\mathbf{J}$  is the source term. As in the scalar case, the source function can be split into a single and a multiple scattering component, and we have the representations

$$\mathbf{J}_{\text{ss}}(\mathbf{r}, \boldsymbol{\Omega}) = F \frac{\omega(\mathbf{r})}{4\pi} e^{-\tau_{\text{ext}}(\mathbf{r}, \mathbf{r}_{\text{max}})} \mathbf{Z}(\mathbf{r}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_{\text{sun}}) \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

and

$$\mathbf{J}_{\text{ms}}(\mathbf{r}, \boldsymbol{\Omega}) = \frac{\omega(\mathbf{r})}{4\pi} \int_{4\pi} \mathbf{Z}(\mathbf{r}, \boldsymbol{\Omega}, \boldsymbol{\Omega}') \mathbf{I}(\mathbf{r}, \boldsymbol{\Omega}') d\Omega',$$

with  $\mathbf{Z}$  being the phase matrix.

The instrumental signal should be simulated with a vector radiative transfer model for two reasons.

First, light reflected from Earth's atmosphere is polarized because of (multiple) scattering of unpolarized light by air molecules and aerosols. Simulations of radiance measurements by a scalar approximation for atmospheric radiative transfer leads to errors of about 10% depending mainly on the viewing scenario (Mishchenko et al., 1994). The scalar radiative transfer errors are small in the spectral regions in which mainly single scattering takes place and significant in the spectral regions in which the amount of multiple scattering

increases because of decreasing gas absorption. For a pseudo-spherical atmosphere, vector radiative transfer models employing the discrete ordinate method (Spurr, 2006, 2008), the successive order of scattering technique (McLinden et al., 2002a) and the discrete ordinate method with matrix exponential (Doicu and Trautmann, 2009b) have been developed. A survey of vector radiative transfer models for a plane-parallel atmosphere can be found in Hansen and Travis (1974).

Second, the different optical devices in the instrument are sensitive to the state of polarization of the incident light. As a result, the radiance that is measured by the detectors, referred to as the polarization-sensitive measurement, is different to the radiance that enters in the instrument. In the calibration process, the instrumental signal is corrected for the polarization sensitivity, whereas the polarization correction factor is determined from broadband on-ground measurements. However, in spectral regions where the state of polarization is varying rapidly with wavelength, the polarization correction is not sufficiently accurate and severely influences the retrieval. To eliminate this drawback, the polarization-sensitive measurement together with the transport of radiation in the atmosphere have been simulated by means of vector radiative transfer models (Hasekamp et al., 2002; McLinden et al., 2002b).

### ***Ring effect***

The filling-in of solar Fraunhofer lines in sky spectra and the telluric filling-in of trace gas absorption features in ultraviolet and visible backscatter spectra are known as the Ring effect. Several studies (Kattawar et al., 1981; Joiner et al., 1995) have demonstrated that the main process responsible for the Ring effect is the rotational Raman scattering by molecular  $O_3$  and  $N_2$ . In backscatter spectra, the Ring effect shows up as small-amplitude distortion, which follows Fraunhofer and absorption lines. For an inelastically scattering atmosphere, the radiative transfer equation includes an additional source term, the Raman source function, and the single and multiple scattering source terms have to be modified accordingly. Several radiative transfer models have been used to simulate the so-called Ring spectrum defined as the ratio of the inelastic and the elastic scattering radiances. These models include a Monte Carlo approach (Kattawar et al., 1981), a successive order of scattering method (Joiner et al., 1995) and a model which treats rotational Raman scattering as a first-order perturbation (Vountas et al., 1998; Landgraf et al., 2004; Spurr et al., 2008).

As Ring structures appear in the polarization signal, a complete simulation of the polarization-sensitive measurement requires a vector radiative transfer model which simulates Ring structures for all relevant Stokes parameters (Aben et al., 2001; Stam et al., 2002; Landgraf et al., 2004). The calculation of Ring spectra with a vector radiative transfer model is numerically expensive and approximation methods are desirable for large data sets. The numerical analysis performed in Landgraf et al. (2004) reveals that

- (1) the polarization Ring spectra of  $Q$  and  $U$  are much weaker than those of the radiance  $I$  due to the low polarization of Raman scattered light;
- (2) the combination of both a vector radiative transfer model, simulating the Stokes vector for an elastic scattering atmosphere, and a scalar radiative transfer approach, simulating the Ring spectrum for the radiance is sufficiently accurate for gas profile retrievals but not for applications involving the retrieval of cloud properties.

### 1.3.4 Radiative transfer in the IR and microwave

Neglecting scattering and assuming local thermodynamical equilibrium, the source function  $J$  is given by the Planck function at temperature  $T$ ,

$$B(\nu, T) = \frac{2hc^2\nu^3}{\exp\left(\frac{h\nu}{k_B T}\right) - 1}. \quad (1.8)$$

The formal solution (1.6), describing the radiance  $I$  at wavenumber  $\nu$  received by an instrument at position  $\mathbf{r}_o$ , is given by the Schwarzschild equation

$$I(\nu, \mathbf{r}_o) = I(\nu, \mathbf{r}_s) \mathcal{T}(\nu, \mathbf{r}_o, \mathbf{r}_s) + \int_{|\mathbf{r}_o - \mathbf{r}|} B(\nu, T(\mathbf{r})) \frac{\partial \mathcal{T}}{\partial s}(\nu, \mathbf{r}_o, \mathbf{r}) ds, \quad (1.9)$$

where  $I(\nu, \mathbf{r}_s)$  is the background contribution at position  $\mathbf{r}_s$ . The monochromatic transmission is computed according to Beer's law as

$$\mathcal{T}(\nu, \mathbf{r}_o, \mathbf{r}) = \exp\left(-\int_{|\mathbf{r}_o - \mathbf{r}|} \sigma_{\text{abs}}(\nu, \mathbf{r}') ds'\right) \quad (1.10)$$

$$= \exp\left(-\int_{|\mathbf{r}_o - \mathbf{r}|} ds' \sum_m C_{\text{abs}m}(\nu, p(\mathbf{r}'), T(\mathbf{r}')) n_m(\mathbf{r}')\right). \quad (1.11)$$

Here,  $\sigma_{\text{abs}}$  is the absorption coefficient,  $p$  is the atmospheric pressure,  $n_m$  is the number density of molecule  $m$ , and  $C_{\text{abs}m}$  is its absorption cross-section.

In general, the molecular absorption cross-section is obtained by summing over the contributions from many lines. For an individual line at position  $\hat{\nu}$ , the cross-section is the product of the temperature-dependent line strength  $S(T)$  and a normalized line shape function  $g(\nu)$  describing the broadening mechanism(s), that is,

$$C_{\text{abs}m}(\nu, p, T) = \sum_l S_{ml}(T) g(\nu, \hat{\nu}_{ml}, \gamma_{ml}(p, T)). \quad (1.12)$$

In the atmosphere, the combined effect of pressure broadening, corresponding to a Lorentzian line shape (indices  $m$  and  $l$  denoting molecule and line will be omitted for simplicity)

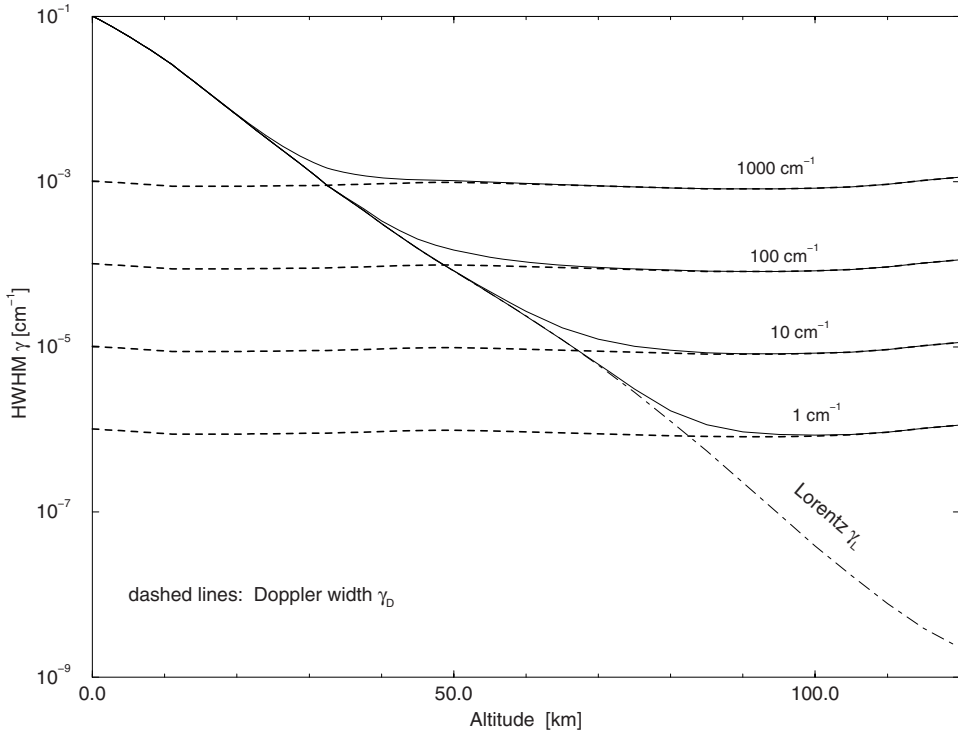
$$g_L(\nu, \hat{\nu}, \gamma_L) = \frac{1}{\pi} \frac{\gamma_L}{(\nu - \hat{\nu})^2 + \gamma_L^2}, \quad (1.13)$$

and Doppler broadening, corresponding to a Gaussian line shape

$$g_D(\nu, \hat{\nu}, \gamma_D) = \frac{1}{\gamma_D} \left(\frac{\log 2}{\pi}\right)^{\frac{1}{2}} \exp\left(-\log 2 \left(\frac{\nu - \hat{\nu}}{\gamma_D}\right)^2\right), \quad (1.14)$$

can be represented by a convolution, i.e., the Voigt line profile  $g_V = g_L \otimes g_D$ . Pressure broadening (air-broadening, with self-broadening neglected) and Doppler broadening half-widths are given by

$$\gamma_L(p, T) = \gamma_{L0} \frac{p}{p_{\text{ref}}} \left(\frac{T_{\text{ref}}}{T}\right)^\alpha$$



**Fig. 1.7.** Lorentz, Gauss and Voigt half-widths (HWHM) as a function of altitude in the Earth atmosphere for a variety of line positions  $\hat{\nu}$ . Pressure and temperature are from US Standard Atmosphere and the molecular mass is 36 amu.

and

$$\gamma_D(T) = \hat{\nu} \sqrt{2 \log 2 \frac{k_B T}{m c^2}},$$

respectively. Here,  $p_{\text{ref}}$  and  $T_{\text{ref}}$  are the reference pressure and temperature of line parameters, respectively,  $m$  denotes the molecular mass, and  $\alpha$  describes the temperature dependence of pressure broadening. Note that pressure broadening dominates in the lower atmosphere; the transition altitude, where Doppler broadening becomes important, moves up from the middle stratosphere to the mesosphere with increasing wavelength (Figure 1.7).

Spectroscopic line parameters required for the calculation of the molecular absorption cross-sections, e.g., the line position  $\hat{\nu}$ , the line strength  $S$ , the temperature exponent  $\alpha$ , the air-broadening half-width  $\gamma_{L0}$ , and the lower state energy  $E$  (required to calculate  $S(T)$  from the database entry  $S(T_{\text{ref}})$ ) have been compiled in various databases such as HITRAN (HIGH-resolution TRANsmision molecular absorption database), GEISA (Gestion et Etude des Informations Spectroscopiques Atmosphériques) and JPL (Jet Propulsion Laboratory) catalog. The latest versions of HITRAN (Rothman et al., 2009) and GEISA (Jacquinot-Husson et al., 2008) list parameters of some million transitions for several dozen molecules from the microwave ( $\hat{\nu} = 10^{-6} \text{ cm}^{-1}$ ) to the ultraviolet ( $\hat{\nu} \approx 25\,232$

and  $\hat{\nu} \approx 35\,877\text{ cm}^{-1}$ , respectively), whereas the JPL catalogue (Pickett et al., 1998) covers millions of rotational transitions in the microwave regime.

At a first glance the forward model appears to be much easier to solve in the infrared than in the ultraviolet as the source function is known. However, for high resolution atmospheric spectroscopy, the line-by-line (lbl) computation of (1.9) and (1.10) remains a challenging task because thousands of spectral lines have to be included in the sum (1.12). Moreover, as the monochromatic wavenumber grid point spacing determined by the half-widths of the spectral lines (cf. Figure 1.7) is very fine, accurate modeling of the spectrum may require thousands or even millions of spectral grid points. Finally, the convolution integral defining the Voigt line profile cannot be solved analytically, and numerical approximations have to be used.

In view of the computational challenges of lbl-modeling, alternative approaches have been used for low to moderate resolution spectra. Band models have been developed since the early days of radiative transfer modeling in meteorology and astrophysics (Goody and Yung, 1989; Liou, 2002; Thomas and Stamnes, 1999; Zdunkowski et al., 2007). More recently, the  $k$ -distribution and correlated  $k$  methods (Fu and Liou, 1992; Lacis and Oinas, 1991) or exponential sum fitting (Wiscombe and Evans, 1977) have been utilized.

Scattering is usually ignored in lbl models. However, if the analysis of data provided by spaceborne infrared sounders would be confined to clear sky observations only, a large fraction of data would be ignored. For nadir sounding, single scattering can be implemented with moderate effort, but multiple scattering, especially for limb sounding geometries, is still a challenging task. Various attempts have been described by Emde et al. (2004), Höpfner et al. (2002), Höpfner and Emde (2005), and Mendrok et al. (2007).

Intercomparisons of high-quality (laboratory and atmospheric) infrared spectra have revealed discrepancies with accurate model spectra obtained with the lbl approach (1.12). These deviations are commonly attributed to the so-called ‘continuum’, and a variety of explanations have been given in the literature, e.g., deviations of the far wing line profile from the Lorentzian line shape, contributions from water dimers ( $\text{H}_2\text{O}$ )<sub>2</sub> etc. For modeling infrared and microwave spectra, the semi-empirical approach developed by Clough et al. (1989) is widely used (see also Clough et al., 2005), whereas the empirical corrections due to Liebe et al. (1993) are frequently employed in the microwave regime.

When local thermodynamic equilibrium (LTE) is assumed, a local temperature can be assigned everywhere in the atmosphere, and thermal emission can be described by Planck’s law of blackbody radiation (1.8). However, because temperature and radiation vary in space and time, the atmosphere is not in thermodynamic equilibrium. Nevertheless, the LTE assumption is justified in the troposphere and stratosphere, where the density of air is sufficiently high so that the mean time between molecular collisions is much smaller than the mean lifetime of an excited state of a radiating molecule. Thus, equilibrium conditions exist between vibrational, rotational and translation energy of the molecule. The breakdown of LTE in the upper atmosphere implies that the source function is no longer given by the Planck function. An adequate description of collisional and radiative processes under non-LTE conditions requires quantum theoretical considerations; see Lopez-Puertas and Taylor (2001) for an in-depth treatment.

### 1.3.5 Instrument aspects

In general, the finite resolution of the spectrometer results in a modification or smearing of the ‘ideal’ spectrum. This effect can be modeled by a convolution of the ‘monochromatic’ spectrum  $S_{\text{mc}}(\nu)$  (radiance  $I$  or transmission  $\mathcal{T}$ ) with an instrument line shape function  $ILS$  (also termed spectral response function  $SRF$ ),

$$S_{\text{obs}}(\nu) = \int_{-\infty}^{\infty} ILS(\nu - \nu') S_{\text{mc}}(\nu') d\nu'. \quad (1.15)$$

The function  $ILS$  clearly depends on the type of the instrument; a Gaussian can be used as a first approximation in many cases, e.g., for a grating instrument. For a Fourier transform spectrometer such as MIPAS (Michelson Interferometer for Passive Atmospheric Sounding) or IASI (Infrared Atmospheric Sounding Interferometer), the finite optical path difference  $L$  of the Michelson interferometer corresponds to a multiplication of the interferogram with a box function, so that (to a first approximation) the line shape function is given by

$$ILS(\nu - \nu') = 2L \text{sinc}(2\pi L(\nu - \nu')) = \frac{\sin(2\pi L(\nu - \nu'))}{\pi(\nu - \nu')}. \quad (1.16)$$

On the other hand, the finite aperture of an instrument results in a superposition of ideal ‘pencil beam’ spectra corresponding to an infinitesimal field of view. Modeling of this finite field of view is especially important for limb geometry and can be done by convolving the pencil-beam spectra with a field-of-view function. Frequently, this function is approximated by box, triangular, or Gauss functions.

### 1.3.6 Derivatives

Often, the radiative transfer models are optimized to deliver in addition to the simulated radiance, the partial derivatives of the radiance with respect to the atmospheric parameters being retrieved. The process of obtaining the set of partial derivatives, which constitute the Jacobian matrix, is commonly referred to as linearization analysis. Several techniques for performing a linearization analysis can be distinguished

In many cases, the Jacobian matrix is computed by *finite differences*, and this calculation is the most time-consuming part of the retrieval. Even more serious is the fact that the amount of perturbation is difficult to predict and an improper choice leads to truncation and/or cancellation errors; see Gill et al. (1981) for a pertinent discussion.

*Analytical calculation* of derivatives is advantageous, both for computational efficiency and accuracy. From (1.6) it is apparent that the partial derivatives of the radiance measured by the instrument are given by the partial derivatives of the single and the multiple scattering radiances. As the multiple scattering radiance depends on the solution of the radiative transfer equation, derivatives calculation can be performed by linearizing the radiative transfer equation with respect to the desired parameters. A linearized radiative transfer model based on an analytical determination of the partial derivatives of the conventional discrete ordinate solution for radiance has been developed by Spurr (2001, 2002, 2008), while a linearized forward approach based on the discrete ordinate method with



matrix exponential has been proposed in Doicu and Trautmann (2009d). For infrared applications, analytic derivatives are implemented in the codes ARTS (Bühler et al., 2005), KOPRA (Stiller et al., 2002) and MOLIERE (Urban et al., 2004). However, calculating the derivatives manually and implementing these in a moderately large code as required for general-purpose radiative transfer is tedious and error-prone. Moreover, no automatic updates of the derivatives calculation in the case of upgrades of the forward model are available.

The measured radiance can be expressed in the framework of a *forward-adjoint approach* as the scalar product of the solution of the adjoint problem and the source function of the forward problem. Employing the linearization technique to the forward and the adjoint problems, analytical expressions for the derivatives in a plane-parallel atmosphere have been derived in Marchuk (1964, 1995), Box (2002), Ustinov (2001, 2005), Rozanov and Rozanov (2007), and Landgraf et al. (2001). For a pseudo-spherical atmosphere, this approach has been applied to nadir viewing geometries in Walter et al. (2004) and to limb geometries in Ustinov (2008), and Doicu and Trautmann (2009c). The forward-adjoint approach is extremely efficient because only two radiative transfer calculations are required for derivative calculations. In this context, Landgraf et al. (2001) reported that under certain conditions a forward-adjoint approach based on the Gauss–Seidel iteration technique is approximately a factor of 20–30 faster than a linearized forward approach based on the conventional discrete ordinate solution.

*Automatic or algorithmic differentiation* provides a pleasant alternative to quickly generate derivative-enhanced versions of computer codes. Automatic differentiation techniques (Griewank and Corliss, 1991; Griewank, 2000) are based on the observation that every model implemented as a computer program is essentially formulated in terms of elementary mathematical operations (sums, products, powers) and elementary functions. In contrast to integration, differentiation is based on a few simple recipes such as the chain rule, and these can be performed automatically by some kind of precompiler, taking a computer code of the forward model as input and delivering a code that additionally produces derivatives with respect to some chosen variables. A number of automatic differentiation tools are available for Fortran and C (cf. the compilation given at <http://www.autodiff.org/>). This approach has been used by Schreier and Schimpf (2001), and Schreier and Boettger (2003) to implement Jacobian matrices in an infrared line-by-line radiative transfer code.

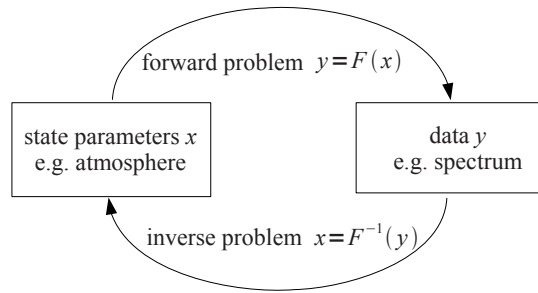
## 1.4 Inverse problems

In atmospheric remote sensing, the relationship between the state parameters  $x$  and collected observations making up some set of data  $y$  is described by a forward model  $F$ . This encapsulates a radiative transfer model and an instrument model, and formally, we write

$$y = F(x).$$

The task of computing the data  $y$  given the state parameters  $x$  is called the forward problem, while the mathematical process to compute  $x$  given  $y$  is called the inverse problem (Figure 1.8). Atmospheric remote sensing deals with the inverse problem. In fact, inverse problems are ubiquitous challenges in almost any field of science and engineering, from astrophysics, helioseismology, geophysics, quantum mechanical scattering the-





**Fig. 1.8.** Forward and inverse problem.

ory and material science to medicine with its large diversity of imaging and tomographic techniques, see, for example, Craig and Brown (1986), Groetsch (1993), Wing (1991) for some introductory surveys, and Baumeister (1987), Engl et al. (2000), Hansen (1998), Kaipio and Somersalo (2005), Kaltenbacher et al. (2008), Kirsch (1996), Menke (1984), Parker (1994), Rieder (2003), Tarantola (2005), Vogel (2002) for more advanced treatments. Inverse problems for atmospheric remote sensing are discussed by Twomey (1977), Stephens (1994) and Rodgers (2000).

The forward model is closely connected with the actual observation being performed and mirrors the physics of the measurement process. In contrast the approach to solving the inverse problem is (to some extent) independent of the physical process and the methods developed throughout this book can be used for inverse problems in other fields as well.

In a general framework, the data  $y$  may be a function of frequency (or wavelength) or it may be a collection of discrete observations. In the first case, the problem is called a continuous problem, while in the second case it is called a semi-discrete problem. When both  $x$  and  $y$  are discrete, the corresponding problem is a discrete problem. In order to avoid possible confusions, vectors will be denoted by bold letters, e.g.,  $\mathbf{x}$  is a vector of state parameters or simply a state vector, while  $x$  is a state parameter function. As any measurement system can deliver only a discrete, finite set of data, the problems arising in atmospheric remote sensing are semi-discrete. Moreover, due to the complexity of the radiative transfer, the forward model has to be computed by a numerical algorithm, which, in turn, requires a discretization of the state parameter function. For these reasons, the atmospheric inverse problems we are dealing with are discrete.

An important issue is that actual observations contain instrumental or measurement noise. We can thus envision data  $\mathbf{y}^\delta$  as generally consisting of noiseless observations  $\mathbf{y}$  from a ‘perfect’ instrument plus a noise component  $\delta$ , i.e.,

$$\mathbf{y}^\delta = \mathbf{y} + \delta.$$

For limb viewing geometries,  $\mathbf{y}^\delta$  is usually a concatenation of spectra corresponding to all limb scans, and the reconstruction of atmospheric profiles from a limb-scanning sequence of spectra is known as the global-fit approach (Carlotti, 1988).

The radiation seen by an observer depends on a large variety of parameters, i.e., spectral range, observation geometry, instrument settings, optical properties of the atmospheric constituents, and the state of the atmosphere characterized by pressure, temperature, and

concentration of molecules and particles. For a complete and accurate modeling of the measurement process, the forward model has to take into account all relevant parameters. However, only a single or a few variables of the atmospheric system can usually be retrieved from the observed data, and all other parameters are assumed to be known. For this reason and following Rodgers (2000), we split the state vector  $\mathbf{x}$  into two components: the first component  $\mathbf{x}_1$  represents the atmospheric profile (temperature or concentration profile of one particular species of interest) to be retrieved, while the second component  $\mathbf{x}_2$  includes all auxiliary parameters or model parameters, which influence the retrieval. It is a common practice to retrieve atmospheric profiles individually in sequence, where the sequence of the target species and temperature retrieval is determined according to the degree of their reciprocal interference. The auxiliary parameters may include

- surface parameters (surface albedo, ground emissivity factor),
- spectral corrections due to the instrumental convolution process (tilt, undersampling, polarization spectra),
- instrumental parameters (pointing offset, wavelength shift and squeeze, *ILS* parameters, baseline shift),
- atmospheric continuum (including all absorption that varies smoothly with the frequency and being represented by a polynomial),
- parameters describing complex physical processes (Ring spectrum, non-LTE/LTE population ratio, temperature and volume mixing ratio gradients).

In general, the auxiliary parameters can be retrieved together with the main atmospheric profile, they can be treated as an observation uncertainty, or they can be assumed to be perfectly known. In the first case, we are talking about a multi-parameter problem, while in the second case, we employ the so-called marginalizing method to solve the inverse problem. Another option is to perform the retrieval in two stages (Rozanov et al., 2005). In the first stage, also known as the pre-processing stage, the scaling factors of the spectral corrections are computed together with the shift and squeeze corrections by considering a linearization of the forward model about a reference state. In the second stage, referenced to as the inversion stage, the scaling factors determined in the pre-processing step are used to compute the corrected measured spectra, and the nonlinear problem is solved for the trace gas profile.

In fact, the true physics of the measurement is described by the so-called forward function  $\mathbf{f}(\mathbf{x})$  (Rodgers, 2000). The forward function is difficult to compute because the real physics is far too complex to deal with explicitly. For example, the correct modeling of aerosol and cloud particles with respect to their shape, size distribution and loading is an impossible task. The forward model errors  $\delta_m$ , defined through the relation

$$\mathbf{f}(\mathbf{x}) = \mathbf{F}(\mathbf{x}) + \delta_m,$$

are difficult to compute and only the norm  $\|\delta_m\|$  can be estimated by an additional computational step.

To get a first idea about the difficulties associated with the solution of inverse problems, we consider an elementary example. Let  $x(z)$  be some function defined on the interval  $[0, z_{\max}]$ , and let us compute the integral  $y(z) = \int_0^z x(t) dt$ . Evidently,  $y$  is an antiderivative of  $x$ , and so, the original function can be rediscovered by taking the derivative of  $y$ , that is,  $x(z) = y'(z)$ . Formally, the integration step is the forward problem,

while the differentiation step is the inverse problem. Now, let  $y^\delta$  be a perturbation of  $y$ . Then, the derivative calculation  $x^\delta(z) = y^{\delta'}(z)$  is an unstable process because, as the opposite of the smoothing effect of integration, differentiation is very sensitive to small perturbations of the function. As a result,  $x^\delta$  may deviate significantly from  $x$ , even though  $y^\delta$  is close to  $y$ .

The goal of this book is to present numerical (regularization) methods for inverse problems involving the reconstruction of atmospheric profiles from (satellite) measurements. The solution of atmospheric inverse problems is not an easy task due to the so-called ill-posedness of the equation describing the measurement process. This concept will be clarified in the next chapter.

# 2

## Ill-posedness of linear problems

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

### 2.1 An illustrative example

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

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

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

$$B(\nu, T(z)) = 2ck_{\text{B}}\nu^2 T(z).$$

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

$$I(\nu) = \int_0^{z_{\max}} k(\nu, z) T(z) dz, \quad (2.1)$$

with

$$k(\nu, z) = 2ck_B\nu^2 \frac{\partial T}{\partial z}(\nu, z) \quad (2.2)$$

being the kernel function.

Equation (2.1), which we rewrite in the generic form ( $y = I$  and  $x = T$ )

$$y(\nu) = \int_0^{z_{\max}} k(\nu, z) x(z) dz, \quad \nu \in [\nu_{\min}, \nu_{\max}], \quad (2.3)$$

is a Fredholm integral equation of the first kind and represents the mathematical model of a continuous problem. To formulate our problem in a general setting, we consider the space of real-valued, square integrable functions on the interval  $[a, b]$ , denoted by  $L^2([a, b])$ . Actually,  $L^2([a, b])$  is a Hilbert space under the inner product  $\langle u, v \rangle = \int_a^b u(t)v(t) dt$  and the induced norm  $\|u\| = \sqrt{\int_a^b u(t)^2 dt}$ . Assuming that  $y$  belongs to the Hilbert space  $Y = L^2([\nu_{\min}, \nu_{\max}])$  and  $x$  belongs to the Hilbert space  $X = L^2([0, z_{\max}])$ , we introduce the linear operator  $K : X \rightarrow Y$  by the relation

$$Kx = \int_0^{z_{\max}} k(\cdot, z) x(z) dz.$$

The integral equation (2.3) can then be written as

$$Kx = y, \quad (2.4)$$

and since (cf. (2.2))  $k \in L^2([\nu_{\min}, \nu_{\max}] \times [0, z_{\max}])$ , the linear operator  $K$  is bounded and compact (the image of any bounded sequence of functions in  $L^2$  has at least one converging subsequence).

A spectral instrument cannot measure a continuous signal and the data is a collection of discrete observations. More specifically, the radiances

$$y(\nu_i) = I(\nu_i),$$

with  $\{\nu_i\}_{i=1, \dots, m}$  being an equidistant set of points in the spectral interval  $[\nu_{\min}, \nu_{\max}]$ , represent the data, and the equation to be solved takes the form

$$y(\nu_i) = \int_0^{z_{\max}} k(\nu_i, z) x(z) dz, \quad i = 1, \dots, m. \quad (2.5)$$

The semi-discrete equation (2.5) is a mathematical model for discrete observations of a physical process and can be expressed in compact form as

$$K_m x = \mathbf{y}_m. \quad (2.6)$$

The data  $\mathbf{y}_m$  is a vector with entries  $[\mathbf{y}_m]_i = y(\nu_i)$ ,  $i = 1, \dots, m$ , and  $K_m$  is a linear operator acting between the Hilbert space  $X$  and the finite-dimensional Euclidean space  $\mathbb{R}^m$ ,

$$[K_m x]_i = (Kx)(\nu_i) = \int_0^{z_{\max}} k(\nu_i, z) x(z) dz.$$

The discretization approach which transforms the continuous equation (2.3) into the semi-discrete equation (2.5) is known as the collocation method. It should be pointed out that the collocation method can be regarded as a projection method with delta functions as basis functions.

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

$$K_m x_n = \mathbf{y}_m. \quad (2.7)$$

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

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

we obtain the system of equations

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

In matrix form, (2.8) can be written as

$$\mathbf{K}_{mn} \mathbf{x}_n = \mathbf{y}_m, \quad (2.9)$$

where  $\mathbf{x}_n = [\xi_1, \dots, \xi_n]^T$  is the coordinate vector and the matrix  $\mathbf{K}_{mn}$ , with entries

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

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

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

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

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

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

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

for  $j = 1, \dots, n$ . Thus, the entries of the coordinate vector are the mean values of the atmospheric profile over each altitude interval (layer).

- (2) For the discretization grid  $\{z_j\}_{j=\overline{0,n+1}}$  with  $z_0 = z_1 = 0$  and  $z_n = z_{n+1} = z_{\max}$ , the piecewise linear functions (or hat functions),

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

can also be chosen as basis functions. Since

$$H_{nj}(z_i) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad i, j = 1, \dots, n,$$

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

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

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

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

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

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

where

$$t_1 \leq t_2 \leq \dots \leq t_{n+k}$$

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

$$t_1 = t_2 = \dots = t_k = 0$$

and

$$t_{n+1} = t_{n+2} = \dots = t_{n+k} = z_{\max}.$$

The second-order B-spline  $B_{n2j}$  coincides with the hat functions, and for this reason,  $B_{n2j}$  is also called a linear B-spline.

## 2.2 Concept of ill-posedness

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

$$Kx = y, \quad (2.10)$$

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

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

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

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

the range space of  $K$  and by

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

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

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

Non-existence can occur in practice because the forward model is approximate or because the data contains noise. Non-uniqueness is introduced by the need for discretization and is a peculiarity of the so-called rank deficient problems, characterized by a matrix  $\mathbf{K}_{mn}$  with a non-trivial null space. In particular, state vectors  $\mathbf{x}_0$  that lie in the null space of  $\mathbf{K}_{mn}$  solve the equation  $\mathbf{K}_{mn}\mathbf{x}_0 = \mathbf{0}$ , and by superposition, any linear combination of these null-space solutions can be added to a particular solution and does not change the fit to the data. Violation of the third Hadamard condition creates serious numerical problems because small errors in the data space can be dramatically amplified in the state space.

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



### 2.3 Analysis of linear discrete equations

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

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

$$\mathbf{K}\mathbf{x} = \mathbf{y}, \quad (2.11)$$

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

#### 2.3.1 Singular value decomposition

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

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

If  $\mathbf{K}$  is of rank  $r$ , and  $\{\sigma_i\}_{i=\overline{1,n}}$  denotes the set of singular values appearing in decreasing order,

$$\sigma_1 > \sigma_2 > \dots > \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0,$$

then there exist the orthonormal sets  $\{\mathbf{v}_i\}_{i=\overline{1,n}} \in \mathbb{R}^n$  and  $\{\mathbf{u}_i\}_{i=\overline{1,m}} \in \mathbb{R}^m$  such that

$$\mathbf{K}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad \mathbf{K}^T \mathbf{u}_i = \sigma_i \mathbf{v}_i, \quad i = 1, \dots, r, \quad (2.12)$$

and

$$\mathbf{K}\mathbf{v}_i = \mathbf{0}, \quad i = r+1, \dots, n, \quad \mathbf{K}^T \mathbf{u}_i = \mathbf{0}, \quad i = r+1, \dots, m. \quad (2.13)$$

Each system  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  with these properties is called a singular system of  $\mathbf{K}$ . The sets  $\{\mathbf{v}_i\}_{i=\overline{1,r}}$  and  $\{\mathbf{v}_i\}_{i=\overline{r+1,n}}$  are orthonormal bases of  $\mathcal{N}(\mathbf{K})^\perp$  and  $\mathcal{N}(\mathbf{K})$ , respectively,

i.e.,

$$\mathcal{N}(\mathbf{K})^\perp = \text{span}\{\mathbf{v}_i\}_{i=\overline{1,r}}, \quad \mathcal{N}(\mathbf{K}) = \text{span}\{\mathbf{v}_i\}_{i=\overline{r+1,n}}, \quad (2.14)$$

while  $\{\mathbf{u}_i\}_{i=\overline{1,r}}$  and  $\{\mathbf{u}_i\}_{i=\overline{r+1,m}}$  are orthonormal bases of  $\mathcal{R}(\mathbf{K})$  and  $\mathcal{R}(\mathbf{K})^\perp$ , respectively, i.e.,

$$\mathcal{R}(\mathbf{K}) = \text{span}\{\mathbf{u}_i\}_{i=\overline{1,r}}, \quad \mathcal{R}(\mathbf{K})^\perp = \text{span}\{\mathbf{u}_i\}_{i=\overline{r+1,m}}. \quad (2.15)$$

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

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

$$\mathbf{K} = \mathbf{U}\Sigma\mathbf{V}^T, \quad (2.16)$$

where  $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_m]$  and  $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n]$  are orthogonal (or orthonormal)  $m \times m$  and  $n \times n$  matrices, respectively, and  $\Sigma$  is an  $m \times n$  matrix of the form

$$\Sigma = \begin{bmatrix} \text{diag}(\sigma_i)_{r \times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

with  $\text{diag}(\sigma_i)_{r \times r}$  being an  $r \times r$  diagonal matrix. The representation (2.16) is called the singular value decomposition (SVD) of the matrix  $\mathbf{K}$ .

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

### 2.3.2 Solvability and ill-posedness

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

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

$$\mathbf{x}^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i. \quad (2.17)$$

To prove this result we first assume that  $\mathbf{x}^\dagger$  is a solution of (2.11), i.e.,  $\mathbf{K}\mathbf{x}^\dagger = \mathbf{y}$ . If  $\mathbf{y}_0 \in \mathcal{N}(\mathbf{K}^T)$ , we see that

$$\mathbf{y}^T \mathbf{y}_0 = \mathbf{x}^{\dagger T} \mathbf{K}^T \mathbf{y}_0 = 0,$$

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

$$\mathbf{y} = \sum_{i=1}^n (\mathbf{u}_i^T \mathbf{y}) \mathbf{u}_i.$$

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

$$\mathbf{K}\mathbf{x}^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{K}\mathbf{v}_i = \sum_{i=1}^n (\mathbf{u}_i^T \mathbf{y}) \mathbf{u}_i,$$

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

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

$$\mathbf{K}^\dagger \mathbf{y} = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i, \quad \mathbf{y} \in \mathbb{R}^m, \quad (2.18)$$

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

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

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

$$\mathbf{x}^\dagger = \mathbf{K}^\dagger \mathbf{y},$$

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

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

$$\mathbf{y}^\delta = \mathbf{y} + \boldsymbol{\delta},$$

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

$$\mathbf{y}^\delta = \sum_{i=1}^m (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i, \quad (2.19)$$

the least squares solution of the linear equation (2.11) is defined by

$$\mathbf{x}^\delta = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i. \quad (2.20)$$

The least squares solution can be characterized as follows:

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

$$\mathbf{K}\mathbf{x}^\delta = P_{\mathcal{R}(\mathbf{K})}\mathbf{y}^\delta = \sum_{i=1}^n (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i;$$

(2)  $\mathbf{x}^\delta$  has the optimality property

$$\mathbf{x}^\delta = \arg \min_{\mathbf{x}} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|;$$

(3)  $\mathbf{x}^\delta$  solves the normal equation

$$\mathbf{K}^T \mathbf{K} \mathbf{x} = \mathbf{K}^T \mathbf{y}^\delta.$$

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

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

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

$$\mathbf{y}^\delta - \mathbf{K}\mathbf{x}^\delta = \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i.$$

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

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

$$\mathbf{x}^\delta = \mathbf{K}^\dagger \mathbf{y}^\delta,$$

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

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

with

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

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

$$\mathbf{x}^\delta = \sum_{i=1}^r \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i,$$

is an element of  $\mathcal{N}(\mathbf{K})^\perp = \text{span}\{\mathbf{v}_i\}_{i=\overline{1,r}}$  and represents the unique least squares solution of equation (2.11) in  $\mathcal{N}(\mathbf{K})^\perp$ . If  $\mathbf{x}_0$  is an arbitrary vector in  $\mathcal{N}(\mathbf{K})$ , then

$$\mathbf{K}(\mathbf{x}^\delta + \mathbf{x}_0) = P_{\mathcal{R}(\mathbf{K})}\mathbf{y}^\delta,$$

and  $\mathbf{x}^\delta + \mathbf{x}_0$  is a least squares solution of equation (2.11) in  $\mathbb{R}^n$ . Using the orthogonality relation  $\mathbf{x}_0^T \mathbf{x}^\delta = 0$ , we observe from

$$\|\mathbf{x}^\delta + \mathbf{x}_0\|^2 = \|\mathbf{x}^\delta\|^2 + 2\mathbf{x}_0^T \mathbf{x}^\delta + \|\mathbf{x}_0\|^2 = \|\mathbf{x}^\delta\|^2 + \|\mathbf{x}_0\|^2,$$

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

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

- (1) as the dimension of  $\mathbf{K}$  increases, the number of small singular values also increases;
- (2) as the singular values  $\sigma_i$  decrease, the corresponding singular vectors  $\mathbf{u}_i$  and  $\mathbf{v}_i$  have more sign changes in their components.

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

$$\mathbf{y}^\delta = \mathbf{y} + \Delta \mathbf{u}_{i^*},$$

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

$$\mathbf{x}^\delta = \mathbf{x}^\dagger + \frac{\Delta}{\sigma_{i^*}} \mathbf{v}_{i^*}$$

and the ratio

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

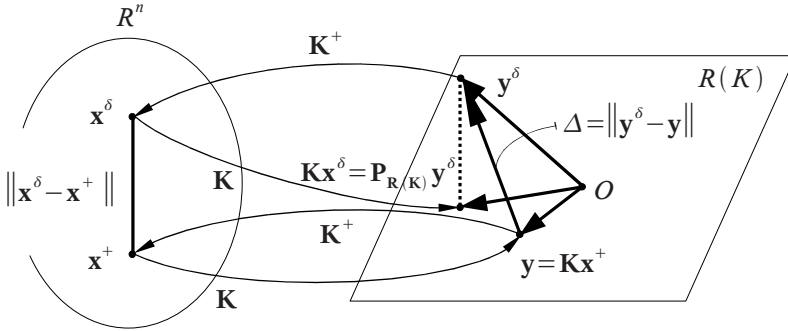
is very large if  $\sigma_{i^*}$  is very small (Figure 2.1). In this context, any naive approach which tries to compute  $\mathbf{x}^\delta$  by using (2.20) will usually return a useless result with extremely large norm.

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

### 2.3.3 Numerical example

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

As water absorption is dominant for frequencies below 40 GHz and above 120 GHz, we assume a single oxygen line at position  $\hat{\nu}_{0_2}$  and ignore other absorbers completely.



**Fig. 2.1.** The generalized inverse  $\mathbf{K}^\dagger$  maps the exact data vector  $\mathbf{y}$  into the exact solution  $\mathbf{x}^\dagger$  and the noisy data vector  $\mathbf{y}^\delta$  into the least squares solution  $\mathbf{x}^\delta$ . The image of  $\mathbf{x}^\delta$  under  $\mathbf{K}$  is the projection of  $\mathbf{y}^\delta$  onto  $\mathcal{R}(\mathbf{K})$ . Although the error in the data space  $\Delta = \|\mathbf{y}^\delta - \mathbf{y}\|$  is small, the error in the state space  $\|\mathbf{x}^\delta - \mathbf{x}^\dagger\|$  can be very large.

Neglecting the temperature-dependence of line strength and pressure broadening, and assuming an observer at infinity gives the absorption optical depth (omitting the gas index)

$$\tau_{\text{abs}}(\nu, z) = \frac{1}{\pi} \int_z^\infty \frac{S\gamma_{\text{L0}}n(z)}{(\nu - \hat{\nu})^2 + \left[\gamma_{\text{L0}} \frac{p(z)}{p_{\text{ref}}}\right]^2} \frac{p(z)}{p_{\text{ref}}} dz. \quad (2.22)$$

The volume mixing ratio  $q$  of  $\text{O}_2$  is constant with altitude, i.e.,  $q = 0.21$ , and the number density depends on altitude through pressure and temperature,

$$n(z) = q \frac{p(z)}{k_{\text{B}}T(z)}. \quad (2.23)$$

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

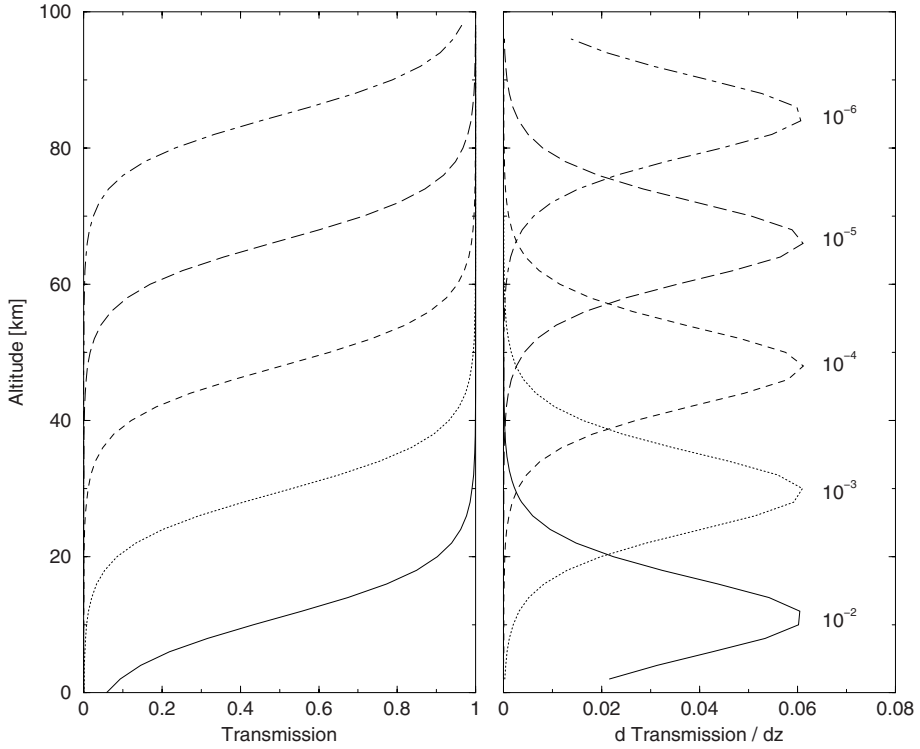
$$\tau_{\text{abs}}(\nu, z) = a \log \frac{(\nu - \hat{\nu})^2 + \gamma_{\text{L0}}^2 \exp\left(-\frac{2z}{\bar{H}}\right)}{(\nu - \hat{\nu})^2},$$

and the transmission is then given by

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

with

$$a = \frac{qSp_0\bar{H}}{2\pi k_{\text{B}}T\gamma_{\text{L0}}} = \frac{qSp_0}{2\pi\gamma_{\text{L0}}mg}.$$



**Fig. 2.2.** Transmission  $\mathcal{T}$  (left) and weighting function  $\partial\mathcal{T}/\partial z$  (right) for a temperature nadir sounding model with exponent  $a = 1.0$ , line position  $\hat{\nu} = 2.0 \text{ cm}^{-1}$  and  $\delta\nu = \nu - \hat{\nu} = 10^{-6}, \dots, 10^{-2} \text{ cm}^{-1}$ .

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

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

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

and making the change of variable

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

the integral equation (2.1) becomes

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

with

$$K(\nu, \zeta) = \frac{\nu^2 (\nu - \hat{\nu})^2 \exp(-\zeta)}{\left[ (\nu - \hat{\nu})^2 + \gamma_{L0}^2 \exp(-\zeta) \right]^2}.$$

Assuming a discretization scheme with piecewise constant functions

$$T_n(\zeta) = \sum_{j=1}^n T(\bar{\zeta}_j) P_{nj}(\zeta)$$

we obtain the discrete equation

$$[\mathbf{I}_m]_i = \sum_{j=1}^n [\mathbf{K}_{mn}]_{ij} [\mathbf{T}_n]_j, \quad i = 1, \dots, m, \quad (2.24)$$

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

$$[\mathbf{K}_{mn}]_{ij} = \int_0^{\zeta_{\max}} K(\bar{\nu}_i, \zeta) P_{nj}(\zeta) d\zeta = \sum_{k=1}^N K(\bar{\nu}_i, \bar{\zeta}_k) P_{nj}(\bar{\zeta}_k) \Delta\zeta_N, \quad (2.25)$$

the data vector

$$[\mathbf{I}_m]_i = I(\bar{\nu}_i),$$

and the state vector

$$[\mathbf{T}_n]_j = T(\bar{\zeta}_j).$$

The layer centerpoints in the data and state spaces are

$$\bar{\nu}_i = \left(i - \frac{1}{2}\right) \Delta\nu_m, \quad i = 1, \dots, m,$$

and

$$\bar{\zeta}_j = \left(j - \frac{1}{2}\right) \Delta\zeta_n, \quad j = 1, \dots, n,$$

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

$$\bar{\zeta}_k = \left(k - \frac{1}{2}\right) \Delta\zeta_N, \quad k = 1, \dots, N,$$

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

$$T^\dagger(\zeta) = \begin{cases} 220 + 28\left(\frac{5}{2} - \zeta\right), & 0 \leq \zeta \leq 2.5, \\ 220, & 2.5 < \zeta \leq 5, \\ 220 + \frac{25}{3}(\zeta - 5), & 5 < \zeta \leq 11, \\ 270, & 11 < \zeta \leq 14, \\ 250 + 20(15 - \zeta), & 14 < \zeta \leq 15. \end{cases} \quad (2.26)$$



Our analysis is organized as follows:

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

$$[\mathbf{I}_m]_i = \int_0^{\zeta_{\max}} K(\bar{\nu}_i, \zeta) T^\dagger(\zeta) d\zeta = \sum_{k=1}^N K(\bar{\nu}_i, \bar{\zeta}_k) T^\dagger(\bar{\zeta}_k) \Delta\zeta_N;$$

- (2) we generate the noisy data vector  $\mathbf{I}_m^\delta$  as

$$\mathbf{I}_m^\delta = \mathbf{I}_m + \delta\mathbf{I},$$

where  $\delta\mathbf{I}$  is a random Gaussian vector with zero mean and covariance  $\mathbf{C}_\delta = \sigma^2\mathbf{I}_m$ ; the noise standard deviation is defined in terms of the signal-to-noise ratio SNR by

$$\sigma = \frac{\|\mathbf{I}_m\|}{\sqrt{m\text{SNR}}};$$

- (3) for different values of the discretization index  $n$ , we compute the least squares solution

$$\mathbf{T}_n^\delta = \mathbf{K}_{mn}^\dagger \mathbf{I}_m^\delta,$$

and determine the solution error

$$\varepsilon_n^2 = \|\mathbf{T}^\dagger - \mathbf{T}_n^\delta\|^2 = \int_0^{\zeta_{\max}} [\mathbf{T}^\dagger(\zeta) - \mathbf{T}_n^\delta(\zeta)]^2 d\zeta,$$

where

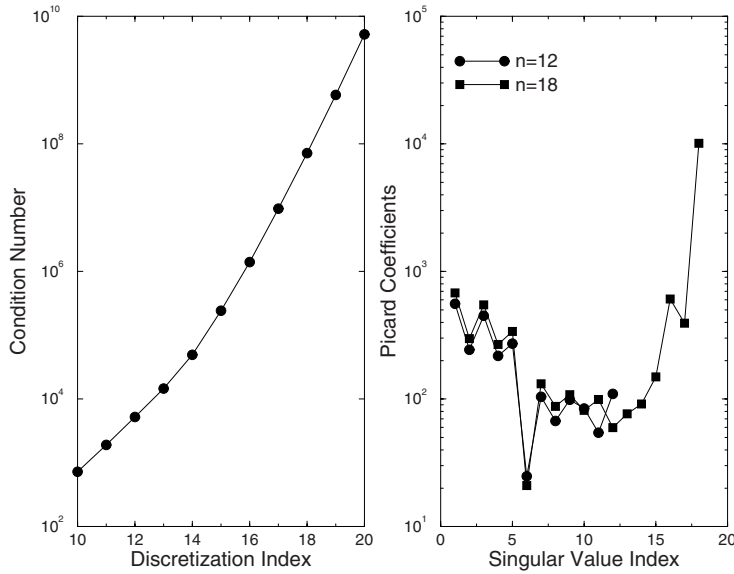
$$\mathbf{T}_n^\delta(\zeta) = \sum_{j=1}^n [\mathbf{T}_n^\delta]_j P_{nj}(\zeta).$$

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

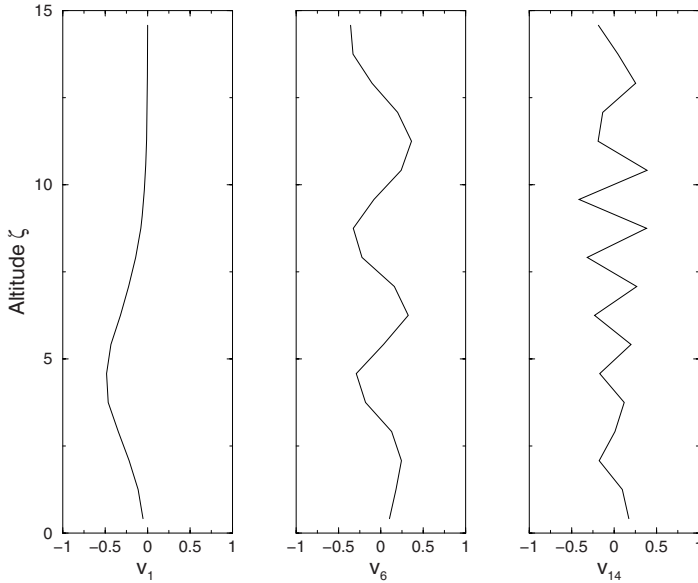
$$P_i^\delta = \frac{|\mathbf{u}_i^T \mathbf{I}_m^\delta|}{\sigma_i},$$

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

Figure 2.4 shows the singular vectors  $\mathbf{v}_1$ ,  $\mathbf{v}_6$  and  $\mathbf{v}_{14}$  corresponding to the singular values  $\sigma_1 = 1.8 \cdot 10^3$ ,  $\sigma_6 = 1.2 \cdot 10^2$  and  $\sigma_{14} = 1.6 \cdot 10^{-1}$ , respectively. The results illustrate that the number of oscillations of the singular vector components increases when the corresponding singular values decrease. Note that fine structures in the profile are reproduced by singular vectors corresponding to smaller singular values, while singular vectors corresponding to larger singular values are responsible for the rough structures (see, e.g., Rodgers, 2000).

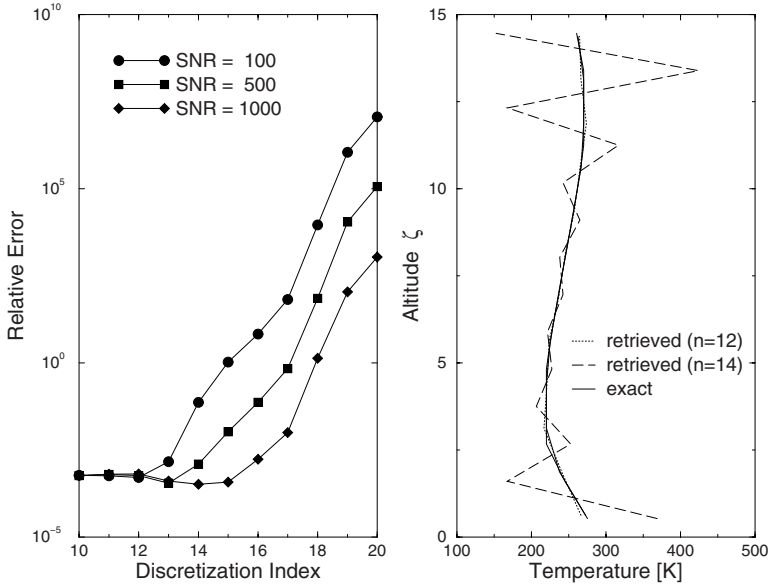


**Fig. 2.3.** Left: condition number versus the discretization index. Right: Picard coefficients for  $n = 12$  and  $n = 18$ , in the case  $\text{SNR} = 100$ .



**Fig. 2.4.** Singular vectors  $\mathbf{v}_1$ ,  $\mathbf{v}_6$  and  $\mathbf{v}_{14}$  for  $n = 18$ .

In the left panel of Figure 2.5 we plot the solution errors versus the discretization index  $n$  for different values of the SNR. The results show that the solution error has a minimum for an optimal value  $n^*$  of the discretization index. These values are  $n^* = 12$  for  $\text{SNR} = 100$ ,  $n^* = 13$  for  $\text{SNR} = 500$ , and  $n^* = 14$  for  $\text{SNR} = 1000$ ; thus  $n^*$  increases



**Fig. 2.5.** Left: relative errors versus the discretization index. Right: retrieved profiles for  $n = 12$  and  $n = 14$ , in the case  $\text{SNR} = 100$ .

when the SNR increases. The solution errors corresponding to the optimal values of the discretization index are reasonable small. This is apparent from the right panel of Figure 2.5, which illustrates the retrieved profile for  $n = 12$  and  $\text{SNR} = 100$ . In contrast, up to the discretization index  $n = 14$ , the least squares solution oscillates and has a large norm.

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

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

where

$$T_n^\dagger(\zeta) = \sum_{j=1}^n [\mathbf{T}_n^\dagger]_j P_{nj}(\zeta)$$

and  $\mathbf{T}_n^\dagger = \mathbf{K}_{mn}^\dagger \mathbf{I}_m$  is the least squares solution in the noise-free case. On the other hand, if the discretization is too fine, then the influence of the small singular values is significant, and the so-called noise error

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

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

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

# 3

## Tikhonov regularization for linear problems

This chapter deals with Tikhonov regularization, which is perhaps the most widely used technique for regularizing discrete ill-posed problems. The reader is encouraged to look at the original papers by Tikhonov (1963a, 1963b) and the monograph by Tikhonov and Arsenin (1977) for the very fundamental results.

We begin our analysis by formulating the method of Tikhonov regularization for linear problems and by making some general remarks on the selection of the regularization matrix. We then summarize the generalized singular value decomposition and discuss a variety of mathematical tools for obtaining more insight into Tikhonov regularization. Afterward, we analyze one- and multi-parameter regularization methods and compare their efficiency by considering numerical examples from atmospheric remote sensing.

### 3.1 Formulation

A problem is called linear if the forward model  $\mathbf{F}$ , which describes the complete physics of the measurement, can be expressed as

$$\mathbf{F}(\mathbf{x}) = \mathbf{K}\mathbf{x}.$$

An example of a linear problem has been considered in the previous chapter. Also encountered in atmospheric remote sensing are the so-called nearly-linear problems. Assuming a linearization of the forward model about some a priori state  $\mathbf{x}_a$ ,

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_a) + \mathbf{K}(\mathbf{x}_a)(\mathbf{x} - \mathbf{x}_a) + O(\|\mathbf{x} - \mathbf{x}_a\|^2), \quad (3.1)$$

where  $\mathbf{K}(\mathbf{x}_a) \in \mathbb{R}^{m \times n}$  is the Jacobian matrix of  $\mathbf{F}(\mathbf{x})$  at  $\mathbf{x}_a$ ,

$$[\mathbf{K}(\mathbf{x}_a)]_{ij} = \frac{\partial [\mathbf{F}]_i}{\partial [\mathbf{x}]_j}(\mathbf{x}_a), \quad i = 1, \dots, m, \quad j = 1, \dots, n,$$

we say that a problem is nearly-linear, when the Taylor remainder or the linearization error is of the same size as the instrumental error. If the actual observations on the forward model

make up the measurement vector  $\mathbf{y}^\delta$ , then by the change of variables  $\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_a) \rightarrow \mathbf{y}^\delta$  and  $\mathbf{x} - \mathbf{x}_a \rightarrow \mathbf{x}$ , we are led to the standard representation of a linear data model

$$\mathbf{y}^\delta = \mathbf{K}\mathbf{x} + \boldsymbol{\delta}, \quad (3.2)$$

with  $\mathbf{K} = \mathbf{K}(\mathbf{x}_a)$ .

For the linear data model (3.2), the Tikhonov solution  $\mathbf{x}_\alpha^\delta$  solves the regularized normal equation

$$(\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L}) \mathbf{x} = \mathbf{K}^T \mathbf{y}^\delta, \quad (3.3)$$

and can be expressed as

$$\mathbf{x}_\alpha^\delta = \mathbf{K}_\alpha^\dagger \mathbf{y}^\delta,$$

where the matrix

$$\mathbf{K}_\alpha^\dagger = (\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T$$

is the regularized generalized inverse. The parameter  $\alpha$  is called the regularization parameter and  $\mathbf{L}$  is known as the regularization matrix. Since  $\mathbf{L}^T \mathbf{L}$  is positive semidefinite, the spectrum of the matrix  $\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L}$  is shifted in the positive direction and the solution of the regularized normal equation is less susceptible to perturbations in the data. If the regularization matrix is chosen so that the Morozov complementary condition  $\|\mathbf{L}\mathbf{x}\| \geq \beta \|\mathbf{x}\|$  is fulfilled for some  $\beta > 0$  and all  $\mathbf{x} \in \mathbb{R}^n$ , then the regularized normal equation has a unique solution  $\mathbf{x}_\alpha^\delta$  which depends continuously on the data  $\mathbf{y}^\delta$ .

Tikhonov regularization can be interpreted as a penalized least squares problem. Taking into account that

$$\begin{aligned} & \|\mathbf{y}^\delta - \mathbf{K}(\mathbf{x} + \Delta\mathbf{x})\|^2 + \alpha \|\mathbf{L}(\mathbf{x} + \Delta\mathbf{x})\|^2 \\ &= \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2 + 2\Delta\mathbf{x}^T [(\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L}) \mathbf{x} - \mathbf{K}^T \mathbf{y}^\delta] \\ & \quad + \|\mathbf{K}\Delta\mathbf{x}\|^2 + \alpha \|\mathbf{L}\Delta\mathbf{x}\|^2 \end{aligned}$$

for all  $\mathbf{x}, \Delta\mathbf{x} \in \mathbb{R}^n$ , we deduce that, if  $\mathbf{x} = \mathbf{x}_\alpha^\delta$  solves equation (3.3), then  $\mathbf{x}_\alpha^\delta$  minimizes the Tikhonov function

$$\mathcal{F}_\alpha(\mathbf{x}) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2. \quad (3.4)$$

The converse result is obvious: the minimizer  $\mathbf{x}_\alpha^\delta$  of the Tikhonov function (3.4) is the solution of the regularized normal equation (3.3). The basic idea of Tikhonov regularization is simple. Minimizing the function (3.4) means to search for some  $\mathbf{x}_\alpha^\delta$  providing at the same time a small residual  $\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2$  and a moderate value of the penalty term  $\|\mathbf{L}\mathbf{x}\|^2$ . The way in which these two terms are balanced depends on the size of  $\alpha$ . If the regularization parameter is chosen too small, equation (3.3) is too close to the original problem and we must expect instabilities; if  $\alpha$  is too large, the equation we solve has only little connection with the original problem.

Regularization methods for solving ill-posed problems can be analyzed in a deterministic or a semi-stochastic setting.

- (1) In a deterministic setting, the solution  $\mathbf{x}^\dagger$  corresponding to the exact data vector  $\mathbf{y}$  is assumed to be deterministic and only information on the noise level  $\Delta$ , defined as  $\|\mathbf{y}^\delta - \mathbf{y}\| \leq \Delta$ , is available.

- (2) In a semi-stochastic setting, the solution  $\mathbf{x}^\dagger$  is deterministic, while the instrumental noise  $\boldsymbol{\delta}$  is assumed to be an  $m$ -dimensional random vector. Usually,  $\boldsymbol{\delta}$  is supposed to be a discrete white noise with zero mean and covariance  $\mathbf{C}_\delta = \sigma^2 \mathbf{I}_m$ , where  $\sigma^2$  is the noise variance. It should be pointed out that a data model with an arbitrary instrumental noise covariance matrix can be transformed into a data model with white noise by using the prewhitening technique (see Chapter 6).

The noise level can be estimated by using the probability distribution of the noise, and we might define  $\Delta = \mathcal{E}\{\|\boldsymbol{\delta}\|\}$  or  $\Delta^2 = \mathcal{E}\{\|\boldsymbol{\delta}\|^2\}$ , where  $\mathcal{E}$  is the expected value operator (or expectation operator). These estimates can be computed either numerically by generating randomly a sample of noise vectors and averaging, or analytically, if the explicit integrals of probability densities are available. In the case of white noise, the second criterion yields

$$\Delta^2 = m\sigma^2.$$

From a practical point of view, the Tikhonov solution does not depend on which setting the problem is treated; differences appear when proving convergence and convergence rate results for different regularization parameter choice methods. Although we are mainly interested in a semi-stochastic analysis, we will not abandon the deterministic analysis; whenever is possible we will evidence the similarity between these two interpretations.

In the presence of forward model errors quantified by  $\boldsymbol{\delta}_m$ , the linear data model can be expressed as

$$\mathbf{y}^\delta = \mathbf{K}\mathbf{x} + \boldsymbol{\delta}_y,$$

where the data error  $\boldsymbol{\delta}_y$  is given by

$$\boldsymbol{\delta}_y = \boldsymbol{\delta}_m + \boldsymbol{\delta}.$$

As  $\boldsymbol{\delta}_m$  is likely to be deterministic,  $\boldsymbol{\delta}_y$  has the mean  $\boldsymbol{\delta}_m$  and the covariance  $\sigma^2 \mathbf{I}_m$ . The presence of forward model errors restricts the class of regularization parameter choice methods and often leads to an erroneous error analysis. Unfortunately, we can only estimate the norm of the forward model errors, but we cannot recover the entire error vector. Under the ‘ideal assumption’ that  $\boldsymbol{\delta}_m$  is known, the appropriate Tikhonov function reads as

$$\mathcal{F}_\alpha(\mathbf{x}) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x} - \boldsymbol{\delta}_m\|^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2,$$

and the regularized solution is given by

$$\mathbf{x}_{m\alpha}^\delta = \mathbf{K}_\alpha^\dagger (\mathbf{y}^\delta - \boldsymbol{\delta}_m). \quad (3.5)$$

### 3.2 Regularization matrices

The penalty term in the expression of the Tikhonov function is called the discrete smoothing norm or the constraint norm and is often, but not always, of the form

$$\Omega(\mathbf{x}) = \|\mathbf{L}\mathbf{x}\|.$$

The discrete smoothing norm takes into account the additional information about the solution and its role is to stabilize the problem and to single out a useful and stable solution.

If we intend to control the magnitude of the solution, then  $\mathbf{L}$  can be chosen as either the identity matrix ( $\mathbf{L}_0 = \mathbf{I}_n$ ) or a diagonal matrix. If the solution should be smooth, then we have to use another measure of the solution, such as the discrete approximations to derivative operators. The use of discrete approximations to derivative operators rather than the identity is recommended by the following argument: the noisy components in the data lead to rough oscillations of the solution which provoke large  $\mathbf{L}$ -norms  $\|\mathbf{L}\cdot\|$ , but do not affect that much the standard norm  $\|\cdot\|$ . The discrete approximations to the first-order ( $\mathbf{L}_1$ ) and the second-order ( $\mathbf{L}_2$ ) derivative operators are frequently used to model this type of additional information. For certain discretizations, the possible forms for the first-order difference regularization matrix are (Gouveia and Scales, 1997)

$$\mathbf{L}_1 = \begin{bmatrix} -1 & 1 & \dots & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1) \times n}, \quad (3.6)$$

and

$$\mathbf{L}_1 = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & \dots & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}. \quad (3.7)$$

There are important differences between the matrix representations (3.6) and (3.7). While they both smooth the solution, the regularization matrix (3.6) maps constant vectors into zero and has the same null space as the continuous first-order derivative operator. Note that for a regularization matrix with  $\text{rank}(\mathbf{L}) < n$ , we have  $\mathcal{N}(\mathbf{L}) \neq \emptyset$ , and  $\|\mathbf{L}\cdot\|$  is said to be a seminorm because it is zero for any vector  $\mathbf{x} \in \mathcal{N}(\mathbf{L})$ , not just for  $\mathbf{x} = \mathbf{0}$ . The matrix (3.6) has a regularizing effect if and only if its null space does not overlap with the null space of the forward model matrix. Indeed, assuming that  $\delta\mathbf{x}$  is a perturbation that happens to be in the null space of  $\mathbf{K}$  and in the null space of  $\mathbf{L}$ , then  $\mathcal{F}_\alpha(\mathbf{x} + \delta\mathbf{x}) = \mathcal{F}_\alpha(\mathbf{x})$  and no improvement of the Tikhonov function is possible. The regularization matrix (3.7) is not singular and has a regularizing effect regardless of the null space of  $\mathbf{K}$ . The first line of this matrix shows that  $\mathbf{L}_1$  controls the magnitude of the first component of the solution. If  $\mathbf{x}$  represents the deviation of an atmospheric profile from the a priori, then this requirement can be regarded as a boundary condition which is imposed on the first component of  $\mathbf{x}$ . In atmospheric remote sensing, this assumption is in general reasonable, because in the upper part of the atmosphere, the gas concentration is very small and the retrieved profile may be close to the a priori profile. As in (3.6) and (3.7), the second-order difference regularization matrix can be expressed as

$$\mathbf{L}_2 = \begin{bmatrix} 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(n-2) \times n},$$

and

$$\mathbf{L}_2 = \begin{bmatrix} -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -2 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 & -2 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

If we have some knowledge about the magnitude of the state vector and we want to constrain the solution to be smooth, we can combine several derivative orders and determine the regularization matrix by the Cholesky factorization (Hansen, 1998)

$$\mathbf{L}^T \mathbf{L} = \omega_0 \mathbf{L}_0^T \mathbf{L}_0 + \omega_1 \mathbf{L}_1^T \mathbf{L}_1,$$

where  $\omega_0$  and  $\omega_1$  are positive weighting factors satisfying the normalization condition

$$\omega_0 + \omega_1 = 1.$$

To compute the regularization matrix  $\mathbf{L}$ , we may consider the QR factorization of the ‘stacked’ matrix,

$$\mathbf{M} = \begin{bmatrix} \sqrt{\omega_1} \mathbf{L}_1 \\ \sqrt{\omega_0} \mathbf{L}_0 \end{bmatrix} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix},$$

with  $\mathbf{Q} \in \mathbb{R}^{(2n-1) \times (2n-1)}$  and  $\mathbf{R} \in \mathbb{R}^{n \times n}$ , and in view of the identity

$$\mathbf{M}^T \mathbf{M} = \omega_0 \mathbf{L}_0^T \mathbf{L}_0 + \omega_1 \mathbf{L}_1^T \mathbf{L}_1 = \mathbf{R}^T \mathbf{R},$$

to set  $\mathbf{L} = \mathbf{R}$ . This triangular QR factor can be computed very efficiently since  $\mathbf{L}_0$  and  $\mathbf{L}_1$  are band matrices. For example, if both  $\omega_0$  and  $\omega_1$  are non-zero, then the sequence of Givens rotations proposed by Elden (1977) for annihilating a diagonal matrix below a bidiagonal matrix can be used.

The regularization matrix can also be constructed by means of statistical information, that is,  $\mathbf{L}$  can be the Cholesky factor of an a priori profile covariance matrix. This construction is legitimated by the similarity between Tikhonov regularization and the Bayesian approach from statistical inversion theory. The covariance matrix  $\mathbf{C}_x$  corresponding to an exponential correlation function is given by

$$[\mathbf{C}_x]_{ij} = \sigma_{xi} \sigma_{xj} [\mathbf{x}_a]_i [\mathbf{x}_a]_j \exp \left( -2 \frac{|z_i - z_j|}{l_i + l_j} \right), \quad i, j = 1, \dots, n,$$

where  $\sigma_{xi}$  are the dimensionless profile standard deviations and  $l_i$  are the lengths which determine the correlation between the parameters at different altitudes  $z_i$ . Defining the diagonal matrices  $\mathbf{\Gamma}$  and  $\mathbf{X}_a$  by  $[\mathbf{\Gamma}]_{ii} = \sigma_{xi}$  and  $[\mathbf{X}_a]_{ii} = [\mathbf{x}_a]_i$ , respectively, and the dense matrix  $\mathbf{R}$  by

$$[\mathbf{R}]_{ij} = \exp \left( -2 \frac{|z_i - z_j|}{l_i + l_j} \right), \quad i, j = 1, \dots, n,$$

we obtain the representation

$$\mathbf{C}_x = (\mathbf{\Gamma} \mathbf{X}_a) \mathbf{R} (\mathbf{\Gamma} \mathbf{X}_a)^T.$$



The inverse of the matrix  $\mathbf{R}$ , which reproduces the shape of the correlation function, can be factorized as

$$\mathbf{R}^{-1} = \mathbf{L}_n^T \mathbf{L}_n,$$

where the Cholesky factor  $\mathbf{L}_n$  is the so-called normalized regularization matrix. To compute the regularization matrix  $\mathbf{L}$  we have two options:

- (1) If the profile standard deviations are known to a certain accuracy, the regularization matrix  $\mathbf{L}$  is defined by the Cholesky factorization

$$\mathbf{C}_x^{-1} = \mathbf{L}^T \mathbf{L},$$

which, in turn, implies that

$$\mathbf{L} = \mathbf{L}_n (\Gamma \mathbf{X}_a)^{-1}.$$

In this case, the regularization parameter  $\alpha$  can be regarded as a scale factor for the matrix  $\Gamma$ . If our assumption on the profile standard deviations is correct, a regularization parameter choice method will yield a scale factor close to one.

- (2) If the profile standard deviations are unknown, it is appropriate to assume that  $\sigma_{xi} = \sigma_x$  for all  $i = 1, \dots, n$ . Consequently,  $\Gamma = \sigma_x \mathbf{I}_n$ , and the covariance matrix can be expressed as

$$\mathbf{C}_x = \sigma_x^2 \mathbf{C}_{nx},$$

where the normalized covariance matrix  $\mathbf{C}_{nx}$  is given by

$$\mathbf{C}_{nx} = \mathbf{X}_a \mathbf{R} \mathbf{X}_a^T.$$

The regularization matrix  $\mathbf{L}$  is then defined as the Cholesky factor of the inverse of the normalized covariance matrix,

$$\mathbf{C}_{nx}^{-1} = \mathbf{L}^T \mathbf{L},$$

and we have the representation

$$\mathbf{L} = \mathbf{L}_n \mathbf{X}_a^{-1}.$$

By this construction, the regularization parameter  $\alpha$  reproduces the profile standard deviation  $\sigma_x$ , and a regularization parameter choice method will yield an estimate for  $\sigma_x$ .

The smoothing property of  $\mathbf{L}$  is reflected by  $\mathbf{L}_n$ . To give a deterministic interpretation of the normalized regularization matrix, we consider an equidistant altitude grid with the step  $\Delta z$  and assume that  $l_i = l$  for all  $i = 1, \dots, n$ . The matrix  $\mathbf{R}$  can then be inverted analytically and the result is (Steck and von Clarmann, 2001),

$$\mathbf{R}^{-1} = \frac{1}{1 - \zeta^2} \begin{bmatrix} 1 & -\zeta & 0 & \dots & 0 & 0 \\ -\zeta & 1 + \zeta^2 & -\zeta & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 + \zeta^2 & -\zeta \\ 0 & 0 & 0 & \dots & -\zeta & 1 \end{bmatrix},$$

with  $\zeta = \exp(-\Delta z/l)$ . This gives

$$\mathbf{L}_n = \frac{1}{\sqrt{1-\zeta^2}} \begin{bmatrix} 1 & -\zeta & 0 & \dots & 0 & 0 \\ 0 & 1 & -\zeta & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -\zeta \\ 0 & 0 & 0 & \dots & 0 & \sqrt{1-\zeta^2} \end{bmatrix},$$

and from a deterministic point of view, we see that  $\mathbf{L}_n \rightarrow \mathbf{L}_0$  as  $l \rightarrow 0$ , and that  $\mathbf{L}_n$  behaves like  $\mathbf{L}_1$  as  $l \rightarrow \infty$ .

In the QPACK tool developed by Eriksson et al. (2005) other types of covariance matrices are considered, e.g., the covariance matrix with a Gaussian correlation function

$$[\mathbf{C}_x]_{ij} = \sigma_{xi}\sigma_{xj} [\mathbf{x}_a]_i [\mathbf{x}_a]_j \exp\left(-4\left(\frac{z_i - z_j}{l_i + l_j}\right)^2\right), \quad i, j = 1, \dots, n,$$

and the covariance matrix with a linearly decreasing correlation function (a tent function)

$$[\mathbf{C}_x]_{ij} = \max\left(0, \sigma_{xi}\sigma_{xj} [\mathbf{x}_a]_i [\mathbf{x}_a]_j \left[1 - 2(1 - e^{-1}) \frac{|z_i - z_j|}{l_i + l_j}\right]\right), \quad i, j = 1, \dots, n.$$

### 3.3 Generalized singular value decomposition and regularized solution

The generalized singular value decomposition (GSVD) of the matrix pair  $(\mathbf{K}, \mathbf{L})$  is a numerical tool which yields important insight into the regularization problem. The use of the SVD and the GSVD in the analysis of discrete ill-posed problems goes back to Hanson (1971) and Varah (1973). In this section, we review this ‘canonical decomposition’ by following the presentation of Hansen (1998).

If  $\mathbf{K}$  is an  $m \times n$  matrix and  $\mathbf{L}$  is a  $p \times n$  matrix, with  $m > n \geq p$ , and further, if  $\text{rank}(\mathbf{L}) = p$  and  $\mathcal{N}(\mathbf{K}) \cap \mathcal{N}(\mathbf{L}) = \emptyset$ , then the GSVD of the matrix pair  $(\mathbf{K}, \mathbf{L})$  is given by

$$\mathbf{K} = \mathbf{U}\Sigma_1\mathbf{W}^{-1}, \quad \mathbf{L} = \mathbf{V}\Sigma_2\mathbf{W}^{-1}, \quad (3.8)$$

where the matrices  $\Sigma_1$  and  $\Sigma_2$  are of the form

$$\Sigma_1 = \begin{bmatrix} \text{diag}(\sigma_i)_{p \times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-p} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \text{diag}(\mu_i)_{p \times p} & \mathbf{0} \end{bmatrix},$$

the matrices  $\mathbf{U}$  and  $\mathbf{V}$ , partitioned as

$$\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_m] \in \mathbb{R}^{m \times m}, \quad \mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_p] \in \mathbb{R}^{p \times p},$$

are orthogonal, i.e.,

$$\mathbf{U}^T \mathbf{U} = \mathbf{U} \mathbf{U}^T = \mathbf{I}_m, \quad \mathbf{V}^T \mathbf{V} = \mathbf{V} \mathbf{V}^T = \mathbf{I}_p,$$

and the matrix

$$\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_n] \in \mathbb{R}^{n \times n}$$

is nonsingular. Moreover,  $\text{diag}(\sigma_i)_{p \times p}$  and  $\text{diag}(\mu_i)_{p \times p}$  are  $p \times p$  diagonal matrices, whose entries are positive and normalized via

$$\sigma_i^2 + \mu_i^2 = 1, \quad i = 1, \dots, p.$$

The generalized singular values of  $(\mathbf{K}, \mathbf{L})$  are defined by

$$\gamma_i = \frac{\sigma_i}{\mu_i},$$

and we shall assume them to be distinct and to appear in decreasing order

$$\gamma_1 > \dots > \gamma_p > 0.$$

From the identities  $\mathbf{KW} = \mathbf{U}\Sigma_1$  and  $\mathbf{LW} = \mathbf{V}\Sigma_2$ , we see that

$$\mathbf{K}\mathbf{w}_i = \sigma_i \mathbf{u}_i, \quad \mathbf{L}\mathbf{w}_i = \mu_i \mathbf{v}_i, \quad i = 1, \dots, p, \quad (3.9)$$

and that

$$\mathbf{K}\mathbf{w}_i = \mathbf{u}_i, \quad \mathbf{L}\mathbf{w}_i = \mathbf{0}, \quad i = p+1, \dots, n.$$

Thus, the set  $\{\mathbf{w}_i\}_{i=\overline{p+1, n}}$  is a basis of  $\mathcal{N}(\mathbf{L})$  and since

$$\mathbf{w}_i^T \mathbf{K}^T \mathbf{K} \mathbf{w}_j = (\mathbf{K}\mathbf{w}_i)^T (\mathbf{K}\mathbf{w}_j) = \mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}, \quad i, j = p+1, \dots, n,$$

where  $\delta_{ij}$  is the Kronecker symbol, we deduce that  $\{\mathbf{w}_i\}_{i=\overline{p+1, n}}$  is  $\mathbf{K}^T \mathbf{K}$ -orthogonal. Scalar multiplying the first and the second equations in (3.9) with  $\mathbf{u}_j$  and  $\mathbf{v}_j$ , respectively, yields further two important relations, namely

$$\mathbf{w}_i^T \mathbf{K}^T \mathbf{u}_j = \sigma_i \delta_{ij}, \quad \mathbf{w}_i^T \mathbf{L}^T \mathbf{v}_j = \mu_i \delta_{ij}, \quad i, j = 1, \dots, p.$$

If  $\mathbf{L}$  is an  $n \times n$  nonsingular matrix ( $p = n$ ), we have

$$\Sigma_1 = \begin{bmatrix} \text{diag}(\sigma_i)_{n \times n} \\ \mathbf{0} \end{bmatrix}, \quad \Sigma_2 = [\text{diag}(\mu_i)_{n \times n}], \quad (3.10)$$

and

$$\mathbf{KL}^{-1} = \mathbf{U}\Sigma_1\Sigma_2^{-1}\mathbf{V}^T, \quad (3.11)$$

with

$$\Sigma_1\Sigma_2^{-1} = \begin{bmatrix} \text{diag}(\gamma_i)_{n \times n} \\ \mathbf{0} \end{bmatrix}.$$

By virtue of (3.11), it is apparent that the SVD of the matrix quotient  $\mathbf{KL}^{-1}$  is given by the GSVD of the matrix pair  $(\mathbf{K}, \mathbf{L})$ . Therefore, instead of the term GSVD, the term quotient SVD is also encountered in the literature (De Moor and Zha, 1991).

If in particular,  $\mathbf{L}$  is the identity matrix  $\mathbf{I}_n$ , then the  $\mathbf{U}$  and  $\mathbf{V}$  of the GSVD coincide with the  $\mathbf{U}$  and  $\mathbf{V}$  of the SVD, and the generalized singular values of  $(\mathbf{K}, \mathbf{L})$  are identical to the singular values of  $\mathbf{K}$ .

The matrix  $\Sigma_1$  reflects the ill-conditioning of  $\mathbf{K}$ . For small  $\sigma_i$ , we have

$$\gamma_i = \frac{\sigma_i}{\sqrt{1 - \sigma_i^2}} \approx \sigma_i,$$

and we see that the generalized singular values decay gradually to zero as the ordinary singular values do. In connection with discrete ill-posed problems, the following features of the GSVD can be evidenced (Hansen, 1998):

- (1) the generalized singular values  $\gamma_i$  decay to zero with no gap in the spectrum, and the number of small generalized singular values increases as the dimension of  $\mathbf{K}$  increases;
- (2) the singular vectors  $\mathbf{u}_i$ ,  $\mathbf{v}_i$  and  $\mathbf{w}_i$  have more sign changes in their components as the corresponding generalized singular values  $\gamma_i$  decrease.

We turn now to the representation of the regularized solution in terms of a generalized singular system of  $(\mathbf{K}, \mathbf{L})$ . By (3.8), we see that the regularized generalized inverse  $\mathbf{K}_\alpha^\dagger$  possesses the factorization

$$\mathbf{K}_\alpha^\dagger = (\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T = \mathbf{W} \Sigma_\alpha^\dagger \mathbf{U}^T, \quad (3.12)$$

with

$$\Sigma_\alpha^\dagger = \begin{bmatrix} \text{diag} \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)_{p \times p} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-p} & \mathbf{0} \end{bmatrix}. \quad (3.13)$$

As a consequence, the regularized solution  $\mathbf{x}_\alpha^\delta$  takes the form

$$\mathbf{x}_\alpha^\delta = \mathbf{K}_\alpha^\dagger \mathbf{y}^\delta = \sum_{i=1}^p \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i + \sum_{i=p+1}^n (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i, \quad (3.14)$$

where the second term

$$\mathbf{x}_0^\delta = \sum_{i=p+1}^n (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i, \quad (3.15)$$

is the component of the solution in the null space of  $\mathbf{L}$ .

If  $p = n$ , the expressions of  $\Sigma_1$  and  $\Sigma_2$  are given by (3.10); these yield

$$\Sigma_\alpha^\dagger = \begin{bmatrix} \text{diag} \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)_{n \times n} & \mathbf{0} \end{bmatrix}, \quad (3.16)$$

and we deduce that the expression of the regularized solution  $\mathbf{x}_\alpha^\delta$  simplifies to

$$\mathbf{x}_\alpha^\delta = \mathbf{K}_\alpha^\dagger \mathbf{y}^\delta = \sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i. \quad (3.17)$$

Further, the factorization

$$\mathbf{K}^\dagger = (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T = \mathbf{W} \Sigma_0^\dagger \mathbf{U}^T$$

gives the following representation of the exact solution:

$$\mathbf{x}^\dagger = \mathbf{K}^\dagger \mathbf{y} = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{w}_i. \quad (3.18)$$

Here, the notation  $\Sigma_0^\dagger$  stands for  $\Sigma_\alpha^\dagger$  with  $\alpha = 0$ . In the data space, we note the useful expansions (cf. (3.9) and (3.17))

$$\mathbf{K} \mathbf{x}_\alpha^\delta = \sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i, \quad (3.19)$$

and (cf. (3.9) and (3.18))

$$\mathbf{K} \mathbf{x}^\dagger = \sum_{i=1}^n (\mathbf{u}_i^T \mathbf{y}) \mathbf{u}_i = \mathbf{y}. \quad (3.20)$$

The computation of the GSVD of  $(\mathbf{K}, \mathbf{L})$  is quite demanding and for this reason, the GSVD is of computational interest only for small- and medium-sized problems. For practical solutions of large-scale problems, algorithms based on standard-form transformation are frequently used. A regularization problem with a discrete smoothing norm  $\Omega(\mathbf{x}) = \|\mathbf{L}\mathbf{x}\|$  is said to be in standard form if the matrix  $\mathbf{L}$  is the identity matrix  $\mathbf{I}_n$ . From a numerical point of view it is much simpler to treat problems in standard form because only one matrix is involved, namely  $\mathbf{K}$ , and the computation of the SVD of the matrix  $\mathbf{K}$  is not so time-consuming. To distinguish the standard-form problem from the general-form problem (3.4), we use bars in our notation, i.e., we are looking for a related minimization problem

$$\min_{\bar{\mathbf{x}}} \mathcal{F}_\alpha(\bar{\mathbf{x}}) = \|\bar{\mathbf{y}}^\delta - \bar{\mathbf{K}}\bar{\mathbf{x}}\|^2 + \alpha \|\bar{\mathbf{x}}\|^2. \quad (3.21)$$

If  $\text{rank}(\bar{\mathbf{K}}) = n$  and  $(\bar{\sigma}_i; \bar{\mathbf{u}}_i, \bar{\mathbf{v}}_i)$  is a singular system of  $\bar{\mathbf{K}}$ , then the regularized solution of (3.21) takes the form

$$\bar{\mathbf{x}}_\alpha^\delta = \bar{\mathbf{K}}_\alpha^\dagger \bar{\mathbf{y}}^\delta = \sum_{i=1}^n \frac{\bar{\sigma}_i^2}{\bar{\sigma}_i^2 + \alpha} \frac{1}{\bar{\sigma}_i} (\bar{\mathbf{u}}_i^T \bar{\mathbf{y}}^\delta) \bar{\mathbf{v}}_i. \quad (3.22)$$

For the simplest case where  $\mathbf{L}$  is square and nonsingular, we put  $\bar{\mathbf{x}} = \mathbf{L}\mathbf{x}$ ; the standard-form transformation is then given by

$$\bar{\mathbf{K}} = \mathbf{K}\mathbf{L}^{-1}, \quad \bar{\mathbf{y}}^\delta = \mathbf{y}^\delta,$$

while the back-transformation becomes

$$\mathbf{x}_\alpha^\delta = \mathbf{L}^{-1} \bar{\mathbf{x}}_\alpha^\delta.$$

For a rectangular (or non-square) regularization matrix, explicit and implicit transformations are given in Appendix B.

In atmospheric remote sensing, the regularization matrix is frequently constructed as the Cholesky factor of some a priori profile covariance matrix and is therefore square and nonsingular. For this reason and in order to simplify our analysis, we will consider the expression of the regularized solution as in (3.17).

### 3.4 Iterated Tikhonov regularization

In the presence of noise, the exact solution of an ill-posed problem can be reconstructed with limited accuracy. In those cases where the Tikhonov solution fails to have optimal accuracy, it is possible to improve it by using the so-called iterated Tikhonov regularization. The first iteration step of iterated Tikhonov regularization is the step of the ordinary method, while at the iteration step  $p \geq 2$ , we evaluate the defect of the linear equation and formulate a new equation in terms of the improved solution step  $\mathbf{p}$ ,

$$\mathbf{K}\mathbf{p} = \mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{\alpha p-1}^\delta. \quad (3.23)$$

Equation (3.23) is again solved by means of Tikhonov regularization, i.e., the improved solution step  $\mathbf{p}_{\alpha p}^\delta$  minimizes the function

$$\mathcal{F}_{\alpha p}(\mathbf{p}) = \|(\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{\alpha p-1}^\delta) - \mathbf{K}\mathbf{p}\|^2 + \alpha \|\mathbf{L}\mathbf{p}\|^2,$$

and the new approximation is given by

$$\mathbf{x}_{\alpha p}^\delta = \mathbf{x}_{\alpha p-1}^\delta + \mathbf{p}_{\alpha p}^\delta.$$

If we iterate  $p$  times, we obtain the iterated Tikhonov regularization of order  $p$ , and the accuracy of the solution increases with every iteration. In fact, for sufficiently large  $p$ , the reconstruction reaches an accuracy that cannot be improved significantly by any other method.

The iterated Tikhonov solution is defined by the regularized normal equation

$$(\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L}) \mathbf{x}_{\alpha p}^\delta = \mathbf{K}^T \mathbf{y}^\delta + \alpha \mathbf{L}^T \mathbf{L} \mathbf{x}_{\alpha p-1}^\delta, \quad (3.24)$$

and can be expressed as

$$\mathbf{x}_{\alpha p}^\delta = \mathbf{K}_\alpha^\dagger \mathbf{y}^\delta + \mathbf{M}_\alpha \mathbf{x}_{\alpha p-1}^\delta,$$

with

$$\mathbf{M}_\alpha = \mathbf{I}_n - \mathbf{K}_\alpha^\dagger \mathbf{K}.$$

Considering a generalized singular value decomposition of the matrix pair  $(\mathbf{K}, \mathbf{L})$ , we find the solution representation (see (3.35) and (3.36) below for computing  $\mathbf{M}_\alpha$ )

$$\mathbf{x}_{\alpha p}^\delta = \left( \sum_{l=0}^{p-1} \mathbf{M}_\alpha^l \right) \mathbf{K}_\alpha^\dagger \mathbf{y}^\delta = \sum_{i=1}^n \left[ 1 - \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^p \right] \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i.$$

Usually, iterated Tikhonov regularization is used with a fixed order  $p$ , but (3.24) can also be regarded as an iterative regularization method when  $p$  is variable and  $\alpha$  depends on  $p$ . The resulting method, in which the iteration index  $p$  plays the role of the regularization parameter, is known as the non-stationary iterated Tikhonov regularization (Hanke and Groetsch, 1998).

### 3.5 Analysis tools

A variety of mathematical tools have been designed to obtain more insight into a discrete ill-posed problem. These tools comprise the filter factors, the errors in the state space and the data space, the mean square error matrix, and the averaging kernels. The discrete Picard condition and several graphical tools as, for instance, the residual curve and the L-curve are also relevant for the analysis of discrete ill-posed problems.

To compute expected values of random vectors we will use the so-called trace lemma (Vogel, 2002). This states that, if  $\delta$  is a discrete white noise with zero mean vector and covariance matrix  $\sigma^2 \mathbf{I}_m$ ,  $\mathbf{y}$  is an  $m$ -dimensional deterministic vector, and  $\mathbf{A}$  is an  $m \times m$  deterministic matrix, there holds

$$\begin{aligned} \mathcal{E} \left\{ \|\mathbf{y} + \mathbf{A}\delta\|^2 \right\} &= \mathcal{E} \left\{ \|\mathbf{y}\|^2 + 2\mathbf{y}^T \mathbf{A}\delta + \|\mathbf{A}\delta\|^2 \right\} \\ &= \|\mathbf{y}\|^2 + \mathcal{E} \left\{ \|\mathbf{A}\delta\|^2 \right\} \\ &= \|\mathbf{y}\|^2 + \sum_{i=1}^m \sum_{j=1}^m [\mathbf{A}^T \mathbf{A}]_{ij} \mathcal{E} \left\{ [\delta]_i [\delta]_j \right\} \\ &= \|\mathbf{y}\|^2 + \sigma^2 \text{trace}(\mathbf{A}^T \mathbf{A}). \end{aligned} \quad (3.25)$$

The following result will be also used in the sequel: if  $\{\mathbf{u}_i\}_{i=\overline{1,m}}$  is an orthonormal basis of  $\mathbb{R}^m$ , we have

$$\mathcal{E} \left\{ (\mathbf{u}_i^T \delta) (\mathbf{u}_j^T \delta) \right\} = \mathcal{E} \left\{ \sum_{k=1}^m \sum_{l=1}^m [\delta]_k [\delta]_l [\mathbf{u}_i]_k [\mathbf{u}_j]_l \right\} = \sigma^2 \mathbf{u}_i^T \mathbf{u}_j = \sigma^2 \delta_{ij}. \quad (3.26)$$

#### 3.5.1 Filter factors

The purpose of a regularization method is to damp or filter out the contributions to the solution corresponding to the small singular values. In general, the regularized solution can be expressed as

$$\mathbf{x}_\alpha^\delta = \sum_{i=1}^n f_\alpha(\gamma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i, \quad (3.27)$$

where  $f_\alpha(\gamma_i^2)$  are the filter factors for a particular regularization method. To damp the contributions  $[(\mathbf{u}_i^T \mathbf{y}^\delta)/\sigma_i] \mathbf{w}_i$  from the smaller  $\sigma_i$ , the filter factors  $f_\alpha(\gamma_i^2)$  must rapidly tend to zero as the  $\sigma_i$  decrease.

The filter factors for Tikhonov regularization and its iterated version are given by

$$f_\alpha(\gamma_i^2) = \frac{\gamma_i^2}{\gamma_i^2 + \alpha}, \quad (3.28)$$

and

$$f_\alpha(\gamma_i^2) = 1 - \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^p, \quad (3.29)$$

respectively. From (3.28) and (3.29) it is apparent that the filter factors are close to 1 for large  $\gamma_i$  and much smaller than 1 for small  $\gamma_i$ . In this way, the contributions to the solution corresponding to the smaller  $\sigma_i$  are filtered out. The filtering effectively sets in for those generalized singular values satisfying  $\gamma_i < \sqrt{\alpha}$ . If the regularization parameter  $\alpha$  is smaller than  $\gamma_n$ , then all the filter factors are approximately 1, and the discrete ill-posed problem is essentially unregularized.

The filter factors can be used to study the influence of the a priori  $\mathbf{x}_a$  on the regularized solution (Hansen, 1998). For this purpose, we choose  $\mathbf{L} = \mathbf{I}_n$ , and express the Tikhonov solution minimizing the function

$$\mathcal{F}_\alpha(\mathbf{x}) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha \|\mathbf{x} - \mathbf{x}_a\|^2$$

as

$$\mathbf{x}_\alpha^\delta = \sum_{i=1}^n \left\{ f_\alpha(\sigma_i^2) \frac{1}{\sigma_i} \mathbf{u}_i^T \mathbf{y}^\delta + [1 - f_\alpha(\sigma_i^2)] \mathbf{v}_i^T \mathbf{x}_a \right\} \mathbf{v}_i. \quad (3.30)$$

The solution representation (3.30) shows that for  $f_\alpha(\sigma_i^2) \approx 1$ , the contribution of the noisy data is dominant, while for  $f_\alpha(\sigma_i^2) \approx 0$ , the contribution of the a priori is dominant. Consequently, for small regularization parameters,  $\mathbf{y}^\delta$  dominates, while for large regularization parameters,  $\mathbf{x}_a$  dominates. This result suggests that the optimal value of the regularization parameter should balance the contributions of the data and the a priori.

### 3.5.2 Error characterization

An error analysis can be performed in the state space by computing the solution error  $\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta$ , or in the data space, by estimating the predictive error  $\mathbf{y} - \mathbf{K}\mathbf{x}_\alpha^\delta$ . Actually, an error analysis is not only a tool for characterizing the accuracy of the solution; it is also the basis for selecting an optimal regularization parameter. In this section we identify the different types of errors and derive representations of the error components in terms of the generalized singular system of the matrix pair  $(\mathbf{K}, \mathbf{L})$ .

#### *Errors in the state space*

Let us express the deviation of the regularized solution from the exact solution as

$$\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta = (\mathbf{x}^\dagger - \mathbf{x}_\alpha) + (\mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta), \quad (3.31)$$

where  $\mathbf{x}_\alpha$  is the regularized solution for the exact data vector  $\mathbf{y}$ , that is,

$$\mathbf{x}_\alpha = \mathbf{K}_\alpha^\dagger \mathbf{y}.$$

Defining the total error by

$$\mathbf{e}_\alpha^\delta = \mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta,$$

and the smoothing and noise errors by

$$\mathbf{e}_{s\alpha} = \mathbf{x}^\dagger - \mathbf{x}_\alpha,$$



and

$$\mathbf{e}_{n\alpha}^\delta = \mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta,$$

respectively, (3.31) becomes

$$\mathbf{e}_\alpha^\delta = \mathbf{e}_{s\alpha} + \mathbf{e}_{n\alpha}^\delta. \quad (3.32)$$

The smoothing error quantifies the loss of information due to the regularization, while the noise error quantifies the loss of information due to the incorrect data.

The smoothing error can be expressed in terms of the exact data vector  $\mathbf{y}$  as

$$\mathbf{e}_{s\alpha} = (\mathbf{K}^\dagger - \mathbf{K}_\alpha^\dagger) \mathbf{y} = \mathbf{W} \left( \Sigma_0^\dagger - \Sigma_\alpha^\dagger \right) \mathbf{U}^T \mathbf{y} = \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{w}_i, \quad (3.33)$$

and in terms of the exact solution  $\mathbf{x}^\dagger$  as

$$\mathbf{e}_{s\alpha} = \mathbf{x}^\dagger - \mathbf{x}_\alpha = (\mathbf{I}_n - \mathbf{K}_\alpha^\dagger \mathbf{K}) \mathbf{x}^\dagger = (\mathbf{I}_n - \mathbf{A}_\alpha) \mathbf{x}^\dagger. \quad (3.34)$$

Here, the  $n \times n$  matrix

$$\mathbf{A}_\alpha = \mathbf{K}_\alpha^\dagger \mathbf{K} = \mathbf{W} \Sigma_\alpha \mathbf{W}^{-1}, \quad (3.35)$$

with

$$\Sigma_\alpha = \left[ \text{diag} \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \right)_{n \times n} \right], \quad (3.36)$$

is called the resolution matrix or the averaging kernel matrix. From (3.34), we deduce that an equivalent expansion for the smoothing error is

$$\mathbf{e}_{s\alpha} = \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha} (\hat{\mathbf{w}}_i^T \mathbf{x}^\dagger) \mathbf{w}_i,$$

where  $\hat{\mathbf{w}}_i^T$  is the  $i$ th row vector of the matrix  $\hat{\mathbf{W}} = \mathbf{W}^{-1}$ . Similarly, the noise error possesses a representation in terms of the noise vector  $\boldsymbol{\delta}$ , that is,

$$\mathbf{e}_{n\alpha}^\delta = \mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta = \mathbf{K}_\alpha^\dagger (\mathbf{y} - \mathbf{y}^\delta) = -\mathbf{K}_\alpha^\dagger \boldsymbol{\delta} = -\sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} (\mathbf{u}_i^T \boldsymbol{\delta}) \mathbf{w}_i. \quad (3.37)$$

In a semi-stochastic setting and for white noise, the smoothing error is deterministic, while the noise error is stochastic with zero mean and covariance

$$\mathbf{C}_{\text{en}} = \sigma^2 \mathbf{K}_\alpha^\dagger \mathbf{K}_\alpha^{\dagger T} = \sigma^2 \mathbf{W} \Sigma_{n\alpha} \mathbf{W}^T, \quad (3.38)$$

where

$$\Sigma_{n\alpha} = \Sigma_\alpha^\dagger \Sigma_\alpha^{\dagger T} = \left[ \text{diag} \left( \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)^2 \right)_{n \times n} \right].$$

If no regularization is applied, the least squares solution  $\mathbf{x}^\delta = \mathbf{K}^\dagger \mathbf{y}^\delta$  is characterized by the noise error covariance matrix

$$\mathbf{C}_{\text{en}0} = \sigma^2 \mathbf{K}^\dagger \mathbf{K}^{\dagger T} = \sigma^2 \mathbf{W} \Sigma_{n0} \mathbf{W}^T. \quad (3.39)$$

From (3.38) and (3.39) we deduce that  $\|\mathbf{C}_{\text{en}}\|$  is generally much smaller than  $\|\mathbf{C}_{\text{eno}}\|$  because the influence from the small  $\sigma_i$  is damped by the corresponding small filter factors  $\gamma_i^2/(\gamma_i^2 + \alpha)$ . Thus, from a stochastic point of view, a regularization method ‘reduces the noise error covariance matrix’ by introducing a bias of the solution (the smoothing error).

The expected value of the total error is given by

$$\mathcal{E} \left\{ \|\mathbf{e}_{\alpha}^{\delta}\|^2 \right\} = \|\mathbf{e}_{\text{s}\alpha}\|^2 + \mathcal{E} \left\{ \|\mathbf{e}_{\text{n}\alpha}^{\delta}\|^2 \right\}, \quad (3.40)$$

whereas the expected value of the noise error is computed as (cf. (3.26) and (3.37))

$$\begin{aligned} \mathcal{E} \left\{ \|\mathbf{e}_{\text{n}\alpha}^{\delta}\|^2 \right\} &= \sum_{i=1}^n \sum_{j=1}^n \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right) \left( \frac{\gamma_j^2}{\gamma_j^2 + \alpha} \frac{1}{\sigma_j} \right) \\ &\quad \times (\mathbf{w}_i^T \mathbf{w}_j) \mathcal{E} \left\{ (\mathbf{u}_i^T \boldsymbol{\delta}) (\mathbf{u}_j^T \boldsymbol{\delta}) \right\} \\ &= \sigma^2 \sum_{i=1}^n \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)^2 \|\mathbf{w}_i\|^2. \end{aligned} \quad (3.41)$$

The smoothing error  $\|\mathbf{e}_{\text{s}\alpha}\|^2$  is an increasing function of  $\alpha$ , while the expected value of the noise error  $\mathcal{E} \left\{ \|\mathbf{e}_{\text{n}\alpha}^{\delta}\|^2 \right\}$  is a decreasing function of  $\alpha$ . Consequently, we may assume that the expected value of the total error  $\mathcal{E} \left\{ \|\mathbf{e}_{\alpha}^{\delta}\|^2 \right\}$  has a minimum for an optimal value of  $\alpha$ . The stability of the linear problem requires a large regularization parameter to keep the noise error small, i.e., to keep the influence of the data errors small. On the other hand, keeping the smoothing error small asks for a small regularization parameter. Obviously, the choice of  $\alpha$  has to be made through a compromise between accuracy and stability.

When the data error  $\boldsymbol{\delta}_y$  is determined by forward model errors and instrumental noise, the regularized solution should be computed as (cf. (3.5))

$$\mathbf{x}_{\text{m}\alpha}^{\delta} = \mathbf{K}_{\alpha}^{\dagger} (\mathbf{y}^{\delta} - \boldsymbol{\delta}_{\text{m}}). \quad (3.42)$$

As  $\boldsymbol{\delta}_{\text{m}}$  is unknown, we can only compute the regularized solution  $\mathbf{x}_{\alpha}^{\delta} = \mathbf{K}_{\alpha}^{\dagger} \mathbf{y}^{\delta}$ ; the relation between  $\mathbf{x}_{\alpha}^{\delta}$  and  $\mathbf{x}_{\text{m}\alpha}^{\delta}$  is given by

$$\mathbf{x}_{\alpha}^{\delta} = \mathbf{x}_{\text{m}\alpha}^{\delta} + \mathbf{K}_{\alpha}^{\dagger} \boldsymbol{\delta}_{\text{m}}.$$

In view of the decomposition

$$\mathbf{x}^{\dagger} - \mathbf{x}_{\alpha}^{\delta} = (\mathbf{x}^{\dagger} - \mathbf{x}_{\alpha}) + (\mathbf{x}_{\alpha} - \mathbf{x}_{\text{m}\alpha}^{\delta}) + (\mathbf{x}_{\text{m}\alpha}^{\delta} - \mathbf{x}_{\alpha}^{\delta}),$$

we introduce the total error in the state space by

$$\mathbf{e}_{\alpha}^{\delta} = \mathbf{e}_{\text{s}\alpha} + \mathbf{e}_{\text{n}\alpha}^{\delta} + \mathbf{e}_{\text{m}\alpha}.$$

Here, the smoothing and noise errors are as in (3.34) and (3.37), respectively, while the new quantity  $\mathbf{e}_{\text{m}\alpha}$ , defined by

$$\mathbf{e}_{\text{m}\alpha} = \mathbf{x}_{\text{m}\alpha}^{\delta} - \mathbf{x}_{\alpha}^{\delta} = -\mathbf{K}_{\alpha}^{\dagger} \boldsymbol{\delta}_{\text{m}}, \quad (3.43)$$

represents the modeling error.

***Constrained errors in the state space***

The error in the solution can also be characterized via the ‘constrained’ total error

$$\mathbf{l}_\alpha^\delta = \mathbf{L} \mathbf{e}_\alpha^\delta = \mathbf{L} (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta).$$

As before, we have the decomposition

$$\mathbf{l}_\alpha^\delta = \mathbf{l}_{s\alpha} + \mathbf{l}_{n\alpha}^\delta,$$

where

$$\mathbf{l}_{s\alpha} = \mathbf{L} \mathbf{e}_{s\alpha} = \mathbf{L} (\mathbf{x}^\dagger - \mathbf{x}_\alpha)$$

is the constrained smoothing error and

$$\mathbf{l}_{n\alpha}^\delta = \mathbf{L} \mathbf{e}_{n\alpha}^\delta = \mathbf{L} (\mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta)$$

is the constrained noise error. Accounting of (3.9) and using (3.33) and (3.37), we find the expansions

$$\mathbf{l}_{s\alpha} = \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha} \frac{1}{\gamma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i, \quad (3.44)$$

and

$$\mathbf{l}_{n\alpha}^\delta = - \sum_{i=1}^n \frac{\gamma_i}{\gamma_i^2 + \alpha} (\mathbf{u}_i^T \boldsymbol{\delta}) \mathbf{v}_i. \quad (3.45)$$

The expected value of the constrained total error is then given by

$$\mathcal{E} \left\{ \|\mathbf{l}_\alpha^\delta\|^2 \right\} = \|\mathbf{l}_{s\alpha}\|^2 + \mathcal{E} \left\{ \|\mathbf{l}_{n\alpha}^\delta\|^2 \right\}, \quad (3.46)$$

with (cf. (3.26), (3.44) and (3.45))

$$\|\mathbf{l}_{s\alpha}\|^2 = \sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 \frac{1}{\gamma_i^2} (\mathbf{u}_i^T \mathbf{y})^2, \quad (3.47)$$

and

$$\mathcal{E} \left\{ \|\mathbf{l}_{n\alpha}^\delta\|^2 \right\} = \sigma^2 \sum_{i=1}^n \left( \frac{\gamma_i}{\gamma_i^2 + \alpha} \right)^2. \quad (3.48)$$

From (3.47) and (3.48) we infer that  $\|\mathbf{l}_{s\alpha}\|^2$  is an increasing function of  $\alpha$  and that  $\mathcal{E}\{\|\mathbf{l}_{n\alpha}^\delta\|^2\}$  is a decreasing function of  $\alpha$ .

When a regularization problem is transformed into the standard form and  $\mathbf{L}$  is nonsingular, we have  $\bar{\mathbf{K}} = \mathbf{K}\mathbf{L}^{-1}$  and  $\bar{\mathbf{x}}_\alpha^\delta = \mathbf{L}\mathbf{x}_\alpha^\delta$ . Thus, the constrained errors corresponding to the general-form solution  $\mathbf{x}_\alpha^\delta$  coincide with the errors corresponding to the standard-form solution  $\bar{\mathbf{x}}_\alpha^\delta$ . As the generalized singular values of  $(\mathbf{K}, \mathbf{L})$  are the singular values of the matrix quotient  $\bar{\mathbf{K}}$ , it is apparent that an analysis involving the constrained errors for the general-form problem is equivalent to an analysis involving the errors for the standard-form problem.

### Errors in the data space

The accuracy of the regularized solution can be characterized via the predictive error or the predictive risk, defined as

$$\mathbf{p}_\alpha^\delta = \mathbf{K} \mathbf{e}_\alpha^\delta = \mathbf{p}_{s\alpha} + \mathbf{p}_{n\alpha}^\delta. \quad (3.49)$$

The predictive smoothing error is given by

$$\mathbf{p}_{s\alpha} = \mathbf{K} \mathbf{e}_{s\alpha} = \mathbf{K} (\mathbf{x}^\dagger - \mathbf{x}_\alpha) = (\mathbf{I}_m - \mathbf{K} \mathbf{K}_\alpha^\dagger) \mathbf{y} = (\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) \mathbf{y}, \quad (3.50)$$

where the  $m \times m$  matrix

$$\hat{\mathbf{A}}_\alpha = \mathbf{K} \mathbf{K}_\alpha^\dagger = \mathbf{U} \hat{\Sigma}_a \mathbf{U}^T, \quad (3.51)$$

with

$$\hat{\Sigma}_a = \begin{bmatrix} \text{diag} \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \right)_{n \times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad (3.52)$$

is called the influence matrix. Essentially, the influence matrix is the counterpart of the resolution matrix and characterizes the smoothing error in the data space. Using the orthogonality relations  $\mathbf{u}_i^T \mathbf{y} = 0$  for  $i = n + 1, \dots, m$ , we obtain the expansion

$$\mathbf{p}_{s\alpha} = \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha} (\mathbf{u}_i^T \mathbf{y}) \mathbf{u}_i. \quad (3.53)$$

For the predictive noise error we find that

$$\mathbf{p}_{n\alpha}^\delta = \mathbf{K} \mathbf{e}_{n\alpha}^\delta = \mathbf{K} (\mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta) = -\mathbf{K} \mathbf{K}_\alpha^\dagger \boldsymbol{\delta} = -\hat{\mathbf{A}}_\alpha \boldsymbol{\delta}, \quad (3.54)$$

and further that

$$\mathbf{p}_{n\alpha}^\delta = -\sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha} (\mathbf{u}_i^T \boldsymbol{\delta}) \mathbf{u}_i. \quad (3.55)$$

Using the representation

$$\mathbf{p}_\alpha^\delta = \mathbf{p}_{s\alpha} - \hat{\mathbf{A}}_\alpha \boldsymbol{\delta}$$

and applying the trace lemma (3.25), we deduce that the expected value of the predictive error is given by

$$\mathcal{E} \left\{ \|\mathbf{p}_\alpha^\delta\|^2 \right\} = \|\mathbf{p}_{s\alpha}\|^2 + \mathcal{E} \left\{ \|\mathbf{p}_{n\alpha}^\delta\|^2 \right\}, \quad (3.56)$$

with

$$\|\mathbf{p}_{s\alpha}\|^2 = \left\| (\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) \mathbf{y} \right\|^2 = \sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y})^2 \quad (3.57)$$

and

$$\mathcal{E} \left\{ \|\mathbf{p}_{n\alpha}^\delta\|^2 \right\} = \sigma^2 \text{trace} \left( \hat{\mathbf{A}}_\alpha^T \hat{\mathbf{A}}_\alpha \right) = \sigma^2 \text{trace} \left( \mathbf{U} \hat{\Sigma}_a^T \hat{\Sigma}_a \mathbf{U}^T \right) = \sigma^2 \sum_{i=1}^n \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \right)^2. \quad (3.58)$$

The monotonicity of the predictive errors is illustrated by (3.57) and (3.58):  $\|\mathbf{p}_{s\alpha}\|^2$  is an increasing function of  $\alpha$  and  $\mathcal{E} \left\{ \|\mathbf{p}_{n\alpha}^\delta\|^2 \right\}$  is a decreasing function of  $\alpha$ . The predictive error is not a computable quantity but it can be approximated with a satisfactory accuracy by the so-called unbiased predictive risk estimator. The minimization of this estimator yields a regularization parameter which balances the smoothing and noise errors in the data space.

### 3.5.3 Mean square error matrix

A measure of the accuracy of the regularized solution is the mean square error matrix defined by (Vinod and Ullah, 1981; O'Sullivan, 1986; Grafarend and Schaffrin, 1993),

$$\begin{aligned} \mathbf{S}_\alpha &= \mathcal{E} \left\{ (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta) (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)^T \right\} \\ &= (\mathbf{x}^\dagger - \mathbf{x}_\alpha) (\mathbf{x}^\dagger - \mathbf{x}_\alpha)^T + \mathcal{E} \left\{ (\mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta) (\mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta)^T \right\} \\ &= (\mathbf{I}_n - \mathbf{A}_\alpha) \mathbf{x}^\dagger \mathbf{x}^{\dagger T} (\mathbf{I}_n - \mathbf{A}_\alpha)^T + \sigma^2 \mathbf{K}_\alpha^\dagger \mathbf{K}_\alpha^{\dagger T}. \end{aligned} \quad (3.59)$$

The first term (bias) in the expression of the mean square error matrix depends on the exact solution  $\mathbf{x}^\dagger$  and is not a computable quantity. Several approximations for this term have been proposed in the literature. Xu and Rummel (1994) suggested the estimate

$$\mathbf{x}^\dagger \mathbf{x}^{\dagger T} \approx \mathbf{x}_\alpha^\delta \mathbf{x}_\alpha^{\delta T}, \quad (3.60)$$

while Grafarend and Schaffrin (1993) proposed the approximation

$$\mathbf{x}^\dagger \mathbf{x}^{\dagger T} \approx \frac{\sigma^2}{\alpha} (\mathbf{L}^T \mathbf{L})^{-1}. \quad (3.61)$$

The estimate (3.61) is justified by the similarity between the mean square error matrix and the a posteriori covariance matrix in statistical inversion theory. In this case, the mean square error matrix becomes

$$\mathbf{S}_\alpha \approx \sigma^2 (\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{L}^T \mathbf{L})^{-1},$$

and coincides with the covariance matrix of the maximum a posteriori estimator (see Chapter 4).

The mean square error matrix can be expressed in terms of the errors in the state space as

$$\mathbf{S}_\alpha = \mathcal{E} \left\{ \mathbf{e}_\alpha^\delta \mathbf{e}_\alpha^{\delta T} \right\} = \mathbf{e}_{s\alpha} \mathbf{e}_{s\alpha}^T + \mathbf{C}_{en},$$

and we have

$$\mathcal{E} \left\{ \|\mathbf{e}_\alpha^\delta\|^2 \right\} = \text{trace}(\mathbf{S}_\alpha).$$

In the presence of forward model errors, the regularized solution is biased by the modeling error in the state space, and  $\mathbf{S}_\alpha$  is given by

$$\mathbf{S}_\alpha = (\mathbf{e}_{s\alpha} + \mathbf{e}_{m\alpha}) (\mathbf{e}_{s\alpha} + \mathbf{e}_{m\alpha})^T + \mathbf{C}_{en}.$$

This relation is useless in practice, and in order to obtain a computable expression, we use the approximation

$$\begin{aligned} \mathbf{S}_\alpha &\approx \mathbf{e}_{s\alpha} \mathbf{e}_{s\alpha}^T + \mathbf{e}_{m\alpha} \mathbf{e}_{m\alpha}^T + \mathbf{C}_{en} \\ &= \mathbf{e}_{s\alpha} \mathbf{e}_{s\alpha}^T + \mathbf{K}_\alpha^\dagger \left( \delta_m \delta_m^T + \sigma^2 \mathbf{I}_m \right) \mathbf{K}_\alpha^{\dagger T}. \end{aligned}$$

The matrix  $\delta_{\mathbf{m}}\delta_{\mathbf{m}}^T + \sigma^2\mathbf{I}_m$ , with diagonal entries

$$\left[\delta_{\mathbf{m}}\delta_{\mathbf{m}}^T + \sigma^2\mathbf{I}_m\right]_{ii} = [\delta_{\mathbf{m}}]_i^2 + \sigma^2, \quad i = 1, \dots, m,$$

is also unknown and we propose the diagonal matrix approximation

$$\delta_{\mathbf{m}}\delta_{\mathbf{m}}^T + \sigma^2\mathbf{I}_m \approx \left(\frac{1}{m} \|\delta_{\mathbf{m}}\|^2 + \sigma^2\right) \mathbf{I}_m. \quad (3.62)$$

According to (3.62), the data error  $\delta_{\mathbf{y}} = \delta_{\mathbf{m}} + \delta$  is replaced by an equivalent white noise  $\delta_{\mathbf{e}}$ , with the variance

$$\sigma_{\mathbf{e}}^2 = \frac{1}{m} \|\delta_{\mathbf{m}}\|^2 + \sigma^2, \quad (3.63)$$

so that

$$\mathcal{E} \left\{ \|\delta_{\mathbf{e}}\|^2 \right\} = \mathcal{E} \left\{ \|\delta_{\mathbf{y}}\|^2 \right\}.$$

The mean square error matrix then becomes

$$\mathbf{S}_{\alpha} \approx \mathbf{e}_{\text{s}\alpha} \mathbf{e}_{\text{s}\alpha}^T + \sigma_{\mathbf{e}}^2 \mathbf{K}_{\alpha}^{\dagger} \mathbf{K}_{\alpha}^{\dagger T}. \quad (3.64)$$

As we will see, the variance  $\sigma_{\mathbf{e}}^2$  can be estimated by computing the norm of the residual  $\mathbf{r}_{\alpha}^{\delta} = \mathbf{y}^{\delta} - \mathbf{K}\mathbf{x}_{\alpha}^{\delta}$  for small values of the regularization parameter  $\alpha$ , and the above equivalence will enable us to perform an approximative error analysis.

### 3.5.4 Resolution matrix and averaging kernels

The mean square error matrix tells us about how precise the regularized solution is. In this section, we consider how much resemblance there is between the exact and the regularized solutions. Representing the regularized solution as

$$\mathbf{x}_{\alpha}^{\delta} = \mathbf{K}_{\alpha}^{\dagger} \mathbf{y}^{\delta} = \mathbf{A}_{\alpha} \mathbf{x}^{\dagger} + \mathbf{K}_{\alpha}^{\dagger} \delta, \quad (3.65)$$

we observe that the first term  $\mathbf{A}_{\alpha} \mathbf{x}^{\dagger}$  is a smoothed version of the exact solution  $\mathbf{x}^{\dagger}$ , while the second term  $\mathbf{K}_{\alpha}^{\dagger} \delta$  reflects the contribution from the noise in the data. Thus, the resolution matrix  $\mathbf{A}_{\alpha}$  quantifies the smoothing of the exact solution by the particular regularization method and describes how well the exact solution is approximated by the regularized solution in the noise-free case.

By virtue of (3.34), it is apparent that the deviation of  $\mathbf{A}_{\alpha}$  from the identity matrix characterizes the smoothing error. Note that although  $\mathbf{A}_{\alpha}$  can deviate significantly from  $\mathbf{I}_n$ , the vector  $\mathbf{A}_{\alpha} \mathbf{x}^{\dagger}$  is still close to  $\mathbf{x}^{\dagger}$  if those spectral components of  $\mathbf{x}^{\dagger}$  which are damped by the multiplication with  $\mathbf{A}_{\alpha}$  are small (Hansen, 1998).

There is more information in the resolution matrix than just a characterization of the smoothing error. If  $\mathbf{a}_{\alpha i}^T$  is the  $i$ th row vector of  $\mathbf{A}_{\alpha}$ , then the  $i$ th component of  $\mathbf{A}_{\alpha} \mathbf{x}^{\dagger}$  is given by  $[\mathbf{A}_{\alpha} \mathbf{x}^{\dagger}]_i = \mathbf{a}_{\alpha i}^T \mathbf{x}^{\dagger}$ , and we see that  $\mathbf{a}_{\alpha i}^T$  expresses  $[\mathbf{A}_{\alpha} \mathbf{x}^{\dagger}]_i$  as a weighted average of all components in  $\mathbf{x}^{\dagger}$ . For this reason, the row vectors  $\mathbf{a}_{\alpha i}^T$  are referred to as the averaging kernels. The  $i$ th averaging kernel has a peak at its  $i$ th component and the width of this

peak depends on the particular regularization method. In atmospheric remote sensing, the averaging kernel is an indication of the vertical resolution of the instrument. According to Rodgers (2000), features in the profile which are much broader than the averaging kernel width will be reproduced well, while features much narrower than the averaging kernel width will be smoothed out.

The width of the averaging kernel can be measured in various ways. A simple measure is the full width of the peak at half of its maximum (FWHM) but this measure does not take into account any ripples on either sides of the main peak. Another way to calculate the width of a function is to use the Backus–Gilbert spread. If we regard  $\mathbf{a}_{\alpha i}$  as a function  $a_i(z)$  of the altitude  $z$ , then the spread of this function about the height  $z_i$  is defined by

$$s(z_i) = c \frac{\int (z - z_i)^2 a_i(z)^2 dz}{\left[ \int a_i(z) dz \right]^2}. \quad (3.66)$$

An alternative form, designed to reduce the problem associated with the presence of negative sidelobes of  $a_i$ , is

$$s(z_i) = c \frac{\int |(z - z_i) a_i(z)| dz}{\int |a_i(z)| dz}, \quad (3.67)$$

while a spread based directly upon the ‘radius of gyration’ of  $a_i^2$  is

$$s(z_i) = \left[ c \frac{\int (z - z_i)^2 a_i(z)^2 dz}{\int a_i(z)^2 dz} \right]^{\frac{1}{2}}. \quad (3.68)$$

The normalization constant  $c$  in the above relations can be chosen so that the spread of a ‘top-hat’ or ‘boxcar’ function is equal to its width. As shown by several authors, the resolution measures (3.66)–(3.68) are in general misleading when the averaging kernels have negative sidelobes of significant amplitudes. In this regard, other measures of resolution derived from the averaging kernels have been proposed by Purser and Huang (1993).

In most applications, the resolution matrix  $\mathbf{A}_\alpha$  is a consequence of the choice of a regularization method. However, in the mollifier method, to be discussed in Chapter 9, the resolution matrix is the starting point for deriving the generalized inverse.

As  $\mathbf{A}_\alpha = \mathbf{K}_\alpha^\dagger \mathbf{K}$  is the resolution matrix for the solution, the influence matrix  $\hat{\mathbf{A}}_\alpha = \mathbf{K} \mathbf{K}_\alpha^\dagger$  is the resolution matrix for the predicted right-hand side, i.e., it describes how well the vector  $\mathbf{K} \mathbf{x}_\alpha^\delta$  predicts the given right-hand side  $\mathbf{y}^\delta$ .

### 3.5.5 Discrete Picard condition

An important analytical tool for analyzing discrete ill-posed problems is the decay of the Fourier coefficients and of the generalized singular values. This topic is strongly connected with the discrete Picard condition. In a continuous setting, Picard’s theorem states that in order for the equation  $Kx = y$  to have a solution  $x^\dagger \in X$ , it is necessary and sufficient that  $y \in \overline{\mathcal{R}(K)}$  and that

$$\sum_{i=1}^{\infty} \frac{\langle y, u_i \rangle^2}{\sigma_i^2} < \infty, \quad (3.69)$$

where  $K$  is a compact operator between the real Hilbert spaces  $X$  and  $Y$ , and  $(\sigma_i; v_i, u_i)$  is a  $K$  system of  $K$ . The infinite sum in (3.69) must converge, which means that the terms in the sum must decay to zero, or equivalently, that the generalized Fourier coefficients  $|\langle y, u_i \rangle|$  must decay faster to zero than the singular values  $\sigma_i$ .

For discrete ill-posed problems there is, strictly speaking, no Picard condition because the solution always exists and is bounded. Nevertheless it makes sense to introduce a discrete Picard condition as follows: the exact data vector  $\mathbf{y}$  of the discrete equation satisfies the discrete Picard condition if the Fourier coefficients  $|\mathbf{u}_i^T \mathbf{y}|$  decay, on the average, to zero faster than the generalized singular values  $\gamma_i$ , that is, the sequence  $\{|\mathbf{u}_i^T \mathbf{y}|/\gamma_i\}$  generally decreases (Hansen, 1990). The discrete Picard condition is not as ‘artificial’ as it may seem; it can be shown that if the underlying continuous equation satisfies the Picard condition, then the discrete equation satisfies the discrete Picard condition (Hansen, 1998). The importance of the discrete Picard condition in the analysis of ill-posed problems has been discussed by Hansen (1992b), and Zha and Hansen (1990).

Let us assume that the Fourier coefficients and the generalized singular values are related by the following model

$$|\mathbf{u}_i^T \mathbf{y}| = C\gamma_i^{\beta+1}, \quad i = 1, \dots, n, \quad (3.70)$$

where  $\beta > 0$  and  $C > 0$ . In addition, if  $p$  is the index defined by

$$(\mathbf{u}_p^T \mathbf{y})^2 = \sigma^2,$$

i.e.,  $C\gamma_p^{\beta+1} = \sigma$ , we suppose that the decay rate of the generalized singular values is such that

$$\begin{aligned} \gamma_i &\gg \gamma_p, \quad i = 1, \dots, p-1, \\ \gamma_i &\ll \gamma_p, \quad i = p+1, \dots, n. \end{aligned} \quad (3.71)$$

Under these assumptions and using the relation (cf. (3.26))

$$\mathcal{E} \left\{ (\mathbf{u}_i^T \boldsymbol{\delta})^2 \right\} = \sigma^2, \quad (3.72)$$

we find that the expected values of the Fourier coefficients, corresponding to the noisy data,

$$F_i^2 = \mathcal{E} \left\{ (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \right\} = (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 = C^2 \left( \gamma_i^{2\beta+2} + \gamma_p^{2\beta+2} \right) \quad (3.73)$$

behave like

$$F_i^2 \propto \begin{cases} \gamma_i^{2\beta+2}, & i = 1, \dots, p-1, \\ \gamma_p^{2\beta+2}, & i = p, \dots, n. \end{cases} \quad (3.74)$$

Thus, for  $i \geq p$ , the Fourier coefficients  $F_i^2$  level off at  $\sigma^2$ . Similarly, for the expected values of the Picard coefficients

$$P_i^2 = \frac{1}{\gamma_i^2} F_i^2, \quad (3.75)$$



there holds

$$P_i^2 \propto \begin{cases} \gamma_i^{2\beta}, & i = 1, \dots, p-1, \\ \gamma_p^{2\beta} \left( \frac{\gamma_p}{\gamma_i} \right)^2, & i = p, \dots, n, \end{cases} \quad (3.76)$$

and we deduce that the Picard coefficients  $P_i^2$  decrease until  $\gamma_p$  and increase afterward.

Another important result, which is also a consequence of assumptions (3.70) and (3.71), states that  $\gamma_p^2$  is close to the optimal regularization parameter for constrained error estimation

$$\bar{\alpha}_{\text{opt}} = \arg \min_{\alpha} \mathcal{E} \left\{ \|\mathbf{l}_{\alpha}^{\delta}\|^2 \right\},$$

where

$$L(\alpha) = \mathcal{E} \left\{ \|\mathbf{l}_{\alpha}^{\delta}\|^2 \right\} = \sum_{i=1}^n \left[ \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 \frac{1}{\gamma_i^2} (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 \left( \frac{\gamma_i}{\gamma_i^2 + \alpha} \right)^2 \right]. \quad (3.77)$$

To justify this assertion we employ a heuristic technique which will be frequently used in the sequel. Let us assume that  $\alpha = \gamma_j^2$  for some  $j = 1, \dots, n$ . Then, using (3.70) and the relation  $\sigma = C\gamma_p^{\beta+1}$ , we obtain

$$L(\gamma_j^2) = C^2 \sum_{i=1}^n \frac{\gamma_i^2 \gamma_j^2}{(\gamma_i^2 + \gamma_j^2)^2} \bar{P}_i^2,$$

where

$$\bar{P}_i^2 = \frac{\gamma_j^2}{\gamma_i^2} \gamma_i^{2\beta} + \frac{1}{\gamma_j^2} \gamma_p^{2\beta+2}.$$

The function  $f(t) = t/(1+t)^2$ , with  $t = \gamma_i^2/\gamma_j^2$ , is common to all the terms in the sum. For  $t \ll 1$ , we have  $f(t) \approx t$ , while for  $t \gg 1$ , we have  $f(t) \approx 1/t$ . Thus,  $f$  is very small if  $t \ll 1$  and  $t \gg 1$ , and we may assume that  $f$  effectively filters out the influence of all  $\bar{P}_i^2$  with  $i \neq j$ . In fact, the validity of this assumption depends on the behavior of the coefficients  $\bar{P}_i^2$ , and, in particular, on the decay rate of the generalized singular values  $\gamma_i$  and the size of the parameter  $\beta$ . We obtain

$$L(\gamma_j^2) \propto \bar{P}_j^2 = P_j^2,$$

and we conclude that  $L(\gamma_j^2)$  has approximately a turning point at  $\gamma_p$  (cf. (3.76)).

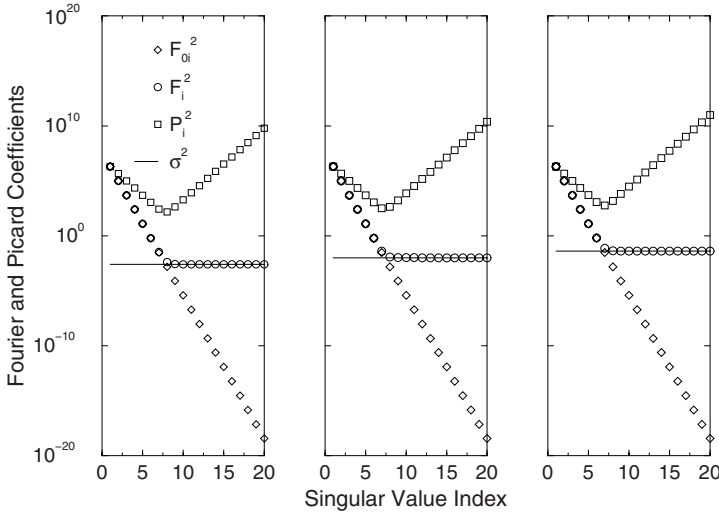
In practice, the computable Fourier coefficients  $(\mathbf{u}_i^T \mathbf{y}^{\delta})^2$  behave like their expected values  $\mathcal{E}\{(\mathbf{u}_i^T \mathbf{y}^{\delta})^2\}$ , and the above results generalize as follows:  $(\mathbf{u}_i^T \mathbf{y}^{\delta})^2$  level off at  $\sigma^2$ , and if  $p$  is the first index satisfying  $(\mathbf{u}_p^T \mathbf{y}^{\delta})^2 \approx \sigma^2$ , then  $\gamma_p^2 \approx \bar{\alpha}_{\text{opt}}$ . The latter result can be used to obtain a rough estimate of the regularization parameter which balances the constrained errors.

In Figure 3.1 we illustrate the Fourier coefficients for exact data  $F_{0i}^2$  together with the expected Fourier and Picard coefficients  $F_i^2$  and  $P_i^2$ , respectively. The results correspond to a synthetic model of a discrete ill-posed problem based on the assumptions

$$\mathbf{L} = \mathbf{I}_n, \quad (3.78)$$

$$\sigma_i = \exp(-\omega i), \quad (3.79)$$

$$|\mathbf{u}_i^T \mathbf{y}| = C\sigma_i^{\beta+1}, \quad (3.80)$$



**Fig. 3.1.** Fourier coefficients for exact data  $F_{0i}^2$  together with the expected Fourier and Picard coefficients  $F_i^2$  and  $P_i^2 = F_i^2/\sigma_i^2$ , respectively. The results correspond to the following values of the noise standard deviation  $\sigma$ : 0.05 (left), 0.1 (middle), and 0.2 (right).

for  $i = 1, \dots, n$ . The parameter  $\beta$ , which controls the decay rate of the Fourier coefficients for exact data, characterizes the smoothness of the exact solution. Note that for  $\mathbf{K} = \mathbf{U}\Sigma\mathbf{V}^T$ , we have  $\mathbf{x}^\dagger = (\mathbf{K}^T\mathbf{K})^{\beta/2}\mathbf{z}$ , with  $\mathbf{z} = C\sum_{i=1}^n \text{sgn}(\mathbf{u}_i^T\mathbf{y})\mathbf{v}_i$ , and the smoothness of  $\mathbf{x}^\dagger$  increases with increasing  $\beta$  (Appendix C). The parameter  $\omega$  characterizes the decay rate of the singular values and since  $\sigma_i = O(e^{-i})$ , we see that the problem is severely ill-posed. In our simulations we choose  $m = 800$ ,  $n = 20$ ,  $\omega = 0.75$  and  $\beta = 1$ , in which case, the condition number of the matrix is  $1.5 \cdot 10^6$ . The plots in Figure 3.1 show that for  $i \geq p$ , where  $p$  is such that  $F_{0p}^2 \approx \sigma^2$ , the expected Fourier coefficients  $F_i^2$  level off at  $\sigma^2$ , while the expected Picard coefficients  $P_i^2$  have a turning point at  $\sigma_p$ . In fact, we cannot recover the singular value components of the solution for  $i > p$ , because the Picard coefficients are dominated by noise. The plots also show that when  $\sigma$  increases,  $p$  decreases, and so,  $\sigma_p^2$  increases. Thus, larger noise standard deviations require larger regularization parameters.

### 3.5.6 Graphical tools

The residual curve for Tikhonov regularization plays a central role in connection with some regularization parameter choice methods as for example, the discrepancy principle and the residual curve method. Furthermore, the residual and the constraint curves determine the L-curve, which is perhaps the most convenient graphical tool for analyzing discrete ill-posed problems. To account on the random character of the noise in the data, it is appropriate to define the expected curves by averaging over noisy data realizations. In this sections, we use the simplified assumptions (3.70) and (3.71) to obtain qualitative information on the behavior of the expected residual and constraint curves, and to understand the L-shape appearance of the L-curve.

**Residual curve**

The residual vector defined by

$$\mathbf{r}_\alpha^\delta = \mathbf{y}^\delta - \mathbf{K}\mathbf{x}_\alpha^\delta \quad (3.81)$$

possesses the generalized singular value expansion (cf. (3.19))

$$\mathbf{r}_\alpha^\delta = \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i. \quad (3.82)$$

The residual norm then becomes

$$\|\mathbf{r}_\alpha^\delta\|^2 = \sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \quad (3.83)$$

and it is apparent that  $\|\mathbf{r}_\alpha^\delta\|^2$  is an increasing function of  $\alpha$ . An equivalent representation for the residual vector in terms of the influence matrix reads as

$$\mathbf{r}_\alpha^\delta = \mathbf{y}^\delta - \mathbf{K}\mathbf{x}_\alpha^\delta = (\mathbf{I}_m - \mathbf{K}\mathbf{K}_\alpha^\dagger) \mathbf{y}^\delta = (\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) \mathbf{y}^\delta. \quad (3.84)$$

Using (3.73) and the identities

$$\mathbf{u}_i^T \mathbf{y} = 0, \quad i = n+1, \dots, m, \quad (3.85)$$

we find that the expected residual is given by

$$R(\alpha) = \mathcal{E} \left\{ \|\mathbf{r}_\alpha^\delta\|^2 \right\} = (m-n)\sigma^2 + \sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 \left[ (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 \right]. \quad (3.86)$$

To analyze the graph  $(\log \alpha, R(\alpha))$ , we make the change of variable  $x = \log \alpha$  and consider the function

$$R_{\log}(x) = R(\exp(x)) = (m-n)\sigma^2 + \sum_{i=1}^n \left( \frac{e^x}{\gamma_i^2 + e^x} \right)^2 \left[ (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 \right].$$

The slope of the curve is

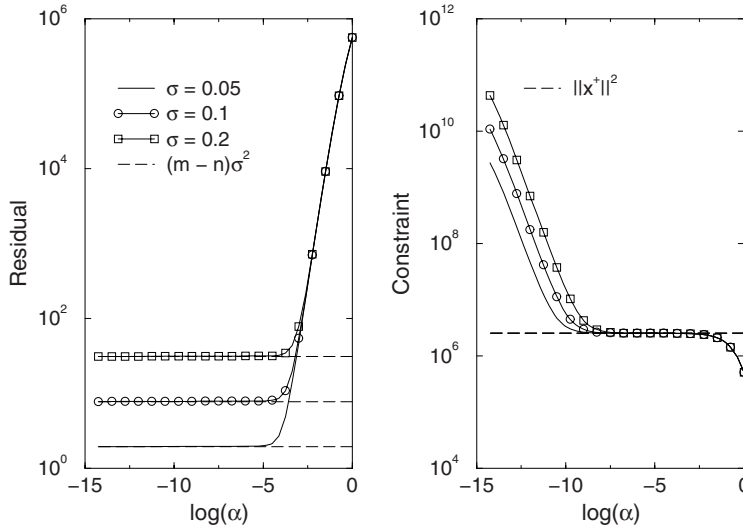
$$R'_{\log}(x) = 2 \sum_{i=1}^n \frac{e^{-x} \gamma_i^2}{(e^{-x} \gamma_i^2 + 1)^3} F_i^2,$$

where  $F_i^2$  are the expected Fourier coefficients (3.73). Setting  $f(t) = t/(1+t)^3$ , with  $t = e^{-x} \gamma_i^2$ , we see that  $f(t) \approx t$  if  $t \ll 1$ , and  $f(t) \approx 1/t^2$  if  $t \gg 1$ . For  $x = x_j = \log(\gamma_j^2)$ ,  $j = 1, \dots, n$ , the filtering property of  $f$  gives

$$R'_{\log}(x_j) \propto F_j^2,$$

and we infer that the slope  $R'_{\log}$  at the discrete points  $x_j$  behaves like the expected Fourier coefficients  $F_j^2$ . From (3.74), it is apparent that the slope of the graph is large for  $j = 1, \dots, p-1$ , and small and constant for  $j = p, \dots, n$ . Supposing that

$$R_{\log}(x_n) = R(\gamma_n^2) \approx \lim_{\alpha \rightarrow 0} R(\alpha) = (m-n)\sigma^2,$$



**Fig. 3.2.** Expected residual (left) and constraint (right) curves for different values of the noise standard deviation  $\sigma$ .

we deduce that  $R_{\log}$  has a plateau at  $(m-n)\sigma^2$  for all  $x_j \leq x_p$  and afterward increases. The plots in the left panel of Figure 3.2 correspond to the synthetic model (3.78)–(3.80) and show that the expected residual is an increasing function of the regularization parameter and has a plateau at  $(m-n)\sigma^2$ .

Let us now assume that the data contains forward model errors and instrumental noise,  $\delta_y = \delta_m + \delta$ , and let us derive an estimate for the equivalent white noise variance (3.63). From (3.83) together with (3.72) and (3.85), we see that

$$\lim_{\alpha \rightarrow 0} \mathcal{E} \left\{ \|\mathbf{r}_\alpha^\delta\|^2 \right\} = \sum_{i=n+1}^m \mathcal{E} \left\{ (\mathbf{u}_i^T \delta_y)^2 \right\} = (m-n)\sigma^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \delta_m)^2.$$

Thus, approximating the expected residual by

$$\mathcal{E} \left\{ \|\mathbf{r}_\alpha^\delta\|^2 \right\} \approx (m-n) \left( \frac{1}{m} \|\delta_m\|^2 + \sigma^2 \right), \quad \alpha \rightarrow 0,$$

we find that the equivalent white noise variance (3.63) can be estimated as

$$\sigma_e^2 \approx \frac{1}{m-n} \mathcal{E} \left\{ \|\mathbf{r}_\alpha^\delta\|^2 \right\} \approx \frac{1}{m-n} \|\mathbf{r}_\alpha^\delta\|^2, \quad \alpha \rightarrow 0.$$

### Constraint curve

The constraint vector is defined as

$$\mathbf{c}_\alpha^\delta = \mathbf{L} \mathbf{x}_\alpha^\delta, \quad (3.87)$$

and we have explicitly

$$\mathbf{c}_\alpha^\delta = \sum_{i=1}^n \frac{\gamma_i}{\gamma_i^2 + \alpha} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i. \quad (3.88)$$

The constraint norm is then given by

$$\|\mathbf{c}_\alpha^\delta\|^2 = \sum_{i=1}^n \left( \frac{\gamma_i}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2, \quad (3.89)$$

and it is readily seen that  $\|\mathbf{c}_\alpha^\delta\|$  is a decreasing function of  $\alpha$ .

We define the expected constraint by

$$C(\alpha) = \mathcal{E} \left\{ \|\mathbf{c}_\alpha^\delta\|^2 \right\} = \sum_{i=1}^n \left( \frac{\gamma_i}{\gamma_i^2 + \alpha} \right)^2 \left[ (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 \right], \quad (3.90)$$

and consider the graph  $(\log \alpha, C(\alpha))$ . As before, we make the change of variable  $x = \log \alpha$ , introduce the function

$$C_{\log}(x) = C(\exp(x)) = \sum_{i=1}^n \left( \frac{\gamma_i}{\gamma_i^2 + e^x} \right)^2 \left[ (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 \right],$$

and compute the slope of the curve as

$$C'_{\log}(x) = -2 \sum_{i=1}^n \frac{(e^{-x} \gamma_i^2)^2}{(e^{-x} \gamma_i^2 + 1)^3} P_i^2,$$

where  $P_i^2$  are the expected Picard coefficients (3.75). The function  $f(t) = t^2/(1+t)^3$ , with  $t = e^{-x} \gamma_i^2$ , behaves like  $f(t) \approx t^2$  if  $t \ll 1$ , and like  $f(t) \approx 1/t$  if  $t \gg 1$ . For  $x = x_j = \log(\gamma_j^2)$ ,  $j = 1, \dots, n$ , the filtering property of  $f$  yields

$$C'_{\log}(x_j) \propto -P_j^2,$$

and we deduce that the slope  $C'_{\log}$  at the discrete points  $x_j$  is reproduced by the expected Picard coefficients  $P_j^2$ . From (3.76), we see that  $|C'_{\log}|$  attains a minimum value at  $j = p$ ; this result together with the inequality  $C'_{\log}(x) < 0$  shows that  $C_{\log}$  is a decreasing function with a plateau in the neighborhood of  $x_p$ .

The plateau of the expected constraint curve appears approximately at

$$C_{\log}(x_p) = C(\gamma_p^2) \approx \|\mathbf{Lx}^\dagger\|^2.$$

To justify this claim, we consider the representation (cf. (3.18))

$$\|\mathbf{Lx}^\dagger\|^2 = \sum_{i=1}^n \frac{1}{\gamma_i^2} (\mathbf{u}_i^T \mathbf{y})^2 = C^2 \left[ \gamma_p^{2\beta} + \sum_{i=1}^{p-1} \gamma_i^{2\beta} + \sum_{i=p+1}^n \gamma_i^{2\beta} \right],$$

and use (3.71) and (3.73) to express (3.90) as

$$C(\gamma_p^2) \approx C^2 \left[ \frac{1}{2} \gamma_p^{2\beta} + \sum_{i=1}^{p-1} \gamma_i^{2\beta} + \gamma_p^{2\beta} \sum_{i=p+1}^n \left( \frac{\gamma_i}{\gamma_p} \right)^2 \right].$$

Hence, neglecting the contribution of all the terms  $\gamma_i$  with  $i \geq p$ , we conclude that

$$C(\gamma_p^2) \approx \|\mathbf{L}\mathbf{x}^\dagger\|^2 \approx C^2 \sum_{i=1}^{p-1} \gamma_i^{2\beta}.$$

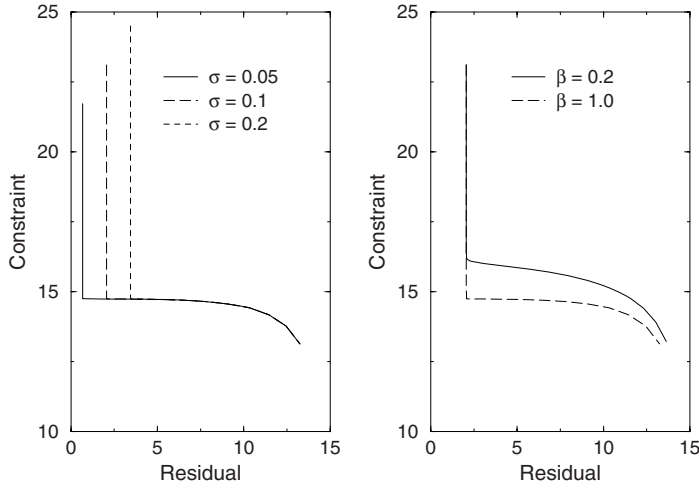
The plots in the right panel of Figure 3.2 illustrate that the expected constraint is a decreasing function with a plateau at  $\|\mathbf{L}\mathbf{x}^\dagger\|^2$ .

### *L-curve*

The L-curve is the plot of the constraint  $\|\mathbf{c}_\alpha^\delta\|^2$  against the residual  $\|\mathbf{r}_\alpha^\delta\|^2$  for a range of values of the regularization parameter  $\alpha$ . The use of such plots for ill-posed problems goes back to Miller (1970), and Lawson and Hanson (1995). The properties of the L-curve in connection with the design of a regularization parameter choice method for linear ill-posed problems have been discussed by Hansen (1992a) and Hansen and O’Leary (1993), and can also be found in Reginska (1996).

When this curve is plotted in log-log scale it has a characteristic L-shape appearance with a distinct corner separating the vertical and the horizontal parts of the curve. To understand the characteristic shape of this curve, we consider the expected L-curve, which is the plot of the expected constraint  $C(\alpha) = \mathcal{E}\{\|\mathbf{c}_\alpha^\delta\|^2\}$  versus the expected residual  $R(\alpha) = \mathcal{E}\{\|\mathbf{r}_\alpha^\delta\|^2\}$ . As we saw before, for small values of the regularization parameters the expected residual curve has a plateau at  $(m - n)\sigma^2$  and after that increases, while for large values of the regularization parameters, the expected constraint curve has a plateau at  $\|\mathbf{L}\mathbf{x}^\dagger\|^2$ . Thus, for small values of the regularization parameters, the L-curve has a vertical part where  $C(\alpha)$  is very sensitive to changes in  $\alpha$ . For large values of the regularization parameters, the L-curve has a horizontal part where  $R(\alpha)$  is most sensitive to  $\alpha$ . If we neglect the forward model errors, the corner of the L-curve appears approximately at  $((m - n)\sigma^2, \|\mathbf{L}\mathbf{x}^\dagger\|^2)$ . Typical expected L-curves are illustrated in Figure 3.3. From the left panel it is apparent that when  $\sigma$  increases, the vertical part of the L-curve moves towards larger  $R$  values. The plots in the right panel show that the faster the Fourier coefficients decay to zero, the smaller the cross-over region between the vertical and horizontal part and, thus, the sharper the L-shaped corner.

The L-curve divides the first quadrant into two regions and any regularized solution must lie on or above this curve (Hansen, 1998). When very little regularization is introduced, the total error is dominated by the noise error. This situation is called under-smoothing, and it corresponds to the vertical part of the L-curve. When a large amount of regularization is introduced, then the total error is dominated by the smoothing error. This situation is called over-smoothing and it corresponds to the horizontal part of the L-curve. For this reason, we may conclude that an optimal regularization parameter balancing the smoothing and noise errors is not so far from the regularization parameter that corresponds to the corner of the L-curve.



**Fig. 3.3.** Expected L-curves for the synthetic model (3.78)–(3.80) with  $m = 800$ ,  $n = 20$  and  $\omega = 0.75$ . The plots in the left panel correspond to  $\beta = 1$  and different values of the noise standard deviation  $\sigma$ , while the plots in the right panel correspond to  $\sigma = 0.1$  and two values of the smoothness parameter  $\beta$ .

### 3.6 Regularization parameter choice methods

The computation of a good approximation  $\mathbf{x}_\alpha^\delta$  of  $\mathbf{x}^\dagger$  depends on the selection of the regularization parameter  $\alpha$ . With too little regularization, reconstructions have highly oscillatory artifacts due to noise amplification. With too much regularization, the reconstructions are too smooth. Ideally, we would like to select a regularization parameter so that the corresponding regularized solution minimizes some indicator of solution fidelity, e.g., some measure of the size of the solution error.

When reliable information about the instrumental noise is available, it is important to make use of this information, and this is the heart of the discrepancy principle and related methods. When no particular information about the instrumental noise is available or when forward model errors are present, the so-called error-free parameter choice methods are a viable alternative.

The formulations of regularization parameter choice methods in deterministic and semi-stochastic settings are very similar. The reason is that the noise level  $\Delta$ , which represents an upper bound for the data error norm  $\|\mathbf{y}^\delta - \mathbf{y}\|$ , can be estimated as  $\Delta^2 = m\sigma^2$ , and this estimate can be used to reformulate a ‘deterministic’ parameter choice method in a semi-stochastic setting. According to the standard deterministic classification (Engl et al., 2000)

- (1) a regularization parameter choice method depending only on  $\Delta$ ,  $\alpha = \alpha(\Delta)$ , is called an a priori parameter choice method;
- (2) a regularization parameter choice method depending on  $\Delta$  and  $\mathbf{y}^\delta$ ,  $\alpha = \alpha(\Delta, \mathbf{y}^\delta)$ , is called an a posteriori parameter choice method;
- (3) a regularization parameter choice method depending only on  $\mathbf{y}^\delta$ ,  $\alpha = \alpha(\mathbf{y}^\delta)$ , is called an error-free parameter choice method.

In this section we review the main regularization parameter choice methods encountered in the literature and compare their efficiency by performing a numerical analysis in a semi-stochastic setting. A deterministic analysis of a priori, a posteriori and error-free parameter choice methods is outlined in Appendix C.

For our numerical simulations we consider the synthetic model (3.78)–(3.80), which is very similar to that considered by Vogel (2002) for analyzing regularization parameter choice methods in a semi-stochastic setting.

### 3.6.1 A priori parameter choice methods

In a deterministic setting, an a priori parameter choice method is of the form  $\alpha \propto \Delta^p$  (Engl et al., 2000; Vogel, 2002; Rieder, 2003), while in a semi-stochastic setting, this selection rule translates into the choice  $\alpha \propto \sigma^p$ . In the next chapter we will see that in the framework of a statistical Bayesian model, the maximum a posteriori estimator is characterized by the a priori selection criterion  $\alpha \propto \sigma^2$ .

In a semi-stochastic setting, we define the optimal regularization parameter for error estimation as the minimizer of the expected error,

$$\bar{\alpha}_{\text{opt}} = \arg \min_{\alpha} \mathcal{E} \left\{ \left\| \mathbf{e}_{\alpha}^{\delta} \right\|^2 \right\}, \quad (3.91)$$

where  $\mathcal{E} \left\{ \left\| \mathbf{e}_{\alpha}^{\delta} \right\|^2 \right\}$  is given by (3.40) together with (3.33) and (3.41). The optimal regularization parameter is not a computable quantity, because the exact solution is unknown, but we may design an a priori parameter choice method by combining this selection criterion with a Monte Carlo technique. The steps of the so-called expected error estimation method can be synthesized as follows:

- (1) perform a random exploration of a domain, in which the exact solution is supposed to lie, by considering a set of state vector realizations  $\{\mathbf{x}_i^{\dagger}\}_{i=1, \dots, N_x}$ , where  $N_x$  is the sample size;
- (2) for each  $\mathbf{x}_i^{\dagger}$ , compute the optimal regularization parameter for error estimation

$$\bar{\alpha}_{\text{opt}i} = \arg \min_{\alpha} \mathcal{E} \left\{ \left\| \mathbf{e}_{\alpha}^{\delta} \left( \mathbf{x}_i^{\dagger} \right) \right\|^2 \right\},$$

and determine the exponent

$$p_i = \frac{\log \bar{\alpha}_{\text{opt}i}}{\log \sigma};$$

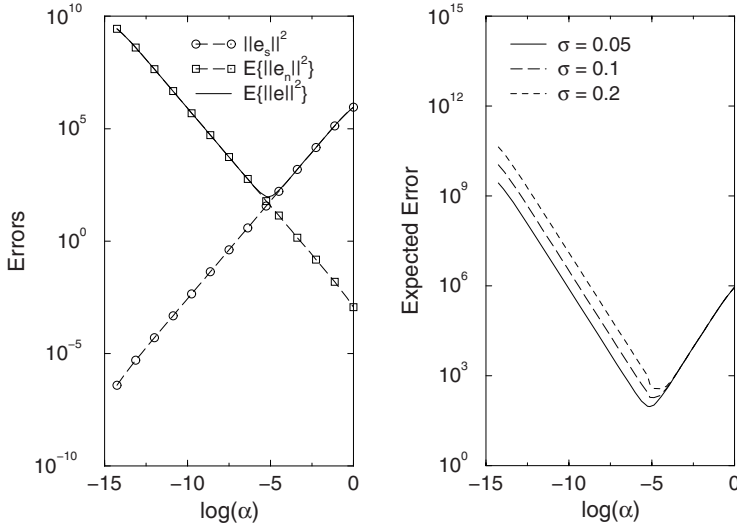
- (3) compute the sample mean exponent

$$\bar{p} = \frac{1}{N_x} \sum_{i=1}^{N_x} p_i;$$

- (4) choose the regularization parameter as  $\alpha_{\mathbf{e}} = \sigma^{\bar{p}}$ .

The idea of the expected error estimation method is very simple; the main problem which has to be solved is the choice of the solution domain. Essentially,  $\{\mathbf{x}_i^{\dagger}\}$  should be a set





**Fig. 3.4.** Left: expected error together with the smoothing and noise errors. Right: expected error for three values of the noise standard deviation  $\sigma$ .

of solutions with physical meaning and stochastic a priori information can be used for an appropriate construction. Assuming that  $\mathbf{x}^\dagger$  is a Gaussian random vector with zero mean and covariance  $\mathbf{C}_x$ , the random exploration of the solution domain is a sampling of the (a priori) probability density. The sampling of a Gaussian probability density is standard and involves the following steps (Bard, 1974):

- (1) given the a priori profile  $\mathbf{x}_a$ , choose the correlation length  $l$  and the profile standard deviation  $\sigma_x$ , and set  $\mathbf{C}_x = \sigma_x^2 \mathbf{C}_{nx}$ , where  $\mathbf{C}_{nx}$  is the normalized covariance matrix defined by

$$[\mathbf{C}_{nx}]_{ij} = [\mathbf{x}_a]_i [\mathbf{x}_a]_j \exp\left(-\frac{|z_i - z_j|}{l}\right), \quad i, j = 1, \dots, n;$$

- (2) compute the SVD of the positive definite matrix  $\mathbf{C}_{nx} = \mathbf{V}_x \Sigma_x \mathbf{V}_x^T$ ;
- (3) generate a random realization  $\mathbf{x}$  of a Gaussian process with zero mean and unit covariance  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ ;
- (4) compute the profile deviation as  $\mathbf{x}^\dagger = \sigma_x \mathbf{V}_x \Sigma_x^{1/2} \mathbf{x}$ .

In Figure 3.4 we plot the expected error  $\mathcal{E}\{\|\mathbf{e}_\alpha^\delta\|^2\}$ . As  $\|\mathbf{e}_{s\alpha}\|^2$  is an increasing function of  $\alpha$  and  $\mathcal{E}\{\|\mathbf{e}_{n\alpha}^\delta\|^2\}$  is a decreasing function of  $\alpha$ ,  $\mathcal{E}\{\|\mathbf{e}_\alpha^\delta\|^2\}$  possesses a minimum. The minimizer of the expected error increases with increasing the noise variance and this behavior is apparent from the right panel of Figure 3.4.

### 3.6.2 A posteriori parameter choice methods

The a posteriori parameter choice methods to be discussed in this section are the discrepancy principle, the generalized discrepancy principle (or the minimum bound method), the

error consistency method, and the unbiased predictive risk estimator method. The first two regularization parameter choice methods can be formulated in deterministic and semi-stochastic settings, while the last two methods make only use of statistical information about the noise in the data.

### *Discrepancy principle*

The most popular a posteriori parameter choice method is the discrepancy principle due to Morozov (1966, 1968). In this method, the regularization parameter is chosen via a comparison between the residual norm (discrepancy)  $\|\mathbf{r}_\alpha^\delta\|$  and the assumed noise level  $\Delta$ ,

$$\|\mathbf{r}_\alpha^\delta\|^2 = \tau \Delta^2, \quad \tau > 1. \quad (3.92)$$

A heuristic motivation for this method is that as long as we have only the noisy data vector  $\mathbf{y}^\delta$  and know that  $\|\mathbf{y}^\delta - \mathbf{y}\| \leq \Delta$ , it does not make sense to ask for an approximate solution  $\mathbf{x}_\alpha^\delta$  with a discrepancy  $\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_\alpha^\delta\| < \Delta$ ; a residual norm in the order of  $\Delta$  is the best we should ask for. In a semi-stochastic setting, the discrepancy principle selects the regularization parameter as the solution of the equation

$$\|\mathbf{r}_\alpha^\delta\|^2 = \tau m \sigma^2, \quad (3.93)$$

which, in terms of a generalized singular system of  $(\mathbf{K}, \mathbf{L})$ , takes the form

$$\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = \tau m \sigma^2, \quad (3.94)$$

with the convention  $\gamma_i = 0$  for  $i = n + 1, \dots, m$ .

### *Generalized discrepancy principle*

In some applications, the discrepancy principle gives a too small regularization parameter and the solution is undersmoothed. An improved variant of the discrepancy principle is the generalized discrepancy principle, which has been considered by Raus (1985) and Gfrerer (1987) in a deterministic setting, and by Lukas (1998b) in a discrete, semi-stochastic setting. For a more general analysis of this regularization parameter choice method we refer to Engl and Gfrerer (1988).

In the generalized version of the discrepancy principle, the regularization parameter is the solution of the equation

$$\|\mathbf{r}_\alpha^\delta\|^2 - \mathbf{r}_\alpha^{\delta T} \hat{\mathbf{A}}_\alpha \mathbf{r}_\alpha^\delta = \tau \Delta^2, \quad \tau > 1. \quad (3.95)$$

As  $\hat{\mathbf{A}}_\alpha$  is positive definite, the left-hand side of this equation is smaller than the residual  $\|\mathbf{r}_\alpha^\delta\|^2$ , and therefore, the regularization parameter computed by the generalized discrepancy principle is larger than the regularization parameter corresponding to the ordinary method. In a semi-stochastic setting, the generalized discrepancy principle seeks the regularization parameter  $\alpha$  solving the equation

$$\|\mathbf{r}_\alpha^\delta\|^2 - \mathbf{r}_\alpha^{\delta T} \hat{\mathbf{A}}_\alpha \mathbf{r}_\alpha^\delta = \tau m \sigma^2. \quad (3.96)$$

Using the relation (cf. (3.84))

$$\|\mathbf{r}_\alpha^\delta\|^2 - \mathbf{r}_\alpha^{\delta T} \hat{\mathbf{A}}_\alpha \mathbf{r}_\alpha^\delta = \mathbf{y}^{\delta T} \left( \mathbf{I}_m - \hat{\mathbf{A}}_\alpha \right)^3 \mathbf{y}^\delta,$$

and the factorization (3.51), we express (3.96) in explicit form as

$$\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^3 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = \tau m \sigma^2, \quad (3.97)$$

with  $\gamma_i = 0$  for  $i = n + 1, \dots, m$ . The difference to the conventional method (compare to (3.94)) is that the factors multiplying the Fourier coefficients  $\mathbf{u}_i^T \mathbf{y}^\delta$  converge more rapidly to zero as  $\alpha$  tends to zero.

An equivalent representation of the generalized discrepancy principle equation is based on the identity

$$\mathbf{I}_m - \hat{\mathbf{A}}_\alpha = \alpha \left[ \mathbf{K} (\mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T + \alpha \mathbf{I}_m \right]^{-1},$$

and is given by

$$\alpha \mathbf{r}_\alpha^{\delta T} \left[ \mathbf{K} (\mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T + \alpha \mathbf{I}_m \right]^{-1} \mathbf{r}_\alpha^\delta = \tau m \sigma^2. \quad (3.98)$$

The generalized discrepancy principle equation can also be formulated in terms of the solution of iterated Tikhonov regularization. In the two-times iterated Tikhonov regularization we compute the improved solution step

$$\mathbf{p}_{\alpha 2}^\delta = \mathbf{K}_\alpha^\dagger (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_\alpha^\delta) = \mathbf{K}_\alpha^\dagger \mathbf{r}_\alpha^\delta,$$

where  $\mathbf{x}_\alpha^\delta = \mathbf{x}_{\alpha 1}^\delta$ , and set  $\mathbf{x}_{\alpha 2}^\delta = \mathbf{x}_\alpha^\delta + \mathbf{p}_{\alpha 2}^\delta$ . Since

$$\mathbf{r}_{\alpha 2}^\delta = \mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{\alpha 2}^\delta = \mathbf{y}^\delta - \mathbf{K} (\mathbf{x}_\alpha^\delta + \mathbf{p}_{\alpha 2}^\delta) = \left( \mathbf{I}_m - \hat{\mathbf{A}}_\alpha \right) \mathbf{r}_\alpha^\delta,$$

we find that

$$\|\mathbf{r}_\alpha^\delta\|^2 - \mathbf{r}_\alpha^{\delta T} \hat{\mathbf{A}}_\alpha \mathbf{r}_\alpha^\delta = \mathbf{r}_\alpha^{\delta T} \left( \mathbf{I}_m - \hat{\mathbf{A}}_\alpha \right) \mathbf{r}_\alpha^\delta = \mathbf{r}_\alpha^{\delta T} \mathbf{r}_{\alpha 2}^\delta.$$

Thus, in terms of the residual at the iterated Tikhonov solution, the generalized discrepancy principle equation takes the form

$$\mathbf{r}_\alpha^{\delta T} \mathbf{r}_{\alpha 2}^\delta = \tau m \sigma^2.$$

The generalized discrepancy principle is equivalent to the so-called minimum bound method. To give a heuristic justification of this equivalence in a deterministic setting and for the choice  $\mathbf{L} = \mathbf{I}_n$ , we consider the error estimate

$$\|\mathbf{e}_\alpha^\delta\|^2 \leq 2 \left( \|\mathbf{e}_{s\alpha}\|^2 + \|\mathbf{e}_{n\alpha}^\delta\|^2 \right).$$

In (3.37) we then employ the inequality

$$\frac{\sigma_i}{\sigma_i^2 + \alpha} \leq \frac{1}{2\sqrt{\alpha}} < \sqrt{\frac{2\tau}{\alpha}}, \quad \tau > 1,$$

and obtain the noise error estimate

$$\|\mathbf{e}_{n\alpha}^\delta\|^2 < \frac{2\tau\Delta^2}{\alpha}; \quad (3.99)$$

this result together with (3.33) yields the following bound for the total error

$$M(\alpha) = 2 \left[ \sum_{i=1}^n \left( \frac{\alpha}{\sigma_i^2 + \alpha} \frac{1}{\sigma_i} \right)^2 (\mathbf{u}_i^T \mathbf{y})^2 + 2\tau \frac{\Delta^2}{\alpha} \right]. \quad (3.100)$$

The regularization parameter of the minimum bound method minimizes  $M(\alpha)$ , whence setting  $M'(\alpha) = 0$ , we obtain the equation

$$\sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^3 (\mathbf{u}_i^T \mathbf{y})^2 = \tau \Delta^2. \quad (3.101)$$

As  $\mathbf{u}_i^T \mathbf{y} = 0$  for  $i = n+1, \dots, m$ , the upper limit of summation in (3.101) can be extended to  $m$ , and in order to obtain an implementable algorithm, we replace  $\mathbf{y}$  by  $\mathbf{y}^\delta$ . The resulting equation is (3.97) with  $\Delta^2$  in place of  $m\sigma^2$ , and the equivalence is proven.

### **Error consistency method**

The error consistency method has been proposed by Ceccherini (2005) and has been successfully applied for MIPAS near-real time data processing. In this method, we impose that the differences between the regularized and the least squares solutions  $\mathbf{x}_\alpha^\delta$  and  $\mathbf{x}^\delta$ , respectively, are on average equal to the error in the least squares solution

$$(\mathbf{x}_\alpha^\delta - \mathbf{x}^\delta)^T \mathbf{C}_e^{-1} (\mathbf{x}_\alpha^\delta - \mathbf{x}^\delta) = n. \quad (3.102)$$

The error in the least squares solution is due to the instrumental noise,

$$\mathbf{e}^\delta = \mathbf{x}^\dagger - \mathbf{x}^\delta = \mathbf{K}^\dagger \mathbf{y} - \mathbf{K}^\dagger \mathbf{y}^\delta = -\mathbf{K}^\dagger \delta,$$

and since  $\mathbf{K}^\dagger = (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T$ , we see that

$$\mathbf{C}_e = \mathcal{E} \{ \mathbf{e}^\delta \mathbf{e}^{\delta T} \} = \sigma^2 \mathbf{K}^\dagger \mathbf{K}^{\dagger T} = \sigma^2 (\mathbf{K}^T \mathbf{K})^{-1}.$$

Using the representation  $\mathbf{x}_\alpha^\delta - \mathbf{x}^\delta = (\mathbf{K}_\alpha^\dagger - \mathbf{K}^\dagger) \mathbf{y}^\delta$ , equation (3.102) becomes

$$\|\mathbf{K} (\mathbf{K}_\alpha^\dagger - \mathbf{K}^\dagger) \mathbf{y}^\delta\|^2 = n\sigma^2,$$

or explicitly,

$$\sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = n\sigma^2.$$

The expected equation of the error consistency method,

$$\mathcal{E} \left\{ \|\mathbf{K} (\mathbf{K}_\alpha^\dagger - \mathbf{K}^\dagger) \mathbf{y}^\delta\|^2 \right\} = n\sigma^2,$$

is identical to the expected equation of the discrepancy principle  $\mathcal{E}\{\|\mathbf{r}_\alpha^\delta\|^2\} = m\sigma^2$  with  $\tau = 1$ , that is, (cf. (3.73) and (3.86))

$$\sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 \left[ (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 \right] = n\sigma^2.$$

For this reason, we anticipate that the regularization parameter of the error consistency method is smaller than the regularization parameter of the discrepancy principle with  $\tau > 1$ .

### ***Unbiased predictive risk estimator method***

The computation of the regularization parameter by analyzing the solution error is not practical, since the exact solution is unknown. Instead, the predictive error can be used as an indicator of the solution fidelity, because it can be accurately estimated in the framework of the unbiased predictive risk estimator method. This approach is also known as the  $C_L$ -method or the predictive mean square error method and was originally developed by Mallows (1973) for model selection in linear regression. For further readings related to the use of the predictive risk as a criterion for choosing the regularization parameter we refer to Golub et al. (1979) and Rice (1986).

In a semi-stochastic setting, the expected value of the predictive error  $\mathcal{E}\{\|\mathbf{p}_\alpha^\delta\|^2\}$  is given by (3.56) together with (3.57) and (3.58). The predictive risk estimator is defined through the relation

$$\pi_\alpha^\delta = \|\mathbf{r}_\alpha^\delta\|^2 + 2\sigma^2 \text{trace}(\hat{\mathbf{A}}_\alpha) - m\sigma^2,$$

and in order to compute its expected value, we write (3.84) as

$$\mathbf{r}_\alpha^\delta = (\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) (\mathbf{y} + \boldsymbol{\delta}),$$

and use the trace lemma (3.25) to obtain

$$\mathcal{E}\{\|\mathbf{r}_\alpha^\delta\|^2\} = \left\| (\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) \mathbf{y} \right\|^2 + \sigma^2 \text{trace}(\hat{\mathbf{A}}_\alpha^T \hat{\mathbf{A}}_\alpha) - 2\sigma^2 \text{trace}(\hat{\mathbf{A}}_\alpha) + m\sigma^2. \quad (3.103)$$

Consequently, we find that

$$\mathcal{E}\{\pi_\alpha^\delta\} = \left\| (\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) \mathbf{y} \right\|^2 + \sigma^2 \text{trace}(\hat{\mathbf{A}}_\alpha^T \hat{\mathbf{A}}_\alpha), \quad (3.104)$$

and by (3.56)–(3.58), we deduce that  $\pi_\alpha^\delta$  is an unbiased estimator for the expected value of the predictive error, that is,

$$\mathcal{E}\{\pi_\alpha^\delta\} = \mathcal{E}\{\|\mathbf{p}_\alpha^\delta\|^2\}.$$

The unbiased predictive risk estimator method chooses the regularization parameter as

$$\alpha_{\text{pr}} = \arg \min_{\alpha} \pi_\alpha^\delta,$$

and, in view of (3.51), (3.52) and (3.83), a computable expression for  $\pi_\alpha^\delta$  reads as

$$\pi_\alpha^\delta = \sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + 2\sigma^2 \sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha} - m\sigma^2,$$

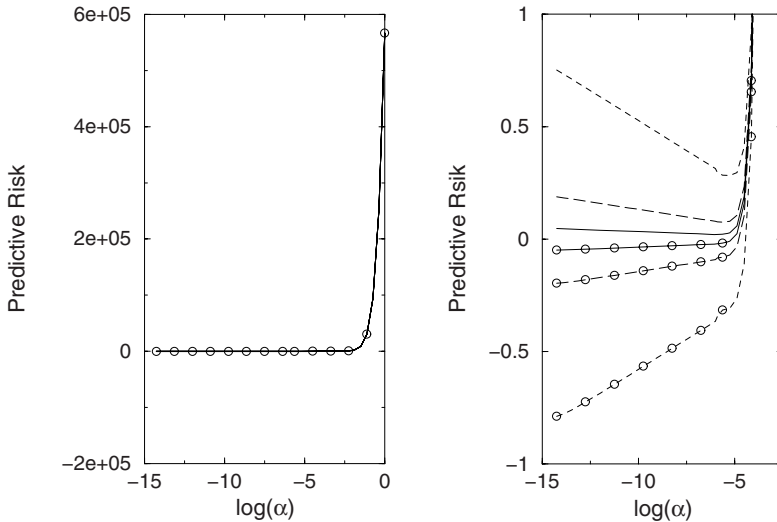
with the standard convention  $\gamma_i = 0$  for  $i = n + 1, \dots, m$ .

Although they have the same expected values it does not necessarily follow that  $\pi_\alpha^\delta$  and  $\|\mathbf{p}_\alpha^\delta\|^2$  have the same minimizers (Vogel, 2002). However, the analysis performed by Lukas (1998a) has shown that these minimizers are close provided that these functions do not have flat minima.

The predictive risk estimator possesses a minimum since  $\|\mathbf{r}_\alpha^\delta\|^2$  is an increasing function of  $\alpha$  and  $\text{trace}(\hat{\mathbf{A}}_\alpha)$  is a decreasing function of  $\alpha$ . However, this minimum can be very flat especially when the trace term is very small as compared to the residual term. For large values of  $\alpha$ ,  $\text{trace}(\hat{\mathbf{A}}_\alpha)$  is very small and the expected predictive risk is reproduced by the expected residual,

$$\mathcal{E}\{\pi_\alpha^\delta\} \approx \mathcal{E}\{\|\mathbf{r}_\alpha^\delta\|^2\} - m\sigma^2, \quad \alpha \rightarrow \infty.$$

In Figure 3.5 we show the expected predictive risk together with its asymptotical approximation. The plots illustrate that  $\mathcal{E}\{\|\mathbf{r}_\alpha^\delta\|^2\} - m\sigma^2$  is a reasonable approximation of  $\mathcal{E}\{\pi_\alpha^\delta\}$  for large values of the regularization parameter  $\alpha$  and small values of the noise standard deviation  $\sigma$ . The expected predictive risk has a flat minimum which moves toward large  $\alpha$  with increasing  $\sigma$  and the flatness of the curves becomes more pronounced as  $\sigma$  decreases.



**Fig. 3.5.** Expected predictive risk and its asymptotical approximation. In the left panel, the curves are plotted over the entire domain of variation of  $\alpha$ , while in the right panel, the  $y$ -axis is zoomed out. The results correspond to  $\sigma = 0.05$  (solid line),  $\sigma = 0.1$  (long dashed line) and  $\sigma = 0.2$  (dashed line). The approximations are marked with circles.

### 3.6.3 Error-free parameter choice methods

Error-free parameter choice methods do not take into account information about the errors in the data and for this reason, these methods do not depend on the setting in which the problem is treated.

#### *Generalized cross-validation*

The generalized cross-validation method is an alternative to the unbiased predictive risk estimator method that does not require the knowledge of the noise variance  $\sigma^2$ . This method was developed by Wahba (1977, 1990) and is a very popular and successful error-free method for choosing the regularization parameter.

The generalized cross-validation function can be derived from the ‘leaving-out-one’ principle (Wahba, 1990). In the ordinary or the ‘leaving-out-one’ cross-validation, we consider models that are obtained by leaving one of the  $m$  data points out of the inversion process. Denoting by  $\mathbf{K}_{(k)}$  the  $(m-1) \times n$  matrix obtained by deleting the  $k$ th row of  $\mathbf{K}$ , and by  $\mathbf{y}_{(k)}^\delta$  the  $(m-1)$ -dimensional vector obtained by deleting the  $k$ th entry of  $\mathbf{y}^\delta$ , we compute  $\mathbf{x}_{(k)\alpha}^\delta$  as the minimizer of the function

$$\mathcal{F}_{(k)\alpha}(\mathbf{x}) = \left\| \mathbf{y}_{(k)}^\delta - \mathbf{K}_{(k)}\mathbf{x} \right\|^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2, \quad (3.105)$$

with

$$\left\| \mathbf{y}_{(k)}^\delta - \mathbf{K}_{(k)}\mathbf{x} \right\|^2 = \sum_{i=1, i \neq k}^m ([\mathbf{y}^\delta]_i - [\mathbf{K}\mathbf{x}]_i)^2.$$

For an appropriate choice of the regularization parameter, the solution  $\mathbf{x}_{(k)\alpha}^\delta$  should accurately predict the missing data value  $[\mathbf{y}^\delta]_k$ . Essentially, the regularization parameter  $\alpha$  is chosen so that on average  $\mathbf{y}^\delta$  and  $\mathbf{K}\mathbf{x}_{(k)\alpha}^\delta$  are very close for all  $k$ , that is,

$$\alpha_{cv} = \arg \min_{\alpha} V_{\alpha},$$

where the ordinary cross-validation function  $V_{\alpha}$  is given by

$$V_{\alpha} = \sum_{k=1}^m \left( [\mathbf{y}^\delta]_k - [\mathbf{K}\mathbf{x}_{(k)\alpha}^\delta]_k \right)^2. \quad (3.106)$$

To compute  $V_{\alpha}$ , we have to solve  $m$  problems of the form (3.105) and this is a very expensive task. The computation can be simplified by defining the modified data vector  $\mathbf{y}_k^\delta$ ,

$$[\mathbf{y}_k^\delta]_i = \begin{cases} [\mathbf{K}\mathbf{x}_{(k)\alpha}^\delta]_k, & i = k, \\ [\mathbf{y}^\delta]_i, & i \neq k, \end{cases} \quad (3.107)$$

which coincides with  $\mathbf{y}^\delta$  except for the  $k$ th component. As  $[\mathbf{y}_k^\delta]_k = [\mathbf{K}\mathbf{x}_{(k)\alpha}^\delta]_k$ , we observe

that  $\mathbf{x}_{(k)\alpha}^\delta$  also minimizes the function

$$\begin{aligned}\tilde{\mathcal{F}}_{(k)\alpha}(\mathbf{x}) &= ([\mathbf{y}_k^\delta]_k - [\mathbf{K}\mathbf{x}]_k)^2 + \mathcal{F}_{(k)\alpha}(\mathbf{x}) \\ &= ([\mathbf{y}_k^\delta]_k - [\mathbf{K}\mathbf{x}]_k)^2 + \sum_{i=1, i \neq k}^m ([\mathbf{y}_k^\delta]_i - [\mathbf{K}\mathbf{x}]_i)^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2 \\ &= \|\mathbf{y}_k^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2,\end{aligned}$$

and so, that  $\mathbf{x}_{(k)\alpha}^\delta = \mathbf{K}_\alpha^\dagger \mathbf{y}_k^\delta$ . This result, which allows us to express  $\mathbf{x}_{(k)\alpha}^\delta$  in terms of the regularized generalized inverse  $\mathbf{K}_\alpha^\dagger$  and the modified data vector  $\mathbf{y}_k^\delta$ , is known as the ‘leaving-out-one’ lemma. To eliminate  $\mathbf{x}_{(k)\alpha}^\delta$  in the expression of the ordinary cross-validation function, we express  $V_\alpha$  as

$$V_\alpha = \sum_{k=1}^m \left( \frac{[\mathbf{K}\mathbf{x}_\alpha^\delta]_k - [\mathbf{y}^\delta]_k}{1 - a_k} \right)^2, \quad (3.108)$$

with

$$a_k = \frac{[\mathbf{K}\mathbf{x}_{(k)\alpha}^\delta]_k - [\mathbf{K}\mathbf{x}_\alpha^\delta]_k}{[\mathbf{K}\mathbf{x}_{(k)\alpha}^\delta]_k - [\mathbf{y}^\delta]_k}.$$

By the ‘leaving-out-one’ lemma we have  $[\mathbf{K}\mathbf{x}_{(k)\alpha}^\delta]_k = [\mathbf{K}\mathbf{K}_\alpha^\dagger \mathbf{y}_k^\delta]_k$ , whence using the identity  $[\mathbf{K}\mathbf{x}_\alpha^\delta]_k = [\mathbf{K}\mathbf{K}_\alpha^\dagger \mathbf{y}^\delta]_k$ , and replacing the divided difference by a derivative, we obtain

$$a_k = \frac{[\mathbf{K}\mathbf{K}_\alpha^\dagger \mathbf{y}_k^\delta]_k - [\mathbf{K}\mathbf{K}_\alpha^\dagger \mathbf{y}^\delta]_k}{[\mathbf{y}_k^\delta]_k - [\mathbf{y}^\delta]_k} \approx \frac{\partial [\mathbf{K}\mathbf{K}_\alpha^\dagger \mathbf{y}^\delta]_k}{\partial [\mathbf{y}^\delta]_k} = [\mathbf{K}\mathbf{K}_\alpha^\dagger]_{kk}.$$

Taking into account that  $\hat{\mathbf{A}}_\alpha = \mathbf{K}\mathbf{K}_\alpha^\dagger$ , and approximating  $[\hat{\mathbf{A}}_\alpha]_{kk}$  by the average value

$$[\hat{\mathbf{A}}_\alpha]_{kk} \approx \frac{1}{m} \text{trace}(\hat{\mathbf{A}}_\alpha),$$

we find that

$$V_\alpha \approx \frac{\sum_{k=1}^m ([\mathbf{K}\mathbf{x}_\alpha^\delta]_k - [\mathbf{y}^\delta]_k)^2}{\left[1 - \frac{1}{m} \text{trace}(\hat{\mathbf{A}}_\alpha)\right]^2} = m^2 \frac{\|\mathbf{r}_\alpha^\delta\|^2}{[\text{trace}(\mathbf{I}_m - \hat{\mathbf{A}}_\alpha)]^2}. \quad (3.109)$$

Thus, in the framework of the generalized cross-validation method, we select the regularization parameter as

$$\alpha_{\text{gcv}} = \arg \min_{\alpha} v_\alpha^\delta,$$

where  $v_\alpha^\delta$  is the generalized cross-validation function (3.109) without the factor  $m^2$ ,

$$v_\alpha^\delta = \frac{\|\mathbf{r}_\alpha^\delta\|^2}{[\text{trace}(\mathbf{I}_m - \hat{\mathbf{A}}_\alpha)]^2}.$$



To obtain an implementable algorithm, we compute  $\|\mathbf{r}_\alpha^\delta\|^2$  according to (3.83) and the trace term by using the relation (cf. (3.51) and (3.52))

$$\text{trace}(\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) = \text{trace}\left(\mathbf{U}\left(\mathbf{I}_m - \hat{\Sigma}_\alpha\right)\mathbf{U}^T\right) = m - n + \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha}. \quad (3.110)$$

It should be pointed out that in statistical inversion theory, the trace term can be viewed as a measure of the degree of freedom for noise in the regularized solution.

Essentially, the generalized cross-validation method seeks to locate the transition point where the residual norm changes from a very slowly varying function of  $\alpha$  to a rapidly increasing function of  $\alpha$ . But instead of working with the residual norm, the generalized cross-validation method uses the ratio of the residual norm and the degree of freedom for noise, which is a monotonically increasing function of  $\alpha$ . As the residual norm is also an increasing function of  $\alpha$ , the generalized cross-validation function has a minimum.

Wahba (1977) showed that if the discrete Picard condition is satisfied, then the minima of the expected generalized cross-validation function and the expected predictive risk are very close. More precisely, if

$$\bar{\alpha}_{\text{gcv}} = \arg \min_{\alpha} \mathcal{E}\{v_\alpha^\delta\},$$

and

$$\bar{\alpha}_{\text{pr}} = \arg \min_{\alpha} \mathcal{E}\{\pi_\alpha^\delta\},$$

then  $\bar{\alpha}_{\text{gcv}}$  is asymptotically equal to  $\bar{\alpha}_{\text{pr}}$  as  $m \rightarrow \infty$ . This result was further examined and extended by Lukas (1998a) and can also be found in Vogel (2002). To reveal the connection between these two methods, we consider the expected value of the generalized cross-validation function (cf. (3.103) and (3.104)),

$$\mathcal{E}\{v_\alpha^\delta\} = \frac{\mathcal{E}\{\|\mathbf{r}_\alpha^\delta\|^2\}}{\left[m - \text{trace}(\hat{\mathbf{A}}_\alpha)\right]^2} = \frac{\mathcal{E}\{\pi_\alpha^\delta\} - 2\sigma^2 \text{trace}(\hat{\mathbf{A}}_\alpha) + m\sigma^2}{m^2 \left[1 - \frac{1}{m} \text{trace}(\hat{\mathbf{A}}_\alpha)\right]^2}.$$

Since

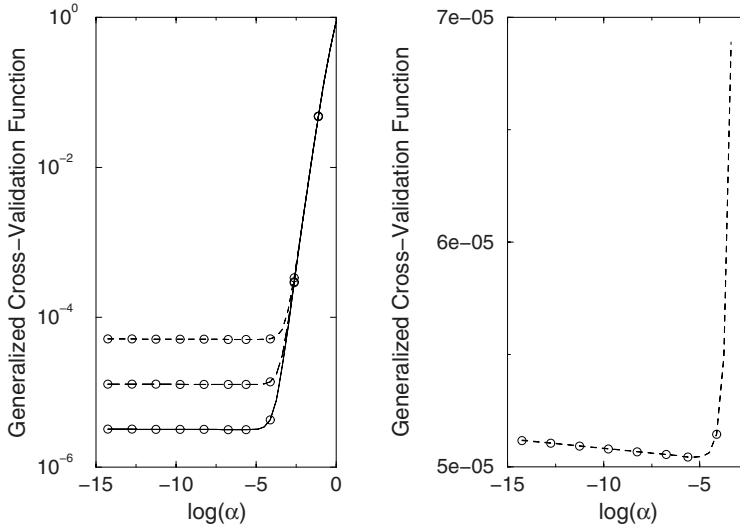
$$0 < \text{trace}(\hat{\mathbf{A}}_\alpha) = \sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha} < n,$$

we see that for  $m \gg n$ , the term  $(1/m) \text{trace}(\hat{\mathbf{A}}_\alpha)$  is small, and therefore

$$\mathcal{E}\{v_\alpha^\delta\} \approx \frac{1}{m^2} \mathcal{E}\{\pi_\alpha^\delta\} + \frac{\sigma^2}{m}. \quad (3.111)$$

Thus, the minimizer of the expected generalized cross-validation function is close to the minimizer of the expected predictive risk. In view of this equivalence, the generalized cross-validation method may suffer from the same drawback as the unbiased predictive risk estimator method: the unique minimum of the generalized cross-validation function can be very flat, thus leading to numerical difficulties in computing the regularization parameter.

In Figure 3.6 we plot the expected generalized cross-validation curve and its approximation (3.111). The agreement between the curves is acceptable over the entire domain of variation of  $\alpha$ .



**Fig. 3.6.** Expected generalized cross-validation function and its approximation. In the right panel, the  $y$ -axis is zoomed out. The results correspond to the same values of the noise standard deviation as in Figure 3.5.

### Maximum likelihood estimation

Based on a Monte Carlo analysis by Thompson et al. (1989) it was observed that the generalized cross-validation function may not have a unique minimum and that the unbiased predictive risk estimator may result in severe undersmoothing. An alternative regularization parameter choice method which overcomes these drawbacks is the maximum likelihood estimation. This selection criterion will be introduced in a stochastic setting, but for the sake of completeness, we include it in the present analysis. In the framework of the maximum likelihood estimation, the regularization parameter is computed as

$$\alpha_{\text{mle}} = \arg \min_{\alpha} \lambda_{\alpha}^{\delta},$$

where  $\lambda_{\alpha}^{\delta}$  is the maximum likelihood function defined by

$$\lambda_{\alpha}^{\delta} = \frac{\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha}) \mathbf{y}^{\delta}}{\sqrt[m]{\det (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha})}} = \frac{\sum_{i=1}^m \frac{(\mathbf{u}_i^T \mathbf{y}^{\delta})^2}{\gamma_i^2 + \alpha}}{\sqrt[m]{\prod_{i=1}^m \frac{1}{\gamma_i^2 + \alpha}}}, \quad (3.112)$$

with  $\gamma_i = 0$  for  $i = n + 1, \dots, m$ . As we shall see in Chapter 4, the minimization of the maximum likelihood function is equivalent to the maximization of the marginal likelihood function when Gaussian densities are assumed (Demoment, 1989; Kitagawa and Gersch, 1985).

### Quasi-optimality criterion

The quasi-optimality criterion is based on the hypothesis of a plateau of  $\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|$  near the optimal regularization parameter, in which case,  $\alpha = \alpha_{\text{qo}}$  is chosen so as to minimize the function

$$\varsigma_\alpha^\delta = \left\| \alpha \frac{d\mathbf{x}_\alpha^\delta}{d\alpha} \right\|^2.$$

This method originates with Tikhonov and Glasko (1965) in a slightly different form, and has been considered by numerous authors thereafter, especially in the Russian literature (Morozov, 1984). As demonstrated by Hansen (1992b), under certain assumptions, this approach also corresponds to finding a balance between the smoothing and noise errors.

To compute  $d\mathbf{x}_\alpha^\delta/d\alpha$ , we consider the regularized normal equation

$$(\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L}) \mathbf{x}_\alpha^\delta = \mathbf{K}^T \mathbf{y}^\delta,$$

take its derivative with respect to  $\alpha$ , and obtain

$$\alpha \frac{d\mathbf{x}_\alpha^\delta}{d\alpha} = (\mathbf{A}_\alpha - \mathbf{I}_n) \mathbf{x}_\alpha^\delta = (\mathbf{A}_\alpha - \mathbf{I}_n) \mathbf{K}_\alpha^\dagger \mathbf{y}^\delta. \quad (3.113)$$

Then, by (3.12), (3.16), (3.35) and (3.36), we find the computable expansion

$$\alpha \frac{d\mathbf{x}_\alpha^\delta}{d\alpha} = \mathbf{W} \Sigma_{\text{qo}} \mathbf{U}^T \mathbf{y}^\delta = - \sum_{i=1}^n \frac{\alpha \gamma_i^2}{(\gamma_i^2 + \alpha)^2} \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i, \quad (3.114)$$

with

$$\Sigma_{\text{qo}} = - \left[ \text{diag} \left( \left( \frac{\gamma_i}{\gamma_i^2 + \alpha} \right)^2 \frac{\alpha}{\sigma_i} \right)_{n \times n} \quad \mathbf{0} \right].$$

The expected quasi-optimality parameter defined by

$$\bar{\alpha}_{\text{qo}} = \arg \min_{\alpha} \mathcal{E} \{ \varsigma_\alpha^\delta \},$$

is related to the turning point of the Picard coefficients. To justify this assertion, we take  $\mathbf{L} = \mathbf{I}_n$  and assume that (3.70) and (3.71) hold with the singular-value index  $p$  being given by  $(\mathbf{u}_p^T \mathbf{y})^2 = \sigma^2$ , or equivalently, by  $C \sigma_p^{\beta+1} = \sigma$ . Using (3.114) with  $\sigma_i$  in place of  $\gamma_i$  and  $\mathbf{v}_i$  in place of  $\mathbf{w}_i$ , we obtain (cf. (3.73) and (3.75))

$$Q(\alpha) = \mathcal{E} \{ \varsigma_\alpha^\delta \} = \sum_{i=1}^n \frac{\left( \frac{\sigma_i^2}{\alpha} \right)^2}{\left( \frac{\sigma_i^2}{\alpha} + 1 \right)^4} P_i^2, \quad (3.115)$$

with  $P_i^2$  being the expected Picard coefficients. The function  $f(t) = t^2/(1+t)^4$ , with  $t = \sigma_i^2/\alpha$ , is very small if  $t \ll 1$  and  $t \gg 1$ . For  $\alpha = \sigma_j^2$ , the contributions of the terms with  $i \neq j$  in (3.115) will get suppressed, and we obtain

$$Q(\sigma_j^2) \propto P_j^2.$$

Hence, the behavior of the expected quasi-optimality function is reproduced by the expected Picard coefficients and we conclude that the turning point  $\sigma_p^2$  is not too far from  $\bar{\alpha}_{qo}$ .

The quasi-optimality criterion can be formulated in terms of the solution of the two-times iterated Tikhonov regularization  $\mathbf{x}_{\alpha 2}^\delta = \mathbf{x}_\alpha^\delta + \mathbf{p}_{\alpha 2}^\delta$ , with

$$\mathbf{p}_{\alpha 2}^\delta = \mathbf{K}_\alpha^\dagger (\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_\alpha^\delta) = (\mathbf{I}_n - \mathbf{A}_\alpha) \mathbf{x}_\alpha^\delta$$

and  $\mathbf{x}_\alpha^\delta = \mathbf{x}_{\alpha 1}^\delta$ . By (3.113) it is apparent that

$$\alpha \frac{d\mathbf{x}_\alpha^\delta}{d\alpha} = -\mathbf{p}_{\alpha 2}^\delta = \mathbf{x}_\alpha^\delta - \mathbf{x}_{\alpha 2}^\delta,$$

and therefore,

$$\zeta_\alpha^\delta = \|\mathbf{x}_\alpha^\delta - \mathbf{x}_{\alpha 2}^\delta\|^2.$$

Thus, assuming that  $\mathbf{x}_{\alpha 2}^\delta$  is a satisfactory approximation of  $\mathbf{x}^\dagger$ , we deduce that a minimizer of  $\zeta_\alpha^\delta$  is also a minimizer of  $\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2$ . In practice, the minimization of the quasi-optimality function is complicated because this function has many local minima.

### *L-curve method*

The L-curve method advocated by Hansen (1992a) is based on the L-curve, which is a parametric plot of the constraint  $\|\mathbf{c}_\alpha^\delta\|^2$  against the residual  $\|\mathbf{r}_\alpha^\delta\|^2$  in log-log scale. The corner of the L-curve appears for regularization parameters close to the optimal parameter that balances the smoothing and noise errors. The notion of a corner originates from a purely visual impression and it is not at all obvious how to translate this impression into a mathematical language. In this regard, the key problem in the L-curve method is to seek a mathematical definition of the L-curve's corner and to use this as a criterion for choosing the regularization parameter.

According to Hansen and O'Leary (1993), the corner of the L-curve is the point of maximum curvature. Defining the L-curve components by

$$x(\alpha) = \log\left(\|\mathbf{r}_\alpha^\delta\|^2\right), \quad y(\alpha) = \log\left(\|\mathbf{c}_\alpha^\delta\|^2\right),$$

we select that value of  $\alpha$  that maximizes the curvature function  $\kappa_{1c\alpha}^\delta$ ,

$$\alpha_{1c} = \arg \max_{\alpha} \kappa_{1c\alpha}^\delta,$$

where

$$\kappa_{1c\alpha}^\delta = \frac{x''(\alpha)y'(\alpha) - x'(\alpha)y''(\alpha)}{\left[x'(\alpha)^2 + y'(\alpha)^2\right]^{\frac{3}{2}}} \quad (3.116)$$

and the prime ( $\prime$ ) denotes differentiation with respect to  $\alpha$ .

In order to simplify the notations, we set  $R_\delta(\alpha) = \|\mathbf{r}_\alpha^\delta\|^2$  and  $C_\delta(\alpha) = \|\mathbf{c}_\alpha^\delta\|^2$ , where  $R_\delta$  and  $C_\delta$  are given by (3.83) and (3.89), respectively. Straightforward differentiation gives

$$R'_\delta(\alpha) = -\alpha C'_\delta(\alpha)$$

and we obtain a simple formula for the curvature depending on  $R_\delta$ ,  $C_\delta$  and  $C'_\delta$ :

$$\kappa_{1c\alpha}^\delta = - \frac{\alpha R_\delta(\alpha) C_\delta(\alpha) [R_\delta(\alpha) + \alpha C'_\delta(\alpha)] + R_\delta(\alpha)^2 C_\delta(\alpha)^2 / C'_\delta(\alpha)}{\left[ R_\delta(\alpha)^2 + \alpha^2 C_\delta(\alpha)^2 \right]^{\frac{3}{2}}}. \quad (3.117)$$

Any one-dimensional optimization routine can be used to locate the regularization parameter  $\alpha_{1c}$  which corresponds to the maximum curvature.

An alternative definition of the corner of the L-curve has been given by Reginska (1996). The point  $C = (x(\alpha_{1c}), y(\alpha_{1c}))$  is the corner of the L-curve if

- (1) the tangent of the curve at  $C$  has slope  $-1$ ;
- (2) in a neighborhood of  $C$ , the points on the curve lie above the tangent.

An implementable algorithm of this selection criterion can be designed by using the following result: the point  $C = (x(\alpha_{1c}), y(\alpha_{1c}))$  is a corner of the L-curve in the aforementioned sense if and only if the function

$$\Psi_{1c}^\delta(\alpha) = R_\delta(\alpha) C_\delta(\alpha)$$

has a local minimum at  $\alpha_{1c}$ , that is,

$$\alpha_{1c} = \arg \min_{\alpha} \Psi_{1c}^\delta(\alpha).$$

For a proof of this equivalence we refer to Engl et al. (2000).

The expected L-curve and its negative curvature are illustrated in Figure 3.7. We recall that the expected L-curve has the components  $x(\alpha) = \log R(\alpha)$  and  $y(\alpha) = \log C(\alpha)$ , where  $R$  and  $C$  are given by (3.86) and (3.90), respectively. Also note that, since  $R'(\alpha) = -\alpha C'(\alpha)$ , the curvature of the expected L-curve can be computed by using (3.117) with  $R$  and  $C$  in place of  $R_\delta$  and  $C_\delta$ , respectively. The plots show that by increasing the noise standard deviation, the regularization parameter also increases and more regularization is introduced.

### ***Residual curve method and its generalized version***

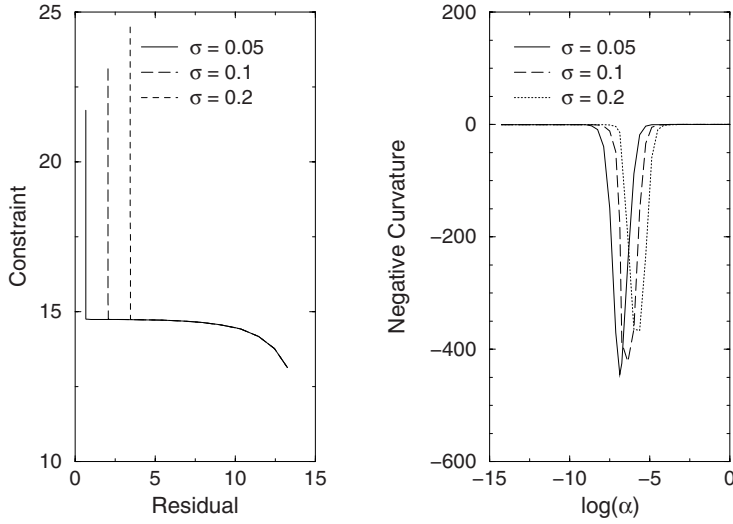
The residual curve is the plot of the log of the residual  $\|\mathbf{r}_\alpha^\delta\|^2$  against the log of regularization parameter  $\alpha$ . This curve typically has a mirror symmetric L-shape and the residual curve method chooses the regularization parameter corresponding to the corner of this curve.

Analogously to the L-curve method, we may define the corner of the residual curve as the point with minimum curvature. Denoting the components of the residual curve by

$$x(\alpha) = \log \alpha, \quad y(\alpha) = \log R_\delta(\alpha), \quad (3.118)$$

we have

$$\alpha_{rc} = \arg \min_{\alpha} \kappa_{rc\alpha}^\delta,$$



**Fig. 3.7.** Expected L-curve and its negative curvature for three values of the noise standard deviation  $\sigma$ .

where

$$\kappa_{\text{rc}\alpha}^{\delta} = - \frac{R_{\delta}(\alpha)^2 [\alpha R'_{\delta}(\alpha) + \alpha^2 R''_{\delta}(\alpha)] - \alpha^2 R'_{\delta}(\alpha)^2 R_{\delta}(\alpha)}{[R_{\delta}(\alpha)^2 + \alpha^2 R'_{\delta}(\alpha)^2]^{\frac{3}{2}}}. \quad (3.119)$$

The corner of the residual curve can also be defined as the point  $C = (x(\alpha_{\text{rc}}), y(\alpha_{\text{rc}}))$  with the following properties:

- (1) the tangent of the curve at  $C$  has slope 1;
- (2) in a neighborhood of  $C$ , the points on the curve lie above the tangent.

This notion of the corner leads to the error-free parameter choice method discussed by Engl et al. (2000): the point  $C = (x(\alpha_{\text{rc}}), y(\alpha_{\text{rc}}))$  is a corner of the residual curve if and only if the function

$$\Psi_{\text{rc}}^{\delta}(\alpha) = \frac{1}{\alpha} R_{\delta}(\alpha)$$

has a local minimum at  $\alpha_{\text{rc}}$ , i.e.,

$$\alpha_{\text{rc}} = \arg \min_{\alpha} \Psi_{\text{rc}}^{\delta}(\alpha). \quad (3.120)$$

To show this equivalence, we observe that by (3.118), we have

$$\Psi_{\text{rc}}^{\delta}(\alpha) = \exp(y(\alpha) - x(\alpha)).$$

If  $\Psi_{\text{rc}}^{\delta}(\alpha)$  has a local extremum at  $\alpha_{\text{rc}}$ , there holds

$$\Psi_{\text{rc}}^{\delta'}(\alpha_{\text{rc}}) = [y'(\alpha_{\text{rc}}) - x'(\alpha_{\text{rc}})] \Psi_{\text{rc}}^{\delta}(\alpha_{\text{rc}}) = 0,$$

which yields

$$y'(\alpha_{rc}) - x'(\alpha_{rc}) = 0.$$

Thus, the tangent of the curve at  $C$  is parallel to the vector  $[1, 1]^T$ , and the equation of the tangent is given by

$$y - x = y(\alpha_{rc}) - x(\alpha_{rc}). \quad (3.121)$$

If  $\alpha_{rc}$  is now a minimizer of  $\Psi_{rc}^\delta(\alpha)$ , then  $\log \Psi_{rc}^\delta(\alpha) = y(\alpha) - x(\alpha)$  also has a local minimum at  $\alpha_{rc}$ , and we have

$$y(\alpha) - x(\alpha) \geq y(\alpha_{rc}) - x(\alpha_{rc}) \quad (3.122)$$

for  $\alpha$  near  $\alpha_{rc}$ . Hence, in the neighborhood of  $\alpha_{rc}$ , the points  $(x(\alpha), y(\alpha))$  lie above the tangent at  $C$ . Conversely, if the tangent of the residual curve at  $C$  has slope 1, then the tangent is given by (3.121), and the condition (3.122) implies that  $\Psi_{rc}^\delta$  has a local minimum at  $\alpha_{rc}$ .

A heuristic justification of the regularization parameter choice method (3.120) relies on the observation that the behaviors of the solution error  $\|\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta\|^2$  and the scaled residual  $(1/\alpha) R_\delta(\alpha)$  as functions of  $\alpha$  are similar, and as a result, their minimizers are very close. In this regard, the function  $(1/\alpha) R_\delta(\alpha)$  is also known as the error indicator function (Rieder, 2003).

In the left and middle panels of Figure 3.8 we plot the expected error indicator function  $\Psi_{rc}(\alpha) = (1/\alpha) R(\alpha)$  and the curvature of the expected residual curve of components  $x(\alpha) = \log \alpha$  and  $y(\alpha) = \log R(\alpha)$ . The plots correspond to the synthetic model (3.78)–(3.80) and show that the curves have unique minimizers which increase with increasing the noise standard deviation.

Wu (2003) proposed a regularization parameter choice method which is very similar to the residual curve method. This method is called the ‘flattest slope method’ and uses the plot of the constraint  $C_\delta(\alpha)$  against  $\log(1/\alpha)$ . The graph of  $(\log(1/\alpha), C_\delta(\alpha))$  has a corner which divides the curve into two pieces: the left piece is flat, while the right piece is very steep. As in the residual curve method, the regularization parameter of the ‘flattest slope method’ corresponds to a point on the flat portion just before the rapid growing.

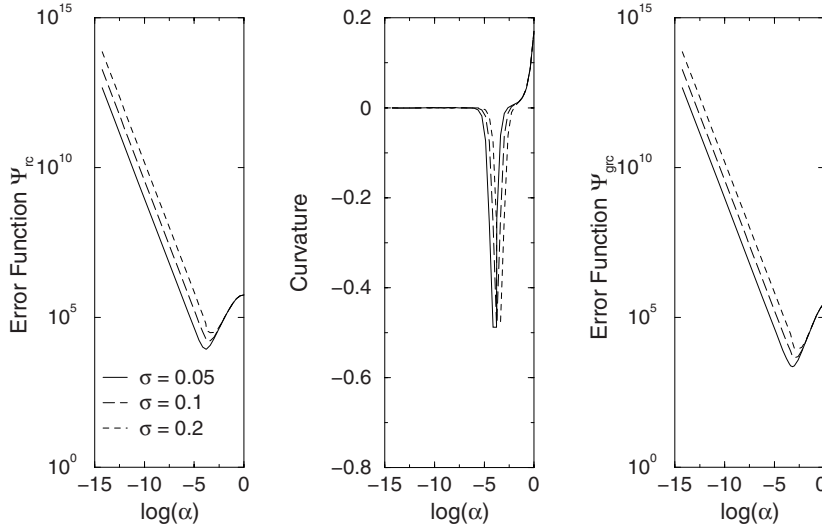
Analogously to the residual curve we may consider the generalized residual curve, which represents the plot of the log of the ‘generalized’ residual  $R_{g\delta}(\alpha) = \|\mathbf{r}_\alpha^\delta\|^2 - \mathbf{r}_\alpha^{\delta T} \hat{\mathbf{A}}_\alpha \mathbf{r}_\alpha^\delta$  against the log of the regularization parameter  $\alpha$ . In view of the slope definition for the corner of the generalized residual curve, we select the regularization parameter  $\alpha_{grc}$  as the minimizer of the error indicator function

$$\Psi_{grc}^\delta(\alpha) = \frac{1}{\alpha} R_{g\delta}(\alpha),$$

with  $R_{g\delta}$  as in (3.97). This regularization parameter choice method has been developed by Hanke and Raus (1996) and can also be found in Engl et al. (2000) and Rieder (2003). The expected error indicator function  $\Psi_{grc}(\alpha) = (1/\alpha) R_g(\alpha)$ , with

$$R_g(\alpha) = \mathcal{E} \{ R_{g\delta}(\alpha) \} = (m - n) \sigma^2 + \sum_{i=1}^n \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^3 \left[ (\mathbf{u}_i^T \mathbf{y})^2 + \sigma^2 \right],$$

is plotted in the right panel of Figure 3.8, and as in the residual curve method, the minimizers increase with increasing the noise standard deviation.



**Fig. 3.8.** Expected error indicator functions  $\Psi_{rc}$  (left) and  $\Psi_{grc}$  (right), and the curvature of the expected residual curve (middle) for different values of the noise standard deviation  $\sigma$ .

### 3.7 Numerical analysis of regularization parameter choice methods

In the first step of our numerical analysis we examine the regularization parameter choice methods discussed in this chapter by considering the synthetic model (3.78)–(3.80) with  $m = 800$ ,  $n = 20$  and  $\omega = 0.75$ . Specifically, we compute the expected regularization parameter  $\bar{\alpha}$  of a particular parameter choice method and estimate the expected relative error

$$\bar{\varepsilon} = \sqrt{\frac{\mathcal{E} \left\{ \left\| \mathbf{e}_{\bar{\alpha}}^{\delta} \right\|^2 \right\}}{\left\| \mathbf{x}^{\dagger} \right\|^2}}. \quad (3.123)$$

The following regularization parameters are considered:

- (1) the optimal regularization parameter for error estimation

$$\bar{\alpha}_{\text{opt}} = \arg \min_{\alpha} \mathcal{E} \left\{ \left\| \mathbf{e}_{\alpha}^{\delta} \right\|^2 \right\};$$

- (2) the expected predictive risk parameter

$$\bar{\alpha}_{\text{pr}} = \arg \min_{\alpha} \mathcal{E} \left\{ \pi_{\alpha}^{\delta} \right\};$$

- (3) the expected discrepancy principle parameter  $\bar{\alpha}_{\text{dp}}$  solving the equation

$$\mathcal{E} \left\{ \left\| \mathbf{r}_{\alpha}^{\delta} \right\|^2 \right\} = \tau m \sigma^2;$$

- (4) the expected generalized discrepancy principle parameter  $\bar{\alpha}_{\text{gdp}}$  solving the equation

$$\mathcal{E} \left\{ \left\| \mathbf{r}_{\alpha}^{\delta} \right\|^2 - \mathbf{r}_{\alpha}^{\delta T} \hat{\mathbf{A}}_{\alpha} \mathbf{r}_{\alpha}^{\delta} \right\} = \tau m \sigma^2;$$



- (5) the expected generalized cross-validation parameter

$$\bar{\alpha}_{\text{gcv}} = \arg \min_{\alpha} \mathcal{E} \{ \nu_{\alpha}^{\delta} \};$$

- (6) the expected L-curve parameter maximizing the curvature of the expected L-curve,

$$\bar{\alpha}_{1c} = \arg \max_{\alpha} \kappa_{1c\alpha};$$

- (7) the expected residual curve parameter
- $\bar{\alpha}_{\text{rc}}$
- minimizing the curvature of the expected residual curve,

$$\bar{\alpha}_{\text{rc}} = \arg \min_{\alpha} \kappa_{\text{rc}\alpha};$$

- (8) the expected residual curve parameters
- $\bar{\alpha}_{\text{rc}}$
- and
- $\bar{\alpha}_{\text{grc}}$
- minimizing the expected error indicator functions

$$\Psi_{\text{rc}}(\alpha) = \frac{1}{\alpha} R(\alpha)$$

and

$$\Psi_{\text{grc}}(\alpha) = \frac{1}{\alpha} R_{\text{g}}(\alpha),$$

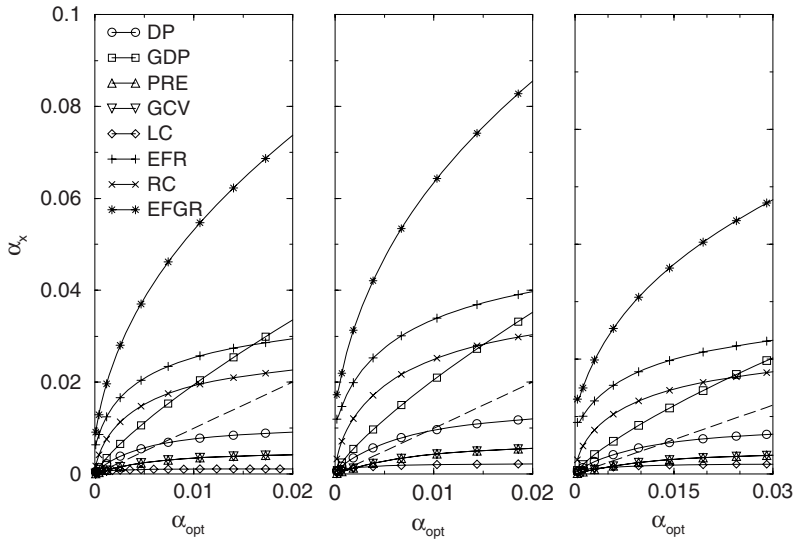
respectively.

The regularization parameters are illustrated in Figure 3.9. For these simulations, we consider three values of  $\sigma$ , and for each  $\sigma$ , we compute the regularization parameters for 20 values of  $\beta$  ranging from 0.2 to 2.0. On the  $y$ -axis, we represent the expected parameter  $\bar{\alpha}$  obtained from a particular parameter choice method, while on the  $x$ -axis, we represent the optimal regularization parameter for error estimation  $\bar{\alpha}_{\text{opt}}$ . The dashed curve in Figure 3.9 is the  $y = x$  line, and the deviation of an error curve from this line serves as an evidence for the deviation of  $\bar{\alpha}$  from  $\bar{\alpha}_{\text{opt}}$ . The plots show that

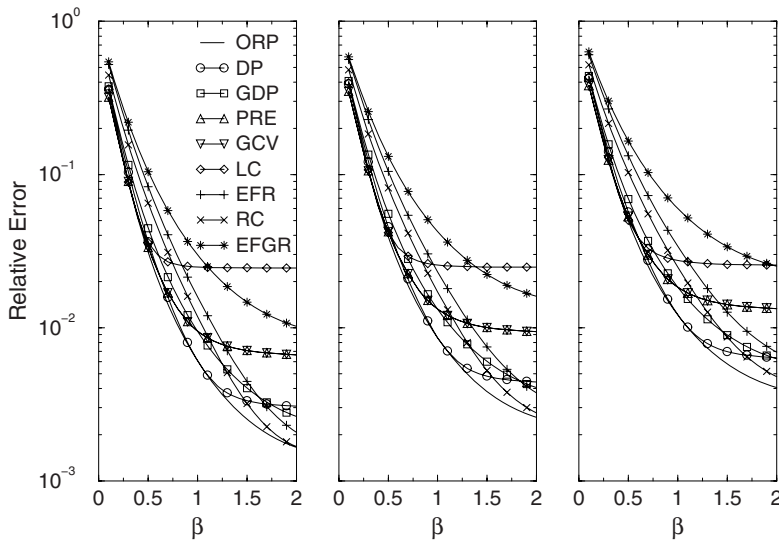
- (1)  $\bar{\alpha}_{\text{pr}} \approx \bar{\alpha}_{\text{gcv}} < \bar{\alpha}_{\text{dp}} < \bar{\alpha}_{\text{opt}};$
- (2)  $\bar{\alpha}_{\text{pr}} \approx \bar{\alpha}_{\text{gcv}} < \bar{\alpha}_{\text{gdp}}.$

The above inequalities have been proven by Lukas (1998a) for a semi-discrete data model and a Picard condition as in (3.70). Note that the inequalities proven by Lukas are  $\bar{\alpha}_{\text{pr}} < \bar{\alpha}_{\text{dp}} < c_1 \bar{\alpha}_{\text{opt}}$  and  $\bar{\alpha}_{\text{pr}} < \bar{\alpha}_{\text{gdp}} < c_2 \bar{\alpha}_{\text{opt}}$ , with  $c_1 > 2$  and  $c_2 > 3$ . The plots also illustrate that the regularization parameter of the L-curve method  $\bar{\alpha}_{1c}$  is significantly smaller than the optimal regularization parameter for error estimation  $\bar{\alpha}_{\text{opt}}$ , and this effect is more pronounced for small values of the noise standard deviation and very smooth solutions. A similar behavior, but in a continuous and deterministic setting, has been reported by Hanke (1996).

The expected relative errors  $\bar{\varepsilon}$  are plotted in Figure 3.10. A general conclusion is that the methods based on the analysis of the expected residual curve and the generalized residual curve yield results with a low accuracy, especially for solutions with a reduced degree of smoothness. Also apparent is that the L-curve method is characterized by a saturation effect, i.e., the relative error does not decrease with decreasing  $\sigma$  and increasing  $\beta$ . For small values of the smoothness parameter  $\beta$ , e.g.,  $0.2 \leq \beta \leq 0.5$ , the expected relative errors are close to the relative error corresponding to error estimation, while for larger values of  $\beta$ , e.g.,  $0.5 \leq \beta \leq 2.0$ , the deviations become more visible. Since in



**Fig. 3.9.** Regularization parameters computed with the discrepancy principle (DP), the generalized discrepancy principle (GDP), the predictive risk estimator (PRE) method, generalized cross-validation (GCV), the L-curve (LC) method, and the methods which minimize the error indicator function  $\Psi_{\text{rc}}$  (EFR), the curvature of the residual curve (RC) and the error indicator function  $\Psi_{\text{grc}}$  (EFGR). The plots correspond to  $\sigma = 0.05$  (left),  $\sigma = 0.1$  (middle) and  $\sigma = 0.2$  (right).



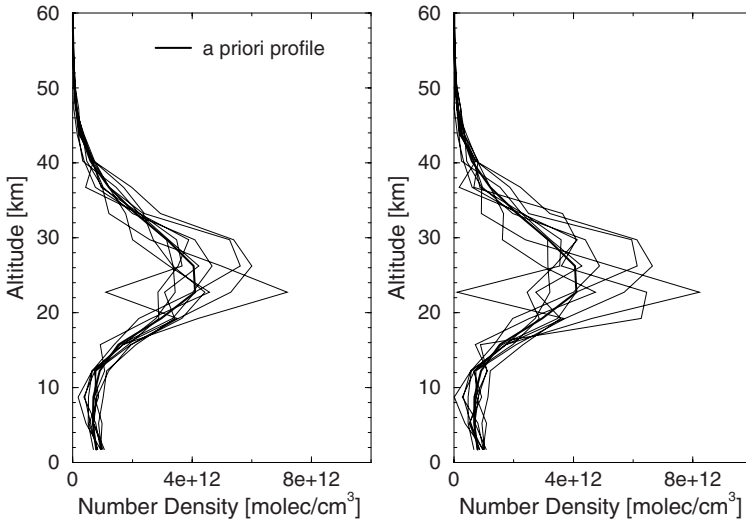
**Fig. 3.10.** Expected relative errors versus the smoothness parameter  $\beta$  for the optimal regularization parameter (ORP) for error estimation and the regularization parameters considered in Figure 3.9.

practice very smooth solutions are not expected, we may conclude that the above methods produce good regularization parameters.

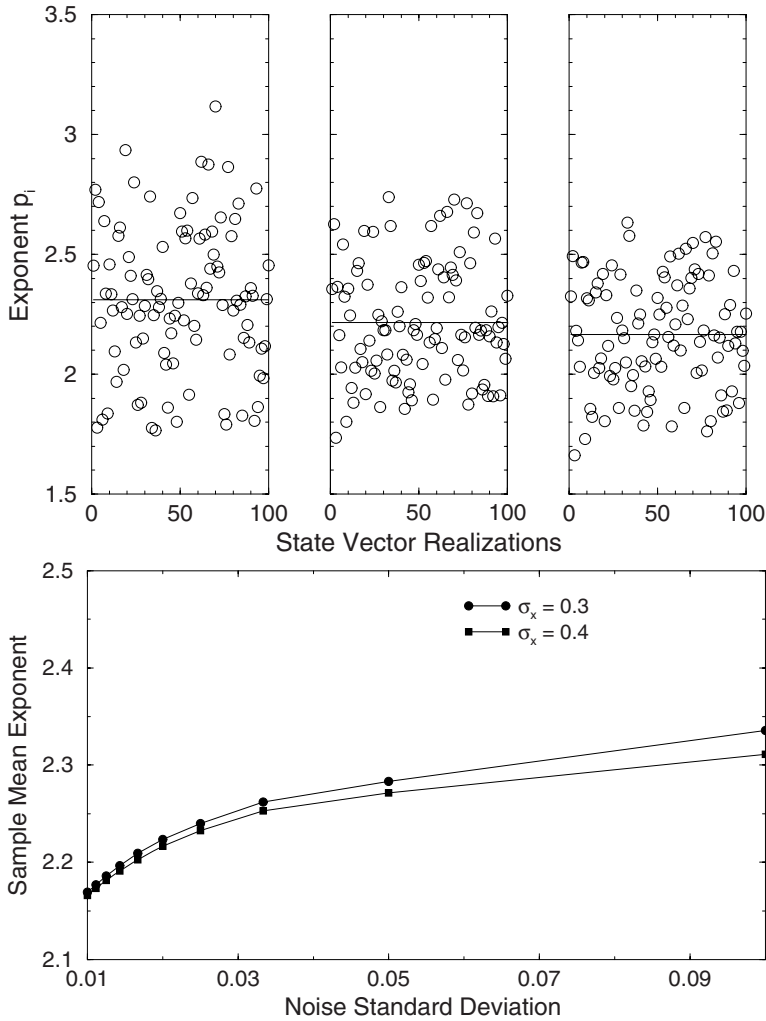
Next, we analyze the performance of the regularization parameter choice methods for an ozone retrieval test problem. The atmospheric ozone profile is retrieved from a sequence of simulated limb spectra in a spectral interval ranging from 323 to 333 nm. The number of limb scans is 11 and the limb tangent height varies between 14 and 49 km in steps of 3.5 km. The atmosphere is discretized with a step of 3.5 km between 0 and 70 km, and a step of 5 km between 70 and 100 km. The problem is assumed to be nearly linear in the sense that a linearization of the forward model about the a priori state is appropriate to find a solution. In view of the transformations discussed in section 3.1, we emphasize that the state vector represents the deviation of the gas profile with respect to the a priori. The forward model assumes piecewise constant interpolation for profile representation, while the regularization matrix is the Cholesky factor of a normalized covariance matrix with an altitude-independent correlation length  $l = 3.5$  km.

To compute the a priori regularization parameter in the framework of the expected error estimation method, we consider 100 realizations of a Gaussian process with zero mean vector and a covariance matrix characterized by the correlation length  $l = 3.5$  km and the profile standard deviations  $\sigma_x = 0.3$  and  $\sigma_x = 0.4$ . Ten realization of the true profile  $\mathbf{x}^\dagger + \mathbf{x}_a$  computed with the generation algorithm described in section 3.6.1 are shown in Figure 3.11.

The results illustrated in the top panel of Figure 3.12 correspond to  $\sigma_x = 0.4$  and represent the exponent  $p_i$  for different state vector realizations and three values of the noise standard deviation  $\sigma$ , namely 0.1, 0.02 and 0.01. In these three situations, the values of the sample mean exponent  $\bar{p}$  are 2.31, 2.21 and 2.17, while the values of the sample standard



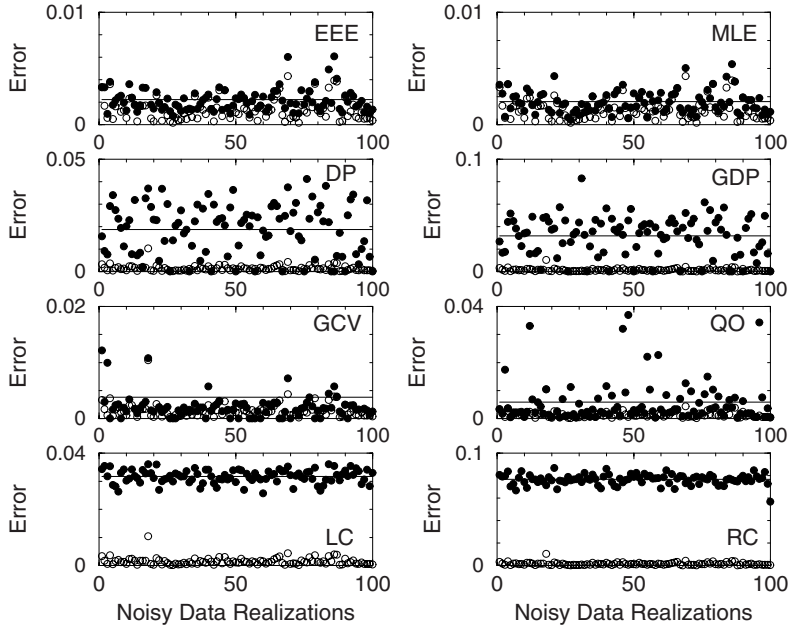
**Fig. 3.11.** Ten realizations of the true ozone profile  $\mathbf{x}^\dagger + \mathbf{x}_a$  for a Gaussian covariance matrix with the correlation length  $l = 3.5$  km and the profile standard deviations  $\sigma_x = 0.3$  (left) and  $\sigma_x = 0.4$  (right).



**Fig. 3.12.** Top: exponent  $p_i$  for different state vector realizations and three values of the noise standard deviation  $\sigma$ : 0.1 (left), 0.02 (middle) and 0.01 (right); the Gaussian covariance matrix is characterized by the correlation length  $l = 3.5$  km and the profile standard deviation  $\sigma_x = 0.4$ . Bottom: sample mean exponent as a function of the noise standard deviation  $\sigma$ .

deviation  $s_p$  are 0.30, 0.24 and 0.23. Thus,  $\bar{p}$  and  $s_p$  slightly increase with increasing  $\sigma$ . The sample mean exponent as a function of the noise standard deviation is shown in the bottom panel of Figure 3.12. As the average values of  $\bar{p}$  over  $\sigma$  are 2.23 for  $\sigma_x = 0.3$  and 2.22 for  $\sigma_x = 0.4$ , we adopt the a priori selection rule  $\alpha_e = \sigma^{2.225}$ .

The exact state vector is now chosen as a translated and a scaled version of a climatological profile with a translation distance of 3 km and a scaling factor of 1.3. For a fixed value of the noise standard deviation  $\sigma$ , we compute the exact data vector  $\mathbf{y}$ , and generate the noisy data vector  $\mathbf{y}_i^\delta = \mathbf{y} + \delta_i$ , with  $\delta_{i=1,N}$  being a random sample of the white noise



**Fig. 3.13.** Relative solution errors for the expected error estimation (EEE) method, the maximum likelihood estimation (MLE), the discrepancy principle (DP), the generalized discrepancy principle (GDP), generalized cross-validation (GCV), the quasi-optimality (QO) criterion, the L-curve (LC) method, and the residual curve (RC) method. The non-filled circles correspond to the optimal values of the regularization parameter. The noise standard deviation is  $\sigma = 0.1$ . For the residual curve method, the regularization parameter is computed by minimizing the curvature of the residual curve.

with the  $N(0, \sigma^2 \mathbf{I}_m)$  distribution. The number of noisy data realizations is 100, and for each  $\mathbf{y}_i^\delta$ , we determine the regularization parameter  $\alpha_i$  by a particular parameter choice method.

In Figure 3.13 we plot the solution errors

$$\varepsilon_i = \frac{\|\mathbf{x}_{\alpha_i}^\delta - \mathbf{x}^\dagger\|}{\|\mathbf{x}^\dagger\|}, \quad (3.124)$$

where  $\mathbf{x}_{\alpha_i}^\delta = \mathbf{K}_{\alpha_i}^\dagger \mathbf{y}_i^\delta$  is the regularized solution of parameter  $\alpha_i$  corresponding to the noisy data vector  $\mathbf{y}_i^\delta$ . Also shown in Figure 3.13 are the solution errors

$$\varepsilon_{\text{opt}i} = \frac{\|\mathbf{x}_{\alpha_{\text{opt}i}}^\delta - \mathbf{x}^\dagger\|}{\|\mathbf{x}^\dagger\|},$$

for the optimal regularization parameter,

$$\alpha_{\text{opt}i} = \arg \min_{\alpha} \|\mathbf{K}_{\alpha}^\dagger \mathbf{y}_i^\delta - \mathbf{x}^\dagger\|^2.$$

The average values of the solution errors over noisy data realizations are given in Table 3.1. It should be remarked that the discrepancy principle, the generalized discrepancy

**Table 3.1.** Average values of the relative solution errors in percent for different regularization parameter choice methods. The noise standard deviation is  $\sigma = 0.1$ .

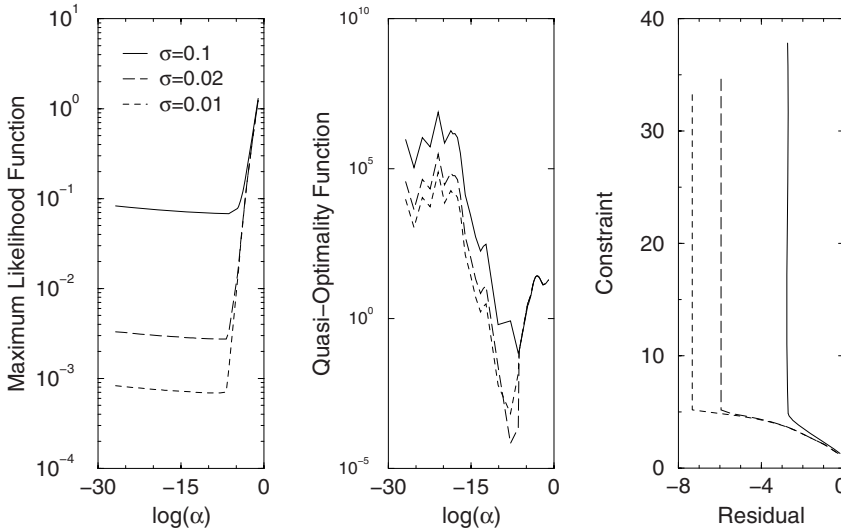
Regularization parameter choice method	Relative solution error
optimal regularization parameter	0.14
expected error estimation method	0.23
maximum likelihood estimation	0.20
discrepancy principle	1.88
generalized discrepancy principle	3.19
generalized cross-validation	0.38
quasi-optimality criterion	0.59
L-curve method	3.17
residual curve method	7.68

principle and generalized cross-validation occasionally fail and produce a very small  $\alpha$ . For the discrepancy principle and its generalized version this happens 11 times, while for generalized cross-validation this happens 17 times. The discrepancy principle fails when the corresponding equation does not have a solution. This never occurs for the ‘expected equation’, but may occur for the ‘noisy equation’. In fact, we can choose the control parameter  $\tau$  sufficiently large, so that the discrepancy principle equation is always solvable, but in this case, the solution errors may become extremely large. In our simulation we optimize the tolerance  $\tau$  by minimizing the error for the first 10 configurations and use the computed value ( $\tau = 1.03$ ) for the rest of the calculation. The failure of generalized cross-validation occurs when the curve has a flat minimum and a very small  $\alpha$  is found as minimizer. The average values of the solution errors reported in Table 3.1 have been computed by disregarding the situations in which the methods fail. The remaining regularization parameter choice methods are robust and can be classified according to their accuracy as follows:

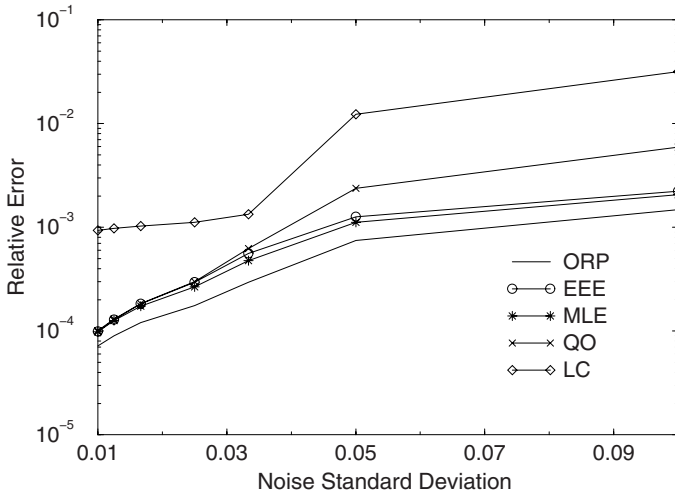
- (1) the maximum likelihood estimation,
- (2) the expected error estimation method,
- (3) the quasi-optimality criterion,
- (4) the L-curve method.

The maximum likelihood function  $\lambda_\alpha^\delta$ , the quasi-optimality function  $\zeta_\alpha^\delta$  and the L-curve are shown in Figure 3.14 for one noisy data realization. The behavior of the maximum likelihood function is somehow similar to the behavior of the generalized cross-validation function, but the minimum is not so extraordinarily flat. The quasi-optimality function has several local minima and to compute the global minimizer, we have to split the interval of variation of the regularization parameter in several subintervals and have to compute a local minimum in each subinterval with a robust minimization routine.

The regularizing effect of the parameter choice methods is illustrated in Figure 3.15. Here, we plot the average values of the solution errors (over noisy data realizations) versus the noise standard deviation. The results show that when the noise standard deviation decreases, the average solution error also decreases, except for the L-curve method which is characterized by a saturation effect in the region of small  $\sigma$ .

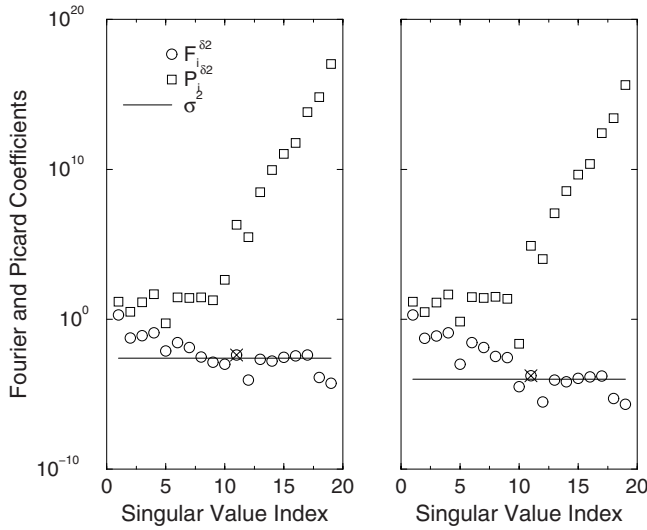


**Fig. 3.14.** Maximum likelihood function (left), quasi-optimality function (middle) and L-curve (right) for different values of the noise standard deviation  $\sigma$ .



**Fig. 3.15.** Average values of the relative solution error over 100 noisy data realizations for the optimal regularization parameter (ORP), the expected error estimation (EEE) method, the maximum likelihood estimation (MLE), the quasi-optimality (QO) criterion, and the L-curve (LC) method.

In Figure 3.16 we plot the Fourier coefficients  $F_i^{\delta 2} = (\mathbf{u}_i^T \mathbf{y}^\delta)^2$  and the Picard coefficients  $P_i^{\delta 2} = F_i^{\delta 2} / \gamma_i^2$  for two noisy data realizations with  $\sigma = 0.1$  and  $\sigma = 0.02$ . In both situations, the Fourier coefficients level off at  $i = 11$ , and we have  $\log \gamma_{11} = -5.13$  and  $\log \gamma_{10} = -2.30$ . As  $\log \sqrt{\alpha_{\text{opt}}} = -2.63$  for  $\sigma = 0.1$ , and  $\log \sqrt{\alpha_{\text{opt}}} = -4.26$  for  $\sigma = 0.02$ , we see that  $\log \gamma_{11} < \log \sqrt{\alpha_{\text{opt}}} < \log \gamma_{10}$ . This result suggests that  $\gamma_{11}^2$  is a rough



**Fig. 3.16.** Fourier and Picard coefficients for  $\sigma = 0.1$  (left) and  $\sigma = 0.02$  (right). The point marked with X corresponds to  $i = 11$  and indicates the plateau of the Fourier coefficients.

approximation of  $\alpha_{\text{opt}}$ .

In our next simulations we analyze the efficiency of the regularization parameter choice methods in the presence of forward model errors. In this case, the noisy data vector is generated as

$$\mathbf{y}^\delta = \mathbf{y} + \boldsymbol{\delta} + \varepsilon_m \mathbf{y},$$

where  $\varepsilon_m$  is a tolerance which controls the magnitude of the forward model error  $\boldsymbol{\delta}_m = \varepsilon_m \mathbf{y}$ .

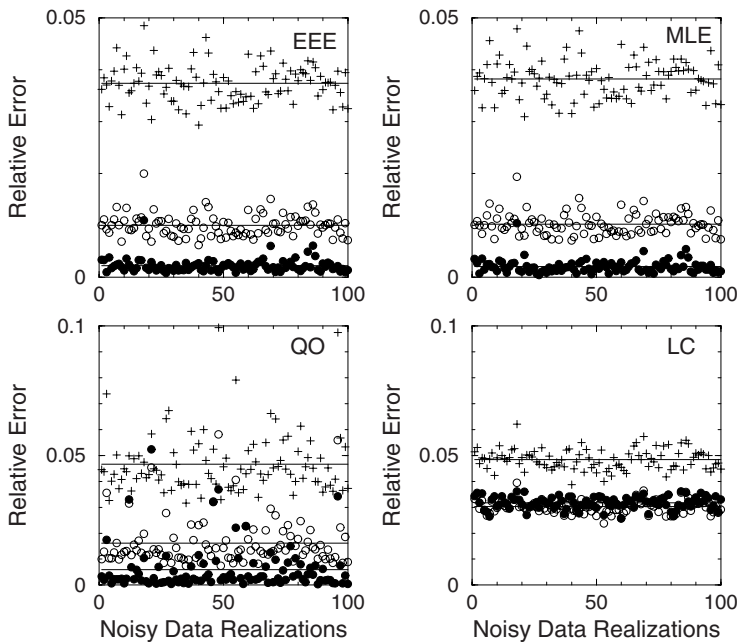
The solution errors for the expected error estimation method, the maximum likelihood estimation, the quasi-optimality criterion and the L-curve method are illustrated in Figure 3.17. The results show that by increasing  $\varepsilon_m$ , the average and the standard deviation of the solution errors also increase.

The average values of the solution errors for different values of the tolerance  $\varepsilon_m$  are given in Table 3.2. It is interesting to note that for large values of  $\varepsilon_m$ , all methods yield the same accuracy. In this regard, we may conclude that the L-curve method is efficient for data with large noise levels.

In actual fact, our numerical simulation reveals that there is no infallible regularization parameter choice method. This is because

- (1) the expected error estimation method requires the knowledge of a solution domain with physical meaning and is time-consuming;
- (2) the discrepancy principle and its generalized version are sensitive to the selection of the control parameter  $\tau$ ;
- (3) the predictive risk, the generalized cross-validation and sometimes the maximum likelihood functions may have very flat minima;
- (4) the quasi-optimality function has several local minima and sometimes it does not have a global minimum at all;





**Fig. 3.17.** Relative solution errors for the expected error estimation (EEE) method, the maximum likelihood estimation (MLE), the quasi-optimality (QO) criterion, and the L-curve (LC) method. The results correspond to  $\sigma = 0.1$  and to three values of  $\varepsilon_m$ : 0 (filled circle), 0.02 (non-filled circle) and 0.04 (plus).

**Table 3.2.** Average values of the relative solution errors in percent for different values of the tolerance  $\varepsilon_m$ . The noise standard deviation is  $\sigma = 0.1$ .

Regularization parameter choice method	Tolerance $\varepsilon_m$		
	0	0.02	0.04
expected error estimation method	0.23	1.01	3.74
maximum likelihood estimation	0.20	1.02	3.82
quasi-optimality criterion	0.59	1.61	4.66
L-curve method	3.17	3.23	4.84

(5) the L-curve may lose its L-shape.

In this context, it is advantageous to monitor several strategies and base the choice of the regularization parameters on the output of all these strategies.

### 3.8 Multi-parameter regularization methods

In many applications, the state vector consists of several components which are assumed to be independent. The statement of a two-component problem reads as

$$\mathbf{y}^\delta = \mathbf{K}_1 \mathbf{x}_1 + \mathbf{K}_2 \mathbf{x}_2 + \boldsymbol{\delta}, \quad (3.125)$$

with

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \quad \mathbf{K} = [\mathbf{K}_1, \mathbf{K}_2].$$

The data model (3.125) may correspond to a linear problem or to a nearly-linear problem, in which case,  $\mathbf{K}_1$  and  $\mathbf{K}_2$  are the Jacobian matrices of the forward model with respect to  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , respectively. Let us assume that for each component  $\mathbf{x}_i$  we are able to construct an appropriate regularization matrix  $\mathbf{L}_i$ . As the components of the state vector are independent, we can assemble the individual regularization matrices into a global regularization matrix with a block-diagonal structure. For a two-component vector, the global regularization matrix can be expressed as

$$\mathbf{L}_\omega = \begin{bmatrix} \sqrt{\omega} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \sqrt{1-\omega} \mathbf{L}_2 \end{bmatrix}, \quad (3.126)$$

while the associated Tikhonov function takes the form

$$\mathcal{F}_{\alpha\omega}(\mathbf{x}) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha \|\mathbf{L}_\omega \mathbf{x}\|^2. \quad (3.127)$$

The parameter  $0 < \omega < 1$  is called the weighting factor and gives the contribution of each individual regularization matrix to the global regularization matrix. In practice, the weighting factor is unknown and we have to use a so-called multi-parameter regularization method to compute both the weighting factor and the regularization parameter. It should be pointed out that a one-component problem with a parameter-dependent regularization matrix  $\mathbf{L}_\omega$  (constructed by means of incomplete statistical information) is also a multi-parameter regularization problem; the parameter  $\omega$  can be the correlation length or the ratio of two altitude-dependent profile standard deviations.

The penalty term can also be expressed as

$$\Omega(\mathbf{x})^2 = \alpha_1 \|\mathbf{H}_1 \mathbf{x}\|^2 + \alpha_2 \|\mathbf{H}_2 \mathbf{x}\|^2, \quad (3.128)$$

with

$$\mathbf{H}_1 = \begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{H}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 \end{bmatrix}, \quad (3.129)$$

whence, in view of the identity

$$\alpha \mathbf{L}_\omega^T \mathbf{L}_\omega = \alpha_1 \mathbf{H}_1^T \mathbf{H}_1 + \alpha_2 \mathbf{H}_2^T \mathbf{H}_2,$$

the equivalence

$$\alpha = \alpha_1 + \alpha_2, \quad \omega = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

readily follows.

Multi-parameter regularization methods can be roughly classified according to the goal of the inversion process. We distinguish between

- (1) complete multi-parameter regularization methods, when we are interested in computing the entire regularized solution;
- (2) incomplete multi-parameter regularization methods, when we are interested in the reconstruction of some components of the state vector, e.g., the retrieval of a main atmospheric gas by considering an auxiliary gas as a contamination.

In this section we treat multi-parameter regularization methods under the simplified assumption that the state vector consists of two components.

### 3.8.1 Complete multi-parameter regularization methods

Most of the one-parameter regularization methods, relying on the minimization of certain objective functions, can be used to handle this problem. The idea is to regard the objective function as a multivariate function and to use an appropriate optimization method to compute the regularization parameter  $\alpha$  and the weighting factor  $\omega$ .

In the one-parameter regularization case, the objective function has been expressed in terms of a generalized singular system of the matrix pair  $(\mathbf{K}, \mathbf{L})$ , and the derivatives with respect to the regularization parameter have been computed in an analytical manner. Unfortunately, in the multi-parameter regularization case, there is no factorization of the form

$$\mathbf{K} = \mathbf{U}\Sigma_0\mathbf{W}^{-1}, \quad \mathbf{H}_1 = \mathbf{V}_1\Sigma_1\mathbf{W}^{-1}, \quad \mathbf{H}_2 = \mathbf{V}_2\Sigma_2\mathbf{W}^{-1},$$

which could reduce the computational complexity preserving the accuracy of computation (Brezinski et al., 2003). Here,  $\mathbf{U}$ ,  $\mathbf{V}_1$  and  $\mathbf{V}_2$  should be orthogonal matrices, while  $\Sigma_0$ ,  $\Sigma_1$  and  $\Sigma_2$  should be ‘diagonal’ matrices. A possible method for solving the underlying minimization problem is to use a conventional multivariate optimization tool as for example, the BFGS (Broyden–Fletcher–Goldfarb–Shanno) method, and to compute the derivatives of the objective function with respect to  $\alpha$  and  $\omega$  by using matrix calculus. The peculiarities of derivative calculations for generalized cross-validation, the quasi-optimality criterion and the maximum likelihood estimation are summarized below.

The selection criterion for generalized cross-validation reads as

$$(\alpha_{\text{gcv}}, \omega_{\text{gcv}}) = \arg \min_{\alpha, \omega} v_{\alpha\omega}^\delta, \quad (3.130)$$

where the multi-parameter generalized cross-validation function is given by

$$v_{\alpha\omega}^\delta = \frac{\|\mathbf{r}_{\alpha\omega}^\delta\|^2}{\left[\text{trace}\left(\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}\right)\right]^2}.$$

Setting

$$\mathbf{M}_{\alpha\omega} = \mathbf{K}^T\mathbf{K} + \alpha\mathbf{L}_\omega^T\mathbf{L}_\omega,$$

and noting that  $\mathbf{K}_{\alpha\omega}^\dagger = \mathbf{M}_{\alpha\omega}^{-1}\mathbf{K}^T$ , we compute the partial derivatives of the residual and the trace term as follows:

$$\frac{\partial}{\partial \lambda} \|\mathbf{r}_{\alpha\omega}^\delta\|^2 = 2\mathbf{y}^{\delta T} \left(\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}\right)^T \frac{\partial}{\partial \lambda} \left(\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}\right) \mathbf{y}^\delta$$

and

$$\frac{\partial}{\partial \lambda} \text{trace} \left( \mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega} \right) = \text{trace} \left( \frac{\partial}{\partial \lambda} \left( \mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega} \right) \right).$$

Here,

$$\frac{\partial}{\partial \lambda} \left( \mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega} \right) = \mathbf{K}_{\alpha\omega}^{\dagger T} \frac{\partial \mathbf{M}_{\alpha\omega}}{\partial \lambda} \mathbf{K}_{\alpha\omega}^{\dagger}, \quad (3.131)$$

where the variable  $\lambda$  stands for  $\alpha$  and  $\omega$ .

The quasi-optimality criterion uses the selection rule

$$(\alpha_{\text{qo}}, \omega_{\text{qo}}) = \arg \min_{\alpha, \omega} \varsigma_{\alpha\omega}^{\delta}, \quad (3.132)$$

where

$$\varsigma_{\alpha\omega}^{\delta} = \left\| (\mathbf{A}_{\alpha\omega} - \mathbf{I}_n) \mathbf{K}_{\alpha\omega}^{\dagger} \mathbf{y}^{\delta} \right\|^2.$$

The derivatives of the quasi-optimality function read as

$$\frac{\partial \varsigma_{\alpha\omega}^{\delta}}{\partial \lambda} = 2 \mathbf{y}^{\delta T} \mathbf{K}_{\alpha\omega}^{\dagger T} (\mathbf{A}_{\alpha\omega} - \mathbf{I}_n)^T \frac{\partial}{\partial \lambda} [(\mathbf{A}_{\alpha\omega} - \mathbf{I}_n) \mathbf{K}_{\alpha\omega}^{\dagger}] \mathbf{y}^{\delta}$$

with

$$\frac{\partial}{\partial \lambda} [(\mathbf{A}_{\alpha\omega} - \mathbf{I}_n) \mathbf{K}_{\alpha\omega}^{\dagger}] = \frac{\partial \mathbf{K}_{\alpha\omega}^{\dagger}}{\partial \lambda} \mathbf{K} \mathbf{K}_{\alpha\omega}^{\dagger} + (\mathbf{A}_{\alpha\omega} - \mathbf{I}_n) \frac{\partial \mathbf{K}_{\alpha\omega}^{\dagger}}{\partial \lambda}$$

and

$$\frac{\partial \mathbf{K}_{\alpha\omega}^{\dagger}}{\partial \lambda} = -\mathbf{M}_{\alpha\omega}^{-1} \frac{\partial \mathbf{M}_{\alpha\omega}}{\partial \lambda} \mathbf{K}_{\alpha\omega}^{\dagger}.$$

The regularization parameter and the weighting factor for the maximum likelihood estimation are given by

$$(\alpha_{\text{ml}}, \omega_{\text{ml}}) = \arg \min_{\alpha, \omega} \lambda_{\alpha\omega}^{\delta}, \quad (3.133)$$

where

$$\lambda_{\alpha\omega}^{\delta} = \frac{\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}) \mathbf{y}^{\delta}}{\sqrt[m]{\det (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega})}}.$$

To compute the partial derivatives of  $\lambda_{\alpha\omega}^{\delta}$  we have to calculate the derivatives of the determinant of the matrix  $\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}$ . For this purpose, we may use Jacobi's formula

$$\frac{\partial}{\partial \lambda} \det (\mathbf{A}) = \text{trace} \left( \text{adj} (\mathbf{A}) \frac{\partial \mathbf{A}}{\partial \lambda} \right),$$

where  $\text{adj} (\mathbf{A})$  is the adjugate of the square matrix  $\mathbf{A}$ . We obtain

$$\frac{\partial}{\partial \lambda} \det (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}) = \det (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}) \text{trace} \left( (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega})^{-1} \frac{\partial}{\partial \lambda} (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}) \right),$$

where the derivatives of the matrix  $\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha\omega}$  are given by (3.131).

The minimization method based on matrix calculus is of general use because it can handle situations with multiple regularization parameters. However, the memory requirement is excessively large and the calculation might be inaccurate, e.g., for small values of  $\alpha$ , the calculation of the inverse  $\mathbf{M}_{\alpha\omega}^{-1}$  is an unstable process due to the large condition number of  $\mathbf{M}_{\alpha\omega}$ . For two-parameter regularization problems, the use of a semi-discrete minimization method seems to be more appropriate. In this approach, we consider a discrete set of weighting factors  $\{\omega_j\}$ , and for each  $\omega_j$ , we use the generalized singular value decomposition of  $(\mathbf{K}, \mathbf{L}_{\omega_j})$  to solve the corresponding one-dimensional minimization problem.

An alternative strategy proposed by Brezinski et al. (2003) is to approximate the multi-parameter solution  $\mathbf{x}_{\alpha_1\alpha_2}^\delta$ , minimizing the Tikhonov function

$$\mathcal{F}_{\alpha_1\alpha_2}(\mathbf{x}) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha_1 \|\mathbf{H}_1\mathbf{x}\|^2 + \alpha_2 \|\mathbf{H}_2\mathbf{x}\|^2,$$

by a linear combination of the one-parameter solutions  $\mathbf{x}_{\alpha_1}^\delta$  and  $\mathbf{x}_{\alpha_2}^\delta$ , minimizing the Tikhonov functions

$$\mathcal{F}_{\alpha_i}(\mathbf{x}) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha_i \|\mathbf{H}_i\mathbf{x}\|^2, \quad i = 1, 2. \quad (3.134)$$

The regularized solutions solve the corresponding normal equations, and we have

$$(\mathbf{K}^T\mathbf{K} + \alpha_1\mathbf{H}_1^T\mathbf{H}_1 + \alpha_2\mathbf{H}_2^T\mathbf{H}_2) \mathbf{x}_{\alpha_1\alpha_2}^\delta = \mathbf{K}^T\mathbf{y}^\delta, \quad (3.135)$$

$$(\mathbf{K}^T\mathbf{K} + \alpha_1\mathbf{H}_1^T\mathbf{H}_1) \mathbf{x}_{\alpha_1}^\delta = \mathbf{K}^T\mathbf{y}^\delta, \quad (3.136)$$

$$(\mathbf{K}^T\mathbf{K} + \alpha_2\mathbf{H}_2^T\mathbf{H}_2) \mathbf{x}_{\alpha_2}^\delta = \mathbf{K}^T\mathbf{y}^\delta. \quad (3.137)$$

Inserting (3.135), (3.136) and (3.137) in the identity

$$\mathbf{K}^T\mathbf{y}^\delta = \xi\mathbf{K}^T\mathbf{y}^\delta + (1 - \xi)\mathbf{K}^T\mathbf{y}^\delta, \quad 0 \leq \xi \leq 1,$$

and setting

$$\mathbf{M}_{\alpha_1\alpha_2} = \mathbf{K}^T\mathbf{K} + \alpha_1\mathbf{H}_1^T\mathbf{H}_1 + \alpha_2\mathbf{H}_2^T\mathbf{H}_2,$$

yields the representation

$$\mathbf{x}_{\alpha_1\alpha_2}^\delta = \mathbf{x}_{\alpha_1\alpha_2}^{\delta*} - \mathbf{M}_{\alpha_1\alpha_2}^{-1}\boldsymbol{\rho}_{\alpha_1\alpha_2}(\xi),$$

with

$$\mathbf{x}_{\alpha_1\alpha_2}^{\delta*} = \xi\mathbf{x}_{\alpha_1}^\delta + (1 - \xi)\mathbf{x}_{\alpha_2}^\delta,$$

and

$$\boldsymbol{\rho}_{\alpha_1\alpha_2}(\xi) = \xi\alpha_2\mathbf{H}_2^T\mathbf{H}_2\mathbf{x}_{\alpha_1}^\delta + (1 - \xi)\alpha_1\mathbf{H}_1^T\mathbf{H}_1\mathbf{x}_{\alpha_2}^\delta.$$

As the minimization of the error between  $\mathbf{x}_{\alpha_1\alpha_2}^\delta$  and  $\mathbf{x}_{\alpha_1\alpha_2}^{\delta*}$  would involve the inverse of the matrix  $\mathbf{M}_{\alpha_1\alpha_2}$  (leading to a considerable computational effort), the choice proposed by Brezinski et al. (2003) is to take  $\xi$  as the minimizer of  $\|\boldsymbol{\rho}_{\alpha_1\alpha_2}\|^2$ , that is,

$$\xi = \frac{\mathbf{q}_2^T(\mathbf{q}_2 - \mathbf{q}_1)}{\|\mathbf{q}_2 - \mathbf{q}_1\|^2},$$

with

$$\mathbf{q}_1 = \alpha_2\mathbf{H}_2^T\mathbf{H}_2\mathbf{x}_{\alpha_1}^\delta, \quad \mathbf{q}_2 = \alpha_1\mathbf{H}_1^T\mathbf{H}_1\mathbf{x}_{\alpha_2}^\delta.$$

The solutions  $\mathbf{x}_{\alpha_1}^\delta$  and  $\mathbf{x}_{\alpha_2}^\delta$  are then computed by using the corresponding generalized singular systems, and  $\mathbf{x}_{\alpha_1\alpha_2}^\delta$  is approximated by  $\mathbf{x}_{\alpha_1\alpha_2}^{\delta*}$ . More precisely, in the aforementioned regularization parameter choice methods, the residual

$$\mathbf{r}_{\alpha_1\alpha_2}^\delta = \mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{\alpha_1\alpha_2}^\delta$$

is replaced by

$$\mathbf{r}_{\alpha_1\alpha_2}^{\delta*} = \mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{\alpha_1\alpha_2}^{\delta*},$$

the influence matrix  $\hat{\mathbf{A}}_{\alpha_1\alpha_2}$ , satisfying  $\mathbf{K}\mathbf{x}_{\alpha_1\alpha_2}^\delta = \hat{\mathbf{A}}_{\alpha_1\alpha_2}\mathbf{y}^\delta$ , by its approximation

$$\hat{\mathbf{A}}_{\alpha_1\alpha_2}^* = \xi \hat{\mathbf{A}}_{\alpha_1} + (1 - \xi) \hat{\mathbf{A}}_{\alpha_2},$$

defined through the relation  $\mathbf{K}\mathbf{x}_{\alpha_1\alpha_2}^{\delta*} = \hat{\mathbf{A}}_{\alpha_1\alpha_2}^*\mathbf{y}^\delta$ , and the averaging kernel matrix  $\mathbf{A}_{\alpha_1\alpha_2}$ , satisfying  $\mathbf{x}_{\alpha_1\alpha_2} = \mathbf{A}_{\alpha_1\alpha_2}\mathbf{x}^\dagger$ , by its approximation

$$\mathbf{A}_{\alpha_1\alpha_2}^* = \xi \mathbf{A}_{\alpha_1} + (1 - \xi) \mathbf{A}_{\alpha_2},$$

defined through the relation  $\mathbf{x}_{\alpha_1\alpha_2}^* = \mathbf{A}_{\alpha_1\alpha_2}^*\mathbf{x}^\dagger$ . It is remarkable to note that in the framework of the generalized cross-validation method and under some additional assumptions, Brezinski et al. (2003) have shown that

$$(\alpha_{1\text{gcv}}, \alpha_{2\text{gcv}}) = \arg \min_{\alpha_1, \alpha_2} v_{\alpha_1\alpha_2}^{\delta*} \approx \arg \min_{\alpha_1, \alpha_2} \left( \sqrt{v_{\alpha_1}^\delta} + \sqrt{v_{\alpha_2}^\delta} \right)^2,$$

which means that this technique corresponds to the simple approach of choosing  $\alpha_1$  and  $\alpha_2$  by applying separately the generalized cross-validation method to each of the one-parameter regularization problems (3.134).

The regularization parameters can also be computed in a generalized L-curve framework by using the concept of the L-surface (Belge et al., 2002). The L-surface components are defined by

$$\begin{aligned} x(\alpha_1, \alpha_2) &= \log \left( \|\mathbf{c}_{1\alpha_1\alpha_2}^\delta\|^2 \right), \\ y(\alpha_1, \alpha_2) &= \log \left( \|\mathbf{c}_{2\alpha_1\alpha_2}^\delta\|^2 \right), \\ z(\alpha_1, \alpha_2) &= \log \left( \|\mathbf{r}_{\alpha_1\alpha_2}^\delta\|^2 \right), \end{aligned}$$

where the constraint vectors are given by  $\mathbf{c}_{i\alpha_1\alpha_2}^\delta = \mathbf{H}_i\mathbf{x}_{\alpha_1\alpha_2}^\delta$ ,  $i = 1, 2$ . The ‘generalized corner’ of the L-surface is the point on the surface around which the surface is maximally wrapped and can be defined as the point maximizing the Gaussian curvature. The Gaussian curvature can be computed given the first- and the second-order partial derivatives of  $z$  with respect to  $x$  and  $y$ . Because this calculation is very time-consuming, the so-called minimum distance function approach can be used instead. The distance function is defined by

$$d(\alpha_1, \alpha_2)^2 = [x(\alpha_1, \alpha_2) - x_0]^2 + [y(\alpha_1, \alpha_2) - y_0]^2 + [z(\alpha_1, \alpha_2) - z_0]^2,$$

where  $x_0$ ,  $y_0$  and  $z_0$  are the coordinates of a properly chosen origin, and the regularization parameters are chosen as

$$(\alpha_{11s}, \alpha_{21s}) = \arg \min_{\alpha_1, \alpha_2} d(\alpha_1, \alpha_2)^2.$$

### 3.8.2 Incomplete multi-parameter regularization methods

For this type of regularization, the parameter choice methods should minimize some measure of the solution error corresponding to the first component of the state vector. As the noisy data vector accounts for both contributions of the state vector components, which cannot be separated (cf. (3.125)), regularization parameter choice methods based on the analysis of the residual or the noisy data cannot be applied. Possible candidates for incomplete multi-parameter regularization are the expected error estimation method, the quasi-optimality criterion, and, with some reticence, the L-curve method.

In the expected error estimation method with a semi-discrete minimization approach, we consider the expected value of the first error component. Specifically, for a discrete set of weighting factors  $\{\omega_j\}$ , we compute the optimal regularization parameter and weighting factor as

$$(\bar{\alpha}_{\text{opt}}, \bar{\omega}_{\text{opt}}) = \arg \min_{\alpha, \omega_j} \mathcal{E} \left\{ \left\| \mathbf{e}_{1\alpha\omega_j}^\delta \right\|^2 \right\}, \quad (3.138)$$

where we have assumed the partition

$$\mathbf{e}_{\alpha\omega}^\delta = \begin{bmatrix} \mathbf{e}_{1\alpha\omega}^\delta \\ \mathbf{e}_{2\alpha\omega}^\delta \end{bmatrix}.$$

To solve the one-dimensional minimization problem with respect to the regularization parameter, we need analytical representations for the error components  $\|\mathbf{e}_{s1\alpha\omega}\|^2$  and  $\mathcal{E}\{\|\mathbf{e}_{n1\alpha\omega}^\delta\|^2\}$  as in (3.33) and (3.41), respectively. If  $(\gamma_{\omega i}; \mathbf{w}_{\omega i}, \mathbf{u}_{\omega i}, \mathbf{v}_{\omega i})$  is a generalized singular system of  $(\mathbf{K}, \mathbf{L}_\omega)$ , the required expansions take the forms

$$\mathbf{e}_{s1\alpha\omega} = \sum_{i=1}^n \frac{\alpha}{\gamma_{\omega i}^2 + \alpha} \frac{1}{\sigma_{\omega i}} (\mathbf{u}_{\omega i}^T \mathbf{y}) \mathbf{w}_{1\omega i},$$

and

$$\mathcal{E} \left\{ \left\| \mathbf{e}_{n1\alpha\omega}^\delta \right\|^2 \right\} = \sigma^2 \sum_{i=1}^n \left( \frac{\gamma_{\omega i}^2}{\gamma_{\omega i}^2 + \alpha} \frac{1}{\sigma_{\omega i}} \right)^2 \|\mathbf{w}_{1\omega i}\|^2, \quad (3.139)$$

where

$$\mathbf{w}_{\omega i} = \begin{bmatrix} \mathbf{w}_{1\omega i} \\ \mathbf{w}_{2\omega i} \end{bmatrix}.$$

The steps of a two-component expected error estimation method can be summarized as follows:

- (1) choose a discrete set of weighting factors  $\{\omega_j\}_{j=1, \overline{N_\omega}}$ , and generate a set of state vectors  $\{\mathbf{x}_i^\dagger\}_{i=1, \overline{N_x}}$  in a random manner;
- (2) for each state vector  $\mathbf{x}_i^\dagger$ , compute the optimal regularization parameter and weighting factor

$$(\bar{\alpha}_{\text{opt}i}, \bar{\omega}_{\text{opt}i}) = \arg \min_{\alpha, \omega_j} \mathcal{E} \left\{ \left\| \mathbf{e}_{1\alpha\omega_j}^\delta (\mathbf{x}_i^\dagger) \right\|^2 \right\},$$

and store the weighting-factor index  $j_i^*$  defined as  $\bar{\omega}_{\text{opt}i} = \omega_{j_i^*}$ ;

(3) count the number of appearances of the index  $j$  over state vector realizations,

$$N_j = \sum_{i \in I_j} 1, \quad I_j = \{i / j_i^* = j\},$$

and determine the index  $\bar{j}$  with maximum frequency of appearance,

$$\bar{j} = \arg \max_j N_j;$$

(4) compute the exponent

$$p_i = \frac{\log \bar{\alpha}_{\text{opt}i}}{\log \sigma}$$

for all  $i \in I_{\bar{j}}$ , and the sample mean exponent

$$\bar{p} = \frac{1}{N_{\bar{j}}} \sum_{i \in I_{\bar{j}}} p_i;$$

(5) set  $\alpha_e = \sigma^{\bar{p}}$  and  $\omega_e = \omega_{\bar{j}}$ .

The regularization parameter and weighting factor for the quasi-optimality criterion are defined by

$$(\alpha_{\text{qo}}, \omega_{\text{qo}}) = \arg \min_{\alpha, \omega_j} \varsigma_{1\alpha\omega_j}^{\delta},$$

where

$$\varsigma_{1\alpha\omega}^{\delta} = \left\| \alpha \frac{\partial \mathbf{x}_{1\alpha\omega}^{\delta}}{\partial \alpha} \right\|^2$$

and

$$\alpha \frac{\partial \mathbf{x}_{1\alpha\omega}^{\delta}}{\partial \alpha} = - \sum_{i=1}^n \frac{\alpha \gamma_{\omega i}^2}{(\gamma_{\omega i}^2 + \alpha)^2} \frac{1}{\sigma_{\omega i}} (\mathbf{u}_{\omega i}^T \mathbf{y}^{\delta}) \mathbf{w}_{1\omega i}.$$

The difficulty associated with this selection criterion is that the quasi-optimality function may have several minima, which are difficult to locate.

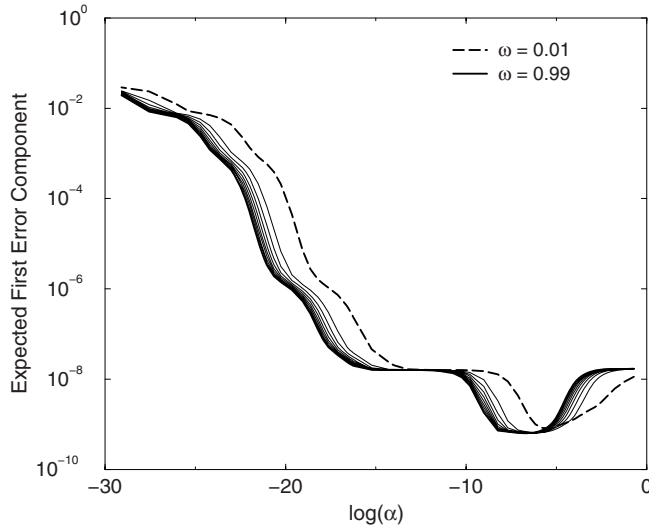
A heuristic regularization parameter choice rule can be designed by combining the L-curve method with the minimum distance function approach. The idea is to consider an  $\omega$ -dependent family of L-curves, and for each L-curve to determine the regularization parameter by maximizing its curvature. The final values of  $\omega$  and  $\alpha$  are then computed by selecting the point with minimum residual and constraint norms. Thus, for each  $\omega_j$ , we consider the L-curve of components

$$x_j(\alpha) = \log \left( \left\| \mathbf{r}_{\alpha\omega_j}^{\delta} \right\|^2 \right), \quad y_j(\alpha) = \log \left( \left\| \mathbf{c}_{1\alpha\omega_j}^{\delta} \right\|^2 \right),$$

and determine the value of the regularization parameter that maximizes the curvature function  $\kappa_{1\text{c}\alpha j}^{\delta}$ ,

$$\alpha_{1\text{c}j} = \arg \max_{\alpha} \kappa_{1\text{c}\alpha j}^{\delta}.$$





**Fig. 3.18.** Expected value of the first error component  $\mathcal{E}\{\|\mathbf{e}_{1\alpha\omega}^\delta\|^2\}$  for the noise standard deviation  $\sigma = 5 \cdot 10^{-2}$  and one state vector realization.

Defining the distance

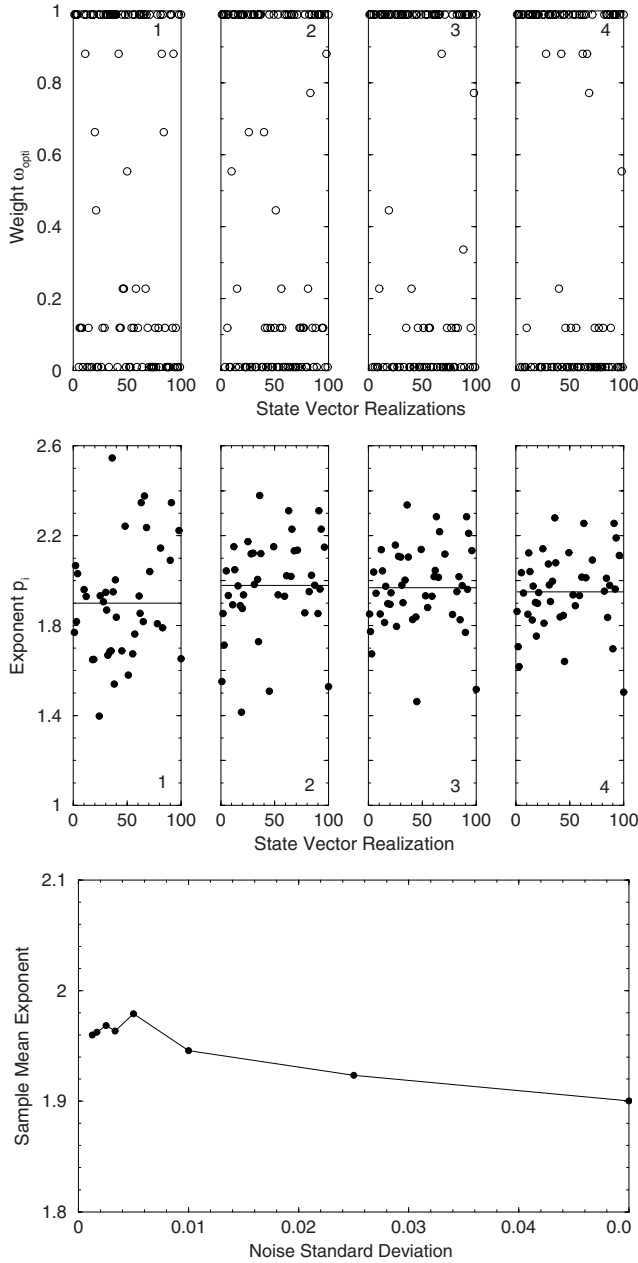
$$d_j^2 = [x_j(\alpha_{1cj}) - x_0]^2 + [y_j(\alpha_{1cj}) - y_0]^2,$$

we compute  $j^* = \arg \min_j d_j^2$ , and set  $\alpha_{1c} = \alpha_{1cj^*}$  and  $\omega_{1c} = \omega_{j^*}$ .

A numerical example dealing with a BrO retrieval test problem may clarify the peculiarities of multi-parameter regularization methods. The retrieval scenario is similar to that considered in section 3.7. The spectral domain of analysis ranges between 337 and 357 nm and in addition to BrO,  $O_3$  is considered as an active gas. The first component of the state vector is the BrO profile, while the second component is the  $O_3$  profile. The discrete set of weighting factors consists of 10 equidistant values between 0.01 and 0.99.

In the expected error estimation method, we generate 100 Gaussian profiles with a correlation length  $l = 3.5$  km and a profile standard deviation  $\sigma_x = 0.4$ . The expected value of the first error component is shown in Figure 3.18 for different values of the weighting factor  $\omega$ . The plots illustrate that the minimum value of  $\mathcal{E}\{\|\mathbf{e}_{1\alpha\omega}^\delta\|^2\}$  with respect to  $\alpha$  does not vary significantly with  $\omega$ . As a result, we may expect that the selection of the weighting factor is not so critical for the inversion process.

In the top panel of Figure 3.19 we plot the optimal weighting factors for error estimation  $\bar{\omega}_{opti}$  for different state vector realizations. The weighting factor with the maximum frequency of appearance is independent of the noise standard deviation  $\sigma$ , and its value is  $\omega_{\bar{j}} = 0.99$ . It should be pointed out that the frequencies of appearance of the weighting factors 0.01 and 0.99 are similar, and these situations correspond to a regularization of one gas species only. Considering the subset  $I_{\bar{j}}$  of all state vector realizations related to the weighting factor with maximum frequency of appearance  $\omega_{\bar{j}}$ , we plot in the middle panel of Figure 3.19 the exponent  $p_i$  for  $i \in I_{\bar{j}}$ . The values of the sample mean exponent are  $\bar{p} = 1.90$  for  $\sigma = 5 \cdot 10^{-2}$ ,  $\bar{p} = 1.97$  for  $\sigma = 5 \cdot 10^{-3}$ ,  $\bar{p} = 1.96$  for  $\sigma = 2.5 \cdot 10^{-3}$ ,



**Fig. 3.19.** Top: optimal weighting factors  $\bar{\omega}_{\text{opt},i}$  for the following values of the noise standard deviation  $\sigma$ :  $5 \cdot 10^{-2}$  (1),  $5 \cdot 10^{-3}$  (2),  $2.5 \cdot 10^{-3}$  (3), and  $1.25 \cdot 10^{-3}$  (4). Middle: exponent  $p_i$  for the state vector realizations corresponding to  $\omega_{\bar{j}} = 0.99$ . Bottom: sample mean exponent as a function of the noise standard deviation  $\sigma$ .

and  $\bar{p} = 1.94$  for  $\sigma = 1.25 \cdot 10^{-3}$ . The sample mean exponent as a function of the noise standard deviation is shown in the bottom panel of Figure 3.19. It is apparent that  $\bar{p}$  does not vary significantly with  $\sigma$ , and its average value is about 1.95.

Next, we choose the exact state vectors  $\mathbf{x}_1^\dagger$  and  $\mathbf{x}_2^\dagger$  as translated and scaled climatological profiles with a translation distance of 2 km and a scaling factor of 1.3, and generate 50 noisy data vectors  $\mathbf{y}_i^\delta$  with the white noise  $\delta \sim N(0, \sigma^2 \mathbf{I}_m)$ . In Figure 3.20 we plot the solution errors for the expected error estimation method with  $\alpha_e = \sigma^{1.95}$  and  $\omega_e = 0.99$ ,

$$\varepsilon_{ei} = \frac{\|\mathbf{x}_{1\alpha_e\omega_e i}^\delta - \mathbf{x}_1^\dagger\|}{\|\mathbf{x}_1^\dagger\|}, \quad \mathbf{x}_{1\alpha_e\omega_e i}^\delta = [\mathbf{K}_{\alpha_e\omega_e}^\dagger \mathbf{y}_i^\delta]_1,$$

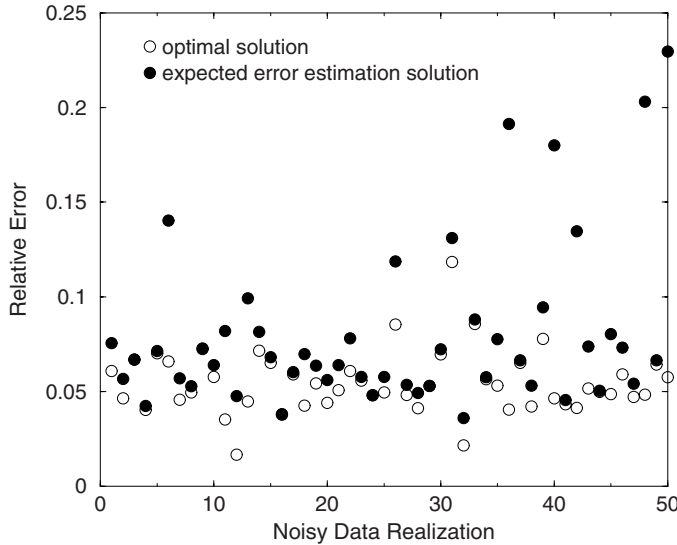
together with the solution errors

$$\varepsilon_{opti} = \frac{\|\mathbf{x}_{1\alpha_{opti}\omega_{opti}}^\delta - \mathbf{x}_1^\dagger\|}{\|\mathbf{x}_1^\dagger\|},$$

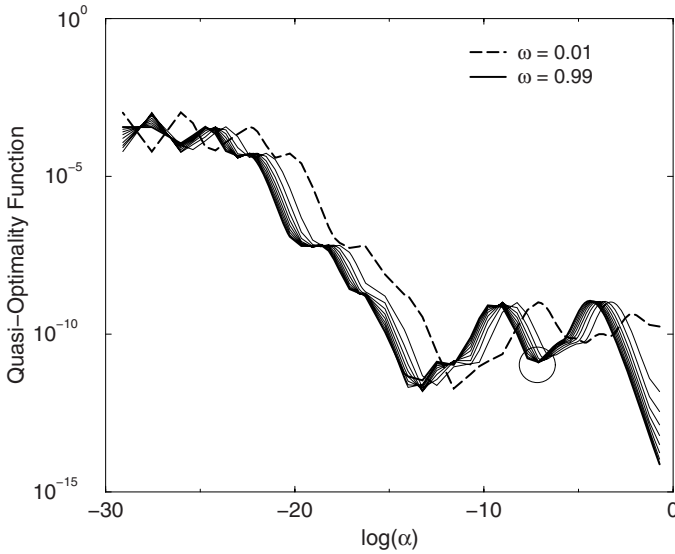
corresponding to the optimal parameters,

$$(\alpha_{opti}, \omega_{opti}) = \arg \min_{\alpha, \omega_j} \left\| [\mathbf{K}_{\alpha\omega_j}^\dagger \mathbf{y}_i^\delta]_1 - \mathbf{x}_1^\dagger \right\|^2.$$

The solution errors for the expected error estimation method are in general comparable with the errors in the optimal solution, but for some noisy data realizations, the errors may exceed 40%.



**Fig. 3.20.** Relative errors in the expected error estimation solution and the optimal solution. The noise standard deviation is  $\sigma = 5 \cdot 10^{-2}$  and 50 noisy data realizations are considered.



**Fig. 3.21.** Expected quasi-optimality function  $\mathcal{E}\{\varsigma_{1\alpha\omega}^\delta\}$  for the noise standard deviation  $\sigma = 5 \cdot 10^{-2}$  and one state vector realization. The circle indicates the minimizer of the expected value of the first error component  $\mathcal{E}\{\|\mathbf{e}_{1\alpha\omega}^\delta\|^2\}$ .

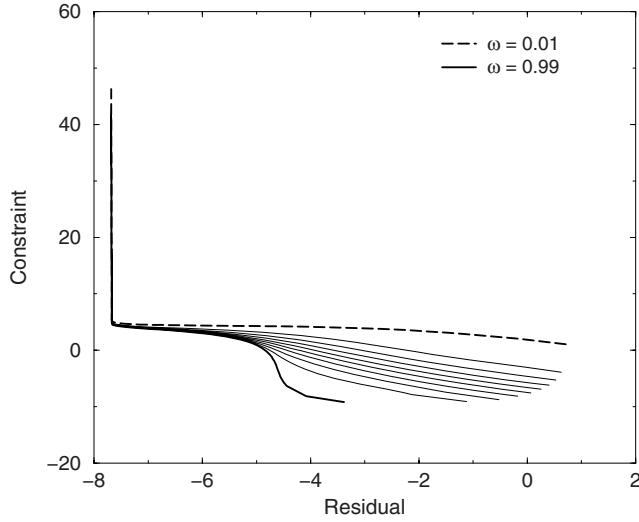
The expected quasi-optimality function  $\mathcal{E}\{\varsigma_{1\alpha\omega}^\delta\}$  is shown in Figure 3.21 for different values of the weighting factor  $\omega$ . The plots evidence that the minimizer of the expected value of the first error component is only a local minimizer and not a global minimizer of the expected quasi-optimality function. This fact disqualifies the quasi-optimality criterion for the present application.

In Figure 3.22 we illustrate the  $\omega$ -dependent family of expected L-curves. The results show that the plateau  $C_{1\omega}(\alpha) = \mathcal{E}\{\|\mathbf{c}_{1\alpha\omega}^\delta\|^2\}$  decreases with increasing  $\omega$ , and for  $\omega = 0.99$ , we obtain a corner with a small constraint norm. Thus, the expected L-curve method and the expected error estimation method predict the same value of the weighting factor.

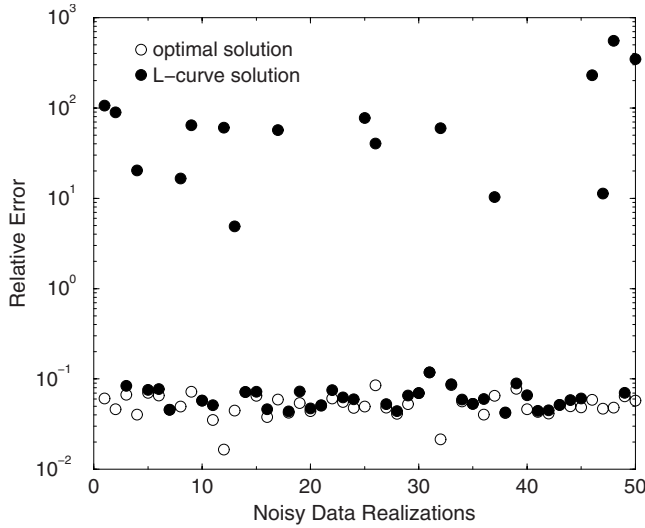
In Figure 3.23 we plot the relative errors in the L-curve solution and the optimal solution for 50 realizations of the noisy data vector. The main drawback of the L-curve method is that in some situations, the L-curve loses its L-shape and the estimation of the regularization parameter is erroneous. In contrast to the expected error estimation method, the failure of the L-curve method is accompanied by extremely large solution errors.

### 3.9 Mathematical results and further reading

Convergence and convergence rate results for Tikhonov regularization with different regularization parameter choice methods can be found in Engl et al. (2000), Groetsch (1984) and Rieder (2003). In a deterministic setting and for a continuous problem given by the operator equation  $Kx = y^\delta$ , the parameter choice rule with  $\alpha = \alpha(\Delta, y^\delta)$  is said to be



**Fig. 3.22.** Expected L-curves for the noise standard deviation  $\sigma = 5 \cdot 10^{-2}$  and one state vector realization.



**Fig. 3.23.** Relative errors in the L-curve solution and the optimal solution. The noise standard deviation is  $\sigma = 5 \cdot 10^{-2}$  and 50 noisy data realizations are considered.

convergent if

$$\left\| x_{\alpha(\Delta, y^\delta)}^\delta - x^\dagger \right\| \rightarrow 0 \text{ as } \Delta \rightarrow 0. \quad (3.140)$$

A regularization method together with a convergent parameter choice rule is called a convergent regularization method. The rate of convergence of a parameter choice method is expressed in terms of the rate with which the regularized solution  $x_{\alpha(\Delta, y^\delta)}^\delta$  converges to  $x^\dagger$

as the noise level  $\Delta$  tends to zero. Rates for regularization parameter choice methods are given under an additional assumption which concerns the smoothness of the solution  $x^\dagger$ . For the so-called Hölder-type source condition

$$x^\dagger = (K^* K)^\mu z, \quad (3.141)$$

with  $\mu > 0$  and  $z \in X$ , a regularization parameter choice method is said to be of optimal order, if the estimate

$$\|x^\dagger - x_\alpha^\delta\| = O\left(\|z\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad \Delta \rightarrow 0, \quad (3.142)$$

holds. A deterministic analysis of the general regularization method

$$\mathbf{x}_\alpha^\delta = g_\alpha(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta, \quad (3.143)$$

in a discrete setting and for the choice  $\mathbf{L} = \mathbf{I}_n$  is given in Appendix C. The function  $g_\alpha$  is related to the filter function  $f_\alpha$  by the relation

$$f_\alpha(\lambda) = \lambda g_\alpha(\lambda),$$

and for  $\mathbf{K} = \mathbf{U}\Sigma\mathbf{V}^T$ , the matrix function in (3.143) should be understood as

$$g_\alpha(\mathbf{K}^T \mathbf{K}) = \mathbf{V} \left[ \text{diag}(g_\alpha(\sigma_i^2))_{n \times n} \right] \mathbf{V}^T. \quad (3.144)$$

In particular, Tikhonov regularization and its iterated version are characterized by the choices

$$g_\alpha(\lambda) = \frac{1}{\lambda + \alpha},$$

and

$$g_\alpha(\lambda) = \frac{1}{\lambda} \left[ 1 - \left( \frac{\alpha}{\lambda + \alpha} \right)^p \right],$$

respectively. The conclusions of this analysis are as follows:

- (1) the a priori parameter choice method  $\alpha = (\Delta/\|\mathbf{z}\|)^{2/(2\mu+1)}$ , the generalized discrepancy principle and the generalized residual curve method are of optimal order for  $0 < \mu \leq \mu_0$ ;
- (2) the discrepancy principle and the residual curve method are of optimal order for  $0 < \mu \leq \mu_0 - 1/2$ .

The index  $\mu_0$  is the qualification of the regularization method, and we have  $\mu_0 = 1$  for Tikhonov regularization and  $\mu_0 = p$  for the  $p$ -times iterated Tikhonov regularization. Thus, in the case of Tikhonov regularization, the best convergence rate which can be achieved by the first group of methods is  $O(\Delta^{2/3})$ , while  $O(\Delta^{1/2})$  is the best convergence rate of the discrepancy principle and the residual curve method.

Although the formulations of regularization parameter choice methods in a deterministic and a semi-stochastic setting are very similar, the convergence analyses differ significantly. In a semi-stochastic setting we are dealing with the semi-discrete data model

$$\mathbf{y}_m^\delta = K_m x + \boldsymbol{\delta}_m,$$

where  $K_m$  is a linear operator between the state space  $X$  and the finite-dimensional Euclidean space  $\mathbb{R}^m$ , and  $\delta_m$  is an  $m$ -dimensional vector whose components are each a random variable. If the noise components are assumed to be uncorrelated with zero mean and common variance  $\sigma^2$ , the analysis is carried out under the assumptions that  $\sigma$  is fixed and that  $m$  tends to infinity. In this regard, denoting by  $x_\alpha^\delta$  the minimizer of the Tikhonov functional

$$\mathcal{F}_{m\alpha}(x) = \|\mathbf{y}_m^\delta - K_m x\|^2 + \alpha \|x\|^2,$$

a regularization parameter choice method is said to be convergent if it yields a parameter  $\alpha = \alpha(m)$  with the property

$$\mathcal{E} \left\{ \left\| x^\dagger - x_{\alpha(m)}^\delta \right\|^2 \right\} \rightarrow 0 \text{ as } m \rightarrow \infty.$$

This type of convergence is often referred to as convergence in mean square. In addition to convergence, other concepts have been introduced to quantify the optimality properties of regularization parameter choice methods (Vogel, 2002). To be more concrete, if  $\bar{\alpha}_{\text{opt}}(m)$  is the optimal regularization parameter for error estimation, then, a regularization parameter choice method yielding an expected parameter  $\bar{\alpha}(m)$  is called

- (1) e-optimal if there exists  $m_0$  so that  $\bar{\alpha}(m) = \bar{\alpha}_{\text{opt}}(m)$ , whenever  $m \geq m_0$ ;
- (2) asymptotically e-optimal if  $\bar{\alpha}(m) \approx \bar{\alpha}_{\text{opt}}(m)$  as  $m \rightarrow \infty$ ;
- (3) order e-optimal if there exists a positive constant  $r$ , called the order constant, so that  $\bar{\alpha}(m) \approx r \bar{\alpha}_{\text{opt}}(m)$  as  $m \rightarrow \infty$ .

A pertinent analysis of regularization parameter choice methods by assuming specific decay rates for the singular values of the semi-discrete operator and for the Fourier coefficients has been given by Vogel (2002). For Tikhonov regularization, the proofs are extremely technical, but the results can be summarized as follows:

- (1) the discrepancy principle, the unbiased predictive risk estimator method, and generalized cross-validation are convergent;
- (2) the L-curve does not give a value of  $\alpha$  that yields mean square convergence, i.e., the L-curve method is non-convergent.

In fact, the convergence properties of the L-curve method has been studied by Hanke (1996) and Vogel (1996). In the first work, the problem is continuous and the analysis is carried out in a deterministic setting, while in the second work, the problem is semi-discrete and the analysis is performed in a semi-stochastic setting. As a consequence, the results established in these two papers are quite different. In a deterministic setting it is shown that the regularization parameter determined by the L-curve method decays too rapidly to zero as the noise level tends to zero. This behavior leads to an undersmoothing, which is more pronounced for small noise levels and very smooth solutions (see Figures 3.9 and 3.10). In a semi-stochastic setting, the regularization parameter computed by the L-curve method stagnates as  $m \rightarrow \infty$ , and for this reason, the regularized solution is oversmoothed. Despite these results, the L-curve method has been successfully used in numerous applications.

# 4

## Statistical inversion theory

The majority of retrieval approaches currently used in atmospheric remote sensing belong to the category of statistical inversion methods (Rodgers, 2000). The goal of this chapter is to reveal the similarity between classical regularization and statistical inversion regarding

- (1) the regularized solution representation,
- (2) the error analysis,
- (3) the design of one- and multi-parameter regularization methods.

In statistical inversion theory all variables included in the model are absolutely continuous random variables and the degree of information concerning their realizations is coded in probability densities. The solution of the inverse problem is the a posteriori density, which makes possible to compute estimates of the unknown atmospheric profile.

In the framework of Tikhonov regularization we have considered the linear data model

$$\mathbf{y}^\delta = \mathbf{K}\mathbf{x} + \boldsymbol{\delta}, \quad (4.1)$$

where  $\mathbf{y}^\delta$  is the noisy data vector and  $\boldsymbol{\delta}$  is the noise vector. In statistical inversion theory all parameters are viewed as random variables, and since in statistics random variables are denoted by capital letters and their realizations by lowercase letters, the stochastic version of the data model (4.1) is

$$\mathbf{Y}^\delta = \mathbf{K}\mathbf{X} + \boldsymbol{\Delta}. \quad (4.2)$$

The random vectors  $\mathbf{Y}^\delta$ ,  $\mathbf{X}$  and  $\boldsymbol{\Delta}$  represent the data, the state and the noise, respectively; their realizations are denoted by  $\mathbf{Y}^\delta = \mathbf{y}^\delta$ ,  $\mathbf{X} = \mathbf{x}$  and  $\boldsymbol{\Delta} = \boldsymbol{\delta}$ , respectively.

### 4.1 Bayes theorem and estimators

The data model (4.2) gives a relation between the three random vectors  $\mathbf{Y}^\delta$ ,  $\mathbf{X}$  and  $\boldsymbol{\Delta}$ , and therefore, their probability densities depend on each other. The following probability densities are relevant for our analysis:



- (1) the a priori density  $p_a(\mathbf{x})$ , which encapsulates our presumable information about  $\mathbf{X}$  before performing the measurement of  $\mathbf{Y}^\delta$ ;
- (2) the likelihood density  $p(\mathbf{y}^\delta | \mathbf{x})$ , which represents the conditional probability density of  $\mathbf{Y}^\delta$  given the state  $\mathbf{X} = \mathbf{x}$ ;
- (3) the a posteriori density  $p(\mathbf{x} | \mathbf{y}^\delta)$ , which represents the conditional probability density of  $\mathbf{X}$  given the data  $\mathbf{Y}^\delta = \mathbf{y}^\delta$ .

The choice of the a priori density  $p_a(\mathbf{x})$  is perhaps the most important part of the inversion process. Different a priori models yield different objective functions, and in particular, the classical regularization terms correspond to Gaussian a priori models. Gaussian densities are widely used in statistical inversion theory because they are easy to compute and often lead to explicit estimators. Besides Gaussian densities other types of a priori models, as for instance the Cauchy density and the entropy density can be found in the literature (Kaipio and Somersalo, 2005).

The construction of the likelihood density  $p(\mathbf{y}^\delta | \mathbf{x})$  depends on the noise assumption. The data model (4.2) operates with additive noise, but other explicit noise models including multiplicative noise models and models with an incompletely known forward model matrix can be considered. If the noise is additive and is independent of the atmospheric state, the probability density  $p_n(\boldsymbol{\delta})$  of  $\boldsymbol{\Delta}$  remains unchanged when conditioned on  $\mathbf{X} = \mathbf{x}$ . Thus,  $\mathbf{Y}^\delta$  conditioned on  $\mathbf{X} = \mathbf{x}$  is distributed like  $\boldsymbol{\Delta}$ , and the likelihood density becomes

$$p(\mathbf{y}^\delta | \mathbf{x}) = p_n(\mathbf{y}^\delta - \mathbf{K}\mathbf{x}). \quad (4.3)$$

Assuming that after analyzing the measurement setting and accounting of the additional information available about all variables we have found the joint probability density  $p(\mathbf{x}, \mathbf{y}^\delta)$  of  $\mathbf{X}$  and  $\mathbf{Y}^\delta$ , then the a priori density is given by

$$p_a(\mathbf{x}) = \int_{\mathbb{R}^m} p(\mathbf{x}, \mathbf{y}^\delta) d\mathbf{y}^\delta,$$

while the likelihood density and the a posteriori density can be expressed as

$$p(\mathbf{y}^\delta | \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y}^\delta)}{p_a(\mathbf{x})}, \quad (4.4)$$

and

$$p(\mathbf{x} | \mathbf{y}^\delta) = \frac{p(\mathbf{x}, \mathbf{y}^\delta)}{p(\mathbf{y}^\delta)}, \quad (4.5)$$

respectively.

The following result known as the Bayes theorem of inverse problems relates the a posteriori density to the likelihood density (cf. (4.4) and (4.5)):

$$p(\mathbf{x} | \mathbf{y}^\delta) = \frac{p(\mathbf{y}^\delta | \mathbf{x}) p_a(\mathbf{x})}{p(\mathbf{y}^\delta)}. \quad (4.6)$$

In (4.6), the marginal density  $p(\mathbf{y}^\delta)$  computed as

$$p(\mathbf{y}^\delta) = \int_{\mathbb{R}^n} p(\mathbf{x}, \mathbf{y}^\delta) d\mathbf{x} = \int_{\mathbb{R}^n} p(\mathbf{y}^\delta | \mathbf{x}) p_a(\mathbf{x}) d\mathbf{x},$$

plays the role of a normalization constant and is usually ignored. However, as we will see, this probability density is of particular importance in the design of regularization parameter choice methods.

The knowledge of the a posteriori density allows the calculation of different estimators and spreads of solution. A popular statistical estimator is the maximum a posteriori estimator

$$\hat{\mathbf{x}}_{\text{map}} = \arg \max_{\mathbf{x}} p(\mathbf{x} | \mathbf{y}^\delta),$$

and the problem of finding the maximum a posteriori estimator requires the solution of an optimization problem. Another estimator is the conditional mean of  $\mathbf{X}$  conditioned on the data  $\mathbf{Y}^\delta = \mathbf{y}^\delta$ ,

$$\hat{\mathbf{x}}_{\text{cm}} = \int_{\mathbb{R}^n} \mathbf{x} p(\mathbf{x} | \mathbf{y}^\delta) d\mathbf{x}, \quad (4.7)$$

and the problem of finding the conditional mean estimator requires to solve an integration problem. The maximum likelihood estimator

$$\hat{\mathbf{x}}_{\text{ml}} = \arg \max_{\mathbf{x}} p(\mathbf{y}^\delta | \mathbf{x})$$

is not a Bayesian estimator but it is perhaps the most popular estimator in statistics. For ill-posed problems, the maximum likelihood estimator corresponds to solving the inverse problem without regularization, and is therefore of little importance for our analysis.

## 4.2 Gaussian densities

An  $n$ -dimensional random vector  $\mathbf{X}$  has a (non-degenerate) Gaussian, or normal, distribution, if its probability density has the form

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C}_x)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{C}_x^{-1}(\mathbf{x} - \bar{\mathbf{x}})\right).$$

In the above relation,

$$\bar{\mathbf{x}} = \mathcal{E}\{\mathbf{X}\} = \int_{\mathbb{R}^n} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \quad (4.8)$$

is the mean vector or the expected value of  $\mathbf{X}$  and

$$\mathbf{C}_x = \mathcal{E}\left\{(\mathbf{X} - \mathcal{E}\{\mathbf{X}\})(\mathbf{X} - \mathcal{E}\{\mathbf{X}\})^T\right\} = \int_{\mathbb{R}^n} (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T p(\mathbf{x}) d\mathbf{x}$$

is the covariance matrix of  $\mathbf{X}$ . These parameters characterize the Gaussian density and we indicate this situation by writing  $\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{C}_x)$ . In this section, we derive Bayesian estimators for Gaussian densities and characterize the solution error following the treatment of Rodgers (2000). We then discuss two measures of the retrieval quality, the degree of freedom for signal and the information content.

### 4.2.1 Estimators

Under the assumption that  $\mathbf{X}$  and  $\Delta$  are independent Gaussian random vectors, characterized by  $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x)$  and  $\Delta \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\delta)$ , the a priori density can be expressed as

$$p_a(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C}_x)}} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{C}_x^{-1} \mathbf{x}\right), \quad (4.9)$$

while by virtue of (4.3), the likelihood density takes the form

$$p(\mathbf{y}^\delta | \mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m \det(\mathbf{C}_\delta)}} \exp\left(-\frac{1}{2}(\mathbf{y}^\delta - \mathbf{Kx})^T \mathbf{C}_\delta^{-1} (\mathbf{y}^\delta - \mathbf{Kx})\right). \quad (4.10)$$

With this information, the Bayes formula yields the following expression for the a posteriori density:

$$p(\mathbf{x} | \mathbf{y}^\delta) \propto \exp\left(-\frac{1}{2}(\mathbf{y}^\delta - \mathbf{Kx})^T \mathbf{C}_\delta^{-1} (\mathbf{y}^\delta - \mathbf{Kx}) - \frac{1}{2}\mathbf{x}^T \mathbf{C}_x^{-1} \mathbf{x}\right). \quad (4.11)$$

Setting

$$p(\mathbf{x} | \mathbf{y}^\delta) \propto \exp\left(-\frac{1}{2}V(\mathbf{x} | \mathbf{y}^\delta)\right),$$

where the a posteriori potential  $V(\mathbf{x} | \mathbf{y}^\delta)$  is defined by

$$V(\mathbf{x} | \mathbf{y}^\delta) = (\mathbf{y}^\delta - \mathbf{Kx})^T \mathbf{C}_\delta^{-1} (\mathbf{y}^\delta - \mathbf{Kx}) + \mathbf{x}^T \mathbf{C}_x^{-1} \mathbf{x},$$

we see that the maximum a posteriori estimator  $\hat{\mathbf{x}}_{\text{map}}$  maximizing the conditional probability density  $p(\mathbf{x} | \mathbf{y}^\delta)$  also minimizes the potential  $V(\mathbf{x} | \mathbf{y}^\delta)$ , that is,

$$\hat{\mathbf{x}}_{\text{map}} = \arg \min_{\mathbf{x}} V(\mathbf{x} | \mathbf{y}^\delta).$$

The solution to this minimization problem is given by

$$\hat{\mathbf{x}}_{\text{map}} = \hat{\mathbf{G}} \mathbf{y}^\delta, \quad (4.12)$$

where

$$\hat{\mathbf{G}} = (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-1} \quad (4.13)$$

is known as the gain matrix or the contribution function matrix (Rodgers, 2000). Equation (4.12) reveals that the gain matrix corresponds to the regularized generalized inverse appearing in the framework of Tikhonov regularization. An alternative representation for the gain matrix can be derived from the relation

$$(\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-1} = \mathbf{C}_x \mathbf{K}^T (\mathbf{C}_\delta + \mathbf{K} \mathbf{C}_x \mathbf{K}^T)^{-1}, \quad (4.14)$$

and the result is

$$\hat{\mathbf{G}} = \mathbf{C}_x \mathbf{K}^T (\mathbf{C}_\delta + \mathbf{K} \mathbf{C}_x \mathbf{K}^T)^{-1}. \quad (4.15)$$

To prove (4.14), we multiply this equation from the left and from the right with the matrices  $\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1}$  and  $\mathbf{C}_\delta + \mathbf{K} \mathbf{C}_x \mathbf{K}^T$ , respectively, and use the identity

$$\mathbf{K}^T + \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} \mathbf{C}_x \mathbf{K}^T = (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1}) \mathbf{C}_x \mathbf{K}^T \quad (4.16)$$

to conclude.

The a posteriori density  $p(\mathbf{x} | \mathbf{y}^\delta)$  can be expressed as a Gaussian density

$$p(\mathbf{x} | \mathbf{y}^\delta) \propto \exp \left( -\frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\mathbf{C}}_x^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \right), \quad (4.17)$$

where the mean vector  $\bar{\mathbf{x}}$  and the covariance matrix  $\hat{\mathbf{C}}_x$  can be obtained directly from (4.11) and (4.17) by equating like terms (see, e.g., Rodgers, 2000). Equating the terms quadratic in  $\mathbf{x}$  leads to the following expression for the a posteriori covariance matrix:

$$\hat{\mathbf{C}}_x = (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1}.$$

To obtain the expression of the a posteriori mean vector, we equate the terms linear in  $\mathbf{x}$  and obtain  $\bar{\mathbf{x}} = \hat{\mathbf{x}}_{\text{map}}$ . On the other hand, by (4.7), (4.8) and (4.17), we see that the a posteriori mean coincides with the conditional mean, and we conclude that in the purely Gaussian case there holds

$$\bar{\mathbf{x}} = \hat{\mathbf{x}}_{\text{map}} = \hat{\mathbf{x}}_{\text{cm}}.$$

Due to this equivalence and in order to simplify the writing, the maximum a posteriori estimator will be simply denoted by  $\hat{\mathbf{x}}$ .

An alternative expression for the a posteriori covariance matrix follows from the identity (cf. (4.14))

$$\begin{aligned} \mathbf{C}_x - \mathbf{C}_x \mathbf{K}^T (\mathbf{C}_\delta + \mathbf{K} \mathbf{C}_x \mathbf{K}^T)^{-1} \mathbf{K} \mathbf{C}_x \\ = \mathbf{C}_x - (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} \mathbf{C}_x \\ = (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1}, \end{aligned} \quad (4.18)$$

which yields (cf. (4.15))

$$\hat{\mathbf{C}}_x = \mathbf{C}_x - \hat{\mathbf{G}} \mathbf{K} \mathbf{C}_x = (\mathbf{I}_n - \mathbf{A}) \mathbf{C}_x \quad (4.19)$$

with  $\mathbf{A} = \hat{\mathbf{G}} \mathbf{K}$  being the averaging kernel matrix.

For Gaussian densities with covariance matrices of the form

$$\mathbf{C}_\delta = \sigma^2 \mathbf{I}_m, \quad \mathbf{C}_x = \sigma_x^2 \mathbf{C}_{\text{nx}} = \sigma_x^2 (\mathbf{L}^T \mathbf{L})^{-1}, \quad (4.20)$$

we find that

$$\hat{\mathbf{x}} = (\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T \mathbf{y}^\delta,$$

where we have set

$$\alpha = \frac{\sigma^2}{\sigma_x^2}.$$

As in section 3.2,  $\sigma$  is the white noise standard deviation,  $\sigma_x$  is the profile standard deviation,  $\mathbf{C}_{nx}$  is the normalized a priori covariance matrix, and  $\alpha$  and  $\mathbf{L}$  are the regularization parameter and the regularization matrix, respectively. Thus, under assumptions (4.20), the maximum a posteriori estimator coincides with the Tikhonov solution. The regularization parameter is the ratio of the noise variance to the profile variance in our a priori knowledge, and in an engineering language,  $\alpha$  can be interpreted as the noise-to-signal ratio. We can think of our a priori knowledge in terms of ellipsoids of constant probability of the a priori, whose shape and orientation are determined by  $\mathbf{C}_{nx}$  and whose size is determined by  $\sigma_x^2$ . The number  $\sigma_x$  then, represents the a priori confidence we have in our initial guess of the state vector, confidence being measured through the Mahalanobis norm with covariance  $\mathbf{C}_{nx}$ . The correspondence between the Bayesian approach and Tikhonov regularization, which has been recognized by several authors, e.g., Golub et al. (1979), O'Sullivan and Wahba (1985), Fitzpatrick (1991), Vogel (2002), Kaipio and Somersalo (2005), allows the construction of natural schemes for estimating  $\sigma_x^2$ .

#### 4.2.2 Error characterization

In a semi-stochastic setting the total error in the state space has a deterministic component, the smoothing error, and a stochastic component, the noise error. In a stochastic setting, both error components are random vectors. To introduce the random errors, we express the maximum a posteriori estimator as (see (3.65))

$$\hat{\mathbf{x}} = \hat{\mathbf{G}}\mathbf{y}^\delta = \hat{\mathbf{G}}(\mathbf{K}\mathbf{x}^\dagger + \delta) = \mathbf{A}\mathbf{x}^\dagger + \hat{\mathbf{G}}\delta.$$

and find that

$$\mathbf{x}^\dagger - \hat{\mathbf{x}} = (\mathbf{I}_n - \mathbf{A})\mathbf{x}^\dagger - \hat{\mathbf{G}}\delta. \quad (4.21)$$

In view of (4.21), we define the random total error by

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} = (\mathbf{I}_n - \mathbf{A})\mathbf{X} - \hat{\mathbf{G}}\Delta, \quad (4.22)$$

where

$$\hat{\mathbf{X}} = \hat{\mathbf{G}}\mathbf{Y}^\delta$$

is an estimator of  $\mathbf{X}$ . In (4.22),  $\mathbf{X}$  should be understood as the true state, and a realization of  $\mathbf{X}$  is the exact solution of the linear equation in the noise-free case.

The random smoothing error is defined by

$$\mathbf{E}_s = (\mathbf{I}_n - \mathbf{A})\mathbf{X},$$

and it is apparent that the statistics of  $\mathbf{E}_s$  is determined by the statistics of  $\mathbf{X}$ . If  $\mathcal{E}\{\mathbf{X}\} = \mathbf{0}$  and  $\mathbf{C}_{xt} = \mathcal{E}\{\mathbf{X}\mathbf{X}^T\}$  is the covariance matrix of the true state, then the mean vector and the covariance matrix of  $\mathbf{E}_s$  become

$$\mathcal{E}\{\mathbf{E}_s\} = \mathbf{0}, \quad \mathbf{C}_{es} = (\mathbf{I}_n - \mathbf{A})\mathbf{C}_{xt}(\mathbf{I}_n - \mathbf{A})^T.$$

In practice, the statistics of the true state is unknown and, as in a semi-stochastic setting, the statistics of the smoothing error is unknown.

The random noise error is defined as

$$\mathbf{E}_n = -\hat{\mathbf{G}}\mathbf{\Delta}$$

and the mean vector and the covariance matrix of  $\mathbf{E}_n$  are given by

$$\mathcal{E}\{\mathbf{E}_n\} = \mathbf{0}, \quad \mathbf{C}_{en} = \hat{\mathbf{G}}\mathbf{C}_\delta\hat{\mathbf{G}}^T.$$

As  $\mathbf{X}$  and  $\mathbf{\Delta}$  are independent random vectors, the random total error has zero mean and covariance

$$\mathbf{C}_e = \mathbf{C}_{es} + \mathbf{C}_{en}.$$

When computing the maximum a posteriori estimator we use an ad hoc a priori covariance matrix  $\mathbf{C}_x$  because the covariance matrix of the true state  $\mathbf{C}_{xt}$  is not available. It should be pointed out, that only for  $\mathbf{C}_x = \mathbf{C}_{xt}$ , the total error covariance matrix coincides with the a posteriori covariance matrix. To prove this claim, we construct the total error covariance matrix as

$$\begin{aligned} \mathbf{C}_e &= (\mathbf{I}_n - \hat{\mathbf{G}}\mathbf{K}) \mathbf{C}_x (\mathbf{I}_n - \hat{\mathbf{G}}\mathbf{K})^T + \hat{\mathbf{G}}\mathbf{C}_\delta\hat{\mathbf{G}}^T \\ &= \mathbf{C}_x - \mathbf{C}_x\mathbf{K}^T\hat{\mathbf{G}}^T - \hat{\mathbf{G}}\mathbf{K}\mathbf{C}_x + \hat{\mathbf{G}}\mathbf{K}\mathbf{C}_x\mathbf{K}^T\hat{\mathbf{G}}^T + \hat{\mathbf{G}}\mathbf{C}_\delta\hat{\mathbf{G}}^T, \end{aligned}$$

and use the result (cf. (4.13) and (4.16))

$$\begin{aligned} \hat{\mathbf{G}}\mathbf{C}_\delta + \hat{\mathbf{G}}\mathbf{K}\mathbf{C}_x\mathbf{K}^T &= (\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K} + \mathbf{C}_x^{-1})^{-1} \mathbf{K}^T\mathbf{C}_\delta^{-1} (\mathbf{C}_\delta + \mathbf{K}\mathbf{C}_x\mathbf{K}^T) \\ &= (\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K} + \mathbf{C}_x^{-1})^{-1} (\mathbf{K}^T\mathbf{C}_\delta^{-1}\mathbf{K} + \mathbf{C}_x^{-1}) \mathbf{C}_x\mathbf{K}^T \\ &= \mathbf{C}_x\mathbf{K}^T \end{aligned}$$

to obtain (cf. (4.19))

$$\mathbf{C}_e = \mathbf{C}_x - \hat{\mathbf{G}}\mathbf{K}\mathbf{C}_x = \hat{\mathbf{C}}_x.$$

The main conclusion which can be drawn is that an error analysis based on the a posteriori covariance matrix is correct only if the a priori covariance matrix approximates sufficiently well the covariance matrix of the true state.

### 4.2.3 Degrees of freedom

In classical regularization theory, the expected residual  $\mathcal{E}\{\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_\alpha^\delta\|^2\}$  and the expected constraint  $\mathcal{E}\{\|\mathbf{L}\mathbf{x}_\alpha^\delta\|^2\}$  are important tools for analyzing discrete ill-posed problems. In statistical inversion theory, the corresponding quantities are the degree of freedom for noise and the degree of freedom for signal.

To introduce these quantities, we consider the expression of the a posteriori potential  $V(\mathbf{x} | \mathbf{y}^\delta)$  and define the random variable

$$\hat{V} = (\mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}})^T \mathbf{C}_\delta^{-1} (\mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}}) + \hat{\mathbf{X}}^T \mathbf{C}_x^{-1} \hat{\mathbf{X}}, \quad (4.23)$$

where, as before,  $\hat{\mathbf{X}} = \hat{\mathbf{G}}\mathbf{Y}^\delta$ . The random variable  $\hat{V}$  is Chi-square distributed with  $m$  degrees of freedom, and therefore, the expected value of  $\hat{V}$  is equal to the number of measurements  $m$  (Appendix D). This can be divided into the degrees of freedom for signal and noise, defined by

$$d_s = \mathcal{E} \left\{ \hat{\mathbf{X}}^T \mathbf{C}_x^{-1} \hat{\mathbf{X}} \right\}$$

and

$$d_n = \mathcal{E} \left\{ \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right)^T \mathbf{C}_\delta^{-1} \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right) \right\},$$

respectively, and evidently we have

$$d_s + d_n = m.$$

The degree of freedom for signal measures that part of  $\mathcal{E}\{\hat{V}\}$  corresponding to the state vector, while the degree of freedom for noise that part corresponding to the measurement.

Using the identity

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \text{trace}(\mathbf{x} \mathbf{x}^T \mathbf{A}),$$

which holds true for a symmetric matrix  $\mathbf{A}$ , we express the degree of freedom for signal as

$$d_s = \mathcal{E} \left\{ \hat{\mathbf{X}}^T \mathbf{C}_x^{-1} \hat{\mathbf{X}} \right\} = \mathcal{E} \left\{ \text{trace} \left( \hat{\mathbf{X}} \hat{\mathbf{X}}^T \mathbf{C}_x^{-1} \right) \right\} = \text{trace} \left( \mathcal{E} \left\{ \hat{\mathbf{X}} \hat{\mathbf{X}}^T \right\} \mathbf{C}_x^{-1} \right),$$

where the covariance of the estimator  $\hat{\mathbf{X}}$  is related to the covariance of the data  $\mathbf{Y}^\delta$  by the relation

$$\mathcal{E} \left\{ \hat{\mathbf{X}} \hat{\mathbf{X}}^T \right\} = \hat{\mathbf{G}} \mathcal{E} \left\{ \mathbf{Y}^\delta \mathbf{Y}^{\delta T} \right\} \hat{\mathbf{G}}^T.$$

To compute the covariance of the data, we assume that the covariance matrix of the true state is adequately described by the a priori covariance matrix, and obtain

$$\mathcal{E} \left\{ \mathbf{Y}^\delta \mathbf{Y}^{\delta T} \right\} = \mathbf{K} \mathcal{E} \left\{ \mathbf{X} \mathbf{X}^T \right\} \mathbf{K}^T + \mathcal{E} \left\{ \Delta \Delta^T \right\} = \mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta. \quad (4.24)$$

By (4.13) and (4.15), we then have

$$\mathcal{E} \left\{ \hat{\mathbf{X}} \hat{\mathbf{X}}^T \right\} = \mathbf{C}_x \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} \left( \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1} \right)^{-1}, \quad (4.25)$$

whence using the identities  $\text{trace}(\mathbf{B}^{-1} \mathbf{A} \mathbf{B}) = \text{trace}(\mathbf{A})$  and  $\text{trace}(\mathbf{A}) = \text{trace}(\mathbf{A}^T)$ , which hold true for a square matrix  $\mathbf{A}$  and a nonsingular matrix  $\mathbf{B}$ , we find that

$$\begin{aligned} d_s &= \text{trace} \left( \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} \left( \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1} \right)^{-1} \right) \\ &= \text{trace} \left( \left( \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1} \right)^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} \right) \\ &= \text{trace} \left( \hat{\mathbf{G}} \mathbf{K} \right) \\ &= \text{trace}(\mathbf{A}). \end{aligned} \quad (4.26)$$

Hence, the degree of freedom for signal is the trace of the averaging kernel matrix. Consequently, the diagonal of the averaging kernel matrix  $\mathbf{A}$  may be thought of as a measure of

the number of degrees of freedom per layer (level), and thus as a measure of information, while its reciprocal may be thought of as the number of layers per degree of freedom, and thus as a measure of resolution. The degree of freedom for signal can also be interpreted as a measure of the minimum number of parameters that could be used to define a state vector without loss of information (Mateer, 1965); Rodgers, 2000).

The degree of freedom for noise can be expressed in terms of the influence matrix  $\hat{\mathbf{A}} = \mathbf{K}\hat{\mathbf{G}}$  as (cf. (4.24))

$$\begin{aligned} d_n &= \mathcal{E} \left\{ \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right)^T \mathbf{C}_\delta^{-1} \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right) \right\} \\ &= \mathcal{E} \left\{ \text{trace} \left( \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right) \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right)^T \mathbf{C}_\delta^{-1} \right) \right\} \\ &= \text{trace} \left( \left( \mathbf{I}_m - \hat{\mathbf{A}} \right) \mathcal{E} \left\{ \mathbf{Y}^\delta \mathbf{Y}^{\delta T} \right\} \left( \mathbf{I}_m - \hat{\mathbf{A}} \right)^T \mathbf{C}_\delta^{-1} \right) \\ &= \text{trace} \left( \left( \mathbf{I}_m - \hat{\mathbf{A}} \right) \left( \mathbf{K}\mathbf{C}_x\mathbf{K}^T + \mathbf{C}_\delta \right) \left( \mathbf{I}_m - \hat{\mathbf{A}} \right)^T \mathbf{C}_\delta^{-1} \right), \end{aligned} \quad (4.27)$$

whence using the identity

$$\left( \mathbf{I}_m - \hat{\mathbf{A}} \right) \left( \mathbf{K}\mathbf{C}_x\mathbf{K}^T + \mathbf{C}_\delta \right) = \mathbf{C}_\delta, \quad (4.28)$$

we obtain

$$d_n = \text{trace} \left( \mathbf{I}_m - \hat{\mathbf{A}} \right). \quad (4.29)$$

Note that the term ‘degree of freedom for noise’ has been used by Craven and Wahba (1979) and later on by Wahba (1985) to designate the denominator of the generalized cross-validation function.

Under assumptions (4.20), we have

$$\text{trace}(\mathbf{A}) = \text{trace}(\hat{\mathbf{A}}) = \sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha}, \quad (4.30)$$

where  $\gamma_i$  are the generalized singular values of the matrix pair  $(\mathbf{K}, \mathbf{L})$ . By (4.26), (4.29) and (4.30), it is apparent that the degree of freedom for signal is a decreasing function of the regularization parameter, while the degree of freedom for noise is an increasing function of the regularization parameter. Thus, when very little regularization is introduced, the degree of freedom for signal is very large and approaches  $n$ , and when a large amount of regularization is introduced, the degree of freedom for noise is very large and approaches  $m$ . As in classical regularization theory, an optimal regularization parameter should balance the degrees of freedom for signal and noise.

The degree of freedom for signal can be expressed in terms of the so-called information matrix  $\mathbf{R}$  defined by

$$\mathbf{R} = \mathbf{C}_x^{\frac{1}{2}} \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} \mathbf{C}_x^{\frac{1}{2}}. \quad (4.31)$$

Using the identity

$$\mathbf{A} = \mathbf{C}_x^{\frac{1}{2}} (\mathbf{I}_n + \mathbf{R})^{-1} \mathbf{R} \mathbf{C}_x^{-\frac{1}{2}}, \quad (4.32)$$



we find that

$$d_s = \text{trace}(\mathbf{A}) = \text{trace}\left((\mathbf{I}_n + \mathbf{R})^{-1} \mathbf{R}\right), \quad (4.33)$$

whence assuming the singular value decomposition of the positive definite matrix  $\mathbf{R}$ ,

$$\mathbf{R} = \mathbf{V}_r \Sigma_r \mathbf{V}_r^T, \quad \Sigma_r = [\text{diag}(\omega_i)_{n \times n}], \quad (4.34)$$

we obtain the representation

$$d_s = \sum_{i=1}^n \frac{\omega_i}{\omega_i + 1}.$$

The degree of freedom for signal  $d_s$  remains unchanged under linear transformations of the state vector or of the data vector, and as a result,  $d_s$  is an invariant of the retrieval. Purser and Huang (1993) showed that the degree of freedom for signal, regarded as a real-valued function over sets of independent data, obeys a positive monotonic subadditive algebra. In order to understand these properties from a practical point of view, we consider a set of  $m_1$  data  $\mathbf{Y}_1^\delta = \mathbf{y}_1^\delta$ , and an independent set of  $m_2$  data  $\mathbf{Y}_2^\delta = \mathbf{y}_2^\delta$ . For the  $i$ th set of measurements, the data model is

$$\mathbf{Y}_i^\delta = \mathbf{K}_i \mathbf{X} + \Delta_i, \quad i = 1, 2,$$

and the maximum a posteriori estimator is computed as

$$\hat{\mathbf{x}}_i = \arg \min_{\mathbf{x}} \left( (\mathbf{y}_i^\delta - \mathbf{K}_i \mathbf{x})^T \mathbf{C}_{\delta i}^{-1} (\mathbf{y}_i^\delta - \mathbf{K}_i \mathbf{x}) + \mathbf{x}^T \mathbf{C}_x^{-1} \mathbf{x} \right).$$

The corresponding information matrix and the degree of freedom for signal are given by

$$\mathbf{R}_i = \mathbf{C}_x^{\frac{1}{2}} \mathbf{K}_i^T \mathbf{C}_{\delta i}^{-1} \mathbf{K}_i \mathbf{C}_x^{\frac{1}{2}}$$

and

$$d_{si} = \text{trace}\left((\mathbf{I}_n + \mathbf{R}_i)^{-1} \mathbf{R}_i\right),$$

respectively. For the full set of  $m_{12} = m_1 + m_2$  measurements, we consider the data model

$$\begin{bmatrix} \mathbf{Y}_1^\delta \\ \mathbf{Y}_2^\delta \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{bmatrix} \mathbf{X} + \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix},$$

and compute the maximum a posteriori estimator as

$$\begin{aligned} \hat{\mathbf{x}}_{12} = \arg \min_{\mathbf{x}} & \left( (\mathbf{y}_1^\delta - \mathbf{K}_1 \mathbf{x})^T \mathbf{C}_{\delta 1}^{-1} (\mathbf{y}_1^\delta - \mathbf{K}_1 \mathbf{x}) \right. \\ & \left. + (\mathbf{y}_2^\delta - \mathbf{K}_2 \mathbf{x})^T \mathbf{C}_{\delta 2}^{-1} (\mathbf{y}_2^\delta - \mathbf{K}_2 \mathbf{x}) + \mathbf{x}^T \mathbf{C}_x^{-1} \mathbf{x} \right). \end{aligned}$$

When the data are treated jointly, the information matrix and the degree of freedom for signal are given by

$$\mathbf{R}_{12} = \mathbf{C}_x^{\frac{1}{2}} (\mathbf{K}_1^T \mathbf{C}_{\delta 1}^{-1} \mathbf{K}_1 + \mathbf{K}_2^T \mathbf{C}_{\delta 2}^{-1} \mathbf{K}_2) \mathbf{C}_x^{\frac{1}{2}} = \mathbf{R}_1 + \mathbf{R}_2$$

and

$$d_{s12} = \text{trace}\left((\mathbf{I}_n + \mathbf{R}_1 + \mathbf{R}_2)^{-1} (\mathbf{R}_1 + \mathbf{R}_2)\right),$$

respectively. In this context, the monotonicity of the degree of freedom for signal means that  $d_{s12}$  of the full  $m_{12}$  measurements is never less than either  $d_{s1}$  or  $d_{s2}$ , i.e.,

$$d_{s12} \geq \max(d_{s1}, d_{s2}), \quad (4.35)$$

while the subadditivity means that  $d_{s12}$  can never exceed  $d_{s1} + d_{s2}$ , i.e.,

$$d_{s12} \leq d_{s1} + d_{s2}. \quad (4.36)$$

These assertions are the result of the following theorem: considering a monotonic, strictly increasing, and strictly concave function  $f(x)$  with  $f(0) = 0$ , and defining the associated scalar function  $F$  of  $\mathbf{R} \in \mathcal{S}_n$  by

$$F(\mathbf{R}) = \sum_{i=1}^n f(\omega_i),$$

where  $\mathcal{S}_n$  is the set of all semi-positive definite matrices of order  $n$ , and  $\omega_i$  are the singular values of  $\mathbf{R}$ , we have

$$\mathbf{R}_2 \geq \mathbf{R}_1 \Rightarrow F(\mathbf{R}_2) \geq F(\mathbf{R}_1) \quad (\text{monotonicity}), \quad (4.37)$$

and

$$F(\mathbf{R}_1) + F(\mathbf{R}_2) \geq F(\mathbf{R}_1 + \mathbf{R}_2) \quad (\text{subadditivity}), \quad (4.38)$$

for all  $\mathbf{R}_1, \mathbf{R}_2 \in \mathcal{S}_n$ . Here, we write  $\mathbf{R}_2 \geq \mathbf{R}_1$  if  $\mathbf{R}_2 - \mathbf{R}_1 \in \mathcal{S}_n$ . Since the degree of freedom for signal  $d_s$  can be expressed in terms of the information matrix  $\mathbf{R}$  as a scalar function  $F(\mathbf{R})$  with  $f(x) = x/(1+x)$ , (4.37) and (4.38) yield (4.35) and (4.36), respectively. A rigorous proof of this theorem has been given by Purser and Huang (1993) by taking into account that  $F(\mathbf{R})$  is invariant to orthogonal transformations. However, (4.35) and (4.36) can simply be justified when

$$m_1 = m_2 = m, \quad \mathbf{K}_1 = \mathbf{K}_2, \quad \mathbf{C}_{\delta 1} = \mathbf{C}_{\delta 2}. \quad (4.39)$$

In this case, we obtain

$$\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}, \quad \mathbf{R}_{12} = 2\mathbf{R},$$

and further,

$$d_{s1} = d_{s2} = \sum_{i=1}^n \frac{\omega_i}{\omega_i + 1}, \quad d_{s12} = \sum_{i=1}^n \frac{2\omega_i}{2\omega_i + 1}.$$

Then, from

$$\frac{2\omega_i}{2\omega_i + 1} \geq \frac{\omega_i}{\omega_i + 1}, \quad \frac{2\omega_i}{2\omega_i + 1} \leq \frac{2\omega_i}{\omega_i + 1}, \quad i = 1, \dots, n,$$

the conclusions are apparent. The deficit  $m_{12} - d_{s12}$  may be interpreted as the internal redundancy of the set of data, while the deficit  $d_{s1} + d_{s2} - d_{s12}$  may be thought as the mutual redundancy between two pooled sets.

Another statistics of a linear retrieval is the effective data density. Whereas the degree of freedom for signal is a measure that indicates the number of independent pieces of information, the effective data density is a measure that indicates the density of effectively

independent pieces of information. The data density at the  $i$ th layer of thickness  $\Delta z_i$  is defined by

$$\rho_i = \frac{[\mathbf{A}]_{ii}}{\Delta z_i}, \quad (4.40)$$

and it is apparent that the ‘integral’ of the effective data density is the degree of freedom for signal,

$$d_s = \sum_{i=1}^n \rho_i \Delta z_i.$$

This estimate together with the degree of freedom for signal can be used to interpret the quality of the retrieval and the effectiveness of the measurements.

#### 4.2.4 Information content

An alternative criterion for the quality of a measurement is the information content or the incremental gain in information. The information content is defined in terms of the change in entropy that expresses a priori and a posteriori knowledge of the atmospheric state. This measure of performance has been proposed in the context of retrieval by Peckham (1974) and has also been discussed by Rodgers (1976) and Eyre (1990).

In information theory, the Shannon entropy or the absolute entropy is a measure of uncertainty associated with a random variable. The Shannon entropy of a discrete random vector  $\mathbf{X}$ , which can take the values  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , is defined by

$$H(p) = - \sum_{i=1}^n p_i \log p_i, \quad (4.41)$$

where the probability mass function of  $\mathbf{X}$  is given by

$$p(\mathbf{x}) = \begin{cases} p_i, & \mathbf{X} = \mathbf{x}_i, \\ 0, & \text{otherwise,} \end{cases} \quad \sum_{i=1}^n p_i = 1.$$

$H$  is positive and attains its global maximum  $H_{\max} = \log n$  for a uniform distribution, i.e., when all  $p_i$  are the same. On the other hand, the lowest entropy level,  $H_{\min} = 0$ , is attained when all probabilities  $p_i$  but one are zero. Shannon (1949) showed that  $H(p)$  defined by (4.41) satisfies the following desiderata:

- (1)  $H$  is continuous in  $(p_1, \dots, p_n)$  (continuity);
- (2)  $H$  remains unchanged if the outcomes  $\mathbf{x}_i$  are re-ordered (symmetry);
- (3) if all the outcomes are equally likely, then  $H$  is maximal (maximum);
- (4) the amount of entropy is the same independently of how the process is regarded as being divided into parts (additivity).

These properties guarantee that the Shannon entropy is well-behaved with regard to relative information comparisons. For a continuous density  $p(\mathbf{x})$ , the following entropy formula also satisfies the properties enumerated above:

$$H(p) = - \int_{\mathbb{R}^n} p(\mathbf{x}) \log p(\mathbf{x}) \, d\mathbf{x}. \quad (4.42)$$

For a Gaussian random vector with covariance matrix  $\mathbf{C}$ , the integral in (4.42) can be analytically computed and the result is

$$H(p) = \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log(\det(\mathbf{C})).$$

As the a priori density  $p_a(\mathbf{x})$  describes knowledge before a measurement and the a posteriori density  $p(\mathbf{x} | \mathbf{y}^\delta)$  describes it afterwards, the information content of the measurement  $\Delta H$  is the reduction in entropy (e.g., Rodgers, 2000)

$$\Delta H = H(p_a(\mathbf{x})) - H(p(\mathbf{x} | \mathbf{y}^\delta)).$$

For Gaussian densities with the a priori and the a posteriori covariance matrices  $\mathbf{C}_x$  and  $\hat{\mathbf{C}}_x$ , respectively, the information content then becomes

$$\Delta H = -\frac{1}{2} \log\left(\det\left(\hat{\mathbf{C}}_x \mathbf{C}_x^{-1}\right)\right) = -\frac{1}{2} \log(\det(\mathbf{I}_n - \mathbf{A})).$$

By virtue of (4.32), which relates the information matrix  $\mathbf{R}$  and the averaging kernel matrix  $\mathbf{A}$ , we obtain the representation

$$\Delta H = \frac{1}{2} \log(\det(\mathbf{I}_n + \mathbf{R})),$$

and further

$$\Delta H = \frac{1}{2} \sum_{i=1}^n \log(1 + \omega_i).$$

Similar to the degree of freedom for signal, the information content obeys a positive monotonic subadditive algebra (Huang and Purser, 1996). By ‘monotonic’ we mean that the addition of independent data does not decrease (on average) the information content, while by ‘subadditive’ we mean that any two sets of data treated jointly never yield more of the information content than the sum of the amounts yielded by the sets treated singly. These results follow from (4.37) and (4.38) by taking into account that the information content  $\Delta H$  can be expressed in terms of the information matrix  $\mathbf{R}$  as a scalar function  $F(\mathbf{R})$  with  $f(x) = (1/2) \log(1 + x)$ , or, in the simple case (4.39), they follow from the obvious inequalities

$$\log(1 + 2\omega_i) \geq \log(1 + \omega_i), \quad \log(1 + 2\omega_i) \leq 2 \log(1 + \omega_i), \quad i = 1, \dots, n.$$

A density of information can be defined by employing the technique which has been used to define the effective data density. For this purpose, we seek an equivalent matrix  $\mathbf{A}_h$ , whose trace is the information content  $\Delta H$ , so that the diagonal elements of this matrix can be used as in (4.40) to define the density of information at each layer,

$$\rho_{hi} = \frac{[\mathbf{A}_h]_{ii}}{\Delta z_i}.$$

The matrix  $\mathbf{A}_h$  is chosen as

$$\mathbf{A}_h = \mathbf{V}_r \Sigma_{ah} \mathbf{V}_r^T,$$

where  $\mathbf{V}_r$  is the orthogonal matrix in (4.34) and

$$\Sigma_{\text{ah}} = \left[ \text{diag} \left( \frac{1}{2} \log(1 + \omega_i) \right)_{n \times n} \right].$$

The information content is used as a selection criterion in the framework of the so-called information operator method. Assuming (4.20) and considering a generalized singular value decomposition of the matrix pair  $(\mathbf{K}, \mathbf{L})$ , the maximum a posteriori estimator and the information content of the measurement can be expressed as

$$\hat{\mathbf{x}}_{\text{map}} = \sum_{i=1}^n f_{\alpha}(\gamma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^{\delta}) \mathbf{w}_i,$$

and

$$\Delta H = \frac{1}{2} \sum_{i=1}^n \log \left( 1 + \frac{\gamma_i^2}{\alpha} \right),$$

respectively, where

$$f_{\alpha}(\gamma_i^2) = \frac{\gamma_i^2}{\gamma_i^2 + \alpha}, \quad i = 1, \dots, n,$$

are the filter factors for Tikhonov regularization and  $\alpha = \sigma^2 / \sigma_x^2$ . In the information operator method, only the generalized singular values  $\gamma_i$  larger than  $\sqrt{\alpha}$  are considered to give a relevant contribution to the information content. Note that  $\alpha$  should not be regarded as a regularization parameter whose value should be optimized; rather  $\alpha$  is completely determined by the profile variance  $\sigma_x^2$  which we take to be fixed. The state space spanned by the singular vectors associated with the relevant singular values gives the effective state space accessible with the measurement (Kozlov, 1983; Rozanov, 2001). If  $p$  is the largest index  $i$  so that

$$\gamma_i^2 \geq \alpha = \frac{\sigma^2}{\sigma_x^2}, \quad i = 1, \dots, p,$$

then the information operator solution can be expressed as

$$\hat{\mathbf{x}}_{\text{io}} = \sum_{i=1}^p f_{\alpha}(\gamma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^{\delta}) \mathbf{w}_i.$$

Essentially, the filter factors of the information operator method are given by

$$f_{\alpha}(\gamma_i^2) = \begin{cases} \gamma_i^2, & \gamma_i^2 \geq \alpha, \\ 0, & \gamma_i^2 < \alpha, \end{cases}$$

and we see that the information operator method has sharper filter factors than Tikhonov regularization.

### 4.3 Regularization parameter choice methods

Under assumptions (4.20), the Bayesian approach is equivalent to Tikhonov regularization in the sense that the maximum a posteriori estimator simultaneously minimizes the potential

$$V(\mathbf{x} | \mathbf{y}^\delta) = \frac{1}{\sigma^2} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \frac{1}{\sigma_x^2} \|\mathbf{L}\mathbf{x}\|^2,$$

and the Tikhonov function

$$\mathcal{F}_\alpha(\mathbf{x}) = \sigma^2 V(\mathbf{x} | \mathbf{y}^\delta) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2, \quad \alpha = \frac{\sigma^2}{\sigma_x^2}.$$

When the profile variance  $\sigma_x^2$  is unknown, it seems to be justified to ask for a reliable estimator  $\hat{\sigma}_x^2$  of  $\sigma_x^2$ , or equivalently, for a plausible estimator  $\hat{\alpha}$  of  $\alpha$ . For this reason, in statistical inversion theory, a regularization parameter choice method can be regarded as an approach for estimating  $\sigma_x^2$ .

#### 4.3.1 Expected error estimation method

In a semi-stochastic setting, the expected error estimation method has been formulated in the following way: given the exact profile  $\mathbf{x}^\dagger$ , compute the optimal regularization parameter  $\bar{\alpha}_{\text{opt}}$  as the minimizer of the expected error  $\mathcal{E}\{\|\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta\|^2\}$ , with  $\mathbf{x}_\alpha^\delta$  being the Tikhonov solution of regularization parameter  $\alpha$ . In statistical inversion theory, an equivalent formulation may read as follows: given the covariance matrix of the true state  $\mathbf{C}_{\text{xt}}$ , compute the profile variance  $\sigma_x^2$  as the minimizer of the expected error

$$\mathcal{E}\{\|\mathbf{E}\|^2\} = \text{trace}\left((\mathbf{I}_n - \mathbf{A})\mathbf{C}_{\text{xt}}(\mathbf{I}_n - \mathbf{A})^T\right) + \sigma^2 \text{trace}\left(\hat{\mathbf{G}}\hat{\mathbf{G}}^T\right), \quad (4.43)$$

where the a priori covariance matrix in the expressions of  $\hat{\mathbf{G}}$  and  $\mathbf{A}$  is given by  $\mathbf{C}_x = \sigma_x^2 \mathbf{C}_{\text{nx}}$ . If the covariance matrix of the true state is expressed as  $\mathbf{C}_{\text{xt}} = \sigma_{\text{xt}}^2 \mathbf{C}_{\text{nx}}$ , then the minimization of the expected error (4.43), yields  $\sigma_x = \sigma_{\text{xt}}$ . To prove this result under assumptions (4.20), we take  $\mathbf{L} = \mathbf{I}_n$ , and obtain

$$E(\alpha) = \mathcal{E}\{\|\mathbf{E}\|^2\} = \sigma_{\text{xt}}^2 \sum_{i=1}^n \left[ \left( \frac{\alpha}{\sigma_i^2 + \alpha} \right)^2 + \alpha_{\text{t}} \left( \frac{\sigma_i}{\sigma_i^2 + \alpha} \right)^2 \right], \quad \alpha_{\text{t}} = \frac{\sigma^2}{\sigma_{\text{xt}}^2}.$$

Setting  $E'(\alpha) = 0$  gives

$$\sum_{i=1}^n \left[ \frac{\alpha \sigma_i^2}{(\sigma_i^2 + \alpha)^3} - \frac{\alpha_{\text{t}} \sigma_i^2}{(\sigma_i^2 + \alpha)^3} \right] = 0,$$

which further implies that  $\alpha = \alpha_{\text{t}}$ , or equivalently that  $\sigma_x = \sigma_{\text{xt}}$ . Thus, the maximum a posteriori estimator is given by  $\hat{\mathbf{x}}_{\text{map}} = \hat{\mathbf{G}}\mathbf{y}^\delta$  with

$$\hat{\mathbf{G}} = (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_{\text{xt}}^{-1})^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-1}. \quad (4.44)$$

The selection rule based on the minimization of (4.43) simply states that if the covariance matrix of the true state is known, then this information should be used to construct the a priori density.

In statistical inversion theory, the minimization of the expected error is not formulated in terms of the profile variance (or the regularization parameter), but rather in terms of the inverse matrix  $\mathbf{G}$ . The resulting method, which is known as the minimum variance method, possesses the following formulation: if the statistics of the true state is known,

$$\mathcal{E}\{\mathbf{X}\} = \mathbf{0}, \quad \mathcal{E}\{\mathbf{X}\mathbf{X}^T\} = \mathbf{C}_{\text{xt}}, \quad (4.45)$$

then for the affine estimation rule  $\hat{\mathbf{x}} = \mathbf{G}\mathbf{y}^\delta$ , the matrix  $\hat{\mathbf{G}}$  minimizing the expected error

$$\hat{\mathbf{G}} = \arg \min_{\mathbf{G}} \mathcal{E} \left\{ \left\| \mathbf{X} - \mathbf{G}\mathbf{y}^\delta \right\|^2 \right\} \quad (4.46)$$

is given by (4.44), and the minimum variance estimator  $\hat{\mathbf{x}}_{\text{mv}} = \hat{\mathbf{G}}\mathbf{y}^\delta$  coincides with the maximum a posteriori estimator  $\hat{\mathbf{x}}_{\text{map}}$ . To justify this claim, we look at the behavior of the expected error when  $\mathbf{G}$  is replaced by a candidate solution  $\mathbf{G} + \mathbf{H}$ . Using the result

$$\begin{aligned} & \left\| \mathbf{X} - (\mathbf{G} + \mathbf{H})\mathbf{y}^\delta \right\|^2 \\ &= \left\| \mathbf{X} - \mathbf{G}\mathbf{y}^\delta \right\|^2 - 2 \left( \mathbf{X} - \mathbf{G}\mathbf{y}^\delta \right)^T \mathbf{H}\mathbf{y}^\delta + \left\| \mathbf{H}\mathbf{y}^\delta \right\|^2 \\ &= \left\| \mathbf{X} - \mathbf{G}\mathbf{y}^\delta \right\|^2 - 2 \text{trace} \left( \mathbf{H}\mathbf{y}^\delta \left( \mathbf{X} - \mathbf{G}\mathbf{y}^\delta \right)^T \right) + \left\| \mathbf{H}\mathbf{y}^\delta \right\|^2, \end{aligned}$$

we obtain

$$\begin{aligned} & \mathcal{E} \left\{ \left\| \mathbf{X} - (\mathbf{G} + \mathbf{H})\mathbf{y}^\delta \right\|^2 \right\} \\ &= \mathcal{E} \left\{ \left\| \mathbf{X} - \mathbf{G}\mathbf{y}^\delta \right\|^2 \right\} - 2 \text{trace} \left( \mathbf{H} \mathcal{E} \left\{ \mathbf{y}^\delta \left( \mathbf{X} - \mathbf{G}\mathbf{y}^\delta \right)^T \right\} \right) + \mathcal{E} \left\{ \left\| \mathbf{H}\mathbf{y}^\delta \right\|^2 \right\}. \end{aligned}$$

The trace term vanishes for the choice

$$\mathbf{G} = \hat{\mathbf{G}} = \mathcal{E} \{ \mathbf{X}\mathbf{y}^{\delta T} \} \left( \mathcal{E} \{ \mathbf{y}^\delta \mathbf{y}^{\delta T} \} \right)^{-1}, \quad (4.47)$$

since

$$\mathcal{E} \left\{ \mathbf{y}^\delta \left( \mathbf{X} - \hat{\mathbf{G}}\mathbf{y}^\delta \right)^T \right\} = \mathcal{E} \{ \mathbf{y}^\delta \mathbf{X}^T \} - \mathcal{E} \{ \mathbf{y}^\delta \mathbf{y}^{\delta T} \} \hat{\mathbf{G}}^T = \mathbf{0}.$$

Under assumptions (4.45), we find that

$$\mathcal{E} \{ \mathbf{X}\mathbf{y}^{\delta T} \} = \mathcal{E} \{ \mathbf{X}\mathbf{X}^T \} \mathbf{K}^T = \mathbf{C}_{\text{xt}}\mathbf{K}^T,$$

whence using (4.24), (4.47) becomes

$$\hat{\mathbf{G}} = \mathbf{C}_{\text{xt}}\mathbf{K}^T \left( \mathbf{K}\mathbf{C}_{\text{xt}}\mathbf{K}^T + \mathbf{C}_\delta \right)^{-1} = \left( \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_{\text{xt}}^{-1} \right)^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-1}.$$

Hence, we have

$$\begin{aligned} \mathcal{E} \left\{ \left\| \mathbf{X} - (\hat{\mathbf{G}} + \mathbf{H}) \mathbf{Y}^\delta \right\|^2 \right\} &= \mathcal{E} \left\{ \left\| \mathbf{X} - \hat{\mathbf{G}} \mathbf{Y}^\delta \right\|^2 \right\} + \mathcal{E} \left\{ \left\| \mathbf{H} \mathbf{Y}^\delta \right\|^2 \right\} \\ &\geq \mathcal{E} \left\{ \left\| \mathbf{X} - \hat{\mathbf{G}} \mathbf{Y}^\delta \right\|^2 \right\} \end{aligned}$$

for any  $\mathbf{H} \in \mathbb{R}^{n \times m}$ , and therefore,  $\mathcal{E} \{ \left\| \mathbf{X} - \mathbf{G} \mathbf{Y}^\delta \right\|^2 \}$  is minimal for  $\mathbf{G} = \hat{\mathbf{G}}$ .

The minimum variance estimator minimizes the expected error, which represents the trace of the a posteriori covariance matrix. Instead of minimizing the trace of the a posteriori covariance matrix we may formulate a minimization problem involving the entire a posteriori covariance matrix. For this purpose, we define the random total error

$$\mathbf{E} = \mathbf{X} - \mathbf{G} \mathbf{Y}^\delta = (\mathbf{I}_n - \mathbf{G} \mathbf{K}) \mathbf{X} - \mathbf{G} \mathbf{\Delta},$$

for some  $\mathbf{G} \in \mathbb{R}^{n \times m}$ . The covariance matrices of the smoothing and noise errors  $\mathbf{E}_s = (\mathbf{I}_n - \mathbf{G} \mathbf{K}) \mathbf{X}$  and  $\mathbf{E}_n = -\mathbf{G} \mathbf{\Delta}$ , can be expressed in terms of the matrix  $\mathbf{G}$ , as

$$\mathbf{C}_{es} = (\mathbf{I}_n - \mathbf{G} \mathbf{K}) \mathbf{C}_{xt} (\mathbf{I}_n - \mathbf{G} \mathbf{K})^T$$

and

$$\mathbf{C}_{en} = \mathbf{G} \mathbf{C}_\delta \mathbf{G}^T,$$

respectively. Then, it is readily seen that the minimizer of the error covariance matrix

$$\hat{\mathbf{G}} = \arg \min_{\mathbf{G}} (\mathbf{C}_{es} + \mathbf{C}_{en}), \quad (4.48)$$

solves the equation

$$\frac{\partial}{\partial \mathbf{G}} (\mathbf{C}_{xt} - \mathbf{C}_{xt} \mathbf{K}^T \mathbf{G}^T - \mathbf{G} \mathbf{K} \mathbf{C}_{xt} + \mathbf{G} \mathbf{K} \mathbf{C}_{xt} \mathbf{K}^T \mathbf{G}^T + \mathbf{G} \mathbf{C}_\delta \mathbf{G}^T) = \mathbf{0} \quad (4.49)$$

and is given by (4.44).

Because in statistical inversion theory, the conventional expected error estimation method is not beneficial, we design a regularization parameter choice method by looking only at the expected value of the noise error. Under assumptions (4.20), the noise error covariance matrix is given by (cf. (3.38))

$$\mathbf{C}_{en} = \sigma^2 \hat{\mathbf{G}}^T \hat{\mathbf{G}} = \sigma^2 \mathbf{W} \Sigma_{n\alpha} \mathbf{W}^T,$$

with

$$\Sigma_{n\alpha} = \left[ \text{diag} \left( \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)^2 \right)_{n \times n} \right],$$

and the expected value of the noise error (cf. (3.41)),

$$\mathcal{E} \left\{ \left\| \mathbf{E}_n \right\|^2 \right\} = \text{trace} (\mathbf{C}_{en}) = \sigma^2 \sum_{i=1}^n \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)^2 \left\| \mathbf{w}_i \right\|^2,$$



is a decreasing function of  $\alpha$ . To improve the degree of freedom for signal we need to chose a small value of the regularization parameter. But when the regularization parameter is too small, the noise error may explode. Therefore, we select the smallest regularization parameter so that the expected value of the noise error is below a specific level. Recalling that  $\mathbf{x}$  is the deviation of the retrieved profile from the a priori profile  $\mathbf{x}_a$ , we define the regularization parameter for noise error estimation  $\hat{\alpha}_{ne}$  as the solution of the equation

$$\mathcal{E} \left\{ \|\mathbf{E}_n\|^2 \right\} = \varepsilon_n \|\mathbf{x}_a\|^2,$$

for some relative error level  $\varepsilon_n$ . In atmospheric remote sensing, the expected noise error estimation method has been successfully applied for ozone retrieval from nadir sounding spectra measured by the Tropospheric Emission Spectrometer (TES) on the NASA Aura platform (Steck, 2002).

### 4.3.2 Discrepancy principle

In a semi-stochastic setting, the discrepancy principle selects the regularization parameter as the solution of the equation

$$\|\mathbf{r}_\alpha^\delta\|^2 = \tau m \sigma^2. \quad (4.50)$$

Under assumptions (4.20), equation (4.50) reads as

$$\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = \tau m \sigma^2, \quad (4.51)$$

with the convention  $\gamma_i = 0$  for  $i = n + 1, \dots, m$ .

The regularization parameter choice method (4.50) with  $\tau = 1$  is known as the constrained least squares method (Hunt, 1973; Trussel, 1983; Trussel and Civanlar, 1984). It has been observed and reported by a number of researchers, e.g., Demoment (1989), that the constrained least squares method yields an oversmooth solution. To ameliorate this problem, Wahba (1983), and Hall and Titterton (1987) proposed, in analogy to regression, the equivalent degree of freedom method. In a stochastic setting, this method takes into account that the expected value of the residual is equal to the trace of the matrix  $\mathbf{I}_m - \hat{\mathbf{A}}$ , that is, (cf. (4.27) and (4.29))

$$\mathcal{E} \left\{ \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right)^T \mathbf{C}_\delta^{-1} \left( \mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} \right) \right\} = \text{trace} \left( \mathbf{I}_m - \hat{\mathbf{A}} \right).$$

The resulting equation for computing the regularization parameter is then given by

$$(\mathbf{y}^\delta - \mathbf{K}\hat{\mathbf{x}})^T \mathbf{C}_\delta^{-1} (\mathbf{y}^\delta - \mathbf{K}\hat{\mathbf{x}}) = \text{trace} \left( \mathbf{I}_m - \hat{\mathbf{A}} \right),$$

or equivalently, by

$$\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = \sigma^2 \sum_{i=1}^m \frac{\alpha}{\gamma_i^2 + \alpha}.$$

On the other hand, the random variable  $\hat{V}$ , defined by (4.23), is Chi-square distributed with  $m$  degrees of freedom. In this regard, we may choose the regularization parameter as the solution of the equation

$$(\mathbf{y}^\delta - \mathbf{K}\hat{\mathbf{x}})^T \mathbf{C}_\delta^{-1} (\mathbf{y}^\delta - \mathbf{K}\hat{\mathbf{x}}) + \hat{\mathbf{x}}^T \mathbf{C}_\mathbf{x}^{-1} \hat{\mathbf{x}} = m,$$

that is,

$$\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = m\sigma^2.$$

As compared to (4.51), the factors multiplying the Fourier coefficients  $\mathbf{u}_i^T \mathbf{y}^\delta$  converge more slowly to zero as  $\alpha$  tends to zero, and therefore, this selection rule yields a larger regularization parameter than the discrepancy principle with  $\tau = 1$ .

### 4.3.3 Hierarchical models

In the Bayesian framework, all unknown parameters of the model are included in the retrieval and this applies also for parameters describing the a priori density. The resulting model is then known as hierarchical or hyperpriori model (Kaipio and Somersalo, 2005).

For the a priori covariance matrix  $\mathbf{C}_\mathbf{x} = \sigma_\mathbf{x}^2 \mathbf{C}_{\mathbf{n}\mathbf{x}}$ , we suppose that the a priori density is conditioned on the knowledge of  $\sigma_\mathbf{x}$ , i.e.,

$$p_\mathbf{a}(\mathbf{x} | \sigma_\mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma_\mathbf{x}^2)^n \det(\mathbf{C}_{\mathbf{n}\mathbf{x}})}} \exp\left(-\frac{1}{2\sigma_\mathbf{x}^2} \mathbf{x}^T \mathbf{C}_{\mathbf{n}\mathbf{x}}^{-1} \mathbf{x}\right). \quad (4.52)$$

For the parameter  $\sigma_\mathbf{x}$ , we assume the Gaussian density

$$p_\mathbf{a}(\sigma_\mathbf{x}) = \frac{1}{\sqrt{2\pi\Delta\sigma_\mathbf{x}^2}} \exp\left(-\frac{1}{2\Delta\sigma_\mathbf{x}^2} (\sigma_\mathbf{x} - \bar{\sigma}_\mathbf{x})^2\right),$$

where the mean  $\bar{\sigma}_\mathbf{x}$  and the variance  $\Delta\sigma_\mathbf{x}^2$  are considered to be known. The joint probability density of  $\mathbf{X}$  and  $\sigma_\mathbf{x}$  is then given by

$$\begin{aligned} p_\mathbf{a}(\mathbf{x}, \sigma_\mathbf{x}) &= p_\mathbf{a}(\mathbf{x} | \sigma_\mathbf{x}) p_\mathbf{a}(\sigma_\mathbf{x}) \\ &\propto \frac{1}{(\sigma_\mathbf{x}^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_\mathbf{x}^2} \mathbf{x}^T \mathbf{C}_{\mathbf{n}\mathbf{x}}^{-1} \mathbf{x} - \frac{1}{2\Delta\sigma_\mathbf{x}^2} (\sigma_\mathbf{x} - \bar{\sigma}_\mathbf{x})^2\right), \end{aligned}$$

the Bayes formula conditioned on the data  $\mathbf{Y}^\delta = \mathbf{y}^\delta$  takes the form

$$\begin{aligned} p(\mathbf{x}, \sigma_\mathbf{x} | \mathbf{y}^\delta) &\propto \frac{1}{(\sigma_\mathbf{x}^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2} (\mathbf{y}^\delta - \mathbf{K}\mathbf{x})^T \mathbf{C}_\delta^{-1} (\mathbf{y}^\delta - \mathbf{K}\mathbf{x}) \right. \\ &\quad \left. - \frac{1}{2\sigma_\mathbf{x}^2} \mathbf{x}^T \mathbf{C}_{\mathbf{n}\mathbf{x}}^{-1} \mathbf{x} - \frac{1}{2\Delta\sigma_\mathbf{x}^2} (\sigma_\mathbf{x} - \bar{\sigma}_\mathbf{x})^2\right), \end{aligned}$$

and the maximum a posteriori estimators  $\hat{\mathbf{x}}$  and  $\hat{\sigma}_\mathbf{x}$  are found by minimizing the a posteriori potential

$$\begin{aligned} V(\mathbf{x}, \sigma_\mathbf{x} | \mathbf{y}^\delta) &= (\mathbf{y}^\delta - \mathbf{K}\mathbf{x})^T \mathbf{C}_\delta^{-1} (\mathbf{y}^\delta - \mathbf{K}\mathbf{x}) \\ &\quad + \frac{1}{\sigma_\mathbf{x}^2} \mathbf{x}^T \mathbf{C}_{\mathbf{n}\mathbf{x}}^{-1} \mathbf{x} + \frac{1}{\Delta\sigma_\mathbf{x}^2} (\sigma_\mathbf{x} - \bar{\sigma}_\mathbf{x})^2 + n \log \sigma_\mathbf{x}^2. \end{aligned}$$

#### 4.3.4 Maximum likelihood estimation

In the Bayes theorem

$$p(\mathbf{x} | \mathbf{y}^\delta) = \frac{p(\mathbf{y}^\delta | \mathbf{x}) p_a(\mathbf{x})}{p(\mathbf{y}^\delta)}, \quad (4.53)$$

the denominator  $p(\mathbf{y}^\delta)$  gives the probability that the data  $\mathbf{Y}^\delta = \mathbf{y}^\delta$  is observed. The marginal density  $p(\mathbf{y}^\delta)$  is obtained by integrating the joint probability density  $p(\mathbf{x}, \mathbf{y}^\delta)$  with respect to  $\mathbf{x}$ , that is,

$$p(\mathbf{y}^\delta) = \int_{\mathbb{R}^n} p(\mathbf{x}, \mathbf{y}^\delta) d\mathbf{x} = \int_{\mathbb{R}^n} p(\mathbf{y}^\delta | \mathbf{x}) p_a(\mathbf{x}) d\mathbf{x}. \quad (4.54)$$

By (4.53) and (4.54), we see that  $p(\mathbf{x} | \mathbf{y}^\delta)$  integrates to 1 as all legitimate probability densities should and that the marginal density  $p(\mathbf{y}^\delta)$  is nothing more than a normalization constant. Despite of this fact,  $p(\mathbf{y}^\delta)$  plays an important role in the design of regularization parameter choice methods and in particular, of the maximum likelihood estimation.

Assuming that the likelihood density  $p(\mathbf{y}^\delta | \mathbf{x})$  and the a priori density  $p_a(\mathbf{x})$  depend on additional parameters, which can be cast in the form of a parameter vector  $\boldsymbol{\theta}$ , we express the marginal density  $p(\mathbf{y}^\delta; \boldsymbol{\theta})$  as

$$p(\mathbf{y}^\delta; \boldsymbol{\theta}) = \int_{\mathbb{R}^n} p(\mathbf{y}^\delta | \mathbf{x}; \boldsymbol{\theta}) p_a(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}. \quad (4.55)$$

The marginal density  $p(\mathbf{y}^\delta; \boldsymbol{\theta})$  is also known as the marginal likelihood function and the maximum likelihood estimator  $\hat{\boldsymbol{\theta}}$  is defined by

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \log p(\mathbf{y}^\delta; \boldsymbol{\theta}).$$

Let us derive the maximum likelihood estimator for Gaussian densities with the covariance matrices (4.20) when  $\sigma^2$  and  $\alpha = \sigma^2/\sigma_x^2$  are unknown, that is, when  $\boldsymbol{\theta}$  is of the form  $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$  with  $\theta_1 = \sigma^2$  and  $\theta_2 = \alpha$ . The a priori density  $p_a(\mathbf{x}; \sigma^2, \alpha)$  and the conditional probability density  $p(\mathbf{y}^\delta | \mathbf{x}; \sigma^2)$  are given by (cf. (4.9) and (4.10))

$$p_a(\mathbf{x}; \sigma^2, \alpha) = \frac{1}{\sqrt{(2\pi\sigma^2)^n \det((\alpha \mathbf{L}^T \mathbf{L})^{-1})}} \exp\left(-\frac{\alpha}{2\sigma^2} \|\mathbf{L}\mathbf{x}\|^2\right)$$

and

$$p(\mathbf{y}^\delta | \mathbf{x}; \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^m}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2\right), \quad (4.56)$$

respectively. Taking into account that

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha \|\mathbf{L}\mathbf{x}\|^2 = (\mathbf{x} - \hat{\mathbf{x}})^T (\mathbf{K}^T \mathbf{K} + \alpha \mathbf{L}^T \mathbf{L}) (\mathbf{x} - \hat{\mathbf{x}}) + \mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta,$$

where  $\hat{\mathbf{x}} = \hat{\mathbf{G}}\mathbf{y}^\delta$  and  $\hat{\mathbf{A}} = \mathbf{K}\hat{\mathbf{G}}$ , we express the integrand in (4.55) as

$$\begin{aligned} & p(\mathbf{y}^\delta | \mathbf{x}; \sigma^2) p_{\mathbf{a}}(\mathbf{x}; \sigma^2, \alpha) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^{n+m} \det((\alpha\mathbf{L}^T\mathbf{L})^{-1})}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \hat{\mathbf{x}})^T (\mathbf{K}^T\mathbf{K} + \alpha\mathbf{L}^T\mathbf{L})(\mathbf{x} - \hat{\mathbf{x}})\right) \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta\right). \end{aligned}$$

Using the normalization condition

$$\begin{aligned} & \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(\mathbf{x} - \hat{\mathbf{x}})^T \left[\sigma^2 (\mathbf{K}^T\mathbf{K} + \alpha\mathbf{L}^T\mathbf{L})^{-1}\right]^{-1} (\mathbf{x} - \hat{\mathbf{x}})\right) d\mathbf{x} \\ &= \sqrt{(2\pi\sigma^2)^n \det((\mathbf{K}^T\mathbf{K} + \alpha\mathbf{L}^T\mathbf{L})^{-1})} \end{aligned}$$

we obtain

$$\begin{aligned} p(\mathbf{y}^\delta; \sigma^2, \alpha) &= \int_{\mathbb{R}^n} p(\mathbf{y}^\delta | \mathbf{x}; \sigma^2) p_{\mathbf{a}}(\mathbf{x}; \sigma^2, \alpha) d\mathbf{x} \\ &= \sqrt{\frac{\det((\mathbf{K}^T\mathbf{K} + \alpha\mathbf{L}^T\mathbf{L})^{-1})}{(2\pi\sigma^2)^m \det((\alpha\mathbf{L}^T\mathbf{L})^{-1})}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta\right). \end{aligned}$$

Taking the logarithm and using the identity

$$\frac{\det((\mathbf{K}^T\mathbf{K} + \alpha\mathbf{L}^T\mathbf{L})^{-1})}{\det((\alpha\mathbf{L}^T\mathbf{L})^{-1})} = \det((\mathbf{K}^T\mathbf{K} + \alpha\mathbf{L}^T\mathbf{L})^{-1} \alpha\mathbf{L}^T\mathbf{L}) = \det(\mathbf{I}_n - \mathbf{A}),$$

yields

$$\begin{aligned} & \log p(\mathbf{y}^\delta; \sigma^2, \alpha) \\ &= -\frac{m}{2} \log(2\pi\sigma^2) + \frac{1}{2} \log(\det(\mathbf{I}_n - \mathbf{A})) - \frac{1}{2\sigma^2} \mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta. \end{aligned} \quad (4.57)$$

Computing the derivative of (4.57) with respect to  $\sigma^2$  and setting it equal to zero gives

$$\hat{\sigma}^2 = \frac{1}{m} \mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta. \quad (4.58)$$

Substituting (4.58) back into (4.57), and using the result

$$\det(\mathbf{I}_n - \mathbf{A}) = \det(\mathbf{I}_m - \hat{\mathbf{A}}) = \prod_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha},$$

we find that

$$\log p(\mathbf{y}^\delta | \hat{\sigma}^2, \alpha) = -\frac{m}{2} \left[ \log(\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta) - \frac{1}{m} \log(\det(\mathbf{I}_m - \hat{\mathbf{A}})) \right] + c,$$

where  $c$  does not depend on  $\alpha$ . Thus, the regularization parameter  $\hat{\alpha}_{\text{mle}}$  which maximizes the log of the marginal likelihood function also minimizes the maximum likelihood function

$$\lambda_{\alpha}^{\delta} = \frac{\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^{\delta}}{\sqrt[m]{\det (\mathbf{I}_m - \hat{\mathbf{A}})}},$$

and we indicate this situation by writing

$$\hat{\alpha}_{\text{mle}} = \arg \min_{\alpha} \lambda_{\alpha}^{\delta}.$$

The numerical simulations performed in the preceding chapter have shown that the maximum likelihood estimation is superior to the generalized cross-validation method in the sense that the minimum of the objective function is not very flat and the estimated regularization parameter is closer to the optimum.

#### 4.3.5 Expectation minimization

The Expectation Minimization (EM) algorithm is an alternative to the maximum likelihood estimation in which the negative of the log of the marginal likelihood function is minimized by an iterative approach. The formulation of the expected minimization as a regularization parameter choice method has been provided by Fitzpatrick (1991), while a very general development can be found in Dempster et al. (1977), and McLachlan and Krishnan (1997). In this section we present a version of the EM algorithm by following the analysis of Vogel (2002).

Taking into account that the a posteriori density  $p(\mathbf{x} | \mathbf{y}^{\delta}; \boldsymbol{\theta})$  is normalized,

$$\int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^{\delta}; \boldsymbol{\theta}) d\mathbf{x} = 1, \quad (4.59)$$

and representing the joint probability density  $p(\mathbf{x}, \mathbf{y}^{\delta}; \boldsymbol{\theta})$  as

$$p(\mathbf{x}, \mathbf{y}^{\delta}; \boldsymbol{\theta}) = p(\mathbf{x} | \mathbf{y}^{\delta}; \boldsymbol{\theta}) p(\mathbf{y}^{\delta}; \boldsymbol{\theta}),$$

we see that for any fixed  $\boldsymbol{\theta}_0$ , the negative of the log of the marginal likelihood function can be expressed as

$$\begin{aligned} -\log p(\mathbf{y}^{\delta}; \boldsymbol{\theta}) &= -\log p(\mathbf{y}^{\delta}; \boldsymbol{\theta}) \int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^{\delta}; \boldsymbol{\theta}_0) d\mathbf{x} \\ &= -\int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^{\delta}; \boldsymbol{\theta}_0) \log p(\mathbf{y}^{\delta}; \boldsymbol{\theta}) d\mathbf{x} \\ &= -\int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^{\delta}; \boldsymbol{\theta}_0) \log \left( \frac{p(\mathbf{x}, \mathbf{y}^{\delta}; \boldsymbol{\theta})}{p(\mathbf{x} | \mathbf{y}^{\delta}; \boldsymbol{\theta})} \right) d\mathbf{x} \\ &= Q(\mathbf{y}^{\delta}, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - H(\mathbf{y}^{\delta}, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \end{aligned}$$

with

$$Q(\mathbf{y}^\delta, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = - \int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}_0) \log p(\mathbf{x}, \mathbf{y}^\delta; \boldsymbol{\theta}) \, d\mathbf{x}$$

and

$$H(\mathbf{y}^\delta, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = - \int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}_0) \log p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}) \, d\mathbf{x}.$$

To evaluate the difference

$$H(\mathbf{y}^\delta, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - H(\mathbf{y}^\delta, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = - \int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}_0) \log \left( \frac{p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta})}{p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}_0)} \right) \, d\mathbf{x},$$

we use the Jensen inequality

$$\int \varphi(g(\mathbf{x})) f(\mathbf{x}) \, d\mathbf{x} \geq \varphi \left( \int g(\mathbf{x}) f(\mathbf{x}) \, d\mathbf{x} \right)$$

for the convex function  $\varphi(u) = -\log u$ , that is,

$$- \int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}_0) \log \left( \frac{p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta})}{p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}_0)} \right) \, d\mathbf{x} \geq - \log \left( \int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^\delta; \boldsymbol{\theta}) \, d\mathbf{x} \right) = 0,$$

and obtain

$$-H(\mathbf{y}^\delta, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq -H(\mathbf{y}^\delta, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0).$$

Assuming that  $\boldsymbol{\theta}$  is such that

$$Q(\mathbf{y}^\delta, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq Q(\mathbf{y}^\delta, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0),$$

it follows that

$$-\log p(\mathbf{y}^\delta; \boldsymbol{\theta}) \leq -\log p(\mathbf{y}^\delta; \boldsymbol{\theta}_0).$$

The EM algorithm seeks to minimize  $-\log p(\mathbf{y}^\delta; \boldsymbol{\theta})$  by iteratively applying the following two steps:

- (1) *Expectation step.* Calculate the function  $Q(\mathbf{y}^\delta, \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_k)$  for the a posteriori density under the current estimator  $\widehat{\boldsymbol{\theta}}_k$ ,

$$Q(\mathbf{y}^\delta, \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_k) = - \int_{\mathbb{R}^n} p(\mathbf{x} | \mathbf{y}^\delta; \widehat{\boldsymbol{\theta}}_k) \log (p(\mathbf{y}^\delta | \mathbf{x}; \boldsymbol{\theta}) p_a(\mathbf{x}; \boldsymbol{\theta})) \, d\mathbf{x}.$$

- (2) *Minimization step.* Find the parameter vector  $\widehat{\boldsymbol{\theta}}_{k+1}$  which minimizes this function, that is,

$$\widehat{\boldsymbol{\theta}}_{k+1} = \arg \min_{\boldsymbol{\theta}} Q(\mathbf{y}^\delta, \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_k).$$

Two main peculiarities of the EM algorithm can be evidenced:

- (1) Even if the algorithm has a stable point, there is no guarantee that this stable point is a global minimum of  $-\log p(\mathbf{y}^\delta; \boldsymbol{\theta})$ , or even a local minimum. If the function  $Q(\mathbf{y}^\delta, \boldsymbol{\theta}, \boldsymbol{\theta}')$  is continuous, convergence to a stationary point of  $-\log p(\mathbf{y}^\delta; \boldsymbol{\theta})$  is guaranteed.
- (2) The solution generally depends on the initialization.

To illustrate how the EM algorithm works, we consider Gaussian densities with the covariance matrices (4.20), and choose the parameter vector  $\theta$  as  $\theta = [\theta_1, \theta_2]^T$  with  $\theta_1 = \sigma_x^2$  and  $\theta_2 = \sigma^2$ . The a priori density  $p_a(\mathbf{x}; \sigma_x^2)$  and the conditional probability density  $p(\mathbf{y}^\delta | \mathbf{x}; \sigma^2)$  are given by (4.52) and (4.56), respectively. Using the results

$$\begin{aligned} \frac{\partial}{\partial \sigma_x^2} \log(p(\mathbf{y}^\delta | \mathbf{x}; \sigma^2) p_a(\mathbf{x}; \sigma_x^2)) &= -\frac{n}{2\sigma_x^2} + \frac{1}{2\sigma_x^4} \mathbf{x}^T \mathbf{C}_{\mathbf{nx}}^{-1} \mathbf{x}, \\ \frac{\partial}{\partial \sigma^2} \log(p(\mathbf{y}^\delta | \mathbf{x}; \sigma^2) p_a(\mathbf{x}; \sigma_x^2)) &= -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2, \end{aligned}$$

we deduce that the EM iteration step yields the recurrence relations

$$\hat{\sigma}_{xk+1}^2 = \frac{1}{n} \int_{\mathbb{R}^n} \mathbf{x}^T \mathbf{C}_{\mathbf{nx}}^{-1} \mathbf{x} p(\mathbf{x} | \mathbf{y}^\delta; \hat{\sigma}_{xk}^2, \hat{\sigma}_k^2) d\mathbf{x}, \quad (4.60)$$

$$\hat{\sigma}_{k+1}^2 = \frac{1}{m} \int_{\mathbb{R}^n} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 p(\mathbf{x} | \mathbf{y}^\delta; \hat{\sigma}_{xk}^2, \hat{\sigma}_k^2) d\mathbf{x}. \quad (4.61)$$

To compute the  $n$ -dimensional integrals in (4.60) and (4.61) we may use the Monte Carlo method (Tarantola, 2005). As the a posteriori density under the current estimator is Gaussian, the integration process involves the following steps:

- (1) for  $\hat{\sigma}_{xk}^2$  and  $\hat{\sigma}_k^2$ , compute the maximum a posteriori estimator  $\hat{\mathbf{x}}_k$  and the a posteriori covariance matrix  $\hat{\mathbf{C}}_{xk}$ ;
- (2) generate a random sample  $\{\mathbf{x}_{ki}\}_{i=1, \overline{N}}$  of a Gaussian distribution with mean vector  $\hat{\mathbf{x}}_k$  and covariance matrix  $\hat{\mathbf{C}}_{xk}$ ;
- (3) estimate the integrals as

$$\begin{aligned} \int_{\mathbb{R}^n} \mathbf{x}^T \mathbf{C}_{\mathbf{nx}}^{-1} \mathbf{x} p(\mathbf{x} | \mathbf{y}^\delta; \hat{\sigma}_{xk}^2, \hat{\sigma}_k^2) d\mathbf{x} &\approx \frac{1}{N} \sum_{i=1}^N \mathbf{x}_{ki}^T \mathbf{C}_{\mathbf{nx}}^{-1} \mathbf{x}_{ki}, \\ \int_{\mathbb{R}^n} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 p(\mathbf{x} | \mathbf{y}^\delta; \hat{\sigma}_{xk}^2, \hat{\sigma}_k^2) d\mathbf{x} &\approx \frac{1}{N} \sum_{i=1}^N \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{ki}\|^2. \end{aligned}$$

This integration process is quite demanding, and as a result, the method may become very time-consuming.

#### 4.3.6 A general regularization parameter choice method

In this section we present a general technique for constructing regularization parameter choice methods in statistical inversion theory. Our analysis follows the treatment of Neumaier (1998) and enables us to introduce the generalized cross-validation method and the maximum likelihood estimation in a natural way.

Assuming Gaussian densities with the covariance matrices (4.20) and considering a generalized singular value decomposition of the matrix pair  $(\mathbf{K}, \mathbf{L})$ , i.e.,  $\mathbf{K} = \mathbf{U}\Sigma_1\mathbf{W}^{-1}$  and  $\mathbf{L} = \mathbf{V}\Sigma_2\mathbf{W}^{-1}$ , we express the covariance matrix of the data  $\mathbf{Y}^\delta$  as (cf. (4.24))

$$\mathcal{E}\{\mathbf{Y}^\delta \mathbf{Y}^{\delta T}\} = \mathbf{K}\mathbf{C}_x\mathbf{K}^T + \mathbf{C}_\delta = \sigma_x^2 \mathbf{K}(\mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T + \sigma^2 \mathbf{I}_m = \mathbf{U}\Sigma_y \mathbf{U}^T,$$

where

$$\begin{aligned}\Sigma_y &= \sigma_x^2 \Sigma_1 (\Sigma_2^T \Sigma_2)^{-1} \Sigma_1^T + \sigma^2 \mathbf{I}_m \\ &= \begin{bmatrix} \text{diag}(\sigma_x^2 \gamma_i^2 + \sigma^2)_{n \times n} & \mathbf{0} \\ \mathbf{0} & \text{diag}(\sigma^2)_{(m-n) \times (m-n)} \end{bmatrix}.\end{aligned}$$

Next, we define the scaled data

$$\bar{\mathbf{Y}}^\delta = \mathbf{U}^T \mathbf{Y}^\delta,$$

and observe that  $\bar{\mathbf{Y}}^\delta$  has a diagonal covariance matrix, which is given by

$$\mathcal{E} \{ \bar{\mathbf{Y}}^\delta \bar{\mathbf{Y}}^{\delta T} \} = \mathcal{E} \{ \mathbf{U}^T \mathbf{Y}^\delta \mathbf{Y}^{\delta T} \mathbf{U} \} = \Sigma_y.$$

If  $\sigma_x$  and  $\sigma$  correctly describe the covariance matrix of the true state and the instrumental noise covariance matrix, respectively, we must have

$$\mathcal{E} \{ \bar{Y}_i^{\delta 2} \} = \sigma_x^2 \gamma_i^2 + \sigma^2, \quad i = 1, \dots, m, \quad (4.62)$$

where  $\bar{Y}_i^\delta = \mathbf{u}_i^T \mathbf{Y}^\delta$  for  $i = 1, \dots, m$ , and  $\gamma_i = 0$  for  $i = n + 1, \dots, m$ . If  $\sigma_x$  and  $\sigma$  are unknown, we can find the estimators  $\hat{\sigma}_x$  and  $\hat{\sigma}$  from the equations

$$\mathcal{E} \{ \bar{Y}_i^{\delta 2} \} = \hat{\sigma}_x^2 \gamma_i^2 + \hat{\sigma}^2, \quad i = 1, \dots, m. \quad (4.63)$$

However, since only one realization of the random vector  $\bar{\mathbf{Y}}^\delta$  is known, the calculation of these estimators may lead to erroneous results and we must replace (4.63) by another selection criterion. For this purpose, we set (cf. (4.62))

$$a_i(\boldsymbol{\theta}) = \theta_1 \gamma_i^2 + \theta_2, \quad (4.64)$$

with  $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$ ,  $\theta_1 = \sigma_x^2$  and  $\theta_2 = \sigma^2$ , and define the function

$$f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta}) = \sum_{i=1}^m \psi(a_i(\boldsymbol{\theta})) + \psi'(a_i(\boldsymbol{\theta})) [\bar{Y}_i^{\delta 2} - a_i(\boldsymbol{\theta})],$$

with  $\psi$  being a strictly concave function. The expected value of  $f$  is given by

$$\mathcal{E} \{ f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta}) \} = \sum_{i=1}^m \psi(a_i(\boldsymbol{\theta})) + \psi'(a_i(\boldsymbol{\theta})) [\mathcal{E} \{ \bar{Y}_i^{\delta 2} \} - a_i(\boldsymbol{\theta})],$$

whence, defining the estimator  $\hat{\boldsymbol{\theta}}$  through the relation

$$\mathcal{E} \{ \bar{Y}_i^{\delta 2} \} = a_i(\hat{\boldsymbol{\theta}}), \quad i = 1, \dots, m, \quad (4.65)$$

$\mathcal{E} \{ f \}$  can be expressed as

$$\mathcal{E} \{ f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta}) \} = \sum_{i=1}^m \psi(a_i(\boldsymbol{\theta})) + \psi'(a_i(\boldsymbol{\theta})) [a_i(\hat{\boldsymbol{\theta}}) - a_i(\boldsymbol{\theta})].$$



Then, we obtain

$$\begin{aligned} \mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\} - \mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \hat{\boldsymbol{\theta}})\} &= \sum_{i=1}^m \psi(a_i(\boldsymbol{\theta})) - \psi(a_i(\hat{\boldsymbol{\theta}})) \\ &\quad + \psi'(a_i(\boldsymbol{\theta})) [a_i(\hat{\boldsymbol{\theta}}) - a_i(\boldsymbol{\theta})]. \end{aligned}$$

Considering the second-order Taylor expansion

$$\psi(a_i(\boldsymbol{\theta})) - \psi(a_i(\hat{\boldsymbol{\theta}})) + \psi'(a_i(\boldsymbol{\theta})) [a_i(\hat{\boldsymbol{\theta}}) - a_i(\boldsymbol{\theta})] = -\frac{1}{2} \psi''(\xi_i) [a_i(\hat{\boldsymbol{\theta}}) - a_i(\boldsymbol{\theta})]^2$$

with some  $\xi_i$  between  $a_i(\boldsymbol{\theta})$  and  $a_i(\hat{\boldsymbol{\theta}})$ , and taking into account that  $\psi$  is strictly concave, we deduce that each term in the sum is non-negative and vanishes only for  $a_i(\boldsymbol{\theta}) = a_i(\hat{\boldsymbol{\theta}})$ . Thus, we have

$$\mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\} \geq \mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \hat{\boldsymbol{\theta}})\},$$

for all  $\boldsymbol{\theta}$ . If, in addition,  $\hat{\boldsymbol{\theta}}$  is determined uniquely by (4.65), then  $\hat{\boldsymbol{\theta}}$  is the unique global minimizer of  $\mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\}$ , and we propose a regularization parameter choice method in which the estimator  $\hat{\boldsymbol{\theta}}$  is computed as

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\}. \quad (4.66)$$

Different regularization parameter choice methods can be obtained by choosing the concave function  $\psi$  in an appropriate way.

### Generalized cross-validation

For the choice

$$\psi(a) = 1 - \frac{1}{a},$$

we obtain

$$\mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\} = m + \sum_{i=1}^m \left[ \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{a_i(\boldsymbol{\theta})^2} - \frac{2}{a_i(\boldsymbol{\theta})} \right]. \quad (4.67)$$

As  $\hat{\boldsymbol{\theta}}$  is the unique global minimizer of  $\mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\}$ , the gradient

$$\nabla \mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\} = -2 \sum_{i=1}^m \left[ \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{a_i(\boldsymbol{\theta})^3} - \frac{1}{a_i(\boldsymbol{\theta})^2} \right] \nabla a_i(\boldsymbol{\theta})$$

vanishes at  $\hat{\boldsymbol{\theta}}$ . Thus,

$$\hat{\boldsymbol{\theta}}^T \nabla \mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \hat{\boldsymbol{\theta}})\} = 0, \quad (4.68)$$

and since (cf. (4.64))

$$\boldsymbol{\theta}^T \nabla a_i(\boldsymbol{\theta}) = a_i(\boldsymbol{\theta}),$$

we deduce that

$$\sum_{i=1}^m \left[ \frac{\mathcal{E} \{ \bar{Y}_i^{\delta 2} \}}{a_i(\hat{\boldsymbol{\theta}})^2} - \frac{1}{a_i(\hat{\boldsymbol{\theta}})} \right] = 0. \quad (4.69)$$

Equation (4.69) together with the relation

$$a_i(\hat{\boldsymbol{\theta}}) = \hat{\sigma}_x^2 \gamma_i^2 + \hat{\sigma}^2,$$

gives

$$\hat{\sigma}_x^2 = \frac{p(\hat{\alpha})}{q(\hat{\alpha})}, \quad \hat{\sigma}^2 = \hat{\alpha} \hat{\sigma}_x^2, \quad (4.70)$$

where

$$p(\alpha) = \sum_{i=1}^m \frac{\mathcal{E} \{ \bar{Y}_i^{\delta 2} \}}{(\gamma_i^2 + \alpha)^2}$$

and

$$q(\alpha) = \sum_{i=1}^m \frac{1}{\gamma_i^2 + \alpha}.$$

From (4.70), it is apparent that  $\hat{\sigma}_x^2$  and  $\hat{\sigma}^2$  are expressed in terms of the single parameter  $\hat{\alpha}$ , and by (4.67) and (4.69), we find that

$$-\mathcal{E} \{ f(\bar{\mathbf{Y}}^\delta, \hat{\boldsymbol{\theta}}) \} + m = \sum_{i=1}^m \frac{1}{a_i(\hat{\boldsymbol{\theta}})} = \frac{1}{\hat{\sigma}_x^2} q(\hat{\alpha}) = \frac{q(\hat{\alpha})^2}{p(\hat{\alpha})}. \quad (4.71)$$

Now, if  $\hat{\alpha}$  minimizes the function

$$v_\alpha = \frac{p(\alpha)}{q(\alpha)^2} = \frac{\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 \mathcal{E} \{ \bar{Y}_i^{\delta 2} \}}{\left( \sum_{i=1}^m \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2},$$

then by (4.71),  $\hat{\alpha}$  maximizes  $-\mathcal{E} \{ f \}$ , or equivalently,  $\hat{\alpha}$  minimizes  $\mathcal{E} \{ f \}$ . In practice, the expectation  $\mathcal{E} \{ \bar{Y}_i^{\delta 2} \}$  cannot be computed since only a single realization  $\bar{y}_i^\delta = \mathbf{u}_i^T \mathbf{y}^\delta$  of  $\bar{Y}_i^\delta$  is known. To obtain a practical regularization parameter choice method, instead of  $v_\alpha$  we consider the function

$$v_\alpha^\delta = \frac{\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2}{\left( \sum_{i=1}^m \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2} = \frac{\| \mathbf{y}^\delta - \mathbf{K} \hat{\mathbf{x}} \|^2}{\left[ \text{trace}(\mathbf{I}_m - \hat{\mathbf{A}}) \right]^2},$$

which represents the generalized cross-validation function discussed in Chapter 3.

Note that for  $\psi(a) = (1 - 1/a^q)/q$  with  $q > -1$  and  $q \neq 0$ , we obtain

$$\mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\} = \frac{m}{q} + \sum_{i=1}^m \left[ \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{a_i(\boldsymbol{\theta})^{q+1}} - \frac{1 + \frac{1}{q}}{a_k(\boldsymbol{\theta})^q} \right],$$

and we are led to a generalization of the cross-validation function of the form

$$v_{\alpha q}^\delta = \frac{\sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^{q+1} (\mathbf{u}_i^T \mathbf{y}^\delta)^{2q}}{\left[ \sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^q \right]^{q+1}}.$$

### Maximum likelihood estimation

For the choice

$$\psi(a) = \log a,$$

we obtain

$$\mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \boldsymbol{\theta})\} = -m + \sum_{i=1}^m \left[ \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{a_i(\boldsymbol{\theta})} + \log a_i(\boldsymbol{\theta}) \right], \quad (4.72)$$

and the minimization condition (4.68) yields

$$\sum_{i=1}^m \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{a_i(\hat{\boldsymbol{\theta}})} = m. \quad (4.73)$$

As before, equation (4.73) implies that  $\hat{\sigma}_x^2$  and  $\hat{\sigma}^2$  can be expressed in terms of the single parameter  $\hat{\alpha}$  through the relations

$$\hat{\sigma}_x^2 = \frac{1}{m} \sum_{i=1}^m \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{\gamma_i^2 + \hat{\alpha}}, \quad \hat{\sigma}^2 = \hat{\alpha} \hat{\sigma}_x^2, \quad (4.74)$$

and we find that

$$\begin{aligned} \mathcal{E} \{f(\bar{\mathbf{Y}}^\delta, \hat{\boldsymbol{\theta}})\} + m \log m &= m \log m + \sum_{i=1}^m \log a_i(\hat{\boldsymbol{\theta}}) \\ &= m \log m + m \log \hat{\sigma}_x^2 + \sum_{i=1}^m \log (\gamma_i^2 + \hat{\alpha}) \\ &= m \log \left( \sum_{i=1}^m \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{\gamma_i^2 + \hat{\alpha}} \right) + \sum_{i=1}^m \log (\gamma_i^2 + \hat{\alpha}) \\ &= m \left[ \log \left( \sum_{i=1}^m \frac{\mathcal{E} \{\bar{Y}_i^{\delta 2}\}}{\gamma_i^2 + \hat{\alpha}} \right) - \frac{1}{m} \log \left( \prod_{i=1}^m \frac{1}{\gamma_i^2 + \hat{\alpha}} \right) \right]. \end{aligned}$$

Hence, if  $\hat{\alpha}$  minimizes the function

$$\lambda_{\alpha} = \frac{\sum_{i=1}^m \mathcal{E}\{\bar{Y}_i^{\delta 2}\}}{\sqrt{\prod_{i=1}^m \frac{1}{\gamma_i^2 + \alpha}}},$$

then  $\hat{\alpha}$  minimizes  $\mathcal{E}\{f\}$ . In practice, we replace  $\mathcal{E}\{\bar{Y}_i^{\delta 2}\}$  by  $(\mathbf{u}_i^T \mathbf{y}^{\delta})^2$  and minimize the maximum likelihood function

$$\lambda_{\alpha}^{\delta} = \frac{\sum_{i=1}^m \frac{(\mathbf{u}_i^T \mathbf{y}^{\delta})^2}{\gamma_i^2 + \alpha}}{\sqrt{\prod_{i=1}^m \frac{1}{\gamma_i^2 + \alpha}}} = \frac{\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^{\delta}}{\sqrt{\det(\mathbf{I}_m - \hat{\mathbf{A}})}}. \quad (4.75)$$

An equivalent interpretation of the maximum likelihood estimation can be given as follows. Let us consider the scaled data  $\bar{\mathbf{Y}}^{\delta} = \mathbf{U}^T \mathbf{Y}^{\delta}$  and let us compute the maximum likelihood estimator  $\hat{\boldsymbol{\theta}}$  as

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \log p(\bar{\mathbf{y}}^{\delta}; \boldsymbol{\theta}),$$

with

$$p(\bar{\mathbf{y}}^{\delta}; \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma_{\bar{\mathbf{y}}}(\boldsymbol{\theta}))}} \exp\left(-\frac{1}{2} \bar{\mathbf{y}}^{\delta T} \Sigma_{\bar{\mathbf{y}}}^{-1}(\boldsymbol{\theta}) \bar{\mathbf{y}}^{\delta}\right),$$

and

$$\Sigma_{\bar{\mathbf{y}}}(\boldsymbol{\theta}) = \begin{bmatrix} \text{diag}(\theta_1 \gamma_i^2 + \theta_2)_{n \times n} & \mathbf{0} \\ \mathbf{0} & \text{diag}(\theta_2)_{(m-n) \times (m-n)} \end{bmatrix}.$$

Then, taking into account that

$$\begin{aligned} \log p(\bar{\mathbf{y}}^{\delta}; \boldsymbol{\theta}) &= -\frac{1}{2} \bar{\mathbf{y}}^{\delta T} \Sigma_{\bar{\mathbf{y}}}^{-1}(\boldsymbol{\theta}) \bar{\mathbf{y}}^{\delta} - \frac{1}{2} \log(\det \Sigma_{\bar{\mathbf{y}}}(\boldsymbol{\theta})) + c \\ &= -\frac{1}{2} \left[ \sum_{i=1}^m \frac{\bar{y}_i^{\delta 2}}{\theta_1 \gamma_i^2 + \theta_2} + \log \left( \prod_{i=1}^m (\theta_1 \gamma_i^2 + \theta_2) \right) \right] + c, \end{aligned}$$

where  $c$  does not depend on  $\boldsymbol{\theta}$ , we see that the maximization of  $\log p(\bar{\mathbf{y}}^{\delta}; \boldsymbol{\theta})$  is equivalent to the minimization of  $f(\bar{\mathbf{Y}}^{\delta}, \boldsymbol{\theta})$  as in (4.72).

#### 4.3.7 Noise variance estimators

In a semi-stochastic setting, we have estimated the noise variance by looking at the behavior of the residual norm in the limit of small  $\alpha$ . This technique considers the solution of the inverse problem without regularization and requires an additional computational step.

In this section we present methods for estimating the noise variance, which do not suffer from this inconvenience.

In the analysis of the generalized cross-validation method and the maximum likelihood estimation we considered the parameter vector  $\theta$ , whose components depend on the regularization parameter  $\alpha$  and the noise variance  $\sigma^2$ . In fact, these methods are ideal candidates for estimating both the regularization parameter and the noise variance.

In the the generalized cross-validation method, the second relation in (4.70) gives the noise variance estimator

$$\hat{\sigma}_{\text{gcv}}^2 = \hat{\alpha}_{\text{gcv}} \frac{p(\hat{\alpha}_{\text{gcv}})}{q(\hat{\alpha}_{\text{gcv}})} \approx \frac{\sum_{i=1}^m \left( \frac{\hat{\alpha}_{\text{gcv}}}{\gamma_i^2 + \hat{\alpha}_{\text{gcv}}} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2}{\sum_{i=1}^m \frac{\hat{\alpha}_{\text{gcv}}}{\gamma_i^2 + \hat{\alpha}_{\text{gcv}}}} = \frac{\|\mathbf{y}^\delta - \mathbf{K}\hat{\mathbf{x}}\|^2}{\text{trace}(\mathbf{I}_m - \hat{\mathbf{A}})}, \quad (4.76)$$

where  $\hat{\mathbf{x}}$  and  $\hat{\mathbf{A}}$  are computed for the regularization parameter  $\hat{\alpha}_{\text{gcv}}$ . The noise variance estimator (4.76) has been proposed by Wahba (1983) and numerical experiments presented by a number of researchers support the choice of this estimator (Fessler, 1991; Nychka, 1988; Thompson et al., 1991).

In the maximum likelihood estimation, a noise variance estimator can be constructed by using (4.58); the result is

$$\hat{\sigma}_{\text{mle}}^2 = \frac{1}{m} \mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta,$$

where  $\hat{\mathbf{A}}$  is computed for the regularization parameter  $\hat{\alpha}_{\text{mle}}$ . Numerical experiments where this estimator is tested has been reported by Galatsanos and Katsaggelos (1992).

An estimator which is similar to (4.76) can be derived in the framework of the unbiased predictive risk estimator method. This selection criterion chooses the regularization parameter  $\hat{\alpha}_{\text{pr}}$  as the minimizer of the function

$$\pi_\alpha^\delta = \sum_{i=1}^m \left( \frac{\alpha}{\gamma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + 2\sigma^2 \sum_{i=1}^n \frac{\gamma_i^2}{\gamma_i^2 + \alpha} - m\sigma^2.$$

Taking the derivative of  $\pi_\alpha^\delta$  with respect to  $\alpha$ , and setting it equal to zero gives

$$\sigma^2 \sum_{i=1}^n \frac{\gamma_i^2}{(\gamma_i^2 + \alpha)^2} = \sum_{i=1}^n \frac{\alpha \gamma_i^2}{(\gamma_i^2 + \alpha)^3} (\mathbf{u}_i^T \mathbf{y}^\delta)^2. \quad (4.77)$$

By straightforward calculation we find that

$$\text{trace}(\hat{\mathbf{A}}(\mathbf{I}_m - \hat{\mathbf{A}})) = \sum_{i=1}^n \frac{\alpha \gamma_i^2}{(\gamma_i^2 + \alpha)^2}$$

and that

$$\mathbf{y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}})^T \hat{\mathbf{A}} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta = \sum_{i=1}^n \frac{\alpha^2 \gamma_i^2}{(\gamma_i^2 + \alpha)^3} (\mathbf{u}_i^T \mathbf{y}^\delta)^2.$$

Now, taking into account that  $\hat{\alpha}_{\text{pr}}$  and  $\hat{\alpha}_{\text{gcv}}$  are asymptotically equivalent, equation (4.77) can be used to estimate the noise variance; we obtain

$$\hat{\sigma}_{\text{pr}}^2 = \frac{\mathbf{y}^{\delta T} \left( \mathbf{I}_m - \hat{\mathbf{A}} \right)^T \hat{\mathbf{A}} \left( \mathbf{I}_m - \hat{\mathbf{A}} \right) \mathbf{y}^{\delta}}{\text{trace} \left( \hat{\mathbf{A}} \left( \mathbf{I}_m - \hat{\mathbf{A}} \right) \right)},$$

where  $\hat{\mathbf{A}}$  is computed for the generalized cross-validation parameter  $\hat{\alpha}_{\text{gcv}}$ . Since

$$\mathbf{y}^{\delta} - \mathbf{K}\hat{\mathbf{x}} = \left( \mathbf{I}_m - \hat{\mathbf{A}} \right) \mathbf{y}^{\delta},$$

we see that this estimator is similar to (4.76); the only difference is the multiplication with the influence matrix in both the numerator and denominator.

#### 4.4 Marginalizing method

In a stochastic setting, a two-component data model reads as

$$\mathbf{Y}^{\delta} = \mathbf{K}_1 \mathbf{X}_1 + \mathbf{K}_2 \mathbf{X}_2 + \Delta, \quad (4.78)$$

where  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are assumed to be independent Gaussian random vectors characterized by  $\mathbf{X}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{x}1})$  and  $\mathbf{X}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{x}2})$ . The dimensions of the random vectors  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are  $n_1$  and  $n_2$ , respectively, and we have  $n_1 + n_2 = n$ . The maximum a posteriori estimator  $\hat{\mathbf{x}}$  of the state

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

is obtained from the Bayes theorem

$$p(\mathbf{x}_1, \mathbf{x}_2 | \mathbf{y}^{\delta}) = \frac{p(\mathbf{y}^{\delta} | \mathbf{x}_1, \mathbf{x}_2) p_{\mathbf{a}}(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{y}^{\delta})} = \frac{p(\mathbf{y}^{\delta} | \mathbf{x}_1, \mathbf{x}_2) p_{\mathbf{a}}(\mathbf{x}_1) p_{\mathbf{a}}(\mathbf{x}_2)}{p(\mathbf{y}^{\delta})}, \quad (4.79)$$

where the a priori densities and the likelihood density are given by

$$p_{\mathbf{a}}(\mathbf{x}_i) \propto \exp \left( -\frac{1}{2} \mathbf{x}_i^T \mathbf{C}_{\mathbf{x}i}^{-1} \mathbf{x}_i \right), \quad i = 1, 2, \quad (4.80)$$

and

$$p(\mathbf{y}^{\delta} | \mathbf{x}_1, \mathbf{x}_2) \propto \exp \left( -\frac{1}{2} (\mathbf{y}^{\delta} - \mathbf{K}_1 \mathbf{x}_1 - \mathbf{K}_2 \mathbf{x}_2)^T \mathbf{C}_{\delta}^{-1} (\mathbf{y}^{\delta} - \mathbf{K}_1 \mathbf{x}_1 - \mathbf{K}_2 \mathbf{x}_2) \right), \quad (4.81)$$

respectively.

To show the equivalence between classical regularization and statistical inversion, we assume Gaussian densities with covariance matrices of the form

$$\mathbf{C}_{\delta} = \sigma^2 \mathbf{I}_m, \quad \mathbf{C}_{\mathbf{x}i} = \sigma_{\mathbf{x}i}^2 \mathbf{C}_{\mathbf{n}xi} = \sigma_{\mathbf{x}i}^2 (\mathbf{L}_i^T \mathbf{L}_i)^{-1}, \quad i = 1, 2, \quad (4.82)$$

and write the penalty term in the expression of  $\sigma^2 V(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{y}^\delta)$  as

$$\sigma^2 \left( \frac{1}{\sigma_{x1}^2} \|\mathbf{L}_1 \mathbf{x}_1\|^2 + \frac{1}{\sigma_{x2}^2} \|\mathbf{L}_2 \mathbf{x}_2\|^2 \right) = \alpha \left[ \omega \|\mathbf{L}_1 \mathbf{x}_1\|^2 + (1 - \omega) \|\mathbf{L}_2 \mathbf{x}_2\|^2 \right].$$

Then, it is readily seen that the regularization parameter  $\alpha$  and the weighting factor  $\omega$  are given by

$$\alpha = \frac{\sigma^2}{\sigma_x^2}, \quad \omega = \frac{\sigma_x^2}{\sigma_{x1}^2}, \quad (4.83)$$

where

$$\frac{1}{\sigma_x^2} = \frac{1}{\sigma_{x1}^2} + \frac{1}{\sigma_{x2}^2}.$$

In the framework of classical regularization theory we discussed multi-parameter regularization methods for computing  $\alpha$  and  $\omega$ , or equivalently, for estimating  $\sigma_{x1}$  and  $\sigma_{x2}$ . An interesting situation occurs when the statistics of  $\mathbf{X}_2$  is known, and only  $\sigma_{x1}$  is the parameter of the retrieval. In this case we can reduce the dimension of the minimization problem by using the so-called marginalizing technique. The idea is to formulate a minimization problem for the first component of the state vector by taking into account the statistics of the second component. The maximum a posteriori estimator for the first component of the state vector is defined as

$$\hat{\mathbf{x}}_1 = \arg \max_{\mathbf{x}_1} p(\mathbf{x}_1 \mid \mathbf{y}^\delta).$$

To compute the marginal a posteriori density  $p(\mathbf{x}_1 \mid \mathbf{y}^\delta)$ , we must integrate the density  $p(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{y}^\delta)$  over  $\mathbf{x}_2$ ,

$$p(\mathbf{x}_1 \mid \mathbf{y}^\delta) = \int_{\mathbb{R}^{n_2}} p(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{y}^\delta) d\mathbf{x}_2 = \frac{p_a(\mathbf{x}_1)}{p(\mathbf{y}^\delta)} \int_{\mathbb{R}^{n_2}} p(\mathbf{y}^\delta \mid \mathbf{x}_1, \mathbf{x}_2) p_a(\mathbf{x}_2) d\mathbf{x}_2, \quad (4.84)$$

where the a priori densities and the likelihood density are given by (4.80) and (4.81), respectively. To evaluate the integral, we have to arrange the argument of the exponential function as a quadratic function in  $\mathbf{x}_2$ . For this purpose, we employ the technique which we used to derive the mean vector and the covariance matrix of the a posteriori density  $p(\mathbf{x} \mid \mathbf{y}^\delta)$  in the one-parameter case, that is,

$$\begin{aligned} & [(\mathbf{y}^\delta - \mathbf{K}_1 \mathbf{x}_1) - \mathbf{K}_2 \mathbf{x}_2]^T \mathbf{C}_\delta^{-1} [(\mathbf{y}^\delta - \mathbf{K}_1 \mathbf{x}_1) - \mathbf{K}_2 \mathbf{x}_2] + \mathbf{x}_2^T \mathbf{C}_{x2}^{-1} \mathbf{x}_2 \\ &= (\mathbf{x}_2 - \bar{\mathbf{x}}_2)^T \hat{\mathbf{C}}_{x2}^{-1} (\mathbf{x}_2 - \bar{\mathbf{x}}_2) + (\mathbf{y}^\delta - \mathbf{K}_1 \mathbf{x}_1)^T (\mathbf{K}_2 \mathbf{C}_{x2} \mathbf{K}_2^T + \mathbf{C}_\delta)^{-1} (\mathbf{y}^\delta - \mathbf{K}_1 \mathbf{x}_1), \end{aligned}$$

with

$$\bar{\mathbf{x}}_2 = \mathbf{G}_2 (\mathbf{y}^\delta - \mathbf{K}_1 \mathbf{x}_1), \quad \mathbf{G}_2 = (\mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 + \mathbf{C}_{x2}^{-1})^{-1} \mathbf{K}_2^T \mathbf{C}_\delta^{-1},$$

and

$$\hat{\mathbf{C}}_{x2} = (\mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 + \mathbf{C}_{x2}^{-1})^{-1}.$$

Using the normalization condition for the Gaussian density

$$\exp \left( -\frac{1}{2} (\mathbf{x}_2 - \bar{\mathbf{x}}_2)^T \hat{\mathbf{C}}_{x2}^{-1} (\mathbf{x}_2 - \bar{\mathbf{x}}_2) \right),$$

we obtain

$$p(\mathbf{x}_1 | \mathbf{y}^\delta) \propto \exp \left( -\frac{1}{2} (\mathbf{y}^\delta - \mathbf{K}_1 \mathbf{x}_1)^T (\mathbf{K}_2 \mathbf{C}_{x2} \mathbf{K}_2^T + \mathbf{C}_\delta)^{-1} (\mathbf{y}^\delta - \mathbf{K}_1 \mathbf{x}_1) - \frac{1}{2} \mathbf{x}_1^T \mathbf{C}_{x1}^{-1} \mathbf{x}_1 \right),$$

and it is apparent that  $\hat{\mathbf{x}}_1$  is given by (4.12) and (4.13), with  $\mathbf{K}$  replaced by  $\mathbf{K}_1$  and  $\mathbf{C}_\delta$  replaced by

$$\mathbf{C}_{\delta_y} = \mathbf{C}_\delta + \mathbf{K}_2 \mathbf{C}_{x2} \mathbf{K}_2^T. \quad (4.85)$$

Thus, when retrieving the first component of the state vector we may interpret the data error covariance matrix as being the sum of the instrumental noise covariance matrix plus a contribution due to the second component (Rodgers, 2000).

Actually, the marginalizing method can be justified more simply as follows: express the data model (4.78) as

$$\mathbf{Y}^\delta = \mathbf{K}_1 \mathbf{X}_1 + \Delta_y,$$

where the random data error  $\Delta_y$  is given by

$$\Delta_y = \mathbf{K}_2 \mathbf{X}_2 + \Delta,$$

and use the result  $\mathcal{E}\{\Delta_y\} = \mathbf{0}$  to conclude that the covariance matrix  $\mathbf{C}_{\delta_y} = \mathcal{E}\{\Delta_y \Delta_y^T\}$  is given by (4.85). In the state space, the marginalizing method yields the random model parameter error

$$\mathbf{E}_{\text{mp}} = -\hat{\mathbf{G}} \mathbf{K}_2 \mathbf{X}_2,$$

characterized by

$$\mathcal{E}\{\mathbf{E}_{\text{mp}}\} = \mathbf{0}, \quad \mathbf{C}_{\text{emp}} = \hat{\mathbf{G}} \mathbf{K}_2 \mathbf{C}_{x2} \mathbf{K}_2^T \hat{\mathbf{G}}^T.$$

Finally, we present a general derivation of the marginalizing method, which is not restricted to a stochastic setting. The maximum a posteriori estimator, written explicitly as

$$\begin{aligned} \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \end{bmatrix} &= \left( \begin{bmatrix} \mathbf{K}_1^T \\ \mathbf{K}_2^T \end{bmatrix} \mathbf{C}_\delta^{-1} [\mathbf{K}_1, \mathbf{K}_2] + \begin{bmatrix} \mathbf{C}_{x1}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{x2}^{-1} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{K}_1^T \\ \mathbf{K}_2^T \end{bmatrix} \mathbf{C}_\delta^{-1} \mathbf{y}^\delta \\ &= \begin{bmatrix} \mathbf{K}_1^T \mathbf{C}_\delta^{-1} \mathbf{K}_1 + \mathbf{C}_{x1}^{-1} & \mathbf{K}_1^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 \\ \mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_1 & \mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 + \mathbf{C}_{x2}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{K}_1^T \\ \mathbf{K}_2^T \end{bmatrix} \mathbf{C}_\delta^{-1} \mathbf{y}^\delta, \end{aligned} \quad (4.86)$$

is equivalent to the Tikhonov solution under assumptions (4.82). Setting

$$\mathbf{A} = \mathbf{K}_1^T \mathbf{C}_\delta^{-1} \mathbf{K}_1 + \mathbf{C}_{x1}^{-1}, \quad \mathbf{B} = \mathbf{K}_1^T \mathbf{C}_\delta^{-1} \mathbf{K}_2, \quad \mathbf{C} = \mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 + \mathbf{C}_{x2}^{-1},$$

we compute the inverse matrix in (4.86) by using the following result (Tarantola, 2005): if  $\mathbf{A}$  and  $\mathbf{C}$  are symmetric matrices, then

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{\mathbf{A}} & \tilde{\mathbf{B}} \\ \tilde{\mathbf{B}}^T & \tilde{\mathbf{C}} \end{bmatrix},$$

with

$$\tilde{\mathbf{A}} = (\mathbf{A} - \mathbf{B} \mathbf{C}^{-1} \mathbf{B}^T)^{-1}, \quad \tilde{\mathbf{C}} = (\mathbf{C} - \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B})^{-1}, \quad \tilde{\mathbf{B}} = -\tilde{\mathbf{A}} \mathbf{B} \mathbf{C}^{-1} = -\mathbf{A}^{-1} \mathbf{B} \tilde{\mathbf{C}}.$$



The first component of the state vector is then given by

$$\hat{\mathbf{x}}_1 = \tilde{\mathbf{A}} \mathbf{K}_1^T \mathbf{C}_\delta^{-1} \mathbf{y}^\delta - \tilde{\mathbf{A}} \mathbf{B} \mathbf{C}^{-1} \mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{y}^\delta = \tilde{\mathbf{A}} (\mathbf{K}_1^T - \mathbf{B} \mathbf{C}^{-1} \mathbf{K}_2^T) \mathbf{C}_\delta^{-1} \mathbf{y}^\delta.$$

By straightforward calculation we obtain

$$\begin{aligned} & \tilde{\mathbf{A}} (\mathbf{K}_1^T - \mathbf{B} \mathbf{C}^{-1} \mathbf{K}_2^T) \mathbf{C}_\delta^{-1} \\ &= (\mathbf{A} - \mathbf{B} \mathbf{C}^{-1} \mathbf{B}^T)^{-1} (\mathbf{K}_1^T - \mathbf{B} \mathbf{C}^{-1} \mathbf{K}_2^T) \mathbf{C}_\delta^{-1} \\ &= \left\{ \mathbf{K}_1^T \mathbf{C}_\delta^{-\frac{1}{2}} \left[ \mathbf{I}_m - \mathbf{C}_\delta^{-\frac{1}{2}} \mathbf{K}_2 (\mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 + \mathbf{C}_{x2}^{-1})^{-1} \mathbf{K}_2^T \mathbf{C}_\delta^{-\frac{1}{2}} \right] \mathbf{C}_\delta^{-\frac{1}{2}} \mathbf{K}_1 \right. \\ & \quad \left. + \mathbf{C}_{x1}^{-1} \right\} \mathbf{K}_1^T \mathbf{C}_\delta^{-\frac{1}{2}} \left[ \mathbf{I}_m - \mathbf{C}_\delta^{-\frac{1}{2}} \mathbf{K}_2 (\mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 + \mathbf{C}_{x2}^{-1})^{-1} \mathbf{K}_2^T \mathbf{C}_\delta^{-\frac{1}{2}} \right] \mathbf{C}_\delta^{-\frac{1}{2}} \end{aligned}$$

and

$$\mathbf{I}_m - \mathbf{C}_\delta^{-\frac{1}{2}} \mathbf{K}_2 (\mathbf{K}_2^T \mathbf{C}_\delta^{-1} \mathbf{K}_2 + \mathbf{C}_{x2}^{-1})^{-1} \mathbf{K}_2^T \mathbf{C}_\delta^{-\frac{1}{2}} = \mathbf{C}_\delta^{\frac{1}{2}} (\mathbf{C}_\delta + \mathbf{K}_2 \mathbf{C}_{x2} \mathbf{K}_2^T)^{-1} \mathbf{C}_\delta^{\frac{1}{2}},$$

which then yields

$$\hat{\mathbf{x}}_1 = \left( \mathbf{K}_1^T \mathbf{C}_{\delta_y}^{-1} \mathbf{K}_1 + \mathbf{C}_{x1}^{-1} \right)^{-1} \mathbf{K}_1^T \mathbf{C}_{\delta_y}^{-1} \mathbf{y}^\delta,$$

with  $\mathbf{C}_{\delta_y}$  as in (4.85). This derivation clearly shows that the solution for the full state vector will give the same results for each of the partial state vectors as their individual solutions.

# 5

## Iterative regularization methods for linear problems

The iterative solution of linear systems of equations arising from the discretization of ill-posed problems is the method of choice when the dimension of the problem is so large that factorization of the matrix is either too time-consuming or too memory-demanding.

The ill-conditioning of the coefficient matrix for these linear systems is so extremely large that some sort of regularization is needed to guarantee that the computed solution is not dominated by errors in the data. In the framework of iterative methods, the regularizing effect is obtained by stopping the iteration prior to convergence to the solution of the linear system. This form of regularization is referred to as regularization by truncated iteration. The idea behind regularization by truncated iteration is that in the first few iteration steps, the iterated solution includes the components  $[(\mathbf{u}_i^T \mathbf{y}^\delta)/\sigma_i] \mathbf{v}_i$  corresponding to the largest singular values and approaches a regularized solution. As the iteration continues, the iterated solution is dominated by amplified noise components and converges to some undesirable solution (often the least squares solution). This phenomenon is referred to as semi-convergence. In this context, it is apparent that the iteration index plays the role of the regularization parameter, and a stopping rule plays the role of a parameter choice method.

In this chapter we first review some classical iterative methods and then focus on the conjugate gradient method and a related algorithm based on Lanczos bidiagonalization. The classical iterative methods to be discussed include the Landweber iteration and semi-iterative methods.

### 5.1 Landweber iteration

The Landweber iteration is based on the transformation of the normal equation

$$\mathbf{K}^T \mathbf{K} \mathbf{x} = \mathbf{K}^T \mathbf{y}^\delta$$

into an equivalent fixed point equation

$$\mathbf{x} = \mathbf{x} + \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}),$$

that is

$$\mathbf{x}_k^\delta = \mathbf{x}_{k-1}^\delta + \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{k-1}^\delta), \quad k = 1, 2, \dots \quad (5.1)$$

The slight inconvenience with the Landweber iteration is that it requires the norm of  $\mathbf{K}$  to be less than or equal to one, otherwise the method either diverges or converges too slowly. If this is not the case, we introduce a relaxation parameter  $\chi$ , chosen as  $0 < \chi \leq \|\mathbf{K}\|^{-1}$ , to obtain

$$\mathbf{x}_k^\delta = \mathbf{x}_{k-1}^\delta + \chi^2 \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{k-1}^\delta), \quad k = 1, 2, \dots$$

This construction has the same effect as multiplying the equation  $\mathbf{K} \mathbf{x} = \mathbf{y}^\delta$  by  $\chi$  and iterating with (5.1). In the present analysis we assume that the problem has been scaled appropriately, so that  $\|\mathbf{K}\| \leq 1$ , and drop the relaxation parameter  $\chi$ .

The initial guess  $\mathbf{x}_0^\delta = \mathbf{x}_a$  plays the same role as in Tikhonov regularization: it selects the particular solution which will be obtained in the case of ambiguity. The iterate  $\mathbf{x}_k^\delta$  can be expressed non-recursively through

$$\mathbf{x}_k^\delta = \mathbf{M}^k \mathbf{x}_0^\delta + \sum_{l=0}^{k-1} \mathbf{M}^l \mathbf{K}^T \mathbf{y}^\delta, \quad (5.2)$$

where

$$\mathbf{M} = \mathbf{I}_n - \mathbf{K}^T \mathbf{K}.$$

This result can be proven by induction. For  $k = 1$ , there holds

$$\mathbf{x}_1^\delta = \mathbf{x}_0^\delta + \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_0^\delta) = \mathbf{M} \mathbf{x}_0^\delta + \mathbf{K}^T \mathbf{y}^\delta,$$

while under assumption (5.2), we obtain

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_k^\delta) = \mathbf{M} \mathbf{x}_k^\delta + \mathbf{K}^T \mathbf{y}^\delta = \mathbf{M}^{k+1} \mathbf{x}_0^\delta + \sum_{l=0}^k \mathbf{M}^l \mathbf{K}^T \mathbf{y}^\delta.$$

To obtain more transparent results concerning the regularizing property of the Landweber iteration, we assume that  $\mathbf{x}_0^\delta = \mathbf{0}$ . Using the result

$$\mathbf{M}^l \mathbf{K}^T \mathbf{y}^\delta = \sum_{i=1}^n (1 - \sigma_i^2)^l \sigma_i (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i, \quad l \geq 0,$$

where  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of  $\mathbf{K}$ , we deduce that the iterate  $\mathbf{x}_k^\delta$  can be expressed as

$$\mathbf{x}_k^\delta = \sum_{i=1}^n \left[ 1 - (1 - \sigma_i^2)^k \right] \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i, \quad (5.3)$$

and the regularized solution for the exact data vector  $\mathbf{y}$  as

$$\mathbf{x}_k = \sum_{i=1}^n \left[ 1 - (1 - \sigma_i^2)^k \right] \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i.$$

Accounting for the expression of the exact solution  $\mathbf{x}^\dagger$ ,

$$\mathbf{x}^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i,$$

we find that the smoothing error norm is given by

$$\|\mathbf{e}_{sk}\|^2 = \|\mathbf{x}^\dagger - \mathbf{x}_k\|^2 = \sum_{i=1}^n (1 - \sigma_i^2)^{2k} \frac{1}{\sigma_i^2} (\mathbf{u}_i^T \mathbf{y})^2. \quad (5.4)$$

Since by assumption  $\|\mathbf{K}\| \leq 1$ , it follows that  $\sigma_i \leq 1$  for all  $i = 1, \dots, n$ , and therefore,  $\|\mathbf{e}_{sk}\| \rightarrow 0$  as  $k \rightarrow \infty$ . On the other hand, the noise error norm

$$\|\mathbf{e}_{nk}\|^2 = \|\mathbf{x}_k - \mathbf{x}_k^\delta\|^2 = \sum_{i=1}^n \left[1 - (1 - \sigma_i^2)^k\right]^2 \frac{1}{\sigma_i^2} (\mathbf{u}_i^T \boldsymbol{\delta})^2 \quad (5.5)$$

converges to

$$\|\mathbf{K}^\dagger \boldsymbol{\delta}\|^2 = \sum_{i=1}^n \frac{1}{\sigma_i^2} (\mathbf{u}_i^T \boldsymbol{\delta})^2$$

as  $k \rightarrow \infty$ . Since  $\mathbf{K}$  possesses small singular values, the noise error is extremely large in this limit. The noise error can be estimated by using the inequality

$$\sup_{0 \leq x \leq 1} \frac{1 - (1 - x^2)^k}{x} \leq \sqrt{k}, \quad k \geq 1,$$

and the result is

$$\|\mathbf{e}_{nk}\|^2 \leq k \Delta^2. \quad (5.6)$$

From (5.4) and (5.6), we see that the smoothing error converges slowly to 0, while the noise error is of the same order of at most  $\sqrt{k} \Delta$ . For small values of  $k$ , the noise error is negligible and the iterate  $\mathbf{x}_k^\delta$  seems to converge to the exact solution  $\mathbf{x}^\dagger$ . When  $\sqrt{k} \Delta$  reaches the order of magnitude of the smoothing error, the noise error is no longer covered in  $\mathbf{x}_k^\delta$  and the approximation changes to worse. This semi-convergent behavior requires a reliable stopping rule for detecting the transition from convergence to divergence.

The regularizing effect of the Landweber iteration is reflected by the filter factors of the computed solution. From (5.3), we infer that the  $k$ th iterate can be expressed as

$$\mathbf{x}_k^\delta = \sum_{i=1}^n f_k(\sigma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i,$$

with the filter factors being given by

$$f_k(\sigma_i^2) = 1 - (1 - \sigma_i^2)^k.$$

For  $\sigma_i \ll 1$ , we have  $f_k(\sigma_i^2) \approx k \sigma_i^2$ , while for  $\sigma_i \approx 1$ , there holds  $f_k(\sigma_i^2) \approx 1$ . Thus, for small values of  $k$ , the contributions of the small singular values to the solution are effectively filtered out, and when  $k$  increases, more components corresponding to small singular values are included in the solution. Therefore, an optimal value of  $k$  should reflect a trade-off between accuracy and stability.

## 5.2 Semi-iterative regularization methods

The major drawback of the Landweber iteration is its slow rate of convergence, this means, too many iterations are required to reduce the residual norm to the order of the noise level. More sophisticated methods have been developed on the basis of the so-called semi-iterative methods.

To introduce semi-iterative methods, we consider again the Landweber iteration and define the function  $g_k(\lambda)$  in terms of the filter function

$$f_k(\lambda) = 1 - (1 - \lambda)^k$$

by the relation

$$g_k(\lambda) = \frac{1}{\lambda} f_k(\lambda) = \frac{1}{\lambda} [1 - (1 - \lambda)^k]. \quad (5.7)$$

In terms of  $g_k$ , the Landweber iterate reads as

$$\mathbf{x}_k^\delta = g_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta, \quad (5.8)$$

where

$$g_k(\mathbf{K}^T \mathbf{K}) = \mathbf{V} \left[ \text{diag}(g_k(\sigma_i^2))_{n \times n} \right] \mathbf{V}^T.$$

Evidently,  $g_k(\lambda)$  is a polynomial of degree  $k - 1$ , which converges pointwise to  $1/\lambda$  on  $(0, 1]$  as  $k \rightarrow \infty$ . This property guarantees that in the noise-free case, the regularized solution converge to the exact solution, that is,  $\lim_{k \rightarrow \infty} \|\mathbf{x}_k - \mathbf{x}^\dagger\| = 0$ , where  $\mathbf{x}_k = g_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}$ .

Any sequence of polynomials  $\{g_k\}$ , with  $g_k$  having the degree  $k - 1$ , defines a semi-iterative method. The idea is that polynomials  $g_k$  different from the one given by (5.7) may converge faster to  $1/\lambda$ , and may thus lead to accelerated Landweber methods. In the case of semi-iterative methods, the polynomials  $g_k$  are called iteration polynomials, while the polynomials

$$r_k(\lambda) = 1 - \lambda g_k(\lambda)$$

are called residual polynomials. The residual polynomials are uniformly bounded on  $[0, 1]$  and converge pointwise to 0 on  $(0, 1]$  as  $k \rightarrow \infty$ . In addition, they are normalized in the sense that  $r_k(0) = 1$ .

If the residual polynomials form an orthogonal sequence with respect to some measure over  $\mathbb{R}_+$ , then they satisfy the three-term recurrence relation

$$r_k(\lambda) = r_{k-1}(\lambda) + \mu_k [r_{k-1}(\lambda) - r_{k-2}(\lambda)] - \omega_k \lambda r_{k-1}(\lambda), \quad k \geq 2. \quad (5.9)$$

By virtue of (5.9) and taking into account that

$$\mathbf{x}_k^\delta = \sum_{i=1}^n [1 - r_k(\sigma_i^2)] \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i$$

and that

$$\mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{k-1}^\delta) = \sum_{i=1}^n [\sigma_i^2 r_{k-1}(\sigma_i^2)] \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i,$$

we deduce that the iterates of the associated semi-iterative method satisfy the recurrence relation

$$\mathbf{x}_k^\delta = \mathbf{x}_{k-1}^\delta + \mu_k (\mathbf{x}_{k-1}^\delta - \mathbf{x}_{k-2}^\delta) + \omega_k \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{k-1}^\delta), \quad k \geq 2. \quad (5.10)$$

Note that because the  $k$ th iterate does not depend only on the  $(k-1)$ th iterate, the iterative approach (5.10) is termed semi-iterative. As in the case of the Landweber iteration,  $\mathbf{K}$  must be scaled so that  $\|\mathbf{K}\| \leq 1$ , and for this reason, systems of polynomials defined on the interval  $[0, 1]$  have to be considered.

The Chebyshev method of Stiefel uses the residual polynomials (Rieder, 2003)

$$r_k(\lambda) = \frac{U_k(1 - 2\lambda)}{k + 1},$$

where  $U_k$  are the Chebyshev polynomials of the second kind

$$U_k(\lambda) = \frac{\sin((k+1) \arccos \lambda)}{\sin(\arccos \lambda)}.$$

Due to the orthogonality of  $U_k$  in the interval  $[-1, 1]$  with respect to the weight function  $\sqrt{1 - \lambda^2}$ , it follows that the  $r_k$  are orthogonal in the interval  $[0, 1]$  with respect to the weight function  $\sqrt{\lambda/(1 - \lambda)}$ . The three-term recurrence relation reads as

$$\mathbf{x}_k^\delta = \frac{2k}{k+1} \mathbf{x}_{k-1}^\delta - \frac{k-1}{k+1} \mathbf{x}_{k-2}^\delta + \frac{4k}{k+1} \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{k-1}^\delta), \quad k \geq 2,$$

with

$$\mathbf{x}_1^\delta = \mathbf{x}_0^\delta + 2\mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_0^\delta).$$

In the Chebyshev method of Nemirovskii and Polyak (1984), the residual polynomials are given by

$$r_k(\lambda) = \frac{(-1)^k T_{2k+1}(\sqrt{\lambda})}{(2k+1) \sqrt{\lambda}},$$

where  $T_k$  are the Chebyshev polynomials of the first kind

$$T_k(\lambda) = \cos(k \arccos \lambda).$$

As before, the orthogonality of  $T_k$  in the interval  $[-1, 1]$  with respect to the weight function  $1/\sqrt{1 - \lambda^2}$  implies the orthogonality of the  $r_k$  in the interval  $[0, 1]$  with respect to the weight function  $\sqrt{\lambda/(1 - \lambda)}$ . The recursion of the Chebyshev method of Nemirovskii and Polyak takes the form

$$\mathbf{x}_k^\delta = 2 \frac{2k-1}{2k+1} \mathbf{x}_{k-1}^\delta - \frac{2k-3}{2k+1} \mathbf{x}_{k-2}^\delta + 4 \frac{2k-1}{2k+1} \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{k-1}^\delta), \quad k \geq 2,$$

with

$$\mathbf{x}_1^\delta = \frac{2}{3} \mathbf{x}_0^\delta + \frac{4}{3} \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_0^\delta).$$

The  $\nu$ -method of Brakhage (1987) uses the residual polynomials

$$r_{\nu k}(\lambda) = \frac{P_k^{(2\nu-\frac{1}{2}, -\frac{1}{2})}(1-2\lambda)}{P_k^{(2\nu-\frac{1}{2}, -\frac{1}{2})}(1)},$$

where  $P_k^{(\alpha, \beta)}$  are the Jacobi polynomials. The parameter  $\nu$  is fixed and is chosen as  $0 < \nu < 1$ . The orthogonality of the Jacobi polynomials in the interval  $[-1, 1]$  with respect to the weight function  $(1-\lambda)^\alpha (1+\lambda)^\beta$ , where  $\alpha > -1$  and  $\beta > -1$ , yields the orthogonality of the residual polynomials in the interval  $[0, 1]$  with respect to the weight function  $\lambda^{2\nu+1/2} (1-\lambda)^{-1/2}$ . The three-term recurrence relation of the Jacobi polynomials leads to the following recursion of the  $\nu$ -method

$$\mathbf{x}_k^\delta = \mathbf{x}_{k-1}^\delta + \mu_k (\mathbf{x}_{k-1}^\delta - \mathbf{x}_{k-2}^\delta) + \omega_k \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_{k-1}^\delta), \quad k \geq 2,$$

with

$$\mathbf{x}_1^\delta = \mathbf{x}_0^\delta + \omega_1 \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_0^\delta)$$

and

$$\begin{aligned} \mu_k &= \frac{(k-1)(2k-3)(2k+2\nu-1)}{(k+2\nu-1)(2k+4\nu-1)(2k+2\nu-3)}, \quad k \geq 2, \\ \omega_k &= 4 \frac{(2k+2\nu-1)(k+\nu-1)}{(k+2\nu-1)(2k+4\nu-1)}, \quad k \geq 1. \end{aligned}$$

### 5.3 Conjugate gradient method

Semi-iterative regularization methods are much more efficient than the classical Landweber iteration but require the scaling of  $\mathbf{K}$ . The conjugate gradient method due to Hestenes and Stiefel (1952) is scaling-free and is faster than any other semi-iterative method.

The conjugate gradient method is applied to the normal equation

$$\mathbf{K}^T \mathbf{K} \mathbf{x} = \mathbf{K}^T \mathbf{y}^\delta$$

of an ill-posed problem, in which case, the resulting algorithm is known as the conjugate gradient for normal equations (CGNR). In contrast to other iterative regularization methods, CGNR is not based on a fixed sequence of polynomials  $\{g_k\}$  and  $\{r_k\}$ ; these polynomials depend on the given right-hand side. This has the advantage of a greater flexibility of the method, but at the price of the iterates depending nonlinearly on the data,

$$\mathbf{x}_k^\delta = g_k(\mathbf{K}^T \mathbf{K}, \mathbf{y}^\delta) \mathbf{K}^T \mathbf{y}^\delta.$$

To formulate the CGNR method we first consider a preliminary definition. If  $\mathbf{A}$  is a real  $n \times n$  matrix and  $\mathbf{x}$  is an element of  $\mathbb{R}^n$ , then the  $k$ th Krylov subspace  $\mathcal{K}_k(\mathbf{x}, \mathbf{A})$  is defined as the linear space

$$\mathcal{K}_k(\mathbf{x}, \mathbf{A}) = \text{span} \{ \mathbf{x}, \mathbf{A}\mathbf{x}, \dots, \mathbf{A}^{k-1}\mathbf{x} \}.$$

Using (5.8) and taking into account that  $g_k$  is a polynomial of degree  $k - 1$ , we deduce that the  $k$ th iterate of any semi-iterative method belongs to the  $k$ th Krylov subspace

$$\mathcal{K}_k(\mathbf{K}^T \mathbf{y}^\delta, \mathbf{K}^T \mathbf{K}) = \text{span} \left\{ \mathbf{K}^T \mathbf{y}^\delta, (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta, \dots, (\mathbf{K}^T \mathbf{K})^{k-1} \mathbf{K}^T \mathbf{y}^\delta \right\}.$$

If  $\text{rank}(\mathbf{K}) = r$ , there holds

$$(\mathbf{K}^T \mathbf{K})^{k-1} \mathbf{K}^T \mathbf{y}^\delta = \sum_{i=1}^r \sigma_i^{2(k-1)+1} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i, \quad k \geq 1,$$

and we infer that

$$\mathcal{K}_k \subseteq \mathcal{N}(\mathbf{K})^\perp = \text{span} \{ \mathbf{v}_i \}_{i=\overline{1,r}}, \quad k \geq 1, \quad (5.11)$$

where, for notation simplification,  $\mathcal{K}_k$  stands for  $\mathcal{K}_k(\mathbf{K}^T \mathbf{y}^\delta, \mathbf{K}^T \mathbf{K})$ .

The  $k$ th iterate of the CGNR method is defined as the minimizer of the residual norm in the corresponding Krylov subspace; assuming a zero initial guess, i.e.,  $\mathbf{x}_0^\delta = \mathbf{0}$ , we have

$$\mathbf{x}_k^\delta = \arg \min_{\mathbf{x}_k \in \mathcal{K}_k} \|\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_k\|^2. \quad (5.12)$$

By virtue of (5.12) and the fact that the  $k$ th iterate of any semi-iterative belongs to  $\mathcal{K}_k$ , we may expect that CGNR requires the fewest iteration steps among all semi-iterative methods. Going further, we define the  $k$ th subspace

$$\mathcal{L}_k = \mathbf{K} \mathcal{K}_k = \{ \mathbf{y}_k / \mathbf{y}_k = \mathbf{K} \mathbf{x}_k, \mathbf{x}_k \in \mathcal{K}_k \}, \quad (5.13)$$

and in view of (5.12), we consider the minimizer

$$\mathbf{y}_k^\delta = \arg \min_{\mathbf{y}_k \in \mathcal{L}_k} \|\mathbf{y}^\delta - \mathbf{y}_k\|. \quad (5.14)$$

The element  $\mathbf{y}_k^\delta$  gives the best approximation of  $\mathbf{y}^\delta$  among all elements of  $\mathcal{L}_k$ , that is,

$$\mathbf{y}_k^\delta = P_k \mathbf{y}^\delta, \quad (5.15)$$

where  $P_k$  is the orthogonal projection operator onto the (linear) subspace  $\mathcal{L}_k$ . The uniqueness of the orthogonal projection implies that  $\mathbf{y}_k^\delta$  is uniquely determined and that

$$\mathbf{y}_k^\delta = \mathbf{K} \mathbf{x}_k^\delta. \quad (5.16)$$

If  $\{ \mathbf{u}_i \}_{i=\overline{1,k}}$  is an orthogonal basis of the (finite-dimensional) subspace  $\mathcal{L}_k$ , then  $\mathbf{y}_k^\delta$  can be expressed as

$$\mathbf{y}_k^\delta = \sum_{i=1}^k \frac{\mathbf{u}_i^T \mathbf{y}^\delta}{\|\mathbf{u}_i\|^2} \mathbf{u}_i. \quad (5.17)$$

Let us now define the vectors

$$\mathbf{s}_k = \mathbf{K}^T \mathbf{r}_k^\delta, \quad k \geq 0,$$

with  $\mathbf{r}_0^\delta = \mathbf{y}^\delta$ . As the residual vector at the  $k$ th iteration step,

$$\mathbf{r}_k^\delta = \mathbf{y}^\delta - \mathbf{y}_k^\delta = (\mathbf{I}_m - P_k) \mathbf{y}^\delta, \quad k \geq 1, \quad (5.18)$$



is orthogonal to  $\mathcal{L}_k$ , the identity

$$\mathbf{s}_k^T \mathbf{x}_k = (\mathbf{K}^T \mathbf{r}_k^\delta)^T \mathbf{x}_k = \mathbf{r}_k^{\delta T} \mathbf{y}_k = 0, \quad (5.19)$$

which holds true for all  $\mathbf{x}_k \in \mathcal{K}_k$  and  $\mathbf{y}_k = \mathbf{K}\mathbf{x}_k \in \mathcal{L}_k$ , yields

$$\mathbf{s}_k \perp \mathcal{K}_k, \quad k \geq 1. \quad (5.20)$$

The finite-dimensional subspaces  $\mathcal{K}_k$  and  $\mathcal{L}_k$  can be characterized by appropriate orthogonal bases. For the  $k$ th Krylov subspace we note the following result: the system  $\{\mathbf{s}_i\}_{i=\overline{0, k-1}}$  is an orthogonal basis of  $\mathcal{K}_k$ , that is,

$$\mathcal{K}_k = \text{span}\{\mathbf{s}_i\}_{i=\overline{0, k-1}}, \quad \mathbf{s}_i^T \mathbf{s}_j = \delta_{ij} \|\mathbf{s}_i\|^2, \quad i, j = 0, \dots, k-1. \quad (5.21)$$

This assertion can be proven by induction on  $k$  (Rieder, 2003). For  $k = 1$ , the result  $\mathcal{K}_1 = \text{span}\{\mathbf{s}_0\}$ , with  $\mathbf{s}_0 = \mathbf{K}^T \mathbf{y}^\delta$ , is evidently true. Now, let us assume that (5.21) holds for  $k$ , i.e.,  $\mathcal{K}_k = \text{span}\{\mathbf{s}_i\}_{i=\overline{0, k-1}}$ , and let  $\{\mathbf{u}_i\}_{i=\overline{1, k}}$  be an orthogonal basis of  $\mathcal{L}_k$ . As  $\mathcal{L}_k = \mathbf{K}\mathcal{K}_k$ ,  $\{\mathbf{u}_i\}_{i=\overline{1, k}}$  can be generated by orthogonalizing the set of vectors  $\{\mathbf{K}\mathbf{s}_i\}_{i=\overline{0, k-1}}$ . From (5.17), we have

$$\mathbf{y}_k^\delta = \sum_{i=1}^k \frac{\mathbf{u}_i^T \mathbf{y}^\delta}{\|\mathbf{u}_i\|^2} \mathbf{u}_i = \mathbf{y}_{k-1}^\delta + \alpha_k \mathbf{u}_k, \quad k \geq 1, \quad (5.22)$$

with  $\mathbf{y}_0^\delta = \mathbf{0}$ ,

$$\mathbf{y}_{k-1}^\delta = P_{k-1} \mathbf{y}^\delta = \sum_{i=1}^{k-1} \frac{\mathbf{u}_i^T \mathbf{y}^\delta}{\|\mathbf{u}_i\|^2} \mathbf{u}_i,$$

and

$$\alpha_k = \frac{\mathbf{u}_k^T \mathbf{y}^\delta}{\|\mathbf{u}_k\|^2}. \quad (5.23)$$

Then, by (5.18) and (5.22), we obtain

$$\mathbf{r}_k^\delta = \mathbf{y}^\delta - \mathbf{y}_k^\delta = (\mathbf{y}^\delta - \mathbf{y}_{k-1}^\delta) - \alpha_k \mathbf{u}_k = \mathbf{r}_{k-1}^\delta - \alpha_k \mathbf{u}_k, \quad k \geq 1, \quad (5.24)$$

and further,

$$\mathbf{s}_k = \mathbf{s}_{k-1} - \alpha_k \mathbf{K}^T \mathbf{u}_k, \quad k \geq 1. \quad (5.25)$$

For  $\mathbf{u}_k \in \mathcal{L}_k = \mathbf{K}\mathcal{K}_k$ , there exists  $\mathbf{v}_k \in \mathcal{K}_k$  such that  $\mathbf{u}_k = \mathbf{K}\mathbf{v}_k$ , and we deduce that

$$\mathbf{K}^T \mathbf{u}_k = \mathbf{K}^T \mathbf{K} \mathbf{v}_k \in \mathcal{K}_{k+1}. \quad (5.26)$$

Since by induction hypothesis  $\mathbf{s}_{k-1} \in \mathcal{K}_k \subset \mathcal{K}_{k+1}$ , (5.25) gives  $\mathbf{s}_k \in \mathcal{K}_{k+1}$ . This result together with the orthogonality relation (5.20) yields the (orthogonal) sum representation  $\mathcal{K}_{k+1} = \mathcal{K}_k \oplus \text{span}\{\mathbf{s}_k\}$ , and the proof is finished. As  $\dim(\mathcal{K}_k) = k$ ,  $\dim(\mathcal{N}(\mathbf{K})^\perp) = r$ , and  $\mathcal{K}_k \subseteq \mathcal{N}(\mathbf{K})^\perp$ , we find that for  $k = r$ ,  $\mathcal{K}_r = \mathcal{N}(\mathbf{K})^\perp$  and, in particular, that the CGNR iterate  $\mathbf{x}_r^\delta = \arg \min_{\mathbf{x} \in \mathcal{N}(\mathbf{K})^\perp} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2$  is the least squares minimal norm

solution of the equation  $\mathbf{K}\mathbf{x} = \mathbf{y}^\delta$ . Since  $\mathbf{x}_r^\delta$  solves the normal equation  $\mathbf{K}^T \mathbf{K}\mathbf{x} = \mathbf{K}^T \mathbf{y}^\delta$ , we obtain

$$\mathbf{s}_r = \mathbf{K}^T \mathbf{r}_r^\delta = \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_r^\delta) = \mathbf{0}.$$

Thus, by the CGNR method we construct a sequence of iterates which approaches the least squares minimal norm solution, and we have to stop at some iteration step  $k < r$  in order to obtain a reliable solution. The set of orthogonal vectors  $\{\mathbf{u}_k\}_{k \geq 1}$  is generated by applying the Gram–Schmidt orthogonalization procedure to the set of vectors  $\{\mathbf{K}\mathbf{s}_k\}_{k \geq 0}$ , that is,

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{K}\mathbf{s}_0, \\ \mathbf{u}_k &= \mathbf{K}\mathbf{s}_{k-1} - \sum_{i=1}^{k-1} \frac{\mathbf{u}_i^T \mathbf{K}\mathbf{s}_{k-1}}{\|\mathbf{u}_i\|^2} \mathbf{u}_i, \quad \mathbf{s}_{k-1} \neq \mathbf{0}, \quad k \geq 2. \end{aligned} \quad (5.27)$$

The special form of the finite-dimensional subspaces  $\mathcal{K}_k$  and  $\mathcal{L}_k$  allows us to derive a recurrence relation for the orthogonal vectors  $\mathbf{u}_k$ . Since, for  $k > 2$  and  $i = 1, \dots, k-2$ , we have  $\mathbf{s}_{k-1} \perp \mathcal{K}_{i+1} \subseteq \mathcal{K}_{k-1}$  and  $\mathbf{K}^T \mathbf{u}_i \in \mathcal{K}_{i+1}$  (cf. (5.26)), we infer that

$$\mathbf{u}_i^T \mathbf{K}\mathbf{s}_{k-1} = (\mathbf{K}^T \mathbf{u}_i)^T \mathbf{s}_{k-1} = 0.$$

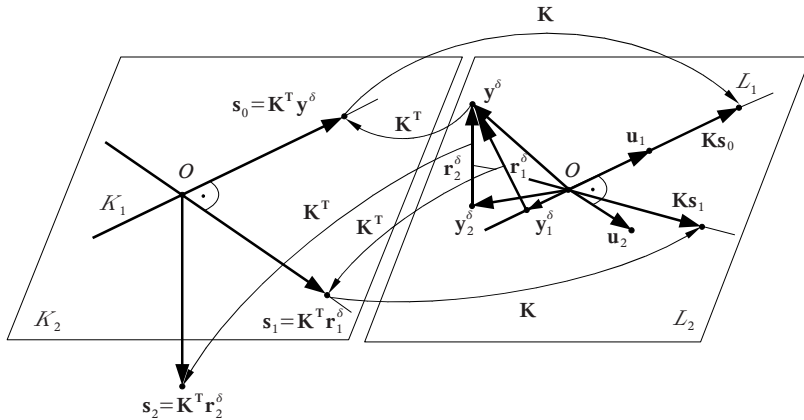
The basis vector  $\mathbf{u}_k$  defined by (5.27) can then be expressed as

$$\mathbf{u}_k = \mathbf{K}\mathbf{s}_{k-1} + \beta_{k-1} \mathbf{u}_{k-1}, \quad k \geq 1, \quad (5.28)$$

with

$$\beta_{k-1} = -\frac{\mathbf{u}_{k-1}^T \mathbf{K}\mathbf{s}_{k-1}}{\|\mathbf{u}_{k-1}\|^2} \quad (5.29)$$

and the convention  $\beta_0 = 0$ . The first orthogonal vectors  $\mathbf{s}_k$  and  $\mathbf{u}_k$  are illustrated in Figure 5.1.



**Fig. 5.1.** The first orthogonal vectors  $\mathbf{s}_k$  and  $\mathbf{u}_k$ . The construction is as follows: (1)  $\mathbf{r}_0^\delta = \mathbf{y}^\delta \rightarrow \mathbf{s}_0 = \mathbf{K}^T \mathbf{r}_0^\delta$ ,  $\mathcal{K}_1 = \text{span}\{\mathbf{s}_0\} \rightarrow \mathcal{L}_1 = \mathbf{K}\mathcal{K}_1$ ; (2)  $\mathbf{r}_1^\delta = \mathbf{y}^\delta - P_{\mathcal{L}_1} \mathbf{y}^\delta \rightarrow \mathbf{s}_1 = \mathbf{K}^T \mathbf{r}_1^\delta$ ,  $\mathcal{K}_2 = \text{span}\{\mathbf{s}_0, \mathbf{s}_1\} \rightarrow \mathcal{L}_2 = \mathbf{K}\mathcal{K}_2$ ; (3)  $\mathbf{r}_2^\delta = \mathbf{y}^\delta - P_{\mathcal{L}_2} \mathbf{y}^\delta \rightarrow \mathbf{s}_2 = \mathbf{K}^T \mathbf{r}_2^\delta$ , and so on.

The preimages  $\mathbf{v}_k \in \mathcal{K}_k$  of the orthogonal vectors  $\mathbf{u}_k \in \mathcal{L}_k$ , already defined by

$$\mathbf{u}_k = \mathbf{K}\mathbf{v}_k, \quad (5.30)$$

satisfy the recurrence relation (cf. (5.11), (5.28) and (5.30))

$$\mathbf{v}_k = \mathbf{s}_{k-1} + \beta_{k-1}\mathbf{v}_{k-1}, \quad k \geq 1. \quad (5.31)$$

Besides that, the residual vector  $\mathbf{r}_k^\delta$  can be computed recursively by using (5.24), while a recurrence relation for the iterates  $\mathbf{x}_k^\delta$  can be obtained from (5.22) in conjunction with (5.11), (5.16) and (5.30); the result is

$$\mathbf{x}_k^\delta = \mathbf{x}_{k-1}^\delta + \alpha_k \mathbf{v}_k, \quad k \geq 1. \quad (5.32)$$

The coefficients  $\alpha_k$  and  $\beta_k$ , defined by (5.23) and (5.29), respectively, can be computed efficiently as follows:

- (1) For  $k \geq 2$ , we have  $\mathbf{u}_k \perp \mathcal{L}_{k-1}$  and  $\mathbf{K}\mathbf{x}_{k-1}^\delta \in \mathcal{L}_{k-1}$ , and we find that  $\mathbf{u}_k^T \mathbf{K}\mathbf{x}_{k-1}^\delta = 0$  for  $k \geq 1$ . Then, by (5.16), (5.18), (5.30), (5.31), and the orthogonality relation  $\mathbf{s}_{k-1} \perp \mathbf{v}_{k-1} \in \mathcal{K}_{k-1}$ , (5.23) yields

$$\begin{aligned} \alpha_k \|\mathbf{u}_k\|^2 &= \mathbf{u}_k^T (\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta) \\ &= (\mathbf{K}\mathbf{v}_k)^T \mathbf{r}_{k-1}^\delta \\ &= \mathbf{v}_k^T \mathbf{s}_{k-1} \\ &= \|\mathbf{s}_{k-1}\|^2 + \beta_{k-1} \mathbf{v}_{k-1}^T \mathbf{s}_{k-1} \\ &= \|\mathbf{s}_{k-1}\|^2, \end{aligned}$$

and so,

$$\alpha_k = \frac{\|\mathbf{s}_{k-1}\|^2}{\|\mathbf{u}_k\|^2}, \quad k \geq 1.$$

- (2) By (5.24) and the orthogonality relation  $\mathbf{s}_k \perp \mathbf{s}_{k-1}$ , we have

$$-\alpha_k \mathbf{u}_k^T \mathbf{K}\mathbf{s}_k = (\mathbf{r}_k^\delta - \mathbf{r}_{k-1}^\delta)^T \mathbf{K}\mathbf{s}_k = (\mathbf{s}_k - \mathbf{s}_{k-1})^T \mathbf{s}_k = \|\mathbf{s}_k\|^2,$$

and (5.29) gives

$$\beta_k = \frac{\|\mathbf{s}_k\|^2}{\alpha_k \|\mathbf{u}_k\|^2} = \frac{\|\mathbf{s}_k\|^2}{\|\mathbf{s}_{k-1}\|^2}, \quad k \geq 1.$$

Collecting all results, we summarize the  $k$ th iteration step of the CGNR method as follows: given  $\mathbf{x}_{k-1}^\delta, \mathbf{r}_{k-1}^\delta, \mathbf{s}_{k-1} \neq \mathbf{0}$  and  $\mathbf{v}_k$ , compute

$$\begin{aligned} \mathbf{u}_k &= \mathbf{K}\mathbf{v}_k, \\ \alpha_k &= \|\mathbf{s}_{k-1}\|^2 / \|\mathbf{u}_k\|^2, \\ \mathbf{x}_k^\delta &= \mathbf{x}_{k-1}^\delta + \alpha_k \mathbf{v}_k, \\ \mathbf{r}_k^\delta &= \mathbf{r}_{k-1}^\delta - \alpha_k \mathbf{u}_k, \\ \mathbf{s}_k &= \mathbf{K}^T \mathbf{r}_k^\delta, \\ \beta_k &= \|\mathbf{s}_k\|^2 / \|\mathbf{s}_{k-1}\|^2, \\ \mathbf{v}_{k+1} &= \mathbf{s}_k + \beta_k \mathbf{v}_k. \end{aligned}$$

Even the best implementation of the CGNR method suffers from some loss of accuracy due to the implicit use of the cross-product matrix  $\mathbf{K}^T \mathbf{K}$ . An alternative iterative method which avoids  $\mathbf{K}^T \mathbf{K}$  completely is the LSQR algorithm of Paige and Saunders (1982). This method is based on the Lanczos bidiagonalization procedure of Golub and Kahan (1965) and is analytically equivalent to the CGNR method.

The Lanczos bidiagonalization algorithm is initialized with

$$\beta_1 \bar{\mathbf{u}}_1 = \mathbf{y}^\delta, \quad \alpha_1 \bar{\mathbf{v}}_1 = \mathbf{K}^T \bar{\mathbf{u}}_1, \quad (5.33)$$

and the iteration step  $k \geq 1$  has the form

$$\beta_{k+1} \bar{\mathbf{u}}_{k+1} = \mathbf{K} \bar{\mathbf{v}}_k - \alpha_k \bar{\mathbf{u}}_k, \quad (5.34)$$

$$\alpha_{k+1} \bar{\mathbf{v}}_{k+1} = \mathbf{K}^T \bar{\mathbf{u}}_{k+1} - \beta_{k+1} \bar{\mathbf{v}}_k. \quad (5.35)$$

The scalars  $\alpha_k > 0$  and  $\beta_k > 0$  are chosen such that

$$\|\bar{\mathbf{u}}_k\| = \|\bar{\mathbf{v}}_k\| = 1;$$

for example, the representation  $\alpha_1 \bar{\mathbf{v}}_1 = \mathbf{K}^T \bar{\mathbf{u}}_1$  assumes the calculations

$$\mathbf{v}_1 = \mathbf{K}^T \bar{\mathbf{u}}_1, \quad \alpha_1 = \|\mathbf{v}_1\|, \quad \bar{\mathbf{v}}_1 = (1/\alpha_1) \mathbf{v}_1.$$

Defining the dense matrices

$$\bar{\mathbf{U}}_{k+1} = [\bar{\mathbf{u}}_1, \dots, \bar{\mathbf{u}}_{k+1}] \in \mathbb{R}^{m \times (k+1)}, \quad \bar{\mathbf{V}}_k = [\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_k] \in \mathbb{R}^{n \times k},$$

and the bidiagonal matrix

$$\mathbf{B}_k = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ \beta_2 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_k \\ 0 & 0 & \dots & \beta_{k+1} \end{bmatrix} \in \mathbb{R}^{(k+1) \times k},$$

we rewrite the recurrence relations (5.33)–(5.35) as

$$\beta_1 \bar{\mathbf{U}}_{k+1} \mathbf{e}_1^{(k+1)} = \mathbf{y}^\delta, \quad (5.36)$$

$$\mathbf{K} \bar{\mathbf{V}}_k = \bar{\mathbf{U}}_{k+1} \mathbf{B}_k, \quad (5.37)$$

$$\mathbf{K}^T \bar{\mathbf{U}}_{k+1} = \bar{\mathbf{V}}_k \mathbf{B}_k^T + \alpha_{k+1} \bar{\mathbf{v}}_{k+1} \mathbf{e}_{k+1}^{(k+1)T}, \quad (5.38)$$

where  $\mathbf{e}_j^{(k+1)}$  is the  $j$ th canonical vector in  $\mathbb{R}^{k+1}$ ,

$$\left[ \mathbf{e}_j^{(k+1)} \right]_i = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

The columns  $\bar{\mathbf{u}}_1, \dots, \bar{\mathbf{u}}_{k+1}$  of  $\bar{\mathbf{U}}_{k+1}$  and  $\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_k$  of  $\bar{\mathbf{V}}_k$  are called the left and the right Lanczos vectors, respectively. In exact arithmetics,  $\bar{\mathbf{U}}_{k+1}$  and  $\bar{\mathbf{V}}_k$  are orthogonal matrices, and we have

$$\bar{\mathbf{U}}_{k+1}^T \bar{\mathbf{U}}_{k+1} = \mathbf{I}_{k+1}, \quad \bar{\mathbf{V}}_k^T \bar{\mathbf{V}}_k = \mathbf{I}_k.$$

As a result,  $\mathbf{B}_k^T \mathbf{B}_k$  can be expressed as

$$\mathbf{B}_k^T \mathbf{B}_k = \bar{\mathbf{V}}_k^T (\mathbf{K}^T \mathbf{K}) \bar{\mathbf{V}}_k,$$

and we infer that

$$(\mathbf{B}_k^T \mathbf{B}_k)^j = \bar{\mathbf{V}}_k^T (\mathbf{K}^T \mathbf{K})^j \bar{\mathbf{V}}_k, \quad j \geq 1.$$

Using the relations

$$\mathbf{K}^T \mathbf{y}^\delta = \alpha \bar{\mathbf{v}}_1 = \alpha \bar{\mathbf{V}}_k \mathbf{e}_1^{(k)}, \quad \alpha = \|\mathbf{K}^T \mathbf{y}^\delta\|,$$

and

$$(\mathbf{K}^T \mathbf{K})^j \mathbf{K}^T \mathbf{y}^\delta = \alpha (\mathbf{K}^T \mathbf{K})^j \bar{\mathbf{v}}_1 = \alpha (\mathbf{K}^T \mathbf{K})^j \bar{\mathbf{V}}_k \mathbf{e}_1^{(k)} = \alpha \bar{\mathbf{V}}_k (\mathbf{B}_k^T \mathbf{B}_k)^j \mathbf{e}_1^{(k)},$$

and setting

$$\mathbf{K}_k = \left[ \mathbf{K}^T \mathbf{y}^\delta, (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta, \dots, (\mathbf{K}^T \mathbf{K})^{k-1} \mathbf{K}^T \mathbf{y}^\delta \right] \in \mathbb{R}^{n \times k}$$

and

$$\mathbf{E}_k = \alpha \left[ \mathbf{e}_1^{(k)}, (\mathbf{B}_k^T \mathbf{B}_k) \mathbf{e}_1^{(k)}, \dots, (\mathbf{B}_k^T \mathbf{B}_k)^{k-1} \mathbf{e}_1^{(k)} \right] \in \mathbb{R}^{k \times k}$$

we find that

$$\mathbf{K}_k = \bar{\mathbf{V}}_k \mathbf{E}_k. \quad (5.39)$$

Thus, (5.39) resembles the QR factorization of the matrix  $\mathbf{K}_k$ , and as  $\mathcal{R}(\mathbf{K}_k) = \mathcal{K}_k$ , we deduce that  $\{\bar{\mathbf{v}}_i\}_{i=1,k}$  is an orthonormal basis of  $\mathcal{K}_k$ . Therefore, the LSQR method can be regarded as a method for constructing an orthonormal basis for the  $k$ th Krylov subspace  $\mathcal{K}_k$ . To solve the least squares problem

$$\min_{\mathbf{x}_k \in \text{span}\{\bar{\mathbf{v}}_i\}_{i=1,k}} \|\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_k\|^2,$$

we proceed as follows. First, we set

$$\mathbf{x}_k = \bar{\mathbf{V}}_k \mathbf{z}_k,$$

for some  $\mathbf{z}_k \in \mathbb{R}^k$ . Then, we express the ‘residual’

$$\mathbf{r}_k = \mathbf{y}^\delta - \mathbf{K} \mathbf{x}_k,$$

as (cf. (5.36) and (5.37))

$$\mathbf{r}_k = \bar{\mathbf{U}}_{k+1} \mathbf{t}_{k+1},$$

with

$$\mathbf{t}_{k+1} = \beta_1 \mathbf{e}_1^{(k+1)} - \mathbf{B}_k \mathbf{z}_k.$$

As we want  $\|\mathbf{r}_k\|^2$  to be small, and since  $\bar{\mathbf{U}}_{k+1}$  is theoretically orthogonal, we minimize  $\|\mathbf{t}_{k+1}\|^2$ . Hence, in the  $k$ th iteration step of the LSQR method we solve the least squares problem

$$\min_{\mathbf{z}_k \in \mathbb{R}^k} \left\| \beta_1 \mathbf{e}_1^{(k+1)} - \mathbf{B}_k \mathbf{z}_k \right\|^2. \quad (5.40)$$

If  $\mathbf{z}_k^\delta$  is the least squares solution of (5.40), then the vector

$$\mathbf{x}_k^\delta = \bar{\mathbf{V}}_k \mathbf{z}_k^\delta = \beta_1 \bar{\mathbf{V}}_k \mathbf{B}_k^\dagger \mathbf{e}_1^{(k+1)},$$

which belongs to the  $k$ th Krylov subspace  $\mathcal{K}_k = \text{span} \{\bar{\mathbf{v}}_i\}_{i=1,k}$ , is the iterate of the LSQR method. Computationally, the least squares problem (5.40) is solved by means of a QR factorization of  $\mathbf{B}_k$ , which is updated efficiently at each iteration step. The QR factorization then yields a simple recurrence relation for  $\mathbf{x}_k^\delta$  in terms of  $\mathbf{x}_{k-1}^\delta$ , and neither  $\bar{\mathbf{U}}_{k+1}$  nor  $\bar{\mathbf{V}}_k$  need to be stored.

For discrete problems that do not require regularization, LSQR is likely to obtain more accurate results in fewer iteration steps as compared to CGNR (Paige and Saunders, 1982). However, for discrete ill-posed problems, where the iteration is stopped before convergence, both iterative methods yield results with comparable accuracies (Hansen, 1998).

In practice, the convergence of CGNR and LSQR is delayed due to the influence of finite precision arithmetic. Specifically,  $\mathbf{x}_k^\delta$  stays almost unchanged for a few steps, then changes to a new vector and stays unchanged again for some steps, and so on. To prevent this delay and to simulate exact arithmetic, it is possible to incorporate some reorthogonalization techniques as for instance, the modified Gram–Schmidt algorithm or the Householder transformation. In LSQR we can orthogonalize the Lanczos vectors  $\bar{\mathbf{u}}_i$  and  $\bar{\mathbf{v}}_i$ , while in CGNR we can orthogonalize the residual vectors  $\mathbf{s}_i = \mathbf{K}^T \mathbf{r}_i^\delta$  (Hansen, 1998). The orthogonalization methods are illustrated in Algorithm 1.

For a deeper insight into the regularizing properties of the LSQR method, we consider the representation of the residual polynomial as given in Appendix F,

$$r_k(\lambda) = \prod_{j=1}^k \frac{\lambda_{k,j} - \lambda}{\lambda_{k,j}},$$

where

$$0 < \lambda_{k,k} < \lambda_{k,k-1} < \dots < \lambda_{k,1},$$

are the eigenvalues of the matrix  $\mathbf{B}_k^T \mathbf{B}_k$ . The eigenvalues  $\lambda_{k,j}$  are called Ritz values and for this reason,  $r_k$  is also known as the Ritz polynomial. The spectral filtering of the LSQR method is controlled by the convergence of the Ritz values to the eigenvalues of the matrix  $\mathbf{K}^T \mathbf{K}$  (Hansen, 1998). This, in turn, is related to the number  $k$  of iteration steps. If, after  $k$  steps, a large eigenvalue  $\sigma_i^2$  has been captured by the corresponding Ritz value  $\lambda_{k,i}$ , i.e.,  $\sigma_i^2 \approx \lambda_{k,i}$ , then the corresponding filter factor is  $f_k(\sigma_i^2) = 1 - r_k(\sigma_i^2) \approx 1$  (Appendix F). On the other hand, for an eigenvalue  $\sigma_i^2$  much smaller than the smallest Ritz value, i.e.,  $\sigma_i^2 \ll \lambda_{k,k}$ , the estimate

$$r_k(\sigma_i^2) = \prod_{j=1}^k \left(1 - \frac{\sigma_i^2}{\lambda_{k,j}}\right) \approx 1 - \sigma_i^2 \sum_{j=1}^k \frac{1}{\lambda_{k,j}},$$

yields

$$f_k(\sigma_i^2) \approx \sigma_i^2 \sum_{j=1}^k \frac{1}{\lambda_{k,j}},$$

---

**Algorithm 1.** Orthogonalization algorithms. (1) Modified Gram–Schmidt orthogonalization routine (MGSOrth): at the iteration step  $k$ , the new vector  $\mathbf{p}$  is added to the set of orthonormal vectors stored in the columns of  $\mathbf{P}$ . (2) Householder orthogonalization routine (HOrth): at the iteration step  $k$ , the candidate vector  $\mathbf{p}$  is transformed into a normalized vector  $\bar{\mathbf{p}}$  orthogonal to the previous vectors; the vectors  $\mathbf{v}_k$  and the scalars  $\beta_k$ , defining the reflection matrix  $\mathbf{P}_k = \mathbf{I}_n - \beta_k \mathbf{v}_k \mathbf{v}_k^T$ , are stored in the columns of the matrix  $\mathbf{P}$  and in the array  $\boldsymbol{\pi}$ , respectively.

---

subroutine MGSOrth ( $k, n, \mathbf{P}; \mathbf{p}$ )

**for**  $i = 1, k - 1$  **do**  
 $a \leftarrow \sum_{j=1}^n [\mathbf{p}]_j [\mathbf{P}]_{ji}; \{ \text{compute } \mathbf{p}^T [\mathbf{P}]_{\cdot i} \}$   
**for**  $j = 1, n$  **do**  $[\mathbf{p}]_j \leftarrow [\mathbf{p}]_j - a [\mathbf{P}]_{ji};$  **end for**  
**end for**

subroutine HOrth ( $k, n, \boldsymbol{\pi}, \mathbf{P}, \mathbf{p}; \bar{\mathbf{p}}, p_{\text{norm}}^{\text{sgn}}$ )

{transformation  $\mathbf{p} \leftarrow \mathbf{P}_{k-1} \mathbf{P}_{k-2} \dots \mathbf{P}_1 \mathbf{p}$ }

**for**  $i = 1, k - 1$  **do**  
 $a \leftarrow \sum_{j=i}^n [\mathbf{p}]_j [\mathbf{P}]_{ji}; \{ \text{compute } [\mathbf{p}]_{i:n}^T [\mathbf{P}]_{i:n,i} \}$   
**for**  $j = i, n$  **do**  $[\mathbf{p}]_j \leftarrow [\mathbf{p}]_j - a [\boldsymbol{\pi}]_i [\mathbf{P}]_{ji};$  **end for**  
**end for**

{Householder reflection matrix  $\mathbf{P}_k$ }

$p \leftarrow \sqrt{\sum_{j=k}^n [\mathbf{p}]_j^2}; [\boldsymbol{\pi}]_k \leftarrow 1 / (p^2 + |[\mathbf{p}]_k| p);$

$[\mathbf{P}]_{kk} \leftarrow [\mathbf{p}]_k + \text{sgn}([\mathbf{p}]_k) p; \text{ for } j = k + 1, n \text{ do } [\mathbf{P}]_{jk} \leftarrow [\mathbf{p}]_j; \text{ end for}$

$p_{\text{norm}}^{\text{sgn}} \leftarrow -\text{sgn}([\mathbf{p}]_k) p;$

{transformation  $\bar{\mathbf{p}} \leftarrow \mathbf{P}_1 \mathbf{P}_2 \dots \mathbf{P}_k \mathbf{e}_k$ , where  $\bar{\mathbf{p}}$  is normalized}

$\bar{\mathbf{p}} \leftarrow \mathbf{0}, [\bar{\mathbf{p}}]_k \leftarrow 1;$

**for**  $i = k, 1, -1$  **do**  
 $a \leftarrow \sum_{j=i}^n [\bar{\mathbf{p}}]_j [\mathbf{P}]_{ji}; \{ [\bar{\mathbf{p}}]_{i:n}^T [\mathbf{P}]_{i:n,i} \}$   
**for**  $j = i, n$  **do**  $[\bar{\mathbf{p}}]_j \leftarrow [\bar{\mathbf{p}}]_j - a [\boldsymbol{\pi}]_i [\mathbf{P}]_{ji};$  **end for**  
**end for**

---

and we see that these filter factors decay like  $\sigma_i^2$ . Thus, if the Ritz values approximate the eigenvalues in natural order, starting from the largest, then the iteration index plays the role of the regularization parameter, and the filter factors behave like the Tikhonov filter factors.

## 5.4 Stopping rules and preconditioning

Stopping the iteration prior to the inclusion of amplified noise components in the solution is an important aspect of iterative regularization methods. Also relevant is the preconditioning of the system of equations in order to improve the convergence rate. These topics are discussed below.

### 5.4.1 Stopping rules

The most widespread stopping rule for iterative regularization methods is the discrepancy principle. According to the discrepancy principle, the algorithm is terminated with  $k^*$  when

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k^*}^\delta\|^2 \leq \tau\Delta^2 < \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\|^2, \quad 0 \leq k < k^*. \quad (5.41)$$

In a semi-stochastic setting and for white noise with variance  $\sigma^2$ , the expected value of the noise  $\mathcal{E}\{\|\delta\|^2\} = m\sigma^2$  is used instead of the noise level  $\Delta^2$ .

Error-free parameter choice methods can also be formulated as stopping rules. In this case we have to store each iterate together with the corresponding objective function, e.g., the generalized cross-validation function, and to perform a sufficient number of iteration steps in order to detect the minimum of the objective function. For iterative regularization methods, the use of the generalized cross-validation and the maximum likelihood estimation requires the knowledge of the influence matrix, which, in turn, requires the knowledge of the generalized inverse. This is a difficult task because neither a canonical decomposition of  $\mathbf{K}$  nor the filter factors  $f_k$  are available (recall that iterative methods are preferred when a factorization of the matrix is infeasible).

More promising for iterative regularization methods is the use of the L-curve criterion. For the CGNR method, the monotonic behavior of both the solution norm  $\|\mathbf{x}_k^\delta\|$  and the residual norm  $\|\mathbf{r}_k^\delta\|$  recommends this approach. In the framework of Tikhonov regularization, the components of the L-curve are defined by some analytical formulas and the calculation of the curvature is straightforward. In the case of iterative methods, we are limited to knowing only a finite number of points on the L-curve (corresponding to different values of the iteration index). Unfortunately, these points are clustered giving fine-grained details that are not relevant for the determination of the corner. To eliminate this inconvenience, Hansen (1998) defined a differentiable smooth curve associated with the discrete points in such a way that fine-grained details are eliminated while the overall shape of the L-curve is maintained. The approximating curve is determined by fitting a cubic spline curve to the discrete points of the L-curve. Since a cubic spline curve does not have the desired local smoothing property, the following algorithm is employed:

- (1) perform a local smoothing of the L-curve, that is, for each interior point  $k = q + 1, \dots, P - q$ , where  $P$  is the number of discrete points of the L-curve and  $q$  is the half-width of the local smoothing interval, fit a polynomial of degree  $p$  to the points  $k - q, \dots, k + q$ , and store the corresponding  $k$ th ‘smoothed’ point situated on the fitting polynomial;
- (2) construct a cubic spline curve by using the smoothed points as control points;
- (3) compute the corner of the spline curve by maximizing its curvature;
- (4) select the point on the original discrete curve that is closest to the spline curve’s corner.

Another method which couples a geometrical approach to identify the corner of the L-curve with some heuristics rules has been proposed by Rodriguez and Theis (2005). The main steps of this approach can be summarized as follows:



- (1) compute the vectors  $\mathbf{a}_k = [x_{k+1} - x_k, y_{k+1} - y_k]^T$ ,  $k = 1, \dots, P - 1$ , where  $x_k = \log(\|\mathbf{r}_k^\delta\|^2)$  and  $y_k = \log(\|\mathbf{x}_k^\delta\|^2)$ ;
- (2) eliminate the clusters by deleting all the ‘short’ vectors;
- (3) normalize the remaining  $V$  vectors;
- (4) select the corner of the L-curve as that point which minimizes the scalar triple product between two successive vectors, i.e.,  $k^* = \arg \min_{k=1, \dots, V-1} w_k$ , where  $w_k = (\mathbf{a}_k \times \mathbf{a}_{k+1}) \cdot \mathbf{e}_3$ , and  $\mathbf{e}_3$  is the Cartesian unit vector codirectional with the  $z$ -axis.

### 5.4.2 Preconditioning

In general, the aim of preconditioning is to improve the convergence rate of iterative methods for solving large systems of equations. When preconditioning from the right, the linear system of equations

$$\mathbf{K}\mathbf{x} = \mathbf{y}^\delta, \quad (5.42)$$

is replaced by

$$\mathbf{K}\mathbf{M}\bar{\mathbf{x}} = \mathbf{y}^\delta, \quad \mathbf{M}\bar{\mathbf{x}} = \mathbf{x},$$

with  $\mathbf{M}$  being a nonsingular matrix. If (5.42) is solved by using an iterative method for normal equations,  $\mathbf{M}$  should be chosen such that the condition number of  $\mathbf{M}^T \mathbf{K}^T \mathbf{K} \mathbf{M}$  is smaller than that of  $\mathbf{K}^T \mathbf{K}$ . This spectral property then yields faster convergence for the iterative method.

For discrete ill-posed problems, the preconditioner should not be regarded as a convergence accelerator, but rather as an enhancer of solution quality, since convergence is never achieved. In fact, there is no point in improving the condition of  $\mathbf{K}$  because only a part of the singular values contributes to the regularized solution (Hansen, 1998).

By right preconditioning we control the solution with a different norm as in the case of Tikhonov regularization with a regularization matrix  $\mathbf{L}$ . Therefore, there is no practical restriction to use a regularization matrix  $\mathbf{L}$  in connection with iterative methods (Hanke and Hansen, 1993; Hansen, 1998). Regularization matrices, when used as right preconditioners, affect the solution of an iterative method in a similar way as they affect the solution of Tikhonov regularization. The system of equations preconditioned from the right by the nonsingular regularization matrix  $\mathbf{L}$  then takes the form

$$\mathbf{K}\mathbf{L}^{-1}\bar{\mathbf{x}} = \mathbf{y}^\delta, \quad \mathbf{L}^{-1}\bar{\mathbf{x}} = \mathbf{x}. \quad (5.43)$$

To obtain more insight into right preconditioning by regularization matrices, we recall that in the framework of Tikhonov regularization, we transformed a general-form problem (with  $\mathbf{L} \neq \mathbf{I}_n$ ) into a standard-form problem (with  $\mathbf{L} = \mathbf{I}_n$ ) by using the transformation  $\bar{\mathbf{K}} = \mathbf{K}\mathbf{L}^{-1}$  and the back-transformation  $\mathbf{x} = \mathbf{L}^{-1}\bar{\mathbf{x}}$ . In terms of the standard-form variables, equation (5.43) expressed as

$$\bar{\mathbf{K}}\bar{\mathbf{x}} = \mathbf{y}^\delta, \quad \mathbf{L}^{-1}\bar{\mathbf{x}} = \mathbf{x},$$

---

**Algorithm 2.**  $\nu$ -algorithm with preconditioning. The control parameters of the algorithm are the maximum number of iterations  $N_{\text{iter}}$ , the noise level  $\Delta$ , and the tolerance  $\tau$ . The notation  $\|\mathbf{A}\|_{\text{F}}$  stands for the Frobenius norm of the matrix  $\mathbf{A}$ .

---

```

 $\chi \leftarrow 1 / \|\mathbf{KL}^{-1}\|_{\text{F}}; \{\text{relaxation parameter}\}$ 
 $\mathbf{x}^\delta \leftarrow \mathbf{0}; \quad \mathbf{r}^\delta \leftarrow \chi (\mathbf{y}^\delta - \mathbf{Kx}^\delta);$ 
 $\{\text{step } k = 1\}$ 
 $\omega \leftarrow \frac{4\nu+2}{4\nu+1};$ 
 $\mathbf{q} \leftarrow \omega \mathbf{r}^\delta; \quad \mathbf{x}^\delta \leftarrow \mathbf{x}^\delta + \chi (\mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T \mathbf{q}; \quad \mathbf{r}^\delta \leftarrow \chi (\mathbf{y}^\delta - \mathbf{Kx}^\delta);$ 
if  $\|\mathbf{r}^\delta\|^2 \leq \tau \chi^2 \Delta^2$  stop;  $\{\text{residual smaller than the prescribed tolerance}\}$ 
 $\{\text{steps } k \geq 2\}$ 
for  $k = 2, N_{\text{iter}}$  do
     $\omega \leftarrow 4 \frac{(2k+2\nu-1)(k+\nu-1)}{(k+2\nu-1)(2k+4\nu-1)};$ 
     $\mu \leftarrow 0.25 \frac{(k-1)(2k-3)}{(k+\nu-1)(2k+2\nu-3)} \omega;$ 
     $\mathbf{q} \leftarrow \mu \mathbf{q} + \omega \mathbf{r}^\delta; \quad \mathbf{x}^\delta \leftarrow \mathbf{x}^\delta + \chi (\mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}^T \mathbf{q}; \quad \mathbf{r}^\delta \leftarrow \chi (\mathbf{y}^\delta - \mathbf{Kx}^\delta);$ 
    if  $\|\mathbf{r}^\delta\|^2 \leq \tau \chi^2 \Delta^2$  exit;  $\{\text{residual smaller than the prescribed tolerance}\}$ 
end for

```

---

reveals that solving the right preconditioned system of equations is equivalent to solving the standard-form problem without preconditioning. In practice, the multiplication with  $\mathbf{L}^{-1}$  is built into the iterative schemes, and the back-transformation is avoided. The  $\nu$ -method, as well as the CGNR and the LSQR methods with preconditioning and using the discrepancy principle as stopping rule are outlined in Algorithms 2–4.

---

**Algorithm 3.** CGNR algorithm with preconditioning and reorthogonalization. The control parameters of the algorithm are the maximum number of iterations  $N_{\text{iter}}$ , the noise level  $\Delta$ , the tolerance  $\tau$ , and the logical variables *TypeOrth*. The values of *TypeOrth* are as follows: 0 if no reorthogonalization is applied, 1 for Householder orthogonalization, and 2 for the modified Gram–Schmidt orthogonalization.

---

```

 $\mathbf{x}^\delta \leftarrow \mathbf{0}$ ;
 $\mathbf{r}^\delta \leftarrow \mathbf{y}^\delta - \mathbf{K}\mathbf{x}^\delta$ ;
if TypeOrth  $\neq 0$   $\mathbf{S} \leftarrow \mathbf{0}$ ;
 $\mathbf{q} \leftarrow \mathbf{K}^T \mathbf{r}^\delta$ ;
 $\mathbf{s} \leftarrow \mathbf{L}^{-T} \mathbf{q}$ ;
{initialization of arrays  $\mathbf{S}$  and  $\boldsymbol{\sigma}$ }
if TypeOrth = 1 then
     $\boldsymbol{\sigma} \leftarrow \mathbf{0}$ ;  $s \leftarrow \|\mathbf{s}\|$ ;  $[\boldsymbol{\sigma}]_1 \leftarrow 1 / (s^2 + |[\mathbf{s}]_1| s)$ ;
     $[\mathbf{S}]_{11} \leftarrow [\mathbf{s}]_1 + \text{sgn}([\mathbf{s}]_1) s$ ;
    for  $i = 2, n$  do  $[\mathbf{S}]_{i1} \leftarrow [\mathbf{s}]_i$ ; end for
     $s_{\text{nrm}} \leftarrow -\text{sgn}([\mathbf{s}]_1) s$ ;
{initialization of array  $\mathbf{S}$ }
else if TypeOrth = 2 then
     $s_{\text{nrm}} \leftarrow \|\mathbf{s}\|$ ;
    for  $i = 1, n$  do  $[\mathbf{S}]_{i1} \leftarrow [\mathbf{s}]_i / s_{\text{nrm}}$ ; end for
else
     $s_{\text{nrm}} \leftarrow \|\mathbf{s}\|$ ;
end if
 $\mathbf{v} \leftarrow \mathbf{L}^{-1} \mathbf{s}$ ;
for  $k = 2, N_{\text{iter}}$  do
     $\mathbf{u} \leftarrow \mathbf{K}\mathbf{v}$ ;
     $\alpha \leftarrow s_{\text{nrm}}^2 / \|\mathbf{u}\|^2$ ;
     $\mathbf{x}^\delta \leftarrow \mathbf{x}^\delta + \alpha \mathbf{v}$ ;
     $\mathbf{r}^\delta \leftarrow \mathbf{r}^\delta - \alpha \mathbf{u}$ ;
    if  $\|\mathbf{r}^\delta\|^2 \leq \tau \Delta^2$  exit; {residual smaller than the prescribed tolerance}
     $\mathbf{q} \leftarrow \mathbf{K}^T \mathbf{r}^\delta$ ;
     $\mathbf{s} \leftarrow \mathbf{L}^{-T} \mathbf{q}$ ;
    if TypeOrth = 1 then
        call HOrth( $k, n, \boldsymbol{\sigma}, \mathbf{S}, \mathbf{s}; \bar{\mathbf{s}}, s_{\text{nrm1}}$ );  $\mathbf{s} \leftarrow s_{\text{nrm1}} \bar{\mathbf{s}}$ ;
    else if TypeOrth = 2 then
        call MGSOrth( $k, n, \mathbf{S}; \mathbf{s}$ );  $s_{\text{nrm1}} \leftarrow \|\mathbf{s}\|$ ;
        for  $i = 1, n$  do  $[\mathbf{S}]_{ik} \leftarrow [\mathbf{s}]_i / s_{\text{nrm1}}$ ; end for
    else
         $s_{\text{nrm1}} \leftarrow \|\mathbf{s}\|$ ;
    end if
     $\beta \leftarrow s_{\text{nrm1}}^2 / s_{\text{nrm}}^2$ ;
     $s_{\text{nrm}} \leftarrow s_{\text{nrm1}}$ ;
     $\mathbf{v} \leftarrow \mathbf{L}^{-1} \mathbf{s} + \beta \mathbf{v}$ ;
end for

```

---

---

**Algorithm 4.** LSQR algorithm with preconditioning and reorthogonalization.
 

---

```

 $\mathbf{x}^\delta \leftarrow \mathbf{0}$ ; if  $TypeOrth \neq 0$  then  $\mathbf{P} \leftarrow \mathbf{0}$ ;  $\mathbf{Q} \leftarrow \mathbf{0}$ ; end if
if  $TypeOrth = 1$  then {initialization of arrays  $\mathbf{P}$  and  $\boldsymbol{\pi}$ }
   $\boldsymbol{\pi} \leftarrow \mathbf{0}$ ;  $p \leftarrow \|\mathbf{y}^\delta\|$ ;  $[\boldsymbol{\pi}]_1 \leftarrow 1/(p^2 + |[\mathbf{y}^\delta]_1|p)$ ;
   $[\mathbf{P}]_{11} \leftarrow [\mathbf{y}^\delta]_1 + \text{sgn}([\mathbf{y}^\delta]_1)p$ ; for  $i = 2, m$  do  $[\mathbf{P}]_{i1} \leftarrow [\mathbf{y}^\delta]_i$ ; end for
   $\beta \leftarrow -\text{sgn}([\mathbf{y}^\delta]_1)p$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta)\mathbf{y}^\delta$ ;
else if  $TypeOrth = 2$  then {initialization of array  $\mathbf{P}$ }
   $\beta \leftarrow \|\mathbf{y}^\delta\|$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta)\mathbf{y}^\delta$ ; for  $i = 1, m$  do  $[\mathbf{P}]_{i1} \leftarrow [\bar{\mathbf{u}}]_i$ ; end for
else
   $\beta \leftarrow \|\mathbf{y}^\delta\|$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta)\mathbf{y}^\delta$ ;
end if
 $\mathbf{q} \leftarrow \mathbf{L}^{-T}\mathbf{K}^T\bar{\mathbf{u}}$ ;
if  $TypeOrth = 1$  then {initialization of arrays  $\mathbf{Q}$  and  $\boldsymbol{\nu}$ }
   $\boldsymbol{\nu} \leftarrow \mathbf{0}$ ;  $q \leftarrow \|\mathbf{q}\|$ ;  $[\boldsymbol{\nu}]_1 \leftarrow 1/(q^2 + |[\mathbf{q}]_1|q)$ ;
   $[\mathbf{Q}]_{11} \leftarrow [\mathbf{q}]_1 + \text{sgn}([\mathbf{q}]_1)q$ ; for  $i = 2, n$  do  $[\mathbf{Q}]_{i1} \leftarrow [\mathbf{q}]_i$ ; end for
   $\alpha \leftarrow -\text{sgn}([\mathbf{q}]_1)q$ ;  $\bar{\mathbf{v}} \leftarrow (1/\alpha)\mathbf{q}$ ;
else if  $TypeOrth = 2$  then {initialization of array  $\mathbf{Q}$ }
   $\alpha \leftarrow \|\mathbf{q}\|$ ;  $\bar{\mathbf{v}} \leftarrow (1/\alpha)\mathbf{q}$ ; for  $i = 1, n$  do  $[\mathbf{Q}]_{i1} \leftarrow [\bar{\mathbf{v}}]_i$ ; end for
else
   $\alpha \leftarrow \|\mathbf{q}\|$ ;  $\bar{\mathbf{v}} \leftarrow (1/\alpha)\mathbf{q}$ ;
end if
 $\mathbf{w} \leftarrow \mathbf{v}$ ;  $\bar{\phi} \leftarrow \beta$ ;  $\bar{\rho} \leftarrow \alpha$ ;
for  $k = 2, N_{\text{iter}}$  do
   $\mathbf{p} \leftarrow \mathbf{K}\mathbf{L}^{-1}\bar{\mathbf{v}} - \alpha\bar{\mathbf{u}}$ ;
  if  $TypeOrth = 1$  then
    call HOrth( $k, m, \boldsymbol{\pi}, \mathbf{P}, \mathbf{p}$ ;  $\bar{\mathbf{u}}, \beta$ );
  else if  $TypeOrth = 2$  then
    call MGSOrth( $k, m, \mathbf{P}$ ;  $\mathbf{p}$ );  $\beta \leftarrow \|\mathbf{p}\|$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta)\mathbf{p}$ ;
  else
     $\beta \leftarrow \|\mathbf{p}\|$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta)\mathbf{p}$ ;
  end if
   $\mathbf{q} \leftarrow \mathbf{L}^{-T}\mathbf{K}^T\bar{\mathbf{u}} - \beta\bar{\mathbf{v}}$ ;
  if  $TypeOrth = 1$  then
    call HOrth( $k, n, \boldsymbol{\nu}, \mathbf{Q}, \mathbf{q}$ ;  $\bar{\mathbf{v}}, \alpha$ );
  else if  $TypeOrth = 2$  then
    call MGSOrth( $k, n, \mathbf{Q}$ ;  $\mathbf{q}$ );  $\alpha \leftarrow \|\mathbf{q}\|$ ;  $\bar{\mathbf{v}} \leftarrow (1/\alpha)\mathbf{q}$ ;
  else
     $\alpha \leftarrow \|\mathbf{q}\|$ ;  $\bar{\mathbf{v}} \leftarrow (1/\alpha)\mathbf{q}$ ;
  end if
  if  $TypeOrth = 2$  store  $\bar{\mathbf{u}}$  in column  $k$  of  $\mathbf{P}$  and  $\bar{\mathbf{v}}$  in column  $k$  of  $\mathbf{Q}$ ;
   $\rho \leftarrow \sqrt{\bar{\rho}^2 + \beta^2}$ ;  $c \leftarrow \bar{\rho}/\rho$ ;  $s \leftarrow \beta/\rho$ ;  $\theta \leftarrow s\alpha$ ;  $\bar{\rho} \leftarrow -c/\alpha$ ;
   $\phi \leftarrow c\bar{\phi}$ ;  $\bar{\phi} \leftarrow s\bar{\phi}$ ;  $\|\mathbf{r}^\delta\| \leftarrow \bar{\phi}$ ;  $\mathbf{x}^\delta \leftarrow \mathbf{x}^\delta + (\phi/\rho)\mathbf{w}$ ;  $\mathbf{w} \leftarrow \bar{\mathbf{v}} - (\theta/\rho)\mathbf{w}$ ;
  if  $\|\mathbf{r}^\delta\|^2 \leq \tau\Delta^2$  exit; {residual smaller than the prescribed tolerance}
end for
 $\mathbf{x}^\delta \leftarrow \mathbf{L}^{-1}\mathbf{x}^\delta$ ;

```

---

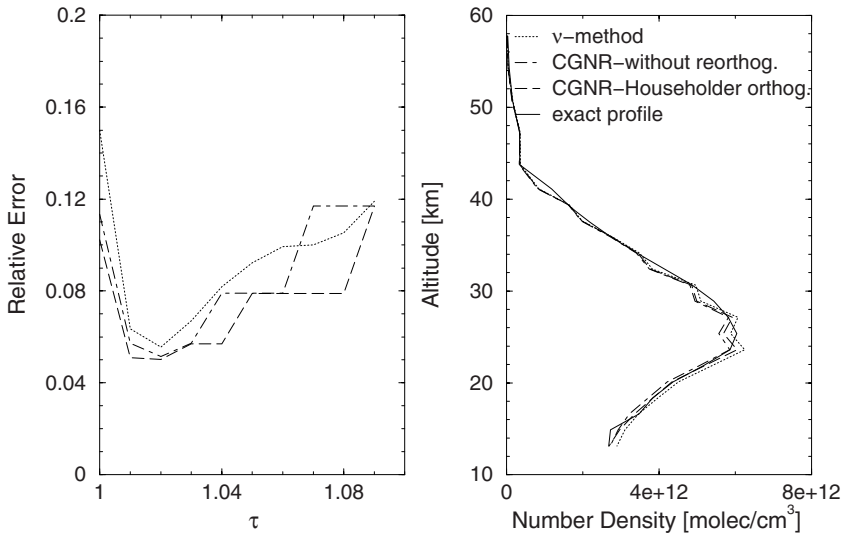
## 5.5 Numerical analysis

To analyze the performance of iterative regularization methods we consider the same retrieval scenario as in Chapter 3, but retrieve the  $O_3$  profile together with the  $NO_2$  profile in a spectral interval ranging from 520 to 580 nm. The atmosphere is discretized with a step of 1 km between 0 and 60 km, and a step of 5 km between 60 and 100 km. The number of unknowns of the inverse problem is  $n = 100$ . In our first simulation, we choose the discrepancy principle as stopping rule. As CGNR and LSQR yield identical results, only the CGNR results are reported here.

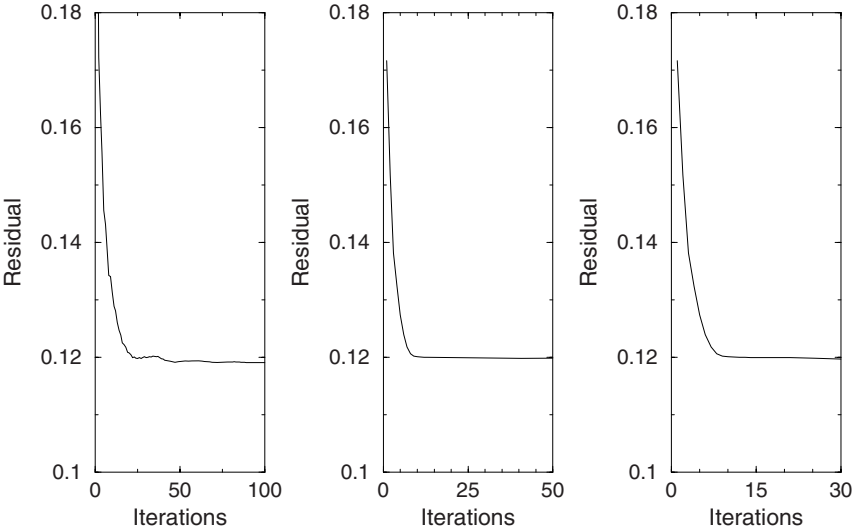
The solution errors for different values of the control parameter  $\tau$  (cf. (5.41)) are illustrated in the left panel of Figure 5.2. The error curves possess a minimum for an optimal value of the control parameter: the smallest errors are  $5.56 \cdot 10^{-2}$  for the  $\nu$ -method,  $5.20 \cdot 10^{-2}$  for CGNR without reorthogonalization and  $5.02 \cdot 10^{-2}$  for CGNR with Householder orthogonalization. Note that the stepwise behavior of the error curves for the CGNR method is a consequence of the discrete nature of the stopping rule. The retrieved profiles are shown in the right panel of Figure 5.2, and a sensible superiority of CGNR with Householder orthogonalization can be observed in the lower part of the atmosphere.

Although the methods are of comparable accuracies, the convergence rates are completely different (Figure 5.3). To reduce the residual norm to the order of the noise level, 100 iteration steps are required by the  $\nu$ -method, 50 by CGNR without reorthogonalization and 30 by CGNR with Householder orthogonalization.

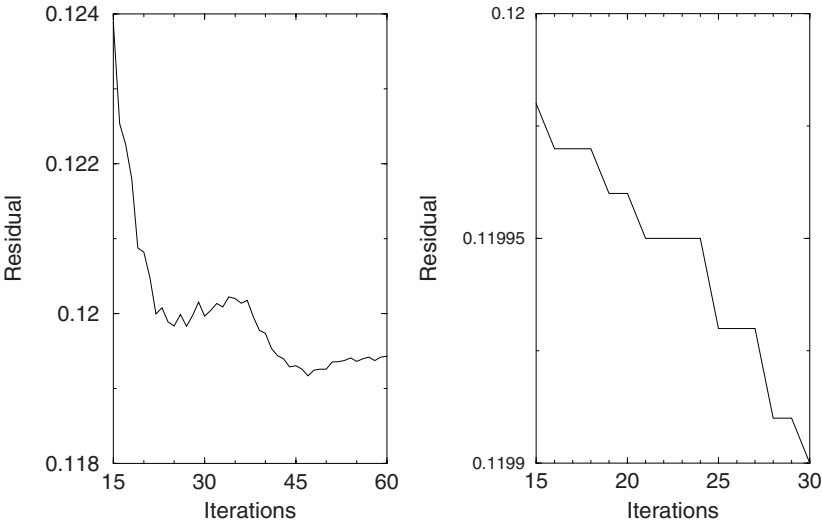
The non-monotonic behavior of the residual curve in the case of the  $\nu$ -method is apparent in the left panel of Figure 5.4, while the delay of CGNR without reorthogonalization



**Fig. 5.2.** Left: relative solution errors for different values of the control parameter  $\tau$ . Right: retrieved profiles corresponding to the optimal values of  $\tau$ . The results are computed with the  $\nu$ -method ( $\nu = 0.5$ ), CGNR without reorthogonalization, and CGNR with Householder orthogonalization.



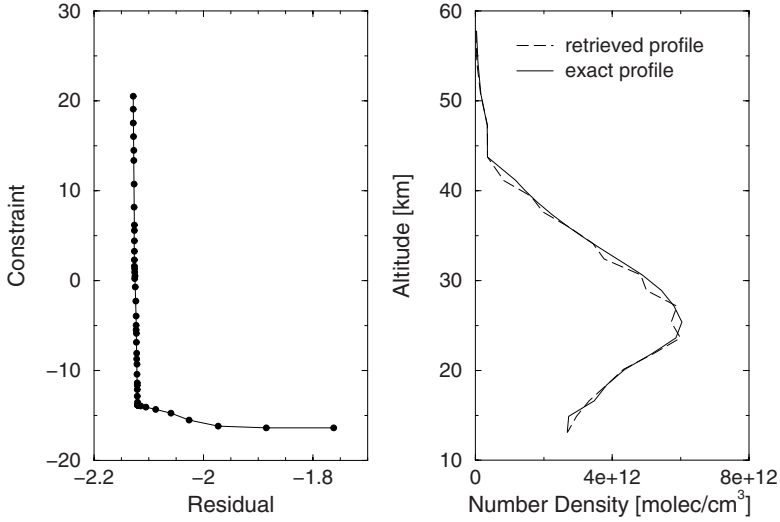
**Fig. 5.3.** Histories of the residual norm corresponding to the  $\nu$ -method (left), CGNR without reorthogonalization (middle), and CGNR with Householder orthogonalization (right).



**Fig. 5.4.** Left: non-monotonic behavior of the residual curve corresponding to the  $\nu$ -method. Right: delay of CGNR without reorthogonalization reflected in the residual curve.

(the iterate stays almost unchanged for a few steps) is evidenced in the right panel of Figure 5.4.

The discrete L-curve for the CGNR method illustrated in Figure 5.5 has a pronounced L-shape with a distinct corner. The inversion performance of CGNR with the L-curve method are slightly better than those of CGNR with the discrepancy principle; the retrieved profile in Figure 5.5 is characterized by a solution error of  $4.52 \cdot 10^{-2}$ .



**Fig. 5.5.** Discrete L-curve for CGNR with Householder orthogonalization (left) and the corresponding retrieved profile (right).

## 5.6 Mathematical results and further reading

A deterministic analysis of the Landweber iteration and of semi-iterative methods equipped with the discrepancy principle as stopping rule is presented in the first part of Appendix E. For the source condition  $\mathbf{x}^\dagger = (\mathbf{K}^T \mathbf{K})^\mu \mathbf{z}$ , with  $\mu > 0$  and  $\mathbf{z} \in \mathbb{R}^n$ , the Landweber iteration is order-optimal for all  $\mu > 0$ , while the  $\nu$ -method is order-optimal for  $0 < \mu \leq \nu - 1/2$ . Despite its optimal convergence rate, the Landweber iteration is rarely used in practice, since it usually requires far too many iteration steps until the stopping criterion (5.41) is met; the stopping index for the Landweber iteration is  $k^* = O(\Delta^{-2/(2\mu+1)})$ , and the exponent  $2/(2\mu+1)$  cannot be improved in general (Engl et al., 2000).

The convergence rate of the CGNR method using the discrepancy principle as stopping rule is derived in the second part of Appendix E. This method is order-optimal for  $\mu > 0$ , and so, no saturation effect occurs. In general, the number of iteration steps of the CGNR method is  $k^* = O(\Delta^{-1/(2\mu+1)})$ , and in particular, we have

$$k^* = O\left(\Delta^{-\frac{1}{(2\mu+1)(\beta+1)}}\right)$$

for the polynomial ill-posedness  $\sigma_i = O(i^{-\beta})$  with  $\beta > 0$ , and

$$k^* = O\left(\left|\log \Delta\right|^{\frac{1}{2\mu+1}}\right)$$

for the exponential ill-posedness  $\sigma_i = O(q^i)$  with  $q \in (0, 1)$ . In any case, the CGNR method requires significantly less iteration steps for the same order of accuracy than the Landweber iteration or the  $\nu$ -method. A detailed analysis of conjugate gradient type methods for ill-posed problems can be found in Hanke (1995), while for a pertinent treatment of preconditioned iterative regularization methods we refer to Hanke et al. (1993).

# 6

## Tikhonov regularization for nonlinear problems

Most of the inverse problems arising in atmospheric remote sensing are nonlinear. In this chapter we discuss the practical aspects of Tikhonov regularization for solving the nonlinear equation

$$\mathbf{F}(\mathbf{x}) = \mathbf{y}. \quad (6.1)$$

As in the linear case, equation (6.1) is the representation of a so-called discrete ill-posed problem because the underlying continuous problem is ill-posed. If we accept a characterization of ill-posedness via linearization, the condition number of the Jacobian matrix  $\mathbf{K}$  of  $\mathbf{F}$  may serve as a quantification of ill-posedness.

Nonlinear problems are treated in the same framework as linear problems. The right-hand side  $\mathbf{y}$  is supposed to be contaminated by instrumental noise, and we have the representation

$$\mathbf{y}^\delta = \mathbf{y} + \boldsymbol{\delta},$$

where  $\mathbf{y}^\delta$  is the noisy data vector and  $\boldsymbol{\delta}$  is the noise vector. In a deterministic setting, the data error is characterized by the noise level  $\Delta$ , while in a semi-stochastic setting,  $\boldsymbol{\delta}$  is assumed to be a discrete white noise with the covariance matrix  $\mathbf{C}_\delta = \sigma^2 \mathbf{I}_m$ .

The formulation of Tikhonov regularization for nonlinear problems is straightforward: the nonlinear equation (6.1) is replaced by a minimization problem involving the objective function

$$\mathcal{F}_\alpha(\mathbf{x}) = \frac{1}{2} \left[ \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2 + \alpha \|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2 \right]. \quad (6.2)$$

For a positive regularization parameter, minimizers of the Tikhonov function always exist, but are not unique, and a global minimizer  $\mathbf{x}_\alpha^\delta$  is called a regularized solution (Seidman and Vogel, 1989).

This chapter begins with a description of four retrieval test problems which, throughout the rest of the book, will serve to illustrate the various regularization algorithms and techniques. We then review appropriate optimization methods for minimizing the Tikhonov function, discuss practical algorithms for computing the iterates and characterize the error in the solution. Finally, we analyze the numerical performance of Tikhonov regularization with a priori, a posteriori and error-free parameter choice methods.



## 6.1 Four retrieval test problems

To investigate the efficiency of nonlinear regularization methods we consider the limb retrieval test problems illustrated in Table 6.1. The last problem is an exotic exercise, because temperature retrieval is usually performed in a thermal infrared  $\text{CO}_2$  or a  $\text{O}_2$  band. However, this problem will enable us to reveal some interesting features of the regularization methods under examination. The limb tangent height varies between 13.6 and 59.8 km in steps of 3.3 km. The atmosphere is discretized with a step of 1.75 km between 0 and 42 km, a step of 3.5 km between 42 and 70 km, and a step of 10 km between 70 and 100 km. The total number of levels is 36, and the spectral resolution is 0.25 nm.

**Table 6.1.** Four retrieval test problems. The auxiliary components with label 1 are included in the retrieval, while the auxiliary components with label 2 are not. The tangent altitude is expressed in km, while the spectral domain is expressed in nm for the first two retrieval problems, and in  $\text{cm}^{-1}$  for the last two retrieval problems.

Main component	Auxiliary component <sup>1</sup>	Auxiliary component <sup>2</sup>	Spectral domain	Tangent altitude	SNR
$\text{O}_3$	$\text{NO}_2$	–	520...580	13.6...49.9	300
BrO	$\text{O}_3$	–	337...357	13.6...43.3	$10^3$
CO	$\text{CH}_4$	$\text{H}_2\text{O}$	4280...4300	13.6...40.0	$10^3$
Temperature	–	$\text{CO}, \text{CH}_4, \text{H}_2\text{O}$	4280...4300	13.6...59.8	$10^4$

An efficient and flexible retrieval algorithm should include a preprocessing step comprising:

- (1) the selection of the forward model by estimating the degree of nonlinearity of the problem;
- (2) a sensitivity analysis revealing our expectations on the inversion process;
- (3) the derivation of a data model with white noise by using the prewhitening technique.

### 6.1.1 Forward models and degree of nonlinearity

The forward model for the retrieval problems in the infrared spectral domain is the radiance model

$$I_{\text{meas}}(\nu, h) \approx P_{\text{scl}}(\nu, \mathbf{p}_{\text{scl}}(h)) I_{\text{sim}}(\nu, \mathbf{x}, h) + P_{\text{off}}(\nu, \mathbf{p}_{\text{off}}(h)), \quad (6.3)$$

where  $\nu$  is the wavenumber,  $h$  is the tangent height, and  $P_{\text{scl}}$  and  $P_{\text{off}}$  are polynomials of low order with coefficients  $\mathbf{p}_{\text{scl}}$  and  $\mathbf{p}_{\text{off}}$ , respectively. The scale polynomial  $P_{\text{scl}}$  accounts on the multiplicative calibration error, while  $P_{\text{off}}$  is a polynomial baseline shift (zero-level calibration correction) accounting for the self-emission of the instrument, scattering of light into the instrument or higher-order nonlinearities of the detectors. The measured spectrum is the convolution of the radiance spectrum with the instrumental line shape, for the latter of which a Gaussian function is assumed in our simulations.

For the retrieval problems in the ultraviolet and visible spectral regions we consider two forward models. The first forward model is the radiance model,

$$R_{\text{meas}}(\lambda, h) \approx P_{\text{sc1}}(\lambda, \mathbf{p}_{\text{sc1}}(h)) R_{\text{sim}}(\lambda, \mathbf{x}, h), \quad (6.4)$$

where  $\lambda$  is the wavelength and  $R$  stands for the ‘scan-ratioed’ radiance ratio, that is, the radiance spectrum normalized with respect to a reference tangent height,

$$R(\cdot, h) = \frac{I(\cdot, h)}{I(\cdot, h_{\text{ref}})}. \quad (6.5)$$

The normalization procedure minimizes the influence of the solar Fraunhofer structure and avoids the need of absolute radiometric calibration of the instrument. In addition, there is a reduction in the effect of surface reflectance and clouds that can influence the diffuse radiation even at high altitudes. The normalization procedure does not completely remove the effect of the surface albedo, but does reduce the accuracy to which the algorithm must model this effect. The scale polynomial  $P_{\text{sc1}}$  is intended to account for the contribution of aerosols with smooth spectral signature. The second forward model is the differential radiance model

$$\log \bar{R}_{\text{meas}}(\lambda, h) \approx \log \bar{R}_{\text{sim}}(\lambda, \mathbf{x}, h), \quad (6.6)$$

with

$$\log \bar{R}_{\text{sim}}(\lambda, \mathbf{x}, h) = \log R_{\text{sim}}(\lambda, \mathbf{x}, h) - P_{\text{sim}}(\lambda, \mathbf{p}_{\text{sim}}(\mathbf{x}, h))$$

and

$$\log \bar{R}_{\text{meas}}(\lambda, h) = \log R_{\text{meas}}(\lambda, h) - P_{\text{meas}}(\lambda, \mathbf{p}_{\text{meas}}(h)).$$

For a state vector  $\mathbf{x}$  and a tangent height  $h$ , the coefficients of the smoothing polynomials  $P_{\text{sim}}$  and  $P_{\text{meas}}$  are computed as

$$\mathbf{p}_{\text{sim}}(\mathbf{x}, h) = \arg \min_{\mathbf{p}} \|\log R_{\text{sim}}(\cdot, \mathbf{x}, h) - P_{\text{sim}}(\cdot, \mathbf{p})\|^2,$$

and

$$\mathbf{p}_{\text{meas}}(h) = \arg \min_{\mathbf{p}} \|\log R_{\text{meas}}(\cdot, h) - P_{\text{meas}}(\cdot, \mathbf{p})\|^2,$$

respectively. In general, a smoothing polynomial is assumed to account for the low-order frequency structure due to scattering mechanisms, so that  $\log \bar{R}$  will mainly reflect the absorption process due to gas molecules (Platt and Stutz, 2008). For the sake of simplicity, the spectral corrections, also referenced as pseudo-absorbers, have been omitted in (6.4) and (6.6). The spectral corrections are auxiliary functions containing spectral features which are not attributed to the retrieved atmospheric species. They describe different kinds of instrumental effects, e.g., polarization correction spectra, undersampling spectrum (Slijkhuis et al., 1999), tilt spectrum (Sioris et al., 2003),  $I_0$ -correction (Aliwell et al., 2002), and more complex physical phenomena, e.g., Ring spectrum.

The choice of the forward model is crucial for the retrieval process, because it may substantially influence the nonlinearity of the problem to be solved.

The degree of nonlinearity can be estimated in a deterministic or a stochastic setting. In a deterministic framework, the degree of nonlinearity can be characterized by using curvature measures of nonlinearity from differential geometry (Bates and Watts, 1988). To present these concepts, we follow the analysis of Huiskes (2002). The  $m$ -dimensional

vector  $\mathbf{F}(\mathbf{x})$  defines an  $n$ -dimensional surface, the so-called measurement surface or expectation surface. To define the curvature measures, we consider the second-order Taylor expansion of the  $k$ th component of  $\mathbf{F}$  about  $\mathbf{x}_a$ ,

$$\begin{aligned} [\mathbf{F}(\mathbf{x}_a + \mathbf{p})]_k &= [\mathbf{F}(\mathbf{x}_a)]_k + \sum_{i=1}^n \frac{\partial [\mathbf{F}]_k}{\partial [\mathbf{x}]_i}(\mathbf{x}_a) [\mathbf{p}]_i \\ &+ \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2 [\mathbf{F}]_k}{\partial [\mathbf{x}]_i \partial [\mathbf{x}]_j}(\mathbf{x}_a) [\mathbf{p}]_i [\mathbf{p}]_j + O(\|\mathbf{p}\|^3). \end{aligned} \quad (6.7)$$

For notation simplification, we introduce the full derivative arrays  $\mathbf{K}$  and  $\mathbf{K}'$  by the relations

$$[\mathbf{K}(\mathbf{x}_a)]_{ki} = \frac{\partial [\mathbf{F}]_k}{\partial [\mathbf{x}]_i}(\mathbf{x}_a), \quad [\mathbf{K}'(\mathbf{x}_a)]_{kij} = \frac{\partial^2 [\mathbf{F}]_k}{\partial [\mathbf{x}]_i \partial [\mathbf{x}]_j}(\mathbf{x}_a),$$

where  $\mathbf{K} \in \mathbb{R}^{m \times n}$  is the Jacobian matrix of  $\mathbf{F}$  and  $\mathbf{K}' \in \mathbb{R}^{m \times n \times n}$  is a three-dimensional array. In general, for an array  $\mathbf{A}$  with three indices, left multiplication by a matrix  $\mathbf{B}$  means a multiplication by summation over the first index of the array,

$$[\mathbf{BA}]_{lij} = \sum_k [\mathbf{B}]_{lk} [\mathbf{A}]_{kij},$$

while right multiplication by two vectors  $\mathbf{c}$  and  $\mathbf{d}$  means a multiplication by summation over the vector indices,

$$[\mathbf{Acd}]_k = \sum_{ij} [\mathbf{A}]_{kij} [\mathbf{c}]_i [\mathbf{d}]_j.$$

If the three-dimensional array  $\mathbf{A}$  is symmetric with respect to the second and third index, i.e.,  $[\mathbf{A}]_{kij} = [\mathbf{A}]_{kji}$ , then right multiplication does not depend on the order of the vectors  $\mathbf{c}$  and  $\mathbf{d}$ ; we will write  $\mathbf{Ac}^2$  for  $\mathbf{Acc}$ . With these notations, the second-order Taylor expansion (6.7) can be expressed as

$$\mathbf{F}(\mathbf{x}_a + \mathbf{p}) = \mathbf{F}(\mathbf{x}_a) + \mathbf{K}(\mathbf{x}_a) \mathbf{p} + \frac{1}{2} \mathbf{K}'(\mathbf{x}_a) \mathbf{p}^2 + O(\|\mathbf{p}\|^3),$$

while the first-order Taylor expansion reads as

$$\mathbf{F}(\mathbf{x}_a + \mathbf{p}) = \mathbf{F}(\mathbf{x}_a) + \mathbf{K}(\mathbf{x}_a) \mathbf{p} + O(\|\mathbf{p}\|^2). \quad (6.8)$$

The range of  $\mathbf{K}$  is the tangent plane to the measurement surface at the point  $\mathbf{x}_a$ , and the linear approximation (6.8) amounts to approximating the measurement surface in a neighborhood of  $\mathbf{x}_a$  by this plane. The tangent plane is a good approximation to the measurement surface if the norm of the quadratic term  $\|\mathbf{K}'\mathbf{p}^2\|$  is negligible compared to the norm of the linear term  $\|\mathbf{K}\mathbf{p}\|$ . It is useful to decompose the quadratic term into two orthogonal components, the projection onto the tangent plane and the component normal to the tangent plane. If  $\mathbf{P} = \mathbf{K}(\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T$  is the projection matrix onto the tangent plane at  $\mathbf{x}_a$ , then the tangential and normal components of  $\mathbf{K}'$  can be expressed as  $\mathbf{K}'_t = \mathbf{P}\mathbf{K}'$  and

$\mathbf{K}'_n = (\mathbf{I}_m - \mathbf{P}) \mathbf{K}'$ , respectively. In view of the decomposition  $\mathbf{K}'\mathbf{p}^2 = \mathbf{K}'_t\mathbf{p}^2 + \mathbf{K}'_n\mathbf{p}^2$ , the nonlinearity measures defined by Bates and Watts (1988) are given by

$$\kappa_t = \frac{\|\mathbf{K}'_t\mathbf{p}^2\|}{\|\mathbf{K}\mathbf{p}\|^2}, \quad \kappa_n = \frac{\|\mathbf{K}'_n\mathbf{p}^2\|}{\|\mathbf{K}\mathbf{p}\|^2}.$$

The quantities  $\kappa_t$  and  $\kappa_n$  are known as the parameter-effects curvature and the intrinsic curvature, respectively. If the intrinsic curvature is high, the model is highly nonlinear and the linear tangent plane approximation is not appropriate.

The curvature measures can be expressed in terms of the so-called curvature arrays. To obtain the curvature arrays, we must transform the vector function  $\mathbf{F}$  into a vector function  $\tilde{\mathbf{F}}$  such that its tangent plane at  $\mathbf{x}_a$  aligns with the first  $n$  axes of a rotated coordinate system. The projection of the second-order derivative of  $\tilde{\mathbf{F}}$  on the tangent plane and its orthogonal complement will be the parameter-effects and the intrinsic curvature arrays, respectively. To derive the curvature arrays, we consider a QR factorization of the Jacobian matrix

$$\mathbf{K} = \mathbf{Q}\mathbf{R} = \begin{bmatrix} \mathbf{Q}_t & \mathbf{Q}_n \end{bmatrix} \begin{bmatrix} \mathbf{R}_t \\ \mathbf{0} \end{bmatrix},$$

where the column vectors of the  $m \times n$  matrix  $\mathbf{Q}_t$  are the basis vectors of the tangent plane ( $\mathcal{R}(\mathbf{K})$ ) and the column vectors of the  $m \times (m - n)$  matrix  $\mathbf{Q}_n$  are the basis vectors of the orthogonal complement of the tangent plane ( $\mathcal{R}(\mathbf{K})^\perp$ ). The  $n \times n$  matrix  $\mathbf{R}_t$  is nonsingular and upper triangular. The vector function  $\tilde{\mathbf{F}}$  is then defined by

$$\tilde{\mathbf{F}}(\tilde{\mathbf{x}}_a) = \mathbf{Q}^T \mathbf{F}(\mathbf{T}\tilde{\mathbf{x}}_a),$$

where  $\mathbf{T} = \mathbf{R}_t^{-1}$  and  $\mathbf{x}_a = \mathbf{T}\tilde{\mathbf{x}}_a$ . As  $\mathbf{Q}$  is an orthogonal matrix, multiplication by  $\mathbf{Q}^T$  can be interpreted as a rotation by which the basis vectors of the tangent plane are mapped into the first  $n$  unit vectors, and the basis vectors of the orthogonal complement of the tangent plane are mapped into the last  $m - n$  basis vectors of the transformed coordinate system. The Jacobian matrix of  $\tilde{\mathbf{F}}$  becomes

$$\tilde{\mathbf{K}}(\tilde{\mathbf{x}}_a) = \mathbf{Q}^T \mathbf{K}(\mathbf{x}_a) \mathbf{T} = \mathbf{Q}^T \mathbf{Q}_t = \begin{bmatrix} \mathbf{I}_n \\ \mathbf{0} \end{bmatrix},$$

and it is apparent that in the transformed coordinate system, projection on the tangent plane consists of taking the first  $n$  components and setting the remaining components to zero. For the second-order derivative, we have explicitly

$$[\tilde{\mathbf{K}}'(\tilde{\mathbf{x}}_a)]_{kij} = \frac{\partial^2 [\tilde{\mathbf{F}}]_k}{\partial [\tilde{\mathbf{x}}]_i \partial [\tilde{\mathbf{x}}]_j}(\tilde{\mathbf{x}}_a) = \sum_{k_1, i_1, j_1} [\mathbf{Q}]_{k_1 k} [\mathbf{K}'(\mathbf{x}_a)]_{k_1 i_1 j_1} [\mathbf{T}]_{i_1 i} [\mathbf{T}]_{j_1 j}.$$

The parameter-effects curvature array  $\mathbf{A}_t$  and the intrinsic curvature array  $\mathbf{A}_n$  are defined as the projection of  $\tilde{\mathbf{K}}'$  on the tangent plane and its orthogonal complement, respectively, that is,

$$\begin{aligned} [\mathbf{A}_t]_{kij} &= [\tilde{\mathbf{K}}'_t]_{kij} = \sum_{k_1, i_1, j_1} [\mathbf{Q}_t]_{k_1 k} [\mathbf{K}']_{k_1 i_1 j_1} [\mathbf{T}]_{i_1 i} [\mathbf{T}]_{j_1 j}, \\ [\mathbf{A}_n]_{kij} &= [\tilde{\mathbf{K}}'_n]_{kij} = \sum_{k_1, i_1, j_1} [\mathbf{Q}_n]_{k_1 k} [\mathbf{K}']_{k_1 i_1 j_1} [\mathbf{T}]_{i_1 i} [\mathbf{T}]_{j_1 j}. \end{aligned}$$

As the curvature measures do not depend on the length of the step vector, we choose  $\mathbf{p} = \mathbf{T}\mathbf{e}$ , with  $\|\mathbf{e}\| = 1$ . Then, using the result  $\mathbf{P} = \mathbf{Q}_t \mathbf{Q}_t^T$  and taking into account that vector norms are invariant under orthogonal transformations, i.e.,  $\|\mathbf{Q}_t \mathbf{x}\| = \|\mathbf{x}\|$ , we obtain

$$\kappa_t = \frac{\|\mathbf{K}'_t \mathbf{p}^2\|}{\|\mathbf{K} \mathbf{p}\|^2} = \frac{\|\mathbf{Q}_t \mathbf{Q}_t^T \mathbf{K}' (\mathbf{T}\mathbf{e})^2\|}{\|\mathbf{K} \mathbf{T}\mathbf{e}\|^2} = \frac{\|\mathbf{Q}_t \mathbf{A}_t \mathbf{e}^2\|}{\|\mathbf{Q}_t \mathbf{e}\|^2} = \|\mathbf{A}_t \mathbf{e}^2\|$$

and similarly,

$$\kappa_n = \|\mathbf{A}_n \mathbf{e}^2\|.$$

The computation of curvature arrays requires the calculation of the first- and second-order derivatives  $\mathbf{K}$  and  $\mathbf{K}'$ . The calculation of  $\mathbf{K}'$  can be performed by using finite differences schemes or automatic differentiation algorithms, but these processes are computationally very expensive (Huiskes, 2002). An efficient approach for computing the curvature arrays by using a symmetric storage scheme is given in Bates et al. (1983).

In a stochastic framework, the degree of nonlinearity can be examined by comparing the forward model with its linearization within the a priori variability (Rodgers, 2000). For this purpose, we assume that  $\mathbf{x}$  is a random vector characterized by a Gaussian a priori density with mean  $\mathbf{x}_a$  and covariance  $\mathbf{C}_x$ . In the  $\mathbf{x}$ -space, the ellipsoid

$$(\mathbf{x} - \mathbf{x}_a)^T \mathbf{C}_x^{-1} (\mathbf{x} - \mathbf{x}_a) = 1$$

represents the contour of the a priori covariance, outlining the region within which the state vector is supposed to lie. Considering the linear transformation

$$\mathbf{z} = \Sigma_x^{-\frac{1}{2}} \mathbf{V}_x^T (\mathbf{x} - \mathbf{x}_a),$$

for  $\mathbf{C}_x = \mathbf{V}_x \Sigma_x \mathbf{V}_x^T$ , we observe that in the  $\mathbf{z}$ -space, the contour of the a priori covariance is a sphere of radius 1 centered at the origin, that is,  $\mathbf{z}^T \mathbf{z} = 1$ . The points  $\mathbf{z}_k^\pm = [0, \dots, \pm 1, \dots, 0]^T$  are the intersection points of the sphere with the coordinate axes and delimit the region to which the state vector belongs. In the  $\mathbf{x}$ -space, these boundary points are given by

$$\mathbf{x}_k^\pm = \mathbf{x}_a + \mathbf{V}_x \Sigma_x^{\frac{1}{2}} \mathbf{z}_k^\pm = \mathbf{x}_a \pm \mathbf{c}_k,$$

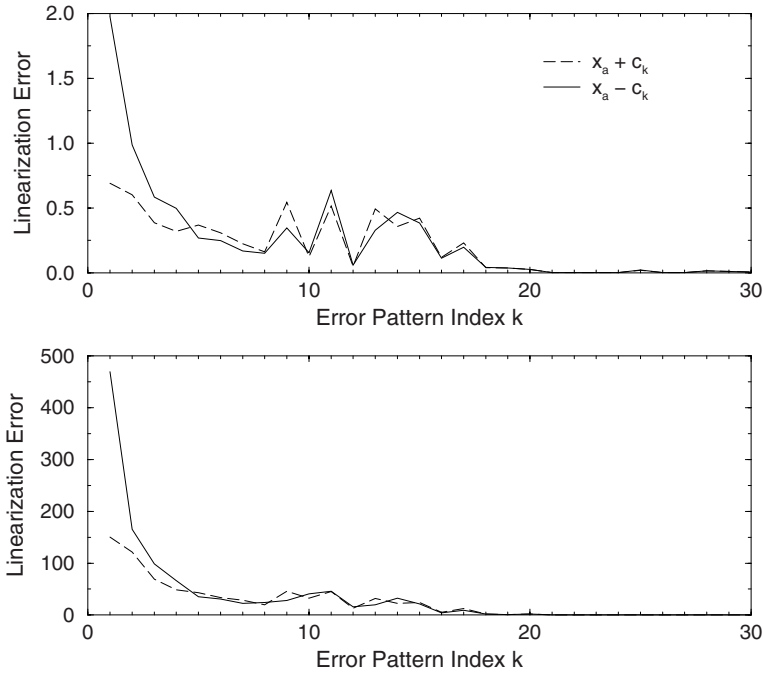
where the vectors  $\mathbf{c}_k$ , defined by the partition  $\mathbf{V}_x \Sigma_x^{1/2} = [\mathbf{c}_1, \dots, \mathbf{c}_n]$ , represent the error patterns for the covariance matrix  $\mathbf{C}_x$ . The size of the linearization error

$$\mathbf{R}(\mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_a) - \mathbf{K}(\mathbf{x}_a)(\mathbf{x} - \mathbf{x}_a),$$

can be evaluated through the quantity

$$\varepsilon_{\text{link}}^2 = \frac{1}{m\sigma^2} \|\mathbf{R}(\mathbf{x}_a \pm \mathbf{c}_k)\|^2.$$

If  $\varepsilon_{\text{link}} \leq 1$  for all  $k$ , then the problem is said to be linear to the accuracy of the measurements within the assumed range of variation of the state. The results plotted in Figure 6.1 show that the differential radiance model (6.6) is characterized by a smaller linearization error than the radiance model (6.4). For this reason, the differential radiance model is adopted in our simulations.



**Fig. 6.1.** Linearization errors for the  $O_3$  retrieval test problem corresponding to the differential radiance model (top) and the radiance model (bottom).

### 6.1.2 Sensitivity analysis

The sensitivity of the forward model with respect to components of the state vector is described by the Jacobian matrix. To be more precise, let us consider a linearization of the forward model about the a priori

$$\mathbf{F}(\mathbf{x}) \approx \mathbf{F}(\mathbf{x}_a) + \mathbf{K}(\mathbf{x}_a)(\mathbf{x} - \mathbf{x}_a).$$

For a change in the  $k$ th component of the state vector about the a priori,  $\Delta \mathbf{x}_k = \mathbf{x} - \mathbf{x}_a$ , with

$$[\Delta \mathbf{x}_k]_j = \begin{cases} \varepsilon [\mathbf{x}_a]_k, & j = k, \\ 0, & j \neq k, \end{cases}$$

the change in the forward model is given by

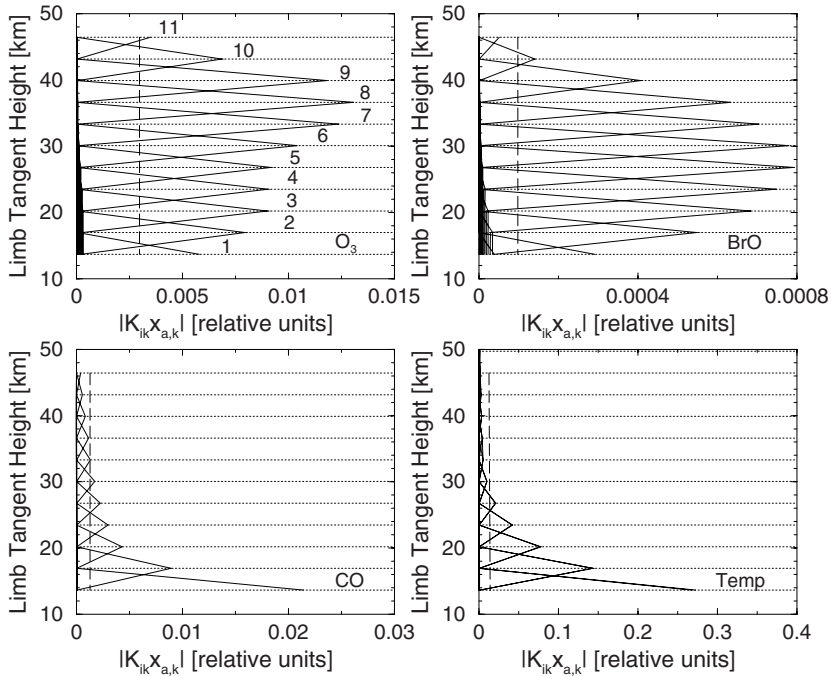
$$\Delta \mathbf{F}_k = \mathbf{F}(\mathbf{x}_a + \Delta \mathbf{x}_k) - \mathbf{F}(\mathbf{x}_a) = \mathbf{K}(\mathbf{x}_a) \Delta \mathbf{x}_k,$$

or componentwise, by

$$[\Delta \mathbf{F}_k]_i = \varepsilon [\mathbf{K}(\mathbf{x}_a)]_{ik} [\mathbf{x}_a]_k, \quad i = 1, \dots, m.$$

In this context, we say that the instrument is sensitive over the ‘entire’ spectral domain to a  $\pm\varepsilon$ -variation in the  $k$ th component of the state vector about the a priori, if  $|[\Delta \mathbf{F}_k]_i| > \sigma$  for all  $i = 1, \dots, m$ .

The complexity of the retrieval test problems in Table 6.1 is increased by assuming that the unknowns of the inverse problems are the layer values of the number density or the temperature. As a result, the limb radiances are mainly sensitive to those quantities which correspond to the layers traversed by the limb scans (Figure 6.2). This means that the retrieval of intermediate layer quantities is essentially based on information coming from the a priori and not from the measurement. In practice, this unfavorable situation can be overcome by considering the level quantities as unknowns of the inversion problem, or by choosing a rougher retrieval grid.



**Fig. 6.2.** Variations of the limb radiances at the first spectral point for a +20%-variation of the number density/temperature in the layers characterized by the central heights 13.125 (1), 16.625 (2), 20.125 (3), 23.625 (4), 27.125 (5), 30.625 (6), 34.125 (7), 35.875 (8), 39.375 (9), 43.750 (10) and 47.250 km (11). The variations of the limb radiances with respect to variations of the number density/temperature in the layers situated at 14.875, 18.375, 21.875, 25.375, 28.875, 32.375, 37.625 and 41.125 km are very small and cannot be distinguished. The horizontal dotted lines indicate the limb tangent heights (13.6, 16.9, 20.2, 23.5, 26.8, 30.1, 33.4, 36.7, 40.0, 43.3 and 46.6 km), while the vertical dashed lines delimit the noise domain.

Other relevant aspects of the sensitivity analysis can be inferred from Figure 6.2.

- (1) For the BrO retrieval test problem, the variations of the limb radiances with respect to the  $O_3$  concentrations are one order of magnitude higher than those corresponding to the BrO concentrations. This fact explains the large value of the signal-to-noise ratio considered in the simulation.

- (2) For the CO retrieval test problem, the variations of the limb radiances are larger than the noise level only for the layers situated between 13 and 30 km. However, above 30 km, the gas concentration is small and of no significant importance for the observed radiance signal.
- (3) For the temperature retrieval test problem, the low sensitivity of the forward model with respect to layer values of the temperature in the upper region of the atmosphere requires an extremely large signal-to-noise ratio. Anyway, large reconstruction errors are expected in the region above 30 km.

### 6.1.3 Prewhitening

When the instrumental noise covariance matrix has non-zero off-diagonal elements, we may use the prewhitening technique to transform noise into white noise. To explain this technique, we consider the data model

$$\mathbf{y}^\delta = \mathbf{F}(\mathbf{x}) + \boldsymbol{\delta}, \quad (6.9)$$

where the instrumental noise  $\boldsymbol{\delta}$  is supposed to have a zero mean vector and a positive definite covariance matrix  $\mathbf{C}_\delta = \mathcal{E}\{\boldsymbol{\delta}\boldsymbol{\delta}^T\}$ . The standard prewhitening approach involves the following steps:

- (1) compute the SVD of the covariance matrix  $\mathbf{C}_\delta$ ,

$$\mathbf{C}_\delta = \mathbf{U}_\delta \boldsymbol{\Sigma}_\delta \mathbf{U}_\delta^T;$$

- (2) define the ‘equivalent’ white noise variance

$$\sigma^2 = \frac{1}{m} \text{trace}(\mathbf{C}_\delta);$$

- (3) compute the preconditioner

$$\mathbf{P} = \sigma \boldsymbol{\Sigma}_\delta^{-\frac{1}{2}} \mathbf{U}_\delta^T.$$

Multiplying the data model (6.9) by  $\mathbf{P}$  we obtain

$$\bar{\mathbf{y}}^\delta = \bar{\mathbf{F}}(\mathbf{x}) + \bar{\boldsymbol{\delta}},$$

with  $\bar{\mathbf{y}}^\delta = \mathbf{P}\mathbf{y}^\delta$ ,  $\bar{\mathbf{F}}(\mathbf{x}) = \mathbf{P}\mathbf{F}(\mathbf{x})$  and  $\bar{\boldsymbol{\delta}} = \mathbf{P}\boldsymbol{\delta}$ . Then it is readily seen that  $\mathcal{E}\{\bar{\boldsymbol{\delta}}\} = \mathbf{0}$ , and that

$$\mathbf{C}_{\bar{\boldsymbol{\delta}}} = \mathcal{E}\{\bar{\boldsymbol{\delta}}\bar{\boldsymbol{\delta}}^T\} = \mathbf{P}\mathbf{C}_\delta\mathbf{P}^T = \sigma^2\mathbf{I}_m.$$

Note that the choice of the equivalent white noise variance is arbitrary and does not influence the retrieval or the error analysis; the representation used in step 2 is merely justified for the common situation of a diagonal noise covariance matrix  $\mathbf{C}_\delta = \boldsymbol{\Sigma}_\delta$ , when the preconditioner is also a diagonal matrix, i.e.,  $\mathbf{P} = \sigma\boldsymbol{\Sigma}_\delta^{-1/2}$ .

Multi-parameter regularization problems, treated in the framework of the marginalizing method, deal with a data error which includes the instrumental noise and the contribution due to the auxiliary parameters of the retrieval. If the auxiliary parameters are



encapsulated in the  $n_2$ -dimensional vector  $\mathbf{x}_2$  and the instrumental noise covariance matrix is the diagonal matrix  $\Sigma_\delta$ , the data error

$$\delta_y = \mathbf{K}_2 (\mathbf{x}_2 - \mathbf{x}_{a2}) + \delta \quad (6.10)$$

has the covariance

$$\mathbf{C}_{\delta_y} = \mathcal{E} \left\{ \delta_y \delta_y^T \right\} = \Sigma_\delta + \mathbf{K}_2 \mathbf{C}_{x_2} \mathbf{K}_2^T, \quad (6.11)$$

where  $\mathbf{C}_{x_2} \in \mathbb{R}^{n_2 \times n_2}$  is the a priori covariance matrix of  $\mathbf{x}_2$  and  $\mathbf{K}_2 \in \mathbb{R}^{m \times n_2}$  is the Jacobian matrix corresponding to  $\mathbf{x}_2$  evaluated at the a priori. Note that for nonlinear problems, the representations (6.10) and (6.11) tacitly assume that  $\mathbf{K}_2$  does not vary significantly during the iterative process. As for large-scale problems the computation of the SVD of the covariance matrix  $\mathbf{C}_{\delta_y} \in \mathbb{R}^{m \times m}$  by using the standard prewhitening technique is quite demanding, we propose the following algorithm:

- (1) perform the Cholesky factorization of the a priori covariance matrix  $\mathbf{C}_{x_2} = \mathbf{L}_2 \mathbf{L}_2^T$ ;
- (2) compute the SVD of the  $m \times n_2$  matrix  $\Sigma_\delta^{-1/2} \mathbf{K}_2 \mathbf{L}_2$ ,

$$\Sigma_\delta^{-\frac{1}{2}} \mathbf{K}_2 \mathbf{L}_2 = \mathbf{U}_2 \Sigma_2 \mathbf{V}_2^T;$$

- (3) define the equivalent white noise variance

$$\sigma^2 = \frac{1}{m} \text{trace} (\mathbf{C}_{\delta_y});$$

- (4) compute the preconditioner

$$\mathbf{P} = \sigma \Sigma^{-\frac{1}{2}} \mathbf{U}_2^T \Sigma_\delta^{-\frac{1}{2}},$$

with

$$\Sigma = \mathbf{I}_m + \Sigma_2 \Sigma_2^T.$$

To justify this approach, we use the result

$$\mathbf{C}_{\delta_y} = \Sigma_\delta + \mathbf{K}_2 \mathbf{C}_{x_2} \mathbf{K}_2^T = \Sigma_\delta^{\frac{1}{2}} \left( \mathbf{I}_m + \Sigma_\delta^{-\frac{1}{2}} \mathbf{K}_2 \mathbf{L}_2 \mathbf{L}_2^T \mathbf{K}_2^T \Sigma_\delta^{-\frac{1}{2}} \right) \Sigma_\delta^{\frac{1}{2}} = \Sigma_\delta^{\frac{1}{2}} \mathbf{U}_2 \Sigma \mathbf{U}_2^T \Sigma_\delta^{\frac{1}{2}},$$

and set  $\bar{\delta} = \mathbf{P} \delta_y$  to conclude that

$$\mathbf{C}_{\bar{\delta}} = \mathbf{P} \mathbf{C}_{\delta_y} \mathbf{P}^T = \sigma^2 \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}} = \sigma^2 \mathbf{I}_m.$$

The treatment of the auxiliary parameters as an extra source of error requires the multiplication of the preconditioner with the Jacobian matrix at each iteration step. For large-scale problems, this process is time-consuming and it is more preferable to include the auxiliary parameters in the retrieval or to account on them only when performing an error analysis (Eriksson et al., 2005).

For the rest of our analysis, prewhitening is implicitly assumed, and we will write  $\mathbf{F}$  for  $\mathbf{P}\mathbf{F}$  and  $\mathbf{y}^\delta$  for  $\mathbf{P}\mathbf{y}^\delta$ .

## 6.2 Optimization methods for the Tikhonov function

In the framework of Tikhonov regularization, the regularized solution  $\mathbf{x}_\alpha^\delta$  is a minimizer of the objective function

$$\mathcal{F}_\alpha(\mathbf{x}) = \frac{1}{2} \left[ \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2 + \alpha \|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2 \right], \quad (6.12)$$

where the factor  $1/2$  has been included in order to avoid the appearance of a factor two in the derivatives. The minimization of the Tikhonov function can be formulated as the least squares problem

$$\min_{\mathbf{x}} \mathcal{F}_\alpha(\mathbf{x}) = \frac{1}{2} \|\mathbf{f}_\alpha(\mathbf{x})\|^2, \quad (6.13)$$

where the augmented vector  $\mathbf{f}_\alpha$  is given by

$$\mathbf{f}_\alpha(\mathbf{x}) = \begin{bmatrix} \mathbf{F}(\mathbf{x}) - \mathbf{y}^\delta \\ \sqrt{\alpha} \mathbf{L}(\mathbf{x} - \mathbf{x}_a) \end{bmatrix}.$$

The regularized solution can be computed by using optimization methods for unconstrained minimization problems. Essentially, optimization tools are iterative methods, which use the Taylor expansion to compute approximations to the objective function at all points in the neighborhood of the current iterate. For Newton-type methods, the quadratic model

$$\mathcal{M}_\alpha(\mathbf{p}) = \mathcal{F}_\alpha(\mathbf{x}) + \mathbf{g}_\alpha(\mathbf{x})^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \mathbf{G}_\alpha(\mathbf{x}) \mathbf{p} \quad (6.14)$$

is used as a reasonable approximation to the objective function. In (6.14),  $\mathbf{g}_\alpha$  and  $\mathbf{G}_\alpha$  are the gradient and the Hessian of  $\mathcal{F}_\alpha$ , that is,

$$\mathbf{g}_\alpha(\mathbf{x}) = \nabla \mathcal{F}_\alpha(\mathbf{x}) = \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x})^T \mathbf{f}_\alpha(\mathbf{x}),$$

and

$$\mathbf{G}_\alpha(\mathbf{x}) = \nabla^2 \mathcal{F}_\alpha(\mathbf{x}) = \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x})^T \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}) + \mathbf{Q}_\alpha(\mathbf{x}),$$

respectively, where

$$\mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}) = \begin{bmatrix} \mathbf{K}(\mathbf{x}) \\ \sqrt{\alpha} \mathbf{L} \end{bmatrix}$$

is the Jacobian matrix of  $\mathbf{f}_\alpha(\mathbf{x})$ ,

$$\mathbf{Q}_\alpha(\mathbf{x}) = \sum_{i=1}^m [\mathbf{f}_\alpha(\mathbf{x})]_i \mathbf{G}_{\alpha i}(\mathbf{x}),$$

is the second-order derivative term and  $\mathbf{G}_{\alpha i}$  is the Hessian of  $[\mathbf{f}_\alpha]_i$ . Although the objective function (6.13) can be minimized by a general method, in most circumstances, the special forms of the gradient and the Hessian make it worthwhile to use methods designed specifically for least squares problems.

Nonlinear optimization methods can be categorized into two broad classes: step-length methods and trust-region methods. In this section we summarize the relevant features of an optimization method by following the analysis of Dennis and Schnabel (1996), and Gill et al. (1981).

### 6.2.1 Step-length methods

For an iterative method it is important to have a measure of progress in order to decide whether a new iterate  $\mathbf{x}_{\alpha k+1}^\delta$  is ‘better’ than the current iterate  $\mathbf{x}_{\alpha k}^\delta$ . A natural measure of progress is to require a decrease of the objective function at each iteration step, and to impose the descent condition

$$\mathcal{F}_\alpha(\mathbf{x}_{\alpha k+1}^\delta) < \mathcal{F}_\alpha(\mathbf{x}_{\alpha k}^\delta).$$

A method that imposes this condition is termed a descent method. A step-length procedure requires the computation of a vector  $\mathbf{p}_{\alpha k}^\delta$  called the search direction, and the calculation of a positive scalar  $\tau_k$ , the step length, for which it holds that

$$\mathcal{F}_\alpha(\mathbf{x}_{\alpha k}^\delta + \tau_k \mathbf{p}_{\alpha k}^\delta) < \mathcal{F}_\alpha(\mathbf{x}_{\alpha k}^\delta).$$

To guarantee that the objective function  $\mathcal{F}_\alpha$  can be reduced at the iteration step  $k$ , the search direction  $\mathbf{p}_{\alpha k}^\delta$  should be a descent direction at  $\mathbf{x}_{\alpha k}^\delta$ , that is, the inequality

$$\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{p}_{\alpha k}^\delta < 0$$

should hold true.

#### *Search direction*

In the steepest-descent method characterized by a linear convergence rate, the objective function is approximated by a linear model and the search direction is taken as

$$\mathbf{p}_{\alpha k}^\delta = -\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta).$$

The negative gradient  $-\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)$  is termed the direction of steepest descent, and evidently, the steepest-descent direction is indeed a descent direction (unless the gradient vanishes) since

$$\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{p}_{\alpha k}^\delta = -\|\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)\|^2 < 0.$$

In the Newton method, the objective function is approximated by the quadratic model (6.14) and the search direction  $\mathbf{p}_{\alpha k}^\delta$ , which minimizes the quadratic function, is the solution of the Newton equation

$$\mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) \mathbf{p} = -\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta). \quad (6.15)$$

For a general nonlinear function, Newton’s method converges quadratically to the minimizer  $\mathbf{x}_\alpha^\delta$  if the initial guess is sufficiently close to  $\mathbf{x}_\alpha^\delta$ , the Hessian matrix is positive definite at  $\mathbf{x}_\alpha^\delta$ , and the step lengths  $\{\tau_k\}$  converge to unity. Note that when  $\mathbf{G}_\alpha$  is always positive definite, the solution of (6.15) is a descent direction, since

$$\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{p}_{\alpha k}^\delta = -\mathbf{p}_{\alpha k}^{\delta T} \mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) \mathbf{p}_{\alpha k}^\delta < 0.$$

In the Gauss–Newton method for least squares problems, it is assumed that the first-order term  $\mathbf{K}_{f\alpha}^T \mathbf{K}_{f\alpha}$  in the expression of the Hessian dominates the second-order term  $\mathbf{Q}_\alpha$ . This assumption is not justified when the residuals at the solution are very large, i.e.,

roughly speaking, when the residual  $\|\mathbf{f}_\alpha(\mathbf{x}_\alpha^\delta)\|$  is comparable to the largest eigenvalue of  $\mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_\alpha^\delta)^T \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_\alpha^\delta)$ . For small residual problems, the search direction solves the equation

$$\mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_{\alpha k}^\delta) \mathbf{p} = -\mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{f}_\alpha(\mathbf{x}_{\alpha k}^\delta), \quad (6.16)$$

and possesses the variational characterization

$$\mathbf{p}_{\alpha k}^\delta = \arg \min_{\mathbf{p}} \|\mathbf{f}_\alpha(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_{\alpha k}^\delta) \mathbf{p}\|^2. \quad (6.17)$$

The vector solving (6.16) is called the Gauss–Newton direction, and if  $\mathbf{K}_{\mathbf{f}_\alpha}$  is of full column rank, then the Gauss–Newton direction is uniquely determined and approaches the Newton direction.

For large-residual problems, the term  $\|\mathbf{f}_\alpha(\mathbf{x}_\alpha^\delta)\|$  is not small, and the second-order term  $\mathbf{Q}_\alpha$  cannot be neglected. In fact, a large-residual problem is one in which the residual  $\|\mathbf{f}_\alpha(\mathbf{x}_\alpha^\delta)\|$  is large relative to the small eigenvalues of  $\mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_\alpha^\delta)^T \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_\alpha^\delta)$ , but not with respect to its largest eigenvalue. One possible strategy for large-residual problems is to include a quasi-Newton approximation  $\bar{\mathbf{Q}}_\alpha$  to the second-order derivative term  $\mathbf{Q}_\alpha$ , and to compute the search direction by solving the equation

$$\left[ \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_{\alpha k}^\delta) + \bar{\mathbf{Q}}_\alpha(\mathbf{x}_{\alpha k}^\delta) \right] \mathbf{p} = -\mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{f}_\alpha(\mathbf{x}_{\alpha k}^\delta). \quad (6.18)$$

Quasi-Newton methods are based on the idea of building up curvature information as the iteration proceeds using the observed behavior of the objective function and of the gradient. The initial approximation of the second-order derivative term is usually taken as zero, and with this choice, the first iteration step of the quasi-Newton method is equivalent to an iteration of the Gauss–Newton method. After  $\mathbf{x}_{\alpha k+1}^\delta$  has been computed, a new approximation of  $\bar{\mathbf{Q}}_\alpha(\mathbf{x}_{\alpha k+1}^\delta)$  is obtained by updating  $\bar{\mathbf{Q}}_\alpha(\mathbf{x}_{\alpha k}^\delta)$  to take into account the newly-acquired curvature information. An update formula reads as

$$\bar{\mathbf{Q}}_\alpha(\mathbf{x}_{\alpha k+1}^\delta) = \bar{\mathbf{Q}}_\alpha(\mathbf{x}_{\alpha k}^\delta) + \mathbf{U}_{\alpha k},$$

where the update matrix  $\mathbf{U}_{\alpha k}$  is usually chosen as a rank-one matrix. The standard condition for updating  $\bar{\mathbf{Q}}_\alpha$  is known as the quasi-Newton condition, and requires that the Hessian should approximate the curvature of the objective function along the change in  $\mathbf{x}$  during the current iteration step. The most widely used quasi-Newton scheme, which satisfies the quasi-Newton condition and possesses the property of hereditary symmetry, is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) update,

$$\begin{aligned} \bar{\mathbf{Q}}_\alpha(\mathbf{x}_{\alpha k+1}^\delta) = & \bar{\mathbf{Q}}_\alpha(\mathbf{x}_{\alpha k}^\delta) - \frac{1}{\mathbf{s}_{\alpha k}^T \mathbf{W}_\alpha(\mathbf{x}_{\alpha k}^\delta) \mathbf{s}_{\alpha k}} \mathbf{W}_\alpha(\mathbf{x}_{\alpha k}^\delta) \mathbf{s}_{\alpha k} \mathbf{s}_{\alpha k}^T \mathbf{W}_\alpha(\mathbf{x}_{\alpha k}^\delta) \\ & + \frac{1}{\mathbf{h}_{\alpha k}^T \mathbf{s}_{\alpha k}} \mathbf{h}_{\alpha k} \mathbf{h}_{\alpha k}^T, \end{aligned} \quad (6.19)$$

where

$$\mathbf{s}_{\alpha k} = \mathbf{x}_{\alpha k+1}^\delta - \mathbf{x}_{\alpha k}^\delta$$

is the change in  $\mathbf{x}$  during the current iteration step,

$$\mathbf{h}_{\alpha k} = \mathbf{g}_{\alpha}(\mathbf{x}_{\alpha k+1}^{\delta}) - \mathbf{g}_{\alpha}(\mathbf{x}_{\alpha k}^{\delta})$$

is the change in the gradient, and

$$\mathbf{W}_{\alpha}(\mathbf{x}_{\alpha k}^{\delta}) = \mathbf{K}_{\mathbf{f}\alpha}(\mathbf{x}_{\alpha k+1}^{\delta})^T \mathbf{K}_{\mathbf{f}\alpha}(\mathbf{x}_{\alpha k+1}^{\delta}) + \bar{\mathbf{Q}}_{\alpha}(\mathbf{x}_{\alpha k}^{\delta}).$$

### Step length

A step-length procedure is frequently included in Newton-type methods because a step length of unity along the Newton direction will not necessarily reduce the objective function. The main requirements of a step-length procedure can be summarized as follows: if  $\mathbf{x}$  and  $\mathbf{p}$  denote the actual iterate and the search direction, respectively, then

- (1) the average rate of decrease from  $\mathcal{F}_{\alpha}(\mathbf{x})$  to  $\mathcal{F}_{\alpha}(\mathbf{x} + \tau\mathbf{p})$  should be at least some prescribed fraction  $\varepsilon_{\mathbf{f}} > 0$  of the initial rate of decrease in that direction,

$$\mathcal{F}_{\alpha}(\mathbf{x} + \tau\mathbf{p}) \leq \mathcal{F}_{\alpha}(\mathbf{x}) + \varepsilon_{\mathbf{f}}\tau\mathbf{g}_{\alpha}(\mathbf{x})^T\mathbf{p};$$

- (2) the rate of decrease of  $\mathcal{F}_{\alpha}$  in the direction  $\mathbf{p}$  at  $\mathbf{x} + \tau\mathbf{p}$  should be larger than some prescribed fraction  $\varepsilon_{\mathbf{g}} > 0$  of the rate of decrease in the direction  $\mathbf{p}$  at  $\mathbf{x}$ ,

$$\mathbf{g}_{\alpha}(\mathbf{x} + \tau\mathbf{p})^T\mathbf{p} \geq \varepsilon_{\mathbf{g}}\mathbf{g}_{\alpha}(\mathbf{x})^T\mathbf{p}.$$

The first condition guarantees a sufficient decrease in  $\mathcal{F}_{\alpha}$  values relative to the length of the step, while the second condition avoids too small steps relative to the initial rate of decrease of  $\mathcal{F}_{\alpha}$ . The condition  $\varepsilon_{\mathbf{g}} > \varepsilon_{\mathbf{f}}$  implies that both conditions can be satisfied simultaneously. In practice, the second condition is not needed because the use of a backtracking strategy avoids excessively small steps.

Since computational experience has shown the importance of taking a full step length whenever possible, the modern strategy of a step-length algorithm is to start with  $\tau = 1$ , and then, if  $\mathbf{x} + \mathbf{p}$  is not acceptable, ‘backtrack’ (reduce  $\tau$ ) until an acceptable  $\mathbf{x} + \tau\mathbf{p}$  is found. The backtracking step-length algorithm 5 uses only condition (1) and is based on quadratic and cubic interpolation (Dennis and Schnabel, 1996). On the first backtracking, the new step length is selected as the minimizer of the quadratic interpolation function  $m_{\mathbf{q}}(\tau)$ , defined by

$$m_{\mathbf{q}}(0) = \mathcal{F}_{\alpha}(\mathbf{x}), \quad m'_{\mathbf{q}}(0) = \mathbf{g}_{\alpha}(\mathbf{x})^T\mathbf{p}, \quad m_{\mathbf{q}}(1) = \mathcal{F}_{\alpha}(\mathbf{x} + \mathbf{p}),$$

but being constrained to be larger than  $\varepsilon_1 = 0.1$  of the old step length. On all subsequent backtracks, the new step length is chosen by using the values of the objective function at the last two values of the step length. Essentially, if  $\tau$  and  $\tau_{\text{prv}}$  are the last two values of the step length, the new step length is computed as the minimizer of the cubic interpolation function  $m_{\mathbf{c}}(\tau)$ , defined by

$$m_{\mathbf{c}}(0) = \mathcal{F}_{\alpha}(\mathbf{x}), \quad m'_{\mathbf{c}}(0) = \mathbf{g}_{\alpha}(\mathbf{x})^T\mathbf{p},$$

and

$$m_{\mathbf{c}}(\tau) = \mathcal{F}_{\alpha}(\mathbf{x} + \tau\mathbf{p}), \quad m_{\mathbf{c}}(\tau_{\text{prv}}) = \mathcal{F}_{\alpha}(\mathbf{x} + \tau_{\text{prv}}\mathbf{p}),$$

but being constrained to be larger than  $\varepsilon_1 = 0.1$  and smaller than  $\varepsilon_2 = 0.5$  of the old step length.

---

**Algorithm 5.** Step-length algorithm. Given the actual iterate  $\mathbf{x}$  and the search direction  $\mathbf{p}$ , the algorithm computes the new iterate  $\mathbf{x}_{\text{new}}$ . The control parameters can be chosen as  $\varepsilon_f = 10^{-4}$ ,  $\varepsilon_1 = 0.1$  and  $\varepsilon_2 = 0.5$ .

---

```

 $\mathcal{F}_\alpha \leftarrow 0.5 \|\mathbf{f}_\alpha(\mathbf{x})\|^2$ ;  $\mathbf{g}_\alpha \leftarrow \mathbf{K}_{\mathbf{f}_\alpha}(\mathbf{x})^T \mathbf{f}_\alpha(\mathbf{x})$ ;
estimate  $\tau_{\min}$ ;
 $\tau \leftarrow 1$ ; stop  $\leftarrow$  false;
while stop = false do
     $\mathbf{x}_{\text{new}} \leftarrow \mathbf{x} + \tau \mathbf{p}$ ;  $\mathcal{F}_{\alpha \text{new}} \leftarrow 0.5 \|\mathbf{f}_\alpha(\mathbf{x}_{\text{new}})\|^2$ ;
    {satisfactory  $\mathbf{x}_{\text{new}}$  found}
    if  $\mathcal{F}_{\alpha \text{new}} \leq \mathcal{F}_\alpha + \varepsilon_f \tau \mathbf{g}_\alpha^T \mathbf{p}$  then
        stop  $\leftarrow$  true;
        {no satisfactory  $\mathbf{x}_{\text{new}}$  can be found distinctly from  $\mathbf{x}$ }
    else if  $\tau < \tau_{\min}$  then
         $\mathbf{x}_{\text{new}} \leftarrow \mathbf{x}$ ; stop  $\leftarrow$  true;
        {reduce  $\tau$ }
    else
        {quadratic interpolation}
        if  $\tau = 1$  then
             $\tau_{\text{tmp}} \leftarrow -0.5 \mathbf{g}_\alpha^T \mathbf{p} / (\mathcal{F}_{\alpha \text{new}} - \mathcal{F}_\alpha - \mathbf{g}_\alpha^T \mathbf{p})$ ;
        {cubic interpolation}
        else
             $\begin{bmatrix} a \\ b \end{bmatrix} \leftarrow \frac{1}{\tau - \tau_{\text{prv}}} \begin{bmatrix} 1/\tau^2 & -1/\tau_{\text{prv}}^2 \\ -\tau_{\text{prv}}/\tau^2 & \tau/\tau_{\text{prv}}^2 \end{bmatrix} \begin{bmatrix} \mathcal{F}_{\alpha \text{new}} - \mathcal{F}_\alpha - \tau \mathbf{g}_\alpha^T \mathbf{p} \\ \mathcal{F}_{\alpha \text{prv}} - \mathcal{F}_\alpha - \tau_{\text{prv}} \mathbf{g}_\alpha^T \mathbf{p} \end{bmatrix}$ ;
             $\Delta \leftarrow b^2 - 3a \mathbf{g}_\alpha^T \mathbf{p}$ ;
            if  $a = 0$  then
                 $\tau_{\text{tmp}} \leftarrow -\mathbf{g}_\alpha^T \mathbf{p} / (2b)$ ; {cubic is a quadratic}
            else
                 $\tau_{\text{tmp}} \leftarrow (-b + \sqrt{\Delta}) / (3a)$ ; {true cubic}
            end if
            if  $\tau_{\text{tmp}} > \varepsilon_2 \tau$   $\tau_{\text{tmp}} \leftarrow \varepsilon_2 \tau$ ;
        end if
         $\tau_{\text{prv}} \leftarrow \tau$ ;  $\mathcal{F}_{\alpha \text{prv}} \leftarrow \mathcal{F}_{\alpha \text{new}}$ ;
        if  $\tau_{\text{tmp}} \leq \varepsilon_1 \tau$  then
             $\tau \leftarrow \varepsilon_1 \tau$ ;
        else
             $\tau \leftarrow \tau_{\text{tmp}}$ ;
        end if
    end if
end while

```

---

### 6.2.2 Trust-region methods

In a trust-region method, the step length is taken as unity, so that the new iterate is defined by

$$\mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_{\alpha k}^\delta + \mathbf{p}_{\alpha k}^\delta.$$

For this reason, the term ‘step’ is often used to designate the search direction  $\mathbf{p}_{\alpha k}^\delta$ . In order to ensure that the descent condition holds, it is necessary to compute several trial steps before finding a satisfactory  $\mathbf{p}_{\alpha k}^\delta$ . The most common mathematical formulation of this idea computes the trial step  $\mathbf{p}_{\alpha k}^\delta$  by solving the constrained minimization problem

$$\begin{aligned} \min_{\mathbf{p}} \mathcal{M}_{\alpha k}(\mathbf{p}) \\ \text{subject to } \|\mathbf{p}\| \leq \Gamma_k, \end{aligned} \quad (6.20)$$

where  $\mathcal{M}_{\alpha k}$  is the quadratic model (6.14) at the current iterate  $\mathbf{x}_{\alpha k}^\delta$ , and  $\Gamma_k$  is the trust-region radius. Thus, as opposite to a step-length method, in which we retain the same step direction and choose a shorter step length without making use of the quadratic model, in a trust-region method, we first select a shorter step length and then use the quadratic model to choose the step direction.

Assuming that the solution occurs on the boundary of the constraint region, the first-order optimality conditions for the Lagrangian function

$$\mathcal{L}(\mathbf{p}, \lambda) = \mathcal{F}_\alpha(\mathbf{x}_{\alpha k}^\delta) + \mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) \mathbf{p} + \frac{1}{2} \lambda \left( \|\mathbf{p}\|^2 - \Gamma_k^2 \right),$$

yield

$$[\mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) + \lambda \mathbf{I}_n] \mathbf{p}_\lambda = -\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta) \quad (6.21)$$

and

$$\|\mathbf{p}_\lambda\|^2 = \Gamma_k^2. \quad (6.22)$$

Particularizing the trust-region method for general minimization to least squares problems with a Gauss–Newton approximation to the Hessian, we deduce that the trial step solves the equation

$$[\mathbf{K}_{\mathbf{f}\alpha}(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{K}_{\mathbf{f}\alpha}(\mathbf{x}_{\alpha k}^\delta) + \lambda \mathbf{I}_n] \mathbf{p}_\lambda = -\mathbf{K}_{\mathbf{f}\alpha}(\mathbf{x}_{\alpha k}^\delta)^T \mathbf{f}_\alpha(\mathbf{x}_{\alpha k}^\delta), \quad (6.23)$$

while the Lagrange multiplier  $\lambda$  solves equation (6.22). For comparison with a step-length method, we note that the solution of (6.23) is a solution of the regularized least squares problem

$$\mathbf{p}_{\alpha k}^\delta = \arg \min_{\mathbf{p}} \left( \|\mathbf{f}_\alpha(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\mathbf{f}\alpha}(\mathbf{x}_{\alpha k}^\delta) \mathbf{p}\|^2 + \lambda \|\mathbf{p}\|^2 \right). \quad (6.24)$$

If  $\lambda$  is zero,  $\mathbf{p}_{\alpha k}^\delta$  is the Gauss–Newton direction; as  $\lambda \rightarrow \infty$ ,  $\mathbf{p}_{\alpha k}^\delta$  becomes parallel to the steepest-descent direction  $-\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)$ .

Generally, a trust-region algorithm uses the predictive reduction in the linearized model (6.14),

$$\Delta \mathcal{F}_{\alpha k}^{\text{pred}} = \mathcal{M}_{\alpha k}(0) - \mathcal{M}_{\alpha k}(\mathbf{p}_{\alpha k}^\delta) \quad (6.25)$$

and the actual reduction in the objective function

$$\Delta \mathcal{F}_{\alpha k} = \mathcal{F}_{\alpha}(\mathbf{x}_{\alpha k}^{\delta}) - \mathcal{F}_{\alpha}(\mathbf{x}_{\alpha k}^{\delta} + \mathbf{p}_{\alpha k}^{\delta}) \quad (6.26)$$

to decide whether the trial step  $\mathbf{p}_{\alpha k}^{\delta}$  is acceptable and how the next trust-region radius is chosen. The heuristics to update the size of the trust region usually depends on the ratio of the actual change in  $\mathcal{F}_{\alpha}$  to the predicted change.

The trust-region algorithm 6 finds a new iterate and produces a trust-region radius for the next iteration step (Dennis and Schnabel, 1996). The algorithm starts with the calculation of the trial step  $\mathbf{p}$  for the actual trust-region radius  $\Gamma$  (cf. (6.22) and (6.23)), and with the computation of the prospective iterate  $\mathbf{x}_{\text{new}} = \mathbf{x} + \mathbf{p}$  and the objective function  $\mathcal{F}_{\alpha}(\mathbf{x}_{\text{new}})$ . Then, depending on the average rate of decrease of the objective function, the following situations may appear.

- (1) If  $\mathcal{F}_{\alpha}(\mathbf{x}_{\text{new}}) > \mathcal{F}_{\alpha}(\mathbf{x}) + \varepsilon_{\text{f}} \mathbf{g}_{\alpha}(\mathbf{x})^T \mathbf{p}$ , then the step is unacceptable. In this case, if the trust-region radius is too small, the algorithm terminates with  $\mathbf{x}_{\text{new}} = \mathbf{x}$ . If not, the step length  $\tau_{\min}$  is computed as the minimizer of the quadratic interpolation function  $m_{\text{q}}(\tau)$ , defined by

$$m_{\text{q}}(0) = \mathcal{F}_{\alpha}(\mathbf{x}), \quad m'_{\text{q}}(0) = \mathbf{g}_{\alpha}(\mathbf{x})^T \mathbf{p}, \quad m_{\text{q}}(1) = \mathcal{F}_{\alpha}(\mathbf{x} + \mathbf{p}),$$

and the new radius is chosen as  $\tau_{\min} \|\mathbf{p}\|$  but constrained to be between  $\varepsilon_{1\Gamma} = 0.1$  and  $\varepsilon_{2\Gamma} = 0.5$  of the old radius.

- (2) If  $\mathcal{F}_{\alpha}(\mathbf{x}_{\text{new}}) \leq \mathcal{F}_{\alpha}(\mathbf{x}) + \varepsilon_{\text{f}} \mathbf{g}_{\alpha}(\mathbf{x})^T \mathbf{p}$ , then the step is acceptable, and the reduction of the objective function predicted by the quadratic Gauss–Newton model

$$\Delta \mathcal{F}_{\alpha}^{\text{pred}} = -\mathbf{g}_{\alpha}(\mathbf{x})^T \mathbf{p} - \frac{1}{2} \|\mathbf{K}_{\text{f}\alpha}(\mathbf{x}) \mathbf{p}\|^2$$

is computed. If  $\Delta \mathcal{F}_{\alpha} = \mathcal{F}_{\alpha}(\mathbf{x}) - \mathcal{F}_{\alpha}(\mathbf{x}_{\text{new}})$  and  $\Delta \mathcal{F}_{\alpha}^{\text{pred}}$  agree to within a prescribed tolerance, or negative curvature is indicated, then the trust-region radius is increased and the while loop is continued. If not,  $\mathbf{x}_{\text{new}}$  is accepted as the new iterate, and the trust-region radius is updated for the next iteration step.

A simplified version of a trust-region method is widely used in atmospheric remote sensing (Rodgers, 2000; Eriksson et al., 2005; von Clarmann et al., 2003). In these implementations, the step  $\mathbf{p}_{\alpha k}^{\delta}$  is computed by solving equation (6.24) and a heuristic strategy is used to update the Lagrange multiplier  $\lambda$  at each iteration step.

### 6.2.3 Termination criteria

In a deterministic setting, the standard termination criteria for unconstrained minimization are the X-convergence test (Dennis and Schnabel, 1996)

$$\max_i \left( \frac{|\mathbf{x}_{\alpha k+1}^{\delta}]_i - [\mathbf{x}_{\alpha k}^{\delta}]_i|}{\max \left( |\mathbf{x}_{\alpha k+1}^{\delta}]_i|, \text{typ}[\mathbf{x}]_i \right)} \right) \leq \varepsilon_{\text{x}} \quad (6.27)$$



---

**Algorithm 6.** Trust-region algorithm. Given the actual iterate  $\mathbf{x}$  and the trust-region radius  $\Gamma$ , the algorithm computes the new iterate  $\mathbf{x}_{\text{new}}$  and produces a starting trust-region radius  $\Gamma$  for the next iteration step. The control parameters can be chosen as  $\varepsilon_f = 10^{-4}$ ,  $\varepsilon_{1\Gamma} = 0.1$ ,  $\varepsilon_{2\Gamma} = 0.5$ ,  $\delta = 0.01$ ,  $c_a = 2$ ,  $c_r = 0.5$ ,  $\varepsilon_r = 0.1$  and  $\varepsilon_a = 0.75$ .

---

```

 $\mathcal{F}_\alpha \leftarrow 0.5 \|\mathbf{f}_\alpha(\mathbf{x})\|^2$ ;  $\mathbf{g}_\alpha \leftarrow \mathbf{K}_{f_\alpha}(\mathbf{x})^T \mathbf{f}_\alpha(\mathbf{x})$ ;
estimate  $\Gamma_{\min}$  and  $\Gamma_{\max}$ ;  $retcode \leftarrow 4$ ;
while  $retcode > 1$  do
    compute the trial step  $\mathbf{p}$  for the trust-region radius  $\Gamma$ ;
     $\mathbf{x}_{\text{new}} \leftarrow \mathbf{x} + \mathbf{p}$ ;  $\mathcal{F}_{\alpha\text{new}} \leftarrow 0.5 \|\mathbf{f}_\alpha(\mathbf{x}_{\text{new}})\|^2$ ;  $\Delta\mathcal{F}_\alpha \leftarrow \mathcal{F}_\alpha - \mathcal{F}_{\alpha\text{new}}$ ;
    {if  $retcode = 3$ , reset  $\mathbf{x}_{\text{new}}$  to  $\mathbf{x}_{\text{prv}}$  and terminate the while loop}
    if  $retcode = 3$  and  $(\mathcal{F}_{\alpha\text{new}} \geq \mathcal{F}_{\alpha\text{prv}}$  or  $\mathcal{F}_{\alpha\text{new}} > \mathcal{F}_\alpha + \varepsilon_f \mathbf{g}_\alpha^T \mathbf{p})$  then
         $retcode \leftarrow 0$ ;  $\mathbf{x}_{\text{new}} \leftarrow \mathbf{x}_{\text{prv}}$ ;  $\mathcal{F}_{\alpha\text{new}} \leftarrow \mathcal{F}_{\alpha\text{prv}}$ ;
    {objective function is too large; reduce  $\Gamma$  and continue the while loop}
    else if  $\mathcal{F}_{\alpha\text{new}} > \mathcal{F}_\alpha + \varepsilon_f \mathbf{g}_\alpha^T \mathbf{p}$  then
        if  $\Gamma < \Gamma_{\min}$  then
             $retcode \leftarrow 1$ ;  $\mathbf{x}_{\text{new}} \leftarrow \mathbf{x}$ ;  $\mathcal{F}_{\alpha\text{new}} \leftarrow \mathcal{F}_\alpha$ ;
        else
             $retcode \leftarrow 2$ ;  $\Gamma_{\text{tmp}} \leftarrow 0.5 (\mathbf{g}_\alpha^T \mathbf{p}) \|\mathbf{p}\| / (\Delta\mathcal{F}_\alpha + \mathbf{g}_\alpha^T \mathbf{p})$ ;
            if  $\Gamma_{\text{tmp}} < \varepsilon_{1\Gamma} \Gamma$  then
                 $\Gamma \leftarrow \varepsilon_{1\Gamma} \Gamma$ ;
            else if  $\Gamma_{\text{tmp}} > \varepsilon_{2\Gamma} \Gamma$  then
                 $\Gamma \leftarrow \varepsilon_{2\Gamma} \Gamma$ ;
            else
                 $\Gamma \leftarrow \Gamma_{\text{tmp}}$ ;
            end if
        end if
    {objective function is sufficiently small}
    else
         $\Delta\mathcal{F}_\alpha^{\text{pred}} \leftarrow -\mathbf{g}_\alpha^T \mathbf{p} - 0.5 \|\mathbf{K}_{f_\alpha} \mathbf{p}\|^2$ ;
        {increase  $\Gamma$  and continue the while loop}
        if  $retcode \neq 2$  and  $(|\Delta\mathcal{F}_\alpha^{\text{pred}} - \Delta\mathcal{F}_\alpha| \leq \delta \Delta\mathcal{F}_\alpha$  or  $\mathcal{F}_{\alpha\text{new}} \leq \mathcal{F}_\alpha + \mathbf{g}_\alpha^T \mathbf{p})$ 
            and  $\Gamma < \Gamma_{\max}$  then
                 $retcode \leftarrow 3$ ;  $\mathbf{x}_{\text{prv}} \leftarrow \mathbf{x}_{\text{new}}$ ;  $\mathcal{F}_{\alpha\text{prv}} \leftarrow \mathcal{F}_{\alpha\text{new}}$ ;  $\Gamma \leftarrow \min(c_a \Gamma, \Gamma_{\max})$ ;
                {accept  $\mathbf{x}_{\text{new}}$  as new iterate and update  $\Gamma$  for the next iteration step}
            else
                 $retcode \leftarrow 0$ ;
                if  $\Delta\mathcal{F}_\alpha \leq \varepsilon_r \Delta\mathcal{F}_\alpha^{\text{pred}}$  then
                     $\Gamma \leftarrow \max(c_r \Gamma, \Gamma_{\min})$ ; {reduce  $\Gamma$ }
                else if  $\Delta\mathcal{F}_\alpha \geq \varepsilon_a \Delta\mathcal{F}_\alpha^{\text{pred}}$  then
                     $\Gamma \leftarrow \min(c_a \Gamma, \Gamma_{\max})$ ; {increase  $\Gamma$ }
                end if
            end if
        end if
    end while

```

---

and the relative gradient test

$$\max_i \left( \left| [\mathbf{g}_\alpha(\mathbf{x}_{\alpha k+1}^\delta)]_i \right| \frac{\max \left( \left| [\mathbf{x}_{\alpha k+1}^\delta]_i \right|, \text{typ}[\mathbf{x}]_i \right)}{\max(\mathcal{F}_\alpha(\mathbf{x}_{\alpha k+1}^\delta), \text{typ} \mathcal{F})} \right) \leq \varepsilon_g. \quad (6.28)$$

The first condition checks whether the sequence  $\{\mathbf{x}_{\alpha k}^\delta\}$  is converging, while the second criterion reflects the optimality condition  $\mathbf{g}_\alpha(\mathbf{x}_\alpha^\delta) \approx \mathbf{0}$ . The relative gradient test (6.28) is a modification of the conventional gradient test  $\|\mathbf{g}_\alpha(\mathbf{x}_{\alpha k+1}^\delta)\|_\infty \leq \varepsilon_g$ , which is strongly dependent on the scaling of both  $\mathcal{F}_\alpha$  and  $\mathbf{x}$ . The relative gradient of  $\mathcal{F}_\alpha$  at  $\mathbf{x}$ , defined as the ratio of the relative rate of change in  $\mathcal{F}_\alpha$  to the relative rate of change in  $\mathbf{x}$ , is independent of any change in the units of  $\mathcal{F}_\alpha$  and  $\mathbf{x}$ .

The termination criteria (6.27) and (6.28) are formulated in terms of the infinity norm (or the maximum norm) rather than in terms of the two-norm (or the Euclidean norm). The reason is that for large  $n$ , the number of terms contributing to the magnitude of the two-norm may cause these tests to be extremely severe.

It should be mentioned that the problem of measuring relative changes when the argument  $z$  is near zero is addressed by substituting  $z$  with  $\max(|z|, \text{typ } z)$ , where  $\text{typ } z$  is an estimate of a typical magnitude of  $z$ . Otherwise, we may substitute  $1 + |z|$  for  $z$ , in which case, the X-convergence test becomes (Gill et al., 1981)

$$\max_i \left( \left| [\mathbf{x}_{\alpha k+1}^\delta]_i - [\mathbf{x}_{\alpha k}^\delta]_i \right| \right) \leq \varepsilon_x \left[ 1 + \max_i \left( \left| [\mathbf{x}_{\alpha k+1}^\delta]_i \right| \right) \right].$$

As a matter of fact, a third strategy is adopted in the PORT optimization routines (Dennis et al., 1981). Here, the problem of measuring relative changes in  $\mathbf{x}$  is addressed by formulating the X-convergence test as

$$\frac{\max_i \left( \left| [\mathbf{x}_{\alpha k+1}^\delta]_i - [\mathbf{x}_{\alpha k}^\delta]_i \right| \right)}{\max_i \left( \left| [\mathbf{x}_{\alpha k+1}^\delta]_i \right| + \left| [\mathbf{x}_{\alpha k}^\delta]_i \right| \right)} \leq \varepsilon_x. \quad (6.29)$$

It is apparent from the above discussion that the termination criteria are based on an implicit definition of ‘small’ and ‘large’, and that variables with large varying orders of magnitude may cause difficulties for some minimization algorithms. This problem can be remedied by scaling the variables through a linear transformation. The goal of the scaling process is to make all the variables of a similar order of magnitude in the region of interest. If typical values of all variables are known (e.g., an a priori atmospheric profile), we may pass from the original variable  $\mathbf{x}$  to the transformed variable  $\hat{\mathbf{x}}$  through the linear transformation  $\hat{\mathbf{x}} = \mathbf{D}\mathbf{x}$ , where  $\mathbf{D}$  is a diagonal matrix with entries  $[\mathbf{D}]_{ii} = 1/\text{typ}[\mathbf{x}]_i$ ,  $i = 1, \dots, n$ . Sometimes the scaling by a diagonal matrix only has the disadvantage that the magnitude of a variable may vary substantially during the minimization process and that some accuracy may be lost. This situation can be overcome if a range of values, that a variable is likely to assume, is known. For example, if we know that  $l_i \leq [\mathbf{x}]_i \leq u_i$  for  $i = 1, \dots, n$ , then the transformed variable  $\hat{\mathbf{x}}$  is defined by (Gill et al., 1981)

$$[\hat{\mathbf{x}}]_i = \frac{2[\mathbf{x}]_i}{u_i - l_i} - \frac{u_i + l_i}{u_i - l_i}, \quad i = 1, \dots, n.$$

In a stochastic framework, the X-convergence test involves the change in the iterate scaled by its estimated error, that is, (Rodgers, 2000; Eriksson et al., 2005)

$$\frac{(\mathbf{x}_{\alpha k+1}^\delta - \mathbf{x}_{\alpha k}^\delta)^T \hat{\mathbf{C}}_{\mathbf{x}k}^{-1} (\mathbf{x}_{\alpha k+1}^\delta - \mathbf{x}_{\alpha k}^\delta)}{n} \leq \varepsilon_{\mathbf{x}}, \quad (6.30)$$

where

$$\hat{\mathbf{C}}_{\mathbf{x}k} = (\mathbf{K}_{\alpha k}^T \mathbf{C}_\delta^{-1} \mathbf{K}_{\alpha k} + \mathbf{C}_{\mathbf{x}}^{-1})^{-1}$$

is the a posteriori covariance matrix at the iteration step  $k$  and  $\mathbf{K}_{\alpha k} = \mathbf{K}(\mathbf{x}_{\alpha k}^\delta)$ . The idea behind criterion (6.30) is that, if  $\mathbf{x}_\alpha^\delta$  is the minimizer of the Tikhonov function, and  $\mathbf{C}_\delta$  and  $\mathbf{C}_{\mathbf{x}}$  accurately reproduce the covariance matrices of the errors in the data and of the true state, respectively, then the random variable

$$(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)^T \hat{\mathbf{C}}_{\mathbf{x}}^{-1} (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta),$$

with  $\hat{\mathbf{C}}_{\mathbf{x}}$  corresponding to  $\mathbf{x}_\alpha^\delta$ , is Chi-square distributed with  $n$  degrees of freedom (Appendix D). Here, the true state  $\mathbf{x}^\dagger$  and its estimator  $\mathbf{x}_\alpha^\delta$  should be regarded as random variables distributed as  $\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{C}}_{\mathbf{x}})$ . Essentially, condition (6.30) requires that instead of the infinity norm, the Mahalanobis norm between two successive iterates  $\|\mathbf{x}_{\alpha k+1}^\delta - \mathbf{x}_{\alpha k}^\delta\|_{\hat{\mathbf{C}}_{\mathbf{x}k}^{-1}}^2$  scaled by  $n$  is smaller than the prescribed tolerance  $\varepsilon_{\mathbf{x}}$ .

A termination criterion, which is frequently used in conjunction with a regularization parameter choice method, is the relative function convergence test (Rodgers, 2000; Carissimo et al., 2005)

$$\frac{r_k^\delta - r_{k+1}^\delta}{r_k^\delta} \leq \varepsilon_{\mathbf{f}\mathbf{r}}, \quad (6.31)$$

where

$$r_k^\delta = [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\alpha k}^\delta)]^T \mathbf{C}_{\mathbf{r}k}^{-1} [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\alpha k}^\delta)],$$

is the ‘residual’, and

$$\mathbf{C}_{\mathbf{r}k} = \mathbf{C}_\delta (\mathbf{K}_{\alpha k} \mathbf{C}_{\mathbf{x}} \mathbf{K}_{\alpha k}^T + \mathbf{C}_\delta)^{-1} \mathbf{C}_\delta$$

is the covariance matrix of the residual at the iteration step  $k$ . If the noise and the a priori covariance matrices properly describe the errors in the data and the true state, respectively, and moreover, if the iterate  $\mathbf{x}_\alpha^\delta$ , satisfying the relative function convergence test, is a minimizer of the Tikhonov function  $\mathcal{F}_\alpha$ , then the corresponding residual  $r^\delta$  is Chi-square distributed with  $m$  degrees of freedom (Appendix D). In this regard, to test the ‘correct’ convergence, we check the condition

$$m - \sqrt{2m}z_{t/2} < r^\delta < m + \sqrt{2m}z_{t/2},$$

where  $z_{t/2}$  is the relevant  $z$ -value for a Chi-square distribution with  $m$  degrees of freedom, and  $t$  is the significance level. A similar test can be performed in the state space by considering the ‘constraint’

$$c_k^\delta = (\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a)^T \mathbf{C}_{\mathbf{x}k} (\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a),$$

with

$$\mathbf{C}_{\hat{\mathbf{x}}k} = \mathbf{C}_{\mathbf{x}} \mathbf{K}_{\alpha k}^T \mathbf{C}_{\delta}^{-1} \mathbf{K}_{\alpha k} (\mathbf{K}_{\alpha k}^T \mathbf{C}_{\delta}^{-1} \mathbf{K}_{\alpha k} + \mathbf{C}_{\mathbf{x}}^{-1})^{-1},$$

and by taking into account that at the minimizer  $\mathbf{x}_{\alpha}^{\delta}$ , the constraint  $c^{\delta}$  is Chi-square distributed with  $n$  degrees of freedom. It should be pointed out that for  $\mathbf{C}_{\delta} = \sigma^2 \mathbf{I}_m$ , the relative function convergence test, formulated in terms of the residual  $\|\mathbf{r}_{\alpha k}^{\delta}\|^2 = \|\mathbf{y}^{\delta} - \mathbf{F}(\mathbf{x}_{\alpha k}^{\delta})\|^2$ , plays a significant role in the framework of iterative regularization methods.

### 6.2.4 Software packages

The Gauss–Newton model of the Hessian is used, usually with enhancements, in much of the software for nonlinear least squares as for example, MINPACK, NAG, TENSOLVE and PORT. For a survey of optimization software we recommend the monograph by Moré and Wright (1993).

The algorithms in MINPACK (Moré et al., 1980) are based on the trust-region concept and employ either a finite-difference or an analytical Jacobian matrix.

The NAG routines (NAG Fortran Library Manual, 1993) use a Gauss–Newton search direction whenever a sufficiently large decrease in the objective function is attained. Otherwise, second-order derivative information is obtained from user-supplied function evaluation routines, quasi-Newton approximations, or difference approximations. Using this information, the software attempts to find a more accurate approximation to the Newton direction than the Gauss–Newton direction is able to provide.

The TENSOLVE software (Bouaricha and Schnabel, 1997) augments the Gauss–Newton model with a low-rank tensor approximation to the second-order term. It has been observed to converge faster than standard Gauss–Newton on many problems, particularly when the Jacobian matrix is rank deficient at the solution.

The optimization algorithms implemented in the PORT library use a trust-region method in conjunction with a Gauss–Newton model and a quasi-Newton model to compute the trial step (Dennis et al., 1981). When the first trial step fails, the alternate model gets a chance to make a trial step with the same trust-region radius. If the alternate model fails to suggest a more successful step, then the current model is maintained for the duration of the present iteration step. The trust-region radius is then decreased until the new iterate is determined or the algorithm fails.

## 6.3 Practical methods for computing the new iterate

A step-length method for minimizing the Tikhonov function is of the form of the following model algorithm:

- (1) compute the search direction;
- (2) compute the step length by using Algorithm 5;
- (3) terminate the iterative process according to the X-convergence test.

The step-length procedure is optional, but our experience demonstrates that this technique improves the stability of the method and reduces the number of iteration steps. In this sec-

tion we are concerned with the computation of the search direction  $\mathbf{p}_{\alpha k}^\delta$ , or more precisely, with the computation of the new iterate  $\mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_{\alpha k}^\delta + \mathbf{p}_{\alpha k}^\delta$ . Certainly, if a step-length procedure is part of the inversion algorithm, then  $\mathbf{x}_{\alpha k+1}^\delta$  is the prospective iterate, but we prefer to use the term ‘new iterate’ because it is frequently encountered in the remote sensing community.

Using the explicit expressions of the augmented vector  $\mathbf{f}_\alpha$  and of the Jacobian matrix  $\mathbf{K}_{\mathbf{f}_\alpha}$ , we deduce that the Gauss–Newton direction  $\mathbf{p}_{\alpha k}^\delta$  solves the equation (cf. (6.16))

$$(\mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k} + \alpha \mathbf{L}^T \mathbf{L}) \mathbf{p} = -\mathbf{K}_{\alpha k}^T [\mathbf{F}(\mathbf{x}_{\alpha k}^\delta) - \mathbf{y}^\delta] - \alpha \mathbf{L}^T \mathbf{L} (\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a),$$

with  $\mathbf{K}_{\alpha k} = \mathbf{K}(\mathbf{x}_{\alpha k}^\delta)$ . Passing from the unknown  $\mathbf{p} = \mathbf{x} - \mathbf{x}_{\alpha k}^\delta$  to the unknown  $\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_a$  yields the regularized normal equation

$$(\mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k} + \alpha \mathbf{L}^T \mathbf{L}) \Delta \mathbf{x} = \mathbf{K}_{\alpha k}^T \mathbf{y}_k^\delta,$$

with

$$\mathbf{y}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\alpha k}(\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a). \quad (6.32)$$

The new iterate is then given by

$$\mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_a + \mathbf{K}_{\alpha k}^\dagger \mathbf{y}_k^\delta, \quad (6.33)$$

where

$$\mathbf{K}_{\alpha k}^\dagger = (\mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k} + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}_{\alpha k}^T$$

is the regularized generalized inverse at the iteration step  $k$ .

In order to give a more practical interpretation of the Gauss–Newton iterate (6.33), we consider a linearization of  $\mathbf{F}$  about  $\mathbf{x}_{\alpha k}^\delta$ ,

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\alpha k}(\mathbf{x} - \mathbf{x}_{\alpha k}^\delta) + \mathbf{R}(\mathbf{x}, \mathbf{x}_{\alpha k}^\delta),$$

where  $\mathbf{R}$  is the remainder term of the first-order Taylor expansion or the linearization error about  $\mathbf{x}_{\alpha k}^\delta$ . If  $\mathbf{x}^\dagger$  is a solution of the nonlinear equation with exact data  $\mathbf{F}(\mathbf{x}) = \mathbf{y}$ , then  $\mathbf{x}^\dagger$  is defined by the equation

$$\mathbf{K}_{\alpha k}(\mathbf{x}^\dagger - \mathbf{x}_a) = \mathbf{y}_k$$

where

$$\mathbf{y}_k = \mathbf{y} - \mathbf{F}(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\alpha k}(\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a) - \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k}^\delta).$$

Because  $\mathbf{y}_k$  is unknown, we consider the equation

$$\mathbf{K}_{\alpha k}(\mathbf{x} - \mathbf{x}_a) = \mathbf{y}_k^\delta, \quad (6.34)$$

with  $\mathbf{y}_k^\delta$  being given by (6.32). Evidently, the errors in the data  $\mathbf{y}_k^\delta$  are due to the instrumental noise and the linearization error, and we have the representation

$$\mathbf{y}_k^\delta - \mathbf{y}_k = \boldsymbol{\delta} + \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k}^\delta).$$

Because the nonlinear problem is ill-posed, its linearization is also ill-posed, and we solve the linearized equation (6.34) by means of Tikhonov regularization with the penalty term

$\|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2$  and the regularization parameter  $\alpha$ . The Tikhonov function for the linearized equation takes the form

$$\mathcal{F}_{1\alpha k}(\mathbf{x}) = \|\mathbf{y}_k^\delta - \mathbf{K}_{\alpha k}(\mathbf{x} - \mathbf{x}_a)\|^2 + \alpha \|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2,$$

and its minimizer is given by (6.33). Thus, the solution of a nonlinear ill-posed problem by means of Tikhonov regularization is equivalent to the solution of a sequence of ill-posed linearizations of the forward model about the current iterate.

The new iterate can be computed by using the GSVD of the matrix pair  $(\mathbf{K}_{\alpha k}, \mathbf{L})$ . Although, the GSVD is of great theoretical interest for analyzing general-form regularization problems, it is of computational interest only for small- and medium-scale problems. The reason is that the computation of the GSVD of the matrix pair  $(\mathbf{K}_{\alpha k}, \mathbf{L})$  is quite demanding; the conventional implementation requires about  $2m^2n + 15n^3$  operations (Hanke and Hansen, 1993). For practical solutions of large-scale problems it is much simpler to deal with standard-form problems in which the regularization matrix is the identity matrix and only the SVD of the transformed Jacobian matrix is required. The regularization in standard form relies on the solution of the equation

$$\bar{\mathbf{K}}_{\alpha k} \triangle \bar{\mathbf{x}} = \mathbf{y}_k^\delta, \quad (6.35)$$

with  $\bar{\mathbf{K}}_{\alpha k} = \mathbf{K}_{\alpha k} \mathbf{L}^{-1}$  and  $\triangle \mathbf{x} = \mathbf{L}^{-1} \triangle \bar{\mathbf{x}}$ , by means of Tikhonov regularization with  $\mathbf{L} = \mathbf{I}_n$ . If  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of  $\bar{\mathbf{K}}_{\alpha k}$ , the solution of the standard-form problem expressed by the regularized normal equation

$$(\bar{\mathbf{K}}_{\alpha k}^T \bar{\mathbf{K}}_{\alpha k} + \alpha \mathbf{I}_n) \triangle \bar{\mathbf{x}} = \bar{\mathbf{K}}_{\alpha k}^T \mathbf{y}_k^\delta \quad (6.36)$$

reads as

$$\triangle \bar{\mathbf{x}}_{\alpha k+1}^\delta = \sum_{i=1}^n \frac{\sigma_i}{\sigma_i^2 + \alpha} (\mathbf{u}_i^T \mathbf{y}_k^\delta) \mathbf{v}_i.$$

An efficient implementation of Tikhonov regularization for large-scale problems, which also takes into account that we wish to solve (6.36) several times for various regularization parameters, is described in Hanke and Hansen (1993). In this approach, the standard-form problem is treated as a least squares problem of the form (cf. (6.36))

$$\min_{\bar{\mathbf{x}}} \left\| \begin{bmatrix} \bar{\mathbf{K}}_{\alpha k} \\ \sqrt{\alpha} \mathbf{I}_n \end{bmatrix} \triangle \bar{\mathbf{x}} - \begin{bmatrix} \mathbf{y}_k^\delta \\ \mathbf{0} \end{bmatrix} \right\|^2.$$

The matrix  $\bar{\mathbf{K}}_{\alpha k}$  is transformed into an upper bidiagonal matrix  $\mathbf{J}$ ,

$$\bar{\mathbf{K}}_{\alpha k} = \mathbf{U} \begin{bmatrix} \mathbf{J} \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T,$$

by means of orthogonal transformations from the left and from the right, with  $\mathbf{U} \in \mathbb{R}^{m \times m}$ ,  $\mathbf{J} \in \mathbb{R}^{n \times n}$  and  $\mathbf{V} \in \mathbb{R}^{n \times n}$ . The bidiagonal matrix  $\mathbf{J}$  is computed explicitly, while the orthogonal matrices  $\mathbf{U}$  and  $\mathbf{V}$  are represented by series of orthogonal transformations, which are usually stored in appropriate arrays and later used when matrix-vector multiplications, e.g.,  $\mathbf{U}^T \mathbf{x}$  and  $\mathbf{V} \mathbf{x}$ , are needed. Making the change of variables

$$\boldsymbol{\xi} = \mathbf{V}^T \triangle \bar{\mathbf{x}}, \quad \mathbf{z}^\delta = \mathbf{U}^T \mathbf{y}_k^\delta,$$

and assuming the partition

$$\mathbf{z}^\delta = \begin{bmatrix} \mathbf{z}_1^\delta \\ \mathbf{z}_2^\delta \end{bmatrix}, \quad \mathbf{z}_1^\delta \in \mathbb{R}^n,$$

we are led to an equivalent minimization problem expressed as

$$\min_{\xi} \left\| \begin{bmatrix} \mathbf{J} \\ \sqrt{\alpha} \mathbf{I}_n \end{bmatrix} \xi - \begin{bmatrix} \mathbf{z}_1^\delta \\ \mathbf{0} \end{bmatrix} \right\|^2.$$

As shown by Elden (1977), the above minimization problem can be solved very efficiently by means of  $O(n)$  operations. Essentially, for each value of the regularization parameter, we compute the QR factorization

$$\begin{bmatrix} \mathbf{J} \\ \sqrt{\alpha} \mathbf{I}_n \end{bmatrix} = \mathbf{Q}_\alpha \begin{bmatrix} \mathbf{T}_\alpha \\ \mathbf{0} \end{bmatrix}, \quad (6.37)$$

by means of  $2n - 1$  Givens rotations, where  $\mathbf{T}_\alpha \in \mathbb{R}^{n \times n}$  is an upper bidiagonal matrix, and  $\mathbf{Q}_\alpha \in \mathbb{R}^{2n \times 2n}$  is a product of Givens rotations. Further, defining the vector

$$\zeta_\alpha^\delta = \mathbf{Q}_\alpha^T \begin{bmatrix} \mathbf{z}_1^\delta \\ \mathbf{0} \end{bmatrix},$$

and partitioning  $\zeta_\alpha^\delta$  as

$$\zeta_\alpha^\delta = \begin{bmatrix} \zeta_{\alpha 1}^\delta \\ \zeta_{\alpha 2}^\delta \end{bmatrix}, \quad \zeta_{\alpha 1}^\delta \in \mathbb{R}^n,$$

we obtain

$$\xi_\alpha^\delta = \mathbf{T}_\alpha^{-1} \zeta_{\alpha 1}^\delta$$

and finally,

$$\Delta \bar{\mathbf{x}}_{\alpha k+1}^\delta = \mathbf{V} \xi_\alpha^\delta.$$

This solution method, relying on a bidiagonalization of the Jacobian matrix, is outlined in Algorithm 7.

The standard-form problem can be formulated as the augmented normal equation (cf. (6.36))

$$\bar{\mathbf{K}}_{\mathbf{f}\alpha}^T \bar{\mathbf{K}}_{\mathbf{f}\alpha} \Delta \bar{\mathbf{x}} = \bar{\mathbf{K}}_{\mathbf{f}\alpha}^T \bar{\mathbf{f}}, \quad (6.38)$$

with

$$\bar{\mathbf{f}} = \begin{bmatrix} \mathbf{y}_k^\delta \\ \mathbf{0} \end{bmatrix}, \quad \bar{\mathbf{K}}_{\mathbf{f}\alpha} = \begin{bmatrix} \bar{\mathbf{K}}_{\alpha k} \\ \sqrt{\alpha} \mathbf{I}_n \end{bmatrix},$$

and the linear equation  $\bar{\mathbf{K}}_{\mathbf{f}\alpha} \Delta \bar{\mathbf{x}} = \bar{\mathbf{f}}$  can be solved by using iterative methods for normal equations like the CGNR and the LSQR algorithms. For large-scale problems, the computational efficiency can be increased by using an appropriate preconditioner. The preconditioner  $\mathbf{M}$  for the normal equation (6.38) should be chosen so that the condition number of  $\mathbf{M}^T \bar{\mathbf{K}}_{\mathbf{f}\alpha}^T \bar{\mathbf{K}}_{\mathbf{f}\alpha} \mathbf{M}$  is small. The construction of a preconditioner based on the close connection between the Lanczos algorithm and the conjugate gradient method has been described by Hohage (2001). Assuming the singular value decomposition  $\bar{\mathbf{K}}_{\alpha k} = \mathbf{U} \Sigma \mathbf{V}^T$ , we have

$$\bar{\mathbf{K}}_{\mathbf{f}\alpha}^T \bar{\mathbf{K}}_{\mathbf{f}\alpha} = \mathbf{V} \left[ \text{diag}(\sigma_i^2 + \alpha)_{n \times n} \right] \mathbf{V}^T,$$

---

**Algorithm 7.** Implementation of Tikhonov regularization with Jacobian bidiagonalization. The bidiagonalization of  $\bar{\mathbf{K}}_{\alpha k}$  can be performed by using the routine DGEHRD from the LAPACK library (Anderson et al., 1995), while the products  $\mathbf{U}^T \mathbf{y}_k^\delta$  and  $\mathbf{V} \xi_\alpha^\delta$  can be computed by using the routine DORMBR from the same library. The notation  $\mathbf{J} = \text{bidiag}[\mathbf{d}, \mathbf{e}]$  means that  $\mathbf{d}$  and  $\mathbf{e}$  are the diagonal and superdiagonal of the bidiagonal matrix  $\mathbf{J}$ .

---

$$\bar{\mathbf{K}}_{\alpha k} \leftarrow \mathbf{K}_{\alpha k} \mathbf{L}^{-1};$$

$$\text{bidiagonalize } \bar{\mathbf{K}}_{\alpha k} = \mathbf{U} \begin{bmatrix} \mathbf{J} \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T \text{ with } \mathbf{J} = \text{bidiag}[\mathbf{d}, \mathbf{e}];$$

$$\mathbf{z}^\delta \leftarrow \mathbf{U}^T \mathbf{y}_k^\delta; \text{ partition } \mathbf{z}^\delta = \begin{bmatrix} \mathbf{z}_1^\delta \\ \mathbf{z}_2^\delta \end{bmatrix} \text{ with } \mathbf{z}_1^\delta \in \mathbb{R}^n; \quad \bar{\mathbf{z}} \leftarrow \begin{bmatrix} \mathbf{z}_1^\delta \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{2n};$$

$$\{\text{QR factorization } \begin{bmatrix} \mathbf{J} \\ \sqrt{\alpha} \mathbf{I}_n \end{bmatrix} = \mathbf{Q}_\alpha \begin{bmatrix} \mathbf{T}_\alpha \\ \mathbf{0} \end{bmatrix}, \mathbf{T}_\alpha = \text{bidiag}[\mathbf{d}, \mathbf{e}]; \bar{\mathbf{z}} \leftarrow \mathbf{Q}_\alpha^T \bar{\mathbf{z}}\}$$

**for**  $i = 1, n$  **do**  $[\mathbf{s}]_i \leftarrow \sqrt{\alpha}$ ; **end for** {diagonal of the regularization matrix  $\sqrt{\alpha} \mathbf{I}_n$ }

**for**  $i = 1, n$  **do**

{rotation in  $(i, i + n)$ -plane: angle of rotation}

$$\rho \leftarrow \sqrt{[\mathbf{d}]_i^2 + [\mathbf{s}]_i^2}; \quad \sin \theta \leftarrow [\mathbf{s}]_i / \rho; \quad \cos \theta \leftarrow [\mathbf{d}]_i / \rho;$$

{rotation in  $(i, i + n)$ -plane:  $[\mathbf{d}]_i$  and  $[\mathbf{e}]_i$ }

**if**  $i \leq n - 1$  **then**

$$[\mathbf{e}]_i \leftarrow [\mathbf{d}]_i [\mathbf{e}]_i / \rho; \quad \lambda \leftarrow [\mathbf{s}]_i [\mathbf{e}]_i / \rho;$$

**end if**

$$[\mathbf{d}]_i \leftarrow \rho;$$

$$[\mathbf{s}]_i \leftarrow 0;$$

{rotation in  $(i, i + n)$ -plane:  $\bar{\mathbf{z}} \leftarrow \mathbf{Q}_\alpha^T \bar{\mathbf{z}}$ }

$$w_1 \leftarrow \cos \theta [\bar{\mathbf{z}}]_i + \sin \theta [\bar{\mathbf{z}}]_{i+n}; \quad w_2 \leftarrow -\sin \theta [\bar{\mathbf{z}}]_i + \cos \theta [\bar{\mathbf{z}}]_{i+n};$$

$$[\bar{\mathbf{z}}]_i \leftarrow w_1; \quad [\bar{\mathbf{z}}]_{i+n} \leftarrow w_2;$$

**if**  $i \leq n - 1$  **then**

{rotation in  $(i + n, i + n + 1)$ -plane: angle of rotation}

$$\rho \leftarrow \sqrt{\lambda^2 + [\mathbf{s}]_{i+1}^2}; \quad \sin \theta \leftarrow \lambda / \rho; \quad \cos \theta \leftarrow [\mathbf{s}]_{i+1} / \rho;$$

$$[\mathbf{s}]_{i+1} \leftarrow \rho;$$

{rotation in  $(i + n, i + n + 1)$ -plane:  $\bar{\mathbf{z}} \leftarrow \mathbf{Q}_\alpha^T \bar{\mathbf{z}}$ }

$$w_1 \leftarrow \cos \theta [\bar{\mathbf{z}}]_{i+n} + \sin \theta [\bar{\mathbf{z}}]_{i+n+1}; \quad w_2 \leftarrow -\sin \theta [\bar{\mathbf{z}}]_{i+n} + \cos \theta [\bar{\mathbf{z}}]_{i+n+1};$$

$$[\bar{\mathbf{z}}]_{i+n} \leftarrow w_1; \quad [\bar{\mathbf{z}}]_{i+n+1} \leftarrow w_2;$$

**end if**

**end for**

$$\{\text{solve } \mathbf{T}_\alpha \xi_\alpha^\delta = \bar{\mathbf{z}}_1, \text{ where } \mathbf{T}_\alpha = \text{bidiag}[\mathbf{d}, \mathbf{e}] \text{ and } \bar{\mathbf{z}} = \begin{bmatrix} \bar{\mathbf{z}}_1 \\ \bar{\mathbf{z}}_2 \end{bmatrix} \text{ with } \bar{\mathbf{z}}_1 \in \mathbb{R}^n\}$$

$$[\xi_\alpha^\delta]_n \leftarrow [\bar{\mathbf{z}}]_n / [\mathbf{d}]_n;$$

$$\textbf{for } i = 1, n - 1 \textbf{ do } [\xi_\alpha^\delta]_{n-i} \leftarrow \left( [\bar{\mathbf{z}}]_{n-i} - [\mathbf{e}]_{n-i} [\xi_\alpha^\delta]_{n-i+1} \right) / [\mathbf{d}]_{n-i}; \textbf{ end do}$$

{new iterate}

$$\Delta \bar{\mathbf{x}}_\alpha^\delta \leftarrow \mathbf{V} \xi_\alpha^\delta; \quad \mathbf{x}_{\alpha k+1}^\delta \leftarrow \mathbf{L}^{-1} \Delta \bar{\mathbf{x}}_\alpha^\delta + \mathbf{x}_a;$$


---



and for a fixed index  $r$ , the preconditioner can be constructed as

$$\mathbf{M} = \mathbf{V} \begin{bmatrix} \text{diag} \left( \frac{1}{\sqrt{\sigma_i^2 + \alpha}} \right)_{r \times r} & \mathbf{0} \\ \mathbf{0} & \text{diag} \left( \frac{1}{\sqrt{\alpha}} \right)_{(n-r) \times (n-r)} \end{bmatrix} \mathbf{V}^T.$$

We then obtain

$$\mathbf{M}^T \bar{\mathbf{K}}_{f\alpha}^T \bar{\mathbf{K}}_{f\alpha} \mathbf{M} = \mathbf{V} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \text{diag} \left( \frac{\sigma_i^2 + \alpha}{\alpha} \right)_{(n-r) \times (n-r)} \end{bmatrix} \mathbf{V}^T,$$

and the condition number of  $\mathbf{M}^T \bar{\mathbf{K}}_{f\alpha}^T \bar{\mathbf{K}}_{f\alpha} \mathbf{M}$  is  $1 + \sigma_{r+1}^2 / \alpha$ . If  $\sigma_{r+1}^2$  is not much larger than  $\alpha$ , then the condition number is small and very few iteration steps are required to compute the new iterate. Turning now to practical implementation issues we mention that iterative algorithms are coded without explicit reference to  $\mathbf{M}$ ; only the matrix-vector product  $\mathbf{M}\mathbf{x}$  is involved. Since

$$\mathbf{M}\mathbf{x} = \frac{1}{\sqrt{\alpha}} \mathbf{x} + \sum_{i=1}^r \left( \frac{1}{\sqrt{\sigma_i^2 + \alpha}} - \frac{1}{\sqrt{\alpha}} \right) (\mathbf{v}_i^T \mathbf{x}) \mathbf{v}_i,$$

we observe that the calculation of  $\mathbf{M}\mathbf{x}$  requires the knowledge of the first  $r$  singular values and right singular vectors of  $\bar{\mathbf{K}}_{f\alpha}$ , and these quantities can be efficiently computed by the Lanczos Algorithm 8. The steps of computing the  $r$  singular values and right singular vectors of an  $m \times n$  matrix  $\mathbf{A}$  can be synthesized as follows:

- (1) apply  $r$  steps of the Lanczos bidiagonalization algorithm with Householder orthogonalization to produce a lower  $(r+1) \times r$  bidiagonal matrix  $\mathbf{B}$ , an  $n \times r$  matrix  $\bar{\mathbf{V}}$  containing the right singular vectors, and an  $m \times (r+1)$  matrix  $\bar{\mathbf{U}}$  containing the left singular vectors,

$$\mathbf{A} \bar{\mathbf{V}} = \bar{\mathbf{U}} \mathbf{B};$$

- (2) compute the QR factorization of the bidiagonal matrix  $\mathbf{B}$ ,

$$\mathbf{B} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix},$$

where  $\mathbf{Q}$  is an  $(r+1) \times (r+1)$  orthogonal matrix, and  $\mathbf{R}$  is an upper  $r \times r$  bidiagonal matrix;

- (3) compute the SVD of the bidiagonal matrix  $\mathbf{R}$ ,

$$\mathbf{R} = \mathbf{U}_R \Sigma \mathbf{V}_R^T;$$

- (4) the first  $r$  singular values are the diagonal entries of  $\Sigma$ , while the corresponding right singular vectors are the column vectors of the  $n \times r$  matrix

$$\mathbf{V} = \bar{\mathbf{V}} \mathbf{V}_R.$$

---

**Algorithm 8.** Lanczos bidiagonalization algorithm for estimating the first  $r$  singular values and right singular vectors of a matrix  $\mathbf{A}$ . The  $r$  singular values are stored in the diagonal of  $\Sigma$ , while the corresponding right singular vectors are stored in the columns of  $\mathbf{V}$ . The algorithm uses the LAPACK routine DLARTG to generate a plane rotation. The SVD of the bidiagonal matrix  $\mathbf{R}$  can be computed by using the routine DBDSDC from the LAPACK library. The routine HOrth is given in Chapter 5.

---

```

 $\mathbf{p} \leftarrow \mathbf{0}$ ;  $[\mathbf{p}]_1 \leftarrow 1$ ; {choose  $\mathbf{p}$  arbitrarily, e.g., the first Cartesian unit vector}
 $\mathbf{d} \leftarrow \mathbf{0}$ ;  $\mathbf{e} \leftarrow \mathbf{0}$ ;  $\bar{\mathbf{v}} \leftarrow \mathbf{0}$ ;
 $\boldsymbol{\pi} \leftarrow \mathbf{0}$ ;  $\mathbf{P} \leftarrow \mathbf{0}$ ;  $\boldsymbol{\nu} \leftarrow \mathbf{0}$ ;  $\mathbf{Q} \leftarrow \mathbf{0}$ ;
{initialization of arrays  $\mathbf{P}$  and  $\boldsymbol{\pi}$ }
 $p \leftarrow \|\mathbf{p}\|$ ;  $[\boldsymbol{\pi}]_1 \leftarrow 1/(p^2 + \|[\mathbf{p}]_1\|p)$ ;
 $[\mathbf{P}]_{11} \leftarrow [\mathbf{p}]_1 + \text{sgn}([\mathbf{p}]_1)p$ ; for  $k = 2, m$  do  $[\mathbf{P}]_{k1} \leftarrow [\mathbf{p}]_k$ ; end for
 $\beta \leftarrow -\text{sgn}([\mathbf{p}]_1)p$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta)\mathbf{p}$ ;
for  $i = 1, r$  do
     $\mathbf{q} \leftarrow \mathbf{A}^T \bar{\mathbf{u}} - \beta \bar{\mathbf{v}}$ ;
    call HOrth ( $i, n, \boldsymbol{\nu}, \mathbf{Q}, \mathbf{q}$ ;  $\bar{\mathbf{v}}, \alpha$ );
     $[\mathbf{d}]_i \leftarrow \alpha$ ; for  $k = 1, n$  do  $[\bar{\mathbf{V}}]_{ki} \leftarrow [\bar{\mathbf{v}}]_k$ ; end for {store  $\alpha$  and  $\bar{\mathbf{v}}$ }
     $\mathbf{p} \leftarrow \mathbf{A} \bar{\mathbf{v}} - \alpha \bar{\mathbf{u}}$ ;
    call HOrth ( $i + 1, m, \boldsymbol{\pi}, \mathbf{P}, \mathbf{p}$ ;  $\bar{\mathbf{u}}, \beta$ );
     $[\mathbf{e}]_i \leftarrow \beta$ ; {store  $\beta$ }
end for
{compute the QR factorization of  $\mathbf{B} = \text{bidiag}[\mathbf{d}, \mathbf{e}]$ }
for  $i = 1, r - 1$  do
    call DLARTG( $[\mathbf{d}]_i, [\mathbf{e}]_i$ ;  $c, s, \rho$ );
     $[\mathbf{d}]_i \leftarrow \rho$ ;  $[\mathbf{e}]_i \leftarrow s[\mathbf{d}]_{i+1}$ ;  $[\mathbf{d}]_{i+1} \leftarrow c[\mathbf{d}]_{i+1}$ ;
end for
if  $r < \min(m, n)$  then
    call DLARTG( $[\mathbf{d}]_r, [\mathbf{e}]_r$ ;  $c, s, \rho$ );
     $[\mathbf{d}]_r \leftarrow \rho$ ;  $[\mathbf{e}]_r \leftarrow 0$ ;
end if
compute the SVD  $\mathbf{R} = \mathbf{U}_R \Sigma \mathbf{V}_R^T$ , where  $\mathbf{R} = \text{bidiag}[\mathbf{d}, \mathbf{e}]$ ;
 $\mathbf{V} \leftarrow \bar{\mathbf{V}} \mathbf{V}_R$ ;

```

---

The above computational steps yield

$$\mathbf{A} = \bar{\mathbf{U}} \mathbf{B} \bar{\mathbf{V}}^T = \bar{\mathbf{U}} \mathbf{Q} \begin{bmatrix} \mathbf{U}_R \Sigma \mathbf{V}_R^T \\ \mathbf{0} \end{bmatrix} \bar{\mathbf{V}}^T = \bar{\mathbf{U}} \mathbf{Q} \begin{bmatrix} \mathbf{U}_R & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \Sigma \\ \mathbf{0} \end{bmatrix} (\bar{\mathbf{V}} \mathbf{V}_R)^T;$$

whence, taking into account that the product of two matrices with orthonormal columns is also a matrix with orthonormal columns, we deduce that Algorithm 8 serves the desired purpose.

The regularized normal equation (6.36) can be expressed as

$$\mathbf{A} \triangle \bar{\mathbf{x}} = \mathbf{b}, \quad (6.39)$$

with

$$\mathbf{A} = \bar{\mathbf{K}}_{\alpha k}^T \bar{\mathbf{K}}_{\alpha k} + \alpha \mathbf{I}_n$$

**Table 6.2.** Computation time in min:ss format for different solution methods.

Problem	Solution method			
	GSVD	SVD	Bidiagonalization	CGNR
O <sub>3</sub>	0:25	0:16	0:14	0:14
BrO	0:32	0:21	0:18	0:20
CO	5:04	3:23	2:43	3:19
Temperature	5:28	3:37	2:54	3:32

and

$$\mathbf{b} = \bar{\mathbf{K}}_{\alpha k}^T \mathbf{y}_k^\delta.$$

As  $\mathbf{A}$  is symmetric, the system of equations (6.39) can be solved by using standard iterative solvers, as for example, the Conjugate Gradient Squared (CGS) or the Biconjugate Gradient Stabilized (Bi-CGSTAB) methods (Barrett et al., 1994). A relevant practical aspect is that for iterative methods, the matrix  $\mathbf{A}$  is never formed explicitly as only matrix-vector products with  $\mathbf{A}$  and eventually with  $\mathbf{A}^T$  are required. The calculation of the matrix-vector product  $\mathbf{A}\mathbf{x}$  demands the calculation of  $\bar{\mathbf{K}}_{\alpha k}^T \bar{\mathbf{K}}_{\alpha k} \mathbf{x}$ , and, clearly, this should be computed as  $\bar{\mathbf{K}}_{\alpha k}^T (\bar{\mathbf{K}}_{\alpha k} \mathbf{x})$  and not by forming the cross-product matrix  $\bar{\mathbf{K}}_{\alpha k}^T \bar{\mathbf{K}}_{\alpha k}$ . The reason for avoiding explicit formation of the cross-product matrix is the loss of information due to round-off errors. A right preconditioner for the system of equations (6.39), i.e.,

$$\mathbf{A}\mathbf{M}_a \triangle \bar{\mathbf{x}}' = \mathbf{b}, \quad \mathbf{M}_a \triangle \bar{\mathbf{x}}' = \triangle \bar{\mathbf{x}},$$

can also be constructed by using the Lanczos algorithm. For  $\bar{\mathbf{K}}_{\alpha k} = \mathbf{U}\Sigma\mathbf{V}^T$ , the right preconditioner is given by

$$\mathbf{M}_a = \mathbf{V} \begin{bmatrix} \text{diag} \left( \frac{1}{\sigma_i^2 + \alpha} \right)_{r \times r} & \mathbf{0} \\ \mathbf{0} & \text{diag} \left( \frac{1}{\alpha} \right)_{(n-r) \times (n-r)} \end{bmatrix} \mathbf{V}^T,$$

in which case, the condition number of  $\mathbf{A}\mathbf{M}_a$  is  $1 + \sigma_{r+1}^2/\alpha$ , and we have

$$\mathbf{M}_a \mathbf{x} = \frac{1}{\alpha} \mathbf{x} + \sum_{i=1}^r \left( \frac{1}{\sigma_i^2 + \alpha} - \frac{1}{\alpha} \right) (\mathbf{v}_i^T \mathbf{x}) \mathbf{v}_i.$$

The comparison of the numerical effort of the methods for computing the new iterate can be inferred from Table 6.2. The fastest method is the approach relying on a bidiagonalization of the Jacobian matrix, and as expected, the slowest method is the approach based on the GSVD of the matrix pair  $(\mathbf{K}_{\alpha k}, \mathbf{L})$ .

## 6.4 Error characterization

An important part of a retrieval is to assess the accuracy of the regularized solution by performing an error analysis. The error representation depends on the solution method which is used to compute a minimizer of the Tikhonov function, or more precisely, on the Hessian approximation.

### 6.4.1 Gauss–Newton method

The Gauss–Newton iterate  $\mathbf{x}_{\alpha k+1}^\delta$  is the regularized solution of the linearized equation (6.34) and its expression is given by (6.33). For the exact data vector  $\mathbf{y}$ , the Gauss–Newton iterate possesses a similar representation, namely

$$\mathbf{x}_{\alpha k+1} = \mathbf{x}_a + \mathbf{K}_{\alpha k}^\dagger \mathbf{y}_k,$$

where, in order to avoid an abundance of notations,  $\mathbf{y}_k$  is now given by

$$\mathbf{y}_k = \mathbf{y} - \mathbf{F}(\mathbf{x}_{\alpha k}) + \mathbf{K}(\mathbf{x}_{\alpha k})(\mathbf{x}_{\alpha k} - \mathbf{x}_a).$$

As in the linear case, we consider the representation

$$\mathbf{x}^\dagger - \mathbf{x}_{\alpha k+1}^\delta = (\mathbf{x}^\dagger - \mathbf{x}_{\alpha k+1}) + (\mathbf{x}_{\alpha k+1} - \mathbf{x}_{\alpha k+1}^\delta) \quad (6.40)$$

and try to estimate each term in the right-hand side of (6.40). Using the linearizations of the forward model about  $\mathbf{x}_{\alpha k}$  and  $\mathbf{x}_{\alpha k}^\delta$ ,

$$\mathbf{y} = \mathbf{F}(\mathbf{x}_{\alpha k}) + \mathbf{K}(\mathbf{x}_{\alpha k})(\mathbf{x}^\dagger - \mathbf{x}_{\alpha k}) + \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k})$$

and

$$\mathbf{y} = \mathbf{F}(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\alpha k}(\mathbf{x}^\dagger - \mathbf{x}_{\alpha k}^\delta) + \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k}^\delta),$$

respectively, and assuming that  $\mathbf{K}_{\alpha k} = \mathbf{K}(\mathbf{x}_{\alpha k}^\delta) \approx \mathbf{K}(\mathbf{x}_{\alpha k})$ , we express the first term in the right-hand side of (6.40) as

$$\mathbf{x}^\dagger - \mathbf{x}_{\alpha k+1} = (\mathbf{x}^\dagger - \mathbf{x}_a) - \mathbf{K}_{\alpha k}^\dagger \mathbf{y}_k = (\mathbf{I}_n - \mathbf{A}_{\alpha k})(\mathbf{x}^\dagger - \mathbf{x}_a) - \mathbf{K}_{\alpha k}^\dagger \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k})$$

and the second term as

$$\mathbf{x}_{\alpha k+1} - \mathbf{x}_{\alpha k+1}^\delta = \mathbf{K}_{\alpha k}^\dagger (\mathbf{y}_k - \mathbf{y}_k^\delta) = -\mathbf{K}_{\alpha k}^\dagger \boldsymbol{\delta} - \mathbf{K}_{\alpha k}^\dagger [\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k}^\delta) - \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k})],$$

with  $\mathbf{A}_{\alpha k} = \mathbf{K}_{\alpha k}^\dagger \mathbf{K}_{\alpha k}$  being the averaging kernel matrix. Inserting the above relations in (6.40) we find that

$$\mathbf{x}^\dagger - \mathbf{x}_{\alpha k+1}^\delta = (\mathbf{I}_n - \mathbf{A}_{\alpha k})(\mathbf{x}^\dagger - \mathbf{x}_a) - \mathbf{K}_{\alpha k}^\dagger \boldsymbol{\delta} - \mathbf{K}_{\alpha k}^\dagger \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_{\alpha k}^\delta). \quad (6.41)$$

Assuming that the sequence  $\{\mathbf{x}_{\alpha k}^\delta\}$  converges to  $\mathbf{x}_\alpha^\delta$  and that  $\mathbf{F}$  is continuously differentiable, we let  $k \rightarrow \infty$  in (6.41), and obtain

$$\mathbf{e}_\alpha^\delta = \mathbf{e}_{s\alpha} + \mathbf{e}_{n\alpha}^\delta + \mathbf{e}_{1\alpha}, \quad (6.42)$$

where

$$\mathbf{e}_\alpha^\delta = \mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta \quad (6.43)$$

is the total error in the solution,

$$\mathbf{e}_{s\alpha} = (\mathbf{I}_n - \mathbf{A}_\alpha)(\mathbf{x}^\dagger - \mathbf{x}_a) \quad (6.44)$$

is the smoothing error,

$$\mathbf{e}_{n\alpha}^\delta = -\mathbf{K}_\alpha^\dagger \boldsymbol{\delta} \quad (6.45)$$

is the noise error, and

$$\mathbf{e}_{1\alpha} = -\mathbf{K}_\alpha^\dagger \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_\alpha^\delta)$$

is the nonlinearity error. In the above relations, the generalized inverse  $\mathbf{K}_\alpha^\dagger$  and the averaging kernel matrix  $\mathbf{A}_\alpha$  are evaluated at  $\mathbf{x}_\alpha^\delta$ .

The expression of the total error can also be derived by using the fact that  $\mathbf{x}_\alpha^\delta$  is a minimizer of the Tikhonov function  $\mathcal{F}_\alpha$ . The stationary condition for  $\mathcal{F}_\alpha$  at  $\mathbf{x}_\alpha^\delta$ ,

$$\nabla \mathcal{F}_\alpha(\mathbf{x}_\alpha^\delta) = \mathbf{g}_\alpha(\mathbf{x}_\alpha^\delta) = \mathbf{0},$$

shows that  $\mathbf{x}_\alpha^\delta$  solves the Euler equation

$$\mathbf{K}_\alpha^T [\mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{y}^\delta] + \alpha \mathbf{L}^T \mathbf{L} (\mathbf{x}_\alpha^\delta - \mathbf{x}_a) = \mathbf{0}; \quad (6.46)$$

whence, assuming a linearization of  $\mathbf{F}$  about  $\mathbf{x}_\alpha^\delta$ ,

$$\mathbf{y} = \mathbf{F}(\mathbf{x}_\alpha^\delta) + \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta) + \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_\alpha^\delta), \quad (6.47)$$

we find that

$$(\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{L}^T \mathbf{L}) (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta) = \alpha \mathbf{L}^T \mathbf{L} (\mathbf{x}^\dagger - \mathbf{x}_a) - \mathbf{K}_\alpha^T \boldsymbol{\delta} - \mathbf{K}_\alpha^T \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_\alpha^\delta).$$

Further, using the identity

$$(\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{L}^T \mathbf{L})^{-1} \alpha \mathbf{L}^T \mathbf{L} = \mathbf{I}_n - (\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}_\alpha^T \mathbf{K}_\alpha,$$

we obtain

$$\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta = (\mathbf{I}_n - \mathbf{A}_\alpha) (\mathbf{x}^\dagger - \mathbf{x}_a) - \mathbf{K}_\alpha^\dagger \boldsymbol{\delta} - \mathbf{K}_\alpha^\dagger \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_\alpha^\delta), \quad (6.48)$$

which is the explicit form of (6.42). Thus, the error representations in the nonlinear and the linear case are similar, except for an additional term, which represents the nonlinearity error. If the minimizer  $\mathbf{x}_\alpha^\delta$  is sufficiently close to the exact solution  $\mathbf{x}^\dagger$ , the nonlinearity error can be neglected, and the agreement is complete.

In a semi-stochastic framework, we suppose that  $\mathbf{K}_\alpha$  is deterministic, and as a result, the total error  $\mathbf{e}_\alpha^\delta$  is stochastic with mean  $\mathbf{e}_{s\alpha}$  and covariance  $\mathbf{C}_{en} = \sigma^2 \mathbf{K}_\alpha^\dagger \mathbf{K}_\alpha^{\dagger T}$ . As in the linear case, we define the mean square error matrix

$$\begin{aligned} \mathbf{S}_\alpha &= \mathbf{e}_{s\alpha} \mathbf{e}_{s\alpha}^T + \mathbf{C}_{en} \\ &= (\mathbf{I}_n - \mathbf{A}_\alpha) (\mathbf{x}^\dagger - \mathbf{x}_a) (\mathbf{x}^\dagger - \mathbf{x}_a)^T (\mathbf{I}_n - \mathbf{A}_\alpha)^T + \sigma^2 \mathbf{K}_\alpha^\dagger \mathbf{K}_\alpha^{\dagger T} \end{aligned} \quad (6.49)$$

to quantify the dispersion of the regularized solution  $\mathbf{x}_\alpha^\delta$  about the exact solution  $\mathbf{x}^\dagger$ . The rank-one matrix  $(\mathbf{x}^\dagger - \mathbf{x}_a) (\mathbf{x}^\dagger - \mathbf{x}_a)^T$  can be approximated by (cf. (3.60))

$$(\mathbf{x}^\dagger - \mathbf{x}_a) (\mathbf{x}^\dagger - \mathbf{x}_a)^T \approx (\mathbf{x}_\alpha^\delta - \mathbf{x}_a) (\mathbf{x}_\alpha^\delta - \mathbf{x}_a)^T \quad (6.50)$$

or by (cf. (3.61))

$$(\mathbf{x}^\dagger - \mathbf{x}_a)(\mathbf{x}^\dagger - \mathbf{x}_a)^T \approx \frac{\sigma^2}{\alpha} (\mathbf{L}^T \mathbf{L})^{-1}. \quad (6.51)$$

The approximation (6.50) yields the so-called semi-stochastic representation of  $\mathbf{S}_\alpha$ , while the approximation (6.51) yields the stochastic representation of  $\mathbf{S}_\alpha$ , since in this case,  $\mathbf{S}_\alpha$  coincides with the a posteriori covariance matrix in statistical inversion theory.

In order to assess the validity of the semi-stochastic and stochastic representations of the mean square error matrix, we perform a numerical analysis for the  $\text{O}_3$  retrieval test problem. In Figure 6.3 we plot the average values of the solution error

$$\bar{\varepsilon}_\alpha = \sqrt{\frac{1}{N} \sum_{i=1}^N \varepsilon_{\alpha i}^2}, \quad \varepsilon_{\alpha i}^2 = \frac{\|\mathbf{x}^\dagger - \mathbf{x}_{\alpha i}^\delta\|^2}{\|\mathbf{x}^\dagger\|^2},$$

and of the expected error

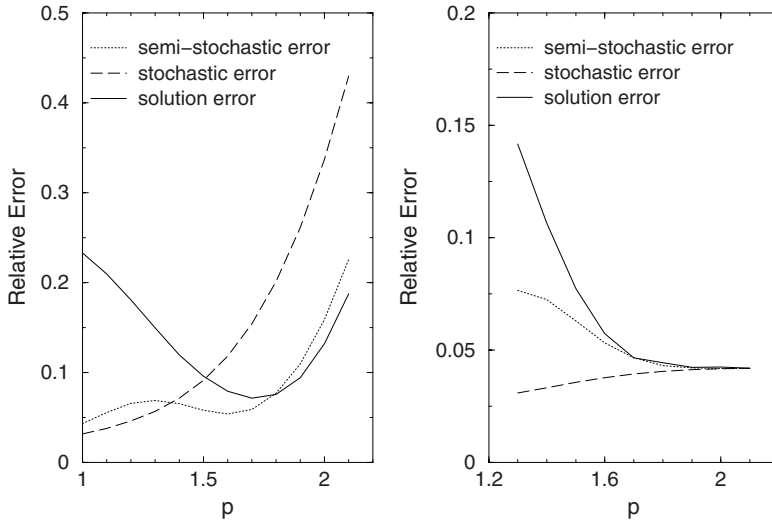
$$\bar{\varepsilon}_{e\alpha} = \sqrt{\frac{1}{N} \sum_{i=1}^N \varepsilon_{e\alpha i}^2}, \quad \varepsilon_{e\alpha i}^2 = \frac{\mathcal{E}\{\|\mathbf{e}_{\alpha i}^\delta\|^2\}}{\|\mathbf{x}^\dagger\|^2} = \frac{\text{trace}(\mathbf{S}_{\alpha i})}{\|\mathbf{x}^\dagger\|^2}$$

for a set of noisy data vectors  $\{\mathbf{y}_i^\delta\}_{i=1, \dots, N}$ , with  $N = 100$ . Here,  $\mathbf{x}_{\alpha i}^\delta$  and  $\mathbf{S}_{\alpha i}$  are the Tikhonov solution and the mean square error matrix corresponding to the noisy data vector  $\mathbf{y}_i^\delta$ , respectively. Because the simulation is performed for a single state vector realization, the numerical analysis is semi-stochastic. The main conclusions are briefly summarized below.

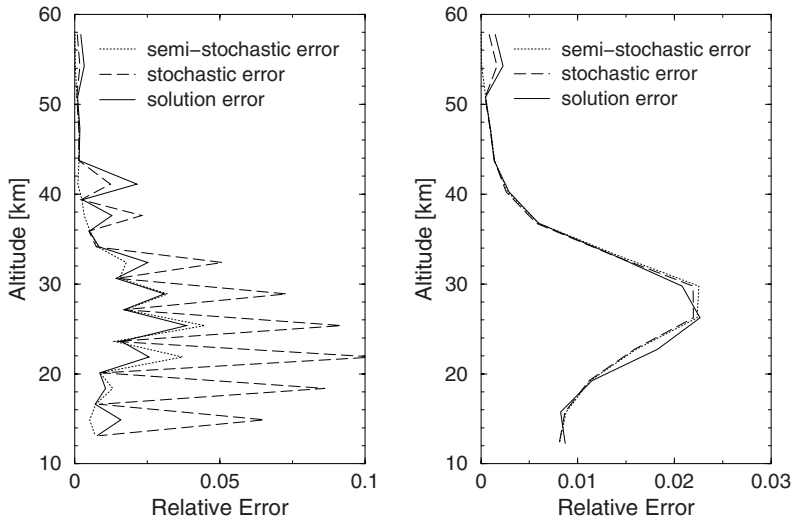
- (1) The (expected) semi-stochastic error, with the smoothing error given by (6.50), approximates sufficiently well the solution error for small values of the regularization parameter and in the neighborhood of the minimizer. For large values of the regularization parameter, the approximation becomes worse because the regularized solution is close to the a priori. As a result, the smoothing error is not a monotonically decreasing function of the regularization parameter, and the semi-stochastic error may not have a unique minimum.
- (2) If the retrieval is not sensitive to some components of the state vector, the (expected) stochastic error, with the smoothing error given by (6.51), is not an appropriate approximation of the solution error. The smoothing error explodes and so, the stochastic error is very large. If the retrieval is sensitive to all components of the state vector, the approximation is satisfactory for that values of the regularization parameter which are close to the minimizer. A typical feature of the stochastic error is that it is a decreasing function of the regularization parameter.

The plots in Figure 6.4 illustrate the distributions of the average errors with respect to the altitude. If the retrieval is sensitive to all components of the state vector, both error representations yields accurate results. If this is not the case, the semi-stochastic representation appears to be superior to the stochastic representation.

An appropriate diagnostic of the retrieval is the comparison of the smoothing and noise errors (Figure 6.5). In general, the minimizer of the solution error is close to the regularization parameter which roughly yields a trade-off between the two error components.

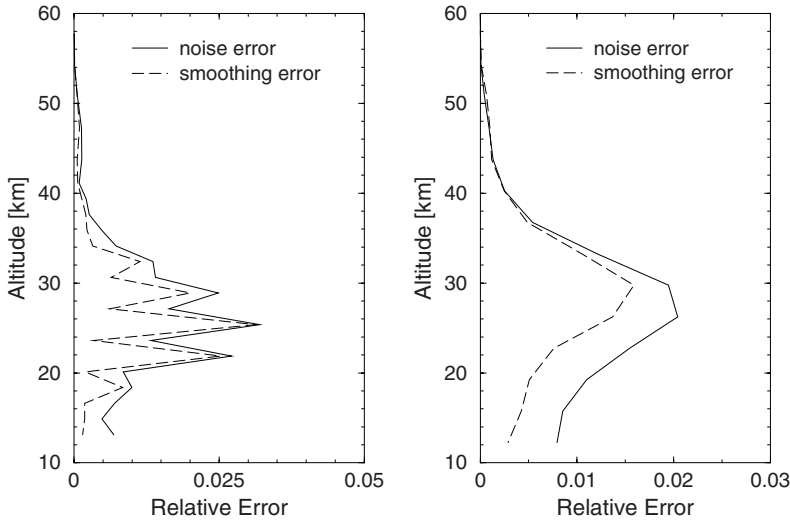


**Fig. 6.3.** Average errors for the  $O_3$  retrieval test problem. The plots in the left panel correspond to an altitude retrieval grid with 36 levels, while the plots in the right panel correspond to an altitude retrieval grid with 24 levels. The regularization parameter is given by  $\alpha = \sigma^p$ , with  $\sigma$  being the noise standard deviation. Since  $\sigma < 1$ , small values of  $\alpha$  correspond to large values of  $p$ .



**Fig. 6.4.** Distributions of the average errors with respect to the altitude for the  $O_3$  retrieval test problem. The plots in the left panel correspond to  $p_{opt} = 1.7$  and an altitude retrieval grid with 36 levels, while the plots in the right panel correspond to  $p_{opt} = 1.9$  and an altitude retrieval grid with 24 levels.

Consequently, if the smoothing and noise errors are of the same order of magnitude, we may conclude that the regularization parameter is close to the minimizer of the solution error.



**Fig. 6.5.** Distributions of the smoothing and noise errors with respect to the altitude for the  $O_3$  retrieval test problem. The curves correspond to the semi-stochastic error representation and to one noisy data realization. The parameters of calculation are as in Figure 6.4.

Accounting for all assumptions employed it is readily seen that a linearized error analysis can be performed when

- (1) the regularization parameter is not too far from the minimizer of the solution error;
- (2) the sequence of iterates  $\{\mathbf{x}_{\alpha k}^\delta\}$  converges;
- (3) the linearization error  $\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_\alpha^\delta)$  is small;
- (4) the errors in the data are correctly modelled.

If one of these assumptions is violated the error analysis is erroneous. The first requirement is the topic of the next section, while the second requirement can be satisfied by using an appropriate termination criterion. Let us pay attention to the last two conditions.

The linearity assumption can be verified at the boundary of a confidence region for the solution (Rodgers, 2000). For this purpose, we consider the SVD of the positive definite mean square error matrix  $\mathbf{S}_\alpha = \mathbf{V}_s \Sigma_s \mathbf{V}_s^T$ , and define the normalized error patterns  $\mathbf{s}_k$  for  $\mathbf{S}_\alpha$  from the partition  $\mathbf{V}_s \Sigma_s^{1/2} = [\mathbf{s}_1, \dots, \mathbf{s}_n]$ . The linearization error

$$\mathbf{R}(\mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x} - \mathbf{x}_\alpha^\delta),$$

can be estimated by comparing

$$\varepsilon_{\text{lin}k}^2 = \frac{1}{m\sigma^2} \|\mathbf{R}(\mathbf{x}_\alpha^\delta \pm \mathbf{s}_k)\|^2 \approx 1,$$

for all  $k = 1, \dots, n$ .

The knowledge of the errors in the data is perhaps the most important problem of an error analysis. If the data error  $\delta_y$  contains only the instrumental noise  $\delta$ , application of



(6.47) and (6.48) gives

$$\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) = \mathbf{K}_\alpha (\mathbf{I}_n - \mathbf{A}_\alpha) (\mathbf{x}^\dagger - \mathbf{x}_a) + (\mathbf{I}_m - \widehat{\mathbf{A}}_\alpha) \boldsymbol{\delta} + (\mathbf{I}_m - \widehat{\mathbf{A}}_\alpha) \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_\alpha^\delta), \quad (6.52)$$

with  $\widehat{\mathbf{A}}_\alpha = \mathbf{K}_\alpha \mathbf{K}_\alpha^\dagger$  being the influence matrix at  $\mathbf{x}_\alpha^\delta$ . As  $\alpha$  approaches 0, the averaging kernel matrix  $\mathbf{A}_\alpha$  approaches the identity matrix  $\mathbf{I}_n$ ; whence, neglecting the linearization error  $\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_\alpha^\delta)$ , we find that the residual  $\mathbf{r}_\alpha^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)$  is given by

$$\mathbf{r}_\alpha^\delta = (\mathbf{I}_m - \widehat{\mathbf{A}}_\alpha) \boldsymbol{\delta} = \sum_{i=n+1}^m (\mathbf{u}_i^T \boldsymbol{\delta}) \mathbf{u}_i, \quad \alpha \rightarrow 0.$$

For  $\boldsymbol{\delta} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ , we then obtain

$$\mathcal{E} \left\{ \|\mathbf{r}_\alpha^\delta\|^2 \right\} = (m - n) \sigma^2, \quad \alpha \rightarrow 0.$$

Equivalently, (6.52) shows that for problems with a small degree of nonlinearity and whenever  $\alpha \rightarrow 0$ , the random variable  $\mathbf{r}_\alpha^{\delta T} \mathbf{C}_\delta^{-1} \mathbf{r}_\alpha^\delta$  with  $\mathbf{C}_\delta = \sigma^2 \mathbf{I}_m$ , is Chi-square distributed with  $m - n$  degrees of freedom (Appendix D). If the contribution of the forward model error  $\boldsymbol{\delta}_m$  in the data error  $\boldsymbol{\delta}_y$  is significant, we have instead

$$\mathcal{E} \left\{ \|\mathbf{r}_\alpha^\delta\|^2 \right\} \approx (m - n) \left( \frac{1}{m} \|\boldsymbol{\delta}_m\|^2 + \sigma^2 \right), \quad \alpha \rightarrow 0.$$

The forward model errors introduce an additional bias in the solution. To handle this type of errors, we may proceed as in the linear case, that is, we may replace the data error  $\boldsymbol{\delta}_y$  by an equivalent white noise  $\boldsymbol{\delta}_e$  with variance

$$\sigma_e^2 = \frac{1}{m} \|\boldsymbol{\delta}_m\|^2 + \sigma^2,$$

so that

$$\mathcal{E} \left\{ \|\boldsymbol{\delta}_e\|^2 \right\} = \mathcal{E} \left\{ \|\boldsymbol{\delta}_y\|^2 \right\}.$$

The noise variance estimate

$$\sigma_e^2 \approx \frac{1}{m - n} \mathcal{E} \left\{ \|\mathbf{r}_\alpha^\delta\|^2 \right\} \approx \frac{1}{m - n} \|\mathbf{r}_\alpha^\delta\|^2, \quad \alpha \rightarrow 0,$$

can then be used to perform an error analysis with the equivalent white noise covariance matrix  $\mathbf{C}_{\delta_e} = \sigma_e^2 \mathbf{I}_m$ . It is apparent that by this equivalence we increase the noise error variance and eliminate the bias due to forward model errors.

## 6.4.2 Newton method

In the framework of the Newton method, the search direction is the solution of the equation (cf. (6.15))

$$\mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) \mathbf{p} = -\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta).$$

To perform an error analysis we rewrite the Newton equation in terms of the a priori profile deviation  $\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_a$ , that is,

$$\mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) \Delta \mathbf{x} = \mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) (\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a) - \mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta)$$

and approximate the right-hand side of the resulting equation as

$$\begin{aligned} & \mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) (\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a) - \mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta) \\ &= [\mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) - (\mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k} + \alpha \mathbf{L}^T \mathbf{L})] (\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a) + \mathbf{K}_{\alpha k}^T \mathbf{y}_k^\delta \\ &\approx \mathbf{K}_{\alpha k}^T \mathbf{y}_k^\delta, \end{aligned}$$

where  $\mathbf{y}_k^\delta$  is given by (6.32). Then, employing the same arguments as in the derivation of (6.42), we find that the smoothing and noise errors are given by

$$\mathbf{e}_{s\alpha} = (\mathbf{I}_n - \mathbf{G}_\alpha^{-1} \mathbf{K}_\alpha^T \mathbf{K}_\alpha) (\mathbf{x}^\dagger - \mathbf{x}_a) \quad (6.53)$$

and

$$\mathbf{e}_{n\alpha}^\delta = -\mathbf{G}_\alpha^{-1} \mathbf{K}_\alpha^T \boldsymbol{\delta}, \quad (6.54)$$

respectively. Hereafter, the notation  $\mathbf{G}_\alpha$  stands for  $\mathbf{G}_\alpha(\mathbf{x}_\alpha^\delta)$ . Thus, the mean vector and the covariance matrix of the total error  $\mathbf{e}_\alpha^\delta$  are the smoothing error  $\mathbf{e}_{s\alpha}$  and the noise error covariance matrix

$$\mathbf{C}_{en} = \sigma^2 \mathbf{G}_\alpha^{-1} \mathbf{K}_\alpha^T \mathbf{K}_\alpha \mathbf{G}_\alpha^{-1}.$$

Similar expressions for the smoothing and noise errors can be derived if we regard the state vector and the data vector as independent variables and consider a linearization of the gradient of the objective function about  $(\mathbf{x}_\alpha^\delta, \mathbf{y}^\delta)$ . Setting  $\mathbf{x}^\dagger = \mathbf{x}_\alpha^\delta - \Delta \mathbf{x}_\alpha^\delta$  and  $\mathbf{y} = \mathbf{y}^\delta - \boldsymbol{\delta}$ , we have

$$\mathbf{g}_\alpha(\mathbf{x}^\dagger, \mathbf{y}) = \mathbf{g}_\alpha(\mathbf{x}_\alpha^\delta, \mathbf{y}^\delta) - \frac{\partial \mathbf{g}_\alpha}{\partial \mathbf{x}}(\mathbf{x}_\alpha^\delta, \mathbf{y}^\delta) \Delta \mathbf{x}_\alpha^\delta - \frac{\partial \mathbf{g}_\alpha}{\partial \mathbf{y}}(\mathbf{x}_\alpha^\delta, \mathbf{y}^\delta) \boldsymbol{\delta} + \mathbf{R}(\mathbf{x}^\dagger, \mathbf{y}; \mathbf{x}_\alpha^\delta, \mathbf{y}^\delta),$$

where  $\mathbf{R}$  is the remainder term of the first-order Taylor expansion of the gradient. Using the stationary condition  $\mathbf{g}_\alpha(\mathbf{x}_\alpha^\delta, \mathbf{y}^\delta) = \mathbf{0}$  and taking into account that

$$\mathbf{g}_\alpha(\mathbf{x}^\dagger, \mathbf{y}) = \alpha \mathbf{L}^T \mathbf{L} (\mathbf{x}^\dagger - \mathbf{x}_a), \quad \frac{\partial \mathbf{g}_\alpha}{\partial \mathbf{x}}(\mathbf{x}_\alpha^\delta, \mathbf{y}^\delta) = \mathbf{G}_\alpha, \quad \frac{\partial \mathbf{g}_\alpha}{\partial \mathbf{y}}(\mathbf{x}_\alpha^\delta, \mathbf{y}^\delta) = -\mathbf{K}_\alpha^T,$$

we find that

$$\mathbf{x}^\dagger - \mathbf{x}_a^\delta = \alpha \mathbf{G}_\alpha^{-1} \mathbf{L}^T \mathbf{L} (\mathbf{x}^\dagger - \mathbf{x}_a) - \mathbf{G}_\alpha^{-1} \mathbf{K}_\alpha^T \boldsymbol{\delta} - \mathbf{G}_\alpha^{-1} \mathbf{R}(\mathbf{x}^\dagger, \mathbf{y}; \mathbf{x}_\alpha^\delta, \mathbf{y}^\delta). \quad (6.55)$$

Finally, employing the approximation

$$\mathbf{G}_\alpha^{-1} (\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{L}^T \mathbf{L}) \approx \mathbf{I}_n,$$

we obtain the expressions of the smoothing and noise errors as in (6.53) and (6.54), respectively.

If instead of the Newton method, the quasi-Newton method is used to compute a minimizer of the Tikhonov function, an additional step involving the calculation of the Hessian at the solution has to be performed. The reason is that the quasi-Newton approximation  $\bar{\mathbf{Q}}(\mathbf{x}_{\alpha k}^\delta)$  is a very crude estimate of the second-order derivative term  $\mathbf{Q}(\mathbf{x}_{\alpha k}^\delta)$ , which is not even certain to converge to the true  $\mathbf{Q}(\mathbf{x}_\alpha^\delta)$  as  $\mathbf{x}_{\alpha k}^\delta$  approaches  $\mathbf{x}_\alpha^\delta$ . For the Hessian calculation, we consider the Taylor expansion of the Tikhonov function about  $\mathbf{x}_\alpha^\delta$ ,

$$\mathcal{F}_\alpha(\mathbf{x}) \approx \mathcal{F}_\alpha(\mathbf{x}_\alpha^\delta) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_\alpha^\delta)^T \mathbf{G}_\alpha(\mathbf{x} - \mathbf{x}_\alpha^\delta), \quad (6.56)$$

where by definition, the entries of the Hessian are given by

$$[\mathbf{G}_\alpha]_{ij} = \frac{\partial^2 \mathcal{F}_\alpha}{\partial [\mathbf{x}]_i \partial [\mathbf{x}]_j}(\mathbf{x}_\alpha^\delta). \quad (6.57)$$

Equations (6.56) and (6.57) suggest that we may use finite differences for computing  $\mathbf{G}_\alpha$ . Denoting by  $\Delta x_i$  the displacement in the  $i$ th component of  $\mathbf{x}$ , we calculate the diagonal entries of  $\mathbf{G}_\alpha$  by using (6.56), that is,

$$[\mathbf{G}_\alpha]_{ii} = 2 \frac{\mathcal{F}_\alpha([\mathbf{x}_\alpha^\delta]_i + \Delta x_i) - \mathcal{F}_\alpha([\mathbf{x}_\alpha^\delta]_i)}{(\Delta x_i)^2}, \quad (6.58)$$

and the off-diagonal entries by using (6.57) with central differences, that is,

$$\begin{aligned} [\mathbf{G}_\alpha]_{ij} = & \left[ \mathcal{F}_\alpha([\mathbf{x}_\alpha^\delta]_i + \Delta x_i, [\mathbf{x}_\alpha^\delta]_j + \Delta x_j) - \mathcal{F}_\alpha([\mathbf{x}_\alpha^\delta]_i - \Delta x_i, [\mathbf{x}_\alpha^\delta]_j + \Delta x_j) \right. \\ & \left. - \mathcal{F}_\alpha([\mathbf{x}_\alpha^\delta]_i + \Delta x_i, [\mathbf{x}_\alpha^\delta]_j - \Delta x_j) + \mathcal{F}_\alpha([\mathbf{x}_\alpha^\delta]_i - \Delta x_i, [\mathbf{x}_\alpha^\delta]_j - \Delta x_j) \right] \\ & / (4\Delta x_i \Delta x_j). \end{aligned} \quad (6.59)$$

In (6.58) and (6.59) only the relevant arguments of the Tikhonov function are indicated; the omitted arguments remain unchanged during the calculation.

The computation of the Hessian by using finite differences requires an adequate choice of the step sizes  $\Delta x_i$ . The difficulty associated with the step size selection stems from the fact that in the  $\mathbf{x}$ -space, the Tikhonov function may vary slowly in some directions and rapidly in other. Small step sizes have to be used in steep directions of the Tikhonov function and large step sizes in flat directions. The iterative Algorithm 9 which significantly improves the reliability of the Hessian matrix calculation has been proposed by Pumplin et al. (2001). The method is based on the following result: if  $\mathbf{G}_\alpha$  is the exact Hessian with the singular value decomposition  $\mathbf{G}_\alpha = \mathbf{V}_g \Sigma_g \mathbf{V}_g^T$ , then the linear transformation

$$\mathbf{x} = \mathbf{V}_g \Sigma_g^{-\frac{1}{2}} \mathbf{z}, \quad (6.60)$$

implies that in the  $\mathbf{z}$ -space, the surface of constant  $\mathcal{F}_\alpha$ -values is a sphere, i.e.,

$$\mathcal{F}_\alpha(\mathbf{z}) - \mathcal{F}_\alpha(\mathbf{z}_\alpha^\delta) = \frac{1}{2}(\mathbf{z} - \mathbf{z}_\alpha^\delta)^T (\mathbf{z} - \mathbf{z}_\alpha^\delta). \quad (6.61)$$

The computation of the pseudo-Hessian  $\Phi$  in Algorithm 9 is performed in the  $\mathbf{z}$ -space by using (6.58) and (6.59), and this process is more stable than a Hessian calculation in the  $\mathbf{x}$ -space. The step sizes  $\Delta z_i$  are chosen so that the variations  $\mathcal{F}_\alpha([\mathbf{z}_\alpha^\delta]_i + \Delta z_i) - \mathcal{F}_\alpha([\mathbf{z}_\alpha^\delta]_i)$  in (6.58) are approximately equal to one.

**Algorithm 9.** Iterative algorithm for Hessian calculation.

---

```

compute the Hessian approximation  $\mathbf{G}_\alpha = \mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{L}^T \mathbf{L}$  at  $\mathbf{x}_\alpha^\delta$ ;
stop  $\leftarrow$  false;
while stop = false do
    compute the SVD  $\mathbf{G}_\alpha = \mathbf{V}_g \Sigma_g \mathbf{V}_g^T$ ;
     $\mathbf{T} \leftarrow \Sigma_g^{1/2} \mathbf{V}_g^T$ ;
     $\mathbf{z}_\alpha^\delta \leftarrow \mathbf{T} \mathbf{x}_\alpha^\delta$ ;
    compute the pseudo-Hessian  $\Phi$  from
         $\mathcal{F}_\alpha(\mathbf{z}) - \mathcal{F}_\alpha(\mathbf{z}_\alpha^\delta) = 0.5 (\mathbf{z} - \mathbf{z}_\alpha^\delta)^T \Phi (\mathbf{z} - \mathbf{z}_\alpha^\delta)$ ;
    if  $\Phi \approx \mathbf{I}_n$  then
        stop  $\leftarrow$  true;
    else
         $\mathbf{G}_\alpha \leftarrow \mathbf{T}^T \Phi \mathbf{T}$ ;
    end if
end while

```

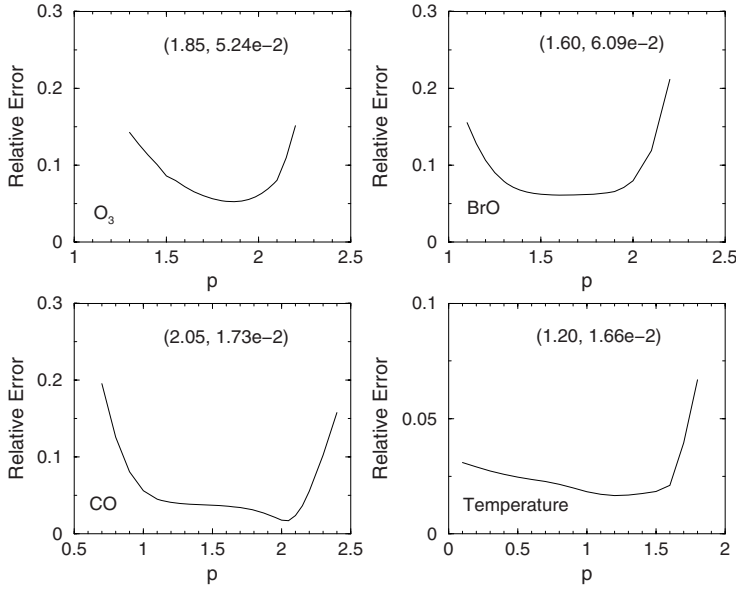
---

It should be pointed out that even though the Gauss–Newton method is used to compute a minimizer of the Tikhonov function, the error analysis can be performed by employing the Hessian approach. In atmospheric remote sensing with infrared spectroscopy, the benefit of computing the a posteriori covariance matrix by means of the Hessian method instead of the Gauss–Newton method has been evidenced by Tsidu (2005).

## 6.5 Regularization parameter choice methods

As for linear problems, the choice of the regularization parameter plays an important role in computing a reliable approximation of the solution. In this section we first extend the expected error estimation method to the nonlinear case. Then, we present selection criteria with variable and constant regularization parameters. In the first case, the regularization parameter is estimated at each iteration step, while in the second case, the minimization of the Tikhonov function is done a few times with different regularization parameters.

In order to judge the accuracy of parameter choice methods, we solve the retrieval test problems for various regularization parameters  $\alpha = \sigma^p$ , where  $\sigma$  is the noise standard deviation. The solution errors  $\|\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta\| / \|\mathbf{x}^\dagger\|$  for different values of the exponent  $p$  and for a single realization of the noisy data vector are illustrated in Figure 6.6. The plots show that all error curves possess a minimum, and by convention, the minimizers of the solution errors represent the optimal values of the regularization parameter. For the  $\text{O}_3$  and the CO retrieval test problems, the minima are relatively sharp, while for the BrO and the temperature retrieval test problems, the minima are flat. The latter situation is beneficial for the inversion process, because acceptable solutions correspond to a large domain of variation of the regularization parameter. The accuracy of a parameter choice method will be estimated by comparing the predicted value of the regularization parameter with the optimal value.



**Fig. 6.6.** Relative solution errors for different values of the exponent  $p$ , where  $\alpha = \sigma^p$  and  $\sigma$  is the noise standard deviation. The numbers in parentheses indicate the minimizer  $p_{\text{opt}}$  and the minimum value of the relative solution error  $\varepsilon_{\text{opt}}$ .

### 6.5.1 A priori parameter choice methods

In the linear case, the expected error estimation method has been formulated as an a priori parameter selection criterion. The idea was to perform a random exploration of a domain in which the solution is supposed to lie, and for each state vector realization  $\mathbf{x}_i^\dagger$ , to compute the optimal regularization parameter for error estimation

$$\bar{\alpha}_{\text{opt}i} = \arg \min_{\alpha} \mathcal{E} \left\{ \left\| \mathbf{e}_{\alpha}^{\delta} (\mathbf{x}_i^{\dagger}) \right\|^2 \right\},$$

and the exponent  $p_i = \log \bar{\alpha}_{\text{opt}i} / \log \sigma$ . The regularization parameter is then chosen as  $\alpha_e = \sigma^{\bar{p}}$ , where

$$\bar{p} = \frac{1}{N_x} \sum_{i=1}^{N_x} p_i$$

is the sample mean exponent and  $N_x$  is the sample size.

The expected error estimation method can be formulated for nonlinear problems, by representing the expected error at the solution as

$$\mathcal{E} \left\{ \left\| \mathbf{e}_{\alpha}^{\delta} \right\|^2 \right\} = \left\| \mathbf{e}_{\text{s}\alpha} \right\|^2 + \mathcal{E} \left\{ \left\| \mathbf{e}_{\text{n}\alpha}^{\delta} \right\|^2 \right\},$$

with

$$\mathbf{e}_{\text{s}\alpha} = (\mathbf{I}_n - \mathbf{A}_{\alpha}) (\mathbf{x}^{\dagger} - \mathbf{x}_a) = \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha} [\hat{\mathbf{w}}_i^T (\mathbf{x}^{\dagger} - \mathbf{x}_a)] \mathbf{w}_i \quad (6.62)$$

**Table 6.3.** Exponent  $p$  of the regularization parameter and relative errors in the Tikhonov solutions computed with the expected error estimation method.

Problem	$p$	$p_{\text{opt}}$	$\varepsilon$	$\varepsilon_{\text{opt}}$
O <sub>3</sub>	1.75	1.85	5.56e-2	5.24e-2
BrO	1.62	1.60	6.15e-2	6.09e-2
CO	1.35	2.05	3.84e-2	1.73e-2
Temperature	1.23	1.20	1.67e-2	1.66e-2

and

$$\mathcal{E} \left\{ \|\mathbf{e}_{n\alpha}^\delta\|^2 \right\} = \sigma^2 \text{trace} (\mathbf{K}_\alpha^\dagger \mathbf{K}_\alpha^{\dagger T}) = \sigma^2 \sum_{i=1}^n \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)^2 \|\mathbf{w}_i\|^2. \quad (6.63)$$

In (6.62) and (6.63),  $\gamma_i$  are the generalized singular values of the matrix pair  $(\mathbf{K}_\alpha, \mathbf{L})$ ,  $\mathbf{w}_i$  is the  $i$ th column vector of the nonsingular matrix  $\mathbf{W}$ , and  $\hat{\mathbf{w}}_i^T$  is the  $i$ th row vector of the matrix  $\hat{\mathbf{W}} = \mathbf{W}^{-1}$ . Because the Jacobian matrix  $\mathbf{K}_\alpha$  is evaluated at the solution, the generalized singular system depends on  $\alpha$ , and as a result, the optimal regularization parameter for error estimation has to be computed for each state vector realization by repeatedly solving the nonlinear minimization problem. The resulting algorithm is extremely computationally expensive and in order to ameliorate this drawback, we approximate the Jacobian matrix at the solution by the Jacobian matrix at the a priori state. This is a realistic assumption for problems with a small degree of nonlinearity. The a priori parameter choice method is then equivalent to the expected error estimation method applied to a linearization of the forward model about the a priori state.

The solution errors shown in Table 6.3 demonstrate that the expected error estimation method yields accurate results except for the CO retrieval test problem. In this case, the algorithm identifies a substantially smaller regularization parameter, but the solution error is still acceptable. The retrieved profiles are illustrated in Figure 6.7 together with the results obtained by using the Bayesian estimate  $p = 2$ . For the temperature retrieval test problem, the Bayesian estimate yields an undersmoothed profile with large oscillations around the exact profile.

---

**Algorithm 10.** Iterated expected error estimation method.

---

choose initial  $\bar{\alpha}$ ;

**for**  $i = 1, N_{\text{iter}}$  **do**

    compute the Tikhonov solution of parameter  $\bar{\alpha}$ ,  $\mathbf{x}_{\bar{\alpha}}^\delta$ ;

    compute the GSVD  $\mathbf{K}_{\bar{\alpha}} = \mathbf{U}\Sigma_1\mathbf{W}^{-1}$  and  $\mathbf{L} = \mathbf{V}\Sigma_2\mathbf{W}^{-1}$ ;

    compute  $\bar{\alpha}_{\text{opt}} = \arg \min_{\alpha} \mathcal{E} \left\{ \|\mathbf{e}_{n\alpha}^\delta\|^2 \right\}$ , with

$$\mathbf{e}_{s\alpha} = \sum_{i=1}^n \frac{\alpha}{\gamma_i^2 + \alpha} [\hat{\mathbf{w}}_i^T (\mathbf{x}_{\bar{\alpha}}^\delta - \mathbf{x}_a)] \mathbf{w}_i \text{ and}$$

$$\mathcal{E} \left\{ \|\mathbf{e}_{n\alpha}^\delta\|^2 \right\} = \sigma^2 \sum_{i=1}^n \left( \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \frac{1}{\sigma_i} \right)^2 \|\mathbf{w}_i\|^2;$$

**if**  $|\bar{\alpha}_{\text{opt}} - \bar{\alpha}| < \text{tol}$  **then**

**exit**;

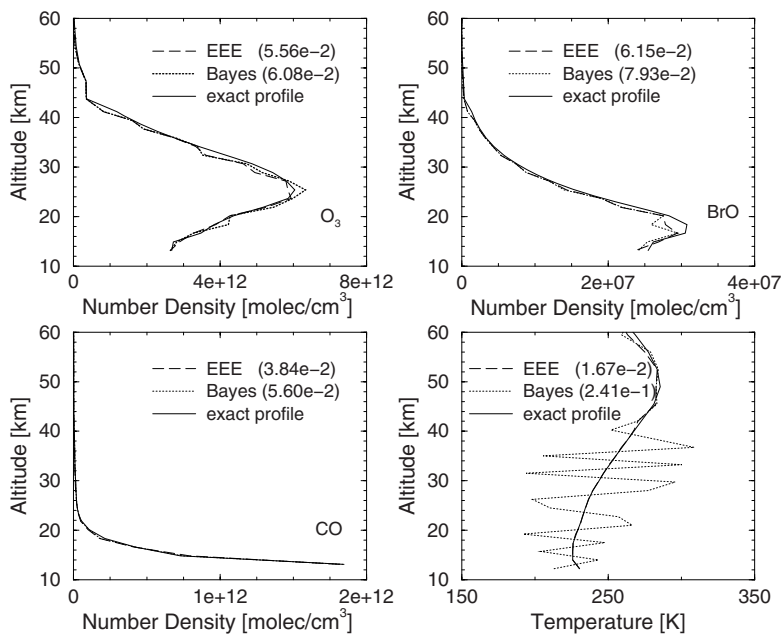
**else**

$$\bar{\alpha} \leftarrow \bar{\alpha}_{\text{opt}};$$

**end if**

**end for**

---



**Fig. 6.7.** Tikhonov solutions computed with the expected error estimation (EEE) method and the Bayesian estimate  $p = 2$ . The numbers in parentheses indicate the relative solution errors.

Another version of the expected error estimation method can be designed by assuming a semi-stochastic error representation and by using an iterative algorithm for minimizing the expected error (Algorithm 10). Two main drawbacks reduce the performance of the so-called iterated expected error estimation method:

- (1) the semi-stochastic error representation is valid if the regularization parameter lies in the neighborhood of the optimal regularization parameter, and for this reason, the solution strongly depends on the initialization;
- (2) the minimizer of the expected error is in general larger than the optimal regularization parameter (see Figure 6.3).

The results shown in Table 6.4 demonstrate that for all test problems, the retrieved profiles are oversmoothed.

**Table 6.4.** Regularization parameters and relative errors in the Tikhonov solutions computed with the iterated expected error estimation method. The numbers in parentheses indicate the exponent of the regularization parameter.

Problem	$\alpha$	$\alpha_{\text{opt}}$	$\varepsilon$	$\varepsilon_{\text{opt}}$
O <sub>3</sub>	6.14e-5 (1.67)	2.09e-5 (1.85)	6.35e-2	5.24e-2
BrO	6.20e-6 (1.31)	3.84e-7 (1.60)	7.89e-2	6.09e-2
CO	2.36e-4 (1.28)	1.17e-6 (2.05)	3.97e-2	1.73e-2
Temperature	2.39e-5 (1.18)	1.41e-5 (1.20)	1.68e-2	1.66e-2

### 6.5.2 Selection criteria with variable regularization parameters

As the solution of a nonlinear ill-posed problem by means of Tikhonov regularization is equivalent to the solution of a sequence of ill-posed linearizations of the forward model about the current iterate, parameter choice methods for linear problems can be used to compute the regularization parameter at each iteration step.

The errors in the right-hand side of the linearized equation (6.34) are due to the instrumental noise and the linearization error. Because the linearization error cannot be estimated, we propose a heuristic version of the discrepancy principle as follows: at the iteration step  $k$ , compute the regularization parameter as the solution of the equation

$$\|\mathbf{r}_{1\alpha k}^\delta\|^2 = \tau \|\mathbf{r}_{1\min k}^\delta\|^2, \quad \tau > 1,$$

where  $\mathbf{r}_{1\alpha k}^\delta$  is the linearized residual vector,

$$\mathbf{r}_{1\alpha k}^\delta = (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha k}) \mathbf{y}_k^\delta,$$

$\hat{\mathbf{A}}_{\alpha k} = \mathbf{K}_{\alpha k} \mathbf{K}_{\alpha k}^\dagger$  is the influence matrix, and  $\|\mathbf{r}_{1\min k}^\delta\|$  is the minimum value of  $\|\mathbf{r}_{1\alpha k}^\delta\|$  corresponding to the smallest generalized singular value of  $(\mathbf{K}_{\alpha k}, \mathbf{L})$ .

Due to the difficulties associated with the data error estimation, error-free parameter choice methods (based only on information about the noisy data) are more attractive. In this context, we mention that the generalized cross-validation method has been applied to the linearized equation (6.34) by Haber (1997), Haber and Oldenburg (2000), and Farquharson and Oldenburg (2004). Selection of the regularization parameter by using the L-curve criterion has been reported by Schimpf and Schreier (1997), Li and Oldenburg (1999), Farquharson and Oldenburg (2004), and Hasekamp and Landgraf (2001). In our retrieval algorithm, we use the following regularization parameter choice methods:

- (1) the generalized cross-validation method,

$$\alpha_{\text{gcv}k} = \arg \min_{\alpha} v_{\alpha k}^\delta,$$

with

$$v_{\alpha k}^\delta = \frac{\|\mathbf{r}_{1\alpha k}^\delta\|^2}{\left[\text{trace}(\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha k})\right]^2}, \quad (6.64)$$

- (2) the maximum likelihood estimation,

$$\alpha_{\text{mle}k} = \arg \min_{\alpha} \lambda_{\alpha k}^\delta,$$

with

$$\lambda_{\alpha k}^\delta = \frac{\mathbf{y}_k^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha k}) \mathbf{y}_k^\delta}{\sqrt[m]{\det(\mathbf{I}_m - \hat{\mathbf{A}}_{\alpha k})}}, \quad (6.65)$$



(3) the L-curve method,

$$\alpha_{1ck} = \arg \max_{\alpha} \kappa_{1c\alpha k}^{\delta},$$

with

$$\kappa_{1c\alpha k}^{\delta} = \frac{x_k''(\alpha) y_k'(\alpha) - x_k'(\alpha) y_k''(\alpha)}{\left[ x_k'(\alpha)^2 + y_k'(\alpha)^2 \right]^{\frac{3}{2}}}$$

and

$$x_k(\alpha) = \log \left( \left\| \mathbf{r}_{1\alpha k}^{\delta} \right\|^2 \right), \quad y_k(\alpha) = \log \left( \left\| \mathbf{c}_{\alpha k}^{\delta} \right\|^2 \right).$$

Note that in the L-curve method, the constraint vector is computed for each value of the regularization parameter by using the relation  $\mathbf{c}_{\alpha k}^{\delta} = \mathbf{L}\mathbf{K}_{\alpha k}^{\dagger} \mathbf{y}_k^{\delta}$ .

In practice, the following recommendations for choosing the regularization parameter have to be taken into account:

- (1) at the beginning of the iterative process, large  $\alpha$ -values should be used to avoid local minima and to get well-conditioned linear problems to solve;
- (2) during the iteration, the regularization parameter should be decreased slowly to achieve a stable solution.

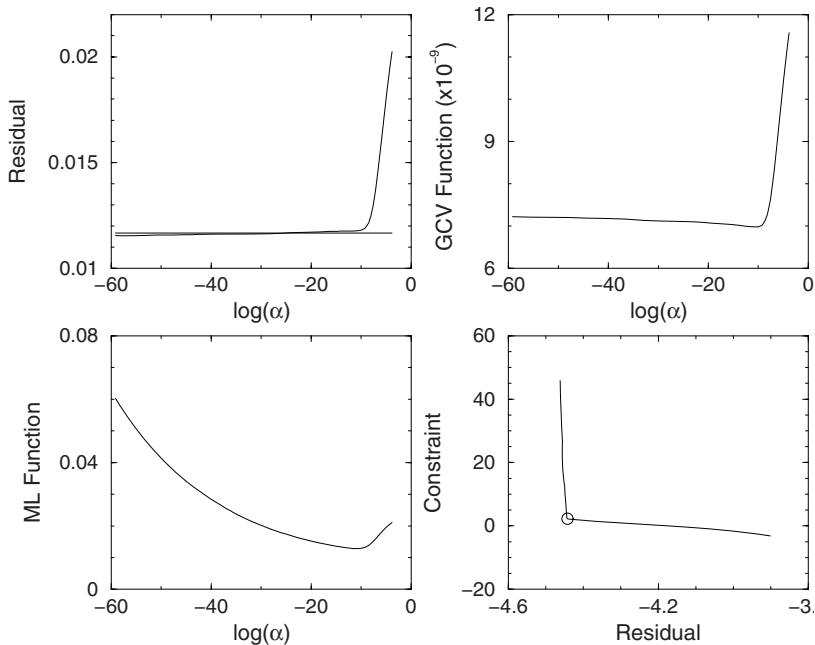
Numerical experiments have shown that a brutal use of the regularization parameter computed by one of the above parameter choice methods may lead to an oscillation sequence of  $\alpha$ -values. A heuristic formula that deals with this problem has been proposed by Eriksson (1996): at the iteration step  $k$ , the regularization parameter  $\alpha_k$  is the weighted sum between the previous regularization parameter  $\alpha_{k-1}$  and the regularization parameter  $\alpha$  computed by one of the above parameter choice methods, that is,

$$\alpha_k = \begin{cases} \xi \alpha_{k-1} + (1 - \xi) \alpha, & \alpha < \alpha_{k-1}, \\ \alpha_{k-1}, & \alpha \geq \alpha_{k-1}, \end{cases}$$

with  $0 < \xi < 1$  being a priori chosen. This selection rule guarantees a descending sequence of regularization parameters, and the resulting method is very similar to the iteratively regularized Gauss–Newton method to be discussed in the next chapter.

For the  $O_3$  retrieval test problem, the residual and the L-curves, as well as the generalized cross-validation and the maximum likelihood functions are shown in Figure 6.8. The curves have the same behaviors as in the linear case: the generalized cross-validation function has a flat minimum, the maximum likelihood function has a distinct minimum, and the L-curve has a sharp corner.

The solution errors listed in Table 6.5 show that Tikhonov regularization with variable regularization parameter yields accurate results, and that the maximum likelihood estimation is superior to the other regularization parameter choice methods.



**Fig. 6.8.** Residual curve, generalized cross-validation (GCV) function, maximum likelihood (ML) function and L-curve for the  $O_3$  retrieval test problem. The curves are computed at the first iteration step.

**Table 6.5.** Relative solution errors for Tikhonov regularization with variable regularization parameters corresponding to the following selection criteria: the discrepancy principle (DP), the maximum likelihood estimation (MLE), generalized cross-validation (GCV), and the L-curve (LC) method.

Problem	Method	$\varepsilon$	$\varepsilon_{\text{opt}}$
$O_3$	DP	6.01e-2	5.24e-2
	MLE	5.24e-2	
	GCV	5.37e-2	
	LC	5.64e-2	
BrO	DP	6.11e-2	6.09e-2
	MLE	6.26e-2	
	GCV	6.28e-2	
	LC	6.22e-2	
CO	DP	3.42e-2	1.73e-2
	MLE	2.08e-2	
	GCV	2.55e-2	
	LC	3.66e-2	
Temperature	DP	1.82e-2	1.66e-2
	MLE	1.66e-2	
	GCV	1.67e-2	
	LC	2.22e-2	

### 6.5.3 Selection criteria with constant regularization parameters

The numerical realization of these parameter choice methods requires us to solve the nonlinear minimization problem several times for different regularization parameters. Each minimization is solved with a regularization parameter  $\alpha$  and a solution  $\mathbf{x}_\alpha^\delta$  is obtained. If the solution is satisfactory as judged by these selection criteria, then the inverse problem is considered to be solved. The discrete values of the regularization parameters are chosen as  $\alpha_i = \sigma^{p_i}$ , where  $\{p_i\}$  is an increasing sequence of positive numbers. Since  $\sigma < 1$ , the sequence of regularization parameters  $\{\alpha_i\}$  is then in decreasing order.

In the framework of the discrepancy principle, the regularization parameter is the solution of the equation

$$\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 = \tau \Delta^2, \quad (6.66)$$

with  $\tau > 1$ . Because, for nonlinear problems, the discrepancy principle equation only has a solution under very strong restrictive assumptions (Kravaris and Seinfeld, 1985), we use a simplified version of this selection criterion: if  $\{\alpha_i\}$  is a decreasing sequence of regularization parameters, we choose the largest  $\alpha_{i^*}$  such that the residual norm is below the noise level, that is,

$$\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\alpha_{i^*}}^\delta)\|^2 \leq \tau \Delta^2 < \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\alpha_i}^\delta)\|^2, \quad 0 \leq i < i^*.$$

Note that this version of the discrepancy principle is typical for iterative regularization methods.

The generalized discrepancy principle can also be formulated as an a posteriori parameter choice method for the nonlinear Tikhonov regularization. A heuristic justification of this regularization parameter choice method can be given in a deterministic setting by using the error estimate

$$\|\mathbf{e}_\alpha^\delta\|^2 \leq 2 \left( \|\mathbf{e}_{s\alpha}\|^2 + \|\mathbf{e}_{n\alpha}^\delta\|^2 \right),$$

together with the noise error bound (3.99),

$$\|\mathbf{e}_{n\alpha}^\delta\|^2 < \frac{2\tau\Delta^2}{\alpha}, \quad \tau > 1.$$

To estimate the smoothing error we assume  $\mathbf{L} = \mathbf{I}_n$ , and consider the unperturbed solution  $\mathbf{x}_\alpha$  corresponding to the exact data vector  $\mathbf{y}$ . The stationary condition for the Tikhonov function at  $\mathbf{x}_\alpha$  yields

$$\mathbf{K}_\alpha^T [\mathbf{F}(\mathbf{x}_\alpha) - \mathbf{y}] + \alpha (\mathbf{x}_\alpha - \mathbf{x}_a) = \mathbf{0}, \quad (6.67)$$

with  $\mathbf{K}_\alpha = \mathbf{K}(\mathbf{x}_\alpha)$ . Employing the same arguments as in the derivation of (6.48), we obtain

$$\mathbf{e}_{s\alpha} = (\mathbf{I}_n - \mathbf{A}_\alpha) (\mathbf{x}^\dagger - \mathbf{x}_a),$$

with the averaging kernel matrix  $\mathbf{A}_\alpha$  being evaluated at  $\mathbf{x}_\alpha$ . Taking into account that for any  $\mathbf{x}$ , there holds

$$\|(\mathbf{I}_n - \mathbf{A}_\alpha) \mathbf{x}\|^2 = \sum_{i=1}^n \left( \frac{\alpha}{\sigma_i^2 + \alpha} \right)^2 (\mathbf{v}_i^T \mathbf{x})^2 \leq \sum_{i=1}^n (\mathbf{v}_i^T \mathbf{x})^2 = \|\mathbf{x}\|^2,$$

we deduce that a bound for the total error is given by

$$M(\alpha) = 4 \left( \frac{1}{2} \|\mathbf{x}^\dagger - \mathbf{x}_\alpha\|^2 + \tau \frac{\Delta^2}{\alpha} \right).$$

To derive the necessary condition for a minimum of the estimate  $M(\alpha)$ , we consider the function

$$f(\alpha) = \frac{1}{2} \|\mathbf{x}^\dagger - \mathbf{x}_\alpha\|^2, \quad (6.68)$$

and compute the derivative

$$f'(\alpha) = -(\mathbf{x}^\dagger - \mathbf{x}_\alpha)^T \frac{d\mathbf{x}_\alpha}{d\alpha}. \quad (6.69)$$

Formal differentiation of the Euler equation (6.67) with respect to  $\alpha$  yields

$$\frac{d\mathbf{K}_\alpha^T}{d\alpha} [\mathbf{F}(\mathbf{x}_\alpha) - \mathbf{y}] + \mathbf{K}_\alpha^T \mathbf{K}_\alpha \frac{d\mathbf{x}_\alpha}{d\alpha} + \alpha \frac{d\mathbf{x}_\alpha}{d\alpha} = -(\mathbf{x}_\alpha - \mathbf{x}_a); \quad (6.70)$$

whence, neglecting the first term in the left-hand side of (6.70) and using (6.67), we obtain

$$\frac{d\mathbf{x}_\alpha}{d\alpha} \approx -(\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{I}_n)^{-1} (\mathbf{x}_\alpha - \mathbf{x}_a) = -\frac{1}{\alpha} \mathbf{K}_\alpha^\dagger [\mathbf{y} - \mathbf{F}(\mathbf{x}_\alpha)].$$

The linear approximation

$$\mathbf{y} \approx \mathbf{F}(\mathbf{x}_\alpha) + \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha)$$

and the matrix identity

$$(\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{I}_n)^{-1} \mathbf{K}_\alpha^T = \mathbf{K}_\alpha^T (\mathbf{K}_\alpha \mathbf{K}_\alpha^T + \alpha \mathbf{I}_m)^{-1},$$

then give

$$\begin{aligned} f'(\alpha) &\approx \frac{1}{\alpha} (\mathbf{x}^\dagger - \mathbf{x}_\alpha)^T \mathbf{K}_\alpha^\dagger [\mathbf{y} - \mathbf{F}(\mathbf{x}_\alpha)] \\ &= \frac{1}{\alpha} [\mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha)]^T (\mathbf{K}_\alpha \mathbf{K}_\alpha^T + \alpha \mathbf{I}_m)^{-1} [\mathbf{y} - \mathbf{F}(\mathbf{x}_\alpha)] \\ &\approx \frac{1}{\alpha} [\mathbf{y} - \mathbf{F}(\mathbf{x}_\alpha)]^T (\mathbf{K}_\alpha \mathbf{K}_\alpha^T + \alpha \mathbf{I}_m)^{-1} [\mathbf{y} - \mathbf{F}(\mathbf{x}_\alpha)]. \end{aligned}$$

Setting  $M'(\alpha) = 0$  and replacing  $\mathbf{x}_\alpha$  by  $\mathbf{x}_\alpha^\delta$  and  $\mathbf{y}$  by  $\mathbf{y}^\delta$ , we obtain the generalized discrepancy principle equation in the form (see (3.98))

$$\alpha [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)]^T (\mathbf{K}_\alpha \mathbf{K}_\alpha^T + \alpha \mathbf{I}_m)^{-1} [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)] = \tau \Delta^2.$$

Error-free methods with constant regularization parameter are natural extensions of the corresponding selection criteria for linear problems; the most popular are the maximum likelihood estimation, generalized cross-validation and the nonlinear L-curve method.

Applications of generalized cross-validation in conjunction with the method of Tikhonov regularization for solving a temperature retrieval problem and an inverse scattering

problem have been reported by O'Sullivan and Wahba (1985) and Vogel (1985), respectively. To formulate the generalized cross-validation method and the maximum likelihood estimation, we employ some heuristic arguments, while for a more rigorous treatment we refer to O'Sullivan and Wahba (1985). At the iteration step  $k$ , the generalized cross-validation function  $v_{\alpha k}^\delta$  and the maximum likelihood function  $\lambda_{\alpha k}^\delta$ , given by (6.64) and (6.65), respectively, depend on the influence matrix  $\hat{\mathbf{A}}_{\alpha k}$ , the linearized residual  $\mathbf{r}_{1\alpha k}^\delta$  and the noisy data vector  $\mathbf{y}_k^\delta$ . If the iterates  $\mathbf{x}_{\alpha k}^\delta$  converge to  $\mathbf{x}_\alpha^\delta$  and  $\mathbf{F}$  is continuously differentiable, then we may assume that  $\hat{\mathbf{A}}_{\alpha k}$  converges to the influence matrix at the solution  $\hat{\mathbf{A}}_\alpha = \mathbf{K}_\alpha \mathbf{K}_\alpha^\dagger, \mathbf{r}_{1\alpha}^\delta$  to the nonlinear residual  $\mathbf{r}_\alpha^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)$  and  $\mathbf{y}_k^\delta$  to

$$\mathbf{y}_\alpha^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) + \mathbf{K}_\alpha (\mathbf{x}_\alpha^\delta - \mathbf{x}_a).$$

Thus, as  $k \rightarrow \infty$ , the generalized cross-validation and the maximum likelihood functions become

$$v_\alpha^\delta = \frac{\|\mathbf{r}_\alpha^\delta\|^2}{\left[\text{trace}(\mathbf{I}_m - \hat{\mathbf{A}}_\alpha)\right]^2},$$

and

$$\lambda_\alpha = \frac{\mathbf{y}_\alpha^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}}_\alpha) \mathbf{y}_\alpha^\delta}{\sqrt[m]{\det(\mathbf{I}_m - \hat{\mathbf{A}}_\alpha)}},$$

respectively.

The use of the L-curve for nonlinear problems has been suggested by Eriksson (1996). The nonlinear L-curve is the plot of the constraint  $\|\mathbf{c}_\alpha^\delta\|^2 = \|\mathbf{L}(\mathbf{x}_\alpha^\delta - \mathbf{x}_a)\|^2$  against the residual  $\|\mathbf{r}_\alpha^\delta\|^2 = \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2$  for a range of values of the regularization parameter  $\alpha$ . This curve is monotonically decreasing and convex as shown by Gulliksson and Wedin (1999). In a computational sense, the nonlinear L-curve consists of a number of discrete points corresponding to the different values of the regularization parameter and in practice, the following techniques can be used for choosing the regularization parameter:

- (1) As for iterative regularization methods, we fit a cubic spline curve to the discrete points of the L-curve  $(x(\alpha_i), y(\alpha_i))$ , with  $x(\alpha) = \log(\|\mathbf{r}_\alpha^\delta\|^2)$  and  $y(\alpha) = \log(\|\mathbf{c}_\alpha^\delta\|^2)$ , and determine the point on the original discrete curve that is closest to the spline curve's corner.
- (2) In the framework of the minimum distance function approach (Belge et al., 2002), we compute

$$\alpha_{1c} = \arg \min_i d(\alpha_i)^2$$

for the distance function

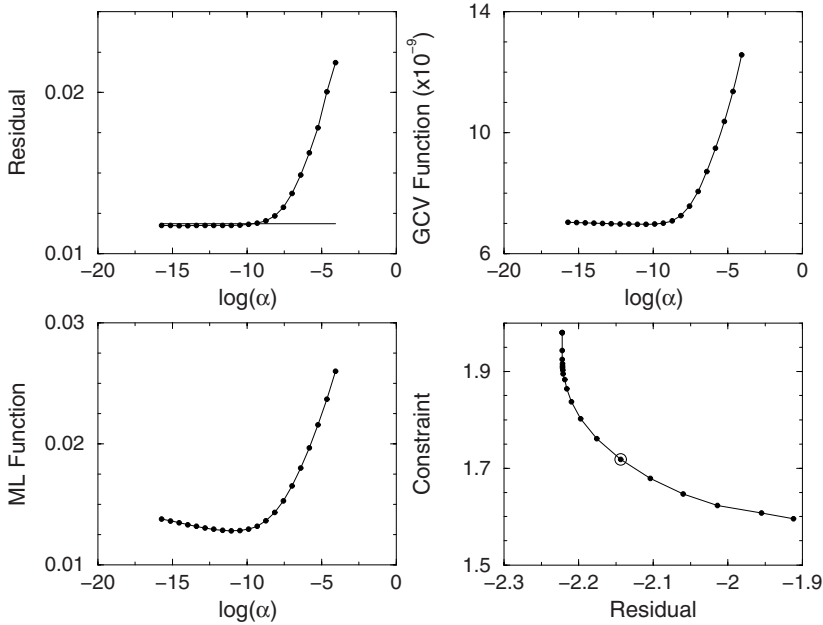
$$d(\alpha)^2 = [x(\alpha) - x_0]^2 + [y(\alpha) - y_0]^2,$$

with  $x_0 = \min_i x(\alpha_i)$  and  $y_0 = \min_i y(\alpha_i)$ .

- (3) Relying on the definition of the corner of the L-curve as given by Reginska (1996), we determine the regularization parameter as

$$\alpha_{1c} = \arg \min_i (x(\alpha_i) + y(\alpha_i)),$$

that is, we detect the minimum of the logarithmic L-curve rotated by  $\pi/4$  radians.



**Fig. 6.9.** Nonlinear residual curve, generalized cross-validation (GCV) function, maximum likelihood (ML) function and L-curve for the  $O_3$  retrieval test problem.

The curves corresponding to the nonlinear parameter choice methods with a constant regularization parameter are illustrated in Figure 6.9. The plots show that the maximum likelihood function has a sharper minimum than the generalized cross-validation function, and that the L-curve corner is not distinctive. The solution errors listed in Table 6.6 indicate that the best results correspond to the maximum likelihood estimation, and that the worst results correspond to the L-curve method. Especially noteworthy is the failure of the L-curve method for the  $O_3$  retrieval test problem: the predicted value of the regularization parameter is considerably larger than the optimal value, and the retrieved profile is close to the a priori (Figure 6.10).

## 6.6 Iterated Tikhonov regularization

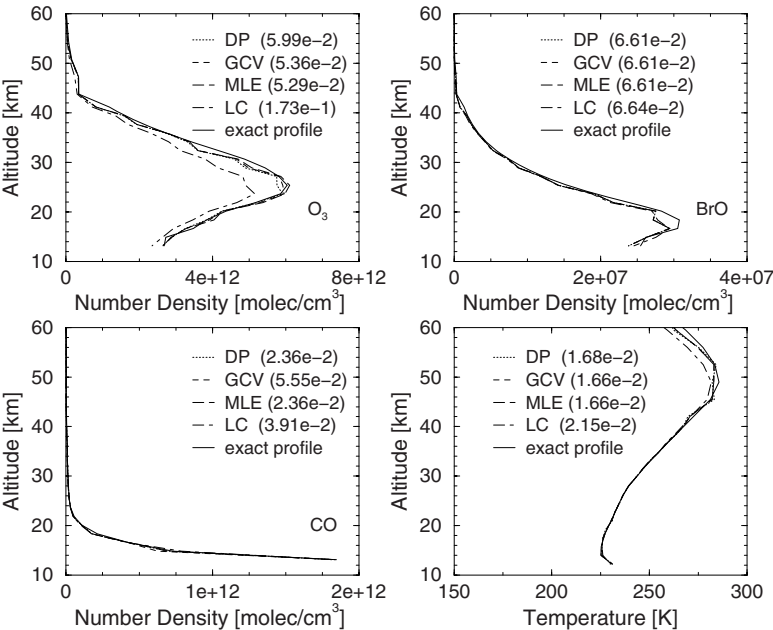
To obtain a higher convergence rate, iterated Tikhonov regularization has been considered for nonlinear ill-posed problems by Scherzer (1993), and Jin and Hou (1997). The  $p$ -times iterated Tikhonov regularization is defined inductively in the following way: the regularized solution at the first iteration step is the ordinary Tikhonov solution  $\mathbf{x}_{\alpha 1}^\delta = \mathbf{x}_\alpha^\delta$ , while the regularized solution  $\mathbf{x}_{\alpha p}^\delta$  at the iteration step  $p \geq 2$  minimizes the objective function

$$\mathcal{F}_{\alpha p}(\mathbf{x}) = \frac{1}{2} \left[ \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2 + \alpha \|\mathbf{L}(\mathbf{x} - \mathbf{x}_{\alpha p-1}^\delta)\|^2 \right].$$

Iterated Tikhonov regularization can also be used to improve the regularized solution

**Table 6.6.** Exponent  $p$  and relative solution errors for Tikhonov regularization with a constant regularization parameter corresponding to the discrepancy principle (DP), the maximum likelihood estimation (MLE), generalized cross-validation (GCV), and the L-curve (LC) method.

Problem	Method	$p$	$p_{\text{opt}}$	$\varepsilon$	$\varepsilon_{\text{opt}}$
O <sub>3</sub>	DP	1.7		5.99e-2	
	MLE	1.9	1.85	5.29e-2	5.24e-2
	GCV	1.8		5.36e-2	
	LC	1.2		1.73e-1	
BrO	DP	1.9		6.61e-2	
	MLE	1.9	1.60	6.61e-2	6.09e-2
	GCV	1.9		6.61e-2	
	LC	1.4		6.64e-2	
CO	DP	2.1		2.36e-2	
	MLE	2.1	2.05	2.36e-2	1.73e-2
	GCV	2.2		5.55e-2	
	LC	1.3		3.91e-2	
Temperature	DP	1.3		1.68e-2	
	MLE	1.2	1.20	1.66e-2	1.66e-2
	GCV	1.2		1.66e-2	
	LC	0.8		2.15e-2	



**Fig. 6.10.** Retrieval results for Tikhonov regularization with a constant regularization parameter computed by using the discrepancy principle (DP), the maximum likelihood estimation (MLE), generalized cross-validation (GCV), and the L-curve (LC) method.

of the linearized equation

$$\mathbf{K}_{\alpha k} \Delta \mathbf{x} = \mathbf{y}_k^\delta,$$

with  $\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_a$ . The solution refinement is based on the following defect iteration: at the iteration step  $l$ , the linear equation

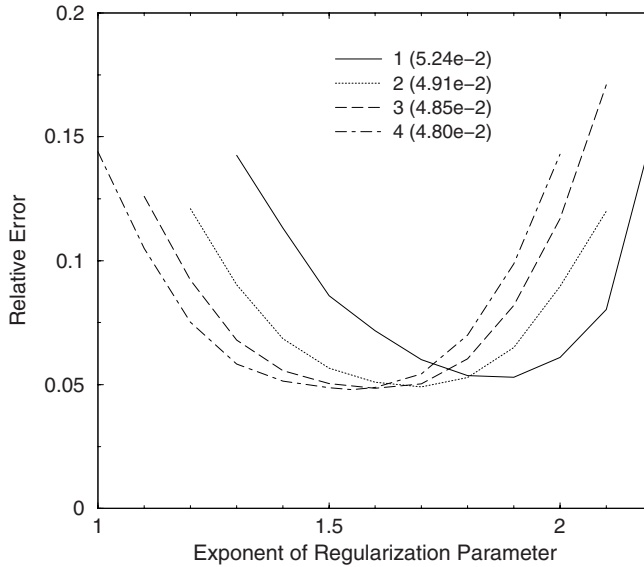
$$\mathbf{K}_{\alpha k} (\Delta \mathbf{x} - \Delta \mathbf{x}_{\alpha k l-1}^\delta) = \mathbf{y}_k^\delta - \mathbf{K}_{\alpha k} \Delta \mathbf{x}_{\alpha k l-1}^\delta$$

is solved by means of Tikhonov regularization with the penalty term  $\|\mathbf{L}(\Delta \mathbf{x} - \Delta \mathbf{x}_{\alpha k l-1}^\delta)\|^2$  and the regularization parameter  $\alpha$ . The algorithm for solution improvement then takes the form

$$\begin{aligned} \Delta \mathbf{x}_{\alpha k 0}^\delta &= \mathbf{0}, \\ \Delta \mathbf{x}_{\alpha k l}^\delta &= \Delta \mathbf{x}_{\alpha k l-1}^\delta + \mathbf{K}_{\alpha k}^\dagger (\mathbf{y}_k^\delta - \mathbf{K}_{\alpha k} \Delta \mathbf{x}_{\alpha k l-1}^\delta), \quad 1 \leq l \leq p, \\ \mathbf{x}_{\alpha k+1}^\delta &= \mathbf{x}_a + \Delta \mathbf{x}_{\alpha k p}^\delta. \end{aligned} \quad (6.71)$$

Essentially, this method consists of an outer Newton iteration for the nonlinear equation, and an inner iteration, the  $p$ -times iterated Tikhonov regularization for the linearized equation. The order of iterated Tikhonov regularization is a control parameter of the algorithm and must be chosen in advance.

The plots in Figure 6.11 show that by increasing the order of iterated Tikhonov regularization, the minimizer of the solution error also increases, the error curve becomes flatter, and the minimum solution error decreases.



**Fig. 6.11.** Relative errors in the iterated Tikhonov solution for the  $\text{O}_3$  retrieval test problem. The order of iterated Tikhonov regularization (the number of iteration steps of the inner scheme) varies between 1 and 4. The numbers in parentheses indicate the minimum value of the relative solution error.



## 6.7 Constrained Tikhonov regularization

Constrained versions of Tikhonov regularization can be developed by making use of additional information about the solution. For example, we may impose that on some layers  $i$ , the entries  $[\mathbf{x}]_i$  of the state vector  $\mathbf{x}$  are bounded,

$$l_i \leq [\mathbf{x}]_i \leq u_i,$$

in which case, the optimization problem involves the minimization of the Tikhonov function subject to simple bounds on the variables. In this section we introduce the constrained Tikhonov regularization by considering a practical example, namely the retrieval of ozone profiles from nadir sounding measurements performed by instruments such as GOME, SCIAMACHY, OMI and GOME-2. The constraints are imposed on the vertical column, which represents the integrated ozone profile. Thus, in this version of Tikhonov regularization, we control the smoothness of the profile through the regularization matrix and the magnitude of the profile through the vertical column. Only equality constraints will be the topic of the present analysis; the incorporation of inequality constraints into the iteratively regularized Gauss–Newton method will be the subject of the next chapter.

In order to simplify our presentation we assume that the entry  $[\mathbf{x}]_i$  of  $\mathbf{x}$  is the partial column of ozone on the layer  $i$ . The number of layers is  $n$  and the vertical column is then given by  $\sum_{i=1}^n [\mathbf{x}]_i$ . The layer  $i = 1$  is situated at the top of the atmosphere, while the layer  $i = n$  is situated at the Earth's surface. The main idea of formulating the equality-constrained Tikhonov regularization relies on the observation that the a priori profile deviation  $\Delta \mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_{\alpha k+1}^\delta - \mathbf{x}_a$  minimizing the Tikhonov function

$$\mathcal{F}_{1\alpha k}(\Delta \mathbf{x}) = \|\mathbf{y}_k^\delta - \mathbf{K}_{\alpha k} \Delta \mathbf{x}\|^2 + \alpha \|\mathbf{L} \Delta \mathbf{x}\|^2,$$

also minimizes the quadratic function

$$\mathcal{Q}(\Delta \mathbf{x}) = \mathbf{g}^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{G} \Delta \mathbf{x}, \quad (6.72)$$

with

$$\mathbf{G} = \mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k} + \alpha \mathbf{L}^T \mathbf{L}, \quad (6.73)$$

and

$$\mathbf{g} = -\mathbf{K}_{\alpha k}^T \mathbf{y}_k^\delta. \quad (6.74)$$

The equality-constrained Tikhonov regularization possesses the following formulation: at the iteration step  $k$ , compute the a priori profile deviation  $\Delta \mathbf{x}_{\alpha k+1}^\delta$  by solving the quadratic programming problem

$$\min_{\Delta \mathbf{x}} \mathcal{Q}(\Delta \mathbf{x}) = \mathbf{g}^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{G} \Delta \mathbf{x} \quad (6.75)$$

$$\text{subject to } \sum_{i=1}^n [\Delta \mathbf{x}]_i = c. \quad (6.76)$$

Here,  $c$  is the vertical column corresponding to  $\Delta \mathbf{x}$ , and by convention,  $c$  will be referred to as the relative vertical column with respect to the a priori.

For solving the quadratic programming problem (6.75)–(6.76), the null-space or the range-space methods can be employed (Gill et al., 1981; Nocedal and Wright, 2006). In the framework of the null-space method, the matrix  $\mathbf{Z} \in \mathbb{R}^{n \times (n-1)}$ , whose column vectors are a basis for the null space of the constraint matrix  $\mathbf{A} = [1, \dots, 1]$  (cf. (6.76)), plays an important role. In general, the matrix  $\mathbf{Z}$  can be computed by using the QR factorization of  $\mathbf{A}^T$ , or it can be derived by using the variable-reduction technique (Appendix J). In the present analysis we adopt the variable-reduction technique, in which case, the algorithm involves the following steps:

- (1) compute a feasible point satisfying the linear constraint, e.g.,

$$\Delta \bar{\mathbf{x}} = c \Delta \bar{\mathbf{x}}_n, \quad \Delta \bar{\mathbf{x}}_n = \frac{1}{n} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix};$$

- (2) compute the gradient of  $\mathcal{Q}$  at  $\Delta \bar{\mathbf{x}}$ ,

$$\bar{\mathbf{g}} = c \mathbf{g}_n + \mathbf{g}, \quad \mathbf{g}_n = \mathbf{G} \Delta \bar{\mathbf{x}}_n,$$

and construct the matrix  $\mathbf{Z}$  as

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \end{bmatrix} \in \mathbb{R}^{n \times (n-1)};$$

- (3) determine the feasible step

$$\mathbf{p} = -\mathbf{H} \bar{\mathbf{g}} = -c \mathbf{H} \mathbf{g}_n - \mathbf{H} \mathbf{g},$$

where

$$\mathbf{H} = \mathbf{Z} (\mathbf{Z}^T \mathbf{G} \mathbf{Z})^{-1} \mathbf{Z}^T$$

is the reduced inverse Hessian of  $\mathcal{Q}$  subject to the constraint;

- (4) compute the solution of the constrained minimization problem as

$$\Delta \mathbf{x}_{\alpha k+1}^\delta(c) = \Delta \bar{\mathbf{x}} + \mathbf{p} = c(\Delta \bar{\mathbf{x}}_n - \mathbf{H} \mathbf{g}_n) - \mathbf{H} \mathbf{g}. \quad (6.77)$$

The above solution representation explicitly indicates the dependency on the relative vertical column, and this representation is beneficial in practice. The reason is that  $c$  is considered as a free parameter of the retrieval ranging in a chosen interval  $[c_{\min}, c_{\max}]$ . The problem to be solved is the computation of the strengths of the constraints, or more precisely, of the regularization parameter, which controls the smoothness of the solution, and of the relative vertical column, which controls the magnitude of the solution. Essentially, we must solve a multi-parameter regularization problem. In this case we adopt a simple strategy: we use an a priori chosen regularization parameter but compute the relative vertical column by using the minimum distance function approach. Two regularization methods with a dynamical selection criterion for the vertical column can be designed.

- (1) *Equality-constrained Tikhonov regularization with constant vertical column.* For each  $c \in [c_{\min}, c_{\max}]$ , we compute the solution  $\mathbf{x}_\alpha^\delta(c)$  of the nonlinear constrained minimization problem, and calculate the residual

$$R_\delta(c) = \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta(c))\|^2$$

and the constraint

$$C_\delta(c) = \|\mathbf{L}[\mathbf{x}_\alpha^\delta(c) - \mathbf{x}_a]\|^2$$

at the solution. Then, we determine the optimal value of the relative vertical column as the minimizer of the (normalized) distance function

$$d(c)^2 = \frac{R_\delta(c)}{R_{\delta \max}} + \frac{C_\delta(c)}{C_{\delta \max}} \quad (6.78)$$

over the interval  $[c_{\min}, c_{\max}]$ , where  $R_{\delta \max} = \max_c R_\delta(c)$  and  $C_{\delta \max} = \max_c C_\delta(c)$ .

- (2) *Equality-constrained Tikhonov regularization with variable total column.* At the iteration step  $k$ , we compute  $\Delta \mathbf{x}_{\alpha k+1}^\delta(c)$  for all  $c \in [c_{\min}, c_{\max}]$ , and evaluate the residual and the constraint for the linearized equation

$$R_\delta(c) = \|\mathbf{y}_k^\delta - \mathbf{K}_{\alpha k} \Delta \mathbf{x}_{\alpha k+1}^\delta(c)\|^2$$

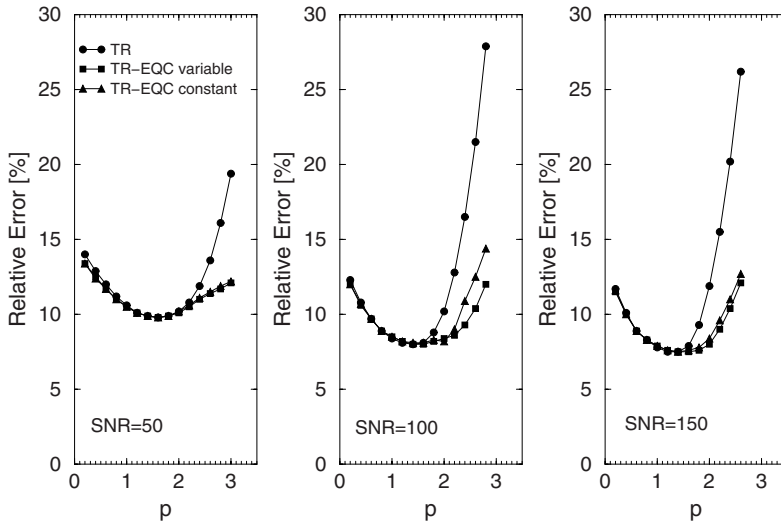
and

$$C_\delta(c) = \|\mathbf{L} \Delta \mathbf{x}_{\alpha k+1}^\delta(c)\|^2,$$

respectively. The optimal value of the total column at the current iteration step is the minimizer of the distance function (6.78) over the interval  $[c_{\min}, c_{\max}]$ .

Noting that the minimization of the distance function is usually performed by using a discrete search algorithm it is readily seen that the first solution method is more time-consuming than the second one. By virtue of (6.77), the computation of  $\Delta \mathbf{x}_{\alpha k+1}^\delta$  involves only a scalar-vector multiplication and the summation of two vectors. As a result, the computational effort of the equality-constrained Tikhonov regularization with variable total column is not much higher than that of the ordinary method.

The performance of the equality-constrained Tikhonov regularization will be analyzed from a numerical point of view. The ozone profile is retrieved from nadir synthetic data by considering 375 equidistant points in the spectral interval ranging from 290 to 335 nm. In this spectral interval,  $\text{O}_3$  and  $\text{NO}_2$  are considered as active gases. The atmosphere is discretized with a step of 3.5 km between 0 and 70 km, and a step of 10 km between 70 and 100 km. The exact state vector is chosen as a translated and a scaled version of a climatological profile with a translation distance of 3 km and a scaling factor of 1.3. The exact relative vertical column of ozone is  $c = 110$  DU (Dobson unit), and we choose  $c_{\min} = 80$  DU and  $c_{\max} = 125$  DU. To compute the minimizer of the distance function by a discrete search algorithm, 80 values of the relative vertical column are considered in the interval  $[c_{\min}, c_{\max}]$ . The reason for choosing this large interval of variation is that we have to guarantee that the distance function has a minimum for low values of the signal-to-noise ratio. The solar zenith angle is  $40^\circ$ , while the zenith and azimuthal angles of the line of sight are  $20^\circ$  and  $90^\circ$ , respectively. The regularization matrix is chosen as the Cholesky



**Fig. 6.12.** Relative solution errors for Tikhonov regularization (TR) and the equality-constrained Tikhonov regularization (TR-EQC) with variable and constant total column. The regularization parameter is  $\alpha = \sigma^p$ , where  $\sigma$  is the noise standard deviation.

factor of a normalized covariance matrix with an altitude-independent correlation length  $l = 3.5$  km.

In Figure 6.12 we plot the solution errors for Tikhonov regularization and the equality-constrained Tikhonov regularization for three values of the signal-to-noise ratio, namely 50, 100 and 150. The results show that for large values of  $p$  (small values of the regularization parameter), the solution errors for the constrained method are smaller than the solution errors for the ordinary method, while for small values of  $p$ , the solution errors are comparable. Thus, the equality constraint comes into effect for underestimations of the regularization parameter. The plots also indicate a slight superiority of the selection criterion with variable total column over that with constant total column.

The normalized constraint, residual and distance function are illustrated in Figure 6.13. The dependency of these quantities on the relative vertical column is similar to their dependency on the regularization parameter. For small values of the relative vertical column, the profile may have oscillatory artifacts around the a priori, so that the mean profile is essentially close to the a priori. Thus, for small values of  $c$ ,  $C_\delta(c)$  is large, while  $R_\delta(c)$  is small. However, in contrast to the regularization parameter dependency,  $C_\delta(c)$  is not a monotonically decreasing function on  $c$ . As  $R_\delta(c)$  is a monotonically increasing function on  $c$ , the minimizer of  $d(c)^2$  is shifted to the left of the minimizer of  $C_\delta(c)$ .

The retrieval results illustrated in Figure 6.14 correspond to a small value of the regularization parameter and two values of the signal-to-noise ratio. The profiles computed by Tikhonov regularization deviate significantly from the a priori, while the retrieved profiles computed by using the equality-constrained Tikhonov regularization are smoother and approximate the exact profile sufficiently well (especially in the troposphere).

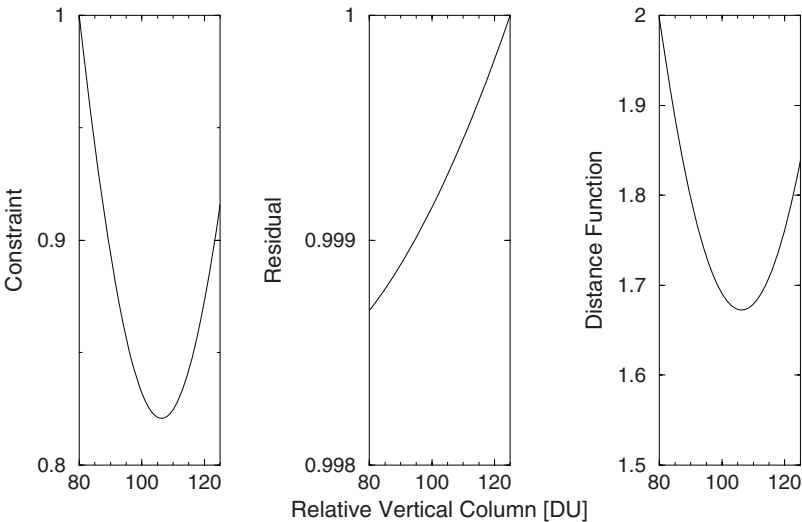


Fig. 6.13. Normalized constraint (left), residual (middle) and distance function (right).

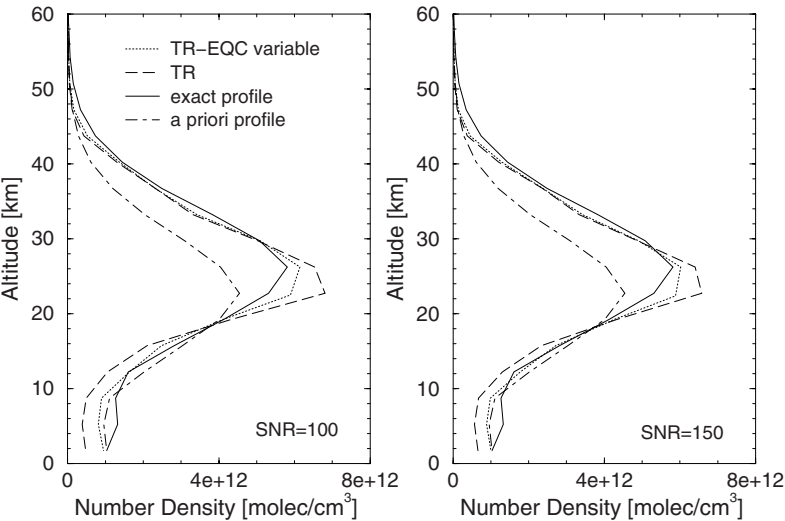


Fig. 6.14. Retrieved profiles computed by using Tikhonov regularization (TR) and the equality-constrained Tikhonov regularization (TR-EQC) with variable total column, in the case  $p = 2.4$ .

The comparison of the numerical effort of the methods can be inferred from Table 6.7. It is pleasant to observe that the computation times of Tikhonov regularization and of the equality-constrained Tikhonov regularization with variable total column are almost the same.

The main conclusions emerging from our numerical analysis is that the equality-constrained Tikhonov regularization is more stable than Tikhonov regularization with re-

**Table 6.7.** Computation time in min:ss format for Tikhonov regularization (TR) and the equality-constrained Tikhonov regularization (TR-EQC) with variable and constant total column. The numbers in parentheses represent the number of iteration steps and the relative solution error expressed in percent.

$p$	Method		
	TR	TR-EQC variable total column	TR-EQC constant total column
2.4	0:20 (4; 16.5)	0:21 (4; 9.2)	12:28 (4; 9.3)
0.2	0:20 (4; 12.3)	0:21 (4; 12.9)	12:28 (4; 12.8)

spect to underestimations of the regularization parameter. The interval of variation of the relative vertical column should be chosen so that the distance function has a minimum for the assumed values of the signal-to-noise ratio. To get some idea of where this interval lies, we may use as guide the value of the total column delivered by an independent retrieval. Evidently, this additional information is for reference only.

## 6.8 Mathematical results and further reading

In a continuous setting, nonlinear inverse problems can be cast into the abstract framework of nonlinear operator equations

$$F(x) = y, \quad (6.79)$$

where the operator  $F$  acts between the Hilbert spaces  $X$  and  $Y$ . Assuming that the nonlinear equation is solvable, i.e., that there exists  $x^\dagger \in \mathcal{D}(F)$  such that  $F(x^\dagger) = y$ , then the problem (6.79) is considered to be ill-posed if  $x^\dagger$  is not an isolated solution of the nonlinear equation or  $x^\dagger$  does not depend continuously on the data. In the first case, the solution cannot be locally reconstructed from the data, while in the second case, the reconstruction from noisy data does not yield reliable solutions.

For linear problems, the equation  $Kx = y$  is ill-posed if and only if  $\mathcal{R}(K)$  is not closed in  $Y$ . As linear equations with compact operators are ill-posed, an ill-posedness criterion for nonlinear equations should involve the compactness of the operators. However, since for nonlinear operators, compactness does not imply continuity, a reasonable demand is to suppose that  $F$  is completely continuous, which means that  $F$  is continuous and compact. In this regard, if  $F$  is completely continuous and weakly sequentially closed, and  $X$  is separable and infinite dimensional, then the problem (6.79) is ill-posed in  $x^\dagger$  (Rieder, 2003; Hofmann, 1997; Engl et al., 1989).

It would be helpful to characterize the stability of the nonlinear equation (6.79) through conditions on its linearization  $F'(x^\dagger)x = y$ , where  $F'$  is the Frechet derivative of  $F$ . The following results are due to Hofmann und Scherzer (1994, 1998) and can also be found in (Rieder, 2003): if  $F$  is Frechet-differentiable and  $F'$  is Lipschitz continuous in  $x^\dagger$ , then the locally ill-posedness of the nonlinear equation  $F(x) = y$  in  $x^\dagger$  implies the locally ill-posedness of its linearization  $F'(x^\dagger)x = y$  in all  $x \in X$ . This means that  $\mathcal{R}(F'(x^\dagger))$  is not closed or that  $F'(x^\dagger)$  is not injective. Unfortunately, the converse result does not hold,

that is, the ill-posedness of the linearization does not imply the ill-posedness of the nonlinear equation. In this context it is apparent that the connection between the ill-posedness of a nonlinear problem and its linearization is not as strong as one might think. This is a consequence of the Taylor expansion

$$F(x) = F(x^\dagger) + F'(x^\dagger)(x - x^\dagger) + R(x, x^\dagger), \quad (6.80)$$

which provides only little information on the local behavior of a nonlinear problem. The linearization error  $R(x, x^\dagger)$  behaves like  $o(\|x - x^\dagger\|)$  as  $x \rightarrow x^\dagger$ , but if  $F$  is completely continuous, then  $F'(x^\dagger)$  is compact and  $F'(x^\dagger)(x - x^\dagger)$  can be significantly smaller than  $R(x, x^\dagger)$ . This situation can be overcome when the linearization error is controlled by the nonlinear residual. Assuming that there exist  $\rho > 0$  and  $0 < \eta < 1$  such that

$$\|R(x, x')\| = \|F(x) - F(x') - F'(x')(x - x')\| \leq \eta \|F(x) - F(x')\| \quad (6.81)$$

for all  $x$  and  $x'$  in a ball  $B_\rho(x^\dagger)$  of radius  $\rho$  around  $x^\dagger$ , then the nonlinear equation  $F(x) = y$  is ill-posed in  $x^\dagger$  if and only if  $\mathcal{N}(F'(x^\dagger)) \neq 0$  or  $\mathcal{R}(F'(x^\dagger))$  is not closed. Conditions like (6.81), which restrict the nonlinearity of the operator, are frequently assumed in the analysis of regularization methods and are crucial for deriving convergence rate results. Examples of nonlinear ill-posed problems with well-posed linearizations and of well-posed nonlinear problems with ill-posed linearizations can be found in Engl et al. (1989).

A typical convergence result for Tikhonov regularization in a deterministic setting can be formulated as follows: under the assumptions

$$\alpha(\Delta) \rightarrow 0, \quad \frac{\Delta^2}{\alpha(\Delta)} \rightarrow 0 \quad \text{as } \Delta \rightarrow 0,$$

the regularized solution  $x_\alpha^\delta$  depends continuously on the data for  $\alpha$  fixed, and  $x_\alpha^\delta$  converges towards a solution of  $F(x) = y$  in a set-valued sense (Seidman and Vogel, 1989; Engl et al., 2000; Rieder, 2003). Although a deterministic theory of Tikhonov regularization for nonlinear problems is relatively complete, the development of a semi-stochastic theory is at the beginning. Whereas there exists a huge literature on linear inverse problems with random noise, only a few results have been published on nonlinear problems of this kind (Snieder, 1991; Wahba, 1990; Weese, 1993). Rigorous consistency and convergence rate results for nonlinear problems with random noise are available in a benchmark paper by O'Sullivan (1990), while more recently, Bissantz et al. (2004) derived rates of convergence for nonlinear Tikhonov regularization in a semi-stochastic setting.

A basic result on convergence rates for Tikhonov regularization with an a priori parameter choice method has been given by Neubauer (1989) and Engl et al. (1989). The main assumptions are that  $F'$  is Lipschitz continuous,

$$\|F'(x) - F'(x^\dagger)\| \leq L \|x - x^\dagger\|, \quad L > 0, \quad (6.82)$$

for all  $x \in B_\rho(x^\dagger)$ , and that there exists  $u \in Y$  such that

$$x^\dagger - x_a = F'(x^\dagger)^* u. \quad (6.83)$$

Then, if  $L \|u\| < 1$ , the a priori selection criterion  $\alpha \propto \Delta$ , yields the convergence rate

$$\|x_\alpha^\delta - x^\dagger\| = O(\sqrt{\Delta}).$$

If furthermore  $x^\dagger - x_a$  satisfies the Hölder-type source condition

$$x^\dagger - x_a = \left[ F' (x^\dagger)^* F' (x^\dagger) \right]^\mu z, \quad z \in X, \quad (6.84)$$

for some  $1/2 \leq \mu \leq 1$ , then the choice  $\alpha \propto \Delta^{2/(2\mu+1)}$  yields the convergence rate  $O(\Delta^{2\mu/(2\mu+1)})$ . The disadvantage of this regularization parameter choice method is that  $\alpha$  depends on the smoothing index  $\mu$  of the exact solution  $x^\dagger$  which is not known in practice. A slight variant of Tikhonov regularization which allows to prove the rate  $O(\sqrt{\Delta})$  for the choice  $\alpha \propto \Delta^2$  (now independent on the unknown  $\mu$ ) and under assumptions (6.82) and (6.83) with  $L \|u\| < 1$  can also be found in Engl et al. (1989). In this case,  $x_\alpha^\delta$  is the minimizer of the function

$$\mathcal{F}_\alpha(x) = \frac{1}{2} \left[ (\|y^\delta - F(x)\| - \Delta)^2 + \alpha \|x - x_a\|^2 \right],$$

and this choice avoids multiple minima of the Tikhonov function. In a semi-stochastic setting, this a priori parameter choice method takes the form  $\alpha \propto \sigma^2$  and coincides with the Bayesian selection criterion.

The convergence rate  $O(\sqrt{\Delta})$  has been proven by Engl et al. (1989) for Tikhonov regularization with the discrepancy principle. The proof relies on the assumption that the discrepancy equation has a solution  $\alpha(\Delta)$  for  $\Delta > 0$  sufficiently small, and that (6.82) and (6.83) hold with  $L \|u\| < 1$ . Another version of the discrepancy principle, which is very simply to implement for a discrete set of regularization parameters, selects that value of the regularization parameter  $\alpha$  satisfying

$$\tau_{\text{dp}} \Delta \leq \|y^\delta - F(x_\alpha^\delta)\| \leq (\tau_{\text{dp}} + \varepsilon) \Delta, \quad (6.85)$$

with  $\tau_{\text{dp}} > 1$  and  $\varepsilon > 0$ . The introduction of the positive number  $\varepsilon$  copes with the fact that the residual norm as a function of the regularization parameter is generally not strong monotonically increasing and not continuous (Tikhonov and Arsenin, 1977). The efficiency of this version of the discrepancy principle for a general regularization method (which includes Tikhonov regularization as a special case) has been demonstrated by Tautenhahn (1997).

For the Hölder-type source condition (6.84) with  $0 < \mu \leq 1$ , the generalized discrepancy principle yields the optimal convergence rate  $O(\Delta^{2\mu/(2\mu+1)})$ . This result has been proven by Scherzer et al. (1993) by assuming a series of restrictive conditions on  $F$ . The same convergence rate has been evidenced by Jin and Hou (1999) under the nonlinearity conditions

$$\begin{aligned} [F'(x) - F'(x')]z &= F'(x')h(x, x', z) \\ \|h(x, x', z)\| &\leq c_R \|x - x'\| \|z\|, \quad c_R > 0, \end{aligned}$$

for all  $x, x' \in \mathcal{B}_\rho(x^\dagger)$ .



For the nonlinear  $p$ -times iterated Tikhonov regularization, the optimal convergence rate  $O(\Delta^{2p/(2p+1)})$  has been established by Scherzer (1993), by comparing the iterated regularized solution of the nonlinear problem with the iterated regularized solution of its linearization.

In a discrete setting and for the choice  $\mathbf{L} = \mathbf{I}_n$ , Tikhonov regularization can be cast into a general framework of a regularization method based on the iteration

$$\mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_a + g_\alpha (\mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k}) \mathbf{K}_{\alpha k}^T \mathbf{y}_k^\delta, \quad k = 0, 1, \dots \quad (6.86)$$

For the sake of completeness we include in Appendix G convergence rate results for the general regularization method (6.86). The following conclusions arising from this analysis can be drawn: if for all  $\mathbf{x} \in B_\rho(\mathbf{x}^\dagger)$ ,  $\mathbf{F}$  satisfies the nonlinearity condition

$$\|\mathbf{F}(\mathbf{x}^\dagger) - \mathbf{F}(\mathbf{x}) - \mathbf{K}(\mathbf{x})(\mathbf{x}^\dagger - \mathbf{x})\| \leq \eta \|\mathbf{F}(\mathbf{x}^\dagger) - \mathbf{F}(\mathbf{x})\|, \quad 0 < \eta < 1,$$

and the source condition

$$\mathbf{x}^\dagger - \mathbf{x}_a = \left[ \mathbf{K}(\mathbf{x})^T \mathbf{K}(\mathbf{x}) \right]^\mu \mathbf{z}, \quad \mu > 0, \quad \mathbf{z} \in \mathbb{R}^n,$$

holds, then the a priori parameter choice method  $\alpha = (\Delta / \|\mathbf{z}\|)^{2/(2\mu+1)}$  and the discrepancy principle are of optimal order for  $0 < \mu \leq \mu_0/2$ . The index  $\mu_0$  represents the qualification of the regularization method, and for the method of Tikhonov regularization, we have  $\mu_0 = 1$ . As in the linear case, we observe that the best convergence rate of Tikhonov regularization equipped with the discrepancy principle as a posteriori parameter choice method is  $O(\sqrt{\Delta})$ .

# 7

## Iterative regularization methods for nonlinear problems

Finding a global minimizer of the Tikhonov function is in general not an easy task. Numerical experience shows that the Tikhonov function has usually many local minima and a descent method for solving the optimization problem may tend to get stuck especially for severely ill-posed problems. Since furthermore, the computation of an appropriate regularization parameter can require high computational effort, iterative regularization methods are an attractive alternative.

For iterative regularization methods, the number of iteration steps  $k$  plays the role of the regularization parameter, and the iterative process has to be stopped after an appropriate number of steps  $k^*$  in order to avoid an uncontrolled expansion of the noise error. In fact, a mere minimization of the residual, i.e., an ongoing iteration, leads to a semi-convergent behavior of the iterated solution: while the error in the residual decreases as the number of iteration steps increases, the error in the solution starts to increase after an initial decay. A widely used a posteriori choice for the stopping index  $k^*$  in dependence of the noise level  $\Delta$  and the noisy data vector  $\mathbf{y}^\delta$  is the discrepancy principle, that is, the iterative process is stopped after  $k^*$  steps such that

$$\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{k^*}^\delta)\|^2 \leq \tau \Delta^2 < \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta)\|^2, \quad 0 \leq k < k^*, \quad (7.1)$$

with  $\tau > 1$  chosen sufficiently large. In a semi-stochastic setting and for white noise with variance  $\sigma^2$ , the expected value of the noise  $\mathcal{E}\{\|\delta\|^2\} = m\sigma^2$  is used instead of the noise level  $\Delta^2$ .

In this chapter we review the relevant iterative regularization methods and discuss practical implementation issues. We first examine an extension of the Landweber iteration to nonlinear ill-posed problems, and then address practical aspects of Newton-type methods. The application of asymptotic regularization methods to the solution of nonlinear ill-posed problems will conclude our analysis.

### 7.1 Nonlinear Landweber iteration

There are several ways to extend the Landweber iteration to the nonlinear case. Interpreting the Landweber iteration for the linear equation  $\mathbf{K}\mathbf{x} = \mathbf{y}^\delta$  as a fixed point iteration  $\mathbf{x}_{k+1} = \Phi(\mathbf{x}_k)$  with the fixed point function  $\Phi(\mathbf{x}) = \mathbf{x} + \mathbf{K}^T(\mathbf{y}^\delta - \mathbf{K}\mathbf{x})$ , we replace  $\mathbf{K}\mathbf{x}$  by  $\mathbf{F}(\mathbf{x})$  in the expression of  $\Phi(\mathbf{x})$ , and obtain the so-called nonlinear Landweber iteration

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{K}_k^T \mathbf{r}_k^\delta, \quad k = 0, 1, \dots, \quad (7.2)$$

where  $\mathbf{K}_k = \mathbf{K}(\mathbf{x}_k^\delta)$  and

$$\mathbf{r}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta). \quad (7.3)$$

Alternatively, the nonlinear Landweber iteration can be regarded as a method of steepest descent, in which the negative gradient of the nonlinear residual

$$\mathcal{F}(\mathbf{x}) = \frac{1}{2} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2$$

determines the update direction for the current iterate.

As in the linear case, the nonlinear Landweber iteration can only converge if the equation  $\mathbf{F}(\mathbf{x}) = \mathbf{y}^\delta$  is properly scaled in the sense that

$$\|\mathbf{K}(\mathbf{x})\| \leq 1, \quad \mathbf{x} \in \mathcal{B}_\rho(\mathbf{x}_a),$$

where  $\mathcal{B}_\rho(\mathbf{x}_a)$  is a ball of radius  $\rho$  around  $\mathbf{x}_a$ . The scaling condition can be fulfilled in practice when both sides of the nonlinear equation are multiplied by a sufficiently small constant

$$0 < \chi \leq \left[ \max_{\mathbf{x} \in \mathcal{B}_\rho(\mathbf{x}_a)} \|\mathbf{K}(\mathbf{x})\| \right]^{-1},$$

which then in (7.2) appears as a relaxation parameter,

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \chi^2 \mathbf{K}_k^T \mathbf{r}_k^\delta, \quad k = 0, 1, \dots$$

The nonlinear Landweber iteration (7.2) corresponds to standard-form problems with  $\mathbf{L} = \mathbf{I}_n$ , while for general-form problems, the iteration takes the form

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + (\mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}_k^T \mathbf{r}_k^\delta, \quad k = 0, 1, \dots,$$

where  $\mathbf{L}$  is a square and nonsingular regularization matrix.

This method requires a large number of iteration steps to reduce the residual norm beyond the noise level. Although several modifications of the conventional method have been proposed to ameliorate this problem (Scherzer, 1998), the computational effort remains extremely high.

### 7.2 Newton-type methods

For ill-posed problems, the basic concepts of the Newton method provide a reliable basis for the development of iterative regularization methods. The key idea of any Newton-type

method consists in repeatedly linearizing the nonlinear equation about some approximate solution  $\mathbf{x}_k^\delta$ , solving the linearized equation

$$\mathbf{K}_k \mathbf{p} = \mathbf{r}_k^\delta, \quad (7.4)$$

for the Newton step  $\mathbf{p}_k^\delta$ , and updating the approximate solution according to the relation

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{p}_k^\delta. \quad (7.5)$$

Equation (7.4) is typically ill-posed and to obtain a reasonable solution some sort of regularization is necessary. The type of regularization employed, or the procedure which is used to compute the Newton step, characterizes a specific iterative method.

### 7.2.1 Iteratively regularized Gauss–Newton method

The iteratively regularized Gauss–Newton method relies on the solution of the linearized equation

$$\mathbf{K}_k (\mathbf{x} - \mathbf{x}_a) = \mathbf{y}_k^\delta, \quad (7.6)$$

with

$$\mathbf{y}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k (\mathbf{x}_k^\delta - \mathbf{x}_a),$$

by means of Tikhonov regularization with the penalty term  $\|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2$  and the regularization parameter  $\alpha_k$ . The new iterate minimizes the function

$$\mathcal{F}_{1k}(\mathbf{x}) = \|\mathbf{y}_k^\delta - \mathbf{K}_k (\mathbf{x} - \mathbf{x}_a)\|^2 + \alpha_k \|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2,$$

and is given by

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_a + \mathbf{K}_k^\dagger \mathbf{y}_k^\delta,$$

where  $\mathbf{K}_k^\dagger = (\mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}_k^T$  is the regularized generalized inverse at the iteration step  $k$ . At first glance, this method seems to be identical to the method of Tikhonov regularization with a variable regularization parameter, but the following differences exist:

- (1) the regularization parameters are the terms of a decreasing sequence satisfying the requirements

$$\alpha_k > 0, \quad 1 < \frac{\alpha_k}{\alpha_{k+1}} \leq c, \quad \lim_{k \rightarrow \infty} \alpha_k = 0; \quad (7.7)$$

- (2) the iterative process is stopped according to the discrepancy principle (7.1) instead of requiring the convergence of iterates and employing the discrepancy principle as an a posteriori parameter choice method.

Several strategies for selecting the regularization parameters  $\alpha_k$  can be considered. In our retrieval algorithm we use the selection criterion

$$\alpha_k = q_k \alpha_{k-1},$$

where  $q_k$  can be chosen as the ratio of a geometric sequence, i.e.,  $q_k = q < 1$  is constant, or as

$$q_k = \frac{\tau \Delta^2}{\|\mathbf{r}_k^\delta\|^2}, \quad (7.8)$$

and

$$q_k = 1 - \frac{\tau \Delta^2}{\|\mathbf{r}_k^\delta\|^2}. \quad (7.9)$$

With the choice (7.8) the regularization parameter decreases very fast at the beginning of iteration, while the scheme (7.9) allows enough regularization to be applied at the beginning of iteration and then to be gradually decreased.

Any iterative method using the discrepancy principle as stopping rule requires the knowledge of the noise level or of its statistical estimate  $\mathcal{E}\{\|\delta\|^2\}$ . Because in many practical problems arising in atmospheric remote sensing, the errors in the data cannot be estimated (due to the forward model errors), we propose the following stopping rules:

- (1) For a geometric sequence of regularization parameters, we store all iterates  $\mathbf{x}_k^\delta$  and require the convergence of the nonlinear residuals  $\|\mathbf{r}_k^\delta\|$  within a prescribed tolerance. If  $\|\mathbf{r}^\delta\|$  is the residual at the last iteration step, we choose the solution  $\mathbf{x}_{k^*}^\delta$ , with  $k^*$  being given by

$$\|\mathbf{r}_{k^*}^\delta\|^2 \leq \tau \|\mathbf{r}^\delta\|^2 < \|\mathbf{r}_k^\delta\|^2, \quad 0 \leq k < k^*, \quad \tau > 1.$$

- (2) For the selection rules (7.8) and (7.9), we first estimate the noise level. For this purpose, we minimize the sum of squares

$$\mathcal{F}(\mathbf{x}) = \frac{1}{2} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2$$

by requiring relative function convergence, compute the equivalent noise variance

$$\sigma_e^2 = \frac{1}{m - n} \|\mathbf{r}^\delta\|^2,$$

where  $\|\mathbf{r}^\delta\|$  is the residual at the last iteration step, and then set  $\Delta^2 = m\sigma_e^2$ .

The above heuristic stopping rules do not have any mathematical justification but work sufficiently well in practice. To our knowledge there is a lack in the mathematical literature dealing with this topic and, for the time being, we do not see other viable alternatives for practical applications.

Although, from a mathematical point of view, the iteratively regularized Gauss–Newton method does not require a step-length procedure, its use may prevent the iterative process from yielding an undesirable solution. Taking into account that the Newton step  $\mathbf{p}_k^\delta = \mathbf{x}_{k+1}^\delta - \mathbf{x}_k^\delta$  solves the equation

$$\mathbf{K}_{\mathbf{f}k}(\mathbf{x}_k^\delta)^T \mathbf{K}_{\mathbf{f}k}(\mathbf{x}_k^\delta) \mathbf{p} = -\mathbf{g}_k(\mathbf{x}_k^\delta),$$

where  $\mathbf{g}_k$  is the gradient of the objective function

$$\mathcal{F}_k(\mathbf{x}) = \frac{1}{2} \|\mathbf{f}_k(\mathbf{x})\|^2, \quad \mathbf{f}_k(\mathbf{x}) = \begin{bmatrix} \mathbf{F}(\mathbf{x}) - \mathbf{y}^\delta \\ \sqrt{\alpha_k} \mathbf{L}(\mathbf{x} - \mathbf{x}_a) \end{bmatrix},$$

and  $\mathbf{K}_{\mathbf{f}_k}$  is the Jacobian matrix of  $\mathbf{f}_k$ , we deduce that

$$\mathbf{g}_k(\mathbf{x}_k^\delta)^T \mathbf{p}_k^\delta = -\|\mathbf{K}_{\mathbf{f}_k}(\mathbf{x}_k^\delta) \mathbf{p}_k^\delta\|^2 < 0,$$

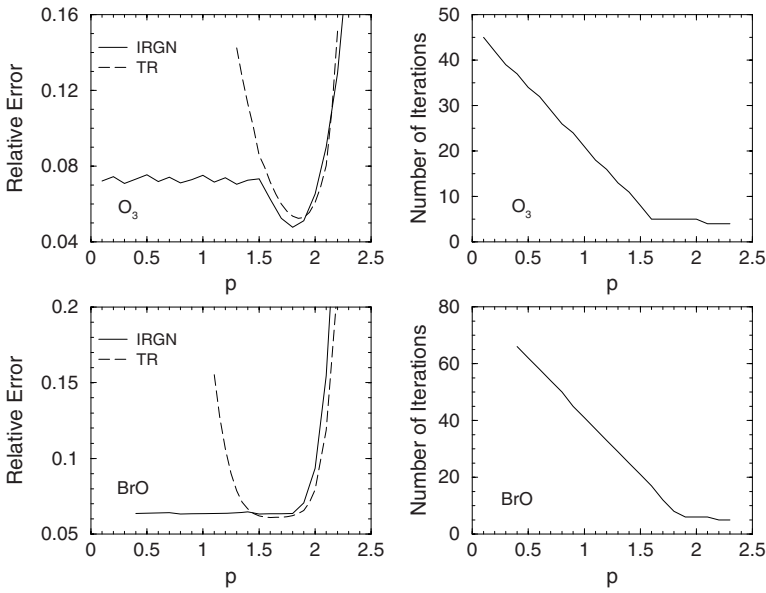
and so,  $\mathbf{p}_k^\delta$  is a descent direction for  $\mathcal{F}_k$ . Thus, the step-length procedure outlined in Algorithm 5 can be applied at each iteration step for the Tikhonov function  $\mathcal{F}_k$ .

In Figure 7.1 we illustrate the solution errors for the iteratively regularized Gauss–Newton method and Tikhonov regularization. In the iteratively regularized Gauss–Newton method, the exponent  $p$  characterizes the initial value of the regularization parameter,  $\alpha_0 = \sigma^p$ , while at all subsequent iteration steps, the regularization parameters are the terms of a geometric sequence with the ratio  $q = 0.8$ . The plots show that the iteratively regularized Gauss–Newton method still yields reliable results for small values of the exponent  $p$ , or equivalently, for large initial values of the regularization parameter. Evidently, a stronger regularization at the beginning of the iterative process requires a larger number of iteration steps as can be seen in the right panels of Figure 7.1. The main conclusion of this numerical simulation is that the iteratively regularized Gauss–Newton method is more stable than Tikhonov regularization with respect to overestimations of the regularization parameter.

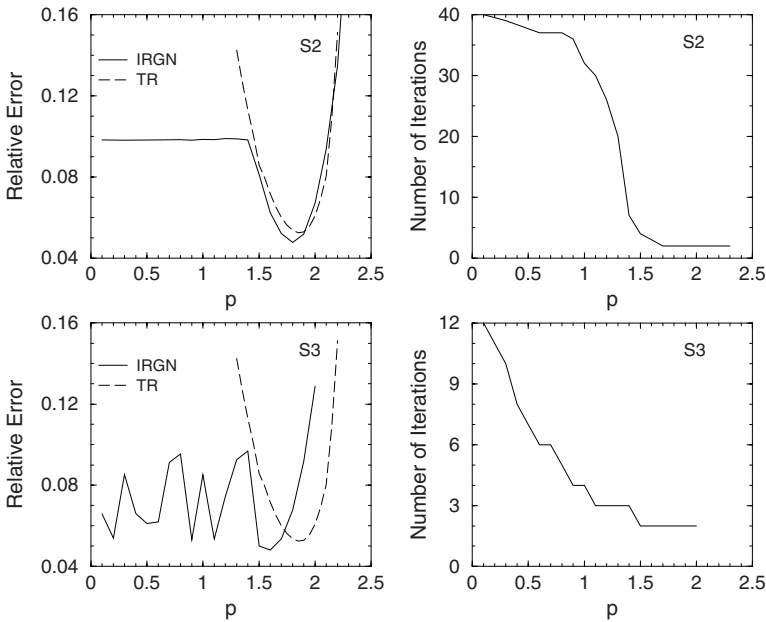
The same results are shown in Figure 7.2 for the dynamical selection criteria (7.8) and (7.9). The selection criterion (7.8) maintains the stability of the regularization method, but the errors at small  $p$ -values are almost two times larger than those corresponding to a geometric sequence. As a result, the retrieved profiles oscillate around the exact profiles and are undersmoothed. Although the selection criterion (7.9) requires a small number of iteration steps, it is less stable with respect to overestimations of the regularization parameter. This is because we cannot find a unique value of the control parameter  $\tau$  yielding accurate results over the entire domain of variation of  $p$ . For example, in the case  $p = 0.3$  and the choice  $\tau = 1.01$ , the solution error is 0.08. Choosing  $\tau = 1.05$ , we reduce the solution error to 0.05, but we increase the solution error at  $p = 0.5$  from 0.06 to 0.09. Thus, for the applications considered here, a dynamical selection of the regularization parameters is less reliable than an a priori selection rule using a geometric sequence (with constant ratio).

An important aspect of any iterative method using the discrepancy principle as stopping rule is the choice of the control parameter  $\tau$ . From a theoretical point of view,  $\tau$  should be larger than 4, but in many practical applications this choice leads to a premature termination of the iterative process. As we do not use the standard version of the discrepancy principle with known noise level, we determine the optimal value of  $\tau$  by minimizing the solution error. The results plotted in Figure 7.3 show that for the  $\text{O}_3$  and the BrO retrieval test problems, the optimal value of  $\tau$  is close to 1, and we find that a good choice for  $\tau$  is 1.01.

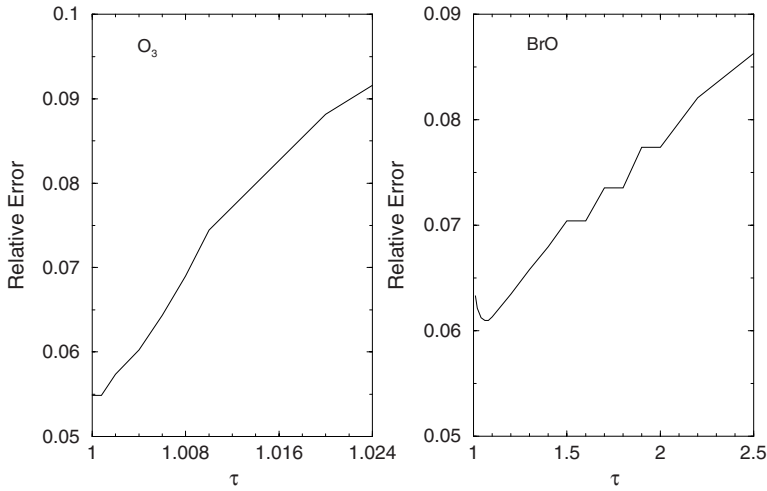
In Figure 7.4 we plot the histories of regularization parameters and residual norms for different initial values of the exponent  $p$ . The plots show that the limiting values of the sequences of regularization parameters and residual norms are comparable whatever the initial values of the regularization parameter are. These values of the regularization



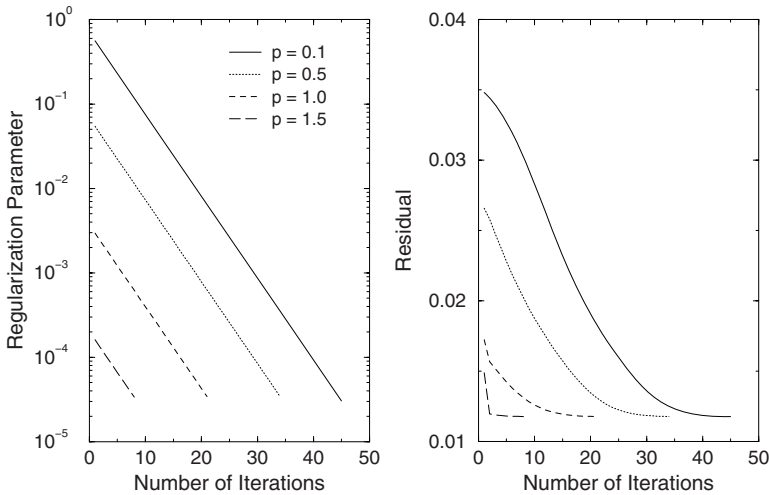
**Fig. 7.1.** Relative solution errors and the number of iteration steps for different values of the exponent  $p$ . The results are computed with the iteratively regularized Gauss–Newton (IRGN) method and Tikhonov regularization (TR).



**Fig. 7.2.** The same as in Figure 7.1 but for the selection criteria (7.8) ( $S_2$ ) and (7.9) ( $S_3$ ). The control parameter  $\tau$  is 1.01.



**Fig. 7.3.** Relative solution errors for different values of the control parameter  $\tau$ .

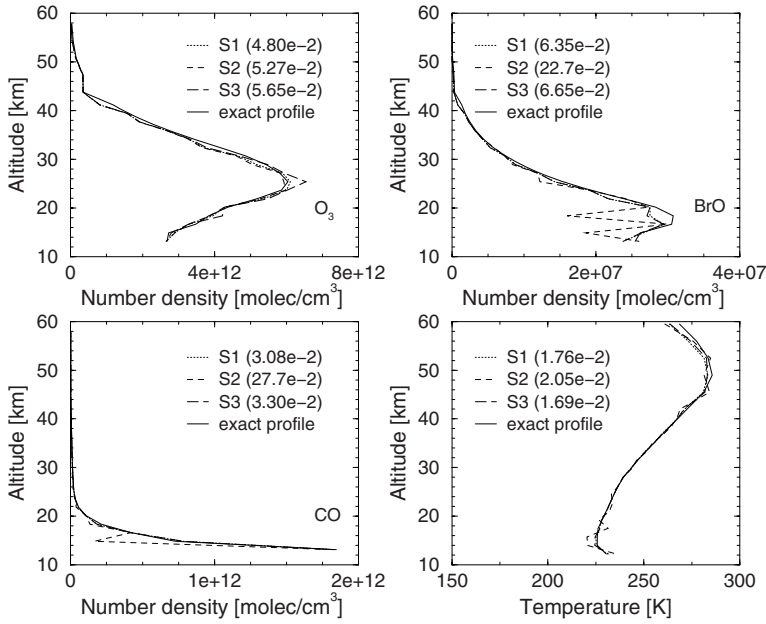


**Fig. 7.4.** Histories of regularization parameters and residual norms for different values of the exponent  $p$ .

parameter are  $3.04 \cdot 10^{-5}$  for  $p = 0.1$ ,  $3.44 \cdot 10^{-5}$  for  $p = 0.5$ ,  $3.40 \cdot 10^{-5}$  for  $p = 1.0$ , and  $3.37 \cdot 10^{-5}$  for  $p = 1.5$ . It is interesting to note that Tikhonov regularization using these limiting values as a priori regularization parameters, yields small solution errors; for the average value  $\alpha = 3.31 \cdot 10^{-5}$  in Figure 7.4, the solution error for Tikhonov regularization is  $5 \cdot 10^{-2}$ . This equivalence suggests that we may perform an error analysis at the solution with the final value of the regularization parameter.

The retrieved profiles for the four test problems are shown in Figure 7.5. The under-smoothing effect of the selection criterion (7.8) is more pronounced for the BrO and the CO retrieval test problems.





**Fig. 7.5.** Retrieved profiles computed with the iteratively regularized Gauss–Newton method. The results correspond to a geometric sequence of regularization parameters with a ratio of 0.8 (S1), and the selection criteria (7.8) (S2) and (7.9) (S3).

The incorporation of additional constraints into the iteratively regularized Gauss–Newton method, hereafter abbreviated as IRGN method, results in a regularization method which is less susceptible to the selection of the regularization parameter over a large range of values. For the ozone nadir sounding problem discussed in the preceding chapter, the equality-constrained IRGN method can be designed by replacing the unconstrained minimization problem

$$\min_{\Delta \mathbf{x}} \mathcal{Q}(\Delta \mathbf{x}) = \mathbf{g}^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{G} \Delta \mathbf{x},$$

by the quadratic programming problem (cf. (6.75) and (6.76))

$$\begin{aligned} \min_{\Delta \mathbf{x}} \mathcal{Q}(\Delta \mathbf{x}) &= \mathbf{g}^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{G} \Delta \mathbf{x} \\ \text{subject to } \sum_{i=1}^n [\Delta \mathbf{x}]_i &= c. \end{aligned}$$

Here, the Hessian and the gradient of  $\mathcal{Q}$  are given by  $\mathbf{G} = \mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{L}^T \mathbf{L}$  and  $\mathbf{g} = -\mathbf{K}_k^T \mathbf{y}_k^\delta$ , respectively. The quadratic programming problem is solved in the framework of the null-space method by using an explicit representation of the solution in terms of the vertical column. As opposed to the constrained Tikhonov regularization, both strengths of the constraints are now computed internally: the regularization parameter, which controls the smoothness of the solution, is decreased during the Newton iteration by a constant

factor, and the vertical column, which controls the magnitude of the solution, is determined by using the minimum distance function approach. As in general, iterative methods require more iteration steps than Tikhonov regularization, only the equality-constrained IRGN method with variable total column is appropriate for practical applications.

An inequality-constrained IRGN method can be derived if the total column is known with sufficiently accuracy. The information on the total column should be the result of an independent retrieval, which can be performed in a distinct spectral interval by using an appropriate algorithm like the DOAS approach (Van Roozendaal et al., 2006; Balis et al., 2007). The proposed inequality-constrained IRGN method is of the form of the following model algorithm: at the iteration step  $k$ , compute the a priori profile deviation  $\Delta \mathbf{x}_{k+1\alpha}^\delta = \mathbf{x}_{k+1\alpha}^\delta - \mathbf{x}_a$  by solving the quadratic programming problem

$$\min_{\Delta \mathbf{x}} Q(\Delta \mathbf{x}) = \mathbf{g}^T \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^T \mathbf{G} \Delta \mathbf{x} \quad (7.10)$$

$$\text{subject to } \sum_{i=1}^{n_t} [\Delta \mathbf{x}]_i \leq c_{\max}, \quad (7.11)$$

$$\sum_{i=1}^n [\Delta \mathbf{x}]_i \geq c_{\min}. \quad (7.12)$$

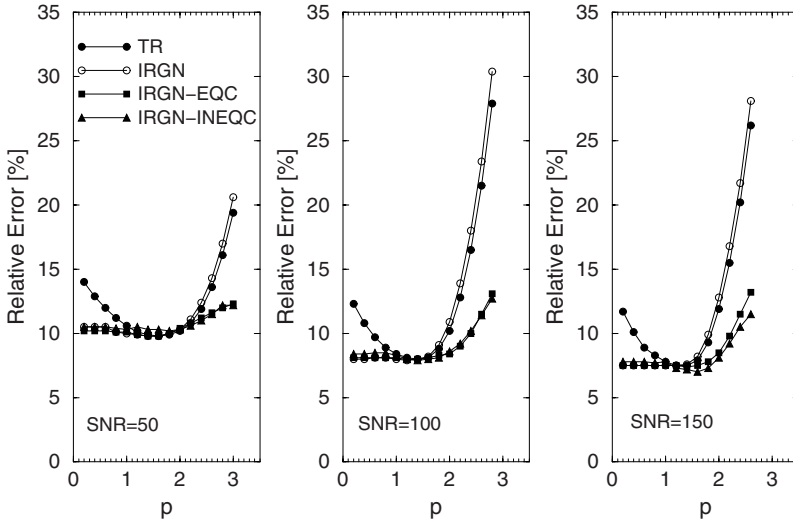
The layer  $n_t < n$ , delimits the tropospheric region from above, and the reasons for the choice (7.11)–(7.12) are the following:

- (1) the constraints should be linearly independent since otherwise one of the constraints can be omitted without altering the solution;
- (2) as the nadir radiance is less sensitive to variations of gas concentrations in the troposphere, the condition (7.11) does not allow large profile deviations in the sensitivity region above the troposphere;
- (3) the condition (7.12) guarantees a sufficiently large deviation of the profile (with respect to the a priori) over the entire altitude range.

If  $c$  is the relative vertical column delivered by an independent retrieval and  $\Delta c$  is the associated uncertainty, we may choose  $c_{\min} = c - \varepsilon_{\min} \Delta c$  with  $\varepsilon_{\min} \geq 1$ , and  $c_{\max} = c$ . This choice of the upper bound is reasonable since  $c_{\max}$  in (7.11) controls only the vertical column above the troposphere. The quadratic programming problem (7.10)–(7.12) can be solved by using primal and dual active set methods. The dual active set method of Goldfarb and Idnani (1983) generates dual-feasible iterates by keeping track of an active set of constraints (Appendix J). An implementation of the method of Goldfarb and Idnani is the routine ‘solve.qp’ from the optimization package ‘quadprog’, which is available free through the internet (CRAN-Package quadprog, 2007).

Considering the same retrieval scenario as in the preceding chapter and taking into account that the exact relative vertical column for ozone is  $c = 110$  DU, we choose  $c_{\min} = 80$  DU and  $c_{\max} = 125$  DU for equality constraints, and  $c_{\min} = 105$  DU and  $c_{\max} = 110$  DU for inequality constraints.

In Figure 7.6 we plot the solution errors for Tikhonov regularization and the constrained and unconstrained IRGN methods. For these simulations, three values of the signal-to-noise ratio have been considered, namely 50, 100 and 150. The plots show that



**Fig. 7.6.** Relative solution errors for Tikhonov regularization (TR), the IRGN method, and the equality- and inequality-constrained IRGN (IRGN-EQC and IRGN-INEQC) methods.

the constrained IRGN methods yield acceptable reconstruction errors over the entire domain of variation of the regularization parameter. The main drawback of the inequality-constrained IRGN method is its sensitivity to the selection of the bounds  $c_{\min}$  and  $c_{\max}$ . The reason is that the method does not use an internal selection criterion for the relative vertical column and the information on  $c$  should be sufficiently accurate. Especially, the choice of the bound  $c_{\min}$  is critical; we found that values smaller than 105 DU lead to large solution errors.

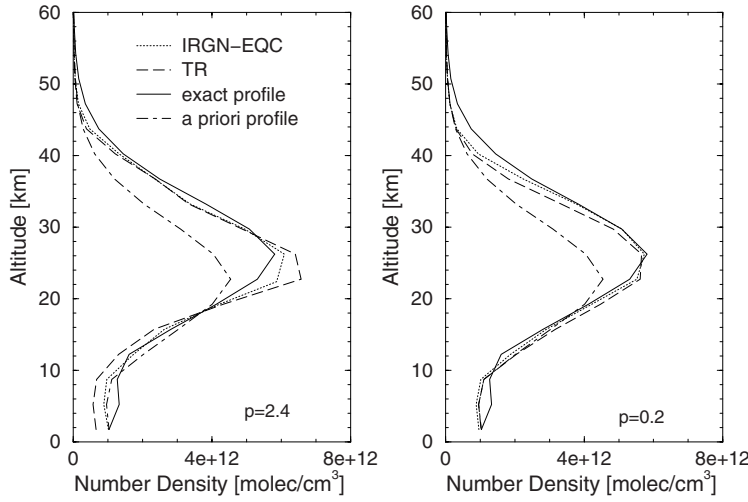
The retrieved profiles computed with the equality-constrained IRGN method and Tikhonov regularization are shown in Figure 7.7. For  $p = 2.4$ , the Tikhonov solution is undersmoothed, while for  $p = 0.2$ , the solution is oversmoothed in the sense that mainly the scaling and less the translation of the a priori profile is reproduced. In both situations, the profiles computed with the equality-constrained IRGN method are better approximations of the exact profile.

The computation times of the methods are outlined in Table 7.1. For  $p = 0.2$ , Tikhonov regularization is by a factor of 2 faster than the constrained IRGN methods, while for  $p = 2.4$  their efficiencies are comparable. This enhancement of computation time is the price that we have to pay for obtaining stable approximations of the solution over a large range of values of the regularization parameter.

We conclude this section by referring to a stopping rule which can be used in conjunction with any iterative regularization method, namely the Lepskij stopping rule (Bauer and Hohage, 2005). This criterion is based on monitoring the total error

$$\mathbf{e}_k^\delta = \mathbf{e}_{sk} + \mathbf{e}_{nk}^\delta,$$

where the smoothing and noise errors are given by  $\mathbf{e}_{sk} = (\mathbf{I}_n - \mathbf{A}_{k-1}) (\mathbf{x}^\dagger - \mathbf{x}_a)$  and  $\mathbf{e}_{nk}^\delta = -\mathbf{K}_{k-1}^\dagger \delta$ , respectively. The idea of the Lepskij stopping rule is to use the noise



**Fig. 7.7.** Retrieval results corresponding to Tikhonov regularization (TR) and the equality-constrained IRGN (IRGN-EQC) method in the case SNR = 100.

**Table 7.1.** Computation time in min:ss format for the regularization methods in Figure 7.6. The numbers in parentheses represent the number of iteration steps and the relative solution errors expressed in percent.

$p$	Method			
	TR	IRGN	IRGN-EQC	IRGN-INEQC
2.4	0:20 (4;16.5)	0:23 (5;18.0)	0:26 (5;9.8)	0:24 (5;9.9)
0.2	0:20 (4;12.3)	0:39 (12;8.1)	0:50 (12;8.1)	0:42 (12;8.3)

error bound

$$\|\mathbf{e}_{nk}^\delta\| \leq c_n \frac{\Delta}{2\sqrt{\alpha_{k-1}}}, \quad c_n \geq 1, \quad (7.13)$$

to detect the iteration step after which the total error is dominated by the noise error. By convention, the optimal stopping index  $k_{\text{opt}}$  is the iteration index yielding roughly a trade-off between the smoothing and noise errors. To estimate  $k_{\text{opt}}$ , we assume that the total error can be bounded as

$$\|\mathbf{x}_k^\delta - \mathbf{x}^\dagger\| \leq \mathfrak{E}(k) \Delta, \quad k = k_{\text{opt}}, \dots, k_{\text{max}},$$

where  $\mathfrak{E} : \mathbb{N} \rightarrow [0, \infty)$  is a known increasing function. Then, using the result

$$\|\mathbf{x}_{k_{\text{opt}}}^\delta - \mathbf{x}_k^\delta\| \leq \|\mathbf{x}_{k_{\text{opt}}}^\delta - \mathbf{x}^\dagger\| + \|\mathbf{x}_k^\delta - \mathbf{x}^\dagger\| \leq \mathfrak{E}(k_{\text{opt}}) \Delta + \mathfrak{E}(k) \Delta \leq 2\mathfrak{E}(k) \Delta$$

for all  $k = k_{\text{opt}} + 1, \dots, k_{\text{max}}$ , we deduce that the optimal stopping index  $k_{\text{opt}}$  can be approximated by the first index  $k^*$  with the property

$$\|\mathbf{x}_{k^*}^\delta - \mathbf{x}_k^\delta\| \leq 2\mathfrak{E}(k) \Delta, \quad k = k^* + 1, \dots, k_{\text{max}}. \quad (7.14)$$

The stopping index  $k^*$  is called the Lepskij stopping index and (7.14) is called the Lepskij stopping rule. The main problem which has to be solved is the choice of the function  $\mathfrak{E}$ . Taking into account that

$$\|\mathbf{e}_{\mathbf{S}k_{\text{opt}}}\| \approx \|\mathbf{e}_{\mathbf{n}k_{\text{opt}}}^\delta\|,$$

and that

$$\|\mathbf{e}_{\mathbf{S}k}\| \leq \|\mathbf{e}_{\mathbf{S}k_{\text{opt}}}\| \approx \|\mathbf{e}_{\mathbf{n}k_{\text{opt}}}^\delta\| \leq \|\mathbf{e}_{\mathbf{n}k}^\delta\|, \quad k = k_{\text{opt}}, \dots, k_{\text{max}},$$

we obtain

$$\|\mathbf{x}_k^\delta - \mathbf{x}^\dagger\| \leq 2 \|\mathbf{e}_{\mathbf{n}k}^\delta\|, \quad k = k_{\text{opt}}, \dots, k_{\text{max}}. \quad (7.15)$$

Thus, in a deterministic setting we may choose (cf. (7.13) and (7.15))

$$\mathfrak{E}(k) = \frac{c}{\sqrt{\alpha_{k-1}}}, \quad c \geq 1,$$

while in a semi-stochastic setting, the estimate

$$\mathcal{E} \left\{ \|\mathbf{e}_{\mathbf{n}k}^\delta\|^2 \right\} = \sigma^2 \text{trace} \left( \mathbf{K}_{k-1}^\dagger \mathbf{K}_{k-1}^{\dagger T} \right)$$

together with (7.15) suggests the choice

$$\mathfrak{E}(k) = c \sqrt{\frac{1}{m} \text{trace} \left( \mathbf{K}_{k-1}^\dagger \mathbf{K}_{k-1}^{\dagger T} \right)}, \quad c \geq 2.$$

## 7.2.2 Regularizing Levenberg–Marquardt method

In the regularizing Levenberg–Marquardt method, the linearized equation

$$\mathbf{K}_k (\mathbf{x} - \mathbf{x}_k^\delta) = \mathbf{r}_k^\delta, \quad (7.16)$$

with  $\mathbf{r}_k^\delta$  being given by (7.3), is solved by means of Tikhonov regularization with the penalty term  $\|\mathbf{L}(\mathbf{x} - \mathbf{x}_k^\delta)\|^2$  and the regularization parameter  $\alpha_k$ . The new iterate minimizing the Tikhonov function

$$\mathcal{F}_{1k}(\mathbf{x}) = \|\mathbf{r}_k^\delta - \mathbf{K}_k(\mathbf{x} - \mathbf{x}_k^\delta)\|^2 + \alpha_k \|\mathbf{L}(\mathbf{x} - \mathbf{x}_k^\delta)\|^2, \quad (7.17)$$

is given by

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{K}_k^\dagger \mathbf{r}_k^\delta. \quad (7.18)$$

The difference from the iteratively regularized Gauss–Newton method consists in the penalty term which now depends on the previous iterate instead of the a priori.

The parameter choice rule  $\alpha_k = q_k \alpha_{k-1}$  with  $q_k < 1$ , designed for the iteratively regularized Gauss–Newton method, can be used for the regularizing Levenberg–Marquardt method as well. Otherwise, the regularization parameter can be selected by applying the discrepancy principle to the linearized equation (7.16) (Hanke, 1997): if  $\mathbf{p}_{\alpha k}^\delta = \mathbf{K}_{\alpha k}^\dagger \mathbf{r}_k^\delta$  with

$$\mathbf{K}_{\alpha k}^\dagger = (\mathbf{K}_k^T \mathbf{K}_k + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}_k^T,$$

denotes the minimizer of the Tikhonov function (7.17) for an arbitrary  $\alpha$ , the Levenberg–Marquardt parameter  $\alpha_k$  is chosen as the solution of the ‘discrepancy principle’ equation

$$\|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha k}^\delta\|^2 = \theta \|\mathbf{r}_k^\delta\|^2, \quad 0 < \theta < 1, \quad (7.19)$$

and the Newton step is taken as  $\mathbf{p}_k^\delta = \mathbf{p}_{\alpha_k k}^\delta$ . The regularization parameter can also be chosen according to the generalized discrepancy principle, in which case,  $\alpha_k$  is the solution of the equation

$$\|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha k}^\delta\|^2 - (\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha k}^\delta)^T \hat{\mathbf{A}}_{\alpha k} (\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha k}^\delta) = \theta \|\mathbf{r}_k^\delta\|^2,$$

where  $\hat{\mathbf{A}}_{\alpha k} = \mathbf{K}_k \mathbf{K}_{\alpha k}^\dagger$  is the influence matrix.

As in the iteratively regularized Gauss–Newton method, a step-length procedure can be used to assure a decrease of the nonlinear residual at each iteration step. Considering the nonlinear residual

$$\mathcal{F}(\mathbf{x}) = \frac{1}{2} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2,$$

and taking into account that the gradient of  $\mathcal{F}$  at  $\mathbf{x}$  is given by

$$\mathbf{g}(\mathbf{x}) = -\mathbf{K}(\mathbf{x})^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})] = -\mathbf{K}(\mathbf{x})^T \mathbf{r}^\delta(\mathbf{x}),$$

we deduce that  $\mathbf{p}_k^\delta$ , solving the regularized normal equation

$$(\mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{L}^T \mathbf{L}) \mathbf{p} = \mathbf{K}_k^T \mathbf{r}_k^\delta,$$

satisfies the inequality

$$\mathbf{g}(\mathbf{x}_k^\delta)^T \mathbf{p}_k^\delta = - \left( \|\mathbf{K}_k \mathbf{p}_k^\delta\|^2 + \alpha_k \|\mathbf{L} \mathbf{p}_k^\delta\|^2 \right) < 0.$$

Thus,  $\mathbf{p}_k^\delta$  is a descent direction for  $\mathcal{F}$ , and the objective function in Algorithm 5 is the nonlinear residual.

Instead of a step-length algorithm, a trust-region algorithm can be used to guarantee the descent condition at each iteration step. This choice is justified by the equivalence between the regularizing Levenberg–Marquardt method and a trust-region method: for a general-form regularization, the  $k$ th iteration step of the optimization problem

$$\min_{\mathbf{x}} \mathcal{F}(\mathbf{x}) = \frac{1}{2} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2,$$

involves the solution of the trust-region problem

$$\begin{aligned} \min_{\mathbf{p}} \mathcal{M}_k(\mathbf{p}) \\ \text{subject to } \|\mathbf{L} \mathbf{p}\| \leq \Gamma_k, \end{aligned} \quad (7.20)$$

where

$$\mathcal{M}_k(\mathbf{p}) = \mathcal{F}(\mathbf{x}_k^\delta) - \mathbf{r}_k^{\delta T} \mathbf{K}_k \mathbf{p} + \frac{1}{2} \mathbf{p}^T \mathbf{K}_k^T \mathbf{K}_k \mathbf{p}, \quad (7.21)$$

---

**Algorithm 11.** Regularizing Levenberg–Marquardt method with a trust-region algorithm. Given the actual iterate  $\mathbf{x}$  and the regularization parameter  $\alpha$ , the algorithm computes the new iterate  $\mathbf{x}_{\text{new}}$  to assure a sufficient decrease of the objective function. The control parameters can be chosen as  $\varepsilon_f = 10^{-4}$ ,  $\varepsilon_{1\Gamma} = 0.1$  and  $\varepsilon_{2\Gamma} = 0.5$ .

---

```

 $\mathcal{F} \leftarrow 0.5 \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2$ ;  $\mathbf{g} \leftarrow -\mathbf{K}(\mathbf{x})^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})]$ ;
compute the step  $\mathbf{p}$  for  $\alpha$ ;
 $\Gamma \leftarrow \|\mathbf{L}\mathbf{p}\|$ ; {trust-region radius for this step}
estimate  $\Gamma_{\min}$ ; retcode  $\leftarrow 2$ ; firstcall  $\leftarrow \text{true}$ ;
while retcode  $> 1$  do
    if firstcall = false compute the trial step  $\mathbf{p}$  for the trust-region radius  $\Gamma$ ;
     $\mathbf{x}_{\text{new}} \leftarrow \mathbf{x} + \mathbf{p}$ ;  $\mathcal{F}_{\text{new}} \leftarrow 0.5 \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\text{new}})\|^2$ ;  $\Delta\mathcal{F} \leftarrow \mathcal{F} - \mathcal{F}_{\text{new}}$ ;
    {objective function is too large; reduce  $\Gamma$  and continue the while loop}
    if  $\mathcal{F}_{\text{new}} > \mathcal{F} + \varepsilon_f \mathbf{g}^T \mathbf{p}$  then
        if  $\Gamma < \Gamma_{\min}$  then
            retcode  $\leftarrow 1$ ;  $\mathbf{x}_{\text{new}} \leftarrow \mathbf{x}$ ;  $\mathcal{F}_{\text{new}} \leftarrow \mathcal{F}$ ;
        else
            retcode  $\leftarrow 2$ ;  $\Gamma_{\text{tmp}} \leftarrow 0.5 (\mathbf{g}^T \mathbf{p}) \|\mathbf{L}\mathbf{p}\| / (\Delta\mathcal{F} + \mathbf{g}^T \mathbf{p})$ ;
            if  $\Gamma_{\text{tmp}} < \varepsilon_{1\Gamma} \Gamma$  then
                 $\Gamma \leftarrow \varepsilon_{1\Gamma} \Gamma$ ;
            else if  $\Gamma_{\text{tmp}} > \varepsilon_{2\Gamma} \Gamma$  then
                 $\Gamma \leftarrow \varepsilon_{2\Gamma} \Gamma$ ;
            else
                 $\Gamma \leftarrow \Gamma_{\text{tmp}}$ ;
            end if
        end if
    {objective function is sufficiently small}
    else
        retcode  $\leftarrow 0$ ;
    end if
    firstcall  $\leftarrow \text{false}$ ;
end while

```

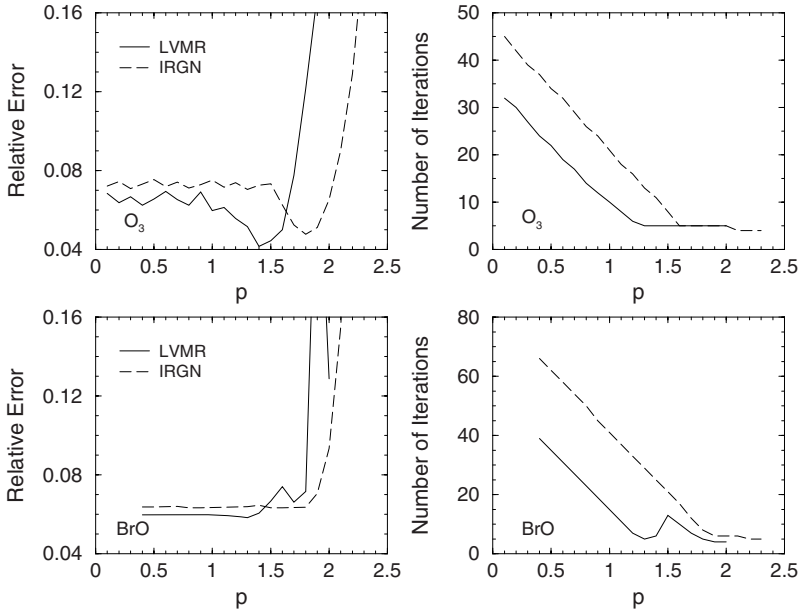
---

is the quadratic Gauss–Newton model about the current iterate and  $\Gamma_k$  is the trust-region radius. The regularizing Levenberg–Marquardt method with a trust-region procedure is illustrated in Algorithm 11. In contrast to the standard implementation (Algorithm 6), the regularization parameter (or the Lagrange multiplier) is chosen a priori and is not determined by the trust-region radius. Only if the descent condition is violated, the trust-region radius is reduced, and the new step is computed accordingly. To compute the trial step  $\mathbf{p}_k^\delta$  for the trust-region radius  $\Gamma_k$ , we consider the standard-form problem

$$(\bar{\mathbf{K}}_k^T \bar{\mathbf{K}}_k + \alpha \mathbf{I}_n) \bar{\mathbf{p}} = \bar{\mathbf{K}}_k^T \mathbf{r}_k^\delta,$$

with  $\bar{\mathbf{K}}_k = \mathbf{K}_k \mathbf{L}^{-1}$  and  $\bar{\mathbf{p}} = \mathbf{L}\mathbf{p}$ , solve the trust-region equation

$$\sum_{i=1}^n \left( \frac{\sigma_i}{\sigma_i^2 + \alpha} \right)^2 (\mathbf{u}_i^T \mathbf{r}_k^\delta)^2 = \Gamma_k^2, \quad (7.22)$$



**Fig. 7.8.** Relative solution errors and the number of iteration steps for different values of the exponent  $p$ . The results correspond to the regularizing Levenberg–Marquardt (LVMR) method and the iteratively regularized Gauss–Newton (IRGN) method.

for  $\bar{\alpha}$ , where  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of  $\bar{\mathbf{K}}_k$ , and then set  $\mathbf{p}_k^\delta = \mathbf{L}^{-1} \bar{\mathbf{p}}_{\bar{\alpha}k}^\delta$ , where  $\bar{\mathbf{p}}_{\bar{\alpha}k}^\delta = \bar{\mathbf{K}}_{\bar{\alpha}k}^\dagger \mathbf{r}_k^\delta$ .

The regularizing Levenberg–Marquardt method is also insensitive to overestimations of the regularization parameter. The results in Figure 7.8 show that the regularizing Levenberg–Marquardt method is superior to the iteratively regularized Gauss–Newton method: for large initial values of the regularization parameter, the number of iteration steps as well as the solution errors are smaller.

The retrieved profiles illustrated in Figure 7.9 give evidence that for the  $BrO$  retrieval test problem, the undersmoothing effect of the selection criterion (7.8) is not so pronounced as in the case of the iteratively regularized Gauss–Newton method.

The results listed in Table 7.2 demonstrate that for the  $BrO$  and the  $CO$  retrieval test problems, the solution errors corresponding to the trust-region algorithm are on average smaller than those corresponding to the step-length algorithm.

The standard trust-region implementation of the Levenberg–Marquardt method (Algorithm 6) is also a regularization, in which the regularization parameter is adjusted by the trust-region radius (Wang and Yuan, 2005). However, we found that this method is very sensitive to the selection of the model parameters, especially to the choice of the amplification factor  $c_a$ , which controls the increase of the trust-region radius. The results in Figure 7.10 show that for large initial values of the regularization parameter we have to increase the amplification factor in order to obtain reasonable accuracies. Acceptable solutions correspond to a small domain of variations of the initial regularization parameter, and the solution errors are in general slightly larger than those corresponding to the regularizing Levenberg–Marquardt method.



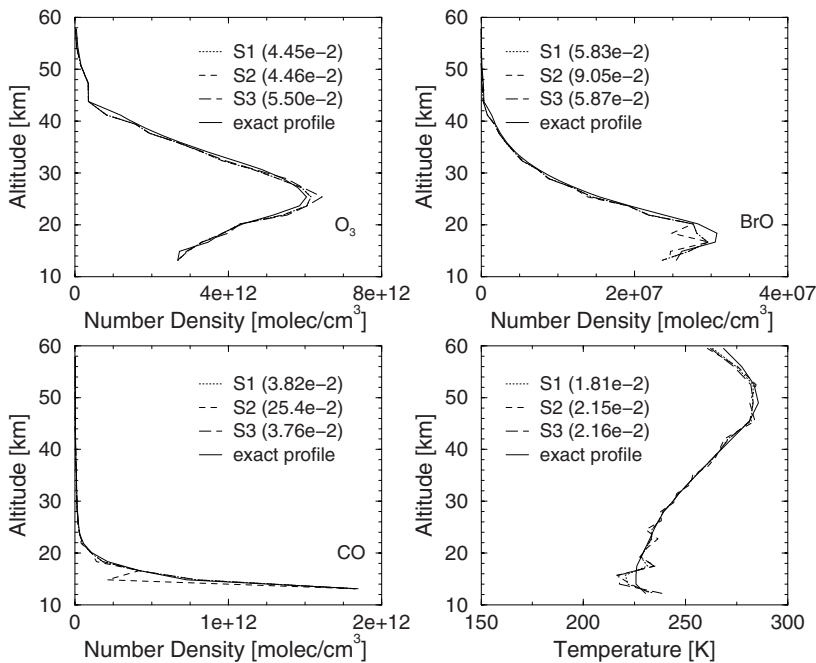
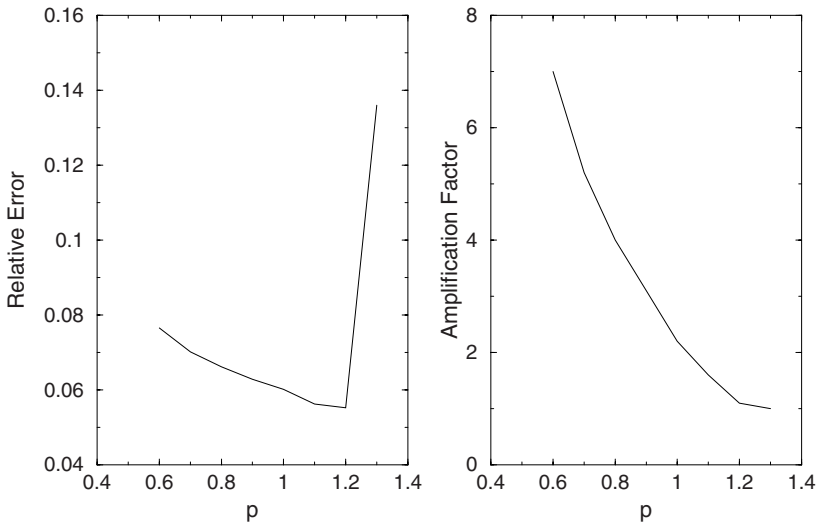


Fig. 7.9. The same as in Figure 7.5 but for the regularizing Levenberg–Marquardt method.

**Table 7.2.** Relative solution errors for the regularizing Levenberg–Marquardt method with the step-length and trust-region algorithms. The results correspond to a geometric sequence of regularization parameters with a ratio of 0.8 (S1), and the selection criteria (7.8) (S2) and (7.9) (S3).

Problem	Procedure	Selection criterion		
		S1	S2	S3
O <sub>3</sub>	step-length	4.45e-2	4.46e-2	5.50e-2
	trust-region	4.41e-2	4.42e-2	5.01e-2
BrO	step-length	5.83e-2	9.05e-2	5.87e-2
	trust-region	3.54e-2	4.44e-2	3.92e-2
CO	step-length	3.82e-2	2.54e-1	3.76e-2
	trust-region	3.12e-2	1.21e-1	1.82e-2
Temperature	step-length	1.81e-2	2.16e-2	2.16e-2
	trust-region	1.96e-2	3.01e-2	2.13e-2



**Fig. 7.10.** Left: relative solution errors versus the exponent  $p$  specifying the initial value of the regularization parameter. Right: amplification factor  $c_a$  which controls the increase of the trust-region radius. The results correspond to the  $O_3$  retrieval test problem.

### 7.2.3 Newton–CG method

The Newton–CG method relies on the solution of the linearized equation

$$\mathbf{K}_k \mathbf{p} = \mathbf{r}_k^\delta, \quad (7.23)$$

by means of the conjugate gradient for normal equations and by using the nonsingular regularization matrix  $\mathbf{L}$  as right preconditioner. For this purpose, the CGNR or the LSQR algorithms discussed in Chapter 5 can be employed. The main peculiarity of this solution method is that the linearized equation is not solved completely; only a number of  $p_k$  iterations are performed at the Newton step  $k$ . In this regard it is apparent that the number of iteration steps  $p_k$  plays the role of the regularization parameter  $\alpha_k$ . The resulting algorithm belongs to the class of the so-called REGINN (REGularization based on INexact Newton iteration) methods (Rieder, 1999; 2003). The term inexact Newton method refers to an approach consisting of two components:

- (1) an outer Newton iteration which updates the current iterate;
- (2) an inner iteration which provides the update by approximately solving a linearized version of the nonlinear equation.

It should be pointed out that other iterative methods as for example, the Landweber iteration or the  $\nu$ -method, can be used for solving the linearized equation (7.23).

---

**Algorithm 12.** REGINN (REGularization based on INexact Newton iteration) algorithm. The control parameters of the algorithm are  $\theta_0, \theta_{\max}, q$  and  $\tau$ .

---

```

set  $\Delta^2 = \mathcal{E} \left\{ \|\delta\|^2 \right\} = m\sigma^2$  or estimate  $\Delta^2$ ;
 $k \leftarrow 0, \quad \mathbf{x}_0^\delta \leftarrow \mathbf{x}_a$ ;
compute  $\mathbf{F}(\mathbf{x}_0^\delta)$  and  $\mathbf{K}_0 = \mathbf{K}(\mathbf{x}_0^\delta); \quad \mathbf{r}_0^\delta \leftarrow \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_0^\delta);$ 
 $\tilde{\theta}_0 \leftarrow \theta_0; \quad \tilde{\theta}_1 \leftarrow \theta_0$ ;
while  $\|\mathbf{r}_k^\delta\|^2 > \tau\Delta^2$  do {discrepancy principle for the outer iteration}
    if  $k > 1$  compute  $\tilde{\theta}_k$  by using (7.26) ;
     $\theta_k \leftarrow \theta_{\max} \max \left( \tau\Delta^2 / \|\mathbf{r}_k^\delta\|^2, \tilde{\theta}_k \right)$ ;
     $l \leftarrow 0$  ;
    repeat
         $l \leftarrow l + 1$ ;
        compute  $\mathbf{p}_{lk}^\delta$ ;
        until  $\|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{lk}^\delta\|^2 \leq \theta_k \|\mathbf{r}_k^\delta\|^2$  {discrepancy principle for the inner iteration}
         $p_k \leftarrow l$ ;
         $\mathbf{x}_{k+1}^\delta \leftarrow \mathbf{x}_k^\delta + \mathbf{p}_{p_k k}^\delta$ ;
        compute  $\mathbf{F}(\mathbf{x}_{k+1}^\delta)$  and  $\mathbf{K}_{k+1} = \mathbf{K}(\mathbf{x}_{k+1}^\delta); \quad \mathbf{r}_{k+1}^\delta \leftarrow \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{k+1}^\delta);$ 
         $k \leftarrow k + 1$ ;
    end while

```

---

The REGINN method outlined in Algorithm 12 is due to Rieder (1999; 2003). The outer Newton iteration (the while loop) is stopped according to the discrepancy principle (7.1). The number of iteration steps  $p_k$  of the inner scheme (the repeat loop) is chosen according to the discrepancy principle for the linearized equation (7.23) (compare to (7.19))

$$\|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{p_k k}^\delta\|^2 \leq \theta_k \|\mathbf{r}_k^\delta\|^2 < \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{lk}^\delta\|^2, \quad 1 \leq l < p_k, \quad (7.24)$$

while the following selection criterion is used for the tolerances  $\theta_k$ :

- (1) choose  $\theta_0 \in (0, 1)$  and  $q \in (0, 1]$ ;
- (2) set  $\tilde{\theta}_0 = \tilde{\theta}_1 = \theta_0$ ;
- (3) compute

$$\theta_k = \theta_{\max} \max \left( \frac{\tau\Delta^2}{\|\mathbf{r}_k^\delta\|^2}, \tilde{\theta}_k \right), \quad (7.25)$$

where  $\tilde{\theta}_k$  is given by

$$\tilde{\theta}_k = \begin{cases} 1 - \frac{p_k - 2}{p_{k-1}} (1 - \theta_{k-1}), & p_{k-1} \geq p_{k-2}, \\ q\theta_{k-1}, & p_{k-1} < p_{k-2}, \end{cases} \quad k \geq 2, \quad (7.26)$$

and  $\theta_{\max} \in (\theta_0, 1)$  bounds the  $\theta_k$  away from 1 (uniformly in  $k$  and  $\Delta$ ).

The parameter  $\theta_{\max}$  should be very close to 1, for instance, the choice  $\theta_{\max} = 0.999$  is reasonable. The general idea of the selection rule (7.25)–(7.26) is to start with a small tolerance and to increase it during the Newton iteration. However, the level  $\theta_k \|\mathbf{r}_k^\delta\|^2$  should

decrease during the iterative process, so that on average, the number of iteration steps  $p_k$  of the inner scheme should increase with increasing  $k$ . In the starting phase of the algorithm, the nonlinear residual is relatively large and as a result, the level  $\theta_k \|\mathbf{r}_k^\delta\|^2$  is not very small even for small values of the tolerances. Thus, in spite of small tolerances, the repeat loop will terminate. From (7.26) it is apparent that the tolerance is increased when the number of passes through the repeat loop of two successive Newton steps increases significantly, and it is decreased by a constant factor whenever the consecutive numbers of passes through the repeat loop drop. However, a rapid decrease of the tolerances should be avoided (the repeat loop may not terminate) and the choice of  $q$  in the interval  $[0.9, 1]$  is appropriate. In (7.25), a safeguarding technique to prevent oversolving of the discrepancy principle (especially in the final Newton step) is incorporated: at each Newton step there holds  $\theta_k \|\mathbf{r}_k^\delta\|^2 \geq \tau \Delta^2$ .

In our retrieval algorithm we use an a priori selection rule instead of the dynamical selection criterion (7.24): the number of iteration steps of the inner scheme is assumed to vary linearly between  $p_{\min}$  and  $p_{\max}$ ,

$$p_k = \xi^k p_{\min} + (1 - \xi^k) p_{\max}, \quad 0 < \xi < 1, \quad (7.27)$$

or according to the exponential law

$$p_k = p_{\max} - (p_{\max} - p_{\min}) e^{-\xi^k}. \quad (7.28)$$

In Figure 7.11 we illustrate the solution error for the selection criterion (7.27) as a function of the initial number of iteration steps of the inner scheme  $p_0 = p_{\min}$ . The main conclusions emerging from this simulation are summarized below.

- (1) For each value of  $p_{\max}$ , there exists a large interval of variation of  $p_{\min}$  yielding acceptable solution errors.
- (2) Large values of both control parameters  $p_{\min}$  and  $p_{\max}$  mean large values of the number of iteration steps  $p_k$ . In this case, the regularization decreases very fast at the beginning of the iterative process, the retrieved profiles are undersmoothed and the solution errors are large.
- (3) For a fixed value of  $p_{\min}$ , small values of  $p_{\max}$  yield small values of  $p_k$ . The regularization applied at each Newton step is large, and therefore, the number of Newton steps is also large.

At each Newton step  $k$ , the number of iterations  $p_k$  of the selection criterion (7.28) is smaller than that of the selection criterion (7.27), and as a result, the number of Newton steps is larger (Figure 7.12).

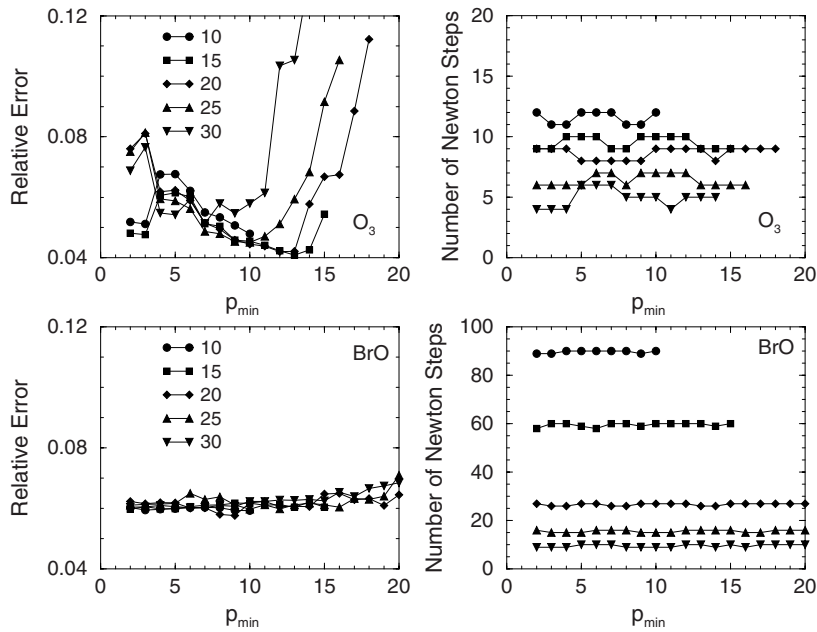
### 7.3 Asymptotic regularization

Asymptotic regularization can be regarded as a continuous analog of the Landweber iteration,

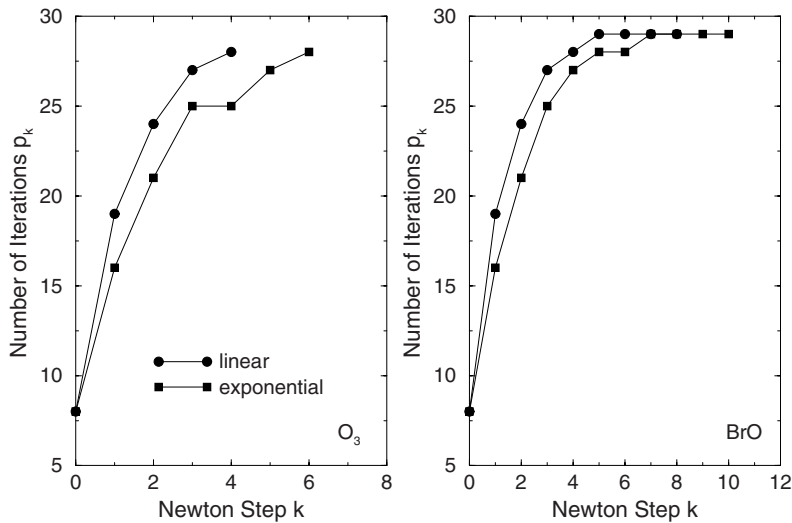
$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{K}_k^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta)], \quad k = 0, 1, \dots$$

In this method, a regularized approximation  $\mathbf{x}^\delta(T)$  of the exact solution  $\mathbf{x}^\dagger$  is obtained by solving the initial value problem (Showalter differential equation)

$$\dot{\mathbf{x}}^\delta(t) = \mathbf{K}(\mathbf{x}^\delta(t))^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}^\delta(t))], \quad 0 < t \leq T, \quad \mathbf{x}^\delta(0) = \mathbf{x}_a, \quad (7.29)$$



**Fig. 7.11.** Relative solution errors and the number of Newton steps versus  $p_{min}$  for the selection criterion (7.27) with  $\xi = 0.5$ . Each curve corresponds to a value of  $p_{max}$  ranging between 10 and 30.



**Fig. 7.12.** Number of iterations  $p_k$  at each Newton step  $k$  for the linear and exponential selection rules (7.27) and (7.28), respectively.

where  $T$  plays the role of the regularization parameter. The basic property of asymptotic regularization states that  $\mathbf{x}(T) \rightarrow \mathbf{x}^\dagger$  as  $T \rightarrow \infty$ , where  $\mathbf{x}(t)$  is the solution of the noise-free problem with the exact data vector  $\mathbf{y}$ . For linear problems, this result is straightforward: the solution of the initial value problem

$$\dot{\mathbf{x}}(t) = \mathbf{K}^T [\mathbf{y} - \mathbf{K}\mathbf{x}(t)], \quad 0 < t \leq T, \quad \mathbf{x}(0) = \mathbf{x}_a,$$

is given by

$$\mathbf{x}(t) = e^{-\mathbf{K}^T \mathbf{K} t} \mathbf{x}_a + (\mathbf{K}^T \mathbf{K})^{-1} (\mathbf{I}_n - e^{-\mathbf{K}^T \mathbf{K} t}) \mathbf{K}^T \mathbf{y},$$

whence letting  $T \rightarrow \infty$ , we obtain

$$\mathbf{x}(T) \rightarrow (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{y} = \mathbf{x}^\dagger.$$

For the nonlinear case, convergence results for the unperturbed and perturbed problems in a continuous setting have been established by Tautenhahn (1994). Applying the family of Runge–Kutta methods to the initial value problem (7.29), several iterative regularization methods have been developed by Böckmann and Pornsawad (2008). Similarly, Hochbruck et al. (2009) proposed an exponential Euler regularization method for solving the Showalter differential equation. In this section we analyze the computational efficiency of the Runge–Kutta regularization method and of the exponential Euler regularization method.

In the framework of Runge–Kutta methods, an approximate solution of the initial value problem

$$\dot{\mathbf{x}}(t) = \Psi(t, \mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_a,$$

is computed as

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \tau_k \sum_{i=1}^s b_i \Psi(t + c_i \tau_k, \mathbf{v}_i), \\ \mathbf{v}_i &= \mathbf{x}_k + \tau_k \sum_{j=1}^s a_{ij} \Psi(t + c_j \tau_k, \mathbf{v}_j), \quad i = 1, \dots, s, \quad k = 0, 1, \dots, \end{aligned}$$

where  $\mathbf{x}_0 = \mathbf{x}_a$ ,  $s$  is the number of stages,  $\tau_k$  is the step length at the actual iteration step and the coefficients  $a_{ij}$ ,  $b_i$  and  $c_i$  with  $i, j = 1, \dots, s$ , depend on the particular method employed. These coefficients are usually arranged in a mnemonic device, known as the Butcher tableau (Figure 7.13).

For our purpose, we consider consistent Runge–Kutta methods with the property

$$\sum_{i=1}^s b_i = 1. \tag{7.30}$$

Applying the above scheme to the initial value problem (7.29) yields

$$\begin{aligned} \mathbf{x}_{k+1}^\delta &= \mathbf{x}_k^\delta + \tau_k \sum_{i=1}^s b_i \mathbf{K}(\mathbf{v}_i)^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{v}_i)], \\ \mathbf{v}_i &= \mathbf{x}_k^\delta + \tau_k \sum_{j=1}^s a_{ij} \mathbf{K}(\mathbf{v}_j)^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{v}_j)], \quad i = 1, \dots, s, \quad k = 0, 1, \dots \end{aligned}$$

$$\begin{array}{ccc}
\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \end{array} & \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} & \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \\
\hline
\begin{array}{c|ccc} & b_1 & \dots & b_s \end{array} & & \\
(1) & (2) & (3)
\end{array}$$
  

$$\begin{array}{ccc}
\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ \hline 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array} & \begin{array}{c|cc} 0 & 1/2 & -1/2 \\ \hline 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} & \\
(4) & (5) & 
\end{array}$$

**Fig. 7.13.** General form of a Butcher tableau (1) and specific Butcher tableaus for the explicit Euler method (2), the implicit Euler method (3), the Radau method (4), and the Lobatto method (5).

Setting

$$\mathbf{z}_i = \mathbf{v}_i - \mathbf{x}_k^\delta = \tau_k \sum_{j=1}^s a_{ij} \mathbf{K}(\mathbf{v}_j)^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{v}_j)],$$

and using the linearization

$$\mathbf{F}(\mathbf{v}_j) = \mathbf{F}(\mathbf{x}_k^\delta + \mathbf{z}_j) \approx \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}(\mathbf{x}_k^\delta) \mathbf{z}_j,$$

and the approximation

$$\mathbf{K}(\mathbf{v}_j) = \mathbf{K}(\mathbf{x}_k^\delta + \mathbf{z}_j) \approx \mathbf{K}(\mathbf{x}_k^\delta),$$

we obtain

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \tau_k \sum_{i=1}^s b_i \mathbf{K}_k^T (\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{z}_i), \quad (7.31)$$

$$\mathbf{z}_i = \tau_k \sum_{j=1}^s a_{ij} \mathbf{K}_k^T (\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{z}_j), \quad i = 1, \dots, s, \quad k = 0, 1, \dots, \quad (7.32)$$

with  $\mathbf{K}_k = \mathbf{K}(\mathbf{x}_k^\delta)$  and  $\mathbf{r}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta)$ . To express (7.31) and (7.32) in a compact form we introduce the matrices

$$\mathbf{A} = \mathbf{A} \otimes \mathbf{I}_n, \quad \mathbf{K}_k = \mathbf{I}_s \otimes \mathbf{K}_k, \quad \mathbf{B} = \mathbf{b}^T \otimes \mathbf{I}_n, \quad \mathbf{I} = \mathbf{I}_s \otimes \mathbf{I}_n,$$

and the vectors

$$\mathbf{r}_k^\delta = \mathbf{1}_s \otimes \mathbf{r}_k^\delta, \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_s \end{bmatrix} \in \mathbb{R}^{sn},$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \dots & a_{ss} \end{bmatrix} \in \mathbb{R}^{s \times s}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_s \end{bmatrix} \in \mathbb{R}^s, \quad \mathbf{1}_s = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^s,$$

and the notation  $\mathbf{X} \otimes \mathbf{Y}$  stands for the Kronecker product of the matrices  $\mathbf{X} \in \mathbb{R}^{m \times n}$  and  $\mathbf{Y} \in \mathbb{R}^{p \times q}$  defined as

$$\mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11}\mathbf{Y} & \dots & x_{1n}\mathbf{Y} \\ \vdots & \ddots & \vdots \\ x_{m1}\mathbf{Y} & \dots & x_{mn}\mathbf{Y} \end{bmatrix} \in \mathbb{R}^{mp \times nq}, \quad [\mathbf{X}]_{ij} = x_{ij}.$$

The use of the Kronecker product enables us to derive a transparent solution representation in a straightforward manner. When working with the Kronecker product, the following calculation rules have to be taken into account: for compatible matrices  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{Z}$  and  $\mathbf{W}$ , there hold

$$(\mathbf{X} \otimes \mathbf{Y})(\mathbf{Z} \otimes \mathbf{W}) = \mathbf{XZ} \otimes \mathbf{YW}, \quad (7.33)$$

$$(\mathbf{X} \otimes \mathbf{Y})^T = \mathbf{X}^T \otimes \mathbf{Y}^T, \quad (7.34)$$

$$(\mathbf{X} \otimes \mathbf{Y})^{-1} = \mathbf{X}^{-1} \otimes \mathbf{Y}^{-1}. \quad (7.35)$$

Moreover, if  $\mathbf{A} = \mathbf{A} \otimes \mathbf{I}_n$  with  $\mathbf{A} \in \mathbb{R}^{s \times s}$  and  $\mathbf{X} = \mathbf{I}_s \otimes \mathbf{X}$  with  $\mathbf{X} \in \mathbb{R}^{n \times n}$ , then the representations

$$\mathbf{AX} = (\mathbf{A} \otimes \mathbf{I}_n)(\mathbf{I}_s \otimes \mathbf{X}) = \mathbf{A} \otimes \mathbf{X}$$

and

$$\mathbf{XA} = (\mathbf{I}_s \otimes \mathbf{X})(\mathbf{A} \otimes \mathbf{I}_n) = \mathbf{A} \otimes \mathbf{X},$$

yield the symmetry relation

$$\mathbf{AX} = \mathbf{XA}. \quad (7.36)$$

Now, using the consistency relation (7.30), (7.31) and (7.32) become

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \tau_k \mathbf{K}_k^T \mathbf{r}_k^\delta - \tau_k \mathbf{BK}_k^T \mathbf{K}_k \mathbf{z}, \quad (7.37)$$

$$(\tau_k \mathbf{AK}_k^T \mathbf{K}_k + \mathbf{I}) \mathbf{z} = \tau_k \mathbf{AK}_k^T \mathbf{r}_k^\delta. \quad (7.38)$$

Equation (7.38) is solved for  $\mathbf{z}$ ,

$$\mathbf{z} = \tau_k (\tau_k \mathbf{AK}_k^T \mathbf{K}_k + \mathbf{I})^{-1} \mathbf{AK}_k^T \mathbf{r}_k^\delta,$$

and is rearranged in the form

$$\mathbf{BK}_k^T \mathbf{K}_k \mathbf{z} = \mathbf{BK}_k^T \mathbf{r}_k^\delta - \mathbf{BA}^{-1} (\tau_k \mathbf{AK}_k^T \mathbf{K}_k + \mathbf{I})^{-1} \mathbf{AK}_k^T \mathbf{r}_k^\delta. \quad (7.39)$$

Since  $\mathbf{b}^T \mathbf{1}_s = 1$  and  $\mathbf{X} = \mathbf{1} \otimes \mathbf{X}$ , we have

$$\mathbf{BK}_k^T \mathbf{r}_k^\delta = (\mathbf{b}^T \otimes \mathbf{I}_n) (\mathbf{I}_s \otimes \mathbf{K}_k^T) (\mathbf{1}_s \otimes \mathbf{r}_k^\delta) = \mathbf{K}_k^T \mathbf{r}_k^\delta, \quad (7.40)$$

and by virtue of (7.39) and (7.40), (7.37) can be written as

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \tau_k \mathbf{BA}^{-1} (\tau_k \mathbf{AK}_k^T \mathbf{K}_k + \mathbf{I})^{-1} \mathbf{AK}_k^T \mathbf{r}_k^\delta.$$

Finally, introducing the regularization parameter  $\alpha_k$  by

$$\alpha_k = \frac{1}{\tau_k},$$



and using the symmetry relation (cf. (7.36) and the identity  $\mathbf{K}_k^T \mathbf{K}_k = \mathbf{I}_s \otimes \mathbf{K}_k^T \mathbf{K}_k$ )

$$\mathbf{A} (\mathbf{K}_k^T \mathbf{K}_k) = (\mathbf{K}_k^T \mathbf{K}_k) \mathbf{A},$$

which yields,

$$\mathbf{A}^{-1} (\mathbf{A} \mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{I})^{-1} \mathbf{A} = (\mathbf{A} \mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{I})^{-1},$$

we obtain the iteration of the Runge–Kutta regularization method

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{B} (\mathbf{A} \mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{I})^{-1} \mathbf{K}_k^T \mathbf{r}_k^\delta, \quad k = 0, 1, \dots \quad (7.41)$$

It is remarkable to note that for the explicit Euler iteration ( $s = 1, a_{11} = 0, b_1 = 1$ ) we are led to  $\mathbf{z}_1 = \mathbf{0}$ , and (7.31) is the iteration of the nonlinear Landweber method (with a relaxation parameter  $\tau_k$ ). Furthermore, for the implicit Euler method ( $s = 1, a_{11} = 1, b_1 = 1$ ) there holds

$$\mathbf{A} = \mathbf{B} = \mathbf{I}_n, \quad \mathbf{K}_k = \mathbf{K}_k, \quad \mathbf{r}_k^\delta = \mathbf{r}_k^\delta,$$

and (7.41) is the iteration of the regularizing Levenberg–Marquardt method with  $\mathbf{L} = \mathbf{I}_n$ , i.e.,

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + (\mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{I}_n)^{-1} \mathbf{K}_k^T \mathbf{r}_k^\delta, \quad k = 0, 1, \dots \quad (7.42)$$

The regularizing property of any inversion method discussed up to now is reflected by the filter factors. This concept can be generalized by introducing the so-called filter matrix. For example, if  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of the matrix  $\mathbf{K}_k$ , then the iterate of the regularizing Levenberg–Marquardt method (7.42) can be expressed as

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{V} \mathbf{F}_k \begin{bmatrix} \frac{1}{\sigma_1} \mathbf{u}_1^T \mathbf{r}_k^\delta \\ \vdots \\ \frac{1}{\sigma_n} \mathbf{u}_n^T \mathbf{r}_k^\delta \end{bmatrix}, \quad (7.43)$$

where the diagonal matrix

$$\mathbf{F}_k = \left[ \text{diag} (f_{\alpha_k} (\sigma_i^2))_{n \times n} \right], \quad f_k (\sigma_i^2) = \frac{\sigma_i^2}{\sigma_i^2 + \alpha_k}, \quad (7.44)$$

represents the filter matrix. Evidently, for very small values of the regularization parameter,  $\mathbf{F}_k \approx \mathbf{I}_n$ , while for very large values of the regularization parameter  $\mathbf{F}_k \approx (1/\alpha_k) [\text{diag} (\sigma_i^2)_{n \times n}]$ . For the Runge–Kutta regularization method, the filter matrix is not diagonal because  $\mathbf{A}$  is not diagonal. To derive the expression of the filter matrix in this case, we first employ the relations (cf. (7.33))

$$\mathbf{A} \mathbf{K}_k^T \mathbf{K}_k = \mathbf{A} \otimes (\mathbf{K}_k^T \mathbf{K}_k) = (\mathbf{I}_s \otimes \mathbf{V}) \left( \mathbf{A} \otimes [\text{diag} (\sigma_i^2)_{n \times n}] \right) (\mathbf{I}_s \otimes \mathbf{V}^T)$$

and

$$\alpha_k \mathbf{I} = (\mathbf{I}_s \otimes \mathbf{V}) (\alpha_k \mathbf{I}) (\mathbf{I}_s \otimes \mathbf{V}^T)$$

to obtain

$$\mathbf{A} \mathbf{K}_k^T \mathbf{K}_k + \alpha_k \mathbf{I} = (\mathbf{I}_s \otimes \mathbf{V}) \left( \mathbf{A} \otimes [\text{diag} (\sigma_i^2)_{n \times n}] + \alpha_k \mathbf{I} \right) (\mathbf{I}_s \otimes \mathbf{V}^T).$$

Then, we use

$$\mathbf{K}_k^T \mathbf{r}_k^\delta = \mathbf{1}_s \otimes (\mathbf{K}_k^T \mathbf{r}_k^\delta) = (\mathbf{I}_s \otimes \mathbf{V}) \left( \mathbf{I}_s \otimes \left[ \text{diag}(\sigma_i^2)_{n \times n} \right] \right) \left( \mathbf{1}_s \otimes \begin{bmatrix} \frac{1}{\sigma_1} \mathbf{u}_1^T \mathbf{r}_k^\delta \\ \vdots \\ \frac{1}{\sigma_n} \mathbf{u}_n^T \mathbf{r}_k^\delta \end{bmatrix} \right),$$

and

$$\mathbf{B}(\mathbf{I}_s \otimes \mathbf{V}) = (\mathbf{b}^T \otimes \mathbf{I}_n) (\mathbf{I}_s \otimes \mathbf{V}) = \mathbf{b}^T \otimes \mathbf{V},$$

together with (cf. (7.35))

$$(\mathbf{I}_s \otimes \mathbf{V}^T)^{-1} = \mathbf{I}_s \otimes \mathbf{V},$$

to conclude that

$$\begin{aligned} \mathbf{x}_{k+1}^\delta &= \mathbf{x}_k^\delta + (\mathbf{b}^T \otimes \mathbf{V}) \left( \mathbf{A} \otimes \left[ \text{diag}(\sigma_i^2)_{n \times n} \right] + \alpha_k \mathbf{I} \right)^{-1} \\ &\quad \times \left( \mathbf{I}_s \otimes \left[ \text{diag}(\sigma_i^2)_{n \times n} \right] \right) \left( \mathbf{1}_s \otimes \begin{bmatrix} \frac{1}{\sigma_1} \mathbf{u}_1^T \mathbf{r}_k^\delta \\ \vdots \\ \frac{1}{\sigma_n} \mathbf{u}_n^T \mathbf{r}_k^\delta \end{bmatrix} \right). \end{aligned} \quad (7.45)$$

The iterate (7.45) can be expressed as in (7.43) by taking into account that  $\mathbf{X} = \mathbf{1} \otimes \mathbf{X}$  and  $\mathbf{x} = \mathbf{1} \otimes \mathbf{x}$ . The result is

$$\mathbf{F}_k = (\mathbf{b}^T \otimes \mathbf{I}_n) \left( \mathbf{A} \otimes \left[ \text{diag}(\sigma_i^2)_{n \times n} \right] + \alpha_k \mathbf{I} \right)^{-1} \left( \mathbf{1}_s \otimes \left[ \text{diag}(\sigma_i^2)_{n \times n} \right] \right).$$

Two extreme situations can be considered. If  $\alpha_k$  is very small, then by virtue of the identity  $\mathbf{b}^T \mathbf{A}^{-1} \mathbf{1}_s = 1$ , which holds true for the Radau and Lobatto methods illustrated in Figure 7.13, we obtain  $\mathbf{F}_k \approx \mathbf{I}_n$ . On the other hand, if  $\alpha_k$  is very large, the consistency relation  $\mathbf{b}^T \mathbf{1}_s = 1$ , yields  $\mathbf{F}_k \approx (1/\alpha_k) [\text{diag}(\sigma_i^2)_{n \times n}]$ . Thus, the filter matrix of the Runge–Kutta regularization method behaves like the ‘Tikhonov filter matrix’.

The exponential Euler method is based on the variation-of-constants formula which allows us to integrate the linear part of semilinear differential equations exactly. For the Showalter differential equation (7.29), Hochbruck et al. (2009) proposed the following modification of the original exponential Euler scheme:

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \tau_k \varphi(-\tau_k \mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T \mathbf{r}_k^\delta, \quad k = 0, 1, \dots,$$

with

$$\varphi(z) = \frac{\mathbf{e}^z - 1}{z}.$$

Assuming the singular value decomposition  $\mathbf{K}_k = \mathbf{U} \Sigma \mathbf{V}^T$  and setting  $\alpha_k = 1/\tau_k$ , the matrix function can be expressed as

$$\varphi(-\alpha_k^{-1} \mathbf{K}_k^T \mathbf{K}_k) = \alpha_k \mathbf{V} \left[ \text{diag} \left( \frac{1 - \exp\left(-\frac{\sigma_i^2}{\alpha_k}\right)}{\sigma_i^2} \right)_{n \times n} \right] \mathbf{V}^T, \quad (7.46)$$

and the iteration takes the form

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \sum_{i=1}^n \left[ 1 - \exp\left(-\frac{\sigma_i^2}{\alpha_k}\right) \right] \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{r}_k^\delta) \mathbf{v}_i, \quad k = 0, 1, \dots \quad (7.47)$$

The exponential Euler regularization method is very similar to the regularizing Levenberg–Marquardt method in which the Tikhonov filter factors (7.44) are replaced by the filter factors

$$f_k(\sigma_i^2) = 1 - \exp\left(-\frac{\sigma_i^2}{\alpha_k}\right).$$

Obviously, the filter factors for the exponential Euler regularization method are close to 1 for large  $\sigma_i$  and much smaller than 1 for small  $\sigma_i$ .

The algorithmic implementation of asymptotic regularization methods resembles that of the regularizing Levenberg–Marquardt method. The main features are as follows:

- (1) the iterations (7.41) and (7.47) are applied to the standard-form problem;
- (2) the regularization parameters are chosen as the terms of a decreasing sequence  $\alpha_k = q_k \alpha_{k-1}$  with constant or variable ratio  $q_k$ ;
- (3) a step-length procedure for the nonlinear residual is used to improve the stability of the method.

Note that the step-length procedure can be used because the Newton step  $\mathbf{p}_k^\delta$  can be expressed as  $\mathbf{p}_k^\delta = \hat{\mathbf{G}}_k \mathbf{K}_k^T \mathbf{r}_k^\delta$ , where  $\hat{\mathbf{G}}_k$  is a positive definite matrix; for example, in the exponential Euler regularization method, we have  $\hat{\mathbf{G}}_k = \alpha_k^{-1} \varphi(-\alpha_k^{-1} \mathbf{K}_k^T \mathbf{K}_k)$ , with  $\varphi(-\alpha_k^{-1} \mathbf{K}_k^T \mathbf{K}_k)$  being given by (7.46).

The numerical performance of asymptotic regularization methods and of the regularizing Levenberg–Marquardt are comparable; for large initial values of the regularization parameters, the solution errors as well as the number of iteration steps are similar (Figure 7.14).

The asymptotic regularization methods yield results of comparable accuracies, although the solution errors given in Table 7.3 indicate a slight superiority of the Radau regularization method, especially for the  $O_3$  retrieval test problem.

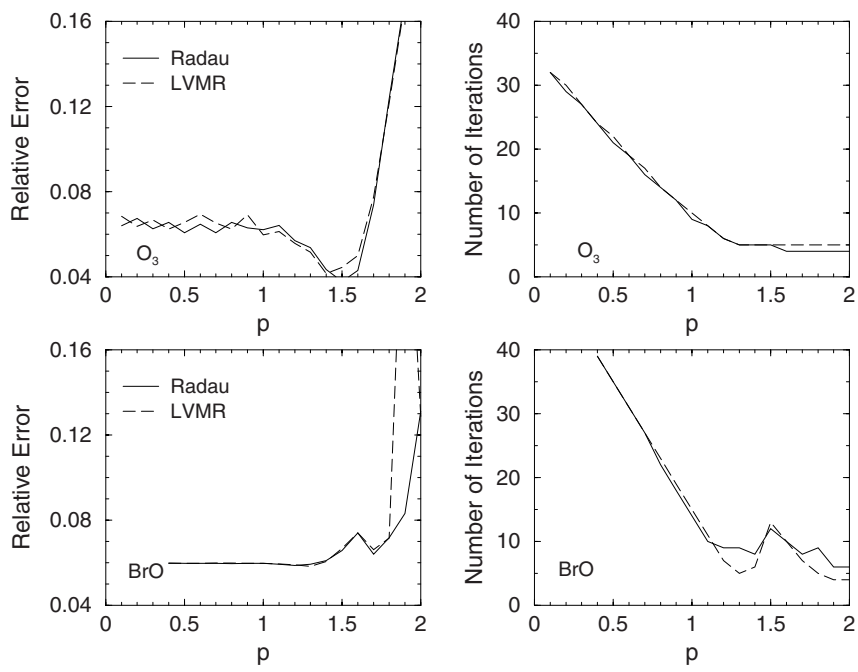
## 7.4 Mathematical results and further reading

The convergence of the nonlinear Landweber iteration is expressed by the following result (Hanke et al., 1995): if  $x^\dagger$  is a solution of the equation  $F(x) = y$  in the ball  $\mathcal{B}_\rho(x_a)$  of radius  $\rho$  about  $x_a$ ,  $F$  has the local property

$$\|F(x) - F(x') - F'(x')(x - x')\| \leq \eta \|F(x) - F(x')\|, \quad 0 < \eta < \frac{1}{2}, \quad (7.48)$$

for all  $x, x' \in \mathcal{B}_{2\rho}(x_a)$ , and the equation  $F(x) = y$  is properly scaled in the sense that

$$\|F'(x)\| \leq 1, \quad x \in \mathcal{B}_{2\rho}(x_a),$$



**Fig. 7.14.** Relative solution errors and the number of iteration steps for the Radau regularization method and the regularizing Levenberg–Marquardt (LVMR) method.

**Table 7.3.** Relative solution errors for Radau, Lobatto and exponential Euler regularization methods. The initial value of the regularization parameter is  $\alpha = \sigma^p$ .

Problem	Method	$p$	Selection criterion		
			S1	S2	S3
$O_3$	Radau	1.5	3.78e-2	3.78e-2	4.11e-2
	Lobatto		4.19e-2	4.19e-2	4.73e-2
	Euler		4.08e-2	4.08e-2	4.51e-2
$BrO$	Radau	1.2	5.88e-2	7.01e-2	5.82e-2
	Lobatto		5.87e-2	6.89e-2	6.92e-2
	Euler		5.88e-2	6.91e-2	6.09e-2
$CO$	Radau	1.0	3.72e-2	31.9e-2	3.75e-2
	Lobatto		3.79e-2	31.5e-2	3.59e-2
	Euler		3.78e-2	34.4e-2	3.27e-2
Temperature	Radau	0.9	1.80e-2	2.07e-2	2.06e-2
	Lobatto		1.80e-2	2.13e-2	2.10e-2
	Euler		1.80e-2	2.17e-2	2.07e-2

then  $x_{k^*}^\delta \rightarrow x^\dagger$  as  $\Delta \rightarrow 0$ , where  $k^* = k^*(\Delta)$  is the stopping index of the discrepancy principle

$$\|y^\delta - F(x_{k^*}^\delta)\| \leq \tau_{\text{dp}} \Delta < \|y^\delta - F(x_k^\delta)\|, \quad 0 \leq k < k^*,$$

and

$$\tau_{\text{dp}} > 2 \frac{1 + \eta}{1 - 2\eta} > 2.$$

In contrast to Tikhonov regularization, the source condition

$$x^\dagger - x_a = \left[ F'(x^\dagger)^* F'(x^\dagger) \right]^\mu z, \quad (7.49)$$

with  $\mu > 0$  and  $z \in X$ , is not sufficient to obtain convergence rates. In Hanke et al. (1995), the convergence rate  $O(\Delta^{2\mu/(2\mu+1)})$  with  $0 < \mu \leq 1/2$ , has been proven under the additional assumption that, for all  $x \in \mathcal{B}_{2\rho}(x_a)$ ,  $F$  satisfies

$$\begin{aligned} F'(x) &= R_x F'(x^\dagger), \\ \|I - R_x\| &\leq c_R \|x - x^\dagger\|, \quad c_R > 0, \end{aligned} \quad (7.50)$$

where  $\{R_x / x \in \mathcal{B}_{2\rho}(x_a)\}$  is a family of bounded linear operators  $R_x : Y \rightarrow Y$ .

The iteratively regularized Gauss–Newton method was introduced by Bakushinsky. In Bakushinsky (1992) local convergence was proven under the source condition (7.49) with  $\mu \geq 1$ , provided that  $F'$  is Lipschitz continuous, i.e.,

$$\|F'(x) - F'(x')\| \leq L \|x - x'\|, \quad L > 0,$$

for all  $x, x' \in \mathcal{B}_{2\rho}(x_a)$ . Lipschitz continuity of  $F'$  suffices to prove convergence rates for the case  $\mu \geq 1/2$ , but if  $\mu < 1/2$  further conditions, that guarantee that the linearization is not too far away from the nonlinear operator, are required. In Blaschke et al. (1997), the convergence rates

$$\|x_{k^*}^\delta - x^\dagger\| = \begin{cases} o\left(\Delta^{\frac{2\mu}{2\mu+1}}\right), & 0 < \mu < 1/2, \\ O\left(\sqrt{\Delta}\right), & \mu = 1/2, \end{cases} \quad (7.51)$$

with  $k^* = k^*(\Delta)$  being the stopping index of the discrepancy principle, have been derived by assuming the following restrictions on the nonlinearity of  $F$ :

$$\begin{aligned} F'(x) &= R(x, x') F'(x') + Q(x, x'), \\ \|I - R(x, x')\| &\leq c_R, \\ \|Q(x, x')\| &\leq c_Q \|F'(x^\dagger)(x - x')\|, \quad c_R, c_Q > 0, \end{aligned} \quad (7.52)$$

for all  $x, x' \in \mathcal{B}_{2\rho}(x_a)$ . Similarly, the optimal error bound  $O(\Delta^{2\mu/(2\mu+1)})$  for  $0 < \mu < 1/2$  has been proven by Bauer and Hohage (2005) for the Lepskij stopping rule and the nonlinearity assumptions (7.52). As the best convergence rate of the discrepancy principle is  $O(\sqrt{\Delta})$ , the generalized discrepancy principle

$$\alpha_{k^*} \left\langle y^\delta - F(x_{k^*}^\delta), \left[ F'(x_{k^*}^\delta) F'(x_{k^*}^\delta)^* + \alpha_{k^*} I \right]^{-1} [y^\delta - F(x_{k^*}^\delta)] \right\rangle \leq \tau \Delta^2, \quad \tau > 1,$$

has been considered in Jin (2000), where the optimal convergence rate  $O(\Delta^{2\mu/(2\mu+1)})$  with  $0 < \mu \leq 1$  has been established under the nonlinearity assumptions:

$$\begin{aligned} [F'(x) - F'(x')]z &= F'(x')h(x, x', z), \\ \|h(x, x', z)\| &\leq c_R \|x - x'\| \|z\|, \quad c_R > 0, \end{aligned}$$

for all  $x, x' \in \mathcal{B}_\rho(x^\dagger)$ .

Results on convergence rates under logarithmic source conditions can be found in Hohage (1997) for the iteratively regularized Gauss–Newton method, and in Deuffhard et al. (1998) for the nonlinear Landweber iteration.

For a general regularization method of the form

$$x_{k+1}^\delta = x_a + g_{\alpha_k} \left( F'(x_k^\delta)^* F'(x_k^\delta) \right) F'(x_k^\delta)^* [y^\delta - F(x_k^\delta) + F'(x_k^\delta)(x_k^\delta - x_a)], \quad (7.53)$$

the convergence rates (7.51) have been derived by Kaltenbacher (1997, 1998) for the modified discrepancy principle

$$\max(\|y^\delta - F(x_{k^*}^\delta)\|, r_{1k^*}) \leq \tau_{\text{dp}} \Delta < \max(\|y^\delta - F(x_{k-1}^\delta)\|, r_{1k}), \quad 1 \leq k < k^*, \quad (7.54)$$

with

$$r_{1k} = y^\delta - F(x_{k-1}^\delta) - F'(x_{k-1}^\delta)(x_k^\delta - x_{k-1}^\delta),$$

provided that  $\tau_{\text{dp}} > 1$  is sufficiently large, the nonlinearity conditions (7.52) hold, and the sequence  $\{\alpha_k\}$  satisfies (7.7). Note that the stopping rule (7.54) is essentially equivalent to the termination criterion

$$\begin{aligned} \max(\|y^\delta - F(x_{k^*}^\delta)\|, \|y^\delta - F(x_{k^*}^\delta)\|) \\ \leq \tau'_{\text{dp}} \Delta < \max(\|y^\delta - F(x_{k-1}^\delta)\|, \|y^\delta - F(x_k^\delta)\|), \quad 1 \leq k < k^*, \end{aligned}$$

which stops the iteration as soon as the residual norms at two subsequent iteration steps are below  $\tau'_{\text{dp}} \Delta$ . Examples of iterative methods having the form (7.53) are the iteratively regularized Gauss–Newton method with

$$g_\alpha(\lambda) = \frac{1}{\lambda + \alpha}$$

and the Newton–Landweber iteration with

$$g_\alpha(\lambda) = \frac{1}{\lambda} [1 - (1 - \lambda)^p], \quad \alpha = \frac{1}{p}.$$

Hanke (1997) established the convergence of the regularizing Levenberg–Marquardt method by using the local nonlinearity assumption

$$\|F(x) - F(x') - F'(x')(x - x')\| \leq c \|x - x'\| \|F(x) - F(x')\|, \quad c > 0,$$

for all  $x, x' \in \mathcal{B}_{2\rho}(x_a)$ , and by choosing the regularization parameter  $\alpha_k$  as the solution of the ‘discrepancy principle’ equation (cf. (7.19))

$$\|y^\delta - F(x_k^\delta) - F'(x_k^\delta)[x_{k+1}^\delta(\alpha) - x_k^\delta]\| = \theta \|y^\delta - F(x_k^\delta)\|,$$

for some  $\theta \in (0, 1)$ .

The regularizing trust-region method was analyzed by Wang and Yuan (2005). Convergence results have been proven under the nonlinearity assumption (7.48) with  $0 < \eta < 1$ , provided that the iterative process is stopped according to the discrepancy principle with

$$\tau_{\text{dp}} > \frac{1 + \eta}{1 - \eta}.$$

Convergence rates for the regularized inexact Newton iteration method

$$x_{k+1}^\delta = x_k^\delta + g_{\alpha_k} \left( F' (x_k^\delta)^* F' (x_k^\delta) \right) F' (x_k^\delta)^* [y^\delta - F (x_k^\delta)], \quad (7.55)$$

and the source condition (7.49), have been established by Rieder (1999, 2003). The general iteration method (7.55) includes the regularizing Levenberg–Marquardt method, and Newton-type methods using as inner iteration the CGNR method, the Landweber iteration and the  $\nu$ -method.

The convergence of the Runge–Kutta regularization method has been proven by Böckmann and Pornsawad (2008) under the nonlinearity assumption (7.48).

The recent monograph by Kaltenbacher et al. (2008) provides an exhaustive and pertinent analysis of iterative regularization methods for nonlinear ill-posed problems. In addition to the methods discussed in this chapter, convergence and convergence rate results can be found for the modified Landweber methods (iteratively regularized Landweber iteration, Landweber–Kaczmarz method), Broyden’s method, multilevel methods and level set methods.

In Appendix H we derive convergence rate results for the general regularization methods (7.53) and (7.55) in a discrete setting. The regularization scheme (7.53) corresponds to the so-called Newton-type methods with a priori information, e.g., the iteratively regularized Gauss–Newton method, while the regularization scheme (7.55) corresponds to the Newton-type methods without a priori information, e.g., the regularizing Levenberg–Marquardt method.

# 8

## Total least squares

In atmospheric remote sensing, near real-time software processors frequently use approximations of the Jacobian matrix in order to speed up the calculation. If the forward model  $\mathbf{F}(\mathbf{x})$  depends on the state vector  $\mathbf{x}$  through some model parameters  $\mathbf{b}_k$ ,

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{b}_1(\mathbf{x}), \dots, \mathbf{b}_N(\mathbf{x})),$$

then, an approximate expression of the Jacobian matrix

$$\mathbf{K} = \sum_{k=1}^N \frac{\partial \mathbf{F}}{\partial \mathbf{b}_k} \frac{\partial \mathbf{b}_k}{\partial \mathbf{x}},$$

can be obtained by assuming that some  $\mathbf{b}_k$  are insensitive to  $\mathbf{x}$ , i.e.,  $\partial \mathbf{b}_k / \partial \mathbf{x} = \mathbf{0}$ . For example, the limb radiance measured by a detector in the ultraviolet or visible spectral domains can be expressed as

$$I(\lambda, \mathbf{x}) = I_{ss}(\lambda, \mathbf{x}) + I_{ms}(\lambda, \mathbf{x}) = I_{ss}(\lambda, \mathbf{x}) [1 + c_{ms}(\lambda, \mathbf{x})], \quad (8.1)$$

where  $I_{ss}$  and  $I_{ms}$  are the single and multiple scattering terms,  $\lambda$  is the wavelength, and  $c_{ms}$  is a correction factor accounting for the multiple scattering contribution. As the computation of the derivative of  $c_{ms}$  is quite demanding, the Jacobian matrix calculation may involve only the derivative of  $I_{ss}$ . Similarly, in a line-by-line model, the absorption coefficient  $C_{absm}$  of the gas molecule  $m$  is the product of the line strength  $S_{ml}$  and the normalized line shape function  $g_{ml}$  (cf. (1.12)),

$$C_{absm}(\nu, T) = \sum_l S_{ml}(T) g_{ml}(\nu, T),$$

where  $\nu$  is the wavenumber,  $T$  is the temperature, and the summation is over all lines  $l$ . As the most important temperature dependence stems from the line strength, the derivative of the line shape function with respect to the temperature is sometimes ignored.

The total least squares (TLS) method is devoted to the solution of linear problems in which both the coefficient matrix and the data are subject to errors. The linear data model can be expressed as

$$\mathbf{y}^\delta = (\mathbf{K}_\Lambda - \mathbf{\Lambda}) \mathbf{x} + \boldsymbol{\delta},$$



where the matrix  $\mathbf{K}_\Lambda$  is a perturbation of the exact (unknown) matrix  $\mathbf{K}$ ,  $\mathbf{K}_\Lambda = \mathbf{K} + \Lambda$ , and the data are affected by the instrumental noise  $\delta$ .

The TLS method was independently derived in several bodies of work by Golub and Van Loan (1980, 1996), and Van Huffel and Vanderwalle (1991). This literature has advanced the algorithmic and theoretical understanding of the method, as well as its application for computing stable solutions of linear systems of equations with highly ill-conditioned coefficient matrices. In this section we review the truncated and the regularized TLS methods for solving linear ill-posed problems, and reveal the similarity with the Tikhonov regularization. We then present a first attempt to extend the regularized TLS to nonlinear ill-posed problems.

### 8.1 Formulation

The linear model which encapsulates the uncertainties in the data vector and the coefficient matrix is of the form  $\mathbf{K}_\Lambda \mathbf{x} \approx \mathbf{y}^\delta$ . To sketch the TLS method, we introduce the augmented matrix  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  and consider the homogeneous system of equations

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}. \quad (8.2)$$

We then assume a singular value decomposition of the  $m \times (n+1)$  matrix,

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} = \bar{\mathbf{U}} \bar{\Sigma} \bar{\mathbf{V}}^T, \quad (8.3)$$

and partition the matrices  $\bar{\mathbf{V}}$  and  $\bar{\Sigma}$  as follows:

$$\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_{n+1}] = \begin{bmatrix} \bar{\mathbf{V}}_{11} & \bar{\mathbf{v}}_{12} \\ \bar{\mathbf{v}}_{21}^T & \bar{v}_{22} \end{bmatrix}, \quad \bar{\mathbf{V}}_{11} \in \mathbb{R}^{n \times n}, \quad \bar{\mathbf{v}}_{12}, \bar{\mathbf{v}}_{21} \in \mathbb{R}^n, \quad (8.4)$$

and

$$\bar{\Sigma} = \begin{bmatrix} \bar{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \bar{\sigma}_{n+1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \bar{\Sigma}_1 = [\text{diag}(\bar{\sigma}_i)_{n \times n}],$$

respectively. If  $\bar{\sigma}_{n+1} \neq 0$ , then  $\text{rank}(\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}) = n+1$ , and the solution of the homogeneous system of equations (8.2) is the trivial solution. Thus, the last component of the solution vector is not  $-1$ , and to solve (8.2) it is necessary to reduce the rank of the augmented matrix from  $n+1$  to  $n$ . This can be achieved by approximating the rank- $(n+1)$  matrix  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  by a rank- $n$  matrix  $\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix}$ . As  $\text{rank}(\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix}) = n$ , we may assume that the last column vector of the matrix  $\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix}$  is a linear combination of the first  $n$  column vectors, i.e.,

$$\mathbf{y}_n = \sum_{i=1}^n x_i \mathbf{k}_i,$$

with  $\mathbf{K}_n = [\mathbf{k}_1, \dots, \mathbf{k}_n]$ , or equivalently that,

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

with  $\mathbf{x} = [x_1, \dots, x_n]^T$ . The (matrix) approximation problem can be expressed as the constrained minimization problem

$$\begin{aligned} \min_{[\tilde{\mathbf{K}} \tilde{\mathbf{y}}] \in \mathbb{R}^{m \times (n+1)}} & \left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{K}} & \tilde{\mathbf{y}} \end{bmatrix} \right\|_F^2 \\ \text{subject to } & \tilde{\mathbf{K}}\mathbf{x} = \tilde{\mathbf{y}}, \end{aligned} \quad (8.5)$$

where the Frobenius norm of the  $m \times n$  matrix  $\mathbf{A}$  is defined by

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n [\mathbf{A}]_{ij}^2}.$$

It should be pointed out that the ordinary least squares method minimizes the norm of the residual vector  $\mathbf{y}^\delta - \tilde{\mathbf{y}}$  under the assumption that  $\mathbf{K}_\Lambda = \tilde{\mathbf{K}}$ .

The solution to the minimization problem (8.5) is given by the Eckart–Young–Mirsky theorem (Golub and Van Loan, 1996): the matrix

$$\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} = \sum_{i=1}^n \bar{\sigma}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^T \quad (8.6)$$

is the closest rank- $n$  matrix to  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$ , and we have

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} = \bar{\sigma}_{n+1} \bar{\mathbf{u}}_{n+1} \bar{\mathbf{v}}_{n+1}^T,$$

yielding

$$\left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} \right\|_F = \bar{\sigma}_{n+1}.$$

The homogeneous system of equations (8.2) is then replaced by a homogeneous system of equations involving the rank- $n$  matrix  $\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix}$ , that is,

$$\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}. \quad (8.7)$$

Since (cf. (8.6))

$$\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} \bar{\mathbf{v}}_{n+1} = \sum_{i=1}^n \bar{\sigma}_i (\bar{\mathbf{v}}_i^T \bar{\mathbf{v}}_{n+1}) \bar{\mathbf{u}}_i = \mathbf{0}, \quad (8.8)$$

we see that the vector  $a\bar{\mathbf{v}}_{n+1}$  is the general solution of the homogeneous system of equations (8.7) and that the scalar  $a$  is (uniquely) determined by imposing that the last component of the solution vector is  $-1$ . We obtain

$$\begin{bmatrix} \mathbf{x}_\Lambda^\delta \\ -1 \end{bmatrix} = -\frac{1}{[\bar{\mathbf{v}}_{n+1}]_{n+1}} \bar{\mathbf{v}}_{n+1}, \quad (8.9)$$

provided that  $[\bar{\mathbf{v}}_{n+1}]_{n+1} \neq 0$ . From (8.4), we find that the TLS solution can be expressed as

$$\mathbf{x}_\Lambda^\delta = -\frac{1}{\bar{v}_{22}} \bar{\mathbf{v}}_{12}. \quad (8.10)$$

Note that if  $\bar{\sigma}_{n+1}$  is a simple singular value, we have (cf. (8.8))  $\mathcal{N}(\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix}) = \text{span}\{\bar{\mathbf{v}}_{n+1}\}$ , and the TLS solution is unique.

## 8.2 Truncated total least squares

The truncated TLS method, which in general is devoted to numerically rank deficient problems, is also a suitable regularization method for discrete ill-posed problems. This technique is similar to the truncated SVD that treats small singular values of  $\mathbf{K}$  as zeros. In both methods, the redundant information in  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  and  $\mathbf{K}$ , respectively, associated to the small singular values, is discarded and the original ill-posed problem with a full rank matrix is replaced by a well-posed problem with a rank-deficient matrix. This approximation is achieved by means of the Eckart–Young–Mirsky theorem. For example, in the truncated SVD, the matrix  $\mathbf{K}$  with  $\text{rank}(\mathbf{K}) = n$  and singular value decomposition

$$\mathbf{K} = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

is replaced by the matrix

$$\mathbf{K}_p = \sum_{i=1}^p \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

with  $\text{rank}(\mathbf{K}_p) = p$ , and the regularized solution takes the form

$$\mathbf{x}_p^\delta = \arg \min_{\mathbf{x}} \|\mathbf{y}^\delta - \mathbf{K}_p \mathbf{x}\|^2 = \sum_{i=1}^p \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i.$$

The major difference between the two methods lies in the way in which the approximation is performed: in the truncated SVD, the modification depends only on  $\mathbf{K}$ , while in the truncated TLS, the modification depends on both  $\mathbf{K}_\Lambda$  and  $\mathbf{y}^\delta$ . Thus, in the framework of the truncated TLS method we approximate the matrix  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  by the rank- $p$  matrix

$$\begin{bmatrix} \mathbf{K}_p & \mathbf{y}_p \end{bmatrix} = \sum_{i=1}^p \bar{\sigma}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^T.$$

To determine the number  $p$  of large singular values or the truncation index, we may require a user-specified threshold or determine  $p$  adaptively. The null space of the approximation matrix is

$$\mathcal{N}(\begin{bmatrix} \mathbf{K}_p & \mathbf{y}_p \end{bmatrix}) = \text{span} \{ \bar{\mathbf{v}}_{p+1}, \dots, \bar{\mathbf{v}}_{n+1} \},$$

whence accounting for the partition

$$\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_{n+1}] = \begin{bmatrix} \bar{\mathbf{V}}_{11} & \bar{\mathbf{V}}_{12} \\ \bar{\mathbf{v}}_{21}^T & \bar{\mathbf{v}}_{22}^T \end{bmatrix}, \quad (8.11)$$

with  $\bar{\mathbf{V}}_{11} \in \mathbb{R}^{n \times p}$ ,  $\bar{\mathbf{V}}_{12} \in \mathbb{R}^{n \times (n-p+1)}$ , and

$$\begin{aligned} \bar{\mathbf{v}}_{21} &= \left[ [\bar{\mathbf{v}}_1]_{n+1}, \dots, [\bar{\mathbf{v}}_p]_{n+1} \right]^T \in \mathbb{R}^p, \\ \bar{\mathbf{v}}_{22} &= \left[ [\bar{\mathbf{v}}_{p+1}]_{n+1}, \dots, [\bar{\mathbf{v}}_{n+1}]_{n+1} \right]^T \in \mathbb{R}^{n-p+1}, \end{aligned}$$

we seek the solution as

$$\begin{bmatrix} \mathbf{x}_{\Lambda p}^\delta \\ -1 \end{bmatrix} = \sum_{i=p+1}^{n+1} a_i \bar{\mathbf{v}}_i = \begin{bmatrix} \bar{\mathbf{V}}_{12} \\ \bar{\mathbf{v}}_{22}^T \end{bmatrix} \mathbf{a}, \quad (8.12)$$

with  $\mathbf{a} = [a_{p+1}, \dots, a_{n+1}]^T \in \mathbb{R}^{n-p+1}$ . From the last equation we find that

$$\bar{\mathbf{v}}_{22}^T \mathbf{a} = -1,$$

or equivalently that

$$\sum_{i=p+1}^{n+1} a_i [\bar{\mathbf{v}}_i]_{n+1} = -1.$$

Since (cf. (8.12))

$$\left\| \begin{bmatrix} \mathbf{x}_{\Lambda p}^\delta \\ -1 \end{bmatrix} \right\|^2 = 1 + \|\mathbf{x}_{\Lambda p}^\delta\|^2 = \sum_{i=p+1}^{n+1} a_i^2, \quad (8.13)$$

we see that the minimum norm solution  $\mathbf{x}_{\Lambda p}^\delta$  requires a minimum value of  $\sum_{i=p+1}^{n+1} a_i^2$ . This can be obtained by solving the constrained minimization problem

$$\begin{aligned} & \min_{a_i} \sum_{i=p+1}^{n+1} a_i^2 \\ & \text{subject to } \sum_{i=p+1}^{n+1} a_i [\bar{\mathbf{v}}_i]_{n+1} = -1. \end{aligned}$$

In the framework of the Lagrange multiplier formalism, the first-order optimality conditions for the Lagrangian function

$$\mathcal{L}(\mathbf{a}, \lambda) = \frac{1}{2} \sum_{i=p+1}^{n+1} a_i^2 + \lambda \left( \sum_{i=p+1}^{n+1} a_i [\bar{\mathbf{v}}_i]_{n+1} + 1 \right),$$

yield

$$\begin{aligned} a_i + \lambda [\bar{\mathbf{v}}_i]_{n+1} &= 0, \quad i = p+1, \dots, n+1, \\ \sum_{i=p+1}^{n+1} a_i [\bar{\mathbf{v}}_i]_{n+1} &= -1, \end{aligned}$$

and we obtain

$$\mathbf{a} = -\frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \bar{\mathbf{v}}_{22}. \quad (8.14)$$

Hence, from (8.12) and (8.14), the minimum norm solution is given by

$$\mathbf{x}_{\Lambda p}^\delta = -\frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \bar{\mathbf{V}}_{12} \bar{\mathbf{v}}_{22}. \quad (8.15)$$

By (8.13), (8.14) and the Eckart–Young–Mirsky theorem, we have

$$\|\mathbf{x}_{\Lambda p}^\delta\|^2 = \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} - 1,$$

and

$$\|\mathbf{R}_{\Lambda p}^\delta\|_F^2 = \left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \mathbf{K}_p & \mathbf{y}_p \end{bmatrix} \right\|_F^2 = \bar{\sigma}_{p+1}^2 + \dots + \bar{\sigma}_{n+1}^2,$$

showing that the solution norm  $\|\mathbf{x}_{\Lambda p}^\delta\|$  increases monotonically with  $p$ , while the residual norm  $\|\mathbf{R}_{\Lambda p}^\delta\|_F$  decreases monotonically with  $p$ . These results recommend the discrepancy principle and the L-curve method for computing the truncation index.

In order to demonstrate the regularizing property of the truncated TLS method, we express  $\mathbf{x}_{\Lambda p}^\delta$  as the filtered sum

$$\mathbf{x}_{\Lambda p}^\delta = \sum_{i=1}^n f_i \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i, \quad (8.16)$$

where  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of  $\mathbf{K}_\Lambda$ . In Appendix I it is shown that if  $\text{rank}(\mathbf{K}_\Lambda) = n$  and  $\text{rank} \left( \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} \right) = n + 1$ , and furthermore, if  $\mathbf{u}_i^T \mathbf{y}^\delta \neq 0$  for all  $i = 1, \dots, n$ , then the filter factors are given by

$$f_i = \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=1}^p \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2, \quad (8.17)$$

and the estimates

$$1 < f_i \leq 1 + \left( \frac{\bar{\sigma}_{p+1}}{\sigma_i} \right)^2 + O \left( \frac{\bar{\sigma}_{p+1}^4}{\sigma_i^4} \right), \quad i = 1, \dots, p, \quad (8.18)$$

and

$$0 < f_i \leq \frac{1 - \|\bar{\mathbf{v}}_{22}\|^2}{\|\bar{\mathbf{v}}_{22}\|^2} \left( \frac{\sigma_i}{\bar{\sigma}_p} \right)^2 \left[ 1 + O \left( \frac{\sigma_i^2}{\bar{\sigma}_p^2} \right) \right], \quad i = p + 1, \dots, n \quad (8.19)$$

hold. From (8.18), (8.19) and the interlacing property of the singular values of  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  and  $\mathbf{K}_\Lambda$ ,

$$\bar{\sigma}_1 > \sigma_1 > \dots > \bar{\sigma}_p > \sigma_p > \bar{\sigma}_{p+1} > \sigma_{p+1} > \dots > \sigma_n > \bar{\sigma}_{n+1},$$

we see that for  $i \ll p$ ,  $(\bar{\sigma}_{p+1}/\sigma_i)^2 \ll 1$  and the filter factors are close to 1, while for  $i \gg p$ ,  $(\sigma_i/\bar{\sigma}_p)^2 \ll 1$  and the filter factors are very small. Thus, the filter factors of the truncated TLS method resemble the Tikhonov filter factors, and  $\mathbf{x}_{\Lambda p}^\delta$  is a filtered solution, with the truncation index  $p$  playing the role of the regularization parameter.

When the dimension of  $\mathbf{K}_\Lambda$  is not too large, the singular value decomposition of the augmented matrix  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  can be computed directly. For large-scale problems, this approach is computationally expensive and an iterative algorithm based on Lanczos bidiagonalization can be used instead (Fierro et al., 1997). The so-called Lanczos truncated TLS

algorithm uses the Lanczos bidiagonalization of the matrix  $\mathbf{K}_\Lambda$  to obtain, after  $p$  iteration steps, the factorization

$$\mathbf{K}_\Lambda \bar{\mathbf{V}}_p = \bar{\mathbf{U}}_{p+1} \mathbf{B}_p, \quad (8.20)$$

and projects the TLS problem onto the subspace spanned by  $\bar{\mathbf{U}}_{p+1} \in \mathbb{R}^{m \times (p+1)}$  and  $\bar{\mathbf{V}}_p \in \mathbb{R}^{n \times p}$ . The projection is a consequence of the assumption that for a sufficiently large  $p$ , all the large singular values of  $\mathbf{K}_\Lambda$ , which contribute to the regularized solution, have been captured. The projected TLS problem reads as

$$\begin{aligned} \min_{[\tilde{\mathbf{K}}_p \tilde{\mathbf{y}}_p] \in \mathbb{R}^{m \times (n+1)}} & \left\| \bar{\mathbf{U}}_{p+1}^T ([\mathbf{K}_\Lambda \quad \mathbf{y}^\delta] - [\tilde{\mathbf{K}}_p \quad \tilde{\mathbf{y}}_p]) \begin{bmatrix} \bar{\mathbf{V}}_p & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \right\|_{\mathbf{F}}^2 \\ \text{subject to } & \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{K}}_p \bar{\mathbf{V}}_p \mathbf{z}_p = \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{y}}_p, \end{aligned}$$

where we have set  $\mathbf{x} = \bar{\mathbf{V}}_p \mathbf{z}_p$  for some  $\mathbf{z}_p \in \mathbb{R}^p$ . Using the result (cf. (8.20) and (5.36))

$$\bar{\mathbf{U}}_{p+1}^T [\mathbf{K}_\Lambda \quad \mathbf{y}^\delta] \begin{bmatrix} \bar{\mathbf{V}}_p & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} = [\bar{\mathbf{U}}_{p+1}^T \mathbf{K}_\Lambda \bar{\mathbf{V}}_p \quad \bar{\mathbf{U}}_{p+1}^T \mathbf{y}^\delta] = [\mathbf{B}_p \quad \beta_1 \mathbf{e}_1^{(p+1)}],$$

the constrained minimization problem can be rewritten as

$$\begin{aligned} \min_{[\tilde{\mathbf{B}}_p \tilde{\mathbf{e}}_p] \in \mathbb{R}^{(p+1) \times (p+1)}} & \left\| \begin{bmatrix} \mathbf{B}_p & \beta_1 \mathbf{e}_1^{(p+1)} \end{bmatrix} - [\tilde{\mathbf{B}}_p \quad \tilde{\mathbf{e}}_p] \right\|_{\mathbf{F}}^2 \\ \text{subject to } & \tilde{\mathbf{B}}_p \mathbf{z}_p = \tilde{\mathbf{e}}_p, \end{aligned} \quad (8.21)$$

where we have put  $\tilde{\mathbf{B}}_p = \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{K}}_p \bar{\mathbf{V}}_p$  and  $\tilde{\mathbf{e}}_p = \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{y}}_p$ . Thus, in each Lanczos step, we use the TLS algorithm for the small-scale problem (8.21) to compute a truncated TLS solution  $\mathbf{x}_{\Lambda p}^\delta$ . More precisely, assuming the singular value decomposition

$$\begin{bmatrix} \mathbf{B}_p & \beta_1 \mathbf{e}_1^{(p+1)} \end{bmatrix} = \bar{\bar{\mathbf{U}}} \bar{\bar{\Sigma}} \bar{\bar{\mathbf{V}}},$$

with

$$\bar{\bar{\mathbf{V}}} = \begin{bmatrix} \bar{\bar{\mathbf{V}}}_{11} & \bar{\bar{\mathbf{v}}}_{12} \\ \bar{\bar{\mathbf{v}}}_{21}^T & \bar{\bar{v}}_{22} \end{bmatrix}, \quad \bar{\bar{\mathbf{V}}}_{11} \in \mathbb{R}^{p \times p}, \quad \bar{\bar{\mathbf{v}}}_{12}, \bar{\bar{\mathbf{v}}}_{21} \in \mathbb{R}^p,$$

the TLS solution to (8.21) is (cf. (8.10))

$$\mathbf{z}_{\Lambda p}^\delta = -\frac{1}{\bar{\bar{v}}_{22}} \bar{\bar{\mathbf{v}}}_{12},$$

and the truncated TLS solution takes the form

$$\mathbf{x}_{\Lambda p}^\delta = \bar{\mathbf{V}}_p \mathbf{z}_{\Lambda p}^\delta = -\frac{1}{\bar{\bar{v}}_{22}} \bar{\mathbf{V}}_p \bar{\bar{\mathbf{v}}}_{12}.$$

In the Lanczos truncated TLS algorithm, the solution norm and the residual norm also possess monotonic behavior, i.e.,  $\|\mathbf{x}_{\Lambda p}^\delta\|$  is an increasing function of  $p$ , while  $\|\mathbf{r}_{\Lambda p}^\delta\|_{\mathbf{F}}$  is a decreasing function of  $p$  (Fierro et al., 1997).

Regularization parameter choice methods for truncated TLS are discrete methods. If explicit knowledge about the errors in  $\mathbf{K}_\Lambda$  and  $\mathbf{y}^\delta$  is available, the discrepancy principle can be used to compute the truncation index. When the errors in  $\mathbf{K}_\Lambda$  and  $\mathbf{y}^\delta$  are not available, error-free parameter choice methods can be employed. In this context, we mention that Sima and Van Huffel (2006) formulated the generalized cross-validation in the framework of the Lanczos truncated TLS, while the L-curve method has been applied by Fierro et al. (1997).

The truncated solution  $\mathbf{x}_{\Lambda p}^\delta$  is a filtered solution whose main contributions come from the first  $p$  singular vectors of  $\mathbf{K}_\Lambda$  (Appendix I). Because these vectors are not always the best basis vectors for a regularized solution, we may implicitly include regularization in general form with  $\mathbf{L} \neq \mathbf{I}_n$ . This is done by transforming the problem involving  $\mathbf{K}_\Lambda$  and  $\mathbf{L}$  into a standard-form problem with the matrix  $\bar{\mathbf{K}}_\Lambda = \mathbf{K}_\Lambda \mathbf{L}^{-1}$ . Then, we apply the truncated TLS method to the standard-form problem to obtain a regularized solution  $\bar{\mathbf{x}}_{\Lambda p}^\delta$ , and finally, we transform  $\bar{\mathbf{x}}_{\Lambda p}^\delta$  back to the general-form setting by computing  $\mathbf{x}_{\Lambda p}^\delta = \mathbf{L}^{-1} \bar{\mathbf{x}}_{\Lambda p}^\delta$ . The conventional and the Lanczos versions of the truncated TLS method are outlined in Algorithms 13 and 14. It should be remarked that Algorithm 13 computes simultaneously the truncated SVD solution and the truncated TLS solution for a fixed value of the truncation index  $p$ .

### 8.3 Regularized total least squares for linear problems

Tikhonov regularization has been recast in the framework of the regularized TLS by Golub et al. (1999). To stress the differences and the similarities between the conventional Tikhonov regularization and the regularized TLS, we first note that Tikhonov regulariza-

---

**Algorithm 13.** Algorithm for computing the truncated SVD solution  $\mathbf{x}_p^\delta$  and the truncated TLS solution  $\mathbf{x}_{\Lambda p}^\delta$  for a fixed value of the truncation index  $p$ .

---

```

 $\bar{\mathbf{K}}_\Lambda \leftarrow \mathbf{K}_\Lambda \mathbf{L}^{-1};$ 
{truncated SVD solution}
compute the SVD  $\bar{\mathbf{K}}_\Lambda = \mathbf{U} \Sigma \mathbf{V}^T$ ;
 $\bar{\mathbf{x}}_p^\delta \leftarrow \sum_{i=1}^p (1/\sigma_i) (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i$ ;
 $\mathbf{x}_p^\delta \leftarrow \mathbf{L}^{-1} \bar{\mathbf{x}}_p^\delta$ ;
{truncated TLS solution}
compute the SVD  $\begin{bmatrix} \bar{\mathbf{K}}_\Lambda & \mathbf{y}^\delta \end{bmatrix} = \bar{\mathbf{U}} \bar{\Sigma} \bar{\mathbf{V}}^T$ ;
partition  $\bar{\mathbf{V}} = \begin{bmatrix} \bar{\mathbf{V}}_{11} & \bar{\mathbf{V}}_{12} \\ \bar{\mathbf{v}}_{21}^T & \bar{\mathbf{v}}_{22}^T \end{bmatrix}$  with  $\bar{\mathbf{V}}_{11} \in \mathbb{R}^{n \times p}$ ;
 $\bar{\mathbf{x}}_{\Lambda p}^\delta \leftarrow - \left( 1 / \|\bar{\mathbf{v}}_{22}\|^2 \right) \bar{\mathbf{V}}_{12} \bar{\mathbf{v}}_{22}$ ;
 $\mathbf{x}_{\Lambda p}^\delta \leftarrow \mathbf{L}^{-1} \bar{\mathbf{x}}_{\Lambda p}^\delta$ ;

```

---

---

**Algorithm 14.** Lanczos truncated TLS algorithm with  $p_{\max} > 1$  iterations.
 

---

$\beta_1 \leftarrow \|\mathbf{y}^\delta\|$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta_1) \mathbf{y}^\delta$ ;  
 $\mathbf{q} \leftarrow \mathbf{L}^{-T} \mathbf{K}^T \bar{\mathbf{u}}$ ;  $\alpha_1 \leftarrow \|\mathbf{q}\|$ ;  $\bar{\mathbf{v}}_1 \leftarrow (1/\alpha_1) \mathbf{q}$ ;  
**for**  $p = 1, p_{\max}$  **do**  
      $\mathbf{p} \leftarrow \mathbf{K} \mathbf{L}^{-1} \bar{\mathbf{v}}_p - \alpha_p \bar{\mathbf{u}}$ ;  $\beta_{p+1} \leftarrow \|\mathbf{p}\|$ ;  $\bar{\mathbf{u}} \leftarrow (1/\beta_{p+1}) \mathbf{p}$ ;  
     **if**  $p > 1$  **then**  
         set  $\mathbf{A} = \begin{bmatrix} \mathbf{B}_p & \beta_1 \mathbf{e}_1^{(p+1)} \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 & \beta_1 \\ \beta_2 & \alpha_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \alpha_p & 0 \\ 0 & 0 & \dots & \beta_{p+1} & 0 \end{bmatrix}$ ;  
         compute the SVD  $\mathbf{A} = \bar{\bar{\mathbf{U}}} \bar{\bar{\Sigma}} \bar{\bar{\mathbf{V}}}$ ;  
         partition  $\bar{\bar{\mathbf{V}}} = \begin{bmatrix} \bar{\bar{\mathbf{V}}}_{11} & \bar{\bar{\mathbf{V}}}_{12} \\ \bar{\bar{\mathbf{v}}}_{21}^T & \bar{\bar{v}}_{22} \end{bmatrix}$  with  $\bar{\bar{\mathbf{V}}}_{11} \in \mathbb{R}^{p \times p}$ ;  
          $\bar{\mathbf{x}}_{\Lambda p}^\delta \leftarrow -(1/\bar{\bar{v}}_{22}) \sum_{j=1}^p [\bar{\bar{\mathbf{v}}}_{12}]_j \bar{\bar{\mathbf{v}}}_j$ ;  
          $\mathbf{x}_{\Lambda p}^\delta \leftarrow \mathbf{L}^{-1} \bar{\mathbf{x}}_{\Lambda p}^\delta$ ;  
     **end if**  
     **if**  $p < p_{\max}$  **then**  
          $\mathbf{q} \leftarrow \mathbf{L}^{-T} \mathbf{K}^T \bar{\mathbf{u}} - \beta_{p+1} \bar{\mathbf{v}}_p$ ;  $\alpha_{p+1} \leftarrow \|\mathbf{q}\|$ ;  $\bar{\mathbf{v}}_{p+1} \leftarrow (1/\alpha_{p+1}) \mathbf{q}$ ;  
     **end if**  
**end for**

---

tion has an important equivalent formulation as

$$\begin{aligned} \min_{\mathbf{x}} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 \\ \text{subject to } \|\mathbf{L}\mathbf{x}\| \leq \varepsilon, \end{aligned} \quad (8.22)$$

where  $\varepsilon$  is a positive constant. The constrained least squares problem (8.22) can be solved by using the Lagrange multiplier formalism. Considering the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \alpha) = \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\|^2 + \alpha (\|\mathbf{L}\mathbf{x}\|^2 - \varepsilon^2),$$

it can be shown that if  $\varepsilon \leq \|\mathbf{L}\mathbf{x}^\delta\|$ , where  $\mathbf{x}^\delta$  is the least squares solution of the equation  $\mathbf{K}\mathbf{x} = \mathbf{y}^\delta$ , then the solution  $\mathbf{x}_\varepsilon^\delta$  to (8.22) is identical to the Tikhonov solution  $\mathbf{x}_\alpha^\delta$ , with  $\alpha$  solving the equation

$$\|\mathbf{L}\mathbf{x}_\alpha^\delta\|^2 = \varepsilon^2. \quad (8.23)$$

To carry this idea over to the TLS setting, we add the bound  $\|\mathbf{L}\mathbf{x}\| \leq \varepsilon$  to the ordinary problem (8.5), in which case, the new problem statement becomes

$$\begin{aligned} \min_{[\tilde{\mathbf{K}} \tilde{\mathbf{y}}] \in \mathbb{R}^{m \times (n+1)}} \left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{K}} & \tilde{\mathbf{y}} \end{bmatrix} \right\|_F^2 \\ \text{subject to } \tilde{\mathbf{K}}\mathbf{x} = \tilde{\mathbf{y}} \text{ and } \|\mathbf{L}\mathbf{x}\| \leq \varepsilon. \end{aligned} \quad (8.24)$$



The corresponding Lagrangian function is

$$\mathcal{L}(\tilde{\mathbf{K}}, \mathbf{x}, \alpha) = \left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{K}} & \tilde{\mathbf{K}}\mathbf{x} \end{bmatrix} \right\|_F^2 + \alpha \left( \|\mathbf{L}\mathbf{x}\|^2 - \varepsilon^2 \right),$$

and the Lagrange multiplier  $\alpha$  is non-zero if the inequality constraint is active. In fact, the solution  $\mathbf{x}_{\Lambda\varepsilon}^\delta$  to (8.24) is different from the TLS solution  $\mathbf{x}_\Lambda^\delta$ , whenever  $\varepsilon$  is less than  $\|\mathbf{L}\mathbf{x}_\Lambda^\delta\|$ .

To characterize  $\mathbf{x}_{\Lambda\varepsilon}^\delta$ , we set the partial derivatives of the Lagrangian function to zero. Differentiation with respect to the entries in  $\tilde{\mathbf{K}}$  yields

$$\tilde{\mathbf{K}} - \mathbf{K}_\Lambda - \mathbf{r}\mathbf{x}^T = \mathbf{0}, \quad (8.25)$$

with  $\mathbf{r} = \mathbf{y}^\delta - \tilde{\mathbf{K}}\mathbf{x}$ , while differentiation with respect to the entries in  $\mathbf{x}$  gives

$$-\tilde{\mathbf{K}}^T \mathbf{r} + \alpha \mathbf{L}^T \mathbf{L} \mathbf{x} = \mathbf{0}. \quad (8.26)$$

Setting the partial derivative with respect to  $\alpha$  to zero also yields

$$\|\mathbf{L}\mathbf{x}\|^2 = \varepsilon^2. \quad (8.27)$$

Making use of the expression of  $\mathbf{r}$ , we rearrange (8.26) as

$$\left( \tilde{\mathbf{K}}^T \tilde{\mathbf{K}} + \alpha \mathbf{L}^T \mathbf{L} \right) \mathbf{x} = \tilde{\mathbf{K}}^T \mathbf{y}^\delta. \quad (8.28)$$

Now, by (8.25) and (8.26), we have  $\mathbf{K}_\Lambda = \tilde{\mathbf{K}} - \mathbf{r}\mathbf{x}^T$  and  $\tilde{\mathbf{K}}^T \mathbf{r} = \alpha \mathbf{L}^T \mathbf{L} \mathbf{x}$ , respectively, and so, we obtain

$$\mathbf{K}_\Lambda^T \mathbf{K}_\Lambda = \tilde{\mathbf{K}}^T \tilde{\mathbf{K}} - \alpha \mathbf{x} \mathbf{x}^T \mathbf{L}^T \mathbf{L} + \|\mathbf{r}\|^2 \mathbf{x} \mathbf{x}^T - \alpha \mathbf{L}^T \mathbf{L} \mathbf{x} \mathbf{x}^T \quad (8.29)$$

and

$$\mathbf{K}_\Lambda^T \mathbf{y}^\delta = \tilde{\mathbf{K}}^T \mathbf{y}^\delta - (\mathbf{r}^T \mathbf{y}^\delta) \mathbf{x}. \quad (8.30)$$

Inserting (8.29) and (8.30) into (8.28), and using the identities (cf. (8.27))

$$\mathbf{x} \mathbf{x}^T \mathbf{L}^T \mathbf{L} \mathbf{x} = \varepsilon^2 \mathbf{x}, \quad \|\mathbf{r}\|^2 \mathbf{x} \mathbf{x}^T \mathbf{x} = \|\mathbf{r}\|^2 \|\mathbf{x}\|^2 \mathbf{x},$$

and

$$\mathbf{L}^T \mathbf{L} \mathbf{x} \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|^2 \mathbf{L}^T \mathbf{L} \mathbf{x},$$

we arrive at

$$(\mathbf{K}_\Lambda^T \mathbf{K}_\Lambda + \alpha_I \mathbf{I}_n + \alpha_L \mathbf{L}^T \mathbf{L}) \mathbf{x} = \mathbf{K}_\Lambda^T \mathbf{y}^\delta, \quad (8.31)$$

with

$$\alpha_I = \alpha \varepsilon^2 - \|\mathbf{r}\|^2 \|\mathbf{x}\|^2 - \mathbf{r}^T \mathbf{y}^\delta \quad (8.32)$$

and

$$\alpha_L = \alpha \left( 1 + \|\mathbf{x}\|^2 \right). \quad (8.33)$$

The next step of our derivation is the elimination of the Lagrange multiplier  $\alpha$  in the expressions of  $\alpha_I$  and  $\alpha_L$ . First, we use the relation (cf. (8.25))

$$\mathbf{r} = \mathbf{y}^\delta - \tilde{\mathbf{K}}\mathbf{x} = \mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x} - \|\mathbf{x}\|^2 \mathbf{r},$$

to obtain

$$(1 + \|\mathbf{x}\|^2) \mathbf{r} = \mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}, \quad (8.34)$$

and further,

$$(1 + \|\mathbf{x}\|^2) \|\mathbf{r}\|^2 = \frac{\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2}{1 + \|\mathbf{x}\|^2}. \quad (8.35)$$

On the other hand, scalar multiplying (8.26) by  $\mathbf{x}$  gives

$$\alpha = \frac{\mathbf{x}^T \tilde{\mathbf{K}}^T \mathbf{r}}{\|\mathbf{L}\mathbf{x}\|^2} = \frac{1}{\varepsilon^2} (\mathbf{r}^T \mathbf{y}^\delta - \|\mathbf{r}\|^2). \quad (8.36)$$

Considering the parameter  $\alpha_I$ , we insert (8.35) and (8.36) into (8.32), and find that

$$\alpha_I = -\frac{\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2}{1 + \|\mathbf{x}\|^2}. \quad (8.37)$$

Turning now to the parameter  $\alpha_L$ , we use (8.33) and (8.36) to get

$$\alpha_L = \alpha (1 + \|\mathbf{x}\|^2) = \frac{1}{\varepsilon^2} (\mathbf{r}^T \mathbf{y}^\delta - \|\mathbf{r}\|^2) (1 + \|\mathbf{x}\|^2). \quad (8.38)$$

Finally, a relationship connecting  $\alpha_L$  and  $\alpha_I$  can be derived as follows: by (8.35) and (8.37), we have  $\alpha_I = -\|\mathbf{r}\|^2 (1 + \|\mathbf{x}\|^2)$ , whence using (8.34), (8.38) becomes

$$\alpha_L = \frac{1}{\varepsilon^2} [\mathbf{y}^{\delta T} (\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}) + \alpha_I]. \quad (8.39)$$

To evaluate the approximation error  $\|[\mathbf{K}_\Lambda \quad \mathbf{y}^\delta] - [\tilde{\mathbf{K}} \quad \tilde{\mathbf{y}}]\|_F$ , we use the relation (cf. (8.25))

$$[\mathbf{K}_\Lambda \quad \mathbf{y}^\delta] - [\tilde{\mathbf{K}} \quad \tilde{\mathbf{K}}\mathbf{x}] = [\mathbf{K}_\Lambda - \tilde{\mathbf{K}} \quad \mathbf{r}] = \begin{bmatrix} -\mathbf{r}\mathbf{x}^T & \mathbf{r} \end{bmatrix} = -\mathbf{r} \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix}^T,$$

together with (8.35) and (8.37), to obtain

$$\|[\mathbf{K}_\Lambda \quad \mathbf{y}^\delta] - [\tilde{\mathbf{K}} \quad \tilde{\mathbf{y}}]\|_F^2 = (1 + \|\mathbf{x}\|^2) \|\mathbf{r}\|^2 = -\alpha_I. \quad (8.40)$$

Collecting all results we conclude that  $\mathbf{x}_{\Lambda\varepsilon}^\delta$  is the solution of equation (8.31) with  $\alpha_I$  and  $\alpha_L$  given by (8.37) and (8.39), respectively. The main features of the regularized TLS are presented below (Golub et al., 1999).

- (1) If the matrix  $\alpha_I \mathbf{I}_n + \alpha_L \mathbf{L}^T \mathbf{L}$  is positive definite, then the regularized TLS solution corresponds to the Tikhonov solution with the penalty term  $\alpha_I \|\mathbf{x}\|^2 + \alpha_L \|\mathbf{L}\mathbf{x}\|^2$ . If the matrix  $\alpha_I \mathbf{I}_n + \alpha_L \mathbf{L}^T \mathbf{L}$  is indefinite or negative definite, there is no equivalent interpretation.

- (2) For a given  $\varepsilon$ , there are several pairs of parameters  $\alpha_I$  and  $\alpha_L$  and thus several solutions  $\mathbf{x}_{\Lambda\varepsilon}^\delta$  that satisfy (8.31), (8.37) and (8.39). However, from (8.40), we see that only the solution with the smallest value of  $|\alpha_I|$  solves the constrained minimization problem (8.24).
- (3) If  $\varepsilon < \|\mathbf{L}\mathbf{x}_\Lambda^\delta\|$ , where  $\mathbf{x}_\Lambda^\delta$  is the TLS solution (8.10), the inequality constraint is binding, the Lagrange multiplier  $\alpha$  is positive and by (8.33), it follows that  $\alpha_L > 0$ . From (8.37) it is apparent that  $\alpha_I$  is always negative and thus adds some deregularization to the solution. The residual (8.40) is a monotonically decreasing function of  $\varepsilon$ , and so,  $\alpha_I$  is a monotonically increasing function of  $\varepsilon$ . If  $\varepsilon = \|\mathbf{L}\mathbf{x}_\Lambda^\delta\|$ , the Lagrange multiplier  $\alpha$  is zero and the regularized TLS solution  $\mathbf{x}_{\Lambda\varepsilon}^\delta$  coincides with the TLS solution  $\mathbf{x}_\Lambda^\delta$ ; for larger  $\varepsilon$ , the constraint is never again binding and so, the solution remains unchanged.

To compute the regularized TLS solution  $\mathbf{x}_{\Lambda\varepsilon}^\delta$  we have to solve a nonlinear problem, and several techniques have been proposed in the literature. In Golub et al. (1999),  $\alpha_L$  is considered as free parameter, a corresponding value is computed for  $\alpha_I$ , and the system of equations (8.31) is solved in an efficient way. The idea is to transform (8.31) into the augmented system of equations

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{0} & \mathbf{K}_\Lambda \\ \mathbf{0} & \mathbf{I}_n & \sqrt{\alpha_L}\mathbf{L} \\ \mathbf{K}_\Lambda^T & \sqrt{\alpha_L}\mathbf{L}^T & -\alpha_I\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{s} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{y}^\delta \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

to reduce  $\mathbf{K}_\Lambda$  to an  $n \times n$  bidiagonal form by means of orthogonal transformations, to apply Elden's algorithm to annihilate the matrix term containing the factor  $\sqrt{\alpha_L}$ , and finally, to use a symmetric perfect shuffle reordering to obtain a symmetric, tridiagonal, indefinite matrix of size  $2n \times 2n$  containing the parameter  $\alpha_I$  on the main diagonal.

In Guo and Renault (2002), a shifted inverse power method is used to obtain the eigen-pair

$$\left( \lambda, \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} \right)$$

for the problem

$$\mathbf{B}(\mathbf{x}) \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix}, \quad (8.41)$$

where

$$\mathbf{B}(\mathbf{x}) = \begin{bmatrix} \mathbf{K}_\Lambda^T \mathbf{K}_\Lambda + \alpha_L(\mathbf{x}) \mathbf{L}^T \mathbf{L} & \mathbf{K}_\Lambda^T \mathbf{y}^\delta \\ \mathbf{y}^{\delta T} \mathbf{K}_\Lambda & -\alpha_L(\mathbf{x}) \varepsilon^2 + \mathbf{y}^{\delta T} \mathbf{y}^\delta \end{bmatrix}$$

is an  $(n+1) \times (n+1)$  matrix,  $\lambda = -\alpha_I$ , and  $\alpha_L$  is given by (cf. (8.37) and (8.39))

$$\alpha_L(\mathbf{x}) = \frac{1}{\varepsilon^2} \left[ \mathbf{y}^{\delta T} (\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}) - \frac{\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2}{1 + \|\mathbf{x}\|^2} \right]. \quad (8.42)$$

In Renault and Guo (2005), the solution of the eigenproblem (8.41) is considered together with the solution of a nonlinear equation which guarantees the bound  $\|\mathbf{L}\mathbf{x}\| = \varepsilon$ . To describe the main features of this algorithm, we consider the decomposition

$$\mathbf{B}(\alpha_L) = \mathbf{M} + \alpha_L \mathbf{N},$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{K}_\Lambda^T \mathbf{K}_\Lambda & \mathbf{K}_\Lambda^T \mathbf{y}^\delta \\ \mathbf{y}^{\delta T} \mathbf{K}_\Lambda & \mathbf{y}^{\delta T} \mathbf{y}^\delta \end{bmatrix}, \quad \mathbf{N} = \begin{bmatrix} \mathbf{L}^T \mathbf{L} & \mathbf{0} \\ \mathbf{0} & -\varepsilon^2 \end{bmatrix},$$

and denote by

$$\left( \lambda_{\alpha_L}, \begin{bmatrix} \mathbf{x}_{\alpha_L} \\ -1 \end{bmatrix} \right)$$

the eigenpair corresponding to the smallest eigenvalue of  $\mathbf{B}(\alpha_L)$ . For a fixed  $\varepsilon$ , we introduce the function

$$g(\mathbf{x}) = \frac{\|\mathbf{L}\mathbf{x}\|^2 - \varepsilon^2}{1 + \|\mathbf{x}\|^2},$$

and compute  $\hat{\alpha}_L$  such that  $\mathbf{x}_{\hat{\alpha}_L}$  solves the equation

$$g(\mathbf{x}_{\alpha_L}) = 0; \quad (8.43)$$

$\mathbf{x}_{\hat{\alpha}_L}$  is then the regularized TLS solution of (8.24). To justify this algorithm, we assume that  $\mathbf{x}_{\hat{\alpha}_L}$  satisfies the eigensystem equation

$$\mathbf{B}(\hat{\alpha}_L) \begin{bmatrix} \mathbf{x}_{\hat{\alpha}_L} \\ -1 \end{bmatrix} = \lambda_{\hat{\alpha}_L} \begin{bmatrix} \mathbf{x}_{\hat{\alpha}_L} \\ -1 \end{bmatrix}, \quad (8.44)$$

and is also a solution of equation (8.43). The first block equation of the eigenvalue problem (8.44) gives (8.31) with  $\hat{\alpha}_I = -\lambda_{\hat{\alpha}_L}$ , while the second block equation yields (8.39). Multiplying the eigensystem equation by  $[\mathbf{x}_{\hat{\alpha}_L}^T, -1]$ , we find that

$$\lambda_{\hat{\alpha}_L} = \frac{1}{1 + \|\mathbf{x}_{\hat{\alpha}_L}\|^2} \left[ \|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}_{\hat{\alpha}_L}\|^2 + \hat{\alpha}_L \left( \|\mathbf{L}\mathbf{x}_{\hat{\alpha}_L}\|^2 - \varepsilon^2 \right) \right]. \quad (8.45)$$

Since  $g(\mathbf{x}_{\hat{\alpha}_L}) = 0$ , it follows that  $\|\mathbf{L}\mathbf{x}_{\hat{\alpha}_L}\|^2 = \varepsilon^2$ , and (8.45) becomes

$$\lambda_{\hat{\alpha}_L} = \frac{\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}_{\hat{\alpha}_L}\|^2}{1 + \|\mathbf{x}_{\hat{\alpha}_L}\|^2}; \quad (8.46)$$

thus  $\hat{\alpha}_I = -\lambda_{\hat{\alpha}_L}$  satisfies indeed (8.37). In summary,  $\mathbf{x}_{\hat{\alpha}_L}$  solves equation (8.31) with  $\hat{\alpha}_I$  as in (8.37) and  $\hat{\alpha}_L$  as in (8.39). Since  $\lambda_{\hat{\alpha}_L}$  is the smallest eigenvalue of  $\mathbf{B}$ , the present approach explicitly computes a solution with the smallest value of  $|\alpha_I|$ .

For a practical implementation of the method of Renault and Guo we note the following results:

- (1) if  $\lambda_{n+1} > 0$  is the smallest eigenvalue of the matrix  $\mathbf{B}$  and  $\mathbf{v}_{n+1}$  is the corresponding eigenvector, then  $\lambda_{\alpha_L} = \lambda_{n+1}$  and

$$\begin{bmatrix} \mathbf{x}_{\alpha_L} \\ -1 \end{bmatrix} = -\frac{1}{[\mathbf{v}_{n+1}]_{n+1}} \mathbf{v}_{n+1};$$

- (2)  $g(\mathbf{x}_{\alpha_L})$  is a monotonically decreasing function of  $\alpha_L$ , and there exists only one solution  $\hat{\alpha}_L$  of the equation  $g(\mathbf{x}_{\alpha_L}) = 0$ .

Algorithm 15 computes the Tikhonov solution and the regularized TLS solution for a fixed value of the parameter  $\alpha$  corresponding to the method of Tikhonov regularization. Both solutions are related to each other through the constraint norms. The input parameter  $\alpha$  is used to determine the bound  $\varepsilon$  and to estimate a bisection interval for  $\alpha_L$ . The algorithm also computes the ‘equivalent’ regularization matrix defined as

$$\alpha \mathbf{L}_{\text{eq}}^T \mathbf{L}_{\text{eq}} = \hat{\alpha}_I \mathbf{I}_n + \hat{\alpha}_L \mathbf{L}^T \mathbf{L}. \quad (8.47)$$

This factorization is performed by using the Cholesky method with added multiple of identity, which takes into account that for large negative values of  $\hat{\alpha}_I$ , the matrix  $\hat{\alpha}_I \mathbf{I}_n + \hat{\alpha}_L \mathbf{L}^T \mathbf{L}$  may not be positive definite. Note that strategies based on modifying a Cholesky factorization or a symmetric indefinite factorization of a non-positive definite Hessian are standard approaches in the framework of Newton’s method (Nocedal and Wright, 2006).

---

**Algorithm 15.** Algorithm for computing the regularized TLS solution by solving the eigenvalue problem (8.41). The regularization parameter  $\alpha$  corresponds to the method of Tikhonov regularization. The algorithm computes the solution  $\mathbf{x}_{\hat{\alpha}_L}$ , the regularization parameters  $\hat{\alpha}_L$  and  $\hat{\alpha}_I$ , and the equivalent regularization matrix  $\mathbf{L}_{\text{eq}}$ .

---

```

compute the Tikhonov solution  $\mathbf{x}_\alpha^\delta$  for  $\alpha$ , i.e.,  $\mathbf{x}_\alpha^\delta = (\mathbf{K}_\Lambda^T \mathbf{K}_\Lambda + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}_\Lambda^T \mathbf{y}^\delta$ ;
 $\varepsilon \leftarrow \|\mathbf{L} \mathbf{x}_\alpha^\delta\|$ ;
compute the matrices  $\mathbf{M}$  and  $\mathbf{N}$ ;
estimate a bisection interval  $[\alpha_{L \min}, \alpha_{L \max}]$  for  $\alpha_L$  around  $\alpha$ ;
solve  $g(\alpha_L) = 0$  in  $[\alpha_{L \min}, \alpha_{L \max}]$  using FuncEval ( $\alpha_L, \varepsilon, \mathbf{M}, \mathbf{N}; g, \mathbf{x}_{\alpha_L}, \alpha_I$ );
store the solution  $\hat{\alpha}_L$  and the corresponding  $\mathbf{x}_{\hat{\alpha}_L}$  and  $\hat{\alpha}_I$ ;
{regularization matrix using Cholesky factorization with added multiple of identity}
choose the tolerance  $\varepsilon_\alpha$ , e.g.,  $\varepsilon_\alpha = 0.001$ ;
 $\Delta \alpha \leftarrow \varepsilon_\alpha |\hat{\alpha}_I|$ ; stop  $\leftarrow$  false;
while stop = false do
    attempt to apply the Cholesky factorization to obtain  $\mathbf{L}_{\text{eq}}^T \mathbf{L}_{\text{eq}} = \hat{\alpha}_I \mathbf{I}_n + \hat{\alpha}_L \mathbf{L}^T \mathbf{L}$ ;
    if factorization is successful then
        stop  $\leftarrow$  true;
    else
         $\hat{\alpha}_I \leftarrow \hat{\alpha}_I + \Delta \alpha$ ;
    end if
end while
 $\mathbf{L}_{\text{eq}} \leftarrow (1/\sqrt{\alpha}) \mathbf{L}_{\text{eq}}$ .

{for given  $\alpha_L$ , the routine computes  $g(\alpha_L)$ ,  $\mathbf{x}_{\alpha_L}$  and  $\alpha_I$ }
routine FuncEval ( $\alpha_L, \varepsilon, \mathbf{M}, \mathbf{N}; g, \mathbf{x}_{\alpha_L}, \alpha_I$ )
 $\mathbf{B} \leftarrow \mathbf{M} + \alpha_L \mathbf{N}$ ;
compute the smallest eigenvalue  $\lambda_{n+1}$  of  $\mathbf{B}$  and the eigenvector  $\mathbf{v}_{n+1}$ ;
compute  $\mathbf{x}_{\alpha_L}$  as  $\begin{bmatrix} \mathbf{x}_{\alpha_L} \\ -1 \end{bmatrix} = -(1/[\mathbf{v}_{n+1}]_{n+1}) \mathbf{v}_{n+1}$ ;
 $\alpha_I \leftarrow -\lambda_{n+1}$ ;
 $g \leftarrow (\|\mathbf{L} \mathbf{x}_{\alpha_L}\|^2 - \varepsilon^2) / (1 + \|\mathbf{x}_{\alpha_L}\|^2)$ .

```

---

In Sima et al. (2003), the objective function is the so-called orthogonal distance, and the constrained minimization problem takes the form (cf. (8.37) and (8.40))

$$\min_{\mathbf{x}} \frac{\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2}{1 + \|\mathbf{x}\|^2}$$

subject to  $\|\mathbf{L}\mathbf{x}\| \leq \varepsilon$ .

The first-order optimality conditions for the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \lambda) = \frac{\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2}{1 + \|\mathbf{x}\|^2} + \lambda \left( \|\mathbf{L}\mathbf{x}\|^2 - \varepsilon^2 \right),$$

yield

$$\mathbf{D}(\mathbf{x}) \mathbf{x} + \lambda \mathbf{L}^T \mathbf{L} \mathbf{x} = \mathbf{d}(\mathbf{x}), \quad \|\mathbf{L}\mathbf{x}\|^2 = \varepsilon^2, \quad (8.48)$$

with

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{K}_\Lambda^T \mathbf{K}_\Lambda}{1 + \|\mathbf{x}\|^2} - \frac{\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2}{\left(1 + \|\mathbf{x}\|^2\right)^2} \mathbf{I}_n, \quad \mathbf{d}(\mathbf{x}) = \frac{\mathbf{K}_\Lambda^T \mathbf{y}^\delta}{1 + \|\mathbf{x}\|^2}.$$

The problem (8.48) is first transformed into the standard form and then solved iteratively by using a fixed point iteration method. Assuming that  $\mathbf{L}$  is square and nonsingular, the transformation to the standard form gives

$$(\mathbf{W} + \lambda \mathbf{I}_n) \bar{\mathbf{x}} = \mathbf{h}, \quad \|\bar{\mathbf{x}}\|^2 = \varepsilon^2, \quad (8.49)$$

with  $\bar{\mathbf{x}} = \mathbf{L}\mathbf{x}$ ,  $\mathbf{W} = \mathbf{L}^{-T} \mathbf{D} \mathbf{L}^{-1}$  and  $\mathbf{h} = \mathbf{L}^{-T} \mathbf{d}$ . Note that since  $\mathbf{D}$  is a symmetric matrix,  $\mathbf{W}$  is also a symmetric matrix. Let us now consider the problem

$$(\mathbf{W} + \lambda \mathbf{I}_n)^2 \mathbf{u} = \mathbf{h}, \quad \mathbf{h}^T \mathbf{u} = \varepsilon^2 \quad (8.50)$$

for  $\mathbf{u} \in \mathbb{R}^n$ . Setting

$$\bar{\mathbf{x}} = (\mathbf{W} + \lambda \mathbf{I}_n) \mathbf{u},$$

and taking into account that, due to the symmetry of  $\mathbf{W} + \lambda \mathbf{I}_n$ , there holds

$$\varepsilon^2 = \mathbf{h}^T \mathbf{u} = \mathbf{u}^T (\mathbf{W} + \lambda \mathbf{I}_n)^2 \mathbf{u} = \|\bar{\mathbf{x}}\|^2,$$

we see that the problems (8.49) and (8.50) are equivalent. Further, using the identity

$$\mathbf{h} = \frac{1}{\varepsilon^2} (\mathbf{h}^T \mathbf{u}) \mathbf{h} = \frac{1}{\varepsilon^2} \mathbf{h} \mathbf{h}^T \mathbf{u},$$

we deduce that (8.50) can be transformed into the quadratic eigenvalue problem

$$\left( \lambda^2 \mathbf{I}_n + 2\lambda \mathbf{W} + \mathbf{W}^2 - \frac{1}{\varepsilon^2} \mathbf{h} \mathbf{h}^T \right) \mathbf{u} = \mathbf{0}. \quad (8.51)$$

This quadratic eigenvalue problem is solved in order to find the largest eigenvalue  $\lambda$  and the corresponding eigenvector  $\mathbf{u}$  scaled so that  $\mathbf{h}^T \mathbf{u} = \varepsilon^2$ . As all matrices in (8.51) are

real and symmetric, the quadratic eigenvalues are real and come in complex conjugate pairs. Moreover, the special form of the quadratic eigenvalue problem (8.51) implies that the rightmost (largest real) eigenvalue is real and positive. The solution of the original problem is then recovered by first computing  $\bar{\mathbf{x}} = (\mathbf{W} + \lambda \mathbf{I}_n) \mathbf{u}$  and then  $\mathbf{x} = \mathbf{L}^{-1} \bar{\mathbf{x}}$ .

---

**Algorithm 16.** Algorithm for computing the regularized TLS solution by solving the quadratic eigenvalue problem (8.51). The regularization parameter  $\alpha$  corresponds to the method of Tikhonov regularization. The algorithm computes the solution  $\mathbf{x}$ , the regularization parameters  $\alpha_L$  and  $\alpha_I$ , and the equivalent regularization matrix  $\mathbf{L}_{\text{eq}}$ .

---

choose the tolerances  $\varepsilon_1$  and  $\varepsilon_x$  for the convergence test;

compute the Tikhonov solution  $\mathbf{x}_\alpha^\delta$  for  $\alpha$ , i.e.,  $\mathbf{x}_\alpha^\delta = (\mathbf{K}_\Lambda^T \mathbf{K}_\Lambda + \alpha \mathbf{L}^T \mathbf{L})^{-1} \mathbf{K}_\Lambda^T \mathbf{y}^\delta$ ;

$\varepsilon \leftarrow \|\mathbf{L} \mathbf{x}_\alpha^\delta\|$ ;

$\bar{\mathbf{K}}_\Lambda \leftarrow \mathbf{K}_\Lambda \mathbf{L}^{-1}$ ;

$\text{stop} \leftarrow \text{false}$ ;  $k \leftarrow 0$ ;  $\mathbf{x} \leftarrow \mathbf{x}_\alpha^\delta$ ; {starting vector}

**while**  $\text{stop} = \text{false}$  **do**

$r \leftarrow \|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2 / (1 + \|\mathbf{x}\|^2)$ ;  $c \leftarrow 1 / (1 + \|\mathbf{x}\|^2)$ ;

$\mathbf{W} \leftarrow c \bar{\mathbf{K}}_\Lambda^T \bar{\mathbf{K}}_\Lambda - r c \mathbf{L}^{-T} \mathbf{L}^{-1}$ ;  $\mathbf{h} \leftarrow c \bar{\mathbf{K}}_\Lambda^T \mathbf{y}^\delta$ ;

set  $\mathbf{A} = \begin{bmatrix} -2\mathbf{W} & -\mathbf{W}^2 + \varepsilon^{-2} \mathbf{h} \mathbf{h}^T \\ \mathbf{I}_n & \mathbf{0} \end{bmatrix}$ ;

compute the largest eigenvalue  $\lambda$  and the corresponding eigenvector  $\begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix}$  of  $\mathbf{A}$ ;

$\mathbf{u} \leftarrow (\varepsilon^2 / \mathbf{h}^T \mathbf{u}) \mathbf{u}$ ; {scale  $\mathbf{u}$ }

$\mathbf{W} \leftarrow \mathbf{W} + \lambda \mathbf{I}_n$ ;

$\mathbf{x} \leftarrow \mathbf{L}^{-1} \mathbf{W} \mathbf{u}$ ;

{convergence test}

**if**  $k > 0$  **and**  $|\lambda - \lambda_{\text{prv}}| \leq \varepsilon_1 \lambda$  **and**  $\|\mathbf{x} - \mathbf{x}_{\text{prv}}\| \leq \varepsilon_x \|\mathbf{x}\|$  **then**

$\text{stop} \leftarrow \text{true}$ ;

**else**

$\lambda_{\text{prv}} \leftarrow \lambda$ ;  $\mathbf{x}_{\text{prv}} \leftarrow \mathbf{x}$ ;

$k \leftarrow k + 1$ ;

**end if**

**end while**

$\alpha_L \leftarrow \lambda (1 + \|\mathbf{x}\|^2)$ ;  $\alpha_I \leftarrow -\|\mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}\|^2 / (1 + \|\mathbf{x}\|^2)$ ;

compute  $\mathbf{L}_{\text{eq}}$  as in Algorithm 15

---

The quadratic eigenvalue problem (8.51) is equivalent to the linear eigenvalue problem

$$\begin{bmatrix} -2\mathbf{W} & -\mathbf{W}^2 + \frac{1}{\varepsilon^2} \mathbf{h} \mathbf{h}^T \\ \mathbf{I}_n & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix},$$

and this can be solved by using for example, the routine DGEEV from the LAPACK library (Anderson et al., 1995), or the routine DNAUPD from the ARPACK library (Maschhoff and Sorensen, 1996). The DNAUPD routine is more efficient because it calculates only the largest eigenvalue and the corresponding eigenvector by using Arnoldi's method (Arnoldi,

1951). The Algorithm 16 generates a sequence  $\{(\lambda_k, \mathbf{x}_k)\}$  by solving the quadratic eigenvalue problem (8.51) at each iteration step  $k$ . From the analysis of Sima et al. (2003) we infer the following results:

- (1)  $\mathbf{x}_k$  should correspond to the largest eigenvalue  $\lambda_k > 0$  since only then the algorithm converges;
- (2) the orthogonal distance decreases at each iteration step;
- (3) any limit point of the sequence  $\{(\lambda_k, \mathbf{x}_k)\}$  solves equation (8.48).

The last result suggests that instead of requiring the convergence of the sequence  $\{(\lambda_k, \mathbf{x}_k)\}$  we may check if equation (8.48) is satisfied within a prescribed tolerance at each iteration step.

#### 8.4 Regularized total least squares for nonlinear problems

As stated in Chapter 6, the solution of a nonlinear ill-posed problem by means of Tikhonov regularization is equivalent to the solution of a sequence of ill-posed linearizations of the forward model about the current iterate. Essentially, at the iteration step  $k$ , we solve the linearized equation

$$\mathbf{K}_{\alpha k} \Delta \mathbf{x} = \mathbf{y}_k^\delta, \quad (8.52)$$

with  $\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_a$ ,  $\mathbf{K}_{\alpha k} = \mathbf{K}(\mathbf{x}_{\alpha k}^\delta)$ , and

$$\mathbf{y}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\alpha k}(\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a),$$

via Tikhonov regularization with the penalty term  $\|\mathbf{L}\Delta \mathbf{x}\|^2$  and the regularization parameter  $\alpha$ . If  $\Delta \mathbf{x}_{\alpha k}^\delta$  is the minimizer of the Tikhonov function

$$\mathcal{F}_{1\alpha k}(\Delta \mathbf{x}) = \|\mathbf{y}_k^\delta - \mathbf{K}_{\alpha k} \Delta \mathbf{x}\|^2 + \alpha \|\mathbf{L}\Delta \mathbf{x}\|^2, \quad (8.53)$$

the new iterate is given by  $\mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_a + \Delta \mathbf{x}_{\alpha k}^\delta$ , and the constraint norm can be readily computed as

$$\varepsilon = \|\mathbf{L}\Delta \mathbf{x}_{\alpha k}^\delta\|. \quad (8.54)$$

In the framework of the regularized TLS, we assume that  $\mathbf{K}_{\alpha k}$  is contaminated by errors, and instead of minimizing (8.53) we solve the problem

$$\min_{[\tilde{\mathbf{K}} \tilde{\mathbf{y}}] \in \mathbb{R}^{m \times (n+1)}} \left\| \begin{bmatrix} \mathbf{K}_{\alpha k} & \mathbf{y}_k^\delta \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{K}} & \tilde{\mathbf{y}} \end{bmatrix} \right\|_{\mathbf{F}}^2 \quad (8.55)$$

subject to  $\tilde{\mathbf{K}}\Delta \mathbf{x} = \tilde{\mathbf{y}}$  and  $\|\mathbf{L}\Delta \mathbf{x}\| \leq \varepsilon$ ,

with  $\varepsilon$  being given by (8.54). The free parameter of the method is the Tikhonov regularization parameter  $\alpha$ , and the Algorithms 15 and 16 can be used to compute both the Tikhonov solution and the regularized TLS solution. Although the numerical implementation of the regularized TLS is very similar to that of Tikhonov regularization, the use of a step-length procedure is problematic. In principle it can be applied for the objective function

$$\mathcal{F}_\alpha(\mathbf{x}) = \frac{1}{2} \|\mathbf{f}_\alpha(\mathbf{x})\|^2, \quad \mathbf{f}_\alpha(\mathbf{x}) = \begin{bmatrix} \mathbf{F}(\mathbf{x}) - \mathbf{y}^\delta \\ \sqrt{\alpha} \mathbf{L}_{\text{eq}}(\mathbf{x} - \mathbf{x}_a) \end{bmatrix}, \quad (8.56)$$



but solving (8.55) is not equivalent to minimizing (8.56) at the iteration step  $k$  because  $\mathbf{L}_{\text{eq}}$  may not be the exact Cholesky factor of  $\hat{\alpha}_T \mathbf{I}_n + \hat{\alpha}_L \mathbf{L}^T \mathbf{L}$  (cf. (8.47)).

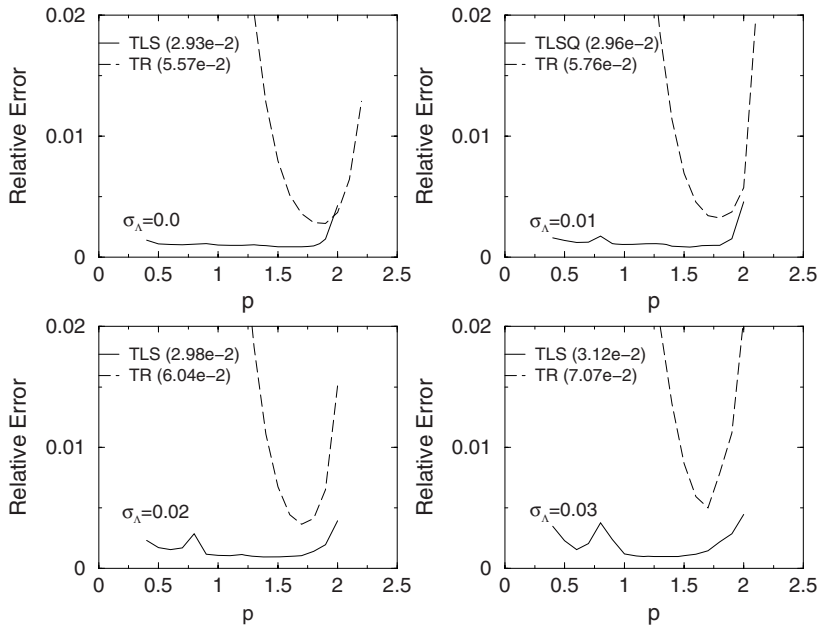
In our numerical analysis, we consider the  $\text{O}_3$  retrieval test problem and compute the Jacobian matrix  $\mathbf{K}_{ss}$  by assuming only the single scattering contribution (cf. (8.1)). Furthermore, at each iteration step, we perturb this matrix as

$$[\mathbf{K}_{k\alpha}]_{ij} = [\mathbf{K}_{ssk\alpha}]_{ij} + \sigma_\Lambda \varepsilon_{ij} [\mathbf{K}_{ssk\alpha}]_{ij},$$

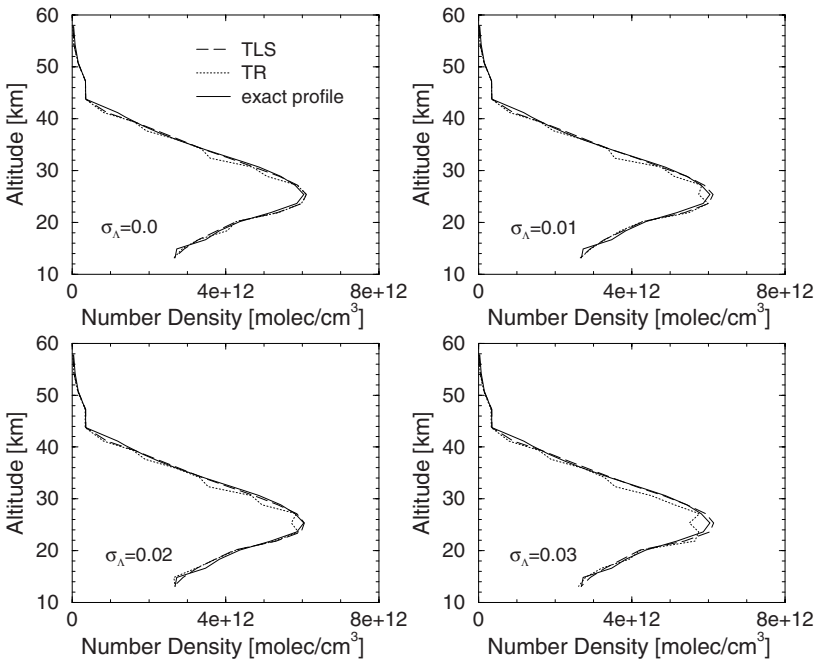
where the elements  $\varepsilon_{ij}$  are from a normal distribution with zero mean and unit variance. Figure 8.1 shows the relative errors in the Tikhonov and the regularized TLS solutions for four values of the standard deviation  $\sigma_\Lambda$ , namely 0, 0.01, 0.02 and 0.03. In all situations, the minimum solution error for the regularized TLS is clearly smaller than that for Tikhonov regularization. Even in the case  $\sigma_\Lambda = 0$  there is a solution improvement due to the approximate Jacobian calculation. The plots also show that the minima of the TLS errors are flat and this situation is beneficial for the inversion process.

In Figure 8.2 we plot the Tikhonov and the regularized TLS solutions, corresponding to the minimizers of the error curves in Figure 8.1. In fact, the improvement of the TLS error as compared to the Tikhonov error is due to the additional term  $\alpha_T \mathbf{I}_n$  in Eq. (8.31).

From the point of view of their accuracy, the regularized TLS algorithms solving the eigenvalue problem (8.41) and the quadratic eigenvalue problem (8.51) are completely



**Fig. 8.1.** Relative errors in the Tikhonov and the regularized TLS solutions as a function of the exponent  $p$ , where  $\alpha = \sigma^p$  and  $\sigma$  is the noise standard deviation. The results correspond to the  $\text{O}_3$  retrieval test problem and are computed with the regularized TLS algorithm solving the quadratic eigenvalue problem (8.51). The numbers in parentheses indicate the minimum values of the relative solution error.



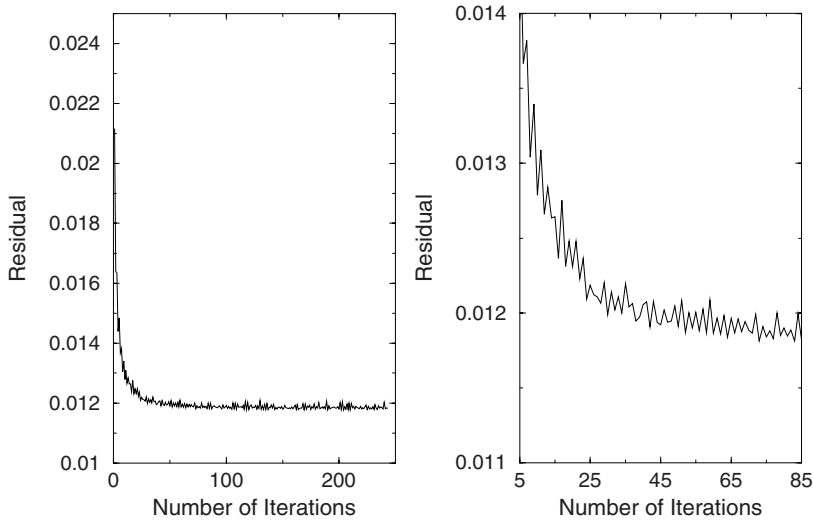
**Fig. 8.2.** Tikhonov (TR) and regularized TLS solutions corresponding to the minimizers of the error curves in Figure 8.1.

equivalent. However, the computation time of the algorithm based on a quadratic eigenvalue problem is on average 6 times smaller (Table 8.1). The main drawback of the regularized TLS is the extraordinarily large number of iteration steps (and so, computation time) as compared to Tikhonov regularization. The decrease of the solution error by a factor of 4–5 is accompanied by an increase of the computation time by a factor of 7–8.

The large number of iteration steps is also a consequence of the fact that we do not use a step-length procedure to guarantee a monotonic decrease of the residual norm (Figure 8.3). A step-length algorithm stops the iterative process too early (because the search direction is not a descent direction for the Tikhonov function), and as a result, the solu-

**Table 8.1.** Computation time in min:ss format. The numbers in parentheses indicate the number of iteration steps for Tikhonov regularization (TR) and the regularized TLS algorithms solving the eigenvalue problem (8.41) (TLS-EP) and the quadratic eigenvalue problem (8.51) (TLS-QEP).

Method	Standard deviation $\sigma_\Lambda$			
	0	0.01	0.02	0.03
TR	0:14 (4)	0:15 (6)	0:18 (8)	0:24 (16)
TLS-QEP	1:24 (108)	1:37 (124)	2:23 (202)	2:58 (243)
TLS-EP	8:01 (108)	9:57 (124)	13:13 (202)	19:17 (243)



**Fig. 8.3.** History of the residual norm in the case  $\sigma_{\Lambda} = 0.03$ . In the left panel the curves are plotted for all iteration steps, while in the right panel, the  $y$ -axis is zoomed out.

tion errors are not sufficiently small. For example, in the case  $\sigma_{\Lambda} = 0.03$ , the regularized TLS with a step-length algorithm terminates after 19 iteration steps with a solution error of  $1.56 \cdot 10^{-2}$ , and if the step-length algorithm is turned off, it terminates after 243 iteration steps with a solution error of  $9.77 \cdot 10^{-4}$ .

The design of an efficient regularized TLS algorithm for nonlinear problems is far from being complete. The selection of an optimal value of the regularization parameter by an a posteriori method will dramatically increase the computational effort, while the use of a variable regularization parameter computed for example, by using the L-curve method, is also problematic. In our numerical simulations, the L-curve either does not have a distinctive L-shape, or it predicts values of the regularization parameter that are too small.

The regularized TLS has been applied to atmospheric trace gas profile retrievals by Koner and Drummond (2008). In this work, the regularized TLS algorithm solving the quadratic eigenvalue problem (8.51) is used for the automatic determination of the regularization strength.

# 9

## Two direct regularization methods

In this chapter we present two direct regularization methods, namely the Backus–Gilbert method and the maximum entropy regularization. Although these approaches have been designed for linear problems they can be applied to nonlinear problems as well.

### 9.1 Backus–Gilbert method

In the framework of Tikhonov regularization, the generalized inverse is not explicitly computed and is merely an analysis tool. The goal of the so-called mollifier methods is the computation of an approximate generalized inverse, which can then be used to obtain an approximate solution. Mollifier methods have been introduced by Louis and Maass (1990) in a continuous setting, and applied for discrete problems by Rieder and Schuster (2000).

To describe mollifier methods, we consider a semi-discrete Fredholm integral equation of the first kind

$$y_i = \int_0^{z_{\max}} k_i(z) x(z) \, dz, \quad i = 1, \dots, m, \quad (9.1)$$

and introduce a smoothing operator  $A_\mu : X \rightarrow X$  by the relation

$$(A_\mu x)(z_0) = \int_0^{z_{\max}} a_\mu(z_0, z) x(z) \, dz. \quad (9.2)$$

The parameter-dependent function  $a_\mu$  in (9.2) is called mollifier and it is chosen such that  $A_\mu x \rightarrow x$  as  $\mu \rightarrow 0$  for all  $x \in X$ . Next, we assume that  $a_\mu$  can be expressed as

$$a_\mu(z_0, z) = \sum_{i=1}^m k_i(z) k_{\mu i}^\dagger(z_0), \quad (9.3)$$

where  $k_{\mu i}^\dagger$  are referred to as the contribution functions. In the framework of mollifier methods we choose a mollifier  $\bar{a}_\mu$  and compute the contribution functions  $k_{\mu i}^\dagger$  as the solution of

the constrained minimization problem

$$\begin{aligned} \min_{k_{\mu i}^\dagger} \int_0^{z_{\max}} [\bar{a}_\mu(z_0, z) - a_\mu(z_0, z)]^2 dz \\ \text{subject to } \int_0^{z_{\max}} a_\mu(z_0, z) dz = 1, \end{aligned} \quad (9.4)$$

with  $a_\mu$  being given by (9.3). The normalization condition in (9.4) just means that for  $x \equiv 1$ ,  $A_\mu x \equiv 1$  (cf. (9.2)). Once the contribution functions are known, we use the representation (cf. (9.1), (9.2) and (9.3))

$$(A_\mu x)(z_0) = \sum_{i=1}^m \left[ \int_0^{z_{\max}} k_i(z) x(z) dz \right] k_{\mu i}^\dagger(z_0) = \sum_{i=1}^m k_{\mu i}^\dagger(z_0) y_i, \quad (9.5)$$

to compute the mollified solution of the linear equation (9.1) with noisy data  $y_i^\delta$  as

$$x_\mu^\delta(z_0) = \sum_{i=1}^m k_{\mu i}^\dagger(z_0) y_i^\delta. \quad (9.6)$$

Thus, in the framework of mollifier methods, instead of solving (9.1), we choose the mollifier and solve (9.3) with respect to the contribution functions as in (9.4). Equation (9.3) is also ill-posed as soon as equation (9.1) is, but the calculation of the mollified solution, according to (9.4) and (9.6), is expected to be a stable process because there are no errors in the data.

The transpose vector  $\mathbf{k}_\mu^{\dagger T} = [k_{\mu 1}^\dagger, \dots, k_{\mu m}^\dagger]$  reproduces the row vector of the generalized inverse  $\mathbf{K}_\mu^\dagger$  corresponding to the altitude height  $z_0$ , and  $a_\mu(z_0, z)$  can be interpreted as a continuous version of the averaging kernel matrix  $\mathbf{K}_\mu^\dagger \mathbf{K}$ .

The function  $a_\mu(z_0, z)$  determines the resolution of the mollifier method at  $z_0$ , and for  $x_\mu^\delta(z_0)$  to be meaningful,  $a_\mu(z_0, z)$  should peak around  $z_0$ . To make  $a_\mu(z_0, z)$  as localized as possible about the point  $z_0$ , we have to choose the mollifiers as smooth regular functions approximating a Dirac distribution. In fact, the choice of mollifiers depends on the peculiarities of the solution, and frequently used choices are (Louis and Maass, 1990)

$$\begin{aligned} \bar{a}_\mu(z_0, z) &= \begin{cases} c, & |z - z_0| \leq \mu, \\ 0, & \text{otherwise,} \end{cases} \\ \bar{a}_\mu(z_0, z) &= c \operatorname{sinc}(\mu(z - z_0)), \\ \bar{a}_\mu(z_0, z) &= c \exp\left(-\frac{(z - z_0)^2}{2\mu^2}\right), \end{aligned}$$

where the parameter  $\mu$  controls the width of the  $\delta$ -like functions and  $c$  is a normalization constant.

Another variant of mollifier methods is the Backus–Gilbert method, also known as the method of optimally localized averages (Backus and Gilbert, 1967, 1968, 1970). In this approach, the averaging kernel function  $a_\mu(z_0, z)$  is controlled by specifying a positive

$\delta^{-1}$ -like function  $d_\mu(z_0, z)$  and then solving the constrained minimization problem

$$\begin{aligned} \min_{k_{\mu i}^\dagger} \int_0^{z_{\max}} d_\mu(z_0, z) a_\mu(z_0, z)^2 dz \\ \text{subject to } \int_0^{z_{\max}} a_\mu(z_0, z) dz = 1. \end{aligned} \quad (9.7)$$

The function  $d_\mu$  can be chosen as

$$d_\mu(z_0, z) = \left| \frac{z - z_0}{l} \right|^\mu \quad (9.8)$$

or as

$$d_\mu(z_0, z) = 1 - \exp\left(-\frac{1}{2} \left| \frac{z - z_0}{l} \right|^\mu\right), \quad (9.9)$$

where  $l$  is the correlation length and as before,  $\mu$  is a parameter which controls the width of the  $\delta^{-1}$ -like function.

Although the Backus–Gilbert method has been designed for linear problems, its extension to nonlinear problems is straightforward. Let us consider the update formula

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{p}_k^\delta, \quad k = 0, 1, \dots,$$

where  $\mathbf{p}_k^\delta$  is the Newton step and  $\mathbf{x}_0^\delta = \mathbf{x}_a$ . Further, let  $\mathbf{p}_k^\dagger = \mathbf{x}^\dagger - \mathbf{x}_k^\delta$  be the exact step, where  $\mathbf{x}^\dagger$  is a solution of the nonlinear equation with exact data  $\mathbf{F}(\mathbf{x}) = \mathbf{y}$ . It is quite obvious that  $\mathbf{p}_k^\dagger$  solves the equation (see Appendix H)

$$\mathbf{K}_k \mathbf{p} = \mathbf{r}_k, \quad (9.10)$$

with

$$\mathbf{r}_k = \mathbf{y} - \mathbf{F}(\mathbf{x}_k^\delta) - \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta) \quad (9.11)$$

and  $\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta)$  being the linearization error. As  $\mathbf{r}_k$  is unknown, and only

$$\mathbf{r}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta), \quad (9.12)$$

is available, we consider the equation

$$\mathbf{K}_k \mathbf{p} = \mathbf{r}_k^\delta, \quad (9.13)$$

and compute  $\mathbf{p}_k^\delta$  as

$$\mathbf{p}_k^\delta = \mathbf{K}_k^\dagger \mathbf{r}_k^\delta. \quad (9.14)$$

In (9.14), the generalized inverse  $\mathbf{K}_k^\dagger$  is unknown and its row vectors will be determined one by one. Before doing this, we observe that the  $i$ th entry of  $\mathbf{p}_k^\delta$  is given by

$$[\mathbf{p}_k^\delta]_i = \mathbf{k}_i^{\dagger T} \mathbf{r}_k^\delta, \quad (9.15)$$

where  $\mathbf{k}_i^{\dagger T}$  is the  $i$ th row vector of  $\mathbf{K}_k^{\dagger}$ , partitioned as

$$\mathbf{K}_k^{\dagger} = \begin{bmatrix} \mathbf{k}_1^{\dagger T} \\ \vdots \\ \mathbf{k}_n^{\dagger T} \end{bmatrix}. \quad (9.16)$$

Now, defining as usual the averaging kernel matrix  $\mathbf{A}_k$  by

$$\mathbf{A}_k = \mathbf{K}_k^{\dagger} \mathbf{K}_k, \quad (9.17)$$

and assuming the partitions

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{bmatrix}, \quad \mathbf{K}_k = [\mathbf{k}_1, \dots, \mathbf{k}_n],$$

we obtain

$$[\mathbf{a}_i]_j = \mathbf{k}_i^{\dagger T} \mathbf{k}_j, \quad i, j = 1, \dots, n. \quad (9.18)$$

To compute the row vector  $\mathbf{k}_i^{\dagger T}$  we proceed to formulate the constrained minimization problem (9.7) in terms of the averaging kernel  $\mathbf{a}_i^T$ . For this purpose, we discretize the altitude interval  $[0, z_{\max}]$  in  $n$  layers and put  $[\mathbf{a}_i]_j = a_{\mu}(z_i, z_j)$ , where  $z_i$  is the centerpoint of the layer  $i$ . The objective function in (9.7) can then be expressed as (cf. (9.18))

$$\begin{aligned} s(z_i) &= \int_0^{z_{\max}} d_{\mu}(z_i, z) a_{\mu}(z_i, z)^2 dz \\ &= \sum_{j=1}^n d_{\mu}(z_i, z_j) a_{\mu}(z_i, z_j)^2 \Delta z_j \\ &= \sum_{j=1}^n d_{\mu}(z_i, z_j) [\mathbf{a}_i]_j^2 \Delta z_j \\ &= \mathbf{k}_i^{\dagger T} \mathbf{Q}_{ki} \mathbf{k}_i^{\dagger}, \end{aligned}$$

where  $\Delta z_i$  is the geometrical thickness of the layer  $i$ , and

$$\mathbf{Q}_{ki} = \mathbf{K}_k \left[ \text{diag}(d_{\mu}(z_i, z_j) \Delta z_j)_{n \times n} \right] \mathbf{K}_k^T.$$

For the choice (9.8) with  $\mu = 2$ ,  $s(z_i)$  represents the spread of the averaging kernel around  $z_i$ , and by minimizing the spread we intend to guarantee that the resolution of the retrieval is as high as possible. The normalization condition in (9.7) takes the form (cf. (9.18))

$$1 = \int_0^{z_{\max}} a_{\mu}(z_i, z) dz = \sum_{j=1}^n a_{\mu}(z_i, z_j) \Delta z_j = \sum_{j=1}^n [\mathbf{a}_i]_j \Delta z_j = \mathbf{k}^T \mathbf{k}_i^{\dagger},$$

with

$$\mathbf{k} = \sum_{j=1}^n \mathbf{k}_j \Delta z_j,$$

and the constrained minimization problem to be solved reads as

$$\begin{aligned} \min_{\mathbf{k}^\dagger} \mathbf{k}^{\dagger T} \mathbf{Q}_{ki} \mathbf{k}^\dagger \\ \text{subject to } \mathbf{k}^T \mathbf{k}^\dagger = 1. \end{aligned} \quad (9.19)$$

Via the Lagrange multiplier formalism, the row vector  $\mathbf{k}_i^{\dagger T}$  is determined by minimizing the Lagrangian function

$$\mathcal{L}(\mathbf{k}^\dagger, \lambda) = \frac{1}{2} \mathbf{k}^{\dagger T} \mathbf{Q}_{ki} \mathbf{k}^\dagger + \lambda (\mathbf{k}^T \mathbf{k}^\dagger - 1), \quad (9.20)$$

and the result is

$$\mathbf{k}_i^\dagger = \frac{1}{\mathbf{q}_i^T \mathbf{k}} \mathbf{q}_i. \quad (9.21)$$

with

$$\mathbf{q}_i = \mathbf{Q}_{ki}^{-1} \mathbf{k}.$$

In practice it is necessary to add regularization when the problem (9.19) is solved numerically, due to the ill-conditioning of the matrix  $\mathbf{Q}_{ki}$ . Neglecting the linearization error  $\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta)$ , the Newton step  $\mathbf{p}_k^\delta$  can be expressed as (cf. (9.10)–(9.12) and (9.14))

$$\mathbf{p}_k^\delta = \mathbf{K}_k^\dagger \mathbf{r}_k^\delta = \mathbf{K}_k^\dagger (\mathbf{r}_k + \boldsymbol{\delta}) = \mathbf{A}_k \mathbf{p}_k^\dagger + \mathbf{K}_k^\dagger \boldsymbol{\delta},$$

and it is apparent that the spread accounts only for the smoothed component  $\mathbf{A}_k \mathbf{p}_k^\dagger$  of  $\mathbf{p}_k^\delta$ . The  $i$ th entry of the noise error vector  $\mathbf{e}_{nk}^\delta = -\mathbf{K}_k^\dagger \boldsymbol{\delta}$  is

$$[\mathbf{e}_{nk}^\delta]_i = -\mathbf{k}_i^{\dagger T} \boldsymbol{\delta},$$

and for white noise with covariance  $\mathbf{C}_\delta = \sigma^2 \mathbf{I}_m$ , the expected value of the noise error is given by

$$n(z_i) = \mathcal{E} \left\{ [\mathbf{e}_{nk}^\delta]_i^2 \right\} = \sigma^2 \left\| \mathbf{k}_i^\dagger \right\|^2. \quad (9.22)$$

In this regard, we construct an objective function reflecting a trade-off between spread and noise error, that is, we consider the constrained minimization problem

$$\begin{aligned} \min_{\mathbf{k}^\dagger} \left( \mathbf{k}^{\dagger T} \mathbf{Q}_{ki} \mathbf{k}^\dagger + \alpha \left\| \mathbf{k}^\dagger \right\|^2 \right) \\ \text{subject to } \mathbf{k}^T \mathbf{k}^\dagger = 1. \end{aligned} \quad (9.23)$$

The objective function in (9.23) is as in (9.19), but with  $\mathbf{Q}_{ki} + \alpha \mathbf{I}_m$  in place of  $\mathbf{Q}_{ki}$ ; the solution of (9.23) is then

$$\mathbf{k}_{\alpha i}^\dagger = \frac{1}{\mathbf{q}_{\alpha i}^T \mathbf{k}} \mathbf{q}_{\alpha i}, \quad (9.24)$$

with

$$\mathbf{q}_{\alpha i} = (\mathbf{Q}_{ki} + \alpha \mathbf{I}_m)^{-1} \mathbf{k}. \quad (9.25)$$



Once the row vectors of the generalized inverse have been computed, the Backus–Gilbert step is determined via (cf. (9.15))

$$[\mathbf{p}_{k\alpha}^\delta]_i = \mathbf{k}_{\alpha i}^\dagger \mathbf{r}_k^\delta = \frac{\mathbf{q}_{\alpha i}^T \mathbf{r}_k^\delta}{\mathbf{q}_{\alpha i}^T \mathbf{k}}, \quad i = 1, \dots, n. \quad (9.26)$$

Let us discuss some practical implementation issues by following the analysis of Hansen (1994). Defining the diagonal matrices

$$\mathbf{D}_i = \left[ \text{diag} \left( \sqrt{d_\mu(z_i, z_j)} \right)_{n \times n} \right], \quad \mathbf{Z} = \left[ \text{diag}(\Delta z_j)_{n \times n} \right],$$

and denoting by  $\mathbf{e}$  the  $n$ -dimensional vector of all ones, i.e.,  $\mathbf{e} = [1, \dots, 1]^T$ , we express  $\mathbf{Q}_{ki}$  as

$$\mathbf{Q}_{ki} = \mathbf{K}_k \mathbf{D}_i \mathbf{Z} \mathbf{D}_i \mathbf{K}_k^T.$$

Setting

$$\bar{\mathbf{K}}_{ki} = \mathbf{K}_k \mathbf{D}_i \mathbf{Z}^{\frac{1}{2}}, \quad \mathbf{e}_i = \mathbf{D}_i^{-1} \mathbf{Z}^{\frac{1}{2}} \mathbf{e},$$

and noting that

$$\mathbf{k} = \mathbf{K}_k \mathbf{Z} \mathbf{e} = \bar{\mathbf{K}}_{ki} \mathbf{e}_i,$$

we write  $\mathbf{q}_{\alpha i}$  as (cf. (9.25))

$$\mathbf{q}_{\alpha i} = (\bar{\mathbf{K}}_{ki} \bar{\mathbf{K}}_{ki}^T + \alpha \mathbf{I}_m)^{-1} \bar{\mathbf{K}}_{ki} \mathbf{e}_i. \quad (9.27)$$

Moreover, we have (cf. (9.24))

$$\mathbf{k}_{\alpha i}^\dagger = \frac{1}{\mathbf{q}_{\alpha i}^T \bar{\mathbf{K}}_{ki} \mathbf{e}_i} \mathbf{q}_{\alpha i}$$

and (cf. (9.26))

$$[\mathbf{p}_{k\alpha}^\delta]_i = \frac{\mathbf{q}_{\alpha i}^T \mathbf{r}_k^\delta}{\mathbf{q}_{\alpha i}^T \bar{\mathbf{K}}_{ki} \mathbf{e}_i}. \quad (9.28)$$

Note that the singularity of  $\mathbf{D}_i^{-1}$  at  $j = i$  can be removed in practice by approximating

$$d_\mu(z_i, z_i) \approx d_\mu(z_i, z_i + \Delta z),$$

with  $\Delta z$  sufficiently small, e.g.,  $\Delta z = 1$  m. An inspection of (9.27) reveals that  $\mathbf{q}_{\alpha i}$  minimizes the Tikhonov function

$$\mathcal{F}_\alpha(\mathbf{q}) = \|\mathbf{e}_i - \bar{\mathbf{K}}_{ki}^T \mathbf{q}\|^2 + \alpha \|\mathbf{q}\|^2.$$

Thus, if  $(\bar{\sigma}_j; \bar{\mathbf{v}}_j, \bar{\mathbf{u}}_j)$  is a singular system of  $\bar{\mathbf{K}}_{ki}$ , we obtain the representation

$$\mathbf{q}_{\alpha i} = \sum_{j=1}^n \frac{\bar{\sigma}_j}{\bar{\sigma}_j^2 + \alpha} (\bar{\mathbf{v}}_j^T \mathbf{e}_i) \bar{\mathbf{u}}_j,$$

and the useful expansions

$$\mathbf{q}_{\alpha i}^T \bar{\mathbf{K}}_{ki} \mathbf{e}_i = \sum_{j=1}^n \frac{\bar{\sigma}_j^2}{\bar{\sigma}_j^2 + \alpha} (\bar{\mathbf{v}}_j^T \mathbf{e}_i)^2, \quad (9.29)$$

and

$$\mathbf{q}_{\alpha i}^T \mathbf{r}_k^\delta = \sum_{j=1}^n \frac{\bar{\sigma}_j^2}{\bar{\sigma}_j^2 + \alpha} \frac{1}{\bar{\sigma}_j} (\bar{\mathbf{v}}_j^T \mathbf{e}_i) (\bar{\mathbf{u}}_j^T \mathbf{r}_k^\delta). \quad (9.30)$$

The Backus–Gilbert solution can be computed for any value of the regularization parameter  $\alpha$ , by inserting (9.29) and (9.30) into (9.28).

To reveal the regularizing effect of the Backus–Gilbert method we mention that the characteristic features of the singular vectors of  $\mathbf{K}_k$  carry over to the singular vectors of  $\bar{\mathbf{K}}_{ki}$ , and that the filter factors in (9.30) damp out the noisy components in the data as the Tikhonov filter factors do.

To compute the regularization parameter we may impose that the noise error (9.22) has a prescribed value, that is,

$$n_\alpha(z_i) = \varepsilon_n [\mathbf{x}_a]_i^2, \quad (9.31)$$

for some relative error level  $\varepsilon_n$ . Another selection criterion can be designed by taking into account that the spread is an increasing function of  $\alpha$  and that the noise error is a decreasing function of  $\alpha$ . Thus, we may follow the idea of the L-curve method, and compute the regularization parameter which balances the spread and noise error. For any value of  $\alpha$ , the computable expressions of the quantities of interest are

$$s_\alpha(z_i) = \frac{1}{(\mathbf{q}_{\alpha i}^T \bar{\mathbf{K}}_{ki} \mathbf{e}_i)^2} \sum_{j=1}^n \left( \frac{\bar{\sigma}_j^2}{\bar{\sigma}_j^2 + \alpha} \bar{\mathbf{v}}_j^T \mathbf{e}_i \right)^2,$$

$$n_\alpha(z_i) = \frac{\sigma^2}{(\mathbf{q}_{\alpha i}^T \bar{\mathbf{K}}_{ki} \mathbf{e}_i)^2} \sum_{j=1}^n \left( \frac{\bar{\sigma}_j}{\bar{\sigma}_j^2 + \alpha} \bar{\mathbf{v}}_j^T \mathbf{e}_i \right)^2,$$

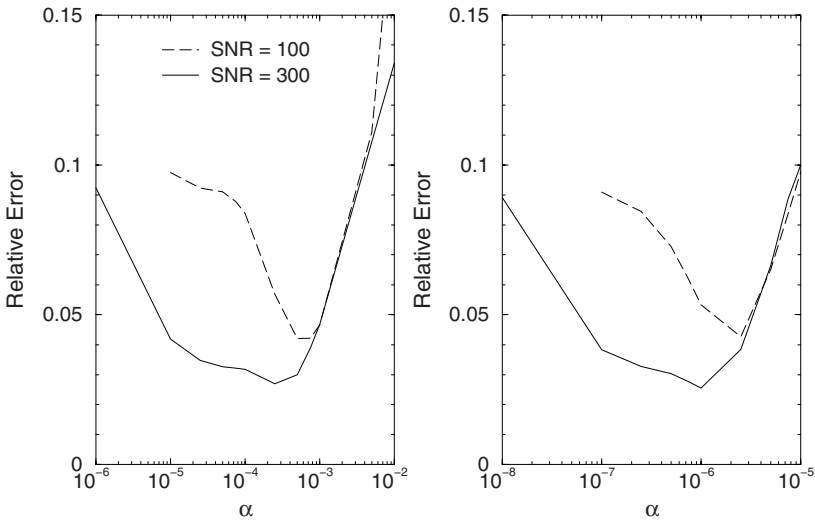
and the regularization parameter, corresponding to the point on the curve at which the tangent has the slope  $-1$ , is chosen as the minimizer of the function (Reginska, 1996),

$$\beta(\alpha) = x(\alpha) + y(\alpha), \quad (9.32)$$

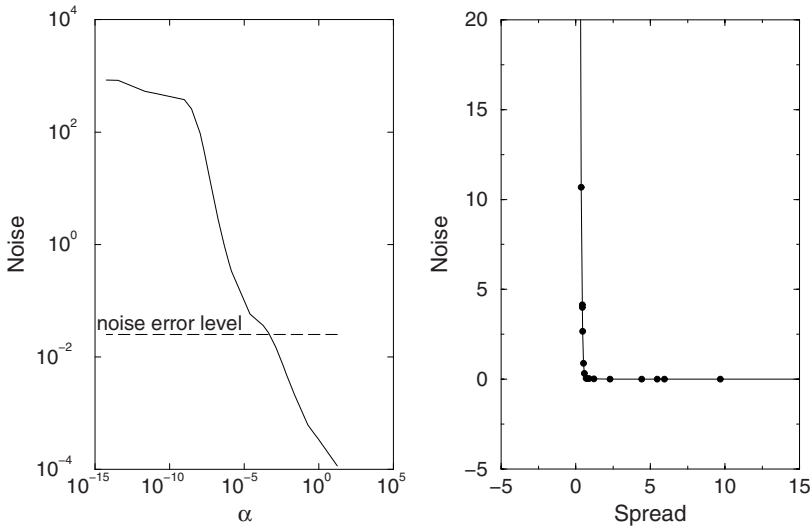
with  $x(\alpha) = s_\alpha$  and  $y(\alpha) = n_\alpha$ .

In Figure 9.1 we plot the solution errors for the  $\text{O}_3$  retrieval test problem. The  $\delta^{-1}$ -like functions (9.8) and (9.9) yield similar accuracies but for different domains of variation of the regularization parameter. The regularizing effect of the Backus–Gilbert method is also apparent in this figure: by increasing the signal-to-noise ratio, the minimum solution error as well as the optimal value of the regularization parameter (the minimizer) decrease.

In our numerical analysis we used a discrete version of the regularization parameter choice methods (9.31) and (9.32), that is, for the set  $\{\alpha_j\}$  with  $\alpha_j = \bar{\sigma}_j^2$ ,  $j = 1, \dots, n$ , we chose the regularization parameter  $\alpha_{j^*}$  as the smallest  $\alpha_j$  satisfying  $n_{\alpha_{j^*}}(z_i) \leq \varepsilon_n [\mathbf{x}_a]_i^2$ , or as the minimizer of  $\beta(\alpha_j)$ . The plots in Figure 9.2 illustrate that the noise error is



**Fig. 9.1.** Relative solution errors for the Backus–Gilbert method with the quadratic function (9.8) (left) and the exponential function (9.9) (right). The parameters of calculation are  $\mu = 2$  and  $l = 1.0$  km for the quadratic function, and  $\mu = 2$  and  $l = 10.0$  km for the exponential function.



**Fig. 9.2.** Noise error curve (left) and L-curve (right) for a layer situated at 30.6 km.

a decreasing function of the regularization parameter and that the L-curve has a distinct corner.

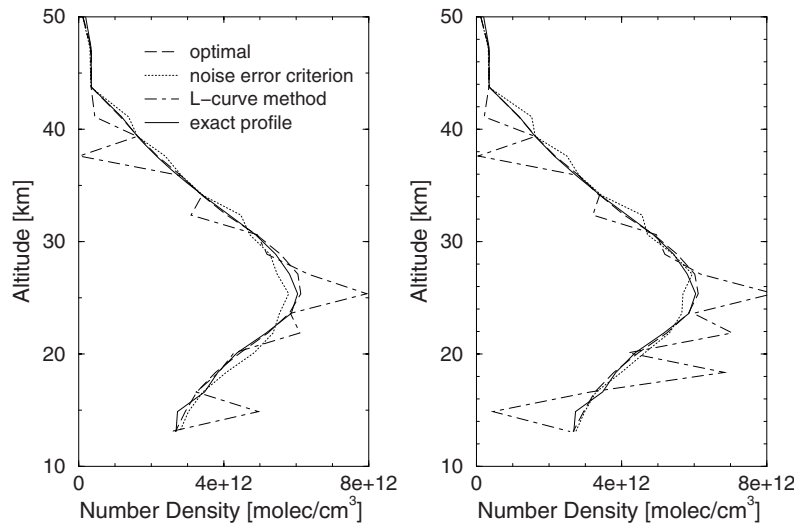
The solution errors given in Table 9.1 show that the noise error criterion yields sufficiently accurate results. By contrast, the L-curve method predicts a value of the regularization parameter which is considerably smaller than the optimal value. As a result,

**Table 9.1.** Relative solution errors for the Backus–Gilbert method with the noise error (NE) criterion and the L-curve (LC) method.

$\delta^{-1}$ -like function	SNR	Method	$\varepsilon$	$\varepsilon_{\text{opt}}$
quadratic	100	NE	6.65e-2	4.21e-2
		LC	2.30e-1	
	300	NE	5.74e-2	2.51e-2
		LC	1.42e-1	
exponential	100	NE	5.88e-2	4.26e-2
		LC	3.15e-1	
	300	NE	5.41e-2	2.55e-2
		LC	2.42e-1	

the retrieved profiles are undersmoothed (Figure 9.3). Note that the failure of the L-curve method is because we use a very rough discrete search procedure to minimize  $\beta$ .

In the framework of mollifier methods, the approximate generalized inverse is determined independently of the data, and therefore, mollifier methods can be viewed as being equivalent to Tikhonov regularization with an a priori parameter choice method. In practice, the methods are computationally very expensive because for each layer, we have to solve an optimization problem. However, for the operational usage of a near real-time software processor, this drawback is only apparent; when the approximate generalized inverse



**Fig. 9.3.** Retrieved profiles computed with the Backus–Gilbert method using the quadratic function (9.8) (left) and the exponential function (9.9) (right). The curves correspond to the optimal value of the regularization parameter (the minimizer in Figure 9.1), the noise error criterion and the L-curve method. The signal-to-noise ratio is SNR = 300 and the parameters of calculation are as in Figure 9.1.

is (a priori) computed and stored, the processing of data is much faster than, for example, Tikhonov regularization with an a posteriori parameter choice method, because it involves only matrix-vector multiplications.

## 9.2 Maximum entropy regularization

First proposed as a general inference procedure by Jaynes (1957) on the basis of Shannon's axiomatic characterization of the amount of information (Shannon, 1949; Shannon and Weaver, 1949), the maximum entropy principle emerged as a successful regularization technique due to the contributions of Frieden (1972), and Gull and Daniel (1978). Although the conventional maximum entropy regularization operates with the concept of absolute entropy (or Shannon entropy), we describe a formulation based on relative and cross entropies, which allows a better exploitation of the available a priori information (Engl et al., 2000).

To sketch the maximum entropy regularization we consider a discrete random variable  $X$  with a finite number of realizations  $x_1, \dots, x_n$ , and suppose that we make some a priori assumptions about the probability mass function of  $X$ ,

$$p_a(x) = \begin{cases} p_{ai}, & X = x_i, \\ 0, & \text{otherwise,} \end{cases} \quad \sum_{i=1}^n p_{ai} = 1.$$

By measurements we obtain additional information on  $X$ , which lets us change our a priori probability mass function into the a posteriori probability mass function,

$$p(x) = \begin{cases} p_i, & X = x_i, \\ 0, & \text{otherwise,} \end{cases} \quad \sum_{i=1}^n p_i = 1.$$

We recall that in statistical inversion theory, the a posteriori probability mass function represents the conditional probability density of  $X$  given the measurement data. The goal of our analysis is the computation of the a posteriori probability mass function by considering the new data.

In information theory, a natural distance measure from the probability mass function  $p$  to the probability mass function  $p_a$  is the Kullback–Leibler divergence defined by

$$D(p; p_a) = \sum_{i=1}^n p_i \log \left( \frac{p_i}{p_{ai}} \right).$$

Essentially, the Kullback–Leibler divergence signifies the amount of useful information about  $X$ , that can be obtained given the measurements. The negative of the Kullback–Leibler divergence represents the relative entropy

$$H_r(p; p_a) = - \sum_{i=1}^n p_i \log \left( \frac{p_i}{p_{ai}} \right).$$

Note that as opposed to the absolute entropy  $H(p) = - \sum_{i=1}^n p_i \log p_i$ , the relative entropy  $H_r$  is negative (cf. (9.35) below) and attains its global maximum  $H_{r\max} = 0$  at  $p = p_a$ .

To compute the a posteriori probability mass function, we minimize the Kullback–Leibler divergence  $D$  (or maximize the relative entropy  $H_r$ ) with the data and the normalization condition  $\sum_{i=1}^n p_i = 1$  as constraints. If  $\mathbf{x}$  is the state vector to be retrieved and  $\mathbf{x}_a$  is the a priori state, we define the normalized vectors

$$\bar{\mathbf{x}} = \frac{1}{\sum_{i=1}^n [\mathbf{x}]_i} \mathbf{x}, \quad \bar{\mathbf{x}}_a = \frac{1}{\sum_{i=1}^n [\mathbf{x}_a]_i} \mathbf{x}_a,$$

and under the assumptions  $[\mathbf{x}]_i > 0$  and  $[\mathbf{x}_a]_i > 0$  for  $i = 1, \dots, n$ , we interpret the components of these vectors as the probabilities  $p_i$  and  $p_{ai}$ , respectively. As data we consider the nonlinear model  $\mathbf{y}^\delta = \mathbf{F}(\mathbf{x}) + \boldsymbol{\delta}$ , and impose the feasibility constraint

$$\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2 \leq \Delta^2.$$

The constrained minimization problem then takes the form

$$\begin{aligned} \min_{\mathbf{x}} \Lambda_r(\mathbf{x}) &= \sum_{i=1}^n [\bar{\mathbf{x}}]_i \log \left( \frac{[\bar{\mathbf{x}}]_i}{[\bar{\mathbf{x}}_a]_i} \right) \\ \text{subject to } &\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2 \leq \Delta^2. \end{aligned} \quad (9.33)$$

By virtue of the Lagrange multiplier formalism, the problem (9.33) is equivalent to the minimization of the Tikhonov function

$$\mathcal{F}_\alpha(\mathbf{x}) = \frac{1}{2} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x})\|^2 + \alpha \Lambda_r(\mathbf{x}). \quad (9.34)$$

Using the inequality

$$\log z \geq 1 - \frac{1}{z}, \quad z > 0, \quad (9.35)$$

we find that

$$\Lambda_r(\mathbf{x}) \geq \sum_{i=1}^n ([\bar{\mathbf{x}}]_i - [\bar{\mathbf{x}}_a]_i) = 0.$$

Evidently, the global minimizer of  $\Lambda_r$  is attained for  $\bar{\mathbf{x}} = \bar{\mathbf{x}}_a$ , which reiterates the role of  $\mathbf{x}_a$  as a priori information.

If  $\mathbf{x}$  and  $\mathbf{x}_a$  are not normalized, the non-negative functions (Eggermont, 1993)

$$\Lambda_B(\mathbf{x}) = \sum_{i=1}^n \left[ \log \left( \frac{[\mathbf{x}]_i}{[\mathbf{x}_a]_i} \right) + \frac{[\mathbf{x}_a]_i}{[\mathbf{x}]_i} - 1 \right]$$

and

$$\Lambda_c(\mathbf{x}) = \sum_{i=1}^n \left[ \frac{[\mathbf{x}]_i}{[\mathbf{x}_a]_i} \log \left( \frac{[\mathbf{x}]_i}{[\mathbf{x}_a]_i} \right) - \frac{[\mathbf{x}]_i}{[\mathbf{x}_a]_i} + 1 \right],$$

representing the negative of the Burg's entropy and the cross entropy, respectively, can be used as penalty terms. A Taylor expansion of the cross entropy about the a priori yields

$$\Lambda_c(\mathbf{x}) = \frac{1}{2} (\mathbf{x} - \mathbf{x}_a)^T \left[ \text{diag} \left( \frac{1}{[\mathbf{x}_a]_i^2} \right)_{n \times n} \right] (\mathbf{x} - \mathbf{x}_a) + O(\|\mathbf{x} - \mathbf{x}_a\|^3),$$

and we see that in the neighborhood of the a priori, the cross entropy regularization matrix behaves like a diagonal matrix.

Ramos et al. (1999), following the work of Landl and Anderson (1996), developed two entropic regularization techniques by using penalty functions which are similar to the discrete difference operators. The first-order penalty function (corresponding to the entropy of the vector of first-order differences of  $\mathbf{x}$ ) is defined by

$$\Lambda_1(\mathbf{x}) = \sum_{i=1}^{n-1} \frac{(n-1)d_{1i}}{\sum_{i=1}^{n-1} d_{1i}} \log \left( \frac{(n-1)d_{1i}}{\sum_{i=1}^{n-1} d_{1i}} \right),$$

where the  $d_{1i}$  can be chosen as

$$d_{1i} = ([\mathbf{x}]_{i+1} - [\mathbf{x}]_i) + (x_{\max} - x_{\min}) + \varsigma, \quad i = 1, \dots, n-1, \quad (9.36)$$

or as

$$d_{1i} = |[\mathbf{x}]_{i+1} - [\mathbf{x}]_i| + \varsigma, \quad i = 1, \dots, n-1. \quad (9.37)$$

Here,  $\varsigma$  is a small positive constant, while  $x_{\min}$  and  $x_{\max}$  are the lower and the upper bounds of all entries in  $\mathbf{x}$ , that is, and  $x_{\min} \leq [\mathbf{x}]_i \leq x_{\max}$ ,  $i = 1, \dots, n$ . By (9.35), we have

$$\frac{(n-1)d_{1i}}{d_1} \log \left( \frac{(n-1)d_{1i}}{d_1} \right) \geq \frac{(n-1)d_{1i}}{d_1} - 1,$$

with  $d_1 = \sum_{i=1}^{n-1} d_{1i}$ , and we infer that  $\Lambda_1 \geq 0$ . The minimum value of  $\Lambda_1$  is attained when all  $d_{1i}$  are the same, and the solutions to (9.34) approach the discrete approximation of a first-order polynomial as  $\alpha \rightarrow \infty$ . The second-order penalty function (corresponding to the entropy of the vector of second-order differences of  $\mathbf{x}$ ) is given by

$$\Lambda_2(\mathbf{x}) = \sum_{i=2}^{n-1} \frac{(n-2)d_{2i}}{\sum_{i=2}^{n-1} d_{2i}} \log \left( \frac{(n-2)d_{2i}}{\sum_{i=2}^{n-1} d_{2i}} \right),$$

with

$$d_{2i} = ([\mathbf{x}]_{i+1} - 2[\mathbf{x}]_i + [\mathbf{x}]_{i-1}) + 2(x_{\max} - x_{\min}) + \varsigma, \quad i = 2, \dots, n-1, \quad (9.38)$$

or

$$d_{2i} = |[\mathbf{x}]_{i+1} - 2[\mathbf{x}]_i + [\mathbf{x}]_{i-1}| + \varsigma, \quad i = 2, \dots, n-1. \quad (9.39)$$

As before,  $\Lambda_2 \geq 0$  attains its minimum when all  $d_{2i}$  coincide, and the solutions to (9.34) approach the discrete approximation of a second-order polynomial as  $\alpha \rightarrow \infty$ . In comparison, under similar conditions, Tikhonov regularization with the first- and second-order difference regularization matrices will yield a constant solution and a straight line, respectively.

The minimization of the Tikhonov function (9.34) can be performed by using the Newton method with

$$\mathbf{g}_\alpha(\mathbf{x}) = \nabla \mathcal{F}_\alpha(\mathbf{x}) = \mathbf{K}(\mathbf{x})^T [\mathbf{F}(\mathbf{x}) - \mathbf{y}^\delta] + \alpha \nabla \Lambda(\mathbf{x}),$$

and the Hessian approximation

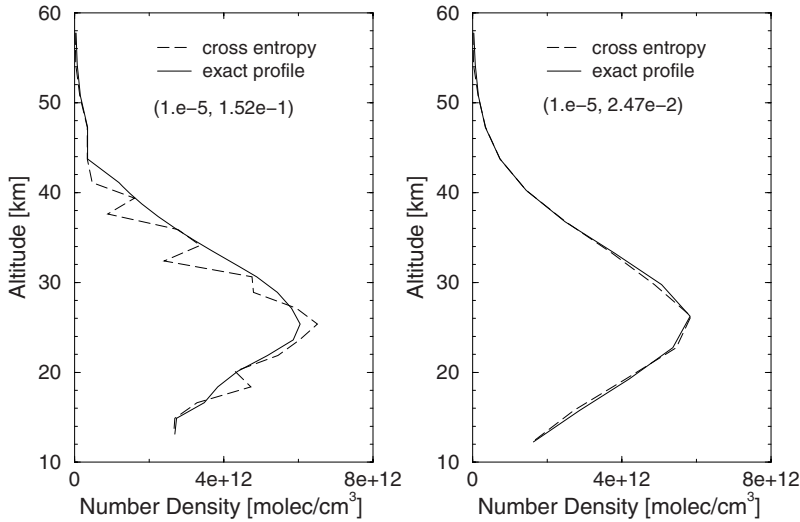
$$\mathbf{G}_\alpha(\mathbf{x}) = \nabla^2 \mathcal{F}_\alpha(\mathbf{x}) \approx \mathbf{K}(\mathbf{x})^T \mathbf{K}(\mathbf{x}) + \alpha \nabla^2 \Lambda(\mathbf{x}).$$

To be more concrete, at the iteration step  $k$ , the search direction  $\mathbf{p}_{\alpha k}^\delta$  is computed as the solution of the Newton equation

$$\mathbf{G}_\alpha(\mathbf{x}_{\alpha k}^\delta) \mathbf{p} = -\mathbf{g}_\alpha(\mathbf{x}_{\alpha k}^\delta),$$

the step length  $\tau_k$  is determined by imposing the descent condition, and the new iterate is taken as  $\mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_{\alpha k}^\delta + \tau_k \mathbf{p}_{\alpha k}^\delta$ .

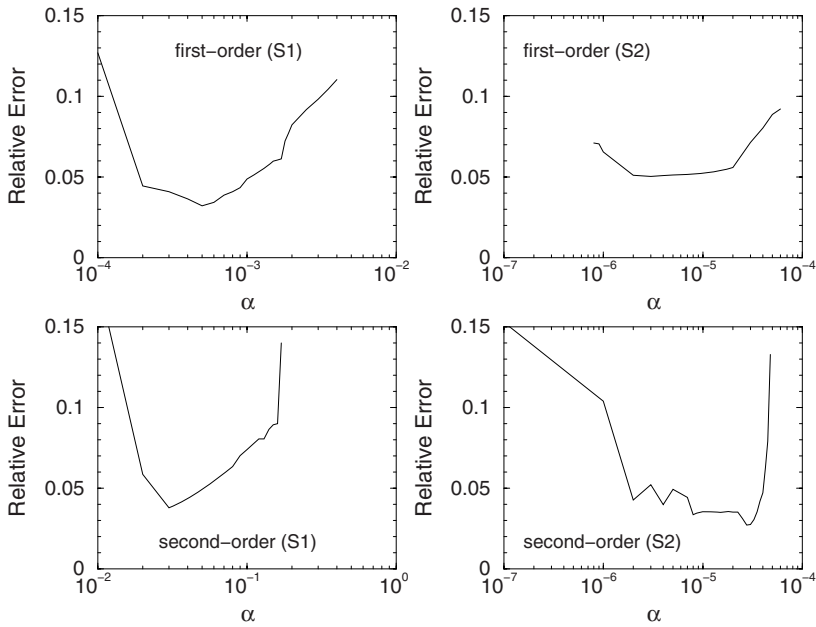
In Figure 9.4 we plot the retrieved  $\text{O}_3$  profiles for the cross entropy regularization with the penalty term  $\Lambda_c$ . Because in this case, the regularization matrix acts like a diagonal matrix, the solution errors may become extremely large. Specifically, on a fine grid, the number densities, with respect to which the retrieval is insensitive, are close to the a priori.



**Fig. 9.4.** Retrieved  $\text{O}_3$  profiles computed with the cross entropy regularization on a retrieval grid with 36 levels (left) and on a retrieval grid with 24 levels (right). The numbers in parentheses represent the values of the regularization parameter and of the relative solution error.

The plots in Figure 9.5 illustrate the solution errors for the first- and second-order entropy regularization with the penalty terms  $\Lambda_1$  and  $\Lambda_2$ , respectively. As for Tikhonov regularization, the error curves possess a minimum for an optimal value of the regularization parameter. The minima of the solution errors are  $3.32 \cdot 10^{-2}$  and  $5.05 \cdot 10^{-2}$  for the first-order entropy regularization with the selection criteria (9.36) and (9.37), respectively, and  $3.79 \cdot 10^{-2}$  and  $2.73 \cdot 10^{-2}$  for the second-order entropy regularization with the selection criteria (9.38) and (9.39), respectively. Comparing both regularization methods we observe that





**Fig. 9.5.** Relative solution errors for the first-order entropy regularizations with the selection criteria (9.36) (S1) and (9.37) (S2), and the second-order entropy regularization with the selection criteria (9.38) (S1) and (9.39) (S2).

- (1) the first- and second-order entropy regularizations yield results of comparable accuracies;
- (2) the domains of variation of the regularization parameter with acceptable reconstruction errors are larger for the selection criteria (9.37) and (9.39).

A pertinent analysis of the maximum entropy regularization can be found in Engl et al. (2000), while for applications of the second-order entropy regularization in atmospheric remote sensing we refer to Steinwagner et al. (2006).

# A

## Analysis of continuous ill-posed problems

In this appendix we analyze the ill-posedness of the Fredholm integral equation of the first kind

$$y(\nu) = \int_0^{z_{\max}} k(\nu, z) x(z) \, dz, \quad \nu \in [\nu_{\min}, \nu_{\max}],$$

written in operator form as

$$Kx = y. \tag{A.1}$$

We begin our presentation by recalling some fundamental results of functional analysis.

### A.1 Elements of functional analysis

Let  $X$  be a real vector space. The function  $\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{R}$  is called a Hermitian form if

- (1)  $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$  (linearity),
- (2)  $\langle x, y \rangle = \langle y, x \rangle$  (symmetry),

for all  $x, y, z \in X$  and all  $\alpha, \beta \in \mathbb{R}$ . A Hermitian form with the properties

- (1)  $\langle x, x \rangle \geq 0$  (positivity),
- (2)  $\langle x, x \rangle = 0$  if and only if  $x = 0$  (definiteness),

is called a scalar product. A vector space with a specified scalar product is called a pre-Hilbert space. In terms of the scalar product in  $X$ , the norm  $\|x\| = \sqrt{\langle x, x \rangle}$  can be introduced, after which  $X$  becomes a normed space.

Given a sequence  $\{x_n\}_{n \in \mathbb{N}}$  of elements of a normed space  $X$ , we say that  $x_n$  converges to an element  $x$  of  $X$ , if  $\|x_n - x\| \rightarrow 0$  as  $n \rightarrow \infty$ . A sequence  $\{x_n\}_{n \in \mathbb{N}}$  of elements of a normed space  $X$  is called a Cauchy sequence, if  $\|x_n - x_m\| \rightarrow 0$  as  $n, m \rightarrow \infty$ . Any convergent sequence is a Cauchy sequence, but the converse result is not true in general.

A subset  $U$  of a normed space  $X$  is called complete if any Cauchy sequence of elements of  $U$  converges to an element of  $U$ . A normed space is called a Banach space if it is complete, and a pre-Hilbert space is called a Hilbert space if it is complete.

A subset  $U$  of a normed space  $X$  is said to be closed if it contains all its limit points. For any set  $U$  in a normed space  $X$ , the closure of  $U$  is the union of  $U$  with the set of all limit points of  $U$ , and the closure of  $U$  is written as  $\overline{U}$ . Obviously,  $U$  is contained in  $\overline{U}$ , and  $U = \overline{U}$  if  $U$  is closed. Note that complete sets are closed, and any closed subset of a complete set is complete.

A subset  $U$  of a normed space  $X$  is said to be dense in  $X$  if, for any  $x \in X$ , there exists a sequence  $\{x_n\}_{n \in \mathbb{N}}$  in  $U$  such that  $\|x_n - x\| \rightarrow 0$  as  $n \rightarrow \infty$ . Any set  $U$  is dense in its closure  $\overline{U}$ , and  $\overline{U}$  is the largest set in which  $U$  is dense (if  $U$  is dense in  $V$ , then  $V \subset \overline{U}$ ). If  $U$  is dense in a Hilbert space  $X$ , then  $\overline{U} = X$ , and conversely, if  $\overline{U} = X$ , then  $U$  is dense in the Hilbert space  $X$ .

Two elements  $x$  and  $y$  of a Hilbert space  $X$  are called orthogonal if  $\langle x, y \rangle = 0$ ; we then write  $x \perp y$ . If an element  $x$  is orthogonal to any element of a set  $U$ , we call it orthogonal to the set  $U$ , and write  $x \perp U$ . Similarly, if any element of a set  $U$  is orthogonal to any element of the set  $V$ , we call these sets orthogonal, and write  $U \perp V$ .

If the subset  $U$  of a Hilbert space  $X$  is dense in  $X$  ( $\overline{U} = X$ ), and  $x$  is orthogonal to  $U$ , then  $x$  is the zero element of  $X$ , i.e.,  $x = 0$ .

A set in a Hilbert space is called orthogonal, if any two elements of the set are orthogonal. If, moreover, the norm of any element is one, the set is called orthonormal.

$U \subset X$  is a linear subspace of  $X$  if  $\alpha x + \beta y \in U$  for any scalars  $\alpha$  and  $\beta$ , and any  $x, y \in U$ . Let  $X$  be a Hilbert space,  $U$  a complete linear subspace of  $X$ , and  $x$  an element of  $X$ . Since, for any  $y \in U$ , we have  $\|x - y\| \geq 0$ , we see that the set  $\{\|x - y\| / y \in U\}$  possesses an infimum. Setting  $d = \inf \{\|x - y\| / y \in U\}$  and taking into account that  $U$  is complete, it can be shown that there exists a unique element  $z \in U$  such that  $d = \|x - z\|$ . The element  $z$  gives the best approximation of  $x$  among all elements of  $U$ . The operator  $P : X \rightarrow U$  mapping  $x$  onto its best approximation  $Px = z$  is a bounded linear operator with the properties  $P^2 = P$  and  $\langle Px_1, x_2 \rangle = \langle x_1, Px_2 \rangle$  for any  $x_1, x_2 \in X$ . It is called the orthogonal projection operator from  $X$  onto  $U$ , and  $Px$  is called the projection of  $x$  onto  $U$ . Occasionally, we will write  $P_U$  to denote the orthogonal projection operator onto the complete linear subspace  $U$ . If  $U$  is a finite-dimensional linear subspace of the Hilbert space  $X$  with the orthonormal basis  $\{u_n\}_{n \in \mathbb{N}}$ , then the orthogonal projection operator is given by  $Px = \sum_{k=1}^n \langle x, u_k \rangle u_k$  for  $x \in X$ .

The set of all elements orthogonal to a subset  $U$  of a Hilbert space  $X$  is called the orthogonal complement of  $U$ ,  $U^\perp = \{x \in X / x \perp U\}$ , and we note that  $U^\perp$  is a complete linear subspace of  $X$ . If  $U$  is a complete linear subspace of the Hilbert space  $X$  and  $P$  is the orthogonal projection operator from  $X$  onto  $U$ , then any element  $x \in X$  can be uniquely decomposed as  $x = Px + x^\perp$ , where  $x^\perp \in U^\perp$ . This result is known as the theorem of orthogonal projection. We also note the decomposition  $X = U \oplus U^\perp$  for any complete linear subspace  $U$  of  $X$ .

A system of elements  $\{u_n\}_{n \in \mathbb{N}}$  is called closed in the Hilbert space  $X$  if there are no elements in  $X$  orthogonal to any element of the set except the zero element, i.e.,  $\langle x, u_n \rangle = 0$  for  $n \in \mathbb{N}$  implies  $x = 0$ . A system of elements  $\{u_n\}_{n \in \mathbb{N}}$  is called complete in the Hilbert space  $X$  if the linear span of  $\{u_n\}_{n \in \mathbb{N}}$ ,

$$\text{span} \{u_n\}_{n \in \mathbb{N}} = \left\{ x = \sum_{n=1}^N \alpha_n u_n / \alpha_n \in \mathbb{R}, N \in \mathbb{N} \right\}$$

is dense in  $X$ , that is,  $\overline{\text{span}\{u_n\}_{n \in \mathbb{N}}} = X$ . The following result connects closedness and completeness in Hilbert spaces: a system of elements  $\{u_n\}_{n \in \mathbb{N}}$  is complete in a Hilbert space  $X$  if and only if it is closed in  $X$ .

A map  $K : X \rightarrow Y$  between the Hilbert spaces  $X$  and  $Y$  is called linear if  $K$  transforms linear combinations of elements into the same linear combination of their images,

$$K(\alpha_1 x_1 + \dots + \alpha_n x_n) = \alpha_1 K(x_1) + \dots + \alpha_n K(x_n).$$

Linear maps are also called linear operators and in linear algebra we usually write arguments without brackets,  $K(x) = Kx$ . The linearity of a map is a very strong condition which is shown by the following equivalent statements:

(1)  $K$  is bounded, i.e., there exists a positive constant  $m$  such that

$$\|Kx\| \leq m \|x\|, \quad x \in X;$$

(2)  $K$  is continuous.

Each number for which the above inequality holds is called a bound for  $K$ , and the induced operator norm defined as

$$\|K\| = \sup_{x \in X, x \neq 0} \frac{\|Kx\|}{\|x\|} = \sup_{\|x\|=1} \|Kx\|$$

is the smallest bound for  $K$ . The range space of  $K$ ,  $\{Kx/x \in X\}$ , will be denoted by  $\mathcal{R}(K)$  and the null space of  $K$ ,  $\{x \in X/Kx = 0\}$ , will be denoted by  $\mathcal{N}(K)$ .

For any bounded linear operator  $K : X \rightarrow Y$  acting from the Hilbert space  $X$  into the Hilbert space  $Y$ , there exists a bounded linear operator  $K^* : Y \rightarrow X$  called the adjoint operator of  $K$  satisfying the requirement

$$\langle Kx, y \rangle = \langle x, K^*y \rangle$$

for all  $x \in X$  and  $y \in Y$ . The following relations between the range and null spaces of  $K$  and  $K^*$  hold:

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

and

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

Note that  $K$  is a bijective operator, if and only if  $K^*$  is bijective.

$\mathbb{R}^n$  is a Hilbert space under the Euclidean inner product,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{k=1}^n [\mathbf{x}]_k [\mathbf{y}]_k,$$

and the induced norm is the Euclidean norm,

$$\|\mathbf{x}\| = \sqrt{\sum_{k=1}^n [\mathbf{x}]_k^2}.$$

The space of real-valued, square integrable functions on the interval  $[a, b]$ , denoted by  $L^2([a, b])$ , is a Hilbert space under the inner product

$$\langle x, y \rangle = \int_a^b x(t)y(t) dt,$$

and the induced norm

$$\|x\| = \sqrt{\int_a^b x(t)^2 dt}.$$

The Fredholm integral operator of the first kind

$$(Kx)(s) = \int_a^b k(s, t)x(t) dt, \quad s \in [a, b], \quad (\text{A.2})$$

is bounded if

$$m = \int_a^b \int_a^b k(s, t)^2 x(t) ds dt < \infty,$$

in which case,  $\|K\| \leq \sqrt{m}$ . The adjoint of  $K$  is given by

$$(K^*y)(t) = \int_a^b k(s, t)y(s) ds, \quad t \in [a, b],$$

and  $K$  is self-adjoint, if and only if  $k(s, t) = k(t, s)$ .

An operator  $K : X \rightarrow Y$  between the Hilbert spaces  $X$  and  $Y$  is called compact, if and only if the image of any bounded set is a relatively compact set, i.e., if the closure of its image is a compact subset of  $Y$ .

The analysis of the Fredholm integral equation of the first kind  $Kx = y$  with  $K$  as in (A.2) relies on the following fundamental result: if  $k \in L^2([a, b] \times [a, b])$ , then the integral operator  $K$  is bounded and compact.

## A.2 Least squares solution and generalized inverse

Before introducing the concepts of least squares solution and generalized inverse for equation (A.1), we note that  $\mathcal{R}(K)$  is closed if  $\mathcal{R}(K) = \overline{\mathcal{R}(K)}$ , but in general  $\mathcal{R}(K)$  is not closed, and so,  $\mathcal{R}(K) \subset \overline{\mathcal{R}(K)}$ . Similarly,  $\mathcal{R}(K)$  is dense in  $Y$  if  $\overline{\mathcal{R}(K)} = Y$ , but we cannot expect that  $\mathcal{R}(K)$  is dense in  $Y$ , and therefore,  $\overline{\mathcal{R}(K)} \subset Y$ . Thus,  $\mathcal{R}(K) \subset \overline{\mathcal{R}(K)} \subset Y$ , and, in view of the orthogonal projection theorem, we have  $Y = \overline{\mathcal{R}(K)} \oplus \mathcal{R}(K)^\perp$ .

**Theorem A.1.** *Let  $y \in Y$ . The following statements are equivalent:*

- (1)  $x \in X$  has the property  $P_{\overline{\mathcal{R}(K)}}y = Kx$ ;
- (2)  $x \in X$  is a least squares solution of equation (A.1), i.e.,

$$\|y - Kx\| = \inf \{\|y - Kz\| : z \in X\};$$

(3)  $x \in X$  solves the normal equation

$$K^* Kx = K^* y. \quad (\text{A.3})$$

*Proof.* To justify these equivalences, we will prove the implications: (1)  $\Rightarrow$  (2)  $\Rightarrow$  (3)  $\Rightarrow$

(1). Let  $x \in X$  be such that  $P_{\overline{\mathcal{R}(K)}} y = Kx$ . Since

$$\begin{aligned} \|y - P_{\overline{\mathcal{R}(K)}} y\| &= \inf \left\{ \|y - y'\| / y' \in \overline{\mathcal{R}(K)} \right\} \\ &\leq \inf \left\{ \|y - y'\| / y' \in \mathcal{R}(K) \right\} \\ &= \inf \left\{ \|y - Kz\| / z \in X \right\}, \end{aligned}$$

we deduce that  $x$  is a least squares solution of (A.1). Let us consider now the quadratic polynomial

$$F(\lambda) = \|y - K(x + \lambda z)\|^2$$

for some  $z \in X$ . The derivative of  $F$  with respect to  $\lambda$  is given by

$$F'(\lambda) = 2\lambda \|Kz\|^2 - 2\langle z, K^*(y - Kx) \rangle,$$

and if  $x$  is a least squares solution of (A.1), then  $F'(0) = 0$  for all  $z \in X$ . Thus,  $x$  solves the normal equation (A.3). Finally, let  $x$  be a solution of the normal equation (A.3), i.e.,  $K^*(y - Kx) = 0$ . Then, we have  $y - Kx \in \mathcal{N}(K^*) = \mathcal{R}(K)^\perp$ , and further,  $y - Kx \perp \mathcal{R}(K)$ . From  $y - P_{\overline{\mathcal{R}(K)}} y \perp \mathcal{R}(K)$  and the uniqueness of the orthogonal projection  $P_{\overline{\mathcal{R}(K)}} y$ , it follows that  $Kx = P_{\overline{\mathcal{R}(K)}} y$ , and the proof is finished. Note that the name of equation (A.3) comes from the fact that the residual  $y - Kx$  is orthogonal (normal) to  $\mathcal{R}(K)$ .  $\square$

**Theorem A.2.** *The normal equation (A.3) has solutions, if and only if  $y \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ .*

*Proof.* Let us assume that  $x$  is a solution of (A.3). From the decomposition  $y = Kx + (y - Kx)$  and the result  $y - Kx \in \mathcal{R}(K)^\perp$  (which follows from the proof of implication (3)  $\Rightarrow$  (1) in Theorem A.1) we infer that  $y \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ . Conversely, let  $y \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ . Then, there exists  $x \in X$  and  $y^\perp \in \mathcal{R}(K)^\perp$  such that  $y = Kx + y^\perp$ . Application of the projection operator  $P_{\overline{\mathcal{R}(K)}}$  and the result  $Kx \in \mathcal{R}(K) \subset \overline{\mathcal{R}(K)}$ , yield  $P_{\overline{\mathcal{R}(K)}} y = Kx$ , whence, by Theorem A.1, we deduce that  $x$  solves the normal equation (A.3).  $\square$

Thus, if  $y \notin \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ , no solution of the normal equation (A.3) exists, or equivalently, no least squares solution of equation (A.1) exists.

From the whole set of least squares solutions to equation (A.1), the element of minimal norm is called the least squares minimum norm solution. Essentially, the least squares minimum norm solution  $x$  is the least squares solution of equation (A.1) in  $\mathcal{N}(K)^\perp$ . If  $x_0$  is an arbitrary element in  $\mathcal{N}(K)$ , that is,  $Kx_0 = 0$ , then  $K(x + x_0) = P_{\overline{\mathcal{R}(K)}} y$  and by Theorem A.1,  $x + x_0$  is a least squares solution of equation (A.1) in  $X$ . Therefore, the set of all least squares solutions is  $x + \mathcal{N}(K)$ .

The linear map  $K^\dagger : \mathcal{D}(K^\dagger) \rightarrow X$  with the domain  $\mathcal{D}(K^\dagger) = \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ , which maps any  $y \in \mathcal{D}(K^\dagger)$  into the least squares minimum norm solution  $x$  of equation (A.1), that is,  $x = K^\dagger y$ , is called the generalized inverse or the Moore–Penrose inverse. Note that  $K^\dagger$  is a linear operator which is defined on all  $Y$  if  $\mathcal{R}(K)$  is closed ( $\mathcal{R}(K) = \overline{\mathcal{R}(K)}$ ). .

**Theorem A.3.** *The generalized inverse  $K^\dagger : \mathcal{D}(K^\dagger) \rightarrow X$  is bounded (continuous), if and only if  $\mathcal{R}(K)$  is closed.*

*Proof.* First we assume that  $\mathcal{R}(K)$  is closed, so that  $\mathcal{D}(K^\dagger) = Y$ . Then, in view of the closed graph theorem,  $K^\dagger$  is bounded. Conversely, let  $K^\dagger$  be bounded. Since  $\mathcal{D}(K^\dagger)$  is dense in  $Y$ ,  $K^\dagger$  has a unique continuous extension  $\overline{K^\dagger}$  to  $Y$ , such that  $K\overline{K^\dagger} = P_{\overline{\mathcal{R}(K)}}$ . Then, for  $y \in \overline{\mathcal{R}(K)}$ , we have  $y = P_{\overline{\mathcal{R}(K)}}y = K\overline{K^\dagger}y$ , which shows that  $y \in \mathcal{R}(K)$ . Hence,  $\overline{\mathcal{R}(K)} \subseteq \mathcal{R}(K)$ , and  $\mathcal{R}(K)$  is closed.  $\square$

Thus, for a compact operator with a non-closed range space, the least squares minimum norm solution  $x$  does not depend continuously on the data  $y$ .

**Theorem A.4.** *The range space  $\mathcal{R}(K)$  of a compact operator  $K$  is closed, if and only if it is finite-dimensional.*

*Proof.* If  $\mathcal{R}(K)$  is closed, then it is complete (as a subset of the Hilbert space  $Y$ ), and by Banach's open mapping theorem, the operator  $K|_{\mathcal{N}(K)^\perp} : \mathcal{N}(K)^\perp \rightarrow \mathcal{R}(K)$  is bijective and continuously invertible. Then, the identity operator

$$I = K \left( K|_{\mathcal{N}(K)^\perp} \right)^{-1} : \mathcal{R}(K) \rightarrow \mathcal{R}(K)$$

is compact, since the product of a compact and a continuous operator is compact. The conclusion then follows by taking into account that the identity operator  $I : \mathcal{R}(K) \rightarrow \mathcal{R}(K)$  is compact, if and only if  $\mathcal{R}(K)$  is of finite dimension.  $\square$

The important conclusion is that if  $K$  is a compact operator acting between the infinite-dimensional Hilbert spaces  $X$  and  $Y$ , and  $\mathcal{R}(K)$  is infinite-dimensional (e.g., an integral operator with a non-degenerate kernel), then  $\mathcal{R}(K)$  is not closed, and as a result, the linear equation (A.1) is ill-posed in the sense that the first and the third Hadamard conditions are violated.

### A.3 Singular value expansion of a compact operator

Any compact operator between Hilbert spaces admits a singular value expansion. For a compact operator  $K : X \rightarrow Y$  and its compact adjoint operator  $K^* : Y \rightarrow X$ , the non-negative square roots of the eigenvalues of the self-adjoint compact operator  $K^*K : X \rightarrow X$  are called the singular values of  $K$ . If  $\{\sigma_i\}_{i \in \mathbb{N}}$  denotes the sequence of non-zero singular values of  $K$  appearing in decreasing order, then there exist the orthonormal sequences  $\{v_i\}_{i \in \mathbb{N}}$  in  $X$  and  $\{u_i\}_{i \in \mathbb{N}}$  in  $Y$  such that

$$Kv_i = \sigma_i u_i, \quad K^*u_i = \sigma_i v_i, \quad i \in \mathbb{N}.$$

The countable set of triples  $\{(\sigma_i; v_i, u_i)\}_{i \in \mathbb{N}}$  is called the singular system of the compact operator  $K$ . The right singular vectors  $\{v_i\}_{i \in \mathbb{N}}$  form an orthonormal basis for  $\mathcal{N}(K)^\perp$ ,

$$\mathcal{N}(K)^\perp = \overline{\text{span}\{v_i\}_{i \in \mathbb{N}}}$$

while the left singular vectors  $\{u_i\}_{i \in \mathbb{N}}$  form an orthonormal basis for  $\overline{\mathcal{R}(K)}$ ,

$$\overline{\mathcal{R}(K)} = \overline{\text{span}\{u_i\}_{i \in \mathbb{N}}}.$$

If  $\mathcal{R}(K)$  is infinite-dimensional, there holds

$$\lim_{i \rightarrow \infty} \sigma_i = 0,$$

and for any  $x \in X$ , we have the singular value expansions

$$x = \sum_{i=1}^{\infty} \langle x, v_i \rangle v_i + P_{\mathcal{N}(K)} x, \quad (\text{A.4})$$

and

$$Kx = \sum_{i=1}^{\infty} \sigma_i \langle x, v_i \rangle u_i.$$

#### A.4 Solvability and ill-posedness of the linear equation

In this section we analyze the solvability of the linear equation (A.1) by making use of the singular value expansion of the compact operator  $K$ . To simplify our presentation, we assume that  $K$  is injective, in which case,  $\mathcal{N}(K) = \emptyset$ . If  $K$  is injective and  $x$  is a least squares solution of equation (A.1), then from  $Kx = P_{\overline{\mathcal{R}(K)}} y$ , we deduce that  $x$  is unique. Therefore, instead of using the appellation least squares minimum norm solution,  $x$  will be simply called the least squares solution of equation (A.1).

**Theorem A.5.** *The linear equation (A.1) is solvable, if and only if  $y \in \overline{\mathcal{R}(K)}$  and  $y$  satisfies the Picard condition*

$$\sum_{i=1}^{\infty} \frac{\langle y, u_i \rangle^2}{\sigma_i^2} < \infty. \quad (\text{A.5})$$

*In this case, the solution is given by*

$$x = \sum_{i=1}^{\infty} \frac{1}{\sigma_i} \langle y, u_i \rangle v_i. \quad (\text{A.6})$$

*Proof.* Let  $x$  be the solution of equation (A.1), i.e.,  $Kx = y$  for  $y \in Y$ , and let  $y_0 \in \mathcal{N}(K^*)$ . Then, from

$$\langle y, y_0 \rangle = \langle Kx, y_0 \rangle = \langle x, K^* y_0 \rangle = 0,$$



the necessity of condition  $y \in \overline{\mathcal{R}(K)} = \mathcal{N}(K^*)^\perp$  follows. As  $x$  is an element of  $X$ ,  $x$  possesses the representation (A.4) with the Fourier coefficients

$$\langle x, v_i \rangle = \frac{1}{\sigma_i} \langle x, K^* u_i \rangle = \frac{1}{\sigma_i} \langle Kx, u_i \rangle = \frac{1}{\sigma_i} \langle y, u_i \rangle. \quad (\text{A.7})$$

Then, from

$$\sum_{i=1}^{\infty} \frac{1}{\sigma_i^2} \langle y, u_i \rangle^2 = \sum_{i=1}^{\infty} \langle x, v_i \rangle^2 \leq \|x\|^2,$$

we see that the series (A.5) converges, and the necessity of the Picard condition is apparent. Conversely, let  $y \in \overline{\mathcal{R}(K)}$  and assume that  $y$  satisfies the Picard condition. Then, by considering the partial sums in (A.5), we deduce that the series  $\sum_{i=1}^{\infty} (1/\sigma_i) \langle y, u_i \rangle v_i$  converges in the Hilbert space  $X$ . Let  $x$  be the sum of this series, that is, let  $x$  be given by (A.6). Application of the operator  $K$  to  $x$  yields

$$Kx = \sum_{i=1}^{\infty} \langle y, u_i \rangle u_i = P_{\overline{\mathcal{R}(K)}} y. \quad (\text{A.8})$$

As  $y \in \overline{\mathcal{R}(K)}$ , we have  $P_{\overline{\mathcal{R}(K)}} y = y$ , and we infer that  $Kx = y$ .  $\square$

The Picard condition, which guarantees the solvability of equation (A.1), states that the generalized Fourier coefficients  $|\langle y, u_i \rangle|$  must decay faster to zero than the singular values  $\sigma_i$ . Essentially, for  $y \in \overline{\mathcal{R}(K)}$ , the Picard condition implies that  $y \in \mathcal{R}(K)$ . As stated by the next theorem, the converse result is also true.

**Theorem A.6.** *If  $y \in \mathcal{R}(K)$ , then the Picard condition (A.5) is satisfied. As a result, the solution of equation (A.1) exists and is given by (A.6).*

*Proof.* Let  $y \in \mathcal{R}(K)$ . Then, there exists  $x \in X$ , such that  $Kx = y$ . By (A.4), we may represent  $x$  in terms of the orthonormal basis  $\{v_i\}_{i \in \mathbb{N}}$  as  $x = \sum_{i=1}^{\infty} \langle x, v_i \rangle v_i$ . Taking into account that (cf. (A.7))  $\langle x, v_i \rangle = (1/\sigma_i) \langle y, u_i \rangle$ , we find that  $x$  is given by (A.6). Consequently, the series (A.5) converges, and the sum of this series is  $\|x\|^2$ .  $\square$

In practice we are dealing with noisy data for which the requirement  $y \in \overline{\mathcal{R}(K)}$  is not satisfied. In general,  $y \in Y$ , and since  $\mathcal{R}(K)$  is not dense in  $Y$ , we have  $\overline{\mathcal{R}(K)} \subset Y$ . Therefore, equation (A.1) is not solvable for arbitrary noisy data. However, by Theorem A.2, we know that equation (A.1) has a least squares solution, if and only if  $y \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ . The existence of the least squares solution to equation (A.1) is given by the following theorem.

**Theorem A.7.** *If  $y \in Y = \overline{\mathcal{R}(K)} \oplus \mathcal{R}(K)^\perp$  and  $y$  satisfies the Picard condition, then the least squares solution of equation (A.1) exists and is given by (A.6).*

*Proof.* Let  $y \in Y$  satisfy the Picard condition (A.5). We employ the same arguments as in the proof of Theorem A.5: by virtue of (A.5), the series  $\sum_{i=1}^{\infty} (1/\sigma_i) \langle y, u_i \rangle v_i$  converges, and if  $x$  is the sum of this series, from (A.8), we have  $Kx = P_{\overline{\mathcal{R}(K)}} y$ . Then, by Theorem A.1, we conclude that  $x$  given by (A.6) is the least squares solution of equation (A.1).  $\square$

The next result is the analog of Theorem A.6.

**Theorem A.8.** *If  $y \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ , then the Picard condition is satisfied. As a result, the least squares solution of equation (A.1) exists and is given by (A.6).*

*Proof.* Let us assume that  $y \in \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ . Then, there exists  $x \in X$  and  $y^\perp \in \mathcal{R}(K)^\perp$  such that  $y = Kx + y^\perp$ . By (A.4), we can expand  $x = \sum_{i=1}^{\infty} \langle x, v_i \rangle v_i$ . Taking into account that (cf. (A.7))  $\langle x, v_i \rangle = (1/\sigma_i) \langle y - y^\perp, u_i \rangle$  and that  $\langle y^\perp, u_i \rangle = 0$  (since  $y^\perp \in \mathcal{R}(K)^\perp$  and  $u_i \in \mathcal{R}(K)$ ), we deduce that  $x$  has the expansion (A.6) and that the sum of the series (A.5) is  $\|x\|^2$ .  $\square$

Turning now to the question of stability, we first observe that, in view of (A.6), the series representation for the generalized inverse  $K^\dagger : \mathcal{D}(K^\dagger) \rightarrow X$  is given by

$$K^\dagger = \sum_{i=1}^{\infty} \frac{1}{\sigma_i} \langle \cdot, u_i \rangle v_i. \quad (\text{A.9})$$

To prevent inherent ambiguities,  $y \in \mathcal{R}(K)$  will be called the exact data and in the presence of the noise  $\delta$ ,  $y^\delta = y + \delta \in Y$  will be referred to as the noisy data. The following conclusions arising from our above analysis can be drawn:

- (1) If  $y \in \mathcal{R}(K)$ , then by Theorem A.6, the exact solution  $x^\dagger$  exists and is given by

$$x^\dagger = K^\dagger y = \sum_{i=1}^{\infty} \frac{1}{\sigma_i} \langle y, u_i \rangle v_i.$$

- (2) If  $y^\delta \in Y$  does not satisfy the Picard condition, then by Theorem A.8, we have  $y^\delta \notin \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ , and further, by Theorem A.2, we deduce that the least squares solution does not exist.
- (3) If  $y^\delta \in Y$  satisfies the Picard condition, then by Theorem A.7, the least squares solution exists and is given by

$$x^\delta = K^\dagger y^\delta = \sum_{i=1}^{\infty} \frac{1}{\sigma_i} \langle y^\delta, u_i \rangle v_i.$$

The least squares solution  $x^\delta$  can be regarded as an approximation of the exact solution  $x^\dagger$ , and the above series representations show how errors in the data affect the solution. Assuming that  $y^\delta = y + \Delta u_i$  for some noise level  $\Delta \in \mathbb{R}$ , we obtain the least squares solution  $x^\delta = x^\dagger + (\Delta/\sigma_i) v_i$ . Thus,  $\|x^\delta - x^\dagger\| = \Delta/\sigma_i$ , and two situations can be distinguished:

- (1) If  $\mathcal{R}(K)$  is of finite dimension, e.g.,  $K$  is an integral operator with a degenerate kernel, then there are finitely many singular values. In this case, the solution error  $\Delta/\sigma_i$  corresponding to a small singular value is bounded but may become unacceptably large.
- (2) If  $\mathcal{R}(K)$  is infinite-dimensional, e.g.,  $K$  is an integral operator with a non-degenerate kernel, then  $\lim_{i \rightarrow \infty} \sigma_i = 0$ . As a result,  $\Delta/\sigma_i \rightarrow \infty$  as  $i \rightarrow \infty$ , and the solution error increases without bound.

# B

## Standard-form transformation for rectangular regularization matrices

In this appendix we discuss the transformation to the standard form when the rectangular regularization matrix  $\mathbf{L} \in \mathbb{R}^{p \times n}$ ,  $p < n$ , has full row rank, i.e.,  $\text{rank}(\mathbf{L}) = p$ . The transformation to the standard form depends on the type of the regularization method employed (Hansen, 1998). For direct regularization methods, we need to compute the matrix  $\tilde{\mathbf{K}}$  of the standard-form problem, while for iterative regularization methods we merely need to compute matrix-vector products with  $\tilde{\mathbf{K}}$  and  $\tilde{\mathbf{K}}^T$ .

### B.1 Explicit transformations

Explicit standard-form transformations can be derived by using the GSVD of the matrix pair  $(\mathbf{K}, \mathbf{L})$  or by means of orthogonal transformations.

Let us consider the GSVD of  $(\mathbf{K}, \mathbf{L})$ , i.e.,

$$\mathbf{K} = \mathbf{U}\Sigma_1\mathbf{W}^{-1}, \quad \mathbf{L} = \mathbf{V}\Sigma_2\mathbf{W}^{-1}, \quad (\text{B.1})$$

with

$$\Sigma_1 = \begin{bmatrix} \text{diag}(\sigma_i)_{p \times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-p} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \text{diag}(\mu_i)_{p \times p} & \mathbf{0} \end{bmatrix}.$$

As  $\mathbf{L}$  is of full row rank, the right inverse  $\mathbf{L}^\dagger$  of  $\mathbf{L}$ , satisfying  $\mathbf{L}\mathbf{L}^\dagger = \mathbf{I}_p$ , can be constructed as

$$\mathbf{L}^\dagger = \mathbf{W} \begin{bmatrix} \text{diag}\left(\frac{1}{\mu_i}\right)_{p \times p} \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T. \quad (\text{B.2})$$

The augmented residual vector

$$\mathbf{r} = \begin{bmatrix} \mathbf{K} \\ \sqrt{\alpha}\mathbf{L} \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{y}^\delta \\ \mathbf{0} \end{bmatrix}, \quad (\text{B.3})$$

can be expressed in terms of a new variable  $\bar{\mathbf{x}}$ , defined through the transformation

$$\mathbf{x} = \mathbf{L}^\dagger \bar{\mathbf{x}} + \mathbf{x}_0^\delta, \quad \bar{\mathbf{x}} \in \mathbb{R}^p,$$

as

$$\mathbf{r} = \begin{bmatrix} \mathbf{K}\mathbf{L}^\dagger \\ \sqrt{\alpha}\mathbf{I}_p \end{bmatrix} \bar{\mathbf{x}} - \begin{bmatrix} \mathbf{y}^\delta - \mathbf{K}\mathbf{x}_0^\delta \\ \mathbf{0} \end{bmatrix}.$$

Here,

$$\mathbf{x}_0^\delta = \sum_{i=p+1}^n (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{w}_i \quad (\text{B.4})$$

is the component of the solution in the null space of  $\mathbf{L}$ , that is,  $\mathbf{L}\mathbf{x}_0^\delta = \mathbf{0}$ . In this regard, the standard-form transformation takes the form

$$\begin{aligned} \bar{\mathbf{K}} &= \mathbf{K}\mathbf{L}^\dagger, \\ \bar{\mathbf{y}}^\delta &= \mathbf{y}^\delta - \mathbf{K}\mathbf{x}_0^\delta, \end{aligned}$$

and, if  $\bar{\mathbf{x}}_\alpha^\delta$  is the solution of the standard-form problem, the back-transformation is given by

$$\mathbf{x}_\alpha^\delta = \mathbf{L}^\dagger \bar{\mathbf{x}}_\alpha^\delta + \mathbf{x}_0^\delta. \quad (\text{B.5})$$

A very efficient standard-form transformation relying on two QR factorizations has been proposed by Elden (1977). First, we consider the QR factorization

$$\mathbf{L}^T = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}, \quad (\text{B.6})$$

where  $\mathbf{Q}_1 \in \mathbb{R}^{n \times p}$ ,  $\mathbf{Q}_2 \in \mathbb{R}^{n \times (n-p)}$  and  $\mathbf{R} \in \mathbb{R}^{p \times p}$ . Since  $\mathbf{Q}$  is an orthogonal matrix, the relation  $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}_n$  yields

$$\mathbf{Q}_1^T \mathbf{Q}_1 = \mathbf{I}_p, \quad \mathbf{Q}_2^T \mathbf{Q}_2 = \mathbf{I}_{n-p}, \quad \mathbf{Q}_1^T \mathbf{Q}_2 = \mathbf{0},$$

and we infer that

$$\mathbf{L}\mathbf{Q}_1 = \mathbf{R}^T, \quad \mathbf{L}\mathbf{Q}_2 = \mathbf{0}. \quad (\text{B.7})$$

Second, we make the change of variable

$$\mathbf{x} = \mathbf{Q}_1 \mathbf{x}_1 + \mathbf{Q}_2 \mathbf{x}_2, \quad \mathbf{x}_1 \in \mathbb{R}^p, \quad \mathbf{x}_2 \in \mathbb{R}^{n-p}, \quad (\text{B.8})$$

and express the augmented residual vector (B.3) as (cf. (B.7))

$$\mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{K}\mathbf{Q}_1 & \mathbf{K}\mathbf{Q}_2 \\ \sqrt{\alpha}\mathbf{R}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{y}^\delta \\ \mathbf{0} \end{bmatrix}.$$

Third, we perform the QR factorization

$$\mathbf{K}\mathbf{Q}_2 = \mathbf{S} \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix}, \quad (\text{B.9})$$

where  $\mathbf{S}_1 \in \mathbb{R}^{m \times (n-p)}$ ,  $\mathbf{S}_2 \in \mathbb{R}^{m \times (m-n+p)}$  and  $\mathbf{T} \in \mathbb{R}^{(n-p) \times (n-p)}$ . Multiplying the equation

$$\mathbf{r}_1 = \mathbf{KQ}_1 \mathbf{x}_1 + \mathbf{KQ}_2 \mathbf{x}_2 - \mathbf{y}^\delta$$

by  $\mathbf{S}^T$ , we obtain

$$\mathbf{S}^T \mathbf{r}_1 = \begin{bmatrix} \mathbf{S}_1^T \mathbf{r}_1 \\ \mathbf{S}_2^T \mathbf{r}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^T \mathbf{KQ}_1 & \mathbf{S}_1^T \mathbf{KQ}_2 \\ \mathbf{S}_2^T \mathbf{KQ}_1 & \mathbf{S}_2^T \mathbf{KQ}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{S}_1^T \mathbf{y}^\delta \\ \mathbf{S}_2^T \mathbf{y}^\delta \end{bmatrix}.$$

Further, using the relation

$$\mathbf{S}^T \mathbf{KQ}_2 = \begin{bmatrix} \mathbf{S}_1^T \mathbf{KQ}_2 \\ \mathbf{S}_2^T \mathbf{KQ}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix},$$

which yields  $\mathbf{S}_1^T \mathbf{KQ}_2 = \mathbf{T}$  and  $\mathbf{S}_2^T \mathbf{KQ}_2 = \mathbf{0}$ , and setting

$$\mathbf{S}^T \mathbf{r}_1 = \begin{bmatrix} \mathbf{r}'_1 \\ \mathbf{r}''_1 \end{bmatrix},$$

we find that

$$\begin{bmatrix} \mathbf{S}^T \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{r}'_1 \\ \mathbf{r}''_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^T \mathbf{KQ}_1 & \mathbf{T} \\ \mathbf{S}_2^T \mathbf{KQ}_1 & \mathbf{0} \\ \sqrt{\alpha} \mathbf{R}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{S}_1^T \mathbf{y}^\delta \\ \mathbf{S}_2^T \mathbf{y}^\delta \\ \mathbf{0} \end{bmatrix},$$

or explicitly, that

$$\begin{aligned} \mathbf{r}'_1(\mathbf{x}_1, \mathbf{x}_2) &= \mathbf{S}_1^T \mathbf{KQ}_1 \mathbf{x}_1 + \mathbf{T} \mathbf{x}_2 - \mathbf{S}_1^T \mathbf{y}^\delta, \\ \mathbf{r}''_1(\mathbf{x}_1) &= \mathbf{S}_2^T \mathbf{KQ}_1 \mathbf{x}_1 - \mathbf{S}_2^T \mathbf{y}^\delta, \\ \mathbf{r}_2(\mathbf{x}_1) &= \sqrt{\alpha} \mathbf{R}^T \mathbf{x}_1. \end{aligned}$$

Now, as  $\mathbf{S}$  is an orthogonal matrix, there holds

$$\|\mathbf{r}_1\|^2 = \|\mathbf{S}^T \mathbf{r}_1\|^2 = \|\mathbf{r}'_1\|^2 + \|\mathbf{r}''_1\|^2,$$

and the  $n$ -dimensional minimization problem

$$\min_{\mathbf{x}_1, \mathbf{x}_2} \|\mathbf{r}(\mathbf{x}_1, \mathbf{x}_2)\|^2 = \|\mathbf{r}'_1(\mathbf{x}_1, \mathbf{x}_2)\|^2 + \|\mathbf{r}''_1(\mathbf{x}_1)\|^2 + \|\mathbf{r}_2(\mathbf{x}_1)\|^2$$

can be split into the  $p$ -dimensional minimization problem

$$\min_{\mathbf{x}_1} \left( \|\mathbf{r}''_1(\mathbf{x}_1)\|^2 + \|\mathbf{r}_2(\mathbf{x}_1)\|^2 \right) \quad (\text{B.10})$$

for  $\mathbf{x}_1$ , and the equation

$$\mathbf{r}'_1(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{0} \quad (\text{B.11})$$

for  $\mathbf{x}_2$ . The minimization problem (B.10) can be written as

$$\min_{\mathbf{x}_1} \left\| \begin{bmatrix} \mathbf{S}_2^T \mathbf{KQ}_1 \\ \sqrt{\alpha} \mathbf{R}^T \end{bmatrix} \mathbf{x}_1 - \begin{bmatrix} \mathbf{S}_2^T \mathbf{y}^\delta \\ \mathbf{0} \end{bmatrix} \right\|^2, \quad (\text{B.12})$$

while the solution of equation (B.11) is

$$\mathbf{x}_2 = \mathbf{T}^{-1} \mathbf{S}_1^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{Q}_1 \mathbf{x}_1). \quad (\text{B.13})$$

Finally, by the transformation

$$\mathbf{x}_1 = \mathbf{R}^{-T} \bar{\mathbf{x}}, \quad \bar{\mathbf{x}} \in \mathbb{R}^p, \quad (\text{B.14})$$

the minimization problem (B.12) can be expressed in the standard form

$$\min_{\bar{\mathbf{x}}} \left\| \begin{bmatrix} \bar{\mathbf{K}} \\ \sqrt{\alpha} \mathbf{I}_p \end{bmatrix} \bar{\mathbf{x}} - \begin{bmatrix} \bar{\mathbf{y}}^\delta \\ \mathbf{0} \end{bmatrix} \right\|^2,$$

with

$$\bar{\mathbf{K}} = \mathbf{S}_2^T \mathbf{K} \mathbf{Q}_1 \mathbf{R}^{-T}$$

and

$$\bar{\mathbf{y}}^\delta = \mathbf{S}_2^T \mathbf{y}^\delta.$$

Combining (B.8), (B.13) and (B.14), we find that the back-transformation is given by

$$\mathbf{x}_\alpha^\delta = \mathbf{Q}_1 \mathbf{R}^{-T} \bar{\mathbf{x}}_\alpha^\delta + \mathbf{Q}_2 \mathbf{T}^{-1} \mathbf{S}_1^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{Q}_1 \mathbf{R}^{-T} \bar{\mathbf{x}}_\alpha^\delta). \quad (\text{B.15})$$

Taking into account that  $\mathbf{L}^\dagger = \mathbf{L}^T (\mathbf{L} \mathbf{L}^T)^{-1} = \mathbf{Q}_1 \mathbf{R}^{-T}$ , and that (cf. the relation  $\mathbf{L} \mathbf{Q}_2 = \mathbf{0}$ )

$$\mathbf{x}_0^\delta = \mathbf{Q}_2 \mathbf{T}^{-1} \mathbf{S}_1^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{Q}_1 \mathbf{R}^{-T} \bar{\mathbf{x}}_\alpha^\delta) \in \mathcal{N}(\mathbf{L}),$$

we see that (B.15) is similar to (B.5). The main steps of this explicit transformation are illustrated in Algorithm 17.

---

**Algorithm 17.** Explicit transformation for computing  $\bar{\mathbf{K}}$ ,  $\bar{\mathbf{y}}^\delta$  and  $\mathbf{x}_\alpha^\delta$ . The solution of the standard-form problem  $\bar{\mathbf{x}}_\alpha^\delta$  is an input parameter of the algorithm.

---

compute the QR factorization  $\mathbf{L}^T = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix};$

compute the QR factorization  $\mathbf{K} \mathbf{Q}_2 = \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix};$

$\mathbf{L}^\dagger \leftarrow \mathbf{Q}_1 \mathbf{R}^{-T}$ ;  $\hat{\mathbf{K}} \leftarrow \mathbf{K} \mathbf{L}^\dagger$ ;  
 {standard-form transformation}

$\bar{\mathbf{K}} \leftarrow \mathbf{S}_2^T \hat{\mathbf{K}}$ ;

$\bar{\mathbf{y}}^\delta \leftarrow \mathbf{S}_2^T \mathbf{y}^\delta$ ;

{back-transformation}

$\mathbf{x}_0^\delta \leftarrow \mathbf{Q}_2 \mathbf{T}^{-1} \mathbf{S}_1^T (\mathbf{y}^\delta - \hat{\mathbf{K}} \bar{\mathbf{x}}_\alpha^\delta)$ ;

$\mathbf{x}_\alpha^\delta \leftarrow \mathbf{L}^\dagger \bar{\mathbf{x}}_\alpha^\delta + \mathbf{x}_0^\delta$ ;

---

## B.2 Implicit transformations

For iterative methods, it is not practical to form  $\bar{\mathbf{K}}$  and  $\bar{\mathbf{K}}^T$  explicitly. Rather, we need to compute  $\mathbf{x}_0^\delta$  and the matrix-vector products  $\mathbf{L}^\dagger \mathbf{z}$  and  $\mathbf{L}^{\dagger T} \mathbf{z}$  efficiently. The implicit transformation presented in this section is due to Hanke and Hansen (1993) and can also be found in Hansen (1998).

Let  $\mathbf{W}$  be the nonsingular matrix of the GSVD (B.1). From Chapter 3 we know that

$$\mathbf{K}\mathbf{w}_i = \sigma_i \mathbf{u}_i, \quad \mathbf{L}\mathbf{w}_i = \mu_i \mathbf{v}_i, \quad i = 1, \dots, p, \quad (\text{B.16})$$

and that

$$\mathbf{K}\mathbf{w}_i = \mathbf{u}_i, \quad \mathbf{L}\mathbf{w}_i = \mathbf{0}, \quad i = p+1, \dots, n. \quad (\text{B.17})$$

Moreover, we have

$$\mathbf{w}_i^T \mathbf{K}^T \mathbf{K} \mathbf{w}_j = (\mathbf{K}\mathbf{w}_i)^T (\mathbf{K}\mathbf{w}_j) = \mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}, \quad i, j = p+1, \dots, n, \quad (\text{B.18})$$

and

$$\mathbf{w}_i^T \mathbf{K}^T \mathbf{K} \mathbf{w}_j = (\mathbf{K}\mathbf{w}_i)^T (\mathbf{K}\mathbf{w}_j) = \sigma_i \mathbf{u}_i^T \mathbf{u}_j = 0, \quad i = 1, \dots, p, \quad j = p+1, \dots, n. \quad (\text{B.19})$$

The second equation in (B.17) shows that the set  $\{\mathbf{w}_i\}_{i=p+1, \dots, n}$  is a basis of  $\mathcal{N}(\mathbf{L})$ , while relation (B.18) shows that this set is  $\mathbf{K}^T \mathbf{K}$ -orthogonal.

Let us consider the partition

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_2 \end{bmatrix},$$

with  $\mathbf{W}_1 = [\mathbf{w}_1, \dots, \mathbf{w}_p] \in \mathbb{R}^{n \times p}$  and  $\mathbf{W}_2 = [\mathbf{w}_{p+1}, \dots, \mathbf{w}_n] \in \mathbb{R}^{n \times (n-p)}$ , and let us define the matrix

$$\mathbf{P} = \mathbf{W}_2 (\mathbf{K}^T \mathbf{K} \mathbf{W}_2)^T. \quad (\text{B.20})$$

Then, using the representation

$$\mathbf{K}^T \mathbf{K} \mathbf{W}_2 = [\mathbf{K}^T \mathbf{K} \mathbf{w}_{p+1}, \dots, \mathbf{K}^T \mathbf{K} \mathbf{w}_n],$$

we find that, for any  $\mathbf{x} \in \mathbb{R}^n$ , there holds

$$\mathbf{P}\mathbf{x} = (\mathbf{x}^T \mathbf{K}^T \mathbf{K} \mathbf{w}_{p+1}) \mathbf{w}_{p+1} + \dots + (\mathbf{x}^T \mathbf{K}^T \mathbf{K} \mathbf{w}_n) \mathbf{w}_n. \quad (\text{B.21})$$

Assuming the expansion

$$\mathbf{x} = x_1 \mathbf{w}_1 + \dots + x_p \mathbf{w}_p + x_{p+1} \mathbf{w}_{p+1} + \dots + x_n \mathbf{w}_n,$$

for some  $x_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ , and using the orthogonality relations (B.18) and (B.19), (B.21) yields

$$\mathbf{P}\mathbf{x} = x_{p+1} \mathbf{w}_{p+1} + \dots + x_n \mathbf{w}_n;$$

this implies that

$$\mathbf{x} - \mathbf{P}\mathbf{x} = x_1 \mathbf{w}_1 + \dots + x_p \mathbf{w}_p. \quad (\text{B.22})$$

In view of the identities  $\mathbf{P}\mathbf{w}_i = \mathbf{0}$ ,  $i = 1, \dots, p$ , we see that  $\mathbf{P}$  can be interpreted as an (oblique) projection matrix with the range space  $\mathcal{R}(\mathbf{P}) = \text{span}\{\mathbf{w}_{p+1}, \dots, \mathbf{w}_n\}$  and the null space  $\mathcal{N}(\mathbf{P}) = \text{span}\{\mathbf{w}_1, \dots, \mathbf{w}_p\}$ .

To compute the matrix-vector product  $\mathbf{L}^\dagger \mathbf{z}$  for  $\mathbf{z} \in \mathbb{R}^p$ , we look at the equation

$$\mathbf{L}\mathbf{x} = \mathbf{z}. \quad (\text{B.23})$$

By virtue of (B.2), it is apparent that  $\mathbf{L}^\dagger \mathbf{z}$  is the unique solution of (B.23) in the subspace  $\mathcal{R}(\mathbf{L}^\dagger) = \text{span}\{\mathbf{w}_1, \dots, \mathbf{w}_p\}$ . On the other hand, from (B.17), it follows that  $\mathbf{L}\mathbf{W}_2 = \mathbf{0}$ , and further, by (B.20), that  $\mathbf{L}\mathbf{P} = \mathbf{0}$ . In this regard, if  $\hat{\mathbf{x}}$  is an arbitrary solution to (B.23), i.e.,  $\mathbf{L}\hat{\mathbf{x}} = \mathbf{z}$ , then,  $\hat{\mathbf{x}} - \mathbf{P}\hat{\mathbf{x}}$  also solves (B.23), and by (B.22), we have  $\hat{\mathbf{x}} - \mathbf{P}\hat{\mathbf{x}} \in \text{span}\{\mathbf{w}_1, \dots, \mathbf{w}_p\}$ . Thus,  $\mathbf{L}^\dagger \mathbf{z}$  can be identified with  $\hat{\mathbf{x}} - \mathbf{P}\hat{\mathbf{x}}$ , that is,

$$\mathbf{L}^\dagger \mathbf{z} = (\mathbf{I}_n - \mathbf{P}) \hat{\mathbf{x}} = \left[ \mathbf{I}_n - \mathbf{W}_2 (\mathbf{K}^T \mathbf{K} \mathbf{W}_2)^T \right] \hat{\mathbf{x}}.$$

Partitioning  $\mathbf{L}$  as

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{bmatrix}, \quad \mathbf{L}_1 \in \mathbb{R}^{p \times p}, \quad \mathbf{L}_2 \in \mathbb{R}^{p \times (n-p)},$$

and supposing that  $\mathbf{L}_1$  is nonsingular, we may choose  $\hat{\mathbf{x}}$  as

$$\hat{\mathbf{x}} = \begin{bmatrix} \mathbf{L}_1^{-1} \mathbf{z} \\ \mathbf{0} \end{bmatrix}.$$

In practice, the matrix

$$\mathbf{T} = (\mathbf{K}^T \mathbf{K} \mathbf{W}_2)^T$$

is computed at the beginning of the iterative process, and  $\mathbf{L}^\dagger \mathbf{z}$  is calculated as

$$\mathbf{L}^\dagger \mathbf{z} = (\mathbf{I}_n - \mathbf{W}_2 \mathbf{T}) \hat{\mathbf{x}}.$$

In a similar manner, it can be shown that the computation of the matrix-vector product  $\mathbf{L}^{\dagger T} \mathbf{z}$  for  $\mathbf{z} \in \mathbb{R}^n$ , involves the steps

$$\hat{\mathbf{z}} = (\mathbf{I}_n - \mathbf{T}^T \mathbf{W}_2^T) \mathbf{z}$$

and

$$\mathbf{L}^{\dagger T} \mathbf{z} = \begin{bmatrix} \mathbf{L}_1^{-T} & \mathbf{0} \end{bmatrix} \hat{\mathbf{z}}.$$

The null-space component of the solution is computed according to

$$\mathbf{x}_0^\delta = \mathbf{W}_2 (\mathbf{K} \mathbf{W}_2)^T \mathbf{y}^\delta. \quad (\text{B.24})$$

To prove this claim, we use the explicit form of (B.24),

$$\mathbf{x}_0^\delta = (\mathbf{y}^{\delta T} \mathbf{K} \mathbf{w}_{p+1}) \mathbf{w}_{p+1} + \dots + (\mathbf{y}^{\delta T} \mathbf{K} \mathbf{w}_n) \mathbf{w}_n,$$

together with (B.17) to obtain (B.4).

From the above analysis it is apparent that the computation of  $\mathbf{L}^\dagger \mathbf{z}$ ,  $\mathbf{L}^{\dagger T} \mathbf{z}$  and  $\mathbf{x}_0^\delta$  requires the knowledge of the matrix  $\mathbf{W}_2$ , whose column vectors form a  $\mathbf{K}^T \mathbf{K}$ -orthogonal



---

**Algorithm 18.** Implicit transformation for computing  $\mathbf{x}_0^\delta$ ,  $\mathbf{L}^\dagger \mathbf{z}$  and  $\mathbf{L}^{\dagger T} \mathbf{z}$ . The vector  $\mathbf{z}$  is an input parameter of the algorithm.

---

compute the QR factorization  $\mathbf{L}^T = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$ ;

compute the QR factorization  $\mathbf{KQ}_2 = \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix}$ ;

$\mathbf{W}_2 \leftarrow \mathbf{Q}_2 \mathbf{T}^{-1}$ ;

$\mathbf{T} \leftarrow (\mathbf{K}^T \mathbf{K} \mathbf{W}_2)^T$ ;

partition  $\mathbf{L} = \begin{bmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{bmatrix}$  with  $\mathbf{L}_1 \in \mathbb{R}^{p \times p}$ ;

partition  $\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{T}_2 \end{bmatrix}$  with  $\mathbf{T}_1 \in \mathbb{R}^{(n-p) \times p}$ ;

{null-space component of the solution solution}

$\mathbf{x}_0^\delta \leftarrow \mathbf{W}_2 (\mathbf{K} \mathbf{W}_2)^T \mathbf{y}^\delta$ ;

{matrix-vector product  $\mathbf{z}_L = \mathbf{L}^\dagger \mathbf{z}$ , with  $\mathbf{z} \in \mathbb{R}^p$ }

$\mathbf{z} \leftarrow \mathbf{L}_1^{-1} \mathbf{z}$ ;

$\mathbf{z}_L \leftarrow \begin{bmatrix} \mathbf{z} \\ \mathbf{0} \end{bmatrix} - \mathbf{W}_2 \mathbf{T}_1 \mathbf{z}$ ;

{matrix-vector product  $\mathbf{z}_{LT} = \mathbf{L}^{\dagger T} \mathbf{z}$ , with  $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \in \mathbb{R}^n$ ,  $\mathbf{z}_1 \in \mathbb{R}^p$ }

$\mathbf{z}_1 \leftarrow \mathbf{z}_1 - \mathbf{T}_1^T \mathbf{W}_2^T \mathbf{z}$ ;

$\mathbf{z}_{LT} \leftarrow \mathbf{L}_1^{-T} \mathbf{z}_1$ ;

---

basis for  $\mathcal{N}(\mathbf{L})$ . To compute  $\mathbf{W}_2$ , we first consider the QR factorization of  $\mathbf{L}^T$  as in (B.6). Evidently, by (B.7), we have  $\mathbf{LQ}_2 = \mathbf{0}$ . Then, we perform a QR factorization of the matrix  $\mathbf{KQ}_2$  as in (B.9), and choose

$$\mathbf{W}_2 = \mathbf{Q}_2 \mathbf{T}^{-1}.$$

To justify this choice, we use the relation  $\mathbf{KQ}_2 \mathbf{T}^{-1} = \mathbf{S}_1$  to obtain  $\mathbf{KW}_2 = \mathbf{S}_1$ ; this yields

$$(\mathbf{KW}_2)^T (\mathbf{KW}_2) = \mathbf{S}_1^T \mathbf{S}_1 = \mathbf{I}_{n-p}.$$

Thus, the column vectors of  $\mathbf{W}_2$  are  $\mathbf{K}^T \mathbf{K}$ -orthogonal, and since  $\mathbf{LW}_2 = \mathbf{LQ}_2 \mathbf{T}^{-1} = \mathbf{0}$ , they span  $\mathcal{N}(\mathbf{L})$ . The implicit transformation is outlined in Algorithm 18.

The iterative algorithms presented in Chapter 5 assume a square and nonsingular regularization matrix  $\mathbf{L}$ . If  $\mathbf{L}$  is a rectangular matrix, then these algorithms should be modified as follows: the initialization step  $\mathbf{x}^\delta = \mathbf{0}$  has to be replaced by  $\mathbf{x}^\delta = \mathbf{x}_0^\delta$ , the matrix  $\mathbf{L}^{-1}$  by  $\mathbf{L}^\dagger$ , and the matrix  $\mathbf{L}^{-T}$  by  $\mathbf{L}^{\dagger T}$ .

# C

## A general direct regularization method for linear problems

In this appendix we analyze a general regularization method for linear problems by particularizing the results established in Engl et al. (2000) and Rieder (2003) to a discrete setting. This approach includes Tikhonov regularization and its iterated version as particular regularization methods.

To simplify our analysis we assume regularization in standard form, characterized by the choice  $\mathbf{L} = \mathbf{I}_n$ . The regularization parameter choice methods to be discussed comprise an a priori selection criterion, the discrepancy principle, the generalized discrepancy principle, and two error-free parameter choice methods, namely the residual curve method and its generalized version.

### C.1 Basic assumptions

Let us assume that the regularized solution possesses the general-form representation

$$\mathbf{x}_\alpha^\delta = \sum_{i=1}^n f_\alpha(\sigma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i, \quad (\text{C.1})$$

where  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of the matrix  $\mathbf{K}$  and the filter function  $f_\alpha(\lambda)$  is positive and continuous on the interval  $[0, \sigma_1^2]$ . The regularized solution for the exact data vector  $\mathbf{y}$  and the exact solution are then given by

$$\mathbf{x}_\alpha = \sum_{i=1}^n f_\alpha(\sigma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i \quad (\text{C.2})$$

and

$$\mathbf{x}^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i, \quad (\text{C.3})$$

respectively. Defining the parameter-dependent family of auxiliary functions  $g_\alpha$  by

$$g_\alpha(\lambda) = \frac{1}{\lambda} f_\alpha(\lambda), \quad (\text{C.4})$$

and the residual functions  $r_\alpha$  by

$$r_\alpha(\lambda) = 1 - f_\alpha(\lambda) = 1 - \lambda g_\alpha(\lambda), \quad (\text{C.5})$$

we express the smoothing and noise errors as

$$\mathbf{e}_{\text{sa}} = \mathbf{x}^\dagger - \mathbf{x}_\alpha = \sum_{i=1}^n r_\alpha(\sigma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i \quad (\text{C.6})$$

and

$$\mathbf{e}_{\text{na}}^\delta = \mathbf{x}_\alpha - \mathbf{x}_\alpha^\delta = - \sum_{i=1}^n [\sigma_i^2 g_\alpha(\sigma_i^2)] \frac{1}{\sigma_i} (\mathbf{u}_i^T \boldsymbol{\delta}) \mathbf{v}_i, \quad (\text{C.7})$$

respectively, and the residual vector as

$$\mathbf{r}_\alpha^\delta = \mathbf{y}^\delta - \mathbf{K} \mathbf{x}_\alpha^\delta = \sum_{i=1}^n r_\alpha(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{u}_i. \quad (\text{C.8})$$

The following assumptions are adopted in our analysis:

$$0 \leq g_\alpha(\lambda) \leq \frac{c_1}{\alpha}, \quad (\text{C.9})$$

$$0 \leq r_\alpha(\lambda) \leq 1, \quad (\text{C.10})$$

$$0 \leq \lambda^\mu r_\alpha(\lambda) \leq c_2 \alpha^\mu, \quad 0 < \mu \leq \mu_0, \quad (\text{C.11})$$

for all  $\alpha > 0$ ,  $\lambda \in [0, \sigma_1^2]$ , and suitable positive constants  $c_1$  and  $c_2$ . The index  $\mu_0$  is the qualification of the regularization method and represents the largest value of  $\mu$  such that the inequality (C.11) holds. The function  $g_\alpha(\lambda)$  is supposed to be right continuous at  $\lambda = 0$ ; setting  $g_\alpha(0) = \lim_{\lambda \rightarrow 0} g_\alpha(\lambda)$ ,  $g_\alpha(\lambda)$  extends to a continuous function in  $[0, \sigma_1^2]$ . Furthermore, we assume the normalization condition

$$r_\alpha(0) = 1, \quad (\text{C.12})$$

and the asymptotic result

$$\lim_{\alpha \rightarrow 0} g_\alpha(\lambda) = \frac{1}{\lambda}, \quad \lambda \in (0, \sigma_1^2]. \quad (\text{C.13})$$

In this context, (C.5) and (C.10) yield

$$0 \leq \lambda g_\alpha(\lambda) \leq 1, \quad (\text{C.14})$$

while (C.5) and (C.13) give

$$\lim_{\alpha \rightarrow 0} r_\alpha(\lambda) = 0, \quad \lambda \in (0, \sigma_1^2]. \quad (\text{C.15})$$

## C.2 Source condition

Let  $\mathbf{K}$  be the discrete version of a smoothing (integral) operator  $K$ . Assuming that  $\mathbf{x}_0$  is the discrete version of a  $k$ -times differentiable function  $x_0$ , then  $\mathbf{y}_1 = \mathbf{K}\mathbf{x}_0$  is the discrete version of a  $(k+1)$ -times differentiable function  $y_1$ . Moreover, as the transpose matrix  $\mathbf{K}^T$  is the discrete version of the adjoint operator  $K^*$ , which is also a smoothing operator,  $\mathbf{x}_1 = \mathbf{K}^T \mathbf{y}_1$  is the discrete version of a  $(k+2)$ -times differentiable function  $x_1$ . Thus, we get ‘smoother and smoother’ vectors by repeating applications of  $\mathbf{K}$  and  $\mathbf{K}^T$  to some vector  $\mathbf{z} \in \mathbb{R}^n$  corresponding to a continuous function only. In this regard, the assumption that the solution  $\mathbf{x}^\dagger$  is smooth is equivalent to the validity of the so-called source condition

$$\mathbf{x}^\dagger = (\mathbf{K}^T \mathbf{K})^\mu \mathbf{z}, \quad (\text{C.16})$$

where  $\mu > 0$  and  $\mathbf{z} \in \mathbb{R}^n$ . Note that under the common assumption that  $\mathbf{K}$  is injective, we have  $\mathbb{R}^n = \mathcal{N}(\mathbf{K})^\perp$ ; if this is not the case, we take  $\mathbf{z} \in \mathcal{N}(\mathbf{K})^\perp$ . In terms of the singular system of the matrix  $\mathbf{K}$ , there holds

$$\mathbf{K}^T \mathbf{K} = \mathbf{V} \left[ \text{diag}(\sigma_i^2)_{n \times n} \right] \mathbf{V}^T,$$

and the source condition (C.16) reads as

$$\mathbf{x}^\dagger = \sum_{i=1}^n \sigma_i^{2\mu} (\mathbf{v}_i^T \mathbf{z}) \mathbf{v}_i = \sum_{i=1}^n \sigma_i^{2\mu} \zeta_i \mathbf{v}_i, \quad (\text{C.17})$$

where we have set  $\zeta_i = \mathbf{v}_i^T \mathbf{z}$ . As  $\{\mathbf{v}_i\}_{i=1,n}$  is an orthonormal basis of  $\mathbb{R}^n$ ,  $\mathbf{x}^\dagger$  can be expressed as

$$\mathbf{x}^\dagger = \sum_{i=1}^n \xi_i \mathbf{v}_i, \quad (\text{C.18})$$

with  $\xi_i = \mathbf{v}_i^T \mathbf{x}^\dagger$ . By (C.17) and (C.18), it is apparent that the source condition (C.16) is equivalent to the following assumption on the Fourier coefficients of the exact solution:

$$\xi_i = \sigma_i^{2\mu} \zeta_i, \quad i = 1, \dots, n. \quad (\text{C.19})$$

For the source condition (C.19), the Fourier coefficients of the exact data vector can be expressed as

$$\mathbf{u}_i^T \mathbf{y} = (\mathbf{K}^T \mathbf{u}_i)^T \mathbf{x}^\dagger = \sum_{j=1}^n \xi_j (\mathbf{K}^T \mathbf{u}_i)^T \mathbf{v}_j = \xi_i \sigma_i = \sigma_i^{2\mu+1} \zeta_i, \quad i = 1, \dots, n, \quad (\text{C.20})$$

and this result will be frequently used in the sequel.

Some basic definitions are now in order. The rate of convergence of a regularization parameter choice method is the rate with which  $\|\mathbf{e}_\alpha^\delta\| \rightarrow 0$  as  $\Delta \rightarrow 0$ . Since

$$\|\mathbf{e}_\alpha^\delta\| \leq \|\mathbf{e}_{\text{sa}}\| + \|\mathbf{e}_{\text{na}}^\delta\|, \quad (\text{C.21})$$

we see that the convergence rate is given by the individual convergence rates of the smoothing and noise errors. A regularization parameter choice method is called of optimal order if, for the source condition (C.19), the estimate

$$\|\mathbf{e}_\alpha^\delta\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad \Delta \rightarrow 0, \quad (\text{C.22})$$

with  $\|\mathbf{z}\|^2 = \sum_{i=1}^n \zeta_i^2$ , holds true.

### C.3 Error estimates

A bound for the noise error can be expressed in terms of the noise level  $\Delta$  and the regularization parameter  $\alpha$ .

**Proposition C.1.** *Let assumptions (C.9) and (C.10) be satisfied. Then there holds*

$$\|\mathbf{e}_{\text{n}\alpha}^\delta\| \leq c_n \frac{\Delta}{\sqrt{\alpha}}, \quad (\text{C.23})$$

with a suitable constant  $c_n > 0$ .

*Proof.* By (C.7), the norm of the noise error vector is given by

$$\|\mathbf{e}_{\text{n}\alpha}^\delta\|^2 = \sum_{i=1}^n \sigma_i^2 g_\alpha^2(\sigma_i^2) (\mathbf{u}_i^T \boldsymbol{\delta})^2.$$

Using (C.9) and (C.14), we find that

$$\sigma_i^2 g_\alpha^2(\sigma_i^2) \leq \frac{c_1}{\alpha},$$

and, because of  $\|\boldsymbol{\delta}\| \leq \Delta$ , we deduce that (C.23) holds with  $c_n = \sqrt{c_1}$ .  $\square$

For the smoothing error (C.6), we use the source condition (C.20) to derive the expansion

$$\|\mathbf{e}_{\text{s}\alpha}\|^2 = \sum_{i=1}^n r_\alpha^2(\sigma_i^2) \sigma_i^{4\mu} \zeta_i^2. \quad (\text{C.24})$$

This representation will be particularized for each regularization parameter choice method under examination.

### C.4 A priori parameter choice method

Convergence of the general regularization method can be established for an a priori parameter choice rule without any assumption on the smoothness of  $\mathbf{x}^\dagger$ .

**Proposition C.2.** *Let assumptions (C.9), (C.10) and (C.13) hold. For the a priori parameter choice rule  $\alpha = \Delta^p$  with  $0 < p < 2$ , we have  $\|\mathbf{e}_\alpha^\delta\| \rightarrow 0$  as  $\Delta \rightarrow 0$ .*

*Proof.* From (C.6) and (C.15), it is apparent that  $\|\mathbf{e}_{s\alpha}\|$  approaches 0 as  $\alpha$  approaches 0. For  $\alpha = \Delta^p$  with  $p > 0$ , we see that  $\alpha \rightarrow 0$  as  $\Delta \rightarrow 0$ , and so,  $\|\mathbf{e}_{s\alpha}\| \rightarrow 0$  as  $\Delta \rightarrow 0$ . On the other hand, the noise error estimate (C.23) yields  $\|\mathbf{e}_{n\alpha}^\delta\|^2 \leq c_n^2 \Delta^{2-p}$ , and, since  $0 < p < 2$ , we deduce that  $\|\mathbf{e}_{n\alpha}^\delta\| \rightarrow 0$  as  $\Delta \rightarrow 0$ . Thus,  $\|\mathbf{e}_\alpha^\delta\|$  approaches 0 as  $\Delta$  approaches 0.  $\square$

Turning now to the convergence rate we state the following result:

**Theorem C.3.** *Let assumptions (C.9)–(C.11) hold and let  $\mathbf{x}^\dagger$  satisfy the source condition (C.19). For the a priori parameter choice method*

$$\alpha = \left( \frac{\Delta}{\|\mathbf{z}\|} \right)^{\frac{2}{2\mu+1}}, \quad (\text{C.25})$$

we have the error estimate

$$\|\mathbf{e}_\alpha^\delta\| = O \left( \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}} \right), \quad 0 < \mu \leq \mu_0.$$

*Proof.* Assumption (C.11) yields

$$\sigma_i^{4\mu} r_\alpha^2(\sigma_i^2) \leq c_2^2 \alpha^{2\mu}, \quad 0 < \mu \leq \mu_0,$$

and the smoothing error (C.24) can be bounded as

$$\|\mathbf{e}_{s\alpha}\|^2 \leq c_2^2 \alpha^{2\mu} \|\mathbf{z}\|^2. \quad (\text{C.26})$$

By virtue of (C.23) and (C.26), the a priori parameter choice rule (C.25) gives

$$\|\mathbf{e}_{n\alpha}^\delta\|^2 \leq c_n^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}$$

and

$$\|\mathbf{e}_{s\alpha}\|^2 \leq c_2^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}},$$

respectively. Thus, (C.25) is of optimal order for  $0 < \mu \leq \mu_0$ .  $\square$

## C.5 Discrepancy principle

Let us define the matrix  $\mathbf{H}_{\text{dp}\alpha}$  by

$$\mathbf{H}_{\text{dp}\alpha} \mathbf{w} = \sum_{i=1}^m r_\alpha(\sigma_i^2) (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i, \quad \mathbf{w} \in \mathbb{R}^m. \quad (\text{C.27})$$

In the above relation, we have assumed that  $\sigma_i = 0$  for  $i = n+1, \dots, m$ , so that the normalization condition (C.12) yields  $r_\alpha(\sigma_i^2) = 1$  for  $i = n+1, \dots, m$ . The following properties of  $\mathbf{H}_{\text{dp}\alpha}$  are apparent:

(1)  $\|\mathbf{H}_{\text{dp}\alpha}\| \leq 1$ , that is, under assumption (C.10), there holds

$$\|\mathbf{H}_{\text{dp}\alpha} \mathbf{w}\|^2 = \sum_{i=1}^m r_\alpha^2 (\sigma_i^2) (\mathbf{u}_i^T \mathbf{w})^2 \leq \sum_{i=1}^m (\mathbf{u}_i^T \mathbf{w})^2 = \|\mathbf{w}\|^2 \quad (\text{C.28})$$

for all  $\mathbf{w} \in \mathbb{R}^m$ ;

(2) for the exact data vector  $\mathbf{y}$ , the orthogonality relations  $\mathbf{u}_i^T \mathbf{y} = 0$ ,  $i = n+1, \dots, m$ , together with the source condition (C.20) yield

$$\|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}\|^2 = \sum_{i=1}^n r_\alpha^2 (\sigma_i^2) (\mathbf{u}_i^T \mathbf{y})^2 = \sum_{i=1}^n r_\alpha^2 (\sigma_i^2) \sigma_i^{4\mu+2} \zeta_i^2, \quad (\text{C.29})$$

whence, using the estimate (cf. (C.11))

$$0 \leq (\sigma_i^2)^{\mu+\frac{1}{2}} r_\alpha (\sigma_i^2) \leq c_2 \alpha^{\mu+\frac{1}{2}}, \quad 0 < \mu \leq \mu_0 - \frac{1}{2},$$

we infer that

$$\|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}\|^2 \leq c_2^2 \alpha^{2\mu+1} \|\mathbf{z}\|^2, \quad 0 < \mu \leq \mu_0 - \frac{1}{2}. \quad (\text{C.30})$$

The representation

$$\|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}^\delta\|^2 = \sum_{i=1}^n r_\alpha^2 (\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2$$

shows that (cf. (C.8)),

$$\|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}^\delta\|^2 = \|\mathbf{r}_\alpha^\delta\|^2, \quad (\text{C.31})$$

and the regularization parameter defined via the discrepancy principle is the solution of the equation

$$\|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}^\delta\|^2 = \tau \Delta^2, \quad (\text{C.32})$$

with  $\tau > 1$ . Setting  $R_\delta(\alpha) = \|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}^\delta\|^2$  and using (C.15), we obtain

$$\lim_{\alpha \rightarrow 0} R_\delta(\alpha) = \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = \|P_{\mathcal{R}(\mathbf{K})^\perp} \mathbf{y}^\delta\|^2.$$

For the exact data vector  $\mathbf{y} \in \mathcal{R}(\mathbf{K})$ , we have  $P_{\mathcal{R}(\mathbf{K})^\perp} \mathbf{y} = \mathbf{0}$ , and so,

$$\|P_{\mathcal{R}(\mathbf{K})^\perp} \mathbf{y}^\delta\| = \|P_{\mathcal{R}(\mathbf{K})^\perp} (\mathbf{y}^\delta - \mathbf{y})\| \leq \|\mathbf{y}^\delta - \mathbf{y}\| \leq \Delta.$$

Thus,  $\lim_{\alpha \rightarrow 0} R_\delta(\alpha) \leq \Delta^2$ , and we infer that, for  $\tau > 1$ , there exists  $\alpha_0$  such that  $R_\delta(\alpha) < \tau \Delta^2$  for all  $0 < \alpha \leq \alpha_0$ . The above arguments allow us to introduce a practical version of the discrepancy principle as follows: if  $\{\alpha_k\}$  is a geometric sequence of regularization parameters with ratio  $q < 1$ , i.e.,  $\alpha_{k+1} = q\alpha_k$ , the regularization parameter  $\alpha_{k^*}$  of the discrepancy principle is chosen as

$$\|\mathbf{H}_{\text{dp}\alpha_{k^*}} \mathbf{y}^\delta\|^2 \leq \tau \Delta^2 < \|\mathbf{H}_{\text{dp}\alpha_k} \mathbf{y}^\delta\|^2, \quad 0 \leq k < k^*. \quad (\text{C.33})$$

**Theorem C.4.** *Let assumptions (C.9)–(C.13) hold and let  $\mathbf{x}^\dagger$  satisfy the source condition (C.19). If the regularization parameter is chosen according to the discrepancy principle (C.33) with  $\tau > 1$ , we have the error estimate*

$$\|\mathbf{e}_{\alpha_{k^*}}^\delta\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \mu_0 - \frac{1}{2}.$$

*Proof.* In the first step of our proof, we derive an estimate for the smoothing error (C.24), while in the second step, we combine this estimate with the noise error estimate (C.23) to derive a convergence rate result.

(a) Applying the Hölder inequality to the right-hand side of (C.24), that is,

$$\sum_{i=1}^n a_i b_i \leq \left(\sum_{i=1}^n a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^n b_i^q\right)^{\frac{1}{q}}, \quad \frac{1}{p} + \frac{1}{q} = 1, \quad a_i, b_i \geq 0, \quad (\text{C.34})$$

with

$$p = \frac{2\mu + 1}{2\mu}, \quad q = 2\mu + 1,$$

and

$$\begin{aligned} a_i &= [r_\alpha^2 (\sigma_i^2)]^{\frac{2\mu}{2\mu+1}} (\sigma_i^2)^{2\mu} (\zeta_i^2)^{\frac{2\mu}{2\mu+1}}, \\ b_i &= [r_\alpha^2 (\sigma_i^2)]^{\frac{1}{2\mu+1}} (\zeta_i^2)^{\frac{1}{2\mu+1}}, \end{aligned}$$

and taking into account that (cf. (C.10))

$$\sum_{i=1}^n b_i^q = \sum_{i=1}^n r_\alpha^2 (\sigma_i^2) \zeta_i^2 \leq \|\mathbf{z}\|^2,$$

we obtain

$$\|\mathbf{e}_{\text{s}\alpha}\|^2 \leq \left(\|\mathbf{z}\|^2\right)^{\frac{1}{2\mu+1}} \left[\sum_{i=1}^n r_\alpha^2 (\sigma_i^2) \sigma_i^{4\mu+2} \zeta_i^2\right]^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.35})$$

and further (cf. (C.29))

$$\|\mathbf{e}_{\text{s}\alpha}\|^2 \leq \left(\|\mathbf{z}\|^2\right)^{\frac{1}{2\mu+1}} \left(\|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}\|^2\right)^{\frac{2\mu}{2\mu+1}}. \quad (\text{C.36})$$

The smoothing error estimate (C.36) together with the result (cf. (C.28) and (C.33))

$$\|\mathbf{H}_{\text{dp}\alpha_{k^*}} \mathbf{y}\| \leq \|\mathbf{H}_{\text{dp}\alpha_{k^*}} \mathbf{y}^\delta\| + \|\mathbf{H}_{\text{dp}\alpha_{k^*}} \boldsymbol{\delta}\| \leq (1 + \sqrt{\tau}) \Delta$$

yields

$$\|\mathbf{e}_{\alpha_{k^*}}\|^2 \leq c_{\text{sdp}}^2 \left(\|\mathbf{z}\|^2\right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.37})$$

with

$$c_{\text{sdp}} = (1 + \sqrt{\tau})^{\frac{2\mu}{2\mu+1}}.$$



(b) To estimate the noise error, we first look for a lower bound for  $\alpha_{k^*}$ . From (C.33) and the boundedness of  $\|\mathbf{H}_{\text{dp}\alpha}\|$ , we deduce that, for  $k = 0, \dots, k^* - 1$ ,

$$\sqrt{\tau}\Delta < \|\mathbf{H}_{\text{dp}\alpha_k}\mathbf{y}^\delta\| \leq \|\mathbf{H}_{\text{dp}\alpha_k}\mathbf{y}\| + \Delta,$$

and therefore

$$\|\mathbf{H}_{\text{dp}\alpha_k}\mathbf{y}\| > (\sqrt{\tau} - 1)\Delta, \quad \tau > 1. \quad (\text{C.38})$$

On the other hand, from (C.30), there holds

$$\|\mathbf{H}_{\text{dp}\alpha_{k^*-1}}\mathbf{y}\| \leq c_2 \alpha_{k^*-1}^{\mu+\frac{1}{2}} \|\mathbf{z}\| = c_2 \left(\frac{\alpha_{k^*}}{q}\right)^{\mu+\frac{1}{2}} \|\mathbf{z}\|, \quad 0 < \mu \leq \mu_0 - \frac{1}{2},$$

and we obtain the bound

$$\alpha_{k^*} > q \left(\frac{\sqrt{\tau} - 1}{c_2}\right)^{\frac{2}{2\mu+1}} \left(\frac{\Delta}{\|\mathbf{z}\|}\right)^{\frac{2}{2\mu+1}}. \quad (\text{C.39})$$

Hence, the noise error estimate (C.23) gives

$$\|\mathbf{e}_{\text{n}\alpha_{k^*}}^\delta\|^2 < c_{\text{ndp}}^2 \left(\|\mathbf{z}\|^2\right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.40})$$

with

$$c_{\text{ndp}} = \frac{c_{\text{n}}}{\sqrt{q}} \left(\frac{c_2}{\sqrt{\tau} - 1}\right)^{\frac{1}{2\mu+1}}.$$

By (C.37) and (C.40), it is readily seen that the convergence rate is optimal for  $0 < \mu \leq \mu_0 - 1/2$ .  $\square$

## C.6 Generalized discrepancy principle

The analysis of the generalized discrepancy principle in a general setting requires an appropriate formulation of this selection criterion. For this purpose, we introduce a parameter-dependent family of positive, continuous functions  $s_\alpha$ , satisfying

$$c_{1s} \left(\frac{\alpha}{\alpha + \lambda}\right)^{2\mu_0+1} \leq s_\alpha(\lambda) \leq c_{2s} \left(\frac{\alpha}{\alpha + \lambda}\right)^{2\mu_0+1} \quad (\text{C.41})$$

for  $\alpha > 0$ ,  $\lambda \in [0, \sigma_1^2]$  and  $c_{1s}, c_{2s} > 0$ , and assume the normalization condition

$$s_\alpha(0) = 1. \quad (\text{C.42})$$

Next, we define the matrix  $\mathbf{H}_{\text{gdp}\alpha}$  through the relation

$$\mathbf{H}_{\text{gdp}\alpha} \mathbf{w} = \sum_{i=1}^m s_\alpha^{\frac{1}{2}}(\sigma_i^2) (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i, \quad \mathbf{w} \in \mathbb{R}^m.$$

As before, the convention  $\sigma_i = 0$  for  $i = n + 1, \dots, m$ , together with the normalization condition (C.42) gives  $s_\alpha(\sigma_i^2) = 1$  for  $i = n + 1, \dots, m$ . In this context, the regularization parameter defined via the generalized discrepancy principle is the solution of the equation

$$\|\mathbf{H}_{\text{gdp}\alpha} \mathbf{y}^\delta\|^2 = \tau \Delta^2,$$

with  $\tau$  sufficiently large and

$$\|\mathbf{H}_{\text{gdp}\alpha} \mathbf{y}^\delta\|^2 = \sum_{i=1}^n s_\alpha(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2. \quad (\text{C.43})$$

The following properties of  $\mathbf{H}_{\text{gdp}\alpha}$  can be evidenced:

- (1)  $\|\mathbf{H}_{\text{gdp}\alpha}\| \leq \sqrt{c_{2s}}$ , that is, under assumption (C.41), there holds

$$\|\mathbf{H}_{\text{gdp}\alpha} \mathbf{w}\|^2 = \sum_{i=1}^m s_\alpha(\sigma_i^2) (\mathbf{u}_i^T \mathbf{w})^2 \leq c_{2s} \sum_{i=1}^m (\mathbf{u}_i^T \mathbf{w})^2 = c_{2s} \|\mathbf{w}\|^2$$

for all  $\mathbf{w} \in \mathbb{R}^m$ ;

- (2) for the exact data vector  $\mathbf{y}$ , the source condition (C.20) implies that

$$\|\mathbf{H}_{\text{gdp}\alpha} \mathbf{y}\|^2 = \sum_{i=1}^n s_\alpha(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y})^2 = \sum_{i=1}^n s_\alpha(\sigma_i^2) \sigma_i^{4\mu+2} \zeta_i^2; \quad (\text{C.44})$$

whence, taking into account that, for  $0 < \mu \leq \mu_0$ ,

$$\begin{aligned} \lambda^{2\mu+1} s_\alpha(\lambda) &\leq c_{2s} \lambda^{2\mu+1} \left( \frac{\alpha}{\alpha + \lambda} \right)^{2\mu_0+1} \\ &\leq c_{2s} \lambda^{2\mu+1} \left( \frac{\alpha}{\alpha + \lambda} \right)^{2\mu+1} \\ &= c_{2s} \alpha^{2\mu+1} \left( \frac{\lambda}{\alpha + \lambda} \right)^{2\mu+1} \\ &\leq c_{2s} \alpha^{2\mu+1}, \end{aligned}$$

we obtain

$$\|\mathbf{H}_{\text{gdp}\alpha} \mathbf{y}\|^2 \leq c_{2s} \alpha^{2\mu+1} \|\mathbf{z}\|^2, \quad 0 < \mu \leq \mu_0. \quad (\text{C.45})$$

A bound for the smoothing error (C.24) can be derived in terms of the functions  $s_\alpha$ . To do this, we first observe that, for  $\alpha \leq \lambda$ , assumption (C.11) yields

$$r_\alpha(\lambda) \leq c_2 \left( \frac{\alpha}{\lambda} \right)^{\mu_0} \leq c_2 \left( \frac{2\alpha}{\alpha + \lambda} \right)^{\mu_0},$$

while, for  $\alpha > \lambda$ , assumption (C.10) gives

$$r_\alpha(\lambda) \leq 1 < \left( \frac{2\alpha}{\alpha + \lambda} \right)^{\mu_0}.$$

Consequently,  $r_\alpha$  can be bounded as

$$r_\alpha(\lambda) \leq 2^{\mu_0} \max(1, c_2) \left( \frac{\alpha}{\alpha + \lambda} \right)^{\mu_0}.$$

Then, because of (cf. (C.41)),

$$\left( \frac{\alpha}{\alpha + \lambda} \right)^{\mu_0} \leq \left[ \frac{s_\alpha(\lambda)}{c_{1s}} \right]^{\frac{\mu_0}{2\mu_0+1}},$$

we find that

$$0 \leq r_\alpha(\lambda) \leq c_r [s_\alpha(\lambda)]^{\frac{\mu_0}{2\mu_0+1}}, \quad (C.46)$$

with

$$c_r = 2^{\mu_0} c_{1s}^{-\frac{\mu_0}{2\mu_0+1}} \max(1, c_2).$$

Thus, by (C.46), the smoothing error (C.24) can be estimated as

$$\|\mathbf{e}_{s\alpha}\|^2 \leq c_r^2 \sum_{i=1}^n [s_\alpha(\sigma_i^2)]^{\frac{2\mu_0}{2\mu_0+1}} \sigma_i^{4\mu} \zeta_i^2. \quad (C.47)$$

As for the discrepancy principle, we use the following practical selection criterion: if  $\{\alpha_k\}$  is a geometric sequence of regularization parameters with ratio  $q < 1$ , the regularization parameter  $\alpha_{k^*}$  of the generalized discrepancy principle is chosen as

$$\|\mathbf{H}_{\text{gdp}\alpha_{k^*}} \mathbf{y}^\delta\|^2 \leq \tau \Delta^2 < \|\mathbf{H}_{\text{gdp}\alpha_k} \mathbf{y}^\delta\|^2, \quad 0 \leq k < k^*. \quad (C.48)$$

**Theorem C.5.** *Let the assumptions of Theorem C.4 hold. If the regularization parameter is chosen according to the generalized discrepancy principle (C.48) with  $\tau > c_{2s}$ , we have the error estimate*

$$\|\mathbf{e}_{\alpha_{k^*}}^\delta\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \mu_0.$$

*Proof.* We estimate the smoothing error bound (C.47) by using the Hölder inequality (C.34), with

$$p = \frac{2\mu+1}{2\mu}, \quad q = 2\mu+1,$$

and

$$\begin{aligned} a_i &= [s_\alpha(\sigma_i^2)]^{\frac{2\mu}{2\mu+1}} (\sigma_i^2)^{2\mu} (\zeta_i^2)^{\frac{2\mu}{2\mu+1}}, \\ b_i &= [s_\alpha(\sigma_i^2)]^{\frac{2\mu_0}{2\mu_0+1} - \frac{2\mu}{2\mu+1}} (\zeta_i^2)^{\frac{1}{2\mu+1}}. \end{aligned}$$

Since

$$\sum_{i=1}^n b_i^q = \sum_{i=1}^n [s_\alpha(\sigma_i^2)]^{\frac{2(\mu_0-\mu)}{2\mu_0+1}} \zeta_i^2,$$

we use the result (cf. (C.41))

$$[s_\alpha(\sigma_i^2)]^{\frac{2(\mu_0-\mu)}{2\mu_0+1}} \leq c_{2s}^{\frac{2(\mu_0-\mu)}{2\mu_0+1}} \left( \frac{\alpha}{\alpha + \sigma_i^2} \right)^{2(\mu_0-\mu)} \leq c_{2s}^{\frac{2(\mu_0-\mu)}{2\mu_0+1}}, \quad 0 < \mu \leq \mu_0,$$

to obtain

$$\|\mathbf{e}_{\text{sa}}\|^2 \leq c_{\text{r}}^2 c_{2\text{s}}^{\frac{2(\mu_0-\mu)}{(2\mu_0+1)(2\mu+1)}} \left(\|\mathbf{z}\|^2\right)^{\frac{1}{2\mu+1}} \left[ \sum_{i=1}^n s_{\alpha}(\sigma_i^2) \sigma_i^{4\mu+2} \zeta_i^2 \right]^{\frac{2\mu}{2\mu+1}}. \quad (\text{C.49})$$

Further, by (C.44), we see that

$$\|\mathbf{e}_{\text{sa}}\|^2 \leq c_{\text{r}}^2 c_{2\text{s}}^{\frac{2(\mu_0-\mu)}{(2\mu_0+1)(2\mu+1)}} \left(\|\mathbf{z}\|^2\right)^{\frac{1}{2\mu+1}} \left(\|\mathbf{H}_{\text{gdp}\alpha} \mathbf{y}\|^2\right)^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.50})$$

and employing the same arguments as in the derivation of (C.37), we find that

$$\|\mathbf{e}_{\text{sa}\alpha_k}\|^2 \leq c_{\text{sgdp}}^2 \left(\|\mathbf{z}\|^2\right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.51})$$

with

$$c_{\text{sgdp}} = c_{\text{r}} c_{2\text{s}}^{\frac{\mu_0-\mu}{(2\mu_0+1)(2\mu+1)}} \left(\sqrt{c_{2\text{s}}} + \sqrt{\tau}\right)^{\frac{2\mu}{2\mu+1}}.$$

Taking into account the similarity between (C.37) and (C.51), and (C.30) and (C.45), and moreover, using the boundedness of  $\|\mathbf{H}_{\text{gdp}\alpha}\|$  and the assumption  $\tau > c_{2\text{s}}$ , we conclude that the generalized discrepancy principle is of optimal order for  $0 < \mu \leq \mu_0$ .  $\square$

## C.7 Error-free parameter choice methods

The following discrete version of the residual curve method is considered in the present analysis: if  $\{\alpha_k\}$  is a geometric sequence of regularization parameters with ratio  $q < 1$ , the regularization parameter  $\alpha_{\bar{k}}$  of the residual curve method is computed as

$$\alpha_{\bar{k}} = \arg \min_k \Psi_{\text{rc}}^{\delta}(\alpha_k), \quad (\text{C.52})$$

where  $\Psi_{\text{rc}}^{\delta}$  is the error indicator function

$$\Psi_{\text{rc}}^{\delta}(\alpha) = \frac{1}{\alpha} \|\mathbf{r}_{\alpha}^{\delta}\|^2 = \frac{1}{\alpha} \|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}^{\delta}\|^2. \quad (\text{C.53})$$

To simplify our analysis we assume that  $\Psi_{\text{rc}}^{\delta}$  has a unique minimizer  $\alpha_{\bar{k}} > 0$  and that  $\|\mathbf{r}_{\alpha_{\bar{k}}}^{\delta}\| \neq 0$ .

**Theorem C.6.** *Let the assumptions of Theorem C.4 hold. If the regularization parameter  $\alpha_{\bar{k}}$  is chosen according to the parameter choice rule (C.52) and the residual  $\|\mathbf{r}_{\alpha_{\bar{k}}}^{\delta}\|$  is of the order of the noise level  $\Delta$ , we have the error estimate*

$$\|\mathbf{e}_{\alpha_{\bar{k}}}^{\delta}\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \mu_0 - \frac{1}{2}.$$

*Proof.* The identity  $\|\mathbf{H}_{\text{dp}\alpha} \mathbf{y}^{\delta}\| = \|\mathbf{r}_{\alpha}^{\delta}\|$  together with the boundedness of  $\|\mathbf{H}_{\text{dp}\alpha}\|$  yields

$$\|\mathbf{H}_{\text{dp}\alpha_{\bar{k}}} \mathbf{y}\| = \|\mathbf{H}_{\text{dp}\alpha_{\bar{k}}} (\mathbf{y}^{\delta} - \boldsymbol{\delta})\| \leq \|\mathbf{r}_{\alpha_{\bar{k}}}^{\delta}\| + \Delta \leq 2 \max\left(\|\mathbf{r}_{\alpha_{\bar{k}}}^{\delta}\|, \Delta\right),$$

and (C.36) gives

$$\|\mathbf{e}_{\text{src}}\|^2 \leq c_{\text{src}}^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} \left[ \max \left( \|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|^2, \Delta^2 \right) \right]^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.54})$$

with

$$c_{\text{src}} = 2^{\frac{2\mu}{2\mu+1}}.$$

By (C.53), the noise error estimate (C.23) can be written as

$$\|\mathbf{e}_{\text{n}\alpha_{\bar{k}}}^\delta\|^2 \leq c_{\text{n}}^2 \frac{\Delta^2}{\alpha_{\bar{k}}} = c_{\text{n}}^2 \frac{\Delta^2}{\|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|^2} \Psi_{\text{rc}}^\delta(\alpha_{\bar{k}}). \quad (\text{C.55})$$

As  $\alpha_{\bar{k}}$  is the minimizer of  $\Psi_{\text{rc}}^\delta$ , we deduce that  $\Psi_{\text{rc}}^\delta(\alpha_{\bar{k}}) \leq \Psi_{\text{rc}}^\delta(\alpha_k)$  for all  $k$ . From the set  $\{\alpha_k\}$  we consider the regularization parameter  $\alpha_{k^*}$  chosen according to the discrepancy principle,

$$\|\mathbf{r}_{\alpha_{k^*}}^\delta\|^2 \leq \tau \Delta^2 < \|\mathbf{r}_{\alpha_k}^\delta\|^2, \quad 0 \leq k < k^*,$$

with  $\tau > 1$ . Then, we have

$$\Psi_{\text{rc}}^\delta(\alpha_{\bar{k}}) \leq \Psi_{\text{rc}}^\delta(\alpha_{k^*}) = \frac{1}{\alpha_{k^*}} \|\mathbf{r}_{\alpha_{k^*}}^\delta\|^2 \leq \tau \frac{\Delta^2}{\alpha_{k^*}},$$

and, by (C.39), which is valid for  $0 < \mu \leq \mu_0 - 1/2$ , we obtain

$$\begin{aligned} \Psi_{\text{rc}}^\delta(\alpha_{\bar{k}}) &< c_{\Psi}^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}} \\ &\leq c_{\Psi}^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} \left[ \max \left( \|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|^2, \Delta^2 \right) \right]^{\frac{2\mu}{2\mu+1}}, \end{aligned} \quad (\text{C.56})$$

with

$$c_{\Psi} = \sqrt{\frac{\tau}{q}} \left( \frac{c_2}{\sqrt{\tau} - 1} \right)^{\frac{1}{2\mu+1}}.$$

Consequently, (C.55) and (C.56) yield

$$\|\mathbf{e}_{\text{n}\alpha_{\bar{k}}}^\delta\|^2 < c_{\text{n}}^2 c_{\Psi}^2 \frac{\Delta^2}{\|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|^2} \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} \left[ \max \left( \|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|^2, \Delta^2 \right) \right]^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.57})$$

whence, by (C.54) and (C.57), we infer that

$$\|\mathbf{e}_{\alpha_{\bar{k}}}^\delta\| < C_{\text{rc}} \left( 1 + \frac{\Delta}{\|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|} \right) \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \left[ \max \left( \|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|, \Delta \right) \right]^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.58})$$

with  $C_{\text{rc}} = \max(c_{\text{n}} c_{\Psi}, c_{\text{src}})$ . The error bound (C.58) shows that the regularization parameter choice method (C.52) is of optimal order for  $0 < \mu \leq \mu_0 - 1/2$ , provided that  $\|\mathbf{r}_{\alpha_{\bar{k}}}^\delta\|$  has the order of  $\Delta$ .  $\square$

To understand the significance of the error estimate (C.58), we assume that

$$C_1 \Delta^{1+\beta} \leq \left\| \mathbf{r}_{\alpha_{\bar{k}}}^\delta \right\| \leq C_2 \Delta^{1+\beta}, \quad \beta \geq 0, \quad 0 < C_1 < C_2,$$

whenever  $\Delta \rightarrow 0$ . For  $\Delta$  sufficiently small, there holds

$$\left\| \mathbf{e}_{\alpha_{\bar{k}}}^\delta \right\| < \frac{2C_{\text{rc}}}{C_1} \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}-\beta}, \quad (\text{C.59})$$

and three situations can be distinguished:

- (1) if  $\beta = 0$ , the convergence rate is optimal;
- (2) if  $\beta < 2\mu/(2\mu + 1)$ , the convergence rate is suboptimal;
- (3) if  $\beta \geq 2\mu/(2\mu + 1)$ , the bound in (C.59) does not converge to zero, and as a result,  $\mathbf{x}_{\alpha_{\bar{k}}}^\delta$  may diverge.

Therefore, if  $\left\| \mathbf{r}_{\alpha_{\bar{k}}}^\delta \right\|$  is much smaller than  $\Delta$ , the regularized solution  $\mathbf{x}_{\alpha_{\bar{k}}}^\delta$  should be disregarded.

A similar error-free parameter choice method can be defined by considering the error indicator function

$$\Psi_{\text{grc}}^\delta(\alpha) = \frac{1}{\alpha} \left\| \mathbf{H}_{\text{gdp}\alpha} \mathbf{y}^\delta \right\|^2,$$

and by selecting the regularization parameter as the minimizer of  $\Psi_{\text{grc}}^\delta$ . The analysis is analog to the treatment of the previous selection criterion; we obtain

$$\left\| \mathbf{e}_{\alpha_{\bar{k}}}^\delta \right\| < C_{\text{grc}} \left( 1 + \frac{\Delta}{\left\| \mathbf{H}_{\text{gdp}\alpha_{\bar{k}}} \mathbf{y}^\delta \right\|} \right) \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \left[ \max \left( \left\| \mathbf{H}_{\text{gdp}\alpha_{\bar{k}}} \mathbf{y}^\delta \right\|, \Delta \right) \right]^{\frac{2\mu}{2\mu+1}}, \quad (\text{C.60})$$

and this regularization parameter choice method is of optimal order for  $0 < \mu \leq \mu_0$ , provided that  $\left\| \mathbf{H}_{\text{gdp}\alpha_{\bar{k}}} \mathbf{y}^\delta \right\|$  has the order of  $\Delta$ .

We conclude our analysis by verifying the assumptions of the general regularization method for Tikhonov regularization and its iterated version.

In the case of Tikhonov regularization, we have

$$f_\alpha(\lambda) = \frac{\lambda}{\lambda + \alpha}, \quad g_\alpha(\lambda) = \frac{1}{\lambda + \alpha}, \quad r_\alpha(\lambda) = \frac{\alpha}{\lambda + \alpha}.$$

It is readily seen that assumption (C.9) is satisfied with  $c_1 = 1$  and that assumptions (C.10), (C.12) and (C.13) are also fulfilled. In order to determine the qualification of Tikhonov regularization, we have to estimate the function

$$h_\mu(\lambda) = \lambda^\mu \frac{\alpha}{\lambda + \alpha}.$$

For  $\mu < 1$ , the function attains its maximum at

$$\lambda = \frac{\alpha\mu}{1 - \mu},$$

and there holds

$$h_\mu(\lambda) \leq \mu^\mu (1 - \mu)^{1-\mu} \alpha^\mu.$$

For  $\mu \geq 1$ , the function is strictly increasing and attains its largest value in the interval  $[0, \sigma_1^2]$  at  $\lambda = \sigma_1^2$ . In this case,

$$h_\mu(\lambda) \leq \sigma_1^{2\mu} \frac{\alpha}{\sigma_1^2 + \alpha} < \sigma_1^{2(\mu-1)} \alpha,$$

and we obtain

$$0 \leq \lambda^\mu r_\alpha(\lambda) \leq \begin{cases} c_2 \alpha^\mu, & \mu < 1, \\ c'_2 \alpha, & \mu \geq 1, \end{cases}$$

with

$$c_2 = \mu^\mu (1 - \mu)^{1-\mu},$$

and  $c'_2 = \sigma_1^{2(\mu-1)}$ . Thus, assumption (C.11) holds for  $\mu \in (0, 1]$  and the qualification of Tikhonov regularization is  $\mu_0 = 1$ . The parameter-dependent family of functions  $s_\alpha$ , appearing in the framework of the generalized discrepancy principle, is chosen as

$$s_\alpha(\lambda) = \left( \frac{\alpha}{\alpha + \lambda} \right)^3,$$

in which case (see Chapter 3),

$$\|\mathbf{H}_{\text{gdp}\alpha} \mathbf{y}^\delta\|^2 = \sum_{i=1}^m \left( \frac{\alpha}{\sigma_i^2 + \alpha} \right)^3 (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = \|\mathbf{r}_\alpha^\delta\|^2 - \mathbf{r}_\alpha^{\delta T} \hat{\mathbf{A}}_\alpha \mathbf{r}_\alpha^\delta.$$

The  $p$ -times iterated Tikhonov regularization is characterized by

$$f_\alpha(\lambda) = 1 - \left( \frac{\alpha}{\lambda + \alpha} \right)^p, \quad g_\alpha(\lambda) = \frac{1}{\lambda} \left[ 1 - \left( \frac{\alpha}{\lambda + \alpha} \right)^p \right], \quad r_\alpha(\lambda) = \left( \frac{\alpha}{\lambda + \alpha} \right)^p.$$

To check assumption (C.9), we use the inequality

$$1 - \left( \frac{1}{x+1} \right)^p \leq px, \quad x \geq 0,$$

and find that

$$g_\alpha(\lambda) = \frac{1}{\lambda} \left[ 1 - \left( \frac{\alpha}{\lambda + \alpha} \right)^p \right] \leq \frac{p}{\alpha}.$$

Hence, (C.9) is satisfied with  $c_1 = p$ , and it is apparent that assumptions (C.10), (C.12) and (C.13) are also fulfilled. To determine the qualification of the method, we consider the function

$$h_\mu(\lambda) = \lambda^\mu \left( \frac{\alpha}{\lambda + \alpha} \right)^p.$$

As in the case of the ordinary Tikhonov regularization, for  $\mu < p$ , the function attains its maximum at

$$\lambda = \frac{\alpha\mu}{p - \mu},$$

and we have

$$h_\mu(\lambda) \leq \left( \frac{\mu}{p} \right)^\mu \left( 1 - \frac{\mu}{p} \right)^{p-\mu} \alpha^\mu,$$

while, for  $\mu \geq p$ , the function is strictly increasing and we have

$$h_{\mu}(\lambda) \leq \sigma_1^{2\mu} \left( \frac{\alpha}{\sigma_1^2 + \alpha} \right)^p < \sigma_1^{2(\mu-p)} \alpha^p.$$

We obtain

$$0 \leq \lambda^{\mu} r_{\alpha}(\lambda) \leq \begin{cases} c_2 \alpha^{\mu}, & \mu < p, \\ c_2' \alpha^p, & \mu \geq p, \end{cases}$$

with

$$c_2 = \left( \frac{\mu}{p} \right)^{\mu} \left( 1 - \frac{\mu}{p} \right)^{p-\mu}$$

and  $c_2' = \sigma_1^{2(\mu-p)}$ . Thus, assumption (C.11) holds for  $\mu \in (0, p]$  and the qualification of the  $p$ -times iterated Tikhonov regularization is  $\mu_0 = p$ .



# D

## Chi-square distribution

The random variable  $X$  is Chi-square distributed with  $m$  degrees of freedom and we write  $X \sim \chi^2(m)$  if its probability density is given by (Tarantola, 2005)

$$p_m(x) = \frac{1}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} x^{\frac{m}{2}-1} e^{-\frac{x}{2}}.$$

Here,  $\Gamma$  is the Gamma function having closed-form values at the half-integers. Sometimes the random variable  $X$  is denoted by  $\chi^2$ , but this notation may lead to ambiguity. The mean of the distribution is equal to the number of degrees of freedom and the variance is equal to two times the number of degrees of freedom. For large values of  $m$ , the Chi-square probability density can be roughly approximated near its maximum by a Gaussian density with mean  $m$  and standard deviation  $\sqrt{2m}$ .

The next theorem, also known as the Fisher–Cochran theorem, states under which conditions quadratic forms for normal variables are Chi-square distributed.

**Theorem D.1.** *Let  $\mathbf{X}$  be an  $n$ -dimensional Gaussian random vector with zero mean and unit covariance, i.e.,  $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ , and let  $\mathbf{P}$  be an  $n \times n$  matrix. A necessary and sufficient condition that the random variable  $X = \mathbf{X}^T \mathbf{P} \mathbf{X}$  has a Chi-square distribution is that  $\mathbf{P}$  is idempotent, that is,  $\mathbf{P}^2 = \mathbf{P}$ . In this case we have  $X \sim \chi^2(n)$  with  $n = \text{trace}(\mathbf{P}) = \text{rank}(\mathbf{P})$ .*

A direct consequence of this theorem is the following result:

**Proposition D.2.** *Let  $\mathbf{X}$  be an  $n$ -dimensional Gaussian random vector with zero mean and covariance  $\mathbf{C}$ . Then, the random variable  $X = \mathbf{X}^T \mathbf{C}^{-1} \mathbf{X}$  is Chi-square distributed with  $n$  degrees of freedom.*

*Proof.* Making the change of variable  $\mathbf{Z} = \mathbf{C}^{-1/2} \mathbf{X}$ , we express  $X$  as  $X = \mathbf{Z}^T \mathbf{Z}$ . From  $\mathcal{E}\{\mathbf{Z} \mathbf{Z}^T\} = \mathbf{I}_n$  we obtain  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ , and we conclude that  $X \sim \chi^2(n)$ .  $\square$

The Fisher–Cochran theorem is a basic tool for analyzing the statistics of regularized and unregularized least squares problems. First, we prove that the a posteriori potential from statistical inversion theory (or the Tikhonov function from classical regularization theory) is Chi-square distributed.

**Theorem D.3.** *Let*

$$\mathbf{Y}^\delta = \mathbf{K}\mathbf{X} + \Delta$$

*be a stochastic data model with  $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x)$  and  $\Delta \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\delta)$ , and let  $\hat{\mathbf{X}} = \hat{\mathbf{G}}\mathbf{Y}^\delta$  be the maximum a posteriori estimator of  $\mathbf{X}$  with*

$$\hat{\mathbf{G}} = (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-1} = \mathbf{C}_x \mathbf{K}^T (\mathbf{C}_\delta + \mathbf{K} \mathbf{C}_x \mathbf{K}^T)^{-1}.$$

*Then, the random variable*

$$\hat{V} = (\mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}})^T \mathbf{C}_\delta^{-1} (\mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}}) + \hat{\mathbf{X}}^T \mathbf{C}_x^{-1} \hat{\mathbf{X}}$$

*is Chi-square distributed with  $m$  degrees of freedom.*

*Proof.* The identity

$$\hat{\mathbf{G}}^T (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1}) = \mathbf{C}_\delta^{-1} \mathbf{K}$$

yields

$$\hat{\mathbf{G}}^T \mathbf{C}_x^{-1} = \mathbf{C}_\delta^{-1} \mathbf{K} - \hat{\mathbf{G}}^T \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K},$$

and further,

$$\hat{\mathbf{G}}^T \mathbf{C}_x^{-1} \hat{\mathbf{G}} = \mathbf{C}_\delta^{-1} \mathbf{K} \hat{\mathbf{G}} - \hat{\mathbf{G}}^T \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} \hat{\mathbf{G}} = \mathbf{C}_\delta^{-1} \hat{\mathbf{A}} - \hat{\mathbf{A}}^T \mathbf{C}_\delta^{-1} \hat{\mathbf{A}}, \quad (\text{D.1})$$

with  $\hat{\mathbf{A}} = \mathbf{K} \hat{\mathbf{G}}$  being the influence matrix. Using (D.1) and the representation

$$\mathbf{Y}^\delta - \mathbf{K}\hat{\mathbf{X}} = (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{Y}^\delta, \quad (\text{D.2})$$

we obtain

$$\begin{aligned} \hat{V} &= \mathbf{Y}^{\delta T} (\mathbf{I}_m - \hat{\mathbf{A}})^T \mathbf{C}_\delta^{-1} (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{Y}^\delta + \mathbf{Y}^{\delta T} (\mathbf{C}_\delta^{-1} \hat{\mathbf{A}} - \hat{\mathbf{A}}^T \mathbf{C}_\delta^{-1} \hat{\mathbf{A}}) \mathbf{Y}^\delta \\ &= \mathbf{Y}^{\delta T} (\mathbf{C}_\delta^{-1} - \hat{\mathbf{A}}^T \mathbf{C}_\delta^{-1}) \mathbf{Y}^\delta. \end{aligned}$$

In terms of the symmetric influence matrix  $\hat{\mathbf{A}}_\delta$ , defined through the relation

$$\hat{\mathbf{A}}_\delta = \mathbf{C}_\delta^{-\frac{1}{2}} \hat{\mathbf{A}} \mathbf{C}_\delta^{\frac{1}{2}} = \mathbf{C}_\delta^{-\frac{1}{2}} \mathbf{K} (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1} \mathbf{K}^T \mathbf{C}_\delta^{-\frac{1}{2}},$$

$\hat{V}$  can be expressed as

$$\hat{V} = \mathbf{Y}^{\delta T} \mathbf{C}_\delta^{-\frac{1}{2}} (\mathbf{I}_m - \hat{\mathbf{A}}_\delta) \mathbf{C}_\delta^{-\frac{1}{2}} \mathbf{Y}^\delta.$$

Then, setting

$$\mathbf{W}^\delta = (\mathbf{I}_m - \hat{\mathbf{A}}_\delta)^{\frac{1}{2}} \mathbf{C}_\delta^{-\frac{1}{2}} \mathbf{Y}^\delta,$$

$\hat{V}$  takes the form

$$\hat{V} = \mathbf{W}^{\delta T} \mathbf{W}^\delta.$$

Assuming that the covariance matrix of the true state is adequately described by the a priori covariance matrix, we have (cf. (4.24))

$$\mathcal{E}\{\mathbf{Y}^\delta\} = \mathbf{0}, \quad \mathcal{E}\{\mathbf{Y}^\delta \mathbf{Y}^{\delta T}\} = \mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta; \quad (\text{D.3})$$

this result together with the identity

$$\mathbf{I}_m - \widehat{\mathbf{A}}_\delta = \mathbf{C}_\delta^{\frac{1}{2}} (\mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta)^{-1} \mathbf{C}_\delta^{\frac{1}{2}}$$

gives  $\mathcal{E}\{\mathbf{W}^\delta\} = \mathbf{0}$  and

$$\mathcal{E}\{\mathbf{W}^\delta \mathbf{W}^{\delta T}\} = \left(\mathbf{I}_m - \widehat{\mathbf{A}}_\delta\right)^{\frac{1}{2}} \mathbf{C}_\delta^{-\frac{1}{2}} (\mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta) \mathbf{C}_\delta^{-\frac{1}{2}} \left(\mathbf{I}_m - \widehat{\mathbf{A}}_\delta\right)^{\frac{1}{2}} = \mathbf{I}_m.$$

Thus,  $\mathbf{W}^\delta \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ , and by Theorem D.1 the conclusion readily follows.  $\square$

Each term appearing in the expression of the a posteriori potential has a special characterization as stated by the following theorem:

**Theorem D.4.** *Let the assumptions of Theorem D.3 hold. Then, the random variable*

$$\widehat{R} = \left(\mathbf{Y}^\delta - \mathbf{K} \widehat{\mathbf{X}}\right)^T \mathbf{C}_r^{-1} \left(\mathbf{Y}^\delta - \mathbf{K} \widehat{\mathbf{X}}\right),$$

with

$$\mathbf{C}_r = \mathbf{C}_\delta (\mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta)^{-1} \mathbf{C}_\delta, \quad (\text{D.4})$$

is Chi-square distributed with  $m$  degrees of freedom, and the random variable

$$\widehat{C} = \widehat{\mathbf{X}}^T \mathbf{C}_{\widehat{\mathbf{x}}}^{-1} \widehat{\mathbf{X}},$$

with

$$\mathbf{C}_{\widehat{\mathbf{x}}} = \mathbf{C}_x \mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} (\mathbf{K}^T \mathbf{C}_\delta^{-1} \mathbf{K} + \mathbf{C}_x^{-1})^{-1}, \quad (\text{D.5})$$

is Chi-square distributed with  $n$  degrees of freedom.

*Proof.* By (D.2), (D.3), and the identity (cf. (4.28))

$$\mathbf{I}_m - \widehat{\mathbf{A}} = \mathbf{C}_\delta (\mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta)^{-1},$$

we get  $\mathcal{E}\{\mathbf{Y}^\delta - \mathbf{K} \widehat{\mathbf{X}}\} = \mathbf{0}$  and

$$\begin{aligned} \mathbf{C}_r &= \mathcal{E}\left\{\left(\mathbf{Y}^\delta - \mathbf{K} \widehat{\mathbf{X}}\right) \left(\mathbf{Y}^\delta - \mathbf{K} \widehat{\mathbf{X}}\right)^T\right\} \\ &= \left(\mathbf{I}_m - \widehat{\mathbf{A}}\right) (\mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta) \left(\mathbf{I}_m - \widehat{\mathbf{A}}\right)^T \\ &= \mathbf{C}_\delta (\mathbf{K} \mathbf{C}_x \mathbf{K}^T + \mathbf{C}_\delta)^{-1} \mathbf{C}_\delta. \end{aligned}$$

Thus,  $\mathbf{Y}^\delta - \mathbf{K} \widehat{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_r)$ . On the other hand, we have  $\widehat{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\widehat{\mathbf{x}}})$ , since by virtue of (4.25),  $\mathbf{C}_{\widehat{\mathbf{x}}}$ , as given by (D.5), is the covariance matrix of the estimator  $\widehat{\mathbf{X}}$ . The assertions now follow from Proposition D.2.  $\square$

The next result is due to Rao (1973) and is also known as the first fundamental theorem of least squares theory. Although this result deals with unregularized least squares problems, it is of significant importance in statistics.

**Theorem D.5.** *Let*

$$\mathbf{y}^\delta = \mathbf{K}\mathbf{x} + \boldsymbol{\delta}$$

*be a semi-stochastic data model with  $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ , and let  $\mathbf{x}^\delta$  be the least squares solution of the equation  $\mathbf{K}\mathbf{x} = \mathbf{y}^\delta$ . Then, the random variable*

$$r^\delta = \frac{1}{\sigma^2} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}^\delta\|^2$$

*is Chi-square distributed with  $m - n$  degrees of freedom.*

*Proof.* The least squares solution of the equation  $\mathbf{K}\mathbf{x} = \mathbf{y}^\delta$  is given by

$$\mathbf{x}^\delta = \mathbf{K}^\dagger \mathbf{y}^\delta,$$

with

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

and

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

for  $\mathbf{K} = \mathbf{U} \Sigma \mathbf{V}^T$ . The influence matrix possesses the factorization

$$\hat{\mathbf{A}} = \mathbf{K} \mathbf{K}^\dagger = \mathbf{U} \Sigma \Sigma^\dagger \mathbf{U}^T = \mathbf{U} \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T,$$

and we have

$$\mathbf{I}_m - \hat{\mathbf{A}} = \mathbf{U} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-n} \end{bmatrix} \mathbf{U}^T. \quad (\text{D.6})$$

For the exact data vector  $\mathbf{y} \in \mathcal{R}(\mathbf{K}) = \text{span}\{\mathbf{u}_i\}_{i=1, n}$ , (D.6) gives

$$(\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y} = \mathbf{0},$$

and the noisy data vector representation  $\mathbf{y}^\delta = \mathbf{y} + \boldsymbol{\delta}$  then yields

$$\mathbf{y}^\delta - \mathbf{K}\mathbf{x}^\delta = (\mathbf{I}_m - \hat{\mathbf{A}}) \mathbf{y}^\delta = (\mathbf{I}_m - \hat{\mathbf{A}}) \boldsymbol{\delta}.$$

By the change of variable  $\boldsymbol{\delta}_n = (1/\sigma) \boldsymbol{\delta}$ , we obtain

$$r^\delta = \boldsymbol{\delta}_n^T \mathbf{P} \boldsymbol{\delta}_n,$$

with

$$\mathbf{P} = (\mathbf{I}_m - \hat{\mathbf{A}})^T (\mathbf{I}_m - \hat{\mathbf{A}}) = \mathbf{U} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-n} \end{bmatrix} \mathbf{U}^T. \quad (\text{D.7})$$

From (D.7) we deduce that  $\mathbf{P}$  is idempotent ( $\mathbf{P}^2 = \mathbf{P}$ ), and that trace( $\mathbf{P}$ ) =  $m - n$ . Since  $\boldsymbol{\delta}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ , it follows immediately that  $r^\delta \sim \chi^2(m - n)$ .  $\square$

# E

## A general iterative regularization method for linear problems

In this appendix we introduce a general framework for analyzing iterative regularization methods. The treatment is similar to the analysis of direct regularization methods but is restricted to the application of the discrepancy principle as stopping rule. This deterministic analysis can be applied only to linear regularization methods, such as the Landweber iteration and semi-iterative methods, while for nonlinear regularization methods, e.g., the conjugate gradient method, a different technique will be used.

### E.1 Linear regularization methods

In a general framework of iterative methods, the regularized solutions are given by (cf. (C.1) and (C.2))

$$\mathbf{x}_k^\delta = \sum_{i=1}^n f_k(\sigma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i, \quad (\text{E.1})$$

$$\mathbf{x}_k = \sum_{i=1}^n f_k(\sigma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i, \quad (\text{E.2})$$

and the smoothing and noise errors by (cf. (C.6) and (C.7))

$$\mathbf{e}_{\text{sk}} = \mathbf{x}^\dagger - \mathbf{x}_k = \sum_{i=1}^n r_k(\sigma_i^2) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}) \mathbf{v}_i, \quad (\text{E.3})$$

$$\mathbf{e}_{\text{nk}}^\delta = \mathbf{x}_k - \mathbf{x}_k^\delta = - \sum_{i=1}^n [\sigma_i^2 g_k(\sigma_i^2)] \frac{1}{\sigma_i} (\mathbf{u}_i^T \boldsymbol{\delta}) \mathbf{v}_i. \quad (\text{E.4})$$

In (E.1)–(E.4),  $f_k(\lambda)$  are the filter polynomials,  $g_k(\lambda) = f_k(\lambda)/\lambda$  are the iteration polynomials, and  $r_k(\lambda) = 1 - f_k(\lambda)$  are the residual polynomials. To simplify our analysis we consider iterative regularization methods with  $\mathbf{x}_0^\delta = \mathbf{0}$ .

In addition to the assumption  $\|\mathbf{K}\| \leq 1$ , we suppose that (compare to (C.9)–(C.11))

$$0 \leq g_k(\lambda) \leq c_1 k, \quad (\text{E.5})$$

$$|r_k(\lambda)| \leq 1, \quad r_k(0) = 1, \quad (\text{E.6})$$

$$\lambda^\mu |r_k(\lambda)| \leq \frac{c_2}{k^\mu}, \quad 0 < \mu \leq \mu_0, \quad (\text{E.7})$$

for all  $\lambda \in [0, 1]$ ,  $k \geq 1$ , and  $c_1, c_2 > 0$ . By virtue of (E.6), the iteration polynomials are bounded as

$$0 \leq \lambda g_k(\lambda) \leq 2. \quad (\text{E.8})$$

The analysis will be carried out under the standard source condition

$$\xi_i = \sigma_i^{2\mu} \zeta_i, \quad i = 1, \dots, n, \quad (\text{E.9})$$

where  $\xi_i$  are the Fourier coefficients of the exact solution  $\mathbf{x}^\dagger$ , i.e.,  $\mathbf{x}^\dagger = \sum_{i=1}^n \xi_i \mathbf{v}_i$ , and  $\zeta_i$  are the Fourier coefficients of a vector  $\mathbf{z} \in \mathbb{R}^n$  of reasonable norm, i.e.,  $\mathbf{z} = \sum_{i=1}^n \zeta_i \mathbf{v}_i$ .

As for direct regularization methods, we define the matrix  $\mathbf{H}_{\text{dp}k}$  through the relation

$$\mathbf{H}_{\text{dp}k} \mathbf{w} = \sum_{i=1}^m r_k(\sigma_i^2) (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i, \quad \mathbf{w} \in \mathbb{R}^m, \quad (\text{E.10})$$

with the convention  $r_k(\sigma_i^2) = 1$  for  $i = n+1, \dots, m$ . Note that the norm of the matrix  $\mathbf{H}_{\text{dp}k}$  is smaller than or equal to one, i.e.,

$$\|\mathbf{H}_{\text{dp}k} \mathbf{w}\|^2 \leq \|\mathbf{w}\|^2, \quad \mathbf{w} \in \mathbb{R}^m,$$

and, for the exact data vector  $\mathbf{y}$ , the ‘residual’

$$\|\mathbf{H}_{\text{dp}k} \mathbf{y}\|^2 = \sum_{i=1}^n r_k^2(\sigma_i^2) \sigma_i^{4\mu+2} \zeta_i^2 \quad (\text{E.11})$$

can be estimated as (cf. (E.7))

$$\|\mathbf{H}_{\text{dp}k} \mathbf{y}\|^2 \leq c_2^2 \frac{\|\mathbf{z}\|^2}{k^{2\mu+1}}, \quad 0 < \mu \leq \mu_0 - \frac{1}{2}. \quad (\text{E.12})$$

In view of the identity

$$\|\mathbf{H}_{\text{dp}k} \mathbf{y}^\delta\|^2 = \|\mathbf{r}_k^\delta\|^2, \quad (\text{E.13})$$

the discrepancy principle for iterative methods can be formulated as follows: the iteration is terminated for  $k = k^*$  when

$$\|\mathbf{H}_{\text{dp}k^*} \mathbf{y}^\delta\|^2 \leq \tau \Delta^2 < \|\mathbf{H}_{\text{dp}k} \mathbf{y}^\delta\|^2, \quad 0 < k < k^*. \quad (\text{E.14})$$

**Theorem E.1.** *Let assumptions (E.5)–(E.7) hold and let  $\mathbf{x}^\dagger$  satisfy the source condition (E.9). If  $k^*$  is the stopping index of the discrepancy principle (E.14) with  $\tau > 1$ , we have the error estimate*

$$\|\mathbf{e}_{k^*}^\delta\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \mu_0 - \frac{1}{2}.$$

*Proof.* First, we derive estimates for the noise and smoothing errors. By (E.5) and (E.8), we have

$$\sigma_i^2 g_k^2 (\sigma_i^2) \leq 2c_1 k,$$

and a noise error estimate is then given by (compare to (C.23))

$$\|\mathbf{e}_{nk}^\delta\|^2 \leq c_n^2 k \Delta^2, \quad (\text{E.15})$$

with  $c_n = \sqrt{2c_1}$ . Employing the same arguments as in Theorem C.4, we find that, for the source condition (E.9), there holds (compare to (C.37))

$$\|\mathbf{e}_{sk^*}\|^2 \leq c_{\text{sdp}}^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}, \quad (\text{E.16})$$

with

$$c_{\text{sdp}} = (1 + \sqrt{\tau})^{\frac{2\mu}{2\mu+1}}.$$

A bound for the termination index can be derived by using the inequality (cf. (C.38)),

$$\|\mathbf{H}_{\text{dp}k^*-1}\mathbf{y}\| > (\sqrt{\tau} - 1) \Delta, \quad \tau > 1, \quad (\text{E.17})$$

and the estimate (cf. (E.12))

$$\|\mathbf{H}_{\text{dp}k^*-1}\mathbf{y}\| \leq c_2 \frac{\|\mathbf{z}\|}{(k^* - 1)^{\mu+\frac{1}{2}}}, \quad 0 < \mu \leq \mu_0 - \frac{1}{2}. \quad (\text{E.18})$$

From (E.17) and (E.18), we obtain

$$k^* - 1 < \left( \frac{c_2}{\sqrt{\tau} - 1} \right)^{\frac{2}{2\mu+1}} \left( \frac{\|\mathbf{z}\|}{\Delta} \right)^{\frac{2}{2\mu+1}}$$

and since

$$\frac{k^*}{2} \leq k^* - 1, \quad k^* > 1,$$

it follows that

$$k^* < 2 \left( \frac{c_2}{\sqrt{\tau} - 1} \right)^{\frac{2}{2\mu+1}} \left( \frac{\|\mathbf{z}\|}{\Delta} \right)^{\frac{2}{2\mu+1}}.$$

The noise error estimate (E.15) then becomes

$$\|\mathbf{e}_{nk^*}^\delta\|^2 < c_{\text{ndp}}^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}, \quad (\text{E.19})$$

with

$$c_{\text{ndp}} = \sqrt{2}c_n \left( \frac{c_2}{\sqrt{\tau} - 1} \right)^{\frac{1}{2\mu+1}}.$$

In view of (E.16) and (E.19), we deduce that the discrepancy principle is an order-optimal stopping rule for  $0 < \mu \leq \mu_0 - 1/2$ .  $\square$

We proceed now to check assumptions (E.5)–(E.7) for the Landweber iteration and semi-iterative methods.

The Landweber iteration is characterized by

$$f_k(\lambda) = 1 - (1 - \lambda)^k, \quad g_k(\lambda) = \frac{1}{\lambda} \left[ 1 - (1 - \lambda)^k \right], \quad r_k(\lambda) = (1 - \lambda)^k.$$

Taking into account that, for  $\lambda \in [0, 1]$ ,

$$g_k(\lambda) = \frac{1 - (1 - \lambda)^k}{\lambda} = \sum_{l=0}^{k-1} (1 - \lambda)^l \leq k,$$

we see that assumption (E.5) holds with  $c_1 = 1$ . To determine the qualification of the Landweber iteration, we use the inequality

$$(1 - \lambda)^k \leq e^{-\lambda k}, \quad 0 \leq \lambda \leq 1,$$

and find that

$$\lambda^\mu r_k(\lambda) = \lambda^\mu (1 - \lambda)^k \leq \lambda^\mu e^{-\lambda k} = \frac{s^\mu e^{-s}}{k^\mu},$$

with  $\mu > 0$ ,  $k \geq 1$  and  $s = \lambda k \geq 0$ . The function

$$h_\mu(s) = s^\mu e^{-s}$$

attains its maximum at  $s = \mu$ , and we obtain

$$0 \leq \lambda^\mu r_k(\lambda) \leq \frac{\mu^\mu e^{-\mu}}{k^\mu}.$$

Thus, assumption (E.7) holds for  $\mu > 0$ , with

$$c_2 = \mu^\mu e^{-\mu},$$

and we say that the qualification of the Landweber iteration is  $\mu_0 = \infty$ .

In all semi-iterative methods which can be found in the literature, assumption (E.6) holds, that is, we have  $|r_k(\lambda)| \leq 1$  for all  $\lambda \in [0, 1]$ , and  $r_k(0) = 1$ . The residual polynomials have additional properties which lead to a reduced set of assumptions as compared to (E.5)–(E.7). One such property is the Markov inequality,

$$|r'_k(\lambda)| \leq 2k^2, \quad 0 \leq \lambda \leq 1.$$

Taking into account that  $r_k(0) = 1$  and using the mean value theorem, we obtain

$$g_k(\lambda) = \frac{1 - r_k(\lambda)}{\lambda} = -\frac{1}{\lambda} \int_0^\lambda r'_k(x) \, dx = -r'_k(\lambda_0)$$

for some  $\lambda_0 \in [0, \lambda]$ . Then, we find that

$$0 \leq g_k(\lambda) \leq \sup_{0 \leq \lambda_0 \leq 1} |r'_k(\lambda_0)| \leq 2k^2, \quad (\text{E.20})$$



and this result is similar to assumption (E.5) with  $k^2$  in place of  $k$ . In agreement with (E.20), we change assumption (E.7) and require that, for  $k \geq 1$ ,

$$\lambda^\mu |r_k(\lambda)| \leq \frac{c_2}{k^{2\mu}}, \quad 0 < \mu \leq \mu_0. \quad (\text{E.21})$$

Employing the same arguments as in Theorem E.1, we can show that, under assumption (E.21), a semi-iterative method is of optimal order for  $0 < \mu \leq \mu_0 - 1/2$ , provided the iteration is stopped according to the discrepancy principle. The  $\nu$ -method of Brakhage (1987) has the qualification  $\mu_0 = \nu$ , and as a result, the regularized solutions obtained with the discrepancy principle are order-optimal for  $0 < \mu \leq \nu - 1/2$  and  $\nu > 1/2$ . Note that in contrast to the Landweber iteration, the  $\nu$ -method has a finite qualification and the solution error does not longer decrease with optimal rate when  $\mu > \nu - 1/2$ .

## E.2 Conjugate gradient method

The regularizing property of the conjugate gradient for normal equations (CGNR) will be established by particularizing the results derived in Rieder (2003) to a discrete setting. To simplify our analysis we assume that  $\text{rank}(\mathbf{K}) = n$ .

The iterates of the CGNR method can be expressed in terms of the iteration polynomials  $g_k$  of degree  $k - 1$  as

$$\mathbf{x}_k^\delta = g_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta,$$

where

$$g_k(\mathbf{K}^T \mathbf{K}) = \mathbf{V} \left[ \text{diag}(g_k(\sigma_i^2))_{n \times n} \right] \mathbf{V}^T, \quad (\text{E.22})$$

for  $\mathbf{K} = \mathbf{U} \Sigma \mathbf{V}^T$ . The residual polynomials  $r_k(\lambda) = 1 - \lambda g_k(\lambda)$ , satisfying the normalization condition  $r_k(0) = 1$ , are polynomials of degree  $k$ . Both the iteration polynomials and the residual polynomials depend on  $\mathbf{y}^\delta$  and for this reason, CGNR is a nonlinear regularization method.

Before proceeding, we derive some matrix identities which will be frequently used in the sequel. By virtue of (E.22), we have the matrix factorization

$$\mathbf{K} g_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T = \mathbf{U} \begin{bmatrix} \text{diag}(\sigma_i^2 g_k(\sigma_i^2))_{n \times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T. \quad (\text{E.23})$$

This gives

$$\mathbf{I}_m - \mathbf{K} g_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T = \mathbf{U} \left[ \text{diag}(r_k(\sigma_i^2))_{m \times m} \right] \mathbf{U}^T \quad (\text{E.24})$$

and

$$\mathbf{K}^T [\mathbf{I}_m - \mathbf{K} g_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T] = \mathbf{V} \left[ \text{diag}(\sigma_i r_k(\sigma_i^2))_{n \times n} \quad \mathbf{0} \right] \mathbf{U}^T, \quad (\text{E.25})$$

with the convention  $r_k(\sigma_i^2) = 1$  for  $i = n + 1, \dots, m$ . Setting

$$r_k(\mathbf{K}^T \mathbf{K}) = \mathbf{V} \left[ \text{diag}(r_k(\sigma_i^2))_{n \times n} \right] \mathbf{V}^T \quad (\text{E.26})$$

and

$$r_k(\mathbf{K}\mathbf{K}^T) = \mathbf{U} \left[ \text{diag}(r_k(\sigma_i^2))_{m \times m} \right] \mathbf{U}^T, \quad (\text{E.27})$$

we express (E.24) and (E.25) as

$$\mathbf{I}_m - \mathbf{K}g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T = r_k(\mathbf{K}\mathbf{K}^T) \quad (\text{E.28})$$

and

$$\mathbf{K}^T [\mathbf{I}_m - \mathbf{K}g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T] = r_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T, \quad (\text{E.29})$$

respectively. As a result, we find that

$$\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta = [\mathbf{I}_m - \mathbf{K}g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T] \mathbf{y}^\delta = r_k(\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta, \quad (\text{E.30})$$

and that

$$\mathbf{K}^T \mathbf{r}_k^\delta = \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta) = \mathbf{K}^T [\mathbf{I}_m - \mathbf{K}g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T] \mathbf{y}^\delta = r_k(\mathbf{K}^T\mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta. \quad (\text{E.31})$$

We also note the matrix factorizations

$$\mathbf{I}_n - g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T\mathbf{K} = \mathbf{V} \left[ \text{diag}(r_k(\sigma_i^2))_{n \times n} \right] \mathbf{V}^T = r_k(\mathbf{K}^T\mathbf{K}) \quad (\text{E.32})$$

and

$$g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T = \mathbf{V} \left[ \text{diag}(\sigma_i g_k(\sigma_i^2))_{n \times n} \quad \mathbf{0} \right] \mathbf{U}^T. \quad (\text{E.33})$$

## E.2.1 CG-polynomials

The CG-polynomials possess some interesting properties which we now describe.

Assuming a zero initial guess, i.e.,  $\mathbf{x}_0^\delta = \mathbf{0}$ , the  $k$ th iterate of the CGNR method is defined by

$$\mathbf{x}_k^\delta = \arg \min_{\mathbf{x}_k \in \mathcal{K}_k} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k\|^2.$$

Thus, we have

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\| \leq \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k\| \quad (\text{E.34})$$

for all  $\mathbf{x}_k \in \mathcal{K}_k$ , where

$$\mathcal{K}_k = \text{span} \left\{ \mathbf{K}^T \mathbf{y}^\delta, (\mathbf{K}^T\mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta, \dots, (\mathbf{K}^T\mathbf{K})^{k-1} \mathbf{K}^T \mathbf{y}^\delta \right\}$$

is  $k$ th Krylov subspace. For any vector  $\mathbf{x}_k \in \mathcal{K}_k$ , there exist the scalars  $\varsigma_l$ ,  $l = 0, \dots, k-1$ , so that  $\mathbf{x}_k$  can be expanded as

$$\mathbf{x}_k = \sum_{l=0}^{k-1} \varsigma_l (\mathbf{K}^T\mathbf{K})^l \mathbf{K}^T \mathbf{y}^\delta = \left[ \sum_{l=0}^{k-1} \varsigma_l (\mathbf{K}^T\mathbf{K})^l \right] \mathbf{K}^T \mathbf{y}^\delta = g(\mathbf{K}^T\mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta, \quad (\text{E.35})$$

where

$$g(\lambda) = \sum_{l=0}^{k-1} \varsigma_l \lambda^l$$

is a polynomial of degree  $k - 1$ . Thus, for any vector  $\mathbf{x}_k \in \mathcal{K}_k$ , there exists a polynomial  $g$  of degree  $k - 1$  so that (E.35) holds. In this regard, (E.30) together with (E.34), shows that the residual polynomial  $r_k$  has the optimality property

$$\|r_k (\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\| \leq \|r (\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\| \quad (\text{E.36})$$

for all  $r \in \mathcal{P}_k^0$ , where  $\mathcal{P}_k^0$  is the set of normalized polynomials of degree  $k$ ,

$$\mathcal{P}_k^0 = \{p \in \mathcal{P}_k / p(0) = 1\},$$

and  $\mathcal{P}_k$  is the set of polynomials of degree  $k$ .

From the derivation of the CGNR algorithm, we know that the vectors

$$\mathbf{s}_0 = \mathbf{K}^T \mathbf{y}^\delta, \quad \mathbf{s}_k = \mathbf{K}^T \mathbf{r}_k^\delta, \quad k \geq 1,$$

are orthogonal, that is,

$$\mathbf{s}_k^T \mathbf{s}_l = 0, \quad k \neq l.$$

By (E.25) and (E.31), we have, for  $k \geq 0$  and the convention  $r_0(\lambda) = 1$ ,

$$\mathbf{s}_k = r_k (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta = \sum_{i=1}^n \sigma_i r_k (\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i,$$

and the orthogonality relation yields

$$\sum_{i=1}^n \sigma_i^2 r_k (\sigma_i^2) r_l (\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = 0, \quad k, l \geq 0, \quad k \neq l. \quad (\text{E.37})$$

From the theory of orthogonal polynomials, we note two important results:

- (1) the residual polynomial  $r_k$  has simple real zeros  $\lambda_{k,j}$ ,  $j = 1, \dots, k$ , assumed to be in decreasing order

$$0 < \lambda_{k,k} < \lambda_{k,k-1} < \dots < \lambda_{k,1}; \quad (\text{E.38})$$

- (2) the zeros of  $r_k$  and  $r_{k-1}$  are interlacing, i.e.,

$$0 < \lambda_{k,k} < \lambda_{k-1,k-1} < \lambda_{k,k-1} < \dots < \lambda_{k,2} < \lambda_{k-1,1} < \lambda_{k,1}. \quad (\text{E.39})$$

The normalization condition  $r_k(0) = 1$  yields the representation

$$r_k(\lambda) = \prod_{j=1}^k \left(1 - \frac{\lambda}{\lambda_{k,j}}\right) = \prod_{j=1}^k \frac{\lambda_{k,j} - \lambda}{\lambda_{k,j}}. \quad (\text{E.40})$$

To analyze the behavior of the residual polynomials, we need to compute the derivatives  $r'_k$  and  $r''_k$ . The first-order derivative of  $r_k$  is given by

$$r'_k(\lambda) = - \sum_{j=1}^k \frac{1}{\lambda_{k,j}} \prod_{i \neq j}^k \left(1 - \frac{\lambda}{\lambda_{k,i}}\right), \quad (\text{E.41})$$

and we have

$$r'_k(0) = -\sum_{j=1}^k \frac{1}{\lambda_{k,j}}. \quad (\text{E.42})$$

To compute the second-order derivative, we set

$$r'_k(\lambda) = -\sum_{j=1}^k \frac{1}{\lambda_{k,j}} R_j(\lambda),$$

with

$$R_j(\lambda) = \prod_{i \neq j}^k \left(1 - \frac{\lambda}{\lambda_{k,i}}\right),$$

and use the result

$$R'_j(\lambda) = -\sum_{i \neq j}^k \frac{1}{\lambda_{k,i}} \prod_{l \neq i, l \neq j}^k \left(1 - \frac{\lambda}{\lambda_{k,l}}\right)$$

to obtain

$$\begin{aligned} r''_k(\lambda) &= -\sum_{j=1}^k \frac{1}{\lambda_{k,j}} R'_j(\lambda) \\ &= r_k(\lambda) \sum_{j=1}^k \sum_{i \neq j}^k \frac{1}{\lambda_{k,i} - \lambda} \frac{1}{\lambda_{k,j} - \lambda} \\ &= r_k(\lambda) \left[ \left( \sum_{j=1}^k \frac{1}{\lambda_{k,j} - \lambda} \right)^2 - \sum_{j=1}^k \frac{1}{(\lambda_{k,j} - \lambda)^2} \right]. \end{aligned} \quad (\text{E.43})$$

In the proof of the convergence rate we will restrict our analysis to the interval  $[0, \lambda_{k,k}]$ . It is therefore useful to study the behavior of the polynomials  $r_k$  and  $g_k$  in this interval. From (E.40), we have

$$0 \leq r_k(\lambda) \leq 1, \quad \lambda \in [0, \lambda_{k,k}], \quad (\text{E.44})$$

while from (E.41) and (E.43), we obtain

$$r'_k(\lambda) \leq 0, \quad \lambda \in [0, \lambda_{k,k}],$$

and

$$r''_k(\lambda) \geq 0, \quad \lambda \in [0, \lambda_{k,k}],$$

respectively.

By the definition of the residual polynomials, there holds

$$r'_k(\lambda) = -g_k(\lambda) - \lambda g'_k(\lambda),$$

and so (cf. (E.42)),

$$g_k(0) = -r'_k(0) = \sum_{j=1}^k \frac{1}{\lambda_{k,j}}. \quad (\text{E.45})$$

The iteration polynomial  $g_k$  is monotonically decreasing in  $[0, \lambda_{k,k}]$ . To prove this result, we will show that

$$g'_k(\lambda) = -\frac{r'_k(\lambda)}{\lambda} - \frac{g_k(\lambda)}{\lambda} = -\frac{r'_k(\lambda)}{\lambda} - \frac{1}{\lambda} \frac{1 - r_k(\lambda)}{\lambda} \quad (\text{E.46})$$

is non-positive in  $[0, \lambda_{k,k}]$ . For the function

$$-\frac{1 - r_k(\lambda)}{\lambda} = \frac{r_k(\lambda) - r_k(0)}{\lambda},$$

we use the mean value theorem to obtain

$$-\frac{1 - r_k(\lambda)}{\lambda} = \frac{1}{\lambda} \int_0^\lambda r'_k(x) \, dx = r'_k(\lambda_0),$$

for some  $\lambda_0 \in [0, \lambda]$ . Then, as  $r'_k$  is monotonically increasing ( $r''_k \geq 0$ ), we find that

$$-\frac{1 - r_k(\lambda)}{\lambda} = r'_k(\lambda_0) \leq r'_k(\lambda). \quad (\text{E.47})$$

Combining (E.46) and (E.47) yields

$$g'_k(\lambda) = -\frac{r'_k(\lambda)}{\lambda} - \frac{1}{\lambda} \frac{1 - r_k(\lambda)}{\lambda} \leq -\frac{r'_k(\lambda)}{\lambda} + \frac{r'_k(\lambda)}{\lambda} = 0,$$

and so,  $g_k$  is monotonically decreasing in  $[0, \lambda_{k,k}]$ . As a result, (E.45) can be expressed in a more general form as

$$0 < g_k(\lambda) \leq g_k(0) = -r'_k(0) = \sum_{j=1}^k \frac{1}{\lambda_{k,j}}, \quad \lambda \in [0, \lambda_{k,k}]. \quad (\text{E.48})$$

The zeros of the residual polynomial  $r_k$  are related to the singular values of the matrix  $\mathbf{K}$  via

$$\lambda_{n,j} = \sigma_n^2, \quad j = 1, \dots, n \quad (\text{E.49})$$

and

$$\sigma_n^2 < \lambda_{k,k} < \lambda_{k,1} < \sigma_1^2, \quad k = 1, \dots, n-1. \quad (\text{E.50})$$

To prove the first assertion we define the polynomial

$$r(\lambda) = \prod_{j=1}^n \left( 1 - \frac{\lambda}{\sigma_j^2} \right) \in \mathcal{P}_n^0,$$

and use the optimality property (E.36) of  $r_n$  and the identities  $r(\sigma_i^2) = 0$ , for  $i = 1, \dots, n$ , to obtain

$$\begin{aligned} \|r_n(\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\|^2 &= \sum_{i=1}^n r_n^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &\leq \|r(\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\|^2 = \sum_{i=1}^n r^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &= \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2, \end{aligned}$$

that is,

$$\sum_{i=1}^n r_n^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \leq 0. \quad (\text{E.51})$$

From (E.51) we get  $r_n^2(\sigma_i^2) = 0$  for all  $i = 1, \dots, n$ , and the proof is finished. The second assertion follows from (E.49) and the interlacing property of the zeros of the residual polynomials given by (E.39).

## E.2.2 Discrepancy principle

In this section we derive the convergence rate of the CGNR method when the discrepancy principle is used as stopping rule, i.e., when the iteration is terminated with  $k = k^*$  so that

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k^*}^\delta\| \leq \tau_{\text{dp}} \Delta < \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\|, \quad 0 \leq k < k^*. \quad (\text{E.52})$$

For the exact solution  $\mathbf{x}^\dagger$ , we assume the source representation

$$\mathbf{x}^\dagger = (\mathbf{K}^T \mathbf{K})^\mu \mathbf{z}, \quad (\text{E.53})$$

with  $\mu > 0$  and  $\mathbf{z} \in \mathbb{R}^n$ .

If  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of  $\mathbf{K}$ , we define the orthogonal projection matrices

$$\mathbf{E}_\Lambda \mathbf{x} = \sum_{\sigma_i^2 \leq \Lambda} (\mathbf{v}_i^T \mathbf{x}) \mathbf{v}_i, \quad \mathbf{x} \in \mathbb{R}^n,$$

and

$$\mathbf{F}_\Lambda \mathbf{w} = \sum_{\sigma_i^2 \leq \Lambda} (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i, \quad \mathbf{w} \in \mathbb{R}^m,$$

for some  $\Lambda > 0$ . For the matrix  $\mathbf{F}_\Lambda$ , we note the equivalent representation

$$\mathbf{F}_\Lambda \mathbf{w} = \sum_{\sigma_i^2 \leq \Lambda} (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i + P_{\mathcal{R}(\mathbf{K})^\perp} \mathbf{w}, \quad \mathbf{w} \in \mathbb{R}^m,$$

yielding

$$\mathbf{F}_\Lambda \mathbf{w} = \sum_{\sigma_i^2 \leq \Lambda} (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i, \quad \mathbf{w} \in \mathcal{R}(\mathbf{K}),$$

and the result

$$\mathbf{w} - \mathbf{F}_\Lambda \mathbf{w} = \sum_{\sigma_i^2 > \Lambda} (\mathbf{u}_i^T \mathbf{w}) \mathbf{u}_i, \quad \mathbf{w} \in \mathbb{R}^m.$$

By virtue of the identities  $\mathbf{K}^T \mathbf{u}_i = \sigma_i \mathbf{v}_i$ ,  $i = 1, \dots, n$ , we have, for  $0 < \Lambda < \sigma_1^2$ ,

$$\|(\mathbf{I}_n - \mathbf{E}_\Lambda) \mathbf{x}\|^2 = \sum_{\sigma_i^2 > \Lambda} \frac{1}{\sigma_i^2} (\mathbf{u}_i^T \mathbf{K} \mathbf{x})^2 < \frac{1}{\Lambda} \sum_{\sigma_i^2 > \Lambda} (\mathbf{u}_i^T \mathbf{K} \mathbf{x})^2 = \frac{1}{\Lambda} \|(\mathbf{I}_m - \mathbf{F}_\Lambda) \mathbf{K} \mathbf{x}\|^2,$$

that is,

$$\|(\mathbf{I}_n - \mathbf{E}_\Lambda) \mathbf{x}\| < \frac{1}{\sqrt{\Lambda}} \|(\mathbf{I}_m - \mathbf{F}_\Lambda) \mathbf{K} \mathbf{x}\|. \quad (\text{E.54})$$

A general bound for the iteration error is stated by the following result.

**Proposition E.2.** *Let  $\mathbf{x}^\dagger$  satisfy the source condition (E.53). Then, for  $0 < \Lambda \leq \lambda_{k,k}$  and  $0 < k \leq n$ , there holds*

$$\|\mathbf{x}^\dagger - \mathbf{x}_k^\delta\| < \frac{1}{\sqrt{\Lambda}} (\Delta + \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\|) + \|\mathbf{z}\| \Lambda^\mu + \Delta g_k^{\frac{1}{2}}(0). \quad (\text{E.55})$$

*Proof.* The inequality  $\lambda_{k,k} < \sigma_1^2$  (cf. (E.50)) yields  $0 < \Lambda < \sigma_1^2$ , and by (E.54), the iteration error can be estimated as

$$\begin{aligned} \|\mathbf{x}^\dagger - \mathbf{x}_k^\delta\| &\leq \|(\mathbf{I}_n - \mathbf{E}_\Lambda)(\mathbf{x}^\dagger - \mathbf{x}_k^\delta)\| + \|\mathbf{E}_\Lambda(\mathbf{x}^\dagger - \mathbf{x}_k^\delta)\| \\ &< \frac{1}{\sqrt{\Lambda}} \|(\mathbf{I}_m - \mathbf{F}_\Lambda)\mathbf{K}(\mathbf{x}^\dagger - \mathbf{x}_k^\delta)\| + \|\mathbf{E}_\Lambda(\mathbf{x}^\dagger - \mathbf{x}_k^\delta)\|. \end{aligned} \quad (\text{E.56})$$

The terms in the right-hand side of (E.56) can be bounded as

$$\begin{aligned} \|(\mathbf{I}_m - \mathbf{F}_\Lambda)\mathbf{K}(\mathbf{x}^\dagger - \mathbf{x}_k^\delta)\| &= \|(\mathbf{I}_m - \mathbf{F}_\Lambda)(\mathbf{y} - \mathbf{K}\mathbf{x}_k^\delta)\| \\ &\leq \|\mathbf{y} - \mathbf{K}\mathbf{x}_k^\delta\| \\ &\leq \Delta + \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\|, \end{aligned} \quad (\text{E.57})$$

and

$$\begin{aligned} &\|\mathbf{E}_\Lambda(\mathbf{x}^\dagger - \mathbf{x}_k^\delta)\| \\ &= \|\mathbf{E}_\Lambda[\mathbf{x}^\dagger - g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T\mathbf{y}^\delta]\| \\ &\leq \|\mathbf{E}_\Lambda[\mathbf{x}^\dagger - g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T\mathbf{K}\mathbf{x}^\dagger]\| + \|\mathbf{E}_\Lambda[g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T(\mathbf{y}^\delta - \mathbf{y})]\|. \end{aligned} \quad (\text{E.58})$$

We are now concerned with the estimation of the two terms in the right-hand side of (E.58). For the first term, the source condition (E.53) and the representation (E.32) yield the factorization

$$[\mathbf{I}_n - g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T\mathbf{K}](\mathbf{K}^T\mathbf{K})^\mu = \mathbf{V} \left[ \text{diag} \left( \sigma_i^{2\mu} r_k(\sigma_i^2) \right)_{n \times n} \right] \mathbf{V}^T,$$

and we find that

$$\|\mathbf{E}_\Lambda[\mathbf{x}^\dagger - g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T\mathbf{K}\mathbf{x}^\dagger]\|^2 = \sum_{\sigma_i^2 \leq \Lambda} \left[ \sigma_i^{2\mu} r_k(\sigma_i^2) \right]^2 (\mathbf{v}_i^T \mathbf{z})^2.$$

For  $0 \leq \lambda \leq \Lambda \leq \lambda_{k,k}$ , there holds (cf. (E.44))

$$0 \leq \lambda^\mu r_k(\lambda) \leq \lambda^\mu \leq \Lambda^\mu,$$

and we conclude that

$$\|\mathbf{E}_\Lambda[\mathbf{x}^\dagger - g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T\mathbf{K}\mathbf{x}^\dagger]\| \leq \|\mathbf{z}\| \Lambda^\mu. \quad (\text{E.59})$$

For the second term, we use (E.33) to obtain

$$\|\mathbf{E}_\Lambda[g_k(\mathbf{K}^T\mathbf{K})\mathbf{K}^T(\mathbf{y}^\delta - \mathbf{y})]\|^2 = \sum_{\sigma_i^2 \leq \Lambda} \sigma_i^2 g_k^2(\sigma_i^2) (\mathbf{u}_i^T \boldsymbol{\delta})^2.$$

Moreover, from (E.44) and (E.48), we have

$$\lambda g_k^2(\lambda) = [1 - r_k(\lambda)] g_k(\lambda) \leq g_k(0), \quad \lambda \in [0, \lambda_{k,k}],$$

and we end up with

$$\|\mathbf{E}_\Lambda [g_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{y})]\| \leq \Delta g_k^{\frac{1}{2}}(0). \quad (\text{E.60})$$

Now, the conclusion follows from (E.56)–(E.60).  $\square$

For the discrepancy principle index  $k^*$ , the error estimate (E.55) becomes

$$\|\mathbf{x}^\dagger - \mathbf{x}_{k^*}^\delta\| < (1 + \tau_{\text{dp}}) \frac{\Delta}{\sqrt{\Lambda}} + \|\mathbf{z}\| \Lambda^\mu + \Delta g_{k^*}^{\frac{1}{2}}(0). \quad (\text{E.61})$$

Let us evaluate this estimate for the choice

$$\Lambda = \min \left( \left( \frac{\Delta}{\|\mathbf{z}\|} \right)^{\frac{2}{2\mu+1}}, g_{k^*}^{-1}(0) \right). \quad (\text{E.62})$$

Before doing this, we observe that (E.62) gives

$$0 < \Lambda \leq g_{k^*}^{-1}(0) = \left( \sum_{j=1}^{k^*} \frac{1}{\lambda_{k^*,j}} \right)^{-1} < \lambda_{k^*,k^*},$$

and we are in the setting in which (E.61) holds. Now, from (E.62), the second term of the estimate (E.61) can be bounded as

$$\|\mathbf{z}\| \Lambda^\mu \leq \|\mathbf{z}\| \left( \frac{\Delta}{\|\mathbf{z}\|} \right)^{\frac{2\mu}{2\mu+1}} = \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}.$$

To evaluate the first term of the estimate (E.61), we observe that, for

$$\left( \frac{\Delta}{\|\mathbf{z}\|} \right)^{\frac{2}{2\mu+1}} \leq g_{k^*}^{-1}(0),$$

we have

$$\frac{\Delta}{\sqrt{\Lambda}} = \Delta \left( \frac{\|\mathbf{z}\|}{\Delta} \right)^{\frac{1}{2\mu+1}} = \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}},$$

while, for

$$g_{k^*}^{-1}(0) < \left( \frac{\Delta}{\|\mathbf{z}\|} \right)^{\frac{2}{2\mu+1}},$$

we have

$$\frac{\Delta}{\sqrt{\Lambda}} = \Delta g_{k^*}^{\frac{1}{2}}(0).$$



Thus, the solution error can be bounded as

$$\|\mathbf{x}^\dagger - \mathbf{x}_{k^*}^\delta\| < (1 + \tau_{\text{dp}}) \max \left( \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}, \Delta g_{k^*}^{\frac{1}{2}}(0) \right) + \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}} + \Delta g_{k^*}^{\frac{1}{2}}(0). \quad (\text{E.63})$$

From (E.63), it is apparent that the optimal convergence rate can be derived if we are able to prove that

$$\Delta g_{k^*}^{\frac{1}{2}}(0) = O \left( \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}} \right).$$

First, we need an auxiliary result.

**Proposition E.3.** *Let  $\mathbf{x}^\dagger$  satisfy the source condition (E.53). Then, for  $0 < k \leq n$ , there holds*

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\| < \Delta + (1 + 2\mu)^{\mu+\frac{1}{2}} \|\mathbf{z}\| g_k^{-(\mu+\frac{1}{2})}(0).$$

*Proof.* Let us define the polynomial

$$r(\lambda) = \frac{r_k(\lambda)}{1 - \frac{\lambda}{\lambda_{k,k}}} = \lambda_{k,k} \frac{r_k(\lambda)}{\lambda_{k,k} - \lambda}. \quad (\text{E.64})$$

As  $r(\lambda) \in \mathcal{P}_{k-1}^0 = \text{span}\{1, r_1, \dots, r_{k-1}\}$  and  $\lambda_{k,k} > 0$ , the orthogonality relation (E.37) yields

$$\sum_{i=1}^n \sigma_i^2 r_k(\sigma_i^2) \frac{r_k(\sigma_i^2)}{\lambda_{k,k} - \sigma_i^2} (\mathbf{u}_i^T \mathbf{y}^\delta)^2 = 0, \quad (\text{E.65})$$

and we obtain

$$\begin{aligned} \sum_{\sigma_i^2 \leq \lambda_{k,k}} r_k^2(\sigma_i^2) \frac{\sigma_i^2}{\lambda_{k,k} - \sigma_i^2} (\mathbf{u}_i^T \mathbf{y}^\delta)^2 &= \sum_{\sigma_i^2 > \lambda_{k,k}} r_k^2(\sigma_i^2) \frac{\sigma_i^2}{\sigma_i^2 - \lambda_{k,k}} (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &> \sum_{\sigma_i^2 > \lambda_{k,k}} r_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2. \end{aligned} \quad (\text{E.66})$$

Note that for  $k = 1$ , (E.37) is applied with  $r(\lambda) = r_0(\lambda) = 1$  and  $r_1(\lambda) = 1 - \lambda/\lambda_{1,1}$ . Going further, from (E.27), (E.30) and (E.66), we find that

$$\begin{aligned} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\|^2 &= \sum_{\sigma_i^2 \leq \lambda_{k,k}} r_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{\sigma_i^2 > \lambda_{k,k}} r_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &< \sum_{\sigma_i^2 \leq \lambda_{k,k}} \left( 1 + \frac{\sigma_i^2}{\lambda_{k,k} - \sigma_i^2} \right) r_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &= \sum_{\sigma_i^2 \leq \lambda_{k,k}} \varphi_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &= \|\mathbf{F}_{\lambda_{k,k}} \varphi_k(\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\|^2. \end{aligned} \quad (\text{E.67})$$

In (E.67), the function  $\varphi_k$  is defined in terms of the residual polynomial  $r_k$  as

$$\varphi_k(\lambda) = r_k(\lambda) \left(1 + \frac{\lambda}{\lambda_{k,k} - \lambda}\right)^{\frac{1}{2}} = r_k(\lambda) \left(\frac{\lambda_{k,k}}{\lambda_{k,k} - \lambda}\right)^{\frac{1}{2}}, \quad (\text{E.68})$$

and we have the matrix factorization

$$\varphi_k(\mathbf{K}\mathbf{K}^T) = \mathbf{U} \left[ \text{diag}(\varphi_k(\sigma_i^2))_{m \times m} \right] \mathbf{U}^T,$$

with  $\varphi_k(\sigma_i^2) = 1$  for  $i = n+1, \dots, m$ . Application of the triangle inequality to the estimate (E.67) then gives

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\| < \|\mathbf{F}_{\lambda_{k,k}} \varphi_k(\mathbf{K}\mathbf{K}^T) \mathbf{y}\| + \|\mathbf{F}_{\lambda_{k,k}} \varphi_k(\mathbf{K}\mathbf{K}^T) (\mathbf{y}^\delta - \mathbf{y})\|. \quad (\text{E.69})$$

To evaluate the second term in the right-hand side of (E.69), we try to bound  $\varphi_k$  in  $[0, \lambda_{k,k}]$ . From the representation (cf. (E.40) and (E.68))

$$\varphi_k(\lambda) = \left(\frac{\lambda_{k,k} - \lambda}{\lambda_{k,k}}\right)^{\frac{1}{2}} \prod_{j=1}^{k-1} \frac{\lambda_{k,j} - \lambda}{\lambda_{k,j}}$$

and the inequality (cf. (E.38))

$$0 \leq 1 - \frac{\lambda}{\lambda_{k,j}} \leq 1, \quad \lambda \in [0, \lambda_{k,k}], \quad j = 1, \dots, k,$$

we obtain

$$0 \leq \varphi_k(\lambda) \leq 1, \quad \lambda \in [0, \lambda_{k,k}], \quad (\text{E.70})$$

and so,

$$\|\mathbf{F}_{\lambda_{k,k}} \varphi_k(\mathbf{K}\mathbf{K}^T) (\mathbf{y}^\delta - \mathbf{y})\|^2 = \sum_{\sigma_i^2 \leq \lambda_{k,k}} \varphi_k^2(\sigma_i^2) (\mathbf{u}_i^T \boldsymbol{\delta})^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \boldsymbol{\delta})^2 \leq \Delta^2. \quad (\text{E.71})$$

To evaluate the first term in the right-hand side of (E.69), we consider the function

$$\Phi(\lambda) = \lambda^\eta \varphi_k^2(\lambda) = \lambda^\eta \frac{\lambda_{k,k} - \lambda}{\lambda_{k,k}} \prod_{j=1}^{k-1} \left(\frac{\lambda_{k,j} - \lambda}{\lambda_{k,j}}\right)^2$$

for some  $\eta > 1$ . As  $\Phi(0) = \Phi(\lambda_{k,k}) = 0$  and  $\Phi(\lambda) \geq 0$  in  $[0, \lambda_{k,k}]$ , we deduce that, according to Rolle's theorem, there exists an extreme point  $\lambda_\star \in (0, \lambda_{k,k})$  of  $\Phi(\lambda)$ , i.e.,  $\Phi'(\lambda_\star) = 0$ . To compute  $\Phi'$ , we see that, for  $\lambda \in (0, \lambda_{k,k})$ , we have  $\Phi(\lambda) > 0$ , and we may write

$$\log \Phi(\lambda) = \eta \log \lambda + \log \left(\frac{\lambda_{k,k} - \lambda}{\lambda_{k,k}}\right) + 2 \sum_{j=1}^{k-1} \log \left(\frac{\lambda_{k,j} - \lambda}{\lambda_{k,j}}\right).$$

Taking the derivative with respect to  $\lambda$  yields

$$\frac{\Phi'(\lambda)}{\Phi(\lambda)} = \frac{\eta}{\lambda} + \frac{1}{\lambda_{k,k} - \lambda} - 2 \sum_{j=1}^k \frac{1}{\lambda_{k,j} - \lambda},$$

and further

$$\Phi'(\lambda) = \lambda^{\eta-1} \varphi_k^2(\lambda) \left[ \eta + \lambda \left( \frac{1}{\lambda_{k,k} - \lambda} - 2 \sum_{j=1}^k \frac{1}{\lambda_{k,j} - \lambda} \right) \right].$$

Hence,  $\Phi'(\lambda_*) = 0$  gives

$$\eta + \lambda_* \left( \frac{1}{\lambda_{k,k} - \lambda_*} - 2 \sum_{j=1}^k \frac{1}{\lambda_{k,j} - \lambda_*} \right) = 0,$$

and we infer that (cf. (E.48))

$$\begin{aligned} \eta &= \lambda_* \left( 2 \sum_{j=1}^k \frac{1}{\lambda_{k,j} - \lambda_*} - \frac{1}{\lambda_{k,k} - \lambda_*} \right) \\ &> \lambda_* \sum_{j=1}^k \frac{1}{\lambda_{k,j} - \lambda_*} \\ &> \lambda_* \sum_{j=1}^k \frac{1}{\lambda_{k,j}} \\ &= \lambda_* g_k(0). \end{aligned}$$

Thus,

$$\lambda_* < \frac{\eta}{g_k(0)},$$

and because of (E.70), we obtain

$$\lambda^\eta \varphi_k^2(\lambda) \leq \lambda_*^\eta \varphi_k^2(\lambda_*) < \eta^\eta g_k^{-\eta}(0). \quad (\text{E.72})$$

Now, since  $\varphi_k(\mathbf{K}\mathbf{K}^T)\mathbf{y} \in \mathcal{R}(\mathbf{K})$ , the representation

$$\|\mathbf{F}_{\lambda_{k,k}} \varphi_k(\mathbf{K}\mathbf{K}^T)\mathbf{y}\|^2 = \sum_{\sigma_i^2 \leq \lambda_{k,k}} \varphi_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y})^2$$

together with the source condition, written as

$$\mathbf{u}_i^T \mathbf{y} = \mathbf{u}_i^T \mathbf{K} (\mathbf{K}^T \mathbf{K})^\mu \mathbf{z} = \sigma_i^{2\mu+1} \mathbf{v}_i^T \mathbf{z}, \quad (\text{E.73})$$

and the inequality (E.72) with  $\eta = 2\mu + 1 > 1$ , yields

$$\begin{aligned} \|\mathbf{F}_{\lambda_{k,k}} \varphi_k(\mathbf{K}\mathbf{K}^T)\mathbf{y}\|^2 &= \sum_{\sigma_i^2 \leq \lambda_{k,k}} (\sigma_i^2)^{2\mu+1} \varphi_k^2(\sigma_i^2) (\mathbf{v}_i^T \mathbf{z})^2 \\ &< (2\mu + 1)^{2\mu+1} \|\mathbf{z}\|^2 g_k^{-(2\mu+1)}(0). \end{aligned} \quad (\text{E.74})$$

The desired estimate follows from (E.69), (E.71) and (E.74).  $\square$

The key point in our derivation is the following result.

**Proposition E.4.** *Let  $\mathbf{x}^\dagger$  satisfy the source condition (E.53). Then, for any  $\theta \in (0, 1)$ , there exists  $a_\theta$  depending on  $\theta$  and  $\mu$ , so that, for all  $0 < k \leq n$ , there holds*

$$\theta \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\| < \Delta + a_\theta \|\mathbf{z}\| g_k^{-(\mu+\frac{1}{2})}(0). \quad (\text{E.75})$$

*Proof.* For an arbitrary  $\theta \in (0, 1)$ , we set

$$\varsigma = \frac{2 - \theta}{1 - \theta} > 2 \quad (\text{E.76})$$

and

$$q = 1 + \frac{\varsigma}{2} > 2. \quad (\text{E.77})$$

In the first part of the proof we assume that  $k > 1$  and distinguish two cases.

Case 1:  $g_k(0) < qg_{k-1}(0)$ . Using the preceding proposition, we find that

$$\begin{aligned} \theta \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\| &< \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\| \\ &< \Delta + (2\mu + 1)^{\mu+\frac{1}{2}} \|\mathbf{z}\| g_{k-1}^{-(\mu+\frac{1}{2})}(0) \\ &< \Delta + a_\theta \|\mathbf{z}\| g_k^{-(\mu+\frac{1}{2})}(0), \end{aligned}$$

with

$$a_\theta = [q(2\mu + 1)]^{\mu+\frac{1}{2}}. \quad (\text{E.78})$$

Case 2:  $g_k(0) \geq qg_{k-1}(0)$ . We analyze for the moment some consequences of this assumption. Using the interlacing property of the zeros of  $r_k$  and  $r_{k-1}$ ,

$$\lambda_{k-1,j} < \lambda_{k,j}, \quad j = 1, \dots, k-1, \quad (\text{E.79})$$

and employing (E.48), we obtain

$$g_k(0) = \frac{1}{\lambda_{k,k}} + \sum_{j=1}^{k-1} \frac{1}{\lambda_{k,j}} < \frac{1}{\lambda_{k,k}} + \sum_{j=1}^{k-1} \frac{1}{\lambda_{k-1,j}} = \frac{1}{\lambda_{k,k}} + g_{k-1}(0). \quad (\text{E.80})$$

The assumption  $g_k(0) \geq qg_{k-1}(0)$  then yields

$$g_k(0) < \frac{1}{\lambda_{k,k}} + \frac{1}{q} g_k(0),$$

and further

$$\lambda_{k,k} < \frac{q}{q-1} g_k^{-1}(0). \quad (\text{E.81})$$

Moreover, the same assumption together with (E.79) and (E.80) gives

$$(q-1) \frac{1}{\lambda_{k,k-1}} < (q-1) \frac{1}{\lambda_{k-1,k-1}} \leq (q-1) \sum_{j=1}^{k-1} \frac{1}{\lambda_{k-1,j}} = (q-1) g_{k-1}(0) < \frac{1}{\lambda_{k,k}},$$

and we infer that (cf. (E.77))

$$\varsigma \lambda_{k,k} < 2\lambda_{k,k-1}. \quad (\text{E.82})$$

Defining the polynomial  $r(\lambda) \in \mathcal{P}_{k-1}^0$  as in (E.64), that is,

$$r(\lambda) = \frac{r_k(\lambda)}{1 - \frac{\lambda}{\lambda_{k,k}}} = \prod_{j=1}^{k-1} \frac{\lambda_{k,j} - \lambda}{\lambda_{k,j}}, \quad (\text{E.83})$$

and using the optimality property (E.36) of  $r_{k-1}$ , we obtain

$$\begin{aligned} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\| &= \|r_{k-1}(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\| \\ &\leq \|r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\| \\ &\leq \|(\mathbf{I}_m - \mathbf{F}_{\varsigma\lambda_{k,k}})r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\| + \|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\|. \end{aligned} \quad (\text{E.84})$$

The first term in the right-hand side of (E.84) can be bounded as (cf. (E.27), (E.30) and (E.83))

$$\begin{aligned} \|(\mathbf{I}_m - \mathbf{F}_{\varsigma\lambda_{k,k}})r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\|^2 &= \sum_{\sigma_i^2 > \varsigma\lambda_{k,k}} \frac{r_k^2(\sigma_i^2)}{\left(\frac{\sigma_i^2}{\lambda_{k,k}} - 1\right)^2} (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &< \frac{1}{(\varsigma - 1)^2} \sum_{\sigma_i^2 > \varsigma\lambda_{k,k}} r_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &\leq \frac{1}{(\varsigma - 1)^2} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\|^2. \end{aligned} \quad (\text{E.85})$$

Since  $\mathbf{x}_{k-1}^\delta \in \mathcal{K}_{k-1} \subset \mathcal{K}_k$ , (E.34) gives

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_k^\delta\| \leq \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\|,$$

and (E.84) becomes

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\| < \|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\| + \frac{1}{\varsigma - 1} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\|.$$

Thus,

$$\frac{\varsigma - 2}{\varsigma - 1} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\| < \|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\|. \quad (\text{E.86})$$

Now, we need a bound for  $\|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\|$ . This bound will be derived by making use of the triangle inequality

$$\|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)\mathbf{y}^\delta\| \leq \|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)\mathbf{y}\| + \|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)(\mathbf{y}^\delta - \mathbf{y})\|. \quad (\text{E.87})$$

Using the source representation (E.73) and taking into account that  $r(\mathbf{K}\mathbf{K}^T)\mathbf{y} \in \mathcal{R}(\mathbf{K})$ , we estimate the first term in the right-hand side of (E.87) as

$$\begin{aligned} \|\mathbf{F}_{\varsigma\lambda_{k,k}}r(\mathbf{K}\mathbf{K}^T)\mathbf{y}\|^2 &= \sum_{\sigma_i^2 \leq \varsigma\lambda_{k,k}} (\sigma_i^2)^{2\mu+1} r^2(\sigma_i^2) (\mathbf{v}_i^T \mathbf{z})^2 \\ &\leq (\varsigma\lambda_{k,k})^{2\mu+1} \sum_{\sigma_i^2 \leq \varsigma\lambda_{k,k}} r^2(\sigma_i^2) (\mathbf{v}_i^T \mathbf{z})^2 \end{aligned} \quad (\text{E.88})$$

and express the second term as

$$\|\mathbf{F}_{\varsigma\lambda_{k,k}} r(\mathbf{K}\mathbf{K}^T) (\mathbf{y}^\delta - \mathbf{y})\|^2 = \sum_{\sigma_i^2 \leq \varsigma\lambda_{k,k}} r^2(\sigma_i^2) (\mathbf{u}_i^T \boldsymbol{\delta})^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \boldsymbol{\delta})^2. \quad (\text{E.89})$$

To bound these two terms we look at the behavior of  $r(\lambda)$  in  $[0, \varsigma\lambda_{k,k}]$ . From (E.82) we obtain

$$\frac{\varsigma\lambda_{k,k}}{\lambda_{k,j}} < 2, \quad j = 1, \dots, k-1,$$

and further,

$$r^2(\lambda) = \prod_{j=1}^{k-1} \left(1 - \frac{\lambda}{\lambda_{k,j}}\right)^2 \leq 1, \quad \lambda \in [0, \varsigma\lambda_{k,k}].$$

Consequently, (E.87) takes the form

$$\|\mathbf{F}_{\varsigma\lambda_{k,k}} r(\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\| \leq (\varsigma\lambda_{k,k})^{\mu+\frac{1}{2}} \|\mathbf{z}\| + \Delta,$$

and, by virtue of (E.81), (E.86) becomes

$$\frac{\varsigma-2}{\varsigma-1} \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k-1}^\delta\| < \Delta + \left(\frac{q\varsigma}{q-1}\right)^{\mu+\frac{1}{2}} \|\mathbf{z}\| g_k^{-(\mu+\frac{1}{2})}(0).$$

Since (cf. (E.76))

$$\theta = \frac{\varsigma-2}{\varsigma-1}, \quad (\text{E.90})$$

we conclude that (E.75) holds with

$$a_\theta = \left(\frac{q\varsigma}{q-1}\right)^{\mu+\frac{1}{2}}. \quad (\text{E.91})$$

For  $k=1$ , we have  $\mathbf{x}_0^\delta = \mathbf{0}$ ,  $r_1(\lambda) = 1 - \lambda/\lambda_{1,1}$ ,  $r(\lambda) = 1$  and  $g_1(\lambda) = 1/\lambda_{1,1}$ . In this case, we consider the estimate

$$\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_0^\delta\| \leq \|(\mathbf{I}_m - \mathbf{F}_{\varsigma\lambda_{1,1}}) \mathbf{y}^\delta\| + \|\mathbf{F}_{\varsigma\lambda_{1,1}} \mathbf{y}^\delta\|,$$

and proceed as in (E.85)–(E.89); we obtain (E.75) with  $\theta$  as in (E.90) and  $a_\theta = \varsigma^{\mu+\frac{1}{2}}$ .  $\square$

The above proposition allows us to derive the required bound for  $\Delta g_{k^*}^{1/2}(0)$ . For a prescribed tolerance  $\tau_{\text{dp}} > 1$ , we choose  $\theta \in (0, 1)$  so that  $\theta\tau_{\text{dp}} > 1$ . For this  $\theta$ , we compute  $\varsigma$  and  $q$  by using (E.76) and (E.77), respectively, and take  $a_\theta$  as the maximum of the values given by (E.78) and (E.91). In this context, the discrepancy principle condition (E.52) yields

$$\theta\tau_{\text{dp}}\Delta < \theta \|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}_{k^*-1}^\delta\| < \Delta + a_\theta \|\mathbf{z}\| g_{k^*}^{-(\mu+\frac{1}{2})}(0),$$

and we obtain

$$\Delta g_{k^*}^{\frac{1}{2}}(0) < C \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}},$$

with

$$C = \left( \frac{a_\theta}{\theta\tau_{\text{dp}} - 1} \right)^{\frac{1}{2\mu+1}}.$$

We are now in the position to formulate the convergence rate result.

**Theorem E.5.** *Let  $\mathbf{x}^\dagger$  satisfy the source condition (E.53). If  $k^\star$  is the stopping index of the discrepancy principle (E.52) with  $\tau_{\text{dp}} > 1$ , then there holds*

$$\|\mathbf{x}^\dagger - \mathbf{x}_{k^\star}^\delta\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right).$$

The above theorem shows that the CGNR method using the discrepancy principle as stopping rule is an order-optimal regularization method for all  $\mu > 0$ . Thus, there is no saturation effect as in the case of Tikhonov regularization or the  $\nu$ -method.

# F

## Residual polynomials of the LSQR method

The residual polynomials of the LSQR method are normalized polynomials of degree  $k$ . At the iteration step  $k \geq 1$ , the vector  $\mathbf{s}_k = \mathbf{K}^T \mathbf{r}_k^\delta$ , with  $\mathbf{r}_k^\delta = \mathbf{y}^\delta - \mathbf{K} \mathbf{x}_k^\delta$ , can be expressed in terms of the residual polynomial  $r_k$  as (cf. (E.31))

$$\mathbf{s}_k = r_k (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta.$$

As  $\mathbf{s}_k$  is orthogonal to the  $k$ th Krylov subspace  $\mathcal{K}_k$  (see Chapter 5), we have

$$r_k (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta \perp \mathcal{K}_k. \quad (\text{F.1})$$

Let  $\mathbf{B}_k$  be the bidiagonal matrix of the LSQR method at the iteration step  $k$  and let  $(\lambda_{k,j}, \mathbf{w}_{k,j})$  be an eigenpair of the matrix  $\mathbf{B}_k^T \mathbf{B}_k \in \mathbb{R}^k$ , that is,

$$(\mathbf{B}_k^T \mathbf{B}_k) \mathbf{w}_{k,j} = \lambda_{k,j} \mathbf{w}_{k,j}, \quad j = 1, \dots, k. \quad (\text{F.2})$$

The eigenvalues  $\lambda_{k,j}$  are called Ritz values, while the eigenvectors  $\mathbf{w}_{k,j}$  are called Ritz vectors. In exact arithmetic, the representation

$$\mathbf{B}_k^T \mathbf{B}_k = \bar{\mathbf{V}}_k^T (\mathbf{K}^T \mathbf{K}) \bar{\mathbf{V}}_k, \quad (\text{F.3})$$

holds, and we obtain

$$(\mathbf{K}^T \mathbf{K}) \bar{\mathbf{w}}_{k,j} = \lambda_{k,j} \bar{\mathbf{w}}_{k,j}, \quad j = 1, \dots, k, \quad (\text{F.4})$$

with

$$\bar{\mathbf{w}}_{k,j} = \bar{\mathbf{V}}_k \mathbf{w}_{k,j}. \quad (\text{F.5})$$

Before we state the main result of this appendix, let us prove the assertion

$$\bar{\mathbf{w}}_{k,j}^T \mathbf{K}^T \mathbf{y}^\delta \neq 0, \quad j = 1, \dots, k. \quad (\text{F.6})$$

By virtue of (F.4), the following set of equalities holds true:

$$\begin{aligned} \bar{\mathbf{w}}_{k,j}^T (\mathbf{K}^T \mathbf{K})^{k-1} \mathbf{K}^T \mathbf{y}^\delta &= [(\mathbf{K}^T \mathbf{K}) \bar{\mathbf{w}}_{k,j}]^T (\mathbf{K}^T \mathbf{K})^{k-2} \mathbf{K}^T \mathbf{y}^\delta \\ &= \lambda_{k,j} \bar{\mathbf{w}}_{k,j}^T (\mathbf{K}^T \mathbf{K})^{k-2} \mathbf{K}^T \mathbf{y}^\delta = \dots = \lambda_{k,j}^{k-1} \bar{\mathbf{w}}_{k,j}^T \mathbf{K}^T \mathbf{y}^\delta. \end{aligned} \quad (\text{F.7})$$



Now, if we assume that  $\bar{\mathbf{w}}_{k,j}^T \mathbf{K}^T \mathbf{y}^\delta = 0$ , then (F.7) implies that

$$\bar{\mathbf{w}}_{k,j} \perp \mathcal{K}_k = \text{span} \left\{ \mathbf{K}^T \mathbf{y}^\delta, (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta, \dots, (\mathbf{K}^T \mathbf{K})^{k-1} \mathbf{K}^T \mathbf{y}^\delta \right\}.$$

But this result is contradictory since, by (F.5) and the fact that the column vectors of  $\bar{\mathbf{V}}_k$  span  $\mathcal{K}_k$ , we have  $\bar{\mathbf{w}}_{k,j} \in \mathcal{K}_k$ . Thus, (F.6) holds true.

**Theorem F.1.** *Let  $\mathbf{B}_k$  be the bidiagonal matrix of the LSQR method at the iteration step  $k \geq 1$  and let  $\{\lambda_{k,j}\}_{j=1,k}$  be the eigenvalues of  $\mathbf{B}_k^T \mathbf{B}_k$ . Then, the residual polynomial of the LSQR method is given by*

$$r_k(\lambda) = \prod_{j=1}^k \frac{\lambda_{k,j} - \lambda}{\lambda_{k,j}}. \quad (\text{F.8})$$

*Proof.* Assuming the representation  $r_k(\lambda) = \sum_{l=0}^k c_l \lambda^l$  and using the result (cf. (F.3))

$$(\mathbf{B}_k^T \mathbf{B}_k)^l = \bar{\mathbf{V}}_k^T (\mathbf{K}^T \mathbf{K})^l \bar{\mathbf{V}}_k, \quad l \geq 0,$$

we obtain

$$r_k(\mathbf{B}_k^T \mathbf{B}_k) = \sum_{l=0}^k c_l (\mathbf{B}_k^T \mathbf{B}_k)^l = \bar{\mathbf{V}}_k^T \left[ \sum_{l=0}^k c_l (\mathbf{K}^T \mathbf{K})^l \right] \bar{\mathbf{V}}_k = \bar{\mathbf{V}}_k^T r_k(\mathbf{K}^T \mathbf{K}) \bar{\mathbf{V}}_k. \quad (\text{F.9})$$

Combining (F.1) and (F.9) gives

$$r_k(\mathbf{B}_k^T \mathbf{B}_k) \bar{\mathbf{V}}_k^T \mathbf{K}^T \mathbf{y}^\delta = \bar{\mathbf{V}}_k^T r_k(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{y}^\delta = 0. \quad (\text{F.10})$$

On the other hand, (F.2), written in matrix form as

$$\mathbf{B}_k^T \mathbf{B}_k = \mathbf{W}_k \mathbf{\Lambda}_k \mathbf{W}_k^T,$$

with  $\mathbf{W}_k = [\mathbf{w}_{k,1}, \dots, \mathbf{w}_{k,k}]$  and  $\mathbf{\Lambda}_k = [\text{diag}(\lambda_{k,j})_{k \times k}]$ , yields

$$r_k(\mathbf{B}_k^T \mathbf{B}_k) = \mathbf{W}_k \left[ \text{diag}(r_k(\lambda_{k,j}))_{k \times k} \right] \mathbf{W}_k^T. \quad (\text{F.11})$$

Using (F.11) and setting  $\bar{\mathbf{W}}_k = [\bar{\mathbf{w}}_{k,1}, \dots, \bar{\mathbf{w}}_{k,k}]$ , where the  $\bar{\mathbf{w}}_{k,j}$  are defined by (F.5), we express (F.10) as

$$\mathbf{W}_k \left[ \text{diag}(r_k(\lambda_{k,j}))_{k \times k} \right] \bar{\mathbf{W}}_k^T \mathbf{K}^T \mathbf{y}^\delta = 0, \quad (\text{F.12})$$

and further as

$$\sum_{j=1}^k r_k(\lambda_{k,j}) (\bar{\mathbf{w}}_{k,j}^T \mathbf{K}^T \mathbf{y}^\delta) \mathbf{w}_{k,j} = 0.$$

As  $\mathbf{W}_k$  is orthogonal, we find that

$$r_k(\lambda_{k,j}) (\bar{\mathbf{w}}_{k,j}^T \mathbf{K}^T \mathbf{y}^\delta) = 0, \quad j = 1, \dots, k,$$

and in view of (F.6), that

$$r_k(\lambda_{k,j}) = 0, \quad j = 1, \dots, k.$$

This result together with the normalization condition  $r_k(0) = 1$  shows that the residual polynomial is given by (F.8).  $\square$

The above theorem simply states that the zeros of the residual polynomial are the Ritz values. Relationships between the Ritz values, assumed to be distinct and in decreasing order,

$$0 < \lambda_{k,k} < \lambda_{k,k-1} < \dots < \lambda_{k,1}, \quad (\text{F.13})$$

and the eigenvalues  $\sigma_j^2$ ,  $j = 1, \dots, n$ , of the positive definite matrix  $\mathbf{K}^T \mathbf{K}$  can be established by making use of fundamental results from the theory of orthogonal polynomials. In particular, we have (Van der Sluis and Van der Vorst, 1986):

- (1) for any fixed  $j$ ,  $\lambda_{k,j}$  increases and  $\lambda_{k,k-j+1}$  decreases as  $k$  increases from  $j$  to  $n$ ;
- (2) if  $\sigma_{j+1}^2 \leq \lambda_{k,j} \leq \sigma_j^2$  for a certain value of  $k$ , then also for all larger values of  $k$ ;
- (3) any two Ritz values  $\lambda_{k,j+1}$  and  $\lambda_{k,j}$  are separated by at least one eigenvalue  $\sigma_i^2$ ;
- (4)  $\lambda_{n,j} = \sigma_j^2$  for all  $j = 1, \dots, n$  (see Appendix E).

The first and the last property show that for any fixed  $j$ , the increasing sequence  $\{\lambda_{k,j}\}_{k=j, n}$  attains its maximum  $\sigma_j^2$  at  $k = n$ , and definitely, we may write

$$\lambda_{k,j} < \lambda_{n,j} = \sigma_j^2, \quad k = j, \dots, n-1. \quad (\text{F.14})$$

For ill-posed problems, this result is even stronger: if the eigenvalues of  $\mathbf{K}^T \mathbf{K}$  are well separated and do not decay too slowly, and moreover, if the discrete Picard condition is satisfied, then the first Ritz values  $\lambda_{k,j}$  approximate the largest eigenvalues  $\sigma_j^2$  in their natural order (Hansen, 1998). To heuristically explain this assertion, we assume that the discrete Picard condition (see Chapter 3)

$$|\mathbf{u}_i^T \mathbf{y}^\delta| = C \sigma_i^{\beta+1}, \quad i = 1, \dots, n, \quad (\text{F.15})$$

with  $\beta > 0$  and  $C > 0$ , is satisfied, and that the eigenvalues  $\sigma_i^2$  decay very rapidly as  $i$  increases, e.g.,

$$\sigma_{i+1}^2 = q_i \sigma_i^2, \quad q_i \ll 1. \quad (\text{F.16})$$

Defining the polynomial

$$r(\lambda) = \prod_{j=1}^k \left(1 - \frac{\lambda}{\sigma_j^2}\right) \in \mathcal{P}_k^0,$$

and using the optimality property (E.36) of  $r_k$  and the identities  $r(\sigma_i^2) = 0$ ,  $i = 1, \dots, k$ , we obtain, for  $k < n$ ,

$$\begin{aligned} \|r_k(\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\|^2 &= \sum_{i=1}^n r_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &\leq \|r(\mathbf{K}\mathbf{K}^T) \mathbf{y}^\delta\|^2 = \sum_{i=1}^n r^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \\ &= \sum_{i=k+1}^n r^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{y}^\delta)^2, \end{aligned}$$

that is,

$$\sum_{i=1}^n r_k^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2 \leq \sum_{i=k+1}^n r^2(\sigma_i^2) (\mathbf{u}_i^T \mathbf{y}^\delta)^2.$$

By making use of the discrete Picard condition (F.15), we rewrite the above inequality as

$$\sum_{i=1}^n r_k^2 (\sigma_i^2) \sigma_i^{2\beta+2} \leq \sum_{i=k+1}^n r^2 (\sigma_i^2) \sigma_i^{2\beta+2}. \quad (\text{F.17})$$

In view of assumption (F.16), we get

$$r(\sigma_i^2) = \left(1 - \frac{\sigma_i^2}{\sigma_1^2}\right) \dots \left(1 - \frac{\sigma_i^2}{\sigma_k^2}\right) \lesssim 1, \quad i = k+1, \dots, n,$$

and further,

$$\sum_{i=k+1}^n r^2 (\sigma_i^2) \sigma_i^{2\beta+2} \leq (n-k) \sigma_{k+1}^{2\beta+2}.$$

As a result, (F.17) implies that

$$r_k^2 (\sigma_i^2) \sigma_i^{2\beta+2} \leq (n-k) \sigma_{k+1}^{2\beta+2}, \quad i = 1, \dots, k. \quad (\text{F.18})$$

Let us now analyze the consequences of condition (F.18). For  $i = 1$ , we have

$$r_k (\sigma_1^2) = \left(1 - \frac{\sigma_1^2}{\lambda_{k,1}}\right) \dots \left(1 - \frac{\sigma_1^2}{\lambda_{k,k}}\right),$$

and from (cf. (F.13) and (F.14))

$$\lambda_{k,k} < \lambda_{k,k-1} < \dots < \lambda_{k,1} < \sigma_1^2,$$

yielding

$$\frac{\sigma_1^2}{\lambda_{k,1}} - 1 < \frac{\sigma_1^2}{\lambda_{k,2}} - 1 < \dots < \frac{\sigma_1^2}{\lambda_{k,k}} - 1,$$

we obtain

$$r_k^2 (\sigma_1^2) \sigma_1^{2\beta+2} > (\theta_1 - 1)^{2k} \sigma_1^{2\beta+2},$$

where  $\theta_1 = \sigma_1^2 / \lambda_{k,1}$ . Then, condition (F.18) gives

$$(\theta_1 - 1)^{2k} < (n-k) \left( \frac{\sigma_{k+1}^2}{\sigma_1^2} \right)^{\beta+1},$$

and since by assumption,  $\sigma_1^2 \gg \sigma_{k+1}^2$ , we deduce that  $\theta_1 \approx 1$ , that is,  $\lambda_{k,1} \approx \sigma_1^2$ . For  $i = 2$ , we proceed analogously; we write

$$r_k (\sigma_2^2) = \left(1 - \frac{\sigma_2^2}{\lambda_{k,1}}\right) \left(1 - \frac{\sigma_2^2}{\lambda_{k,2}}\right) \dots \left(1 - \frac{\sigma_2^2}{\lambda_{k,k}}\right) = \varepsilon_1 \left(1 - \frac{\sigma_2^2}{\lambda_{k,2}}\right) \dots \left(1 - \frac{\sigma_2^2}{\lambda_{k,k}}\right),$$

with  $\varepsilon_1 = 1 - \sigma_2^2 / \lambda_{k,1} \approx 1 - q_1 \approx 1$ , and use the inequalities

$$\lambda_{k,k} < \lambda_{k,k-1} < \dots < \lambda_{k,2} < \sigma_2^2$$

to conclude that

$$r_k^2 (\sigma_2^2) \sigma_2^{2\beta+2} > \varepsilon_1^2 (\theta_2 - 1)^{2(k-1)} \sigma_2^{2\beta+2},$$

where  $\theta_2 = \sigma_2^2 / \lambda_{k,2}$ . As before, condition (F.18) gives

$$(\theta_2 - 1)^{2(k-1)} < \frac{n-k}{\varepsilon_1^2} \left( \frac{\sigma_{k+1}^2}{\sigma_2^2} \right)^{\beta+1},$$

and we infer that  $\lambda_{k,2} \approx \sigma_2^2$ . Repeating these arguments for all  $i \leq k$ , we conclude that under assumptions (F.15) and (F.16), we have  $\lambda_{k,j} \approx \sigma_j^2$  for all  $j = 1, \dots, k$ .

# G

## A general direct regularization method for nonlinear problems

A general regularization method for solving ill-posed problems given by the nonlinear equation

$$\mathbf{F}(\mathbf{x}) = \mathbf{y}^\delta, \quad (\text{G.1})$$

has been proposed by Tautenhahn (1997). In this appendix, we particularize Tautenhahn's analysis to a discrete setting and for the choice  $\mathbf{L} = \mathbf{I}_n$ . The method is based on the iteration

$$\mathbf{x}_{\alpha k+1}^\delta = \mathbf{x}_a + g_\alpha (\mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k}) \mathbf{K}_{\alpha k}^T \mathbf{y}_k^\delta, \quad k = 0, 1, \dots, \quad (\text{G.2})$$

with  $\mathbf{K}_{\alpha k} = \mathbf{K}(\mathbf{x}_{\alpha k}^\delta)$ ,  $\mathbf{x}_0^\delta = \mathbf{x}_a$ ,

$$\mathbf{y}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{\alpha k}^\delta) + \mathbf{K}_{\alpha k}(\mathbf{x}_{\alpha k}^\delta - \mathbf{x}_a),$$

and

$$g_\alpha (\mathbf{K}_{\alpha k}^T \mathbf{K}_{\alpha k}) = \mathbf{V} \left[ \text{diag} (g_\alpha (\sigma_i^2))_{n \times n} \right] \mathbf{V}^T \quad (\text{G.3})$$

for  $\mathbf{K}_{\alpha k} = \mathbf{U} \Sigma \mathbf{V}^T$ .

If for any  $\alpha$  this iteration method converges, then the limit  $\mathbf{x}_\alpha^\delta$  solves the equation

$$\mathbf{x} = \mathbf{x}_a + g_\alpha \left( \mathbf{K}(\mathbf{x})^T \mathbf{K}(\mathbf{x}) \right) \mathbf{K}(\mathbf{x})^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}) + \mathbf{K}(\mathbf{x})(\mathbf{x} - \mathbf{x}_a)]. \quad (\text{G.4})$$

For linear problems,  $\mathbf{x}_\alpha^\delta$  is given by

$$\mathbf{x}_\alpha^\delta = \mathbf{x}_a + g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T (\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_a), \quad (\text{G.5})$$

and (G.5) is the general regularization method discussed in Appendix C.

As in the linear case, we suppose that the iteration function  $g_\alpha$  satisfies the conditions

$$0 \leq g_\alpha (\lambda) \leq \frac{1}{\alpha}, \quad (\text{G.6})$$

$$0 \leq 1 - \lambda g_\alpha (\lambda) \leq \alpha g_\alpha (\lambda), \quad (\text{G.7})$$

$$0 \leq \lambda^\mu [1 - \lambda g_\alpha (\lambda)] \leq c_2 \alpha^\mu, \quad 0 < \mu \leq \mu_0, \quad (\text{G.8})$$

for all  $\alpha > 0$ ,  $\lambda \in [0, \sigma_{\max}^2]$  and  $c_2 > 0$ . The index  $\mu_0$  is the qualification of the regularization method and  $\sigma_{\max}^2$  is a bound for  $\|\mathbf{K}(\mathbf{x})^T \mathbf{K}(\mathbf{x})\|$  in a ball  $B_\rho(\mathbf{x}^\dagger)$  of radius  $\rho$  about  $\mathbf{x}^\dagger$ . Here,  $\mathbf{x}^\dagger$  is a solution of the nonlinear equation with exact data  $\mathbf{F}(\mathbf{x}) = \mathbf{y}$ . The iteration function  $g_\alpha(\lambda)$  is continuously extended at  $\lambda = 0$  by defining  $g_\alpha(0) = \lim_{\lambda \rightarrow 0} g_\alpha(\lambda)$ .

In particular,  $g_\alpha$  may correspond to Tikhonov regularization,

$$g_\alpha(\lambda) = \frac{1}{\lambda + \alpha}, \quad \mu_0 = 1, \quad (\text{G.9})$$

the method of asymptotic regularization,

$$g_\alpha(\lambda) = \frac{1}{\lambda} \left(1 - e^{-\frac{\lambda}{\alpha}}\right), \quad \mu_0 = \infty, \quad (\text{G.10})$$

and the Landweber iteration,

$$g_\alpha(\lambda) = \frac{1}{\lambda} [1 - (1 - \lambda)^p], \quad \mu_0 = \infty, \quad \alpha = \frac{1}{p}. \quad (\text{G.11})$$

The approach with the iteration function (G.10) is the exponential Euler regularization method discussed in Chapter 7; in this case, assumption (G.8) holds for  $\mu > 0$  with  $c_2 = \mu^\mu e^{-\mu}$ . The approach with the iteration function (G.11) solves at each Newton step  $k$  the linearized equation

$$\mathbf{K}_{\alpha k} \Delta \mathbf{x} = \mathbf{y}_k^\delta, \quad (\text{G.12})$$

by using the Landweber iteration with zero initial guess, that is,

$$\begin{aligned} \Delta \mathbf{x}_{\alpha k 0}^\delta &= \mathbf{0}, \\ \Delta \mathbf{x}_{\alpha k l}^\delta &= \Delta \mathbf{x}_{\alpha k l-1}^\delta + \mathbf{K}_{\alpha k}^T (\mathbf{y}_k^\delta - \mathbf{K}_{\alpha k} \Delta \mathbf{x}_{\alpha k l-1}^\delta), \quad 1 \leq l \leq p, \\ \mathbf{x}_{\alpha k+1}^\delta &= \mathbf{x}_a + \Delta \mathbf{x}_{\alpha k p}^\delta. \end{aligned} \quad (\text{G.13})$$

It should be pointed out that for the method of Tikhonov regularization, we have

$$g_\alpha(\mathbf{K}_\alpha^T \mathbf{K}_\alpha) = (\mathbf{K}_\alpha^T \mathbf{K}_\alpha + \alpha \mathbf{I}_n)^{-1},$$

with  $\mathbf{K}_\alpha = \mathbf{K}(\mathbf{x}_\alpha^\delta)$ , and equation (G.4) represents the stationary condition for the Tikhonov function, or the so-called Euler equation.

## G.1 Error estimates

To derive a bound for the solution error  $\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|$  we first prove two auxiliary results.

**Proposition G.1.** *Let  $\mathbf{x}_\alpha^\delta$  be given by (G.5) and let assumptions (G.6) and (G.7) hold. Then, for all  $\mathbf{x} \in \mathbb{R}^n$ , we have*

$$\begin{aligned} &\|\mathbf{y}^\delta - \mathbf{K} \mathbf{x}_\alpha^\delta\|^2 + \alpha \|\mathbf{x}_\alpha^\delta - \mathbf{x}\|^2 \\ &\leq \|\mathbf{y}^\delta - \mathbf{K} \mathbf{x}\|^2 + \alpha (\mathbf{x} - \mathbf{x}_a)^T [\mathbf{I}_n - g_\alpha(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] (\mathbf{x} - \mathbf{x}_a). \end{aligned} \quad (\text{G.14})$$

*Proof.* Using the expression of  $\mathbf{x}_\alpha^\delta$  given by (G.5), and setting  $\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_a$  and  $\Delta \mathbf{y}^\delta = \mathbf{y}^\delta - \mathbf{K} \mathbf{x}_a$ , we have to show that

$$\begin{aligned} & \left\| [\mathbf{I}_m - \mathbf{K} g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T] \Delta \mathbf{y}^\delta \right\|^2 + \alpha \left\| g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \Delta \mathbf{y}^\delta - \Delta \mathbf{x} \right\|^2 \\ & \leq \left\| \Delta \mathbf{y}^\delta - \mathbf{K} \Delta \mathbf{x} \right\|^2 + \alpha \Delta \mathbf{x}^T [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] \Delta \mathbf{x}. \end{aligned} \quad (\text{G.15})$$

If  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of the matrix  $\mathbf{K}$ , we use (G.3) to obtain

$$\begin{aligned} \left\| [\mathbf{I}_m - \mathbf{K} g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T] \Delta \mathbf{y}^\delta \right\|^2 &= \sum_{i=1}^n [1 - \sigma_i^2 g_\alpha (\sigma_i^2)]^2 (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2 \\ &\quad + \sum_{i=n+1}^m (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2, \\ \left\| g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \Delta \mathbf{y}^\delta - \Delta \mathbf{x} \right\|^2 &= \sum_{i=1}^n [\sigma_i g_\alpha (\sigma_i^2) \mathbf{u}_i^T \Delta \mathbf{y}^\delta - \mathbf{v}_i^T \Delta \mathbf{x}]^2, \end{aligned}$$

and

$$\begin{aligned} \left\| \Delta \mathbf{y}^\delta - \mathbf{K} \Delta \mathbf{x} \right\|^2 &= \sum_{i=1}^n (\sigma_i \mathbf{v}_i^T \Delta \mathbf{x} - \mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2, \\ \Delta \mathbf{x}^T [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] \Delta \mathbf{x} &= \sum_{i=1}^n [1 - \sigma_i^2 g_\alpha (\sigma_i^2)] (\mathbf{v}_i^T \Delta \mathbf{x})^2. \end{aligned}$$

Inserting the above relations into (G.15) and rearranging the terms, we are led to the inequalities

$$\begin{aligned} & [1 - \sigma_i^2 g_\alpha (\sigma_i^2)]^2 (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2 + \alpha \sigma_i^2 g_\alpha^2 (\sigma_i^2) (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2 \\ & \leq \alpha g_\alpha (\sigma_i^2) (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2 + [1 - \alpha g_\alpha (\sigma_i^2)] (\sigma_i \mathbf{v}_i^T \Delta \mathbf{x} - \mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2, \quad i = 1, \dots, n, \end{aligned} \quad (\text{G.16})$$

which we must prove to be true. The last term in the right-hand side of (G.16) is positive due to assumption (G.6). By (G.7), we have

$$[1 - \sigma_i^2 g_\alpha (\sigma_i^2)]^2 \leq [1 - \sigma_i^2 g_\alpha (\sigma_i^2)] \alpha g_\alpha (\sigma_i^2)$$

and the left-hand side of (G.16) can be bounded as

$$[1 - \sigma_i^2 g_\alpha (\sigma_i^2)]^2 (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2 + \alpha \sigma_i^2 g_\alpha^2 (\sigma_i^2) (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2 \leq \alpha g_\alpha (\sigma_i^2) (\mathbf{u}_i^T \Delta \mathbf{y}^\delta)^2$$

for  $i = 1, \dots, n$ . Thus, (G.16) is satisfied and the proof is finished.  $\square$

Let us define the matrix  $\mathbf{R}_\alpha$  by the relation

$$\mathbf{R}_\alpha = \mathbf{I}_n - g_\alpha (\mathbf{K}_\alpha^T \mathbf{K}_\alpha) \mathbf{K}_\alpha^T \mathbf{K}_\alpha. \quad (\text{G.17})$$

For  $\mathbf{K}_\alpha = \mathbf{U}\Sigma\mathbf{V}^T$ , we have

$$\mathbf{R}_\alpha = \mathbf{V} \left[ \text{diag} \left( 1 - \sigma_i^2 g \left( \sigma_i^2 \right) \right)_{n \times n} \right] \mathbf{V}^T, \quad (\text{G.18})$$

and from assumptions (G.6) and (G.7), we see that  $\|\mathbf{R}_\alpha\| \leq 1$ . The matrix  $\mathbf{R}_\alpha$  can be expressed in terms of the residual function  $r_\alpha(\lambda) = 1 - \lambda g_\alpha(\lambda)$  as (cf. (E.32))  $\mathbf{R}_\alpha = r_\alpha(\mathbf{K}_\alpha^T \mathbf{K}_\alpha)$ , and for this reason,  $\mathbf{R}_\alpha$  is also known as the residual matrix.

**Proposition G.2.** *Let  $\mathbf{x}_\alpha^\delta$  be a solution of equation (G.4) and let assumptions (G.6) and (G.7) hold. Then, for all  $\mathbf{x} \in \mathbb{R}^n$ , we have*

$$\begin{aligned} & \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 + \alpha \|\mathbf{x}_\alpha^\delta - \mathbf{x}\|^2 \\ & \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x} - \mathbf{x}_\alpha^\delta)\|^2 + \alpha(\mathbf{x} - \mathbf{x}_a)^T \mathbf{R}_\alpha(\mathbf{x} - \mathbf{x}_a). \end{aligned} \quad (\text{G.19})$$

*Proof.* In (G.4) we put

$$\Delta \mathbf{y}^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) + \mathbf{K}_\alpha \mathbf{x}_\alpha^\delta,$$

and observe that  $\mathbf{x}_\alpha^\delta$  is as in (G.5) with  $\Delta \mathbf{y}^\delta$  in place of  $\mathbf{y}^\delta$  and  $\mathbf{K}_\alpha$  in place of  $\mathbf{K}$ . We now apply Proposition G.1 and use the results

$$\|\Delta \mathbf{y}^\delta - \mathbf{K} \mathbf{x}_\alpha^\delta\|^2 = \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2$$

and

$$\|\Delta \mathbf{y}^\delta - \mathbf{K} \mathbf{x}\|^2 = \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x} - \mathbf{x}_\alpha^\delta)\|^2$$

to conclude.  $\square$

Next, we introduce the following local property of  $\mathbf{F}$ :

$$\|\mathbf{F}(\mathbf{x}^\dagger) - \mathbf{F}(\mathbf{x}) - \mathbf{K}(\mathbf{x})(\mathbf{x}^\dagger - \mathbf{x})\| \leq \eta \|\mathbf{F}(\mathbf{x}^\dagger) - \mathbf{F}(\mathbf{x})\|, \quad 0 < \eta < 1, \quad (\text{G.20})$$

for all  $\mathbf{x} \in B_\rho(\mathbf{x}^\dagger)$ . This condition is a restriction on the nonlinearity of  $\mathbf{F}$ , and by the triangle inequality, we have

$$\|\mathbf{K}(\mathbf{x})(\mathbf{x}^\dagger - \mathbf{x})\| \leq (1 + \eta) \|\mathbf{F}(\mathbf{x}^\dagger) - \mathbf{F}(\mathbf{x})\|, \quad \mathbf{x} \in B_\rho(\mathbf{x}^\dagger).$$

A bound for the solution error is stated by the following result.

**Proposition G.3.** *Let  $\mathbf{x}_\alpha^\delta \in B_\rho(\mathbf{x}^\dagger)$  be a solution of equation (G.4) and let assumptions (G.6), (G.7) and (G.20) hold. Then we have*

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2 \leq c_n^2 \frac{\Delta^2}{\alpha} + (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_a), \quad (\text{G.21})$$

with  $c_n > 0$ .

*Proof.* With  $\mathbf{F}(\mathbf{x}^\dagger) = \mathbf{y}$  and  $\mathbf{x} = \mathbf{x}_\alpha^\delta$ , the nonlinearity assumption (G.20) reads as

$$\|\mathbf{y} - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\| \leq \eta \|\mathbf{y} - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|.$$



The ‘linearization error’ at  $\mathbf{x}_\alpha^\delta$  can then be estimated as

$$\begin{aligned} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\| &\leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\| + \Delta \\ &\leq \eta \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| + (1 + \eta) \Delta, \end{aligned} \quad (\text{G.22})$$

and we find that

$$\begin{aligned} &\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\|^2 - \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 \\ &\leq (\eta^2 - 1) \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 + 2\eta(1 + \eta) \Delta \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| + (1 + \eta)^2 \Delta^2. \end{aligned} \quad (\text{G.23})$$

The inequality

$$2ab \leq a^2 + b^2,$$

with

$$a = \sqrt{1 - \eta^2} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|, \quad b = \frac{\eta(1 + \eta) \Delta}{\sqrt{1 - \eta^2}}, \quad 0 < \eta < 1,$$

yields

$$2\eta(1 + \eta) \Delta \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| \leq (1 - \eta^2) \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 + \eta^2 \frac{1 + \eta}{1 - \eta} \Delta^2,$$

and (G.23) becomes

$$\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\|^2 - \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 \leq \frac{1 + \eta}{1 - \eta} \Delta^2. \quad (\text{G.24})$$

From (G.19) with  $\mathbf{x} = \mathbf{x}^\dagger$  and (G.24), we have

$$\alpha \|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2 \leq \frac{1 + \eta}{1 - \eta} \Delta^2 + \alpha (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a),$$

and we conclude that (G.21) holds with

$$c_n = \sqrt{\frac{1 + \eta}{1 - \eta}}.$$

□

## G.2 A priori parameter choice method

To derive convergence rate results, we impose a source condition which is similar to (C.16): for all  $\mathbf{x} \in B_\rho(\mathbf{x}^\dagger)$ , there holds

$$\mathbf{x}^\dagger - \mathbf{x}_a = \left[ \mathbf{K}(\mathbf{x})^T \mathbf{K}(\mathbf{x}) \right]^\mu \mathbf{z}, \quad (\text{G.25})$$

with  $\mu > 0$  and  $\mathbf{z} \in \mathbb{R}^n$ . This condition can be interpreted as an abstract smoothness condition for the difference  $\mathbf{x}^\dagger - \mathbf{x}_a$ , where the smoothing properties of  $\mathbf{K}(\mathbf{x})^T \mathbf{K}(\mathbf{x})$  should be ‘uniform’ in some sense and do not change very much when  $\mathbf{x}$  varies in a small ball around the exact solution. Actually, we will use the source condition (G.25) for  $\mathbf{x} = \mathbf{x}_\alpha^\delta$ , and this representation is justified if  $\mathbf{x}_\alpha^\delta$  is not too far from  $\mathbf{x}^\dagger$ .

**Theorem G.4.** Let  $\mathbf{x}_\alpha^\delta \in B_\rho(\mathbf{x}^\dagger)$  be a solution of equation (G.4) and let assumptions (G.6), (G.7), (G.8), (G.20) and (G.25) hold. Then, for the a priori parameter choice method

$$\alpha = \left( \frac{\Delta}{\|\mathbf{z}\|} \right)^{\frac{2}{2\mu+1}}, \quad (\text{G.26})$$

we have the error estimate

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \frac{\mu_0}{2}. \quad (\text{G.27})$$

*Proof.* We start by evaluating the term

$$(\mathbf{x}^\dagger - \mathbf{x}_\alpha)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha) = \sum_{i=1}^n [1 - \sigma_i^2 g_\alpha(\sigma_i^2)] [\mathbf{v}_i^T (\mathbf{x}^\dagger - \mathbf{x}_\alpha)]^2. \quad (\text{G.28})$$

The source condition (G.25), written as

$$\mathbf{x}^\dagger - \mathbf{x}_\alpha = (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu \mathbf{z} = \sum_{j=1}^n \sigma_j^{2\mu} (\mathbf{v}_j^T \mathbf{z}) \mathbf{v}_j,$$

together with the orthogonality relation  $\mathbf{v}_i^T \mathbf{v}_j = \delta_{ij}$  and assumption (G.8), gives

$$(\mathbf{x}^\dagger - \mathbf{x}_\alpha)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha) = \sum_{i=1}^n \sigma_i^{4\mu} [1 - \sigma_i^2 g_\alpha(\sigma_i^2)] (\mathbf{v}_i^T \mathbf{z})^2 \leq c_2 \alpha^{2\mu} \|\mathbf{z}\|^2 \quad (\text{G.29})$$

for  $0 < \mu \leq \mu_0/2$ . Inserting this estimate into (G.21) yields

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2 \leq c_n^2 \frac{\Delta^2}{\alpha} + c_2 \alpha^{2\mu} \|\mathbf{z}\|^2,$$

and, by (G.26), the conclusion readily follows.  $\square$

### G.3 Discrepancy principle

A simplified version of the discrepancy principle is used in the present analysis. The residual norm at the solution is captured by a lower and an upper bound depending on the noise level, that is, for  $\varepsilon > 0$ , we assume that there exists at least one regularization parameter  $\alpha > 0$  so that

$$\tau_{\text{dp}} \Delta \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| \leq (\tau_{\text{dp}} + \varepsilon) \Delta. \quad (\text{G.30})$$

Before proceeding, we recall some matrix identities which we used in Appendix E in a slightly different form. For  $\mathbf{K} = \mathbf{U}\Sigma\mathbf{V}^T$ , we have, analogous to

$$g_\alpha(\mathbf{K}^T \mathbf{K}) = \mathbf{V} \left[ \text{diag}(g_\alpha(\sigma_i^2))_{n \times n} \right] \mathbf{V}^T,$$

the representation

$$g_\alpha (\mathbf{K}\mathbf{K}^T) = \mathbf{U} \left[ \text{diag} (g_\alpha (\sigma_i^2))_{m \times m} \right] \mathbf{U}^T, \quad (\text{G.31})$$

with the convention  $g_\alpha (\sigma_i^2) = g_\alpha (0) = \lim_{\lambda \rightarrow 0} g_\alpha (\lambda)$  for  $i = n + 1, \dots, m$ . Then, we find that (see (E.23))

$$\mathbf{K} g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T = \mathbf{U} \begin{bmatrix} \text{diag} (\sigma_i^2 g_\alpha (\sigma_i^2))_{n \times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T \quad (\text{G.32})$$

and that

$$g_\alpha (\mathbf{K}\mathbf{K}^T) \mathbf{K}\mathbf{K}^T = \mathbf{U} \begin{bmatrix} \text{diag} (\sigma_i^2 g_\alpha (\sigma_i^2))_{n \times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{U}^T. \quad (\text{G.33})$$

From (G.32) and (G.33), we obtain

$$\mathbf{K} g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T = g_\alpha (\mathbf{K}\mathbf{K}^T) \mathbf{K}\mathbf{K}^T, \quad (\text{G.34})$$

which then yields

$$\mathbf{K} [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] = [\mathbf{I}_m - g_\alpha (\mathbf{K}\mathbf{K}^T) \mathbf{K}\mathbf{K}^T] \mathbf{K}. \quad (\text{G.35})$$

Using the representations

$$\mathbf{K} [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] \mathbf{x} = \sum_{i=1}^n \sigma_i [1 - \sigma_i^2 g (\sigma_i^2)] (\mathbf{v}_i^T \mathbf{x}) \mathbf{u}_i$$

and

$$(\mathbf{K}^T \mathbf{K})^{\frac{1}{2}} [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] \mathbf{x} = \sum_{i=1}^n \sigma_i [1 - \sigma_i^2 g_\alpha (\sigma_i^2)] (\mathbf{v}_i^T \mathbf{x}) \mathbf{v}_i,$$

we deduce that

$$\|\mathbf{K} [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] \mathbf{x}\| = \|(\mathbf{K}^T \mathbf{K})^{\frac{1}{2}} [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] \mathbf{x}\|; \quad (\text{G.36})$$

this together with (G.35) then gives

$$\|[\mathbf{I}_m - g_\alpha (\mathbf{K}\mathbf{K}^T) \mathbf{K}\mathbf{K}^T] \mathbf{K}\mathbf{x}\| = \|(\mathbf{K}^T \mathbf{K})^{\frac{1}{2}} [\mathbf{I}_n - g_\alpha (\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}] \mathbf{x}\|. \quad (\text{G.37})$$

The following moment inequality, which is a consequence of the Hölder inequality, will be frequently used in the sequel.

**Proposition G.5.** *Let  $\mathbf{A} \in \mathbb{R}^{n \times n}$  be a positive definite matrix. Then there holds the moment inequality*

$$\|\mathbf{A}^r \mathbf{x}\| \leq \|\mathbf{A}^s \mathbf{x}\|^{\frac{r}{s}} \|\mathbf{x}\|^{1 - \frac{r}{s}}, \quad 0 \leq r \leq s. \quad (\text{G.38})$$

*Proof.* For  $r = s$  we have equality and we consider the case  $r < s$ . If  $\mathbf{A} = \mathbf{V}\Sigma\mathbf{V}^T$  is a singular value decomposition of the positive definite matrix  $\mathbf{A}$ , we have

$$\mathbf{A}^r \mathbf{x} = \sum_{i=1}^n \sigma_i^r (\mathbf{v}_i^T \mathbf{x}) \mathbf{v}_i,$$

and therefore,

$$\|\mathbf{A}^r \mathbf{x}\|^2 = \sum_{i=1}^n \sigma_i^{2r} (\mathbf{v}_i^T \mathbf{x})^2. \quad (\text{G.39})$$

Similarly, we have

$$\|\mathbf{A}^s \mathbf{x}\|^{\frac{2r}{s}} = \left[ \sum_{i=1}^n \sigma_i^{2s} (\mathbf{v}_i^T \mathbf{x})^2 \right]^{\frac{r}{s}}, \quad (\text{G.40})$$

and, from  $\mathbf{x} = \sum_{i=1}^n (\mathbf{v}_i^T \mathbf{x}) \mathbf{v}_i$ , there holds

$$\|\mathbf{x}\|^{2(1-\frac{r}{s})} = \left[ \sum_{i=1}^n (\mathbf{v}_i^T \mathbf{x})^2 \right]^{1-\frac{r}{s}}. \quad (\text{G.41})$$

We consider now the Hölder inequality

$$\sum_{i=1}^n a_i b_i \leq \left( \sum_{i=1}^n a_i^p \right)^{\frac{1}{p}} \left( \sum_{i=1}^n b_i^q \right)^{\frac{1}{q}}, \quad \frac{1}{p} + \frac{1}{q} = 1, \quad a_i, b_i \geq 0,$$

with

$$p = \frac{s}{r}, \quad q = \frac{s}{s-r},$$

and

$$a_i = \sigma_i^{2r} (\mathbf{v}_i^T \mathbf{x})^{\frac{2r}{s}}, \quad b_i = (\mathbf{v}_i^T \mathbf{x})^{\frac{2(s-r)}{s}}.$$

Since

$$a_i b_i = \sigma_i^{2r} (\mathbf{v}_i^T \mathbf{x})^2, \quad a_i^p = \sigma_i^{2s} (\mathbf{v}_i^T \mathbf{x})^2, \quad b_i^q = (\mathbf{v}_i^T \mathbf{x})^2,$$

we obtain

$$\sum_{i=1}^n \sigma_i^{2r} (\mathbf{v}_i^T \mathbf{x})^2 \leq \left[ \sum_{i=1}^n \sigma_i^{2s} (\mathbf{v}_i^T \mathbf{x})^2 \right]^{\frac{r}{s}} \left[ \sum_{i=1}^n (\mathbf{v}_i^T \mathbf{x})^2 \right]^{1-\frac{r}{s}},$$

and, by (G.39)–(G.41), we see that (G.38) holds.  $\square$

**Theorem G.6.** *Let the assumptions of Theorem G.4 hold. Then, if we select the regularization parameter from the discrepancy principle (G.30) with*

$$\tau_{\text{dp}} > \frac{1 + \eta + \eta\varepsilon}{1 - \eta}, \quad (\text{G.42})$$

*we have the error estimate*

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \min\left(\frac{1}{2}, \mu_0 - \frac{1}{2}\right). \quad (\text{G.43})$$

*Proof.* The proof relies on the error bound (G.21), written for convenience as

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2 \leq e_s^2 + e_n^2, \quad (\text{G.44})$$

with

$$e_s^2 = (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a)$$

and

$$e_n^2 = c_n^2 \frac{\Delta^2}{\alpha}, \quad c_n > 0.$$

The quantities  $e_s$  and  $e_n$  can be interpreted as bounds for the smoothing and noise errors, respectively. As in the linear case, we estimate  $e_s$  and  $e_n$  separately.

(a). To estimate  $e_s$ , we first consider the source condition (G.25) for  $\mathbf{x} = \mathbf{x}_\alpha^\delta$  and exploit the symmetry of the matrix  $(\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu$  to obtain

$$\begin{aligned} (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) &= \mathbf{z}^T (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \\ &\leq \left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\| \|\mathbf{z}\|. \end{aligned} \quad (\text{G.45})$$

Assuming  $0 < \mu \leq 1/2$  and applying the moment inequality (G.38) with

$$r = 2\mu, \quad s = 1,$$

and

$$\mathbf{A} = (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}}, \quad \mathbf{x} = \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a),$$

gives

$$\left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\| \leq \left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^{2\mu} \left\| \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^{1-2\mu},$$

whence, (G.45) becomes

$$(\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \leq \|\mathbf{z}\| \left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^{2\mu} \left\| \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^{1-2\mu}. \quad (\text{G.46})$$

On the other hand, the condition (cf. (G.6) and (G.7))

$$0 \leq 1 - \lambda g_\alpha(\lambda) \leq 1, \quad (\text{G.47})$$

together with (G.18) and (G.28) yields

$$\begin{aligned} \left\| \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^2 &= \sum_{i=1}^n [1 - \sigma_i^2 g_\alpha(\sigma_i^2)]^2 [\mathbf{v}_i^T (\mathbf{x}^\dagger - \mathbf{x}_a)]^2 \\ &\leq \sum_{i=1}^n [1 - \sigma_i^2 g_\alpha(\sigma_i^2)] [\mathbf{v}_i^T (\mathbf{x}^\dagger - \mathbf{x}_a)]^2 \\ &= (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a). \end{aligned} \quad (\text{G.48})$$

Combining (G.46) and (G.48), we obtain

$$\|R_\alpha(\mathbf{x}^\dagger - \mathbf{x}_a)\| \leq \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} R_\alpha(\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^{\frac{2\mu}{2\mu+1}},$$

and inserting this result back into (G.46), we find a first estimate for  $e_s$ ,

$$e_s^2 \leq \|\mathbf{z}\|^{\frac{2}{2\mu+1}} \left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} R_\alpha(\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^{\frac{4\mu}{2\mu+1}}. \quad (\text{G.49})$$

To express this estimate in terms of the noise level  $\Delta$  we proceed to derive a bound for  $\left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{1/2} R_\alpha(\mathbf{x}^\dagger - \mathbf{x}_a) \right\|$ . For this purpose, we consider the Euler equation (G.4),

$$\mathbf{x}_\alpha^\delta = \mathbf{x}_a + g_\alpha (\mathbf{K}_\alpha^T \mathbf{K}_\alpha) \mathbf{K}_\alpha^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) + \mathbf{K}_\alpha(\mathbf{x}_\alpha^\delta - \mathbf{x}_a)].$$

Multiplying this equation by  $\mathbf{K}_\alpha$  and using (G.34), gives

$$\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) = [\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T] [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) + \mathbf{K}_\alpha(\mathbf{x}_\alpha^\delta - \mathbf{x}_a)], \quad (\text{G.50})$$

and further

$$\begin{aligned} & [\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T] \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \\ &= \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - [\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T] [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)]. \end{aligned} \quad (\text{G.51})$$

By (G.22) and (G.30), the last factor in the right-hand side of (G.51) can be bounded as

$$\begin{aligned} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\| &\leq \eta \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| + (1 + \eta) \Delta \\ &\leq [\eta (\tau_{\text{dp}} + \varepsilon) + 1 + \eta] \Delta. \end{aligned} \quad (\text{G.52})$$

This result together with the matrix norm equality (cf. (G.33) and (G.47))

$$\|\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T\| = 1, \quad (\text{G.53})$$

leads to the following estimate for the left-hand side of (G.51) (cf. (G.30)),

$$\begin{aligned} & \|\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T\| \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \| \\ & \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| + \|\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T\| \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\| \\ & \leq (1 + \eta) (1 + \tau_{\text{dp}} + \varepsilon) \Delta. \end{aligned} \quad (\text{G.54})$$

By (G.37), (G.54) becomes

$$\left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} R_\alpha(\mathbf{x}^\dagger - \mathbf{x}_a) \right\| \leq (1 + \eta) (1 + \tau_{\text{dp}} + \varepsilon) \Delta, \quad (\text{G.55})$$

and (G.49) takes the form

$$e_s^2 \leq c_{\text{sdp}}^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}, \quad 0 < \mu \leq 1/2, \quad (\text{G.56})$$

with

$$c_{\text{sdp}} = \left[ (1 + \eta) (1 + \tau_{\text{dp}} + \varepsilon) \right]^{\frac{2\mu}{2\mu+1}}.$$

(b) To derive an estimate for  $e_n$ , we look at a lower bound for  $\alpha$ . Taking into account that

$$\left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^2 = \sum_{i=1}^n (\sigma_i^2)^{2\mu+1} [1 - \sigma_i^2 g(\sigma_i^2)]^2 (\mathbf{v}_i^T \mathbf{z})^2,$$

and that (cf. (G.8))

$$(\sigma_i^2)^{\mu+\frac{1}{2}} [1 - \sigma_i^2 g(\sigma_i^2)] \leq c_2 \alpha^{\mu+\frac{1}{2}}, \quad 0 < \mu \leq \mu_0 - \frac{1}{2},$$

we obtain

$$\left\| (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\|^2 \leq c_2^2 \alpha^{2\mu+1} \|\mathbf{z}\|^2, \quad (\text{G.57})$$

and further (cf. (G.37))

$$\left\| [\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T] \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\| \leq c_2 \alpha^{\mu+\frac{1}{2}} \|\mathbf{z}\|. \quad (\text{G.58})$$

Moreover, from (G.30), (G.51), (G.52) and (G.53), we have

$$\begin{aligned} \tau_{\text{dp}} \Delta &\leq \left\| \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) \right\| \leq \left\| [\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T] \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\| \\ &\quad + \left\| \mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T \right\| \\ &\quad \times \left\| \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a^\delta) \right\| \\ &\leq \left\| [\mathbf{I}_m - g_\alpha (\mathbf{K}_\alpha \mathbf{K}_\alpha^T) \mathbf{K}_\alpha \mathbf{K}_\alpha^T] \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \right\| \\ &\quad + [\eta (\tau_{\text{dp}} + \varepsilon) + 1 + \eta] \Delta, \end{aligned} \quad (\text{G.59})$$

and as a result, (G.42) and (G.58) yield

$$\alpha \geq \left[ \frac{\tau_{\text{dp}} (1 - \eta) - (1 + \eta + \eta \varepsilon)}{c_2} \right]^{\frac{2}{2\mu+1}} \left( \frac{\Delta}{\|\mathbf{z}\|} \right)^{\frac{2}{2\mu+1}}. \quad (\text{G.60})$$

Thus,

$$e_n^2 \leq c_{\text{ndp}}^2 \left( \|\mathbf{z}\|^2 \right)^{\frac{1}{2\mu+1}} (\Delta^2)^{\frac{2\mu}{2\mu+1}}, \quad 0 < \mu \leq \mu_0 - \frac{1}{2}, \quad (\text{G.61})$$

with

$$c_{\text{ndp}} = c_n \left[ \frac{c_2}{\tau_{\text{dp}} (1 - \eta) - (1 + \eta + \eta \varepsilon)} \right]^{\frac{1}{2\mu+1}}. \quad (\text{G.62})$$

The conclusion now follows from (G.56) and (G.61).  $\square$

To estimate  $e_n$  we used assumptions (G.8) with  $\mu \leq \mu_0 - 1/2$  and (G.42). In fact, we can avoid this computational step and disregard the condition  $\mu \leq \mu_0 - 1/2$  by assuming that

$$\tau_{\text{dp}} \geq \frac{1 + \eta}{1 - \eta}.$$

To see this, we use (G.19) with  $\mathbf{x} = \mathbf{x}^\dagger$  and (G.23) to obtain

$$\begin{aligned}
 & \|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2 \\
 & \leq (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) + \frac{1}{\alpha} \left[ \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\|^2 \right. \\
 & \quad \left. - \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 \right] \\
 & \leq (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) + \frac{1+\eta}{\alpha} \left[ (\eta-1) \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 \right. \\
 & \quad \left. + 2\eta\Delta \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| + (1+\eta)\Delta^2 \right].
 \end{aligned}$$

By the above assumption and the discrepancy principle condition (G.30) we have

$$\frac{1+\eta}{1-\eta}\Delta \leq \tau_{\text{dp}}\Delta \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|;$$

this yields

$$\begin{aligned}
 & (\eta-1) \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 + 2\eta\Delta \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\| + (1+\eta)\Delta^2 \\
 & \leq \left[ (\eta-1) + 2\eta\frac{1-\eta}{1+\eta} + \frac{(1-\eta)^2}{1+\eta} \right] \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta)\|^2 = 0,
 \end{aligned}$$

and we find that

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2 \leq (\mathbf{x}^\dagger - \mathbf{x}_a)^T \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a). \quad (\text{G.63})$$

Thus, to estimate the solution error  $\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|$ , we have only to evaluate  $e_s$  as in (G.56). In this regard, we can formulate the following result.

**Theorem G.7.** *Let the assumptions of Theorem G.4 excepting assumption (G.8) hold. Then, if we select the regularization parameter from the discrepancy principle (G.30) with*

$$\tau_{\text{dp}} \geq \frac{1+\eta}{1-\eta}, \quad (\text{G.64})$$

*we have the error estimate*

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \frac{1}{2}. \quad (\text{G.65})$$

The main drawback of the convergence rate (G.43) is that it suffers from a saturation effect: for regularization methods with  $\mu_0 > 1$ , (G.43) holds for  $\mu \leq 1/2$ , and a convergence rate better than  $O(\sqrt{\Delta})$  cannot be achieved. To eliminate this inconvenience, we suppose that the iteration function  $g_\alpha$  satisfies the additional condition

$$0 \leq \sqrt{\lambda} g_\alpha(\lambda) \leq \frac{c_0}{\sqrt{\alpha}} \quad (\text{G.66})$$

for all  $\alpha > 0$ ,  $\lambda \in [0, \sigma_{\text{max}}^2]$  and  $c_0 > 0$ . Assumption (G.66) holds for the method of Tikhonov regularization with  $c_0 = 1/2$ , and for the regularization methods (G.10) and (G.11) with  $c_0 = 1$ .



**Theorem G.8.** *Let the assumptions of Theorem G.4 together with assumption (G.66) hold. Then, if we select the regularization parameter from the discrepancy principle (G.30) with  $\tau_{\text{dp}}$  as in (G.42), we have the error estimate*

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \mu_0 - \frac{1}{2}. \quad (\text{G.67})$$

*Proof.* First, we proceed to derive another error bound than in (G.21). For this purpose, we express the Euler equation (G.4) as

$$\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger = \mathbf{R}_\alpha (\mathbf{x}_a - \mathbf{x}^\dagger) + g_\alpha (\mathbf{K}_\alpha^T \mathbf{K}_\alpha) \mathbf{K}_\alpha^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)]. \quad (\text{G.68})$$

For any  $\mathbf{w} \in \mathbb{R}^m$ , assumption (G.66) yields

$$\|g_\alpha (\mathbf{K}_\alpha^T \mathbf{K}_\alpha) \mathbf{K}_\alpha^T \mathbf{w}\|^2 = \sum_{i=1}^n \sigma_i^2 g_\alpha^2 (\sigma_i^2) (\mathbf{u}_i^T \mathbf{w})^2 \leq \frac{c_0^2}{\alpha} \|\mathbf{w}\|^2, \quad (\text{G.69})$$

and this result together with (G.52) and (G.68) gives

$$\begin{aligned} \|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\| &\leq \|\mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a)\| + \frac{c_0}{\sqrt{\alpha}} \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_\alpha^\delta) - \mathbf{K}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_\alpha^\delta)\| \\ &\leq \|\mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a)\| + c_{n1} \frac{\Delta}{\sqrt{\alpha}}, \end{aligned} \quad (\text{G.70})$$

with

$$c_{n1} = c_0 [\eta (\tau_{\text{dp}} + \varepsilon) + 1 + \eta].$$

Thus, the solution error can be bounded as

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\| \leq e_s + e_n, \quad (\text{G.71})$$

where

$$e_s = \|\mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a)\|, \quad e_n = c_{n1} \frac{\Delta}{\sqrt{\alpha}}.$$

(a) To estimate  $e_s$ , we use the symmetry relation

$$\mathbf{R}_\alpha (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu = (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu \mathbf{R}_\alpha$$

to obtain

$$\|\mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a)\| = \|(\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu \mathbf{R}_\alpha \mathbf{z}\|. \quad (\text{G.72})$$

By the moment inequality (G.38), with

$$r = \mu, \quad s = \mu + \frac{1}{2},$$

and

$$\mathbf{A} = \mathbf{K}_\alpha^T \mathbf{K}_\alpha, \quad \mathbf{x} = \mathbf{R}_\alpha \mathbf{z},$$

we find that

$$\|(\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^\mu \mathbf{R}_\alpha \mathbf{z}\| \leq \|(\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\mu+\frac{1}{2}} \mathbf{R}_\alpha \mathbf{z}\|^{\frac{2\mu}{2\mu+1}} \|\mathbf{R}_\alpha \mathbf{z}\|^{\frac{1}{2\mu+1}}. \quad (\text{G.73})$$

The estimate  $\|\mathbf{R}_\alpha\| \leq 1$  together with the identity

$$(\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\mu+\frac{1}{2}} \mathbf{R}_\alpha \mathbf{z} = (\mathbf{K}_\alpha^T \mathbf{K}_\alpha)^{\frac{1}{2}} \mathbf{R}_\alpha (\mathbf{x}^\dagger - \mathbf{x}_a) \quad (\text{G.74})$$

and the relations (G.55), (G.72) and (G.73) then yields

$$e_s \leq c_{\text{sdp}} \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}, \quad (\text{G.75})$$

with  $c_{\text{sdp}}$  as in (G.56).

(b) To estimate  $e_n$  we proceed as in Theorem G.6, and obtain

$$e_n \leq c_{\text{ndp}} \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}, \quad 0 < \mu \leq \mu_0 - \frac{1}{2}, \quad (\text{G.76})$$

with  $c_{\text{ndp}}$  depending now on  $c_{n1}$  instead of  $c_n$ . The desired error estimate follows then from (G.75) and (G.76).  $\square$

If instead of (G.42) we assume (G.64), we have the following theorem.

**Theorem G.9.** *Under the same assumptions as in Theorem G.8, if we select the regularization parameter from the discrepancy principle (G.30) with  $\tau_{\text{dp}}$  as in (G.64), then there holds the error estimate*

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\| = O\left(\|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \frac{\mu_0}{2}. \quad (\text{G.77})$$

*Proof.* We distinguish two cases. In the first case, we assume that

$$\alpha \leq \left(\frac{\Delta}{\|\mathbf{z}\|}\right)^{\frac{2}{2\mu+1}}. \quad (\text{G.78})$$

For the choice (G.64), the solution error is bounded as in (G.63). Then, as in the proof of Theorem G.4, assumption (G.8) yields the error estimate

$$\|\mathbf{x}_\alpha^\delta - \mathbf{x}^\dagger\|^2 \leq c_2 \alpha^{2\mu} \|\mathbf{z}\|^2, \quad 0 < \mu \leq \frac{\mu_0}{2},$$

and, by (G.78), the conclusion readily follows. In the second case, we suppose that

$$\alpha > \left(\frac{\Delta}{\|\mathbf{z}\|}\right)^{\frac{2}{2\mu+1}}. \quad (\text{G.79})$$

Assumption (G.66) gives the error bound (G.70). Then, for estimating  $e_s = \|\mathbf{R}_\alpha(\mathbf{x}^\dagger - \mathbf{x}_a)\|$ , we proceed as in Theorem G.8 and derive the bound (G.75), while for estimating  $e_n$ , we use (G.79) to obtain

$$e_n = c_{n1} \frac{\Delta}{\sqrt{\alpha}} < c_{n1} \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}.$$

$\square$

We conclude our analysis by mentioning that condition (G.6) is too sharp for a regularization method which uses as inner iteration the  $p$ -times iterated Tikhonov regularization. At each Newton step  $k$ , this approach applies the  $p$ -times iterated Tikhonov regularization (with fixed  $p$ ) to the linearized equation (G.12), that is,

$$\begin{aligned}\Delta \mathbf{x}_{\alpha k 0}^{\delta} &= \mathbf{0}, \\ \Delta \mathbf{x}_{\alpha k l}^{\delta} &= \Delta \mathbf{x}_{\alpha k l-1}^{\delta} + \mathbf{K}_{\alpha k}^{\dagger} (\mathbf{y}_k^{\delta} - \mathbf{K}_{\alpha k} \Delta \mathbf{x}_{\alpha k l-1}^{\delta}), \quad 1 \leq l \leq p, \\ \mathbf{x}_{\alpha k+1}^{\delta} &= \mathbf{x}_a + \Delta \mathbf{x}_{\alpha k p}^{\delta},\end{aligned}\tag{G.80}$$

in which case,

$$g_{\alpha}(\lambda) = \frac{1}{\lambda} \left[ 1 - \left( \frac{\alpha}{\lambda + \alpha} \right)^p \right], \quad \mu_0 = p,\tag{G.81}$$

and

$$0 \leq g_{\alpha}(\lambda) \leq \frac{p}{\alpha}.$$

However, the proof of Theorem G.8 reveals that assumption (G.6) is only used in conjunction with assumption (G.7) to derive the estimate

$$0 \leq 1 - \lambda g_{\alpha}(\lambda) \leq 1,\tag{G.82}$$

which then yields  $\|\mathbf{R}_{\alpha}\| \leq 1$ . Therefore, if instead of (G.6) and (G.7) we assume that (G.82) holds, and furthermore, if we take into account that, for  $g_{\alpha}$  as in (G.81), condition (G.66) is satisfied with  $c_0 = p$ , we deduce that a regularization method using as inner iteration the  $p$ -times iterated Tikhonov regularization is of optimal order for  $0 < \mu \leq \mu_0 - 1/2$ .

# H

## A general iterative regularization method for nonlinear problems

The iteratively regularized Gauss–Newton method belongs to the class of Newton-type methods with a priori information, in which case, the linearized equation is solved by means of Tikhonov regularization with a penalty term depending on the a priori. By contrast, the regularizing Levenberg–Marquardt method can be categorized as a Newton-type method without a priori information, since the penalty term depends on the previous iterate and not on the a priori. In this appendix we analyze both regularization methods in a general setting.

### H.1 Newton-type methods with a priori information

A regularization method accounting for a priori information uses the iteration

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_a + g_{\alpha_k} (\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T \mathbf{y}_k^\delta, \quad k = 0, 1, \dots,$$

where  $\mathbf{K}_k = \mathbf{K}(\mathbf{x}_k^\delta)$ ,  $\mathbf{x}_0^\delta = \mathbf{x}_a$ ,

$$\mathbf{y}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k(\mathbf{x}_k^\delta - \mathbf{x}_a),$$

and  $\{\alpha_k\}$  is a monotonically decreasing sequence satisfying the requirements

$$1 < \frac{\alpha_k}{\alpha_{k+1}} \leq c, \quad \alpha_k > 0. \tag{H.1}$$

The iteration function  $g_\alpha$  fulfills assumptions (G.8), (G.66) and (G.82). In particular, when

$$g_\alpha(\lambda) = \frac{1}{\lambda + \alpha},$$

we obtain the iteratively regularized Gauss–Newton method; otherwise,  $g_\alpha$  may correspond to iterated Tikhonov regularization with fixed order (cf. (G.80) and (G.81)) and the Landweber iteration (cf. (G.11) and (G.13)). To prove convergence rate results, we closely follow the studies of Deuffhard et al. (1998) and Kaltenbacher et al. (2008).

Our analysis will be carried out under the nonlinearity assumption

$$\|\mathbf{K}(\mathbf{x}) - \mathbf{K}(\mathbf{x}^\dagger)\| \leq c_K \|\mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\|, \quad \mathbf{x} \in B_\rho(\mathbf{x}^\dagger), \quad (\text{H.2})$$

where  $\mathbf{x}^\dagger$  is a solution of the nonlinear equation with exact data  $\mathbf{F}(\mathbf{x}) = \mathbf{y}$ . The linearization error can be estimated as

$$\begin{aligned} & \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger) - \mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\| \\ & \leq \int_0^1 \|\mathbf{K}(\mathbf{x}^\dagger + t(\mathbf{x} - \mathbf{x}^\dagger)) - \mathbf{K}(\mathbf{x}^\dagger)\|(\mathbf{x} - \mathbf{x}^\dagger)\| dt \\ & \leq \|\mathbf{x} - \mathbf{x}^\dagger\| \int_0^1 \|\mathbf{K}(\mathbf{x}^\dagger + t(\mathbf{x} - \mathbf{x}^\dagger)) - \mathbf{K}(\mathbf{x}^\dagger)\| dt \\ & \leq \frac{c_K}{2} \|\mathbf{x} - \mathbf{x}^\dagger\| \|\mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\|, \end{aligned} \quad (\text{H.3})$$

while application of the triangle inequality yields

$$\begin{aligned} & \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger) - \mathbf{K}(\mathbf{x})(\mathbf{x} - \mathbf{x}^\dagger)\| \\ & \leq \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger) - \mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\| + \|\mathbf{K}(\mathbf{x}^\dagger) - \mathbf{K}(\mathbf{x})\|(\mathbf{x} - \mathbf{x}^\dagger)\| \\ & \leq \frac{3c_K}{2} \|\mathbf{x} - \mathbf{x}^\dagger\| \|\mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\|. \end{aligned} \quad (\text{H.4})$$

Convergence rate results will be derived by assuming the source condition

$$\mathbf{x}^\dagger - \mathbf{x}_a = \left[ \mathbf{K}(\mathbf{x}^\dagger)^T \mathbf{K}(\mathbf{x}^\dagger) \right]^\mu \mathbf{z}, \quad (\text{H.5})$$

with  $\mu > 0$  and  $\mathbf{z} \in \mathbb{R}^n$ . Note that the Jacobian matrix in (H.5) is evaluated at  $\mathbf{x}^\dagger$ , while the Jacobian matrix in (G.25) is evaluated at  $\mathbf{x}$  and does not change very much when  $\mathbf{x}$  varies in a small ball around  $\mathbf{x}^\dagger$ .

The iteration error defined by

$$\mathbf{e}_{k+1}^\delta = \mathbf{x}_{k+1}^\delta - \mathbf{x}^\dagger$$

can be expressed as (compare to (G.68))

$$\mathbf{e}_{k+1}^\delta = \mathbf{R}_k(\mathbf{x}_a - \mathbf{x}^\dagger) + g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta] \quad (\text{H.6})$$

where the residual matrix  $\mathbf{R}_k$  is as in (G.17), i.e.,

$$\mathbf{R}_k = \mathbf{I}_n - g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T \mathbf{K}_k.$$

Setting  $\mathbf{K} = \mathbf{K}(\mathbf{x}^\dagger)$  and

$$\mathbf{R} = \mathbf{I}_n - g_{\alpha_k}(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T \mathbf{K}$$

we summarize some results of Appendix G, which will be used in the sequel.

(1) Analogously to (G.69), assumption (G.66) gives

$$\|g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T \mathbf{w}\| \leq \frac{c_0}{\sqrt{\alpha_k}} \|\mathbf{w}\| \quad (\text{H.7})$$

for all  $\mathbf{w} \in \mathbb{R}^m$ .

(2) Assumption (G.8) and the source condition (H.5) yield, for  $0 < \mu \leq \mu_0$ ,

$$\|\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger)\|^2 = \sum_{i=1}^n \sigma_i^{4\mu} [1 - \sigma_i^2 g_{\alpha_k}(\sigma_i^2)]^2 (\mathbf{v}_i^T \mathbf{z})^2 \leq c_2^2 \alpha_k^{2\mu} \|\mathbf{z}\|^2, \quad (\text{H.8})$$

where  $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$  is a singular system of  $\mathbf{K}$ .

(3) Assumption (G.8) and the source condition (H.5) give a relation similar to (G.57); this together with (G.36) implies that

$$\|\mathbf{K}\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger)\| \leq c_2 \alpha_k^{\mu + \frac{1}{2}} \|\mathbf{z}\|, \quad 0 < \mu \leq \mu_0 - \frac{1}{2}. \quad (\text{H.9})$$

(4) The matrix factorization (G.32) and assumption (G.82) give

$$\|\mathbf{K}_k g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T\| \leq 1. \quad (\text{H.10})$$

(5) The source condition (H.5) in conjunction with the relations (G.36), (G.72), (G.73), (G.74), and the estimate (cf. (G.82))  $\|\mathbf{R}\| \leq 1$  yields

$$\|\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger)\| \leq \|\mathbf{K}\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger)\|^{\frac{2\mu}{2\mu+1}} \|\mathbf{z}\|^{\frac{1}{2\mu+1}}. \quad (\text{H.11})$$

Replacing the source condition (G.25) by (H.5) requires further assumptions on  $g_\alpha$ , namely

$$\|\mathbf{R}_1 - \mathbf{R}_2\| \leq \frac{c_R}{\sqrt{\alpha}} \|\mathbf{K}_1 - \mathbf{K}_2\| \quad (\text{H.12})$$

and

$$\|\mathbf{K}_2(\mathbf{R}_1 - \mathbf{R}_2)\| \leq c_R \|\mathbf{K}_1 - \mathbf{K}_2\| \quad (\text{H.13})$$

for all  $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{m \times n}$ , and  $\mathbf{R}_i = \mathbf{I}_n - g_\alpha(\mathbf{K}_i^T \mathbf{K}_i) \mathbf{K}_i^T \mathbf{K}_i$ ,  $i = 1, 2$ . These conditions have been verified for Tikhonov regularization, iterated Tikhonov regularization and the Landweber iteration in Kaltenbacher et al. (2008). It is remarkable to note that Kaltenbacher et al. (2008) considered a regularization method with the non-stationary iterated Tikhonov regularization, when the order  $p_k$  is variable and  $\alpha_k$  depends on  $p_k$ . By (H.12), (H.13), the local property (H.2), and the assumption  $\mathbf{x}_a \in B_\rho(\mathbf{x}^\dagger)$ , we deduce that

$$\|(\mathbf{R}_k - \mathbf{R})(\mathbf{x}_a - \mathbf{x}^\dagger)\| \leq \frac{c_R}{\sqrt{\alpha_k}} \|\mathbf{K}_k - \mathbf{K}\| \|\mathbf{x}_a - \mathbf{x}^\dagger\| \leq \frac{c_R c_K}{\sqrt{\alpha_k}} \rho \|\mathbf{K} \mathbf{e}_k^\delta\| \quad (\text{H.14})$$

and that

$$\|\mathbf{K}(\mathbf{R}_k - \mathbf{R})(\mathbf{x}_a - \mathbf{x}^\dagger)\| \leq c_R \|\mathbf{K}_k - \mathbf{K}\| \|\mathbf{x}_a - \mathbf{x}^\dagger\| \leq c_R c_K \rho \|\mathbf{K} \mathbf{e}_k^\delta\|. \quad (\text{H.15})$$

### H.1.1 Error estimates

Before proving convergence rates we need to derive estimates for  $\|\mathbf{e}_{k+1}^\delta\|$  and  $\|\mathbf{K}\mathbf{e}_{k+1}^\delta\|$ .

**Proposition H.1.** *Let assumptions (G.8), (G.66), (G.82), (H.2), (H.5), (H.12) and (H.13) hold. Then, if  $\mathbf{x}_a \in B_\rho(\mathbf{x}^\dagger)$ , we have the estimates*

$$\|\mathbf{e}_{k+1}^\delta\| \leq c_2 \|\mathbf{z}\| \alpha_k^\mu + \frac{c_{\mathbf{R}} c_{\mathbf{K}}}{\sqrt{\alpha_k}} \rho \|\mathbf{K}\mathbf{e}_k^\delta\| + \frac{c_0}{\sqrt{\alpha_k}} \left( \frac{3c_{\mathbf{K}}}{2} \|\mathbf{e}_k^\delta\| \|\mathbf{K}\mathbf{e}_k^\delta\| + \Delta \right) \quad (\text{H.16})$$

and

$$\begin{aligned} \|\mathbf{K}\mathbf{e}_{k+1}^\delta\| &\leq c_2 \|\mathbf{z}\| \alpha_k^{\mu+\frac{1}{2}} + c_{\mathbf{R}} c_{\mathbf{K}} \rho \|\mathbf{K}\mathbf{e}_k^\delta\| \\ &\quad + \left( \frac{c_{\mathbf{K}} c_0}{\sqrt{\alpha_k}} \|\mathbf{K}\mathbf{e}_k^\delta\| + 1 \right) \left( \frac{3c_{\mathbf{K}}}{2} \|\mathbf{e}_k^\delta\| \|\mathbf{K}\mathbf{e}_k^\delta\| + \Delta \right), \end{aligned} \quad (\text{H.17})$$

for  $0 < \mu \leq \mu_0 - 1/2$ .

*Proof.* The iteration error (H.6) can be expressed as

$$\begin{aligned} \mathbf{e}_{k+1}^\delta &= \mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger) + (\mathbf{R}_k - \mathbf{R})(\mathbf{x}_a - \mathbf{x}^\dagger) \\ &\quad + g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta], \end{aligned} \quad (\text{H.18})$$

whence, using (H.7), (H.8), (H.14), and the result (cf. (H.4))

$$\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta\| \leq \|\mathbf{y} - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta\| + \Delta \leq \frac{3c_{\mathbf{K}}}{2} \|\mathbf{e}_k^\delta\| \|\mathbf{K}\mathbf{e}_k^\delta\| + \Delta, \quad (\text{H.19})$$

we obtain (H.16). Similarly, the representation

$$\begin{aligned} \mathbf{K}\mathbf{e}_{k+1}^\delta &= \mathbf{K}\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger) + \mathbf{K}(\mathbf{R}_k - \mathbf{R})(\mathbf{x}_a - \mathbf{x}^\dagger) \\ &\quad + \mathbf{K}_k g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta] \\ &\quad + (\mathbf{K} - \mathbf{K}_k) g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta] \end{aligned} \quad (\text{H.20})$$

together with (H.9), (H.10), (H.15), (H.19) and (cf. (H.2) and (H.7))

$$\begin{aligned} &\|(\mathbf{K} - \mathbf{K}_k) g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T [\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta]\| \\ &\leq \frac{c_{\mathbf{K}} c_0}{\sqrt{\alpha_k}} \|\mathbf{K}\mathbf{e}_k^\delta\| \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k \mathbf{e}_k^\delta\| \end{aligned}$$

yields (H.17). □

### H.1.2 A priori stopping rule

Similarly to the a priori parameter choice method (G.26), we consider the stopping rule:

$$\vartheta \alpha_{k^*}^{\mu+\frac{1}{2}} \leq \Delta \leq \vartheta \alpha_k^{\mu+\frac{1}{2}}, \quad 0 \leq k < k^*, \quad (\text{H.21})$$

with  $\vartheta > 0$  and  $0 < \mu \leq \mu_0 - 1/2$ .

**Theorem H.2.** *Let assumptions (G.8), (G.66), (G.82), (H.2), (H.5), (H.12) and (H.13) hold. Moreover, let  $\alpha_k$  be chosen as in (H.1) with  $\alpha_0$  satisfying*

$$\|\mathbf{K}(\mathbf{x}_a - \mathbf{x}^\dagger)\| \leq \alpha_0^{\mu+\frac{1}{2}}, \quad (\text{H.22})$$

*and assume that  $\|\mathbf{z}\|$ ,  $\rho$  and  $\vartheta$  are sufficiently small such that*

$$Cc^{\mu+\frac{1}{2}} \leq 1, \quad (\text{H.23})$$

*with*

$$C = c_2 \|\mathbf{z}\| + \vartheta + c_K \left( c_R + \frac{3c_K c_0}{2} \alpha_0^\mu + \frac{3}{2} \right) \rho + c_K c_0 \vartheta \alpha_0^\mu,$$

*holds. If  $k^*$  is chosen according to the stopping rule (H.21) and  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  for all  $k = 0, \dots, k^* - 1$ , then we have the estimate*

$$\|\mathbf{K}\mathbf{e}_k^\delta\| \leq \alpha_k^{\mu+\frac{1}{2}}, \quad k = 0, \dots, k^*, \quad (\text{H.24})$$

*and the convergence rate*

$$\|\mathbf{x}_{k^*}^\delta - \mathbf{x}^\dagger\| = O\left(\Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \mu_0 - \frac{1}{2}. \quad (\text{H.25})$$

*Proof.* The assumption  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  gives  $\|\mathbf{e}_k^\delta\| \leq \rho$  for  $k = 0, \dots, k^* - 1$ . Then, using (H.21), we express the estimates (H.16) and (H.17), for  $k = 0, \dots, k^* - 1$ , as

$$\|\mathbf{e}_{k+1}^\delta\| \leq (c_2 \|\mathbf{z}\| + c_0 \vartheta) \alpha_k^\mu + c_K \left( c_R + \frac{3c_0}{2} \right) \frac{\rho}{\sqrt{\alpha_k}} \|\mathbf{K}\mathbf{e}_k^\delta\| \quad (\text{H.26})$$

and

$$\begin{aligned} \|\mathbf{K}\mathbf{e}_{k+1}^\delta\| &\leq (c_2 \|\mathbf{z}\| + \vartheta) \alpha_k^{\mu+\frac{1}{2}} \\ &\quad + \left[ c_K \left( c_R + \frac{3c_K c_0}{2\sqrt{\alpha_k}} \|\mathbf{K}\mathbf{e}_k^\delta\| + \frac{3}{2} \right) \rho + c_K c_0 \vartheta \alpha_k^\mu \right] \|\mathbf{K}\mathbf{e}_k^\delta\|, \end{aligned} \quad (\text{H.27})$$

respectively. To prove (H.24), we proceed by induction. For  $k = 0$ , the estimate (H.24) is valid due to assumption (H.22). Supposing that (H.24) holds for some  $k \leq k^* - 1$ , and making use of (H.1) and (H.23), we see that (H.27) yields

$$\begin{aligned} \|\mathbf{K}\mathbf{e}_{k+1}^\delta\| &\leq \alpha_k^{\mu+\frac{1}{2}} \left[ c_2 \|\mathbf{z}\| + \vartheta + c_K \left( c_R + \frac{3c_K c_0}{2} \alpha_k^\mu + \frac{3}{2} \right) \rho + c_K c_0 \vartheta \alpha_k^\mu \right] \\ &\leq \alpha_k^{\mu+\frac{1}{2}} \left[ c_2 \|\mathbf{z}\| + \vartheta + c_K \left( c_R + \frac{3c_K c_0}{2} \alpha_0^\mu + \frac{3}{2} \right) \rho + c_K c_0 \vartheta \alpha_0^\mu \right] \\ &\leq C c^{\mu+\frac{1}{2}} \alpha_{k+1}^{\mu+\frac{1}{2}} \\ &\leq \alpha_{k+1}^{\mu+\frac{1}{2}}. \end{aligned}$$



Thus, the estimate (H.24) is valid for all  $k = 0, \dots, k^*$ , and by (H.26) we find that

$$\|\mathbf{e}_{k^*}^\delta\| \leq c^\mu \alpha_{k^*}^\mu \left[ c_2 \|\mathbf{z}\| + c_0 \vartheta + c_K \left( c_R + \frac{3c_0}{2} \right) \rho \right]. \quad (\text{H.28})$$

From (H.21) and (H.28) we obtain

$$\|\mathbf{x}_{k^*}^\delta - \mathbf{x}^\dagger\| = O(\alpha_{k^*}^\mu) = O\left(\left(\alpha_{k^*}^{\mu+\frac{1}{2}}\right)^{\frac{2\mu}{2\mu+1}}\right) = O\left(\Delta^{\frac{2\mu}{2\mu+1}}\right),$$

and the proof is finished.  $\square$

In view of Theorem H.2, the best convergence rate of the iteratively regularized Gauss–Newton method is  $O(\sqrt{\Delta})$ . However, under slightly different assumptions, Kaltenbacher et al. (2008) proved that the best possible rate which can be achieved with the a priori rule (H.21) is  $O(\Delta^{2/3})$ .

### H.1.3 Discrepancy principle

As in (G.30), the following simplified version of the discrepancy principle is used in our analysis: the stopping index  $k^*$  is chosen so that the residual norm at the last iterate falls below  $\tau_{\text{dp}}\Delta$ ,

$$\|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{k^*}^\delta)\| < \tau_{\text{dp}}\Delta \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta)\|, \quad 0 \leq k < k^*, \quad (\text{H.29})$$

and the previous residual norm is of the order of magnitude of the noise level

$$\tau_{\text{dp}}\Delta \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{k^*-1}^\delta)\| \leq (\tau_{\text{dp}} + \varepsilon) \Delta, \quad \varepsilon > 0. \quad (\text{H.30})$$

Under the assumption that  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  for all  $k = 0, \dots, k^*$ , the following partial results can be established:

(1) By virtue of (H.3), we have

$$\begin{aligned} & \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger)\| \\ & \leq \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger) - \mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\| + \|\mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\| \\ & \leq \left(1 + \frac{c_K}{2} \|\mathbf{x} - \mathbf{x}^\dagger\|\right) \|\mathbf{K}(\mathbf{x}^\dagger)(\mathbf{x} - \mathbf{x}^\dagger)\|, \end{aligned}$$

and we obtain, for  $k = 0, \dots, k^* - 1$ ,

$$\tau_{\text{dp}}\Delta \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta)\| \leq \left(1 + \frac{c_K}{2}\rho\right) \|\mathbf{K}\mathbf{e}_k^\delta\| + \Delta.$$

Thus,

$$\Delta \leq \frac{1}{\tau_{\text{dp}} - 1} \left(1 + \frac{c_K}{2}\rho\right) \|\mathbf{K}\mathbf{e}_k^\delta\|, \quad \tau_{\text{dp}} > 1. \quad (\text{H.31})$$

(2) The triangle inequality

$$\begin{aligned} & \| \mathbf{K}(\mathbf{x}^\dagger) (\mathbf{x} - \mathbf{x}^\dagger) \| \\ & \leq \| \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger) \| + \| \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger) - \mathbf{K}(\mathbf{x}^\dagger) (\mathbf{x} - \mathbf{x}^\dagger) \| \end{aligned}$$

together with (H.3) yields

$$\left(1 - \frac{c_K}{2} \|\mathbf{x} - \mathbf{x}^\dagger\|\right) \|\mathbf{K}(\mathbf{x}^\dagger) (\mathbf{x} - \mathbf{x}^\dagger)\| \leq \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}^\dagger)\|,$$

and, by (H.30), we infer that

$$\left(1 - \frac{c_K}{2} \rho\right) \|\mathbf{K}\mathbf{e}_{k^*-1}^\delta\| \leq \|\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{k^*-1}^\delta)\| + \Delta \leq (1 + \tau_{\text{dp}} + \varepsilon) \Delta.$$

Hence,

$$\|\mathbf{K}\mathbf{e}_{k^*-1}^\delta\| \leq \frac{1 + \tau_{\text{dp}} + \varepsilon}{1 - \frac{c_K}{2} \rho} \Delta, \quad 0 < c_K < \frac{2}{\rho}. \quad (\text{H.32})$$

Similarly, from (H.29) we deduce that

$$\|\mathbf{K}\mathbf{e}_{k^*}^\delta\| < \frac{1 + \tau_{\text{dp}}}{1 - \frac{c_K}{2} \rho} \Delta. \quad (\text{H.33})$$

**Theorem H.3.** *Let the assumptions of Theorem H.2 hold, and suppose that  $\|\mathbf{z}\|$  and  $\rho$  are sufficiently small and that  $\tau_{\text{dp}} > 1$  is sufficiently large such that*

$$C c^{\mu + \frac{1}{2}} \leq 1, \quad (\text{H.34})$$

with

$$C = c_2 \|\mathbf{z}\| + c_R c_K \rho + (c_K c_0 \alpha_0^\mu + 1) \left[ \frac{3c_K}{2} \rho + \frac{1}{\tau_{\text{dp}} - 1} \left(1 + \frac{c_K}{2} \rho\right) \right],$$

is fulfilled. Moreover, let  $k^*$  be the stopping index of the discrepancy principle (H.29)–(H.30), and assume that  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  for all  $k = 0, \dots, k^*$ . Then we have the estimate

$$\|\mathbf{K}\mathbf{e}_k^\delta\| \leq \alpha_k^{\mu + \frac{1}{2}}, \quad k = 0, \dots, k^*, \quad (\text{H.35})$$

and the convergence rate

$$\|\mathbf{x}_{k^*}^\delta - \mathbf{x}^\dagger\| = O\left(\Delta^{\frac{2\mu}{2\mu+1}}\right), \quad 0 < \mu \leq \mu_0 - \frac{1}{2}. \quad (\text{H.36})$$

*Proof.* Under the assumption  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  and by virtue of (H.31), the error bounds (H.16) and (H.17) become, for  $k = 0, \dots, k^* - 1$ ,

$$\|\mathbf{e}_{k+1}^\delta\| \leq c_2 \|\mathbf{z}\| \alpha_k^\mu + \left[ c_K \left( c_R + \frac{3c_0}{2} \right) \rho + \frac{c_0}{\tau_{\text{dp}} - 1} \left(1 + \frac{c_K}{2} \rho\right) \right] \frac{1}{\sqrt{\alpha_k}} \|\mathbf{K}\mathbf{e}_k^\delta\| \quad (\text{H.37})$$

and

$$\begin{aligned} \|\mathbf{Ke}_{k+1}^\delta\| &\leq c_2 \|\mathbf{z}\| \alpha_k^{\mu+\frac{1}{2}} + \left\{ c_R c_K \rho + \left( \frac{c_K c_0}{\sqrt{\alpha_k}} \|\mathbf{Ke}_k^\delta\| + 1 \right) \right. \\ &\quad \left. \times \left[ \frac{3c_K}{2} \rho + \frac{1}{\tau_{dp} - 1} \left( 1 + \frac{c_K}{2} \rho \right) \right] \right\} \|\mathbf{Ke}_k^\delta\|, \end{aligned} \quad (\text{H.38})$$

respectively. As in Theorem H.2, the estimate (H.35) is proven by induction using assumption (H.22) and the closeness condition (H.34). Essentially, assuming that (H.35) holds for some  $k \leq k^* - 1$ , we find that

$$\begin{aligned} \|\mathbf{Ke}_{k+1}^\delta\| &\leq \alpha_k^{\mu+\frac{1}{2}} \left\{ c_2 \|\mathbf{z}\| + c_R c_K \rho + (c_K c_0 \alpha_k^\mu + 1) \left[ \frac{3c_K}{2} \rho + \frac{1}{\tau_{dp} - 1} \left( 1 + \frac{c_K}{2} \rho \right) \right] \right\} \\ &\leq \alpha_k^{\mu+\frac{1}{2}} \left\{ c_2 \|\mathbf{z}\| + c_R c_K \rho + (c_K c_0 \alpha_0^\mu + 1) \left[ \frac{3c_K}{2} \rho + \frac{1}{\tau_{dp} - 1} \left( 1 + \frac{c_K}{2} \rho \right) \right] \right\} \\ &\leq \alpha_{k+1}^{\mu+\frac{1}{2}}. \end{aligned}$$

We proceed now to prove the convergence rate (H.36). First, we observe that the estimates (H.31) and (H.35) yield

$$\Delta \leq \frac{1 + \frac{c_K}{2} \rho}{\tau_{dp} - 1} \alpha_{k^*-1}^{\mu+\frac{1}{2}},$$

and therefore,

$$\frac{1}{\sqrt{\alpha_{k^*-1}}} \leq \left( \frac{1 + \frac{c_K}{2} \rho}{\tau_{dp} - 1} \right)^{\frac{1}{2\mu+1}} \Delta^{-\frac{1}{2\mu+1}}.$$

Combining this result with (H.32) gives

$$\frac{1}{\sqrt{\alpha_{k^*-1}}} \|\mathbf{Ke}_{k^*-1}^\delta\| \leq \frac{1 + \tau_{dp} + \varepsilon}{1 - \frac{c_K}{2} \rho} \left( \frac{1 + \frac{c_K}{2} \rho}{\tau_{dp} - 1} \right)^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu}{2\mu+1}}, \quad (\text{H.39})$$

and we deduce that

$$\frac{1}{\sqrt{\alpha_{k^*-1}}} \|\mathbf{Ke}_{k^*-1}^\delta\| = O\left(\Delta^{\frac{2\mu}{2\mu+1}}\right). \quad (\text{H.40})$$

The derivation of the estimate (H.16) relies on the error representation (H.18). Repeating the steps of this derivation but without evaluating the term  $\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger)$ , yields a bound for  $\|\mathbf{e}_{k+1}^\delta\|$ ; this together with (H.31) gives (analogously to (H.37))

$$\begin{aligned} \|\mathbf{e}_{k^*}^\delta\| &\leq \|\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger)\| + \left[ c_K \left( c_R + \frac{3c_0}{2} \right) \rho \right. \\ &\quad \left. + \frac{c_0}{\tau_{dp} - 1} \left( 1 + \frac{c_K}{2} \rho \right) \right] \frac{1}{\sqrt{\alpha_{k^*-1}}} \|\mathbf{Ke}_{k^*-1}^\delta\|. \end{aligned} \quad (\text{H.41})$$

Similarly, the error representation (H.20) yields an estimate for  $\|\mathbf{KR}(\mathbf{x}_a - \mathbf{x}^\dagger)\|$  which, by virtue of (H.31), can be expressed as (analogously to (H.38))

$$\begin{aligned} \|\mathbf{KR}(\mathbf{x}_a - \mathbf{x}^\dagger)\| &\leq \|\mathbf{Ke}_{k^*}^\delta\| + \left\{ c_R c_K \rho + \left( \frac{c_K c_0}{\sqrt{\alpha_{k^*-1}}} \|\mathbf{Ke}_{k^*-1}^\delta\| + 1 \right) \right. \\ &\quad \left. \times \left[ \frac{3c_K}{2} \rho + \frac{1}{\tau_{dp} - 1} \left( 1 + \frac{c_K}{2} \rho \right) \right] \right\} \|\mathbf{Ke}_{k^*-1}^\delta\|. \end{aligned} \quad (\text{H.42})$$

From (H.32) and (H.33) we have  $\|\mathbf{Ke}_{k^*-1}^\delta\| = O(\Delta)$  and  $\|\mathbf{Ke}_{k^*}^\delta\| = O(\Delta)$ , respectively. Inserting these results into (H.42), we obtain

$$\|\mathbf{KR}(\mathbf{x}_a - \mathbf{x}^\dagger)\| = O(\Delta). \quad (\text{H.43})$$

Finally, (H.43) and the moment inequality (H.11) give

$$\|\mathbf{R}(\mathbf{x}_a - \mathbf{x}^\dagger)\| = O\left(\Delta^{\frac{2\mu}{2\mu+1}}\right), \quad (\text{H.44})$$

and the desired convergence rate follows from (H.41) in conjunction with (H.40) and (H.44).  $\square$

## H.2 Newton-type methods without a priori information

An ingenious proof of convergence rate results for Newton-type methods without a priori information has been provided by Rieder (1999, 2003). For the sake of completeness and in order to evidence the elegance of the arguments employed, we present below a simplified version of Rieder's analysis.

Newton-type methods rely on the update formula

$$\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{p}_k^\delta, \quad k = 0, 1, \dots, \quad (\text{H.45})$$

where  $\mathbf{p}_k^\delta$  is the Newton step and  $\mathbf{x}_0^\delta = \mathbf{x}_a$ . If  $\mathbf{x}^\dagger$  is a solution of the nonlinear equation with exact data  $\mathbf{F}(\mathbf{x}) = \mathbf{y}$ , then

$$\mathbf{p}_k^\dagger = \mathbf{x}^\dagger - \mathbf{x}_k^\delta$$

is the exact step, since in this case  $\mathbf{x}_{k+1}^\delta = \mathbf{x}^\dagger$ . Using the Taylor expansion of the forward model about  $\mathbf{x}_k^\delta$ ,

$$\mathbf{F}(\mathbf{x}^\dagger) = \mathbf{F}(\mathbf{x}_k^\delta) + \mathbf{K}_k(\mathbf{x}^\dagger - \mathbf{x}_k^\delta) + \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta),$$

with  $\mathbf{K}_k = \mathbf{K}(\mathbf{x}_k^\delta)$ , and taking into account that  $\mathbf{F}(\mathbf{x}^\dagger) = \mathbf{y}$ , we see that  $\mathbf{p}_k^\dagger$  solves the equation

$$\mathbf{K}_k \mathbf{p} = \mathbf{r}_k, \quad (\text{H.46})$$

with

$$\mathbf{r}_k = \mathbf{y} - \mathbf{F}(\mathbf{x}_k^\delta) - \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta).$$

The exact step  $\mathbf{p}_k^\dagger$  is the least squares solution of equation (H.46), and for  $\mathbf{K}_k = \mathbf{U}\Sigma\mathbf{V}^T$ , we have

$$\mathbf{p}_k^\dagger = \sum_{i=1}^n \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{r}_k) \mathbf{v}_i \quad (\text{H.47})$$

and

$$\mathbf{K}_k \mathbf{p}_k^\dagger = \sum_{i=1}^n (\mathbf{u}_i^T \mathbf{r}_k) \mathbf{u}_i. \quad (\text{H.48})$$

In practice,  $\mathbf{r}_k$  is unknown and only its noisy version,

$$\mathbf{r}_k^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta),$$

is available; the deviation of  $\mathbf{r}_k^\delta$  from  $\mathbf{r}_k$ ,

$$\|\mathbf{r}_k^\delta - \mathbf{r}_k\| = \|\mathbf{y}^\delta - \mathbf{y} + \mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta)\| \leq \Delta + \|\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta)\|, \quad (\text{H.49})$$

accounts of the instrumental noise and the linearization error. In the framework of Newton-type methods without a priori information,  $\mathbf{p}_{\alpha_k k}^\delta$  is computed as the solution of the equation

$$\mathbf{K}_k \mathbf{p} = \mathbf{r}_k^\delta \quad (\text{H.50})$$

by using a general regularization method of the form

$$\mathbf{p}_{\alpha_k k}^\delta = g_{\alpha_k}(\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T \mathbf{r}_k^\delta,$$

and the new iterate is taken as  $\mathbf{x}_{k+1}^\delta = \mathbf{x}_k^\delta + \mathbf{p}_{\alpha_k k}^\delta$ . The iteration function  $g_\alpha$  may correspond to Tikhonov regularization,

$$g_\alpha(\lambda) = \frac{1}{\lambda + \alpha},$$

the  $p$ -times iterated Tikhonov regularization (with fixed  $p$ ),

$$g_\alpha(\lambda) = \frac{1}{\lambda} \left[ 1 - \left( \frac{\alpha}{\lambda + \alpha} \right)^p \right],$$

and the Landweber iteration,

$$g_\alpha(\lambda) = \frac{1}{\lambda} [1 - (1 - \lambda)^p], \quad \alpha = \frac{1}{p}.$$

The last two regularization methods solve the linearized equation (H.50) by using the iterations

$$\begin{aligned} \mathbf{p}_{0k}^\delta &= \mathbf{0}, \\ \mathbf{p}_{lk}^\delta &= \mathbf{p}_{l-1k}^\delta + \mathbf{K}_k^\dagger (\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{l-1k}^\delta), \quad 1 \leq l \leq p, \\ \mathbf{p}_{\alpha_k k}^\delta &= \mathbf{p}_{pk}^\delta, \end{aligned}$$

and

$$\begin{aligned} \mathbf{p}_{0k}^\delta &= \mathbf{0}, \\ \mathbf{p}_{lk}^\delta &= \mathbf{p}_{l-1k}^\delta + \mathbf{K}_k^T (\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{l-1k}^\delta), \quad 1 \leq l \leq p_k, \end{aligned}$$

respectively.

For the iteration and the residual functions, we consider the simplified assumptions

$$0 \leq g_\alpha(\lambda) \leq \frac{c_1}{\alpha}, \quad (\text{H.51})$$

$$0 \leq r_\alpha(\lambda) \leq 1, \quad r_\alpha(0) = 1, \quad (\text{H.52})$$

$$0 \leq \lambda r_\alpha(\lambda) \leq c_2 \alpha, \quad (\text{H.53})$$

for all  $\alpha > 0$ ,  $\lambda \in [0, \sigma_{\max}^2]$  and  $c_1, c_2 > 0$ . As usual,  $\sigma_{\max}^2$  is a bound for  $\|\mathbf{K}(\mathbf{x})^T \mathbf{K}(\mathbf{x})\|$  in  $B_\rho(\mathbf{x}^\dagger)$ , and the iteration function  $g_\alpha(\lambda)$  is continuously extended at  $\lambda = 0$  by setting  $g_\alpha(0) = \lim_{\lambda \rightarrow 0} g_\alpha(\lambda)$ . Conditions (H.51)–(H.53) hold for Tikhonov regularization with  $c_1 = c_2 = 1$ , for the  $p$ -times iterated Tikhonov regularization with  $c_1 = p$  and  $c_2 = (p-1)^{p-1}/p^p$ , and for the Landweber iteration with  $c_1 = 1$  and  $c_2 = \exp(-1)$ .

The regularization method under examination belongs to the class of inexact Newton iterations. It consists of an inner iteration, which provides the regularization parameter, and an outer Newton iteration, which updates the current iterate. At the Newton step  $k$ , the regularization parameter  $\alpha_k$  is chosen as follows: if  $\{\alpha_j\}$  is a geometric sequence of regularization parameters with ratio  $q < 1$ , i.e.,  $\alpha_{j+1} = q\alpha_j$ , we choose  $\alpha_k = \alpha_{j^*(k)}$  such that the linearized residual is of the same order of magnitude with the nonlinear residual,

$$\left\| \mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha_{j^*(k)}k}^\delta \right\| \leq \theta_k \left\| \mathbf{r}_k^\delta \right\| < \left\| \mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha_{jk}}^\delta \right\|, \quad 0 \leq j < j^*(k). \quad (\text{H.54})$$

The Newton iteration is stopped according to the discrepancy principle in order to avoid noise amplification, i.e.,

$$\left\| \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{k^*}^\delta) \right\| \leq \tau_{\text{dp}} \Delta < \left\| \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_k^\delta) \right\|, \quad 0 \leq k < k^*. \quad (\text{H.55})$$

The convergence analysis is performed by assuming the following local property of the forward model:

$$\mathbf{K}(\mathbf{x}_1) = \mathbf{Q}(\mathbf{x}_1, \mathbf{x}_2) \mathbf{K}(\mathbf{x}_2), \quad \|\mathbf{I}_m - \mathbf{Q}(\mathbf{x}_1, \mathbf{x}_2)\| \leq c_Q \|\mathbf{x}_1 - \mathbf{x}_2\|, \quad (\text{H.56})$$

for all  $\mathbf{x}_1, \mathbf{x}_2 \in B_\rho(\mathbf{x}^\dagger)$  and  $c_Q > 0$ . By virtue of (H.56), it is apparent that the norm of the  $m \times m$  matrix  $\mathbf{Q}$  can be bounded in  $B_\rho(\mathbf{x}^\dagger)$  as

$$\|\mathbf{Q}(\mathbf{x}_1, \mathbf{x}_2)\| \leq 1 + \|\mathbf{I}_m - \mathbf{Q}(\mathbf{x}_1, \mathbf{x}_2)\| \leq 1 + c_Q \|\mathbf{x}_1 - \mathbf{x}_2\| \leq \bar{c}_Q, \quad (\text{H.57})$$

with

$$\bar{c}_Q = 1 + 2c_Q \rho,$$

and that

$$\|[\mathbf{K}(\mathbf{x}_1) - \mathbf{K}(\mathbf{x}_2)](\mathbf{x}_1 - \mathbf{x}_2)\| \leq 2c_Q \rho \|\mathbf{K}(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)\|.$$

For the linearization error

$$\mathbf{R}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{F}(\mathbf{x}_1) - \mathbf{F}(\mathbf{x}_2) - \mathbf{K}(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)$$

the estimate

$$\begin{aligned} \|\mathbf{R}(\mathbf{x}_1, \mathbf{x}_2)\| &\leq \int_0^1 \|[\mathbf{K}(\mathbf{x}_2 + t(\mathbf{x}_1 - \mathbf{x}_2)) - \mathbf{K}(\mathbf{x}_2)](\mathbf{x}_1 - \mathbf{x}_2)\| dt \\ &= \int_0^1 \|[\mathbf{Q}(\mathbf{x}_2 + t(\mathbf{x}_1 - \mathbf{x}_2), \mathbf{x}_2) - \mathbf{I}_m] \mathbf{K}(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)\| dt \\ &\leq c_Q \rho \|\mathbf{K}(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)\| \end{aligned}$$

and the triangle inequality

$$\|\mathbf{K}(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)\| \leq \|\mathbf{R}(\mathbf{x}_1, \mathbf{x}_2)\| + \|\mathbf{F}(\mathbf{x}_1) - \mathbf{F}(\mathbf{x}_2)\|$$

give

$$\|\mathbf{R}(\mathbf{x}_1, \mathbf{x}_2)\| \leq \omega \|\mathbf{F}(\mathbf{x}_1) - \mathbf{F}(\mathbf{x}_2)\|, \quad (\text{H.58})$$

with

$$\omega = \frac{c_Q \rho}{1 - c_Q \rho}, \quad 0 < c_Q \rho < 1.$$

Particularizing the above estimate for  $\mathbf{x}_1 = \mathbf{x}^\dagger$  and  $\mathbf{x}_2 = \mathbf{x}_k^\delta$ , we obtain

$$\|\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta)\| \leq \omega \|\mathbf{y} - \mathbf{F}(\mathbf{x}_k^\delta)\| \leq \omega (\Delta + \|\mathbf{r}_k^\delta\|). \quad (\text{H.59})$$

Before going any further, let us show that the selection criterion (H.54) is well defined. For a regularization parameter  $\alpha$ , the linearized residual can be computed as (cf. (E.27) and (E.28))

$$\begin{aligned} \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{P}_{\alpha k}^\delta\|^2 &= \|\mathbf{I}_m - \mathbf{K}_k g_\alpha (\mathbf{K}_k^T \mathbf{K}_k) \mathbf{K}_k^T\| \mathbf{r}_k^\delta\|^2 \\ &= \|r_\alpha (\mathbf{K}_k \mathbf{K}_k^T) \mathbf{r}_k^\delta\|^2 \\ &= \sum_{i=1}^m r_\alpha^2 (\sigma_i^2) (\mathbf{u}_i^T \mathbf{r}_k^\delta)^2, \end{aligned} \quad (\text{H.60})$$

with the convention  $r_\alpha (\sigma_i^2) = 1$  for  $i = n+1, \dots, m$ . Supposing that  $r_\alpha$  is an increasing function of  $\alpha$ , we deduce that the linearized residual is also an increasing function of  $\alpha$ , and the additional assumption (cf. (C.15))  $\lim_{\alpha \rightarrow 0} r_\alpha(\lambda) = 0$ , yields

$$\lim_{\alpha \rightarrow 0} \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{P}_{\alpha k}^\delta\|^2 = \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{P}_{0k}^\delta\|^2 = \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{r}_k^\delta)^2.$$

On the other hand, by virtue of (H.48), we have

$$\begin{aligned} \mathbf{r}_k^\delta - \mathbf{r}_k &= \mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_k^\dagger \\ &= \mathbf{r}_k^\delta - \sum_{i=1}^n (\mathbf{u}_i^T \mathbf{r}_k) \mathbf{u}_i \\ &= \sum_{i=1}^n [\mathbf{u}_i^T (\mathbf{r}_k^\delta - \mathbf{r}_k)] \mathbf{u}_i + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{r}_k^\delta) \mathbf{u}_i, \end{aligned}$$

and clearly,

$$\|\mathbf{r}_k^\delta - \mathbf{r}_k\|^2 = \sum_{i=1}^n [\mathbf{u}_i^T (\mathbf{r}_k^\delta - \mathbf{r}_k)]^2 + \sum_{i=n+1}^m (\mathbf{u}_i^T \mathbf{r}_k^\delta)^2 \geq \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{0k}^\delta\|^2. \quad (\text{H.61})$$

If we define  $\tau_k$  by

$$\tau_k = \frac{\theta_k \|\mathbf{r}_k^\delta\|}{\|\mathbf{r}_k^\delta - \mathbf{r}_k\|},$$

then the selection criterion (H.54) can also be expressed as

$$\|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha_{j^*(k)}k}^\delta\| \leq \tau_k \|\mathbf{r}_k^\delta - \mathbf{r}_k\| < \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha_{jk}}^\delta\|, \quad 0 \leq j < j^*(k). \quad (\text{H.62})$$

From (H.61), we observe that the existence of  $\alpha_k = \alpha_{j^*(k)}$  in (H.62) is guaranteed if  $\tau_k > 1$ . This condition can be satisfied if the control parameters  $\tau_{\text{dp}}$  and  $\theta_k$  are chosen appropriately. By (H.49) and (H.59), we find that

$$\tau_k = \frac{\theta_k \|\mathbf{r}_k^\delta\|}{\|\mathbf{r}_k^\delta - \mathbf{r}_k\|} \geq \frac{\theta_k}{\omega + (1 + \omega) \frac{\Delta}{\|\mathbf{r}_k^\delta\|}},$$

and the discrepancy principle condition (H.55) then gives

$$\tau_k > \frac{\theta_k}{\omega + (1 + \omega) \frac{1}{\tau_{\text{dp}}}}, \quad 0 \leq k < k^*.$$

Assuming that

$$\tau_{\text{dp}} > \frac{1 + \omega}{1 - \omega}, \quad 0 < \omega < 1, \quad (\text{H.63})$$

which yields

$$\omega + (1 + \omega) \frac{1}{\tau_{\text{dp}}} < 1,$$

and choosing the tolerance  $\theta_k$  as

$$\omega + (1 + \omega) \frac{1}{\tau_{\text{dp}}} < \theta_k \leq 1, \quad (\text{H.64})$$

we find that  $\tau_k > 1$ . Thus, conditions (H.63) and (H.64) guarantee that the selection criterion (H.54) is well defined.

The next result states that the nonlinear residuals decrease linearly.

**Proposition H.4.** *For  $0 < \eta < 1$ , assume that*

$$0 < \omega < \frac{\eta}{\eta + 2} \quad (\text{H.65})$$

*is satisfied, and choose the tolerances  $\tau_{\text{dp}}$  and  $\theta_k$  as*

$$\tau_{\text{dp}} > \frac{1 + \omega}{\eta - (2 + \eta)\omega}, \quad \omega + (1 + \omega) \frac{1}{\tau_{\text{dp}}} < \theta_k \leq \eta - (1 + \eta)\omega. \quad (\text{H.66})$$



Then, if  $\mathbf{x}_k^\delta, \mathbf{x}_{k+1}^\delta \in B_\rho(\mathbf{x}^\dagger)$ , there holds

$$\frac{\|\mathbf{r}_{k+1}^\delta\|}{\|\mathbf{r}_k^\delta\|} \leq \frac{\theta_k + \omega}{1 - \omega} \leq \eta. \quad (\text{H.67})$$

*Proof.* Let us first discuss the selection rules for  $\tau_{\text{dp}}$  and  $\theta_k$ . For  $0 < \eta < 1$ , assumption (H.65) yields  $0 < \omega < 1$ . Then, the obvious inequality

$$\eta - 1 < 0 < (1 + \eta)\omega, \quad (\text{H.68})$$

together with assumption (H.65) gives

$$0 < \eta - (2 + \eta)\omega < 1 - \omega$$

and further, by (H.66),

$$\tau_{\text{dp}} > \frac{1 + \omega}{\eta - (2 + \eta)\omega} > \frac{1 + \omega}{1 - \omega}.$$

Thus, assumption (H.63) still holds true. On the other hand, from (H.68) and the first selection rule in (H.66) we have

$$\eta - (1 + \eta)\omega < 1,$$

and

$$\omega + (1 + \omega) \frac{1}{\tau_{\text{dp}}} < \eta - (1 + \eta)\omega,$$

respectively. Hence,  $\theta_k$  can be chosen as in (H.66), and (H.64) still holds true.

Using the identity

$$\mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_{k+1}^\delta) = \mathbf{r}_k^\delta - \mathbf{K}_k(\mathbf{x}_{k+1}^\delta - \mathbf{x}_k^\delta) - \mathbf{R}(\mathbf{x}_{k+1}^\delta, \mathbf{x}_k^\delta),$$

where  $\mathbf{x}_{k+1}^\delta$  is computed for  $\alpha_k = \alpha_{j^*(k)}$ , and employing (H.54) together with (H.58), we find that

$$\begin{aligned} \|\mathbf{r}_{k+1}^\delta\| &\leq \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha_k}^\delta\| + \|\mathbf{R}(\mathbf{x}_{k+1}^\delta, \mathbf{x}_k^\delta)\| \\ &\leq \theta_k \|\mathbf{r}_k^\delta\| + \omega \|\mathbf{F}(\mathbf{x}_{k+1}^\delta) - \mathbf{F}(\mathbf{x}_k^\delta)\| \\ &\leq (\theta_k + \omega) \|\mathbf{r}_k^\delta\| + \omega \|\mathbf{r}_{k+1}^\delta\| \end{aligned}$$

and further that

$$\frac{\|\mathbf{r}_{k+1}^\delta\|}{\|\mathbf{r}_k^\delta\|} \leq \frac{\theta_k + \omega}{1 - \omega}.$$

The upper bound on  $\theta_k$  in (H.66) gives

$$\frac{\theta_k + \omega}{1 - \omega} \leq \eta$$

and the proof is finished.  $\square$

Estimates for the termination index and the regularization parameter of the Newton iteration are given below.

**Proposition H.5.** *Under the same assumptions as in Proposition H.4, the termination index of the regularization method satisfies*

$$k^* < \log_{\eta} \left( \frac{\tau_{\text{dp}} \Delta}{\|\mathbf{r}_0^\delta\|} \right) + 1, \quad (\text{H.69})$$

provided that  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  for all  $k = 0, \dots, k^* - 1$ .

*Proof.* By (H.67), we have

$$\|\mathbf{r}_k^\delta\| \leq \eta \|\mathbf{r}_{k-1}^\delta\| \leq \dots \leq \eta^k \|\mathbf{r}_0^\delta\|,$$

with  $\mathbf{r}_0^\delta = \mathbf{y}^\delta - \mathbf{F}(\mathbf{x}_0^\delta)$  and  $0 < \eta < 1$ . This yields

$$k \leq \log_{\eta} \left( \frac{\|\mathbf{r}_k^\delta\|}{\|\mathbf{r}_0^\delta\|} \right).$$

Using the discrepancy principle condition  $\|\mathbf{r}_{k^*-1}^\delta\| > \tau_{\text{dp}} \Delta$  and the fact that  $\log_{\eta}$  is a monotonic decreasing function, we deduce that (H.69) holds.  $\square$

**Proposition H.6.** *Let assumptions (H.52) and (H.53) be fulfilled and let us suppose that at the Newton step  $k$ , there exists  $\mathbf{w}_k \in \mathbb{R}^m$  such that  $\mathbf{p}_k^\dagger = \mathbf{K}_k^T \mathbf{w}_k$ . Then the regularization parameter  $\alpha_k = \alpha_{j^*(k)}$  satisfies*

$$\alpha_{j^*(k)} > \frac{q(\tau_k - 1)}{c_2} \frac{1}{\|\mathbf{w}_k\|} \|\mathbf{r}_k^\delta - \mathbf{r}_k\|. \quad (\text{H.70})$$

*Proof.* By (H.60), assumption (H.52), and the relation  $\mathbf{r}_k = \mathbf{K}_k \mathbf{p}_k^\dagger$ , we obtain

$$\begin{aligned} \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha k}^\delta\| &= \|r_\alpha (\mathbf{K}_k \mathbf{K}_k^T) \mathbf{r}_k^\delta\| \\ &\leq \|r_\alpha (\mathbf{K}_k \mathbf{K}_k^T) \mathbf{K}_k \mathbf{p}_k^\dagger\| + \|r_\alpha (\mathbf{K}_k \mathbf{K}_k^T) (\mathbf{r}_k^\delta - \mathbf{r}_k)\| \\ &\leq \|r_\alpha (\mathbf{K}_k \mathbf{K}_k^T) \mathbf{K}_k \mathbf{p}_k^\dagger\| + \|\mathbf{r}_k^\delta - \mathbf{r}_k\|, \end{aligned}$$

and further, by assumption (H.53) and the relation  $\mathbf{p}_k^\dagger = \mathbf{K}_k^T \mathbf{w}_k$ , yielding

$$\|r_\alpha (\mathbf{K}_k \mathbf{K}_k^T) \mathbf{K}_k \mathbf{p}_k^\dagger\|^2 = \sum_{i=1}^n [\sigma_i^2 r_\alpha^2 (\sigma_i^2)]^2 (\mathbf{u}_i^T \mathbf{w}_k)^2 \leq c_2^2 \alpha^2 \|\mathbf{w}_k\|^2,$$

we get

$$\|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha k}^\delta\| \leq c_2 \alpha \|\mathbf{w}_k\| + \|\mathbf{r}_k^\delta - \mathbf{r}_k\|.$$

The selection rule (H.62) gives

$$\tau_k \|\mathbf{r}_k^\delta - \mathbf{r}_k\| < \|\mathbf{r}_k^\delta - \mathbf{K}_k \mathbf{p}_{\alpha_{j^*(k)-1} k}^\delta\| \leq c_2 \alpha_{j^*(k)-1} \|\mathbf{w}_k\| + \|\mathbf{r}_k^\delta - \mathbf{r}_k\|,$$

and we infer that

$$\alpha_{j^*(k)-1} > \frac{\tau_k - 1}{c_2} \frac{1}{\|\mathbf{w}_k\|} \|\mathbf{r}_k^\delta - \mathbf{r}_k\|.$$

Finally, the relation  $\alpha_{j^*(k)} = q\alpha_{j^*(k)-1}$  yields (H.70).  $\square$

Crucial for proving convergence rates is the derivation of an estimate for  $\|\mathbf{w}_k\|$ .

**Proposition H.7.** *Let assumptions (H.51), (H.52), (H.53) and (H.56) be fulfilled and let the initial guess  $\mathbf{x}_0^\delta = \mathbf{x}_a \in B_\rho(\mathbf{x}^\dagger)$  be such that*

$$\mathbf{p}_0^\dagger = \mathbf{K}_0^T \mathbf{w}_0 \quad (\text{H.71})$$

for some  $\mathbf{w}_0 \in \mathbb{R}^m$ . If  $k^*$  is the termination index of the regularization method and  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  for all  $k = 1, \dots, k^*$ , then there holds

$$\mathbf{p}_k^\dagger = \mathbf{K}_k^T \mathbf{w}_k, \quad k = 1, \dots, k^*,$$

with

$$\mathbf{w}_k = \mathbf{Q}(\mathbf{x}_0^\delta, \mathbf{x}_k^\delta)^T \mathbf{w}_0 - \sum_{i=0}^{k-1} \mathbf{Q}(\mathbf{x}_i^\delta, \mathbf{x}_k^\delta)^T g_{\alpha_i}(\mathbf{K}_i \mathbf{K}_i^T) \mathbf{r}_i^\delta,$$

and  $\alpha_i = \alpha_{j^*(i)}$ . Moreover, we have

$$\|\mathbf{w}_k\| < \bar{c}_Q (1 + c) (1 + c\bar{c}_Q)^{k-1} \|\mathbf{w}_0\|, \quad (\text{H.72})$$

with

$$c = 1 + \frac{c_1 c_2}{q(\tau_{\min} - 1)}, \quad \tau_{\min} = \min(\tau_0, \dots, \tau_{k^*-1}). \quad (\text{H.73})$$

*Proof.* Using the representation

$$\mathbf{p}_k^\dagger = \mathbf{x}^\dagger - \mathbf{x}_k^\delta = \mathbf{p}_0^\dagger - \sum_{i=0}^{k-1} \mathbf{p}_{\alpha_i i}^\delta, \quad k = 1, \dots, k^*,$$

with  $\alpha_i = \alpha_{j^*(i)}$ , and the relations (cf. (H.56))

$$\mathbf{p}_0^\dagger = \mathbf{K}_0^T \mathbf{w}_0 = \mathbf{K}_k^T \mathbf{Q}(\mathbf{x}_0^\delta, \mathbf{x}_k^\delta)^T \mathbf{w}_0$$

and

$$\mathbf{p}_{\alpha_i i}^\delta = g_{\alpha_i}(\mathbf{K}_i^T \mathbf{K}_i) \mathbf{K}_i^T \mathbf{r}_i^\delta = \mathbf{K}_i^T g_{\alpha_i}(\mathbf{K}_i \mathbf{K}_i^T) \mathbf{r}_i^\delta = \mathbf{K}_k^T \mathbf{Q}(\mathbf{x}_i^\delta, \mathbf{x}_k^\delta)^T g_{\alpha_i}(\mathbf{K}_i \mathbf{K}_i^T) \mathbf{r}_i^\delta,$$

we obtain

$$\mathbf{p}_k^\dagger = \mathbf{K}_k^T \mathbf{Q}(\mathbf{x}_0^\delta, \mathbf{x}_k^\delta)^T \mathbf{w}_0 - \mathbf{K}_k^T \sum_{i=0}^{k-1} \mathbf{Q}(\mathbf{x}_i^\delta, \mathbf{x}_k^\delta)^T g_{\alpha_i}(\mathbf{K}_i \mathbf{K}_i^T) \mathbf{r}_i^\delta, \quad (\text{H.74})$$

and the first assertion is proven. Note that in the derivation of (H.74) we used the identity  $g_\alpha(\mathbf{K}^T \mathbf{K}) \mathbf{K}^T = \mathbf{K}^T g_\alpha(\mathbf{K} \mathbf{K}^T)$ , with  $g_\alpha(\mathbf{K} \mathbf{K}^T)$  being given by (G.31).

The norm of  $\mathbf{w}_k$  can be bounded as

$$\|\mathbf{w}_k\| \leq \left\| \mathbf{Q} (\mathbf{x}_0^\delta, \mathbf{x}_k^\delta)^T \mathbf{w}_0 \right\| + \sum_{i=0}^{k-1} \left\| \mathbf{Q} (\mathbf{x}_i^\delta, \mathbf{x}_k^\delta)^T g_{\alpha_i} (\mathbf{K}_i \mathbf{K}_i^T) \mathbf{r}_i^\delta \right\|;$$

whence, by (H.57) and the result  $\mathbf{r}_i = \mathbf{K}_i \mathbf{p}_i^\dagger = \mathbf{K}_i \mathbf{K}_i^T \mathbf{w}_i$ , we find that

$$\begin{aligned} \|\mathbf{w}_k\| &\leq \bar{c}_Q \left( \|\mathbf{w}_0\| + \sum_{i=0}^{k-1} \|g_{\alpha_i} (\mathbf{K}_i \mathbf{K}_i^T) \mathbf{r}_i^\delta\| \right) \\ &\leq \bar{c}_Q \left( \|\mathbf{w}_0\| + \sum_{i=0}^{k-1} \|g_{\alpha_i} (\mathbf{K}_i \mathbf{K}_i^T) (\mathbf{r}_i^\delta - \mathbf{r}_i)\| + \|g_{\alpha_i} (\mathbf{K}_i \mathbf{K}_i^T) \mathbf{K}_i \mathbf{K}_i^T \mathbf{w}_i\| \right). \end{aligned}$$

Now, by (H.52), there holds

$$\|g_{\alpha_i} (\mathbf{K}_i \mathbf{K}_i^T) \mathbf{K}_i \mathbf{K}_i^T \mathbf{w}_i\| \leq \|\mathbf{w}_i\|,$$

while, for  $\alpha_i = \alpha_{j^*(i)}$ , (H.51) and (H.70) give

$$\|g_{\alpha_i} (\mathbf{K}_i \mathbf{K}_i^T) (\mathbf{r}_i^\delta - \mathbf{r}_i)\| \leq \frac{c_1}{\alpha_i} \|\mathbf{r}_i^\delta - \mathbf{r}_i\| < \frac{c_1 c_2}{q(\tau_i - 1)} \|\mathbf{w}_i\| \leq \frac{c_1 c_2}{q(\tau_{\min} - 1)} \|\mathbf{w}_i\|,$$

where  $\tau_{\min} = \min(\tau_0, \dots, \tau_{k^*-1})$ . Collecting all results we are led to

$$\|\mathbf{w}_k\| < \bar{c}_Q \left( \|\mathbf{w}_0\| + c \sum_{i=0}^{k-1} \|\mathbf{w}_i\| \right),$$

with  $c$  as in (H.73). The assertion (H.72) follows now by an induction argument. Indeed, assuming

$$\|\mathbf{w}_i\| < \bar{c}_Q (1 + c) (1 + c\bar{c}_Q)^{i-1} \|\mathbf{w}_0\|, \quad i = 1, \dots, k,$$

we obtain

$$\begin{aligned} \|\mathbf{w}_{k+1}\| &< \bar{c}_Q \left[ \|\mathbf{w}_0\| + c \sum_{i=0}^k \|\mathbf{w}_i\| \right] \\ &< \bar{c}_Q \left[ (1 + c) \|\mathbf{w}_0\| + c\bar{c}_Q (1 + c) \|\mathbf{w}_0\| \sum_{i=1}^k (1 + c\bar{c}_Q)^{i-1} \right] \\ &= \bar{c}_Q (1 + c) (1 + c\bar{c}_Q)^k \|\mathbf{w}_0\|. \end{aligned}$$

□

In a compact form, the estimate (H.72) can be expressed as

$$\|\mathbf{w}_k\| < c_w \Lambda^k \|\mathbf{w}_0\|, \quad (\text{H.75})$$

with

$$c_w = \frac{\bar{c}_Q (1 + c)}{1 + c\bar{c}_Q}, \quad \Lambda = 1 + c\bar{c}_Q > 1. \quad (\text{H.76})$$

We are now in the position to formulate a convergence rate result.

**Theorem H.8.** For  $\tau > 1$  and  $0 < \eta < 1$ , assume that

$$1 < \Lambda < \frac{1}{\eta} \quad (\text{H.77})$$

and

$$0 < \omega < \frac{\eta}{\eta + \tau + 1} \quad (\text{H.78})$$

are satisfied, and choose the tolerances  $\tau_{\text{dp}}$  and  $\theta_k$  as

$$\tau_{\text{dp}} \geq \frac{\tau(1+\omega)}{\eta - [\eta + (1+\tau)]\omega}, \quad \tau \left[ \omega + (1+\omega) \frac{1}{\tau_{\text{dp}}} \right] < \theta_k \leq \eta - (1+\eta)\omega. \quad (\text{H.79})$$

Suppose that (H.51), (H.52), (H.53) and (H.56) hold, and let the source condition (H.71) be fulfilled for  $\mathbf{x}_0^\delta = \mathbf{x}_a \in B_\rho(\mathbf{x}^\dagger)$ . If  $k^*$  is the termination index of the regularization method and  $\mathbf{x}_k^\delta \in B_\rho(\mathbf{x}^\dagger)$  for all  $k = 1, \dots, k^*$ , then there holds the error estimate

$$\|\mathbf{x}^\dagger - \mathbf{x}_{k^*}^\delta\| = O \left( \|\mathbf{w}_0\|^{\frac{1}{2}} \Delta^{\frac{1 - \log_1/\eta}{2}} \right), \quad (\text{H.80})$$

with

$$0 < \log_{\frac{1}{\eta}} \Lambda < 1. \quad (\text{H.81})$$

*Proof.* For the choice  $\tau > 1$ , assumption (H.78) gives

$$0 < \omega < \frac{\eta}{\eta + \tau + 1} < \frac{\eta}{\eta + 2},$$

while the selection rules (H.79) yield

$$\tau_{\text{dp}} \geq \frac{\tau(1+\omega)}{\eta - [\eta + (1+\tau)]\omega} > \frac{1+\omega}{\eta - (2+\eta)\omega}$$

and

$$\theta_k > \tau \left[ \omega + (1+\omega) \frac{1}{\tau_{\text{dp}}} \right] > \omega + (1+\omega) \frac{1}{\tau_{\text{dp}}}.$$

Hence, assumption (H.65) and the selection rules (H.66) still hold, and Propositions H.4 and H.5 are valid for  $\eta$  satisfying the requirements of the theorem.

The iteration error can be bounded as

$$\|\mathbf{x}^\dagger - \mathbf{x}_k^\delta\|^2 = \|\mathbf{p}_k^\dagger\|^2 = \mathbf{p}_k^{\dagger T} \mathbf{K}_k^T \mathbf{w}_k \leq \|\mathbf{K}_k \mathbf{p}_k^\dagger\| \|\mathbf{w}_k\|.$$

The estimate

$$\|\mathbf{K}_k \mathbf{p}_k^\dagger\| = \|\mathbf{K}_k(\mathbf{x}^\dagger - \mathbf{x}_k^\delta)\| \leq \|\mathbf{F}(\mathbf{x}^\dagger) - \mathbf{F}(\mathbf{x}_k^\delta)\| + \|\mathbf{R}(\mathbf{x}^\dagger, \mathbf{x}_k^\delta)\|$$

together with (H.59) implies that

$$\|\mathbf{K}_k \mathbf{p}_k^\dagger\| \leq (1+\omega) \|\mathbf{y} - \mathbf{F}(\mathbf{x}_k^\delta)\|,$$

and so,

$$\|\mathbf{x}^\dagger - \mathbf{x}_k^\delta\|^2 \leq (1 + \omega) \|\mathbf{w}_k\| \|\mathbf{y} - \mathbf{F}(\mathbf{x}_k^\delta)\| \leq (1 + \omega) \|\mathbf{w}_k\| (\Delta + \|\mathbf{r}_k^\delta\|). \quad (\text{H.82})$$

For  $k = k^*$ , we have  $\|\mathbf{r}_{k^*}^\delta\| \leq \tau_{\text{dp}} \Delta$ , and, by virtue of (H.75), (H.82) becomes

$$\|\mathbf{x}^\dagger - \mathbf{x}_{k^*}^\delta\|^2 < c_w (1 + \omega) (1 + \tau_{\text{dp}}) \|\mathbf{w}_0\| \Lambda^{k^*} \Delta.$$

The estimate of the termination index (H.69) gives

$$\Lambda^{k^*} < \Lambda \Lambda^{\log_\eta \left( \frac{\tau_{\text{dp}} \Delta}{\|\mathbf{r}_0^\delta\|} \right)} = \Lambda \left( \frac{\tau_{\text{dp}} \Delta}{\|\mathbf{r}_0^\delta\|} \right)^{\log_\eta \Lambda},$$

and in view of the identity

$$\log_\eta \Lambda = -\log_{\frac{1}{\eta}} \Lambda,$$

we conclude that (H.80) holds. Since  $0 < \eta < 1$ ,  $\log_{1/\eta}$  is an increasing function and, as a result, (H.77) yields (H.81).  $\square$

It should be remarked that assumption (H.71) gives

$$\mathbf{x}^\dagger - \mathbf{x}_a \in \mathcal{R}(\mathbf{K}_0^T) = \mathcal{R}((\mathbf{K}_0^T \mathbf{K}_0)^{\frac{1}{2}})$$

and represents a source condition imposed on  $\mathbf{x}^\dagger$ . To be more concrete, the existence of  $\mathbf{w}_0 \in \mathbb{R}^m$  so that

$$\mathbf{x}^\dagger - \mathbf{x}_a = \mathbf{K}_0^T \mathbf{w}_0,$$

means that  $\mathbf{x}^\dagger - \mathbf{x}_a$  possesses the representation

$$\mathbf{x}^\dagger - \mathbf{x}_a = \sum_{i=1}^n \sigma_i (\mathbf{u}_i^T \mathbf{w}_0) \mathbf{v}_i,$$

for  $\mathbf{K}_0 = \mathbf{U} \Sigma \mathbf{V}^T$ . Defining  $\mathbf{z} \in \mathbb{R}^n$  by the expansion

$$\mathbf{z} = \sum_{i=1}^n (\mathbf{u}_i^T \mathbf{w}_0) \mathbf{v}_i,$$

which yields

$$\mathbf{v}_i^T \mathbf{z} = \mathbf{u}_i^T \mathbf{w}_0, \quad i = 1, \dots, n,$$

we find that

$$\mathbf{x}^\dagger - \mathbf{x}_a = \sum_{i=1}^n \sigma_i (\mathbf{v}_i^T \mathbf{z}) \mathbf{v}_i = (\mathbf{K}_0^T \mathbf{K}_0)^{\frac{1}{2}} \mathbf{z}.$$

Note that for the general source condition

$$\mathbf{p}_0^\dagger = (\mathbf{K}_0^T \mathbf{K}_0)^\mu \mathbf{z}, \quad \mu > 0, \quad \mathbf{z} \in \mathbb{R}^n,$$

the convergence rate

$$\|\mathbf{x}^\dagger - \mathbf{x}_{k^*}^\delta\| = O \left( \|\mathbf{z}\|^{\frac{1}{2\mu+1}} \Delta^{\frac{2\mu - \log_{1/\eta} \Lambda}{2\mu+1}} \right),$$

has been proven by Rieder (2003).

# I

## Filter factors of the truncated total least squares method

In this appendix we derive the expression of the filter factors for the truncated TLS by following the analysis of Fierro et al. (1997).

Let

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} = \bar{\mathbf{U}} \bar{\Sigma} \bar{\mathbf{V}} \quad (\text{I.1})$$

and

$$\mathbf{K}_\Lambda = \mathbf{U} \Sigma \mathbf{V}^T, \quad (\text{I.2})$$

be the singular value decompositions of the augmented matrix  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  and of the coefficient matrix  $\mathbf{K}_\Lambda$ . First, we first proceed to derive general representations for the singular values  $\bar{\sigma}_j$  and the right singular vectors  $\bar{\mathbf{v}}_j$  of  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  in terms of the singular system  $\{(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)\}$  of  $\mathbf{K}_\Lambda$ . In order not to jumble our presentation with technical details and studies of special cases, we assume that  $\text{rank}(\mathbf{K}_\Lambda) = n$  and  $\text{rank}(\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}) = n + 1$ . Moreover, we suppose that there holds

$$\mathbf{u}_j^T \mathbf{y}^\delta \neq 0, \quad j = 1, \dots, n. \quad (\text{I.3})$$

To derive the desired relationships, we use (I.1) and (I.2) to obtain

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}^T \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} = \bar{\mathbf{V}} \bar{\Sigma}^T \bar{\Sigma} \bar{\mathbf{V}}^T,$$

and

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}^T \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} = \begin{bmatrix} \mathbf{K}_\Lambda^T \mathbf{K}_\Lambda & \mathbf{K}_\Lambda^T \mathbf{y}^\delta \\ \mathbf{y}^{\delta T} \mathbf{K}_\Lambda & \|\mathbf{y}^\delta\|^2 \end{bmatrix} = \bar{\bar{\mathbf{V}}} \bar{\mathbf{S}} \bar{\bar{\mathbf{V}}}^T,$$

with

$$\bar{\bar{\mathbf{V}}} = \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \quad (\text{I.4})$$

and

$$\bar{\mathbf{S}} = \begin{bmatrix} \Sigma^T \Sigma & \Sigma^T \mathbf{U}^T \mathbf{y}^\delta \\ \mathbf{y}^{\delta T} \mathbf{U} \Sigma & \|\mathbf{y}^\delta\|^2 \end{bmatrix}.$$

Performing a singular value decomposition of the positive definite matrix  $\mathbf{S}$ , which we write as

$$\mathbf{S} = \mathbf{V}_s \Sigma_s^T \Sigma_s \mathbf{V}_s^T, \quad \Sigma_s^T \Sigma_s = \left[ \text{diag}(\sigma_{sj}^2)_{(n+1) \times (n+1)} \right], \quad (\text{I.5})$$

we find that

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}^T \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} = \bar{\bar{\mathbf{V}}} \mathbf{V}_s \Sigma_s^T \Sigma_s \mathbf{V}_s^T \bar{\bar{\mathbf{V}}}^T.$$

Thus,  $\bar{\Sigma}^T \bar{\Sigma} = \Sigma_s^T \Sigma_s$  and  $\bar{\mathbf{V}} = \bar{\bar{\mathbf{V}}} \mathbf{V}_s$ . Explicitly, we have

$$\bar{\sigma}_j = \sigma_{sj}, \quad j = 1, \dots, n+1, \quad (\text{I.6})$$

and

$$\bar{\mathbf{v}}_j = \bar{\bar{\mathbf{V}}} \mathbf{v}_{sj}, \quad j = 1, \dots, n+1, \quad (\text{I.7})$$

where the  $\mathbf{v}_{sj}$  are the column vectors of  $\mathbf{V}_s$ .

In the next step of our analysis, we write  $\mathbf{S}$  as

$$\mathbf{S} = \begin{bmatrix} \sigma_1^2 & \dots & 0 & \sigma_1 \mathbf{u}_1^T \mathbf{y}^\delta \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \sigma_n^2 & \sigma_n \mathbf{u}_n^T \mathbf{y}^\delta \\ \sigma_1 \mathbf{u}_1^T \mathbf{y}^\delta & \dots & \sigma_n \mathbf{u}_n^T \mathbf{y}^\delta & \|\mathbf{y}^\delta\|^2 \end{bmatrix} \quad (\text{I.8})$$

and express (I.5) as

$$\mathbf{S} \mathbf{v}_{sj} = \sigma_{sj}^2 \mathbf{v}_{sj}, \quad j = 1, \dots, n+1. \quad (\text{I.9})$$

Then, from (I.8) and (I.9) we obtain

$$(\sigma_{sj}^2 - \sigma_i^2) [\mathbf{v}_{sj}]_i = \sigma_i (\mathbf{u}_i^T \mathbf{y}^\delta) [\mathbf{v}_{sj}]_{n+1}, \quad i = 1, \dots, n, \quad (\text{I.10})$$

$$\sum_{i=1}^n \sigma_i (\mathbf{u}_i^T \mathbf{y}^\delta) [\mathbf{v}_{sj}]_i = (\sigma_{sj}^2 - \|\mathbf{y}^\delta\|^2) [\mathbf{v}_{sj}]_{n+1}. \quad (\text{I.11})$$

The singular system of the matrix  $\mathbf{S}$  has two interesting features, namely, for any  $j = 1, \dots, n+1$ ,

- (1)  $[\mathbf{v}_{sj}]_{n+1} \neq 0$ ;
- (2)  $\sigma_{sj}$  does not coincide with a singular value  $\sigma_i$  of  $\mathbf{K}_\Lambda$ .

To prove the first assertion, we assume that  $[\mathbf{v}_{sj}]_{n+1} = 0$ . In this case, two situations can be distinguished:

- (1) Suppose that there exist  $i_1$  and  $i_2$  such that  $[\mathbf{v}_{sj}]_{i_1} \neq 0$  and  $[\mathbf{v}_{sj}]_{i_2} \neq 0$ . Then, from (I.10), it follows that  $\sigma_{i_1} = \sigma_{i_2} = \sigma_{sj}$ , and this result is contradictory to our assumption that the singular values of  $\mathbf{K}_\Lambda$  are simple.
- (2) Suppose that there exists  $i$  such that  $[\mathbf{v}_{sj}]_i \neq 0$  and that  $[\mathbf{v}_{sj}]_l = 0$  for all  $l \neq i$ . From (I.11) and the assumption  $\mathbf{u}_i^T \mathbf{y}^\delta \neq 0$ , we deduce that  $\sigma_i = 0$ . Since by assumption  $\text{rank}(\mathbf{K}_\Lambda) = n$ , we are again faced with a contradiction.

Hence,  $[\mathbf{v}_{sj}]_{n+1} \neq 0$ . Turning now to the second assertion we assume that there exists  $i$  such that  $\sigma_i = \sigma_{sj}$ . This yields  $\sigma_i (\mathbf{u}_i^T \mathbf{y}^\delta) [\mathbf{v}_{sj}]_{n+1} = 0$ , and, since  $\mathbf{u}_i^T \mathbf{y}^\delta \neq 0$  and



$[\mathbf{v}_{sj}]_{n+1} \neq 0$ , we are led to the contradictory result  $\sigma_i = 0$ . Thus,  $\sigma_{sj}$  does not coincide with a singular value  $\sigma_i$  of  $\mathbf{K}_\Lambda$ , and we have

$$[\mathbf{v}_{sj}]_i = \frac{\sigma_i}{\sigma_{sj}^2 - \sigma_i^2} (\mathbf{u}_i^T \mathbf{y}^\delta) [\mathbf{v}_{sj}]_{n+1}, \quad i = 1, \dots, n.$$

The second assertion above together with (I.6) implies that the interlacing inequalities for the singular values of  $\mathbf{K}_\Lambda$  and  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  are strict, i.e.,

$$\bar{\sigma}_1 > \sigma_1 > \dots > \bar{\sigma}_p > \sigma_p > \bar{\sigma}_{p+1} > \sigma_{p+1} > \dots > \sigma_n > \bar{\sigma}_{n+1}, \quad (\text{I.12})$$

where  $p$  is the truncation index of the truncated TLS method.

We are now in the position to derive a final expression for the right singular vectors of  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$ . By (I.4) and (I.7), we have

$$\bar{\mathbf{v}}_j = \bar{\bar{\mathbf{V}}} \mathbf{v}_{sj} = \begin{bmatrix} \mathbf{V} \begin{bmatrix} [\mathbf{v}_{sj}]_1 \\ \vdots \\ [\mathbf{v}_{sj}]_n \end{bmatrix} \\ [\mathbf{v}_{sj}]_{n+1} \end{bmatrix},$$

and the entries of the right singular vectors  $\bar{\mathbf{v}}_j$  are given by

$$\begin{bmatrix} [\bar{\mathbf{v}}_j]_1 \\ \vdots \\ [\bar{\mathbf{v}}_j]_n \end{bmatrix} = \sum_{i=1}^n \frac{\sigma_i}{\sigma_{sj}^2 - \sigma_i^2} (\mathbf{u}_i^T \mathbf{y}^\delta) [\mathbf{v}_{sj}]_{n+1} \mathbf{v}_i, \quad (\text{I.13})$$

and

$$[\bar{\mathbf{v}}_j]_{n+1} = [\mathbf{v}_{sj}]_{n+1}, \quad (\text{I.14})$$

for  $j = 1, \dots, n+1$ . We summarize the above results in the following theorem.

**Theorem I.1.** *Let (I.2) be the singular value decomposition of the coefficient matrix  $\mathbf{K}_\Lambda$  and suppose that  $\text{rank}(\mathbf{K}_\Lambda) = n$  and  $\text{rank}(\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}) = n+1$ . Furthermore, assume that (I.3) holds. If (I.5) is the singular value decomposition of the matrix  $\mathbf{S}$  defined by (I.8), then the singular values of the augmented matrix  $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$  are given by (I.6), while the entries of the right singular vectors are given by (I.13) and (I.14).*

Next, we proceed to derive the filter factors for the truncated TLS solution

$$\mathbf{x}_{\Lambda p}^\delta = -\frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \bar{\mathbf{V}}_{12} \bar{\mathbf{v}}_{22}, \quad (\text{I.15})$$

where

$$\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_{n+1}] = \begin{bmatrix} \bar{\mathbf{V}}_{11} & \bar{\mathbf{V}}_{12} \\ \bar{\mathbf{v}}_{21}^T & \bar{\mathbf{v}}_{22}^T \end{bmatrix}, \quad (\text{I.16})$$

$\bar{\mathbf{V}}_{11} \in \mathbb{R}^{n \times p}$ ,  $\bar{\mathbf{V}}_{12} \in \mathbb{R}^{n \times (n-p+1)}$ , and

$$\begin{aligned} \bar{\mathbf{v}}_{21} &= \left[ [\bar{\mathbf{v}}_1]_{n+1}, \dots, [\bar{\mathbf{v}}_p]_{n+1} \right]^T \in \mathbb{R}^p, \\ \bar{\mathbf{v}}_{22} &= \left[ [\bar{\mathbf{v}}_{p+1}]_{n+1}, \dots, [\bar{\mathbf{v}}_{n+1}]_{n+1} \right]^T \in \mathbb{R}^{n-p+1}. \end{aligned}$$

**Theorem I.2.** *Under the same assumptions as in Theorem I.1, the filter factors for the truncated TLS solution are given by*

$$f_i = -\frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=p+1}^{n+1} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2 = \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=1}^p \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2 \quad (\text{I.17})$$

for  $i = 1, \dots, n$ .

*Proof.* Using (I.13) together with (I.6) and (I.14), we express the truncated TLS solution (I.15) as

$$\begin{aligned} \mathbf{x}_{\Lambda p}^\delta &= -\frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \bar{\mathbf{V}}_{12} \bar{\mathbf{v}}_{22} \\ &= -\frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=p+1}^{n+1} [\bar{\mathbf{v}}_j]_{n+1} \begin{bmatrix} [\bar{\mathbf{v}}_j]_1 \\ \vdots \\ [\bar{\mathbf{v}}_j]_n \end{bmatrix} \\ &= -\frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{i=1}^n \left( \sum_{j=p+1}^{n+1} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2 \right) \frac{1}{\sigma_i} (\mathbf{u}_i^T \mathbf{y}^\delta) \mathbf{v}_i. \end{aligned}$$

The first representation in (I.17) is then apparent. To derive the second representation in (I.17), we first use the orthogonality relation  $\mathbf{V}_s \mathbf{V}_s^T = \mathbf{I}_{n+1}$  to obtain

$$\sum_{j=1}^{n+1} \mathbf{v}_{sj} \mathbf{v}_{sj}^T = \mathbf{I}_{n+1}.$$

This gives

$$\sum_{j=1}^{n+1} [\mathbf{v}_{sj}]_i [\mathbf{v}_{sj}]_{n+1} = 0, \quad i = 1, \dots, n \quad (\text{I.18})$$

and

$$\sum_{j=1}^{n+1} [\mathbf{v}_{sj}]_{n+1}^2 = 1. \quad (\text{I.19})$$

On the other hand, from (I.10) in conjunction with (I.6), we have

$$\sigma_i (\mathbf{u}_i^T \mathbf{y}^\delta) \frac{[\mathbf{v}_{sj}]_{n+1}^2}{\bar{\sigma}_j^2 - \sigma_i^2} = [\mathbf{v}_{sj}]_i [\mathbf{v}_{sj}]_{n+1}, \quad i = 1, \dots, n. \quad (\text{I.20})$$

Now, (I.18) and (I.20) together with (I.3) and (I.14), yield

$$\sum_{j=1}^{n+1} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2 = 0, \quad i = 1, \dots, n,$$

and so,

$$\sum_{j=p+1}^{n+1} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2 = -\sum_{j=1}^p \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2, \quad i = 1, \dots, n.$$

The proof of the theorem is now complete.  $\square$

The filter factors of the truncated TLS can be bounded as follows.

**Theorem I.3.** *Under the same assumptions as in Theorem I.1, the filter factors satisfy*

$$0 < f_i - 1 \leq \frac{\bar{\sigma}_{p+1}^2}{\sigma_i^2 - \bar{\sigma}_{p+1}^2}, \quad i = 1, \dots, p \quad (\text{I.21})$$

and

$$0 < f_i \leq \frac{1 - \|\bar{\mathbf{v}}_{22}\|^2}{\|\bar{\mathbf{v}}_{22}\|^2} \frac{\sigma_i^2}{\bar{\sigma}_p^2 - \sigma_i^2}, \quad i = p+1, \dots, n. \quad (\text{I.22})$$

*Proof.* For  $i = 1, \dots, p$ , we use the first representation in (I.17) and the result

$$\|\bar{\mathbf{v}}_{22}\|^2 = \sum_{j=p+1}^{n+1} [\bar{\mathbf{v}}_j]_{n+1}^2 \quad (\text{I.23})$$

to obtain

$$\begin{aligned} f_i &= \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=p+1}^{n+1} \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} [\bar{\mathbf{v}}_j]_{n+1}^2 \\ &= 1 + \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=p+1}^{n+1} \frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} [\bar{\mathbf{v}}_j]_{n+1}^2. \end{aligned} \quad (\text{I.24})$$

From the interlacing inequalities for the singular values of  $\mathbf{K}_\Lambda$  and  $[\mathbf{K}_\Lambda \quad \mathbf{y}^\delta]$  given by (I.12), we see that, for  $i = 1, \dots, p$ , we have  $\sigma_i > \bar{\sigma}_{p+1} = \max_{j=p+1, n+1}(\bar{\sigma}_j)$ . Hence, the second term in (I.24) is positive and the left inequality in (I.21) holds true. Going further, from  $\bar{\sigma}_j \leq \bar{\sigma}_{p+1}$  for  $j = p+1, \dots, n+1$ , we deduce that

$$\frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} \leq \frac{\bar{\sigma}_{p+1}^2}{\sigma_i^2 - \bar{\sigma}_{p+1}^2},$$

and so,

$$\sum_{j=p+1}^{n+1} \frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} [\bar{\mathbf{v}}_j]_{n+1}^2 \leq \frac{\bar{\sigma}_{p+1}^2}{\sigma_i^2 - \bar{\sigma}_{p+1}^2} \sum_{j=p+1}^{n+1} [\bar{\mathbf{v}}_j]_{n+1}^2.$$

This result together with (I.23) yields the right inequality in (I.21).

For  $i = p+1, \dots, n$ , we consider the second representation in (I.17), that is,

$$f_i = \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=1}^p \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2.$$

From (I.12), we see that, for  $i = p+1, \dots, n$ , we have  $\sigma_i < \bar{\sigma}_p = \min_{j=1, p}(\bar{\sigma}_j)$ , and therefore, the left inequality in (I.22) is satisfied. Further, from  $\bar{\sigma}_j \geq \bar{\sigma}_p$  for  $j = 1, \dots, p$ , we find that

$$f_i = \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \sum_{j=1}^p \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} [\bar{\mathbf{v}}_j]_{n+1}^2 \leq \frac{1}{\|\bar{\mathbf{v}}_{22}\|^2} \frac{\sigma_i^2}{\bar{\sigma}_p^2 - \sigma_i^2} \sum_{j=1}^p [\bar{\mathbf{v}}_j]_{n+1}^2.$$

Finally, from (I.14) and (I.19) we obtain

$$\sum_{j=1}^p [\bar{\mathbf{v}}_j]_{n+1}^2 = 1 - \sum_{j=p+1}^n [\bar{\mathbf{v}}_j]_{n+1}^2 ,$$

and (I.23) can now be used to conclude. □

# J

## Quadratic programming

In this appendix we analyze methods for solving quadratic programming problems. For equality constraints, we present the basic concepts of the null-space method, while for inequality constraints, we discuss a dual active set method. The theory is based on the works of Gill et al. (1981), Nocedal and Wright (2006), and Goldfarb and Idnani (1983).

### J.1 Equality constraints

Let us consider the equality-constrained quadratic programming problem

$$(P) : \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{G} \mathbf{x} + \mathbf{g}^T \mathbf{x} \quad (\text{J.1})$$
$$\text{subject to } \mathbf{A} \mathbf{x} = \mathbf{b}, \quad (\text{J.2})$$

where  $f$  is the objective function,  $\mathbf{G} \in \mathbb{R}^{n \times n}$  is a positive definite matrix,  $\mathbf{A} \in \mathbb{R}^{r \times n}$  is the constraint matrix,  $n$  is the number of variables, and  $r$  is the number of constraints. The  $i$ th row of  $\mathbf{A}$  contains the coefficients corresponding to the  $i$ th constraint and we assume that the row vectors of  $\mathbf{A}$  are linearly independent. If the row vectors of  $\mathbf{A}$  are linearly dependent then either some constraints can be omitted without changing the solution, or there is no feasible point. A point  $\mathbf{x}$  is said to be feasible if  $\mathbf{A} \mathbf{x} = \mathbf{b}$ .

Let  $\mathbf{Z} \in \mathbb{R}^{n \times (n-r)}$  be a matrix whose column vectors form a basis for the null space of  $\mathbf{A}$ , and let  $\mathbf{Y} \in \mathbb{R}^{n \times r}$  be a matrix whose column vectors form a basis for the range space of  $\mathbf{A}^T$ . Then, any  $n$ -dimensional vector  $\mathbf{x}$  can be expressed as a linear combination of the column vectors of  $\mathbf{Y}$  and  $\mathbf{Z}$ , i.e.,

$$\mathbf{x} = \mathbf{Y} \mathbf{x}_Y + \mathbf{Z} \mathbf{x}_Z, \quad (\text{J.3})$$

with  $\mathbf{x}_Y \in \mathbb{R}^r$  and  $\mathbf{x}_Z \in \mathbb{R}^{n-r}$ . Using the orthogonality relation  $\mathbf{A} \mathbf{Z} = \mathbf{0}$ , we find that

$$\mathbf{A} \mathbf{x} = \mathbf{A} \mathbf{Y} \mathbf{x}_Y,$$

and, if  $\mathbf{x}$  is feasible, that

$$\mathbf{A} \mathbf{Y} \mathbf{x}_Y = \mathbf{b}. \quad (\text{J.4})$$

As  $\mathbf{AY}$  is nonsingular by construction,  $\mathbf{x}_Y$  is uniquely determined by (J.4). Thus, the range-space component  $\mathbf{x}_Y$  is completely determined by the constraints, while the null-space component  $\mathbf{x}_Z$  has to be computed by minimizing the objective function  $f$ . Now, let  $\bar{\mathbf{x}}$  be a feasible point and let  $\bar{\mathbf{p}}$  be the step to the solution  $\mathbf{x}^*$ , i.e.,

$$\mathbf{x}^* = \bar{\mathbf{x}} + \bar{\mathbf{p}}. \quad (\text{J.5})$$

Substituting  $\mathbf{x}$  with  $\bar{\mathbf{x}} + \mathbf{p}$  in (J.1), we deduce that  $\bar{\mathbf{p}}$  solves the equality-constrained quadratic programming problem

$$\min_{\mathbf{p} \in \mathbb{R}^n} \frac{1}{2} \mathbf{p}^T \mathbf{G} \mathbf{p} + \bar{\mathbf{g}}^T \mathbf{p} \quad (\text{J.6})$$

$$\text{subject to } \mathbf{A} \mathbf{p} = \mathbf{0}, \quad (\text{J.7})$$

where

$$\bar{\mathbf{g}} = \mathbf{G} \bar{\mathbf{x}} + \mathbf{g}$$

is the gradient of  $f$  at  $\bar{\mathbf{x}}$ . Setting  $\mathbf{p} = \mathbf{Y} \mathbf{p}_Y + \mathbf{Z} \mathbf{p}_Z$ , then from  $\mathbf{AY} \mathbf{p}_Y = \mathbf{0}$ , we get  $\mathbf{p}_Y = \mathbf{0}$ , and therefore,  $\mathbf{p}$  is a linear combination of the column vectors of  $\mathbf{Z}$ , that is,

$$\mathbf{p} = \mathbf{Z} \mathbf{p}_Z.$$

In this regard, the  $n$ -dimensional constrained minimization problem (J.6)–(J.7) is equivalent to the  $(n - r)$ -dimensional unconstrained minimization problem

$$\min_{\mathbf{p}_Z \in \mathbb{R}^{n-r}} \frac{1}{2} \mathbf{p}_Z^T \mathbf{Z}^T \mathbf{G} \mathbf{Z} \mathbf{p}_Z + \bar{\mathbf{g}}^T \mathbf{Z} \mathbf{p}_Z. \quad (\text{J.8})$$

The solution of (J.8) is defined by the linear system

$$\mathbf{Z}^T \mathbf{G} \mathbf{Z} \bar{\mathbf{p}}_Z = -\mathbf{Z}^T \bar{\mathbf{g}},$$

and we obtain

$$\bar{\mathbf{p}} = \mathbf{Z} \bar{\mathbf{p}}_Z = -\mathbf{Z} (\mathbf{Z}^T \mathbf{G} \mathbf{Z})^{-1} \mathbf{Z}^T \bar{\mathbf{g}}. \quad (\text{J.9})$$

By (J.5) and (J.9), it is readily seen that the computation of the solution requires the knowledge of a feasible point  $\bar{\mathbf{x}}$  and of the matrix  $\mathbf{Z}$ . To compute these quantities, two techniques can be employed.

(1) *QR factorization.* Let us consider the QR factorization of  $\mathbf{A}^T$ ,

$$\mathbf{A}^T = \mathbf{Q} \begin{bmatrix} \mathbf{R}^T \\ \mathbf{0} \end{bmatrix} = [\mathbf{Y} \quad \mathbf{Z}] \begin{bmatrix} \mathbf{R}^T \\ \mathbf{0} \end{bmatrix}, \quad (\text{J.10})$$

where  $\mathbf{Y} \in \mathbb{R}^{n \times r}$  and  $\mathbf{Z} \in \mathbb{R}^{n \times (n-r)}$  have orthonormal columns and  $\mathbf{R} \in \mathbb{R}^{r \times r}$  is a nonsingular lower triangular matrix. The column vectors of  $\mathbf{Y}$  are an orthonormal basis of  $\mathcal{R}(\mathbf{A}^T)$ , the column vectors of  $\mathbf{Z}$  are an orthonormal basis of  $\mathcal{N}(\mathbf{A})$ , and we have (cf. (B.7))

$$\mathbf{AY} = \mathbf{R}, \quad \mathbf{AZ} = \mathbf{0}.$$

Then,  $\mathbf{x}_Y^*$  solving (J.4) is defined by  $\mathbf{R} \mathbf{x}_Y^* = \mathbf{b}$ , and the initial feasible point can be taken as  $\bar{\mathbf{x}} = \mathbf{Y} \mathbf{x}_Y^*$ .

(2) *Variable-reduction technique.* Assuming the partitions

$$\mathbf{A} = \begin{bmatrix} \mathbf{V} & \mathbf{U} \end{bmatrix},$$

and

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_V \\ \mathbf{x}_U \end{bmatrix},$$

with  $\mathbf{V} \in \mathbb{R}^{r \times r}$  being a nonsingular matrix, we see that (J.2) yields

$$\mathbf{V}\mathbf{x}_V + \mathbf{U}\mathbf{x}_U = \mathbf{b},$$

and further,

$$\mathbf{x}_V = \mathbf{V}^{-1}(\mathbf{b} - \mathbf{U}\mathbf{x}_U).$$

Thus, any feasible point can be expressed as

$$\begin{bmatrix} \mathbf{V}^{-1}(\mathbf{b} - \mathbf{U}\mathbf{x}_U) \\ \mathbf{x}_U \end{bmatrix},$$

and one possible choice is

$$\bar{\mathbf{x}} = \begin{bmatrix} \mathbf{V}^{-1}\mathbf{b} \\ \mathbf{0} \end{bmatrix}.$$

By straightforward calculation it can be shown that the matrix  $\mathbf{Z}$  defined by

$$\mathbf{Z} = \begin{bmatrix} -\mathbf{V}^{-1}\mathbf{U} \\ \mathbf{I}_{n-r} \end{bmatrix}$$

satisfies the orthogonality relation  $\mathbf{AZ} = \mathbf{0}$ .

The above approach for solving the quadratic programming problem is known as the null-space method. An alternative approach is the range-space method, which is described in the next section. Both techniques can be regarded as solution methods for the so-called Kuhn–Tucker system of equations, and in order to evidence their similarity, we reformulate the null-space method in this new framework. For the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^T \mathbf{G}\mathbf{x} + \mathbf{g}^T \mathbf{x} + \mathbf{u}^T (\mathbf{A}\mathbf{x} - \mathbf{b}),$$

with  $\mathbf{u} \in \mathbb{R}^r$  being the vector of Lagrange multipliers, the first-order optimality conditions

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{u}) = \mathbf{0},$$

$$\nabla_{\mathbf{u}} \mathcal{L}(\mathbf{x}, \mathbf{u}) = \mathbf{0},$$

lead to the Kuhn–Tucker system of equations

$$\begin{bmatrix} \mathbf{G} & \mathbf{A}^T \\ -\mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = - \begin{bmatrix} \mathbf{g} \\ \mathbf{b} \end{bmatrix}. \quad (\text{J.11})$$

Assuming the QR factorization (J.10), we express  $\mathbf{x}$  as in (J.3) and obtain

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{Y} & \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_r \end{bmatrix} \begin{bmatrix} \mathbf{x}_Y \\ \mathbf{x}_Z \\ \mathbf{u} \end{bmatrix}.$$

The Kuhn–Tucker system of equations can be transformed as

$$\begin{bmatrix} \mathbf{Y} & \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_r \end{bmatrix}^T \begin{bmatrix} \mathbf{G} & \mathbf{A}^T \\ -\mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Y} & \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_r \end{bmatrix} \begin{bmatrix} \mathbf{x}_Y \\ \mathbf{x}_Z \\ \mathbf{u} \end{bmatrix} = - \begin{bmatrix} \mathbf{Y} & \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_r \end{bmatrix}^T \begin{bmatrix} \mathbf{g} \\ \mathbf{b} \end{bmatrix},$$

or further, using the identities  $\mathbf{AZ} = \mathbf{0}$  and  $\mathbf{AY} = \mathbf{R}$ , as

$$\begin{bmatrix} \mathbf{Y}^T \mathbf{G} \mathbf{Y} & \mathbf{Y}^T \mathbf{G} \mathbf{Z} & \mathbf{R}^T \\ \mathbf{Z}^T \mathbf{G} \mathbf{Y} & \mathbf{Z}^T \mathbf{G} \mathbf{Z} & \mathbf{0} \\ -\mathbf{R} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_Y \\ \mathbf{x}_Z \\ \mathbf{u} \end{bmatrix} = - \begin{bmatrix} \mathbf{Y}^T \mathbf{g} \\ \mathbf{Z}^T \mathbf{g} \\ \mathbf{b} \end{bmatrix}. \quad (\text{J.12})$$

To solve (J.12), we proceed by backward substitution; the last and the middle block equations give

$$\mathbf{R} \mathbf{x}_Y = \mathbf{b}, \quad (\text{J.13})$$

and

$$\mathbf{Z}^T \mathbf{G} \mathbf{Z} \mathbf{x}_Z = -\mathbf{Z}^T (\mathbf{G} \mathbf{Y} \mathbf{x}_Y + \mathbf{g}), \quad (\text{J.14})$$

respectively, while the first block equation yields (cf. (J.3))

$$\mathbf{R}^T \mathbf{u} = -\mathbf{Y}^T (\mathbf{G} \mathbf{x} + \mathbf{g}).$$

The solution given by (J.3), (J.13) and (J.14), coincides with the solution given by (J.5) with the step as in (J.9) and the feasible point computed by the QR factorization of  $\mathbf{A}^T$ . As  $\mathbf{R}$  is lower triangular, the system of equations (J.13) is solved by forward substitution. Similarly, as  $\mathbf{Z}^T \mathbf{G} \mathbf{Z}$  is positive definite, the solution to the system of equations (J.14) is found by first considering a Cholesky factorization of the matrix  $\mathbf{Z}^T \mathbf{G} \mathbf{Z}$  and then by solving the resulting triangular systems of equations by backward and forward substitutions.

## J.2 Inequality constraints

Let us consider the inequality-constrained quadratic programming problem

$$(P_R) : \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{G} \mathbf{x} + \mathbf{g}^T \mathbf{x} \quad (\text{J.15})$$

$$\text{subject to } \mathbf{A} \mathbf{x} \leq \mathbf{b}, \quad (\text{J.16})$$

where as before,  $\mathbf{G} \in \mathbb{R}^{n \times n}$  is a positive definite matrix,  $\mathbf{A} \in \mathbb{R}^{r \times n}$  is the constraint matrix, and  $R = \{1, \dots, r\}$  is index set of the constraints. In general, the vector inequality  $\mathbf{x} \leq \mathbf{0}$  means that all entries of the vector  $\mathbf{x}$  are non-positive. The matrix  $\mathbf{A}$  is partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_r^T \end{bmatrix},$$

in which case, the  $i$ th row vector  $\mathbf{a}_i^T$  contains the coefficients corresponding to the  $i$ th constraint. At a feasible point  $\mathbf{x}$ , the constraint  $\mathbf{a}_i^T \mathbf{x} \leq [\mathbf{b}]_i$  is said to be active (or binding)



if  $\mathbf{a}_i^T \mathbf{x} = [\mathbf{b}]_i$ , and inactive if  $\mathbf{a}_i^T \mathbf{x} < [\mathbf{b}]_i$ . If the constraint is active or inactive, then the constraint is said to be satisfied. By contrast, the constraint is said to be violated at  $\mathbf{x}$ , if  $\mathbf{a}_i^T \mathbf{x} > [\mathbf{b}]_i$ .

For the quadratic programming problem  $(P_R)$ , the corresponding Lagrangian function is given by

$$\mathcal{L}(\mathbf{x}, \mathbf{u}) = \frac{1}{2} \mathbf{x}^T \mathbf{G} \mathbf{x} + \mathbf{g}^T \mathbf{x} + \mathbf{u}^T (\mathbf{A} \mathbf{x} - \mathbf{b}), \quad (\text{J.17})$$

where  $\mathbf{u} \in \mathbb{R}^r$  is the vector of Lagrange multipliers. The next result, known as the Kuhn-Tucker theorem, states the necessary and sufficient conditions for  $\mathbf{x}^*$  to solve  $(P_R)$ .

**Theorem J.1.** *Let  $\mathbf{x}^*$  solve  $(P_R)$ . Then, there exists a vector of Lagrange multipliers  $\mathbf{u}^*$  such that the Kuhn-Tucker conditions*

$$\mathbf{G} \mathbf{x}^* + \mathbf{g} + \mathbf{A}^T \mathbf{u}^* = \mathbf{0}, \quad (\text{J.18})$$

$$\mathbf{A} \mathbf{x}^* - \mathbf{b} \leq \mathbf{0}, \quad (\text{J.19})$$

$$\mathbf{u}^* \geq \mathbf{0}, \quad (\text{J.20})$$

$$[\mathbf{u}^*]_i (\mathbf{a}_i^T \mathbf{x}^* - [\mathbf{b}]_i) = 0, \quad i = 1, \dots, r, \quad (\text{J.21})$$

are fulfilled. Conversely, let  $\mathbf{G}$  be a positive definite matrix, and suppose that for some feasible point  $\mathbf{x}^*$  there exists a vector of Lagrange multipliers  $\mathbf{u}^*$  such that the Kuhn-Tucker conditions (J.18)–(J.21) are satisfied. Then,  $\mathbf{x}^*$  solves  $(P_R)$ .

The conditions (J.21) are complementary conditions and just say that either the constraint  $i$  is active or  $[\mathbf{u}^*]_i = 0$ , or possibly both. Obviously, (J.21) yields

$$\mathbf{u}^{*T} (\mathbf{A} \mathbf{x}^* - \mathbf{b}) = 0, \quad (\text{J.22})$$

and note that the Lagrange multipliers corresponding to inactive constraints are zero. Also note that due to the positive definiteness of  $\mathbf{G}$ ,  $\mathbf{x}^*$  is the unique solution of  $(P_R)$ .

The quadratic programming problem (J.15)–(J.16) can be solved by using primal and dual active set methods. In this appendix we present the dual active set method of Goldfarb and Idnani (1983). This method does not have the possibility of cycling and benefits from having an easily calculated feasible starting point.

In general, given the optimization problem  $(P)$  (the primal problem), we can define a related problem  $(D)$  (the dual problem) such that the Lagrange multipliers of  $(P)$  are part of the solution of  $(D)$ . For the quadratic programming problem  $(P_R)$ , the so-called Wolf dual problem can be stated as

$$(D_R) : \max_{\mathbf{u} \in \mathbb{R}^r} \max_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \mathbf{u}) \quad (\text{J.23})$$

$$\text{subject to } \mathbf{G} \mathbf{x} + \mathbf{g} + \mathbf{A}^T \mathbf{u} = \mathbf{0}, \quad (\text{J.24})$$

$$\mathbf{u} \geq \mathbf{0}. \quad (\text{J.25})$$

The following result explains the relationship between the primal and the dual problems.

**Theorem J.2.** *Let  $\mathbf{G}$  be a positive definite matrix and let  $\mathbf{x}^*$  solve  $(P_R)$ . If  $(\mathbf{x}^*, \mathbf{u}^*)$  satisfies the Kuhn-Tucker conditions (J.18)–(J.21), then  $(\mathbf{x}^*, \mathbf{u}^*)$  solves  $(D_R)$ , and conversely.*

*Proof.* The proof relies on the inequality

$$\mathcal{L}(\mathbf{x}_1, \mathbf{u}) \geq \mathcal{L}(\mathbf{x}_2, \mathbf{u}) + (\mathbf{G}\mathbf{x}_2 + \mathbf{g} + \mathbf{A}^T \mathbf{u})^T (\mathbf{x}_1 - \mathbf{x}_2), \quad \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n, \quad (\text{J.26})$$

which is a consequence of the positive definiteness of  $\mathbf{G}$ , i.e.,

$$(\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{G} (\mathbf{x}_1 - \mathbf{x}_2) > 0, \quad \mathbf{x}_1 \neq \mathbf{x}_2.$$

Note that the inequality (J.26) is strict whenever  $\mathbf{x}_1 \neq \mathbf{x}_2$ . Let  $(\mathbf{x}^*, \mathbf{u}^*)$  satisfy the Kuhn-Tucker conditions (J.18)–(J.21), and let  $(\mathbf{x}, \mathbf{u})$  be a pair satisfying the constraints (J.24) and (J.25). Then, condition (J.22) gives  $\mathcal{L}(\mathbf{x}^*, \mathbf{u}^*) = f(\mathbf{x}^*)$ , and we have

$$\begin{aligned} \mathcal{L}(\mathbf{x}^*, \mathbf{u}^*) &= \frac{1}{2} \mathbf{x}^{*T} \mathbf{G} \mathbf{x}^* + \mathbf{g}^T \mathbf{x}^* \\ &\geq \frac{1}{2} \mathbf{x}^{*T} \mathbf{G} \mathbf{x}^* + \mathbf{g}^T \mathbf{x}^* + \mathbf{u}^T (\mathbf{A} \mathbf{x}^* - \mathbf{b}) \\ &= \mathcal{L}(\mathbf{x}^*, \mathbf{u}) \\ &\geq \mathcal{L}(\mathbf{x}, \mathbf{u}) + (\mathbf{G}\mathbf{x} + \mathbf{g} + \mathbf{A}^T \mathbf{u})^T (\mathbf{x}^* - \mathbf{x}) \\ &= \mathcal{L}(\mathbf{x}, \mathbf{u}). \end{aligned}$$

The first inequality follows from (J.19) and (J.25), which give  $\mathbf{u}^T (\mathbf{A} \mathbf{x}^* - \mathbf{b}) \leq 0$ , the second inequality follows from (J.26), while the last equality is a consequence of (J.24). Thus,  $(\mathbf{x}^*, \mathbf{u}^*)$  maximizes  $\mathcal{L}$  over the constraints (J.24) and (J.25), and so,  $(\mathbf{x}^*, \mathbf{u}^*)$  solves  $(D_R)$ .

To prove the converse result we proceed by contradiction. Let  $(\mathbf{x}^*, \mathbf{u}^*)$  solve  $(D_R)$  and let us assume that  $\bar{\mathbf{x}}$  solves  $(P_R)$ . Moreover, let us suppose that  $\mathbf{x}^* \neq \bar{\mathbf{x}}$ . By Theorem J.1, we know there exists the vector of Lagrange multipliers  $\bar{\mathbf{u}}$  such that the pair  $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$  satisfies the Kuhn-Tucker conditions (J.18)–(J.21). Consequently, by the direct theorem, it is apparent that  $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$  also solves  $(D_R)$ , and we have

$$\mathcal{L}(\mathbf{x}^*, \mathbf{u}^*) = \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{u}}). \quad (\text{J.27})$$

From (J.24) and (J.26), we obtain

$$\mathcal{L}(\bar{\mathbf{x}}, \mathbf{u}^*) > \mathcal{L}(\mathbf{x}^*, \mathbf{u}^*) + (\mathbf{G}\mathbf{x}^* + \mathbf{g} + \mathbf{A}^T \mathbf{u}^*)^T (\bar{\mathbf{x}} - \mathbf{x}^*) = \mathcal{L}(\mathbf{x}^*, \mathbf{u}^*),$$

and further, (cf. (J.27))

$$\mathcal{L}(\bar{\mathbf{x}}, \mathbf{u}^*) > \mathcal{L}(\bar{\mathbf{x}}, \bar{\mathbf{u}}). \quad (\text{J.28})$$

Taking into account the expression of the Lagrangian function and making use of (J.22), (J.28) gives

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

But from  $\mathbf{u}^* \geq \mathbf{0}$  and  $\mathbf{A} \bar{\mathbf{x}} - \mathbf{b} \leq \mathbf{0}$ , we have  $\mathbf{u}^{*T} (\mathbf{A} \bar{\mathbf{x}} - \mathbf{b}) \leq 0$ , and we are led to a contradiction. Thus,  $\mathbf{x}^* = \bar{\mathbf{x}}$ , as required. Note that if the constraint vectors are linearly independent, we have  $\mathcal{N}(\mathbf{A}^T) = \emptyset$ , and from (J.24), we infer that  $\mathbf{u}^* = \bar{\mathbf{u}}$ .  $\square$

The above theorem shows that the optimal value  $\mathcal{L}(\mathbf{x}^*, \mathbf{u}^*)$  of the dual problem is equivalent to the optimal value  $f(\mathbf{x}^*)$  of the primal problem, and that the solution of the primal problem can be found by solving the dual problem.

Let  $(\mathbf{x}^*, \mathbf{u}^*)$  solve  $(D_R)$  and let us assume that  $\mathbf{x}^*$  lies on a linearly independent active set of constraints indexed by  $I^* \subseteq R$ , i.e.,  $\mathbf{a}_i^T \mathbf{x}^* = [\mathbf{b}]_i$  for  $i \in I^*$ . By Theorem J.2, the necessary and sufficient conditions for optimality of the dual problem  $(D_R)$  are the Kuhn-Tucker conditions, which we express explicitly as

$$\begin{aligned} \mathbf{G}\mathbf{x}^* + \mathbf{g} + \mathbf{A}_{I^*}^T \mathbf{u}_{I^*}^* + \mathbf{A}_{I_d^*}^T \mathbf{u}_{I_d^*}^* &= \mathbf{0}, \\ \mathbf{A}_{I^*} \mathbf{x}^* &= \mathbf{b}_{I^*}, \quad \mathbf{A}_{I_d^*} \mathbf{x}^* < \mathbf{b}_{I_d^*}, \\ \mathbf{u}_{I^*}^* &\geq \mathbf{0}, \quad \mathbf{u}_{I_d^*}^* = \mathbf{0}. \end{aligned}$$

Here,  $I_d^* = R \setminus I^*$  is the inactive set of constraints and, for a generic set  $I$ , we used the notations  $[\mathbf{A}_I]_{ij} = [\mathbf{a}_i]_j$ ,  $i \in I$ ,  $j = 1, \dots, n$ ,  $[\mathbf{b}_I]_i = [\mathbf{b}]_i$ ,  $i \in I$ , and similarly for  $\mathbf{u}_I$ .

In the framework of a dual active set method, we generate feasible iterates  $(\mathbf{x}, \mathbf{u})$ , which fulfill the conditions (J.24) and (J.25), by keeping track of an active set  $I$ . For the active set  $I$ , we have

$$\mathbf{G}\mathbf{x} + \mathbf{g} + \mathbf{A}_I^T \mathbf{u}_I = \mathbf{0}, \quad (\text{J.29})$$

$$\mathbf{A}_I \mathbf{x} = \mathbf{b}_I, \quad (\text{J.30})$$

$$\mathbf{u}_I \geq \mathbf{0}, \quad (\text{J.31})$$

and the optimality conditions of the dual problem show that the solution  $\mathbf{x} = \mathbf{x}^*$  has been found if

$$\mathbf{A}_{I_d} \mathbf{x} \leq \mathbf{b}_{I_d},$$

with  $I_d = R \setminus I$ . If this is not the case, some violated constraint  $p \in R \setminus I$  exists, i.e.,  $c_p(\mathbf{x}) = \mathbf{a}_p^T \mathbf{x} - [\mathbf{b}]_p > 0$ , and  $(\mathbf{x}, \mathbf{u})$  is not a solution pair. Indeed, from

$$\frac{\partial \mathcal{L}}{\partial [\mathbf{u}]_p}(\mathbf{x}, \mathbf{u}) = c_p(\mathbf{x}) > 0,$$

we see that the Lagrangian function  $\mathcal{L}$  can be increased by increasing the multiplier  $[\mathbf{u}]_p$ . The main idea of a dual active set method is to choose a violated constraint  $c_p(\mathbf{x}) > 0$  from the complementary set  $R \setminus I$  and make it satisfy  $c_p(\mathbf{x}) \leq 0$  by increasing the Lagrangian multiplier  $[\mathbf{u}]_p$ .

A realization of a dual active set method is illustrated in Algorithm 19. The algorithm starts with  $I_0 = \emptyset$  and produces a sequence  $\{I_k\}$  such that

$$\min(P_{I_k}) < \min(P_{I_{k+1}}),$$

but not necessarily that  $I_k \subset I_{k+1}$ . In order to simplify the notations, the vector of Lagrange multipliers corresponding to the active set  $I_k$  is denoted by  $\mathbf{u}_k \in \mathbb{R}^{|I_k|}$  instead of  $\mathbf{u}_{I_k}$ . Because the minimum value of the objective function increases, a problem cannot be run twice and the algorithm must stop after a finite number of steps. If the iterate satisfies all the constraints in the complementary set, then the solution has been found and the algorithm terminates.

**Algorithm 19.** General structure of a dual active set method.

---

```

{unconstrained minimum}
 $I_0 \leftarrow \emptyset$ ;  $\mathbf{x}_0 \leftarrow -\mathbf{G}^{-1}\mathbf{g}$ ;
 $k \leftarrow 0$ ;  $stop \leftarrow false$ ;
while  $stop = false$  do
    { $\mathbf{x}_k$  is the optimal solution of  $(P_R)$ }
    if  $c_i(\mathbf{x}_k) = \mathbf{a}_i^T \mathbf{x}_k - [\mathbf{b}]_i \leq 0$  for all  $i \in R \setminus I_k$  then
         $stop \leftarrow true$ ;
    else
        {choose a violated constraint}
        choose  $p \in R \setminus I_k$  with  $c_p(\mathbf{x}_k) = \mathbf{a}_p^T \mathbf{x}_k - [\mathbf{b}]_p > 0$ ;
        {computational step—Algorithm 20}
        compute  $I_{k+1} \subseteq I_k \cup \{p\}$ ,  $\mathbf{x}_{k+1}$ ,  $\mathbf{u}_{k+1}$  and  $f(\mathbf{x}_{k+1}) > f(\mathbf{x}_k)$ ;
    end if
     $k \leftarrow k + 1$ ;
end while

```

---

Before proceeding, we would like to point out that if the pair  $(\mathbf{x}_k, \mathbf{u}_k)$  satisfies (J.29)–(J.31) for  $I_k$ , then  $\mathbf{x}_k$  solves the problem  $(P_{I_k})$  with the vector of Lagrange multipliers  $\mathbf{u}_k$ . Here,  $(P_{I_k})$  is the quadratic programming problem with the objective function (J.15) subject only to the subset of constraints (J.16) indexed by  $I_k$ .

The computational step of a dual active set method is illustrated in Algorithm 20. The while loop is initialized with the solution of the problem  $(P_{I_k})$  and the  $p$ th constraint is assumed to be violated. Thus, the following optimality conditions are fulfilled at the beginning of the while loop (cf. (J.29)–(J.31)):

$$\mathbf{G}\mathbf{x} + \mathbf{g} + \mathbf{A}_I^T \mathbf{u} = \mathbf{0}, \quad (\text{J.32})$$

$$\mathbf{A}_I \mathbf{x} = \mathbf{b}_I, \quad (\text{J.33})$$

$$\mathbf{a}^T \mathbf{x} > b, \quad (\text{J.34})$$

$$\mathbf{u} \geq \mathbf{0}, \quad (\text{J.35})$$

where  $I = I_k$ ,  $\mathbf{x} = \mathbf{x}_k$ ,  $\mathbf{u} = \mathbf{u}_k$ ,  $\mathbf{a} = \mathbf{a}_p$  and  $b = [\mathbf{b}]_p$ . In addition to assumptions (J.32)–(J.35), we suppose that the constraint vectors in the active set  $\{\mathbf{a}_i/i \in I\}$  are linearly independent.

Let us assume that at a generic step, the conditions which require the continuation of the while loop are

$$\mathbf{G}\mathbf{x} + \mathbf{g} + \mathbf{A}_I^T \mathbf{u} + \theta \mathbf{a} = \mathbf{0}, \quad (\text{J.36})$$

$$\mathbf{A}_I \mathbf{x} = \mathbf{b}_I, \quad (\text{J.37})$$

$$\mathbf{a}^T \mathbf{x} > b, \quad (\text{J.38})$$

$$\mathbf{u} \geq \mathbf{0}, \quad \theta \geq 0, \quad f = f(\mathbf{x}). \quad (\text{J.39})$$

Here,  $I$  is the current active set and  $\theta$  is the Lagrange multiplier of the violated constraint. At the first execution of the while loop, these assumptions are satisfied because of (J.32)–(J.35) and the fact that  $\theta = 0$ . The Lagrange multiplier of the violated constraint  $c(\mathbf{x}) =$

**Algorithm 20.** Computational step of a dual active set method.

---

$\mathbf{a} \leftarrow \mathbf{a}_p, \quad b \leftarrow [\mathbf{b}]_p;$   
 {initialization step;  $\mathbf{x}_k$  solves  $(P_{I_k})$  with  $\mathbf{u}_k$ }  
 $I \leftarrow I_k; \quad \mathbf{x} \leftarrow \mathbf{x}_k; \quad \mathbf{u} \leftarrow \mathbf{u}_k \in \mathbb{R}^{|I|}; \quad \theta \leftarrow 0; \quad f \leftarrow f(\mathbf{x}_k); \quad stop \leftarrow \text{false};$   
**while**  $stop = \text{false}$  **do**  
     {constraint matrix}  
     set  $\mathbf{A}_I^T = [\mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_q}] \in \mathbb{R}^{n \times q}$  with  $I = \{i_1, \dots, i_q\}$  and  $q = |I|$ ;  
     {search direction in the dual space  $\mathbf{d}$ }  
     **if**  $I \neq \emptyset$  **then**  
          $\mathbf{d} \leftarrow (\mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T)^{-1} \mathbf{A}_I \mathbf{G}^{-1} \mathbf{a};$   
     **else**  
          $\mathbf{d} \leftarrow \mathbf{0};$   
     **end if**  
     {search direction in the primal space  $\mathbf{p}$ }  
      $\mathbf{p} \leftarrow \mathbf{G}^{-1} (\mathbf{a} - \mathbf{A}_I^T \mathbf{d});$   
     { $\mathbf{a} \notin \text{span} \{\mathbf{a}_i / i \in I\}$ }  
     **if**  $\mathbf{p} \neq \mathbf{0}$  **then** {Step 1}  
         {full step length}  
          $t_1 \leftarrow (\mathbf{a}^T \mathbf{x} - b) / \mathbf{a}^T \mathbf{p};$   
         {add constraint;  $\mathbf{x}_{k+1}$  solves  $(P_{I_{k+1}})$  with  $\mathbf{u}_{k+1}$ }  
         **if**  $\mathbf{u} - t_1 \mathbf{d} \geq \mathbf{0}$  **or**  $I = \emptyset$  **then** {Step 1a}  
              $I_{k+1} \leftarrow I \cup \{p\}; \quad \mathbf{x}_{k+1} \leftarrow \mathbf{x} - t_1 \mathbf{p}; \quad \mathbf{u}_{k+1} \leftarrow \begin{bmatrix} \mathbf{u} - t_1 \mathbf{d} \\ \theta + t_1 \end{bmatrix};$   
              $f(\mathbf{x}_{k+1}) \leftarrow f + t_1 (\theta + t_1/2) \mathbf{a}^T \mathbf{p}; \quad stop \leftarrow \text{true};$   
             {partial step length  $t_2$ ; drop constraint and update  $\mathbf{x}$  and  $\mathbf{u}$ }  
         **else** {Step 1b}  
              $t_2 \leftarrow \frac{[\mathbf{u}]_l}{[\mathbf{d}]_l} = \min \left\{ \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} / [\mathbf{d}]_i > 0, i \in I \right\};$   
              $I \leftarrow I \setminus \{l\}; \quad \mathbf{x} \leftarrow \mathbf{x} - t_2 \mathbf{p};$   
             **for all**  $i \in I$  **do**  $[\mathbf{u}]_i \leftarrow [\mathbf{u}]_i - t_2 [\mathbf{d}]_i$ ; **end for**  
              $\theta \leftarrow \theta + t_2; \quad f \leftarrow f + t_2 (\theta + t_2/2) \mathbf{a}^T \mathbf{p};$   
         **end if**  
         { $\mathbf{a} \in \text{span} \{\mathbf{a}_i / i \in I\}$ }  
     **else** {Step 2}  
         { $(P_{I \cup \{p\}})$  is infeasible and so,  $(P_R)$  is infeasible}  
         **if**  $\mathbf{d} \leq \mathbf{0}$  **then** {Step 2a}  
              $stop \leftarrow \text{true};$   
             {partial step length  $t_2$ ; drop constraint and update  $\mathbf{u}$ }  
         **else** {Step 2b}  
              $t_2 \leftarrow \frac{[\mathbf{u}]_l}{[\mathbf{d}]_l} = \min \left\{ \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} / [\mathbf{d}]_i > 0, i \in I \right\};$   
              $I \leftarrow I \setminus \{l\};$   
             **for all**  $i \in I$  **do**  $[\mathbf{u}]_i \leftarrow [\mathbf{u}]_i - t_2 [\mathbf{d}]_i$ ; **end for**  
              $\theta \leftarrow t_2;$   
         **end if**  
     **end if**  
**end while**

---

$\mathbf{a}^T \mathbf{x} - b$  should be increased from  $\theta$  to some value  $\theta + t$  that will make the constraint binding. This can be achieved by moving from

$$\left( \mathbf{x}, \begin{bmatrix} \mathbf{u} \\ \theta \end{bmatrix} \right) \text{ to } \left( \mathbf{x}(t), \begin{bmatrix} \mathbf{u}(t) \\ \theta + t \end{bmatrix} \right),$$

where

$$\mathbf{x}(t) = \mathbf{x} + \mathbf{p}(t) \quad (\text{J.40})$$

and

$$\mathbf{u}(t) = \mathbf{u} + \mathbf{d}(t). \quad (\text{J.41})$$

The parameter of the transformation  $t$  should be chosen such that  $\mathbf{x}(t)$  solves  $(P_{I \cup \{p\}})$  with the vector of Lagrange multipliers

$$\begin{bmatrix} \mathbf{u}(t) \\ \theta + t \end{bmatrix} \geq \mathbf{0},$$

that is,

$$\mathbf{G}\mathbf{x}(t) + \mathbf{g} + \mathbf{A}_I^T \mathbf{u}(t) + (\theta + t) \mathbf{a} = \mathbf{0}, \quad (\text{J.42})$$

$$\mathbf{A}_I \mathbf{x}(t) = \mathbf{b}_I, \quad (\text{J.43})$$

$$\mathbf{a}^T \mathbf{x}(t) = b, \quad (\text{J.44})$$

$$\mathbf{u}(t) \geq \mathbf{0}, \quad \theta + t \geq 0. \quad (\text{J.45})$$

The first two equations can be expressed in matrix form as

$$\begin{bmatrix} \mathbf{G} & \mathbf{A}_I^T \\ -\mathbf{A}_I & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{u}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{g} \\ \mathbf{b}_I \end{bmatrix} + (\theta + t) \begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix} = \mathbf{0},$$

whence, using (cf. (J.36) and (J.37))

$$\begin{bmatrix} \mathbf{G} & \mathbf{A}_I^T \\ -\mathbf{A}_I & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} + \begin{bmatrix} \mathbf{g} \\ \mathbf{b}_I \end{bmatrix} + \theta \begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix} = \mathbf{0},$$

we obtain

$$\begin{bmatrix} \mathbf{G} & \mathbf{A}_I^T \\ -\mathbf{A}_I & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{p}(t) \\ \mathbf{d}(t) \end{bmatrix} = -t \begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix}. \quad (\text{J.46})$$

In the case of equality constraints, we solved (J.46) by using a QR factorization of  $\mathbf{A}_I^T$  and by employing a backward substitution for the resulting block matrix equations. Now, we use the following result: if  $\mathbf{G}$  is positive definite and  $\mathbf{A}_I$  has full row rank, the inverse of the augmented matrix in (J.46) is

$$\begin{bmatrix} \mathbf{G}^{-1} - \mathbf{G}^{-1} \mathbf{A}_I^T (\mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T)^{-1} \mathbf{A}_I \mathbf{G}^{-1} & -\mathbf{G}^{-1} \mathbf{A}_I^T (\mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T)^{-1} \\ (\mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T)^{-1} \mathbf{A}_I \mathbf{G}^{-1} & (\mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T)^{-1} \end{bmatrix},$$

and we have

$$\mathbf{x}(t) = \mathbf{x} - t\mathbf{p}, \quad \mathbf{u}(t) = \mathbf{u} - t\mathbf{d},$$

with

$$\mathbf{d} = (\mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T)^{-1} \mathbf{A}_I \mathbf{G}^{-1} \mathbf{a}, \quad (\text{J.47})$$

and

$$\mathbf{p} = \mathbf{G}^{-1} (\mathbf{a} - \mathbf{A}_I^T \mathbf{d}). \quad (\text{J.48})$$

In (J.47) and (J.48),  $\mathbf{d}$  and  $\mathbf{p}$  represent the search directions in the dual and the primal spaces, respectively, while  $t$  is the step length. If  $I = \emptyset$ , then  $\mathbf{A}_I$  is not defined and we set  $\mathbf{d} = \mathbf{0}$ , which, in turn, yields  $\mathbf{p} = \mathbf{G}^{-1} \mathbf{a}$ . Noting that the step length  $t > 0$  will be chosen to make (J.44) satisfied, we establish some basic results which are relevant for our analysis.

- (1) If  $I \neq \emptyset$ , then from (J.37) and (J.43), we see that  $\mathbf{A}_I \mathbf{p} = \mathbf{0}$ . This result together with (J.48) yields

$$\mathbf{p}^T (\mathbf{a} - \mathbf{Gp}) = \mathbf{p}^T \mathbf{A}_I^T \mathbf{d} = (\mathbf{A}_I \mathbf{p})^T \mathbf{d} = 0,$$

and, as  $\mathbf{G}$  is positive definite, we have

$$\mathbf{p}^T \mathbf{a} = \mathbf{p}^T \mathbf{Gp} > 0 \quad (\text{J.49})$$

for  $\mathbf{p} \neq \mathbf{0}$ . Similarly, if  $I = \emptyset$ , then  $\mathbf{d} = \mathbf{0}$ . Hence,  $\mathbf{Gp} = \mathbf{a}$ , and as  $\mathbf{p} \neq \mathbf{0}$ , we obtain  $\mathbf{p}^T \mathbf{a} = \mathbf{p}^T \mathbf{Gp} > 0$ .

- (2) From (J.49), we observe that, for  $t > 0$ , the  $p$ th constraint at  $\mathbf{x} - t\mathbf{p}$ ,

$$c(\mathbf{x} - t\mathbf{p}) = c(\mathbf{x}) - t\mathbf{a}^T \mathbf{p}$$

is decreasing as we move from  $\mathbf{x}$  to  $\mathbf{x} - t\mathbf{p}$ , and this is exactly what we want as it is violated and positive at  $\mathbf{x}$ .

- (3) The objective function at  $\mathbf{x} - t\mathbf{p}$  is given by (cf. (J.36), (J.49) and the relation  $\mathbf{A}_I \mathbf{p} = \mathbf{0}$ )

$$\begin{aligned} f(\mathbf{x} - t\mathbf{p}) &= f(\mathbf{x}) - t(\mathbf{g} + \mathbf{Gx})^T \mathbf{p} + \frac{1}{2} t^2 \mathbf{p}^T \mathbf{Gp} \\ &= f(\mathbf{x}) + t\mathbf{p}^T (\mathbf{A}_I^T \mathbf{u} + \theta \mathbf{a}) + \frac{1}{2} t^2 \mathbf{p}^T \mathbf{a} \\ &= f(\mathbf{x}) + t \left( \theta + \frac{1}{2} t \right) \mathbf{p}^T \mathbf{a}. \end{aligned} \quad (\text{J.50})$$

- (4) The Lagrangian functions at

$$\left( \mathbf{x} - t\mathbf{p}, \begin{bmatrix} \mathbf{u} - t\mathbf{d} \\ \theta + t \end{bmatrix} \right) \text{ and } \left( \mathbf{x}, \begin{bmatrix} \mathbf{u} \\ \theta \end{bmatrix} \right)$$

can be expressed as (cf. (J.17))

$$\mathcal{L}(t) = f(\mathbf{x} - t\mathbf{p}) + (\theta + t) [\mathbf{a}^T (\mathbf{x} - t\mathbf{p}) - b],$$

and

$$\mathcal{L}(0) = f(\mathbf{x}) + \theta (\mathbf{a}^T \mathbf{x} - b),$$

respectively. Then, using (J.50), we obtain

$$\mathcal{L}(t) - \mathcal{L}(0) = - \left[ \frac{1}{2} t^2 \mathbf{p}^T \mathbf{a} - t (\mathbf{a}^T \mathbf{x} - b) \right]. \quad (\text{J.51})$$

We proceed now to analyze the computational step of the dual active set method. Depending on the size of  $\mathbf{p}$ , two situations can occur, namely  $\mathbf{p} \neq \mathbf{0}$  and  $\mathbf{p} = \mathbf{0}$ .

*Step 1:  $\mathbf{p} \neq \mathbf{0}$ .* This branch of the if-statement occurs when  $\mathbf{a} \notin \text{span} \{\mathbf{a}_i / i \in I\}$ . In this case, the full step length  $t_1$ , given by

$$t_1 = \frac{1}{\mathbf{a}^T \mathbf{p}} (\mathbf{a}^T \mathbf{x} - b), \quad (\text{J.52})$$

is well defined, and from (J.38) and (J.49), it follows that  $t_1 > 0$ . For  $t = t_1$ , (J.44) is fulfilled, that is,  $\mathbf{a}^T \mathbf{x}(t_1) = b$ , while for  $0 \leq t \leq t_1$ ,  $\mathcal{L}(t) - \mathcal{L}(0) \geq 0$ , that is, the Lagrangian function increases as we move from

$$\left( \mathbf{x}, \begin{bmatrix} \mathbf{u} \\ \theta \end{bmatrix} \right) \text{ to } \left( \mathbf{x} - t_1 \mathbf{p}, \begin{bmatrix} \mathbf{u} - t_1 \mathbf{d} \\ \theta + t_1 \end{bmatrix} \right).$$

The step length should be chosen so that the Lagrange multipliers are non-negative. In this regard, two situations can be distinguished.

- *Step 1a:  $\mathbf{u} - t_1 \mathbf{d} \geq \mathbf{0}$  or  $I = \emptyset$ .* Putting  $t = t_1$  in (J.42), (J.43) and (J.44) and taking into account that  $t_1 > 0$ , and so,  $t_1 + \theta > 0$ , we deduce that, for  $I_{k+1} = I \cup \{p\}$ ,  $\mathbf{x}_{k+1} = \mathbf{x} - t_1 \mathbf{p}$  is the solution of the primal problem  $(P_{I_{k+1}})$  with the vector of Lagrange multipliers

$$\mathbf{u}_{k+1} = \begin{bmatrix} \mathbf{u} - t_1 \mathbf{d} \\ \theta + t_1 \end{bmatrix} \geq \mathbf{0}.$$

In addition, by (J.49) and (J.50), we have

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}) + t_1 \left( \theta + \frac{1}{2} t_1 \right) \mathbf{p}^T \mathbf{a} > f(\mathbf{x}).$$

- *Step 1b:  $[\mathbf{u}]_i - t_1 [\mathbf{d}]_i < 0$  for all  $i \in \bar{I} \subseteq I$ , and  $I \neq \emptyset$ .* As  $[\mathbf{u}]_i \geq 0$  and  $t_1 > 0$ , it follows that  $[\mathbf{d}]_i > 0$  for all  $i \in \bar{I}$ . Consequently, the partial step length  $t_2$  given by

$$t_2 = \frac{[\mathbf{u}]_l}{[\mathbf{d}]_l} = \min \left\{ \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} / i \in \bar{I} \right\} = \min \left\{ \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} / [\mathbf{d}]_i > 0, i \in I \right\} \quad (\text{J.53})$$

is well defined, and from

$$t_1 > \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} \geq \frac{[\mathbf{u}]_l}{[\mathbf{d}]_l} = t_2, \quad i \in \bar{I}, \quad (\text{J.54})$$

we infer that  $0 \leq t_2 < t_1$ . Note that

$$\left\{ \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} / i \in \bar{I} \right\} \subseteq \left\{ \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} / [\mathbf{d}]_i > 0, i \in I \right\},$$



because the larger set may contain elements  $[\mathbf{u}]_i/[\mathbf{d}]_i$  with  $[\mathbf{u}]_i - t_1[\mathbf{d}]_i \geq 0$ ; for these elements, we have  $t_1 \leq [\mathbf{u}]_i/[\mathbf{d}]_i$  and by (J.54) we deduce that the minimizers of the two sets coincide. The inequality

$$[\mathbf{u}]_i - t_2[\mathbf{d}]_i = [\mathbf{d}]_i \left( \frac{[\mathbf{u}]_i}{[\mathbf{d}]_i} - t_2 \right) \geq 0$$

holds for  $[\mathbf{d}]_i > 0$  and  $[\mathbf{d}]_i \leq 0$ , that is, for all  $i \in I$ . Hence,  $\mathbf{u} - t_2\mathbf{d} \geq \mathbf{0}$ , and in particular,  $[\mathbf{u}]_l - t_2[\mathbf{d}]_l = 0$ . Let  $I^- = I \setminus \{l\}$ ,  $\mathbf{x}^- = \mathbf{x} - t_2\mathbf{p}$ ,  $[\mathbf{u}^-]_i = [\mathbf{u}]_i - t_2[\mathbf{d}]_i$  for  $i \in I^-$ ,  $\theta^- = \theta + t_2$ , and

$$f^- = f + t_2 \left( \theta + \frac{1}{2}t_2 \right) \mathbf{p}^T \mathbf{a} \geq f.$$

Since  $[\mathbf{u}]_l - t_2[\mathbf{d}]_l = 0$ , we have

$$\mathbf{A}_I^T (\mathbf{u} - t_2\mathbf{d}) = \sum_{i \in I} ([\mathbf{u}]_i - t_2[\mathbf{d}]_i) \mathbf{a}_i = \sum_{i \in I^-} ([\mathbf{u}]_i - t_2[\mathbf{d}]_i) \mathbf{a}_i = \mathbf{A}_{I^-}^T \mathbf{u}^-, \quad (\text{J.55})$$

and (J.42) with  $t = t_2$  gives

$$\mathbf{G}\mathbf{x}^- + \mathbf{g} + \mathbf{A}_{I^-}^T \mathbf{u}^- + \theta^- \mathbf{a} = \mathbf{0}.$$

Moreover, from (J.43) with the  $l$ th constraint dropped, we have  $\mathbf{A}_{I^-} \mathbf{x}^- = \mathbf{b}_{I^-}$ , while from (J.49), we find that

$$\mathbf{a}^T \mathbf{x}^- = \mathbf{a}^T \mathbf{x} - t_2 \mathbf{a}^T \mathbf{p} > \mathbf{a}^T \mathbf{x} - t_1 \mathbf{a}^T \mathbf{p} = b.$$

Thus, conditions (J.36)–(J.39) are satisfied for  $I^-$ ,  $\mathbf{x}^-$ ,  $\mathbf{u}^-$ ,  $\theta^-$  and  $f^-$ , and the while loop will continue to run.

*Step 2:  $\mathbf{p} = \mathbf{0}$ .* This branch of the if-statement occurs at the first execution of the while loop when conditions (J.32)–(J.35) are fulfilled, and when  $\mathbf{a} = \mathbf{A}_I^T \mathbf{d}$ , that is, when  $\mathbf{a} \in \text{span}\{\mathbf{a}_i/i \in I\}$ . Depending on the sign of the dual search direction, the algorithm may terminate with infeasibility or it may continue to run with a reduced active set.

- *Step 2a:  $\mathbf{d} \leq \mathbf{0}$ .* Let us assume that there exists  $\mathbf{x}'$  such that  $\mathbf{x} + \mathbf{x}'$  is a feasible solution to  $(P_{I \cup \{p\}})$ . Then, from  $\mathbf{A}_I (\mathbf{x} + \mathbf{x}') \leq \mathbf{b}_I$  and  $\mathbf{A}_I \mathbf{x} = \mathbf{b}_I$  we must have that  $\mathbf{A}_I \mathbf{x}' \leq \mathbf{0}$ . This condition together with  $\mathbf{a} = \mathbf{A}_I^T \mathbf{d}$  and  $\mathbf{d} \leq \mathbf{0}$  shows that  $\mathbf{a}^T \mathbf{x}' = \mathbf{d}^T (\mathbf{A}_I \mathbf{x}') \geq 0$  must hold. On the other hand, the violated constraint should be satisfied and we must have that  $\mathbf{a}^T (\mathbf{x} + \mathbf{x}') \leq b$ , or equivalently that (cf. (J.34)),  $\mathbf{a}^T \mathbf{x}' \leq b - \mathbf{a}^T \mathbf{x} < 0$ . Thus, we are led to a contradiction and we conclude that in this case, the problem  $(P_{I \cup \{p\}})$  is not feasible.
- *Step 2b:  $[\mathbf{d}]_i > 0$  for all  $i \in \bar{I} \subseteq I$ .* Arguing as in Step 1b, we see that  $t_2$  given by (J.53) is well defined and that  $t_2 \geq 0$ . Let  $I^- = I \setminus \{l\}$ ,  $\mathbf{x}^- = \mathbf{x}$ ,  $[\mathbf{u}^-]_i = [\mathbf{u}]_i - t_2[\mathbf{d}]_i$  for  $i \in I^-$ ,  $\theta^- = t_2$ , and  $f^- = f$ . Using (J.55), the result  $\mathbf{a} = \mathbf{A}_I^T \mathbf{d}$ , and (J.32) with  $\mathbf{x} = \mathbf{x}^-$ , yield

$$\mathbf{G}\mathbf{x}^- + \mathbf{g} + \mathbf{A}_{I^-}^T \mathbf{u}^- + \theta^- \mathbf{a} = \mathbf{G}\mathbf{x} + \mathbf{g} + \mathbf{A}_I^T (\mathbf{u} - t_2\mathbf{d}) + t_2 \mathbf{a} = \mathbf{0},$$

while (J.33) and (J.34) with  $\mathbf{x} = \mathbf{x}^-$ , give  $\mathbf{A}_I \mathbf{x}^- = \mathbf{b}_I^-$  and  $\mathbf{a}^T \mathbf{x}^- > b$ , respectively. Hence, conditions (J.36)–(J.39) are satisfied for  $I^-$ ,  $\mathbf{x}^-$ ,  $\mathbf{u}^-$ ,  $\theta^-$  and  $f^-$ , and the while loop does not terminate.

Although at the beginning of Step 2, we have  $\mathbf{a} \in \text{span}\{\mathbf{a}_i/i \in I\}$ , at the end of Step 2b, we have  $\mathbf{a} \notin \text{span}\{\mathbf{a}_i/i \in I^-\}$ . To prove this claim, we assume that  $\mathbf{a} \in \text{span}\{\mathbf{a}_i/i \in I^-\}$  and use the condition  $\mathbf{a} \in \text{span}\{\mathbf{a}_i/i \in I\}$ , written as

$$\mathbf{a} = [\mathbf{d}]_I \mathbf{a}_I + \sum_{i \in I^-} [\mathbf{d}]_i \mathbf{a}_i, \quad (\text{J.56})$$

to conclude that  $\mathbf{a}_I \in \text{span}\{\mathbf{a}_i/i \in I^-\}$ . This result is contradictory to our initial assumption that the vectors  $\{\mathbf{a}_i/i \in I\}$  are linearly independent and the claim is proven. Thus, the branch  $\mathbf{p} = \mathbf{0}$  of the if-statement is executed only once, since at the subsequent runs of the while loop, the active set is reduced and the condition  $\mathbf{a} \notin \text{span}\{\mathbf{a}_i/i \in I\}$  is always fulfilled.

In conclusion, after a sequence of at most  $\min(r, n)$  partial steps (the first of which may occur in Step 2b) and one full step (Step 1a), either the solution to the primal problem ( $P_{I_{k+1}}$ ) is found, or the infeasibility is detected (Step 2a). The Algorithm 2 terminates after a finite number of steps, since  $I_k$  is finite and only one constraint is dropped in Steps 1b and 2b.

We close our presentation with some comments on implementation issues. The search directions in the dual and the primal space can be expressed as (cf. (J.47) and (J.48))

$$\mathbf{d} = \mathbf{A}_I^\dagger \mathbf{a}$$

and

$$\mathbf{p} = \mathbf{H}_I \mathbf{a},$$

respectively, where

$$\mathbf{A}_I^\dagger = (\mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T)^{-1} \mathbf{A}_I \mathbf{G}^{-1} \in \mathbb{R}^{q \times n}, \quad q = |I|,$$

is a left inverse of  $\mathbf{A}_I^T$ , i.e.,  $\mathbf{A}_I^\dagger \mathbf{A}_I^T = \mathbf{I}_q$ , and

$$\mathbf{H}_I = \mathbf{G}^{-1} \left( \mathbf{I}_n - \mathbf{A}_I^T \mathbf{A}_I^\dagger \right) \in \mathbb{R}^{n \times n}$$

is the reduced inverse Hessian of  $f$  subject to the active set of constraints. To compute  $\mathbf{A}_I^\dagger$  and  $\mathbf{H}_I$ , we consider the Cholesky factorization  $\mathbf{G} = \mathbf{L}\mathbf{L}^T$  and the QR factorization of the matrix  $\mathbf{B} = \mathbf{L}^{-1} \mathbf{A}_I^T \in \mathbb{R}^{n \times q}$ , that is,

$$\mathbf{B} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = [\mathbf{Q}_1 \quad \mathbf{Q}_2] \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix},$$

with  $\mathbf{Q}_1 \in \mathbb{R}^{n \times q}$ ,  $\mathbf{Q}_2 \in \mathbb{R}^{n \times (n-q)}$  and  $\mathbf{R} \in \mathbb{R}^{q \times q}$ . Then, using the results

$$\begin{aligned} \mathbf{A}_I \mathbf{G}^{-1} &= \mathbf{B}^T \mathbf{L}^{-1} = [\mathbf{R}^T \quad \mathbf{0}] \mathbf{Q}^T \mathbf{L}^{-1} \\ \mathbf{A}_I \mathbf{G}^{-1} \mathbf{A}_I^T &= \mathbf{B}^T \mathbf{B} = \mathbf{R}^T \mathbf{R} \end{aligned}$$

and

$$\mathbf{I}_n = \mathbf{Q}\mathbf{Q}^T = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{bmatrix} = \mathbf{Q}_1\mathbf{Q}_1^T + \mathbf{Q}_2\mathbf{Q}_2^T$$

we find that

$$\mathbf{A}_I^\dagger = \mathbf{R}^{-1}\mathbf{R}^{-T} \begin{bmatrix} \mathbf{R}^T & \mathbf{0} \end{bmatrix} \mathbf{Q}^T \mathbf{L}^{-1} = \mathbf{R}^{-1}\mathbf{Q}_1^T \mathbf{L}^{-1}$$

and that

$$\mathbf{H}_I = \mathbf{L}^{-T} (\mathbf{I}_n - \mathbf{L}^{-1}\mathbf{A}_I^T \mathbf{R}^{-1}\mathbf{Q}_1^T) \mathbf{L}^{-1} = \mathbf{L}^{-T}\mathbf{Q}_2\mathbf{Q}_2^T \mathbf{L}^{-1}.$$

In terms of the auxiliary matrix  $\mathbf{J} = \mathbf{L}^{-T}\mathbf{Q} \in \mathbb{R}^{n \times n}$ , partitioned as

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{L}^{-T}\mathbf{Q}_1 & \mathbf{L}^{-T}\mathbf{Q}_2 \end{bmatrix},$$

$\mathbf{A}_I^\dagger$  and  $\mathbf{H}_I$  can be expressed as

$$\mathbf{A}_I^\dagger = \mathbf{R}^{-1}\mathbf{J}_1^T$$

and

$$\mathbf{H}_I = \mathbf{J}_2\mathbf{J}_2^T,$$

respectively; whence introducing the vector

$$\mathbf{z} = \mathbf{J}^T \mathbf{a} = \begin{bmatrix} \mathbf{J}_1^T \mathbf{a} \\ \mathbf{J}_2^T \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix},$$

we obtain

$$\mathbf{d} = \mathbf{R}^{-1}\mathbf{z}_1$$

and

$$\mathbf{p} = \mathbf{J}_2\mathbf{z}_2.$$

Thus, the computation of the search directions requires the knowledge of the matrices  $\mathbf{J}$  and  $\mathbf{R}$ . To increase the numerical efficiency of the algorithm, these matrices are updated in an ingenious way whenever a constraint is added to or deleted from the set of active constraints.

Two efficient implementations of the method of Goldfarb and Idnani can be found in Powell (1985). For a general discussion of primal active set methods we refer to Gill and Murray (1978) and Gill et al. (1981).

## References

- Aben, I., Stam, D. M., and Helderman, F. (2001). The Ring effect in skylight polarisation. *Geophys. Res. Lett.* 28, 519–522.
- Aliwell, S. R., Van Roozendaal, M., Johnston, P. V., Richter, A., Wagner, T., Arlander, D. W., Burrows, J. P., Fish, D. J., Jones, R. L., Tornkvist, K. K., Lambert, J.-C., Pfeilsticker, K., and Pundt, I. (2002). Analysis for BrO in zenith-sky spectra: An intercomparison exercise for analysis improvement. *J. Geophys. Res.* 107(D14), 4199. doi:10.1029/2001JD000329.
- Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J. J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S., and Sorensen, D. (1995). *LA-PACK Users' Guide* (2nd Edition). SIAM, Philadelphia, PA.
- Anderson, G. P., Clough, S. A., Kneizys, F. X., Chetwynd, J. H., and Shettle, E. P. (1986). *AFGL Atmospheric Constituent Profiles (0–120 km)*. Technical Report TR-86–0110, U.S. Air Force Geophysics Laboratory, Hanscom Air Force Base, Massachusetts.
- Arnoldi, W. E. (1951). The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. J. Applied Mathematics* 9, 17–29.
- Backus, G. and Gilbert, F. (1967). Numerical applications of a formalism for geophysical inverse problems. *Geophys. J. R. Astron. Soc.* 13, 247–276.
- Backus, G. and Gilbert, F. (1968). The resolving power of gross earth data. *Geophys. J. R. Astron. Soc.* 16, 169–205.
- Backus, G. and Gilbert, F. (1970). Uniqueness in the inversion of inaccurate gross earth data. *Phil. Trans. Roy. Soc.* 266, 123–192.
- Bakushinsky, A. B. (1992). The problem of the convergence of the iteratively regularized Gauss–Newton method. *Comput. Math. Phys.* 32, 1353–1359.
- Balis, D., Lambert, J. C., Van Roozendaal, M., Spurr, R., Loyola, D., Livschitz, Y., Valks, P., Amiridis, V., Gerard, P., Granville, J., and Zehner, C. (2007). Ten years of GOME/ERS-2 total ozone data: the new GOME Data Processor (GDP) Version 4: II. Ground-based validation and comparisons with TOMS V7/V8. *J. Geophys. Res.* 112, D07307.
- Bard, Y. (1974). *Nonlinear Parameter Estimation*. Academic Press, New York.
- Barkstrom, B. (1975). A finite difference method for solving anisotropic scattering problems. *J. Quant. Spectrosc. Radiat. Transfer* 16, 725–739.

- Barrett, R., Berry, M., Chan, T. F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., and Van der Vost, H. (1994). *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, Philadelphia, PA.
- Bates, D. M., Hamilton, D. C., and Watts, D. G. (1983). Calculation of intrinsic and parameter-effects curvature for nonlinear regression models. *Commun. Stat. Simul. Comput.* 12, 71–81.
- Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.
- Bauer, F. and Hohage, T. (2005). A Lepskij-type stopping rule for regularized Newton methods. *Inverse Problems* 21, 1975–1991.
- Baumeister, J. (1987). *Stable Solution of Inverse Problems*. Vieweg, Wiesbaden.
- Belge, M., Kilmer, M. E., and Miller, E. L. (2002). Efficient determination of multiple regularization parameters in a generalized L-curve framework. *Inverse Problems* 18, 1161–1183.
- Berk, A., Bernstein, L. S., and Robertson, D. (1989). *MODTRAN: A Moderate Spectra Resolution LOWTRAN7*. Technical Report GL-TR-89-0122, U.S. Air Force Geophysics Laboratory, Hanscom Air Force Base, Massachusetts.
- Bissantz, N., Hohage, T., and Munk, A. (2004). Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems* 20, 1773–1789.
- Blaschke, B., Neubauer, A., and Scherzer, O. (1997). On convergence rates for the iteratively regularized Gauss–Newton method. *IMA Journal of Numer. Anal.* 17, 421–436.
- de Boor, C. (2001). *A Practical Guide to Splines*. Springer, New York.
- Bouaricha, A. and Schnabel, R. B. (1997). Algorithm 768: TENSOLVE: a software package for solving systems of nonlinear equations and nonlinear least-squares problems using tensor methods. *ACM Transactions on Mathematical Software* 23, 174–195.
- Box, M. A. (2002). Radiative perturbation theory: A review. *Environ. Modelling Software* 17, 95–106.
- Böckmann, C. and Pornsawad, P. (2008). Iterative Runge–Kutta-type methods for nonlinear ill-posed problems. *Inverse Problems* 24. doi:10.1088/0266-5611/24/2/025002.
- Brakhage, H. (1987). On ill-posed problems and the method of conjugate gradients. In Engl, H. W. and Groetsch, C. W. (eds), *Inverse and Ill-Posed Problems*, Academic Press, Orlando, FL, pp. 165–175.
- Brezinski, C., Redivo-Zaglia, M., Rodriguez, G., and Seatzu, S. (2003). Multi-parameter regularization techniques for ill-conditioned linear systems. *Numer. Math.* 94, 203–228.
- Bühler, S. A., Eriksson, P., Kuhn, T., von Engeln, A., and Verdes, C. (2005). ARTS, the atmospheric radiative transfer simulator. *J. Quant. Spectrosc. Radiat. Transfer* 91, 65–93.
- Carissimo, A., De Feis, I., and Serio, C. (2005). The physical retrieval methodology for IASI: The  $\delta$ -IASI code. *Environ. Model. Software* 20, 1111–1126.
- Carlotti, M. (1988). Global-fit approach to the analysis of limb-scanning atmospheric measurements. *Appl. Opt.* 27, 3250–3254.
- Ceccherini, S. (2005). Analytical determination of the regularization parameter in the retrieval of atmospheric vertical profiles. *Opt. Lett.* 30, 2554–2556.

- Chahine, M. T., Pagano, T. S., Aumann, H. H., Atlas, R., Barnet, C., Blaisdell, J., Chen, L., Divakarla, M., Fetzner, E. J., Goldberg, M., Gautier, C., Granger, S., Hannon, S., Irion, F. W., Kakar, R., Kalnay, E., Lambrigtsen, B. H., Lee, S. Y., Le Marshall, J., McMillan, W. W., McMillin, L., Olsen, E. T., Revercomb, H., Rosenkranz, P., Smith, W. L., Staelin, D., Strow, L. L., Susskind, J., Tobin, D., Wolf, W., and Zhou, L. (2006). AIRS: improving weather forecasting and providing new data on greenhouse gases. *Bull. Am. Met. Soc.* 87, 911–926.
- von Clarmann, T., Ceccherini, S., Doicu, A., Dudhia, A., Funke, B., Grabowski, U., Hilgers, S., Jay, V., Linden, A., López-Puertas, M., Martin-Torres, F.-J., Payne, V., Reburn, J., Ridolfi, M., Schreier, F., Schwarz, G., Siddans, R., and Steck T. (2003). A blind test retrieval experiment for infrared limb emission spectrometry. *J. Geophys. Res.* 108(D23), 4746. doi:10.1029/2003JD003835.
- Clough, S. A., Kneizys, F. X., and Davies, R. (1989). Line shape and the water vapor continuum. *Atmos. Res.* 23, 229–241.
- Clough, S. A., Shephard, M. W., Mlawer, E. J., Delamere, J. S., Iacono, M. J., Cady-Pereira, K., Boukabara, S., and Brown, P. D. (2005). Atmospheric radiative transfer modeling: A summary of the AER codes. *J. Quant. Spectrosc. Radiat. Transfer* 91, 233–244.
- Craig, I. J. D. and Brown, J. C. (1986). *Inverse Problems in Astronomy*. Adam Hilger, Bristol.
- CRAN-Package quadprog (2007). *quadprog: Functions to solve Quadratic Programming Problems*. Available at <http://cran.r-project.org/web/packages/quadprog/index.html> (Accessed 16 October 2009).
- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* 31, 377–407.
- Dahlback, A. and Stamnes, K. (1991). A new spherical model for computing the radiation field available for photolysis and heating at twilight. *Planet. Space Sci.* 39, 671–683.
- Demoment, G. (1989). Image reconstruction and restoration: Overview of common estimation problems. *IEEE Trans. Acoust. Speech Signal Processing* 37, 2024–2036.
- De Moor, B. L. R. and Zha, H. (1991). A tree of generalizations of the ordinary singular value decomposition. *Linear Algebra Appl.* 147, 469–500.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.* 39, 1–38.
- Dennis, J. E. Jr. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983; reprinted by SIAM, Philadelphia, PA.
- Dennis, J. E. Jr., Gay, D. M., and Welsch, R. E. (1981). Algorithm 573. NL2SOL—an adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Software* 7, 369–383.
- Deuffhard, P., Engl, H. W., and Scherzer, O. (1998). A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinity invariant conditions. *Inverse Problems* 14, 1081–1106.
- Des Marais, D. J., Harwit, M. O., Jucks, K. W., Kasting, J. F., Lin, D. N. C., Lunine, J. I., Schneider, J., Seager, S., Traub, W. A., and Woolf, N. J. (2002). Remote sensing of planetary properties and biosignatures on extrasolar terrestrial planets. *Astrobiology* 2, 153–181.

- Doicu, A. and Trautmann, T. (2009a). Discrete-ordinate method with matrix exponential for a pseudo-spherical atmosphere: Scalar case. *J. Quant. Spectrosc. Radiat. Transfer* 110, 146–158.
- Doicu, A. and Trautmann, T. (2009b). Discrete-ordinate method with matrix exponential for a pseudo-spherical atmosphere: Vector case. *J. Quant. Spectrosc. Radiat. Transfer* 110, 159–172.
- Doicu, A. and Trautmann, T. (2009c). Adjoint problem of radiative transfer for a pseudo-spherical atmosphere and general viewing geometries. *J. Quant. Spectrosc. Radiat. Transfer* 110, 464–476.
- Doicu, A. and Trautmann, T. (2009d). Two linearization methods for atmospheric remote sensing. *J. Quant. Spectrosc. Radiat. Transfer* 110, 477–490.
- Doicu, A. and Trautmann, T. (2009e). Picard iteration methods for a spherical atmosphere. *J. Quant. Spectrosc. Radiat. Transfer* 110, 1851–1863.
- Eggermont, P. P. B. (1993). Maximum entropy regularization for Fredholm integral equations of the first kind. *SIAM J. Math. Anal.* 24, 1557–1576.
- Elden, L. (1977). Algorithms for the regularization of ill-conditioned least squares problems. *BIT* 17, 134–145.
- Emde, C., Bühler, S. A., Davis, C., Eriksson, P., Sreerekha, T. R., and Teichmann, C. (2004). A polarized discrete ordinate scattering model for simulations of limb and nadir longwave measurements in 1D/3D spherical atmospheres. *J. Geophys. Res.* 109, D24207. doi:10.1029/2004JD005140.1.4.4.
- Engl, H. W. and Gfrerer, H. (1988). A posteriori parameter choice for general regularization methods for solving linear ill-posed problems. *Appl. Num. Math.* 4, 395–417.
- Engl, H. W., Kunisch, K., and Neubauer, A. (1989). Convergence rates for Tikhonov regularization of nonlinear ill-posed problems. *Inverse Problems* 5, 523–540.
- Engl, H. W., Hanke, M., and Neubauer, A. (2000). *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht.
- Eriksson, J. (1996). *Optimization and Regularization of Nonlinear Least Square Problems*. Ph.D. Thesis, Umeå University, Sweden.
- Eriksson, P., Jimenez, C., and Bühler, S. A. (2005). Qpack, a general tool for instrument simulation and retrieval work. *J. Quant. Spectrosc. Radiat. Transfer* 91, 47–64.
- Eyre, J. R. (1990). The information content of data from satellite sounding systems: A simulation study. *Quart. J. Roy. Meteor. Soc.* 116, 401–434.
- Farquharson, C. G. and Oldenburg, D. W. (2004). A comparison of automatic techniques for estimating the regularization parameter in nonlinear inverse problems. *Geophys. J. Int.* 156, 411–425.
- Fessler, J. A. (1991). Nonparametric fixed-interval smoothing with vector splines. *IEEE Trans. Signal Processing* 39, 852–859.
- Fierro, R. D., Golub, G. H., Hansen, P. C., and O’Leary, D. P. (1997). Regularization by truncated total least squares. *SIAM J. Sci. Comput.* 18, 1223–1241.
- Fitzpatrick, B. G. (1991). Bayesian analysis in inverse problems. *Inverse Problems* 7, 675–702.
- Frieden, B. R. (1972). Restoring with maximum likelihood and maximum entropy. *J. Opt. Soc. Am.* 62, 511–518.
- Fu, Q. and Liou, K. -N. (1992). On the correlated  $k$ -distribution method for radiative transfer in nonhomogeneous atmospheres. *J. Atmos. Sci.* 49, 2139–2156.



- Galatsanos, N. P. and Katsaggelos, A. K. (1992). Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Trans. Image Process.* 3, 322–336.
- Gfrerer, H. (1987). An a posteriori parameter choice method for ordinary and iterated Tikhonov regularization of ill-posed problems leading to optimal convergence rates. *Math. Comput.* 49, 507–522.
- Gill, P. E. and Murray, W. (1978). Numerically stable methods for quadratic programming. *Math. Programming* 14, 349–372.
- Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical Optimization*. Academic Press, London.
- Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Programming* 27, 1–33.
- Golub, G. H. and Kahan, W. (1965). Calculating the singular values and pseudoinverse of a matrix. *SIAM J. Numer. Anal.* 2, 205–224.
- Golub, G. H. and Van Loan, C. F. (1980). An analysis of the total least squares problem. *SIAM J. Numer. Anal.* 17, 883–893.
- Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations* (3rd Edition). Johns Hopkins University Press, Baltimore, MD.
- Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21, 215–223.
- Golub, G. H., Hansen P. C., and O’Leary, D. P. (1999). Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.* 21, 185–194.
- Goody, R. M. and Yung, Y. L. (1989). *Atmospheric Radiation – Theoretical Basis*. Oxford University Press, Oxford.
- Gouveia, W. P. and Scales, J. A. (1997). Resolution of seismic waveform inversion: Bayes versus Occam. *Inverse Problems* 13, 323–349.
- Grafarend, E. and Schaffrin, B. (1993). *Ausgleichsrechnung in Linearen Modellen*. BI Wissenschaftsverlag, Mannheim.
- Griewank, A. (2000). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, Philadelphia, PA.
- Griewank, A. and Corliss, G. (eds) (1991). *Automatic Differentiation of Algorithms*. SIAM, Philadelphia, PA.
- Groetsch, C. W. (1984). *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, Boston, MA.
- Groetsch, C. W. (1993). *Inverse Problems in the Mathematical Sciences*. Vieweg, Wiesbaden.
- Gull, S. F. and Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature* 272, 686–690.
- Gulliksson, M. E. and Wedin, P. A. (1999). Optimization tools for inverse problems using the nonlinear L-curve and A-curve. In *Proceedings of the 3rd ASME International Conference on Inverse Problems in Engineering*, Port Ludlow, Washington, 13–18 June.
- Guo, H. and Renault, R. A. (2002). A regularized total least squares algorithm. In Van Huffel, S. and Lemmerling, P. (eds), *Total Least Squares and Errors-in-Variables Modelling: Analysis, Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 57–66.



- Haber, E. (1997). *Numerical Strategies for the Solution of Inverse Problems*. Ph.D. Thesis, The University of British Columbia, Vancouver, Canada.
- Haber, E. and Oldenburg, D. (2000). A GCV based method for nonlinear ill-posed problems. *Computat. Geosci.* 4, 41–63.
- Hall, P. and Titterton, D. (1987). Common structure of techniques for choosing smoothing parameters in regression problems. *J. Roy. Stat. Soc.* 49, 184–198.
- Hanke, M. (1995). *Conjugate Gradient Type Methods for Ill-Posed Problems*. Pitman Research Notes in Mathematics 327, Longman, Harlow.
- Hanke, M. (1996). Limitations of the L-curve method in ill-posed problems. *BIT* 36, 287–301.
- Hanke, M. (1997). A regularizing Levenberg–Marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Problems* 13, 79–95.
- Hanke, M. and Groetsch, C. W. (1998). Nonstationary iterated Tikhonov regularization. *J. Opt. Theory Appl.* 98, 37–53.
- Hanke, M. and Hansen, P. C. (1993). Regularization methods for large-scale problems. *Surveys Math. Indust.* 3, 253–315.
- Hanke, M. and Raus, T. (1996). A general heuristic for choosing the regularization parameter in ill-posed problems. *SIAM J. Sci. Comput.* 17, 956–972.
- Hanke, M., Nagy, J. G., and Plemmons, R. J. (1993). Preconditioned iterative regularization methods for ill-posed problems. In Reichel, L., Ruttan, A., and Varga, R. S. (eds), *Numerical Linear Algebra*, de Gruyter, Berlin, pp 141–163.
- Hanke, M., Neubauer, A., and Scherzer, O. (1995). A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.* 72, 21–37.
- Hansen, J. E. (1971). Multiple scattering of polarized light in planetary atmospheres. Part I: The doubling method. *J. Atmos. Sci.* 28, 120–125.
- Hansen, J. E. and Travis, L. D. (1974). Light scattering in planetary atmospheres. *Space Sci. Rev.* 16, 527–610.
- Hansen, P. C. (1990). The discrete Picard condition for discrete ill-posed problems. *BIT* 30, 658–672.
- Hansen, P. C. (1992a). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev.* 34, 561–580.
- Hansen, P. C. (1992b). Numerical tools for analysis and solution of Fredholm integral equations of the first kind. *Inverse Problems* 8, 849–872.
- Hansen, P. C. (1994). The Backus–Gilbert method: SVD analysis and fast implementation. *Inverse Problems* 10, 895–904.
- Hansen, P. C. (1998). *Rank Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, Philadelphia, PA.
- Hansen, P. C. and O’Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.* 14, 1487–1503.
- Hanson, R. J. (1971). A numerical method for solving Fredholm integral equations of the first kind using singular values. *SIAM J. Numer. Anal.* 8, 616–622.
- Hasekamp, O. and Landgraf, J. (2001). Ozone profile retrieval from backscattered ultraviolet radiances: The inverse problem solved by regularization. *J. Geophys. Res.* 106, 8077–8088.

- Hasekamp, O. P., Landgraf, J., and Van Oss, R. (2002). The need of polarization modeling for ozone profile retrieval from backscattered sunlight. *J. Geophys. Res.* 107(D23), 4692. doi:10.1029/2002JD002387.
- Herman, B. and Browning, S. (1965). A numerical solution to the equation of radiative transfer. *J. Atmos. Sci.* 22, 559–566.
- Hestenes, M. R. and Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards* 49, 409–436.
- Hochbruck, M., Hönig, M., and Ostermann, A. (2009). A convergence analysis of the exponential Euler iteration for nonlinear ill-posed problems. *Inverse Problems* 25. doi: 10.1088/0266-5611/25/7/075009.
- Hofmann, B. (1997). On ill-posedness and local ill-posedness of operator equations in Hilbert spaces. Preprint 97–8, Technische Universität Chemnitz–Zwickau, Chemnitz, Germany.
- Hofmann, B. and Scherzer, O. (1994). Factors influencing the ill-posedness of nonlinear problems. *Inverse Problems* 10, 1277–1297.
- Hofmann, B. and Scherzer, O. (1998). Local ill-posedness and source conditions of operator equations in Hilbert spaces. *Inverse Problems* 14, 1189–1206.
- Hohage, T. (1997). Logarithmic convergence rates of the iteratively regularized Gauss–Newton method for an inverse potential and an inverse scattering problem. *Inverse Problems* 13, 1279–1299.
- Hohage, T. (2001). On the numerical solution of a three-dimensional inverse medium scattering problem. *Inverse Problems* 7, 1743–1763.
- Höpfner, M. and Emde, C. (2005). Comparison of single and multiple scattering approaches for the simulation of limb-emission observations in the mid-IR. *J. Quant. Spectrosc. Radiat. Transfer* 91, 275–285.
- Höpfner, M., Oelhaf, H., Wetzell, G., Friedl-Vallon, F., Kleinert, A., Lengel, A., Maucher, G., Nordmeyer, H., Glatthor, N., Stiller, G. P., von Clarmann, T., Fischer, H., Krüger, C., and Deshler, T. (2002). Evidence of scattering of tropospheric radiation by PSC's in mid-IR limb emission spectra: MIPAS-B observations and KOPRA simulations. *Geophys. Res. Letters* 29. doi:10.1029/2001GL014443.
- Huang, H. L. and Purser, R. J. (1996). Objective measures of the information density of satellite data. *Meteorol. Atmos. Phys.* 60, 105–117.
- Huiskes, M. J. (2002). *Automatic Differentiation Algorithms in Model Analysis*. Ph. D. Thesis, Wageningen University, The Netherlands.
- Hunt, B. R. (1973). The application of constrained least squares estimation to image restoration by digital computers. *IEEE Trans. Comput.* 22, 805–812.
- Jacquinet-Husson, N., Scott, N. A., Chedin, A., Crepeau, L., Armante, R., Capelle, V., Orphal, J., Coustenis, A., Boone, C., Poulet-Crovisier, N., Barbe, A., Birk, M., Brown, L. R., Camy-Peyret, C., Claveau, C., Chance, K., Christidis, N., Clerbaux, C., Coheur, P. F., Dana, V., Daumont, L., De Backer-Barilly, M. R., Di Lonardo, G., Flaud, J. M., Goldman, A., Hamdouni, A., Hess, M., Hurley, M. D., Jacquemart, D., Kleiner, I., Köpke, P., Mandin, J. Y., Massie, S., Mikhailenko, S., Nemtchinov, V., Nikitin, A., Newnham, D., Perrin, A., Perevalov, V. I., Pinnock, S., Regalia-Jarlot, L., Rinsland, C. P., Rublev, A., Schreier, F., Schult, L., Smith, K. M., Tashkun, S. A., Teffo, J. L., Toth, R. A., Tyuterev, V. G., Auwera, J. V., Varanasi, P., and Wagner, G. (2008). The

- GEISA spectroscopic database: Current and future archive for Earth and planetary atmosphere studies. *J. Quant. Spectrosc. Radiat. Transfer* 109, 1043–1059.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* 106, 620–630.
- Jin, Q. N. (2000). On the iteratively regularized Gauss–Newton method for solving nonlinear ill-posed problems. *Math. Comput.* 69, 1603–1623.
- Jin, Q. N. and Hou, Z. Y. (1997). On the choice of the regularization parameter for ordinary and iterated Tikhonov regularization of nonlinear ill-posed problems. *Inverse Problems* 13, 815–827.
- Jin, Q. N. and Hou, Z. Y. (1999). On an a posteriori parameter choice strategy for Tikhonov regularization of nonlinear ill-posed problems. *Numer. Math.* 83, 139–159.
- Joiner, J., Bhartia, P., Cebula, R., Hilsenrath, E., McPeters, R., and Park, H. (1995). Rotational Raman scattering (Ring effect) in satellite backscatter ultraviolet measurements. *Appl. Opt.* 34, 4513–4525.
- Kaipio, J. and Somersalo, E. (2005). *Statistical and Computational Inverse Problems*. Springer, New York.
- Kaltenbacher, B. (1997). Some Newton-type methods for the regularization of nonlinear ill-posed problems. *Inverse Problems* 13, 729–753.
- Kaltenbacher, B. (1998). A posteriori parameter choice strategies for some Newton type methods for the regularization of nonlinear ill-posed problems. *Numer. Math.* 79, 501–528.
- Kaltenbacher, B., Neubauer, A., and Scherzer, O. (2008). *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. Radon Series on Computational and Applied Mathematics, Walter de Gruyter, Berlin.
- Kaplan, L. D. (1959). Inference of atmospheric structure from remote radiance measurements. *J. Opt. Soc. Am.* 49, 1004–1007.
- Kattawar, G., Young, A., and Humphreys, T. (1981). Inelastic scattering in planetary atmospheres: I. The Ring effect, without aerosols. *Astrophys. J.* 243, 1049–1057.
- King, J. I. F. (1956). The radiative heat transfer of planet Earth. In Van Allen, J. A. (ed), *Scientific Uses of Earth Satellites*, University of Michigan Press, Ann Arbor, MI, pp. 133–136.
- Kirsch, A. (1996). *An Introduction to the Mathematical Theory of Inverse Problems*. Springer, Berlin.
- Kitagawa, G. and Gersch, W. (1985). A smoothness priors long AR model method for spectral estimation. *IEEE Trans. Autom. Contr.* 30, 57–65.
- Koner, P. K. and Drummond, J. R. (2008). Atmospheric trace gases profile retrievals using the nonlinear regularized total least squares method. *J. Quant. Spectrosc. Radiat. Transfer* 109, 2045–2059.
- Kozlov, V. P. (1983). Design of experiments related to the inverse problem of mathematical physics. In Ermakov, C. M. (ed), *Mathematical Theory of Experiment Design*, Nauka, Moscow, pp. 216–246.
- Kravaris, C. and Seinfeld, J. H. (1985). Identification of parameters in distributed parameter systems by regularization. *SIAM J. Control Optim.* 23, 217–241.
- Kress, R. (1999). *Linear Integral Equations*. Springer-Verlag, Heidelberg.
- Kuo, K. S., Weger, R. C., Welch, R. M., and Cox, S. K. (1996). The Picard iterative approximation to the solution of the integral equation of radiative transfer. Part II: Three-dimensional geometry. *J. Quant. Spectrosc. Radiat. Transfer* 55, 195–213.

- Lacis, A. A. and Oinas, V. (1991). A description of the correlated  $k$  distribution method for modeling nongray gaseous absorption, thermal emission, and multiple scattering in vertically inhomogeneous atmospheres. *J. Geophys. Res.* 96, 9027–9063.
- Landgraf, J., Hasekamp, O. P., Box, M. A., and Trautmann, T. (2001). A linearized radiative transfer model for ozone profile retrieval using the analytical forward-adjoint perturbation theory approach. *J. Geophys. Res.* 106(D21), 27291–27305.
- Landgraf, J., Hasekamp, O., Van Deelen, R., and Aben, I. (2004). Rotational Raman scattering of polarized light in the Earth atmosphere: A vector radiative transfer model using the radiative transfer perturbation theory approach. *J. Quant. Spectrosc. Radiat. Transfer* 87, 399–433.
- Landl, G. and Anderssen, R. S. (1996). Non-negative differentially constrained entropy-like regularization. *Inverse Problems* 12, 35–53.
- Lawson, C. L. and Hanson, R. J. (1995). *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ, 1974; reprinted by SIAM, Philadelphia, PA.
- Lenoble, J. (1985). *Radiative Transfer in Scattering and Absorbing Atmospheres: Standard Computational Procedures*. Deepak Publishing, Hampton, VA.
- Li, Y. and Oldenburg, D. W. (1999). 3-D inversion of DC resistivity data using an L-curve criterion. In *Extended Abstracts of 69th SEG Meeting*, Houston, 31 October–5 November.
- Liebe, H. J., Hufford G. A., and Cotton, M. G. (1993). Propagation modeling of moist air and suspended water/ice particles at frequencies below 1000 GHz. In *Proceedings of the NATO/AGARD 52nd Specialists' Meeting of the Electromagnetic Wave Propagation Panel*, Palma de Mallorca, Spain, pp. 1–10.
- Liou, K.-N. (2002). *An Introduction to Atmospheric Radiation*. Academic Press, San Diego, CA.
- Lopez-Puertas, M. and Taylor, F. W. (2001). *Non-LTE Radiative Transfer in the Atmosphere*. World Scientific Publishing, Singapore.
- Louis, A. K. (1996). Approximate inverse for linear and some nonlinear problems. *Inverse Problems* 12, 175–190.
- Louis, A. K. and Maass, P. (1990). A mollifier method for linear operator equations of the first kind. *Inverse Problems* 6, 427–440.
- Lukas, M. A. (1998a). Comparison of parameter choice methods for regularization with discrete noisy data. *Inverse Problems* 14, 161–184.
- Lukas, M. A. (1998b). Asymptotic behaviour of the minimum bound method for choosing the regularization parameter. *Inverse Problems* 14, 149–159.
- Mallows, C. L. (1973). Some comments on Cp. *Technometrics* 15, 661–676.
- Marchuk, G. I. (1964). Equation for the value of information from weather satellites and formulation of inverse problems. *Cosmic Res.* 2, 394–409.
- Marchuk, G.I. (1995). *Adjoint Equations and Analysis of Complex Systems*. Kluwer, Amsterdam.
- Marchuk, G. I., Mikhailov, G.A., Nazaraliev, M. A., Darbinjan, R. A., Kargin, B. A., and Elepov, B. S. (1980). *The Monte Carlo Method in Atmospheric Optics*. Springer, Berlin.
- Maschhoff, K. J. and Sorensen, D.C. (1996). P\_ARPACK: An efficient portable large scale eigenvalue package for distributed memory parallel architectures. In Wasniewski, J., Dongarra, J., Madsen, K., and Olesen, D. (eds), *Applied Parallel Computing in Indus-*

- trial Problems and Optimization*, Lecture Notes in Computer Science 1184, Springer, Berlin, pp. 478–486.
- Mateer, C. L. (1965). On the information content of Umkehr observations. *J. Atmos. Sci.* 22, 370–381.
- McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.
- McLinden, C. A., McConnell, J. C., Griffioen, E., and McElroy, C. T. (2002a). A vector radiative-transfer model for the Odin/OSIRIS project. *Can. J. Phys.* 80, 375–393.
- McLinden, C. A., McConnell, J. C., Strong, K., McDade, I. C., Gattinger, R. L., King, R., Solheim, B., Llewellyn, E. J., and Evans, W. J. (2002b). The impact of the OSIRIS grating efficiency on radiance and trace-gas retrievals. *Can. J. Phys.* 80, 469–481.
- Mendrok, J., Schreier, F., and Höpfner, M. (2007). Estimating cirrus cloud properties from MIPAS data. *Geophys. Res. Letters* 34, L08807. doi: 10.1029/2006GL028246.
- Menke, W. (1984). *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, Orlando, FL.
- Miller, K. (1970). Least squares methods for ill-posed problems with a prescribed bound. *SIAM J. Math. Anal.* 1, 52–74.
- Mishchenko, M. I., Lacis, A. A., and Travis, L. D. (1994). Errors induced by the neglect of polarization in radiance calculations for Rayleigh-scattering atmospheres. *J. Quant. Spectrosc. Radiat. Transfer* 51, 491–510.
- Moré, J. J. and Wright, S. J. (1993). *Optimization Software Guide*. SIAM, Philadelphia, PA.
- Moré, J. J., Garbow, B. S., and Hillstom, K. E. (1980). *User Guide for MINPACK-1*. Argonne National Laboratory Report ANL-80-74, Argonne, IL.
- Morozov, V. A. (1966). On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.* 7, 414–417.
- Morozov, V. A. (1968). The error principle in the solution of operational equations by the regularization method. *USSR Comp. Math. Math. Phys.* 8, 63–87.
- Morozov, V. A. (1984). *Methods for Solving Incorrectly Posed Problems*. Springer, New York.
- NAG Fortran Library Manual, Mark 16* (1993). NAG Ltd., Oxford.
- Natterer, F. (1977). Regularisierung schlecht gestellter Probleme durch Projektionsverfahren. *Numer. Math.* 28, 329–341.
- Nemirovskii, A. S. and Polyak, B. T. (1984). Iterative methods for solving linear ill-posed problems under precise information I. *Engrg. Cybernetics* 22, 1–11.
- Neubauer, A. (1989). Tikhonov regularisation for non-linear ill-posed problems: optimal convergence rates and finite-dimensional approximations. *Inverse Problems* 5, 541–557.
- Neumaier, A. (1998). Solving ill-conditioned and singular systems: A tutorial on regularization. *SIAM Review* 40, 636–666.
- Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York.
- Nychka, D. (1988). Confidence intervals for smoothing splines. *J. Amer. Stat. Assoc.* 83, 1134–1143.
- Oikarinen, L., Sihvola, E., and Kyrola, E. (1999). Multiple scattering in limb viewing geometry. *J. Geophys. Res.* 31, 261–274.
- O’Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* 1, 502–527.

- O'Sullivan, F. (1990). Convergence characteristics of a method of regularization estimators for nonlinear operator equations. *SIAM J. Num. Anal.* 27, 1635–1649.
- O'Sullivan, F. and Wahba, G. (1985). A cross validated Bayesian retrieval algorithm for nonlinear remote sensing experiments. *J. Comput. Phys.* 59, 441–455.
- Paige, C. C. and Saunders, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software* 8, 43–71.
- Parker, R. L. (1994). *Geophysical Inverse Theory*. Princeton University Press, Princeton, NJ.
- Peckham, G. (1974). The information content of remote measurements of the atmospheric temperature by satellite IR radiometry and optimum radiometer configurations. *Quart. J. Roy. Meteor. Soc.* 100, 406–419.
- Pickett, H. M., Poynter, R. L., Cohen, E. A., Delitsky, M. L., Pearson, J. C., and Mueller, H. S. P. (1998). Submillimeter, millimeter, and microwave spectral line catalog. *J. Quant. Spectrosc. Radiat. Transfer* 60, 883–890.
- Platt, U. and Stutz, J. (2008). *Differential Optical Absorption Spectroscopy. Principles and Applications*. Springer, Berlin.
- Powell, M. J. D. (1985). On the quadratic programming algorithm of Goldfarb and Idnani. *Math. Programming* 25, 46–61.
- Pumplin, J., Stump, D. R., and Tung, W. K. (2001). Multivariate fitting and the error matrix in global analysis of data. *Phys. Rev.* 65, 014011–1–7.
- Purser, R. J. and Huang, H.-L. (1993). Estimating effective data density in a satellite retrieval or an objective analysis. *J. Appl. Meteor.* 32, 1092–1107.
- Ramos, F. M., Velho, H. F. C., Carvalho, J. C., and Ferreira, N. J. (1999). Novel approaches to entropic regularization. *Inverse Problems* 15, 1139–1148.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.
- Raus, T. (1985). The principle of the residual in the solution of ill-posed problems with non-selfadjoint operators. *Acta Comment. Univ. Tartuensis* 715, 12–20.
- Reginska, T. (1996). A regularization parameter in discrete ill-posed problems. *SIAM J. Sci. Comput.* 17, 740–749.
- Renault, R. A. and Guo, H. (2005). Efficient algorithms for solution of regularized total least squares. *SIAM J. Matrix Anal. Appl.* 26, 457–476.
- Rice, J. A. (1986). Choice of smoothing parameter in deconvolution problems. *Contemporary Mathematics* 59, pp. 137–151.
- Rieder, A. (1999). On the regularization of nonlinear ill-posed problems via inexact Newton iterations. *Inverse Problems* 15, 309–327.
- Rieder, A. (2003). *Keine Probleme mit Inversen Problemen*. Vieweg, Wiesbaden.
- Rieder, A. and Schuster, T. (2000). The approximate inverse in action with an application to computerized tomography. *SIAM J. Numer. Anal.* 37, 1909–9120.
- Rodgers, C. D. (1976). Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Rev. Geophys. Space Phys.* 14, 609–624.
- Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific, Singapore.
- Rodriguez, G. and Theis, D. (2005). An algorithm for estimating the optimal regularization parameter by the L-curve. *Rendiconti di Matematica* 25, 69–84.
- Rothman, L. S., Gordon, I. E., Barbe, A., Benner, D. C., Bernath, P. F., Birk, M., Boudon, V., Brown, L. R., Campargue, A., Champion, J. -P, Chance, K., Coudert, L. H., Dana,



- V., Devi, V. M., Fally, S., Flaud, J. -M., Gamache, R. R., Goldman, A., Jacquemart, D., Kleiner, I., Lacombe, N., Lafferty, W. J., Mandin, J. -Y., Massie, S. T., Mikhailenko, S. N., Miller, C. E., Moazzen-Ahmadi, N., Naumenko, O. V., Nikitin, A. V., Orphal, J., Perevalov, V. I., Perrin, A., Predoi-Cross, A., Rinsland, C. P., Rotger, M., Simeckova, M., Smith, M. A. H., Sung, K., Tashkun, S. A., Tennyson, J., Toth, R. A., Vandaele, A. C., and Auwera, J. V. (2009). The HITRAN 2008 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transfer* 110, 533–572.
- Rozanov, A. (2001). *Modeling of Radiative Transfer through a Spherical Planetary Atmosphere: Application to Atmospheric Trace Gases Retrieval from Occultation- and Limb-Measurements in UV-Vis-NIR*. Ph.D. Thesis, University of Bremen, Germany.
- Rozanov, A. V., Rozanov, V. V., and Burrows, J. P. (2000). Combined differential-integral approach for the radiation field computation in a spherical shell atmosphere: Nonlimb geometry. *J. Geophys. Res.* 105, 22937–22942.
- Rozanov, A., Rozanov, V., and Burrows, J. P. (2001). A numerical radiative transfer model for a spherical planetary atmosphere: Combined differential-integral approach involving the Picard iterative approximation. *J. Quant. Spectrosc. Radiat. Transfer* 69, 491–512.
- Rozanov, A., Rozanov, V., Buchwitz, M., Kokhanovsky, A. and Burrows, J. P. (2005). SCIATRAN 2.0 – A new radiative transfer model for geophysical applications in the 175–2400 nm spectral region. *Adv. Space Res.* 36, 1015–1019.
- Rozanov, V. V. and Rozanov, A. V. (2007). Relationship between different approaches to derive weighting functions related to atmospheric remote sensing problems. *J. Quant. Spectrosc. Radiat. Transfer* 105, 217–242.
- Scherzer, O. (1993). Convergence rates of iterated Tikhonov regularized solutions of nonlinear ill-posed problems. *Numer. Math.* 66, 259–279.
- Scherzer, O. (1998). A modified Landweber iteration for solving parameter estimation problems. *Appl. Math. Optim.* 38, 45–68.
- Scherzer, O., Engl, H. W., and Kunisch, K. (1993). Optimal a posteriori parameter choice for Tikhonov regularization for solving nonlinear ill-posed problems. *SIAM J. Numer. Anal.* 30, 1796–1838.
- Schimpf, B. and Schreier, F. (1997). Robust and efficient inversion of vertical sounding atmospheric high-resolution spectra by means of regularization. *J. Geophys. Res.* 102, 16037–16055.
- Schreier, F. and Boettger, U. (2003). MIRART, a line-by-line code for infrared atmospheric radiation computations incl. derivatives. *Atmos. Ocean. Optics* 16, 262–268.
- Schreier, F. and Schimpf, B. (2001). A new efficient line-by-line code for high resolution atmospheric radiation computations incl. derivatives. In Smith, W. L. and Timofeyev, Y. (eds), *IRS 2000: Current Problems in Atmospheric Radiation*, A. Deepak Publishing, Hampton, VA, pp 381–384.
- Seidman, T. I. and Vogel, C. R. (1989). Well posedness and convergence of some regularization methods for non-linear ill posed problems. *Inverse Problems* 5, 227–238.
- Shannon, C. E. (1949). Communication in the presence of noise. *Proc. ICE* 37, 10–21.
- Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

- Sima, D. and Van Huffel, S. (2006). Level choice in truncated total least squares. In *Proceedings of the Fourth Total Least Squares and Errors-in-Variables Modeling Workshop*, Leuven, Belgium, 21–23 August.
- Sima, D., Van Huffel, S., and Golub, G. H. (2003). *Regularized Total Least Squares Based on Quadratic Eigenvalue Problem Solvers*. Technical Report SCCM–03–03, Stanford University, Stanford, CA.
- Sioris, C. E., Haley, C. S., McLinden, C. A., von Savigny, C., McDade, I. C., McConnell, J. C., Evans, W. F. J., Lloyd, N. D., Llewellyn, E. J., Chance, K. V., Kurosu, T. P., Murtagh, D., Frisk, U., Pfeilsticker, K., Bösch, H., Weidner, F., Strong, K., Stegman, J., and Mégie, G. (2003). Stratospheric profiles of nitrogen dioxide observed by OSIRIS on the Odin satellite. *J. Geophys. Res.* 108(D7), 4215. doi:10.1029/2002JD002672.
- Slijkhuis, S., Bargaen, A., Thomas, W., and Chance, K. (1999). Calculation of undersampling correction spectra for DOAS spectral fitting. In *Proceedings of the European Symposium on Atmospheric Measurements from Space, ESAMS'99*, Noordwijk, The Netherlands, 18–22 January, pp. 563–569.
- Snieder, R. (1991). An extension of Backus–Gilbert theory to nonlinear inverse problems. *Inverse Problems* 7, 409–433.
- Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L. (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge.
- Spurr, R. J. D. (2001). A linearized discrete ordinate radiative transfer model for atmospheric remote-sensing retrieval. *J. Quant. Spectrosc. Radiat. Transfer* 68, 689–735.
- Spurr, R. J. D. (2002). Simultaneous derivation of intensities and weighting functions in a general pseudo-spherical discrete ordinate radiative transfer treatment. *J. Quant. Spectrosc. Radiat. Transfer* 75, 129–175.
- Spurr, R. J. D. (2006). VLIDORT: A linearized pseudo-spherical vector discrete ordinate radiative transfer code for forward model and retrieval studies in multilayer multiple scattering media. *J. Quant. Spectrosc. Radiat. Transfer* 102, 316–342.
- Spurr, R. J. D. (2008). LIDORT and VLIDORT. Linearized pseudo-spherical scalar and vector discrete ordinate radiative transfer models for use in remote sensing retrieval problems. In Kokhanovsky, A. A. (ed), *Light Scattering Reviews 3*, Springer-Praxis, Chichester, pp. 229–271.
- Spurr, R., de Haan, J., Van Oss, R., and Vasikov, A. (2008). Discrete-ordinate radiative transfer in a stratified medium with first-order rotational Raman scattering. *J. Quant. Spectrosc. Radiat. Transfer* 109, 404–425.
- Stam, D. M., Aben, I., and Helderma, F. (2002). Skylight polarization spectra: Numerical simulation of the Ring effect. *J. Geophys. Res.* 107(D20), 4419. doi:10.1029/2001JD000951.
- Stamnes, K., Tsay, S.-C., Wiscombe, W., and Jayaweera, K. (1988). Numerically stable algorithm for discrete-ordinate method radiative transfer in multiple scattering and emitting layered media. *Appl. Opt.* 27, 2502–2509.
- Steck, T. (2002). Methods for determining regularization for atmospheric retrieval problems. *Appl. Opt.* 41, 1788–1797.



- Steck, T. and von Clarmann, T. (2001). Constrained profile retrieval applied to the observation mode of the Michelson Interferometer for Passive Atmospheric Sounding. *Appl. Opt.* 40, 3559–3571.
- Steinwagner, J., Schwarz, G., and Hilgers, S. (2006). Use of maximum entropy method as a regularization technique during the retrieval of trace gas profiles from limb sounding measurements. *J. Atmos. Oceanic Tech.* 23, 1657–1667.
- Stephens, G. (1994). *Remote Sensing of the Lower Atmosphere*. Oxford University Press, Oxford.
- Stiller, G. P., von Clarmann, T., Funke, B., Glatthor, N., Hase, F., Höpfner, M., and Linden, A. (2002). Sensitivity of trace gas abundances retrievals from infrared limb emission spectra to simplifying approximations in radiative transfer modelling. *J. Quant. Spectrosc. Radiat. Transfer* 72, 249–280.
- Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, PA.
- Tautenhahn, U. (1994). On the asymptotical regularization of nonlinear ill-posed problems. *Inverse Problems* 10, 1405–1418.
- Tautenhahn, U. (1997). On a general regularization scheme for nonlinear ill-posed problems. *Inverse Problems* 13, 1427–1437.
- Thomas, G. E. and Stamnes, K. (1999). *Radiative Transfer in the Atmosphere and Ocean*. Cambridge University Press, Cambridge.
- Thompson, A. M., Brown, J. C., Kay, J. W., and Titterton, D. M. (1991). A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 326–339.
- Thompson, A. M., Kay, J. W., and Titterton, D. M. (1989). A cautionary note about the crossvalidatory choice. *J. Statist. Comput. Simul.* 33, 199–216.
- Tikhonov, A. N. (1963a). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* 4, 1035–1038.
- Tikhonov, A. N. (1963b). Regularization of incorrectly posed problems. *Soviet Math. Dokl.* 4, 1624–1627.
- Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*. Wiley, New York.
- Tikhonov, A. N. and Glasko, V. B. (1965). Use of the regularization method in nonlinear problems. *USSR Comp. Math. and Math. Phys.* 5, 93–107.
- Trussel, H. J. (1983). Convergence criteria for iterative restoration methods. *IEEE Trans. Acoust. Speech Signal Processing* 31, 129–136.
- Trussel, H. J. and Civanlar, M. R. (1984). The feasibility solution in signal restoration. *IEEE Trans. Acoust. Speech Signal Processing* 32, 201–212.
- Tsidu, G. M. (2005). On the accuracy of covariance matrix: Hessian versus Gauss–Newton method in atmospheric remote sensing with infrared spectroscopy. *J. Quant. Spectrosc. Radiat. Transfer* 96, 103–121.
- Twomey, S. (1977). *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements*. Elsevier Science, Amsterdam.
- Urban, J., Baron, P., Lautie, N., Schneider, N., Dassas, K., Ricaud, P., and De La Noe, J. (2004). Moliere (v5): A versatile forward- and inversion model for the millimeter and sub-millimeter wavelength range. *J. Quant. Spectrosc. Radiat. Transfer* 83, 529–554.

- Ustinov, E. A. (2001). Adjoint sensitivity analysis of radiative transfer equation: temperature and gas mixing ratio weighting functions for remote sensing of scattering atmospheres in thermal IR. *J. Quant. Spectrosc. Radiat. Transfer* 68, 195–211.
- Ustinov, E. A. (2005). Atmospheric weighting functions and surface partial derivatives for remote sensing of scattering planetary atmospheres in thermal spectral region: General adjoint approach. *J. Quant. Spectrosc. Radiat. Transfer* 92, 351–371.
- Ustinov, E. A. (2008). Adjoint approach to evaluation of weighting functions for remote sensing of scattering planetary atmospheres in thermal spectral region with limb-viewing geometry. In *Geophysical Research Abstracts of EGU General Assembly 2008*, Vienna, 13–18 April, SRef-ID: 1607-7962/gra/EGU2008-A-10664.
- Van Huffel, S. and Vanderwalle, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia, PA.
- Van Roozendaal, M., Loyola, D., Spurr, R., Balis, D., Lambert, J. C., Livschitz, Y., Valks, P., Ruppert, T., Kenter, P., Fayt, C., and Zehner, C. (2006). Ten years of GOME/ERS-2 total ozone data: the new GOME Data Processor (GDP) Version 4: I. Algorithm description. *J. Geophys. Res.* 111, D14311.
- Van der Sluis, A. and Van der Vorst, H. A. (1986). The rate of convergence of conjugate gradients. *Numer. Math.* 48, 543–560.
- Varah, J. M. (1973). On the numerical solutions of ill-conditioned linear systems with applications to ill-posed problems. *SIAM J. Numer. Anal.* 10, 257–267.
- Vinod, H. D. and Ullah, A. (1981). *Recent Advances in Regression Methods*. Dekker, New York.
- Vogel, C. R. (1985). Numerical solution of a non-linear ill-posed problem arising in inverse scattering. *Inverse Problems* 1, 393–403.
- Vogel, C. R. (1996). Non-convergence of the L-curve regularization parameter selection method. *Inverse Problems* 12, 535–547.
- Vogel, C. R. (2002). *Computational Methods for Inverse Problems*. SIAM, Philadelphia, PA.
- Vountas, M., Rozanov, V., and Burrows, J. (1998). Ring effect: Impact of rotational Raman scattering on radiative transfer in Earth's atmosphere. *J. Quant. Spectrosc. Radiat. Transfer* 60, 943–961.
- Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* 14, 651–667.
- Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing splines. *J. Roy. Stat. Soc.* 45, 133–150.
- Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals Statist.* 13, 1378–1402.
- Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.
- Walter, H., Landgraf, J., and Hasekamp, O. P. (2004). Linearization of a pseudo-spherical vector radiative transfer model. *J. Quant. Spectrosc. Radiat. Transfer* 85, 251–283.
- Walter, H. H., Landgraf, J., Spada, F., and Doicu, A. (2005). Linearization of a radiative transfer model in spherical geometry. *J. Geophys. Res.* 111, D24304. doi:10.1029/2005JD007014.
- Wang, Y. and Yuan, Y. (2005). Convergence and regularity of trust region methods for nonlinear ill-posed problems. *Inverse Problems* 21, 821–838.

- Weese, J. (1993). A regularization method for nonlinear ill-posed problems. *Comput. Phys. Commun.* 77, 429–440.
- Wing, G. M. (1991). *A Primer on Integral Equations of the First Kind*. SIAM, Philadelphia, PA.
- Wiscombe, W. J. and Evans, J. W. (1977). Exponential-sum fitting of radiative transmission functions. *J. Chem. Phys.* 24, 416–444.
- Wu, L. (2003). A parameter choice method for Tikhonov regularization. *Electron. Trans. Numer. Anal.* 16, 107–128.
- Xu, P. and Rummel, R. (1994). Generalized ridge regression with applications in determination of potential fields. *Manuscr. Geodaet.* 20, 8–20.
- Zdunkowski, W., Trautmann, T. and Bott, A. (2007). *Radiation in the Atmosphere – A Course in Theoretical Meteorology*. Cambridge University Press, Cambridge.
- Zha, H. and Hansen, P. C. (1990). Regularization and the general Gauss–Markov linear model. *Math. Comp.* 55, 613–624.

# Index

- asymptotic regularization
  - exponential Euler method, 245
  - Runge–Kutta method, 241
- averaging kernel matrix, 52, 57
- Backus–Gilbert method, 273
- Bayes’ theorem, 108
- conjugate gradient for normal equations
  - algorithm, 146
  - convergence rate, 332
- constrained iteratively regularized Gauss–Newton method with
  - equality constraints, 228
  - inequality constraints, 229
- constraint
  - expected value, 64
  - norm, 64
  - vector, 63
- corner
  - L-curve, 79
  - residual curve, 80
- covariance matrix
  - a posteriori, 111
  - a priori profile, 43
  - data error, 139
  - instrumental noise, 41
  - noise error, 52, 113
  - normalized a priori profile, 44
  - smoothing error, 112
  - total error, 113
  - true state, 112
- curvature
  - intrinsic, 167
  - parameter-effect, 167
- data density, 118
- degree of freedom
  - noise, 115
  - signal, 114
- degree of nonlinearity
  - deterministic, 165
  - stochastic, 168
- density of information, 119
- direct regularization method for linear problems
  - a priori parameter choice method, 306
  - discrepancy principle, 307
  - error-free parameter choice methods, 313
  - generalized discrepancy principle, 310
- direct regularization method for nonlinear problems
  - a priori parameter choice method, 353
  - discrepancy principle, 354
- discrepancy principle
  - generalized, 69
  - linear problems, 69
  - nonlinear problems, 203, 206
- entropy
  - relative, 280
  - Shannon, 118
- equality-constrained Tikhonov regularization
  - with
    - constant vertical column, 214
    - variable vertical column, 214
- error
  - forward model, 41

- model parameter, 139
  - random, 112
- error patterns
  - a priori covariance matrix, 168
  - mean square error matrix, 195
- estimators
  - conditional mean, 109
  - maximum a posteriori, 109
  - maximum likelihood, 109
- expectation minimization, 128
- expected error estimation method
  - iterated, 202
  - linear problems, 67
  - multi-parameter problems, 98
  - nonlinear problems, 200
  - statistical inversion, 121
- filter factors
  - information operator method, 120
  - iterated Tikhonov regularization, 50
  - Landweber iteration, 143
  - LSQR method, 153
  - Runge–Kutta regularization method, 244
  - Tikhonov regularization, 50
  - truncated total least squares, 256, 385
- gain matrix, 110
- generalized cross-validation
  - linear problems, 74
  - multi-parameter problems, 94
  - nonlinear problems, 203, 208
  - statistical inversion, 132
- generalized inverse
  - continuous problems, 290
  - discrete problems, 30
  - regularized, 40
- generalized singular value decomposition, 45
- Hadamard's conditions, 27
- hierarchical models, 125
- ill-posedness of
  - continuous problems, 291
  - discrete problems, 29
- influence matrix, 55
- information
  - content, 119
  - matrix, 115
  - operator method, 120
- interpolant with
  - B-splines, 26
  - piecewise constant functions, 25
  - piecewise linear functions, 26
- iterated Tikhonov regularization
  - linear problems, 49
  - nonlinear problems, 209
- iterative regularization method for linear problems, 323
- iterative regularization method for nonlinear problems
  - with a priori information, 365
  - without a priori information, 373
- iteratively regularized Gauss–Newton method, 223
- Krylov subspace, 147
- L-curve, 65
- L-curve method
  - Backus–Gilbert approach, 277
  - discrete, 155
  - linear problems, 79
  - multi-parameter problems, 99
  - nonlinear problems, 204, 208
- L-surface method, 97
- Landweber iteration
  - linear problems, 141
  - nonlinear problems, 222
- least squares solution
  - continuous problems, 288
  - discrete problems, 31
- leaving-out-one lemma, 75
- LSQR method, 151
- marginalizing method, 137
- maximum entropy regularization
  - cross entropy, 281
  - first-order, 282
  - second-order, 282
- maximum likelihood estimation
  - linear problems, 77
  - multi-parameter problems, 95
  - nonlinear problems, 203, 208
  - statistical inversion, 126, 134
- mean square error matrix
  - linear problems, 56
  - nonlinear problems, 192

- minimum bound method
  - linear problems, 70
  - nonlinear problems, 206
- minimum distance function approach
  - multi-parameter problems, 97
  - nonlinear problems, 208
- minimum variance method, 122
- mollifier methods, 271
- multi-parameter regularization methods
  - complete, 94
  - incomplete, 98
- Newton–CG method, 237
- noise error
  - constrained, 54
  - expected value, 53
  - linear problems, 52
  - nonlinear problems, 192
  - random, 113
- noise error method
  - Backus–Gilbert method, 277
  - statistical inversion, 123
- noise variance estimators
  - generalized cross-validation, 136
  - maximum likelihood estimation, 136
  - unbiased predictive risk estimator method, 136
- normal equation
  - continuous problems, 288
  - discrete problems, 31
  - regularized, 40
- optimization methods
  - step-length method, 174
  - trust-region method, 178
- Picard coefficients, 59
- Picard condition
  - continuous problems, 291
  - discrete problems, 58
- preconditioning, 156, 186, 190
- predictive error
  - noise, 55
  - smoothing, 55
  - total, 55
- prewhitening, 171
- projection method, 25
  - inequality-constrained, 394
- quasi-Newton method, 175
- quasi-optimality criterion
  - multi-parameter problems, 95, 99
  - one-parameter problems, 78
- regularization by projection, 38
- regularization matrix
  - a priori profile covariance matrix, 43
  - first-order difference, 42
  - normalized, 44
  - second-order difference, 42
- regularization parameter choice methods
  - a posteriori, 69
  - a priori, 67
  - error-free, 74
- regularizing Levenberg–Marquardt method
  - with
    - step-length procedure, 233
    - trust-region procedure, 233
- residual
  - expected value, 62
  - norm, 62
  - vector, 62
- residual curve method
  - generalized, 82
  - ordinary, 80
- residual polynomials
  - conjugate gradient method, 328
  - LSQR method, 153, 343
  - semi-iterative methods, 144
- Ritz polynomial, 153
- Ritz values, 153
- Schwarzschild equation, 23
- search direction
  - Gauss–Newton method, 175
  - Newton method, 174
  - steepest descent method, 174
- semi-iterative regularization methods
  - Chebyshev method, 145
  - convergence rate, 326
  - $\nu$ -method, 146
- sensitivity analysis, 169
- singular value decomposition, 28
- smoothing error
  - constrained, 54
  - linear problems, 51
  - nonlinear problems, 192
  - random, 112
- quadratic programming
  - equality-constrained, 391

- source condition
  - linear problems, 305
  - nonlinear problems, 353, 366
- spread of averaging kernel, 58
- standard form
  - explicit transformations, 295
  - implicit transformations, 299
  - problem, 48
  - transformation, 48
- step-length procedure, 176
- stopping rules
  - Lepskij, 230
  - linear problems, 155
  - nonlinear problems, 224
- termination criteria
  - relative function convergence test, 182
  - relative gradient test, 179
  - X-convergence test, 179
- Tikhonov iterate computed by
  - bidiagonalization of Jacobian matrix, 185
  - iterative methods for normal equations, 186
  - standard iterative methods, 189
  - SVD, 185
- total error
  - constrained, 54
  - expected value, 53
  - linear problems, 51
  - nonlinear problems, 191
  - random, 112
- total least squares
  - formulation, 252
  - Lanczos truncated, 257
  - regularized for linear problems, 258
  - regularized for nonlinear problems, 267
  - truncated, 254
- trace lemma, 50
- trust-region procedure, 179
- unbiased predictive risk estimator method, 72
- white noise, 41