Yanfei Wang
Anatoly G. Yagola
Changchun Yang
Editors

Optimization and Regularization for Computational Inverse Problems and Applications

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Yanfei Wang
Anatoly G. Yagola
Changchun Yang

Optimization and Regularization for Computational Inverse Problems and Applications

Yanfei Wang
Anatoly G. Yagola
Changchun Yang

Optimization and Regularization for Computational Inverse Problems and Applications

With 41 figures

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Editors

Prof. Dr. Yanfei Wang
Key Laboratory of Petroleum Geophysics
Institute of Geology and Geophysics
Chinese Academy of Sciences
Beijing 100029, China
e-mail: yfwang@mail.iggcas.ac.cn

Prof. Dr. Anatoly G. Yagola
Department of Mathematics
Faculty of Physics
Lomonosov Moscow State University
Moscow 119991, Russia
e-mail: yagola@inverse.phys.msu.ru

Prof. Dr. Changchun Yang
Key Laboratory of Petroleum Geophysics
Institute of Geology and Geophysics
Chinese Academy of Sciences
Beijing 100029, China
e-mail: ccy@mail.iggcas.ac.cn

ISBN 978-7-04-028515-4
Higher Education Press, Beijing

ISBN 978-3-642-13741-9 e-ISBN 978-3-642-13742-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010928203

Cover design: Frido Steinen-Broo, EStudio Calamar, Spain

Printed on acid-free paper

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

Preface by Anatoly G. Yagola

This volume contains the papers presented by invited speakers of the first international workshop "Optimization and Regularization for Computational Inverse Problems and Applications". The workshop was organized under the auspices of the Chinese Academy of Sciences in the Institute of Geology and Geophysics, located in Beijing, the capital of China, and held during July 21–25, 2008, just before the opening of the Olympic Games. The workshop was sponsored by the National Natural Science Foundation of China, China-Russia Cooperative Research Project RFBR-07-01-92103-NFSC and the National "973" Key Basic Research Developments Program of China. The main goal of the workshop was to teach about 60 young Chinese participants (mostly geophysicists) how to solve inverse and ill-posed problems using optimization procedures. Eminent specialists from China, Russia (partially sponsored by the Russian Foundation of Basic Research), USA and Austria were invited to present their lectures. Some of them could not participate personally but all invited speakers found a possibility to write papers especially for this publication.

The book covers many directions in the modern theory of inverse and ill-posed problems – the variational approach, iterative methods, using *a priori* information for constructing regularizing algorithms, etc. But the most important for the papers is to show how these methods can be applied to effectively solving of practical problems in geophysics, astrophysics, vibrational spectroscopy, and image processing. This issue should encourage specialists in the inverse problems field not only to investigate mathematical methods and propose new approaches but also to apply them to processing of real experimental data. I would like to wish all of them great successes!

Lomonosov Moscow State University
Moscow, Russia
March 2010

Anatoly G. Yagola

Preface by Editors

The field of inverse problems has existed in many branches of physics, earth science, engineering and mathematics for a long time. From the beginning of the birth of the inversion theory, inverse problem with its modeling design and optimization becomes a multi-disciplinary subject, which has received much more attention nowadays. The aim of the inverse problems, modeling design and optimization is to provide a better, more accurate, and more efficient simulation in practical applications. Many methodologies for solving inverse problems employs optimization algorithms. At the same time, optimization community that employ methods of inverse modeling design could reduce the number of time-consuming analyses required by the typical optimization algorithms substantially. This book provides readers who do research in computational/applied mathematics, engineering, geophysics, medical science, image processing, remote sensing and atmospheric science a background of using regularization and optimization techniques for solving practical inverse problems.

The book covers advances of inversion theory and recent developments with practical applications. Particularly, it emphasizes combining optimization and regularization for solving inverse problems. The methods include standard regularization theory, Fejér processes for linear and nonlinear problems, balancing principle, extrapolated regularization, nonstandard regularization, nonlinear gradient method, nonmonotone (Barzilai-Borwein) method, subspace method and Lie group method. The practical applications include reconstruction problem for inverse scattering, molecular spectra data processing, quantitative remote sensing inversion, seismic inversion by Lie group method and gravitational lensing problem.

Uniqueness of this book is that it provides novel methods for both standard and nonstandard regularization and practical applications in frontiers of sciences. Each chapter is written by renown researchers in their research field respectively. Illustrations and tables are provided for better understanding of their ideas. Scientists, researchers, engineers and as well as graduate students engaged in applied mathematics, engineering, geophysics, medical science, image processing, remote sensing and atmospheric science will benefit from the contents of the book since

the book incorporates a background of using regularization and optimization techniques for solving practical inverse problems.

Chinese Academy of Sciences, Beijing *Yanfei Wang*
Lomonosov Moscow State University, Moscow *Anatoly G. Yagola*
Chinese Academy of Sciences, Beijing *Changchun Yang*
May 2010

Contents

List of Contributors

Boris Artamonov
Sternberg Astronomical Institute of Moscow State University,
Universitetskiy prospekt, 19, Moscow, Russia.
e-mail: bartamon@sai.msu.ru

Hui Cao
Johann Radon Institute for Computational and Applied Mathematics (RICAM),
Austrian Academy of Science,
Altenbergstrasse 69, 4040 Linz, Austria.
e-mail: hui.cao@ricam.oeaw.ac.at

Jin Cheng
Department of Mathematics,
Fudan University, Shanghai 200433, China.
e-mail: jcheng@fudan.edu.cn

Yuhong Dai
State Key Laboratory of Scientific and Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
P.O.Box 2719, Beijing 100190, China.
e-mail: dyh@lsec.cc.ac.cn

Ekaterina Koptelova
Sternberg Astronomical Institute of Moscow State University,
Universitetskiy prospekt, 19, Moscow, Russia.
e-mail: koptelova@xray.sai.msu.ru

Gulnara Kuramshina
Department of Physical Chemistry, Faculty of Chemistry,
Lomonosov Moscow State University,
Moscow 119991, Russia.
e-mail: kuramshi@phys.chem.msu.ru

Xiaowen Li
Research Center for Remote Sensing and GIS,
Beijing Normal University,
Beijing, 100875, China.
e-mail: lix@bnu.edu.cn

Hong Liu
Key Laboratory of Petroleum Geophysics,
Institute of Geology and Geophysics, Chinese Academy of Sciences,
Beijing 100029, China.
e-mail: liuhong@mail.igcas.ac.cn

Jijun Liu
Department of Mathematics,
Southeast University,
Nanjing, 210096, China.
e-mail: jjliu@seu.edu.cn

Shiqian Ma
Department of Industrial Engineering and Operations Research,
Columbia University,
New York, NY 10027-6902, USA.
e-mail: sm2756@columbia.edu

Qinghua Ma
College of Art and Science,
Beijing Union University,
Beijing, 100083, China.
e-mail: qinghua@ygi.edu.cn

Sergei Pereverzyev
Johann Radon Institute for Computational and Applied Mathematics (RICAM),
Austrian Academy of Science,
Altenbergstrasse 69, 4040 Linz, Austria.
e-mail: sergei.pereverzev@ricam.oeaw.ac.at

Elena Shimanovskaya
Sternberg Astronomical Institute of Moscow State University,
Universitetskiy prospekt, 19, Moscow, Russia.
e-mail: eshim@sai.msu.ru

Guozhong Su
School of Sciences, Hebei University of Technology,
Tianjin 300130, China.
e-mail: guozhong_su@hebut.edu.cn

Vladimir V. Vasin
Department of the Ill–posed Problems,
Institute of Mathematics and Mechanics UB RAS,

Ekaterinburg, 620219, Russia
e-mail: vasin@imm.uran.ru

Haibing Wang
School of Mathematics and Computational Science,
Hunan University of Science and Technology,
Xiangtan 411201, China.
e-mail: wanghb845@yahoo.com.cn

Yanfei Wang
Key Laboratory of Petroleum Geophysics,
Institute of Geology and Geophysics, Chinese Academy of Sciences,
Beijing 100029, China.
e-mail: yfwang@mail.iggcas.ac.cn

Tingyan Xiao
School of Sciences, Hebei University of Technology,
Tianjin 300130, China.
e-mail: ty_xiao@hebut.edu.cn

Pingli Xie
School of Sciences,
Henan University of Technology,
Zhengzhou 450001, China.
e-mail: plxie@fudan.edu.cn

Anatoly G. Yagola
Department of Mathematics, Faculty of Physics,
Lomonosov Moscow State University,
Moscow 119991, Russia.
e-mail: yagola@inverse.phys.msu.ru

Changchun Yang
Key Laboratory of Petroleum Geophysics,
Institute of Geology and Geophysics, Chinese Academy of Sciences,
Beijing 100029, China.
e-mail: ccy@mail.iggcas.ac.cn

Yaxiang Yuan
State Key Laboratory of Scientific/Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
Zhong Guan Cun Donglu 55, Beijing, 100190, China.
e-mail: yyx@lsec.cc.ac.cn

Yuan Zhao
School of Sciences, Hebei University of Technology,
Tianjin 300130, China.
e-mail: zhaoyuan135821@163.com

Part I
Introduction

Chapter 1
Inverse Problems, Optimization and Regularization: A Multi-Disciplinary Subject

Yanfei Wang and Changchun Yang

Abstract. Inverse problems, optimization, regularization and scientific computing as a multi-disciplinary subject are introduced in this introductory chapter.

1.1 Introduction

The field of inverse problems has existed in many branches of physics (geophysics), engineering and mathematics (actually the majority of the natural scientific problems) for a long time. Inverse problems theory has been widely developed within the past decades due partly to its importance of applications, the arrival on the scene of large computers and the reliable numerical methods. Examples like deconvolution in seismic exploration, image reconstruction, tomography and parameter identification, all require powerful computers and reliable solution methods to carry out the computation [27, 12, 17, 19, 1, 24].

Inverse problems consist in using the results of actual observations or indirect measurements to infer the model or the values of the parameters characterizing the system under investigation. A problem is ill-posed according to the French mathematician Hadamard [4] if the solution does not exist, or is not unique, or if it is not a continuous function of the data. Practical inverse problems are typically related with the case that noise in the data may give rise to significant errors in the estimate. Therefore, to suppress the ill-posedness, designing the proper inversion model and developing proper regularization and optimization algorithms play a vital role [12, 19, 15].

From the beginning of the birth of the inversion theory, inverse problem with its modeling design and optimization becomes a multi-disciplinary subject, which

Yanfei Wang and Changchun Yang
Key Laboratory of Petroleum Geophysics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China.
e-mail: {yfwang,ccy}@mail.iggcas.ac.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

has become more and more important nowadays [19] because the modeling design theory in applied science is not well known in the optimization community and there are no optimization algorithms that employ methods of inverse modeling design. The aim of the inverse problems modeling design and optimization is to provide a better, more accurate, and more efficient simulation in practical applications. Many methodologies for solving inverse problems employ optimization algorithms. At the same time, optimization algorithms that employ methods of inverse modeling design could potentially substantially reduce the number of time-consuming analyses required by the typical optimization algorithms substantially.

1.2 Examples about mathematical inverse problems

Example 1.2.1. Fredholm integral equations of the first kind

This is a very useful and practical description of the inverse problems: given the kernel function $k(x, y)$, the input $f(x)$, desire the output $h(x)$ through the integral equation

$$\int k(x, y)f(y) = h(x). \tag{1.2.1}$$

Inverse problem is to infer what f represents through an inverse process of the above equation. Problems like seismic migration, deconvolution, light attenuation require solving the above equations, please refer to [27, 19, 16, 18, 21] for practical examples.

Example 1.2.2. Ill-conditioning linear algebraic problems

This can be regarded as discrete ill-posed problems. Actually any inverse problems, computationally, can be reduced to solve a linear algebraic problem in the form

$$Ax = b, \tag{1.2.2}$$

where $A \in \mathbb{R}^{M\times N}$, $x \in \mathbb{R}^N$ and $b \in \mathbb{R}^M$. There are three cases of the problem:

Case 1: $M = N$, yielding a square system. If A is nonsingular (full rank), then $A^{-1}b$ is the unique solution to the linear equation. On the other hand, if A is singular, the solution to the equation either does not exist, or is not unique.

Case 2: $M > N$, yielding an over-determined system. If the rank of the rectangular matrix A is N, then the solution x^{lse} can be found in the least squares sense to the over-determined system of equation. And, x^{lse} is the unique solution to the problem with minimum norm.

Case 3: $M < N$, yielding an underdetermined system. Underdetermined linear systems involve more unknowns than equations. That is, there are more unknown parameters to be identified than the number of measurements obtained. In this case, the solution is never unique. To solve the problem, *a priori* knowledge or constraints on the solution must be incorporated [20, 22].

Example 1.2.3. The Laplace transform

The Laplace transform is very important for many applications (e.g., synchrotron radiation light attenuation problem [23]). In the notation of integral equations, the problem of Laplace transform of a signal can be written as

$$(Af)(s) = \int_{t_{\min}}^{t_{\max}} k(s,t)f(t)dt = g(s),\ s \in [s_{\min}, s_{\max}], \tag{1.2.3}$$

where the integral kernel is $k(s,t) = \exp(-st)$, f an input signal and g the observations.

Inverse Laplace transform is to estimate the input f from measurements g.

1.3 Examples in applied science and engineering

Example 1.3.1. Atmospheric inversion

We focus on atmospheric aerosols inverse problems. It is well-known that the characteristics of the aerosol particle size, which can be represented as a size distribution function in the mathematical formalism, say $n(r)$, play an important role in climate modeling due to their uncertainty. So, the determination of particle size distribution function becomes a basic task in aerosol research [10, 3].

For sun-photometer, the attenuation of the aerosols can be written as the integral equation of the first kind

$$\begin{aligned} \tau_{aero}(\lambda) &= d(\lambda) + \varrho(\lambda) \\ &= (Kn)(\lambda) + \varrho(\lambda) \\ &:= \int_0^\infty \pi r^2 Q_{ext}(r,\lambda,\eta)n(r)dr + \varrho(\lambda), \end{aligned} \tag{1.3.1}$$

where K defines an operator from parameter space F to the observation space O; r is the particle radius; $n(r)$ is the columnar aerosol size distribution (i.e. the number of particles per unit area per unit radius interval in a vertical column through the atmosphere); η is the complex refractive index of the aerosol particles; λ is the wavelength; $\varrho(\lambda)$ is the error/noise; and $Q_{ext}(r,\lambda,\eta)$ is the extinction efficiency factor from Mie theory. Since aerosol optical thickness (AOT) can be obtained from the measurements of the solar flux density with sun-photometers, one can retrieve the size distribution by the inversion of AOT measurements through the above equation. This type of method is called extinction spectrometry, which is not only the earliest method applying remote sensing to determining atmospheric aerosol size characteristics, but also the most mature method thus far.

Determining the particle size distribution function is an inverse problem, recent advances employing regularization strategies can be found in [16, 18, 21, 11, 2].

Example 1.3.2. Geophysical inversion

Geophysical inverse problems are not new issues. However to put the advanced inverse and optimized techniques into practice is still underdeveloped. Roughly speaking, the inversion in geophysical problems is to adjust models (parameterized models) to minimize the difference between theoretical and observation values as far as possible using different norms in different spaces. Geophysical inverse problems are closely related with deconvolution, where the operators are designed to bring the predicted output to the actual output as close as possible using different norms in different spaces. The essence is to recover the unknowns from the measured data, e.g., in reflection seismics, try to recover the reflectivity function knowing the seismic records (Fig. 1.1). Mathematically, the observed time series $d(t)$ can be expressed by the Fredholm integral equations of the first kind

$$d(t) = \int_{\Omega} K(t,\tau)m(\tau)d\tau, \tag{1.3.2}$$

where $m(t)$ is the model and $k(t,\tau)$ is the kernel function. On the assumption that the mapping kernel $k(t,\tau)$ is time shift-invariant, the above equation becomes

$$d(t) = \int_{\Omega} k(\tau)m(t-\tau)d\tau, \tag{1.3.3}$$

which is equivalent to

$$d(t) = \int_{\Omega} k(t-\tau)m(\tau)d\tau. \tag{1.3.4}$$

If we regard $k(t)$ as a seismic source wavelet, $m(t)$ as an impulse function, then $d(t)$ is a seismic trace which can be recorded by using geophones, and the problem can be formulated as a source-dependent operator form

$$d(t) = K_s(m)(t). \tag{1.3.5}$$

Solving $m(t) = K_s^{-1}d(t)$ is a deconvolution problem. If, sometimes, the source is unknown, then solving $m(t)$ is called the blind source deconvolution. If solve the source $k(t)$ with the knowledge of data $d(t)$ and a proper inverse operator K^{-1}, i.e., $k(t) = K^{-1}d(t)$, then the problem is called an inverse source problem.

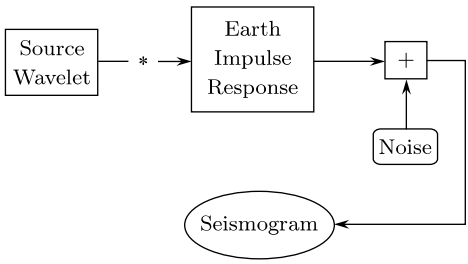


Fig. 1.1 Generating seismic records.

Note that the above model is a general description. Different geophysical inverse problems have different concrete forms and different physical meanings. For example, the operator equations can be referred to as inverse problems in gravity field, magnetic field, electromagnetic field and seismic wave field with different forms of kernel operators, model parameters and data, respectively [27].

Example 1.3.3. Inversion in Lidar sensing

Airborne laser scanning (ALS) is an active remote sensing technique which is also often referred to as lidar or laser radar. Due to the increasing availability of sensors, ALS has been receiving increasing attention in recent years (e.g., see [13]). In ALS a laser emits short infrared pulses towards the Earth's surface and a photodiode records the backscattered echo. With each scan, measurements are taken at the round-trip time of the laser pulse, the received echo power and at the beam angle in the locator's coordinate system. The round-trip time of the laser pulse allows calculating the range (distance) between the laser scanner and the object that generated the backscattered echo. Thereby, information about the geometric structure of the Earth's surface is obtained. The received power provides information about the scattering properties of the targets, which can be exploited for object classification and for modeling of the scattering properties.

Airborne laser scanning utilizes a measurement principle firstly strongly related to radar remote sensing. The fundamental relation to explain the signal strength in both techniques is the radar equation ([14]):

$$P_r(t) = \frac{D_r^2}{4\pi R^4 \beta_t^2} P_t \left(t - \frac{2R}{v_g} \right) \sigma, \tag{1.3.6}$$

where t is the time, R is the range, D_r is the aperture diameter of the receiver optics, β_t is the transmitter beam width, P_t is the transmitted power of the laser and σ denotes the scattering cross-section. The time delay is equal to $t' = 2R/v_g$ where v_g is the group velocity of the laser pulse in the atmosphere. Fig. 1.2 displays the geometry and parameters for laser scanner.

Taking the occurrence of multiple scatterers into account and regarding the impulse response $\Gamma(t)$ of the system's receiver, we get [14]

$$P_r(t) = \sum_{i=1}^{N} \frac{D_r^2}{4\pi R^4 \beta_t^2} P_t(t) \star \sigma_i'(t) \star \Gamma(t), \tag{1.3.7}$$

where $\star$ denotes the convolution operator. Since convolution is commutative, we can set $P_t(t) \star \sigma_i'(t) \star \Gamma(t) = P_t(t) \star \Gamma(t) \star \sigma_i'(t) = S(t) \star \sigma_i'(t)$, i.e. it is possible to combine both the transmitter and the receiver characteristics to a single term $S(t)$. This term is referred to as the *system waveform* ([14]). Thus, we are able to write our problem in the form

$$h(t) = \sum_{i=1}^{N} (f \star g)(t). \tag{1.3.8}$$

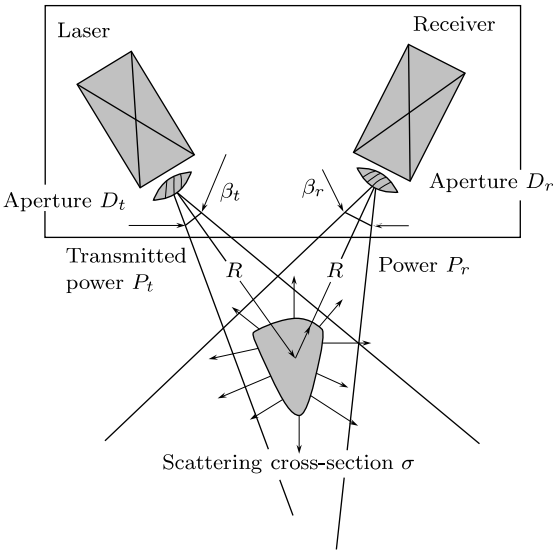


Fig. 1.2 Geometry and parameters for laser scanner.

where h is the incoming signal recorded by the receiver, f denotes a mapping which specifies the kernel function or point spread function and g is the unknown cross-section.

The received pulse consists of an effective wave $h_{eff}(t)$ and an additive noise $n(t)$, i.e.,

$$h(t) = h_{eff}(t) + n(t).$$

Therefore, it is quite important to stably retrieve the cross-section from equation (1.3.8) and suppress the perturbation simultaneously. We may write $h_{eff}(t)$ in the form

$$f \star g_{eff} = h_{eff}, \tag{1.3.9}$$

where g_{eff} denotes the actual backscatter cross-section.

Now the problem is how to deconvolve the convolution equation (1.3.8) to get the approximation to the actual cross-section g_{eff}. If we can identify an operator $b(t)$ which is the inverse of $f(t)$, then

$$b(t) \star h_{eff}(t) = f^{-1}(t) \star f(t) \star g_{eff}(t) = \delta(t) \star g_{eff}(t) = g_{eff}(t),$$

Though this is perfect in theory, this approach may not work well in practice. Numerically, the inverse of $f(t)$ is hard to obtain.

Example 1.3.4. Geochemical inversion

This is also a large topic, it can not be complete. For the marine geophysics and remote sensing, the inverse problem is to determine the bio-geochemical parameters of the water body from the upwelling radiance spectrum, or equivalently, from the spectral normalized water-leaving radiance or the spectral remote sens-

ing reflectance. Of course this is an ill-posed problem since we only know the radiance at the surface in a few directions.

For the subsurface aquifer systems, modeling reactive geochemical transport is a powerful tool for understanding and interpreting geochemical processes in it. A mathematical and numerical methodology for solving the inverse problem of water flow, heat transport, and multicomponent reactive solute transport in variably saturated media is very important in quantitative geochemistry. With proper inverse model, quantitative description of aqueous complexation, acid-base and redox reactions, cation exchange, proton surface complexation, and mineral dissolution and precipitation; identifying relevant geochemical processes; and estimating key reactive transport parameters from available hydrogeochemical data, can be obtained. Inverse modeling and optimized solution methods can provide optimum estimates of transmissivities, leakage rates, dispersivities, cation exchange capacity, cation selectivities, and initial and boundary concentrations of selected chemical components.

For the river water system, geochemistry utilizes (major and trace) elements, isotopes, and equations, to study various Earth and environmental processes. A combination of the experimental tools (elements and isotopes) with theoretical tools (equations) provides penetrating insights into the Earth and environmental processes. For example, with constraints on the parameters like Cl, Na, Sr, Ca and Mg, and solving the budget equations using inversion technique, the quantification of the elements originating from atmosphere and rock weathering can be obtained, and hence the end-members can be identified.

Example 1.3.5. Inverse problems in astronomy

Astronomy is by its essence an inverse problem in the sense that astronomers are attempting to determine the structure of the universe from remotely sensed data. The radio astronomy consists in the determination of the shape of celestial objects emitting radio waves, from the radio waves received by radio telescope on the surface of the Earth.

A typical example we are all familiar with is the astonishing hearten news about Hubble space telescope image restoration. It views the heavens without looking through the Earth's atmosphere. It can take pictures and analyze light.

Discussion about the inverse problems in astronomy is a huge topic. It deals with inverse problems of instrumentation and signal analysis, dynamics and oscillations, source distributions, spectral and polarimetric problems.

Solving the inverse problems in astronomy is difficult and at most of the time the problems are in large-scale. Hence robust optimization techniques are needed.

Example 1.3.6. Inversion in economy and finance

Inverse problems in econometric and finance are very important for human lives. In economics, functions are not usually explicitly given, but are assumed to have characteristic qualitative properties. The problems are meaningful and often generate important insights into economic behavior. Inverse problems in economics are often ill-posed, due to endogeneity (e.g., unobserved product information) and model misspecification problem (poor modeling).

The forward economic model is the generation of data Y_i. The general model is in the form of

$$Y_i = \Gamma(X_i) + \varepsilon_i, \tag{1.3.10}$$

where (Y_i, X_i) for all i are paired observation, $i = 1, \cdots, n$; the ε_i are independent, normally distributed random variables with equal variance σ^2; Γ is an unspecified function, which is usually assumed to be smooth. It can be also non-smooth for some special applications. There is a variety of estimation methods for the function Γ, including local polynomial estimators, spline regression and series estimators. The model (1.3.10) is usually called nonparametric because of the unspecified function Γ. A very familiar model to us is the choice of the function Γ as a polynomial function [24]:

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i + \varepsilon_i \tag{1.3.11}$$

for fixed $p \geq 1$. This is known as a linear parametric regression model.

The inverse problem is to estimate the unknown constant regression coefficients β_i, $i = 1, \cdots, p$.

Nowadays, financial problem has entered human being's daily life. Especially in recent years, financial markets and financial safety have become hot topics. Information on the major securities can be found easily in the daily press or on internet. Researchers also paid a great attention to this problem [19]. The well-known problem is the identification of the price-dependent volatilities from given option price data. A description of the relationship is the stochastic process

$$\frac{dX}{X} = \mu(X)dt + \sigma(X)dW_t, \tag{1.3.12}$$

where X is the price of the underlying asset; μ is the parameters drift; σ is the local volatility; W_t denotes a standard Wiener process. An European call option gives the holder the right to buy the underlying asset at the expiration date (or maturity) T for the strike price (or exercise price) K independent of the actual price X of the asset at time T.

We denote by $f(X, t, K, T)$ the (fair) market price of a call option as a function of the variable asset price X, time $t \geq 0$, strike price K and expiration date $T \geq t$. For fixed strike price K and expiration date T, the market price f satisfies the so-called Black-Scholes equation

$$\frac{\partial f}{\partial t} + r(t)X\frac{\partial f}{\partial X} + \frac{1}{2}\sigma^2(X)X^2\frac{\partial^2 f}{\partial X^2} - r(t)f = 0, \ (t, X) \in [0, T) \times (0, \infty), \tag{1.3.13}$$

where $r(t)$ denotes the interest rate of a risk-less investment and is assumed to be known, and the final condition $f(X, T, K, T) = \max(X - K, 0)$ for $X \in (0, \infty)$ holds.

We see that the volatility σ plays an important role in option pricing. It is suggested to introduce a mapping from the Black-Scholes equation

$$\sigma \longrightarrow f(X, T, K, T). \tag{1.3.14}$$

Now the forward problem is that to calculate option prices by a given volatility σ; the inverse problem is to identify the corresponding volatility function σ from given option price data.

A more general case is the volatility σ is a function as well of the asset price X as of the time t. Knowing the prices of European call options for all strike prices $K > 0$ and all maturities $T > t$ we can determine the corresponding volatility accordingly.

Example 1.3.7. Earth surface parameters inversion

Inverse problems in quantitative remote sensing are nowadays hot topics. Both modeling and model-based inversion are important for quantitative remote sensing [20, 22]. Hundreds of models related to vegetation and radiation have been established during past decades. The model-based inversion in atmospheric science has been well understood. However, the model-based inverse problems for land surface received much attention from scientists only in recent years. Compared to modeling, model-based inversion is still in the stage of exploration. This is because intrinsic difficulties exist in the application of *a priori* information, inverse strategy and inverse algorithm. The appearance of hyperspectral and multiangular remote sensor enhanced the exploration means, and provided us more spectral and spatial dimension information than before. However, how to utilize these information to solve the problems faced in quantitative remote sensing to make remote sensing really enter the time of quantification is still an arduous and urgent task for remote sensing scientists.

Example 1.3.8. Inversion in high energy physics

Using synchrotron radiation for detection of X-Ray Absorption Fine Structure (XAFS) is an important problem in high energy physics. The XAFS refers to modulations in X-ray absorption coefficient around an X-ray absorption edge. XAFS is often divided into "EXAFS" (Extended X-ray Absorption Fine Structure) and "XANES" (X-ray Absorption Near Edge Structure). The physical origin of EXAFS and XANES is basically the same. X-ray absorption probability can be calculated using standard quantum theory. In recent years, attenuation filter method receives much more attention. It is because on one side, the experiment can be simply established, on the other side, numerical simulations using computers can be easily performed. Different experiment filter setup may lead to different operator equations, but the essence is the same. We recall a EXAFS equation for a single-component material established in [7]

$$(Ag)(k) := \int_0^\infty NK(k, r)g(r)dr = \chi(k), \tag{1.3.15}$$

where A is a mapping operator, $g(r)$ is the atomic radial distribution function, N is the coordination number, $\chi(k)$ is the oscillatory part of the observed X-ray spectrum in the EXAFS energy region, $K(k, r)$ is a kernel function given by

$$K(k,r) = \frac{S_0^2(k)F(k,r)}{k}\frac{g(r)}{r^2}e^{-2r/\lambda(k)}\sin(2kr+\phi(k,r)), \tag{1.3.16}$$

where $S_0^2(k)$ is the many-electron overlap factor, $F(k,r)$ is the effective backscattering amplitude, r is the distance from the center of the excited atom, $\lambda(k)$ is the energy-dependent mean free path of the photoelectron, $\phi(k)$ is the phase shift due to the atomic potentials, and k is the photoelectron wave number.

Solving $g(r)$ is an inverse problem. For different types of atoms r_j and the measurement spectrum data vector $\chi(k)$ for wave vector k, the above problem can be solved numerically. However, the solution is highly unstable because of the ill-posedness. Regularization techniques should be included if need be.

Example 1.3.9. Inversion in life sciences

Life sciences is a fast growing field for mathematical modeling. An important step in modeling is to determine parameters from measurements. Inverse protein folding problems are a bit classical inverse problems in life sciences. Solving the inverse problems in life sciences usually leads to large-scale inverse problems, e.g., the simultaneous determination of hundreds of rate constants in very large reaction diffusion systems. We emphasize that many mathematical models in the life sciences are just now being developed, therefore, much more work needs to be done.

1.4 Basic theory

The basic theory for solving the inverse problems is the regularization, which had been established in the last century [8, 9, 6, 5]. It involves solving operator equations of the first kind

$$Kf = h \tag{1.4.1}$$

by regularization strategies, where $K : F \to H$ is a linear or nonlinear compact operator between Hilbert spaces F and H. The details and variations of the Tikhonov regularization will be discussed in the following chapters.

1.5 Scientific computing

Practical inverse problems involve scientific computing, and in many cases it may be large-scale computational problems, e.g., majority of geophysical and atmospherical inverse problems. Scientific computing is the collection of tools, techniques and theories which is required to solve on a computer mathematical model of problems in science and engineering. The primary focus is the development of algorithms and software for solving numerical problems that can enable large-scale modeling and simulations from a variety of disciplines.

Large-scale scientific computing as a relatively new field of science, grows very fast due to a steady increase of collaborative researches conducted by natural scientists and engineers, applied mathematicians and computer scientists, covering a variety of disciplines from the physical and chemical sciences to many engineering disciplines. The large scale scientific computing technique has become a motor of development. A modern engineer needs good knowledge of numerical analysis and scientific computing technique in order to contribute to engineering projects. Large scale scientific computing typically solves very large-scale problems. These problems are "large" not only in enormous variables but also in "mass data", e.g., the problems of geophysics and remote sensing. Solving large-scale problems requires advanced optimization technique and high performance computers to finish the numerical computations fast enough, or to solve a problem with enough precision within a given time [25, 26]. This consideration is more related with economy and commercial production. For example, problems in oil and gas detection, problems in numerical weather prediction, and problems in data assimilation are required to couple biosphere, hydrosphere and atmosphere [19]. The goal of large-scale scientific computing is the design, verification, implementation, and analysis of numerical, symbolic, and geometrical methods for solving large-scale direct and inverse problems with constraints, and their synergetic use in scientific computing for real life problems of high complexity.

1.6 Conclusion

In this chapter, we briefly introduce the basic concepts of inverse problems, multidisciplinary applications in both mathematics and new frontiers of applied science and engineering. Scientific computing problems arising as important tools for numerical inversion are briefly mentioned.

Acknowledgements This work is supported by the National Natural Science Foundation of China under grant numbers 10871191, 40974075 and National "973" Key Basic Research Developments Program of China under grant numbers 2005CB422104 and 2007CB714400.

References

1. M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics Publishing, Bristol and Philadelphia, 1998.
2. C. Bockmann and A. Kirsche, Iterative regularization method for lidar remote sensing, *Computer Physics Communications*, **174**, 607–615, 2006.
3. G. F. Bohren and D. R. Huffman, *Absorption and Scattering of Light by Small Particles*, John Wiley and Sons, New York, 1983.
4. J. Hadamard, *Lectures on the Cauchy Problems in Linear Partial Differential Equations*, Yale University Press, New Haven, 1923.

5. V. K. Ivanov, V. V. Vasin and V. P. Tanana, *Theory of Linear Ill-posed Problems and Its Applications*, Moscow: Nauka, 1978 (in Russian).
6. M. Z. Nashed, *Generalized Inverses and Applications*, Academic Press, New York, 1976.
7. E. A. Stern, D. E. Sayers and F. W. Lytle, Extended x-ray absorption fine structure technique. III. Determination of physical parameters, *Phys. Rev. Lett. B*, **11**, 4836–4846, 1975.
8. A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*, New York, John Wiley and Sons, 1977.
9. A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, Dordrecht, Kluwer, 1995.
10. S. Twomey, *Atmospheric Aerosols*, Elsevier Sci. Publ. Company, Amsterdam etc, 1977.
11. A. Voutilainen and J. P. Kaipio, Statistical inversion of aerosol size distribution data, *J. Aerosol Sci*, **31**(Suppl.1), 767–768, 2000.
12. C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, 2002.
13. W. Wagner, A. Ullrich, T. Melzer, C. Briese and K. Kraus, From single-pulse to full-waveform airborne laser scanners: potential and practical challenges, International Archives of Photogrammetry, *Remote Sensing and Spatial Information Sciences*, **35**, 201–206, 2004.
14. W. Wagner, A. Ullrich, V. Ducic, T. Melzer and N. Studnicka, Gaussian decomposition and calibration of a novel small-footprint full-waveform digitising airborne laser scanner, *ISPRS Journal of Photogrammetry & Remote Sensing*, **60**, 100–112, 2006.
15. Y. F. Wang and Y. X. Yuan, Convergence and regularity of trust region methods for nonlinear ill-posed inverse problems, *Inverse Problems* **21**, 821–838, 2005.
16. Y. F. Wang, S. F. Fan, X. Feng, G. J. Yan and Y. N. Guan, Regularized inversion method for retrieval of aerosol particle size distribution function in W space, *Applied Optics*, **45**(28), 7456–7467, 2006.
17. Y. F. Wang, Z. W. Wen, M. Z. Nashed and Q. Y. Sun, A direct fast method for time-limited signal reconstruction, *Applied Optics*, **45**, 3111–3126, 2006.
18. Y. F. Wang, S. F. Fan and X. Feng, Retrieval of the aerosol particle size distribution function by incorporating *a priori* information, *Journal of Aerosol Science*, **38**(8), 885–901, 2007.
19. Y. F. Wang, *Computational Methods for Inverse Problems and Their Applications*, Higher Education Press, Beijing, 2007.
20. Y. F. Wang, X. W. Li, Z. Nashed, F. Zhao, H. Yang, Y. N. Guan and H. Zhang, Regularized kernel-based BRDF model inversion method for ill-posed land surface parameter retrieval, *Remote Sensing of Environment*, **111**, 36–50, 2007.
21. Y. F. Wang, An efficient gradient method for maximum entropy regularizing retrieval of atmospheric aerosol particle size distribution function, *Journal of Aerosol Science*, **39**(4), 305–322, 2008.
22. Y. F. Wang, C. C. Yang and X. W. Li, A regularizing kernel-based BRDF model inversion method for ill-posed land surface parameter retrieval using smoothness constraint, *Journal of Geophysical Research*, **113**, D13101, doi:10.1029/2007JD009324, 2008.
23. Y. F. Wang, Y. H. Du and T. D. Hu, Projected gradient methods for synchrotron radiation spectra distribution function reconstruction, *Inverse Problems in Science and Engineering*, 2009.
24. T. Y. Xiao, S. G. Yu and Y. F. Wang, *Numerical Methods for the Solution of Inverse Problems*, Science Press, Beijing, 2003.
25. Y. X. Yuan, *Numerical Methods for Nonliear Programming*, Shanghai Science and Technology Publication, Shanghai, 1993.
26. Y. X. Yuan, Subspace methods for large scale nonlinear equations and nonlinear least squares, *Optimizaiton and Engineering*, **10**(2), 2009.
27. M. S. Zhdanov, *Geophysical Inverse Theory and Regularization Problems*, Elsevier, Amsterdam, 2002.

Part II
Regularization Theory and Recent Developments

Chapter 2
Ill-Posed Problems and Methods for Their Numerical Solution

Anatoly G. Yagola

Abstract. In the present chapter, the basic conceptions of the theory of ill-posed problems and numerical methods for their solving under different *a priori* information are described. Hadamard's definition of well-posedness and examples of ill-posed problems are given. Tikhonov's definition of a regularizing algorithm and classification of mathematical problems are described. The main properties of ill-posed problems are discussed. As an example of *a priori* information application for constructing regularizing algorithms an operator equation in Hilbert spaces is considered. If it is known that the exact solution belongs to a compact set then the quasisolution method can be used. An error of an approximate solution can be calculated also. If it is known that there is an *a priori* information concerning sourcewise representability of an exact solution with a completely continuous operator then the method of extending compacts can be applied. There exists a possibility to calculate an *a posteriori* error of an approximate solution. If strong *a priori* constraints are not available then the variational approach based on minimization of the Tikhonov functional with a choice of a regularization parameter, e.g., according to the generalized discrepancy principle is recommended. It is formulated by an equivalence of the generalized discrepancy principle and the generalized discrepancy method resulting in a possibility of the generalized discrepancy principle modification for solving incompatible ill-posed problems. Possible approaches for solving nonlinear ill-posed problems and iterative methods are described briefly.

Anatoly G. Yagola
Department of Mathematics, Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia.
e-mail: yagola@inverse.phys.msu.ru

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

2.1 Well-posed and ill-posed problems

Below we will describe fundamentals of the theory of ill-posed problems so as to constitute numerical methods for their solution if different *a priori* information is available. For simplicity, only linear equations in normed spaces are considered, although, it is clear that all similar definitions can be introduced also for nonlinear problems in more general metric (even also topological) spaces.

Let us consider an operator equation:

$$Az = u,$$

where A is a linear operator acting from a Hilbert space Z into a Hilbert space U. It is required to find a solution of the operator equation z corresponding to a given inhomogeneity (or right-hand side) u.

This equation is a typical mathematical model for many physical so called inverse problems if it is supposed that unknown physical characteristics z cannot be measured directly. As results of experiments, it is possible to obtain only data u connected with z with help of an operator A.

French mathematician J. Hadamard formulated the following conditions of well-posedness of mathematical problems. Let us consider these conditions for the operator equation above. The problem of solving the operator equation is called *well-posed* problem (according to Hadamard) if the following three conditions are fulfilled:

1. the solution holds $\forall u \in U$;
2. the solution is unique;
3. if $u_n \to u$, $Az_n = u_n$, $Az = u$, then $z_n \to z$.

Condition 2) can be realized then and only then operator A is one-to-one (injective). Conditions 1) and 2) imply that an inverse operator A^{-1} exists, and its domain $D(A^{-1})$ (or the range $R(A)$ of operator A) coincides with U. It is equivalent to that operator A is bijective. Condition 3) means that the inverse operator A^{-1} is continuous, i.e., to "small" perturbations of the right-hand side u "small" changes of the solution z correspond to. Moreover, J. Hadamard believed that well-posed problems only can be considered while solving practical problems. However, there are a lot of well-known examples of ill-posed problems that should be numerically solved when practical problems are treated. It should be noted that stability or instability of solutions depends on definition of the space of solutions Z. Usually, a choice of the space of solutions (including a choice of the norm) is determined by requirements of an applied problem. A mathematical problem can be ill-posed or well-posed depending on a choice of a norm in a functional space.

Numerous inverse (including ill-posed) problems can be found in different branches of physics. E.g., it is impossible for an astrophysicist to influence actively the processes in remote stars and galaxies. He is induced to make conclusions about physical characteristics of very remote objects using their indirect

manifestations measured on the Earth surface or near the Earth on space stations. Excellent examples of ill-posed problems are in medicine. Firstly, let us point out computerized tomography. A lot of applications of ill-posed problems are in geophysics. Indeed, it is easier and cheaper to judge about what is going on under the Earth surface for solving inverse problems than drilling deep boreholes. Other examples are in radio astronomy, spectroscopy, nuclear physics, plasma diagnostics, etc., etc.

The Fredholm integral equation of the 1st kind is a very well-known sample of an ill-posed problem. Let an operator A be of the form:

$$Az \equiv \int_a^b K(x,s)z(s)\,\mathrm{d}s = u(x), \qquad x \in [c,d].$$

Let the kernel of the integral operator $K(x,s)$ be a function continuously depending on a set of arguments $x \in [c,d]$, $s \in [a,b]$, and the solution $z(s)$ be continuous on the segment $[a,b]$ function. Then let us consider the operator A as acting between the following spaces: $A : C[a,b] \to C[c,d]$. (The space $C[a,b]$ consists of functions that are continuous on the segment $[a,b]$. The norm in this space of the element $z \in C[a,b]$ is defined as $\|z\|_{C[a,b]} = \max_{s\in[a,b]} |z(s)|$.) Let us show that in this case the problem is ill-posed. It is necessary to check conditions of well-posedness:

1) Existence of a solution for any continuous $[c,d]$ function $u(x)$. In truth, it is not so. There exists an infinite number of continuous functions such that the integral equation has no solutions.

2) Uniqueness of a solution. This condition is true if and only if the kernel of the integral operator is closed.

The first two conditions are equivalent to existence of an inverse operator A^{-1} with the domain $D(A^{-1}) = C[c,d]$. If the kernel of the integral operator is closed then the inverse operator exists but its domain does not coincide with $C[c,d]$.

3) Stability of a solution. It means that for any sequence $u_n \to \bar{u}$ ($Az_n = u_n$, $A\bar{z} = \bar{u}$) the sequence $z_n \to \bar{z}$. Stability is equivalent to continuity of the inverse operator A^{-1} if it exists. In this case it is not true. Let us consider an example. Let the sequence of continuous functions $z_n(s)$, $n = 1, 2, \ldots$, $s \in [a,b]$, be such that $z_n(s) \neq 0$ on the segment $[\frac{a+b}{2} - d_n, \frac{a+b}{2} + d_n]$ and equal to zero out-of-segment, $\max|z_n(s)| = 1$, $s \in [a,b]$, and a numerical sequence $d_n \to 0 + 0$. Such functions could be chosen, e.g., as being piecewise linear. Then for any $x \in [c,d]$

$$|u_n(x)| = \left| \int_a^b K(x,s)z_n(s)\,\mathrm{d}s \right| = \left| \int_{\frac{a+b}{2}-d_n}^{\frac{a+b}{2}+d_n} K(x,s)z_n(s)\,\mathrm{d}s \right| \leq K_0 \cdot 1 \cdot 2d_n \to 0$$

as $n \to \infty$, where $K_0 = \max|K(x,s)|$, $x \in [c,d]$, $s \in [a,b]$.

The functional sequence $u_n(x)$ uniformly (that is in norm of $C[c,d]$) converges to $\bar{u} = 0$. Though the solution of the equation $A\bar{z} = \bar{u}$ in this case is $\bar{z} = 0$ the sequence z_n does not converges to $\bar{z}$ so far as $\|z_n - \bar{z}\|_{C[a,b]} = 1$.

The integral operator A is completely continuous (compact and continuous) while acting from $L_2[a,b]$ into $L_2[c,d]$, from $C[a,b]$ into $L_2[c,d]$ and from $C[a,b]$ into $C[c,d]$. (The functional space $L_2[a,b]$ consists of functions which are square integrable on the segment $[a,b]$. The norm $z \in L_2[a,b]$ is defined as $\|z\|_{L_2[a,b]} = \{\int_a^b z^2(s)\,\mathrm{d}s\}^{1/2}$). It means that the operator transforms any bounded sequence to a compact sequence. By definition, from any subsequence of a compact sequence it is possible to select a converging subsequence. It is easy to indicate a sequence z_n, $\|z_n\|_{L_2[a,b]} = 1$, from which it is impossible to select a converging in $C[a,b]$ subsequence. For instance,

$$z_n(x) = \left(\frac{2}{b-a}\right)^{1/2} \sin\frac{\pi n(x-a)}{b-a}, \qquad n = 1,2,\ldots.$$

Norms of all terms of this sequence are equal to 1 in $L_2[a,b]$, so the sequence is bounded. But from any subsequence of the sequence it is impossible to select a converging subsequence because $\|z_i - z_j\| = \sqrt{2}$, $i \neq j$. Obviously, all functions $z_n(x)$ are continuous on $[a,b]$, and the sequence $z_n(x)$, $n = 1,2,\ldots$, is uniformly (in norm of $C[a,b]$) bounded. From this sequence it is impossible to select a converging in $C[a,b]$ subsequence (then it converges also in $L_2[a,b]$ as far as convergence in average follows from uniform convergence). Let us suppose that the operator A^{-1} is continuous. It is very easy to arrive at a contradiction. If A is an injective operator then an inverse operator exists. Evidently, if an operator B: $C[c,d] \to C[a,b]$, is continuous and an operator A is completely continuous then the operator BA: $C[a,b] \to C[a,b]$, is completely continuous also. Since for any n

$$A^{-1}Az_n = z_n,$$

the sequence z_n is compact, and that is not true. An operator that is inverse to a completely continuous operator cannot be continuous. A similar proof can be provided for any infinite dimensional Banach (i.e. full normed) space.

Since the problem of solving the Fredholm integral equation of the 1st kind is ill-posed in the above mentioned spaces, even very small errors in the right-hand side $u(x)$ can produce a result that the equation is not solvable or the solution exists but it differs *ad lib* strongly from the exact solution.

Thus, a completely continuous operator acting in infinite dimensional Banach spaces has an inverse operator that is not continuous (not bounded). Moreover, a range of a completely continuous operator acting between infinite dimensional Banach spaces is not closed. Therefore, in any neighborhood of the right-hand side $u(x)$ such that the equation has a solution, there exists infinite number of right-hand sides such that the equation is not solvable.

A mathematical problem can be ill-posed in connection also with errors in an operator. The simplest example gives the problem to find a normal pseudosolution of a system of linear algebraic equations. Instability of this problem is determined by errors in a matrix.

Let us consider a system of linear algebraic equations (SLAE):

$$Ax = b, \qquad x \in \mathbb{R}^n, \quad b \in \mathbb{R}^m, \qquad A : \mathbb{R}^n \to \mathbb{R}^m.$$

The system can be solvable or unsolvable. In the beginning of the 19th century Gauss and Legendre independently proposed the least squares method. Instead of solving SLAE, they suggested to minimize a quadratic functional (discrepancy):

$$\Phi(x) = \|Ax - b\|^2 = (A^*Ax, x) - 2 \cdot (A^*b, x) + (b, b),$$

A^* is a conjugate (transposed) matrix. The matrix A^*A is nonnegatively definite, so $\Phi(x)$ is a convex functional. For a convex differentiable functional the problem to find $\min_{x\in\mathbb{R}^n} \Phi(x)$ is equivalent to that to find a stationary point, notably solving an equation $\Phi'(x) = 0$. It easy to see that $\Phi'(x) = 2 \cdot (A^*Ax - A^*b)$, $\Phi''(x) = 2 \cdot A^*A \geq 0$. Then an equation $\Phi'(x) = 0$ (the gradient of the discrepancy is equal to zero) is turned into a system of linear algebraic equations with a square nonnegatively definite matrix (system of normal equations):

$$A^*Ax = A^*b.$$

In a finite dimensional case it easy to prove that the system of normal equations has a solution for any vector b (it maybe does not exist for the original SLAE). This solution is called a pseudosolution (or least squares solution) of the SLAE $Ax = b$. The pseudosolution can be nonunique if the determinant $\det(A^*A) = 0$. If $\det(A^*A) \neq 0$ then the pseudosolution is unique. The set of pseudosolutions is an affine (or linear) subspace in $\mathbb{R}^n$, it is convex and closed.

If the system $Ax = b$ has a solution then it coincides with a solution of the system $A^*Ax = A^*b$. In this case, $\min_{x\in\mathbb{R}^n} \Phi(x) = \mu = 0$. But if $\min_{x\in\mathbb{R}^n} \Phi(x) = \mu > 0$, then the system $Ax = b$ has no solutions, though as shown above there exists a pseudosoltion (maybe nonunique). The number μ is called a measure of incompatibility of the system $Ax = b$.

Definition 2.1.1. *A normal pseudosolution x_n of the system $Ax = b$ is a pseudosolution with a minimal norm, i.e., it is a solution of extreme problem: find minimum* $\min_{x:A^*Ax=A^*b} \|x\|$.

It easy to prove that a normal pseudosolution exists and is unique. We can formulate a problem: it is given a vector b. Let us find a correspondence to its normal pseudolution x_n. An operator A^+ that realizes this correspondence is linear. It is called a pseudoinverse to A operator: $x_n = A^+b$. If a solution of the original system $Ax = b$ exists and is unique for any vector b then $A^+ = A^{-1}$. If there exists a unique solution of the system $A^*Ax = A^*b$ for any vector b (it means that the operator A^*A is invertible) then $A^+ = (A^*A)^{-1} \cdot A^*$. In a general case, an expression for A^+ has the following form: $A^+ = \lim_{\alpha\to 0+0}(A^* \cdot A + \alpha \cdot E)^{-1} \cdot A^*$.

If instead of a vector b a vector $\tilde{b}$ is given such that: $\|\tilde{b} - b\| \leq \delta$, $\delta \geq 0$, and $x_n = A^+b$, $\tilde{x}_n = A^+\tilde{b}$, then $\|\tilde{x}_n - x_n\| \leq \|A^+\| \cdot \delta$ (it proves stability of a normal pseudosolution for disturbances in a right-hand side). So, the problem of finding a normal pseudosolution is well-posed if a pseudoinverse operator can be calculated exactly. But the problem of a pseudoinverse operator calculation can

be ill-posed. It means that the problem to find $x_n = A^+ b$ may be unstable for errors in A.

Example. Let us consider a system:

$$\begin{cases} x + y = 1, \\ x + y = 1. \end{cases}$$

Obviously, the system has an infinite number of solutions, and $(1/2, 1/2)$ is its normal solution. In this case, $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Let the matrix contain an error:

$$\begin{cases} (1 + \varepsilon)x + y = 1, \\ x + y = 1, \end{cases} \qquad \varepsilon \neq 0,$$

Such approximate system has a unique "approximate" solution $x_\varepsilon = 0$, $y_\varepsilon = 1$, which does not depend on ε. Moreover, as $\varepsilon \to 0$ it does not converge to the exact normal pseudosolution $\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$. Assuming that all four entries of the matrix have errors it is easy to construct examples of convergence of "approximate" solutions to different vectors as errors tend to zero.

Let us change a vector b and consider a system:

$$\begin{cases} x + y = 1/2, \\ x + y = 3/2. \end{cases}$$

It has no solutions. Its normal pseudosolution is the same as in the previous case: $\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$. It is no problem to construct examples of its instability.

Sometimes, it is more convenient to consider a problem of solving an operator equation as a problem of calculating values of an unbounded and not everywhere defined operator A^{-1}: $z = A^{-1}u$. In such a form, it is usually considered a problem of differentiation: $z(x) = \frac{\mathrm{d}u}{\mathrm{d}x}$. Obviously, if consider an operator of differentiation as acting from $C[0,1]$ into $C[0,1]$ then the problem of calculating values of its operator is ill-posed because the first and the third conditions of Hadamard are not fulfilled. If consider the same operator as acting from $C^{(1)}[0,1]$ into $C[0,1]$ then the problem of calculating values of its operator is well-posed. (The space $C^{(1)}[a,b]$ consists of continuously differentiable $[a,b]$ functions. The norm of $z \in C^{(1)}[a,b]$ is defined as $\|z\|_{C^{(1)}[a,b]} = \max_{s\in[a,b]} |z(s)| + \max_{s\in[a,b]} |z'(s)|$.)

2.2 Definition of the regularizing algorithm

Let us give an operator equation:

$$Az = u,$$

where A is an operator acting between normed spaces Z and U. In 1963 A. N. Tikhonov formulated a famous definition of the regularizing algorithm (RA) that is a basic conception in the modern theory of ill-posed problems.

Definition 2.2.1. *Regularizing algorithm (regularizing operator) is called an operator $R(\delta, u_\delta) \equiv R_\delta(u_\delta)$ possessing two properties:*

1. *$R_\delta(u_\delta)$ is defined for any $\delta > 0$, $u_\delta \in U$, and is mapping $(0, +\infty) \times U$ into Z.*
2. *For any $z \in Z$ and for any $u_\delta \in U$ such that $Az = u$, $\|u - u_\delta\| \leq \delta$, $\delta > 0$, $z_\delta = R_\delta(u_\delta) \to z$ as $\delta \to 0$.*

A problem of solving an operator equation is called *regularizable* if there exists at least one regularizing algorithm. Directly from the definition it follows that if there exists one regularizing algorithm then number of them is infinite.

At the present time, all mathematical problems can be divided into the following classes:

1. well-posed problems;
2. ill-posed regularizable problems;
3. ill-posed nonregularizable problems.

All well-posed problems are regularizable as it can be taken $R_\delta(u_\delta) = A^{-1}$. Let us note that knowledge of $\delta > 0$ is not obligatory in this case.

Not all ill-posed problems are regularizable, depending on a choice of spaces Z, U. Russian mathematician L. D. Menikhes constructed an example of an integral operator with a continuous closed kernel acting from $C[0, 1]$ into $L_2[0, 1]$ such that an inverse problem (that is, solving a Fredholm integral equation of the 1st kind) is nonregularizable. It depends on properties of the space $C[0, 1]$. Below it would be shown that if Z is the Hilbert space, and an operator A is bounded and injective, then the problem of solving of the operator equation is regularizable. This result is valid for some Banach spaces, not for all (for reflexive Banach spaces only). Particularly, the space $C[0, 1]$ does not belong to such spaces.

An equivalent definition of the regularizing algorithm is the following. Giving an operator (mapping) $R_\delta(u_\delta)$ defined for any $\delta > 0$, $u_\delta \in U$, and reflecting $(0, +\infty) \times U$ into Z. An accuracy of solving an operator equation in a point $z \in Z$ using an operator $R_\delta(u_\delta)$ under condition that the right-hand side defined with an error $\delta > 0$ is defined as $\Delta(R_\delta, \delta, z) = \sup_{u_\delta \in U: \|u_\delta - u\| \leq \delta,\ Az = u} \|R_\delta u_\delta - z\|$. An operator $R_\delta(u_\delta)$ is called a regularizing algorithm (operator) if for any $z \in Z$ $\Delta(R_\delta, \delta, z) \to 0$ as $\delta \to 0$. This definition is equivalent to the definition above.

Similarly, a definition of the regularizing algorithm can be formulated for a problem of calculating values of an operator (see the end of the previous section), that is for a problem of calculating values of mapping G: $D(G) \to Y$, $D(G) \subseteq X$ under condition that an argument of G is specified with an error (X, Y are metric or normed spaces). Of course, if A is an injective operator then a problem of solving an operator equation can be considered as a problem of calculating values of A^{-1}.

It is very important to get an answer to the following question: is it possible to solve an ill-posed problem (i.e., to construct a regularizing algorithm) without knowledge of an error level δ?

Evidently, if a problem is well-posed then a stable method of its solution can be constructed without knowledge of an error δ. E.g., if an operator equation is under consideration then it can be taken $z_\delta = A^{-1}u_\delta \to z = A^{-1}u$ as $\delta \to 0$. It is impossible if a problem is ill-posed. A. B. Bakushinsky proved the following theorem for a problem of calculating values of an operator. An analogous theorem is valid for a problem of solving operator equations.

Theorem 2.2.2. *If there exists a regularizing algorithm for calculating values of an operator G on a set $D(G) \subseteq X$, and the regularizing algorithm does not depend on δ (explicitly), then an extension of G from $D(G) \subseteq X$ to X exists, and this extension is continuous on $D(G) \subseteq X$.*

So, construction of regularizing algorithms not depending on errors explicitly is feasible only for well-posed on its domains problems.

The next very important property of ill-posed problems is impossibility of error estimation for a solution even if an error of a right-hand side of an operator equation or an error of an argument in a problem of calculating values of an operator is known. This basic result was also obtained by A. B. Bakushinsky for solving operator equations.

Theorem 2.2.3. *Let $\Delta(R_\delta, \delta, z) = \sup_{u_\delta \in U: \|u_\delta - u\| \le \delta,\ Az = u} \|R_\delta u_\delta - z\| \le \varepsilon(\delta) \to 0$ as $\delta \to 0$ for any $z \in D \subseteq Z$. Then a contraction of the inverse operator on the set AD: $A^{-1}|_{AD \subseteq U}$ is continuous on AD.*

So, a uniform on z error estimation of an operator equation on a set $D \subseteq Z$ exists then and only then if the inverse operator is continuous on AD. The theorem is valid also for nonlinear operator equations, in metric spaces at that.

From the definition of the regularizing algorithm it follows immediately if one exists then there is infinite number of them. While solving ill-posed problems, it is impossible to choose a regularizing algorithm that finds an approximate solution with the minimal error. It is impossible also to compare different regularizing algorithms according to errors of approximate solutions. Only including *a priori* information in a statement of the problem can give such a possibility, but in this case a reformulated problem is well-posed in fact. We will consider examples below.

Regularizing algorithms for operator equations in infinite dimensional Banach spaces could not be compared also according to convergence rates of approximate solutions to an exact solution as errors of input data tend to zero, which was obtained by V. A. Vinokurov.

In conclusion, let us formulate a definition of the regularizing algorithm in the case when an operator can also contain an error, i.e., instead of an operator A it is given a bounded linear operator A_h: $Z \to U$ such that $\|A_h - A\| \le h$, $h \ge 0$. Briefly, let us note a pair of errors δ, h as $\eta = (\delta, h)$.

Definition 2.2.4. *Regularizing algorithm (regularizing operator) is called an operator* $R(\eta, u_\delta, A_h) \equiv R_\eta(u_\delta, A_h)$ *possessing two properties:*

1. $R_\eta(u_\delta, A_h)$ *is defined for any* $\delta > 0$, $h \geq 0$, $u_\delta \in U$, $A_h \in L(Z,U)$, *and is mapping* $(0,+\infty) \times [0,+\infty) \times U \times L(Z,U)$ *into* Z;
2. *for any* $z \in Z$, *for any* $u_\delta \in U$ *such that* $Az = u$, $\|u - u_\delta\| \leq \delta$, $\delta > 0$ *and for any* $A_h \in L(Z,U)$ *such that* $\|A_h - A\| \leq h$, $h \geq 0$, $z_\eta = R_\eta(u_\delta, A_h) \to z$ *as* $\eta \to 0$.

Here $L(Z,U)$ *is a space of linear bounded operators acting from* Z *into* U *with the usual operator norm.*

Similarly, it is possible to define whether it is a regularizing algorithm if an operator equation is considered on a set $D \subseteq Z$, i.e., *a priori* information that an exact solution $z \in D \subseteq Z$ is available.

For ill-posed SLAE A. N. Tikhonov was the first to prove impossibility to construct a regularizing algorithm that does not depend explicitly on h.

2.3 Ill-posed problems on compact sets

Let us consider an operator equation:

$$Az = u,$$

A is a linear injective operator acting between normed spaces Z and U. Let $\bar{z}$ be an exact solution of an operator equation, $A\bar{z} = \bar{u}$, u is an exact right-hand side, and it is given an approximate right-hand side such that $\|\bar{u} - u_\delta\| \leq \delta$, $\delta > 0$.

A set $Z_\delta = \{z_\delta : \|Az_\delta - u_\delta\| \leq \delta\}$ is a set of approximate solutions of the operator equation. For linear ill-posed problems $\operatorname{diam} Z_\delta = \sup\{\|z_1 - z_2\| : z_1, z_2 \in Z_\delta\} = \infty$ for any $\delta > 0$ since the inverse operator A^{-1} is not bounded.

The question is that: is it possible to use *a priori* information in order to restrict a set of approximate solutions or (it is better) to reformulate a problem to be well-posed? A. N. Tikhonov proposed the following idea: if it is known that the set of solutions is a compact then a problem of solving an operator equation is well-posed under condition that an approximate right-hand side belongs to the image of the compact. A. N. Tikhonov proved this assertion using as basis the following theorem.

Theorem 2.3.1. *Let an injective continuous operator* A *be mapping:* $D \in Z \to AD \in U$, *where* Z, U *are normed spaces,* D *is a compact. Then the inverse operator* A^{-1} *is continuous on* AD.

The theorem is true for nonlinear operators also. So, a problem of solving an operator equation is well-posed under condition that an approximate right-hand side belongs to AD. This idea made M. M. Lavrentiev possible to introduce a conception of a well-posed problem according to A. N. Tikhonov mathematical

problem (it is supposed that a set of well-posedness exists), and V. K. Ivanov possible to define a quasisolution of an ill-posed problem.

The theorem above is not valid if $u_\delta \notin R(A)$. So, it should be generalized.

Definition 2.3.2. *An element* $z_\delta \in D$ *such that* $z_\delta = \arg\min_{z\in D} \|Az - u_\delta\|$ *is called a quasisolution of an operator equation on a compact* D *(*$z_\delta = \arg\min_{z\in D} \|Az - u_\delta\|$ *means that* $\|Az_\delta - u_\delta\| = \min\{\|Az - u_\delta\| : z \in D\}$*).*

A quasisolution exists but maybe is nonunique. Though, any quasisolution tends to an exact solution: $z_\delta \to \bar{z}$ as $\delta \to 0$. In this case, knowledge of an error δ is not obligatory. If δ is known then:

1. any element $z_\delta \in D$ satisfying an inequality: $\|Az_\delta - u_\delta\| \le \delta$, can be chosen as an approximate solution with the same property of convergence to an exact solution (δ-quasisolution);
2. it is possible to find an error of an approximate solution solving an extreme problem:
   find $\max \|z - z_\delta\|$ maximizing on all $z \in D$ satisfying an inequality: $\|Az - u_\delta\| \le \delta$ (it is obvious that an exact solution satisfies the inequality).

Thus, the problem of quasisolving an operator equation does not differ sharply from a well-posed problem. A condition of uniqueness only maybe is not satisfied.

If an operator A is specified with an error then the definition of a quasisolution can be modified changing an operator A to an operator A_h.

Definition 2.3.3. *An element* $z_\eta \in D$ *such that* $z_\eta = \arg\min_{z\in D} \|A_h z - u_\delta\|$ *is called a quasisolution of an operator equation on a compact* D*.*

Any element $z_\eta \in D$ satisfying an inequality: $\|Az_\eta - u_\delta\| \le \delta + h\|z_\eta\|$ can be chosen as an approximate solution (η-quasisolution).

If Z and U are Hilbert spaces then many numerical methods of finding quasisolutions of linear operator equations are based on convexity and differentiability of the discrepancy functional $\|Az - u_\delta\|^2$. If D is a convex compact then finding a quasisolution is a problem of convex programming. The inequalities mentioned above and defining approximate solutions can be used as stopping rules for minimizing the discrepancy procedures. The problem of calculating errors of an approximate solution is a nonstandard problem of convex programming because it is necessary to maximize (not to minimize) a convex functional.

Some sets of correctness are very well known in applied sciences. First of all, if an exact solution belongs to a family of functions depending on finite number of bounded parameters then the problem of finding parameters can be well-posed. The same problem without such *a priori* information can be ill-posed.

If an unknown function $z(s)$, $s \in [a, b]$, is monotonic and bounded then it is sufficient to define a compact set in the space $L_2[a, b]$. After finite-dimensional approximation the problem of finding a quasisolution is a quadratic programming problem. For numerical solving, known methods such a method of projections of conjugate gradients or a method of conditional gradient can be applied. Similar approach can be used also when the solution is monotonic and bounded, or

monotonic and convex, or has given number of maxima and minima. In these cases, an error of an approximate solution can be calculated.

2.4 Ill-posed problems with sourcewise represented solutions

Let an operator A be linear injective continuous and mapping $Z \to U$; Z, U are normed spaces. Let the following *a priori* information be valid: it is known that an exact solution $\bar{z}$ for an equation $\bar{u} = A\bar{z}$ is represented in the form $B\bar{v} = \bar{z}$, $\bar{v} \in V$; $B : V \to Z$; B is an injective completely continuous operator; V is a Hilbert space. Suppose that an approximate right-hand side u_δ is such that $\|\bar{u} - u_\delta\| \leq \delta$, and its error $\delta > 0$ is known. Such *a priori* information is typical of many physical problems.

V. K. Ivanov and I. N. Dombrovskaya proposed an idea of a method of extending compacts. Let's describe a version of this method below.

Let's preset an iteration number $n = 1$, and define a closed ball in the space V: $\overline{S_n}(0) = \{v : \|v\| \leq n\}$. Its image $Z_n = B\overline{S_n}(0)$ is a compact since B is a completely continuous operator and V is a Hilbert space. After that let us find $\min \|Az - u_\delta\|_{z \in B(\overline{S_n}(0))}$, where u_δ is given approximate right-hand side $\|\bar{u} - u_\delta\| \leq \delta$, $\delta > 0$. Existence of the minimum is guaranteed by compactness of Z_n and continuity of A. If $\min_{z \in B(\overline{S_n}(0))} \|Az - u_\delta\| \leq \delta$, then the iteration process should be stopped, and the number $n(\delta) = n$ defined. An approximate solution of the operator equation can be chosen as any element $z_{n(\delta)} : z_{n(\delta)} \in B(\overline{S_{n(\delta)}}(0))$ satisfying $\|Az_{n(\delta)} - u_\delta\| \leq \delta$. If $\min_{z \in B(\overline{S_n}(0))} \|Az - u_\delta\| > \delta$ then the compact should be extended. For this purpose n changes to $n+1$, and the process repeats.

Theorem 2.4.1. *The process described above converges: $n(\delta) < +\infty$. There exists $\delta_0 > 0$ (generally speaking, depending on $\bar{z}$) such that $n(\delta) = n(\delta_0)$ $\forall \delta \in (0, \delta_0]$. Approximate solutions $z_{n(\delta)}$ strongly converge to the exact solution $\bar{z}$ as $\delta \to 0$.*

It is clear why the method is referred to as "an extending compacts method". It appears that using this method, the so-called an *a posteriori* error estimate can be defined. It means that there exists a function $\chi(u_\delta, \delta)$ such that $\chi(u_\delta, \delta) \to 0$ as $\delta \to 0$, and $\chi(u_\delta, \delta) \geq \|z_{n(\delta)} - \bar{z}\|$ at least for sufficiently small $\delta > 0$. As an *a posteriori* error estimate $\chi(u_\delta, \delta) = \max\{\|z_{n(\delta)} - z\| : z \in Z_{n(\delta)}, \|Az - u_\delta\| \leq \delta\}$ can be taken.

An *a posteriori* error estimate is not an error estimate in a general sense, error estimates cannot be constructed for ill-posed problems. However, for sufficiently small $\delta > 0$ (notably $\forall \delta \in (0, \delta_0]$) an *a posteriori* error estimate is an error estimate for a solution of an ill-posed problem if an *a priori* information about sourcewise representability is available.

This approach was generalized to cases when both operators A and B are specified with errors, also to nonlinear ill-posed problems under condition of sourcewise representation of an exact solution.

Numerical methods for solving linear ill-posed problems under condition of sourcewise representation were constructed, including methods for an *a posteriori* error estimation. To use a sequence of natural numbers as radii of balls in the space V is not obligatory. Any unbounded monotonically increasing sequence of positive numbers can be taken.

2.5 Variational approach for constructing regularizing algorithms

The variational approach for constructing regularizing algorithms firstly was proposed by A. N. Tikhonov. Tikhonov regularizing algorithm based on minimization of smoothing functional (or Tikhonov functional) is described below.

Let us give an operator equation:

$$Az = u,$$

where A is a linear injective bounded operator acting from a Hilbert space Z into a Hilbert space U. Suppose that an exact solution $\bar{z} \in D \subseteq Z$, where D is a closed convex set such that $0 \in D$. The set D is defined by known *a priori* constraints. If constraints lack then $D = Z$. In this section the constraints maybe are not so strong as in previous sections. Application of the quasisolutions method or the method of extending compacts is not possible. Though, it is strongly recommended while solving inverse problems to include in a statement of a problem all given constraints. E.g., what are typical of solutions of many physical problems are the following constraints: non-negativeness, boundedness from the above or/and from the below, etc.

Let the exact right-hand side of the operator equation $A\bar{z} = \bar{u}$ be not known; it is given its approximation u_δ such that $\|\bar{u} - u_\delta\| \leq \delta$, and an error $\delta > 0$. Let the operator A be also specified with an error, that is given a linear bounded operator $A_h : Z \to U$; $\|A_h - A\| \leq h$; and an error $h \geq 0$. For brevity, let us note $\eta = (\delta, h)$. The problem of constructing a regularizing algorithm consists in: using given set of data $\{u_\delta, A_h, \eta\}$ construct an approximate solution $z_\eta \in D$ such that $z_\eta \to \bar{z}$ as $\eta \to 0$, or $z_\eta = R_\eta(u_\delta, A_h) \in D$, $z_\eta \to \bar{z}$ as $\eta \to 0$, where $R_\eta(u_\delta, A_h)$ is a regularizing algorithm.

A. N. Tikhonov proposed the following approach to construct regularizing algorithms. He introduced the functional (Tikhonov functional). The simplest form of this functional is: $M^\alpha[z] = \|A_h z - u_\delta\|^2 + \alpha \cdot \|z\|^2$, $\alpha > 0$ is a regularization parameter. Let us consider an extreme problem: find $\min_{z \in D} M^\alpha[z]$. Under conditions formulated in the beginning of the section there exists a unique minimizer $z_\eta^\alpha = \arg\min_{z \in D} M^\alpha[z]$.

Numerical methods for minimizing the Tikhonov functional when a regularization parameter is fixed are based in general on the two following approaches: 1) If no constraints ($D = Z$) then a necessary and sufficient condition for minimum is the equality of the gradient of the functional to zero. So, it appears the Euler equation for the Tikhonov functional: $(M^{\alpha}[z])' = 0$. After finite-dimensional approximation an SLAE turned, it should be solved numerically. For some special problems different transforms can be applied to simplification of SLAE solving, e.g., convolution type equations (typical of image processing), including multidimensional, fast discrete Fourier transform successfully applied. 2) With constraints or without them direct methods for minimization of the Tikhonov functional can be used (method of conjugate gradients, Newton method, etc.).

If constraints are absent then the problem finding a minimizer of the Tikhonov functional can be reduced to solving the equation $A_h^* A_h z + \alpha z = A_h^* u_\delta$. If operator A is positively definite and self-adjoint, and $Z = U$, a regularized approximation can be found from the equation $Az + \alpha z = u_\delta$ (Lavrentiev method).

For constructing a regularizing algorithm, a choice of a regularization parameter α should be determined. The regularization parameter if the problem is ill-posed should depend explicitly on errors. If it is not so then according to theorems of A. B. Bakushinsky and A. N. Tikhonov only well-posed problems can be solved. Numerous examples show that the most known "error-free methods" "L-curve method" and "generalized cross-validation (GCV) method" (GCV) cannot be applied to the solution of ill-posed problems and fail in solving simplest well-posed problems.

Methods of a choice of the regularization parameter can be conditionally divided into *a priori* and *a posteriori*. The first *a priori* method was proposed by A. N. Tikhonov. Some of its generalization is described below. Let us give a rate of decreasing a regularization parameter: a) $\alpha(\eta) \to 0$ as $\eta \to 0$; b) $\frac{(\delta+h)^2}{\alpha(\eta)} \to 0$ as $\eta \to 0$, that is $\alpha(\eta) \to 0$ slower than $(\delta + h)^2$. Then it can be proved that $z_\eta^{\alpha(\eta)} \to \bar{z}$ as $\eta \to 0$. If operator A is not injective then the regularized solution converges to the exact solution with minimal norm (normal solution).

In practice, applications of *a priori* methods of the regularization parameter choice cause great difficulties because while solving applied problems it is necessary to find an approximate solution when errors are fixed. As an example of an *a posteriori* method let us describe the generalized discrepancy principle (GDP) proposed and substantiated by A. V. Goncharsky, A. S. Leonov, A. G. Yagola. GDP is a generalization of V. A. Morozov discrepancy principle worked out for the case of exactly given operator ($h = 0$).

Definition 2.5.1. *The measure of incompatibility of an operator equation with approximate data on the set $D \subset Z$ is called*

$$\mu_\eta(u_\delta, A_h) = \inf_{z \in D} \|A_h z - u_\delta\|.$$

Obviously, $\mu_\eta = 0$ if $u_\delta \in \overline{A_h D}$.

Lemma 2.5.2. *If* $\|u_\delta - \bar{u}\| \le \delta$, $\bar{u} = A\bar{z}$, $\bar{z} \in D$, $\|A - A_h\| \le h$, *then* $\mu_\eta(u_\delta, A_h) \to 0$ *as* $\eta \to 0$.

If the measure of incompatibility is calculated with an error $\kappa \ge 0$ fitted with errors h then instead of $\mu_\eta(u_\delta, A_h)$ it is given $\mu_\eta^\kappa(u_\delta, A_h)$ satisfying inequalities: $\mu_\eta(u_\delta, A_h) \le \mu_\eta^\kappa(u_\delta, A_h) \le \mu_\eta(u_\delta, A_h) + \kappa$; $\kappa = \kappa(\eta) \to 0$ as $\eta \to 0$.

Definition 2.5.3. *The function of the regularization parameter* $\alpha > 0$*:*

$$\rho_\eta^\kappa(\alpha) = \|A_h z_\eta^\alpha - u_\delta\|^2 - (\delta + h\|z_\eta^\alpha\|)^2 - (\mu_\eta^\kappa(u_\delta, A_h))^2$$

is named the generalized discrepancy.

The following method of the regularization parameter choice is named the *generalized discrepancy principle*. Let the condition $\|u_\delta\|^2 > \delta^2 + (\mu_\eta^\kappa(u_\delta, A_h))^2$ be not true, then an approximate solution is $z_\eta = 0$. If the condition above is true then the generalized discrepancy has a positive root $\alpha^* > 0$, that is $\rho_\eta^\kappa(\alpha^*) = 0$, or $\|A_h z_\eta^{\alpha^*} - u_\delta\|^2 = (\delta + h\|z_\eta^{\alpha^*}\|)^2 + (\mu_\eta^\kappa(u_\delta, A_h))^2$.

In this case, an approximate solution is defined as $z_\eta = z_\eta^{\alpha^*}$ and is unique. It can be proved that $z_\eta \to \bar{z}$ as $\eta \to 0$. If operator A is not injective then it takes place convergence to the solution with minimal norm (normal solution).

For finding a root of the generalized discrepancy its properties can be used:

1. $\rho_\eta^\kappa(\alpha)$ is continuous and monotonically nondecreasing for $\alpha > 0$.
2. $\lim\limits_{\alpha \to +\infty} \rho_\eta^\kappa(\alpha) = \|u_\delta\|^2 - \delta^2 - (\mu_\eta^\kappa(u_\delta, A_h))^2$.
3. $\lim\limits_{\alpha \to 0+0} \rho_\eta^\kappa(\alpha) \le -\delta^2$.

From properties 1)–3) it follows immediately that if $\|u_\delta\|^2 > \delta^2 + (\mu_\eta^\kappa(u_\delta, A_h))^2$ then a root of the generalized discrepancy exists. It can be calculated using well known numerical methods of finding monotonic continuous functions (e.g., a secant method).

Let us try to understand which approximate solution is chosen in accordance with the generalized discrepancy principle, and consider the following extremum problem. It is named the *generalized discrepancy method* (GDM): find $\inf \|z\|$,

$$z \in \left\{ z : z \in D, \|A_h z - u_\delta\|^2 \le (\delta + h\|z\|)^2 + (\mu_\eta^\kappa(u_\delta, A_h))^2 \right\}.$$

Theorem 2.5.4. *Let* A, A_h *be linear bounded operators acting from a Hilbert space* Z *into a Hilbert space* U*;* D *is a closed convex set containing* 0, $D \subseteq Z$*;* $\|A - A_h\| \le h$, $\|u_\delta - \bar{u}\| \le \delta$, $\bar{u} = A\bar{z}$, $\bar{z} \in D$. *Then the generalized discrepancy principle and the generalized discrepancy method are equivalent, that is a solution of an operator equation chosen in accordance with GDP and a solution of an extremum problem GDM coincide.*

If consider a set

$$\left\{ z : z \in D, \|A_h z - u_\delta\|^2 \le (\delta + h\|z\|)^2 + (\mu_\eta^\kappa(u_\delta, A_h))^2 \right\}$$

as a set of approximate solutions of an operator equation with approximate data (an exact solution $\bar{z}$ is an element of the set under conditions of the theorem) then an approximate solution chosen by GDP is an approximate solution with a minimal norm (an element with a normal from the set above). Particularly, if $Z = W_2^1[a,b]$ (a norm in this space is defined as $\|z\| = \left\{\int_a^b z^2(s)\,\mathrm{d}s + \int_a^b (z'(s))^2\,\mathrm{d}s\right\}^{1/2}$, a derivative is included into a definition of a norm), then it is possible to say that the "smoothest" solution is chosen.

If there exists *a priori* information that an exact solution is close to a given element z_0 GDP can be easily adapted to the case of finding an approximate solution with shortest distance (norm) from z_0. For this purpose it is sufficient to change an unknown solution to $z - z_0$ changing a right-hand side respectively.

Numerous versions of GDP exist. E.g., a generalized discrepancy can be considered in the form: $\rho_\eta^\kappa(\alpha) = \|A_h z_\eta^\alpha - u_\delta\|^2 - (\delta + h\|z_\eta^\alpha\|)^2$ without calculation of a measure of incompatibility. In this case, existence of a positive root of a generalized discrepancy cannot be guaranteed. If a root lacks then an approximate solution should be found as $z_\eta = \lim_{\alpha\to 0+0} z_\eta^\alpha$. Then GDP and GDM generally speaking are not equivalent.

GDP cannot be applied directly to solving incompatible problems (finding their pseudosolutions or normal pseudosolutions). Though, it can be modified for this case also. It is necessary to change a generalized discrepancy and use an upper estimate of a measure of incompatibility proposed by A. M. Levin. He developed also numerical methods for calculation of this estimate. This problem can be ill-posed. Let us give one algebraic equation with one unknown variable: $0 \cdot x = 1$. Obviously, a measure of incompatibility is equal to 1. But any small changes of an "operator" (a coefficient before an unknown value) result in that a measure of incompatibility is equal to zero. The Levin upper estimate has a form: $\hat{\mu}_\eta = \inf_{z\in D}(\delta + h\|z\| + \|A_h z - u_\delta\|)$.

Let $\bar{z} \in D$ be an exact pseudosolution of an operator equation $Az = u$ on a set D corresponding to a right-hand side $\bar{u}$, that is $\|A\bar{z} - \bar{u}\| = \bar{\mu}$, where $\bar{\mu} = \inf_{z\in D} \|Az - \bar{u}\|$.

Lemma 2.5.5. *$\hat{\mu}_\eta \geq \bar{\mu}$, $\hat{\mu}_\eta \to \bar{\mu}$ as $\eta \to 0$.*

A generalized discrepancy for solving compatible so as incompatible ill-posed problems have a form: $\hat{\rho}_\eta^\kappa(\alpha) = \|A_h z_\eta^\alpha - u_\delta\|^2 - (\delta + h\|z_\eta^\alpha\| + \hat{\mu}_\eta^\kappa(u_\delta, A_h))^2$. As mentioned above, it is supposed that an upper estimate of a measure of incompatibility is calculated with an error $\kappa = \kappa(\eta) \to 0$ as $\eta \to 0$. GDP is formulated in the following way. Let a condition $\|u_\delta\| > \delta + \hat{\mu}_\eta^\kappa(u_\delta, A_h)$ be not valid, then an approximate solution $z_\eta = 0$. Otherwise, a generalized discrepancy has a positive root $\alpha^* > 0$, that is $\hat{\rho}_\eta^\kappa(\alpha^*) = 0$, or $\|A_h z_\eta^{\alpha^*} - u_\delta\| = \delta + h\|z_\eta^{\alpha^*}\| + \hat{\mu}_\eta^\kappa(u_\delta, A_h)$. An approximate solution $z_\eta = z_\eta^{\alpha^*}$ is unique. It can be proved that $z_\eta \to \bar{z}$ as $\eta \to 0$, where $\bar{z} \in D$ is an exact pseudosolution. If an operator A is not injective then $z_\eta \to \bar{z}$ as $\eta \to 0$, where $\bar{z} \in D$ is an exact normal pseudosolution. It is clear that the method can be applied to compatible problems also.

The Tikhonov variational approach can be generalized in the case when Z is a reflexive Banach space, and U is a normed Banach space. Regretfully, generally speaking it cannot be applied if Z is not a reflexive Banach space.

2.6 Nonlinear ill-posed problems

All results formulated above can be generalized in a case when an operator A is nonlinear. If an operator is given exactly then it can be defined as above without any changes no matter whether it is a quasisolution or an approximate solution on a compact set. A definition of a quasisolution in the case when an approximate operator A_h is given is almost the same. Only it is necessary to define if it is an error of an operator. Usually for description of proximity of an approximate operator A_h to an exact operator A a parameter $h \geq 0$ and a function $\psi(h, z) \geq 0$ should be introduced such that $\|A_h z - Az\| \leq \psi(h, z)$, $\psi(h, z) \to 0$ monotonically on h as $h \to 0$, and under special conditions on the second argument depending on a specific problem. In a linear case usually $\psi(h, z) = h\|z\|$. The main difficulty produces minimization of a discrepancy functional on a compact set of constraints because this functional generally is not convex. Regretfully, there are no general recommendations for how to solve this problem. Every time a special investigation should be conducted.

A method of extending compacts if an operator A is nonlinear is investigated completely similarly to a linear case and can be generalized in a case when both operators are specified with errors.

A. N. Tikhonov applied a variational approach based on minimization of a smoothing functional to nonlinear ill-posed problem also. Though, in this case it is not enough to suppose continuity and injectiveness of an operator A only. It is necessary to demand stronger continuity of an operator A (weakly converging sequences in a space Z are transformed by an operator A into strongly converging sequences in a space U) or to use a scheme with three spaces (a scheme of compact embedding): $V \to Z \to U$. An operator $A : Z \to U$ is continuous, a space V is embedded into a space Z, and an embedding operator B is completely continuous at that. After that, the Tikhonov functional $M^\alpha[v] = \|A_h Bv - u_\delta\|^2 + \alpha \cdot \|v\|^2$ can be considered in order to find $z = Bv$. Such scheme was proposed by A. N. Tikhonov in his first publications on the theory of regularization of ill-posed problems for spaces $V = W_2^1[a, b]$, $Z = C[a, b]$, $U = L_2[c, d]$. Later A. B. Bakushinsky proved that for constructing regularizing algorithms for linear ill-posed problems it is sufficient to use two spaces (a space of solutions is Hilbert or reflexive Banach) only.

A choice of regularization parameter in accordance with *a posteriori* principles not only induces a calculation of its value but also finds a certain extremal function because functions of a regularization parameter (e.g., a generalized discrepancy) can be discontinuous, and extremal functions are not unique. In linear case under conditions formulated above Tikhonov functional has a unique ex-

tremal function. This problems are treated in publications by A. V. Goncharsky, A. S. Leonov, A. G. Yagola.

In nonlinear case it is possible to construct regularizing algorithm such as GDM. Enhancement of conditions on an operator or using a scheme of compact embedding is obligatory. Equivalence of GDP and GDM that is valid in a linear case is not true generally.

2.7 Iterative and other methods

Since this chapter has no enough length it has no sense to describe all possible approaches to construct regularizing algorithms. All of them can be described in frame of a general scheme of point-wise approximation of an inverse operator and matching an approximation parameter with data errors.

In a case when it is known (or can be calculated) as a spectral decomposition of an operator then a spectral "cut-off" can be applied, that is reconciliation between "higher frequencies" and data errors.

For SLAE regularized versions of singular values decomposition were developed using agreement between cutting-off minimal singular values with errors in a matrix and a right-hand side. A method of minimal pseudoinverse matrix is very interesting. In this method proposed by A. S. Leonov the matrix with minimal norm should be chosen from a set of pseudoinverse matrices corresponding to a given approximate matrix of SLAE and its error. After that an approximate normal pseudosolution of SLAE can be calculated.

M. M. Lavrentiev, O. M. Alifanov, A. B. Bakushinsky, H. Engl, G. M. Vainikko, V. V. Vasin and many other authors proposed and developed the so-called iterative methods for solving ill-posed problems. For these methods an iteration number can be a "regularization parameter", and a "stopping rule" should be defined connecting an iteration number with data errors. The simplest iterative method is a simple iteration method. Let spaces Z and U be Hilbert, and $Z = U$, an operator A is self-adjoint, positively definite, completely continuous, $\|A\| < 1$, and an equation $Az = u$ is solvable. Then the equation can be rewritten in a form of $z = z - (Az - u)$ and fixing an initial approximation $z^{(0)}$ the following iterative process (*simple iteration method*) can start: $z^{(k+1)} = z^{(k)} - (Az^{(k)} - u)$. The process converges to a normal solution of an operator equation. If $\|A\| \geq 1$, then a normalizing factor should be introduced beforehand. If an operator equation is rewritten in a form: $Az + \beta z = \beta z + u$, $\beta > 0$, $z = (A + \beta I)^{-1}(\beta z + u)$, where I is a unit operator, then it is possible to organize an iterative process: $z^{(k+1)} = (A+\beta I)^{-1}(\beta z^{(k)}+u)$, that is called an *implicit iterative scheme*. It converges to a normal solution of an operator equation without a condition $\|A\| < 1$. If an operator A is not self-adjoint and positively definite, for organizing iterative processes an equation should be transformed to $A^*Az = A^*u$ preliminarily. If a problem is ill-posed and input data are given with errors then stopping rules should be formulated (e.g., a discrepancy principle or a generalized discrepancy

principle). Both iteration processes described above are from a class of linear iterative processes. As examples of nonlinear iterative processes applied to solving ill-posed problems there are generalizations of methods of steepest descent, minimal discrepancies and others. In accordance with a principle of iterative regularization many classical methods intending in general to minimize a discrepancy functional (Newton method, conjugate gradients method and others) can be transformed to regularizing algorithms using regularizing corrections.

Below is a list of basic monographs and textbooks devoted to different parts of the theory of ill-posed and numerical methods of their solution. Of course, this list is not complete.

Acknowledgements The work was supported by the Russian Foundation for Basic Research (grants 07-01-92103-NSFC, 08-01-00160).

References

1. O. M. Alifanov, E. A. Artioukhine and S. V. Rumyantsev, *Extreme Methods for Solving Ill-Posed Problems with Applications to Inverse Heat Transfer Problems*, Begell House Inc., 1995.
2. A. Bakushinsky and A. Goncharsky, *Ill-posed Problems: Theory and Applications*, Kluwer Academic Publishers, 1994.
3. A. B. Bakushinsky and M. Yu. Kokurin, *Iterative Methods for Approximate Solution of Inverse Problems*, Mathematics and Its Applications, Springer, 2005.
4. A. M. Denisov, *Elements of the Theory of Inverse Problems*, VSP, 1999.
5. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, 1996.
6. C. W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg, 1993.
7. V. K. Ivanov, V. V. Vasin and V. P. Tanana, *Theory of Linear Ill-Posed Problems and Its Applications*, VSP, 2002.
8. B. Kaltenbacher, A. Neubauer and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, De Gruyter, 2008.
9. M. M. Lavrentiev, A. V. Avdeev and Jr. M. M. Lavrentiev, *Inverse Problems of Mathematical Physics*, De Gruyter, 2003.
10. M. M. Lavrentiev and L. Ya. Saveliev, *Operator Theory and Ill-Posed Problems*, De Gruyter, 2006.
11. V. A. Morozov, *Regularization Methods for Ill-Posed Problems*, CRC Press, 1993.
12. A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, Kluwer Academic Publishers, 1995.
13. A. N. Tikhonov, A. S. Leonov and A. G. Yagola, *Nonlinear Ill-Posed Problems*, Chapman and Hall, 1998.
14. V. G. Romanov, *Inverse Problems of Mathematical Physics*, VSP, 1986.
15. V. V. Vasin and A. L. Ageev, *Ill-Posed Problems with A Priori Information*, VSP, 1995.

Chapter 3
Inverse Problems with *A Priori* Information

Vladimir V. Vasin

Abstract. For the last thirty years in the theory of ill–posed problems the direction of investigations was formed that joins with solving the ill–posed problems with *a priori* information. This is the class of problems, for which, together with the basic equation, additional information about the solution to be found is known, and this information is given in the form of some relations and restrictions that contains important data about the object under consideration. Inclusion of this information into algorithm plays the crucial role in increasing the accuracy of solution of the ill–posed (unstable) problem. It is especially important in the case when solution is not unique, since it allows one to select a solution that corresponds to reality. In this work, the review of methods for solving such problems is presented. Though the author touches all approaches known to him in this scope, the main attention is paid to the methodology that is developed by the Author and based on iterative processes of the Fejér type, which give flexible and effective realization for a wide class of *a priori* restrictions. In the final section, description of several applied inverse problems with the *a priori* information and numerical algorithms for their solving are given.

3.1 Introduction

Consider a (non)linear operator equation of the first kind as an abstract model of the ill–posed problem

$$Au = f \tag{3.1.1}$$

Vladimir V. Vasin
Department of the ill–posed problems, Institute of Mathematics and Mechanics UB RAS
Ekaterinburg, 620219, Russia.
e-mail: vasin@imm.ura.ru

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

on a pair if Hilbert spaces U and F with discontinuous and, possibly, multi–valued mapping A^{-1} and the solution set $M \neq \varnothing$. Absence of continuous dependence of solution on the input data does not allow one to approximate reliably solution of equation (3.1.1) on the basis of traditional computational algorithms in the frames of usual concept of approximate solution as one of equation (3.1.1) with approximate data.

In the path-breaking works by M. M. Lavren't'ev [22, 23], V.K. Ivanov [18, 19], and A.N. Tikhonov [35-37], the crucial breakthrough was made in solving this problem. The step made joins with introduction of the regularized family of approximate solutions and the regularizing algorithm. This had opened the way to constructing the regular (stable with respect to disturbances) methods for solving ill–posed problems at known level of inaccuracy of the input data. These explorers had also concluded that any method of solving the ill–posed problem can have whatever slow rate of convergence with respect to the controlling parameters, and only having some *a priori* information on belonging solution to the correctness set, can one obtain some approximate solution with the guaranteed accuracy.

At the first stage of developing the ill–posed problems theory, the main approach to solving the problem of stability was contracting the admissible set of solutions; on such a set, operator A of equation (3.1.1) has the continuous inverse one. For this, A.N. Tikhonov [35] used the topological Hausdorf lemma on homeomorphism of a continuous, one–to–one mapping on a compactum and gave examples of such sets to the inverse problems of geophysics.

But the problem continued to be yet unsolved that joins with the presence of disturbances at the right–hand side of equation (3.1.1): the disturbances can get the solution out of the compactum. Under this situation, it is not possible already to guarantee the stability of the approximate solution.

This problem was solved in works [18, 19, 22, 23]. Namely, M.M. Lavren't'ev in [22, 23] suggested the method of regularization by the shift

$$Au + \alpha u = f_\delta, \quad \|f - f_\delta\| \leqslant \delta \tag{3.1.2}$$

that generates the strong approximation of the solution $u_0 = Bv_0$ belonging to the compactum $M = \{u : u = Bv,\ \|v\| \leqslant r\}$ by the sequence of approximate solutions u^α of equation (3.1.2) under $\alpha(\delta) \to 0$, $\delta/\alpha(\delta) \to 0$, and $\delta \to 0$. Under this, the following estimate of the error holds:

$$\|u_0 - u_\alpha\| \leqslant \omega(\alpha) + \delta/\alpha.$$

Here, ω is the modulus of continuity of the operator A^{-1} on the set $N = AM$, A is the positive operator, B is linear completely continuous operator permutational with A, and u_0 is a solution of equation (3.1.1) for some right–hand part $f \in AM$, but, now, the condition of belonging $f_\delta \in AM$ is not necessary.

Note that, in this approach, one assumes the existence of a solution for equation (3.1.1) on the compactum M. It joins with the fact that there do not exist any effective criteria of solvability of the operator equation (3.1.1) on its given right–hand side.

To overcome this difficulty, V.K. Ivanov in [18, 19] generalized the notion of the solution by means of his introduced the *quasi–expansion* as an element $\widehat{u}$ that realizes the minimum of the residual

$$\min\{\|Au - f\| : u \in Q\} = \|A\widehat{u} - f\| \tag{3.1.3}$$

on the compact set Q. Thus, in the case of insolvability of equation(3.1.1), it is possible to work with its natural generalization. i.e., quasi–expansion. Note that this generalization (for the linear one–to–one operator A) satisfies the classical Hadamard conditions of correctness, and, hence, allows one to apply effective methods of its finding. In other words, the passage to the quasi–solution allows one to solve both the problem of solution existence and the problem of constructing the stable approximate solution.

Accomplishment of creation of methodology for solving the ill–posed problems was given in Tikhonov's works [36] and [37], where the notions of the *regularized family of solutions* and the *regularizing algorithm* (see, (3.1.6)) were formulated and, also, the variational method for regularization of the ill–posed problems was suggested that nowadays carries the Tikhonov name

$$\min\{\|Au - f\|^2 + \alpha\Omega(u) : u \in U\}. \tag{3.1.4}$$

Here, the Lebesgue sets $\Omega_C = \{u : \Omega(u) \leqslant C\}$ of the stabilizing functional Ω are compact in some topology. It means that, in the Tikhonov method, some *a priori* qualitative information about the solution is used, since compactness of the sets Ω_C assumes presence of the additional smoothness of the solution in comparison with the elements (functions) of the space U.

If to compare two variational methods (3.1.3) and (3.1.4), then from the point of view of their numerical realization, application of the Tikhonov method (3.1.4) is preferable, since we deal with the problem of unconstrained minimization that allows one in the case of linear operator and quadratic functional $\Omega(u)$ to reduce the problem to solving a system of linear algebraic equations.

In the method of quasi–solutions we are impelled to solve a problem on conditional extremum that, no doubt, is a more difficult problem. But the essential advantage of this method is the opportunity to apply more detailed information about the solution by means of giving corresponding *a priori* set of restrictions Q. Though in all methods for solving the ill–posed problems various information on the initial data and the solution is used. The term "problems with *a priori* information" [39, 44] or "descriptive regularization" [29] has been introduced to select and describe the situation when additional structural information about the solution is known, for example, the fixed sign property, monotonicity, convexity, presence of the δ–wise forms, fractures, discontinuities, and, also, various bounding relations in the forms of equalities or inequalities.

It is also important to underline that the Tikhonov method (3.1.4) gives more wide choice of the stabilizing functional Ω, in particular, in the form of the Sobolev norm

$$\Omega(u) = \int_D (|u(x)|^2 + |u^{(n)}(x)|^2)dx \tag{3.1.5}$$

that corresponds to regularization of the n-th order. In the pithy review by A.N. Tikhonov and F.P. Vasil'ev [38], many other opportunities for choice of the stabilizer Ω ([38], p. 319) are given, for instance, in the form of the non-smooth Lipschitz norm or

$$\Omega(u) = \|u\|_{L_1} + V(u),$$

where $V(u)$ is the function variation over the segment $[a, b]$. So, in our opinion, it is absolutely unjustified to call the variational method (3.1.4) in the choice of $\Omega(u)$ in the form of L_1–norm and the generalized variation of the function not the Tikhonov method but the "bounded variation penalty method" or "regularization by total bounded control", as it is done by authors of other works [1, 6, 49], which regard (on some reason) that the Tikhonov regularization includes only the case of the quadratic functional (3.1.5).

To complete the introduction regarding the foundations of the ill–posed problems theory, let us discuss another important problem, namely, the problem of regularizability. Existence of at least one regularizing family of linear operators $\{R_\delta\}$ means the opportunity of existence of the stable approximation for solutions of the initial equation (3.1.1), i.e.,

$$\lim_{\delta\to 0} \sup_{\|Au-f\|\leqslant\delta} \|R_\delta f - u\| = 0. \tag{3.1.6}$$

In this case, one says about linear Tikhonov regularizability of problem (3.1.1). Otherwise, the problem is called unregularizable.

For the first sight, it seems that the problems that can not be regularizable are not met in practice and are of rather "exotic" type. But it was discovered that there exist very routine nonregularizable ill–posed problems (see, [27, 48]). For example, in work [27] the following interesting fact was established for the integral Fredholm equation

$$Au \equiv \int_0^1 K(t, s)u(s)ds = f(t) \tag{3.1.7}$$

with the smooth kernel $Ker(t, s)$, where $A : C[0, 1] \to L_2[0, 1]$. If the invertible operator A has continuous extension $\overline{A}$ from $C[0, 1]$ onto $L_2[0, 1]$ with the kernel $Ker(\overline{A})$ that is finite–dimensional (in particular, $Ker(\overline{A}) = \{0\}$), then equation (0.7) is regularizable. But if $\dim Ker(\overline{A}) = \infty$, then, as the examples show, this equation can be nonregularizable.

The main goal of the chapter is to describe the basic approaches to solving the ill–posed problems (under conditions when the solution is not unique) that are formulated in the form of a operator equation with the *a priori* information about belonging solution to be sought to some convex closed set Q. The main contents of the chapter are devoted to methodology of solving such problems on the basis of iterative processes that take into account the *a priori* restrictions by

means of the Fejér mappings. Effectiveness of such algorithms is illustrated by examples of several applied inverse problems.

The chapter has the following structures: 3.1. Introduction; 3.2. Formulation of the problem with *a priori* information; 3.3. The main classes of mappings of the Fejér type and their properties; 3.4. Convergence theorems of the successive approximations method for the pseudo–contractive operators; 3.5. Examples of operators of the Fejér type; 3.6. Fejér processes for nonlinear equations; 3.7. Applied problems with the *a priori* information and methods for solving; 3.8. Conclusions.

3.2 Formulation of the problem with *a priori* information

We shall call *basic* the initial equation (3.1.1). This is a linear one with the bounded operator A acting on the pair of the Hilbert spaces U, F, for which the inverse operator A^{-1} is discontinuous and, in general, multi–valued. The additional *a priori* information about the solution is included into the problem by means of a convex closed set Q, to which the solution to be found belongs.

Thus, we come to the following problem: find a solution of the system

$$Au = f, \quad u \in Q. \tag{3.2.1}$$

For simplicity, we assume that only the right–hand side is known with the error $\|f - f_\delta\| \leqslant \delta$. The general case, when operator A is given approximately, does not contain any principal difficulties for building the methods.

In applied problems, inclusion of the *a priori* information is caused by necessity of obtaining more qualitative solution that describes the searching object (phenomenon). The additional restrictions can describe some important characteristics of the solution that join with the form of the object described, more detailed properties of smoothness, various peculiarities, and the subtle structure following from the physical essence of the problem. Inclusion of the *a priori* restrictions is especially important in investigations of models with non–unique solution of the basic equation. They allow one to select (to approximate) a solution satisfying some concrete physical demands from all others in the set of possible solutions.

One of possible approaches to constructing the stable approximate solutions in the problem with the *a priori* information was widely used on the first stages of investigations of this problem. The approach is based on the idea of the *two–step algorithm* (see, [29, 39]).

At the first step, the initial formulation (3.2.1) with the linear operator A is reduced also to the minimization problem (for the *method of quasi–solutions*, see, [18, 19])

$$\min\{\|Au - f_\delta\|^2 : u \in Q\} \tag{3.2.2}$$

with the compact set Q, or to minimization of the Tikhonov functional with restrictions

$$\min\{\|Au - f_\delta\|^2 + \alpha\Omega(u) : u \in Q\}, \tag{3.2.3}$$

where Q is the convex, closed, but not obligatory compact set.

At the second step, to solve well–posed extremal problems, one can apply any of the traditional methods, such as the method of the conditional reduced gradient, the gradient projections method, the method of linearization, or the algorithms specially oriented on some certain class of *a priori* restrictions.

But these methods can be justified and effectively realized only for a narrow class of the *a priori* restrictions, actually, for the sets that are defined by a system of linear inequalities. Here, it is worthy to mention monograph [39], in which problem (3.2.2) was investigated in detail on the sets of the fixed–sing, monotone, and convex (concave) functions (in various combinations) and economical numerical algorithms were described.

Another approach based on the idea of the iterative regularization [4] consists in substitution of problem (3.2.1) by the problem of solving the following variational inequality:

$$\langle Au - f_\delta, u - v\rangle \leqslant 0 \quad \forall \quad v \in Q. \tag{3.2.4}$$

Here, in general, A is the nonlinear monotone operator.

For solving the variational inequality (3.2.4), the iterative process is used in the form

$$u^{k+1} = P_Q[u^k - \gamma_k(Au^k + \alpha_k u^k - f_\delta)], \tag{3.2.5}$$

for which, with special choice of the sequences $\{\gamma_k\}$, $\{\alpha_k\}$ and the stopping rule, convergency of iterations (3.2.5) exists, i.e., the process generates the regularizing algorithm. Note that presence of the projecting operator P_Q in (3.2.5) essentially hampers the numerical realization of the iterative method, since in each step it is necessary to solve additionally the problem of quadratic or convex programming if Q is given by a system of linear or convex inequalities.

In the next section we consider the approach to solving the problems with the *a priori* information that was suggested by the Author in works [40,42] and described in monographs [44, 46]. This approach is based on using the techniques of the Fejér mappings [9–11] to take account of the *a priori* restrictions in the form of convex inequalities. The mappings possessing the Fejér property allow one to construct the iterative processes by taking into account the *a priori* restrictions of rather general form and, in contrast to the metrical projection, adopt effective realization.

The main idea is that the step operator in such iterative methods is constructed in the form of superposition of the step operator of some classical iterative scheme for solving equation (3.1.1) and the Fejér mapping responsible for the *a priori* restrictions. In each step, this mapping implements the shift to direction of the set Q, and, as a result, convergency of iterations to the solution of system (3.2.1) is

provided, i.e., to the solution of equation (3.1.1) satisfying the *a priori* restrictions in the form $x \in Q$.

3.3 The main classes of mappings of the Fejér type and their properties

Before the passage to definition of classes of the nonexpansive mappings, note that the different terminology for definition of the same classes of nonlinear mappings has been formed in the functional analysis [25, 26, 28, 44] and in the mathematical programming [9–11]. Thus, in the definitions suggested below for denotation of various types of quasi–contractivity of operators, we give their names from both mentioned scopes, but in the text, we prefer one or other terminology dependent on the subject under discussion.

Take the denotations $Fix(T) = \{u : u \in U,\ u = T(u)\}$ for the set of fixed points of the operator (mapping) T; we use the writing $T : U \to U$ even in the case when T is defined only on $D(T) \subseteq U$.

Definition 3.3.1. *The mapping $T : U \to U$ is called M–quasi–nonespansive or weak M–Fejér if $M = Fix(T) \neq \varnothing$ and*

$$\|T(u) - z\| \leqslant \|u - z\| \quad \forall u \in X, \quad \forall z \in M;$$

denote this class by $\mathcal{K}_M$.

Definition 3.3.2. *The mapping $T : U \to U$ is called strictly M–quasi–nonexpansive or M–Fejér if $M = Fix(T) \neq \varnothing$ and $\|T(u) - z\| < \|u - z\|$ for any $z \in M$ and $u \in U$, $u \notin M$; denote this class by $\mathcal{F}_M$.*

Definitions 3.3.1 and 3.3.2 have the sense for any arbitrary normed space. In the sequel, if the contrary is not declared, we assume U to be the Hilbert space.

Definition 3.3.3. *The mapping $T : U \to U$ is called M–pseudo–contractive or strongly M–Fejér if $M = Fix(T) \neq \varnothing$ and there exists a constant $\nu > 0$ such that*

$$\|T(u) - z\|^2 \leqslant \|u - z\|^2 - \nu\|u - T(x)\|^2$$

for any $u \in H$ and $z \in M$; we denote this class by $\mathcal{P}_M^\nu$.

Directly from this definition, the inclusions follow $\mathcal{P}_M^\nu \subset \mathcal{F}_M \subset \mathcal{K}_M$ and, as the examples show, they are strict.

In the following lemma, we establish relations between the classes defined above useful in applications [42, 44, 46].

Lemma 3.3.4. *Let $T : U \to U$, $T \in \mathcal{P}_M^\nu$. Then $T = \frac{1}{1+\nu}V + \frac{\nu}{1+\nu}I$, where $V \in \mathcal{K}_M$. Inversely, if $V \in \mathcal{K}_M$, then under $\lambda \in (0, 1)$ the mapping $T = \lambda V + (1 - \lambda)I$ belongs to the class $\mathcal{P}_M^\nu$ under $\nu = (1 - \lambda)/\lambda$.*

Definition 3.3.5. *The mapping $T : U \to U$ is called nonexpansive (nonexpanding) if*

$$\|T(u) - T(v)\| \leqslant \|u - v\| \quad \forall v \in X;$$

we denote this class by $\mathcal{K}$.

Definition 3.3.6. *The operator $T : U \to U$ is called pseudo–contractive if there exists a constant $\nu > 0$ such that the inequality holds*

$$\|T(u) - T(v)\|^2 \leqslant \|u - v\|^2 - \nu\|u - T(u) - (v - T(v))\|^2$$

for any $u, v \in H$; we denote this class by $\mathcal{P}^\nu$.

Evidently, if $M = Fix(T) \neq \varnothing$, then $\mathcal{K}_M \subset \mathcal{K}$, $\mathcal{P}^\nu_M \subset \mathcal{P}^\nu$; moreover, Lemma 3.3.4 is valid by changing the class $\mathcal{P}^\nu_M$ by $\mathcal{P}^\nu$ and $\mathcal{K}_M$ by $\mathcal{K}$, i.e., under $T \in \mathcal{P}^\nu$ the representation holds

$$T = \frac{1}{1+\nu}V + \frac{\nu}{1+\nu}I,$$

where $V \in \mathcal{K}$, but if $V \in \mathcal{K}$, then

$$T = \lambda V + (1-\lambda)I \subset \mathcal{P}^\nu, \quad \nu = (1-\lambda)/\lambda.$$

The latter inequality is widely used in iterative processes. It is stipulated by the fact that $Fix(T) = Fix(V)$ and the operator $T \in \mathcal{P}^\nu_M \subset \mathcal{P}^\nu$ having more strong condition of contractivity generates the converging process (see, Theorem 3.4.1, below)

$$u^{k+1} = T(u^k) \equiv \lambda V(u^k) + (1-\lambda)u^k, \quad T \in \mathcal{P}^{(1-\lambda)/\lambda}, \tag{3.3.1}$$

to some element $\widehat{u} \in Fix(V) = Fix(T)$, but the iterative process

$$u^{k+1} = V(u^k) \tag{3.3.2}$$

with the step operator $V \in \mathcal{K}_M$ is not obligatory convergent.

The class $\mathcal{P}^1$, i.e., $\mathcal{P}^\nu$ under $\nu = 1$ was introduced in work [25], and there, the theorem on the weak convergence was formulated. Investigations of properties of mappings from the more wide class $\mathcal{P}^\nu$ were continued in the Author's works [41, 42].

Note that in Definitions 3.3.1–3.3.3 for classes of nonlinear mappings we use the double terminology, which originates, on one hand, from the works in the scope of numerical functional analysis [25, 26, 28, 44] and, on the other hand, from those in the scope of mathematical programming where the "Fejér mapping" is the widely used term.

Before the passage to investigation of properties of operators from the classes defined above, stop for a little while on the history of appearance of the term the ""Fejér mapping". This appears in works by I.I. Eremin [9, 11, 46] in honor of the widely known Hungarian mathematician L. Fejér (1880–1959). The foundation

for this term was the Fejér work [12] where the following notions and definitions were introduced.

Definition 3.3.7. *Let E_n be the Euclidian space, M be some subset, and p, p_1 be two points from E_n. If*

$$\|p - a\| > \|p_1 - a\| \quad \forall a \in M,$$

then we say that p_1 is closer to M than p. If the point p is such that there do not exist points p_1, which are closer to the set M than p, then the point p is called closest to M.

Having introduced this notion, L. Fejér characterized the set of the points closest to M (this coincides with the closure of convex hull of the set M). From this fact, it follows that if the point p does not belong to $\overline{\mathrm{conv}(M)}$, then it is possible to find the point p_1 that will be closer to the set M than the point p. Hence, if M is the convex set of solutions of some problem, then the principle opportunity appears for constructing an (iterative) algorithm for its solving.

Further, in work [30] authors (with reference to [12]) had given the definition of the monotone Fejér sequence as one satisfying the conditions

$$\|q_i - a\| \geqslant \|q_{i+1} - a\| \quad \forall a \in M, \tag{3.3.3}$$

where $q_i \neq q_{i+1}$, and suggested the relaxation method for solving a system of linear inequalities.

At last, as mentioned above, in works [9–11] the following were introduced the notions of the Fejér mapping (see, Definition 3.3.2), the Fejér method, and the Fejér sequence $\{q_i\}$ as one, for which in (3.3.3) the strict inequality holds. For the author of this chapter it seems more natural that the Fejér property of some operator (or a sequence) is defined by the strict inequality.

From the point of view of convergency of iterative processes, the operators from the classes $\mathcal{P}_M$ and $\mathcal{F}_M$ are of the most interest; for these operators we shall use the general term "Fejér operator (mapping)" or "operator of the Fejér type" and for the corresponding processes with the step operators of the Fejér type we shall use the term "Fejér processes (methods)".

These classes possess a remarkable property, namely, the property of closedness with respect to the operations of superposition and convex summation [42, 44, 46].

Theorem 3.3.8. *Let $T_i : H \to H$, H be the Hilbert space, and $T_i \in \mathcal{P}_{M_i}^{\nu_i}$, $M = \bigcap\limits_{i=1}^{m} M_i \neq \varnothing$. Then*

1) $T = T_m T_{m-1} \dots T_1 \in \mathcal{P}_M^{\nu}$, *where* $\nu = \min\limits_{1 \leqslant i \leqslant m} \{\nu_i\}/2^{m-1}$;

2) $T = \sum\limits_{i=1}^{m} \alpha_i T_i \in \mathcal{P}_M^{\nu}$, *where* $\alpha_i > 0$, $\sum\limits_{i=1}^{m} \alpha_i = 1$, $\nu = \min\{\nu_i\}$.

Theorem 3.3.9. *Let $T_i : X \to X$, $T_i \in \mathcal{F}_{M_i}$, $M = \bigcap_{i=1}^{m} M_i \neq \varnothing$. Then each of the mappings T defined by formulas*

$$T = T_m T_{m-1} \dots T_1, \quad T = \sum_{i=1}^{m} \alpha_i T_i, \quad \left(\alpha_i > 0, \ \sum_{i=1}^{m} \alpha_i = 1\right)$$

belongs to the class of the $\mathcal{F}_M$ M–Fejér mappings.

Remark 3.3.10. *The statements of these theorems are valid under any order of the indices in the superposition $T = T_{i_1} T_{i_2} \dots T_{i_m}$.*

Corollary 3.3.11. *If $T_i \in \mathcal{P}_M^{\nu_i}$ $(T_i \in \mathcal{F}_M)$, then the operators of the form*

$$T = \sum_{i=1}^{m} \alpha_i T_{i_1}^{n_{i_1}} T_{i_2}^{n_{i_2}} \dots T_{i_m}^{n_{i_m}}$$

also belong to the class $\mathcal{P}_M^{\nu}$ (correspondingly, to $\mathcal{F}_M$); here, $\alpha_i > 0$, $\sum_{i=1}^{m} \alpha_i = 1$, n_{i_k} are the integers, and $(i_1, i_2, \dots i_m)$ is an arbitrary transposition of the indices $(1, 2, \dots, m)$.

Remark 3.3.12. *In solving a problem, including an ill–posed one, by the method of successive approximations*

$$u^{k+1} = T(u^k), \tag{3.3.4}$$

$$u^{k+1} = \lambda T(u^k) + (1 - \lambda) u^k \tag{3.3.5}$$

the initial formulation is preliminary reduced to the equivalent problem of finding a fixed point of some mapping T, i.e., to solving the equation

$$u = T(u). \tag{3.3.6}$$

For example, solution of the linear equation

$$Au = f$$

is equivalent to finding the fixed point of the operator

$$T(u) = u - \beta[A^* A u - A^* f], \quad \beta > 0$$

or

$$T(u) = (A^* A + \alpha I)^{-1}(\alpha u + A^* f), \quad \alpha > 0,$$

which are the step operators for the method of simple iteration and the iterated version of the Tikhonov method, respectively.

The problem of minimization of a convex differentiable functional on the convex closed set Q

$$\min\{f(u) : u \in Q\}$$

is reduced to solving the equation

$$u = P_Q[u - \beta\nabla f(u)] \equiv T(u), \quad \beta > 0,$$

i.e., to find a fixed point of the operator T. Particularly, problem (3.1.3) (see, Introduction) of finding a quasi–solution is equivalent to problem of finding the fixed point of the operator T defined by the formula

$$u = P_Q[u - \beta(A^*Au - A^*u)] \equiv T(u),$$

where under $0 < \beta, 2/\|A\|^2$, T is a nonexpansive operator from the class $\mathcal{P}_M^v$.

The problem of finding the saddle point of the convex–concave function $L(u, v)$, $L : U \times U \to R^1$,

$$\min_u \max_v L(u, v)$$

is equivalent to finding the fixed point of the pseudo–contracting mapping [32]

$$\Lambda : (z, w) \to \arg\min_u \max_v \left\{ L(u, v) + 0.5\|u - z\|^2 - 0.5\|v - w\|^2 \right\}. \qquad (3.3.7)$$

The statements presented above open wide opportunity, on one hand, for constructing many new step operators in the method of successive approximations to approximate $u \in Fix(T)$, i.e., for solving equation (3.3.6) (Corollary 3.3.11). On the other hand, from Theorems 3.3.8 and 3.3.9, it follows that if problem (3.3.6) can be presented in the form of the system

$$u = T_i(u), \quad i = 1, 2, \ldots, m,$$

where $T_i \in \mathcal{P}_{M_i}^{\nu_i}$ (or $T_i \in \mathcal{F}_{M_i}$), $\bigcap_{i=1}^{m} M_i = M = Fix(T)$, then the constructions of operators in the form of the superposition T_i and their convex combination containing in these theorems give a wide opportunity for building the algorithms of parallelizing the solving operator T in the method of successive approximations (concrete schemes can be found in [46]).

As a simple but pithy example, consider the system of linear algebraic equations

$$\langle a_i, x\rangle - q_i = 0, \quad i = 1, 2, \ldots, m.$$

Define $T_i(u) = P_{M_i}(u)$, where P_{M_i} is the operator of the metric projection into the hyperplane $L_i = \{u : \langle a_i, x\rangle - q_i = 0\}$, and the operator is described by the explicit formula

$$P_{M_i}(u) = u - (\langle a_i, x\rangle - q_i)a_i/\|a_i\|^2.$$

Then the iterative processes

$$u^{k+1} = T_1(T_2 \dots (T_m(u^k))), \quad u^{k+1} = \sum_{i=1}^{m} \lambda_i T_i(u^k),$$

$$0 < \lambda_i < 1, \quad \sum_{i=1}^{m} \lambda_i = 1$$

give the Kaczmarz algorithms for solving the system of linear algebraic equations [21].

Actually, the same processes are appropriate for solving linear inequalities (see, Section 3.5), since the operator of projecting for them is computed by the same formulas.

3.4 Convergence theorems of the method of successive approximations for the pseudo–contractive operators

In theorems presented below, it is established that for the operators T from the class $\mathcal{P}_M^\nu$, in the general case, it is possible to guarantee only the weak convergence of the successive approximations method (3.3.4) (the symbol "$\rightharpoonup$" is taken for denotation of the weak convergence).

Theorem 3.4.1. *Let the operator $T : U \to U$, $T \in \mathcal{P}_M^\nu$ $(T \in \mathcal{K}_M)$ and satisfy the condition*

$$u^k \rightharpoonup u, \quad T(u^k) - u^k \to 0 \Rightarrow u \in Fix(T). \tag{3.4.1}$$

Then for iterations (3.3.4), (and (3.3.5)) *the following properties are valid:*

1) $u^k \rightharpoonup \widehat{u} \in Fix(T)$;

2) $\inf\limits_z \{ \lim\limits_{k\to\infty} \|u^k - z\| : z \in Fix(T)\} = \lim\limits_{k\to\infty} \|u^k - \widehat{u}\|$;

3) *either $\|u^{k+1} - \widehat{u}\| < \|u^k - \widehat{u}\|$ for any k, or $\{u^k\}$ is stationary beginning from some number k_0;*

4) *the estimate is valid*

$$\sum_{k=0}^{\infty} \|u^{k+1} - u^k\|^2 \leqslant \|u^0 - z\|^2/\nu \quad \forall z \in Fix(T).$$

The proofs of various versions of this theorem can be found in [25, 42, 44].

Corollary 3.4.2. *The theorem statement is valid if instead of $T : U \to U$ the operator $T : D \to D$ is given, where D is a convex closet subset of the space U.*

Corollary 3.4.3. *Let $P_Q^\lambda = I - \lambda(I - P_Q)$, where P_Q is the metric projection onto the convex closed subset Q of the Hilbert space, $0 < \lambda < 2$. Then $Fix(P_Q^\lambda) = Q$, $P_Q^\lambda \in \mathcal{P}_Q^{(2-\lambda)/\lambda}$, condition* (3.1) *is satisfied, and, hence, for $T = P_Q^\lambda$ the conclusion of Theorem 3.4.1 is valid; this guarantees the weak convergence of the iterations. Moreover, if Q is a compact set, then the strong convergence holds.*

Note that the inclusion $P_Q^{\lambda} \in \mathcal{P}_Q^{(2-\lambda)/\lambda}$ was established in [46, Chapter 1, Lemma 3.2]. Condition (3.4.1) for P_Q^{λ} follows from the properties of the projection P_Q, for which, in turn, this property follows from the known fact for the nonexpansive mappings (see, for instance, [44, Chapter 1, Lemma 2.1]).

Corollary 3.4.4. *Let the domain $D(T)$ of definition of the operator T be convex and closed, then Theorem 3.4.1 is valid if instead of condition (3.4.1) the weak closedness of the operator $T : x_k \rightharpoonup x$, $x_k \in D(T)$, $T(x_k) \rightharpoonup y \Rightarrow x \in D(T)$, $T(x) = y$ holds.*

Remark 3.4.5. *Under conditions of Theorem 3.1, in the general case, it is impossible to prove the strong convergence of the iterations. This follows from work* [13]*, where the corresponding example of the convex closed bounded set $D \subset l_2$ and the mapping $T : D \to D$ from the class $\mathcal{P}_M^{\nu}$ was constructed, for which the sequence converges weakly but not strongly.*

Assume that the iterations in process (3.3.4) are calculated with an error, i.e.,

$$z^{k+1} = T(z^k) + \xi_k, \quad \|\xi_k\| \leqslant \varepsilon_k.$$

Theorem 3.4.6. *Let conditions of Theorem 3.4.1 be satisfied for the operator T and $\sum\limits_{k=0}^{\infty} \varepsilon_k < \infty$. Then for the sequence $\{z^k\}$, the following properties are valid:* 1) $z^k \rightharpoonup \widehat{z} \in M = Fix(T)$; 2) $\lim\limits_{k\to\infty} \|z^{k+1} - z^k\| = 0$.

The proof of this statement for various conditions is in [32, 40, 41].

Definition 3.4.7. *The operator $T : U \to F$, in the general case multi–valued, is called closed if its graph $\Gamma = \{(u, T(u)) : u \in U\}$ is closed in $U \times F$.*

Definition 3.4.8. *The operator $T : U \to F$ is called completely continuous if from $u^k \rightharpoonup u$ it follows $\lim\limits_{k\to\infty} \|u^k - \widehat{u}\| = 0$.*

Theorem 3.4.9. *Let $T : U \to U$, where U is the Hilbert space, T is the M–Fejér operator, i.e., $T \in \mathcal{F}_M$ (Definition 3.3.2) and one of the following conditions is satisfied*:

1) *one–to–one operator T is strongly continuous*;

2) *U is finite–dimensional, T is closed, possibly, multi–valued operator.*

Then for $\{u^k\}$ in the successive approximations method defined in the multi–valued case by the relation $u^k \in T(u^k)$, the convergence holds: $\lim\limits_{k\to\infty} \|u^k - \widehat{u}\| = 0$, *where $\widehat{u} \in M$*

Remark 3.4.10. *Investigations joined with the Fejér operators of the class $\mathcal{F}_M$, the iterative processes of the Fejér type, and their applications to a wide circle of problems of mathematical programming were implemented in works* [9 – 11] (*see, also* [46]).

Remark 3.4.11. *In work* [26] (*and, also* [46]) *the sufficient conditions are investigated for systems of nonlinear equations, under which assumptions of Theorem 3.4.9 for the step operators T of the gradient method and the Newton method are satisfied.*

Consider now the main problem formulated in Section 3.2, namely, solving the problem with the *a priori* information (3.2.1) that can be reformulated in the following way.

Problem 3.4.12. *Find the element*

$$u \in M \bigcap Q \neq \varnothing,$$

where M is the set of solutions of the basic equation $Au = f$, Q is the set of a priori restrictions. This problem is reduced to finding a fixed point of some mapping T with $Fix(T) = M \bigcap Q$.

To construct the approximative sequence, consider the iterative processes from [40, 41, 42]

$$u^{k+1} = P(V(u^k)) \equiv T(u^k), \quad u^{k+1} = \lambda P(u^k) + (1-\lambda)V(u^k) \equiv T(u^k), \quad (3.4.2)$$

where $0 < \lambda < 1$, $P, V : U \to U$.

Theorem 3.4.13. *Let $P \in \mathcal{P}_Q^\nu$, $V \in \mathcal{P}_M^\mu$ and these mappings satisfy condition (3.4.1). Then*

a) *the iterations u^k defined by processes (3.4.2) satisfy properties* 1)–4) *from Theorem 3.4.1;*

b) *if iterations (3.4.2) are calculated with the error ξ_k, where $\|\xi_k\| \leqslant \varepsilon_k$, $\varepsilon_k > 0$, $\sum\limits_{k=0}^{\infty} \varepsilon_k < \infty$, then the conclusion of Theorem 3.4.6 is valid.*

As mentioned above (Remark 3.4.5), under conditions of Theorem 3.4.1 (and, hence, of Theorem 3.4.13, point a)), in the general case it is impossible to guarantee convergence by the norm of the space H. But under some additional assumptions it is possible to obtain the strong convergence.

Corollary 3.4.14. *If Q is a compactum, $P = P_Q$ is the metric projection onto the set Q or $P = P_Q^\lambda = I - \lambda(I - P_Q)$ is the projection with the relaxation $(0 < \lambda < 2)$, then under the exact computation of processes (3.4.2) (Theorem 3.4.13, point a)) the strong convergence of the iterations u^k to some $\widehat{u} \in M \bigcap Q$ holds.*

Give now some examples of compact sets.

In the spaces of functions of one variable

$$Q_1 = \{u(x) : u \in W_p^1[a,b],\ \|u\|_{W_p^1} \leqslant r\}, \quad 1 < p < \infty$$

is compact in the spaces $L_p[a,b]$ and $C[a,b]$;

$$Q_2 = \{u(x) : u \in \bigvee_a^b,\ \|u\|_{\bigvee_a^b} \leqslant r\}$$

is compact in the spaces $L_p[a,b]$, $1 < p < \infty$;

$$Q_3 = \{u(x) : |u(x)| \leqslant C,\ \ u \text{ is monotone}\}$$

is compact in the spaces L_p.

In the spaces of functions of n variables

$$Q_4 = \left\{u(x) : u \in C[\Pi],\ \max|u(x)| + \sup_{x_1,x_2\in\Pi} \frac{|u(x_1)-u(x_2)|}{\|x_1-x_2\|^{\mu}_{R^n}} \leqslant r\right\}$$

is compact in the spaces of continuous functions $C[\Pi]$ and, hence, in $L_p[\Pi]$; here, Π is a closed bounded subset in R^n,

$$Q_5 = \{u(x) : \|u\|_{L_1(D)} + J(u) \leqslant r\},$$

where J is the generalized variation (see, [15]),

$$J(u) = \sup\left\{\int_D u(x)\ \mathrm{Div} v(x)dx : v \in C_0^1(D, R^n),\ |v(x)| \leqslant 1\right\}$$

is relatively compact in the space $L_p(D)$ under $p < n(n-1)$.

In each of examples considered under $p = 2$ we obtain the Hilbert space and corresponding (pre)compact sets Q_i in it.

The strong convergence of iterations can be achieved in some cases by modification of processes (3.4.2) with the correcting multipliers [17]

$$u^{k+1} = \gamma_k T(u^k) + (1-\gamma_k)v_0, \tag{3.4.3}$$

where $0 < \gamma_{k-1} < \gamma_k < 1$, $\gamma_k \to 1$ and satisfy some additional conditions [17] (see, also [44]).

Theorem 3.4.15. *Let the step operator T in processes (3.4.3) be nonexpansive and map the convex closed set D into itself, $Fix(T) = M \neq \varnothing$, and γ_k be an admissible sequence [17], $v_0 \in D$.*

Then the iterative sequence u_k converges strongly to the element $\widehat{u} \in M \bigcap Q$ with the minimal deviation from v_0.

Remark 3.4.16. *Process (3.4.3) has the regularizing property, namely, if $\widetilde{T} \in \mathcal{K}$ is a disturbed operator satisfying the condition of approximation*

$$\|\widetilde{T}(x) - T(x)\| \leqslant \varphi(\delta) \quad (\varphi(\delta) \to 0 \quad \textit{under} \quad \delta \to 0)$$

on some ball S_r, then the process with corrupted data

$$\widetilde{u}^{k+1} = \gamma_k \widetilde{T}(\widetilde{u}^k) + (1-\gamma_k)v_0$$

converges to the same solution $\widehat{u} \in M \bigcap Q$ *with choice of the number* $k(\delta)$ *of iterations accordingly to the rule* $\varphi(\delta) \cdot k(\delta) \to 0$ *for* $\delta \to 0$.

Under conditions of Theorem 3.4.15, the sequence of correcting multipliers γ_k is chosen to be *a priori* and some conditions onto the asymptotic of behavior are necessary. Actually, it is possible to suggest *a posteriori* choice of γ_k [46].

Introduce denoting $T_{\gamma_k}(u) = \gamma_k T(u) + (1-\gamma_k)v_0$, $u_{\gamma_k} = \gamma_k T(u_{\gamma_k}) + (1-\gamma_k)v_0$, i.e., u_{γ_k} is the fixed point of the operator T_{γ_k}. Since $0 < \gamma_k < 1$ and $\|T_{\gamma_k}(u) - T_{\gamma_k}(v)\| \leqslant \gamma_k \|u - v\|$, and under fixed γ_k and the initial approximation u^0, the iterative process

$$u_0^n = T_{\gamma_k}^n(u^0), \quad n = 0, 1, \ldots$$

converges to u_{γ_k}. Then by virtue of the inequality $\|T_{\gamma_k}(u) - u_{\gamma_k}\| \leqslant \gamma_k \|u - u_{\gamma_k}\|$ with sufficiently large number n for $d_\gamma(u) = \|u - T_\gamma(u)\|$, the following inequality is provided:

$$d_{\gamma_k}(u_0^n) \leqslant \delta.$$

Lemma 3.4.17. *Let* $d_\gamma(u) \leqslant \delta$ *and* $\gamma, \delta, \varepsilon$ *be obeyed to the relation* $\delta/(1-\gamma) \leqslant \varepsilon$. *Then*

$$\|u - u_\gamma\| \leqslant \varepsilon.$$

The considerations given above allow one to construct the following computational process for approximation of the fixed point of the operator T:

1) give the sequence of positive parameters ε_k, γ_k, and δ_k, satisfying the relation $\delta_k/(1-\gamma_k) \leqslant \delta_k$;

2) give some initial approximation u^0 and compute the element $u_0^{k_1} = T_{\gamma_1}^{k_1}(u^0)$, for which the inequality holds

$$d_{\gamma_1}(u_0^{k_1}) \leqslant \delta_1,$$

and put $u^1 = u_0^{k_1}$;

3) let u^{n-1} be built, then we take u^{n-1} as the initial approximation and calculate the element $u_{n-1}^{k_n} = T_{\gamma_n}^{k_n}(u^{n-1})$, for which the condition is valid

$$d_{\gamma_n}(u_{n-1}^{k_n}) \leqslant \delta_n,$$

and put $u^n = u_{n-1}^{k_n}$.

3.5 Examples of operators of the Fejér type

In the iterative processes (3.4.2), the step operator T is the superposition of two mappings V and P, where V is the step operator of some iterative method for solving equation (3.1.1) with $Fix(U) = M$ and M is the set of solutions of the basic equation $Au = f$. The fact that the operator V belongs to the classes $\mathcal{P}_M$, $\mathcal{F}_M$, or $\mathcal{K}$ allows one to apply the convergence Theorem 3.4.13 or 3.4.15.

Examples of such classical iterative methods with the step operator from the class $\mathcal{P}_M^1$ are given below:

$$V(u) = u - \beta(A^*Au - A^*f), \quad \beta \leqslant 1/\|A\|^2, \tag{3.5.1}$$

this one corresponds to the step operator in the method of simple iteration;

$$V(u) = (A^*A + \alpha I)^{-1}(\alpha u + A^*f), \quad \alpha > 0, \tag{3.5.2}$$

this one defines the iterated version of the Tikhonov method;

$$V(u) = u - \frac{\|S(u))\|^2}{\|AS(u)\|^2}S(u), \quad S(u) = A^*(Au - f), \tag{3.5.3}$$

$V(u) = u$, if $S(u) = 0$, this one gives the method of the steepest descent;

$$V(u) = u - \frac{\|Au - f)\|^2}{\|S(u)\|^2}S(u), \tag{3.5.4}$$

where $V(u) = u$, if $Au - f = 0$, is the operator of the nonlinear method of the minimal error.

If operator A is self–adjoint and positively defined, then

$$V(u) = u - \frac{\langle A^\alpha \Delta, \Delta\rangle}{\langle A^{\alpha+1}\Delta, \Delta\rangle}\Delta, \quad \Delta = Au - f \neq 0, \tag{3.5.5}$$

where $V(u) = u$ with $\Delta = 0$ generates α–processes, which, in particular, includes the methods of the minimal residuals, of the steepest descent, and of the minimal error.

Note that the operators defined by formulas (3.5.1) and (3.5.2) are not only M–Fejér, but also nonexpansive. So, under the condition that the mapping P is also from the class $\mathcal{K}$, modification of methods (3.4.2) is possible with the help of the correcting multipliers (see, (3.4.3)), and such a modification can allow one to obtain the strong convergence (Theorem 3.4.15).

As noted in Corollary 3.4.3, the metric projection or the projection with relaxation could play the role of operator P in processes (3.4.2). This operator is defined by an explicit formula and is easy computed only in some particular cases (of *a priori* restrictions) that often are met in applications

$$\begin{aligned} Q_1 &= \{u : u \in L_2,\ u \geqslant 0\}, & Q_2 &= \{u : u \in U,\ \|u\| \leqslant r\}, \\ Q_3 &= \{u : u \in R^n,\ a \leqslant u \leqslant b\}, & Q_4 &= \{u : u \in U,\ \langle u, v\rangle - q = 0\}, \\ Q_5 &= \{u : u \in U,\ \langle u, v\rangle - q \leqslant 0\}. \end{aligned}$$

Here, U is the Hilbert space, $\langle\ ,\ \rangle$ denotes the inner product.

For example, the metric projection for the set Q_5 is defined by the formula

$$P_{Q_5}(u) = u - \frac{(\langle u, v\rangle - q)^+ v}{\|v\|^2},$$

and, under this, if the *a priori* set is given by the system of linear inequalities

$$Q = \{u : \langle u, v_i\rangle - q_i \leqslant 0, \quad i = 1, 2, \ldots, m\},$$

then Q is the pseudo–contractive (Fejér) mapping and can be built in the form

$$P_Q = P_{Q_1}(P_{Q_2} \ldots (P_{Q_m})) \quad \text{or} \quad P_Q = \sum_{i=1}^{m} \lambda_i P_{Q_i}, \ \sum_{i=1}^{m} \lambda_i = 1, \ 0 < \lambda_i < 1,$$

(see, Theorem 3.3.9), where P_{Q_i} is the projection into the half–space defined by the i–th inequality.

But if the *a priori* restrictions are given by the system of convex inequalities

$$Q = \{u : u \in U, \ g_i(u) \leqslant 0, \ i = 1, 2, \ldots, m\}, \tag{3.5.6}$$

then to find the metric projection, it is necessary to solve a problem of convex programming, so, we need to search for new alternative approaches.

For finding solutions of systems of convex inequalities with differentiable (in the general case, subdifferentiable) functionals, I.I. Eremin in [9–11] had suggested constructions of mappings from the class $\mathcal{F}_Q$, where $Fix(T) = Q$ and Q is the solutions set of system (3.5.6). This allowed one to formulate the convergence theorems for various classes of problems of mathematical programming. It was discovered ([42, 44]) that these mappings have the more strong property of contractivity. Namely, they belong to the class $\mathcal{P}_Q^\nu$. This means that they can be applied to processes (3.4.2) for solving the ill–posed problem with the *a priori* information.

Moreover, just as in $\mathcal{F}_Q$, the class $\mathcal{P}_Q^\nu$ possesses the property of closedness with respect to operations of superposition and convex combination (see, Theorem 3.3.8 and Corollary 3.3.11). Therefore, when applying $P_Q \in \mathcal{P}_Q^\nu$ and $V \in \mathcal{P}_M^\mu$, processes (3.4.2) adopt wide parallelizing of algorithms.

The basic construction of the Fejér mapping from the class $\mathcal{P}_Q^\nu$ for the set Q given by the system of inequalities (3.5.6) has the form

$$T(u) = u - d^+(u)e(u)/\|e(u)^2\|, \tag{3.5.7}$$

where $d(u) \geqslant 0$ is the convex functional and $Fix(T) = \{u : u \in U, \ d(u) \leqslant 0\}$, $e : U \to U$.

Write out concrete realizations of the functional d and the mappings e from formula (3.5.7)

$$d(u) = \sum_{i=1}^{m} k_i [g_i^+(u)]^{\mu_i}, \quad e(u) = \sum_{i=1}^{m} k_i \mu_i [g_i^+(u)]^{\mu_i - 1} \nabla g_i(u), \tag{3.5.8}$$

where $k_i > 0$, $\mu_i \geqslant 1$, ∇g_i is the gradient or subgradient of the functional g_i, and, under this, for $\mu_i = 1$, summation in the latter sum is bounded by the set $S(u) = \{j : g_j(u) > 0\}$,

$$d(u) = \max_{1 \leqslant i \leqslant m} g_i(u), \quad e(u) = \nabla g_{i(u)}, \quad i(u) \in \{j : g_j(u) = d(u)\}. \tag{3.5.9}$$

Taking account of the properties of the mappings T from the class $\mathcal{P}_Q^v$ established in Theorem 3.3.8 and the fact that the set Q in (3.5.6) can be represented in the form $Q = \bigcap_{i=1}^{m} Q_i$, let us write other two families of the mappings $T \in \mathcal{P}_Q^v$:

$$T = T_{j_1} T_{j_2} \dots T_{j_m}, \quad T = \sum_{i=1}^{m} \lambda_i T_i, \quad 0 < \lambda < 1, \tag{3.5.10}$$

where $j_1, j_2, \dots, j_m$, is an arbitrary permutation of the indices $\{1, 2, \dots, m\}$, $0 < \lambda_i < 1$, $\sum_{i=1}^{m} \lambda_i = 1$, and the mapping T_i is defined by the formula

$$T_i(u) = u - \lambda g_i^+(u) \nabla g_i(u) / \|\nabla g_i(u)\|^2, \quad 0 < \lambda < 2. \tag{3.5.11}$$

Note that these are the constructions of the Fejér mappings of forms (3.5.10) and (3.5.11) that allow one to use various technologies of parallelization in realizations of the iterative methods. These topics are in detail considered in book [46, Chapter 4].

3.6 Fejér processes for nonlinear equations

In the previous sections of this article in solving a problem with the *a priori* information, equation (3.1.1) with the linear bounded operator A was discussed as the basic one. Consider now the case when the basic equation is nonlinear

$$A(u) = f \tag{3.6.1}$$

and has the differentiable operator acting on a pair of the Hilbert spaces U and F. As in the linear case, we shall denote by M the set of solutions of equation (3.6.1).

When necessary, one can pass to the more general formulation. i.e., to minimization of the residual

$$\min\{\tfrac{1}{2}\|A(u) - f\|^2 : u \in U\}. \tag{3.6.2}$$

Denoting in (3.6.2) the goal function by $\Phi(u)$, we come to the necessary condition of extremum in problem (3.6.2)

$$\nabla\Phi(u) = A'(u)^*(A(u) - f) = S(u) = 0.$$

Thus, the iterative processes of the form

$$u^{k+1} = u^k - \beta_k[A'(u^k)^*(A(u^k) - f)] \equiv T(u^k) \tag{3.6.3}$$

are to be regarded as the methods of the gradient type.

Since $\beta_k = \beta$ from (3.6.3), we obtain the gradient method with the constant step (the Landweber method).

Using the principles of the choice of the parameter β, from the condition of minimum of the residual and the error for operator A linearized at the point u^k, we come to the processes ($S(u) = A'(u)^*(Au - f)$)

$$u^{k+1} = u^k - \frac{\|S(u^k)\|^2}{\|A'(u^k)S(u^k)\|^2}S(u^k) \equiv T(u^k), \tag{3.6.4}$$

$$u^{k+1} = u^k - \frac{\|A(u^k) - f\|^2}{\|S(u^k)\|^2}S(u^k) \equiv T(u^k) \tag{3.6.5}$$

that, analogically to the linear case, is called the *method of the steepest descent* and the *method of the minimal residual*, respectively (see, [8, 31, 33].

One can formulate the sufficient condition, which guarantees that the step operator in the gradient methods belongs to the class of the pseudo–contractive operators $\mathcal{P}_M^\nu$.

Theorem 3.6.1. *Let* $\sup\{\|A'(u)\| : u \in S_\rho(z)\} \leqslant N_1$ *and under some* $\varkappa > 0$ *in the neighborhood* $S_\rho(z)$ *of the solution* z *of equation (3.6.1), the following condition can be satisfied:*

$$\|A(u) - A(z)\|^2 \leqslant \varkappa\langle A(u) - A(z), A'(u)(u - z)\rangle. \tag{3.6.6}$$

Then under $\beta_k \equiv \beta < \frac{2}{\varkappa N_1^2}$ *in the method of gradients (3.6.3) and under* $\varkappa < 2$ *in methods (3.6.4) and (3.6.5), the step operators belong to the class* $\mathcal{P}_M^\nu$ *under* $\nu = \frac{2}{\beta N_1^2} - 1$ *and under* $\nu = \frac{2}{\varkappa} - 1$, *respectively.*

Consider now the modified Levenberg–Marquardt method (with the additional parameter $\beta > 0$, see, [46])

$$u^{k+1} = u^k - \beta[A'(u^k)^*A'(u^k) + \alpha_k I]^{-1}S(u^k) \equiv T_k(u^k). \tag{3.6.7}$$

Introduce the Hilbert norm

$$\|u\|_k^2 = \langle B_k u, u\rangle, \quad B_k = A'(u^k)^*A(u^k) + \alpha_k I,$$

that is equivalent (uniformly in k) to the main norm of the space U under $\alpha_k \geqslant \alpha > 0$.

Theorem 3.6.2. *Let* $\sup\{\|A'(u)\|\} \leqslant N_1 < \infty$ *and condition (3.6.6) be satisfied.*

Then under $0 < \beta < 2\alpha/\varkappa N_1^2$ *for the sequence of the operators* $T_k(u) = u - \beta B_k^{-1} S(u)$ *in the variable norm, the condition of pseudo–contractivity is satisfied at the iterative points, i.e.,*

$$\|T_k u^k - z\|_k^2 \leqslant \|u^k - z\|_k^2 - \nu \|u^k - T_k(u^k)\|_k^2$$

for some $\nu > 0$.

By Theorems 3.6.1, 3.6.2, and 3.4.1, the following corollaries are valid.

Corollary 3.6.3. *Let the assumptions of Theorems 3.6.1 and 3.6.2 be satisfied and the relation hold*

$$u^k \rightharpoonup \overline{u}, \quad S(u^k) \to 0 \Rightarrow S(\overline{u}) = 0 \quad (A(\overline{u}) = f).$$

Then the iterative sequences $\{u^k\}$ *generated by processes (3.6.4), (3.6.5), and (3.6.7), converges weakly to a solution of equation (3.6.1) and, in addition,* $\lim_{k\to\infty} \|Au^k - f\| = 0$.

Corollary 3.6.4. *If together with the basic equation (3.6.1) we have the a priori information* $\widehat{u} \in Q$ *and* P_Q *is the mapping of the Fejér type, for example, this is from the class* $\mathcal{P}_Q^\nu$ *with condition (3.4.1), then under condition of Theorem 3.6.2 and Corollary 3.6.3 the iterative processes*

$$u^{k+1} = P_Q T_k(u^k), \quad u^{k+1} = \lambda P_Q(u^k) + (1-\lambda) T_k(u^k) \tag{3.6.8}$$

weakly converge to $\widehat{u} \in Q \bigcap M$, *where* M *is the solutions set of equation (3.6.1),* T_k *is the step operator in method (3.6.7).*

Note that even in the case of the unique solvability of equation (3.6.1), processes (3.6.8) could be more effective by the accuracy than method (3.6.7) if the mapping P_Q adopts an economical realization.

Consider the process

$$\widetilde{u}^{k+1} = P_Q T_k(\widetilde{u}^k),$$

where T_k is defined by (3.6.7) or $T_k = T$ from (3.6.4), (3.6.5) and in these processes f is changed by $f_\delta : \|f - f_\delta\| \leqslant \delta$, where Q is the compact set containing the solution $\widehat{u}$.

Lemma 3.6.5. *Let equation (3.6.1) be uniquely solved and for any* $\delta < \delta_0$ *the number* $k(\delta) \to \infty$ *exists that is defined by the relation*

$$\|A(u^{k+1}) - f_\delta\| \leqslant \tau\delta < \|A(\widetilde{u}^k) - f_\delta\|, \quad k = 1, 2, \ldots, k(\delta) - 1, \quad \tau > 1.$$

Then $\lim\limits_{\delta\to 0} \|\widetilde{u}^{k(\delta)} - \widehat{u}\| = 0$.

Remark 3.6.6. *The main condition for operator A that guarantees the pseudo–contractivity and the weak convergence is defined by relation (3.6.6), in which the solution z to be sought enters. This fact hampers the verification in practical problems. But one can demand the more strong condition* [31]

$$\|A(u) - A(\widetilde{u}) - A'(u)(u - \widetilde{u})\|^2 \leqslant \eta\|A(u) - A(\widetilde{u})\| \quad \forall u, \widetilde{u} \in S_\rho(u^0), \qquad (3.6.9)$$

which, as it is easy seen, implies (3.6.6) under $\varkappa = 2/(1-\eta^2)$ *(*$\eta < 1/2$*) and does not contain the solution* z *explicitly.*

Moreover, the following operators satisfy condition (3.6.9) (and, hence, (3.6.6), also) (see, [8, 31, 33]):

— the nonlinear operator A generated by the inverse coefficients problem for the differential equation

$$-(a(s)u(s)_s)_s = f(s), \quad s \in (0,1)$$

with the boundary conditions $u(0) = g_0$, $u(1) = g_1$;

— the operator A in the nonlinear Volterra equation

$$A(u) \equiv \int_0^t \varphi(u(s))ds = f(t), \quad \varphi \in C^2(R), \quad A : W_2^1[0,1] \to L_2[0,1],$$

where W_2^1 is the Sobolev space with the norm (3.1.5) ($n = 1$).

— the integral operator $A : L_2[a,b] \to L_2[a,b]$ in the plane problem of gravimetry [14]

$$A(u) \equiv \int_a^b \ln \frac{(t-s)^2 + H^2}{(t-s)^2 + (H - u(s))^2} ds = f(t);$$

— the nonlinear operator A generated by the problem of identification of the parameter $q(x)$ (for a function) for the differential equation of the second order [20]

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = q(x)u(x,t) + f(x,t), \ u(x,0) = \varphi(x), \ u_t(x,0) = \psi(x).$$

Remark 3.6.7. *In contrast to condition (3.6.6) that implies the weak convergence, the strong convergence of methods (3.6.3)–(3.6.5) and (3.6.7) (see,* [8, 31]*) holds if condition (3.6.9) is satisfied.*

In conclusion, consider the problem of conditional convex minimization

$$\min\{f(u) : u \in Q\}$$

with the non–empty solutions set M. The mapping P given by the relation

$$P : v \to \arg\min\{f(u) + \alpha\|u - v\|^2 : u \in Q\} \quad (\alpha > 0)$$

is called the *prox–mapping* (see, [25, 28, 32, 44]), and under this $P \in \mathcal{P}^1$ and $Fix(P) = M$. This allows one to apply the process $u^{k+1} = P(u^k)$ to approximation of elements $u \in M$. Under $f(u) = \|Au - f\|^2$, $Q = U$, the process passes into the iterated version of the Tikhonov method (see, (3.5.2)). These processes give the opportunity to work (under the constant $\alpha > 0$) with the strong convex

functional, i.e., with the problem that is correct according to Tikhonov method. This is especially important for using the subgradient methods in the case of the non–smooth functional $f(u)$ (for instance, of the functional obtained under the Tikhonov regularization with the non–smooth stabilizer). The mapping Λ defined by formula (3.3.7) has analogous properties.

3.7 Applied problems with *a priori* information and methods for solution

3.7.1 Atomic structure characterization

In investigation of the atomic structure (the close order of the atoms placing) of the one–component non–ordered (amorphous) materials by the Röntgen–spectral analysis, the Fredholm integral equation of the first kind appears

$$Ag \equiv \int_a^b K(k,r)g(r)dr = \mathcal{X}(k), \quad c \leqslant k \leqslant d, \tag{3.7.1}$$

where $\mathcal{X}(k)$ is the absorption coefficient of a bunch of the monochromatic Röntgen rays in the material under investigation made experimentally, $g(r)$ is the function of the radial distribution of the atoms to be defined. This function is the main structural characteristic of the material (see details in [2, 44]).

From its definition and physical sense, the function $g(r)$ satisfies the following conditions:

$$g(r) \geqslant 0, \quad \langle g, v\rangle \equiv \frac{3}{b^3 - a^3}\int_a^b r^2 g(r)dr = 1. \tag{3.7.2}$$

Thorough numerical analysis [44] has shown that the kernel of operator A is nontrivial, hence, equation (3.7.1) has non–unique solution. The approximate solution built on the basis of the classical Tikhonov method

$$g^\alpha : \min\{\|Ag - \mathcal{X}\|^2 + \alpha\|g\|^2_{W_2^1 \ (\text{or } L_2)}\}, \tag{3.7.3}$$

appears to be absolutely non–physical (it is negative on the most part of the argument interval). This means that the normal solution reconstructed by method (3.7.3) does not satisfy conditions (3.7.2).

To obtain *physically meaning* solution, the iterated version of the Tikhonov method together with projecting onto the *a priori* sets (restrictions) was applied. To do this, the prox–mapping (the step operator of the basic method) is defined

$$V : v \to \arg\min\{\|Ag - \mathcal{X}\|^2_{L_2} + \alpha\|g - v\|^2_{W_2^1 \ (\text{or } L_2)}\}$$

and the iterative process is constructed (see (3.4.2))

$$g^{k+1} = PV(g^k) \equiv T(g^k), \tag{3.7.4}$$

where $P = P_{Q_2} P_{Q_1}$, P_{Q_1} is the metric projection onto the set $Q_1 = \{g : \langle g, v\rangle = 1\}$, P_{Q_2} is the projection onto the set $Q_2 = \{g : g \geqslant 0\}$. Here, the step operator T is the superposition of two pseudo–contractive mappings P with $Fix(T) = Q = Q_1 \bigcap Q_2$ and V with $Fix(V) = M$ that is the solutions set of problem (3.7.1); and, under this, $Fix(T) = M \bigcap Q$. Thus, we have satisfied conditions of Theorem 3.4.13.

Representative numerical experiment with model and real data has shown that taking account of the *a priori* restrictions in algorithm (3.7.4) with using the projections, we obtain solution of good quality. Details of the numerical simulation are given in book [44] (see, also, the references there).

It is worthy to note that the solution also is reconstructed with the satisfying accuracy if to use, as the operators V, ones generated by the methods of the steepest descent, the method of the minimal error, and the method of the conjugated gradients with the same mapping P.

3.7.2 Radiolocation of the ionosphere

Consider the oblique radiolocation of the ionosphere from the ground surface when the vertical profile of the electron concentration has to be defined along the epicentral distances (as the function of the beam parameter) obtained on two frequencies. In this process, the so-called *problem of the waveguides* appears as a result of violation of the monotone dependence of the electron concentration along the hight. In this case the uniqueness of the solution is violated, and, moreover, the traditional methodic (inversion of the integral of the Abel type) allows one to define the concentration only at the beginning of the waveguide (i.e., of the path with violation of the monotonicity).

In work [3], the problem of the waveguides has been solved and it was established that it is impossible to define the electron concentration in the unique way in the waveguide, but one can define the "measure of the waveguide", the measure of the Lebesgue sets for the refraction coefficient from the Fredholm–Stiltjes equation of the form

$$\int_a^b K(p, r) dF(r) = R(p), \tag{3.7.5}$$

where $K(p, r)$ is some given function, $R(p)$ is the function to be defined from the experiment, $F(r)$ is the measure of the waveguide to be sought. Moreover, by using the function $F(r)$, the computational procedure was constructed for defining the electron concentration above the waveguide (see details in [3]).

If use the Tikhonov regularization method of the first order in the standard form, the quality of the solution of equation (3.7.5) is low. But if take account of the *a priori* data about the solution F in the form

$$F(a) = 0, \quad F'(b) = 0, \quad F'(r) \geqslant 0, \quad F(r) \geqslant 0$$

and apply the iterative process in form (3.7.4), then it is succeeded to increase sufficiently the quality of the approximate solution. Under this, $P = P_{Q_2} P_{Q_1}$, where P_{Q_1} is the mapping of forms (3.5.7), (3.5.8) for the system of linear inequalities obtained after discretization of the condition $F'(r) \geqslant 0$ and P_{Q_2} is the projection onto the positive orthant.

3.7.3 Image reconstruction

The problem of reconstructing the image is corrupted by the hardware function of the measuring device and the additive error is reduced to solving the two–dimensional integral equation of the first kind of the convolution type

$$Au \equiv \int_0^1 \int_0^1 K(x-t, y-s)u(t,s)dt\,ds = f(x,y). \tag{3.7.6}$$

The numerical experiments (see, for example, [49]) show that in problems (3.7.6) joined with reconstruction of images application of the Tikhonov regularization of the first order (with the stabilizer $\Omega(u) = \|u\|^2_{W_2^1}$) is not always advisable because of the strong smoothing of the non-smooth solution. Application of the stabilizer

$$\Omega(u) = \|u\|^p_{L_p} + J(u) \tag{3.7.7}$$

with the variation $J(u)$ of this or that form (see, [49]) usually leads to the better results. Particularly, the following total variation [15] is used as the functional $J(u)$:

$$J(u) = \sup \left\{ \int_0^1 \int_0^1 u(x,y)\ \mathrm{Div} v(x,y)dx\,dy : \ v \in C_0^1(\Pi, R^2), \ \ |v| \leqslant 1 \right\}.$$

For solving the regularized problem

$$\min\{\|Au - f_\delta\|^2_{L_p} + \alpha\Omega(u) : u \in U\} = \Phi^* \tag{3.7.8}$$

with the nondifferentiable stabilizer $\Omega(u)$ of form (3.7.7), the following approach was suggested in works [43, 45]. Assume that some estimate $\widetilde{\Phi}$ for Φ^* is known, i.e.,

$$\widetilde{\Phi} \leqslant \Phi^* + \varepsilon, \quad \varepsilon > 0,$$

together with the *a priori* information of the form $\widetilde{u} \geqslant 0$. Now, instead of (3.7.8) it is possible to solve the system of convex inequalities

$$\Phi^\alpha(u) - \widetilde{\Phi} \leqslant 0, \quad u \geqslant 0, \tag{3.7.9}$$

where Φ^α is the goal functional in (3.7.8).

The following iterative process of the subgradient type was constructed by the methodic of Section 3.5 for system (3.7.9):

$$u^{k+1} = P^+\left\{u^k - \lambda\frac{[\Phi^\alpha(u^k) - \widetilde{\Phi}]^+ + \partial\Phi^\alpha(u^k)}{\|\partial\Phi^\alpha(u^k)\|^2}\right\}, \quad 0 < \lambda < 2.$$

Here, P^+ is the projection onto the positive orthant and $\partial\Phi^\alpha(u^k)$ is an arbitrary subgradient of the functional Φ^α. This process solves problem (3.7.8) rather sufficiently [45].

Another different approach to the numerical solution of the problem of non–smooth minimization of (3.7.8) (see, [24, 43, 49]) is based on the preliminary approximation of the non–smooth stabilizer Ω by some family of the differentiable functionals Ω_β with further application of the methods of smooth optimization.

3.7.4 Thermal sounding of the atmosphere

In the inverse problems of the thermal sounding the atmosphere that join with defining the temperature $T(h)$ and concentration $n(h)$ of the green-house gases (CO, CO_2, CH_4) as function of the hight, the spectra are used that were measured by the satellite sensor. In these problems, the following nonlinear integral equation has to be solved

$$A(u) \equiv \int_{-\infty}^{+\infty} W(\nu')F(\nu - \nu')d\nu' = \Phi(\nu),$$

which is the convolution of the spectrum of the high resolution W with the hardware function F. The function W depends on the absorption coefficient $B_\nu(T(h), n(h))$ that nonlinearly depends on the parameters $u(h) = (T(h), n(h))$ to be defined.

Usually, only a part of the parameters (for example, the temperature and methane, the temperature, water vapor, and CO_2) are defined, but others are assumed to be fixed and their values are chosen from the data base for the region under investigation and the chosen time and the observation conditions.

Problems of such a type have one peculiarity, namely, there often exists rather precious *a priori* information about the solution to be found in the form of the two-sided inequalities

$$Q = \{u : \quad \underline{u}(h) \leqslant u(h) \leqslant \overline{u}(h)\}. \tag{3.7.10}$$

Sufficiently good results are obtained by using the iteratively regularized methods of the Gauss–Newton type, particularly, by application of the modified Levenberg–Marquart method (see (3.6.7))

$$u^{k+1} = P_Q[u^k - \beta(A'(u^k)^* A'(u^k) + \alpha_k I)^{-1}(A'(u^k)^*(A(u^k) - \Phi))]$$

with the operator P_Q of projection onto the n–dimensional parallelepiped (3.7.10) (see [16, 47]).

3.7.5 Testing a wellbore/reservoir

Under investigation and interpretation of productivity of a wellbore/reservoir system the problem of the solution of the convolution Volterra equation [34] arises

$$\int_0^1 q(t-\tau)g(\tau)\,d\tau = \Delta p(t), \quad \Delta p(t) = p_0 - p(t),\ t \in [0,T]. \tag{3.7.11}$$

Here, $p(t)$ and $q(t)$ are the pressure and the flow rate, respectively; p_0 is the initial reservoir pressure; $g(t) = dp_u(t)/dt$, where $p_u(t)$ is the so-called constant–unit–rate pressure response. The functions $p_u(t)$ and $g(t)$ are unknown and should be reconstructed by $\Delta p(t)$ and $q(t)$ that are given with noise. The functions $p_u(t)$ and $g(t)$ are used for analysis of a wellbore/reservoir system [5].

As it is known [7], the function $g(t)$ satisfies the following constraints

$$C \geqslant g(t) \geqslant 0, \quad g(t) \leqslant 0, \quad g'(t) \leqslant 0, \quad g''(t) \geqslant 0. \tag{3.7.12}$$

So, we must solve the system

$$Ag = \Delta p. \quad g \in M, \tag{3.7.13}$$

where $M = \{\, g : 0 \leqslant g \leqslant C,\ g'(t) \leqslant 0,\ g''(t) \geqslant 0 \,\}$.

Now the following iterative processes

$$\begin{array}{l} \text{either} \\ g^{k+1} = P_M(g^k - \beta(A^* A g^k - A^* \Delta p)),\ 0 \leqslant \beta \leqslant 2/\|A\|^2, \\ \text{or} \\ g^{k+1} = P_M(A^* A + \alpha I)^{-1}(\alpha g^k + A^* \Delta p) \end{array} \tag{3.7.14}$$

are appropriate for solving system (3.7.13). The mapping P_M from (3.7.14) is given, for example, by formulas (3.5.7)–(3.5.9) or (3.5.10), (3.5.11) from Section 3.5.

Since the set of functions $\bar{M}$ given by inequalities (3.7.12) is compact in $L_2[0,T]$, we can also use the Ivanov quasi-solution method [18, 19]

$$\min\{\|Ag - \Delta p\|^2 \,:\, g \in \bar{M}\}. \tag{3.7.15}$$

After discretization of problem (3.7.15) by finite-difference method, we arrive at minimization of quadratic function with linear constraints, which can be solved by methods of the gradient type.

Another approach to solving equation (3.7.11) is based on the preliminary transition from the linear equation (3.7.11) to the nonlinear one

$$A(z) = \int_{-\infty}^{\log_{10} t} q(t - 10^{\sigma})\, 10^{\sigma} d\sigma = \Delta p(t),\ t \in [0, T] \tag{3.7.16}$$

after changing the variable τ and the function $g(t)$ as follows: $\sigma = \log_{10} \tau,\ z(\sigma) = \log_{10}(\tau g(\tau))$. Now for solving (3.7.16), we can apply the iterative processes in the form

$$g^{k+1} = P_M V(g^k),$$

where V is the step operator of the iterative method of the Gauss–Newton type (f.e., (3.6.7)) and P_M is a Fejér mapping responding for constraints (3.7.12) (see Section 3.5).

3.8 Conclusions

This chapter presents the review of methods for solving linear and nonlinear operator equations of the first kind with *a priori* information. The approach discussed in detail is based on iterative processes with operators of the Fejér type. The description of algorithms is given for solving some inverse problems that are met in various scopes of the mathematical physics.

References

1. R. Acar and C. R. Vogel, Analysis of bounded variation penalty method for ill-posed problems, *Inverse Problems*, **10**, 1217–1229, 1994.
2. A. L. Ageev, Yu. A. Babanov, V. V. Vasin and N. V. Ershov, Constructing the regunarizing algorithms for determining the structure of amorphous bodies by the method of the Röntgen-spectral structural analysis, Chislenn. i analitich. metody resheniya zadach mekhaniki sploshnoy sredy, Sverdlovsk, 3–25, 1981.
3. A. L. Ageev, V. V. Vasin, E. N. Bessonova and V. M. Markushevich, Radiolocation of the atmosphere on two frequencies. Algorithmic analysis of the integral Fredholm-Stiljes equation, in: Teoretich. probl. in geofizike, Vychisl. Seysmologiya. Nauka, Moscow, **29**, 100–118, 1997.
4. A. B. Bakushinsky and A. V. Goncharsky, *Iterative Methods for Solving the Ill-Posed Problems*, Nauka, Moscow 1989.
5. J. A. Bourdet, et al., *Use of pressure derivative in well test interpretation*, SPE Paper 12777, SPE Formation Evaluation, 293–302, 1989.
6. E. Chavent and K. Kunish, Regularization of linear least squares problems by total bounded variation control, *J. Optimization and Calculus of Variations*, **2**, 359–376, 1997.

7. K. H. Coats, et al., Determination of aquifer influence functions from field data, *Trans. SPE (AIME)*, **231**, 1417–1424, 1964.
8. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Acad. Publ., Dordrecht ets. 1996.
9. I. I. Eremin, Generalization of the Motzkin-Agmon relaxational method, *Uspekhi Mat. Nauk*, **20** (2), 183–187, 1965.
10. I. I. Eremin, On systems of inequalities with convex functions in the left-hand parts, *Izv. AN SSSR, Ser. Matem.*, **30**(2), 265–278, 1966.
11. I. I. Eremin, Methods of Fejér approximations in the convex programming, *Matem. Zametki*, **3**(2), 217–234, 1968.
12. L. Fejér, Über die Lage der Nullstellen von Polynomen, die aus Minimumforderungen gewisser Art entspringen, *Math. Ann.*, **85**(1), 41–48, 1922.
13. A. Genel and L. Lindenstrauss, An example concerning fixed point, *Israel J. Math.*, **22**(1), 81–86, 1975.
14. S. F. Gilyazov and N. L. Gol'dman, *Regularization of Ill-Posed Problems by Iteration Methods*, Kluwer Acad. Publ., Dordrecht ets, 2000.
15. E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Basel, 1984.
16. K. G. Gribanov, V. I. Zakharov and V. V. Vasin, Iterative regularization in the problem of defining CO_2 in the atmosphere by the data of the satellite sounding, Abstr. of the Internat. Conf. *Algorithmic Analysis of the Unstable Problems*, 119–120, Ekaterinburg, Russia, September 01–06, 2008.
17. B. Halperin, Fixed points of nonexpansive maps, *Bull. Amer. Math. Soc.*, **73**(6), 957–961, 1967.
18. V. K. Ivanov, On linear ill-posed problems, *Dokl. AN SSSR.*, **145**(2), 270–272, 1962.
19. V. K. Ivanov, On the ill-posed problems, *Matemat. Sborn.*, **61**(2), 211–223, 1963.
20. S. I. Kabanikhin, R. Kowar and O. Scherzer, On the Landweber iteration for the solution of a parameter identification problem in a hyperbolic partial differential equation of second order, *J. Inv. Ill-posed Problems*, **6**(5), 403–430, 1998.
21. S. Kaczmarz, Angenaherte auflosung von systemen linearer gleichungen, *Bull. Inst. Acad. Pol. Sci. Leet.*, **35**(4), 355–357, 1937.
22. M. M. Lavren't'ev, On interval equations of the first kind, *Dokl. AN SSSR.*, **127**(1), 31–33, 1959.
23. M. M. Lavren't'ev, On some ill–posed problems of the mathematical physics, *Izd. Sibirsk. Otdel. AN SSSR*, Novosibirsk, 1962.
24. A. C. Leonov, Piecewise-uniform regularization of the two-dimensional ill-posed problems with discontinuous solutions: numerical analysis, *Zhurn. Vychisl. Matem. i Matem. Fiziki*, **39**(12), 1934–1944, 1999.
25. B. Martinet, Determination approachee d'un point fixe d'une application pseudo-contractante, Cas de l'application prox, C. R. Acad. Sci. Paris 274, Ser. A–B, A163–A165, 1972.
26. S. Maruster, Quasi-nonexpansivity and two classical methods for solving nonlinear equations, *Proc. Amer. Math. Soc.*, **62**(1), 119–123, 1977.
27. L. D. Menikhes, On regularizability of mappings inverse to the intergral operators, *Dokl. AN SSSR*, **241**(2), 625–629, 1978.
28. J. J. Moreau, Proximite et dualite dans un espace Hilbertien, *Bull. Soc. Math. France*, **93**(2), 273–299, 1965.
29. V. A. Morozov, N. P Goldman and M. K. Samarin, Method of the descriptive regularization and quality of the approximate solution, *Inzhenern. Fisichesk. Zhurn.*, **38**(6), 117–121, 1977.
30. T. S. Motzkin and J. J. Schoenberg, The relaxation method for linear inequalities, *Canad. J. Math.*, **6**(3), 393–404, 1954.
31. A. Neubauer and O. A. Scherzer, A convergence rate results for a steepest desceut method and minimal error method for the solution of nonlinear ill-posed problems, *J. Anal. Appl.*, **14**(2), 369–378, 1995.

32. R. T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM J. Control and Optim.*, **14**(5), 877–898, 1976.
33. O. Scherzer, Convergence criteria of iterative methods based on Landweber iteration for solving nonlinear problems, *J. Math. Anal. Appl.*, **194**, 911–933, 1995.
34. T. Schroter, et al., Deconvolution of well test data analysis as a nonlinear total least squares problem, SPE Paper 71574, 2001 SPE Annual Technical Conference and Exibition, New Orlean, 30 September–3 October, 2001.
35. A. N. Tikhonov, About stability of the inverse problens, *Dokl. AN SSSR.*, **39**(5), 195–198, 1944.
36. A. N. Tikhonov, On solving the ill-posed problems and the method of regularization, *Dokl. AN SSSR.*, **151**(3), 501–504, 1963.
37. A. N. Tikhonov, About regularization of the ill-posed problems, *Dokl. AN SSSR.*, **153**(1), 49–52, 1963.
38. A. N. Tikhonov and F. P. Vasil'ev, Methods for solving the ill-posed extremal problems, *Math. models and numerical methods*, Banach Center Publ., Warszava 3, 297–342, 1978.
39. A. N. Tikhonov, A. V Goncharsky, V. V Stepanov and A. G. Yagola, *Regularizing Algorithms and A Priori Information*, Nauka, Moscow, 1983.
40. V. V. Vasin, Proximal algorithm with projection in problems of the convex programming, Ural Sci. Center AN SSSR, Instit. Matem. i Mekhan., Sverdlovsk, 1982.
41. V. V. Vasin, Discretization, iterative-approximative algorithms for solving unstable problems and their applications, Dissert. ... Dokt. Fiz. mat. nauk, Vychislit. Tzentr Sib. Branch of AN SSSR, Novosibirsk, 1985.
42. V. V. Vasin, Iterative methods for solving the ill-posed problems with *a priori* infromation in the Hilbert spaces, *Zhurn. Vychisl. Mathem. i Fiziki.*, **28**(7), 971–980, 1988.
43. V. V. Vasin, Regularization and iterative approximation for linear ill-posed problems in the space of functions of bounded variation, *Proc. Stekl. Inst. Math.*, Suppl. 1, S225–S229, 2002.
44. V. V. Vasin and A. L. Ageev, *Ill-Posed Problems with A Priori Information*, VSP, Utrecht, The Netherlands, 1995.
45. V. V. Vasin and T. I. Serezhnikova, Two-stage method for approximation of the nonsmooth solutions and reconstruction of the noised image, *Avtomat. i Telemekh.*, **3**, 392–402, 2004.
46. V. V. Vasin and I. I. Eremin, Operators and iterative processes of the Fejér type, Theory and Applications, Instit. comput. research. – Regular & Chaotic Dynamics, Moscow-Izhevsk, 2005.
47. V. V. Vasin, K. G. Gribanov, V. I. Zakharov, et al., Regular methods of the solution of ill-posed problems for remote sensing of Earth atmosphere using high-resolution spectrometry, Proc. SPE 6880, 65800T, Dec. 12, 2006.
48. V. A. Vinokurov, On one necessary condition of regularizability by the Tikhonov method, *Dokl. AN SSSR*, **195**(3), 530–531, 1970.
49. C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia. 2002.

Chapter 4
Regularization of Naturally Linearized Parameter Identification Problems and the Application of the Balancing Principle

Hui Cao and Sergei Pereverzyev

Abstract. The chapter is a survey on recently proposed technique for parameter identification in partial differential equations. This technique combines natural linearization of an identification problem with the Tikhonov scheme, where the regularization parameter is chosen adaptively by means of the so-called balancing principle. We describe the natural linearization approach and show how it can be treated within the framework of Tikhonov regularization as a problem with noisy operator and noisy data. Then the balancing principle is discussed in the context of such a problem. We demonstrate the performance of proposed technique in some typical parameter identification problems.

4.1 Introduction

Natural linearization (NL) was introduced in [14] as an alternative to output least-squares (OLS) approach to nonlinear parameter identification problems for partial differential equations. OLS are known as a powerful tool for solving such inverse problems, but require a subroutine to solve the forward problem that can handle the nonlinearity of the differential equation, and since hundreds of forward problems may have to be solved during the minimization of OLS cost functional (or its regularized version), this method has a tendency to become very expensive.

NL, as presented in [14], has the following advantages over OLS: original nonlinear inverse problem is reduced to a finite sequence of linear ones; resulting linear problem is one-dimensional while the underlying process can depend on

Hui Cao and Sergei Pereverzyev
Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Science, Altenbergstrasse 69, 4040 Linz, Austria.
e-mail: {hui.cao,sergei.pereverzyev}@ricam.oeaw.ac.at

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

several time and space variables; sometimes only a single call of the solver for corresponding forward problem is necessary.

At first glance these advantages are due to a specific structure of the identification problem discussed in [14], where a parabolic system with unknown nonlinearity in the boundary condition has been considered. But as we will see in the sequel, NL scheme can also be successfully implemented for the identification problems in both elliptic system and parabolic system. It seems now that there is a class of nonlinear parameter identification problems that can be treated by NL technique. At this point it is worthy to stress once again that speaking about natural linearization we have in mind a reduction of original nonlinear inverse problem to a final sequence of linear ill-posed problems, when a regularized solution of preceding problem is used as data for succeeding one. For both such problems no information concerning the smoothness of their solutions is usually given. Then, as a consequence of their ill-posedness, no reliable estimations for the error of the approximate solution can be obtained. But an approximation error for the preceding problem plays a role of data noise level for the next one. As far as it is unknown, no consistent regularization strategy for succeeding problem can be proposed, as it follows from Theorem 3.3 [15] known also as a Bakushinskii's negative result.

At the same time, for some problems from the above mentioned ill-posed problems sequence (at least for the first of them) data noise level is usually given, or can be in principle estimated. Therefore, one can regularize other problems relating regularization parameters with data noise level known for preceding problem. One of the goals of the present work is to propose a regularization strategy based on this idea.

Concisely speaking, an adaptive strategy of regularization parameter choice, which is known as the *balancing principle*, is studied relying on either the estimated noise level (see Section 1.1.3) or the value of perturbation level known for the preceding linear ill-posed problem (see Section 1.1.4). Therefore, balancing principle will be presented first in a basic style, then in an extended form. In both cases, such *a posteriori* regularization parameter choice strategy can provide order-optimal regularization error estimates.

In this context, we choose the following parameter identification problems as prototype problems for the application of Natural linearization, as well as Balancing principle, since they are typical nonlinear inverse problems no matter whether the corresponding forward problems are linear or not.

Problem 4.1.1 (Parameter identification in elliptic equation–(P1))**.** *Consider the boundary value problem*

$$\begin{aligned} -\bigtriangledown (a \bigtriangledown u) &= f \ \textit{in}\ \Omega, \\ u &= g \ \textit{on}\ \partial\Omega. \end{aligned} \tag{4.1.1}$$

We are interested in identifying the coefficient $a = a(x)$ from the noisy measurement u^δ of the solution u. Here Ω is a convex domain with Lipschitz boundary, $f \in L^2(\Omega)$, $g \in H^{\frac{3}{2}}(\partial\Omega)$, and for some fixed noise level δ we have

$$\|u^\delta - u\|_{L^2(\Omega)} \leq \delta. \tag{4.1.2}$$

Problem 4.1.2 (Parameter identification in quasi-linear parabolic system–(P2))**.** *Consider the problem of identifying the diffusion coefficient $a(\tau)$ in a quasi-linear parabolic system*

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla \cdot (a(u)\nabla u) &= G(t,x) \ \textit{in}\ [0,T]\times\Omega, \\ a(u)\frac{\partial u}{\partial n} &= g(t,x) \ \textit{on}\ [0,T]\times\partial\Omega, \\ u(0,x) &= u_{in}(x) \ \textit{in}\ \Omega, \end{aligned} \tag{4.1.3}$$

from noisy short-time observation $u^\delta(t,x)$ of the solution $u(t,x)$, i.e.,

$$u^\delta(t,x) = u(t,x) + \delta\xi(t,x), \quad (t,x) \in [T-\sigma, T]\times\Omega. \tag{4.1.4}$$

Here Ω is a bounded, simply-connected domain with sufficiently smooth boundary $\partial\Omega$ and all the functions G, g and u_{in} are assumed to be known.

The identification problems (P1) and (P2) can find a wide range of applications, since the partial differential systems (4.1.1) and (4.1.3) serve as mathematical models for many processes in engineering and industry. They have been extensively discussed in the literature as model problems for parameter identification (e.g.[2, 3, 13, 16, 19, 28, 47]).

Before further discussion, we briefly make the following assertions for (P1).

- **On the forward problem of solving (4.1.1):** If Ω is a convex domain with Lipschitz boundary, for $a \in L^\infty(\Omega)$ bounded away from zero, and $f \in L^2(\Omega), g \in H^{\frac{3}{2}}(\partial\Omega)$, (4.1.1) has a unique weak solution $u \in H^1(\Omega)$.
- **Uniqueness for the inverse problem in the ideal case:** If a is known on the boundary $\partial\Omega$ and ∇u is bounded away from zero then by the basic theory for hyperbolic boundary value problems (see, e.g. [4]), a is uniquely determined on the whole domain Ω by (4.1.1).

The last assertion is established theoretically in the ideal case. In this context, we only assume that the search-for parameter a for the unperturbed parameter identification problems exists. Based on such assumption, our effort here is devoted to the nonlinear ill-posed problem of identifying $a(x)$ from the noisy measurement u^δ of u with $\|u^\delta - u\|_{L^2(\Omega)} \leq \delta$.

As for (P2), similarly, we have the following facts: (see [23, Chapter 9]).

- **On the forward problem of solving (4.1.3):** Assume that Ω yields to the condition described in (4.1.3). For $G(t,\cdot) \in L^2(\Omega)$, $u_{in} \in L^2(\Omega)$, $g(t,x), \partial g/\partial t \in L^2([0,T]) \otimes H^{\frac{1}{2}}(\partial\Omega)$, (4.1.3) has a unique weak solution $u \in C([0,T]) \otimes H^1(\Omega)$.
- **Uniqueness for the inverse problem in the ideal case:** If $g(t,x), \partial g/\partial t \geq 0$ (and not identically zero) on $(0,T)\times\partial\Omega$, then the coefficient a can be uniquely identified.

4.2 Discretized Tikhonov regularization and estimation of accuracy

In each step of NL, we will meet with the problem of the regualrization of a linear operator equation, where both the operator and the right-hand side term are subject to error. Now, we aim at the estimation of accuracy of such kind of regularization. At first, we will give the general model of the regularization problem and describe the algorithm of the discretized Tikhonov regularization. Then, the theory of operator monotone functions as an auxiliary tool is studied to draw the final result concerning the estimation of accuracy .

Let X and Y be separable Hilbert spaces over the reals. We denote the inner products in these spaces by $\langle\cdot,\cdot\rangle$, and corresponding norms by $\|\cdot\|$. Moreover, a symbol $\|\cdot\|$ stands also for a standard operator norm. It will be always clear from the context whose space or norm is being considered, if not specified.

In this section we briefly summarize a concept of the regularization of linear ill-posed operator equations

$$\bar{A}s = \bar{r} \tag{4.2.1}$$

with approximately known operator and right-hand side term, which means that we are given A, r instead of $\bar{A}$, $\bar{r}$ such that

$$\|\bar{A} - A\| \le \varepsilon_1, \tag{4.2.2}$$

$$\|\bar{r} - r\| \le \varepsilon_2, \tag{4.2.3}$$

where the operators $\bar{A}, A$ act compactively between X and Y, $\bar{r} \in \mathrm{Range}(\bar{A}) \subset Y$, $r \in Y$, $\varepsilon_1, \varepsilon_2 \in (0,1)$. Recall that equation (4.2.1) is classified as essentially ill-posed problem if $\mathrm{Range}(\bar{A})$ is not closed in Y. Facing such an equation, what one usually looks for is a stable approximation for the Moore-Penrose generalized solution of (4.2.1) defined as an element $\bar{s} \in X$ with minimal X-norm such that

$$\|\bar{A}\bar{s} - \bar{r}\| = \inf\{\|\bar{A}s - \bar{r}\|, s \in X\}.$$

4.2.1 Generalized source condition

Using a singular value decomposition

$$\bar{A} = \sum_{k=1}^{\infty} \lambda_k \langle u_k, \cdot\rangle w_k \tag{4.2.4}$$

of the operator $\bar{A}$ one can represent $\bar{s}$ as

$$\bar{s} = \sum_{k=1}^{\infty} \lambda_k^{-1} \langle w_k, \bar{r}\rangle u_k, \tag{4.2.5}$$

where $\{u_k\}$, $\{w_k\}$ are orthonormal systems of eigenvectors of the operators $\bar{A}^*\bar{A} : X \to X$ and $\bar{A}\bar{A}^* : Y \to Y$, $\lambda_k^2, k = 1, 2, \cdots$, are corresponding eigenvalues (decreasing with k), and $\bar{A}^* : Y \to X$ is the adjoint of the operator $\bar{A}$.

In view of the representation (4.2.5) it is natural to measure a smoothness of the Moore-Penrose solution $\bar{s}$ against the decay rate of the Fourier coefficients $\langle w_k, \bar{r}\rangle$. A well-known Picard criterion asserts that $\bar{s}$ has a zero smoothness (merely $\bar{s} \in X$) if and only if

$$\sum_{k=1}^{\infty} \lambda_k^{-2} \langle w_k, \bar{r}\rangle^2 < \infty.$$

If a smoothness of $\bar{s}$ is higher than a conventional one, then the Fourier coefficients $\langle w_k, \bar{r}\rangle$ should decay much faster than λ_k. More precisely, not only Picard criterion but also a stronger condition

$$\sum_{k=1}^{\infty} \frac{\langle w_k, \bar{r}\rangle^2}{\lambda_k^2 \phi^2(\lambda_k^2)} < \infty$$

should be satisfied, where ϕ is some continuous function defined on the interval $[0, a)$. Here we assume that $\phi(0) = 0$ and a is a fixed positive number with $\{\lambda_k\} \subset [0, a)$, i.e., $a > b$, $b = \lambda_1^2 = \|\bar{A}^*A^*\|$. Then

$$w = \sum_{k=1}^{\infty} \frac{\langle w_k, \bar{r}\rangle}{\lambda_k \phi(\lambda_k^2)} u_k \in X,$$

and in view of (4.2.5)

$$\bar{s} = \sum_{k=1}^{\infty} \phi(\lambda_k^2) \langle u_k, w\rangle u_k = \phi(\bar{A}^*\bar{A}) w \in X.$$

Thus, additional smoothness of $\bar{s}$ can be expressed as an inclusion

$$\bar{s} \in \bar{A}_{\phi,R} := \{s \in X : s = \phi(\bar{A}^*\bar{A})w, \quad w \in X, \quad \|w\| \le R\}, \tag{4.2.6}$$

that goes usually under the name of source condition. In this context the function ϕ is called index function. Ill-posed inverse problems under such general source conditions can be found throughout work [24], and also in [1, 12, 21, 48]. Recent progress has been reported in [6, 36, 39, 40].

Source conditions given in terms of powers $\phi(\lambda) = \lambda^\mu$ were studied extensively. Selected references from the huge literature are [15, 49]. For severely ill-posed problems, where $\phi(\lambda) = \log^{-\mu} \frac{1}{\lambda}$, we can refer to [22, 33, 43].

Moreover, recently in [34] it has been observed that for any $s \in X$ and compact operator $A : X \to Y$ there are an index function ϕ and a number R such that $s \in A_{\phi,R}$. From this view point the smoothness assumption (4.2.6) is not restrictive at all.

4.2.2 Discretized Tikhonov regularization

As we know, for obtaining a stable approximate solution of an ill-posed problem (4.2.1) regularization procedure should be employed, and Tikhonov regularization is the most popular one. Recall that in accordance with (4.2.2) only a perturbed equation

$$As = r \tag{4.2.7}$$

is accessible, and it may have no solution.

Within the framework of Tikhonov regularization scheme it is substituted for uniquely solvable equation

$$\alpha s + A^*As = A^*r, \quad \alpha > 0. \tag{4.2.8}$$

But in general this regularization procedure becomes numerically feasible only after appropriate discretization. In the present context a discretization means a substitution of (4.2.8) for a problem in some n-dimensional space $V_n \subset X$. Let P_n be the orthogonal projector from X onto V_n. A discretized Tikhonov regularized solution is defined as $s_{\alpha,n} = g_\alpha(B^*B)B^*r$, where $B = AP_n$, $g_\alpha(\lambda) = 1/(\alpha + \lambda)$ is Tikhonov regularization function. In other words, $s_{\alpha,n} \in V_n$ solves

$$\alpha s + B^*Bs = B^*r, \quad \alpha > 0.$$

Another way to look at this is the following: we seek for the solution of (4.2.8) in a weak formulation: find $s_{\alpha,n} \in V_n$ such that for all $s \in V_n$

$$\langle \alpha s_{\alpha,n} + A^*As_{\alpha,n}, s\rangle = \langle A^*r, s\rangle. \tag{4.2.9}$$

Let $\{\Phi_i\}_{i=1}^n$ be some basis of V_n. Then

$$s_{\alpha,n} = \sum_{i=1}^{n} \beta_i \Phi_i,$$

and we have the following system of linear algebraic equations for the vector $\beta = \{\beta_i\}$ of the coefficients:

$$(M + \alpha G_\Phi)\beta = F, \tag{4.2.10}$$

where

$$\begin{aligned} G_\Phi &= \{\langle \Phi_i, \Phi_j\rangle\}_{i,j=1}^n, \\ M &= \{\langle A\Phi_i, A\Phi_j\rangle\}_{i,j=1}^n, \\ F &= \{\langle A\Phi_i, r\rangle\}_{i=1}^n. \end{aligned} \tag{4.2.11}$$

We would like to note that the adjoint operator A^*, which sometimes has a rather complicated implicit definition, is not involved in the construction of $s_{\alpha,n}$. As to $A\Phi_i$, they can theoretically be either computed exactly or precomputed numerically in advance for any given basis Φ_i. Observe also that we do not need

each element $A\Phi_i$ in an explicit form, but only their inner products as in M and F, which can be computed much more accurately than $A\Phi_i$ itself. In any way, the computation error in M and F can be made much smaller than error levels ε_1 and ε_2.

4.2.3 Operator monotone index functions

Recall the properties of the function $g_\alpha = 1/(\alpha + \lambda)$ associated with Tikhonov regularization. It is well known that

$$\sup_{\lambda>0} \sqrt{\lambda}|g_\alpha(\lambda)| \leq \frac{1}{2\sqrt{\alpha}}, \tag{4.2.12}$$

and

$$\sup_{\lambda>0} \lambda^p|1 - \lambda g_\alpha(\lambda)| \leq \alpha^p \tag{4.2.13}$$

holds only for $0 \leq p \leq 1$.

To proceed further we should specify the assumptions concerning index function ϕ. From [37] it follows that when dealing with the discretized Tikhonov scheme, it is convenient to assume that the smoothness index function ϕ is operator monotone (increasing)(see the definition below), because this assumption covers all types of smoothness studied so far in the theory of Tikhonov method.

Definition 4.2.1. *A function ϕ is operator monotone on $(0, a)$, if for any pair of self-adjoint operators U, V with spectra in $(0, a)$, such that $U \leq V$, we have $\phi(U) \leq \phi(V)$ (i.e. $\forall f \in X$, $\langle \phi(U)f, f\rangle_X \leq \langle \phi(V)f, f\rangle_X$).*

It follows from Löwner Theorem (see, e.g. [20, Section 2]) that each operator monotone function ϕ with $\phi(0) = 0$ admits an integral representation as a Pick function

$$\phi(\zeta) = \tilde{\alpha}\zeta + \int \left[\frac{1}{\lambda - \zeta} - \frac{\lambda}{\lambda^2 + 1}\right] \mu(d\lambda),$$

for some $\tilde{\alpha} \geq 0$ and finite positive measure μ on $\mathbb{R}$, satisfying $\int(\lambda^2+1)^{-2}\mu(d\lambda) < \infty$. This can be considered as a kind of criterion concerning operator monotony. A more applicable version is also given in [20], and tells us that on any interval $(0, a)$, a monotone function is operator monotone if its analytic continuation in the corresponding strip of upper half-plane has imaginary part which is always positive. We refer the detailed analysis on this concept to [20, 38].

Proposition 4.2.2. *If $\phi : [0, a) \to \mathbb{R}^+ \cup \{0\}$ is operator monotone on $(0, a)$ and $\phi(0) = 0$, then*

$$\sup_{0\leq\lambda\leq b} |1 - \lambda g_\alpha(\lambda)|\phi(\lambda) \leq c\phi(\alpha), \quad \alpha \leq b, \tag{4.2.14}$$

where the numbers a and b are defined as before ($b = \|\bar{A}^*\bar{A}\| < a$) and the constant c does not depend on α.

Proof. Considering the function $\lambda \mapsto |1 - \lambda g_\alpha(\lambda)|\phi(\lambda)$, we need to show that on $[0, b]$ it is uniformly bounded by the right-hand side in (4.2.14) for any value of α. We distinguish two cases.

First if $\lambda \le \alpha$, then the required bound follows from monotonicity of ϕ, since $\sup_{\lambda>0} |1 - \lambda g_\alpha(\lambda)| \le 1$.

Consider the case $\alpha \le \lambda \le b$. Since ϕ is operator monotone on $(0, a) \supset (0, b]$ and $\phi(0) = 0$, then as in [38], such ϕ can be represented as a sum of two non-negative functions $\phi = \phi_0 + \phi_1$, where ϕ_0 is a concave function, ϕ_1 meets Lipschitz condition with Lipschitz constant c_1, and $\phi_0(0) = \phi_1(0) = 0$. Then $\phi_0(b)/b \le \phi_0(\lambda)/\lambda \le \phi_0(\alpha)/\alpha$ whenever $\alpha \le \lambda \le b$. Thus, we have

$$\frac{\phi(\lambda)}{\lambda} = \frac{(\phi_0(\lambda) + \phi_1(\lambda))}{\lambda} \le \frac{\phi_0(\lambda)}{\lambda} + c_1.$$

Now, put $c := (c_1 b/\phi_0(b) + 1)$, we conclude that for $\alpha \le \lambda \le b$

$$\frac{\phi(\lambda)}{\lambda} \le (c_1 \frac{b}{\phi_0(b)} + 1)\frac{\phi_0(\alpha)}{\alpha} \le c\frac{\phi(\alpha)}{\alpha} \tag{4.2.15}$$

Then

$$|1 - \lambda g_\alpha(\lambda)|\phi(\lambda) = |1 - \lambda g_\alpha(\lambda)|\lambda\frac{\phi(\lambda)}{\lambda}$$
$$\le \alpha \sup_{\alpha\le\lambda\le b} \frac{\phi(\lambda)}{\lambda} \le \alpha c\frac{\phi(\alpha)}{\alpha} = c\phi(\alpha).$$

The last assertion is based on (4.2.13) with $p = 1$ and (4.2.15). □

Proposition 4.2.3 ([38])**.** *Let $\phi : [0, a) \to \mathbb{R}^+ \cup \{0\}$ be operator monotone on $(0, a)$, satisfying $\phi(0) = 0$. For each $0 < c < a$ there is a constant C, such that for any pair of non-negative self-adjoint operators U, V with $\|U\|, \|V\| \le c$ we have*

$$\|\phi(U) - \phi(V)\| \le \phi(\|U - V\|) + C\|U - V\|, \tag{4.2.16}$$

where the constant C depends only on ϕ.

Proposition 4.2.4. *Under the conditions of the propositions above, for any pair of self-adjoint operators U, V with spectra on $[0, b]$*

$$\|\phi(U) - \phi(V)\| \le d\phi(\|U - V\|), \tag{4.2.17}$$

where d depends only on ϕ.

Proof. As above, we make use of decomposition $\phi = \phi_0 + \phi_1$. From the concavity of ϕ_0, it follows that for any given constant C, there exists another constant $C' = bC/\phi_0(b)$ such that for any $t \in [0, b]$, $Ct \le C'\phi_0(t)$. Thus,

$$C\|U-V\| \le C'\phi_0(\|U-V\|) \le C'\phi(\|U-V\|). \tag{4.2.18}$$

Now (4.2.16) and (4.2.18) lead to (4.2.17) with $d = 1 + C'$. □

In the sequel, we will assume that index function ϕ is operator monotone on $[0, b]$, $b = \|\bar{A}\|^2$. We define the following function class

$$\mathcal{F} := \{\phi, \phi : (0, b] \to \mathbb{R}_+, \phi(0) = 0, \phi \text{ is operator monotone}\}.$$

Then as [36] we assume more specifically, either $\phi^2(\lambda)$ to be concave, or $\phi(\lambda) \le c\sqrt{\lambda}$. The classes of such operator monotone functions will be denoted by $\mathcal{F}_0$ and $\mathcal{F}_{1/2}$ respectively. Observe that up to a certain extent these classes complement each other, because for any $\phi \in \mathcal{F}_0$, $\phi(0) = 0$, $\phi^2(\lambda) \ge \phi^2(b)\lambda/b = c\lambda$, and thus $\phi(\lambda) \ge c\sqrt{\lambda}$. Note that index functions

$$\phi(\lambda) = \lambda^\mu, \lambda > 0, \text{ for } 0 < \mu \le 1,$$
$$\phi(\lambda) = \log^{-p}(1/\lambda), 0 < \lambda < 1, \text{ for } p > 0,$$

traditionally used in the regularization theory [12] are contained in $\mathcal{F}_0 \cup \mathcal{F}_{1/2}$.

For the sake of simplicity we normalize index functions ϕ in such a way that $\phi(b) = \sqrt{b}$. Namely,

$$\mathcal{F}_0 := \left\{\phi \in \mathcal{F}, \phi(b) = \sqrt{b}, \phi^2 \text{ is concave}\right\},$$
$$\mathcal{F}_{1/2} := \left\{\phi \in \mathcal{F}, \phi(b) = \sqrt{b}, \phi(\lambda) \le \sqrt{\lambda}\right\}.$$

4.2.4 Estimation of the accuracy

The following proposition was proven in [36].

Proposition 4.2.5. *Let $\phi(\lambda)$ be any increasing index function from $\mathcal{F}_0 \cup \mathcal{F}_{1/2}$. Then for the orthogonal projector P_n from X onto n-dimensional subspace $V_n \subset X$,*

$$\|P_n\phi(\bar{A}^*\bar{A})P_n - \phi(P_n\bar{A}^*\bar{A}P_n)\| \le d_1\phi(\|\bar{A}(I-P_n)\|^2), \tag{4.2.19}$$

where the constant d_1 depends only on ϕ. Moreover, for $s = \phi(\bar{A}^*\bar{A})\nu, \|\nu\| \le R$,

$$\|(I-P_n)s\| \le \begin{cases} R\phi(\|\bar{A}(I-P_n)\|^2), & \phi \in \mathcal{F}_0; \\ R\|\bar{A}(I-P_n)\|, & \phi \in \mathcal{F}_{1/2}. \end{cases} \tag{4.2.20}$$

Now we arrive at the main result of this chapter.

Theorem 4.2.6. *Assume that the Moore-Penrose generalized solution $\bar{s}$ of the equation (4.2.1) belongs to the set $\bar{A}_{\phi,R}$ given by (4.2.6) with $\phi \in \mathcal{F}_0 \cup \mathcal{F}_{1/2}$. If A, r and n are such that $\|\bar{A} - A\| \le \varepsilon_1$, $\|\bar{r} - r\| \le \varepsilon_2$ and*

$$\|\bar{A}(I-P_n)\| \le \min\left\{\sqrt{\alpha}, \frac{\varepsilon}{\sqrt{\alpha}}\right\},\ \varepsilon = \max\{\varepsilon_1, \varepsilon_2\}, \tag{4.2.21}$$

then the following error bound holds true

$$\|\bar{s} - s_{\alpha,n}\| \le c_1\phi(\alpha) + c_2\phi(\varepsilon) + c_R\frac{\varepsilon}{\sqrt{\alpha}}, \tag{4.2.22}$$

where $c_R = (R\sqrt{b}+3)/2$, *and the constants* c_1, c_2 *do not depend on* α, n *and* ε.

Proof. Note that

$$\begin{aligned}&\|\bar{s} - s_{\alpha,n}\| = \|\bar{s} - g_\alpha(B^*B)B^*r\|\\ &\le \|\bar{s} - g_\alpha(B^*B)B^*B\bar{s}\| + \|g_\alpha(B^*B)B^*(B\bar{s} - r)\|.\end{aligned}$$

Moveover,

$$\begin{aligned}&\|B\bar{s} - r\|\\ &\le \|AP_n\bar{s} - \bar{A}P_n\bar{s}\| + \|\bar{A}P_n\bar{s} - \bar{A}\bar{s}\| + \|\bar{A}\bar{s} - r\|\\ &\le \|\bar{A} - A\|\|\bar{s}\| + \|\bar{r} - r\| + \|\bar{A}(I-P_n)\|\|(I-P_n)\bar{s}\|\\ &\le \sqrt{b}R\varepsilon_1 + \varepsilon_2 + \|\bar{A}(I-P_n)\|\|(I-P_n)\bar{s}\|,\end{aligned}$$

since $\bar{s} \in \bar{A}_{\phi,R}$ implies that $\|\bar{s}\| \le \phi(b)R = \sqrt{b}R$. Then (4.2.12) and Proposition 4.2.5 give us

$$\begin{aligned}&\|g_\alpha(B^*B)B^*(B\bar{s} - r)\|\\ &\le \begin{cases}\dfrac{1}{2\sqrt{\alpha}}\left((\sqrt{b}R+1)\varepsilon + R\|\bar{A}(I-P_n)\|\phi(\|\bar{A}(I-P_n)\|^2)\right), & \phi \in \mathcal{F}_0,\\ \dfrac{1}{2\sqrt{\alpha}}\left((\sqrt{b}R+1)\varepsilon + R\|\bar{A}(I-P_n)\|^2\right), & \phi \in \mathcal{F}_{1/2}.\end{cases}\end{aligned}$$

Keeping in mind (4.2.14), we can continue

$$\begin{aligned}&\|\bar{s} - g_\alpha(B^*B)B^*B\bar{s}\|\\ &\le \|(I-P_n)\bar{s}\| + \|(I - g_\alpha(B^*B)B^*B)\phi(B^*B)\nu\|\\ &+ \|(I - g_\alpha(B^*B)B^*B)(P_n\phi(\bar{A}^*\bar{A}) - \phi(B^*B))\nu\|\\ &\le Rc\phi(\alpha) + \|(P_n\phi(\bar{A}^*\bar{A}) - \phi(B^*B))\nu\| + \|(I-P_n)\bar{s}\|.\end{aligned}$$

The last term has been estimated in Proposition 4.2.5, and we proceed with the remainder as follows:

$$\begin{aligned}&\|(P_n\phi(\bar{A}^*\bar{A}) - \phi(B^*B))\nu\|\\ &\le R\|(I-P_n)\phi(\bar{A}^*\bar{A})\| + R\|P_n\phi(\bar{A}^*\bar{A})P_n - \phi(P_n\bar{A}^*\bar{A}P_n)\|\\ &+ R\|\phi(P_n\bar{A}^*\bar{A}P_n) - \phi(P_nA^*AP_n)\|.\end{aligned}$$

The first two terms here have been also estimated in Proposition 4.2.5, and to estimate the last one we use the property (4.2.17).

$$\begin{aligned}&\|\phi(P_n\bar{A}^*\bar{A}P_n) - \phi(P_nA^*AP_n)\|\\ &\le \phi(\|P_n\bar{A}^*\bar{A}P_n - P_nA^*AP_n\|).\\ &\le \phi(d_2\|\bar{A}-A\|) \le ([d_2]+1)\phi(\varepsilon),\end{aligned}$$

where d_2 is a positive constant, and $[\cdot]$ denotes the integer part of a positive number.

Summing up we obtain the following inequalities

$$\begin{aligned}&\|\bar{s} - s_{\alpha,n}\|\\ &\le Rc\phi(\alpha) + R\phi(\|\bar{A}(I-P_n)\|^2) + d_1R\phi(\|\bar{A}(I-P_n)\|^2) + R([d_2]+1)\phi(\varepsilon)\\ &+ \frac{1}{2\sqrt{\alpha}}\left((\sqrt{b}R+1)\varepsilon + R\|\bar{A}(I-P_n)\|\phi(\|\bar{A}(I-P_n)\|^2)\right), \quad \text{if } \phi\in\mathcal{F}_0,\end{aligned}$$

$$\begin{aligned}&\|\bar{s} - s_{\alpha,n}\|\\ &\le Rc\phi(\alpha) + R\|\bar{A}(I-P_n)\| + d_1R\phi(\|\bar{A}(I-P_n)\|^2) + R([d_2]+1)\phi(\varepsilon)\\ &+ \frac{1}{2\sqrt{\alpha}}\left((\sqrt{b}R+1)\varepsilon + \|\bar{A}(I-P_n)\|^2\right), \quad \text{if } \phi\in\mathcal{F}_{1/2}.\end{aligned}$$

These inequalities together with (4.2.21) give us the statement (4.2.22). □

4.3 Parameter identification in elliptic equation

4.3.1 Natural linearization

Consider the boundary value problem

$$\begin{aligned}-\bigtriangledown(a\bigtriangledown u) &= f \text{ in } \Omega,\\ u &= g \text{ on } \partial\Omega.\end{aligned} \tag{4.3.1}$$

We are interested in recovering the (unknown) diffusion coefficient $a = a(x)$ from noisy measurements u^δ of the solution u. Here Ω is a convex domain with Lipschitz boundary, $f \in L^2(\Omega)$, $g \in H^{\frac{3}{2}}(\partial\Omega)$, and for some fixed noise level δ we have

$$\|u^\delta - u\|_{L^2(\Omega)} \le \delta. \tag{4.3.2}$$

Using an initial guess a_0 which is bounded away from zero, one can rewrite system (4.3.1) as follows:

$$-\bigtriangledown(a_0\bigtriangledown(u-u_0))=\bigtriangledown((a-a_0)\bigtriangledown u)\text{ in }\Omega,$$
$$u-u_0=0\text{ on }\partial\Omega, \tag{4.3.3}$$

where u_0 solves

$$-\bigtriangledown(a_0\bigtriangledown u_0)=f\text{ in }\Omega,$$
$$u_0=g\text{ on }\partial\Omega. \tag{4.3.4}$$

Then we obtain the following linear operator equation:

$$\bar{A}s=\bar{r}, \tag{4.3.5}$$

where $s=a-a_0$ is the difference between unknown parameter a and the initial guess a_0, $\bar{r}=u-u_0$, and the operator $\bar{A}$ maps any given $s\in L^2(\Omega)$ to the weak solution z of

$$-\bigtriangledown(a_0\bigtriangledown z)=\bigtriangledown(s\bigtriangledown u)\text{ in }\Omega,$$
$$z=0\text{ on }\partial\Omega. \tag{4.3.6}$$

Observe that in the implicit definition of the operator $\bar{A}$ and in the right-hand side term of (4.3.6) the unknown exact solution u is involved.

Therefore, replacing u by a smoothed version u^δ_{sm} of u^δ such that $\nabla u^\delta_{sm}\in L^\infty$, we switch to the equation

$$As=r, \tag{4.3.7}$$

with the perturbed operator $A=A(u^\delta_{sm})$ and noisy right-hand side $r=u^\delta-u_0$, where A maps s to the weak solution z of the problem

$$-\bigtriangledown(a_0\bigtriangledown z)=\bigtriangledown(s\bigtriangledown u^\delta_{sm})\text{ in }\Omega,$$
$$z=0\text{ on }\partial\Omega. \tag{4.3.8}$$

The error in the operator A and in the free term r can be in principle estimated in terms of the initial noise level δ. Then the solution of equation (4.3.5) can be recovered from its noisy counterpart (4.3.7) by means of an appropriate regularization strategy, which will be studied later in this section.

Note that in [26] noisy equation (4.3.7) was treated by some modified form of OLS method. As a result, its linearity was not fully utilized there. Particularly, a theoretical justification was done only under *a priori* assumption concerning the smoothness of unknown diffusion coefficient a.

4.3.2 Data smoothing and noise level analysis

To apply the natural linearization (NL), data mollification should be carried out in an appropriate way to make the implicitly defined operator A in (4.3.8) well-defined and to make the noise level for $\|\bar{A} - A\| \leq \varepsilon_1$ and $\|\bar{r} - r\| \leq \varepsilon_2$ be controlled in terms of initial noise level δ. Indeed, data smoothing is already part of the regularization of our ill-posed parameter identification problem, which also determines how far the perturbed equation (4.3.7) deviates from the standard equation (4.3.5).

The following condition should be satisfied.

$$\nabla u_{sm}^{\delta} \in L^{\infty}(\Omega). \tag{4.3.9}$$

Condition (4.3.9) guarantees that the weak solution of (4.3.8) is well-defined for $s \in L^2(\Omega)$. Therefore, As is also well-defined as long as

$$a_0 \in L^{\infty}(\Omega) \text{ with } a_0(x) > \underline{a} > 0 \ \ a.e.$$

Note that it is not necessary to demand the search-for parameter a to satisfy such a condition.

Once As is well defined by the procedure above, we can always seek for the solution z of (4.3.8) in $H_0^1(\Omega)$, which means that A acts from $L^2(\Omega)$ to $L^2(\Omega)$ with $\text{Range}(A) \subseteq H_0^1(\Omega)$. This leads to the compactness of the operator A, and makes (4.3.7) ill-posed.

Now we discuss the estimate $\|\bar{A} - A\| \leq \varepsilon_1$.

If $\{\lambda_k\}$ and $\{u_k\}$ are respectively eigenvalues and orthogonal eigenfunctions of the differential operator $\nabla(a_0 \nabla(\cdot))$ with zero boundary condition on $\partial\Omega$, i.e. $\nabla(a_0 \nabla u_k) = \lambda_k u_k$, $u_k = 0$ on $\partial\Omega$, then

$$(\bar{A} - A)s = \sum_k \lambda_k^{-1} u_k \langle u_k, \nabla(s \nabla(u - u_{sm}^{\delta}))\rangle,$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product in $L^2(\Omega)$. Now it is clear that ε_1 depends on the approximation of ∇u by ∇u_{sm}^{δ}. If, for example, a_0 is such that

$$c^2(a_0, \Omega) := \sum_k \lambda_k^{-2} \|\nabla u_k\|_{L^{\infty}(\Omega)}^2 < \infty, \tag{4.3.10}$$

then

$$\begin{aligned}
&\|(\bar{A}-A)s\|^2_{L^2(\Omega)}\\
&=\sum_k \lambda_k^{-2}\langle u_k, \nabla(s\nabla(u-u^\delta_{sm}))\rangle^2\\
&=\sum_k \lambda_k^{-2}\langle \nabla u_k s, \nabla(u-u^\delta_{sm})\rangle^2\\
&\le c^2(a_0,\Omega)\|s\|^2_{L^2(\Omega)}\|\nabla(u-u^\delta_{sm})\|^2_{L^2(\Omega)},
\end{aligned}$$

and

$$\varepsilon_1 \le c(a_0,\Omega)\|\nabla(u-u^\delta_{sm})\|_{L^2(\Omega)}.$$

Note that (4.3.10) holds, in particular, for $\Omega=[0,2\pi]$, $a_0\equiv 1$, $\nabla(a_0\nabla u)=u''$, because in this case $u_k(x)=\pi^{-1/2}\sin kx$, $\lambda_k=-k^2$, $\|\nabla u_k\|_{L^\infty}=\|u_k'\|_{L^\infty[0,2\pi]}=k\pi^{-1/2}$.

In [26] a smoothed approximation u^δ_{sm} has been constructed in such a way that $\|\nabla(u-u^\delta_{sm})\|_{L^2(\Omega)}\le c\sqrt{\delta}$ under the assumption that the exact solution $u\in H^2(\Omega)\cap W^{1,\infty}(\Omega)$ with triangulation mesh size $h_{sm}\sim\sqrt{\delta}$. In this case one can take $\varepsilon_1=c\sqrt{\delta}$. At the same time, if the above mentioned assumptions are not satisfied, or some other approximation u^δ_{sm} is used, then the relation between ε_1 and δ changes. Therefore, in the sequel we will assume only that ε_1 is known and it is much larger than δ, i.e, $\varepsilon_1 >> \delta$.

On the other hand, in view of the structures of $\bar{r}$ and r

$$\varepsilon_2=\|\bar{r}-r\|_{L^2(\Omega)}=\|u-u^\delta\|\le\delta.$$

As a result, if we apply the estimation of accuracy established in Theorem 4.2.6, $\varepsilon=\max\{\varepsilon_1,\varepsilon_2\}$ in (4.2.21) can be just taken as the order of $\sqrt{\delta}$.

4.3.3 Estimation of the accuracy

Proposition 4.3.1. *Let $\bar{A}$ be an operator defined by (4.3.5) and (4.3.6), where Ω is a convex domain with Lipschitz boundary, a_0 is bounded away from zero, and $\bigtriangledown a_0\in L^\infty(\Omega)$, $u\in H^2(\Omega)\cap W^{1,\infty}(\Omega)$. If P_n is the orthogonal projector from $L^2(\Omega)$ onto n-dimensional space of piece-wise linear continuous functions corresponding to triangulation of Ω with mesh size h_n, then*

$$\|\bar{A}(I-P_n)\|\le ch_n, \tag{4.3.11}$$

where the constant c does not depend on h_n.

Proof. The adjoint operator of $\bar{A}$ is given by

$$\bar{A}^*\psi=\nabla u\cdot\nabla\tilde{\psi}, \tag{4.3.12}$$

where $\tilde{\psi}$ solves boundary value problem

$$-\bigtriangledown(a_0 \bigtriangledown \tilde{\psi}) = \psi \text{ in } \Omega, \qquad \tilde{\psi} = 0 \text{ on } \partial\Omega. \tag{4.3.13}$$

For L^2-right hand side ψ in (4.3.13), in view of the conditions on a_0, $\tilde{\psi}$ belongs to $H^2(\Omega)$. Since $u \in H^2(\Omega) \cap W^{1,\infty}(\Omega)$, we have that $\bar{A}^*$ acts from $L^2(\Omega)$ to $H^1(\Omega)$, as a linear bounded operator. Approximation Theory provides us with the following Jackson type inequality:

$$\|I - P_n\|_{H^1(\Omega)\to L^2(\Omega)} \le dh_n. \tag{4.3.14}$$

Then

$$\begin{aligned} &\|\bar{A}(I - P_n)\| \\ &= \|(I - P_n)\bar{A}^*\| \\ &\le \|I - P_n\|_{H^1(\Omega)\to L^2(\Omega)} \|\bar{A}^*\|_{L^2(\Omega)\to H^1(\Omega)} \\ &\le ch_n. \square \end{aligned}$$

□

We establish the following proposition on estimation of accuracy, which can be seen as the corollary from Theorem 4.2.6 and has more instructive meaning in numerical application.

Proposition 4.3.2. *Let $\alpha \ge \varepsilon^2$, $h_n \sim \min\left\{\sqrt{\alpha}, \dfrac{\varepsilon}{\sqrt{\alpha}}\right\}$ or $h_n \sim \varepsilon$. Then under the conditions of Proposition 4.3.1 and Theorem 4.2.6 the estimation of accuracy (4.2.22) holds true, i.e.,*

$$\|\bar{s} - s_{\alpha,n}\| \le c_1\phi(\alpha) + c_2\phi(\varepsilon) + c_R\frac{\varepsilon}{\sqrt{\alpha}}.$$

Proof. From our assumption it follows that $\min\left\{\sqrt{\alpha}, \dfrac{\varepsilon}{\sqrt{\alpha}}\right\} \ge \varepsilon$. On the other hand, under the condition of Proposition 4.3.1, $\|\bar{A}(I - P_n)\| \sim h_n$, and for h_n chosen as in the statement of the proposition the assumption (4.2.21) is satisfied that gives us (4.2.22). □

Note that the assumption $\alpha \ge \varepsilon^2$ is not restrictive. It simply means that the term $\dfrac{\varepsilon}{\sqrt{\alpha}}$ from the error estimation (4.2.22) is smaller than 1, which is rather natural.

4.3.4 Balancing principle

Here we will develop a strategy for the choice of the regularization parameter α which adapts to unknown smoothness of $s = a - a_0$, i.e., the index function ϕ. The essence is derived from the general adaptive strategy from [36].

Assume that h_n is chosen as in Proposition 4.3.2 with $n = n(\alpha, \varepsilon)$. Let $s_{\alpha,\varepsilon} = s_{\alpha,n(\alpha,\varepsilon)}$.

In view of the expression of error estimation in Theorem 4.2.6 (also in Proposition 4.3.1), the optimal choice of the regularization parameter would be $\alpha = \alpha_{opt}$ for which

$$\Xi_\varepsilon(\alpha_{opt}) = c_R \frac{\varepsilon}{\sqrt{\alpha_{opt}}}, \tag{4.3.15}$$

where $\Xi_\varepsilon(\cdot) := c_1\phi(\cdot) + c_2\phi(\varepsilon)$. Due to the monotonicity of the functions $\Xi_\varepsilon(\cdot)$, $\frac{1}{\sqrt{\cdot}}$, this optimal choice is achieved at the crossing point of two monotonous curves (Fig. 4.1).

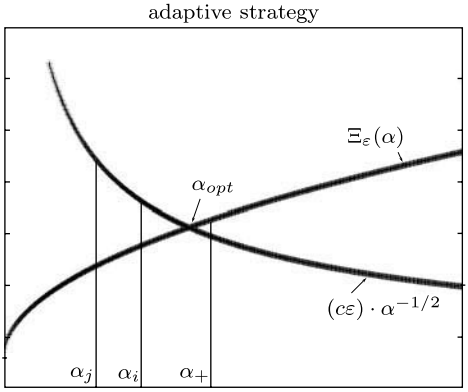


Fig. 4.1 Adaptive choice of α by the balancing principle.

Proposition 4.3.3. *Let* $\Theta_\varepsilon(\alpha) := \Xi_\varepsilon(\alpha)\sqrt{\alpha}$, *then*

$$\|\bar{s} - s_{\alpha_{opt},\varepsilon}\| \le 2\Xi_\varepsilon(\Theta_\varepsilon^{-1}(c_R\varepsilon)). \tag{4.3.16}$$

Of course, for unknown ϕ this optimal choice can not be realized in practice. Thus, an *a posteriori* adaptive strategy is needed.

To describe this strategy we introduce

$$\Delta_N := \{\alpha_k = \alpha_0 q^k, k = 0, 1, \ldots, N\}$$

with $\alpha_0 = \varepsilon^2$, $q > 1$; N is an integer number such that $\alpha_{N-1} \le b \le \alpha_N$.

Then the regularized solutions $s_{\alpha_k,\varepsilon}$ will be studied successively as long as

$$\|s_{\alpha_i,\varepsilon} - s_{\alpha_k,\varepsilon}\| \le c_R\varepsilon\left(\frac{3}{\sqrt{\alpha_i}} + \frac{1}{\sqrt{\alpha_k}}\right) \text{ for } k = 1, \cdots, i-1.$$

The procedure terminates with

$$\alpha_+ = \max\left\{\alpha_i \in \Delta_N : \|s_{\alpha_i,\varepsilon} - s_{\alpha_k,\varepsilon}\| \le c_R\varepsilon\left(\frac{3}{\sqrt{\alpha_i}} + \frac{1}{\sqrt{\alpha_k}}\right),\ \ k = 1, \cdots, i-1\right\}. \tag{4.3.17}$$

Fig. 4.1 illustrates the choice $\alpha = \alpha_+$.

Proposition 4.3.4. *Under the conditions of Proposition 4.3.2*

$$\|\bar{s} - s_{\alpha_+,\varepsilon}\| \le \frac{6\sqrt{q}c_R\varepsilon}{\sqrt{\alpha_{opt}}}.$$

Proof. Since α_{opt} may not be an element of Δ_N we introduce

$$\alpha_* = \max\{\alpha \in \Delta_N, \Theta_\varepsilon(\alpha) \le c_R\delta\}.$$

Then, if $\alpha_* = \alpha_l$ for some l, $1 < l < N$, it is easy to verify that $\alpha_l \le \alpha_{opt} \le \alpha_{l+1}$. Now we show that

$$\alpha_* \le \alpha_+ \tag{4.3.18}$$

In fact, by construction,

$$\begin{aligned}
&\|s_{\alpha_*,\varepsilon} - s_{\alpha_k,\varepsilon}\| \\
&\le \|s - s_{\alpha_*,\varepsilon}\| + \|s - s_{\alpha_k,\varepsilon}\| \\
&\le \Xi_\varepsilon(\alpha_l) + c_R\frac{\varepsilon}{\sqrt{\alpha_l}} + \Xi_\varepsilon(\alpha_k) + c_R\frac{\varepsilon}{\sqrt{\alpha_k}} \\
&\le c_R\varepsilon\left(\frac{3}{\sqrt{\alpha_l}} + \frac{1}{\sqrt{\alpha_k}}\right)
\end{aligned}$$

for $k = 1, \cdots, l-1$. By the definition of α_+, (4.3.18) holds true.

Meanwhile, if $\alpha_+ = \alpha_m$ for some $m \ge l$, then

$$\begin{aligned}
&\|\bar{s} - s_{\alpha_+,\varepsilon}\| \\
&\le \|s - s_{\alpha_*,\varepsilon}\| + \|s_{\alpha_m,\varepsilon} - s_{\alpha_l,\varepsilon}\| \\
&\le \Xi_\varepsilon(\alpha_*) + c_R\frac{\varepsilon}{\sqrt{\alpha_*}} + 3c_R\frac{\varepsilon}{\sqrt{\alpha_m}} + c_R\frac{\varepsilon}{\sqrt{\alpha_l}} \\
&\le 6c_R\frac{\varepsilon}{\sqrt{\alpha_*}} \le 6c_R\frac{\varepsilon}{\sqrt{q^{-1}\alpha_{opt}}}.
\end{aligned}$$

The proof is complete. □

Remark 4.3.5. *An error bound given by Proposition 4.3.4 can be represented in the form*

$$\|\bar{s} - s_{\alpha_+,\varepsilon}\| \le 6\sqrt{q}\Xi_\varepsilon(\Theta_\varepsilon^{-1}(c_R\varepsilon)). \tag{4.3.19}$$

It means that the adaptive regularization parameter choice strategy (4.3.17) provides us with an error bound which is worse only by a constant factor $3\sqrt{q}$ than a benchmark (4.3.16).

Theorem 4.3.6. *Under the conditions of Theorem 4.2.6 and Proposition 4.3.2 for α_+ chosen as in (4.3.19), we have*

$$\|\bar{s} - s_{\alpha_+,\varepsilon}\| \le C(\phi(\Theta^{-1}(\varepsilon)) + \phi(\varepsilon)), \tag{4.3.20}$$

where the constant C does not depend on ε.

Proof. In [35] (see Lemma 3), it has been proven that for any operator monotone index function ϕ and $d > 0$, there is a constant c_d depending only on ϕ and d such that

$$\phi(d\lambda) \le c_d\phi(\lambda). \tag{4.3.21}$$

Note also that

$$\Theta^{-1}(d\lambda) \le \max\{d^2, 1\}\Theta^{-1}(\lambda). \tag{4.3.22}$$

In fact, for $d \in (0, 1)$ this inequality is obvious. If $d > 1$ then $\Theta^{-1}(d\lambda) \ge \Theta^{-1}(\lambda)$, and $\phi(\Theta^{-1}(d\lambda)) \ge \phi(\Theta^{-1}(\lambda))$. Therefore,

$$d = \frac{\Theta(\Theta^{-1}(d\lambda))}{\Theta(\Theta^{-1}(\lambda))} = \frac{\phi(\Theta^{-1}(d\lambda))\sqrt{\Theta^{-1}(d\lambda)}}{\phi(\Theta^{-1}(\lambda))\sqrt{\Theta^{-1}(\lambda)}} \ge \sqrt{\frac{\Theta^{-1}(d\lambda)}{\Theta^{-1}(\lambda)}},$$

which is equivalent to (4.3.22).

Observe now that for any $\lambda \in [0, b]$

$$\Theta_\varepsilon(\lambda) = (c_1\phi(\lambda) + c_2\phi(\varepsilon))\sqrt{\lambda} \ge c_1\phi(\lambda)\sqrt{\lambda} = c_1\Theta(\lambda).$$

Then

$$\Theta(\Theta_\varepsilon^{-1}(c_R\varepsilon)) \le c_1^{-1}\Theta_\varepsilon(\Theta_\varepsilon^{-1}(c_R\varepsilon)) = \frac{c_R}{c_1}\varepsilon,$$

and from (4.3.22) we have

$$\Theta_\varepsilon^{-1}(c_R\varepsilon) \le \Theta^{-1}(\frac{c_R}{c_1}\varepsilon) \le \max\left\{\left(\frac{c_R}{c_1}\right)^2, 1\right\}\Theta^{-1}(\varepsilon) = \bar{c}\Theta^{-1}(\varepsilon). \tag{4.3.23}$$

From (4.3.17), (4.3.21) and (4.3.23) we finally obtain

$$\begin{aligned}
&\|s - s_{\alpha_+,\varepsilon}\| \\
&\le 6\sqrt{q}(c_1\phi(\Theta_\varepsilon^{-1}(c_R\varepsilon)) + c_2\phi(\varepsilon)) \\
&\le 6\sqrt{q}(c_1\phi(\bar{c}\Theta^{-1}(\varepsilon)) + c_2\phi(\varepsilon)) \\
&\le C(\phi(\Theta^{-1}(\varepsilon)) + \phi(\varepsilon)).
\end{aligned}$$

□

The following corollaries specify the estimation of the accuracy in concrete cases.

Corollary 4.3.7. *If $\phi \in \mathcal{F}_{1/2}$, then*

$$\|\bar{s} - s_{\alpha_+,\varepsilon}\| \le C\phi(\Theta^{-1}(\varepsilon)). \tag{4.3.24}$$

Proof. For $\phi \in \mathcal{F}_{1/2}$, $\phi(\varepsilon) \le \sqrt{\varepsilon}$. Thus, $\phi(\varepsilon)\sqrt{\varepsilon} \le \varepsilon$, which means $\varepsilon \le \Theta^{-1}(\varepsilon)$, then (4.3.20) is reduced to (4.3.24). □

Corollary 4.3.8. *If index function $\phi \in \mathcal{F}_0$, then*

$$\|\bar{s} - s_{\alpha_+,\varepsilon}\| \le C\phi(\varepsilon). \tag{4.3.25}$$

At the same time, for $\phi(\lambda) = c\lambda^\mu$, $0 \le \mu < 1/2$, (4.3.24) holds true as well. In this case, $\|\bar{s} - s_{\alpha_+,\varepsilon}\| \le C\varepsilon^{\frac{2\mu}{2\mu+1}}$.

Proof. We prove only the last statement. It is well known (see, e.g. [49], p 93) that for $\phi(\lambda) = c\lambda^\mu$, $0 \le \mu < 1/2$,

$$\|\phi(P_n\bar{A}^*\bar{A}P_n) - \phi(P_nA^*AP_n)\| = c\||\bar{A}P_n)|^{2\mu} - |AP_n|^{2\mu}\| \le C\|(\bar{A} - A)\|^{2\mu},$$

where $|F| = (F^*F)^{1/2}$. Then $\phi(\varepsilon)$ appearing in (4.2.22) and (4.3.20) will be replaced by $\phi(\varepsilon^2) = c\varepsilon^{2\mu}$. Therefore, $\phi(\varepsilon^2) \le \phi(\Theta^{-1}(\varepsilon))$, and (4.3.24) holds true. □

Direct calculations show that the following statement is also true.

Corollary 4.3.9. *If $\phi(\lambda) = c\log^{-p}(1/\lambda)$, $c, p > 0$, then (4.3.24) holds true.*

Remark 4.3.10. *It follows from Theorem 2.3 of [49], on page 99, that under the assumption that $\|\bar{A} - A\| \le \varepsilon$ and $\bar{s} \in \bar{A}_{\phi,R}$, $\phi(\lambda) = \lambda^\mu, \mu > 0$, one can construct a regularized approximation s^ε such that*

$$\|\bar{s} - s^\varepsilon\| \le c\varepsilon^{\frac{2\mu}{2\mu+1}}.$$

This is in agreement with (4.3.20), because for $\phi(\lambda) = \lambda^\mu$, we have $\phi(\Theta^{-1}(\lambda)) = \lambda^{\frac{2\mu}{2\mu+1}}$. Thus, Corollary 4.3.7 can be seen as an extension of Theorem 2.3 of [49] to the case of arbitrary $\phi \in \mathcal{F}_{1/2}$.

4.3.5 Numerical examples

Two numerical examples are provided here to support and verify the theoretical results. We use MATLAB-code in one-dimensional case, where $\Omega = [0, 1]$. As in [25], for such Ω, the situation described in Proposition 4.3.1 is simplified, and the

estimation for $\bar{A}(I - P_n)$ is still valid. At first we take the same example as in [25].

Example 4.3.11. *Consider*

$$a(x) = \begin{cases} 1 + \frac{1}{3}\sin^2(\pi\frac{x-0.5}{0.2}), & x \in [0.3, 0.7], \\ 1, & \textit{else.} \end{cases}$$

$$u(x) = \begin{cases} \frac{x}{1-0.2(2-\sqrt{3})}, & x \in [0, 0.3], \\ \frac{\left(0.3 + \frac{0.2\sqrt{3}}{2\pi}(\arctan(\sqrt{3}\tan(\frac{\pi}{2}\frac{x-0.5}{0.2})) \right.}{1-0.2(2-\sqrt{3})} & \\ \left. + \arctan(\frac{1}{\sqrt{3}}\tan(\frac{\pi}{2}\frac{x-0.5}{0.2})) + \pi\right) & x \in [0.3, 0.7], \\ \frac{x-0.2(2-\sqrt{3})}{1-0.2(2-\sqrt{3})} & x \in [0.7, 1], \end{cases}$$

satisfying the following one-dimensional problem of the form (4.3.1):

$$\begin{aligned} &-(au_x)_x = 0 \ \textit{in}\ (0,1), \\ &u(0) = 0, u(1) = 1. \end{aligned} \tag{4.3.26}$$

For the implementation of natural liberalization, we fix initial guess $a_0 \equiv 1$, which implies $u_0(x) = x$. Such choice of a_0 allows a lot of convenience in computation due to the definition of the operator A.

Here we take the data noise level $\delta = 0.001$, $u^\delta = u + \delta\xi$, where ξ is random variable with uniform distribution in the interval $[-1, 1]$. The data mollification is done by piece-wise linear interpolation. Then in such one-dimensional case (4.3.10) is satisfied. Thus, we have the noise level $\varepsilon \sim \sqrt{\delta}$. The number of piece-wise linear basis elements for projection is $n = 50$. The components in (4.2.11) are computed using MATLAB-code for numerical integration. To obtain the functions $A\Phi_i$, we need to solve a finite number of mixed boundary value problem of the form of (4.3.8), where s is substituted for Φ_i. As long as these components in (4.2.11) are precomputed, for each α, to obtain a regularized solution we simply need to solve the algebraic system (4.2.10).

Adaptive strategy helps us to give the proper choice of the regularization parameter. We also describe the procedure here.

Step 1 Choose α_0 small enough and set $k = 0$.
Step 2 Calculate $s_{\alpha_k,\varepsilon}$ by $s_{\alpha_k,\varepsilon} = (\alpha_k I + P_n A^* A P_n)^{-1} P_n A^* r$.
Step 3 If $\|s_{\alpha_k,\varepsilon} - s_{\alpha_j,\varepsilon}\| \le c_R \varepsilon \left(\frac{3}{\sqrt{\alpha_k}} + \frac{1}{\sqrt{\alpha_j}}\right)$ for $j = 1, \cdots, k-1$, (for $k = 0$, let this hold true) then set $\alpha_k = \alpha_{k+1} = q\alpha_k$ and go to 2; else $\Rightarrow$ break & print $\hat{\alpha} = \alpha_k$.
Step 4 The final parameter $\alpha_+ = \alpha_{k-1}$.

In the considered example, the final regularization parameter $\alpha = 0.00013$ is produced by adaptive procedure described above, where we take $\alpha_0 = 0.00008$, $q = 1.1$ and $N = 26$.

In Fig. 4.2 (b), we enlarge the solution function $u(x)$ by factor 10. In this case relative error becomes smaller, and the adaptive procedure selects $\alpha = 0.00025$.

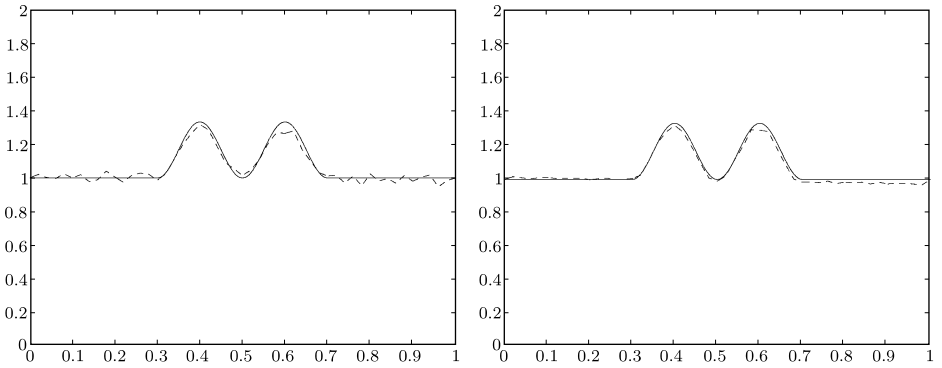


Fig. 4.2 (a) Regularized approximation with $\alpha = 0.00013$ (dash line) and exact parameter (solid line); (b) Regularized approximation with $\alpha = 0.00025$ (dash line) and exact parameter (solid line).

Example 4.3.12. *Consider the problem (4.3.1) with*

$$a(x) = (2x-1)^{\frac{2}{5}},$$
$$u(x) = \frac{1}{2} + \frac{1}{2}(2x-1)^{\frac{3}{5}}.$$

Fig. 4.3 (a) and (b) shows the results of application of the adaptive procedure (4.3.17) with the same parameters as in Example 4.3.11. Fig. 4.3 (b) is again obtained by enlarging the exact solution $u(x)$ by factor 10.

It is worthy to note that in this example the exact coefficient a has a zero point $x = \frac{1}{2}$. It shows that our approach can work without the additional assumption that $a(x)$ is bounded away from zero.

4.4 Parameter identification in parabolic equation

Consider the identification of a diffusion coefficient $a(\tau)$ in a quasi-linear parabolic system of the form

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla \cdot (a(u)\nabla u) &= G(t,x) \text{ in } [0,T] \times \Omega, \\ a(u)\frac{\partial u}{\partial n} &= g(t,x) \text{ on } [0,T] \times \partial\Omega, \\ u(0,x) &= u_{in}(x) \text{ in } \Omega, \end{aligned} \tag{4.4.1}$$

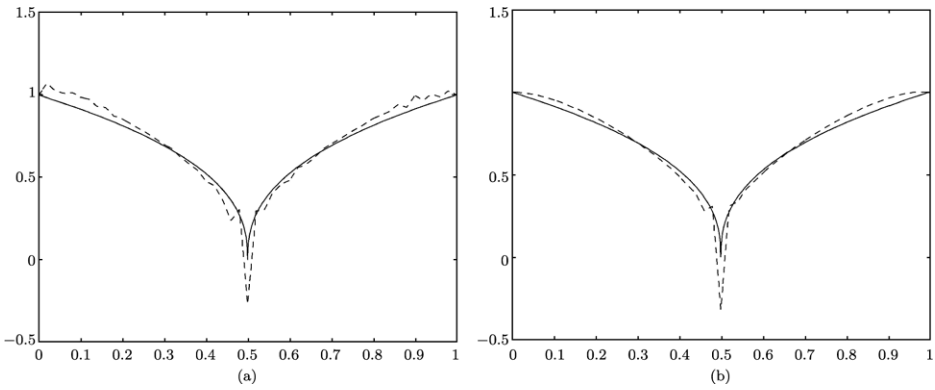


Fig. 4.3 (a) Regularized approximation with $\alpha = 0.0011$ (dash line) and exact parameter (solid line); (b) Regularized approximation with $\alpha = 0.21$ (dash line) and exact parameter (solid line).

from noisy observation $u^\delta(t,x)$ of the solution $u(t,x)$ of (4.4.1) given for a very short time period $[T-\sigma, T]$ such that

$$u^\delta(t,x) = u(t,x) + \delta\xi(t,x), \quad (t,x) \in [T-\sigma, T] \times \Omega, \tag{4.4.2}$$

where δ is the noise level, and ξ is a normalized observation noise, which is assumed to be a square summable function.

Of course, using (4.4.2), one has to be aware of the fact that a coefficient a is identifiable only on the range of the values $u(t,x)$ for $(t,x) \in [T-\sigma, T] \times \Omega$. Unless we have *a priori* information about this range we can only use noisy measurements $u^\delta(t,x)$ and construct an interval $[u_1, u_2]$ large enough such that

$$u_1 \le u^\delta(t,x), u(t,x) \le u_2, \quad (t,x) \in [T-\sigma, T] \times \Omega. \tag{4.4.3}$$

The natural linearization (NL) process for the identification of a coefficient $a = a(u)$ in (4.4.1) from data (4.4.2) can be performed in two steps:

1. NL for recovering $b(x) = a(u(T,x))$.
2. Regularized approximation of a diffusion coefficient $a(u)$ as a function of one variable.

4.4.1 Natural linearization for recovering $b(x) = a(u(T,x))$

The first step is similar to the linearization of the identification of a diffusion coefficient in (4.3.1). Note that the function $v(x) = u(T,x)$, $x \in \Omega$, describing a terminal status of a system (4.4.1), solves the following elliptic boundary value problem

$$\begin{aligned} -\nabla \cdot (b(x)\nabla v(x)) &= G_T(x), \quad x \in \Omega, \\ v(x) &= u(T,x), \quad x \in \partial\Omega, \end{aligned} \tag{4.4.4}$$

where $b(x) = b(T,x) = a(u(T,x))$, $G_T(x) = G(T,x) - \dfrac{\partial u(T,x)}{\partial t}$.

We use initial guess b_0 and rewrite (4.4.4) with $v = u(T,\cdot)$ as follows

$$\begin{aligned} -\nabla \cdot (b_0 \nabla (u(T,\cdot) - v_0)) &= \nabla \cdot ((b - b_0)\nabla u(T,\cdot)) \text{ in } \Omega, \\ u(T,\cdot) - v_0 &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{4.4.5}$$

where v_0 solves the boundary value problem (4.4.4) with b substituted for b_0

$$\begin{aligned} -\nabla \cdot (b_0(x)\nabla v(x)) &= G_T(x), \quad x \in \Omega, \\ v(x) &= u(T,x), \quad x \in \partial\Omega. \end{aligned}$$

Then for $b_0 \in L^\infty(\Omega)$ bounded away from zero we consider the linear operator $\bar{A}$ mapping s to the weak solution z of the boundary value problem

$$\begin{aligned} -\nabla \cdot (b_0 \nabla z) &= \nabla \cdot (s \nabla u(T,\cdot)) \text{ in } \Omega, \\ z &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{4.4.6}$$

Similar to the NL for elliptic problem, under a mild assumption on the exact solution u, this operator is well-defined for any $s \in L^2(\Omega)$.

In view of (4.4.5) the difference $\bar{s} = b - b_0$ between the searched-for parameter b and the initial guess b_0 can be seen as the unique solution of the linear operator equation (4.3.5) with $\bar{r} = u(T,\cdot) - v_0$.

As before, in (4.4.6) we replace $u(T,\cdot)$ by a smoothed version $u^\delta_{sm}(T,\cdot)$ of $u^\delta(T,\cdot)$. Then we consider the operator A mapping s to the solution of the boundary value problem

$$\begin{aligned} -\nabla \cdot (b_0 \nabla z) &= \nabla \cdot (s \nabla u^\delta_{sm}(T,\cdot)) \text{ in } \Omega, \\ z &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{4.4.7}$$

and the perturbation equation

$$As = \tilde{r} \tag{4.4.8}$$

with $\tilde{r}(x) = u^\delta(T,x) - v_0(x)$.

However, $\tilde{r}(x) = u^\delta(T,x) - v_0$ is not available for us, because v_0 depends on the exact data $u(T,\cdot)$. Moreover, since only the values of $u(t,x)$ (but not derivative) are observable, the exact source term $G(T,x) - \frac{\partial u(T,x)}{\partial t}$ is also not available. The latter difficulty can be resolved within the framework of some numerical differentiation procedure applied to noisy data (4.4.2) and producing some $Du^\delta(T,\cdot)$ as an approximation for $\frac{\partial u(T,\cdot)}{\partial t}$ such that

$$\left\| \frac{\partial u(T,\cdot)}{\partial t} - Du^\delta(T,\cdot) \right\|_{L^2} \le \varepsilon.$$

For example, from [15], we know that for $u \in C^2([T-\sigma, T]) \otimes H^2(\Omega)$ one can put $Du^\delta(T,\cdot) = [u^\delta(T,\cdot) - u^\delta(T-h,\cdot)]/h$, $h = \sqrt{\delta} < \sigma$. In this case $\varepsilon = c\sqrt{\delta}$.

Getting $Du^\delta(T,\cdot)$, we can obtain an approximation v_0^δ for v_0 as the solution of the boundary value problem

$$\begin{aligned} -\nabla \cdot (b_0(x)\nabla v(x)) &= G_T^\delta(x), \quad x \in \Omega, \\ v(x) &= u^\delta(T,x), \quad x \in \partial\Omega, \end{aligned}$$

with $G_T^\delta(\cdot) = G(T,x) - Du^\delta(T,\cdot)$.

With this approximation in hand we substitute the right-hand side $\tilde{r}$ of (4.4.8) for its numerically feasible counterpart $r(x) = u^\delta(T,x) - v_0^\delta(x)$ that gives us an analog of equation (4.3.7).

Following the data mollification described in [26] one can smooth the data (4.4.2) in such a way that for $u(T,\cdot) \in H^2(\Omega) \cap W^{1,\infty}(\Omega)$

$$\begin{aligned} \|u_{sm}^\delta(T,\cdot) - u(T,\cdot)\|_{L^2(\Omega)} &\le c\delta, \\ \|\nabla(u_{sm}^\delta(T,\cdot) - u(T,\cdot))\|_{L^2(\Omega)} &\le c\sqrt{\delta}, \\ \|\nabla u_{sm}^\delta(T,\cdot)\|_{L^\infty(\Omega)} &\le c. \end{aligned} \tag{4.4.9}$$

Therefore, $\varepsilon_1 = \|\bar{A} - A\|$ has the order of $\sqrt{\delta}$.

As for the error between $\bar{r} = u(T,\cdot) - v_0$ and $r = u^\delta(T,\cdot) - v_0^\delta$, we can estimate it as

$$\|\bar{r} - r\| \le \|u(T,\cdot) - u^\delta(T,\cdot)\| + \|v_0 - v_0^\delta\|.$$

Since $\|u(T,\cdot) - u^\delta(T,\cdot)\| \le \delta$ is known (initial noise level), now we concentrate on the estimation of $\|v_0 - v_0^\delta\|$. Recalling the definitions of v_0 and v_0^δ, we know that $v_0 - v_0^\delta$ solves the mixed boundary value problem

$$\begin{aligned} -\nabla \cdot (b_0(x)\nabla v(x)) &= \frac{\partial u(T,x)}{\partial t} - Du^\delta(T,x), \quad x \in \Omega, \\ v(x) &= (u^\delta - u)(T,x), \quad x \in \partial\Omega. \end{aligned}$$

Recalling our discussion concerning the error of numerical differentiation, we can fix ε such that $\left\| \dfrac{\partial u(T,x)}{\partial t} - Du^\delta(T,x) \right\| \le \varepsilon$. Using representations of the solutions of boundary value problems in terms of corresponding Green functions one can easily check that $\|v_0 - v_0^\delta\| \le \bar{c}(\varepsilon + \delta)$, where $\bar{c}$ depends on b_0 and Ω. Therefore,

$$\|\bar{r} - r\| \le \delta + \bar{c}(\varepsilon + \delta).$$

In practically interesting cases the error of the numerical differentiation ε is much larger than data error δ, such that $\|\bar{r} - r\| \le c_2\varepsilon$, where the constant c_2 does not depend on δ. From our discussion above it follows, in particular, that for $u \in C^2([T-\sigma, T] \otimes H^2(\Omega))$, $\Omega = [c,d]$, $\sigma > \sqrt{\delta}$, both $\varepsilon_1, \varepsilon_2$ from (4.2.21) are of order $\sqrt{\delta}$, that will be used in our numerical experiments later.

Now we have all necessary ingredients for dealing with a perturbed equation $As = r$ in the considered case. Applying discretized Tikhonov regularization

we obtain a regularized approximate solution $s_{\alpha,n}$, as it has been described in the previous section. Then the approximation for the searched-for parameter $b = \bar{s} + b_0$ can be constructed as $b_{\alpha,n} = s_{\alpha,n} + b_0$.

Then we have the following proposition, which is an analogue of Proposition 4.3.4.

Proposition 4.4.1. *Assume that the conditions of Proposition 4.3.1 are satisfied. Let $n = n(\varepsilon)$ be such that $h_n = O(\varepsilon)$. Suppose that $\bar{s} = b - b_0 \in \bar{A}_{\phi_1,R}$, $\phi_1 \in \mathcal{F}_0 \cup \mathcal{F}_{1/2}$, where $\bar{A}$ is defined in (4.4.6). Then for $\alpha(\varepsilon) = \alpha_+$ chosen as in (4.3.17) and $b_{\alpha(\varepsilon)} = s_{\alpha_+,\varepsilon} + b_0$*

$$\|b - b_{\alpha(\varepsilon)}\| \le \frac{6\sqrt{q}c_R\varepsilon}{\sqrt{\alpha_{opt}}},$$

where $\alpha_{opt} = \Theta_{1,\varepsilon}^{-1}(c_R\varepsilon)$, $\Theta_{1,\varepsilon}(\alpha) = \Xi_{1,\varepsilon}(\alpha)\sqrt{\alpha}$, $\Xi_{1,\varepsilon}(\alpha) = c_1\phi_1(\alpha) + c_2\phi_1(\varepsilon)$.

4.4.2 Regularized identification of the diffusion coefficient $a(u)$

As it has been mentioned in the Introduction we are going to use the equation

$$a(u(t,x)) = b(t,x) \tag{4.4.10}$$

with $t = T$ for the identification of $a(u)$. Keeping in mind a structure of a quasilinear parabolic system (4.4.1) it is natural to suppose that $a(u)$ is a differentiable function. In view of (4.4.3) we will assume more specifically that a belongs at least to the Sobolev space $W_2^1 = W_2^1(u_1, u_2)$.

The problem with equation (4.4.10) is that we know neither the exact parameter $b = b(T,\cdot)$ nor terminal status $u(T,\cdot)$. We therefore have to deal with the following perturbed version of (4.4.10)

$$a(u_{sm}^{\delta}(T,x)) = b_{\alpha(\varepsilon)}(x), \tag{4.4.11}$$

where u_{sm}^{δ}, $b_{\alpha(\varepsilon)}$ have been defined above. Of course, in general we cannot guarantee the existence of the function $a \in W_2^1(u_1, u_2)$ such that (4.4.11) is valid. For example, if $u_{sm}^{\delta}(T,x)$ takes equal value at several $x \in \Omega$, equation (4.4.11) may define a multivalued function a.

As suggested perviously, consider linear operators $\bar{A}$, A which map an arbitrary function $s \in W_2^1(u_1, u_2)$ into the functions

$$x \to s(u(T,x)) \in L^2(\Omega), \quad x \to s(u_{sm}^{\delta}(T,x)) \in L^2(\Omega).$$

Then for $t = T$ equation (4.4.10) can be seen as a linear operator equation $\bar{A}s = \bar{r}$ with $\bar{r} = b = b(T,\cdot)$, and with the exact solution $\bar{s} = a \in W_2^1(u_1, u_2)$. Meanwhile, our problem now consists in solving the perturbed equation $As = r$ with $r = b_{\alpha(\varepsilon)}$ to find an approximation for $\bar{s} = a$.

To use general results of the previous section, we should specify Hilbert spaces X, Y and $n-$dimensional space V_n. In the present context the choice of Hilbert spaces is obvious: $X = W_2^1(u_1, u_2)$, $Y = L^2(\Omega)$.

In order to meet the requirement of (4.2.21) in Theorem 4.2.6, we assume that the trial space $V_n \subset W_2^1(u_1, u_2)$ is chosen such that for the orthogonal projector P_n (now in $W_2^1(u_1, u_2)$) onto V_n the following Jackson type inequality holds true

$$\|I - P_n\|_{W_2^1(u_1,u_2)\to C[u_1,u_2]} \le c_J n^{-1/2}, \tag{4.4.12}$$

where the constant c_J does not depend on n. This assumption is not restrictive. Particularly, it is fulfilled for the standard approximation spaces such as algebraic polynomials or piecewise polynomial splines. Let

$$P_n = \sum_{i=1}^{n} \Psi_i \langle \Psi_i, \cdot \rangle_{W_2^1},$$

where the functions $\Psi_i(u)$, $i = 1, 2, \cdots, n$, form W_2^1-orthonormal basis in V_n. Then for any $s \in W_2^1(u_1, u_2)$ it follows from (4.4.3), (4.4.8) that

$$\begin{aligned}
&\|\bar{A}s - \bar{A}P_n s\|_{L^2(\Omega)} \\
&\le (mes(\Omega))^{1/2} \max_{x\in\Omega} \left| s(u(T,x)) - \sum_{i=1}^{n} \Psi_i(u(T,x)) \langle \Psi_i, s \rangle_{W_2^1} \right| \\
&\le (mes(\Omega))^{1/2} \max_{u\in[u_1,u_2]} \left| s(u) - \sum_{i=1}^{n} \Psi_i(u) \langle \Psi_i, s \rangle_{W_2^1} \right| \\
&= (mes(\Omega))^{1/2} \|s - P_n s\|_C \\
&\le c_J (mes(\Omega))^{1/2} \|s\|_{W_2^1} n^{-1/2}.
\end{aligned}$$

Thus,

$$\|\bar{A}(I - P_n)\|_{W_2^1(u_1,u_2)\to L^2(\Omega)} \le c_J (mes(\Omega))^{1/2} n^{-1/2}, \tag{4.4.13}$$

and the same estimate holds for the operator A. This means, in particular, that $\bar{A}$, A are linear compact operators from $W_2^1(u_1, u_2)$ into $L^2(\Omega)$, because the compact operators of a finite rank $\bar{A}P_n$, AP_n converge to them in the operator norm when $n \to \infty$. Keeping in mind that $\bar{A}$, A are the operators of an infinite rank, we can conclude that equations (4.4.10), (4.4.11) are ill-posed, and the application of discretized Tikhonov regularization is relevant. This regularization scheme seems to be appropriate for equations like (4.4.11), since it does not involve the adjoint operator $A^* : L^2(\Omega) \to W_2^1(u_1, u_2)$ which has a rather complicated structure in considered case (see e.g. Appendix B in [19]).

Note that for this particular problem the entries of the associated stiffness matrix M and of the right-hand side F in (4.2.11) can be easily computed as follows:

$$\langle A\Phi_i, A\Phi_j\rangle = \int_\Omega \Phi_i(u^\delta_{sm}(T,x))\Phi_j(u^\delta_{sm}(T,x))dx,$$
$$\langle A\Phi_i, f\rangle = \int_\Omega \Phi_i(u^\delta_{sm}(T,x))b_{\alpha(\varepsilon)}(x)dx,$$

where $\{\Phi_i\}_{i=1}^n$ is our favorite basis in V_n.

The main question now is connected with the choice of the regularization parameter. To implement the adaptive parameter choice strategy (4.3.17) we again need to know noise level estimates $\|\bar{A} - A\| \le \varepsilon_1$ and $\|\bar{r} - r\| \le \varepsilon_2$. The first of them can be easily obtained. Indeed, for any $s \in W_2^1(u_1, u_2)$ it follows from (4.4.9) that

$$\begin{aligned}
&\|\bar{A}s - As\|^2 \\
&= \int_\Omega |s(u(T,x)) - s(u^\delta_{sm}(T,x))|^2 dx \\
&= \int_\Omega \left| \int_{u^\delta_{sm}(T,x)}^{u(T,x)} s'(u)du \right|^2 dx \\
&\le \int_\Omega \int_{u_1}^{u_2} |s'(u)|^2 du \left| \int_{u^\delta_{sm}(T,x)}^{u(T,x)} du \right| dx \\
&\le \|s\|^2_{W_2^1} \int_\Omega |u^\delta_{sm}(T,x) - u(T,x)|dx \\
&\le \|s\|^2_{W_2^1} (mes(\Omega))^{1/2} \|u^\delta_{sm}(T,\cdot) - u(T,\cdot)\|_{L^2(\Omega)} dx \\
&\le c\|s\|^2_{W_2^1} (mes(\Omega))^{1/2}\delta.
\end{aligned}$$

Thus, in the considered case

$$\|\bar{A} - A\| \le c(mes(\Omega))^{1/4}\sqrt{\delta} = \varepsilon_1. \tag{4.4.14}$$

At the same time, we have a problem with an estimation of $\|\bar{r} - r\|$. Recall that now $\bar{r} = b = b(T,\cdot)$ and $r = b_{\alpha(\varepsilon)}$. Then under the assumptions of the Proposition 4.4.1 we have

$$\|\bar{r} - r\| = \|b - b_{\alpha(\varepsilon)}\| \le \frac{6\sqrt{q}c_R\varepsilon}{\sqrt{\alpha_{opt}}}. \tag{4.4.15}$$

Moreover, from (4.3.16) we know that this estimate is order optimal. Therefore, it is natural to assume that there is a constant $c \in (0,1)$ for which

$$6c\sqrt{q}c_R\frac{\varepsilon}{\sqrt{\alpha_{opt}}} \le \|b - b_{\alpha(\varepsilon)}\| \le 6\sqrt{q}c_R\frac{\varepsilon}{\sqrt{\alpha_{opt}}}. \tag{4.4.16}$$

However, we have no access to the value of α_{opt}, since it is given in terms of unknown index function ϕ. Such situation is typical of the Theory of ill-posed problems, where no error estimations can be given without knowledge of the compacts containing the problem solutions.

On the other hand, the value $\alpha(\varepsilon)$ in (4.4.15) is known. It has been obtained within the framework of adaptive parameter choice strategy (4.3.17), and from the proof of Proposition 4.3.4 we know that $\alpha_{opt} \le q\alpha(\varepsilon)$. Keeping in mind that the choice of the regularization parameter $\alpha = \alpha(\varepsilon)$ leads to the same order of accuracy as the optimal choice $\alpha = \alpha_{opt}$, it is natural to assume that there is a positive constant $\bar{q}$ for which $\alpha(\varepsilon) \le \bar{q}\alpha_{opt}$. Then (4.4.16) can be rewritten as follows

$$\bar{d}\frac{\varepsilon}{\sqrt{\alpha(\varepsilon)}} \le \|b - b_{\alpha(\varepsilon)}\| \le \underline{d}\frac{\varepsilon}{\sqrt{\alpha(\varepsilon)}}. \tag{4.4.17}$$

Of course, the values $\bar{d} = 6cc_R$, $\underline{d} = 6\sqrt{\bar{q}q}c_R$ are not known here. The essence of the two sided estimation (4.4.17) is that the known quantity $\omega = \varepsilon/\sqrt{\alpha(\varepsilon)}$ can be taken as a scale for measuring a perturbation level. As we will show in the next section, under some additional assumptions the estimate (4.4.17) can be used for adaptation to unknown index function in the source condition (4.2.6), and to unknown values of $\bar{d}, \underline{d}$ as well.

4.4.3 Extended balancing principle

Now we propose a strategy for choosing a regularization parameter in the situation when a level of data perturbation is not exactly given.

To be specific, we discuss the application of discretized Tikhonov regularization to equation (4.4.11) treated as a perturbed operator equation $As = r$.

We assume that the exact searched-for parameter $a(u)$ meets a source condition of the form (4.2.6). Keeping in mind that a solves the equation $\bar{A}s = \bar{r}$ for the operator $\bar{A}$ defined in the last section, and $\bar{r} = b(T, \cdot)$, it is natural to give a source condition in terms of this operator. Therefore, we will assume that

$$a \in \bar{A}_{\phi_2,R}, \quad \phi_2 \in \mathcal{F}_0 \cup \mathcal{F}_{1/2}. \tag{4.4.18}$$

Let

$$\begin{aligned} a_{\alpha,n} &= (\alpha I + P_n A^* A P_n)^{-1} P_n A^* b_{\alpha(\varepsilon)}, \\ \bar{a}_{\alpha,n} &= (\alpha I + P_n A^* A P_n)^{-1} P_n A^* b \end{aligned}$$

be discretized Tikhonov approximation given to the perturbed and the exact right-hand side terms respectively. Here $A : a \to a(u^{\delta}_{sm}(T, \cdot))$ is a perturbed version of the operator $\bar{A} : a \to a(u(T, \cdot))$. Using spectral calculus we have

$$\begin{aligned} &\|a_{\alpha,n} - \bar{a}_{\alpha,n}\| \\ &\le \|(\alpha I + P_n A^* A P_n)^{-1} P_n A^*\|_{L^2(\Omega) \to W_2^1(u_1,u_2)} \|b - b_{\alpha(\varepsilon)}\|_{L^2(\Omega)} \\ &\le \sup_{\lambda} \left| \frac{\sqrt{\lambda}}{\alpha + \lambda} \right| \|b - b_{\alpha(\varepsilon)}\| \le \frac{\|b - b_{\alpha(\varepsilon)}\|}{2\sqrt{\alpha}}. \end{aligned} \tag{4.4.19}$$

Moveover, (4.4.13), (4.4.14) and (4.4.18) allow the application of the Theorem 4.2.6 with $\bar{s} = a$, $s_{\alpha,n} = \bar{a}_{\alpha,n}$, $\phi = \phi_2$, $\varepsilon_1 = O(\sqrt{\delta})$, $\varepsilon_2 = 0$, and

$$n = n_\delta = \lceil c_J^2 \mathrm{mes}(\Omega)\delta^{-1} \rceil = O(\delta^{-1}), \tag{4.4.20}$$

where $\lceil c \rceil$ means the smallest integer that is larger than c. Then

$$\|a - \bar{a}_{\alpha,n_\delta}\| \le c_1\phi_2(\alpha) + c_2\phi_2(\varepsilon_1) + \frac{c_R\varepsilon_1}{\sqrt{\alpha}}. \tag{4.4.21}$$

Using this estimate together with the triangle inequality, (4.4.19), (4.4.17) and a trivial estimate $\omega = \varepsilon/\sqrt{\alpha(\varepsilon)} > \varepsilon_1 = O(\sqrt{\delta})$, we arrive at

$$\|a - a_{\alpha,n_\delta}\| \le c_1\phi_2(\alpha) + c_2\phi_2(\varepsilon_1) + d\frac{\omega}{\sqrt{\alpha}}, \tag{4.4.22}$$

where $d = \underline{d} + c_R$. In view of (4.4.19) and (4.4.17) for this d we also have

$$\|a_{\alpha,n_\delta} - \bar{a}_{\alpha,n_\delta}\| \le d\frac{\omega}{\sqrt{\alpha}}. \tag{4.4.23}$$

Again, to maintain the best possible error bound in (4.4.22) it is sufficient to choose $\alpha = \alpha_{opt}(d)$ balancing the values $\Xi_{2,\varepsilon_1}(\alpha) = c_1\phi_2(\alpha) + c_2\phi_2(\varepsilon_1)$ and $d\omega/\sqrt{\alpha}$. In terms of the function $\Theta_{2,\varepsilon_1}(\alpha) = \Xi_{2,\varepsilon_1}(\alpha)\sqrt{\alpha}$ it can be written as $\alpha_{opt}(d) = \Theta_{2,\varepsilon_1}^{-1}(d\omega)$.

Thus, under the source condition (4.4.18) we have

$$\|a - a_{\alpha_{opt}(d),n_\delta}\| \le \frac{2d\omega}{\sqrt{\alpha_{opt}(d)}} = 2\Xi_{2,\varepsilon_1}(\Theta_{2,\varepsilon_1}^{-1}(d\omega)), \tag{4.4.24}$$

where n_δ is chosen as in (4.4.20).

Moreover, with an argument like that in the proof of Proposition 4.3.4, we get the estimate

$$\|a - a_{\alpha(d),n_\delta}\| \le \frac{6\sqrt{q}d\omega}{\sqrt{\alpha_{opt}(d)}} = 3\sqrt{q}\Xi_{2,\varepsilon_1}(\Theta_{2,\varepsilon_1}^{-1}(d\omega)), \tag{4.4.25}$$

for $\alpha(d)$ chosen within the framework of strategy (4.3.17), where $c_R\varepsilon$ is substituted for $d\omega$, i.e.

$$\alpha(d) = \tag{4.4.26}$$
$$\max\left\{\alpha_k \in \Delta_N : \|a_{\alpha_k,n_\delta} - a_{\alpha_i,n_\delta}\| \le d\omega\left(\frac{3}{\sqrt{\alpha_k}} + \frac{1}{\sqrt{\alpha_i}}\right), i = 0, 1, \cdots, k\right\}.$$

At this point it is important to recall that the value of d is unknown for us. On the other hand, it is clear that the estimates (4.4.22)–(4.4.25) remain valid if d is substituted for any larger constant. But since Ξ_{2,ε_1}, $\Theta_{2,\varepsilon}^{-1}$ are increasing functions the estimates (4.4.24) and (4.4.25) worsen when d increases.

Our goal now is to describe a procedure for adaptation to unknown value of d. We find it pertinent to call this procedure the extended balancing principle. It consists in combining the adaptive strategy (4.4.26) with successive testing of the hypothesis that the constant d in (4.4.22)–(4.4.25) is not larger than some term d_j of a fixed geometric sequence

$$D_p = \{d_j = d_0 p^j, j = 0, 1, \cdots, M\}, \quad p > 1.$$

For each of these hypotheses the strategy (4.4.26) selects the value $\alpha(j) = \alpha(d_j) \in \Delta_N$, and it is easy to check that the sequence $\alpha(j), j = 0, 1, \cdots, M$, is non-decreasing:

$$\alpha(0) \le \alpha(1) \le \cdots \le \alpha(j) \le \cdots \le \alpha(M) \le \alpha_N. \tag{4.4.27}$$

To justify the extended balancing principle we need to assume that within the regularization process the data error really propagates with the rate of order $\alpha^{-1/2}$. In view of (4.4.23), (4.4.19) and (4.4.17) it can be written in the form of two-sided stability estimate

$$cd\frac{\omega}{\sqrt{\alpha}} \le \|a_{\alpha,n_\delta} - \bar{a}_{\alpha,n_\delta}\| \le d\frac{\omega}{\sqrt{\alpha}}, \tag{4.4.28}$$

which is assumed to be valid for some $c \in (0, 1)$ and for any $\alpha \in \Delta_N$.

In our subsequent analysis we rely on the assumption that there are two adjacent terms d_l, $d_{l+1} \in D_p$ such that

$$d_l \le cd < d \le d_{l+1}. \tag{4.4.29}$$

This assumption means that the term with (unknown) number $l + 1$ is the best candidate among D_p putting it in place of d in (4.4.22)–(4.4.26). The basic idea of the extended balancing principle is the following. If $d_j \in D_p$ is smaller than the actual value d_l in (4.4.29), then a corresponding $\alpha(j)$ is "too small". It turns out that this can be detected from approximations a_{α_i,n_δ}, $\alpha_i \in \Delta_N$, using a special choice of initial terms α_0, d_0 and denominators q, p in geometric sequences Δ_N, D_p. We argue that if δ is small enough then they always can be chosen such that for a fixed design constant $\kappa > 1$

$$\begin{aligned} \alpha_{opt}(d) = \Theta^{-1}_{2,\varepsilon_1}(d\omega) &> \kappa^2 q \alpha_0 \left(\frac{3p^2+7}{p-1}\right)^2, \\ c_R\varepsilon_1 &< \frac{\kappa-1}{\kappa}(d_l - d_{l-1})\omega. \end{aligned} \tag{4.4.30}$$

For example, if the solution $u(t, x)$ of (4.4.1) belongs to $C^2[T-\sigma, T] \otimes H^2(\Omega)$, and Ω is some interval of the real axis then $\alpha_0 = \delta\kappa^{-2}q^{-1}(p-1)^2/(3p^2+7)^2$ meets the first inequality in (4.4.30). Indeed, without loss of generality we can assume that $\alpha_{opt}(d) > \omega^2$. Such assumptions simply mean that the term $\omega/\sqrt{\alpha_{opt}(d)}$ in the error estimations (4.4.24) is smaller than 1, which is not at all restrictive. On the other hand, based on the previous analysis in the considered case $\varepsilon = O(\sqrt{\delta})$.

Then $\alpha_{opt}(d) > \omega^2 = \varepsilon^2/\alpha(\varepsilon) \gg \delta$, which means that the first inequality in (4.4.30) is satisfied with the above chosen α_0. Concerning the second inequality, it is clear, for example, that for $d_0 = \kappa/((\kappa-1)(p-1))$

$$\frac{\kappa-1}{\kappa}(d_l - d_{l-1})\omega > \frac{\kappa-1}{\kappa}d_0(p-1)\omega = \omega \gg \sqrt{\delta}.$$

At the same time, $c_R\varepsilon_1 = O(\sqrt{\delta})$, and the same argument can be used again. We would like to note also that (4.4.30) is only a technical assumption, and it will be used just for the sake of presentation simplicity.

Lemma 4.4.2. *Let the assumptions (4.4.18), (4.4.28), (4.4.29) hold. Assume that δ is small enough such that the technical assumption (4.4.30) is satisfied. If $\alpha(j) = \alpha(d_j)$ is chosen in accordance with (4.4.26) for $d = d_j \in D_p$, and $j \le l-1$ then*

$$\alpha(j) < 9\kappa^2\left(\frac{p^2+1}{p-1}\right)^2\alpha_0.$$

Proof. Using (4.4.29) we rewrite (4.4.22) as

$$\|a - a_{\alpha,n_\delta}\| \le \Xi_{2,\varepsilon_1}(\alpha) + \frac{d_{l+1}\omega}{\sqrt{\alpha}}.$$

Then from (4.4.21), (4.4.28), and since $\alpha_j \ge \alpha_0$, we obtain

$$\begin{aligned}
&\|a_{\alpha(j),n_\delta} - a_{\alpha_0,n_\delta}\| \\
&\ge \|a_{\alpha_0,n_\delta} - \bar{a}_{\alpha_0,n_\delta}\| - \|a - \bar{a}_{\alpha_0,n_\delta}\| - \|a - a_{\alpha_(j),n_\delta}\| \\
&\ge \frac{d_l\omega}{\sqrt{\alpha_0}} - \Xi_{2,\varepsilon_1}(\alpha_0) - \frac{c_R\varepsilon_1}{\sqrt{\alpha_0}} - \Xi_{2,\varepsilon_1}(\alpha(j)) - \frac{d_{l+1}\omega}{\sqrt{\alpha(j)}} \\
&\ge \frac{d_l\omega - c_R\varepsilon_1}{\sqrt{\alpha_0}} - 2\Xi_{2,\varepsilon_1}(\alpha(j)) - \frac{d_{l+1}\omega}{\sqrt{\alpha(j)}}. \qquad (4.4.31)
\end{aligned}$$

Now we introduce

$$\alpha_* = \max\left\{\alpha_i \in \Delta_N : \Xi_{2,\varepsilon_1}(\alpha) \le \frac{d_{l+1}\omega}{\sqrt{\alpha}}\right\},$$

and consider the cases $\alpha(j) < \alpha_*$ and $\alpha(j) \ge \alpha_*$ separately. For $\alpha(j) < \alpha_*$, $\Xi_{2,\varepsilon_1}(\alpha(j)) \le d_{l+1}\omega/\sqrt{\alpha(j)}$, and using (4.4.26) with $d = d_j$ and $\alpha(j) = \alpha(d_j)$ we can extend the chain of inequalities (4.4.31) as

$$d_j\omega\left(\frac{3}{\sqrt{\alpha(j)}} + \frac{1}{\sqrt{\alpha_0}}\right) \ge \|a_{\alpha_(j),n_\delta} - a_{\alpha_0,n_\delta}\| \ge \frac{d_l\omega - c_R\varepsilon_1}{\sqrt{\alpha_0}} - \frac{3d_{l+1}\omega}{\sqrt{\alpha(j)}}.$$

Keeping in mind that $d_j \le d_{l-1} < d_l$ it is easy to derive the required estimate

$$\begin{aligned}\alpha(j) &\le 9\alpha_0\left(\frac{(d_{l+1}+d_j)\omega}{(d_l-d_j)\omega - c_R\varepsilon_1}\right)^2\\ &\le 9\alpha_0\left(\frac{d_{l-1}(p^2+1)\omega}{d_{l-1}(p-1)\omega - c_R\varepsilon_1}\right)^2\\ &< 9\kappa^2\alpha_0\left(\frac{p^2+1}{p-1}\right)^2.\end{aligned}$$

Note that the technical assumption (4.4.30) is also used here.

Now let us consider the remaining case $\alpha(j) \ge \alpha_*$ and prove that it is impossible.

Note that in (4.4.31) α_j can be substituted for α_*. Then

$$\|a_{\alpha_*,n_\delta} - a_{\alpha_0,n_\delta}\| \ge \frac{d_l\omega - c_R\varepsilon_1}{\sqrt{\alpha_0}} - \frac{3d_{l+1}\omega}{\sqrt{\alpha_*}}.$$

On the other hand, from the definition of $\alpha_j = \alpha(d_j)$, one can derive the estimate

$$\begin{aligned}&\|a_{\alpha_*,n_\delta} - a_{\alpha_0,n_\delta}\|\\ &\le \|a_{\alpha(j),n_\delta} - a_{\alpha_*,n_\delta}\| + \|a_{\alpha_0,n_\delta} - a_{\alpha(j),n_\delta}\|\\ &\le d_j\omega\left(\frac{6}{\sqrt{\alpha(j)}} + \frac{1}{\sqrt{\alpha_*}} + \frac{1}{\sqrt{\alpha_0}}\right)\\ &\le d_j\omega\left(\frac{7}{\sqrt{\alpha_*}} + \frac{1}{\sqrt{\alpha_0}}\right).\end{aligned}$$

Thus,

$$d_j\omega\left(\frac{7}{\sqrt{\alpha_*}} + \frac{1}{\sqrt{\alpha_0}}\right) \ge \frac{d_l\omega - c_R\varepsilon_1}{\sqrt{\alpha}_0} - \frac{3d_{l+1}\omega}{\sqrt{\alpha_*}},$$

and repeating our previous argument we conclude that

$$\alpha_* < \kappa^2\alpha_0\left(\frac{3p^2+7}{p-1}\right)^2. \tag{4.4.32}$$

At the same time, from the definition of α_* it follows that for $q\alpha_* \in \Delta_N$

$$\Xi_{2,\varepsilon_1}(q\alpha_*)\sqrt{q\alpha_*} = \Theta_{2,\varepsilon_1}(q\alpha_*) > d_{l+1}\omega \ge d\omega = \Theta_{2,\varepsilon_1}(\alpha_{opt}(d)).$$

Since $\Theta_{2,\varepsilon_1}(\alpha)$ is an increasing function and δ is assumed to be small enough such that (4.4.30) holds, we get the estimate

$$\alpha_* > q^{-1}\alpha_{opt}(d) > \kappa^2\alpha_0\left(\frac{3p^2+7}{p-1}\right)^2, \tag{4.4.33}$$

which is in contradiction with (4.4.32). □

Let $\hat{\alpha}$ denote the first term of the sequence $\{\alpha(j)\}_{j=0}^{M}$(see (4.4.27)) which is above the threshold indicated in Lemma 4.4.2, i.e.

$$\hat{\alpha} = \min\left\{\alpha(j) : \alpha(j) \geq 9\kappa^2 \left(\frac{p^2+1}{p-1}\right)^2 \alpha_0\right\}.$$

We stress that the unknown index function ϕ_2 from the source condition (4.4.18) and the best (unknown) approximation d_{l+1} of d from (4.4.22), (4.4.23) are not involved in determining $\hat{\alpha}$. By construction $\hat{\alpha} = \alpha(k)$ corresponds to some $d_k \in D_p$.

We show now that using the value of d_k one can detect at least d_{l+2} that is the closest term to the best candidate d_{l+1}.

Theorem 4.4.3. *Let $\hat{\alpha} = \alpha(k) \in \Delta_N$. Then under the conditions of Lemma 4.4.2 $d_{l+1} \leq d_{k+1} \leq d_{l+2}$, and for $\alpha_+ = \alpha(k+1)$ we have either $\alpha_+ = \alpha(l+1)$ or $\alpha_+ = \alpha(l+2)$.*

Proof. Note that similar to inequality (4.3.18), one can prove $\alpha(l+1) = \alpha(d_{l+1}) \geq \alpha_*$. Then from (4.4.33) we have

$$\alpha(l+1) > \kappa^2\alpha_0 \left(\frac{3p^2+7}{p-1}\right)^2 > 9\kappa^2\alpha_0 \left(\frac{p^2+1}{p-1}\right)^2.$$

Therefore, from the definition $\hat{a} = \alpha(k) \leq \alpha(l+1)$. On the other hand, from Lemma 4.4.2 one can derive the estimate $\alpha(k) > \alpha(l-1)$. In view of (4.4.27) it means that $\alpha(k) \geq \alpha(l)$, and as a consequence

$$\alpha(l) \leq \alpha(k) \leq \alpha(l+1) \Rightarrow \alpha(l+1) \leq \alpha(k+1) \leq \alpha(l+2).$$

Thus, $\alpha(k+1)$ can take only the values $\alpha(k+1) = \alpha(l+1)$, or $\alpha(k+1) = \alpha(l+2)$. □

The regularized approximation of the searched-for parameter $a(u)$ we are interested in is now defined as a_{α_+,n_δ}.

Corollary 4.4.4. *Assume that the conditions of Lemma 4.4.2 are satisfied. Then*

$$\|a - a_{\alpha_+,n_\delta}\|_{W_2^1(u_1,u_2)} \leq c_{p,q}\Xi_{2,\varepsilon_1}(\Theta_{2,\varepsilon_1}^{-1}(d\omega)), \tag{4.4.34}$$

where d is the smallest constant for which (4.4.22), (4.4.23) hold, and the factor $c_{p,q}$ depends only on the denominators of the geometric sequences D_p, Δ_N.

Proof. From Theorem 4.4.3, it follows that $d_{k+1} \in D_p$ corresponding to $\alpha_+ = \alpha(k+1)$ is not larger than d_{l+2}. Using (4.4.25) with $d = d_{l+2}$ we have

$$\|a - a_{\alpha_+,n_\delta}\|_{W_2^1(u_1,u_2)} \leq 3\sqrt{q}\Xi_{2,\varepsilon_1}(\Theta_{2,\varepsilon_1}^{-1}(d_{l+2}\omega)).$$

Moreover, from (4.4.29) if follows that $d_{l+2} \le p^2 d$. Now the same argument as in the proofs of (4.3.20), (4.3.22) gives

$$\Xi_{2,\varepsilon_1}(\Theta^{-1}_{2,\varepsilon_1}(d_{l+2}\omega)) \le \Xi_{2,\varepsilon_1}(\Theta^{-1}_{2,\varepsilon_1}(p^2 d\omega))$$
$$\le \Xi_{2,\varepsilon_1}(p^4\Theta^{-1}_{2,\varepsilon_1}(d\omega)) \le c(p)\Xi_{2,\varepsilon_1}(\Theta^{-1}_{2,\varepsilon_1}(d\omega)).$$

□

It is interesting to know how the bound (4.4.34) can be expressed through data noise level δ in (4.4.2). To make it definite we assume that the exact solution $u(t,x)$ of (4.4.2) belongs to $C^2[T-\sigma,T]\otimes H^2(\Omega)$, and Ω is some interval of real axis. Then, as we already mentioned, both ε in (4.4.17) and ε_1 in (4.4.21) are of order $\sqrt{\delta}$. In view of Proposition 4.4.1

$$\omega = \frac{\varepsilon}{\sqrt{\alpha(\varepsilon)}} \le \frac{\varepsilon}{\sqrt{\alpha_{opt}}} \le c\Xi_{1,\varepsilon}(\Theta^{-1}_{1,\varepsilon}(c_R\varepsilon)).$$

Thereon with an argument like that in the proof of Corollary 4.3.1 we get the estimate $\omega \le \phi_1(\Theta_1^{-1}(\sqrt{\delta}))$, and as a consequence

$$\|a - a_{\alpha_+,n_\delta}\|_{W_2^1(u_1,u_2)} \le c\phi_2(\Theta_2^{-1}(\phi_1(\Theta_1^{-1}(\sqrt{\delta})))). \tag{4.4.35}$$

Here and below c is a δ-independent generic constant.

To simplify (4.4.35) we consider a special case $\phi_1(\lambda) = \lambda^{\mu_1}$, $\phi_2(\lambda) = \lambda^{\mu_2}$, $\frac{1}{2} \le \mu_1, \mu_2 \le 1$. Then (4.4.35) can be rewritten as

$$\|a - a_{\alpha_+,n_\delta}\|_{W_2^1(u_1,u_2)} \le c\delta^{\frac{2\mu_1\mu_2}{(2\mu_1+1)(2\mu_2+1)}}. \tag{4.4.36}$$

From this estimate it follows that the best accuracy that in principle can be guaranteed within the framework of proposed approach has the order $\delta^{2/9}$.

To the best of our knowledge there are only three papers [19, 28, 47] containing error bounds for similar parameter identification problems, but none of them can be compared with (4.4.36) directly.

In [28] the error was measured in the mixed boundary weighted norm $\max_t \|(a(u(t,\cdot)) - \tilde{a}(\tilde{u}(t,\cdot)))\nabla u(t,\cdot)\|_{L^2(\Omega)}$ and estimated as $O(\delta^{1/2})$. Here and below $\tilde{a}$ and $\tilde{u}$ are the approximations for the exact coefficient a and the solution u of (4.4.1) respectively.

In [19] the assumption that the observations (4.4.2) are given for the whole time period $[0,T]$ is crucial. It corresponds to the case $\sigma = T$, while for the NL described above σ can be even equal to zero, if as it is assumed in [47], some approximation for $\frac{\partial u(T,\cdot)}{\partial t}$ is a priori given as data. Moreover, the method from [19] has been designed only for the case of one space variable x.

In [47] the error was measured in the standard L^2-norm, with an assumption of the positivity of ∇u, that is not necessary for our NL scheme. However, the result [47] can be used in a heuristic explanation for our error bound $O(\delta^{2/9})$.

Indeed, from Theorem 3.6 [47] it follows that for the diffusion coefficient with a priori known smoothness given in the form of inclusion $a \in W_\infty^3(u_1, u_2)$ one can construct an approximation $\tilde{a}$ in such a way that

$$\|a - \tilde{a}\|_{L^2(u_1,u_2)} \le c\delta^{1/3}. \tag{4.4.37}$$

If $\tilde{a}$ is so smooth that $a - \tilde{a}$ also belongs to $W_\infty^3(u_1, u_2)$ and $\|a - \tilde{a}\|_{W_\infty^3(u_1,u_2)} = O(1)$, then using (4.4.33) together with interpolation inequality one can estimate the error in $W_2^1(u_1, u_2)$ as follows,

$$\|a - \tilde{a}\|_{W_2^1(u_1,u_2)} \le \|a - \tilde{a}\|^{2/3}_{L^2(u_1,u_2)} \|a - \tilde{a}\|^{1/3}_{W_2^3(u_1,u_2)} \le c\delta^{2/9},$$

that coincides with (4.4.36) for $\mu_1 = \mu_2 = 1$.

4.4.4 Numerical examples

We present the results of several numerical experiments demonstrating the performance of the NL for parameter identification in quasi-linear parabolic systems.

Following [47] we will consider only one-dimensional examples, where the domain Ω reduces to the standard unit interval $[0, 1]$, since it already contains all the important aspects of the NL.

We consider (4.4.1) with the diffusion coefficient $a(u) = \exp(u)$ and the exact solution $u(t, x) = \exp(-t + x/2)$. Note that in [47] noisy data were simulated in the form $u^\delta(t, x) = u(t, x) + \xi_\delta \sin(\pi t) \sin(\pi x)$, $t, x \in [0, 1]$, where ξ_δ is some fixed constant.

In our experiments we use the values $u(0.95, x)$, $u(1, x)$, $x \in [0, 1]$, contaminated by additive random noise with the uniform distribution in the interval $[-\delta, \delta]$, $\delta = 0.001$. It corresponds to the observation period $[T - \sigma, T]$, where $T = 1$, $\sigma = 0.05$. A smoothed version $u^\delta_{sm}(1, x)$ is constructed as a piece-wise linear spline interpolating noisy values at points $x_i = i/n$, $i = 0, 1, \cdots, n$, $n = 50$. The approximation for $\frac{\partial u(1,x)}{\partial t}$ is taken in the form of the finite difference with $h = \sigma = 0.05$.

In the first linearization step we approximate the function $b(x) = a(u(1, x))$ by $b_{\alpha,n}(x) = s_{\alpha,n} + b_0(x)$, where $b_0(x) \equiv 1$, and $s_{\alpha,n}(x)$ is a linear combination of $n = 50$ piece-wise linear B-splines $\{\Phi_i(x)\}_{i=1}^n$ corresponding to equidistant knots. The coefficients of this linear combination are found from system (4.2.11). In the considered case the operator A is defined by (4.4.7) with $b_0(x) \equiv 1$, and the functions $A\Phi_i$ are computed symbolically. Then the entries of the matrices (4.2.11) are calculated using MATLAB-code for numerical integration. At the end of the first linearization step we implement the regularization parameter choice strategy (4.3.17) with $\varepsilon = \sqrt{\delta} \approx 0.03$, $c_R = 1$, $\alpha_i = (1.2)^i \alpha_0$, $i = 0, 1, \cdots, 30$, $\alpha_0 = \varepsilon^2 = 0.001$. For these parameters the procedure chooses the value $\alpha(\varepsilon) =$

0.02. The exact function $b(x) = a(u(1,x))$ and its approximation $b_{\alpha(\varepsilon)}(x)$ are displayed in Fig. 4.4 (a).

In the considered case we approximate the diffusion coefficient $a(u) = \exp(u)$ in the interval $[u_1, u_2]$, $u_1 = \exp(-1)$, $u_2 = \exp(-0.5)$, by piece-wise linear spline $a_{\alpha,n}$ with $n = 40$ equidistant knots. Coefficients of this spline are found from system (4.2.11). This time in (4.2.11) A is the operator from the representation of equation (4.4.11) in the form of $As = r$.

In our experiments we test a robustness of the balancing principle using various norms and inner products in (4.4.26), and in the formula for entries of the matrix G_Φ in (4.2.11). In the first experiment we take $\|\cdot\| = \|\cdot\|_{W_2^1(u_1,u_2)}$, as suggested by the theory, and use Δ_N with $\alpha_0 = 3 \cdot 10^{-8}$, $q = 1.3$, $N = 26$, and D_p with $d_0 = 0.02$, $p = 1.05$, $M = 30$. Moreover, taking $\kappa = 1.1$ we define the value of the threshold level $9\kappa^2 \left(\frac{p^2+1}{p-1}\right)^2 \alpha_0 = 5.7767 \cdot 10^{-4}$. Recall that after the first linearization step we have $\alpha(\varepsilon) = 0.02$. Therefore in considered case $\omega = \varepsilon/\sqrt{\alpha(\varepsilon)} = 0.2236$. Applying parameter choice strategy (4.4.26) with $d_j \in D_p$, $j = 0, 1, \cdots, 30$, we obtain the following sequence of $\alpha(j) = \alpha(d_j)$: $\alpha(0) = \alpha(1) = 2.752 \cdot 10^{-5}, \alpha(3) = \alpha(4) = \alpha(5) = 3.5776 \cdot 10^{-5}, \alpha(6) = \alpha(7) = \alpha(8) = 4.6509 \cdot 10^{-5}, \alpha(9) = \alpha(10) = 6.0461 \cdot 10^{-5}, \alpha(11) = \alpha(12) = 7.86 \cdot 10^{-5}, \alpha(13) = 1.0218 \cdot 10^{-4}, \alpha(14) = \alpha(15) = 1.3283 \cdot 10^{-4}, \alpha(16) = 1.7268 \cdot 10^{-4}, \alpha(17) = 2.2449 \cdot 10^{-4}, \alpha(18) = 2.9284 \cdot 10^{-4}, \alpha(19) = 3.7939 \cdot 10^{-4}, \alpha(20) = 6.4116 \cdot 10^{-4}, \alpha(21) = 1.0836 \cdot 10^{-3}, \alpha(22) = 1.8312 \cdot 10^{-3}, \alpha(23) = \cdots = \alpha(26) = 4.0232 \cdot 10^{-3}, \alpha(27) = \cdots = \alpha(30) = 5.2305 \cdot 10^{-3}$.

Note that $\alpha(20) = 6.4116 \cdot 10^{-4}$ is the first term of this sequence which is above the threshold $9\kappa^2 \left(\frac{p^2+1}{p-1}\right)^2 \alpha_0 = 5.7767 \cdot 10^{-4}$. Therefore, as suggested in Theorem 4.4.3, we put $\hat{\alpha} = \alpha(20)$, and choose the value of the regularization parameter $\alpha_+ = \alpha(21) = 1.0836 \cdot 10^{-3}$. The exact coefficient $a(u) = \exp(u)$ and its approximation $a_{\alpha_+,40}(u)$ are displayed in Fig. 4.4 (b).

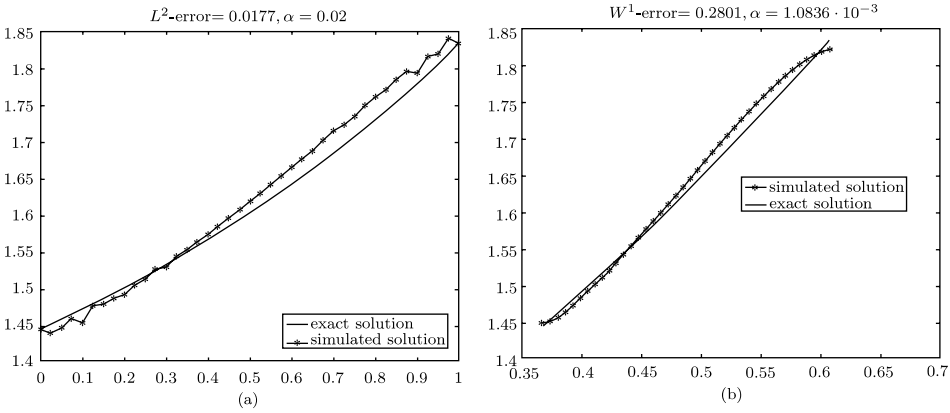


Fig. 4.4 (a) Simulated result for $b(x) = a(u(1,x)) = \exp(\exp(-1 + 0.5x))$; (b) For function $a(x) = \exp(x)$, balancing principle operates with W_2^1-norm.

In the second experiment we apply (4.4.26) with $\|\cdot\| = \|\cdot\|_{L^2(u_1,u_2)}$, but keep the inner product $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_{W_2^1(u_1,u_2)}$, for the entries of the matrix G_Φ. It means that we construct $a_{\alpha,n}$ as an element of $W_2^1(u_1,u_2)$, but the accuracy is measured in L^2-norm. Since this norm is weaker than $\|\cdot\|_{W_2^1(u_1,u_2)}$ the upper bounds (but not the lower estimations) (4.4.21)–(4.4.23) are still valid. We also try a more rough set of regularization parameters Δ_N with $\alpha_0 = 10^{-7}$, $q = 1.3$, $N = 26$. In this case the value of the threshold is $1.9255 \cdot 10^{-3}$, and we find $\alpha_+ = 2.1 \cdot 10^{-3}$. Using Fig. 4.5 (a), one can compare the exact coefficient and its approximation in this case.

We also perform the test with the norm and inner product of $L^2(u_1,u_2)$ used in (4.4.26) and in the formula for the entries of G_Φ. It means that the treatment of equation (4.4.11) is entirely done within the space $L^2(u_1,u_2)$. We again try a more rough parameter sets Δ_N with $\alpha_0 = 1.5 \cdot 10^{-6}$, $q = 1.3$, $N = 26$, and D_p with $d_0 = 0.001$, $p = 1.1$, $M = 30$. Then the parameter choice strategy (4.4.9) gives us the sequence: $\alpha(0) = \alpha(1) = 1.8 \cdot 10^{-3}, \alpha(2) = \cdots = \alpha(6) = 2.3 \cdot 10^{-3}, \alpha(7) = \cdots \alpha(10) = 3 \cdot 10^{-3}, \alpha(11) = \cdots = \alpha(14) = 3.9 \cdot 10^{-3}, \alpha(15) = \cdots = \alpha(18) = 5.1 \cdot 10^{-3}, \alpha(19) = \cdots = \alpha(22) = 6.6 \cdot 10^{-3}, \alpha(23) = \cdots = \alpha(26) = 8.6 \cdot 10^{-3}, \alpha(27) = \cdots = \alpha(30) = 1.12 \cdot 10^{-2}$.

In the considered case the value of the threshold level is $7.9 \cdot 10^{-3}$. Therefore, we choose $\hat{\alpha} = \alpha(23)$ and $\alpha_+ = \alpha(24) = 8.6 \cdot 10^{-3}$. Corresponding approximate solution is displayed in Fig. 4.5 (b) together with the exact parameter $a(u)$.

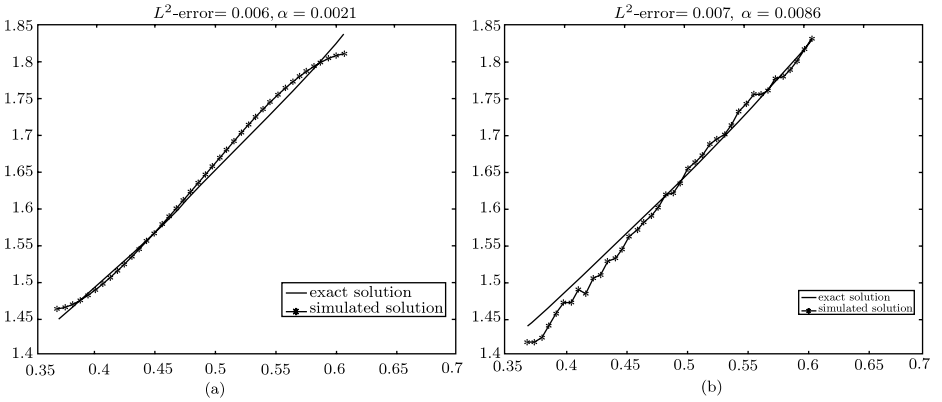


Fig. 4.5 (a) Application of (4.4.26) with L^2-norm; (b) L^2-regularization equipped with L^2-balancing principle.

Moreover, in Fig. 4.6 one can see the exact parameter together with approximate solutions corresponding to $\alpha = 0$ (non-smooth line) and $\alpha = 0.01$ (smooth line).

The latter value is worth a discussion. Recall that in considered case a noise level is estimated as $\omega = \varepsilon\sqrt{\alpha(\varepsilon)} = 0.2236$. If this noise estimate is accepted then

in accordance with the theory of Tikhonov regularization the value ω^2 should be considered as a lower bound for regularization parameters ($\omega/\sqrt{\alpha} \to 0$ as $\omega \to 0$).

At the same time, the values of parameters chosen in accordance with the extended balancing principle are much smaller than $\alpha = 0.01$ or $\omega^2 = 0.049$. Nevertheless, comparing Fig. 4.6 with Figs. 4.4 (b)–4.5 (b), one can see that these values lead to much more accurate reconstruction. It shows evidence of a reliability of the balancing principle in the situation when a noise level is not exactly given.

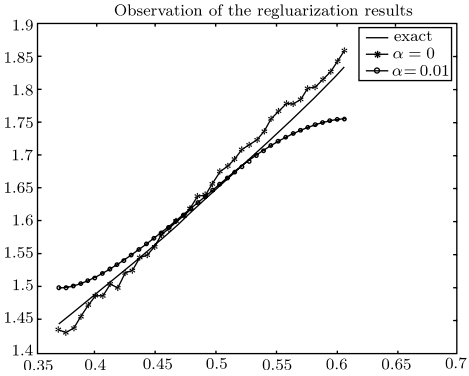


Fig. 4.6 Regularization with parameters chosen on an ad hoc basis.

Our final test demonstrates the ability of the natural linearization to recover a diffusion coefficient $a(u)$ in the situation, when the solution $u(t, x)$ is not a monotonous function. In this case equations (4.4.10), (4.4.11) may define a multivalued function $a(u)$. To avoid such situation the positivity of ∇u was assumed in [47]. This assumption is violated for system (4.4.1) with $a(u) = \exp(u)$ and $u(t, x) = 10e^{-t}x(1 - x)$. To perform our experiment we simulate noisy data and construct $u^{\delta}_{sm}(1, x)$, $Du^{\delta}(1, x)$ as before. Then in the first linearization step the application of the parameter choice strategy (4.3.17) gives the value $\alpha(\varepsilon) = 1.08$. The exact function $b(x)$ and corresponding regularized approximation $b_{\alpha(\varepsilon)}(x)$ are displayed in Fig. 4.7 (a). The second linearization step is performed in the same way as in our first experiment. In accordance with the balancing principle the value $\alpha_+ = 0.006$ is chosen. The exact coefficient $a(u)$ and corresponding regularized approximation are displayed in Fig. 4.7 (b). This test shows evidence of a reliability of the natural linearization.

Acknowledgements This research was supported by the Austrian Fonds Zur Förderung der wissenschaftlichen Forschung (FWF), Project No. P20235-N18.

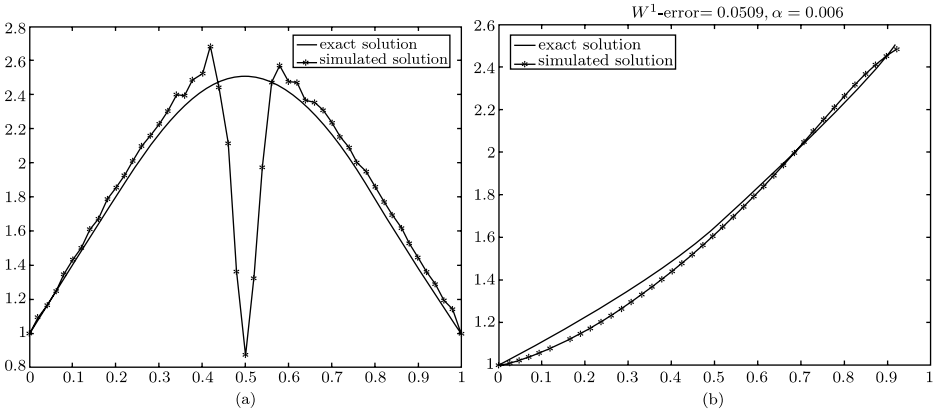


Fig. 4.7 (a) Simulated result for $b(x) = a(u(1,x)) = \exp(10\exp(-1)x(1-x))$; (b) Reconstruction of a diffusion coefficient for the case of non-monotone solution.

References

1. A. B. Bakushinskii, On the rate of convergence of iterative processes for nonlinear operator equations (Russian), *Zh. Vychisl. Mat. Mat. Fiz.*, **38**, 559–563, 1998; translation in *Comput. Math. Math. Phys.*, **38**, 538–542, 1998.
2. H. T. Banks and K. Kunisch, Estimation techniques for distributed parameter systems, *Systems and Control: Foundations & Applications*, **1**, Birkhäuser Boston Inc., 1989.
3. H. T. Banks and K. A. Murphy, Estimation of nonlinearities in parabolic models for growth, predation, and dispersal of populations, *J. Math. Anal. Appl.*, **141**, 580–602, 1989.
4. C. Bardos, D. Brézis and H. Brezis, Perturbations singulières et prolongements maximaux d'opérateurs positifs (French), *Arch. Rational Mech. Anal.*, 69–100, 1973/1974.
5. F. Bauer and T. Hohage, A Lepskij-type stopping rule for regularized Newton methods, *Inverse Problems*, **21**, 1975–1991, 2005.
6. A. Böttcher, B. Hofmann, U. Tautenhahn and M. Yamamoto, Convergence rates for Tikhonov regualrization from different kind of smootheness conditions, *Appl. Anal.*, **85**, 555–578, 2006.
7. G. Bruckner, J. Elschner and M. Yamamoto, An optimization method for grating profile reconstruction, *Progress in Analysis*, Vol. I, II Berlin 1391–1404 World Sci. Publishing River Edge NJ 2003, 2001.
8. A. L. Bukhgeim, J. Cheng and M. Yamamoto, Stability for an inverse boundary problem of determining a part of a boundary, *Inverse Problems*, **15**, 1021–1032, 1999.
9. H. Cao, Discretized Tikhonov regularization for a naturally linearized parameter identification problem, *J. Complexity*, **21**, 864–877, 2005.
10. H. Cao and S. V. Pereverzev, Natural linearization for the identification of a diffusion coefficient in a quasi-linear parabolic system from short-time observations, *Inverse Problems*, **22**, 2311–2330, 2006.
11. H. Cao and S. V. Pereverzev, Balancing principle for the regularization of elliptic Cauchy problems, *Inverse Problems*, **23**, 1943–1961, 2007.
12. P. Deuflhard, H. W. Engl and O. Scherzer, A convergence analysis of itera tive methods for the solution of nonlinear ill-posed problems under affinely invariant conditions, *Inverse Problems*, **14**, 1081–1106, 1998.

13. P. DuChateau, R. Thelwell and G. Butters, Analysis of an adjoint problem approach to the identification of an unknown diffusion coefficient, *Inverse Problems*, **20**, 601–625, 2004.
14. H. W. Engl, P. Fusek and S. V. Pereverzev, Natural linearization for the identification of nonlinear heat transfer laws, *J. Inverse Ill-Posed Probl.*, **13**(3–6), 567–82, 2005.
15. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Dordrecht, Kluwer, 1996.
16. H. W. Engl and J. Zou, A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction, *Inverse Problems*, **16**, 1907–1923, 2000.
17. C. W. Groetsch and A. Neubauer, Convergence of a general projection method for an operator equation of the first kind, *Houston J. Math.*, **14**, 201–208, 1988.
18. M. Hanke, A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems, *Inverse Problems*, **13**(1), 79–95, 1997.
19. M. Hanke and O. Scherzer, Error analysis of an equation error method for the identification of the diffusion coefficient in a quasi-linear parabolic differential equation, *SIAM J. Appl. Math.*, **59**, 1012–1027, 1999.
20. F. Hansen, Operator inequaliteis associated with Jense's inequality, *Survery on Classical Inequalites*, T. M. Rassias, Ed., Kluwer Academic Publishers Group: Dordrecht, 67-98, 2000.
21. M. Hegland, An optimal order regularization mehtod which does not use additional smoothness assumptions, *SIAM J. Numer. Anal.*, **29**, 1446–1461, 1992.
22. T. Hohage, Regularization of exponentially ill-posed problems, *Numer. Funct. Anal. Optim.*, **21**, 439–464, 2000.
23. V. Isakov, Inverse problems for partial differential equations, *Applied Mathematical Sciences*, **127**, Springer-Verlag, New York, 1998.
24. V. Ivanov and T. Korolyuk, Error estimates for solutions of incorrectly posed linear problems, *USSR Compat. Math. Math. Phys.*, **9**, 35–49, 1969.
25. B. Kaltenbacher, A projection-regularized Newton method for nonlinear ill-posed problems and its application to parameter identification problems with finite element discretization, *SIAM J. Numer. Anal.*, **37**(6), 1885–1908, 2000.
26. B. Kaltenbacher and J. Schöberl, A saddle point variational formulation for projection-regularized parameter identification, *Numer. Math.*, **91**(4), 675–697, 2002.
27. B. Kaltenbacher, Some Newton type methods for the regularization of nonlinear ill-posed problems, *PhD Dissertation*, 1996.
28. T. Kärkkäinen, A linearization technique and error estimates for distributed parameter identification in quasilinear problems, *Numer. Funct. Anal. Optim.*, **17**(3–4), 345–364, 1996.
29. P. Kügler, A derivative free Landweber method for parameter identification in elliptic partial differential equations with application to the manufacture of car windshields, *PhD Dissertation*, 2003.
30. Y. Keung and J. Zou, Numerical identifications of parameters in parabolic systems, *Inverse Problems*, **14**, 83–100, 1998.
31. R. D. Lazarov, S. Lu and S. V. Pereverzyev, On the balancing principle for some problems of numerical analysis, *Numer. Math.*, **106**, 659–689, 2007.
32. G. R. Luecke and K. R. Hickey, Convergence of approximate solutions of an operator equation, *Houston J. Math.*, **11**, 345–354, 1985.
33. B. A. Mair and F. H. Ruymgaart, Statistical inverse estimation in Hilbert scales, *SIAM J. Appl. Math.*, **56**, 1424–1444, 1996.
34. P. Mathé and B. Hofmann, How general are general source conditions, *Inverse Problems*, **24**(1), 2008.
35. P. Mathé and S. V. Pereverzev, Regularization of some linear ill-posed problems with discretized random noisy data, *Math. Comp.*, **75**(256), 1913–1929, 2006.
36. P. Mathé and S. V. Pereverzev, Geometry of linear ill-posed problems in variable Hilbert scales, *Inverse Problems*, **19**, 789–803, 2003.

37. P. Mathé and S. V. Pereverzev, Discretization strategy for linear ill-posed problems in variable Hilbert scales, *Inverse Problems*, **19**(6), 1263–1277, 2003.
38. P. Mathé and S. V. Pereverzev, Moduli of continuity for operator valued functions, *Numer. Funct. Anal. Optim.*, **23**(5–6), 623–631, 2002.
39. M. T. Nair, E. Schock and U. Tautenhahn, Morozov's discrepancy principle under general source conditions, *Z. Anal. Anwendungen*, **22**, 199–214, 2003.
40. M. T. Nair, S. V. Pereverzev and U. Tautenhahn, Regularization in Hilbert scales under general smoothing conditions, *Inverse Problems*, **21**, 1851–1869, 2005.
41. F. Natterer, Regularisierung schlecht gestellter probleme durch projektionsverfahren (German), *Numer. Math.*, **28**, 329–341, 1977.
42. S. S. Pereverzyev, R. Pinnau and N. Siedow, Regularized fixed-point iterations for nonlinear, *Inverse Problems*, **22**, 1–22, 2006.
43. S. V. Pereverzev and E. Schock, Morozov's discrepancy principle for Tikhonov regularization of severely ill-posed problems in finite-dimensional subspaces, *Numer. Funct. Anal. Optim.*, **21**, 901–916, 2000.
44. S. V. Pereverzev and E. Schock, On the adaptive selection of the parameter in regularization of ill-posed problems, *SIAM J. Numer. Anal.*, **43**, 2060–2076, 2005.
45. A. Pietsch, Operator ideals, *Mathematische Monographien [Mathematical Monographs]*, **16**, 451, 1978.
46. T. I. Seidman, Nonconvergence results for the application of least-squares estimation to ill-posed problems, *J. Optim. Theory Appl.*, **30**, 535–547, 1980.
47. X. C. Tai and T. Kärkkäinen, Identification of a nonlinear parameter in a parabolic equation from a linear equation, *Mat. Apl. Comput.*, **14**, 157–184, 1995.
48. U. Tautenhahn, Optimality for ill-posed problems under general source conditions, *Numer. Funct. Anal. Optim.*, **19**, 377–398, 1998.
49. G. M. Vainikko and A. Y. Veretennkov, *Iteration Procedures in Ill-Posed Problems*, Nauka, Moscow, 1986 (in Russian).
50. G. M. Vainikko, On the discretization and regularization of ill-posed problems with noncompact operators, *Numer. Funct. Anal. Optim.*, **13**(3–4), 381–396, 1992.

Chapter 5
Extrapolation Techniques of Tikhonov Regularization

Tingyan Xiao, Yuan Zhao and Guozhong Su

Abstract. The numerical solution of inverse problems using Tikhonov's regularization methods requires a huge amount of computations in iterative processes. It can employ extrapolation techniques to accelerate the convergence process or to improve accuracy of the regularized solution. This chapter aims to introduce some main extrapolation methods that have been studied for solving linear inverse problems in detail. Our emphasis is to discuss related technical problems, to propose a new extrapolation algorithm based on the Hermitian interpolation and to present results of numerical experiments for showing the merits of extrapolated regularization methods.

5.1 Introduction

Since the 60's of last century, the theory and techniques of regularization are well developed for linear inverse problems. So far, a great amount of research work has focused on the development of appropriate strategies for selecting regularization parameter with its fast numerical implementation (see [5, 11, 14] and references therein).

Numerically, selecting or determining a reasonable regularization parameter $\alpha^* > 0$ with some choice rules is an iterative process which results in a sequence of the regularized solutions, say, written by $\{u_{\alpha_j}\}_{j=0}^{\infty}$. This is to say, we have a mapping from $\{\alpha_j\}_{j=0}^{\infty}$ to a set of vector-valued functions, $\{u_{\alpha_j}\}_{j=0}^{\infty}$. Of course, we hope that the iterative process can be speeded up, or equivalent, the approximation of the limit function u_{α^*} is with as high accuracy as possible and can be obtained in only a few finite steps. From a viewpoint of scientific computing, the employment of some acceleration techniques can serve for this purpose, since

Tingyan Xiao, Yuan Zhao and Guozhong Su
School of Sciences, Hebei University of Technology, Tianjin 300130, China.
e-mail: ty_xiao@hebut.edu.cn, zhaoyuan135821@163.com, guozhong_su@hebut.edu.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

several extrapolation methods have been widely and successfully used in many well-posed problems [1]. However, as Hämarik pointed out that "extrapolation for increasing the accuracy of regularization methods is much less studied" [7].

The early stage efforts to this end may be traced back to the work of Saidurov, 1973 [13], Groetsch, 1979 [9] and Marchuk, 1983 [12]. Since then, after more than ten years' silence, the above idea again attracts researchers' attention; among which we should speak of Brezinski and Hämarik's constributions [2, 3] and [7, 8].

Although, Marchuk,Brezinski and Hämarik used some kinds of linear combination of the regularized solutions to approximate the desired solution, they represent two different directions for applying the current extrapolation techniques to regularization methods. The first direction is due to Marchuk, 1983 [12] who employs Lagrange interpolation polynomial as the extrapolation tool, and Brezinski, 1998 [2] who adopts rational interpolation polynomial to construct the extrapolation scheme. From the viewpoint of approximation, Brezinski's improvement is obvious since the rational interpolation is usually better than the algebraic polynomial interpolation, so it should be better for the extrapolation.

Hämarik's work represents another direction. His extrapolation schemes are based on a proper selection of the combination coefficients which can eliminate the leading terms in the related error expansions and guarantee higher qualification for extrapolated method than the original regularization method. He not only presented a complete set of extrapolated algorithms for Lavrentiev and Tikhonov regularization methods, and their iterative variant version, but also suggested the rules for choosing appreciate regularization parameters in the extrapolated formulas [7, 8], whereas these rules had not been given previously. Meanwhile, it is reported that a Matlab regularization toolbox, based on the procedures presented in [2], is under construction.

This chapter aims to introduce some major extrapolation methods that have been studied for solving linear inverse problems in detail, including general extrapolation procedures; several concrete extrapolated schemes and the choice criterion of the extrapolation parameters. Moreover, a new extrapolation algorithm based on Hermitian interpolation will be given and the numerical experiments are performed to demonstrate their merits and effectiveness.

The chapter is organized as follows: Section 5.2 lists some notations and preliminaries; Section 5.3 discusses about extrapolated regularization based on vector-valued function approximation; Section 5.4 discusses about extrapolated regularization based on improvement of regularizing qualification; Section 5.5 studies the choice of parameters in the extrapolated regularizing approximation; Section 5.6 performs numerical experiments, and finally, conclusions are given in Section 5.7.

5.2 Notations and preliminaries

Let us consider linear ill-posed inverse problem modelled as an operator equation of the first kind

$$Au = f, \tag{5.2.1}$$

where A is a linear compact operator between Hilbert spaces U and F, and its range $R(A)$ is non-closed. Instead of exact $f \in F$, noisy data f^δ is available with an error level such that $||f - f^\delta|| \leq \delta$. Because of its nature of ill-posedness, solving this kind of problem is a difficult task and some regularization methods (strategies) should be employed.

A class of linear regularization methods can be described by using spectral theory for self-adjoint operators [5]. Let $\{E_\lambda\}$ be a spectral family for A^*A. Then the regularized approximation for noise data f^δ with $\|f - f^\delta\| \leq \delta$ can be formed by

$$u_\alpha^\delta = \int g_\alpha(\lambda) dE_\lambda A^* f^\delta := \mathbf{R}_\alpha f^\delta. \tag{5.2.2}$$

In (5.2.2), the function $g_\alpha(\lambda)$ is called the generating function which is at least piecewise continuous on $[0, \|A\|^2]$ and satisfies the following conditions: for all $\alpha > 0$,

$$|\lambda g_\alpha(\lambda)| \leq c_1, \tag{5.2.3}$$

$$\lim_{\alpha \to 0} g_\alpha(\lambda) = \lambda^{-1}, \text{ for } \lambda \in (0, \|A\|^2], \tag{5.2.4}$$

$$\sup_{0 \leq \lambda \leq a} |g_\alpha(\lambda)| \leq c_2 \alpha^{-1}, \tag{5.2.5}$$

and

$$\sup_{0 \leq \lambda \leq a} \lambda^\mu |r_\alpha(\lambda)| \leq c_3 \alpha^\mu \qquad 0 \leq \mu \leq \mu_0, \tag{5.2.6}$$

where $a = ||A||^2$, μ, μ_0 and c_i $(i = 1, 2, 3)$ are positive constants, and $r_\alpha(\lambda) = 1 - \lambda g_\alpha(\lambda)$ in (5.2.6). Thus u_α^δ can be computed in a stable way.

The maximum value of μ_0, for which the inequality (5.2.6) holds is called the *qualification* of regularization method $(g_\alpha(A^*A)A^*, \alpha)$ or regularization operator $\mathbf{R}_\alpha$.

As we know, one of the most widely used methods in applications is Tikhonov regularization in which the regularized solution is given by

$$u_\alpha^\delta = (\alpha I + A^*A)^{-1} A^* f^\delta \tag{5.2.7}$$

where $g_\alpha(\lambda) = \frac{1}{\alpha + \lambda}$, $\mu_0 = 1$ and I is the identity operator. The accuracy of regularized approximation can be increased by iteration. Let $m \in \mathbb{N}$ be fixed integer number and $u_{\alpha,0}^\delta := 0$ be the initial approximation. Then we get the m-iterated Tikhonov approximation $u_{\alpha,m}^\delta$ through iteratively computing the approximations

$$u_{\alpha,i}^\delta = (\alpha I + A^*A)^{-1}(A^* f^\delta + \alpha u_{\alpha,i-1}^\delta) \quad (i = 1, \ldots, m) \tag{5.2.8}$$

In this method, $g_{\alpha,m}(\lambda) = \frac{(\lambda+\alpha)^m - \alpha^m}{\lambda(\alpha+\lambda)^m}$ and the qualification of $u^\delta_{\alpha,m}$ is $\mu_0 = m$.

Below, we summarize some important properties for the above regularization methods.

Proposition 5.2.1. *With a priori choice rule of the parameter:*

$$\alpha(\delta) = \bar{c}(\delta/\rho))^{\frac{2}{2\mu+1}},$$

the regularization method $(g_\alpha(A^*A)A^*, \alpha)$ is of optimal order in $\mathcal{X}_{\mu,\rho} := \{u \in U |\ u = (A^*A)^\mu w,\ ||w|| \le \rho\}$ for $0 < \mu \le \mu_0$. It means that if $u \in \mathcal{X}_{\mu,\rho}$ and $||Au - f^\delta|| \le \delta$, then the error estimate

$$||g_\alpha(A^*A)A^*f^\delta - u^+|| \le c\delta^{\frac{2\mu}{2\mu+1}}\rho^{\frac{1}{2\mu+1}} \tag{5.2.9}$$

holds for $0 < \mu \le \mu_0$; here u^+ is the generalized solution of (5.2.1).

Proposition 5.2.2. *The regularized solution of Tikhonov's method u^δ_α determined by (5.2.7) is infinitely differentiable at every $\alpha > 0$, which satisfies the following equations:*

$$(A^*A + \alpha I)\frac{d^k u^\delta_\alpha}{d\alpha^k} = -k\frac{d^{(k-1)}u^\delta_\alpha}{d\alpha^{(k-1)}}, \quad k = 1, 2, \cdots. \tag{5.2.10}$$

It is not difficult to verify that the above differentiability of the regularized solutions also holds for the iterated Tikhonov regularization solutions and the discrete regularization solutions. So, the Tikhonov's regularized solution (no matter continuous or discrete type), as function of parameter $\alpha > 0$, can be approximated by some interpolating polynomials which can serve as the basis in constructing extrapolation algorithms for regularization.

Now we consider extrapolation of standard regularization methods. Suppose its generating function $g_\alpha(\lambda)$ satisfies (5.2.3)–(5.2.6) with $\mu_0 < \infty$. Let the sequences of parameters $0 < \alpha_j < \alpha$ $(j = 1, 2 \ldots k)$ be given with $\alpha_j \ne \alpha_i$ for $j \ne i$. According to the principle of general extrapolation [1], let $d_j \in \mathbb{R}$ fulfill $\sum\limits_{j=1}^{k} d_j = 1$ and d_j are independent of α. Consider the linear combination of regularized approximations

$$u^\delta_k(\alpha) = \sum_{j=1}^{k} d_j u^\delta_{\alpha_j} = \int \sum_{j=1}^{k} d_j g_{\alpha_j}(\lambda) dE_\lambda A^* f^\delta. \tag{5.2.11}$$

Denoting that

$$g_{\alpha,k}(\lambda) = \sum_{j=1}^{k} d_j g_{\alpha_j}(\lambda); \quad q_j = \alpha_j/\alpha, \tag{5.2.12}$$

we have $r_{\alpha,k}(\lambda) = 1 - \lambda g_{\alpha,k}(\lambda) = \sum_{j=1}^{k} d_j r_{\alpha_j}(\lambda)$. Assume k is fixed, and α is a regularization parameter. Then the function $g_{\alpha,k}(\lambda)$ satisfies the following

inequality for $0 \le \mu \le \mu_0$:

$$|\lambda g_{\alpha,k}(\lambda)| \le c_1 \sum_{j=1}^{k} |d_j| := c_1 \Lambda_k, \tag{5.2.13}$$

$$\lim_{\alpha \to 0} g_{\alpha,k}(\lambda) = \sum_{j=1}^{k} \lim_{\alpha \to 0} d_j g_{\alpha_j}(\lambda) = 1/\lambda \quad (\lambda \neq 0), \tag{5.2.14}$$

$$\sup_{\lambda \in [0,a]} |g_{\alpha,k}(\lambda)| \le \sum_{j=1}^{k} \sup_{\lambda \in [0,a]} |d_j g_{\alpha_j}(\lambda)| \le c_2 \sum_{j=1}^{k} |d_j| (q_j \alpha)^{-1}, \tag{5.2.15}$$

and

$$\sup_{\lambda \in [0,a]} \lambda^{\mu} |r_{\alpha,k}(\lambda)| \le c_3 \sum_{j=1}^{k} |d_j| \alpha_j^{\mu} = c_3 \sum_{j=1}^{k} (|d_j| q_j^{\mu}) \alpha^{\mu}. \tag{5.2.16}$$

So a proper choice of $d_j = d_j(q_1, q_2, \ldots, q_k)$, $j = 1, 2, \ldots, k$ may guarantee conditions similar to (5.2.3)–(5.2.6) and the qualification (written by μ^E) of the approximation $u_k^{\delta}(\alpha)$ is higher than μ_0, the qualification of initial (non-extrapolated) Tikhonov regularization method. In other words, we will get an extrapolated regularization method $\{g_{\alpha,k}(A^*A)A^*, \alpha\}$, or, an extrapolation-type regularization operator:

$$\mathbf{R}_{\alpha,k}^{E} = \int \sum_{j=1}^{k} d_j g_{\alpha_j}(\lambda) dE_{\lambda} A^* \longrightarrow u_k^{\delta}(\alpha) = \mathbf{R}_{\alpha,k}^{E} f^{\delta}, \tag{5.2.17}$$

the qualification μ^E of $\mathbf{R}_{\alpha,k}^{E}$ will be no less than μ_0. Of course, as a regularization method, the regularization parameter α and the index k must be determined properly.

5.3 Extrapolated regularization based on vector-valued function approximation

Let us consider concrete extrapolation schemes of Tikhonov regularization method, which, in this section, are all based on vector-valued function approximation. For convenience, we describe them in a discrete form.

After the discretization, equation (5.2.1) turns out to be the following ill-posed system of linear equations. For simplicity, we still write it as

$$Au = f, A \in C^{m \times n}, \quad u \in C^n, \ f \in C^m. \tag{5.3.1}$$

For true data f, the regularized Euler's equation is given by

$$(A^*A + \alpha I)u_\alpha = A^*f. \tag{5.3.2}$$

Proceeding from equation (5.3.2), we describe three schemes as follows:

5.3.1 Extrapolated scheme based on Lagrange interpolation

According to Marchuk's original idea [12], one can find a unitary matrix P and a diagonal matrix $\Lambda = diag(\lambda_1 \dots \lambda_n)$, where $\lambda_i \geq 0$, such that $A^*A = P\Lambda P^*$. Let $y_\alpha = P^*u_\alpha$, $F = P^*A^*f$, then (5.3.2) is equivalent to

$$(\Lambda + \alpha I)y_\alpha = F \tag{5.3.3}$$

and the components of the vector y_α are defined by $y_{\alpha,i} = F_i/(\lambda_i + \alpha)$. Note that the function $\gamma(\alpha) = \frac{1}{\lambda+\alpha}$ exists. When $\lambda > 0$, the function $\gamma(\alpha)$ is infinitely differentiable for any $\alpha > 0$ and $|\gamma^{(k)}(\alpha)| \leq k!/\lambda^{k+1}$.

Let $\alpha_1 > \alpha_2 > \cdots > \alpha_{k+1}$ be a decreasing sequence of values of α. Using the Lagrangian formula yields[4]

$$\gamma(\alpha) = \sum_{j=1}^{k} l_j(\alpha)\gamma(\alpha_j) + Q(\alpha), \tag{5.3.4}$$

where

$$l_j(\alpha) = \prod_{i=1\ i\neq j}^{k} \frac{\alpha - \alpha_i}{\alpha_j - \alpha_i}, \quad j = 1, 2, \cdots, k. \tag{5.3.5}$$

Substituting α_{k+1} into (5.3.4), we get

$$\gamma(\alpha_{k+1}) = \sum_{j=1}^{k} l_j(\alpha_{k+1})\gamma(\alpha_j) + Q(\alpha_{k+1}). \tag{5.3.6}$$

and then we can estimate the remainder: $|Q(\alpha_{k+1})| \leq \lambda^{-k-1} \prod_{j=1}^{k} \alpha_j$.

Let $\bar{\lambda} = \min_{\{\lambda_i \neq 0\}}\{\lambda_i\}$, Applying (5.3.6) to the solution of (5.3.3) yields

$$y_{\alpha_{k+1}} = \sum_{j=1}^{k} l_j(\alpha_{k+1})y_{\alpha_j} + q(\alpha_{k+1}), \tag{5.3.7}$$

where

$$\|q\| \leq \bar{\lambda}^{-k-1} \left(\prod_{j=1}^{k} \alpha_j \right) \|F\|.$$

Using the above estimates and the norm-preserving property of unitary matrices result in $\|A^*f\| = \|P^*A^*f\| = \|F\|$, $\|Pq\| = \|q\|$, it follows that the solution of

the regularized problem (5.3.2) obeys the relation

$$u_{\alpha_{k+1}} = \sum_{j=1}^{k} l_j(\alpha_{k+1}) u_{\alpha_j} + r(\alpha_{k+1}), \quad \|r\| \le c_1 \prod_{j=1}^{k} \alpha_j, \tag{5.3.8}$$

where c_1 is a constant which does not depend on α_j. Of course it is more interested in the case of $\alpha_{k+1} = 0$. Since $u^+ = \lim_{\alpha \to 0} u_\alpha = u_0$, $\alpha_{k+1} = 0$ yields

$$u^+ - \sum_{j=1}^{k} \lambda_j(0) u_{\alpha_j} = r(0); \quad \|r(0)\| \le \kappa_1 \prod_{j=1}^{k} \alpha_j, \tag{5.3.9}$$

where κ_1 is a constant independent of α_j.

Now we construct a linear combination of regularized solutions of (5.3.2), $u^\delta_{\alpha_j}$ with noise data f^δ using the sequence $\{\alpha_j\}_{j=1}^k$:

$$v^\delta_{\alpha^*,k} = \sum_{j=1}^{k} d_j u^\delta_{\alpha_j} \quad (d_j = l_j(0), j = 1, 2, \ldots, k), \tag{5.3.10}$$

where $\alpha^* = \max\{\alpha_j\}$. As to the error estimation between $v^\delta_{\alpha^*,k}$ and u^+, we could give a modified version of **Theorem 1.2** of [12] as follows.

Theorem 5.3.1. *Let u^+ be the generalized solution of system (5.3.1) and an approximate equation be in the form $Au = f^\delta$ such that $||f - f^\delta|| \le \delta$. Let $v_{\alpha^*,k} = \sum_{j=1}^{k} d_j u_{\alpha_j}$, $v^\delta_{\alpha^*,k}$ and d_j be as in (5.3.10), respectively. Then*

$$\|u^+ - v^\delta_{\alpha^*,k}\| \le \kappa_1 \prod_{j=1}^{k} \alpha_j + \sum_{j=1}^{k} |d_j| \delta / \sqrt{\alpha_j} \tag{5.3.11}$$

Proof. Obviously, from the following equality

$$(A^*A + \alpha_j I)(u_{\alpha_j} - u^\delta_{\alpha_j}) = A^*(f - f^\delta) \tag{5.3.12}$$

and $\|(A^*A + \alpha_j I)^{-1} A^*\| \le 1/\sqrt{\alpha_j}$, we have

$$\|v_{\alpha^*,k} - v^\delta_{\alpha^*,k}\| = \|\sum_{j=1}^{k} d_j (u_{\alpha_j} - u^\delta_{\alpha_j})\| \le \sum_{j=1}^{k} |d_j| \delta / \sqrt{\alpha_j}. \tag{5.3.13}$$

Combining the inequality in (5.3.9), it follows immediately that

$$\|u^+ - v^\delta_{\alpha^*,k}\| \le \|u^+ - v_{\alpha^*,k}\| + \|v_{\alpha^*,k} - v^\delta_{\alpha^*,k}\| \le \kappa_1 \prod_{j=1}^{k} \alpha_j + \sum_{j=1}^{k} |d_j| \delta / \sqrt{\alpha_j}. \tag{5.3.14}$$

The proof of the theorem is finished. □

Remark 5.3.2. *Obviously, to guarantee the numerical stability of the extrapolated formula (5.3.10), one should choose α_j such that $d_j = l_j(0)$ remains bounded when all of the α_j tend to zero. To this end Marchuk presented a general choice rule:*

$$\alpha_j/\alpha_{j+1} \geq \theta > 1, \quad j = 1, 2, \ldots, k-1, \tag{5.3.15}$$

with θ independent of α_j. We refer (5.3.15) to Marchuk condition. *This condition ensures that [12]*

$$|d_j| \leq \left(\frac{\theta}{\theta - 1}\right)^k, \ j = 1, 2, \cdots, k. \tag{5.3.16}$$

Corollary 5.3.3. *Suppose integer k is fixed and α_j $(j = 1, 2, \cdots)$ satisfy condition (5.3.15). The other notations and conditions are the same as those in Theorem 5.3.1. If we choose $\alpha^* = \alpha^*(\delta) = \delta^\nu, \quad 0 < \nu < 2$, then we obtain*

$$\lim_{\delta \to 0} v^\delta_{\alpha^*(\delta),\, k} = u^+. \tag{5.3.17}$$

Corollary 5.3.4. *Marchuk gave a numerical test for a singular and incompatible system of ten linear equations with ten unknowns. With $\alpha_1 = 0.01$, $\alpha_2 = \alpha_1/2$, $\alpha_3 = \alpha_1/3$, the numerical results are very satisfactory; but he did not give a discussion over how to select parameter α_1 and how to determine the integer k. To simplify notation, we abbreviate this scheme as* **E-Tik-Lag**.

5.3.2 Extrapolated scheme based on Hermitian interpolation

We give another form of the extrapolation method using the differentiability of Tikhonov regularized solution in this subsection.

Letting $du_\alpha = \frac{du_\alpha}{d\alpha}$, from (5.3.2), it follows

$$(A^*A + \alpha I)du_\alpha = -u_\alpha. \tag{5.3.18}$$

Suppose $\{\alpha_j\}_{j=1}^k$ are given for positive integer k. Assume $\{u_{\alpha_j}\}_{j=1}^k$ and their first derivatives $\{du_{\alpha_j}\}_{j=1}^k$ have been prescribed. Since u_α and du_α are vector-valued functions with respect to variable $\alpha > 0$, we can construct related Hermitian interpolation approximation.

Hermitian interpolated polynomial is usually given by

$$H_{2k-1}(\alpha) = \sum_{j=1}^{k} \left(A_j(\alpha)u_{\alpha_j} + B_j(\alpha)du_{\alpha_j}\right), \tag{5.3.19}$$

where $\{A_j(\alpha)\}_{j=1}^k$ and $\{B_j(\alpha)\}_{j=1}^k$ satisfy the following conditions

$$A_j(\alpha_i) = \delta_{ji}, \quad A'_j(\alpha_i) = 0, \ \ (\forall i, j),$$
$$B_j(\alpha_i) = 0, \quad B'_j(\alpha_i) = \delta_{ji}, \ \ (\forall i, j),$$

in which, δ_{ji} denote the Kronecker symbol. With the aid of the functions $l_j(\alpha) = \prod_{i \neq j}(\alpha - \alpha_i)/(\alpha_j - \alpha_i)$, we have

$$A_j(\alpha) = \left[1 - 2(\alpha - \alpha_j)l'_j(\alpha_j)\right] l_j^2(\alpha), \quad j = 1 : k,$$
$$B_j(\alpha) = (\alpha - \alpha_j)l_j^2(\alpha), \quad j = 1 : k.$$

So $H_{2k-1}(\alpha_j) = u_{\alpha_j}$, $H'_{2k-1}(\alpha_j) = du_{\alpha_j}$ for j from 1 to k. Letting $\xi_j = A_j(0)$, $\eta_j = B_j(0)$, we have

$$\xi_j = \left[1 + 2\alpha_j l'_j(\alpha_j)\right] l_j^2(0), \quad \eta_j = -\alpha_j l_j^2(0). \tag{5.3.20}$$

Therefore, a new approximation to u_α can be given by

$$v_{\alpha^*, k} = \sum_{j=1}^{k} \left(\xi_j u_{\alpha_j} + \eta_j du_{\alpha_j}\right). \tag{5.3.21}$$

Similarly, we have

$$(\Lambda + \alpha I)dy_\alpha = -y_\alpha. \tag{5.3.22}$$

The components of the vector dy_α are $dy_{\alpha,i} = -y_{\alpha,i}/(\lambda_i + \alpha)^2$.

In analogy with the inference in §5.3.1, we can obtain the following results:

Theorem 5.3.5. *Let u^+ be the generalized solution of (5.3.1), $v_{\alpha^*, k} = \sum_{j=1}^{k}(\xi_j u_{\alpha_j} + \eta_j du_{\alpha_j})$, where ξ_j, η_j are given in (5.3.20). Then*

$$\|u^+ - v_{\alpha^*, k}\| \leq \kappa_2 \times \prod_{j=1}^{k} \alpha_j^2 \leq \kappa_2 (\alpha^*)^{2k}, \tag{5.3.23}$$

where κ_2 is a constant.

Theorem 5.3.6. *If we know only $Au = f^\delta$, $\|f - f^\delta\| \leq \delta$ and $v_{\alpha^*,k}$ are as in Theorem 5.3.5. Let $v^\delta_{\alpha^*, k} = \sum_{j=1}^{k}(\xi_j u^\delta_{\alpha_j} + \eta_j du^\delta_{\alpha_j})$. Then*

$$||u^+ - v^\delta_{\alpha^*, k}|| \leq \kappa_2 (\alpha^*)^{2k} + \sum_{j=1}^{k} |\xi_j| |\delta/\sqrt{\alpha_j} + \sum_{j=1}^{k} |\eta_j| \delta/\alpha_j, \tag{5.3.24}$$

where κ_2 is a constant as in (5.3.23).

Corollary 5.3.7. *Suppose integer k is fixed, α_j for all j satisfy condition (5.3.15), $\alpha^* = \alpha^*(\delta) = \delta^\nu$, $0 < \nu < 1$ and $v^\delta_{\alpha^*,k}$ is given in Theorem 5.3.6. Then*

$$\lim_{\delta \to 0} v^{\delta}_{\alpha^*(\delta),\, k} = u^{+}. \tag{5.3.25}$$

To be simple, we abbreviate the above scheme as **E-Tik-Her**.

5.3.3 Extrapolation scheme based on rational interpolation

Based on the SVD realization of the Tikhonov regularization, Brezinski presented several extrapolation methods using rational approximation. We just introduce one of them, **Algorithm 1.a: Full case**[2] which possesses lower computational complexity than the complexity of others. Using a rational and vector-valued function as a candidate,

$$R_k(\alpha) = \sum_{i=1}^{k} \frac{1}{b_i + \alpha} w_i, \quad k \le p, \tag{5.3.26}$$

where b_i's and w_i's are unknown scales and vectors. Moreover, $\{w_i\}$ are linearly independent of each other. Suppose that $\{\alpha_i\}_{i=0}^{k}$, $\alpha_i \neq \alpha_j$ for $i \neq j$ are given and $u^{\delta}_{\alpha_i}$ for all i are corresponding solutions of $(A^*A + \alpha_i I)\, u = A^* f^{\delta}$, $i = 0$ *to* k. The above unknowns will be determined by imposing the interpolated conditions of $R_k(\alpha_i) = u^{\delta}_{\alpha_i}$, $i = 0, 1, \ldots, k$.

We will extrapolate (5.3.26) at $\alpha = 0$, i.e., we will compute

$$v_k = R_k(0) = \sum_{i=1}^{k} w_i / b_i. \tag{5.3.27}$$

The main task is to determine $\{b_i\}$ and $\{w_i\}$.

First, reducing the sum in (5.3.26) to the same denominator:

$$R_k(\alpha) = P_{k-1}(\alpha)/Q_k(\alpha), \tag{5.3.28}$$

where

$$\begin{aligned} &Q_k(\alpha) = \textstyle\prod_{i=1}^{k}(b_i + \alpha) = \beta_0 + \cdots + \beta_{k-1}\alpha^{k-1} + \alpha^k,\ \beta_i \in \mathbb{R}, \\ &P_{k-1}(\alpha) = \gamma_0 + \cdots + \gamma_{k-1}\alpha^{k-1},\ \gamma_i \in \mathbb{R}. \end{aligned} \tag{5.3.29}$$

Then it results in the following interpolating equations

$$u^{\delta}_{\alpha_i} Q_k(\alpha_i) = P_{k-1}(\alpha_i),\ i = 0, 1, \ldots, k-1 \tag{5.3.30}$$

by Lagrangian formula. We have

$$\begin{aligned} &Q_k(\alpha) = \textstyle\sum_{i=0}^{k} L_i(\alpha) Q_k(\alpha_i), \\ &P_{k-1}(\alpha) = \textstyle\sum_{i=0}^{k-1} \bar{L}_i(\alpha) P_{k-1}(\alpha_i) = \sum_{i=0}^{k-1} \bar{L}_i(\alpha) Q_k(\alpha_i) u^{\delta}_{\alpha_i}, \end{aligned} \tag{5.3.31}$$

where $\{L_i(\alpha)\}$ and $\{\bar{L}_i(\alpha)\}$ are interpolating basic functions.

For $\alpha_k \neq \alpha_j, j = 0$ *to* $k-1$, the combinations of (5.3.30)–(5.3.31) give that

$$\sum_{i=0}^{k-1} \bar{L}_i(\alpha_k) Q_k(\alpha_i) u_{\alpha_i}^{\delta} = Q_k(\alpha_k) u_{\alpha_k}^{\delta}. \tag{5.3.32}$$

Let $s_1, s_2, \ldots, s_p$ be linearly independent vectors. Scalar multiplying (5.3.32) by s_j for $j = 1, 2, \ldots, p$ and setting $c_i = Q_k(\alpha_i)/Q_k(\alpha_k)$ lead to the following linear system

$$\sum_{i=0}^{k-1} \bar{L}_i(\alpha_i) \left(u_{\alpha_i}^{\delta}, s_j\right) c_i = (u_{\alpha_k}^{\delta}, s_j),\ j = 1, 2, \ldots, p. \tag{5.3.33}$$

Solving this system in the least squares sense gives $c_0, c_1, \ldots, c_{k-1}$. Since the Q_k is a monic polynomial and $c_k = 1$, we have the supplementary condition

$$Q_k(\alpha_k) \sum_{i=0}^{k} \left(c_i / \prod_{j=0, j\neq i}^{k} (\alpha_i - \alpha_j) \right) = 1. \tag{5.3.34}$$

So $Q_k(\alpha_k)$ is determined and we obtain the following results

$$\begin{aligned} &Q_k(\alpha_i) = c_i Q_k(\alpha_k), \\ &\beta_0 = Q_k(0) = \textstyle\sum_{i=0}^{k} L_i(0) Q_k(\alpha_i), \\ &\gamma_0 = P_{k-1}(0) = \textstyle\sum_{i=0}^{k-1} \bar{L}_i(0) Q_k(\alpha_i) u_{\alpha_i}^{\delta}. \end{aligned} \tag{5.3.35}$$

With the above preparations, we get the extrapolated regularization approximation

$$u_k^{\delta} = R_k(0) = \gamma_0/\beta_0. \tag{5.3.36}$$

Remark 5.3.8. *For the setting of* $\{\alpha_j\}$*, Brezenski did not give a general rule; but for the vector of* $\{s_1, s_2, \ldots, s_k\}$*, he chose the canonical basis. We abbreviate this scheme as* **E-Tik-Bre**.

Remark 5.3.9. *To select the index k, Brezenski adopted L−curve rule, and he pointed out that "with respect to extrapolation methods, the L-curve proved to be quite effective in the restricted case, but ineffective at all in the full case,* $\cdots$*, an effective criterion to select the optimal value of k in the full case would be of great interest". Thus this problem warrants a careful consideration.*

5.4 Extrapolated regularization based on improvement of regularizing qualification

Now we introduce more general extrapolation schemes for increasing the qualification of regularization methods.

(1) Suppose integer $k \geq 2$ is fixed and let the function $g_\alpha(\lambda)$ satisfy condition (5.2.3)–(5.2.6) with $\mu_0 < +\infty$;

(2) Take a group of distinct regularization parameters: $0 < \alpha_j \neq \alpha_i < \alpha \ for\ i \neq j;\ i, j = 1, 2, \ldots, k$; for example $\alpha_i = \alpha_{i-1}q,\ 0 < q < 1$;

(3) Let $d_j \in \mathbb{R},\ j = 1, 2, \ldots, k$ and satisfy firstly $\sum_{j=1}^{k} d_j = 1$;

(4) Construct a linear combination of the regularized solutions:

$$v_{\alpha,k}^{\delta} = \sum_{j=1}^{k} d_j u_{\alpha_j}^{\delta} = \int \sum_{j=1}^{k} d_j g_{\alpha_j}(\lambda) dE_\lambda A^* f^{\delta} \tag{5.4.1}$$

where $\{u_{\alpha_j}^{\delta}\}_{j=1}^{k}$ are given by (5.2.7). As §5.2 shows, $g_{\alpha,k}(\lambda) = \sum_{j=1}^{k} d_j g_{\alpha_j}(\lambda)$ satisfies the inequalities (5.2.13)–(5.2.16).

Consider the choice of $d_j = d_j(q_1, q_2, \ldots, q_k),\ j = 1, 2, \ldots, k$. Recalling the Tikhonov regularization method and its m-iterated variant, we have $1 - \lambda g_\alpha(\lambda) = (1 + \lambda/\alpha)^{-m} := (1 + \varepsilon)^{-m}$. Using the Taylor series

$$(1 + \varepsilon)^{-m} = \sum_{j=0}^{\infty} c_j \varepsilon^j, \quad c_j = (-1)^j \frac{(m + j - 1)!}{(m - 1)! j!} \quad (|\varepsilon| < 1) \tag{5.4.2}$$

for $\varepsilon = (\lambda/\alpha)^{-1} < 1$ and $l = j + m$, we obtain

$$\begin{cases} 1 - \lambda g_\alpha(\lambda) = (1 + \varepsilon^{-1})^{-m} = \varepsilon^m (1 + \varepsilon)^{-m} = \sum_{l=m}^{\infty} c_{l-m} \varepsilon^l, \\ 1 - \lambda g_{\alpha,k}(\lambda) = \sum_{i=1}^{k} d_i \sum_{l=m}^{\infty} c_{l-m} q_i^l \varepsilon^l = \sum_{l=m}^{\infty} \left[\sum_{i=1}^{k} d_i q_i^l \right] c_{l-m} \varepsilon^l. \end{cases} \tag{5.4.3}$$

If $d_i = d_i(q_1, q_2, \ldots, q_k), i = 1, 2, \ldots, k$ are chosen as the solution of the system of linear equations

$$\begin{cases} \sum_{i=1}^{k} d_i = 1, \\ \sum_{i=1}^{k} d_i q_i^l = 0,\ l = m, m + 1, \ldots, m + k - 2, \end{cases} \tag{5.4.4}$$

then $1 - \lambda g_{\alpha,k}(\lambda) = \mathcal{O}(\varepsilon^{m+k-1}) = \mathcal{O}(\lambda/\alpha)^{-(m+k-1)})$; hence $\mu^E = m + k - 1$. Hämarik obtained the explicit solution of the above equations as follows:

Theorem 5.4.1 ([7]). *Let $u_{\alpha_i}^{\delta}$ with $\alpha_i = q_i \alpha$ be the approximations in $m-$iterated Tikhonov method. Then the combination coefficients of $\{d_j\}_{j=1}^{k}$ in the extrapolated approximation $v_{\alpha,\,k}^{\delta}$ are given by*

$$d_i = \begin{cases} \prod_{j=1, j\neq i}^{k} (\alpha_j/(\alpha_j - \alpha_i)), & (m = 1), \\ D_i / \sum_{j=1}^{n} D_j, \quad D_i = (-1)^{i-1} \prod_{j=1, j\neq i}^{k} \alpha_j^m \prod_{k>l, k\neq i, l\neq i} (\alpha_k - \alpha_l) & (m > 1). \end{cases} \tag{5.4.5}$$

Remark 5.4.2. *Let μ^E be the qualification of the extrapolated approximation $v_{\alpha,k}^{\delta}$, then it can be shown that $\mu^E = m + k - 1 = k \geq 2$. Notice that in the case of $m = 1$ in the above expression, the deduced combination coefficients $\{d_j\}_{j=1}^{k}$ are just the coefficients of extrapolated (original) Tikhonov regularization.*

Remark 5.4.3. *We abbreviate this scheme as* **E-Tik-Häm** *or* **E-Tik-Häm-m**. *Compared with Section 5.3, it can be seen that the* **E-Tik-Lag** *is actually the special case of* **E-Tik-Häm**, *i.e.,* **E-Tik-Häm-1**.

Remark 5.4.4. *It is proven that if the solution is smooth and the noise level δ is known, a proper choice of $k = k(\delta)$ guarantees the accuracy $\mathcal{O}(\delta^{\frac{2k}{2k+1}})$ for the extrapolated Tikhonov approximation versus the accuracy $\mathcal{O}(\delta^{\frac{2}{3}})$ of the Tikhonov approximation. The choice rules for index k and α in* **E-Tik-Häm**-*algorithms are also given in [7, 8], which will be described in the next section.*

Remark 5.4.5. *The* **E-Tik-Häm**-*schemes with their analysis are also suitable for the regularization method with self-adjoint operator, known as* **Lavrentiev** *regularization*[7, 8]. *We would not go into details here.*

5.5 The choice of parameters in the extrapolated regularizing approximation

As a new type of regularization method, there of course exits the problem of choosing regularization parameters: the positive parameter of α and the index k, the number of terms in the combination of $v_{\alpha,k}^{\delta}$.

Certainly, if a value of index $k \geq 2$ is given, some selection rules, say, discrepancy principle [7], L-curve rule [5] and the monotone error rule [7, 8] for the Tikhonov regularization, should be employed since the extrapolated regularized solution as the combination of regularized solutions should satisfy, for example, the feasible condition: $\|Av_{\alpha,k}^{\delta} - f^{\delta}\| \leq \delta$.

The key to the questions in selecting the extrapolation parameters (α and k), is to give a rule to check whether $v_{\alpha,k+1}^{\delta}$ is a more accurate than $v_{\alpha,k}^{\delta}$ when both of them satisfied the same necessary condition (for example, the discrepancy requirements). To this question, [7, 8] hold that one should consider the cases separately, i.e., when one of parameters k and α is fixed, the other parameter is regularization parameter. We will quote a main theorem from [7] later on.

It is well known that for a posteriori choice of the regularization parameter α in the Tikhonov regularzation method, there exist several well-developed rules.

For discrepancy principle, one chooses α such that $\|Au_\alpha^\delta - f^\delta\| = C\delta$ with $C \geq 1$; for the modified discrepancy (MD) principle [7] α_{MD} and the monotone error rule (ME-rule) α_{ME} are chosen by solving equations

$$\begin{aligned} &\left(Au_\alpha^\delta - f^\delta, Au_{2,\alpha}^\delta - f^\delta\right) = C\delta, \\ &\left(Au_\alpha^\delta - f^\delta, Au_{2,\alpha}^\delta - f^\delta\right)/\|Au_{2,\alpha}^\delta - f^\delta\| = C\delta, \end{aligned} \tag{5.5.1}$$

respectively. Here $u_{2,\alpha}^\delta = (A^*A + \alpha I)^{-1}\left(\alpha u_\alpha^\delta + A^* f_\delta\right)$ is the approximation of the iterated Tikhonov method.

1) Give the sequences $\alpha_1, \alpha_2, \ldots,$ (α is fixed) and consider the choice of index k in extrapolated Tikhonov approximation $v_{\alpha,k}^\delta$. Denote $r_k = Av_{\alpha,k}^\delta - f^\delta$ and $C =$ constant > 1. For **E-Tik-Hämm-m** schemes, we have the following theorem.

Theorem 5.5.1. *Let $v_{\alpha,\,k}^\delta = \sum_{l=1}^k d_l u_{\alpha_l}^\delta$ with d_l be determined by (5.4.5). Then the functions*

$$d_D(k) = \|r_k\| > 0,\ d_{ME}(k) = \frac{(r_k + r_{k+1}, r_{k+1})}{(2\|r_{k+1}\|)} > 0 \tag{5.5.2}$$

are monotonously decreasing and

$$d_D(k+1) < d_{ME}(k) < d_D(k), \quad \forall\ k. \tag{5.5.3}$$

Let k_D and k_{ME} be the first numbers with $d_D(k) \leq C\delta$, $d_{ME}(k) \leq C\delta$, respectively. Then $k_D - 1 \leq k_{ME} \leq k_D$ and

$$\|v_{\alpha,\,k}^\delta - u^+\| < \|v_{\alpha,\,k-1}^\delta - u^+\| \quad \textit{for } k = 1, 2, \ldots, k_{ME}. \tag{5.5.4}$$

If the monotonically decreasing infinite sequences $\alpha_1, \alpha_2, \ldots$ satisfy conditions

$$\sum_{i=1}^{\infty} \alpha_i^{-1} = \infty, \quad \alpha_k \leq const \sum_{i=1}^{k-1} \alpha_i^{-1}, \tag{5.5.5}$$

then the existence of finite k_D and k_{ME} is guaranteed; and for $k \in \{k_D, k_{ME}\}$, $\|v_{\alpha,k}^\delta - u^+\| \to 0 (\delta \to 0)$, the error estimate

$$\|v_{\alpha,k}^\delta - u^+\| \leq const \times \delta^{\mu/(\mu+1)} \tag{5.5.6}$$

holds for all $\mu > 0$ under the source condition of

$$u^* \in \mathcal{R}((A^*A)^{\mu/2}). \tag{5.5.7}$$

2) Let $k \geq 2$ and $q_1, q_2, \ldots, q_{k+1}$ be given. The choice of α in extrapolated Tikhonov approximation $v_{\alpha,k}^\delta$ can be determined by **Theorem 2** of [8]. The interested reader can refer to their work for more information.

So the twice checking processes must be performed alternately. Hämarik, R. Palm and T. Raus made extensive numerical experiments [7, 8] and showed that

the above combined checking rules were quite effective; and "the error of extrapolated approximation was in most cases smaller than the error of the Tikhonov approximation" if $u^+ \in \mathcal{R}(A)$.

Of course, we can fully follow the above rules to realize extrapolating regularization method, but sometimes we could also put it in practice by using a simplified implemental strategy.

Instead of performing the twice checking process, we elaborately select an interval of $[\alpha_{min}, \alpha_{max}]$ and generate a sequence: $\alpha_1, \alpha_2, \cdots$; then choose the index k by Theorem 5.5.1. In other words, we apply Theorem 5.5.1 based on some selected node-location: [*interval,sequence*].

For the interval selection, we suggest two types as candidates:
Interval-1: By Theorem 5.3.1 and Theorem 5.3.5, we choose

$$[\alpha_{min}, \alpha_{max}] = [0.0, \delta^{\lambda}],\ 0 < \lambda < 2 \quad or \quad 0 < \lambda < 1. \tag{5.5.8}$$

Interval-2: From [10], and after some calculations, we choose

$$\alpha_{max} = \|A\|^2\delta/(\|f^{\delta}\| - \delta), \quad \alpha_{min} = \delta^2/(4\|u^{\delta}_{\delta^2}\|^2). \tag{5.5.9}$$

Actually, the $\alpha_{\min}$ can be obtained using the relationships

$$\begin{array}{l} M^{\alpha} := \|Au_{\alpha} - f^{\delta}\|^2 + \alpha\|u_{\alpha}\|^2, \forall \alpha > 0, \\ M^{\alpha} > \|Au_{\alpha} - f^{\delta}\|^2, \quad M^{\alpha} \geq 2\sqrt{\alpha}\|Au_{\alpha} - f^{\delta}\|\|u_{\alpha}\|, \end{array} \tag{5.5.10}$$

and taking $\|Au_{\alpha} - f^{\delta}\| \approx \delta,\ \|u_{\alpha}\| \approx \|u^{\delta}_{\delta^2}\|$, which results in $2\delta\sqrt{\alpha}\|u^{\delta}_{\delta^2}\| \gtrapprox \delta^2$ and we can take $\alpha_{min} = \delta^2/(4\|u^{\delta}_{\delta^2}\|^2)$.

Based on one of the above interval settings, the sequence of α_i can be chosen as the following four groups:

Sequence 1: $\alpha_1 = \alpha_{max}$; $\alpha_i = \alpha_1/i,\ i = 2, 3, \cdots$;

Sequence 2: $\alpha_1 = \alpha_{max}$, $\alpha_i = \alpha_{i-1}q$; $i = 2, 3, \cdots$; $0 < q < 1$;

Sequence 3: $\alpha_1 = \alpha_{min} = c(\delta/\rho)^{2/(2\mu+1)}$; $q = 1.01$; $\alpha_i = \alpha_{i-1} * q$; $i = 2, 3, \cdots$, where $c = c(k)$ and $\rho = \|u^+\| = \|u_{true}\|$ are constants;

Sequence 4: $\alpha_j = \frac{\alpha_{min}+\alpha_{max}}{2} + \frac{\alpha_{max}-\alpha_{min}}{2}\cos(\frac{2j-1}{2k}\pi), (j = 1, \ldots k)$.

It can be verified that the combination coefficients $\{d_j\}$ obtained from the above sequences entirely satisfy the Marchuk's condition (5.3.16). Any combination of the intervals and the sequences can generate a node-location; for example, (2,4) denotes Chebyshev node-system based on the second interval setting. As we know, the polynomial interpolation at Chebyshev node-system is nearly the best approximation [4], so we can expect that the extrapolated schemes based on (2,4) may reveal the merits of the above nearly best approximation.

5.6 Numerical experiments

In order to demonstrate the merits and efficiency of the above extrapolated schemes in this chapter, a large number of numerical experiments is arranged and performed with Matlab 6.5 for several well-known test problems taken from [6].

To test the impact of the smoothness of true solution on the selection of node location ($[\alpha_{min}, \alpha_{max}], sequence$), we set $u_\mu = (A^*A)^\mu u_T$ and $b_\mu = Au_\mu$ where u_T is the true solution of $Au = b_T$, so the smooth true solution u_μ and the related true right-hand side b_μ are obtained. Obviously, $\mu = 0$ results in $u_0 = u_T$ and $b_0 = b_T$. Moreover, the perturbations are added into b_μ:

$$b_\mu^\delta = b_\mu + \tau \|b_\mu\| \cdot randn(m, 1)/\sqrt{m},$$

where $m = length(b_\mu)$, $randn(\bullet)$ is an M-function generating a random noise, the parameter τ controls the degree of the noise, and $\delta = \|b_\mu - b_\mu^\delta\|_2$.

In numerical experiments, we employed three widely used non-extrapolated regularization methods: **Tik-dp, Tik-lc** and **Tik-gcv** which correspond to Tikhonov regularization with the discrepancy principle, L-curve rule and gcv-rule to determine regularization parameters, respectively. And, the aforementioned extrapolated regularization schemes, **E-Tik-Lag**, **E-Tik-Her**, **E-Tik-Bre** and **E-Tik-Hä m-2**, are included to make a comparison. The relative error $\|u_{true} - u_{approx}^\delta\|/\|u_{true}\|$ is used for measuring the approximation. For all of the extrapolated schemes in Tables 5.1–5.3, the best performances of the schemes are denoted by "$\|u_{true} - u_{opt}^\delta\|/\|u_{true}\|$ (index k)", where k in the parentheses is the index for which u_{opt}^δ represents the best approximation among $\{v_{\alpha,i}^\delta;\ i = 2, 3, \ldots, 10\}$ in the sense of relative error.

The results of numerical experiments are summarized in Tables 5.1–5.3. If the digital values are printed in '*italics*', then they are considered to be the best among the related results (Tables 5.1 and 5.2) by all of the (extrapolated and non-extrapolated) regularization algorithms for the corresponding test problems, or among the results by employing the five node-locations (Table 5.3) for each test problem. In all of the tables, the notation N.L. refers to the node location; the values in Table 5.2 marked with "$\underline{\quad\quad}$" are used for comparison.

From Table 5.1 it can be seen, for the four tested problems, that the extrapolated schemes, with the small k (lower computation cost), are better than the non-extrapolated algorithms in most cases; the more smoother the true solution is, the more effective the extrapolated scheme will be. Especially, the **E-Tik-Bre** seems to be the best one among the extrapolated schemes. Another obvious feature is that the **E-Tik-dp** had failed in solving '*shaw*' problem with high smoothness of $\mu = 2$ and '*gravity*' problem with $\mu \geq 2/3$! Table 5.1 also implies that the node location of (2,4) is feasible for solving problems regardless of lower or higher smoothness; this observation inspires us that we can apply it to solving practical problems.

Table 5.1 The relative errors of extrapolated and non-extrapolated regularization solutions ($m = 1000$; $\tau = 0.001$ and for N.L.=(2,4)).

$\mu =$	Algorithm	shaw	heat	foxgood	gravity
	Tik-dp	.28e-02	.77e-02	.12e-02	.27e-02
	Tik-lc	.25e-02	.57e-00	.197e-01	.14e-01
	Tik-gcv	.12e-01	.29e-01	.43e-02	.45e-02
1/2	E-Tik-Lag	*.23e-02*(2)	.62e-02(2)	.50e-03(2)	.23e-02(2)
	E-Tik-Her	.27e-02(2)	.97e-02(2)	.10e-02(2)	.31e-02(2)
	E-Tik-Bre	*.23e-02(3)*	*.49e-02(3)*	*.16e-03(3)*	*.17e-02(2)*
	E-Tik-Häm-2	*.23e-02*(2)	.60e-02(2)	.59e-03(2)	.21e-02(2)
	Tik-dp	.26e-02	.66e-02	.12e-02	fail !!
	Tik-lc	.25e-01	.53e-00	.19e-01	.14e-01
	Tik-gcv	.12e-01	.28e-02	.41e-02	.44e-02
2/3	E-Tik-Lag	.21e-02(2)	.64e-02(2)	.59e-03(2)	.23e-02(2)
	E-Tik-Her	.28e-02(2)	.10e-01(2)	.11e-02(2)	.31e-02(2)
	E-Tik-Bre	*.18e-02(3)*	*.50e-02(2)*	*.28e-03*(3)	*12e-02(2)*
	E-Tik-Häm-2	.19e-02(2)	.57e-02(2)	.58e-03(2)	.21e-02(2)
	Tik-dp	.25e-02	.63e-02	.12e-02	fail !!
	Tik-lc	.25e-01	.51e-00	.19e-01	.14e-01
	Tik-gcv	.12e-01	.28e-01	.41e-02	.44e-02
4/5	E-Tik-Lag	.21e-02(2)	.65e-02(2)	.61e-03(2)	.23e-02(2)
	E-Tik-Her	.28e-02(2)	.10e-01(2)	.11e-02(2)	.31e-02(2)
	E-Tik-Bre	*.13e-02(3)*	*.30e-02(2)*	*.38e-03(3)*	*.11e-02(2)*
	E-Tik-Häm-2	.19e-02(2)	.58e-02(2)	.58e-03(2)	.21e-02(2)
	Tik-dp	.25e-02	.61e-02	.12e-02	fail !!
	Tik-lc	.25e-01	.48e-00	.19e-01	.14e-01
	Tik-gcv	.12e-01	.28e-01	.41e-02	.44e-02
1	E-Tik-Lag	.21e-02(2)	.67e-02(2)	.62e-03(2)	.23e-02(2)
	E-Tik-Her	.28e-02(2)	.11e-01(2)	.11e-02(2)	.31e-02(2)
	E-Tik-Bre	*.97e-03(2)*	*.21e-02(2)*	*.34e-03(2)*	*.96e-03(2)*
	E-Tik-Häm-2	.19e-02(2)	.59e-02(2)	.57e-03(2)	.21e-02(2)
	Tik-dp	fail !!	.58e-02	.12e-02	fail !!
	Tik-lc	.25e-01	.45e-00	.19e-01	.14e-01
	Tik-gcv	.12e-01	.28e-01	.41e-02	.43e-02
2	E-Tik-Lag	.21e-02(2)	.68e-02(2)	.62e-03(2)	.22e-02(2)
	E-Tik-Her	.28e-02(2)	.11e-01(2)	.11e-02(2)	.31e-02(2)
	E-Tik-Bre	*.72e-03(2)*	*.19e-02(2)*	*.15e-03(2)*	*.91e-03(2)*
	E-Tik-Häm-2	.19e-02(2)	.61e-02(2)	.57e-03(2)	.21e-02(2)

Table 5.2 indicates, on the whole, that node location of (2,3) is better than (2,4), and this is affirmative because the former has used a priori information about the true solution. But, node-location (2,3) is closely related to setting of α_{min} in $[\alpha_{min}, \alpha_{max}]$ which reflects smooth properties of underlying solution, and such a choice can rarely be used in practice. Therefore, from the above analysis, it is more practical to employ (2,4) for solving applied problems.

A drawback of (2,4) is that changing a new Chebyshev node-system in (2,4) requires a new computation of the regularized solutions from scratch, this is not worthwhile sometimes; but it appears that this defect can be compensated by its

Table 5.2 The relative errors at two node-locations with and without using the smoothness of true solutions ($m = 1000$; $\tau = 0.001$).

$\mu =$	Algorithm	heat (N.L.=(2,3))	heat (N.L.=(2,4))	foxgood (N.L.=(2,3))	foxgood (N.L.=(2,4))
1/2	Tik-dp	.77e-02	.77e-02	.12e-02	.12e-02
	E-Tik-Lag	.57e-02(8)	.62e-02(2)	.20e-03(7)	.50e-03(2)
	E-Tik-Her	.11e-01(3)	.97e-02(2)	.75e-03(3)	.10e-02(2)
	E-Tik-Bre	.57e-02(2)	*.49e-02*(3)	*.21e-03*(3)	*.16e-03*(3)
	E-Tik-Häm-2	*.50e-02*(6)	.60e-02(2)	.56e-03(2)	.59e-03(2)
2/3	Tik-dp	.66e-02	.66e-02	.12e-02	.12e-02
	E-Tik-Lag	*.33e-02*(9)	.64e-02(2)	.15e-03(7)	.59e-03(2)
	E-Tik-Her	.67e-02(3)	.10e-01(2)	.40e-03(2)	.11e-02(2)
	E-Tik-Bre	.36e-02(3)	*.50e-02*(2)	*.15e-03*(3)	*.28e-03*(3)
	E-Tik-Häm-2	.34e-02(6)	.57e-02(2)	.64e-03(2)	.58e-03(2)
4/5	Tik-dp	.63e-02	.63e-02	.12e-02	.12e-02
	E-Tik-Lag	*.23e-02*(7)	.65e-02(2)	.13e-03(7)	.61e-03(2)
	E-Tik-Her	.48e-02(3)	.10e-01(2)	.27e-03(3)	.11e-02(2)
	E-Tik-Bre	.25e-02(3)	*.30e-02*(2)	*.13e-03*(3)	*.38e-03*(3)
	E-Tik-Häm-2	.26e-02(3)	.58e-02(2)	.76e-03(2)	.58e-03(2)
1	Tik-dp	.61e-02	.61e-02	.12e-02	.12e-02
	E-Tik-Lag	.15e-02(7)	.67e-02(2)	.14e-03(6)	.62e-03(2)
	E-Tik-Her	.31e-02(3)	.11e-01(2)	.17e-03(3)	.11e-02(2)
	E-Tik-Bre	*.17e-02(3)*	*.21e-02*(2)	*.11e-03*(3)	*.34e-03*(2)
	E-Tik-Häm-2	.21e-02(2)	.59e-02(2)	.11e-02(2)	.57e-03(2)
2	Tik-dp	.58e-02	.58e-02	.12e-02	.12e-02
	E-Tik-Lag	.49e-03(7)	.69e-02(2)	.93e-04(6)	.62e-03(2)
	E-Tik-Her	.99e-03(3)	.11e-01(2)	.10e-03(2)	.11e-02(2)
	E-Tik-Bre	*.49e-03*(3)	*.19e-02*(2)	*.89e-04*(2)	*.15e-03*(2)
	E-Tik-Häm-2	.38e-02(2)	.61e-02(2)	.48e-02(2)	.57e-03(2)

good approximation property since the k corresponding to (2,4) is usually smaller than the k corresponding to (2,3).

From Table 5.3 we can see that most of the extrapolated schemes for most of test problems are effective even though the realization of them is done in a simplified way. Moreover, among the extrapolated schemes, the **E-Tik-Bre** seems to be the best one; the **E-Tik-Lag**, **E-Tik-Her** and **E-Tik-Hämm-2** are more or less equality.

It should be pointed out that the choice of k is not full-automatically determined in the tests. For **E-Tik-Lag** and **E-Tik-Hämm-2**, the rule in **Theorem 5.1** is quite effective but it can not always give the optimal index k. For **E-Tik-Bre** and **E-Tik-Her**, we adopt roughly the discrepancy principle but the coefficient C must be taken carefully.

Table 5.3 The application of the extrapolated regularization schemes to several test problems ($m = 1000; \quad \tau = 0.001$).

Problem	N.L.	E-Tik-Lag	E-Tik-Her	E-Tik-Bre	E-Tik-Hämm-2	Tik-dp
	(1,1)	.47e-1(10)	.47e-1(9)	.47e-1(5)	.51e-1(10)	.48e-1
	(1,4)	*.47e-1*(8)	.47e-1(5)	*.46e-1*(10)	*.47e-1*(10)	.48e-1
shaw	(2,1)	.47e-1(10)	.47e-1(10)	.52e-1(10)	.55e-1(10)	.48e-1
($\delta = .0696$)	(2,2)	.47e-1(10)	*.46e-1*(10)	*.46e-1*(9)	.48e-1(10)	.48e-1
	(2 4)	.47e-1(10)	.48e-1(4)	.46e-1(10)	.48e-1(10)	.48e-1
	(1,1)	*.27e-1*(2)	.31e-1(2)	.29e-1(3)	*.28e-1*(3)	.29e-1
	(1,4)	.32e-1(2)	.43e-1(2)	.31e-1(3)	.31e-1(2)	.29e-1
heat	(2,1)	.29e-1(10)	.26e-1(10)	.28e-1(10)	.40e-1(10)	.29e-1
($\delta = .0014$)	(2,2)	.94e-1(2)	.14e-0(3)	.29e-1(3)	.94e-1(2)	.29e-1
	(2,4)	*.27e-1*(8)	*.27e-1*(5)	*.27e-1*(9)	.28e-1(10)	.29e-1
	(1,1)	.22e-2(9)	.23e-2(6)	.56e-2(4)	.35e-2(10)	.38e-2
	(1,4)	*.18e-2*(5)	.25e-2(3)	.57e-2(4)	.21e-2(6)	.38e-2
foxgood	(2,1)	.43e-2(10)	.24e-2(10)	.53e-2(10)	.86e-2(10)	.38e-2
($\delta = .0134$)	(2,2)	.24e-1(2)	.33e-1(2)	.55e-2(3)	.24e-1(2)	.38e-2
	(2,4)	.20e-2(9)	*.21e-2*(6)	.62e-2(4)	*.21e-2*(10)	.38e-2
	(1,1)	.12e-1(9)	.12e-1(6)	.16e-1(8)	.12e-1(9)	.13e-1
	(1,4)	*.12e-1*(5)	.12e-1(3)	.12e-1(7)	*.12e-1*(6)	.13e-1
gravity	(2,1)	.12e-1(9)	.12e-1(9)	.12e-1(8)	.13e-1(10)	.13e-1
($\delta = .1396$)	(2,2)	.52e-1(2)	.86e-1(2)	.12e-1(3)	.52e-1(2)	.13e-1
	(2,4)	*.12e-1*(7)	*.12e-2*(5)	*.11e-1*(8)	.12e-1(10)	.13e-1

5.7 Conclusion

In conclusion, if the true solution of the problems is smooth and the extrapolated parameters are selected properly, the extrapolated regularization methods are usually better than the single Tikhonov regularization method. But for the problems with lower smoothness, as [7] pointed out, "the positive effect of extrapolation was moderate." As a simplified version, the node-location of (2,4) is attractive when the information about smoothness of the problem is not known at all. Finally, more effective rules for selecting the extrapolation parameters must be ulteriorly studied at least for **E-Tik-Bre** and **E-Tik-Her** schemes.

Acknowledgements The first author is grateful for the useful discussion with Professor U. Hämarik in several personal communications. This work was completed with financial support from the National Natural Science Foundation of Hebei Province under grant number A2006000004. The first author was also partially supported by the National Natural Science Foundation of China under grant number 10571039.

References

1. C. Brezinski, A general extrapolation algorithm, *Numerische Mathematik*, **35**, 175–187, 1980.
2. C. Brezinski, M. Redivo-Zaglia, G.Rodriguez and S. Seatzu, Extrapolation techniques for ill-conditioned linear systems, *Numerische Mathematik*, **81**, 1–29, 1998.
3. C. Brezinski, M. Redivo-Zaglia, G. Rodriguez and S. Seatzu, Multi-parameter regularization techniques for for ill-conditioned linear systems, *Numerische Mathematik*, **94**, 203–228, 2003.
4. E. W. Cheney, *Introduction to Approximation Theory*, Chelsea Publ. Co., New York, 1982.
5. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, The Netherlands, 1996.
6. P. C. Hansen, Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems, *Numerical Algorithms*,**6**, 1–35, 1994.
7. U. Hämarik, R. Palm and T. Raus, Use of extrapolation in regularization methods, *Journal of Inverse and Ill-Posed Problems*, **15**, 277–294, 2007.
8. U. Hämarik, R. Palm and T. Raus, Extrapolation of Tikhonov and Lavrentiev regularization methods, *Journal of Physics: Conference Series*, **135**, 012048 (8pp), 2008.
9. C. W. Groetsch and J. T. King, Extrapolation and the method of regularization for generalized inverses, *J. Approx. Theory*, **25**(3), 233–247, 1979.
10. R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.
11. K. Kunisch and J. Zou, Iterative choices of regularization parameter in linear inverse problems, *Inverse Problems*, **14**, 1264–1274, 1998.
12. G. I. Marchuk and V. V. Shaidurov, *Difference Methods and Their Extrapolations*, Springer-Verlag, 1983.
13. V. V. Saidurov, Continuation with respect to the parameter in the method of regularization, *Numerical Methods of Linear Algebra*, 77–85, Novosibirsk, 1973 (In Russian).
14. Y. F. Wang and T. Y. Xiao, The fast realization algorithms for determinig the regularization parameter in linear inverse problems, *Inverse Problems*, **17**, 281–291, 2001.

Chapter 6
Modified Regularization Scheme with Application in Reconstructing Neumann-Dirichlet Mapping

Pingli Xie and Jin Cheng

Abstract. In this chapter, we propose a new regularization method for solving a linear operator equation when the operator and right hand term are both known approximately. The advantage of our method is that we just use the information about the error level instead of assuming the reliable bounds of the unknown solution. The algorithms are presented and the numerical simulation results show the efficiency of our method. As an application, we discuss the problem of reconstructing the Neumann-Dirichlet mapping from the discrete Neumann and Dirichlet data. The Neumann-Dirichlet mappings are used widely in the studying of inverse problems for partial differential equations.

6.1 Introduction

Tikhonov regularization is a widely used tool for solving the linear and nonlinear ill-posed problems in the engineering sciences. There is much research on this topic, for both the theoretical and numerical aspects. We refer to the books [2], [6], [9] and the references in these books.

Our focus in this chapter is solving ill-posed problems of the form

$$A_0 x = y_0, \tag{6.1.1}$$

Pingli Xie
School of Sciences, Henan University of Technology, Zhengzhou 450001, China.
e-mail: plxie@fudan.edu.cn

Jin Cheng
School of Mathematical Sciences, Fudan University, Shanghai 200433, China.
e-mail: jcheng@fudan.edu.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

where $A_0 \in \mathcal{L}(X, Y)$ is a linear, injective and bounded operator with a non-closed range $\mathcal{R}(A_0)$, X, Y are Hilbert spaces with corresponding inner products $(\cdot, \cdot)$ and norms $\|\cdot\|$.

Throughout the context, we assume that the right-hand side y_0 belongs to $\mathcal{R}(A_0)$ so that there exists a unique solution $x^\dagger \in X$ to equation (6.1.1).

Such problems arise, for example, from the discretization of ill-posed problems such as integral equations of the first kind. See, e.g., [1, 2, 4, 5] for examples and details. In real applications, an ideal problem like (6.1.1) is seldom available, and the following situation is more realistic:

1) Instead of the exact right-hand side $y_0 \in \mathcal{R}(A_0)$, noisy data $y_\delta \in Y$ are given with

$$\|y_0 - y_\delta\| \leq \delta. \tag{6.1.2}$$

2) Instead of the exact operator $A_0 \in \mathcal{L}(X, Y)$, some approximate operator A_h is given with

$$\|A_0 - A_h\| \leq h. \tag{6.1.3}$$

As mentioned in [7], for ill-posed problems, regularized total least squares methods (RTLS) take account of additional perturbations in the operator A_0 [3]. This method, however, requires a reliable bound for $\|Bx^\dagger\|$, which is generally unknown. On the other hand, in some applications, reliable bounds for the noise levels δ and h in (6.1.2) and (6.1.3) are known. In this case, it makes sense to look for approximations $(\hat{x}, \hat{y}, \hat{A})$ subject to some conditions with respect to these parameters. The method of dual regularized least squares [7] is out of such an motivation. But then, the solution is obtained only when the constraint inequality is replaced by equality, i.e., the solution must satisfy

$$Ax = y, \quad \|y - y^\delta\| = \delta, \quad \|A - A_h\| = h,$$

which is not always the case in real applications. Notice that the noisy level of $\|y - y^\delta\|$ is of order δ and $\|Bx\|$ is bounded. Motivated by this, we propose a new method to minimize the functional

$$\delta^2 \|Bx\|^2 + \|y - y^\delta\|^2$$

instead of $\|Bx\|$ in [7].

Note that in the modified regularization scheme, we not only obtain the solution, but also get the approximation of the operator. As an application of our method, we discuss the reconstruction of Neumann-Dirichlet mapping for the elliptic equation of the second order in the plane by the Neumann and Dirichlet data.

This chapter is organized as follows: In Section 6.2, we introduce the new method and study its error bounds. Computational aspects are described in Section 6.3, an iterative method for numerically computing the regularization parameters is provided. Section 6.4 contains the numerical simulation results for our method. In Sections 6.5 and 6.6, we show how to apply our method to recon-

structing the Neumann-Dirichlet mapping from the given Neumann and Dirichlet data.

6.2 Regularization method

Our regularization of the ill-posed problems is based on Tikhonov regularization for the linear least squares problems with exact operators. In our method, an estimate (x, y, A) for $(x^\dagger, y_0, A_0)$ from known data (y_δ, A_h) is determined by solving the constrained minimization problem

$$\delta^2\|Bx\|^2 + \|y - y_\delta\|^2 \to \min, \text{ subject to } Ax = y, \|A - A_h\| \le h. \tag{6.2.1}$$

Substituting the constraint $Ax = y$ into the problem (6.2.1), we can get the corresponding Lagrange multiplier formulation

$$L(A, x, \mu) = \delta^2\|Bx\|^2 + \|Ax - y_\delta\|^2 + \mu(\|A - A_h\|^2 - h^2), \tag{6.2.2}$$

where μ is the Lagrange multiplier, zero if the inequality constraint is inactive. In the context, we always assume that is active. The solution to this problem can be characterized by the following theorem.

Theorem 6.2.1. *The solution to (6.2.1), with the inequality constraint replaced by equality, is a solution to the problem*

$$(A_h^T A_h + \alpha B^T B + \beta I)x = A_h^T y_\delta, \tag{6.2.3}$$

where α and β are positive regularization parameters given by

$$\alpha = \frac{\mu + \|x\|_2^2}{\mu}\delta^2, \tag{6.2.4}$$

$$\beta = -\frac{\|A_h x - y_\delta\|_2^2}{\mu + \|x\|_2^2}, \tag{6.2.5}$$

and μ is the Lagrange multiplier in (6.2.2). Moreover, the two parameters can be expressed as

$$\alpha = \frac{\|A_h x - y_\delta\|_2}{\|A_h x - y_\delta\|_2 + h\|x\|_2}\delta^2, \tag{6.2.6}$$

$$\beta = \frac{\|A_h x - y_\delta\|_2}{\|x\|_2}h. \tag{6.2.7}$$

Proof. We eliminate y in problem (6.2.1) and characterize the solution to (6.2.1) by setting the partial derivatives of the Lagrange function L in (6.2.2) to zero. Thus we can obtain

$$L_x = 2\delta^2 B^T Bx + 2A^T(Ax - y_\delta) = 0, \tag{6.2.8}$$

$$L_A = 2(Ax - y_\delta)x^T + 2\mu(A - A_h) = 0, \tag{6.2.9}$$

$$L_\mu = \|A - A_h\|^2 - h^2 = 0. \tag{6.2.10}$$

From (6.2.9), we have $A(xx^T + \mu I) = \mu A_h + y_\delta x^T$, or equivalently,

$$\begin{aligned} A &= (\mu A_h + y_\delta x^T)(xx^T + \mu I)^{-1} \\ &= (\mu A_h + y_\delta x^T)(\frac{1}{\mu} I - \frac{1}{\mu(\mu + \|x\|_2^2)} xx^T) \\ &= A_h - \frac{1}{\mu + \|x\|_2^2}(A_h x - y_\delta)x^T. \end{aligned} \tag{6.2.11}$$

Substituting this into (6.2.8) and gathering the terms, we readily arrive at the equation

$$\delta^2 B^T Bx + \frac{\mu}{\mu + \|x\|_2^2} A_h^T A_h x - \frac{\mu \|A_h x - y_\delta\|_2^2}{(\mu + \|x\|_2^2)^2} x = \frac{\mu}{\mu + \|x\|_2^2} A_h^T y_\delta.$$

Multiplying this equation by $(\mu + \|x\|_2^2)/\mu$ implies the equivalent equation (6.2.3) with α and β given by (6.2.4) and (6.2.5) respectively. It remains proving (6.2.6) and (6.2.7). To this end, we recall equation (6.2.11), due to the property of $\|\cdot\|_F$, i.e.,

$$\|xy^T\|_F = \|x\|_2 \|y\|_2, \text{ for all } x \in \mathbf{R}^n, y \in \mathbf{R}^m,$$

we have

$$\|A - A_h\|_F = \left| \frac{1}{\mu + \|x\|_2^2} \right| \|A_h x - y_\delta\|_2 \|x\|_2.$$

This together with (6.2.10) gives

$$|\mu + \|x\|_2^2| = \|A_h x - y_\delta\|_2 \|x\|_2 / h.$$

Thus, $\mu = \frac{\|A_h x - y_\delta\|_2 \|x\|_2}{h} - \|x\|_2^2$, if $\mu + \|x\|_2^2 > 0$; and $\mu = -\frac{\|A_h x - y_\delta\|_2 \|x\|_2}{h} - \|x\|_2^2$, if $\mu + \|x\|_2^2 < 0$.

Therefore, we have two cases:

(i).

$$\alpha = \frac{\|A_h x - y_\delta\|_2}{\|A_h x - y_\delta\|_2 - h\|x\|_2} \delta^2, \quad \beta = -\frac{\|A_h x - y_\delta\|_2}{\|x\|_2} h; \tag{6.2.12}$$

and (ii)

$$\alpha = \frac{\|A_h x - y_\delta\|_2}{\|A_h x - y_\delta\|_2 + h\|x\|_2} \delta^2, \quad \beta = \frac{\|A_h x - y_\delta\|_2}{\|x\|_2} h. \tag{6.2.13}$$

Since the regularization parameters are both positive, we cast the first case out here. Thus, we get the desired result. □

6.3 Computational aspect

In the following, we first provide the algorithm for our new regularization method.

Algorithm 6.3.1 (Solving problem (6.2.3)). **Input:** $\varepsilon > 0, y_\delta, A_h, B, \delta$ *and* h.
(i) Choose some starting value x_**, and find* α *and* β *with (6.2.6) and (6.2.7).*
(ii) Solve $(A_h^T A_h + \alpha B^T B + \beta I)x = A_h^T y_\delta$.
(iii) Update α_{new} *and* β_{new} *by (6.2.6) and (6.2.7).*
(iv) **if** $|\alpha_{new} - \alpha| + |\beta_{new} - \beta| \geq \varepsilon$, **then** $\alpha := \alpha_{new}, \beta := \beta_{new}$ *and* **goto** *(ii),*
(v) **else** *solve* $(A_h^T A_h + \alpha_{new} B^T B + \beta_{new} I)x = A_h^T y_\delta$.

To reduce the iteration, we propose other two algorithms here. We first solve the minimization problem with respect to operator A:

$$\delta^2 \|Bx\|^2 + \|Ax - y_\delta\|^2 \to \min, \text{subject to } \|A - A_h\| \leq h. \tag{6.3.1}$$

with any fixed point x. Similar to the above Lagrange multiplier analysis, we can get that this functional derives its minimum at $A = A_h + \frac{h}{\|A_h x - y_\delta\|_2 \|x\|_2}(A_h x - y_\delta)x^T$. Substituting it into problem (6.2.1), we obtain the minimization problem with respect to x:

$$\delta^2 \|Bx\|^2 + (\|A_h x - y_\delta\| + h\|x\|)^2 \to \min. \tag{6.3.2}$$

Since there is a cross term $2h\|A_h x - y_\delta\|\|x\|$ in the functional, it may not be differentiable at zero. Here we use approximate terms to replace it. Then employ the usual method for solving minimization problem without iteration. Two different approximations yield the following two algorithms.

Approximating the term with $\|A_h x - y^\delta\|^2 + h^2\|x\|^2$, we have

Algorithm 6.3.2 (Approximate method 1). **Input:** y_δ, A_h, B, δ *and* h.
Solve $(A_h^T A_h + \frac{\delta^2}{2} B^T B + h^2 I)x = A_h^T y_\delta$.

Approximating the term with $h\|A_h x - y^\delta\|^2 + h\|x\|^2$, we have

Algorithm 6.3.3 (Approximate method 2). **Input:** $y_\delta, A_h, B, \delta,$ *and* h.
Solve $(A_h^T A_h + \frac{\delta^2}{1+h} B^T B + hI)x = A_h^T y_\delta$.

The numerical experiments below show that these two methods are not as efficient as Algorithm 6.3.1, but obviously, they have an advantage over other methods in computation time since no iterations are involved.

6.4 Numerical simulation results for the modified regularization

In this section, we present some numerical examples to demonstrate the efficiency of the new regularization method. Our computations are carried out in Matlab

using the Regularization Tools package [5]. Below the weighted operator B is chosen as the discrete approximation matrix to the first derivative operator.

Example 6.4.1. *The first test problem we have chosen to illustrate our algorithm is performed with i_laplace*$((n, 1)$.

The function *i_laplace*$((n, 1)$ is a discretization of the inverse Laplace transformation, written as a Fredholm integral equation of the first kind

$$\int_a^b K(s,t)f(t)dt = g(s), \quad s \in [a,b], \tag{6.4.1}$$

by means of Gauss-Laguerre quadrature. The kernel K is given by

$$K(s,t) = exp(-st),$$

and both integration intervals are $[0, +\infty)$. The corresponding function is implemented with

$$f(t) = exp(-t/2), \quad g(s) = 1/(s + 0.5). \tag{6.4.2}$$

The discretization points number of the kernel and the solution are chosen as $n = 64$ to reduce the size of matrix A and the solution $x^\dagger$. The exact discrete right-hand side is produced by $y_0 = Ax^\dagger$. As in [5], the perturbed right hand side is generated as $y^\delta = (A + h\|E\|_F^{-1}E)x^\dagger + \delta\|e\|_2^{-1}e$, where the perturbation matrix E and the vector e are from a normal distribution with zero mean and unit standard deviation. The exact solution and approximate solutions obtained by different algorithms are plotted in Fig. 6.1 (a). To illustrate the stability of different methods, we perform the experiment with 50 different noisy data y^δ generated separately. The corresponding results are given in Fig. 6.1 (b). Here and after, we use "lu, xx, a2" and "a1" in the legend of each figure to represent

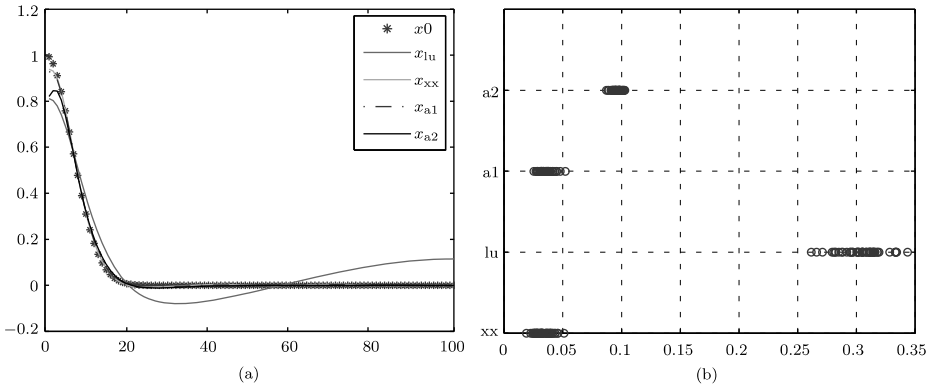


Fig. 6.1 (a) The comparison of different solutions in Example 6.4.1; (b) The corresponding stability results in Example 6.4.1.

the corresponding numerical results derived by the algorithm proposed in [7], Algorithm 6.3.1, Algorithm 6.3.2 and Algorithm 6.3.3, respectively.

Example 6.4.2. *The second example is based on the function baart(n).*

It is a discretization of the Fredholm equation

$$\int_0^{\pi} e^{s\cos t} f(t)dt = 2\frac{\sin s}{s}, \quad s \in [0, \frac{\pi}{2}],$$

with the solution $f(t) = \sin(t)$. The results are plotted in Fig. 6.2. As above, the first figure shows the different solutions and the second provides the stability of the methods.

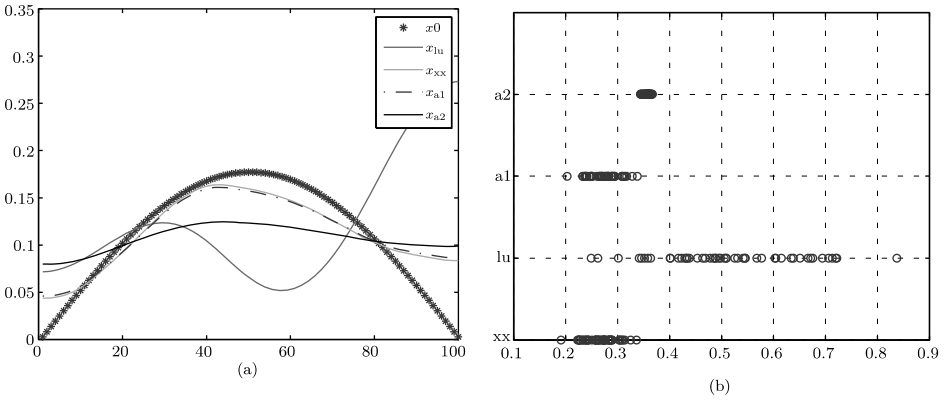


Fig. 6.2 (a) The comparison of different solutions in Example 6.4.2; (b) The stability results in Example 6.4.2.

Example 6.4.3. *The third example is based on the function shaw(n).*

It is a discretization of the integral equation of the first kind

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} k(s,t)f(t)dt = g(s), \quad s \in [-\frac{\pi}{2}, \frac{\pi}{2}],$$

where the kernel and the solution are given by

$$k(s,t) = (\cos(s) + \cos(t))^2 \left(\frac{\sin(u)}{u}\right)^2, \quad u = \pi(\sin(s) + \sin(t)),$$

$$f(t) = a_1 e^{-c_1(t-t_1)^2} + a_2 e^{-c_2(t-t_2)^2},$$

$$a_1 = 2, a_2 = 1, c_1 = 6, c_2 = 2, t_1 = 0.8, t_2 = -0.5.$$

The corresponding results are shown in Fig. 6.3.

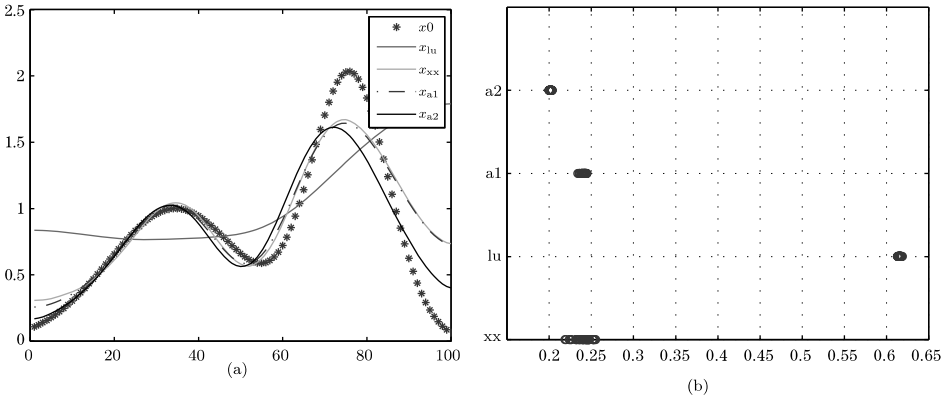


Fig. 6.3 (a) The different solutions in Example 6.4.3; (b) The stability results in Example 6.4.3.

Example 6.4.4. *The final example is based on gravity*$(n, example, 0, 1.0.25)$.

It corresponds to the Fredholm equation (6.4.2) with

$$K(s,t) = 0.25((0.25)^2 + (s-t)^2)^{-3/2}, \quad f(t) = \sin(\pi t) + 0.5\sin(2\pi t)$$

discretized at the interval $[0, 1]$ by means of a collocation scheme with n knots. The corresponding results are shown in Fig. 6.4.

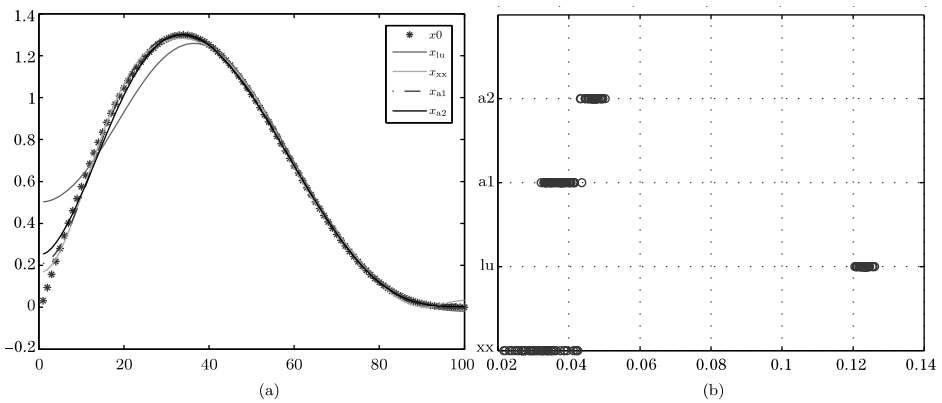


Fig. 6.4 (a) The different solutions in Example 6.4.4; (b) The stability results in Example 6.4.4.

6.5 The Neumann-Dirichlet mapping for elliptic equation of second order

In many practical problems, there are some possibilities that we do know the exact structure of the domain, for example, in the case there are unknown cavities and cracks inside some materials. But we can know the governing equation and we can measure the data on some part of boundary. This kind mathematical formulation has been used by many mathematicians. The corresponding inverse problems have been well studied. But from practical point of view, how to reconstruct the Neumann-Dirichlet mapping is an important topic for applying the well established mathematical theory and algorithms. There are only few results on this topic.

In this section, we will mainly show how to realize the reconstruction by the regularization we proposed in the previous section. The inverse problem we study is the determination of the unknown inclusions inside the fixed domain.

First, we give the definition of Neumann-Dirichlet mapping.

Consider the elliptic equation in the domain $\Omega \subset R^2$:

$$\begin{cases} -\triangle u = 0, & \text{in } \Omega, \\ \frac{\partial u}{\partial n} = g(x,y), & \text{on } \Gamma_N, \\ u = 0, & \text{on } \Gamma_D. \end{cases} \tag{6.5.1}$$

where Γ_D and Γ_N are two subset of $\partial\Omega$, which satisfy that $\Gamma_D \neq \emptyset$, $\overline{\Gamma_D} \cup \overline{\Gamma_N} = \partial\Omega$.

The Neumann-Dirichlet mapping is defined as

$$\Lambda : g(x,y) \mapsto u \mid_{\Gamma_N} .$$

Our purpose is to construct the Neumann-Dirichlet mapping from the Neumann data g and the corresponding Dirichlte data $f = u \mid_{\Gamma_N}$.

Assume that we just have the discrete data $(f_i, g_i), (i = 1, \ldots, m)$, which are the discrete boundary values of the solutions to the above equation.

We denote the discrete Dirichlet boundary value as f_i and the discrete Neumann boundary value as $\frac{\partial u_h}{\partial n} = g_i$.

By the regularization method we proposed, we can find the matrix K which is the minimum solution of the following optimal problem:

$$1/m \sum_{i=1}^{m} \|Kg_i - f_i\|^2 + \alpha \|K\|_F^2 \to \min, \tag{6.5.2}$$

where α is the regularization parameter.

Taking the derivative for the functional with respect to K vanished, we can have

$$1/m\sum_{i=1}^{m} 2(Kg_i - f_i)f_i' + 2\alpha K = 0.$$

Then we have

$$K(1/m\sum_{i=1}^{m} g_i g_i' + \alpha I) = 1/m\sum_{i=1}^{m} f_i g_i'.$$

If we take that $\{g_i\}, i = 1, \cdots, n$ are orthogonal vectors, then

$$\left(1/m\sum_{i=1}^{m} g_i g_i' + \alpha I\right) \text{ is invertable}$$

when $\alpha \neq 1/m\sum g_i' \sum g_i$.

Therefore we can get $K = 1/m\sum_{i=1}^{m} f_i g_i'(1/m\sum_{i=1}^{m} g_i g_i' + \alpha I)^{-1}$.

As for our example, equation (6.5.1) is considered in the domain Ω , which has the outer boundary unit circle. We use the second order triangle finite element to solve the forward problems and get u_h and f_i.

It is well known that any function in $L^2[-\pi, \pi]$ can be expanded to a Fourier series.

6.6 The numerical results of the Neumann-Dirichlet mapping

We notice that from the well-posedness of the forward problem, the integral of $g(x, y)$ on Γ_N is 0.

First we transform the boundary Γ_N to the interval $[0, 2\pi]$. Then the basis functions of g_i can be taken as

$$\cos\theta, \sin\theta, \cos(2\theta), \sin(2\theta), \cdots, cos(n\theta), \sin(n\theta),$$

where $0 \leq \theta \leq 2\pi$,.

It is obvious that these functions are orthogonal. In our examples, we all take $n = 40$. We consider the following examples.

Example 6.6.1. *We take* $\Omega = \{x | \|x\| < 1\}$.

Since the problem becomes a Neumann problem for the Laplace equation, we take $u(-1, 0) = 0$ to ensure a unique solution. The numerical results are shown in Fig. 6.5 (a) and (b).

Example 6.6.2. *We take* $\Omega = \{x | \|x\| < 1\} \setminus \omega$, *where* $\omega = \{x | \|x - (0.3, 0.3)\| < 0.5\}$.

By taking the regularization parameter $\alpha = 10^{-8}$, we have the results shown in Fig. 6.6 (a) and (b).

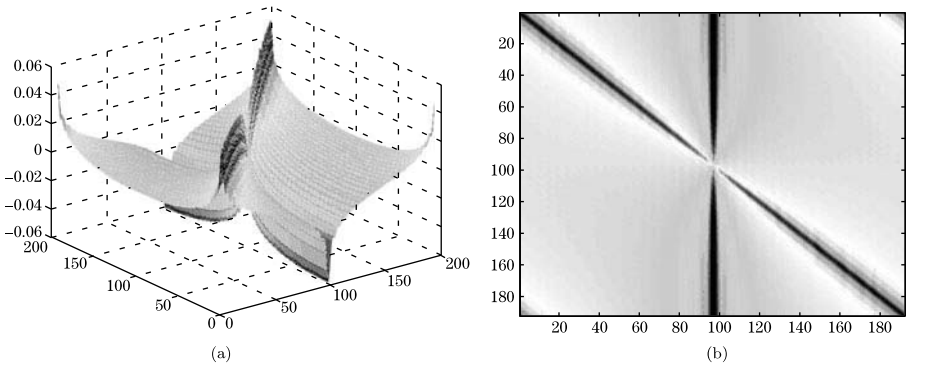


Fig. 6.5 (a) The image of K when $\alpha = 0.8$ in Example 6.6.1; (b) The mesh picture of K when $\alpha = 0.8$ in Example 6.6.1.

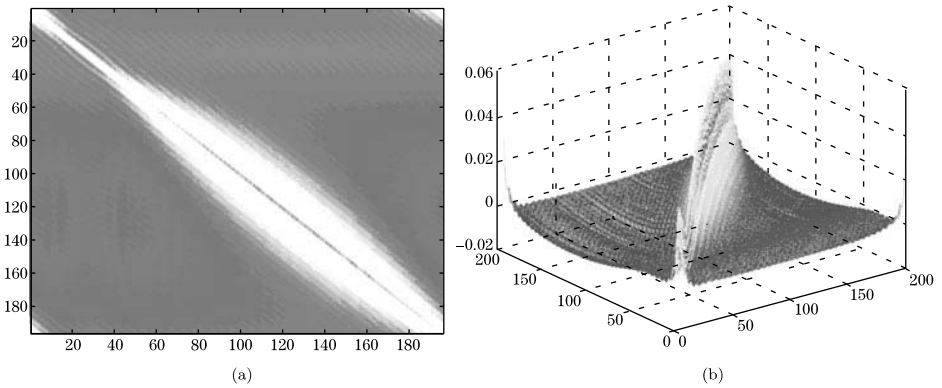


Fig. 6.6 (a) The image of K when $\alpha = 0.8$ in Example 6.6.2; The mesh picture of K when $\alpha = 0.8$ in Example 6.6.2.

Example 6.6.3. *We take* $\Omega = \{x | \|x\| < 1\} \backslash \omega$, *where* $\omega = \{x | \|x - (-0.3, 0.3)\| < 0.5\}$.

By taking the regularization parameter $\alpha = 10^{-8}$, we have the results shown in Fig. 6.7 (a) and (b).

6.7 Conclusion

In this chapter, we propose a modified regularization method for the ill-posed operator equation when both operator and right hand term are known approximately. The numerical algorithms are constructed. The numerical results show that our method is efficient. As an application of our method, we discuss the prob-

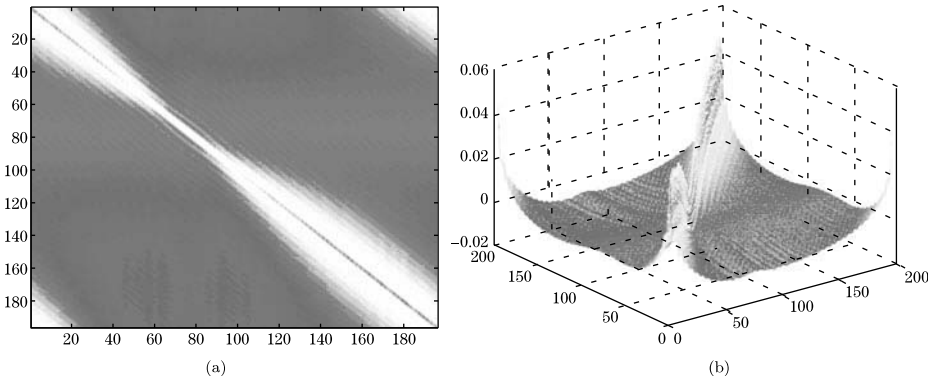


Fig. 6.7 (a) The image of K when $\alpha = 0.8$ in Example 6.6.3; (b) The mesh picture of K when $\alpha = 0.8$ in Example 6.6.3.

lem of reconstructing the Neumann-Dirichlet mapping from the discrete boundary data for two dimensional problems.

References

1. H. W. Engl, Regularization methods for the stable solution of inverse problems, *Surv. Math. Ind.*, **3**, 71–143, 1993.
2. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.
3. G. H. Golub, P. C. Hansen and D. P. O'Leary, Tikhonov regularization and total least squares, *SIAM J. Matrix Anal. Appl.*, **21**, 185–194, 1999.
4. C. W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg, Wiesbaden, 1993.
5. P. C. Hansen, Regularization tools, a Matlab package for analysis of discrete regularization problems, *Numerical Algorithms*, **6**, 1–35, 1994.
6. V. Isakov, *Inverse Problems for Partial Differential Equations*, 2nd Edition, Springer-Verlag, 2006.
7. S. Lu, S. V. Pereverzev and U. Tautenhahn, Dual regularized total least squares and multi-parameter regularization, *Comput. Methods Appl. Math.*, **8**(3), 253–262, 2008.
8. U. Tautenhahn, Regularization of linear ill-posed problems with noisy right hand side and noisy operator, *J. Inv. Ill-Posed Problems*, **16**, 507–523, 2008.
9. A. N. Tikhonov, A. S. Leonov and A. G. Yagola, Nonlinear ill-posed problems, Vol. 1, 2, *Applied Mathematics and Mathematical Computation*, **14**, Chapman & Hall, London, 1998.

Part III
Nonstandard Regularization and Advanced Optimization Theory and Methods

Chapter 7
Gradient Methods for Large Scale Convex Quadratic Functions

Yaxiang Yuan

Abstract. The gradient method is one of the most simple methods for solving unconstrained optimization, it has the advantages of being easy to program and suitable for large scale problems. Different step-lengths give different gradient algorithms. In 1988, Barzilai and Borwein gave two interesting choices for the step-length and established superlinearly convergence results for two-dimensional convex quadratic problems. Barzilai and Borwein's work triggered much research on the gradient method in the past two decades. In this chapter we investigate how the BB method can be further improved. We generalize the convergence result for the gradient method with retards. Our generalization allows more choices for the step-lengths. An intuitive analysis is given on the impact of the step-length for the speed of convergence of the gradient method. We propose a short BB step-length method. Numerical results on random generated problems are given to show that our short step technique can improve the BB method for large scale and ill-conditioned problems, particularly when high accurate solutions are needed.

7.1 Introduction

In this chapter we consider the gradient methods for minimizing large scale convex quadratic functions. Most inverse problems can be formulated as

$$\mathcal{L}x = z, \tag{7.1.1}$$

Yaxiang Yuan
State Key Laboratory of Scientific/Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
Zhong Guan Cun Donglu 55, Beijing, 100190, China.
e-mail: yyx@lsec.cc.ac.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

where z is the given data or observations, and $\mathcal{L}$ is an mapping, and x is the unknown that needs to computed. After discretization and linearization, we will need to solve a set of linear equations

$$Ax = b \quad x \in \Re^n, \tag{7.1.2}$$

where $b \in \Re^n$ and $A \in \Re^{n\times n}$ is a symmetric positive definite matrix. Many inverse problems can be formulated into (7.1.4) with a very large n and A is ill-conditioned (for example, see[17, 18, 19, 20]). It is easy to see that linear system (7.1.2) is equivalent to the following unconstrained optimization problem:

$$\min_{x\in R^n} f(x), \tag{7.1.3}$$

with

$$f(x) = \frac{1}{2}x^T Ax - b^T x. \tag{7.1.4}$$

The gradient method is one of the most simple methods for solving (7.1.3) where $f(x)$ is a continuously differentiable function in $\Re^n$. Assume that $g(x) = \nabla f(x)$ can be obtained at every x. Given an iterate point x_k, the gradient method chooses the next iterate point x_{k+1} in the following form:

$$x_{k+1} = x_k - \alpha_k g_k, \tag{7.1.5}$$

where $g_k = g(x_k)$ is the gradient at x_k and $\alpha_k > 0$ is a step-length. The gradient method has the advantages of being easy to program and suitable for large scale problems. Different step-lengths α_k give different gradient algorithms. If $\alpha_k = \alpha_k^*$ where α_k^* satisfies

$$f(x_k - \alpha_k^* g_k) = \min_{\alpha>0} f(x_k - \alpha g_k), \tag{7.1.6}$$

the gradient method is the steepest descent method, which is also called the Cauchy's method. However, the steepest descent method, though it uses the "best" direction and the "best" step-length, turns out to be a very bad method as it normally converges very slowly, particularly for ill-conditioned problems.

In this chapter, we discuss the gradient method for the special case when $f(x)$ is a strictly convex quadratic function (7.1.4) because this special problem appears in inverse problems very often.

We denote the eigenvalues of A by $\lambda_i (i = 1, 2, \ldots, n)$ and assume that

$$0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n. \tag{7.1.7}$$

The steepest descent method, which uses the exact line search step

$$\alpha_k^* = \frac{g_k^T g_k}{g_k^T A g_k} = \frac{\|Ax_k - b\|_2^2}{(Ax_k - b)^T A(Ax_k - b)}, \tag{7.1.8}$$

turns out to converge very slowly when A is ill-conditioned in the sense that the ratio of λ_1/λ_n is very small.

In 1988, Barzilai and Borwein[1] gave two interesting choices for the step-length α_k:

$$\alpha_k^{BB1} = \frac{\|s_{k-1}\|_2^2}{s_{k-1}^T y_{k-1}}, \tag{7.1.9}$$

$$\alpha_k^{BB2} = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|_2^2}, \tag{7.1.10}$$

where

$$s_{k-1} = x_k - x_{k-1} \quad y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1}). \tag{7.1.11}$$

Barzilai and Borwein[1] establishes superlinear convergence results for two dimensional convex quadratic problems. Moreover, numerical results indicate that for convex quadratic funtions $f(x)$ the BB method performs much better than the steepest descent method. Barzilai and Borwein's work triggered much research on the gradient method in the past two decades. For example, see Dai[2], Dai and Fletcher[3], Dai and Liao[4], Dai et al.[5], Dai and Yuan[6, 7], Dai and Zhang[8], Fletcher[9], Friedlander et al.[10], Hager and Zhang[11], Nocedal et al.[12], Raydan[13, 14], Raydan and Svaiter[15], Vrahatis et al.[16], Yuan[21, 22] and Zhou et al.[23]. Among the convergence properties of the BB step, major results are the following. For $n = 2$, Barzilai and Browein[1] showed that the BB step is Q-superlinear convergence for strictly convex quadratic functions. For general n, Raydan[13] proved the global convergence of the BB method for strict convex quadratic functions. Dai and Liao[4] proved the R-linear convergence results for strictly convex quadratic functions. There are many generalizations of the BB method, including the *alternate minimization* by Dai and Yuan[6], *cyclical BB* by Hager and Zhang[11], *adaptive BB* by Dai et al.[23] and Vrahatis et al.[16], *gradient method with retards* by Friedlander et al.[10], and *monotone BB type method* by Yuan[21].

In this chapter we consider how the the BB method can be further improved. First we generalize the global convergence result of the gradient method with retards, which enables us to use a wider range of step-lengths for the gradient method. Then, we propose to use short BB step-lengths. Numerical results of random generated problems indicate that short BB step-lengths will produce improvement over the standard BB step-length, particularly for large-scale and ill-conditioned problems.

7.2 A generalized convergence result

Barzilai and Borwein[1] proved the superlinear convergence of their method for the case having only two variables. Raydan[13] established the global convergence of the gradient method with BB step-lengths for general strictly convex quadratic functions. The following general global convergence result was given by Friedlander et al.[10]:

Theorem 7.2.1. *Let $f(x)$ be given by (7.1.4) and A is positive definite. Let m be a positive integer and $q_j \geq 1 (j = 1, 2, \ldots, m)$ be m positive numbers. Let $\{x_k\}$ be generated by the gradient method (7.1.5) with the step-length α_k given by*

$$\alpha_k = \frac{(x_{v(k)} - x^*)A^{(\rho(k)-1)}(x_{v(k)} - x^*)}{(x_{v(k)} - x^*)A^{\rho(k)}(x_{v(k)} - x^*)}, \tag{7.2.1}$$

where $x^* = -A^{-1}g$, $\rho(k) \in \{q_1, q_2, \ldots, q_m\}$ and $v(k) \in \{k, k-1, \ldots, \max\{0, k-m\}\}$ for $k = 0, 1, 2, \ldots$. Then either $x_k = x^*$ for some finite k or the sequence $\{x_k\}$ converges to x^*.

For a proof of the above theorem, please see [10]. We can easily generalize the above theorem to the following more general form:

Theorem 7.2.2. *Let $f(x)$ be given by (7.1.4) and A is positive definite. Let m be a positive integer and γ be a positive number. Let $\{x_k\}$ be generated by the gradient method (7.1.5) with the step-length α_k satisfying*

$$\alpha_k \in \left[\min_{\substack{|\rho| \leq \gamma, \\ \max[0, k-m] \leq j \leq k}} \frac{g_j A^{(\rho-1)} g_j}{g_j A^{\rho} g_j}, \max_{\substack{|\rho| \leq \gamma, \\ \max[0, k-m] \leq j \leq k}} \frac{g_j A^{(\rho-1)} g_j}{g_j A^{\rho} g_j} \right]. \tag{7.2.2}$$

Then either $x_k = x^*$ for some finite k or the sequence $\{x_k\}$ converges to x^*.

Proof. Our proof is similar to that of Theorem 7.2.1 which is given by Friedlander et al.[10].

First it is easy to see that (7.2.2) implies that

$$0 < \frac{1}{\lambda_n} \leq \alpha_k \leq \frac{1}{\lambda_1} \tag{7.2.3}$$

holds for all k.

Let the orthogonal decomposition of A be as follows:

$$A = Q \Lambda Q^T, \tag{7.2.4}$$

where $Q = [q_1, q_2, \ldots, q_n]$ is an orthogonal matrix and $\Lambda = Diag[\lambda_1, \lambda_2, \ldots, \lambda_n]$.

For any given initial point x_0, the gradient $g_0 = Ax_0 + g$ can be expressed by

$$g_0 = \sum_{i=1}^{n} \beta_i^{(0)} q_i, \tag{7.2.5}$$

where $\beta_i^{(0)} \in \Re (i = 1, 2, \ldots, n)$. Let $g_k = \sum_{i=1}^{n} \beta_i^{(k)} q_i$ for all $k \geq 0$. It follows from (7.1.5), (7.1.4), (7.2.4) and (7.2.5) that

$$\beta_i^{(k+1)} = (1 - \alpha_k \lambda_i)\beta_i^{(k)} = \prod_{j=0}^{k}(1 - \alpha_j \lambda_i)\beta_i^{(0)} \,. \tag{7.2.6}$$

From (7.2.6), we have that

$$|\beta_1^{(k)}| = |(1-\alpha_{k-1}\lambda_1)\beta_1^{(k-1)}| \le \left|\left(1 - \frac{\lambda_1}{\lambda_n}\right)\beta_1^{(k-1)}\right| \le (1-\lambda_1/\lambda_n)^k |\beta_1^{(0)}| \,. \tag{7.2.7}$$

The above inequality shows that

$$\lim_{k\to\infty} \beta_1^{(k)} = 0 \,. \tag{7.2.8}$$

We see that the theorem is true if we can prove that

$$\lim_{k\to\infty} \beta_i^{(k)} = 0 \,, \tag{7.2.9}$$

for all $i = 1, 2, \ldots, n$. If this were not true, there would exist a positive number $\hat{\delta}$ and an integer $l \in [1, n-1]$ such that (7.2.9) holds for all $i = 1, \ldots, l$ and

$$\limsup_{k\to\infty} |\beta_{l+1}^{(k)}| > \hat{\delta} > 0 \,. \tag{7.2.10}$$

For any given positive number δ, we have that

$$\lim_{|\beta_{l+1}^{(k)}|\ge\delta, k\to\infty} \max_{\substack{|\rho| \le \gamma, \\ k-m \le j \le k}} \frac{g_j^T A^{\rho-1} g_j}{g_j^T A^{\rho} g_j} \tag{7.2.11}$$

$$= \lim_{|\beta_{l+1}^{(k)}|\ge\delta, k\to\infty} \max_{\substack{|\rho| \le \gamma, \\ k-m \le j \le k}} \frac{\sum_{i=1}^{n} (\beta_i^{(j)})^2 \lambda_i^{\rho-1}}{\sum_{i=1}^{n} (\beta_i^{(j)})^2 \lambda_i^{\rho}}$$

$$\le \lim_{|\beta_{l+1}^{(k)}|\ge\delta, k\to\infty} \max_{\substack{|\rho| \le \gamma, \\ k-m \le j \le k}} \frac{\sum_{i=1}^{l+1} (\beta_i^{(j)})^2 \lambda_i^{\rho-1}}{\sum_{i=1}^{l+1} (\beta_i^{(j)})^2 \lambda_i^{\rho}}$$

$$= \frac{1}{\lambda_{l+1}} \,, \tag{7.2.12}$$

due to the fact that (7.2.9) holds for $i = 1, \ldots, l$ and that inequality $|\beta_{l+1}^{(k)}| \ge \delta$ and relation (7.2.6) imply $|\beta_{l+1}^{(j)}| \ge \left(\frac{\lambda_1}{\lambda_{l+1}}\right)^m \delta$ holds for all $j \in [\max[0, k-m], k]$. Therefore, there exists a sufficiently large integer $\hat{k}$ such that

$$\alpha_k \le \frac{11}{10}\frac{1}{\lambda_{l+1}} \tag{7.2.13}$$

for all k satisfying $k \geq \hat{k}$ and

$$|\beta_{l+1}^{(k)}| \geq \frac{\lambda_1}{\lambda_{l+1}} \frac{\hat{\delta}}{2}. \tag{7.2.14}$$

Thus, for any $k \geq \hat{k}$, if (7.2.14) holds, we have that

$$\begin{aligned} |\beta_{l+1}^{(k+1)}| &= |(1 - \alpha_k \lambda_{l+1})||\beta_{l+1}^{(k)}| \\ &\leq \max\left[1 - \frac{\lambda_{l+1}}{\lambda_n}, \quad \left|\lambda_{l+1} \frac{11}{10} \frac{1}{\lambda_{l+1}} - 1\right|\right] |\beta_{l+1}^{(k)}| \\ &\leq \max[1 - \lambda_{l+1}/\lambda_n, \quad 0.1]|\beta_{l+1}^{(k)}|. \end{aligned} \tag{7.2.15}$$

On the other hand, if (7.2.14) does not hold, from (7.2.6) we can show that

$$|\beta_{l+1}^{(k+1)}| \leq \max[1 - \frac{\lambda_{l+1}}{\lambda_n}, \lambda_{l+1}/\lambda_1 - 1]|\beta_{l+1}^{(k)}| \leq \lambda_{l+1}/\lambda_1 |\beta_{l+1}^{(k)}| \leq \frac{\hat{\delta}}{2}. \tag{7.2.16}$$

It follows from (7.2.15) and (7.2.16) that

$$\limsup_{k \to \infty} |\beta_{l+1}^{(k)}| \leq \frac{\hat{\delta}}{2} \tag{7.2.17}$$

which contradicts (7.2.10). This completes our proof. □

It should be pointed out that the above result can also deduce from a more general convergence result of Dai[2] by showing that Property (A) of Dai[2] holds. From Dai's results it can be shown that the gradient method with (7.2.2) converges R-linearly. The reason for giving our direct and simple proof is to avoid unnecessary lengthy analysis.

Though the generalization from (7.2.1) to (7.2.2) is very simple and straightforward, it does contain more choices for the step-lengths. For example, we can let α_k be the mean values of any two Raleigh ratios:

$$\alpha_k = \frac{1}{2}\left[\frac{g_{j_1}^T A^{\rho_1 - 1} g_{j_1}}{g_{j_1}^T A^{\rho_1} g_{j_1}} + \frac{g_{j_2}^T A^{\rho_2 - 1} g_{j_2}}{g_{j_2}^T A^{\rho_2} g_{j_2}}\right], \tag{7.2.18}$$

or

$$\alpha_k = \sqrt{\frac{g_{j_1}^T A^{\rho_1 - 1} g_{j_1}}{g_{j_1}^T A^{\rho_1} g_{j_1}} \frac{g_{j_2}^T A^{\rho_2 - 1} g_{j_2}}{g_{j_2}^T A^{\rho_2} g_{j_2}}}. \tag{7.2.19}$$

Moreover, all the known choices of α_k (see (4.27)–(4.29) of Dai[2]) having Property (A) satisfy our simple condition (7.2.2).

7.3 Short BB steps

When we apply the gradient method to large scale problems, the most important issue is which step-length will give a fast convergence rate. Therefore it is vitally important to find what choices of α_k in the interval (7.2.2) require less number of iterations to reduce the gradient norm to a given tolerance. Much work has been done on this issue. And it seems that up to now the best choice is the adaptive BB step given by Zhou et al.[23] in which

$$\alpha_k = \alpha_k^{ABB} = \begin{cases} \alpha_k^{BB2}, & \text{if } \alpha_k^{BB2}/\alpha_k^{BB1} < \kappa; \\ \alpha_k^{BB1}, & \text{otherwise,} \end{cases} \tag{7.3.1}$$

and $\kappa \in (0, 1)$ is a parameter.

The motivation of the ABB step and its derivation can be found in Zhou et al.[23]. Basically, The ABB step is a hybrid combination of BB1 and BB2 steps, which mainly use the BB1 step unless α_k^{BB2} is much smaller than α^{BB1}.

It is trivial that $\alpha_k^{BB1} \geq \alpha_k^{BB2}$, namely the BB1 step is normally longer than the BB2 step. Numerical results favor the BB1 (the longer BB step). However, to the author's knowledge, there are no sound theoretical results which ensure that BB1 is better than BB2, though most papers choose to study the BB1 step when the BB method is studied. It seems all the theoretical results holding for the BB1 method are also true for the BB2 method. Thus it is very interesting to know why BB1 is better than BB2. It would be nice to produce sound theoretical results to shed light on this question. Unfortunately, we have not yet been able to do so. In the following paragraph, we give an intuitive analysis on the impact of the step-lengths for the gradient method.

In order to obtain a fast convergence, we need to make all the terms

$$\beta_i^{(k+1)} = (1 - \alpha_k \lambda_i)\beta_i^{(k)} = \prod_{j=0}^{k} (1 - \alpha_j \lambda_i)\beta_i^{(0)} \quad (i = 1, 2, \ldots, n), \tag{7.3.2}$$

converge to zero as fast as possible. Due to the relation (7.2.6), remembering that we have $1/\lambda_n \leq \alpha_k \leq 1/\lambda_1$, we can see that a smaller α_k will reduce $\beta_n^{(k+1)}$ more quickly, while a larger α_k will reduce the other $\beta_i^{(k+1)}$ (particularly with smaller i) more quickly. This observation tells us that either the longer BB step (BB1) or the shorter BB step (BB2) has its own advantage. Theoretically speaking, from the proof of Theorem 7.2.2, it is more easy to have $\beta_i^{(k)}$ converging to zero for small i (for example, $|\beta_1^{(k+1)}| \leq (1 - \lambda_1/\lambda_n)|\beta_1^{(k)}|$ for all k). Thus, we should put more weight for reducing $\beta_i(k)$ for large i, which means that we would prefer to use a shorter step-length. This seems to contradict the fact that the larger step (BB1) is better than the smaller step (BB2) based on many numerical tests.

For most test examples, we choose the starting point randomly, which would implies that $|q_i^T(x_0 - x^*)|(i = 1, 2, \ldots, n)$ are more or less of the same magnitude.

Thus, because

$$\beta_i^{(0)} = \lambda_i q_i^T (x_0 - x^*), \quad i = 1, 2, \ldots, n, \tag{7.3.3}$$

we would have that $|\beta_n^{(0)}|$ is much larger than the other $|\beta_i^{(0)}|$ if we assume

$$\lambda_n \gg \lambda_{n-1}. \tag{7.3.4}$$

In this case, we will see that the first iteration (with the exact line search) would give a very small step-length $\alpha_0 \approx 1/\lambda_n$. Consequently, $\beta_n^{(1)} \approx 0$ while $\beta_i^{(1)} \approx \beta_i^{(0)} (i = 1, 2, \ldots, n-1)$. Hence, from the second iteration on, it is more important to reduce the other $\beta_i (i = 1, 2, \ldots, n-1)$ instead of β_n. This may, in some sense, explain that a larger α_k (such as BB1) is better than a smaller α_k (such as BB2).

However, we would have a different picture when an iterate point x_k has the property that $|\beta_i^{(k-1)}| (i = 1, 2, \ldots n)$ are in the same order. For simplicity, we suppose that

$$|\beta_i^{(k-1)}|^2 \approx \|g_{k-1}\|_2^2 / n, \tag{7.3.5}$$

for all $i = 1, 2, \ldots, n$. Thus, we would have

$$\alpha_k^{BB1} \approx \frac{n}{\sum_{j=1}^n \lambda_i}, \tag{7.3.6}$$

and

$$\alpha_k^{BB2} \approx \frac{\sum_{j=1}^n \lambda_i}{\sum_{j=1}^n \lambda_i^2}. \tag{7.3.7}$$

These give that

$$\alpha_k^{BB1} \approx \frac{n}{\lambda_n} >> \frac{1}{\lambda_n} \approx \alpha_k^{BB2}. \tag{7.3.8}$$

In this case, it can be easily seen that normally the shorter BB step $\alpha_k = \alpha_k^{BB2}$ would give a smaller $\|g_{k+1}\|$. Therefore, it is reasonable for us to believe the shorter BB step (BB2) would be efficient if we want to obtain a very accurate solution of a very large-scale and ill-conditioned problem.

Hence, we would like to investigate the behavior of the BB2 step and shorter BB2 steps. What made us to explore the shorter steps is the curiosity on why BB2 in general performs worse than BB1. Another motivation is our belief that for very large scale and ill-conditioned problems a shorter step may be more efficient than a larger step. We consider the short BB2 step:

$$\alpha_k = \alpha_k^{SBB(m)} = \min_{\max[0,k-m] \le j \le k} \alpha_j^{BB2}, \tag{7.3.9}$$

where m is a given non-negative integer. If $m = 0$, the step-length (7.3.9) is nothing but the BB2 step. For $m > 0$, α_k given by (7.3.9) is not larger and may be smaller than α_k^{BB2}. Thus we call the method (7.1.5) with (7.3.9) the short BB method (SBB). The step-length defined by (7.3.9) satisfies (7.2.2), which means that the SBB method always converges for convex quadratic functions.

7.4 Numerical results

In this section, we test our SBB method, namely the gradient method (7.1.5) with (7.3.9). We call our SBB method with parameter m SBB(m). Different parameters $m = 1, 2, 3, 4, 9$ and 19 are used. We compared our algorithms with BB1, BB2 and the adaptive BB (ABB) of Zhou et al.[23]. For the ABB method, $\kappa = 0.25$ is used.

The problem we used to compare the algorithms is the one suggested by Yuan[21]. The function to be minimized has the following form:

$$f(x) = (x - x^*)^T Diag(\lambda_1, \cdots, \lambda_n)(x - x^*). \quad x \in \Re^n. \tag{7.4.1}$$

The diagonal structure of the Hessian of the objective function does not lose generality because the gradient method is invariant with respect to orthogonal transformations. We test problems from small scale to large scale, with $n = 10^i (i = 1, 2, 3, 4, 5, 6)$. The solution vector $x_i^* (i = 1, \ldots, n) \in (-5, 5)$ is randomly generated. We let $\lambda_1 = 1$ and $\lambda_n = Cond (= 10^L, L = 1, 2, 3, 4, 5, 6)$ which is the condition number of the Hessian of function $f(x)$. $\lambda_i (i = 2, \cdots, n-1)$ are randomly chosen in the interval $(1, \lambda_n)$. For all problems the initial point is the zero vector $(0, \cdots, 0)^T$. We use two stop conditions. One is

$$\|g_k\|_2 \leq 10^{-5} \|g_0\|_2, \tag{7.4.2}$$

and the other is

$$\|g_k\|_2 \leq 10^{-5}. \tag{7.4.3}$$

The numerical results with the two different stopping conditions (7.4.2) and (7.4.3) are reported in Table 7.1 and Table 7.2 separately. For each case (different n and different λ_n), 10 runs are made and the average numbers of iterations required by each algorithm are listed. For each case, The least average iteration number is given in bold font to indicate the winner amongst all the algorithms.

When the stopping condition is (7.4.2), from Table 7.1 we find that the ABB method is the winner, as it wins 14 out of the all 36 cases. Ranking by achieving the least number of iterations, the next best algorithms are SBB(2), SBB(1) and BB1, with winning in 7, 6 and 5 cases respectively. Table 7.1 also shows that BB1 is much better than BB2 as expected. If we make a one-to-one comparison between BB1 and BB2 in all the 36 cases, we find that BB2 wins only 5 cases while BB1 wins 31 cases. These results agree with the general belief that BB1 is better than BB2.

Now, let us discuss Table 7.2, where the results with the stopping condition (7.4.3) are given. Since in general the stopping condition (7.4.3) is more strict than (7.4.2), some algorithms fail to find a solution within the maximum-allowed CPU time, which is set to 10 minutes. In Table 7.2, a number marked with a superscript "$*$" is the iteration number of a single run instead of the average of 10 runs. While "Fail" indicates that even a single run failed to find a solution within 10 minutes.

Table 7.1 Iteration numbers of different gradient methods ($\|g_k\|_2 \leq 10^{-5}||g_0||_2$)

n	λ_n	ABB	BB1	BB2	m=2	m=3	m=4	m=5	m=10	m=20
10	10	19.1	19.1	**18.1**	18.7	19.0	19.8	22.0	22.5	29.9
10	10^2	**42.7**	47.8	54.5	45.1	49.9	52.2	46.5	52.9	67.7
10	10^3	**51.9**	103.2	92.9	87.8	77.7	82.6	72.1	73.6	76.2
10	10^4	76.2	169.6	199.3	106.6	102.4	77.4	72.3	**59.0**	62.8
10	10^5	22.2	21.6	24.0	**20.9**	21.0	22.5	22.2	26.5	34.2
10	10^6	22.1	23.2	22.1	22.4	21.8	**21.6**	23.0	28.0	33.9
10^2	10	19.4	19.4	20.0	19.0	**18.9**	20.8	21.2	28.3	29.8
10^2	10^2	**46.6**	50.9	56.8	56.5	53.2	55.7	57.1	57.2	64.8
10^2	10^3	96.3	110.2	109.0	116.0	111.0	103.2	**94.3**	103.5	125.0
10^2	10^4	149.9	171.6	211.1	168.8	172.9	167.7	144.1	**130.3**	141.9
10^2	10^5	**75.8**	92.5	124.9	99.2	93.1	81.0	82.1	85.1	95.1
10^2	10^6	77.9	88.2	92.0	86.1	**74.4**	78.0	81.5	79.0	96.7
10^3	10	19.0	19.0	20.3	**18.7**	19.0	20.5	21.0	26.7	30.0
10^3	10^2	**50.1**	53.7	56.5	56.4	53.4	60.0	59.7	59.9	65.5
10^3	10^3	**103.3**	109.2	111.8	108.7	104.3	**103.3**	111.5	113.2	133.5
10^3	10^4	106.7	109.2	109.3	107.7	105.6	105.5	**105.0**	108.6	129.3
10^3	10^5	111.4	**108.1**	124.9	120.8	117.2	110.4	123.2	119.2	129.3
10^3	10^6	103.2	107.3	114.1	114.8	108.8	**100.6**	109.5	109.2	124.9
10^4	10	19.0	19.0	19.9	**18.8**	19.0	20.2	21.0	26.3	30.0
10^4	10^2	**52.6**	54.0	56.3	54.9	55.2	55.7	56.3	57.9	65.2
10^4	10^3	**101.4**	111.3	117.2	108.1	101.5	103.7	107.7	106.9	126.0
10^4	10^4	116.6	120.0	124.5	127.0	117.5	**115.8**	121.4	118.3	140.9
10^4	10^5	117.7	123.0	119.0	117.5	119.4	112.0	112.3	**111.8**	132.8
10^4	10^6	**107.6**	121.6	116.0	118.3	111.1	116.2	114.7	117.6	132.5
10^5	10	**19.0**	**19.0**	20.0	**19.0**	**19.0**	20.0	21.0	26.0	30.0
10^5	10^2	53.7	52.7	59.0	**52.2**	54.4	55.0	53.6	57.1	65.0
10^5	10^3	101.9	96.5	110.9	108.5	**95.9**	102.0	104.4	116.5	121.0
10^5	10^4	**111.6**	121.8	123.8	118.2	115.5	116.7	121.7	119.9	141.8
10^5	10^5	**110.9**	117.2	120.0	117.7	116.7	124.4	124.1	120.8	142.4
10^5	10^6	117.6	**115.5**	129.6	125.3	123.7	120.9	124.3	120.9	142.6
10^6	10	**19.0**	**19.0**	20.0	**19.0**	**19.0**	20.0	21.0	26.0	30.0
10^6	10^2	52.0	51.3	59.5	49.8	54.4	56.0	**48.1**	57.0	65.0
10^6	10^3	94.5	103.1	110.7	108.4	**94.0**	98.1	102.3	111.3	121.0
10^6	10^4	110.8	116.8	123.6	122.2	113.1	112.4	124.3	120.3	148.7
10^6	10^5	125.5	**114.4**	118.6	120.8	121.1	115.6	122.8	120.3	149.0
10^6	10^6	**113.6**	113.9	123.6	119.1	**113.6**	114.7	122.3	120.3	149.0

It is surprising to find that in Table 7.2 the winner now is SBB (19), which wins 18 cases out of all the 36 cases, particularly it wins all the cases when both n and λ_n are large. Please notice that the only change that makes Table 7.1 and Table 7.2 different is the stopping condition. For a given case, the 10 repeated randomly generated problems for both Table 7.1 and Table 7.2 are also the same. This shows that SBB(19) uses far less numbers of iterations to reduce $\|g_k\|_2$ from $10^{-5}\|g_0\|_2$ to 10^{-5}. Another interesting point is that now BB2 performs much better than BB1, which is unexpected. If we have a one-to-one comparison between BB1 and BB2, we find that BB2 wins 23 cases while BB1 wins only 11 cases. Particularly,

Table 7.2 Iteration numbers of different gradient methods ($\|g_k\|_2 \leq 10^{-5}$)

n	λ_n	ABB	BB1	BB2	m=1	m=2	m=3	m=4	m=9	m=19
10	10	24.4	24.4	25.3	**24.2**	24.9	25.3	26.3	32.7	32.7
10	10^2	66.7	81.4	85.6	69.5	72.9	73.3	69.4	**65.4**	76
10	10^3	98.6	200.3	228.1	183.6	127.7	125.1	105.7	92.4	**91.6**
10	10^4	117.1	679.5	608.0	286.3	158.5	116.7	103.6	**77.4**	85.4
10	10^5	329.8	2235.0	1148.0	356.5	169.7	122.9	105.2	**92.3**	95
10	10^6	1011.6	3052.3	908.4	307.7	152.2	137.1	100.3	**97.6**	102.2
10^2	10	30.3	30.3	30.4	30.6	30.7	30.0	**29.4**	34.1	35.6
10^2	10^2	**85.8**	98.5	106.5	98.3	96.5	93.8	93.4	101.8	111.2
10^2	10^3	197.4	328.8	318.2	300.4	277.5	258.3	223.2	**179**	182.8
10^2	10^4	310.4	989.9	937.1	879.9	640.7	502.5	437.1	251.5	**220.6**
10^2	10^5	774.5	2872.9	1903.9	1261.4	792.5	599.5	480.6	260.3	**225.6**
10^2	10^6	1214.4	4522.3	2038.0	1114.8	672.7	537.7	443.8	258.2	**236.4**
10^3	10	32.2	32.2	**30.5**	31.8	31.8	33	31.6	36.1	38.0
10^3	10^2	**101.1**	109.8	104.1	109.2	106.8	105.5	107.3	106.6	121.1
10^3	10^3	286.8	343.5	385.4	341.2	305.3	319.8	283.5	217.9	**216.1**
10^3	10^4	530.3	1140.3	1149.2	965.8	819.9	694.3	521.0	337.4	**302.3**
10^3	10^5	988.9	2927.7	2443.9	1638.9	1096.4	934.0	717.8	401.5	**346**
10^3	10^6	1587.5	4534.8	2310.5	1590.7	1076.2	858.2	693.9	384.6	**333.7**
10^4	10	34.3	34.3	**31.1**	32.6	33.1	37.7	34.1	37.0	48.2
10^4	10^2	**106.9**	115.0	113.3	111.2	108.8	114.2	113.6	114.0	133.3
10^4	10^3	334.4	380.7	370.2	385.1	351.8	321.7	304.9	**233.1**	233.5
10^4	10^4	921.6	1272.5	1250.8	1071.9	888.8	742.6	646.0	387.4	**338.4**
10^4	10^5	1610.6	3226.1	2574.3	1827.0	1424.7	1014.5	940.8	502.0	**425.1**
10^4	10^6	2316.0	5954.0	3404.7	1993.3	1602.7	1271.6	1069.7	534.6	**454.1**
10^5	10	35.0	35.0	**33.5**	35.7	35.3	39.3	38.4	38.0	57.0
10^5	10^2	**115.3**	119.3	121.7	117.5	115.6	117.7	122.2	120.5	150.3
10^5	10^3	349.5	402.0	416.8	392.2	367.9	325.9	317.9	**239.3**	244.8
10^5	10^4	1055.7	1341.9	1195.8	1057.7	869.4	748.5	645.7	389.3	**357.1**
10^5	10^5	2291.1	2640*	2475.3	2088.8	1527.6	1217.6	979.6	559.4	**461.0**
10^5	10^6	2140*	5120*	3149*	2445.1	1924.8	1571.7	1263.5	654.9	**531.5**
10^6	10	36.9	36.9	**35.2**	36.7	36.5	39.4	41.0	40.0	58.0
10^6	10^2	**121.1**	123.4	131.5	129.3	125.3	123.0	123.1	122.3	150.3
10^6	10^3	369*	376*	461*	470*	410*	387*	303*	243*	**220***
10^6	10^4	1154*	1130*	1325*	1128*	977*	760*	613*	406*	**364***
10^6	10^5	Fail	Fail	Fail	Fail	1557*	1192*	1023*	520*	**456***
10^6	10^6	Fail	Fail	Fail	Fail	Fail	1357*	1343*	652*	**567***

BB2 wins over BB1 for all the very ill-conditioned cases, when $\lambda_n = 10^5$ or $\lambda_n = 10^6$.

Now we consider two specific distributions of the eigenvalues $\lambda_i (i = 2, \ldots, n-1)$. The first case is

$$\frac{\lambda_{i+1}}{\lambda_i} = \left(\frac{\lambda_n}{\lambda_1}\right)^{\frac{1}{n-1}}, \quad i = 1, 2, \ldots, n-1. \tag{7.4.4}$$

And the second case is to have $\lambda_i (i = 2, \ldots, n-1)$ equally distributed in the interval $(1, \lambda_n)$. To be more exact, we let $\lambda_{i+1} - \lambda_i$ be constant:

$$\lambda_{i+1} - \lambda_i = \frac{\lambda_n - \lambda_1}{n-1}, \quad i = 1, 2, \ldots, n-1. \tag{7.4.5}$$

In both cases, the stopping condition is

$$\|g_k\|_2 \leq 10^{-6}\|g_0\|_2. \tag{7.4.6}$$

The solution is chosen by vector $x_i^*(i = 1, \ldots, n) \in (-0.5, 0.5)$ randomly, and the average number of iterations out of 10 runs with the fixed starting point $(0, 0, \ldots, 0)^T$ are given in Table 7.3 and Table 7.4 for the two cases respectively.

From Table 7.3, the numerical results favor our SBB method over the ABB, BB1 and BB2 methods. Particularly, for very ill-conditioned problems, namely problems with $\lambda_n = 10^5$, our SBB method performs much better than ABB, BB1 and BB2. Therefore, we believe that for ill-conditioned problems with $\lambda_{i+1}/\lambda_1 \approx$ *constant*, the SBB method will be much faster than ABB, BB1 and BB2.

Table 7.3 Iteration numbers of different gradient methods when λ_{i+1}/λ_i is constant.

n	λ_n	ABB	BB1	BB2	m=1	m=2	m=3	m=4	m=9	m=19
10	10	25.3	25.3	25.5	**24.7**	25.3	25.5	26.2	34.3	32.4
10^3	10	25.4	25.4	26.3	25.7	25.9	**23.1**	24.3	32.0	32.2
10^5	10	25.0	25.0	26.0	25.0	26.0	**23.0**	25.0	32.0	33.0
10	10^3	178.9	173.5	169.1	149.1	113.5	103.7	**95.5**	101.6	132.9
10^3	10^3	200.3	245.4	246.8	230.4	230.8	211.8	212.2	187.5	**179.7**
10^5	10^3	196.5	215.8	240.0	236.8	202.5	202.4	213.0	**193.7**	222.0
10	10^5	836.9	802.8	720.7	203.6	130.0	**118.7**	127.2	136.7	179.5
10^3	10^5	1262.6	1444.1	1702.1	1391.2	1148.5	966.6	832.8	477.9	**377.6**
10^5	10^5	1297.5	1482.5	1447.0	1406.7	1291.2	1175.5	1025.7	678.8	**437.3**

From Table 7.4, we can see that all the algorithms perform more or less the same. Actually, the ABB method, which wins when $\lambda_n = 10^3$ for $n = 10$, 10^3 and 10^5, can be regarded as the overall best method when the stopping condition is (7.4.6). Similar to the phenomenon revealed in Tables 7.1 and 7.2, we also observe that our SBB method will outperform ABB method if a more accurate

Table 7.4 Iteration numbers of different gradient methods when $\lambda_{i+1} - \lambda_i$ is constant.

n	λ_n	ABB	BB1	BB2	m=1	m=2	m=3	m=4	m=9	m=19
10	10	22.0	22.0	23.9	**21.6**	23.0	24.0	25.0	33.8	32.5
10^3	10	24.4	24.4	25.7	23.5	23.4	**22.9**	24.1	33.0	32.0
10^5	10	24.0	24.0	26.0	**23.0**	**23.0**	**23.0**	24.0	33.0	32.0
10	10^3	**54.8**	130.4	131.7	125.8	96.1	108.4	107.8	96.5	70.5
10^3	10^3	**145.6**	163.7	168.4	157.8	162.3	160.7	155.1	147.1	177.0
10^5	10^3	**144.5**	158.0	163.0	155.9	146.5	152.5	153.0	149.7	168.5
10	10^5	237.9	382.1	549.9	280.3	158.3	111.6	104.5	**77.0**	109.7
10^3	10^5	151.4	164.0	185.7	164.4	165.0	155.0	165.4	**148.8**	159.2
10^5	10^5	262.0	286.4	295.3	274.7	266.9	282.7	268.1	**238.5**	244.1

solution is needed. For example, let us consider the situation when $n = 10^3$ and $\lambda_n = 10^3$, which is the case that ABB wins for the stopping condition (7.4.6). If the stopping condition is replaced by $\|g_k\|_2 \leq 10^{-9}\|g_0\|_2$ the ABB method needs 263.6 iterations while the SBB(9) method needs 243.7 iterations. If we use an even more strict stopping condition $\|g_k\|_2 \leq 10^{-12}\|g_0\|_2$, the ABB method would need 380.2 iterations against 304.3 iterations by the SBB(9) method. Even for the situation when $n = 10$ and $\lambda_n = 10^3$, for which the ABB method preforms much better than the other methods under the stopping condition (7.4.6), we find that the ABB method requires 126.0 iterations comparing 117.2 iterations by SBB(9) method if the stopping condition is replaced by $\|g_k\|_2 \leq 10^{-13}\|g_0\|_2$. Of course, in real applications it is unlikely to require such high accurate solutions.

For many practical problems, matrix A is obtained by finite difference approximation to Laplace's equation[9, 23]. For such A, we can easily see that the differences $\lambda_{i+1} - \lambda_i$ are of the same magnitude for many i. Therefore, we expect that for such problems derived from Laplace equations the best gradient method to use is the ABB method. Indeed, we treated the Laplace1 (b) problem of Fletcher[9] in which A is defined by

$$A = \begin{bmatrix} W & -I & & & \\ -I & W & -I & & \\ & -I & W & \ddots & \\ & & \ddots & \ddots & -I \\ & & & -I & W \end{bmatrix} \in \Re^{10^6 \times 10^6} \tag{7.4.7}$$

where

$$W = \begin{bmatrix} T & -I & & & \\ -I & T & -I & & \\ & -I & T & \ddots & \\ & & \ddots & \ddots & -I \\ & & & -I & T \end{bmatrix} \in \Re^{10^4 \times 10^4}, \qquad T = \begin{bmatrix} 6 & -1 & & & \\ -1 & 6 & -1 & & \\ & -1 & 6 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 6 \end{bmatrix} \in \Re^{10^2 \times 10^2}. \tag{7.4.8}$$

It is known[9] for this matrix A we have $\lambda_n/\lambda_1 \approx 4133.6$. In Table 7.5, we give the numbers of iterations needed for all the algorithms with different stopping conditions

$$\|Ax_k - b\|_2 \leq \theta\|b\|_2, \qquad \theta = 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}. \tag{7.4.9}$$

The starting point is $x_0 = (0, 0, \ldots, 0)^T$ for all the runs.

Our results in Table 7.5 confirm the finding of Zhou et al.[23] that the ABB method is better than the BB1 method. Moreover, for this specific problem, BB1 is much better than BB2. Though our short BB2 steps do show improvements over the original BB2 method, the SBB methods do not perform as good as the ABB method for this Laplace1(b) problem. Therefore it is reasonable for us to

Table 7.5 Iterations for Laplace1 (b) with different stopping conditions.

θ	ABB	BB1	BB2	m=1	m=2	m=3	m=4	m=9	m=19
10^{-4}	173	176	157	178	200	199	**166**	181	225
10^{-5}	**276**	394	392	278	289	298	290	322	417
10^{-6}	387	462	611	374	426	558	**361**	442	605
10^{-7}	460	510	864	478	**458**	760	493	652	701
10^{-8}	**570**	590	1017	737	601	844	645	820	759
10^{-9}	**590**	611	1062	775	819	851	676	881	942

believe that if we want to find a better gradient method than the ABB method for such problems derived from Laplace equations, we might need to explore special step-lengths which make good use of the special eigenvalue distributions of such matrices. This is an interesting and important problem to study because many practical problems are derived from Laplace equations.

7.5 Discussion and conclusion

In this chapter we have generalized the convergence result for the gradient method with retards by Friedlander et al.[10] from (7.2.1) to (7.2.2). Our simple generalization allows more choices of step-lengths α_k. We give an intuitive analysis on the impact of the step-length of the gradient method for large-scale and ill-conditioned problems, and believe that short step-lengths in the interval (7.2.2) should perform better than long step-lengths. We propose the short BB2 (SBB) method, which uses the smallest value of all the BB2 step-lengths in the previous m iterations as the step-length. Numerical results of large-scale and ill-conditioned problems show that the SBB method performs better than the BB methods and the adaptive BB method if a high accurate solution is needed. Our numerical results also reveal that BB2 is better than BB1 when we need to find a very high accurate solution for large-scale and ill-conditioned problems. This is, to some extent, a surprising discovery because in general it has been widely regarded that BB1 is better than BB2. Our numerical results also suggest that corresponding special step-lengths might be needed to construct efficient gradient methods for solving problems with certain special eigenvalue distributions such as those problems derived from Laplace equations.

Acknowledgements This work was partially supported by the National Natural Science Foundation of China under grant No.10831006 and by Chinese Academy of Sciences under grant No.kjcx-yw-s7.

References

1. J. Barzilai and J. M. Borwein, Two point step size gradient methods, *IMA J. Numer. Anal.*, **8**, 141–148, 1988.
2. Y. H. Dai, Alternate step gradient method, *Optimization*, **52**, 395–415, 2003.
3. Y. H. Dai and R. Fletcher, On the asymptotic behaviour of some new gradient methods, *Math. Program*, **13**, 541–559, 2005.
4. Y. H. Dai and L. Z. Liao, R-linear convergence of the Barzilai and Borwein gradient method, *IMA J. Numer. Anal.*, **22**, 1–10, 2002.
5. Y. H. Dai, J. Y. Yuan and Y. Yuan, Modified two-point step-size gradient methods for unconstrained optimization, *Computational Optimization and Applications*, **22**, 103–109, 2002.
6. Y. H. Dai and Y. Yuan, Alternate minimization gradient method, *IMA Journal of Numerical Analysis*, **23**, 377–393, 2003.
7. Y. H. Dai and Y. Yuan, Analysis of monotone gradient methods, *J. Industrial and Management Optimization*, **1**, 181–192, 2005.
8. Y. H. Dai and H. Zhang, An adaptive two-point step-size gradient method, *Numerical Algorithm*, **27**, 377–385, 2001.
9. R. Fletcher, *On the Barzilar-Borwein method*, Research Report, University of Dundee, UK, 2001.
10. A. Friedlander, J. M. Martínez, B. Molina and M. Raydan, Gradient method with retards and generalizations, *SIAM J. Numer. Anal.*, **36**, 275–289, 1999.
11. W. W. Hager and H. Zhang, A new active set algorithm for box constrained optimization, *SIAM J. Optim.*, **17**, 526–557, 2006.
12. J. Nocedal, A. Sartenaer and C. Zhu, On the behavior of the gradient norm in the steepest descent method, *Computational Optimization and Applications*, **22**, 5–35, 2002.
13. M. Raydan, On the Barzilai and Borwein choice of steplength for the gradient method, *IMA J. Numer. Anal.*, **13**, 321–326, 1993.
14. M. Raydan, The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem, *SIAM J. Optim.*, **7**, 26–33, 1997.
15. M. Raydan and B. F. Svaiter, Relaxed steepest descent and Cauchy-Barzilai-Borwein method, *Computational Optimization and Applications*, **21**, 155–167, 2002.
16. M. N. Vrahatis, G. S. Androulakis, J. N. Lambrinos and G. D. Magoulas, A class of gradient unconstrained minimization algorithms with adaptive step-size, *J. Comp. and Appl. Math.*, **114**, 367–386, 2000.
17. Y. F. Wang, Y. Yuan and H. C. Zhang, A trust region-CG algorithm for delurring problem in atmospheric image reconstruction, *Science in China*, **45**, 731–740, 2002.
18. Y. F. Wang and Y. Yuan, On the regularity of a trust region-CG algorithm for nonlinear ill-posed inverse problems, in: T. Sunada, P.W. Sy and L. Yang, Eds., *Proceedings of the Third Asian Mathematical Conference*, 562–580, World Scientific, Singapore, 2002.
19. Y. F. Wang and Y. Yuan, A trust region method for solving distributed parameter identification problems, *Journal of Comp. Math.*, **21**, 759–772, 2003.
20. Y. F. Wang and Y. Yuan, Convergence and regularity of trust region methods for nonlinear ill-posed inverse problems, *Inverse Problems*, **21**, 821–838, 2005.
21. Y. Yuan, A new stepsize for the steepest descent method, *Journal of Comp. Math.*, **24**, 149–156, 2006.
22. Y. Yuan, Step-sizes for the gradient method, in: K. S. Liu, Z.P. Xin and S. T. Yau, eds., *Third International Congress of Chinese Mathematicians*, AMS/IP Studies in Advanced Mathematics, 785–796, 2008.
23. B. Zhou, L. Gao and Y. H. Dai, Gradient methods with adaptive stp-sizes, *Computational Optimization and Applications*, **35**, 69–86, 2006.

Chapter 8
Convergence Analysis of Nonlinear Conjugate Gradient Methods

Yuhong Dai

Abstract. Conjugate gradient methods are a class of important methods for unconstrained optimization and vary only with a scalar β_k. In this chapter, we analyze general conjugate gradient method using the Wolfe line search and propose a condition on the scalar β_k, which is sufficient for the global convergence. An example is constructed, showing that the condition is also necessary in some sense for the global convergence of general conjugate gradient method. To make better use of the condition, we introduce a new property for conjugate gradient methods. It is shown that many conjugate gradient methods have such property, including the FR, PRP, HS, and DY methods and the FR-PRP, and DY-HS hybrid methods. Consequently, convergence results are gained for these methods under mild assumptions. In addition, an analysis is also given to a new conjugate gradient method, which further demonstrates the usefulness of the condition and the new property. Some discussions about the bound in the hybrid conjugate gradient methods are also given.

8.1 Introduction

Conjugate gradient methods are a class of important methods for solving the the unconstrained nonlinear optimization problem

$$\min f(x), \quad x \in R^n, \tag{8.1.1}$$

Yuhong Dai
State Key Laboratory of Scientific and Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
P. O. Box 2719, Beijing 100190, China.
e-mail: dyh@lsec.cc.ac.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

especially if the dimension n is large. The methods without regular restarts are of the form

$$x_{k+1} = x_k + \lambda_k d_k, \tag{8.1.2}$$
$$d_k = -g_k + \beta_k d_{k-1}, \tag{8.1.3}$$

where $d_1 = -g_1 = -\nabla f(x_1)$, λ_k is a step-length obtained by a line search, and β_k is a scalar. The step-length λ_k is often required to satisfy the strong Wolfe conditions, namely,

$$f(x_k + \lambda_k d_k) - f(x_k) \leq \rho \lambda_k g_k^T d_k, \tag{8.1.4}$$
$$|g(x_k + \lambda_k d_k)^T d_k| \leq -\sigma g_k^T d_k, \tag{8.1.5}$$

where $0 < \rho < \sigma < 1$. The scalar β_k should be so chosen that the method (8.1.2)–(8.1.3) reduces to the linear conjugate gradient method in the case when f is a convex quadratic and the line search is exact. Some well-known formulae for β_k are called the FR [9], PRP [15, 16], HS [12] and DY [7] formulae, and are given by

$$\beta_k^{\mathrm{FR}} = ||g_k||^2 / ||g_{k-1}||^2, \tag{8.1.6}$$
$$\beta_k^{\mathrm{PRP}} = g_k^T y_{k-1} / ||g_{k-1}||^2, \tag{8.1.7}$$
$$\beta_k^{\mathrm{HS}} = g_k^T y_{k-1} / d_{k-1}^T y_{k-1}, \tag{8.1.8}$$
$$\beta_k^{DY} = \|g_k\|^2 / d_{k-1}^T y_{k-1}, \tag{8.1.9}$$

respectively, where $\|\cdot\|$ denotes the l_2-norm of R^n, and $y_{k-1} = g_k - g_{k-1}$.

Although all these methods have the quadratic termination property, their convergence properties and numerical performances may be very different for general objective functions. Basically, nonlinear conjugate gradient methods can be divided into the following three categories. The first category includes the FR, and DY methods, etc. In practical computations, this category of methods perform worse than the second, and third categories for they may produce small steps continuously ([17, 8]). However, their convergences can be achieved under mild assumptions. For example, the FR method with the strong Wolfe line search is shown to converge globally for general functions if the scalar σ in (8.1.5) is not greater than 0.5 (for example, see [1, 5]). The DY method is globally convergent provided that λ_k satisfies the Wolfe conditions, namely, (8.1.4) and

$$g(x_k + \lambda_k d_k)^T d_k \geq \sigma g_k^T d_k, \tag{8.1.10}$$

where $0 < \rho < \sigma < 1$ [7]. The second category includes the PRP, and HS methods, etc. If a small step occurs, the methods in this category can generate search directions close to the negative gradient direction and hence avoid the propensity of small steps [17, 10]. As a result, they perform often much better than the first category of methods. However, Powell [17] was able to show that the PRP method with exact line searches can cycle round eight nonstationary points.

The example also applies to the HS method since the two methods are the same in case of exact line searches. Till 1992, Gilbert and Nocedal [10] gave the global convergence of the PRP method with the restriction $\beta_k \geq 0$ for general functions. The third category includes the FR-PRP, and DY-HS hybrid methods, etc. This category was first proposed by Touati-Ahmed and Storey [20]. They suggested the following hybrid method of the FR, and PRP methods:

$$\beta_k = \max\{0, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}. \tag{8.1.11}$$

Gilbert and Nocedal [10] further considered the hybrid method

$$\beta_k = \max\{-\beta_k^{FR}, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}, \tag{8.1.12}$$

that allows negative values. The hybrid methods (8.1.11) and (8.1.12) have the following advantages: (i) their convergences can be achieved similar to FR; and (ii) they can avoid the propensity of small steps like PRP. However, their numerical performances are worse than PRP, though better than FR (for example, see [10]). Dai and Yuan [8]) extended the convergence result of the DY method in [7] to the following hybrid method:

$$\beta_k = \max\{0, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}. \tag{8.1.13}$$

Promising numerical results were also obtained for this hybrid method [8, 4]. For a collection of medium and large-scale problems drawn from CUTE [2], it was shown that the use of the Wolfe line search in the hybrid method (8.1.13) is better than the use of the strong Wolfe line search, and that the method with the Wolfe line search performs better than the PRP method with the strong Wolfe line search for most of the test problems.

Although the already-existing results offer fascinating glimpses into the behavior of conjugate gradient methods, its theory still remains fragmentary [14]. A comprehensive theory of conjugate gradient methods, which is regarded in [14] as one of the outstanding challenges in theoretical optimization, is then to be developed. Reference [3] analyzed general conjugate gradient method with the strong Wolfe line search, and showed that the method converges globally if $\|d_k\|^2$ increases mostly linearly (see Lemma 8.2.3). Since it is possible to get convergence results and develop efficient algorithms in the conjugate gradient field via the Wolfe line search, as stated in the previous paragraph, we will analyze general conjugate gradient method with the Wolfe line search in this chapter. Specifically, since conjugate gradient methods vary only with the scalar β_k, we ask the following question: what condition on β_k can ensure the global convergence of general conjugate gradient method?

We will provide an answer to this question in Section 8.3 after giving some preliminaries in the next section. See (8.3.5) for the condition. An example is also constructed in Section 8.3, which shows that the condition (8.3.5) is necessary in some sense for the global convergence of general conjugate gradient method. To make better use of the condition, we will introduce a new property, namely,

Property (#) , for conjugate gradient methods in Section 8.4. It is shown that all the three categories of conjugate gradient methods may have such property. As a result, convergence results can be obtained conveniently for these methods under suitable assumptions. An analysis is also given to a new conjugate gradient method in Section 8.4, which further demonstrates the usefulness of the condition (8.3.5) and Property (#) . Some discussions about the bound in hybrid methods are made in the last section.

8.2 Some preliminaries

Throughout this chapter, we assume that $g_k \neq 0$ for all $k \geq 1$ for otherwise a stationary point has been found. We also assume that $\beta_k \neq 0$ for all $k \geq 1$. This is because if $\beta_k = 0$, the direction in (8.1.3) reduces to the negative gradient direction. Thus either the method converges globally if $\beta_k = 0$ for infinite number of k, or one can regard some x_k as the new initial point.

We give the following basic assumptions of the objective function.

Assumption 8.2.1. *(i) The level set $\mathcal{L} = \{x \in R^n : f(x) \leq f(x_1)\}$ is bounded, where x_1 is the starting point; (ii) In some neighborhood $\mathcal{N}$ of $\mathcal{L}$, f is continuously differentiable, and its gradient is Lipschitz continuous; namely, there exists a constant $L > 0$ such that*

$$\|g(x) - g(y)\| \leq L\|x - y\|, \quad for\ all\ x, y \in \mathcal{N}. \tag{8.2.1}$$

Under Assumption 8.2.1 on f, we state a very useful result, which was mainly obtained by Zoutendijk [23] and Wolfe [21, 22].

Lemma 8.2.2. *Suppose that Assumption 8.2.1 holds. Consider any iterative method of the form (8.1.2), where d_k satisfies $g_k^T d_k < 0$ and λ_k is obtained by the Wolfe line search. Then*

$$\sum_{k=1}^{\infty} \frac{(g_k^T d_k)^2}{||d_k||^2} < +\infty. \tag{8.2.2}$$

Relation (8.2.2) is normally called the Zoutendijk condition. In the case that the sufficient descent condition holds,

$$-g_k^T d_k \geq c\|g_k\|^2, \quad \text{for some } c > 0 \text{ and all } k \geq 1, \tag{8.2.3}$$

we can conclude from (8.2.2) that if $\|d_k\|^2$ increases at most linearly,

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} = +\infty, \tag{8.2.4}$$

the iterative method converges in the sense that

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{8.2.5}$$

In fact, the sufficient descent condition (8.2.3) is often implied or required in many convergence analyses of conjugate gradient methods, for example see [1, 10, 11, 13, 20]. However, for general method (8.1.2)–(8.1.3) with strong Wolfe line searches, Dai et al. [3] showed that this result still holds even if the sufficient descent condition (8.2.3) is replaced with the descent condition $g_k^T d_k < 0$.

Lemma 8.2.3. *Suppose that Assumption 8.2.1 holds. Consider any iterative method of the form (8.1.2)–(8.1.3), where d_k satisfies $g_k^T d_k < 0$ and λ_k is obtained by the strong Wolfe line search. Then if the condition (8.2.4) holds, then the method gives the convergence relation (8.2.5).*

In the above lemma, the condition (8.2.4) is also necessary in some sense for the global convergence, as will be briefly discussed in Section 8.5.

8.3 A sufficient and necessary condition on β_k

The purpose of this section is to provide a condition on β_k, which is sufficient for the global convergence of general conjugate gradient method with the Wolfe line search. To do so, we will first give some analyses for general conjugate gradient method with the strong Wolfe line search with the help of Lemma 8.2.3 (see §8.3.1). In §8.3.2, we will give a basic lemma for any method in the form of (8.1.2)–(8.1.3) and prove that the condition (8.3.5) can really ensure the global convergence of general conjugate gradient method with the Wolfe line search. An example is constructed in §8.3.3, which shows that the condition (8.3.5) is also necessary in some sense for the global convergence.

8.3.1 Proposition of the condition

In this subsection, we assume that the step-length λ_k satisfies the strong Wolfe conditions (8.1.4)–(8.1.5). By Lemma 8.2.3, we know that if the relation (8.2.4) holds, then the method gives the convergence relation (8.2.5). Otherwise, we have that

$$\sum_{k\geq 1} \frac{1}{\|d_k\|^2} < +\infty, \tag{8.3.1}$$

which gives

$$\lim_{k\to\infty} \|d_k\| = +\infty. \tag{8.3.2}$$

It follows from Assumption 8.2.1 that

$$\|g_k\| \leq \bar{\gamma}, \quad \text{for some } \bar{\gamma} > 0 \text{ and all } k \geq 1. \tag{8.3.3}$$

Then by (8.1.3), (8.3.2) and (8.3.3), we have that

$$\|d_k\| \approx |\beta_k| \, \|d_{k-1}\|. \tag{8.3.4}$$

Thus if the scalar β_k is such that

$$\sum_{k\geq 1} \prod_{j=2}^{k} \beta_j^{-2} = +\infty, \tag{8.3.5}$$

it is possible for us to establish (8.2.4) and then by Lemma 8.2.3 obtain a contradiction to (8.3.1). We formally describe this result as follows and give a strict proof, since the proof here is easy to understand and is quite different from the one for the Wolfe line search.

Theorem 8.3.1. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3) with d_k satisfying $g_k^T d_k < 0$ and with the strong Wolfe line searches (8.1.4) and (8.1.5). If β_k satisfies (8.3.5), we have that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. We write (8.1.3) as follows:

$$d_k + g_k = \beta_k d_{k-1}. \tag{8.3.6}$$

Squaring both sides of (8.3.6), we get that

$$\|d_k\|^2 = -2g_k^T d_k - \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2. \tag{8.3.7}$$

Noting that

$$-2g_k^T d_k - \|g_k\|^2 \leq \frac{(g_k^T d_k)^2}{\|g_k\|^2}, \tag{8.3.8}$$

it follows from this and (8.3.7) that

$$\|d_k\|^2 \leq \frac{(g_k^T d_k)^2}{\|g_k\|^2} + \beta_k^2 \|d_{k-1}\|^2 \tag{8.3.9}$$

Letting θ_k be the angle between $-g_k$ and d_k, namely,

$$\cos\theta_k = \frac{-g_k^T d_k}{\|g_k\| \, \|d_k\|}, \tag{8.3.10}$$

we get from (8.3.9) and (8.3.10) that

$$\begin{aligned}\|d_k\|^2 &\le (1-\cos^2\theta_k)^{-1}\beta_k^2\|d_{k-1}\|^2\\ &\le \cdots\cdots\\ &\le \prod_{j=j_0}^{k}(1-\cos^2\theta_j)^{-1}\,(\prod_{j=j_0}^{k}\beta_j^2)\|d_{j_0-1}\|^2,\end{aligned} \tag{8.3.11}$$

where $j_0 \ge 2$ is any integer. We now assume that (8.2.5) does not hold and hence there exists some constant $\gamma > 0$ such that

$$\|g_k\| \ge \gamma, \quad \text{for all } k \ge 1. \tag{8.3.12}$$

Then it follows from (8.2.2), (8.3.10) and (8.3.12) that

$$\sum_{k\ge 1}\cos^2\theta_k < +\infty. \tag{8.3.13}$$

The above relation clearly implies that

$$\prod_{j\ge j_0}(1-\cos^2\theta_j) \ge c, \quad \text{for some } c>0 \text{ and integer } j_0 \ge 2. \tag{8.3.14}$$

By (8.3.11), (8.3.14) and (8.3.5), we know that (8.2.4) holds. Thus by Lemma 8.2.3, (8.2.5) holds. This with (8.3.12) gives a contradiction, which ends the proof. □

8.3.2 Sufficiency of (8.3.5)

In the above subsection, we propose a condition on (8.3.5) for the global convergence of general conjugate gradient method with the strong Wolfe line search. Its proof is based on Lemma 8.2.3. In this subsection, we prove that Theorem 8.3.1 still holds if the strong Wolfe conditions are replaced with the Wolfe conditions.

At first, we present a basic lemma for general method (8.1.2)–(8.1.3) without line searches (see [6] for a similar lemma).

Lemma 8.3.2. *Consider any method of the form (8.1.2)–(8.1.3). Define ϕ_k and t_k as follows:*

$$\phi_k := \begin{cases}\|g_k\|^2, & \text{for } k=1;\\ \prod_{j=2}^{k}\beta_j^2, & \text{for } k\ge 2\end{cases} \tag{8.3.15}$$

and

$$t_k := \frac{\|d_k\|^2}{\phi_k^2}. \tag{8.3.16}$$

Then we have for all $k \ge 1$,

$$t_k = -2\sum_{i=1}^{k} \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^{k} \frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.17}$$

Proof. Since $d_1 = -g_1$ and $\phi_1 = \|g_1\|^2$, (8.3.17) holds for $k = 1$. For $k \geq 2$, dividing (8.3.7) by ϕ_k^2 and using the definitions of t_k and ϕ_k, we get that

$$t_k = t_{k-1} - \frac{2g_k^T d_k}{\phi_k^2} - \frac{\|g_k\|^2}{\phi_k^2}. \tag{8.3.18}$$

Summing this expression for $k = 2, \ldots, k$, we obtain

$$t_k = t_1 - 2\sum_{i=2}^{k} \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=2}^{k} \frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.19}$$

Since $d_1 = -g_1$ and $t_1 = \|g_1\|^2/\phi_1^2$, the above relation is equivalent to (8.3.17). So (8.3.17) holds for all $k \geq 1$. □

To show our main result, we still require the following lemma. See, for example, Pu and Yu [19] for its proof.

Lemma 8.3.3. *Suppose that $\{a_i\}$ and $\{b_i\}$ are positive number sequences, satisfying*

$$b_k \leq c_1 + c_2 \sum_{i=1}^{k} a_i, \quad \textit{for all } k, \tag{8.3.20}$$

where c_1 and c_2 are positive constants. If the sum $\sum_{k\geq 1} a_k$ is divergent, then $\sum_{k\geq 1} a_k/b_k$ is also divergent.

Now we prove that the condition (8.3.5) on β_k is sufficient for the global convergence of any method of the form (8.1.2)–(8.1.3).

Theorem 8.3.4. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3) with d_k satisfying $g_k^T d_k < 0$ and with the Wolfe line searches (8.1.4) and (8.1.10). If β_k satisfies (8.3.5), we have that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. Define ϕ_k as in (8.3.15). It follows from (8.3.5) that

$$\sum_{k\geq 1} \frac{1}{\phi_k^2} = +\infty. \tag{8.3.21}$$

Using (8.3.8) in (8.3.17), we can get

$$t_k \leq \sum_{i=1}^{k} \frac{(g_i^T d_i)^2}{\|g_i\|^2 \phi_i^2}. \tag{8.3.22}$$

Since $t_k \geq 0$, we also have by (8.3.17) that

$$-2\sum_{i=1}^{k}\frac{g_i^T d_i}{\phi_i^2} \geq \sum_{i=1}^{k}\frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.23}$$

Noting that

$$-4g_k^T d_k - \|g_k\|^2 \leq 4\frac{(g_k^T d_k)^2}{\|g_k\|^2} \tag{8.3.24}$$

for any k, we obtain from this and (8.3.23) that

$$4\sum_{i=1}^{k}\frac{(g_i^T d_i)^2}{\|g_i\|^2\phi_i^2} \geq -4\sum_{i=1}^{k}\frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^{k}\frac{\|g_i\|^2}{\phi_i^2} \geq \sum_{i=1}^{k}\frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.25}$$

Now we proceed by contradiction and assume that (8.3.12) holds. Then by (8.3.25), (8.3.21) and (8.3.12), we have that

$$\sum_{k\geq 1}\frac{(g_k^T d_k)^2}{\|g_k\|^2\phi_k^2} = +\infty. \tag{8.3.26}$$

Using relations (8.3.22), (8.3.26) and Lemma (8.3.3), we then obtain

$$\sum_{k\geq 1}\frac{(g_k^T d_k)^2}{\|g_k\|^2\phi_k^2}\frac{1}{t_k} = \sum_{k\geq 1}\frac{(g_k^T d_k)^2}{\|g_k\|^2\|d_k\|^2} = \sum_{k\geq 1}\cos^2\theta_k = +\infty, \tag{8.3.27}$$

which contradicts (8.3.13). The contradiction shows the truth of (8.2.5). □

Thus we have proved that the condition on β_k is sufficient for the global convergence of general conjugate gradient method with the Wolfe line search. Instead of the sufficient descent condition (8.2.3), only the descent condition $d_k^T g_k < 0$ is used in the Theorem. Since different nonlinear conjugate gradient methods vary with the scalar β_k and condition (8.3.5) only concerns β_k, we believe that Theorem 8.3.4 is very powerful in the convergence analyses of conjugate gradient methods, as will partly be shown in the coming section.

From the proof to Theorem 8.3.4, we can see that the relation (8.3.22) gives an upper bound for the quantity $t_k = \|d_k\|^2/\phi_k^2$, whereas (8.3.25) estimates the lower bound of a quantity related to $g_k^T d_k$. Both the relations (8.3.22) and (8.3.25) are derived from (8.3.17). Then by the two relations and Lemma 8.3.3, we are able to prove the sufficiency of (8.3.5) for the global convergence of general conjugate gradient method with the Wolfe line search.

For some conjugate gradient method, we know that if the sequence $\{\|d_k\|^2\}$ increases mostly linearly,

$$\|d_k\|^2 \leq c_1 + c_2 k, \tag{8.3.28}$$

where c_1 and c_2 are positive constants, and if the sufficient descent condition (8.2.3) holds for all k, then we can conclude the global convergence by the Zou-

tendijk condition (8.2.2) and the contradiction principle. Such approach is often used in the convergence analyses of many conjugate gradient methods, for example, the analyses of the FR method in Al-Baali [1] and the ones of the PRP method in Gilbert and Nocedal [10]. From the proof to Theorem 8.3.4, we see that the sufficient descent condition (8.2.3) is not necessary, but the roles of (8.3.22) and (8.3.25) are similar to those of (8.3.28) and (8.2.3).

For convenience in use, we give the following corollary of Theorem 8.3.4 at the end of this subsection.

Corollary 8.3.5. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3) with d_k satisfying $g_k^T d_k < 0$ and with the Wolfe line searches (8.1.4) and (8.1.10). If there exist nonnegative constants c_1 and c_2 such that*

$$\prod_{j=2}^{k} \beta_j^2 \leq c_1 + c_2 k, \tag{8.3.29}$$

we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. Since (8.3.29) implies that (8.3.5) holds, the statement follows Theorem 8.3.4. □

8.3.3 Necessity of (8.3.5)

In this subsection, we consider the necessity of the condition (8.3.5). To make our analyses more general, we assume that the line search is exact, namely,

$$g_{k+1}^T d_k = 0, \quad \text{for all } k. \tag{8.3.30}$$

We also assume that the iterations $\{x_k; k = 2, 3, \ldots\}$ fall into a region Ω where f is a quadratic function with the unit Hessian,

$$f(x) = \frac{1}{2} x^T x, \quad x \in \Omega \subset R^n. \tag{8.3.31}$$

Then by (8.1.3) and (8.3.30), we have that

$$d_k^T g_k = -\|g_k\|^2, \tag{8.3.32}$$

$$\|d_k\|^2 = \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2. \tag{8.3.33}$$

It follows from (8.1.2), (8.1.3) and (8.3.31) that

$$g_{k+1} = g_k + \alpha_k d_k, \tag{8.3.34}$$

which with (8.3.30) and (8.3.32) gives

$$\alpha_k = -\frac{g_k^T d_k}{\|d_k\|^2} = \frac{\|g_k\|^2}{\|d_k\|^2}. \tag{8.3.35}$$

By squaring both sides of (8.3.34) and using (8.3.32), (8.3.35) and (8.3.33), we get that

$$\|g_{k+1}\|^2 = \|g_k\|^2 + 2\alpha_k g_k^T d_k + \alpha_k^2 \|d_k\|^2 = \|g_k\|^2 \left[1 - \frac{\|g_k\|^2}{\|d_k\|^2}\right] = \beta_k^2 \frac{\|d_{k-1}\|^2}{\|d_k\|^2} \|g_k\|^2. \tag{8.3.36}$$

The recursion of the above relation yields

$$\|g_{k+1}\|^2 = \left(\prod_{j=2}^{k} \beta_j^2\right) \cdot \frac{\|d_1\|^2}{\|d_k\|^2} \cdot \|g_2\|^2. \tag{8.3.37}$$

Still define ϕ_k and t_k as in Lemma 8.3.2. Then we see that (8.3.37) is equivalent to

$$\|g_{k+1}\|^2 = t_k^{-1} \|d_1\|^2 \|g_2\|^2. \tag{8.3.38}$$

On the other hand, we have from (8.3.17), (8.3.32) and $d_1 = -g_1$ that

$$t_k = \sum_{i=1}^{k} \frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.39}$$

Note from the second equality in (8.3.36) that the sequence $\{\|g_k\|^2; k = 2, 3, \ldots\}$ is monotonically decreasing. Thus have that

$$\|g_k\| \le \|g_2\|, \quad \text{for all } k \ge 2. \tag{8.3.40}$$

Therefore if (8.3.5) is false, namely,

$$\sum_{k \ge 2} \prod_{j=2}^{k} \beta_j^{-2} < +\infty, \tag{8.3.41}$$

we have from the definition of ϕ_k, (8.3.40) and (8.3.39) that

$$t_k \le M, \quad \text{for some positive constant } M. \tag{8.3.42}$$

Relations (8.3.42) and (8.3.38) indicate that

$$\|g_k\| \ge M^{-1} \|d_1\|^2 \|g_2\|^2, \quad \text{for all } k, \tag{8.3.43}$$

which implies that the iterations can not approach the unique minimizer $x^* = 0$ of the function in (8.3.31). By contrast, if (8.3.5) is true, we have by the definition of ϕ_k, (8.3.39) and (8.3.38) that $t_k \to +\infty$ and $\lim_{k \to \infty} \|g_k\| = 0$. Therefore, in this

example, any method of the form (8.1.2)–(8.1.3) converges globally if and only if the condition (8.3.5) holds.

8.4 Applications of the condition (8.3.5)

In the above section, we have presented a sufficient condition, namely, (8.3.5), on β_k for the global convergence of general conjugate gradient method using the Wolfe line search. As a matter of fact, the previous analyses do not depend on the choice of β_k and hence apply to any method of the form (8.1.2)–(8.1.3). To make better use of the condition in the conjugate gradient field, we will introduce a new property, namely, Property (#), in §4.1. Such property may apply to all the three categories of conjugate gradient methods. Consequently, by Theorem 8.3.4 and Property (#), convergence results can conveniently be gained for some known conjugate gradient methods (see §4.2). An analysis is also given to a new conjugate gradient method, which further shows the usefulness of Theorem 8.3.4 and Property (#) (see §4.3).

8.4.1 Property (#)

In [10], Gilbert and Nocedal proposed the so-called Property (∗) for the second category of conjugate gradient methods and brought about the convergence results for the PRP, and HS methods with the restriction $\beta_k \geq 0$. The purpose of this subsection is to define a new property, that may apply to all the three categories of conjugate gradient methods.

Denoting $s_{k-1} = x_k - x_{k-1}$, we define Property (#) as follows:

Property (#). *Consider a method of the form (8.1.2)–(8.1.3), and suppose that*

$$0 < \gamma \leq \|g_k\| \leq \bar{\gamma}, \quad \text{for all } k. \tag{8.4.1}$$

Under this assumption we say that the method has Property (#) if there exist a positive and uniformly bounded sequence $\{\psi_k\}$, and constants $b \geq 1$ and $\lambda > 0$ such that for all k:

$$\text{(1) } |\beta_k| \leq b\,\frac{\psi_k}{\psi_{k-1}}; \tag{8.4.2}$$

$$\text{(2) if } \|s_{k-1}\| \leq \lambda, \text{ then } |\beta_k| \leq \frac{1}{b}\,\frac{\psi_k}{\psi_{k-1}}. \tag{8.4.3}$$

The above Property (#) clearly has Property (∗) in [10] as its special case. Under the same assumption (8.4.1), Property (∗) requires that there exist constants $b > 1$ and $\lambda > 0$ such that $|\beta_k| \leq b$ and if $\|s_{k-1}\| \leq \lambda$, then $|\beta_k| \leq \frac{1}{2b}$. So if Property (∗) holds, Property (#) must be true with $\psi_k \equiv 1$.

Similar to [10], we now present an analysis of the PRP method. Let

$$b = \frac{2\bar{\gamma}^2}{\gamma^2}, \quad \lambda = \frac{\gamma^2}{L\bar{\gamma}b}, \quad \psi_k \equiv 1, \tag{8.4.4}$$

where L is the Lipschitz constant in (8.2.1). Then by (8.1.7), (8.4.1) and (8.2.1), we can get that

$$|\beta_k^{PRP}| \le \frac{(\|g_k\| + \|g_{k-1}\|)\|g_k\|}{\|g_{k-1}\|} \le \frac{2\bar{\gamma}^2}{\gamma^2} = b\frac{\psi_k}{\psi_{k-1}}, \tag{8.4.5}$$

and if $\|s_{k-1}\| \le \lambda$,

$$|\beta_k^{PRP}| \le \frac{\|y_{k-1}\|\,\|g_k\|}{\|g_{k-1}\|} \le \frac{L\bar{\gamma}\|s_{k-1}\|}{\gamma^2} \le \frac{L\lambda\bar{\gamma}}{\gamma^2} = \frac{1}{b}\frac{\psi_k}{\psi_{k-1}}. \tag{8.4.6}$$

So Property (#) holds with the b, λ and ψ_k in (8.4.4). If we reduce the above λ by half, then Property $(*)$ in [10] also holds.

Since the ψ_k in Property (#) can be any bounded sequence, and since the factor $\frac{1}{2}$ of Property $(*)$ is missing in (8.4.3), Property (#) may apply to not only the second category of methods but the first and third categories. In fact, for the FR method, which belongs to the first category, we can choose

$$b = 1, \quad \psi_k = \|g_k\|^2, \quad \text{and } \lambda \text{ is any positive number.} \tag{8.4.7}$$

By the definition (8.1.6) of β_k^{FR}, (8.4.2)–(8.4.3) clearly holds. In addition, (8.4.1) implies that $\psi_k = \|g_k\|^2$ is uniformly bounded. Thus the FR method has Property (#). For the DY method with the Wolfe line search, we can get by multiplying (8.1.3) with g_k and using (8.1.9) that

$$g_k^T d_k = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}} g_{k-1}^T d_{k-1}. \tag{8.4.8}$$

The above relation and (8.1.9) give an equivalent formula of β_k^{DY} (see also [7]):

$$\beta_k^{DY} = \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \tag{8.4.9}$$

Further, we have from this, (8.4.8), (8.4.1) and (8.1.10) that

$$-g_k^T d_k \le (1-\sigma)^{-1}\|g_k\|^2 \le (1-\sigma)^{-1}\bar{\gamma}^2. \tag{8.4.10}$$

Thus for the DY method, Property (#) holds with

$$b = 1, \quad \psi_k = -g_k^T d_k, \quad \text{and any } \lambda > 0. \tag{8.4.11}$$

For the FR-PRP, and DY-HS hybrid methods, that belong to the third category, we have by (8.1.11), (8.1.12) and (8.1.13) that $|\beta_k| \le \beta_k^{FR}$ or $\beta_k \in [0, \beta_k^{DY}]$.

Hence these hybrid methods have also Property (#), as will be seen in the proof of Corollaries 8.4.2 and 8.4.3.

To sum up, Property (#) includes Property (∗) in [10] as its special case, and may apply to all the three categories of conjugate gradient methods.

8.4.2 Applications to some known conjugate gradient methods

In this subsection, we will discuss how to use Theorem 8.3.4 and Property (#) to analyze the global convergence of some known conjugate gradient methods.

At first, we have the following theorem for those methods for which Property (#) holds with $b = 1$.

Theorem 8.4.1. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3), where the scalar β_k has Property (#) with $b = 1$. If the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$, then we have that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. We proceed by contradiction, assuming that (8.3.12) holds. Then we know by (8.3.12) and (8.3.3) that (8.4.1) is true. By Property (#) with $b = 1$, we then have that

$$\prod_{j=2}^{k} \beta_j^2 = \frac{\psi_k^2}{\psi_1^2}, \tag{8.4.12}$$

which with the boundedness of ψ_k implies that (8.3.29) holds. Thus by Corollary 8.3.5, we have $\liminf_{k\to\infty} \|g_k\| = 0$, contradicting (8.3.12). The contradiction shows the truth of this theorem. □

By the above theorem, we can analyze the global convergence of some conjugate gradient methods in the first and third categories. For example, we have the following result for the FR method and its related hybrid methods.

Corollary 8.4.2. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with $|\beta_k| \le \beta_k^{FR}$. If the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$, then we have that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. Noting that the method is such that (8.4.2)–(8.4.3) holds with the parameters in (8.4.7), the statement follows Theorem 8.4.1. □

The above corollary clearly covers the FR method and the hybrid methods (8.1.11) and (8.1.12). If the strong Wolfe conditions (8.1.4)–(8.1.5) are used, and if $\sigma \le 0.5$, we can prove any method (8.1.2)–(8.1.3) with $|\beta_k| \le \beta_k^{FR}$ generates a descent direction at every iteration. Then by Corollary 8.4.2, we know that there is the global convergence, and hence obtain again those corresponding results in

[1, 5, 10] for the FR method. For the DY method and its related hybrid method, we also have the following corollary. See also [8] for the result.

Corollary 8.4.3. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with $\beta_k = r_k\beta_k^{DY}$. If the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10), and if*

$$r_k \in \left[\frac{\sigma-1}{1+\sigma}, 1\right], \tag{8.4.13}$$

then we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. First, we prove by induction that $g_k^T d_k < 0$ for all k. In fact, since $d_1 = -g_1$, it is obvious that $g_1^T d_1 < 0$. Suppose that $g_{k-1}^T d_{k-1} < 0$. Then we have by (8.1.10) that

$$l_k := \frac{g_k^T d_{k-1}}{g_{k-1}^T d_{k-1}} \le \sigma. \tag{8.4.14}$$

It follows from (8.1.3) and $\beta_k = r_k\beta_k^{DY}$ that

$$g_k^T d_k = -\|g_k\|^2 + r_k\beta_k^{DY} g_k^T d_{k-1}, \tag{8.4.15}$$

from (8.4.13) and (8.4.14), we obtain

$$g_k^T d_k = \frac{1+(r_k-1)l_k}{l_k-1}\|g_k\|^2 \in [-\frac{1}{1+\sigma}\|g_k\|^2, 0). \tag{8.4.16}$$

So $g_k^T d_k < 0$. By the induction principle, $g_k^T d_k < 0$ for all k.

Further, by $\beta_k = r_k\beta_k^{DY}$, (8.4.15) and the definition of l_k, we get that

$$\beta_k = \frac{r_k}{1+(r_k-1)l_k}\frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}, \tag{8.4.17}$$

from (8.4.13) and (8.4.14), we can obtain

$$|\beta_k| \le \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \tag{8.4.18}$$

This relation and (8.4.16) indicate that Property (#) holds with the parameters in (8.4.11). Therefore the result follows Theorem 8.4.1. □

To use Property (#) to analyze the second category of conjugate gradient methods, we now provide the following general lemma. In the lemma, we denote N^* to be the set of positive integers and

$$\mathcal{K}_{k,\Delta}^{\lambda} := \{i \in N^* : k \le i \le k+\Delta-1, \|s_{i-1}\| \ge \lambda\}, \tag{8.4.19}$$

and let $|\mathcal{K}_{k,\Delta}^{\lambda}|$ be the number of elements of the set $\mathcal{K}_{k,\Delta}^{\lambda}$.

Lemma 8.4.4. *Suppose that Assumption 8.2.1 holds. Consider any method (8.1.2)–(8.1.3) having Property (#), where the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$. If there exist $\Delta \in N^*$ and integer k_0 such that*

$$|\mathcal{K}^{\lambda}_{k,\Delta}| \leq \frac{\Delta}{2}, \quad \textit{for any } k \geq k_0, \tag{8.4.20}$$

we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. For any $i \geq 1$, denote

$$p_i := |\mathcal{K}^{\lambda}_{k_0,i\Delta}|. \tag{8.4.21}$$

It follows by (8.4.20) and the arbitrariness of $k \geq k_0$ in the relation that

$$p_i \leq \frac{i\Delta}{2}. \tag{8.4.22}$$

This means that in the range $[k_0, k_0 + i\Delta - 1)$ there are exactly p_i indices j such that $\|s_{j-1}\| > \lambda$, and thus there are $(i\Delta - p_i)$ indices with $\|s_{j-1}\| < \lambda$. Using this fact, (8.4.2), (8.4.3) and (8.4.22), we obtain for any $i \geq 1$

$$\begin{aligned}\prod_{j=k_0}^{k_0+i\Delta-1} \beta_j^2 &\leq b^{2p_i} \left(\frac{1}{b}\right)^{2(i\Delta-p_i)} \prod_{j=k_0}^{k_0+i\Delta-1} \frac{\psi_j^2}{\psi_{j-1}^2} \\ &\leq b^{2(2p_i-i\Delta)} \frac{\psi_{k_0+i\Delta-1}^2}{\psi_{k_0-1}^2} \leq \frac{\psi_{k_0+i\Delta-1}^2}{\psi_{k_0-1}^2}. \end{aligned} \tag{8.4.23}$$

Since $\{\psi_k\}$ is uniformly bounded, the above relation indicates that (8.3.5) holds. Thus the result follows Theorem 8.3.4. □

Now we prove a general result for any method with Property (#) and $\beta_k \geq 0$. The restriction that $\beta_k \geq 0$ was first suggested by Powell [18] for the PRP method and later used by Gilbert and Nocedal [10] in getting the convergence result for algorithms related to the PRP method.

Theorem 8.4.5. *Suppose that Assumption 8.2.1 holds. Consider any method (8.1.2)–(8.1.3) with Property (#) and $\beta_k \geq 0$. If the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$, then we have either*

$$\liminf_{k\to\infty} \|d_k\| < +\infty, \tag{8.4.24}$$

or the convergence relation (8.2.5) holds.

Proof. Suppose that (8.1.10) is false. Then we have that

$$\lim_{k\to\infty} \|d_k\| = +\infty. \tag{8.4.25}$$

Since $g_k^T d_k < 0$, we have that $d_k \neq 0$. Define $u_k := d_k/\|d_k\|$,

$$\rho_k := \frac{-g_k}{\|d_k\|} \quad \text{and} \quad \delta_k := \frac{\beta_k \|d_{k-1}\|}{\|d_k\|}. \tag{8.4.26}$$

From (8.1.3), we have for $k \geq 2$:

$$u_k = \rho_k + \delta_k u_{k-1}. \tag{8.4.27}$$

Note that $\|u_k\| = \|u_{k-1}\| = 1$ and by (8.3.3) and (8.4.25), $\lim_{k\to\infty} \|\rho_k\| = 0$. Hence, by (8.4.27),

$$\lim_{k\to\infty} |\delta_k| = 1, \tag{8.4.28}$$

which with (8.4.27) and the condition $\delta_k \geq 0$ implies that

$$\lim_{k\to\infty} \|u_k - u_{k-1}\| = \lim_{k\to\infty} \|\rho_k + (\delta_k - 1)u_{k-1}\| = 0. \tag{8.4.29}$$

In addition, using (8.1.2) and the definition of u_k, we can write for any indices l, k, with $l \geq k$:

$$x_l - x_{k-1} = \sum_{i=k}^{l} \|s_{i-1}\| u_{i-1} = \sum_{i=k}^{l} \|s_{i-1}\| u_{k-1} + \sum_{i=k}^{l} \|s_{i-1}\| (u_{i-1} - u_{k-1}). \tag{8.4.30}$$

So we have that

$$\begin{aligned} \sum_{i=k}^{l} \|s_{i-1}\| &\leq \|x_l - x_{k-1}\| + \sum_{i=k}^{l} \|s_{i-1}\| \|u_{i-1} - u_{k-1}\| \\ &\leq 2B + \sum_{i=k}^{l} \|s_{i-1}\| \, \|u_{i-1} - u_{k-1}\|, \end{aligned} \tag{8.4.31}$$

where B is a bound of the level set $\mathcal{L}$.

We now proceed by contradiction and assume that $\liminf_{k\to\infty} \|g_k\| \neq 0$. Then by Lemma 8.4.4, for any Δ and integer k_0, there exists an index $k \geq k_0$ such that

$$|\mathcal{K}_{k,\Delta}^{\lambda}| > \frac{\Delta}{2}. \tag{8.4.32}$$

Let $\Delta := \lceil 8B/\lambda \rceil$. For this Δ, by (8.4.29), we can choose k_0 such that

$$\|u_k - u_{k-1}\| \leq \frac{1}{2\Delta}, \quad \text{for all } k \geq k_0. \tag{8.4.33}$$

Then for any $i \in [k, k + \Delta - 1]$, we have by (8.4.29) and (8.4.33) that

$$\|u_{i-1} - u_{k-1}\| \leq \sum_{j=k}^{i-1} \|u_j - u_{j-1}\| \leq \Delta \cdot \left(\frac{1}{2\Delta}\right) = \frac{1}{2}. \tag{8.4.34}$$

Using (8.4.34) and (8.4.32) in (8.4.31), with $l = k + \Delta - 1$, we obtain

$$2B \geq \frac{1}{2} \sum_{i=k}^{k+\Delta-1} \|s_{i-1}\| > \frac{\lambda}{2} |\mathcal{K}^{\lambda}_{k,\Delta}| > \frac{\lambda\Delta}{4}. \tag{8.4.35}$$

Thus $\Delta < 8B/\lambda$, which contradicts the definition of Δ. Therefore (8.2.5) holds, which ends our proof. □

Note that if (8.4.24) holds, then (8.2.4) must be true. Thus by Lemma 8.2.3, we know that the convergence relation (8.2.5) holds if the Wolfe line search in Theorem 8.4.5 is replaced with the strong Wolfe line search. Consequently, we have the following corollary for the PRP method with $\beta_k \geq 0$. See also [3] for this result.

Corollary 8.4.6. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with* $\beta_k = \max\{\beta_k^{PRP}, 0\}$. *If the step-length* λ_k *satisfies the strong Wolfe conditions (8.1.4)–(8.1.5) and the descent condition* $g_k^T d_k < 0$, *we have that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. Suppose that this corollary is false and (8.3.12) holds. Then, noting that β_k is nonnegative and that Property (#) holds with the parameters in (8.4.4), we have by Theorem 8.4.5 that relation (8.4.24) holds. It follows that (8.2.4) is true. Therefore by Lemma 8.2.3, we obtain $\liminf_{k\to\infty} \|g_k\| = 0$, contradicting (8.3.12). The contradiction shows the truth of this corollary. □

Noting that relations (8.2.2), (8.2.3) and (8.3.12) indicate the truth of (8.4.25), we know that there is also the global convergence if the descent condition $g_k^T d_k < 0$ in Theorem 8.4.5 is replaced with the sufficient descent condition (8.2.3). Thus we can prove the following result for the HS method with $\beta_k \geq 0$ ([10]). The proof here is different from the one in [10].

Corollary 8.4.7. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with* $\beta_k = \max\{\beta_k^{HS}, 0\}$. *If the step-length* λ_k *satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the sufficient descent condition (8.2.3), then we have that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. We proceed by contradiction and assume that (8.3.12) holds. Define

$$b = \frac{2\bar{\gamma}}{c\gamma(1-\sigma)}, \quad \psi_k = -g_k^T d_k, \quad \lambda = \frac{c\gamma(1-\sigma)}{Lb}. \tag{8.4.36}$$

Then it follows by (8.1.8), (8.3.12), (8.3.3), (8.1.10), (8.2.3) and (8.2.1) that

$$|\beta_k^{HS}| \leq \frac{2\bar{\gamma}\|g_k\|}{(\sigma-1)g_{k-1}^T d_{k-1}} \leq \frac{bg_k^T d_k}{g_{k-1}^T d_{k-1}} = b\frac{\psi_k}{\psi_{k-1}}, \tag{8.4.37}$$

and if $\|s_{k-1}\| \leq \lambda$,

$$|\beta_k^{HS}| \leq \frac{L\lambda\|g_k\|}{(\sigma-1)g_{k-1}^T d_{k-1}} \leq \frac{g_k^T d_k}{bg_{k-1}^T d_{k-1}} = \frac{1}{b}\frac{\psi_k}{\psi_{k-1}}. \tag{8.4.38}$$

In addition, the line search condition (8.1.10) and $g_{k-1}^T d_{k-1} < 0$ imply that

$$|g_k^T d_{k-1}| \leq |d_{k-1}^T y_{k-1}|. \tag{8.4.39}$$

By (8.1.3), (8.1.8) and (8.3.3), we have that

$$-g_k^T d_k = \|g_k\|^2 - \frac{g_k^T d_{k-1}}{d_{k-1}^T y_{k-1}} g_k^T y_{k-1} \leq \|g_k\|^2 + |g_k^T y_{k-1}| \leq 3\bar{\gamma}^2. \tag{8.4.40}$$

The above relation implies that the ψ_k is a bounded sequence. So Property (#) holds. Since $\beta_k \geq 0$, we have by Theorem 8.4.5 and (8.3.12) that (8.4.24) is true. However, we have from (8.2.2), (8.2.3) and (8.3.12) that $\lim\limits_{k\to\infty} \|d_k\| = +\infty$, contradicting (8.4.24). The contradiction shows the truth of this corollary. □

8.4.3 Application to a new conjugate gradient method

To further show the usefulness of Property (#) in the convergence analyses of conjugate gradient methods, we will consider a new conjugate gradient method in this subsection.

For any method of the form (8.1.2)–(8.1.3), noting that

$$d_k^T g_k = -\|g_k\|^2 + \beta_k g_k^T d_{k-1}, \tag{8.4.41}$$

we know that d_k is a descent direction if the $(k-1)$-th line search is enough exact. Since exact line searches are expensive, and since the line search is only to minimize the objective function in the one-dimensional subspace $\{x_{k-1}+\alpha d_{k-1}\}$, it is preferable to do some inexact line search in practical computations. Suppose that the Wolfe line search is used and d_{k-1} is a descent direction. In this case, to ensure the descent property of d_k, we know from (8.4.41) that the choice of β_k should satisfy

$$\beta_k g_k^T d_{k-1} < \|g_k\|^2. \tag{8.4.42}$$

Assuming that

$$\beta_k = \frac{\|g_k\|^2}{g_k^T d_{k-1} + \tau_k}, \tag{8.4.43}$$

where τ_k satisfies

$$g_k^T d_{k-1} + \tau_k > 0, \tag{8.4.44}$$

the condition (8.4.42) is equivalent to

$$\tau_k > 0. \tag{8.4.45}$$

If we choose $\tau_k = -g_{k-1}^T d_{k-1}$, then it follows from the descent property of d_{k-1} and the second Wolfe condition (8.1.10) that the relations (8.4.44) and (8.4.45) hold. This method is just the DY method, for which we have proved its descent property and global convergence (see [7] or Corollary 8.4.3). Another possible choice is that

$$\tau_k = \|g_{k-1}\|^2, \tag{8.4.46}$$

which with (8.4.43) gives

$$\beta_k = \frac{\|g_k\|^2}{g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.47}$$

For such method, we can really prove that it can produce a descent direction at every iteration if the parameter σ in (8.1.10) is not greater than 0.25. However, due to the good numerical performances of the hybrid method (8.1.13), we are only interested in the following hybrid method of (8.4.47):

$$\beta_k = \frac{\max\{0, \min\{g_k^T y_{k-1}, \|g_k\|^2\}\}}{g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.48}$$

Under mild assumptions, we can prove that the method (8.4.48) produces a descent direction at every iteration and converges globally. The proof is mainly based on Theorem 4.5.

Theorem 8.4.8. *Suppose that Assumption 8.2.1 holds. Consider the methods (8.1.2), (8.1.3), (8.4.48), where the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10). If the parameter σ is such that*

$$\sigma \leq 0.25, \tag{8.4.49}$$

we have that for all $k \geq 1$,

$$-2\|g_k\|^2 \leq g_k^T d_k < 0. \tag{8.4.50}$$

Further, we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. Defining

$$\xi_k = \max\left\{0, \min\left\{\frac{g_k^T y_{k-1}}{\|g_k\|^2}, 1\right\}\right\}, \tag{8.4.51}$$

the formula (8.4.48) for β_k can be rewritten as

$$\beta_k = \frac{\xi_k \|g_k\|^2}{g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.52}$$

Multiplying (8.1.3) by $-g_k$ and using (8.4.52), we can get

$$-g_k^T d_k = \frac{(1-\xi_k) g_k^T d_{k-1} + \|g_{k-1}\|^2}{g_k^T d_{k-1} + \|g_{k-1}\|^2} \|g_k\|^2. \tag{8.4.53}$$

We now prove by induction that (8.4.50) holds for all $k \geq 1$. In fact, since $d_1 = -g_1$, (8.4.50) holds for $k = 1$. Suppose that (8.4.50) holds for some $k-1$. Then by (8.1.10), (8.4.49) and the induction hypothesis, we get

$$g_k^T d_{k-1} \geq \sigma g_{k-1}^T d_{k-1} \geq -\frac{1}{2}\|g_{k-1}\|^2. \tag{8.4.54}$$

Then it follows from (8.4.53), (8.4.54) and $\xi_k \in [0, 1]$ that

$$\frac{-g_k^T d_k}{\|g_k\|^2} = \frac{(1-\xi_k)\frac{g_k^T d_{k-1}}{\|g_{k-1}\|^2} + 1}{\frac{g_k^T d_{k-1}}{\|g_{k-1}\|^2} + 1} \geq \min\left\{ \frac{\bar{\xi}_k a_k + 1}{a_k + 1} : a_k \geq -\frac{1}{2}, \bar{\xi}_k \in [0,1] \right\} = 2. \tag{8.4.55}$$

The above relation implies that (8.4.50) holds for k. By induction, (8.4.50) holds. So each d_k is a descent search direction.

Now we show that the method has Property (#). In fact, using (8.4.53), we can also write β_k as

$$\beta_k = \frac{\xi_k(-g_k^T d_k)}{(1-\xi_k) g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.56}$$

By (8.4.50), (8.4.54) and the fact that $\xi_k \in [0, 1]$, we have

$$(1-\xi_k) g_k^T d_{k-1} + \|g_{k-1}\|^2 \geq \frac{1}{2}\|g_{k-1}\|^2 \geq \frac{1}{4}(-g_{k-1}^T d_{k-1}), \tag{8.4.57}$$

which with (8.4.56) implies that

$$|\beta_k| \leq 4\xi_k \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \tag{8.4.58}$$

Let $b = 4$, $\lambda = \gamma/(16L)$ and $\psi_k = -g_k^T d_k$. It follows by (8.4.58) and $\xi_k \in [0, 1]$ that (8.4.2) holds. If $\|s_{k-1}\| \leq \lambda$, we have from the definition (8.4.51) of ξ_k, (8.2.1) and (8.4.1) that

$$|\xi_k| \leq \frac{\|y_{k-1}\|}{\|g_k\|} \leq \frac{L\lambda}{\gamma} = \frac{1}{b}. \tag{8.4.59}$$

Thus (8.4.3) is also true. In addition, (8.4.50) and (8.3.3) imply that ψ_k is uniformly bounded. So Property (#) holds.

We now proceed by contradiction, assuming (8.3.12). In this case, by Theorem 8.4.5, we know that $\liminf_{k\to\infty} \|d_k\| < +\infty$ and hence there must exist constant $M > 0$ such that

$$\|d_{k_i}\| \leq M \tag{8.4.60}$$

holds for some infinite subsequence $\{k_i\} \subset N^*$. It follows by this and (8.3.3) that

$$g_{k_i+1}^T d_{k_i} \leq \|g_{k_i+1}\| \, \|d_{k_i}\| \leq \bar{\gamma} M. \tag{8.4.61}$$

Applying (8.4.61) and (8.3.12) in (8.4.53) (with k replaced by $k_i + 1$), we obtain

$$-g_{k_i+1}^T d_{k_i+1} \geq \frac{\gamma^4}{\bar{\gamma} M + \gamma^2}, \tag{8.4.62}$$

which with the Zoutendijk condition means that

$$\lim_{i\to\infty} \|d_{k_i+1}\| = +\infty. \tag{8.4.63}$$

On the other hand, we have by (8.4.56), $\xi_k \in [0, 1]$, (8.4.54), (8.3.12) and (8.3.3) that

$$|\beta_k| \leq \frac{2\|g_k\|^2}{\|g_{k-1}\|^2} \leq \frac{2\bar{\gamma}^2}{\gamma^2}. \tag{8.4.64}$$

By (8.1.3), (8.4.64) and (8.3.3), we can prove

$$\|d_{k_i+1}\| \leq \bar{\gamma} + \frac{2\bar{\gamma}^2}{\gamma^2} \|d_{k_i}\|. \tag{8.4.65}$$

Thus by (8.4.60) and (8.4.63), we obtain a contradiction by letting $i \to \infty$ in (8.4.65). The contradiction shows that $\liminf_{k\to\infty} \|g_k\| = 0$. □

According to our numerical experiences with the hybrid method (8.1.13) with the Wolfe line search [4, 8], the parameter σ in (8.1.10) can be chosen as 0.1. Thus, to some extent, we would be satisfied with the condition (8.4.49) in Theorem 8.4.8.

8.5 Discussion

In this chapter we have analyzed nonlinear conjugate gradient methods, where the step-length is computed by the Wolfe line search under the assumption that all the search directions are descent. A general condition on the scalar β_k, that is (8.3.5), was proposed which is sufficient for the global convergence. Since different conjugate gradient methods vary with the scalar β_k, we believe that the result is very powerful in the convergence analyses of conjugate gradient methods.

To use the result better, we have presented a new property, that is Property (#), for conjugate gradient methods. It was also shown that such property may apply to all the three categories of conjugate gradient methods, including the FR, PRP, HS and DY methods and the hybrid methods (8.1.11), (8.1.12) and (8.1.13). As a result, convergence analyses were provided for these methods under mild assumptions.

The result in Section 8.3 can also be used to analyze the bound in the hybrid conjugate gradient methods. Denote $r_k = \beta_k/\beta_k^{FR}$ and consider any method (8.1.2)–(8.1.3) related to the FR method. Assume that the line search satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$. If

$$\sum_{k\geq 2}\prod_{j=2}^{k} r_j^{-2} = +\infty, \tag{8.5.1}$$

we have by this, the definition of r_k, (8.1.6) and (8.3.3) that

$$\sum_{k\geq 2}\prod_{j=2}^{k} \beta_j^{-2} = \sum_{k\geq 2}\frac{\|g_1\|^4}{\|g_k\|^4}\prod_{j=2}^{k} r_j^{-2} \geq \frac{\|g_1\|^4}{\bar{\gamma}^4}\prod_{j=2}^{k} r_j^{-2} = +\infty. \tag{8.5.2}$$

Then we can conclude the global convergence by Theorem 8.3.4 and the contradiction principle. On the other hand, if

$$\sum_{k\geq 2}\prod_{j=2}^{k} r_j^{-2} < +\infty, \tag{8.5.3}$$

we can make use of the example in §8.3.3 that the method (8.1.2)–(8.1.3) with exact line searches need not converge. In fact, it follows from (8.3.37), the definition of r_k, (8.1.6), the monotonical decreasing of $\|g_k\|$ and $d_1 = -g_1$ that

$$\frac{\|g_k\|^2}{\|d_k\|^2} = \frac{\|g_1\|^2}{\|g_2\|^2}\frac{\|g_{k+1}\|^2}{\|g_k\|^2}\prod_{j=2}^{k} r_j^{-2} \leq \frac{\|g_1\|^2}{\|g_2\|^2}\prod_{j=2}^{k} r_j^{-2}. \tag{8.5.4}$$

Thus we know from the above relation and (8.5.3) that

$$\sum_{k\geq 1}\frac{\|g_k\|^2}{\|d_k\|^2} < +\infty, \tag{8.5.5}$$

which with the second equality of (8.3.36) implies that for all k,

$$\|g_{k+1}\|^2 = \|g_1\|^2\prod_{i=1}^{k}(1 - \frac{\|g_i\|^2}{\|d_i\|^2}) \geq c', \tag{8.5.6}$$

where c' is some positive constant. Therefore (8.5.1) is also necessary in some sense for the global convergence of general method (8.1.2)–(8.1.3) related to the FR method. A direct corollary to this result is that for any $c > 1$, any method (8.1.2)–(8.1.3) with the restriction $|r_k| \leq c$ need not converge. This result is stronger than Proposition 3.3 in [15], where Nocedal proved that there exists some constant $c > 1$ such that the method (8.1.2)–(8.1.3) with $|r_k| \leq c$ need not converge.

Here we also note that if

$$\sum_{k\geq 1} \frac{1}{\|d_k\|^2} < +\infty, \tag{8.5.7}$$

it follows by this and the boundness of $\|g_k\|^2$ that relation (8.5.5) holds. Then we also have (8.5.6). Since exact line searches are used in the example of §8.3.3, we know that the condition (8.2.4) is also necessary in some sense for the global convergence.

As an illustrative example, in this chapter we have also analyzed a hybrid conjugate gradient method, namely, the method (8.1.2)–(8.1.3) with β_k given by (8.4.48). With the help of Theorem 8.4.5, the descent property and global convergence of the method are proved under the Wolfe conditions with $\sigma \leq 0.25$. We wonder whether the method (8.4.48) is also efficient in practice or we can obtain a more efficient conjugate gradient algorithm by combining (8.4.48) and (8.1.13). This question still remains under studies.

Finally, we would like to mention that although most of the analyses in this chapter use the Wolfe line search, they are also efficient for the strong Wolfe line search. As is known, there is still lack of a similar theory for conjugate gradient methods using the strong Wolfe line search. We also expect that this chapter will arouse more attention to the use of the Wolfe line search in conjugate gradient methods, even from the aspect of numerical computation.

Acknowledgements The author thanks Professor Yaxiang Yuan very much for his useful discussion and suggestions. This research was partly supported by the Chinese NSF grants 19801033, 10571171 and 10831006 and the CAS grant kjcx-yw-s7-03.

References

1. M. Al-Baali, Descent property and global convergence of the Fletcher-Reeves method with inexact linesearch, *IMA J. Numer. Anal.*, **5**, 121–124, 1985.
2. I. Bongartz, A. R. Conn, N. Gould and Ph. L. Toint, CUTE: constrained and unconstrained testing environment, *ACM Transactions on Mathematical Software*, **21**, 123–160, 1995.
3. Y. H. Dai, J. Han, G. Liu, D. Sun, H. Yin and Y. Yuan, Convergence properties of nonlinear conjugate gradient methods, *SIAM Journal on Optimization*, **10**(2), 345–358, 1999.

4. Y. H. Dai and Q. Ni, Testing different conjugate gradient methods for large-scale unconstrained optimization, *Journal of Computational Mathematics*, **21**(3), 311–320, 2003.
5. Y. H. Dai and Y. Yuan, Convergence properties of the Fletcher-Reeves method, *IMA J. Numer. Anal.*, **16**, 155–164, 1996.
6. Y. H. Dai and Y. Yuan, A class of globally convergent conjugate gradient methods, *Sciences in China (Series A)*, **46**(2), 251–261, 2003.
7. Y. H. Dai and Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, *SIAM Journal on Optimization*, **10**(1), 177–182, 1999.
8. Y. H. Dai and Y. Yuan, An efficient hybrid conjugate gradient method for unconstrained optimization, *Annals of Operations Research*, **103**, 33–47, 2001.
9. R. Fletcher and C. M. Reeves, Function minimization by conjugate gradients, *Comput. J.*, **7**, 149–154, 1964.
10. J. C. Gilbert and J. Nocedal, Global convergence properties of conjugate gradient methods for optimization, *SIAM J. Optimization*, **2**, 21–42, 1992.
11. L. Grippo and S. Lucidi, A globally convergent version of the Polak-Ribière conjugate gradient method, *Math. Prog.*, **78**, 375–391, 1997.
12. M. R. Hestenes and E. Stiefel, Method of conjugate gradient for solving linear system, *J. Res. Nat. Bur. Stand.*, **49**, 409–436, 1952.
13. Y. F. Hu and C. Storey, Global convergence result for conjugate gradient methods, *J. Optim. Theory Appl.*, **71**, 399–405, 1991.
14. J. Nocedal, Theory of algorithms for unconstrained optimization, *Acta Numerica*, 199–242, 1991.
15. E. Polak and G. Ribière, Note sur la convergence de méthods de directions conjugées, *Revue Française d'Informatique et de Recherche Opiérationnelle*, **16**, 35–43, 1969.
16. B. T. Polyak, Conjugate gradient method in extremal problems, *USSR Comp. Math. and Math. Phys.*, **9**, 94–112, 1969.
17. M. J. D. Powell, Nonconvex minimization calculations and the conjugate gradient method, in: D. F. Griffiths, ed., *Numerical Analysis, Lecture Notes in Mathematics*, **1066**, 122–141, Springer-Verlag, Berlin, 1984.
18. M. J. D. Powell, Convergence properties of algorithms for nonlinear optimization, *SIAM Rev.*, **28**, 487–500, 1986.
19. D. Pu and W. Yu, On the convergence properties of the DFP algorithms, *Annals of Operations Research*, **24**, 175–184, 1990.
20. D. Touati-Ahmed and C. Storey, Efficient hybrid conjugate gradient techniques, *J. Optim. Theory Appl.*, **64**, 379–397, 1990.
21. P. Wolfe, Convergence conditions for ascent methods, *SIAM Rev.*, **11**, 226–235, 1969.
22. P. Wolfe, Convergence conditions for ascent methods II: some corrections, *SIAM Rev.*, **13**, 185–188, 1971.
23. G. Zoutendijk, Nonlinear programming, computational methods, in: J. Abadie, ed., *Integer and Nonlinear Programming*, 37–86, North-holland, Amsterdam, 1970.

Chapter 9
Full Space and Subspace Methods for Large Scale Image Restoration

Yanfei Wang, Shiqian Ma and Qinghua Ma

Abstract. In this chapter, we discuss about the full space and subspace methods for ill-posed image restoration problems. Image restoration refers to minimizing the degradation which is caused by sensing environment, say CCD camera misfocus, nonuniform motion, atmospheric aerosols and atmospheric turbulence. For image restoration problems, a key matter is to solve a quadratic programming problem. We study numerical solution methods in full space by limited memory of BFGS method and the subspace trust region method. We develop a novel approach for reducing the cost of sparse matrix-vector multiplication when applying the full space and subspace methods to atmospheric image restoration. Also the projection technique for the regularized convex quadratic functional is developed in the iteration for ensuring nonnegativity. Numerical experiments indicate that these methods are useful for large-scale image restoration problems.

9.1 Introduction

Image restoration is a major problem in digital image processing, which attracts more and more attention in recent years in different kinds of research fields[11, 10,

Yanfei Wang
Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China.
e-mail: yfwang@mail.iggcas.ac.cn

Shiqian Ma
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6902, USA.
e-mail: sm2756@columbia.edu

Qinghua Ma[1,2]
[1]School of Information, Renmin University of China, Beijing, 100872, China.
[2]College of Art and Science, Beijing Union University, Beijing, 100083, China.
e-mail: qinghua@ygi.edu.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

20, 17, 9, 2, 27, 28]. Image restoration refers to restoration of degradation which may be degraded by senor noise, misfocus of CCD camera, nonuniform motion, atmospheric aerosols, random atmospheric turbulence and so on. For example, in remote sensing applications, we are often required to recover the true signal or image f with resolution $N \times N$ by direct ground measurements or by receiving response from satellite sensors and giving the information about the modulation transfer function (MTF) or the point spread function (PSF) of the system. The perfect case for the PSF of the sensor should be the Gaussian. A key problem in remotely sensed image restoration is to restore the image by solving a blurring model and removing noise. This is because the signal (reflection) transferring from the land surface to the satellite sensors is inevitably interfered by atmosphere, say, atmospheric turbulence, aerosols, this process also leads to the blurring of the original signal. But remotely sensed image recovery is more complicated. Since it needs geometric correction, radiometric correction (removal of clouds and aerosols), we can do the work on restoration of degradation by PSF or MTF filtering.

Astronomical images obtained are usually corrupted or distorted by blurring and noise[19]. The blurring is characterized by a point spread function or impulse response, the noise is usually assumed to be additive, say, Gaussian random noise, Poisson noise, background noise and so forth. For astronomical images acquisition, the cost of a space telescope is high, so image restoration algorithms and/or adaptive optics have become an important research area in astronomical imaging. Considerable progress has been achieved over the past two decades in improving results of Earth-based observation. Telescopic surveillance of objects and scenes at ground level or anywhere in the atmosphere suffers from similar problems more severely; in this case, image restoration (or adaptive optics) is the only means of improvement, since the problems cannot be eliminated simply by leaving the atmosphere behind.

It indicates that the power distribution in the image plane due to a point source in the object plane can be expressed as follows:

$$h(x, y) = k(x, y) \star f(x, y) + n(x, y), \tag{9.1.1}$$

where, $h(x, y)$ denotes the recorded blurred image, $f(x, y)$ denotes the original object, their unique relation in the spatial domain is given by the two-dimensional point spread function (PSF) $k(x, y)$. $\star$ is the convolution operator, x and y are the spatial coordinates and $n(x, y)$ denotes an additive noise term. The above expression is commonly modelled as a first kind integral equation in the form

$$h(x, y) = \int\int_{\mathbb{R}^2} k(x - \xi, y - \eta) f(\xi, \eta) d\xi d\eta + n(x, y). \tag{9.1.2}$$

The image restoration problem is recovering f according to the knowledge of h and k.

In digital image restoration, a discrete model of (9.1.2) is a must. The discretization can be performed by discrete quadrature rule, say, midpoint quadra-

ture method, rectangular quadrature method and so on. We do not want to discuss about the discretization in detail and assume that after discretization, we have obtained the following linear system:

$$\mathbf{h} = \mathcal{K}\mathbf{f} + \mathbf{n}, \tag{9.1.3}$$

where $\mathcal{K} \in \mathbb{R}^{N^2 \times N^2}$, $\mathbf{f}, \mathbf{h}, \mathbf{n} \in \mathbb{R}^{N^2}$. The noise $\mathbf{n}$ cannot be ignored and the matrix $\mathcal{K}$ is usually badly conditioned, so we cannot solve this linear system by algebra strategy easily.

In practice, the vector $\mathbf{f}$ records the image pixel values, so the components of $\mathbf{f}$ must be nonnegative [2, 10, 17]. Thus we can express the image restoration problem as

$$\begin{aligned} &\min\ \Psi(\mathbf{f}) := \tfrac{1}{2}\|\mathcal{K}\mathbf{f} - \mathbf{h}\|^2, \\ &s.t. \qquad\quad \mathbf{f} \geq 0. \end{aligned} \tag{9.1.4}$$

The remaining task is to solve (9.1.4) efficiently and accurately. It is clear that (9.1.4) is equivalent to a constrained convex quadratic programming problem

$$\begin{aligned} &\min\ \Psi(\mathbf{f}) := \tfrac{1}{2}\mathbf{f}^T A\mathbf{f} - \mathbf{h}^T\mathcal{K}\mathbf{f}, \\ &s.t. \qquad\quad \mathbf{f} \geq 0, \end{aligned} \tag{9.1.5}$$

where $A := \mathcal{K}^T\mathcal{K}$. There exists a large amount of methods for solving the convex quadratic programming (see [16] and the references therein). Overall, there exists two groups of methods: the Newton types methods and the gradient types methods. Each kind of method has its own properties. We focus on the Newton type method, particularly the quasi-Newton method in this chapter.

We discuss about the full space and subspace methods for ill-posed image restoration problems. Particularly, we study numerical solution methods in full space by limited memory of BFGS method and the subspace trust region method.

9.2 Image restoration without regularization

The task of image restoration is to recover the original image $\mathbf{f}$ by the given observation vector $\mathbf{h}$ and the blurring operator $\mathcal{K}$. A direct approach, given $\mathcal{K}$, is to find an approximation $\mathbf{f}^{appr}$ to $\mathbf{f}$ which minimizes error in the fit to the observation data through minimization of the norm of the residual $\mathbf{h} - \mathcal{K}\mathbf{f}$. As is noted in Section 1, the approximation $\mathbf{f}^{appr}$ to the true image $\mathbf{f}$ which minimizes the energy of the noise/error $\mathbf{n}$ is formulated as

$$\mathbf{f}^{appr} = \arg\min_{\mathbf{f}} \|\mathbf{n}\| = \arg\min_{\mathbf{f}} \Psi(\mathbf{f}). \tag{9.2.1}$$

For noisy data, however, this does not result in a usable restoration due to the associated severe oscillation of the solution which is amplified by the noise in the

observation and the small singular values of the blurring operator. The reason of the occurrence of this phenomenon is that the inversion process of the image restoration problem is extremely ill-posed [21, 25]. Therefore in order to recover a low noise solution, we have to resort to other *a priori* information about the smoothness of the solution. There are different ways of incorporating *a priori* information, including statistical information and non statistical information. This results in a regularized problem, which receives much more attention in recent years[21, 29].

9.3 Image restoration with regularization

For the regularization for image restoration problem, refer to the following unconstrained minimization problem

$$\mathbf{f}^{Regu} = \arg\min_{\mathbf{f}} \Psi(\mathbf{f}) + \alpha\Omega[\mathbf{f}], \tag{9.3.1}$$

where $\Omega[\mathbf{f}]$ is the regularized term which provides the *a priori* information on $\mathbf{f}$, α is the regularization parameter which governs the tradeoff between the fit to the observation data and the smoothness of the restoration. There are different ways of choosing the regularized term $\Omega[\mathbf{f}]$. We consider the smooth regularizer in this chapter. For example, we can choose $\Omega[\mathbf{f}]$ in the l_2 norm as $\Omega[\mathbf{f}] = \|L\mathbf{f}\|_{l_2}^2$ or $\Omega[\mathbf{f}] = (L\mathbf{f}, \mathbf{f})$, with L being a (positive) semi-definite scale matrix; or choose $\Omega[\mathbf{f}]$ as a discretization of $\|\mathbf{f}\|_{W^{1,2}}^2$ in $W^{1,2}$ Sobolev space.

For smooth regularization, since the role of the regularizer is to suppress the ill-posed nature induced by the small singular values of the discrete PSF kernel, we consider the simple case, i.e., $\Omega[\mathbf{f}] = \frac{1}{2}(L\mathbf{f}, \mathbf{f})$ with $L \equiv I$ (here I is the identity).

In this case, the regularized solution $\mathbf{f}^{Regu}$ can be obtained by minimizing the regularized quadratic functional

$$J[\mathbf{f}] := \frac{1}{2}\mathbf{f}^T A\mathbf{f} + \frac{\alpha}{2}\mathbf{f}^T\mathbf{f} - \mathbf{h}^T\mathcal{K}\mathbf{f}. \tag{9.3.2}$$

The gradient and Hessian of $J[\mathbf{f}]$ are given by

$$g[\mathbf{f}] = (A + \alpha I)\mathbf{f} - \mathcal{K}^T h,$$
$$H[\mathbf{f}] = A + \alpha I.$$

respectively. Many kinds of optimization methods can be used to solve the minimization problem, say, iterative regularization method[14] and the recently developed trust region method and truncated conjugate gradient method (refer to [22, 23] and the references therein).

Considering the physical meaning of the pixel values of the images, i.e., nonnegativity, we solve the constrained regularized problem

$$\begin{array}{l} \min \ J[\mathbf{f}], \\ s.t. \ \mathbf{f} \geq 0. \end{array} \tag{9.3.3}$$

Now the problem remains solving (9.3.3) efficiently.

9.4 Optimization methods for solving the smoothing regularized functional

9.4.1 Minimization of the convex quadratic programming problem with projection

The minimization problem (9.3.3) is a special case of the convex quadratic programming problem[6]

$$\begin{array}{ll} \min q(x) := \frac{1}{2}x^T Ax - b^T x, \\ s.t. \qquad l \leq x \leq u, \end{array} \tag{9.4.1}$$

where $A \in \mathbb{R}^{m\times m}$ is a semi-definite symmetric matrix, and $b,\ l,\ u$ are vectors in $\mathbb{R}^m$. One can easily identify that if we set $x := \mathbf{f}$, $A := A + \alpha I = \mathcal{K}^T\mathcal{K} + \alpha I$, $b := \mathcal{K}^T\mathbf{h}$, $l := 0$ and $u := \infty$, then equation (9.4.1) reduces to (9.3.3). Therefore in the following, we will use equation (9.4.1) for algorithm description.

Let us define the feasible set of (9.4.1) as

$$\Omega = \{x \in \mathbb{R}^n : l \leq x \leq u\}.$$

Ideally, the projection P_Ω should be chosen such that

$$P_\Omega(x) = \arg\min_z \|x - z\|. \tag{9.4.2}$$

As for our problems with $l = 0$ and $u = \infty$, the set $\Omega = \{x : 0 \leq x < \infty\}$ is bounded below and convex, therefore, there exists an orthogonal projection operator P_Ω onto Ω such that

$$P_\Omega : \mathbb{R}^m \to \Omega$$

and

$$P_\Omega^* = P_\Omega,\ P_\Omega^2 = P_\Omega.$$

Our choice of the projection P_Ω is that

$$P_\Omega(x) = \text{argmin}_z\|x - z\|.$$

The projection can be easily calculated. It is related closely to the space-limiting operator or bandlimited operator [1], where the projection operator associated with a bounded domain Ω is defined by

$$(P_\Omega x)(t) = \chi_\Omega(t)x(t)$$

where $\chi_\Omega(t)$ is the characteristic function of the domain Ω. This operator projects onto the subspace of all functions which are zero outside the domain Ω. The i-th component of $P_\Omega(x)$ is

$$[P_\Omega(x)]_i = \max(x_i, 0) = \begin{cases} x_i, & \text{if } x_i \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Assume that the current iterate x_k is feasible, then the next point can be obtained by

$$x_{k+1} = P_\Omega(x_k + \alpha_k d_k), \tag{9.4.3}$$

where d_k is the search direction and α_k is the step length. It deserves noting that this gradient projection method has been considered by several authors in recent researches, e.g., [3, 5] and so forth.

9.4.2 Limited memory BFGS method with projection

There exists a large amount of methods for solving (9.4.1). Perhaps, the largest class of methods belongs to the Newton's and quasi-Newton's methods. Newton's method is an iterative process where the solution to the problem is updated as follows:

$$x_{k+1} = x_k - \alpha_k A^{-1} g_k, \tag{9.4.4}$$

where x_{k+1} is the updated solution at iteration $k+1$, g_k the gradient at the current iterate x_k, α_k the step length computed by a line search that ensures a sufficient decrease of $q(x)$ and A the Hessian. In many circumstances the inverse of the Hessian can not be computed directly. It happens for example when the matrix A is too big or A comes from a discretization of the ill-posed operator equations, which leads to the ill-conditioning of A.

A possible update of the Hessian is given by the BFGS technique [30, 8]. The BFGS update is given by

$$H_{k+1} = H_k - \frac{H_k y_k s_k^T + s_k y_k^T H_k}{y_k^T s_k} + \Big(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\Big)\frac{s_k s_k^T}{s_k^T y_k} \tag{9.4.5}$$

and A^{-1} is approximated by the iteration of H_k. In (9.4.5), the iterates are denoted by x_k, and we define $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$.

The limited memory strategy avoids two bottlenecks specified above in the Newton algorithm when applied to big systems, the storage and the inversion of big matrices. Here although a quasi-Newton algorithm is used, the inverse of the Hessian matrix is never built up, but directly the product of the inverse of the Hessian matrix by the gradient. Then, no Hessian diagonalization is required. What makes this method powerful is that in order to update this matrix product

only information of last m steps is used. In this way only the geometry and gradient of the last m steps have to be stored. When a BFGS update formula is used this procedure is called L-BFGS. This method was developed by [18, 15]

$$H_{k+1} = \Big(I - \frac{s_k y_k^T}{s_k^T y_k}\Big) H_k \Big(I - \frac{y_k s_k^T}{s_k^T y_k}\Big) + \frac{s_k s_k^T}{s_k^T y_k}. \tag{9.4.6}$$

The source code can be obtained free of charge from the web.

Now we give a precise description of the L-BFGS method. We follow the description in the form [30]

$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T, \tag{9.4.7}$$

where $\rho_k = \frac{1}{s_k^T y_k}$, and

$$V_k = I - \rho_k y_k s_k^T.$$

Usually the L-BFGS method is implemented with a line search for the step length α_k to ensure a sufficient decrease of the misfit function. Assume that x_{k+1}^* is an approximate solution for problem (9.4.1) at x_k. Convergence properties of the L-BFGS method are guaranteed if the steplength α_k in equation (9.4.3) satisfies the Wolfe line search conditions along $d_k = x_{k+1}^* - x_k$[13]

$$q(x_k + \alpha_k d_k) \le q(x_k) + \gamma_1 \alpha_k g_k^T d_k, \tag{9.4.8}$$

$$|g(x_k + \alpha_k d_k)^T d_k| \le \gamma_2 |g(x_k)^T d_k|, \tag{9.4.9}$$

where γ_1 and γ_2 are constants to be chosen *a priori*. The line search condition can ensure that the iterates remain in the feasible region.

We give the L-BFGS algorithm as follows:

Algorithm 9.4.1. *(Projected L-BFGS algorithm for image restoration)*

Step 1 *Choose* x_0, m, $0 < \gamma_1 < \frac{1}{2}$, $\gamma_1 < \gamma_2 < 1$, *and a symmetric positive definite starting matrix* H_0; *Set* $k := 0$.

Step 2 *If the stopping rule is satisfied, STOP; Otherwise, GOTO Step 3.*

Step 3 *Compute*

$$d_k = -H_k g_k, \tag{9.4.10}$$

$$x_{k+1} = P_\Omega(x_k + \alpha_k d_k), \tag{9.4.11}$$

where α_k *satisfies the above Wolfe conditions (9.4.8)–(9.4.9).*

Step 4 *Let* $\hat{m} = \min\{k, m-1\}$, *check if* $y_k^T s_k > 0$.
If NO: $H_{k+1} = I$ *(steepest descent step) and delete the pairs*

$$\{y_i, s_i\}_{i=k-\hat{m}}^{k};$$

If YES: Update H_0 $\hat{m}+1$ *times using the pairs* $\{y_i, s_i\}_{i=k-\hat{m}}^{k}$, *i.e., let*

$$
\begin{aligned}
H_{k+1} = {} & (V_k^T V_{k-1}^T \cdots V_{k-\hat{m}}^T) H_0 (V_{k-\hat{m}} \cdots V_{k-1} V_k) \\
& + \rho_{k-\hat{m}} (V_k^T V_{k-1}^T \cdots V_{k-\hat{m}+1}^T) s_{k-\hat{m}} s_{k-\hat{m}}^T (V_{k-\hat{m}+1} \cdots V_{k-1} V_k) \\
& + \rho_{k-\hat{m}+1} (V_k^T V_{k-1}^T \cdots V_{k-\hat{m}+2}^T) s_{k-\hat{m}+1} s_{k-\hat{m}+1}^T (V_{k-\hat{m}+2} \cdots V_{k-1} V_k) \\
& \quad \vdots \\
& + \rho_k s_k s_k^T .
\end{aligned}
$$

Step 5 Set $k := k + 1$ and GOTO Step 2.

Remark 9.4.2. *In Step 1 of Algorithm 9.4.1, the initial guess for the Hessian H_0 is the identity matrix I. In the algorithm proposed by Liu and Nocedal (1989), the initial symmetric positive definite H_0 need to be scaled as follows:*

$$H_0^{'} = \frac{y_0^T s_0}{\|y_0\|^2} H_0 = \frac{y_0^T s_0}{\|y_0\|^2} I.$$

Then after one iteration is completed, all methods update $H_0^{'}$ instead of H_0. This scaling greatly improves the performances of the method.

In Step 2, the choice of the stopping rule must be careful. As is known, the algorithm considered in this chapter can be regarded as a kind of iterative regularization method. Since the image restoration involves controlling the noise propagation, there must be a saturation of the iterations, i.e., it is improbable to solve the true solution such that the norm of the noise equals zero. Our stopping rule is chosen such that the algorithm iterates until the condition

$$\|g(x_k)\| > \zeta \cdot \max\{1, \|g(x_0)\|\}$$

is violated, *where ζ is the tolerance of break of the iteration cycle. This choice of the stopping rule is reasonable in practice. Since for application problems such as image restorations, we can not expect $\|g(x_k)\|$ approach zero since the right-hand side is always noisy, the usual stopping rule* if $\|g(x_k)\| \leq \epsilon$ (here ϵ is a sufficiently small number) then stop the iteration *can not be employed. Therefore, only a low accuracy in the solution is required. Empirically, at least in this chapter, we recommend choosing $\zeta = 1.0e-3$ for small noise level and $\zeta = 1.0e-2$ for large noise level.*

In Step 4, the condition $y_k^T s_k > 0$ guarantees the positive definiteness of the L-BFGS matrix. However, this is not always the case[7]. If it is violated, a simple reparation is conducting a steepest descent step.

Remark 9.4.3. *We note that in the L-BFGS method, the storing of the matrices H_k is unnecessary, instead, a prefixed number (say m) of vectors pairs $\{s_k, y_k\}$ that define define them implicitly are stored. Therefore, during the first m iterations the L-BFGS and the BFGS methods are identical, but when $k > m$ only information from the m previous iterations is used to obtain H_k. The number m of BFGS corrections that must be kept can be specified by the users. Moreover, in the L-BFGS the product $H_k g_k$ which represents the search direction is obtained*

by means of a recursive formula involving g_k and the most recent vectors pairs $\{s_k, y_k\}$.

9.4.3 Subspace trust region methods

Trust region method has been recently proved to be a kind of regularization [23, 24]. We apply it to solving ill-posed image restoration problems.

Trust region methods are usually formulated for non-quadratic nonlinear programming problem. Consider, for example, an unconstrained non-quadratic minimization problem $\min_{\mathbf{f}\in\mathbb{R}^n} \Gamma(\mathbf{f})$. The trust region method requires solving a trust region subproblem

$$\begin{aligned} &\min_s \Upsilon(s) := (g(\mathbf{f}), s) + \frac{1}{2}(H(\mathbf{f})s, s), \\ &\text{s.t. } \|s\| \le \Delta, \end{aligned}$$

where $g(\mathbf{f})$ and $H(\mathbf{f})$ denote the gradient and Hessian of $\Gamma(\mathbf{f})$, respectively. In each step, a trial step s is computed to decide whether it is acceptable or not. The decision rule is based on the ratio ρ between the actual reduction in the objective functional and the predicted reduction in the approximate model. And the trust region iterative step remains unchanged if $\rho \le 0$, where

$$\rho = \frac{Ared(\mathbf{f})}{Pred(\mathbf{f})},$$

and $Ared(\mathbf{f})$ and $Pred(\mathbf{f})$ are defined by $\Gamma(\mathbf{f}) - \Gamma(\mathbf{f}+s)$ and $\Upsilon(0) - \Upsilon(s)$, respectively.

For the model in (9.1.5), since it is in a quadratic form, the ratio ρ is always equal to 1. This means the trial step s, no matter it is good or not, will be always accepted. But the model is ill-posed, this seems to be unreasonable. To overcome this shortcoming, we propose the following modified trust region scheme when the model is a quadratic programming problem. We note that the approximate accuracy is characterized by the discrepancy between the observation and the true data; therefore variations of the norm of the discrepancy may reflect the degree of approximation. Based on these considerations, we propose to accept or reject the trial step s_k at the k-th step by the ratio

$$\rho_k = \frac{\Psi(\mathbf{f}_{k+1})}{\Psi(\mathbf{f}_k)} = \frac{\Psi(\mathbf{f}_k + \mathbf{s}_k)}{\Psi(\mathbf{f}_k)},$$

where $\Psi(\mathbf{f}_{k+1})$ and $\Psi(\mathbf{f}_k)$ are the reductions in norm of the discrepancy at $k+1$-th and k-th steps, respectively.

Because of the nonnegativity of the problem (9.1.4), the first order Karush-Kuhn-Tucker optimality condition for (9.1.4) can be expressed as

$$D(\mathbf{f})\nabla\Psi(\mathbf{f}) = 0, \tag{9.4.12}$$

where $\nabla\Psi(\mathbf{f}) = \mathcal{K}^T(\mathcal{K}\mathbf{f} - \mathbf{h_n})$ is the gradient of $\Psi(\mathbf{f})$ in (9.1.4) and $D(\mathbf{f})$ is a diagonal matrix whose diagonal elements are given by

$$(D(\mathbf{f}))_{ii} = \begin{cases} \mathbf{f}_i & \text{if} \quad (\nabla\Psi(\mathbf{f}))_i > 0, \\ 1 & \text{if} \quad (\nabla\Psi(\mathbf{f}))_i \leq 0. \end{cases} \tag{9.4.13}$$

Our subspace trust region model for (9.1.4) is

$$\begin{aligned} &\min\ \Phi(\mathbf{s}) := g_k^T\mathbf{s} + \tfrac{1}{2}\mathbf{s}^T\mathcal{K}^T\mathcal{K}\mathbf{s}, \\ &\text{s.t. } \|D_k^{-1}\mathbf{s}\| \leq \Delta_k, \\ &\quad \mathbf{s} \in S_k, \end{aligned} \tag{9.4.14}$$

where $g_k = \mathcal{K}^T(\mathcal{K}\mathbf{f}_k - \mathbf{h_n})$, D_k is defined as $D(\mathbf{f}_k)$ which is a scaling matrix to restrict the step s, Δ_k is the radius of the trust region and S_k is a subspace chosen so that (9.4.14) can be solved cheaply.

After getting a trial step $\mathbf{s}_k$ (see [26] for computational details), we decide to accept or reject $\mathbf{s}_k$. Our rule is whether the function value $\Psi(\mathbf{f}_k + \mathbf{s}_k)$ has some reductions compared to $\Psi(\mathbf{f}_k)$. We denote the ratio of $\Psi(\mathbf{f}_k + \mathbf{s}_k)$ and $\Psi(\mathbf{f}_k)$ as ρ_k^f:

$$\rho_k^f = \frac{\Psi(\mathbf{f}_k + \mathbf{s}_k)}{\Psi(\mathbf{f}_k)}. \tag{9.4.15}$$

In our algorithm, we will accept $\mathbf{s}_k$ if $\rho_k^f < \eta$ and reject it otherwise, where $\eta \in (0, 1)$. The reason we adopt this stopping rule is that (9.4.15) uses discrepancy between the observation and the model, which is more physically meaningful ([25]).

It is worth noting that $\bar{\mathbf{s}}_k^{tr}$ will tend to the scaled gradient direction when the radius of the trust region Δ_k approaches zero. So the update of Δ_k is also important when generating the good trial step. But Δ_k should not be expected to restrict the point stay interior. In fact, it is important to allow the trial step to pass the boundary of the trust region as long as it is still satisfying the bound constraints. In our algorithm, we use a similar approach proposed in [4] to update Δ_k.

Based on the above discussion, we recall a subspace trust-region algorithm recently developed in [26] as follows:

Algorithm 9.4.4. *(A subspace trust-region algorithm)*

Step 1 *Give* $\epsilon, \beta_c, \gamma_0 \in (0, 1)$, $1 = \mu_2 \geq \mu_1 \geq \eta > 0$, $\gamma_2 \geq \gamma_1 > 1$, $x_0 \in int(\mathcal{F})$, $\Delta_0 > 0$ *and set* $k := 0$.

Step 2 *If* $\|D_k g_k\| < \epsilon$, *stop.*

Step 3 *Compute* $\Psi(\mathbf{f}_k)$ *and* g_k*;*
Define the quadratic model Φ *as (9.4.14).*

Step 4 *Compute the trial step* $\mathbf{s}_k$ *from (9.4.14).*

Step 5 *Compute* ρ_k^f *by (9.4.15).*

If $\rho_k^f < \eta$ *then set* $\mathbf{f}_{k+1} := \mathbf{f}_k + \mathbf{s}_k$. *Otherwise* $\mathbf{f}_{k+1} := \mathbf{f}_k$.
Step 6 Set $k := k + 1$*; Update* Δ_k *as follows and then go to Step 2.*

- *If* $\rho_k^f \geq \mu_2$ *then* $\Delta_{k+1} = \gamma_0 \Delta_k$.
- *If* $\mu_1 \leq \rho_k^f < \mu_2$ *then* $\Delta_{k+1} \in [\Delta_k, \gamma_1 \Delta_k]$.
- *Otherwise,* $\Delta_{k+1} \in [\gamma_1 \Delta_k, \gamma_2 \Delta_k]$.

In Step 4, the computation for the trial step $\mathbf{s}_k$ from (9.4.14) can be performed in a two dimensional space. We refer to [26] for details. The convergence of the algorithm is proved in [26].

9.5 Matrix-Vector Multiplication (MVM)

We suppose the PSF kernel function in (9.1.2) is spatially invariant, i.e., the kernel is separable and can be reformulated as

$$k(x - \xi, y - \eta) = k_x(x - \xi) k_y(y - \eta). \tag{9.5.1}$$

This indicates that the blurring is identical in all parts of the image and is separated into pure horizontal and pure vertical components.

Numerically, assume that the discretization of k_x and k_y is $\mathcal{K}_x$ and $\mathcal{K}_y$ respectively, then the matrix $\mathcal{K}$ is a tensor of $\mathcal{K}_x$ and $\mathcal{K}_y$, i.e., the Kronecker product of $\mathcal{K}_x$ and $\mathcal{K}_y$,

$$\mathcal{K} = \mathcal{K}_x \otimes \mathcal{K}_y. \tag{9.5.2}$$

"vec$(\cdot)$" notation is a useful tool to simplify the expression of the matrix-vector multiplication. Given an array $U \in \mathbb{R}^{m_x \times m_y}$, one can obtain a vector $\mathbf{U} \in \mathbb{C}^{m_x m_y}$ by stacking the columns of U. This defines a linear mapping vec : $\mathbb{R}^{m_x \times m_y} \longrightarrow \mathbb{R}^{m_x m_y}$,

$$\text{vec}(U) = [U_{11}, \cdots, U_{m_x 1}, U_{12}, \cdots, U_{m_x 2}, \cdots, U_{1 m_y}, \cdots, U_{m_x m_y}]^T.$$

Therefore equation (9.1.4) can be rewritten as

$$\begin{array}{cc} \min \Psi(\mathbf{f}) := \frac{1}{2} \|(\mathcal{K}_x \otimes \mathcal{K}_y)\text{vec}(\mathbf{f}) - \text{vec}(\mathbf{h})\|^2 \\ s.t. \qquad\qquad \text{vec}(\mathbf{f}) \geq 0. \end{array} \tag{9.5.3}$$

9.5.1 MVM: FFT-based method

It is obvious that the main cost of computation in our projected L-BFGS algorithm for image restoration is the matrix-vector multiplication (MVM), so it is necessary to give an efficient algorithm to compute the matrix-vector multipli-

cation. Generally speaking, a classical matrix-vector multiplication accounts for the $2m^2$ flops if we regard the length of the signal f as m.

Note that both $\mathcal{K}_x$ and $\mathcal{K}_y$ are matrices of Toeplitz type. Therefore their Kronecker product $\mathcal{K}$ in (9.1.4) is a block Toeplitz with Toeplitz blocks (BTTB). By extending the BTTB into a block circulant with circulant blocks matrix (BCCB), we can use the two dimensional discrete Fourier transform to compute the matrix vector multiplication.

The BCCB matrix can be decomposed as

$$\mathcal{K} = \mathcal{F}^\star \Lambda \mathcal{F},$$

where $\mathcal{F}$ is the two dimensional discrete Fourier transform matrix, and Λ is a diagonal matrix containing the eigenvalues of $\mathcal{K}$. And the eigenvalues of $\mathcal{K}$ can be obtained by computing a two dimensional discrete Fourier transform (DFT) of the first column of $\mathcal{K}$, so we can compute $\mathcal{K}x$ by $\mathcal{F}^\star \Lambda \mathcal{F} x$.

DFT's can be computed at a low computational cost by utilizing the fast Fourier transform (FFT). The Fourier transform of an m-vector (signal) can be computed in $O(m \log_2 m)$ operations.

9.5.2 MVM with sparse matrix

In atmospheric image restoration, the PSF is usually modelled by Gaussian. Actually, the Gaussian function simulates well the convolution process of the true signal with the PSF operator. Both the blurring by aerosols and turbulence can be taken as Gaussian, which are in the form

$$k(x, y) = \frac{1}{2\pi\sigma^2} \exp\Big(-\frac{1}{2}\big(\frac{x^2 + y^2}{\sigma^2}\big)\Big), \tag{9.5.4}$$

where σ is a positive constant. The larger we choose σ, the more f gets smoothed. So by the same argument, the smaller we choose σ, the more the convolution result resembles f.

In our numerical experiments, we use the Gaussian point spread function as the integral kernel k in (9.1.2), so $\mathcal{K}$ can be represented by a kronecker product of two low order matrices as $\mathcal{K} = A \otimes B$ with $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times n}$. For the blurring process, A and B are usually taken as sparse banded matrices [12], which means only pixels within a distance band $- 1$ contribute to the blurring. So we can use a very economic algorithm proposed by [27, 28]. Let us take band $= 2$ and 3 as examples. Similar discussion can be made for other bands of matrices.

Suppose band $= p$, then A, B are $2p - 1$ diagonal matrices. Pixels within a distance $p-1$ of A and B contribute to the blurring. The different elements of $\mathcal{K}$ are only $C = A(1 : p, 1) \otimes B(1 : p, 1)$. The resulting matrix $\mathcal{K}$ is a sparse BTTB with each block being a $2p - 1$ diagonal matrix. If we define

$$A = \begin{pmatrix} a_0 & a_1 & \cdots & a_{p-1} & \cdots & 0 \\ a_1 & a_0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & a_{p-1} \\ a_{p-1} & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_0 & a_1 \\ 0 & 0 & a_{p-1} & \cdots & a_1 & a_0 \end{pmatrix}, B = \begin{pmatrix} b_0 & b_1 & \cdots & b_{p-1} & \cdots & 0 \\ b_1 & b_0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & b_{p-1} \\ b_{p-1} & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & b_0 & b_1 \\ 0 & 0 & b_{p-1} & \cdots & b_1 & b_0 \end{pmatrix},$$

then $C := (a_0b_0, a_0b_1, \cdots, a_0b_{p-1}, \cdots, a_{p-1}b_0, a_{p-1}b_1, \cdots, a_{p-1}b_{p-1})^T$. Thus, we can write the matrix-vector multiplication $y = (A \otimes B)x$ as follows:

$$y = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ a_{p-1}Bx_1 \\ a_{p-1}Bx_2 \\ \vdots \\ a_{p-1}Bx_{m-p+1} \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ a_1Bx_1 \\ \vdots \\ a_1Bx_{m-1} \end{pmatrix}$$

$$+ \begin{pmatrix} a_0Bx_1 \\ a_0Bx_2 \\ \vdots \\ a_0Bx_m \end{pmatrix} + \begin{pmatrix} a_1Bx_2 \\ \vdots \\ a_1Bx_m \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} a_{p-1}Bx_p \\ a_{p-1}Bx_{p+1} \\ \vdots \\ a_{p-1}Bx_{2p-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where $x = (x_1, x_2, \ldots, x_m)^T, x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$. Then each component of y, for example a_0Bx_1, can be evaluated as

$$a_0b_{p-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ x_{1p} \\ \vdots \\ x_{1,n-p+1} \end{pmatrix} + \cdots + a_0b_1 \begin{pmatrix} 0 \\ x_{11} \\ \vdots \\ x_{1,n-1} \end{pmatrix}$$

$$+ a_0b_0 \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{pmatrix} + a_0b_1 \begin{pmatrix} x_{12} \\ \vdots \\ x_{1n} \\ 0 \end{pmatrix} + \cdots + a_0b_{p-1} \begin{pmatrix} x_{1p} \\ \vdots \\ x_{1n} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

the same to others. Thus the matrix-vector multiplication reduces a lot.

Suppose band = 5, then A, B are 9 diagonal matrices. Pixels within a distance 4 of A and B contribute to the blurring. The different elements of $\mathcal{K}$ are only

$C = A(1:5,1) \otimes B(1:5,1)$. The resulting matrix $\mathcal{K}$ is a sparse BTTB with each block being a five-diagonal matrix. For the 9 diagonal matrices, the cost of the MVM computation would be $25mn$. Whereas for FFT-based matrix-vector computation, the cost is $O(5mn \log_2 mn)$, which is greater than $25mn$ for banded matrix-vector multiplication for large m and n, say, $m = n = 256,\ 512$ or more.

As an example, we suppose band = 2, then A, B are tridiagonal matrices. Pixels within a distance 1 of A and B contribute to the blurring. The different elements of $\mathcal{K}$ are only $C = A(1:2,1) \otimes B(1:2,1)$. The resulting matrix $\mathcal{K}$ is a sparse BTTB with each blocking a tridiagonal matrix. If we define

$$A = \begin{pmatrix} a_0 & a_1 & \cdots & 0 \\ a_1 & a_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & a_1 \\ 0 & 0 & a_1 & a_0 \end{pmatrix}, B = \begin{pmatrix} b_0 & b_1 & \cdots & 0 \\ b_1 & b_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & b_1 \\ 0 & 0 & b_1 & b_0 \end{pmatrix},$$

then $C := (C_1, C_2, C_3, C_4)^T = (a_0b_0, a_0b_1, a_1b_0, a_1b_1)^T$. Thus, we can write the matrix-vector multiplication as follows:

$$y = \begin{pmatrix} 0 \\ a_1Bx_1 \\ \vdots \\ a_1Bx_{m-1} \end{pmatrix} + \begin{pmatrix} a_0Bx_1 \\ a_0Bx_2 \\ \vdots \\ a_0Bx_m \end{pmatrix} + \begin{pmatrix} a_1Bx_2 \\ \vdots \\ a_1Bx_m \\ 0 \end{pmatrix},$$

where

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}, x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}.$$

Then

$$a_0Bx_1 = \begin{pmatrix} 0 \\ C_2x_{11} \\ \vdots \\ C_2x_{1,n-1} \end{pmatrix} + \begin{pmatrix} C_1x_{11} \\ C_1x_{12} \\ \vdots \\ C_1x_{1n} \end{pmatrix} + \begin{pmatrix} C_2x_{12} \\ \vdots \\ C_2x_{1n} \\ 0 \end{pmatrix},$$

so we can give the matrix-vector multiplication algorithm in a Matlab code as follows:

Algorithm 9.5.1. *(Banded BTTB Matrix-Vector Multiplication)*

```
function y=MatVecTri(m,n,C,x)
y=blockmulti(m,n,C(1:2),x);
x=blockmulti(m,n,C(3:4),x);
y(1:n*(m-1))=y(1:n*(m-1))+x(n+1:m*n);
y(n+1:m*n)=y(n+1:m*n)+x(1:n*(m-1));
```

```
function b=blockmulti(m,n,D,x)
b=zeros(m*n,1);
tmp=D(2)*x;
b(1:m*n-1)=tmp(2:m*n);
b(n:n:m*n)=0;
b=D(1)*x+b;
tmp(2:m*n)=tmp(1:m*n-1);
tmp(1:n:m*n)=0;
b=b+tmp;
```

In Algorithm 9.5.1, y=MatVecTri(m,n,C,x) is a main function, which calls the block multiplication function b=blockmulti(m,n,D,x) to finish the BTTB matrix-vector multiplication.

9.6 Numerical experiments

In this section, we give examples for the restoration of atmospheric image. The blurring process is modelled by a Gaussian point spread function:

$$k(x-\xi, y-\eta) = \frac{1}{2\pi\rho\bar{\rho}} \exp(-\frac{1}{2}(\frac{x-\xi}{\rho})^2 - \frac{1}{2}(\frac{y-\eta}{\bar{\rho}})^2). \tag{9.6.1}$$

In our test, we choose $\rho = \bar{\rho} = 0.7$. And the noise level is denoted by *level*, i.e.,

$$\mathbf{n} = \frac{level}{N}\|\mathbf{h}\| \times randn(N^2, 1),$$

where N is the size of the image, $randn(N^2, 1)$ is the Gaussian normal distributed random vector, and we set $randn('state', 0)$ in our Matlab codes to insure the same random vector is generated every time.

The image for testing is a cropland with size equalling 256×256 (Fig. 9.1). The resulting PSF matrix is a BTTB with size equalling $65,536 \times 65,536$. To simulate the blurring, we choose the band which is equal to 5. This induces a severe atmospheric or turbulence blurring. The condition number of the discrete Kronecker kernel $\mathcal{K}$ is equal to 1.2985×10^{30}. Therefore, the matrix $\mathcal{K}$ is very badly conditioned. On occasion that the weather is not too bad, one may choose a small band value. The noisy blurred images for different noise levels are plotted in Figs. 9.2 and 9.3. The deblurred restored images by fast subspace algorithms are illustrated in Figs. 9.4 and 9.5.

In computer computations, we set the maximum iteration number as $k_{max} = 200$. the choice of the maximum iteration number k_{max} is empirical. The reason we choose a maximum iteration number is that we do not consider that further iterations are necessary after reaching such a number k_{max}. However, such a regulation is never activated. Our algorithm converges before reaching the

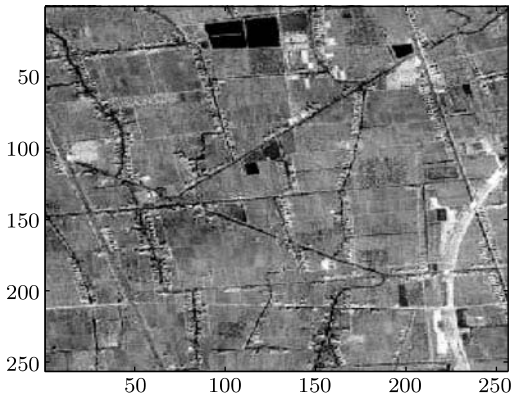


Fig. 9.1 The noise-free remotely sensed image.

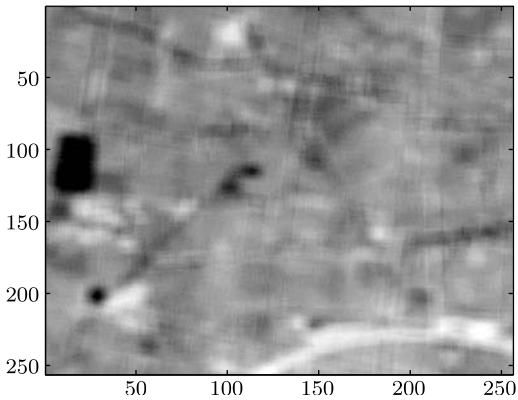


Fig. 9.2 The blurred images for noise level. level = 0.005.

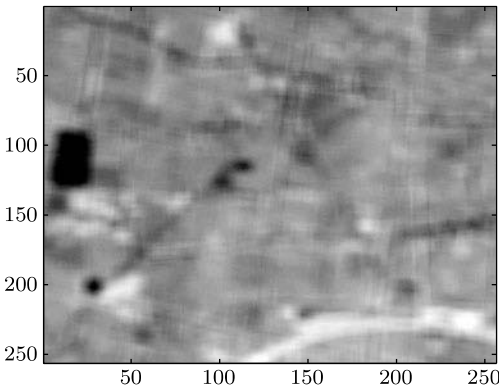


Fig. 9.3 The blurred images for noise level. level = 0.01.

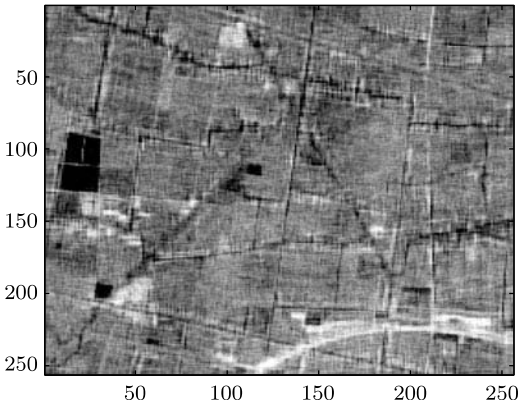


Fig. 9.4 Restored images for $\epsilon = 10^{-3}$ and level = 0.005.

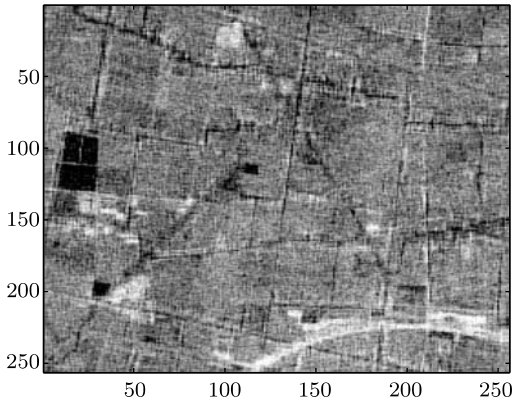


Fig. 9.5 Restored images for $\epsilon = 10^{-3}$ and level = 0.01.

maximum iteration number. The iteration results by the projected L-BFGS method and the subspace trust region method for different noise levels are listed in Table 9.1. Note that the choice of the value of ζ for projected L-BFGS method is related with the noise levels. In our experiments, we find that for small noise levels (less than or equal to 0.01), the ζ equalling 1.0e-3 is large enough, however for large noise levels (greater than 0.01), this value of ζ is not suitable, since more

Table 9.1 Comparison of efficiency of the projected L-BFGS method with the subspace trust region method.

	noise level	steps	CPU (seconds)
Projected L-BFGS	0.005	18	80.5860s
	0.01	23	98.7020s
Subspace method	0.005	2	7.0469 s
	0.01	2	7.5938 s

noise is involved in the iteration than the useful information due to enormous iterations. Therefore, we choose the ζ equalling 1.0e-2. In this case, reasonable relative errors are obtained.

9.7 Conclusions

In this chapter we address a limited memory of BFGS method with projection and a subspace trust region method for solving large scale ill-posed image restoration problems.

It deserves noting that there exists a large group of optimization methods, say, Newton types of methods, gradient types of methods and statistical optimization methods. Each method possesses its own advantages. Therefore, a complete comparison of these methods to the ill-posed image restoration deserves further investigating. Besides, our algorithm is about to restore images with smooth regularizer. Incorporating non-smooth regularizer in the algorithm for special application problems deserves special studying.

Acknowledgements This work was supported by the National Natural Science Foundation of China under grant numbers 10871191, 40974075 and National "973" Key Basic Research Developments Program of China under grant numbers 2005CB422104 and 2007CB714400.

References

1. S. Ando, On some mathematical problems in sensing and measurement, *SICE 2004 Annual Conference*, **2**, 1846–1852, 2004.
2. J. Bardsley and C. R. Vogel, A nonnegatively constrained convex programming method for image reconstruction, *SIAM Journal on Scientific Computing*, **25**, 1326–1343, 2003.
3. D. P. Bertsekas, Projected Newton methods for optimization problems with simple constraints, *SIAM J. Control and Optimization*, **20**, 221–246, 1982.
4. T. F. Coleman and Y. Li, An interior trust region approach for nonlinear minimization subject to bounds, *SIAM Journal on Optimization*, **6**(2), 418–445, 1996.
5. A. R. Conn, N. I. M. Gould and R. B. Schnable, Testing a class of methods for solving minimization problems with simple bounds on the variables, *Mathematics of Computation*, **50**, 399–430, 1988.
6. Y. H. Dai and R. Fletcher, *Projected Barzilai-Borwein Methods for Large-Scale Box-Constrained Quadratic Programming*, University of Dundee Report NA/215, 2003.
7. J. E. Dennis and R. B. Schnable, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Englewood Cliffs, N.J.: Prentice Hall, 1983.
8. R. Fletcher, *Practical Methods of Optimization*, 2nd edition, Chichester: John Wiley and Sons, 1987.
9. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd edition, Beijing: Publishing Hourse of Electronics Industry, 2002.
10. M. Hanke, J. Nagy and C. R. Vogel, Quasi-Newton approach to nonnegative image restorations, *Linear Algebra and Its Applications*, **316**, 223–236, 2000.

11. M. Hanke, J. Nagy, Restoration of atmospherically blurred images by symmetric indefinite conjugate gradient techniques, *Inverse Problems*, **12**, 157–173, 1996.
12. P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, Philadelphia: SIAM, 1998.
13. C. T. Kelley, *Iterative Methods for Optimization*, SIAM in Applied Mathematics, 1999.
14. K. P. Lee and J. G. Nagy, Steepest descent, CG, and iterative regularization of ill-posed problems, *BIT*, **43**, 1003–1017, 2003.
15. D. C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, **45**, 503–528, 1989.
16. J. Moré and G. Toraldo, On the solution of large quadratic programming problems with bound constraints, *SIAM J. Optim.*, **1**, 93–113, 1991.
17. J. Nagy and Z. Strakos, Enforcing nonnegativity in image reconstruction algorithms, in: David C. Wilson, et al. eds., *Mathmatical Modeling, Estimation, and Imaging*, **4121**, 182–190, 2000.
18. J. Nocedal, Updating quasi-Newton matrices with limited storage, *Mathematics of Computation*, **95**, 339–353, 1980.
19. M. Roggemann and B. Welsh, *Imaging Through Turbulence*, Florida: CRC Press, Boca Raton, 1996.
20. C. R. Vogel and M. E. Oman, A fast, robust algorithm for total variation based reconstruction of noisy, blurred images, *IEEE Trans. on Image Processing*, **7**, 813–824, 1998.
21. C. R. Vogel, *Computational Methods for Inverse Problems*, Philadelphia: SIAM, 2002.
22. Y. F. Wang, Y. X. Yuan and H. C. Zhang, A Trust Region-CG Algorithm for Deblurring Problem in Atmospheric Image Reconstruction, *Science in China A*, **45**, 731–740, 2002.
23. Y. F. Wang, On the regularity of trust region-CG algorithm: with application to deconvolution problem, *Science in China A*, **46**, 312–325, 2003.
24. Y. F. Wang and Y. X. Yuan, Convergence and regularity of trust region methods for nonlinear ill-posed inverse problems, *Inverse Problems*, **21**, 821–838, 2005.
25. Y. F. Wang, *Computational Methods for Inverse Problems and Their Applications*, Beijing: Higher Education press, 2007.
26. Y. F. Wang and S. Q. Ma, A fast subspace method for image deblurring, *Applied Mathematics and Computation*, 2009.
27. Z. W. Wen and Y. F. Wang, A trust region method for large scale inverse problems in atmospheric image restoration, in: Y. Yuan ed., *Numerical Linear Algebra and Optimization*, Beijing/New York: Science Press, 275–287, 2004.
28. Z. W. Wen and Y. F. Wang, A new trust region algorithm for image restoration, *Science in China A*, **48**, 169–184, 2005.
29. T. Y. Xiao, Y. S. Yu and Y. F. Wang, *Numerical Methods for Inverse Problems*, Beijing: Science Press, 2003.
30. Y. X. Yuan, *Numerical Methods for Nonliear Programming*, Shanghai: Shanghai Science and Technology Publication, 1993.

Part IV
Numerical Inversion in Geoscience and Quantitative Remote Sensing

Chapter 10
Some Reconstruction Methods for Inverse Scattering Problems

Jijun Liu and Haibing Wang

Abstract. Inverse scattering problems are one of the main research areas in the optimization techniques. The main purpose of inverse scattering problems is to detect the physical properties of an obstacle from some information related to the scattered waves of the obstacle for given incident wave. Generally, if the incident plane waves are given from the finite number of directions, which are indeed the practical situations, there is no uniqueness for reconstructing the obstacle properties such as the boundary shape. In these cases, the optimization techniques can be applied to reconstructing the obstacle approximately. That is, the obstacle shape is approximated by a minimizer of some cost functional which measures the defect between the measurement data of the scattered wave and the computational scattered wave related to the approximate obstacle. Of course, for this optimization problems in infinite dimensional space, some regularizing term should be introduced to the cost functional.

Although these general optimization techniques have been applied widely in the last century, they also suffer from many disadvantages theoretically and numerically. From the theoretical point of view, the lack of uniqueness makes the obtained approximate obstacle from the optimization procedure ambiguous, namely, we do not know whether or not the computational result indeed approximates the true one. From the numerical respects, such an optimization procedure needs to solve the direct scattering problem at each iteration step, which entails huge time cost. Moreover, to get the convergence of the iteration, a good initial guess is required. Even if in the case of convergence, the approximate sequence generally approaches to some local minimizer. These shortcomings of the direct optimiza-

Jijun Liu
Department of Mathematics, Southeast University, Nanjing 210096, China.
e-mail: jjliu@seu.edu.cn

Haibing Wang
School of Mathematics and Computational Science, Hunan University of Science and Technology, Xiangtan 411201, China.
e-mail: wanghb845@yahoo.com.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

tions used in inverse scattering problems cause some problems for the numerical reconstruction of the obstacles.

In recent years, some modified version of optimization algorithms based on the potential methods for inverse scattering problems are proposed, in the sense that the direct problems are not required to solve in the iteration procedures. On the other hand, some new schemes combining the advantages of optimizations with the exact reconstruction formulas have been developed for inverse scattering problems. In this chapter, we give an overview on these two directions. We begin with a brief introduction to the physical background to acoustic scattering problems as well as some well-known inverse scattering models. Then we review some classical and recently developed inversion methods for detecting the information about the unknown scatterer from a knowledge of the far-field pattern u^∞ for one or several incident plane waves. For each method, the basic idea is described, and some main results are presented. To test their validity, the numerical implementation of all inversion methods needs to be studied. So we finally focus on the numerical realizations of these existing methods, pointing out the main difficulties encountered in numerical realization. In addition, the advantages and disadvantages of these methods are also analyzed.

10.1 Introduction

For the purposes of exposition, we focus our attention to the case of acoustic waves. Particularly, we consider an acoustic incident wave propagating in a homogeneous isotropic medium. When this incident wave meets an obstacle D, it will be scattered. In this case, we express the total field outside D as the summation of the original incident wave and the scattered wave. Here the behavior of the scattered wave depends on both the incident wave and the nature of the obstacle.

To be more precise, consider the scattering of a time-harmonic acoustic wave by a bounded object $D \subset \mathbb{R}^3$, embedded in homogeneous isotropic medium with density ρ and sound speed c. Denote by $U(x,t)$ for $x \in \mathbb{R}^3 \setminus \overline{D}$ the velocity potential, then the wave motion is governed by the wave equation

$$\frac{\partial^2 U}{\partial t^2} = c^2 \Delta U, \tag{10.1.1}$$

where Δ is the Laplacian operator in $\mathbb{R}^3$. For time-harmonic acoustic waves with frequency ω, the time dependence is factored into the form $U(x,t) = \Re\left\{u(x)e^{-i\omega t}\right\}$, and thus the complex valued amplitude $u(x)$ satisfies the Helmholtz equation

$$\Delta u + k^2 u = 0 \tag{10.1.2}$$

in $\mathbb{R}^3 \setminus \overline{D}$, where the wave number $k = \omega/c$. To describe the scattering procedure, we must distinguish two cases, namely, D is an impenetrable object and a penetrable one.

Consider the scattering of a given incident wave u^i by an impenetrable obstacle D. Then the total wave $u = u^i + u^s$ with u^s the scattered wave must satisfy the Helmholtz equation outside D. Furthermore, if the obstacle is *sound-soft*, then the total wave vanishes on the boundary, which implies that the total wave meets the Dirichlet boundary condition

$$u = 0 \quad \text{on } \partial D. \tag{10.1.3}$$

Similarly, the scattering from a *sound-hard* obstacle leads to a Neumann boundary condition

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial D, \tag{10.1.4}$$

where ν is the unit outward normal to ∂D, since in this case the normal velocity of the acoustic wave vanishes on the boundary. More generally, if the normal velocity on the boundary is proportional to the excess pressure on the boundary, then we get an impedance boundary condition of the form

$$\frac{\partial u}{\partial \nu} + i\lambda u = 0 \quad \text{on } \partial D \tag{10.1.5}$$

with boundary impedance coefficient $\lambda(x)$ satisfying $\Re\lambda \geq 0$.

In the case of D being a penetrable inhomogeneous obstacle with slowly varying density $\rho_D = \rho_D(x)$ and sound speed $c_D = c_D(x)$ different from ρ and c in the surrounding medium $\mathbb{R}^3 \setminus \overline{D}$, the scattering by D leads to a transmission problem. Here, in addition to the scattered field u^s in $\mathbb{R}^3 \setminus \overline{D}$, we also have a transmitted wave v in D satisfying

$$\Delta v + k^2 n(x) v = 0 \tag{10.1.6}$$

with the wave number $k = \omega/c$, where $n(x) = c^2/c_D^2(x)$ is the index of refraction. From the continuity of the pressure and normal velocity across the obstacle boundary, the following transmission conditions on the boundary hold

$$u = v, \quad \frac{1}{\rho}\frac{\partial u}{\partial \nu} = \frac{1}{\rho_D}\frac{\partial v}{\partial \nu} \quad \text{on } \partial D.$$

For the scattered wave u^s, the Sommerfeld radiation condition

$$\lim_{r\to\infty} r\left(\frac{\partial u^s}{\partial r} - iku^s\right) = 0, \quad r = |x| \tag{10.1.7}$$

is required. Physically, (10.1.7) characterizes the outgoing waves, while from the mathematical point of view this condition ensures the uniqueness for the solutions to the scattering problems.

Using the Green's formula and the radiation condition, it is easy to show that

$$u^s(x) = \int_{\partial D} \left\{ u^s(y) \frac{\partial \Phi(x,y)}{\partial \nu(y)} - \frac{\partial u^s(y)}{\partial \nu(y)} \Phi(x,y) \right\} ds(y), \quad x \in \mathbb{R}^3 \setminus \overline{D}, \quad (10.1.8)$$

where $\Phi(x,y)$ is the radiating fundamental solution to the Helmholtz equation defined by

$$\Phi(x,y) := \frac{1}{4\pi} \frac{e^{ik|x-y|}}{|x-y|}, \quad x \neq y. \quad (10.1.9)$$

From (10.1.8) and the asymptotic behavior of $\Phi(x,y)$, we obtain that

$$u^s(x) = \frac{e^{ik|x|}}{|x|} \left\{ u^\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad \hat{x} = \frac{x}{|x|} \quad (10.1.10)$$

as $|x| \to \infty$, where u^∞ is usually known as the *far-field pattern* or scattering amplitude of the scattered field u^s, with the representation

$$u^\infty(\hat{x}) = \frac{1}{4\pi} \int_{\partial D} \left\{ u^s(y) \frac{\partial e^{-ik\hat{x}\cdot y}}{\partial \nu(y)} - \frac{\partial u^s(y)}{\partial \nu(y)} e^{-ik\hat{x}\cdot y} \right\} ds(y), \quad \hat{x} \in \mathbb{S}(\text{unit sphere}). \quad (10.1.11)$$

Obviously, u^∞ is an analytic function of $\hat{x}$ on the unit sphere $\mathbb{S}$. As one of the most important tools in the scattering theory, the Rellich's lemma establishes the one-to-one correspondence between a radiating scattered wave u^s and its far-field pattern u^∞, which is stated in [17].

Lemma 10.1.1 (Rellich)**.** *Assume that the bounded set D is the open complement of an unbounded domain and let $u \in C^2(\mathbb{R}^3 \setminus \overline{D})$ be a solution to the Helmholtz equation satisfying*

$$\lim_{r\to\infty} \int_{|x|=r} |u(x)|^2 ds = 0. \quad (10.1.12)$$

Then $u \equiv 0$ in $\mathbb{R}^3 \setminus \overline{D}$.

The inverse scattering problems aim to detect the property of an obstacle from the far-field pattern for the incident plane waves. However, the relations between scattered wave for incident plane waves and the one of point-source are of great importance in the recently developed inversion method. In the sequel, we always denote the scattered field for an incident plane wave $u^i(x,d) := e^{ikx\cdot d}$ with unit incident direction $d \in \mathbb{S}$ by $u^s(x,d)$, $x \in \mathbb{R}^3 \setminus \overline{D}$, and the corresponding *far-field pattern* by $u^\infty(\hat{x},d)$, $\hat{x} \in \mathbb{S}$. If the incident field is a point-source $\Phi(\cdot,z)$ with source location $z \in \mathbb{R}^3 \setminus \overline{D}$, then the scattered field is denoted by $\Phi^s(\cdot,z)$ and the corresponding far-field pattern by $\Phi^\infty(\hat{x},z)$, $\hat{x} \in \mathbb{S}$.

We begin with the reciprocity principles for the scattering procedure, which play an important role in some inversion methods, such as point source method and singular sources method.

Theorem 10.1.2 (Far-field reciprocity relation)**.** *The far-field patterns for scattering of plane waves by a sound-soft scatterer satisfy the reciprocity relation*

$$u^\infty(\hat{x}, d) = u^\infty(-d, -\hat{x}), \quad \hat{x}, d \in \mathbb{S}. \tag{10.1.13}$$

For the proof, we refer to [17]. In the derivation of (10.1.13), only the Helmholtz equation for the incident field in $\mathbb{R}^3$ and for the scattered field in $\mathbb{R}^3 \setminus \overline{D}$ and the radiation condition are used, so the reciprocity relation (10.1.13) is valid for the obstacle with three kinds of boundary conditions.

Theorem 10.1.3 (Mixed reciprocity relation [63]). *For acoustic scattering of plane waves $u^i(\cdot, d),\ d \in \mathbb{S}$ and point source $\Phi(\cdot, z),\ z \in \mathbb{R}^3 \setminus \overline{D}$ from an obstacle with sound-soft, sound-hard or impedance boundary condition, we have*

$$\Phi^\infty(\hat{x}, z) = \frac{1}{4\pi} u^s(z, -\hat{x}), \quad z \in \mathbb{R}^3 \setminus \overline{D},\ \hat{x} \in \mathbb{S}. \tag{10.1.14}$$

For the direct scattering problem, the scatterer and the incident wave u^i are assumed to be given, so the aim is to compute the scattered field or its far-field pattern, respectively. However, the inverse scattering problem that we will mainly concern in this chapter is to detect the information about D from a knowledge of $u^\infty(\hat{x}, d)$ for $\hat{x}, d \in \mathbb{S}$ and fixed wave number k. Some classical inverse scattering models can be stated as follows.

- *Model problem I:* Let $u = u^i + u^s$ be a solution to (10.1.2) in $\mathbb{R}^3 \setminus \overline{D}$ satisfying one of the boundary conditions: (10.1.3), (10.1.4) or (10.1.5). From the knowledge of the far-field pattern u^∞, one tries to determine the shape and location of obstacle D.
  Such problems have been researched thoroughly, which will be discussed in next sections.
- *Model problem II:* Let $u = u^i + u^s$ be a solution to (10.1.2) in $\mathbb{R}^3 \setminus \overline{D}$ satisfying the impedance boundary condition (10.1.5) with $\Re\lambda \geq 0$. From the knowledge of the far-field pattern u^∞, determine boundary impedance $\lambda = \lambda(x)$ and ∂D. These problems are related to the antenna design for specific purposes and the hostile decoy in military applications. For example, the hostile decoy in practice can become a perfect conductor coated by a thin dielectric layer, in this case, the shape of the decoy is known in advance. Therefore the surface impedance serves as a target signature [1, 7, 16, 75].
- *Model problem III:* Let $u = u^i + u^s$ be a solution to (10.1.2) in $\mathbb{R}^3 \setminus \overline{D}$. The boundary ∂D has different properties in its different components, that is,

$$u = 0 \quad \text{on } \partial D_D, \tag{10.1.15}$$

and

$$\frac{\partial u}{\partial \nu} + i\lambda u = 0 \quad \text{on } \partial D_I \tag{10.1.16}$$

with some impedance λ, where ∂D_D and ∂D_I are open surfaces on ∂D satisfying

$$\partial D = \overline{\partial D_D} \cup \overline{\partial D_I}, \quad \partial D_D \cap \partial D_I = \emptyset.$$

This mixed boundary value problem typically models the scattering procedure by an obstacle coated by a thin layer of material on part ∂D_I of the boundary. In general, it is not known *a priori* whether or not the scattering obstacle is coated, and if so, what is the extent of the coating? So the corresponding inverse scattering problem is to determine the shape of obstacle D, to identify ∂D_D and ∂D_I, and to reconstruct the surface impedance $\lambda(x)$ on ∂D_I from a knowledge of the far-field data $\{u^\infty(\hat{x}, d) : \hat{x}, d \in \mathbb{S}\}$ [5, 6, 23, 52, 53, 54, 58].

- *Model problem IV:* Let $u = u^i + u^s$ be a solution to the Helmholtz equation

$$\Delta u(x) + k^2 n(x) u(x) = 0 \tag{10.1.17}$$

for $x \in \mathbb{R}^3$, where

$$n(x) := \frac{c_0^2}{c^2(x)} \tag{10.1.18}$$

is the index of refraction, u^s satisfies the Sommerfeld radiating condition (10.1.7). On the boundary ∂D, we assume that both u and its normal derivative are continuous. The inverse problem is to determine the refractive index from a knowledge of u^∞.

This model comes from the media scattering problem. The inhomogeneity of the media is represented by the index of refraction $n(x)$, which is assumed to be the one in $\mathbb{R}^3 \setminus \overline{D}$ and different from the one in D. If the medium is absorbing, then the refractive index has an imaginary component, that is, $n(x)$ is of the form

$$n(x) = n_1(x) + i\frac{n_2(x)}{k}. \tag{10.1.19}$$

For more introductions to the inverse scattering models, we refer to [16, 17, 60].

10.2 Iterative methods and decomposition methods

This section is devoted to the classical iterative methods and decomposition methods, as well as the hybrid methods combining the ideas of Newton iterations and decomposition methods.

10.2.1 Iterative methods

The direct scattering problem for given incident plane wave u^i defines an operator

$$A : \partial D \mapsto u^\infty,$$

which maps the boundary ∂D of the scatterer onto the far-field pattern u^∞ of the scattered wave. In terms of this operator, the inverse scattering problem can

be abstracted as solving a nonlinear ill-posed operator equation

$$A(\partial D) = u^\infty \tag{10.2.1}$$

for the unknown surface ∂D. Hence it is natural to use some iterative methods such as regularized Newton method, Landweber iteration or conjugate gradient method. For this purpose, we need to further investigate the operator A. Choose a fixed reference domain D of class $\mathcal{C}^2$ and consider a set of scatterers D_h with the boundaries of the form

$$\partial D_h = \{x + h(x) : x \in \partial D\},$$

where $h : \partial D \to \mathbb{R}^3$ is of class $\mathcal{C}^2$ and sufficiently small in the sense of C^2-norm on ∂D. Thus, we can consider the operator A as a mapping from $V := \{h \in C^2(\partial D) : \|h\|_{C^2} < a\}$ with a sufficiently small radius $a > 0$ into $L^2(\mathbb{S})$. For simplicity, we devote our attention to the case of a starlike domain D_r with

$$\partial D_r = \{r(\hat{x})\hat{x} : \hat{x} \in \mathbb{S}\}, \tag{10.2.2}$$

where $r : \mathbb{S} \to \mathbb{R}^+$ represents the radial distance from the origin. Then A can be interpreted as a mapping

$$A : r \in \{r \in C^2(\mathbb{S}) : r > 0\} \mapsto u_\infty \in L^2(\mathbb{S}). \tag{10.2.3}$$

Consequently, the inverse obstacle scattering problem is to solve

$$A(r) = u_\infty \tag{10.2.4}$$

for the unknown radial function r.

Based on these parameterizations, the operator A can be proven to be Fréchet differentiable ([17]).

Theorem 10.2.1. *The boundary to far-field mapping $A : r \mapsto u^\infty$ is Fréchet differentiable and the derivative is given by*

$$A'(r) : q \mapsto v^\infty, \tag{10.2.5}$$

where v^∞ is the far-field pattern of the radiation solution v to the Dirichlet problem for the Helmholtz equation in $\mathbb{R}^3 \setminus \overline{D_r}$ satisfying the boundary condition

$$v = -\nu \cdot \hat{x} \frac{\partial u}{\partial \nu} q \quad \text{on } \partial D_r \tag{10.2.6}$$

with $u = u^i + u^s$ being the total wave for scattering from the domain D_r.

Rigorous foundations for the Fréchet differentiability were established by Kirsch [41] in the sense of a domain derivative via variational methods and by Potthast [73] via boundary integral equation techniques. Based on the Green's

theorem and a factorization of the difference of the far-field for the domains D_r and D_{r+q}, Kress and Päiärinta provided an alternative proof in [47].

In terms of the Fréchet derivatives $A'(r)$, we may linearize the nonlinear operator as

$$A(r+q) = A(r) + A'(r)q + O(q^2).$$

Then, for a current approximation r for the solution to (10.2.4), in order to obtain an update $r+q$, we solve the approximate linear equation

$$A(r) + A'(r)q = u^\infty \tag{10.2.7}$$

for q instead of solving the original equation $A(r+q) = u^\infty$. As in the classical Newton iterations this linearized procedure is done until some given stopping criteria are satisfied. Here, it should be noted that the linearized equation (10.2.7) inherits the ill-posedness of the nonlinear equation (10.2.4), therefore the regularization scheme has to be applied. For example, in the standard Tikhonov regularization, (10.2.7) is replaced by

$$\alpha q + [A'(r)]^* A'(r)q = [A'(r)]^* \{u^\infty - A(r)\} \tag{10.2.8}$$

with some positive regularization parameter α and the L^2-adjoint $[A'(r)]^*$ of $A'(r)$. However, in this case, some properties of the operator $A'(r) : L^2(\mathbb{S}) \to L^2(\mathbb{S})$ need to be established. For the Dirichlet and impedance boundary condition with large λ, Kress and Rundell proved that $A'(r)$ is injective and has dense range, while the corresponding result remains open for the Neumann boundary condition [43].

Obviously, a common feature of the above method is of an iterative nature. Hence, for the numerical implementation an efficient forward solver is needed at each step and a good initial guess is required. In addition, this method is computationally costly in practice. However, this approach is conceptually simple, more important, it usually leads to highly accurate reconstructions with reasonable stability against errors in the far-field data.

10.2.2 Decomposition methods

The basic idea of these methods is to break up the inverse scattering problem into two steps. The first step deals with the ill-posedness, while the second step considers the nonlinearity. We confine our analysis to a sound-soft obstacle in $\mathbb{R}^3$. The method can also be carried over for two-dimensional case and other boundary conditions with some evident modifications.

In the classical potential method due to Kirsch and Kress [17], the first step is to reconstruct the scattered wave $u^s(x)$ from its far-field pattern u^∞, which is a linear ill-posed problem. In the second step we determine the unknown boundary

∂D of the scatterer from the boundary condition by solving a nonlinear optimization problem.

Firstly we choose an auxiliary closed $\mathcal{C}^2$ surface Γ contained in the unknown scatterer D based on some *a priori* information about D and represent the scattered field as an acoustic single-layer potential

$$u^s(x) = \int_\Gamma \phi(y)\Phi(x,y)ds(y) \tag{10.2.9}$$

with an unknown density $\phi \in L^2(\Gamma)$. From the asymptotic behavior of $\Phi(x,y)$, it can be seen that the far-field pattern of (10.2.9) has the representation

$$u^\infty(\hat{x}) = \frac{1}{4\pi}\int_\Gamma \phi(y)e^{-ik\hat{x}\cdot y}ds(y), \quad \hat{x} \in \mathbb{S}. \tag{10.2.10}$$

Hence, if we introduce the far-field operator $S^\infty : L^2(\Gamma) \to L^2(\mathbb{S})$ defined by

$$(S^\infty\phi)(\hat{x}) := \frac{1}{4\pi}\int_\Gamma \phi(y)e^{-ik\hat{x}\cdot y}ds(y), \quad \hat{x} \in \mathbb{S}, \tag{10.2.11}$$

then for given far field u^∞, the density function $\phi \in L^2(\Gamma)$ can be solved from

$$(S^\infty\phi)(\hat{x}) = u^\infty(\hat{x}), \quad \hat{x} \in \mathbb{S}. \tag{10.2.12}$$

Obviously, the integral operator S^∞ has an analytic kernel and therefore equation (10.2.12) is severely ill-posed.

Without loss of generality we assume that k^2 is not the Dirichlet eigenvalue for $-\Delta$ in the domain cycled by Γ. Then S^∞ is injective and has dense range in $L^2(\mathbb{S})$. Hence, we can apply the Tikhonov regularization method to solving (10.2.12), i.e.,

$$\alpha\phi_\alpha + S^{\infty,*}S^\infty\phi_\alpha = S^{\infty,*}u^\infty \tag{10.2.13}$$

with a regularization parameter $\alpha > 0$ and the L^2-adjoint $S^{\infty,*}$ of S^∞. Once ϕ_α is solved, an approximation u^s_α for the scattered field is obtained by inserting the density ϕ_α into (10.2.9). Here solving (10.2.13) is equivalent to the minimization of the penalized residual

$$\|S^\infty\phi - u^\infty\|^2_{L^2(\mathbb{S})} + \alpha\|\phi\|^2_{L^2(\Gamma)}$$

over all $\phi \in L^2(\Gamma)$.

Please notice, ϕ_α converges as $\alpha \to 0$ if and only if the original equation (10.2.12) is solvable, while the solvability of (10.2.12) is related to the question of whether or not the scattered wave can be analytically extended as a solution to the Helmholtz equation across the boundary ∂D. More precisely, (10.2.12) is solvable if and only if u^∞ is the far-field pattern of a radiating solution to the Helmholtz equation in the exterior of Γ with the boundary data in the Sobolev space $H^1(\Gamma)$. However, in general, these regularity properties of the scattered

wave cannot be known in advance for an unknown obstacle D. Whereas this problem can be remedied by using the range test to the far-field equation (10.2.12). We will present the details in next section.

Define the single-layer potential operator $S : L^2(\Gamma) \to L^2(\Lambda)$ by

$$(S\phi)(x) = \int_\Gamma \phi(y)\Phi(x,y)ds(y), \quad x \in \Lambda, \tag{10.2.14}$$

where Λ is a closed surface containing Γ in its interior.

Knowing u^s_α, we can determine the boundary ∂D of the scatterer via the location of zeros of $u^i + u^s$ in a minimum norm sense. More precisely, given some suitable class U of admissible surfaces Λ, we can approximate ∂D by solving

$$\min_U \|u^i + u^s_\alpha\|_{L^2(\Lambda)}. \tag{10.2.15}$$

Note that ϕ_α may not converge as $\alpha \to 0$ even if u^∞ is the exact far-field pattern, so, to remedy the problem with the convergence for ∂D, another possible approach is to combine the linear problem for solving (10.2.13) with (10.2.15) into one optimization problem. Thus, the final problem for determining ∂D is to minimize

$$\mu(\phi, \Lambda; \alpha) := \|S^\infty\phi - u^\infty\|^2_{L^2(\mathbb{S})} + \alpha\|\phi\|^2_{L^2(\Gamma)} + \gamma\|u^i + S\phi\|^2_{L^2(\Gamma)} \tag{10.2.16}$$

over$(\phi, \Lambda) \in L^2(\Gamma) \times U$, where $\alpha > 0$ is the regularization parameter and $\gamma > 0$ is a coupling parameter.

For this nonlinear optimization problem, some theoretical results are referred to [17, 42].

There is also the other decomposition method, namely, *dual space method* ([17, 42]). As in the potential method, this method also contains two steps.

In the first step we look for superpositions of incident fields with different directions which lead to simple far-field patterns.

Express an incident wave v^i as a Herglotz wave function with density $g \in L^2(\mathbb{S})$, i.e.,

$$v^i(x) = \int_{\mathbb{S}} e^{ikx\cdot d} g(d)ds(d), \quad x \in \mathbb{R}^3, \tag{10.2.17}$$

then its far-field pattern is given by

$$v^\infty(\hat{x}) = \int_{\mathbb{S}} u^\infty(\hat{x}, d)g(d)ds(d), \quad \hat{x} \in \mathbb{S}. \tag{10.2.18}$$

If we want the scattered wave to be a special radiating solution v^s to the Helmholtz equation with far-field pattern v^∞, then, for given far-field patterns $u^\infty(\cdot, d)$ for all incident directions d, we need to solve the integral equation

$$Fg = v^\infty, \tag{10.2.19}$$

where the operator $F : L^2(\mathbb{S}) \to L^2(\mathbb{S})$ is called the far-field operator, which is defined by

$$(Fg)(\hat{x}) := \int_{\mathbb{S}} u^\infty(\hat{x}, d)g(d)ds(d), \quad \hat{x} \in \mathbb{S}. \tag{10.2.20}$$

Under the assumption that $\mathbb{R}^3 \setminus \overline{D}$ is contained in the domain of definition for v^s, the solvability of (10.2.19) is related to the following interior Dirichlet problem

$$\Delta v^i + k^2 v^i = 0 \quad \text{in } D, \tag{10.2.21}$$

$$v^i + v^s = 0 \quad \text{on } \partial D. \tag{10.2.22}$$

More precisely, (10.2.19) is solvable for $g \in L^2(\mathbb{S})$ if and only if the solution v^i to (10.2.21) and (10.2.22) is a Herglotz wave function with kernel g.

Since u^∞ is an analytic function, (10.2.19) is severely ill-posed. So we again apply Tikhonov regularization to obtaining an approximate solution to (10.2.19), and therefore an approximation for the incident wave v^i is given by (10.2.17).

In the second step we analogously seek the boundary ∂D as the location of the zeros of the total field $v^i + v^s$ in the minimum norm sense.

However, as pointed out in the potential method case, to obtain a satisfactory reconstruction result, we need to combine these two steps together and minimize

$$\mu(g, \Lambda; \alpha) := \|Fg - v^\infty\|^2_{L^2(\mathbb{S})} + \alpha\|Hg\|^2_{L^2(\Lambda_0)} + \gamma\|Hg + v^s\|^2_{L^2(\Lambda)} \tag{10.2.23}$$

over $(g, \Lambda) \in L^2(\mathbb{S}) \times U$, where Λ_0 is a surface containing all surfaces of U ([17]), and $H : L^2(\mathbb{S}) \to L^2(\Lambda)$ is the Herglotz operator defined by

$$(Hg)(x) = \int_{\mathbb{S}} e^{ikx\cdot d} g(d)ds(d), \quad x \in \Lambda. \tag{10.2.24}$$

If we take $v^i(x) = -\Phi(z_0, x)$ for some fixed $z_0 \in D$, then the corresponding scattered field and far-field pattern are given respectively by

$$v^s(x) = \Phi(z_0, x),\ x \in \mathbb{R}^3 \setminus \overline{D} \quad \text{and} \quad v^\infty(\hat{x}) = \frac{1}{4\pi} e^{-ik\hat{x}\cdot z_0},\ \hat{x} \in \mathbb{S}.$$

If $z_0 \to \partial D$, then it can be proven that $\|g\|_{L^2(\mathbb{S})} \to \infty$, which leads to the so-called linear sampling method. The concrete description of this method will be given in the next section.

The point source method proposed by Potthast [61, 62, 63] can also be interpreted as a decomposition method. Here we state the more recent version of this method [51, 55], which is derived from the potential theory rather than the mixed reciprocity principle in the original version.

Firstly, for $x \in \mathbb{R}^3 \setminus \overline{D}$ we choose a domain G such that $x \notin G, \overline{D} \subset G$ and approximate the point source $\Phi(\cdot, x)$ for any fixed $x \in \mathbb{R}^3 \setminus \overline{G}$ by a superposition of plane waves

$$\Phi(z, x) = \int_{\mathbb{S}} e^{ikz\cdot d} g_x(d)ds(d), \quad z \in \partial G. \tag{10.2.25}$$

Secondly, we express $u^s(x)$ outside $\overline{G}$ as a single-layer potential

$$u^s(x) = \int_{\partial G} f(z)\Phi(z,x)ds(z), \quad x \in \mathbb{R}^3 \setminus \overline{G}, \tag{10.2.26}$$

with an unknown density $f \in L^2(\partial G)$, and therefore its far-field pattern is given by

$$u^\infty(\hat{x}) = \frac{1}{4\pi} \int_{\partial G} f(z)e^{-ik\hat{x}\cdot z}ds(z), \quad \hat{x} \in \mathbb{S}. \tag{10.2.27}$$

Finally, by inserting (10.2.25) into (10.2.26) and exchanging the order of integration, it follows from (10.2.27) that

$$u^s(x) = 4\pi \int_{\mathbb{S}} u^\infty(-d)g_x(d)ds(d), \quad x \in \mathbb{R}^3 \setminus \overline{G}. \tag{10.2.28}$$

Using the above steps, we can reconstruct $u^s(x)$ from its far-field pattern. Consequently the boundary ∂D can be determined approximately from the boundary condition.

It should be noted that equation (10.2.25) for the density may not have an exact solution. However, under the assumption that k^2 is not a Dirichlet eigenvalue for the $-\Delta$ in G, the Herglotz wave functions are dense in $L^2(\partial G)$, hence we can get an approximate solution by regularization schemes. For example, the approximation can be obtained by finding the minimum norm solution to (10.2.25) with arbitrary discrepancy $\varepsilon > 0$.

Here, we also would like to mention that the original version of point source method is based on the mixed reciprocity principle, and then in [51] the authors established a relation between the classical potential method and point source method, which suggests that the point source method can be derived from the potential method. Such a new interpretation extends the applicable scope of this reconstruction scheme from incident plane waves to arbitrary incident waves. Recently further quantitative relation between these two methods is explained in [55]. Essentially, these two methods are the same. In fact, if the regularizing parameter for one scheme is chosen as an appropriate constant multiple of that for the other scheme, then these two schemes will yield the exactly same solution. This essential relation originates from the almost adjoint relation between the Herglotz wave operator and the far-field operator, which can be explained as the duality relation of these two schemes.

In conclusion, the basic idea of decomposition method is to separate the ill-posedness from the nonlinearity. The main advantage is that it is not necessary to solve the forward problem in the numerical implementation, which is required for the iterative type method under a general framework. Of course, we cannot also avoid the disadvantages in the optimization procedure. For example, as in the Newton iterative method, some good *a priori* information about the unknown scatterer is still needed. Moreover, the accuracy of the numerical reconstructions is generally *inferior* to that of using the direct Newton iterative method.

10.2.3 Hybrid method

In recent years, a so-called hybrid method has been developed [43, 44, 45, 46, 74], which combines the ideas of iterative methods and decomposition methods, inheriting the advantages of each of them. Using this new algorithm, accurate reconstructions are expected without the forward solver at each step. In principle, this method can be viewed as a modification of the classical potential method, in the sense that the auxiliary surface Γ is adjusted at each iteration step using the known approximate shape ∂D to be updated. Then Γ is updated again by linearizing the boundary condition around Γ. Now we simply state the main idea behind this method.

Given a far-field pattern u^∞ and a current approximation ∂D_r described by (10.2.2) for the boundary surface, we firstly solve the following ill-posed integral equation

$$\frac{1}{4\pi}\int_{\partial D_r} e^{-ik\hat{x}\cdot y}\phi(y)ds(y) = u^\infty(\hat{x}), \quad \hat{x}\in\mathbb{S} \tag{10.2.29}$$

by standard Tikhonov regularization. Set

$$u^s(x) := \int_{\partial D_r} \Phi(x,y)\phi(y)ds(y)$$

for $x\in\mathbb{R}^3\setminus\partial D_r$, then we can compute the boundary values of $u = u^i + u^s$ and its derivatives on ∂D_r by the jump relations of classical potential method. In order to find an update $r+q$, we solve the linearized equation of the boundary condition $u|_{\partial D_{r+q}} = 0$, that is,

$$u|_{\partial D_r} + \hat{x}\cdot \operatorname{grad} u|_{\partial D_r} q = 0 \tag{10.2.30}$$

for q. In this way, an iterative scheme is generated with the advantage of not requiring a forward solver at each step.

In closing this section, we want to mention another modification of the potential method. Its main idea is to formulate the original inverse problem as a system of two equations for two unknowns ([48]).

Firstly, represent the scattered wave u^s by

$$u^s(x) = \int_{\partial D} \phi(y)\Phi(x,y)ds(y) \tag{10.2.31}$$

with an unknown density $\phi\in L^2(\partial D)$, and therefore its far-field pattern is given by

$$u^\infty(\hat{x}) = \frac{1}{4\pi}\int_{\partial D}\phi(y)e^{-ik\hat{x}\cdot y}ds(y), \quad \hat{x}\in\mathbb{S}. \tag{10.2.32}$$

Secondly, in terms of the sound-soft boundary condition, we have

$$u^i(x) + \int_{\partial D} \phi(y)\Phi(x,y)ds(y) = 0, \quad x \in \partial D. \tag{10.2.33}$$

Thus, (10.2.32) and (10.2.33) constitute a system of two equations for two unknowns ∂D and ϕ, which can be solved by regularized Newton iterations.

As an important feature of all inversion methods introduced in this section, the boundary type describing physical properties of boundary should be known. This requirement can be viewed as one of the main shortcomings of optimization type inversion scheme. In many cases, the physical property of the scatterer with the boundary shape to be reconstructed is also unknown. That is, we need to reconstruct both the boundary type and its geometric shape. We will view some inversion schemes in this area in the next section.

10.3 Singular source methods

In this section, we use the term "singular source methods" to represent different methods which apply some indicator to detect the boundary property. Generally, by introducing the point source, these methods construct some indicator function of the obstacle boundary using the far-field data of incident plane waves. Near the boundary, such indicators blow up in some way depending on the boundary type and boundary shape. Thus the boundary property can be detected. The mathematical essence behind these methods is the approximate computation of the related Green function using the far-field data.

We will investigate the probe method, singular sources method, linear sampling method, the factorization method, range test and no response test.

10.3.1 Probe method

In recent years, a new inversion scheme for reconstructing ∂D from the far-field data called probe method has been developed [12, 14, 29, 30, 31, 32, 33, 49], with the advantage of being an exact reconstruction formula. The key idea is to construct the Dirichlet-to-Neumann map using the far-field data of the scattered wave and then to compute an indicator function for suitably chosen approaching domain and boundary value f. In fact f is the trace of the Runge approximation of the fundamental solution to Helmholtz equation, with a detecting point z outside D. When $z \to \partial D$, the indicator function blows up. In this way, the boundary ∂D of the scatterer can be reconstructed.

Here, we state this method for the inverse scattering problem caused by an obstacle in $\mathbb{R}^2$ with impedance boundary condition, which can be found in [12, 49]. For the case of multiple obstacles, we refer to [14].

The inverse scattering problem can be stated as follows: For given incident plane waves $w^i(x) = e^{ikx\cdot d}$ with incident direction $d \in \mathbb{S}$, the total wave $w(x) =$

$w^i(x) + w^s(x)$ outside an impenetrable D with impedance boundary condition satisfies

$$\begin{cases} \Delta w + k^2 w = 0, & x \in \mathbb{R}^2 \setminus \overline{D} \\ \frac{\partial w}{\partial \nu} + i\lambda w = 0, & x \in \partial D \\ w^s(x) \text{ meets the radiation condition at } \infty. \end{cases} \tag{10.3.1}$$

If the far-field data $\{w^\infty(\hat{x}, d) : \hat{x}, d \in \mathbb{S}\}$ of w^s is given, we try to identify ∂D as well as its boundary impedance $\lambda(x)$.

To construct some indicator which can detect ∂D as well as its physical property, the main idea contains the following three steps:

- Construct a Dirichlet-to-Neumann map on the approximate domain boundary.
- Use this map to construct an indicator function with suitably chosen Dirichlet data.
- Compute the Dirichlet-to-Neumann map using the given far-field data.

Let Ω be a given domain containing the unknown obstacle, and k^2 be not a Dirichlet eigenvalue for the operator $-\Delta$ in Ω and $\Omega \setminus \overline{D}$ be connected.

For $f(x) \in H^{\frac{1}{2}}(\partial\Omega)$, consider the following mixed boundary value problem

$$\begin{cases} \Delta u + k^2 u = 0, \ x \in \Omega \setminus \overline{D} \\ \frac{\partial u}{\partial \nu} + i\lambda u = 0, \ x \in \partial D \\ u = f, \qquad x \in \partial\Omega \end{cases} \tag{10.3.2}$$

for $u \in H^1(\Omega \setminus \overline{D})$, where λ is the impedance coefficient with $\Re\lambda > 0$.

By (10.3.2), we can define a Dirichlet-to-Neumann map $\Lambda_{\partial D} : H^{\frac{1}{2}}(\partial\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$ as

$$\Lambda_{\partial D} : f(x) \mapsto \partial_\nu u|_{\partial\Omega}. \tag{10.3.3}$$

Corresponding to the case of $D = \emptyset$, we also introduce a map $\Lambda_0 : H^{\frac{1}{2}}(\partial\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$,

$$\Lambda_0 : f(x) \mapsto \partial_\nu u_1|_{\partial\Omega}, \tag{10.3.4}$$

where $u_1 \in H^1(\Omega)$ satisfies

$$\begin{cases} \Delta u_1 + k^2 u_1 = 0, \ x \in \Omega \\ u_1 = f, \qquad x \in \partial\Omega. \end{cases} \tag{10.3.5}$$

Due to the assumption of Ω, the existence and uniqueness of the solution to (10.3.5) are ensured.

Lemma 10.3.1. *Let u be the solution to problem (10.3.2) for $f(x) \in H^{\frac{1}{2}}(\partial\Omega)$. Then $\frac{\partial u}{\partial \nu}|_{\partial\Omega}$ can be obtained from $f(x)$ and far-field data $\{w^\infty(\hat{x}, d) : \hat{x}, d \in \mathbb{S}\}$.*

Obviously, this lemma implies that the Dirichlet-to-Neumann map can be determined completely from the far-field patterns for all incident directions. From this result, we see that the original inverse problem can be restated as the problem

of reconstructing the shape and boundary impedance of the obstacle from the map $\Lambda_{\partial D}$.

In the sequel, our main goal is to find the appropriate function $f(x) \in H^{\frac{1}{2}}(\partial\Omega)$ and construct the indicator function. To this end, the Runge approximation theorem is needed.

For any continuous curve $c = \{c(t)|0 \leq t \leq 1\}$, if it satisfies $c(0), c(1) \in \partial\Omega$ and $c(t) \in \Omega$ for $0 < t < 1$, we call c a needle in Ω.

Lemma 10.3.2. *Suppose that Γ is an arbitrary open set of $\partial\Omega$. For each $t > 0$, there exists a sequence $\{v_n\}_{n=1,2,\cdots}$ in $H^1(\Omega)$, which satisfies $\Delta v_n + k^2 v_n = 0$, such that $supp(v_n|_{\partial\Omega}) \subset \Gamma$ and $v_n \to \Phi(\cdot - c(t))$ in $H^1_{loc}(\Omega \setminus \{c(t') \mid 0 < t' \leq t\})$, where $\Phi(x, y)$ is the fundamental solution to the Helmholtz equation.*

Since $v_n|_{\partial\Omega}$ depends on $c(t)$, we denote it by $f_n(\cdot, c(t))$, that is, $f_n(\cdot, c(t)) := v_n|_{\partial\Omega} \in H^{1/2}(\partial\Omega)$ with $supp(f_n(\cdot, c(t))) \subset \Gamma$.

For the given needle $c \in \Omega$ and $0 < t < 1$, we construct the indicator function

$$I(t, c) = \lim_{n\to\infty} \langle \overline{(\Lambda_{\partial D} - \Lambda_0) f_n(\cdot, c(t))}, f_n(\cdot, c(t)) \rangle, \tag{10.3.6}$$

where $\langle \cdot, \cdot \rangle$ is the L^2-inner product on $\partial\Omega$.

The main result of the probe method in this case can be stated as follows ([12]):

Theorem 10.3.3. *For a given needle c in Ω, we have*

$$\lim_{c(t)\to\partial D} \Re(I(t, c)) = -\infty \tag{10.3.7}$$

and $I(t, c)$ exists for all $c(t)$ outside D satisfying $|\Re(I(t, c(t)))| < +\infty$.

Theoretically, the probe method can be viewed as a novel method since this reconstruction scheme is an exact formula in the sense that in principle we can obtain the full shape of unknown obstacle D.

The natural far-field version of the probe method is analyzed in [28, 64], where this method is proved to be identical to the singular sources method. This version uses directly the far-field patterns to reconstruct the boundary of the obstacle in one step, and its indicator function is equivalent to that of the original version stated above with respect to the blow-up property.

Once the boundary ∂D is determined, the boundary impedance λ can also be identified. For the reconstruction of $\lambda(x)$ using the moment method, we refer to [12].

10.3.2 Singular sources method

The idea of the singular sources method is to use the scattered field $\Phi^s(z, z)$ for the point source $\Phi(\cdot, z)$ with source location $z \in \mathbb{R}^3 \setminus \overline{D}$ as an indicator function,

which can be computed from the far-field data of the incident plane waves for all incident directions. Under suitable assumptions on the unknown scatterer, the blow-up property

$$|\Phi^s(z,z)| \to \infty, \quad z \to \partial D. \tag{10.3.8}$$

can be shown. Consequently, the boundary ∂D can be reconstructed by probing the area where the scatterer might be located.

In original version of singular sources method ([25, 62, 63, 64, 65]), to compute $\Phi^s(z,z)$ from the far-field pattern, the authors use the back project operator, which is constructed by using the mixed reciprocity relation. However, the calculation of $\Phi^s(z,z)$ can be carried out from a formula justified directly by using the identity (10.1.11). Now we state the modified version of this method.

Let $z \in \mathbb{R}^3 \setminus \overline{D}$. We suitably choose the domain G such that $\overline{D} \subset G$ and $z \in \mathbb{R}^3 \setminus \overline{G}$. Moreover, we assume that the interior Dirichlet problem in G is uniquely solvable. In this case, due to the denseness property of the Herglotz wave functions in $L^2(\partial G)$, we can take $\{f_n\}_{n=1}^{\infty}$ and $\{g_m\}_{m=1}^{\infty}$ as sequences in $L^2(\mathbb{S})$ such that

$$\|v_{f_n} - \Phi(\cdot,x)\|_{L^2(\partial G)} \to 0,\ n \to \infty, \quad \|v_{g_m} - \Phi(\cdot,z)\|_{L^2(\partial G)} \to 0,\ m \to \infty \tag{10.3.9}$$

for $x, z \in \mathbb{R}^3 \setminus \overline{G}$, where v_{f_n} and v_{g_m} are the Herglotz wave functions with densities f_n and g_m respectively.

Noticing that v_{f_n}, v_{g_m}, $\Phi(\cdot,x)$ and $\Phi(\cdot,z)$ satisfy the Helmholtz equation in G, it is deduced from (10.3.9) that

$$\|v_{f_n} - \Phi(\cdot,x)\|_{H^1(D)} \to 0,\ n \to \infty, \quad \|v_{g_m} - \Phi(\cdot,z)\|_{H^1(D)} \to 0,\ m \to \infty. \tag{10.3.10}$$

Multiplying in (10.1.11) by $f_n(\hat{x})g_m(d)$ and integrating $\mathbb{S} \times \mathbb{S}$, we get

$$\begin{aligned}
&\int_{\mathbb{S}}\int_{\mathbb{S}} u^{\infty}(-\hat{x},d) f_n(\hat{x}) g_m(d) ds(\hat{x}) ds(d) \\
&= \frac{1}{4\pi}\int_{\partial D}\left\{\int_{\mathbb{S}} \frac{\partial e^{ik\hat{x}\cdot y}}{\partial \nu(y)} f_n(\hat{x}) ds(\hat{x}) \cdot \int_{\mathbb{S}} u^s(y,d) g_m(d) ds(d)\right. \\
&\quad \left. - \int_{\mathbb{S}} \frac{\partial u^s(y,d)}{\partial \nu(y)} g_m(d) ds(d) \cdot \int_{\mathbb{S}} e^{ik\hat{x}\cdot y} f_n(\hat{x}) ds(\hat{x})\right\} ds(y) \\
&= \frac{1}{4\pi}\int_{\partial D}\left\{\frac{\partial v_{f_n}(y)}{\partial \nu(y)} v^s_{g_m}(y) - \frac{\partial v^s_{g_m}(y)}{\partial \nu(y)} v_{f_n}(y)\right\} ds(y).
\end{aligned} \tag{10.3.11}$$

where

$$v^s_{g_m}(y) := \int_{\mathbb{S}} u^s(y,d) g_m(d) ds(d)$$

is the scattered field related to the incident field v_{g_m}. From (10.3.10) and (10.3.11), we obtain

$$\lim_{n\to\infty}\int_{\mathbb{S}}\int_{\mathbb{S}} u^{\infty}(-\hat{x},d)f_n(\hat{x})g_m(d)ds(\hat{x})ds(d)$$
$$=\frac{1}{4\pi}\int_{\partial D}\left\{\frac{\partial\Phi(y,x)}{\partial\nu(y)}v^s_{g_m}(y)-\frac{\partial v^s_{g_m}(y)}{\partial\nu(y)}\Phi(y,x)\right\}ds(y)=\frac{1}{4\pi}v^s_{g_m}(x). \quad (10.3.12)$$

Hence the well-posedness of the direct scattering problem says $v^s_{g_m}(x)\to\Phi^s(x,z)$ as $m\to\infty$. These arguments lead to

Theorem 10.3.4. *For $x,z\in\mathbb{R}^3\setminus\overline{D}$, we have*

$$\Phi^s(x,z)=4\pi\lim_{m\to\infty}\lim_{n\to\infty}\int_{\mathbb{S}}\int_{\mathbb{S}}u^{\infty}(-\hat{x},d)f_n(\hat{x})g_m(d)ds(\hat{x})ds(d) \quad (10.3.13)$$

where f_n and g_m satisfy (10.3.9).

From the singularity of the fundamental solution, it can be deduced that $\Phi^s(z,z)\to\infty$ as the source point z approaches to the boundary ∂D. Hence, we can use

$$I_{ss}(z):=\Phi^s(z,z) \quad (10.3.14)$$

as an indictor function of the singular sources method to reconstruct the boundary ∂D.

For any probing point z, we firstly calculate $\Phi^s(z,z)$ in terms of (10.3.13), and then decide whether or not z is near to the boundary by observing the value of $\Phi^s(z,z)$. In this way, the rough shape of unknown obstacle can be reconstructed by suitable choosing the probing points z and corresponding approximate domains G.

Here we would like to point out that the probe method suggested by Ikehate follows from the uniqueness proof of Isakov ([34]) based on an energy integral for $\Phi^s(\cdot,z)$, while the singular sources method due to Potthast follows from the uniqueness proof of Kirsch and Kress ([37]) based on a point-wise calculation of $\Phi(z,z)$. In essence, they are two equivalent methods with respect to the blow-up rate of the indicator functions. Moreover, the blow-up rate is of the order of singularity of the fundamental solution, see [28, 57] for the details.

In addition, the computation of $\Phi^s(z,z)$ in original version is meaningful only for $z\in\mathbb{R}^3\setminus\overline{D}$, while the singular sources method for any $z\in\mathbb{R}^3$ is developed in [28]. Now we state it in a simple way.

Definition 10.3.5. *We call a bounded domain G with C^2-regular boundary, such that $\mathbb{R}^3\setminus\overline{G}$ is connected, a non-vibrating domain, if k^2 is not a Dirichlet eigenvalue for $-\Delta$. If this last condition is not satisfied, we say that G is vibrating.*

For any given domain, we can always test whether or not it is non-vibrating by using the Courant min-max principle. For details, we refer the reader to [59].

Definition 10.3.6 (The indicator function)**.** *Let Ω be large enough but bounded domain containing the unknown obstacle D. For $z\in\Omega$, we denote by $\mathcal{C}_z$ the set of continuous curves c_z which join the point z to the boundary $\partial\Omega$. For any curve $c_z\in\mathcal{C}_z$, we define Ω_z to be a C^2-regular domain contained strictly in $\Omega\setminus c_z$. Let*

v_{g_n} be a sequence of Herglotz waves which approximate the point source $\Phi(\cdot, z)$ and set

$$I_{ss}(z, c_z, \Omega_z) := 4\pi \lim_{n\to\infty} \int_{\mathbb{S}} \int_{\mathbb{S}} u^{\infty}(-\hat{x}, d) g_n(\hat{x}) g_n(d) ds(\hat{x}) ds(d).$$

Then, we define the indicator function $I(z)$ by

$$I(z) := \inf_{c_z \in \mathcal{C}_z} \sup_{\{\Omega_z \subset (\Omega \setminus c_z)\}} \{|I_{ss(z,c_z,\Omega_z)}|\}. \tag{10.3.15}$$

In this definition, if Ω_z is a vibrating domain, then we replace it with a larger non-vibrating domain $\tilde{\Omega}_z$ such that $z \notin \tilde{\Omega}_z$ and then approximate $\Phi(\cdot, z)$ on $\tilde{\Omega}_z$ and get also the same approximation on Ω_z. This means that we can define the functional $I_{ss}(z, c_z, \Omega_z)$ for any C^2-regular domains.

Theorem 10.3.7. *The indicator function $I(z)$ holds the two properties:*
(1) If $z \in \Omega \setminus \overline{D}$, then $I(z) = |\Phi^s(z, z)|$;
(2) If $z \in \overline{D}$, then $I(z) = \infty$.
As a conclusion, the obstacle D is characterized by the indicator function $I(z)$ as follows:

$$\overline{D} = \{z \in \Omega : \ I(z) = \infty\}.$$

It is also worth noticing that this theorem is based on a conjecture about the properties of $I_{ss}(z.c_z, \Omega_z)$, see Claim 3.4 in [28], while its full proof is still open.

Notice that, for the case of penetrable scatterers, the higher order multipoles are needed, if we still use the singular sources method to detect some information about the scatterers. Consider the scattering of acoustic waves by some inhomogeneous medium with the refractive index $\chi = 1 - n$. Then the scattered field $\Phi^s(z, z)$ corresponding to incident point source $\Phi(\cdot, z)$ is bounded independently of the location of its source point z ([63]). This implies the behavior of $\Phi^s(z, z)$ does not characterize the unknown boundary. However, it has been shown by Potthast [63, 67, 68] that the scattered field $\Phi^s_\mu(z, z)$ for the multipoles with appropriate order μ holds a singular behavior. Using this singularity we can recover the shape of an inhomogeneous medium. Further, we can also determine the size of the jump of χ at the boundary.

More recently, the dipoles and multipoles of order two are also applied to reconstructing the complex obstacles in Model problem III by probing method from the far-field patterns for many incident plane waves [52, 53, 54, 58]. Using dipoles or multipoles of order two as point sources, we can derive the asymptotic expansion of the indicator functions with respect to the source point. In fact, the first (i.e. highest) order term of the real parts gives the location of this surface and the unit normal vectors on it, while the second order terms involve the curvature coupled, in a clear and simple way, with the imaginary part of the surface impedance. The appearance of the curvature explains the difficulty to reconstruct non-uniform shapes, the non convex ones in particular, without more *a priori* information. In addition, this relation enables us to use the surface

impedance (the coating coefficient) to design obstacles which can be detected in a more (or less) accurate way by using the indicator function-based methods.

Here, we restrict ourselves to two-dimensional Helmholtz model which is a well-known approximate model of practical three-dimensional obstacle, provided that the obstacle in $\mathbb{R}^3$ is of cylinder form [5].

Let D be a bounded domain of $\mathbb{R}^2$ with its boundary $\partial D \in C^{2,1}$ and $\mathbb{R}^2 \setminus \overline{D}$ is connected. We assume that ∂D has the decomposition

$$\partial D = \overline{\partial D_I} \cup \overline{\partial D_D},\ \partial D_I \cap \partial D_D = \emptyset,$$

where ∂D_D and ∂D_I are open curves on ∂D.

For given incident plane wave $u^i(x) = e^{ikd\cdot x}$, we associate the total wave $u(x) = u^i(x) + u^s(x)$ satisfying the following exterior problem

$$\begin{cases} \Delta u + k^2 u = 0 & \text{in } \mathbb{R}^2 \setminus \overline{D}, \\ u = 0 & \text{on } \partial D_D, \\ \frac{\partial u}{\partial \nu} + ik\sigma u = 0 & \text{on } \partial D_I, \end{cases} \tag{10.3.16}$$

with $\nu(x)$ as the outward normal direction of ∂D, where the scattered fields u^s satisfies the Sommerfeld radiation condition

$$\lim_{r\to\infty} \sqrt{r}(\frac{\partial u^s}{\partial r} - iku^s) = 0, \quad r = |x|. \tag{10.3.17}$$

Assume that the surface impedance $\sigma(x) := \sigma^r(x) + i\sigma^i(x)$ is a complex-valued Lipschitz continuous function, and its real part $\sigma^r(x)$ has a uniform lower bound $\sigma_0^r > 0$ on ∂D_I. The part ∂D_I is referred to as the coated part of ∂D and ∂D_D is the non-coated part.

For the above scattering problem by a complex obstacle, the inverse scattering problem is stated as follows.

Given $u^\infty(\cdot,\cdot)$ on $\mathbb{S}\times\mathbb{S}$ for the scattering problem (10.3.16)–(10.3.17), we need to

- Reconstruct the shape of the obstacle D;
- Reconstruct some geometrical properties of ∂D such as normal directions and the curvature;
- Distinguish the coated part ∂D_I from the non-coated part ∂D_D;
- Reconstruct the complex surface impedance $\sigma(x)$ on ∂D_I, including the real and the imaginary parts.

The answers to the above problems are based on the following asymptotic expansion results, see [53, 54] for more details.

The scattered field associated with the Herglotz incident field $v_g^i := v_g(x)$ defined by

$$v_g(x) := \int_{\mathbb{S}} e^{ikx\cdot d} g(d)\, ds(d), \quad x \in \mathbb{R}^2 \tag{10.3.18}$$

with $g \in L^2(\mathbb{S})$ is given by

$$v_g^s(x) := \int_{\mathbb{S}} u^s(x,d)g(d)\ ds(d), \quad x \in \mathbb{R}^2 \setminus \overline{D}, \tag{10.3.19}$$

and its far-field pattern is

$$v_g^\infty(\hat{x}) := \int_{\mathbb{S}} u^\infty(\hat{x},d)g(d)\ ds(d), \quad \hat{x} \in \mathbb{S}. \tag{10.3.20}$$

Denote by $\Phi(x,z)$ the fundamental solution for the Helmholtz equation in $\mathbb{R}^2$. Assume that $\overline{D} \subset\subset \Omega$ for some known Ω with smooth boundary. For $a \in \Omega \setminus D$, denote by $\{z_p\} \subset \Omega \setminus \overline{D}$ a sequence tending to a and starting from $\partial\Omega$. For any z_p, set D_a^p a C^2-regular domain such that $\overline{D} \subset D_a^p$ (resp. $\overline{\partial D} \subset D_a^p$) with $z_q \in \Omega \setminus \overline{D_a^p}$ for every $q = 1, 2, \cdots, p$ and that the Dirichlet interior problem to D_a^p for the Helmholtz equation is uniquely solvable. In this case, the Herglotz wave operator H from $L^2(\mathbb{S})$ to $L^2(\partial D_a^p)$ defined by

$$(Hg)(x) := v_g(x) = \int_{\mathbb{S}} e^{ikx\cdot d} g(d)\ ds(d) \tag{10.3.21}$$

is injective, compacting dense range, see [17].

For every p fixed, we construct three density sequences $\{g_n^p\}$, $\{f_m^{j,p}\}$ and $\{h_l^{j,p}\}$ in $L^2(\mathbb{S})$ with $j = 1, 2$, by the Tikhonov regularization such that

$$\left\| v_{g_n^p} - \Phi(\cdot, z_p) \right\|_{L^2(\partial D_a^p)} \to 0, \quad n \to \infty, \tag{10.3.22}$$

$$\left\| v_{f_m^{j,p}} - \frac{\partial}{\partial x_j}\Phi(\cdot, z_p) \right\|_{L^2(\partial D_a^p)} \to 0, \quad m \to \infty, \tag{10.3.23}$$

$$\left\| v_{h_l^{j,p}} - \frac{\partial}{\partial x_j}\frac{\partial}{\partial x_2}\Phi(\cdot, z_p) \right\|_{L^2(\partial D_a^p)} \to 0, \quad l \to \infty. \tag{10.3.24}$$

Using these three density sequences, we construct the following three indicators

$$I^0(z_p) := \frac{1}{\gamma_2} \lim_{m\to\infty} \lim_{n\to\infty} \int_{\mathbb{S}}\int_{\mathbb{S}} u^\infty(-\hat{x},d)\ g_m^p(d)\ g_n^p(\hat{x})\ ds(\hat{x})ds(d), \tag{10.3.25}$$

$$I_j^1(z_p) := \frac{1}{\gamma_2} \lim_{m\to\infty} \lim_{n\to\infty} \int_{\mathbb{S}}\int_{\mathbb{S}} u^\infty(-\hat{x},d)\ f_m^{j,p}(d)\ g_n^p(\hat{x})\ ds(\hat{x})ds(d), \tag{10.3.26}$$

$$I_j^2(z_p) := \frac{1}{\gamma_2} \lim_{m\to\infty} \lim_{n\to\infty} \int_{\mathbb{S}}\int_{\mathbb{S}} u^\infty(-\hat{x},d)\ h_m^{j,p}(d)\ g_n^p(\hat{x})\ ds(\hat{x})ds(d), \tag{10.3.27}$$

where $\gamma_2 = \frac{e^{\frac{\pi}{4}i}}{\sqrt{8\pi k}}$.

For the points $a \in \partial D$, we choose the sequence $\{z_p\}_{p\in\mathbb{N}}$ included in $C_{a,\theta}$, where $C_{a,\theta}$ is a cone with center a, angle $\theta \in [0, \frac{\pi}{2})$ and axis $\nu(a)$. Denote by $\mathcal{C}(a)$ the curvature of ∂D at point a.

Theorem 10.3.8. *Assume that the boundary ∂D is of class $C^{2,1}$ and $\sigma = \sigma^r + i\sigma^i$ defined on ∂D_I is a complex-valued Lipschitz function with positive real part. Then the above three indicators have the following asymptotic formulas:*

I. For pole $\Phi(x, z)$ as source, it follows that

$$\Re\, I^0(z_p) = \begin{cases} -\dfrac{1}{4\pi} \ln |(z_p - a) \cdot \nu(a)| + O(1), & a \in \partial D_I, \\ +\dfrac{1}{4\pi} \ln |(z_p - a) \cdot \nu(a)| + O(1), & a \in \partial D_D. \end{cases} \tag{10.3.28}$$

$$\Im\, I^0(z_p) = O(1), \quad a \in \partial D. \tag{10.3.29}$$

II. Using dipoles $\frac{\partial}{\partial x_j}\Phi(x, z)$ with $j = 1, 2$ as sources, it follows that

$$\Re\, I_j^1(z_p) =$$

$$\begin{cases} \dfrac{-\nu_j(a)}{4\pi|(z_p - a) \cdot \nu(a)|} - \dfrac{\nu_j(a)(k\sigma^i(a) + \frac{1}{2}\mathcal{C}(a))}{\pi} \ln |(z_p - a) \cdot \nu(a)| + O(1), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad a \in \partial D_I, \\ \dfrac{\nu_j(a)}{4\pi|(z_p - a) \cdot \nu(a)|} - \dfrac{\nu_j(a)}{2\pi}\mathcal{C}(a) \ln |(z_p - a) \cdot \nu(a)| + O(1), \ a \in \partial D_D. \end{cases} \tag{10.3.30}$$

$$\Im\, I_j^1(z_p) = \begin{cases} -\dfrac{\nu_j(a)k\sigma^r(a)}{\pi} \ln |(z_p - a) \cdot \nu(a)| + O(1), & a \in \partial D_I, \\ O(1), \quad a \in \partial D_D. \end{cases} \tag{10.3.31}$$

III. Using multipoles of order two $\frac{\partial}{\partial x_j}\frac{\partial}{\partial x_2}\Phi(x, z)$ with $j = 1, 2$, it follows that

$$\Re\, I_1^2(z_p) =$$

$$\begin{cases} \dfrac{\nu_1(a)\nu_2(a)}{4\pi|(z_p - a) \cdot \nu(a)|^2} - \dfrac{\nu_1(a)\nu_2(a)}{\pi}[k\sigma^i(a) + \dfrac{3}{4}\mathcal{C}(a)]\dfrac{1}{|(z_p - a) \cdot \nu(a)|} + \\ \quad O(\ln |(z_p - a) \cdot \nu(a)|), \quad a \in \partial D_I, \\ \dfrac{-\nu_1(a)\nu_2(a)}{4\pi|(z_p - a) \cdot \nu(a)|^2} - \dfrac{3\nu_1(a)\nu_2(a)}{4\pi}\mathcal{C}(a)\dfrac{1}{|(z_p - a) \cdot \nu(a)|} + \\ \quad O(\ln |(z_p - a) \cdot \nu(a)|), \quad a \in \partial D_D. \end{cases} \tag{10.3.32}$$

$$\Re\, I_2^2(z_p) =$$

$$\begin{cases} \dfrac{\nu_2^2(a)-\nu_1^2(a)}{8\pi|(z_p-a)\cdot\nu(a)|^2} - \dfrac{\nu_2^2(a)-\nu_1^2(a)}{2\pi}[k\sigma^i(a)+\dfrac{3}{4}\mathcal{C}(a)]\dfrac{1}{|(z_p-a)\cdot\nu(a)|}+ \\ \quad O(\ln|(z_p-a)\cdot\nu(a)|), \quad a\in\partial D_I, \\ \dfrac{\nu_1^2(a)-\nu_2^2(a)}{8\pi|(z_p-a)\cdot\nu(a)|^2} - \dfrac{3(\nu_2^2(a)-\nu_1^2(a))}{8\pi}\mathcal{C}(a)\dfrac{1}{|(z_p-a)\cdot\nu(a)|}+ \\ \quad O(\ln|(z_p-a)\cdot\nu(a)|), \quad a\in\partial D_D. \end{cases} \tag{10.3.33}$$

and

$$\Im\, I_1^2(z_p) = \begin{cases} \dfrac{\nu_1(a)\nu_2(a)}{\pi|(z_p-a)\cdot\nu(a)|}k\sigma^r + O(\ln|(z_p-a)\cdot\nu(a)|), & a\in\partial D_I, \\ O(\ln|(z_p-a)\cdot\nu(a)|), & a\in\partial D_D. \end{cases} \tag{10.3.34}$$

$$\Im\, I_2^2(z_p) = \begin{cases} \dfrac{\nu_2^2(a)-\nu_1^2(a)}{2\pi|(z_p-a)\cdot\nu(a)|}k\sigma^r + O(\ln|(z_p-a)\cdot\nu(a)|), & a\in\partial D_I, \\ O(\ln|(z_p-a)\cdot\nu(a)|), & a\in\partial D_D. \end{cases} \tag{10.3.35}$$

By using these formulas, the boundary location, boundary type as well as the boundary impedance can be identified numerically [53, 54].

10.3.3 Linear sampling method

Here we describe this method for *model problem I* with sound-soft boundary.

Let $\Phi_0^\infty(\cdot,z)$ be the far-field pattern of the fundamental solution $\Phi(\cdot,z)$ with source point $z\in\mathbb{R}^3$, then it follows from the asymptotic behavior of $\Phi(\cdot,z)$ that

$$\Phi_0^\infty(\hat{x},z) = \frac{1}{4\pi}e^{-ik\hat{x}\cdot z}, \quad \hat{x}\in\mathbb{S}. \tag{10.3.36}$$

Please notice that $\Phi_0^\infty(\cdot,z)$ is different from $\Phi^\infty(\cdot,z)$, which denotes the far-field pattern of the scattered field $\Phi^s(\cdot,z)$ for the incident point source $\Phi(\cdot,z)$. Generally, they are not the same. But we have $\Phi^\infty(\cdot,z) = -\Phi_0^\infty(\cdot,z)$ for $z\in D$ since in this case $\Phi^s(\cdot,z) = -\Phi(\cdot,z)$.

The main idea of linear sampling method is to approximately solve the far-field equation

$$(Fg_z)(\hat{x}) = \Phi_0^\infty(\hat{x},z), \quad \hat{x}\in\mathbb{S} \tag{10.3.37}$$

for z on a grid of points containing D, where the far-field operator F is defined by (10.2.20), and then to look for the points where $\|g_z\|_{L^2(\mathbb{S})}$ becomes unbounded.

Note that, as pointed out in the overview on dual space method, the solvability of (10.3.37) is equivalent to the existence of the solution as a Herglotz wave

function with the density g_z to the interior Dirichlet problems (10.2.21) and (10.2.22). However, the solution to (10.2.21) and (10.2.22) has an extension as a Herglotz wave function across the boundary ∂D only in some very special cases. Hence, in general there does not exist a solution to (10.3.37). Nevertheless, due to the denseness property of the Herglotz wave functions, we can obtain an approximate solution to (10.3.37) by applying the Tikhonov regularization.

Now the main result of the linear sampling method can be stated as follows [22]:

Theorem 10.3.9. *Assume that k^2 is not a Dirichlet eigenvalue for $-\Delta$ in D. Then*

(1) For $z \in D$ and a given $\epsilon > 0$, there exists a function $g_z^\epsilon \in L^2(\mathbb{S})$ such that

$$\|Fg_z^\epsilon - \Phi_0^\infty\|_{L^2(\mathbb{S})} < \epsilon \tag{10.3.38}$$

and the corresponding Herglotz wave function Hg_z^ϵ converges to a solution of

$$\begin{cases} \Delta u + k^2 u = 0, \text{ in } D \\ u = -\Phi(\cdot, z) \quad \text{on } \partial D \end{cases} \tag{10.3.39}$$

in $H^1(D)$ as $\epsilon \to 0$.

(2) For $z \in D$ and a fixed $\epsilon > 0$, we have

$$\lim_{z \to \partial D} \|Hg_z^\epsilon\|_{H^1(D)} = \infty \quad \text{and} \quad \lim_{z \to \partial D} \|g_z^\epsilon\|_{L^2(\mathbb{S})} = \infty.$$

(3) For $z \in \mathbb{R}^3 \setminus \overline{D}$ and a given $\epsilon > 0$, every $g_z^\epsilon \in L^2(\mathbb{S})$ that satisfies (10.3.38) is such that

$$\lim_{\epsilon \to 0} \|Hg_z^\epsilon\|_{H^1(D)} = \infty \quad \text{and} \quad \lim_{\epsilon \to 0} \|g_z^\epsilon\|_{L^2(\mathbb{S})} = \infty.$$

From this theorem, by solving the integral equations (10.3.37) for many sampling points z and scanning the values for $\|g_z\|_{L^2(\mathbb{S})}$, we can expect to obtain an approximation for ∂D. In fact, we can take it as a set of points where the norm of g_z is large enough.

The correlative discussion on the inverse medium scattering is given in [22]. For more cases and their applications, we refer to [2, 7, 8, 9, 10, 11, 18, 19, 20, 21, 23, 76].

10.3.4 Factorization method

The drawback that the integral equation (10.3.37), in general, is not solvable is remedied by the factorization method due to Kirsch [35]. In this method, (10.3.37) is replaced by

$$(F^*F)^{1/4} g_z(\cdot) = \Phi_0^\infty(\cdot, z), \tag{10.3.40}$$

where F^* is the L^2-adjoint of F.

Here we describe the main idea of this method only for the *Model problem I* with Dirichlet boundary condition ([35]).

Define an operator $G : L^2(\partial D) \to L^2(\mathbb{S})$ by $Gf = w^\infty$, where w^∞ is the far-field pattern of the radiating solution to the exterior Dirichlet problem with the boundary data $f \in L^2(\partial D)$. Recall the single-layer potential operator $S : L^2(\partial D) \to L^2(\partial D)$ defined by (10.2.14), the relation between the operators F, G and S is given by

Lemma 10.3.10. *The operator F has the decomposition of $F = -4\pi GS^*G^*$, where G^* and S^* are the L^2-adjoint operators of G and S, respectively.*

Assume that k^2 is not a Dirichlet eigenvalue for $-\Delta$ in D. Then the operator F is normal and one-to-one (Theorem 4.4 and Theorem 4.5 in [18]), and hence there exist eigenvalues $\lambda_j \in \mathbb{C}$ of F with $\lambda_j \neq 0$ for $j = 1, 2, \cdots$. The corresponding eigenfunctions $\psi_j \in L^2(\mathbb{S})$ form a complete orthogonal system in $L^2(\mathbb{S})$. We note that $|\lambda_j|$ are the singular values of F and $\{|\lambda_j|,\ \psi_j\ , \text{sign}(\lambda_j)\psi_j\}$ is a singular system of F, where $\text{sign}(\lambda_j) = \frac{\lambda_j}{|\lambda_j|}$. From the above lemma, it follows that

$$-4\pi GS^*G^*\psi_j = \lambda_j\psi_j, \quad j = 1, 2, \cdots .$$

Define the functions $\phi_j \in L^2(\partial D)$ by $G^*\psi_j = -\sqrt{\lambda_j}\phi_j,\ j = 1, 2, \cdots$, where we choose the branch of $\sqrt{\lambda_j}$ such that $\Im\sqrt{\lambda_j} > 0$, then we have

$$GS^*\phi_j = \frac{\sqrt{\lambda_j}}{4\pi}\psi_j, \quad j = 1, 2, \cdots . \tag{10.3.41}$$

It can be shown that the functions ϕ_j form a Riesz basis in the Sobolev space $H^{-\frac{1}{2}}(\partial D)$, i.e., $H^{-\frac{1}{2}}(\partial D)$ consists exactly of functions ϕ of the form

$$\phi = \sum_{j=1}^{\infty} \alpha_j\phi_j \quad \text{with } \sum_{j=1}^{\infty} |\alpha_j|^2 < \infty.$$

Furthermore, there exists a constant $c > 1$ with

$$\frac{1}{c^2}\|\phi\|^2_{H^{-\frac{1}{2}}(\partial D)} \leq \sum_{j=1}^{\infty} |\alpha_j|^2 \leq c^2\|\phi\|^2_{H^{-\frac{1}{2}}(\partial D)}, \quad \forall \phi \in H^{-\frac{1}{2}}(\partial D). \tag{10.3.42}$$

Now we can state the result for factorization method ([35]).

Theorem 10.3.11. *Assume that k^2 is not a Dirichlet eigenvalue for $-\Delta$ in D. Then the ranges of $G : H^{\frac{1}{2}}(\partial D) \to L^2(\mathbb{S})$ are given by*

$$\mathcal{R}(G) = \left\{ \sum_{j=1}^{\infty} \rho_j\psi_j : \sum_{j=1}^{\infty} \frac{|\rho_j|^2}{\sigma_j} < \infty \right\} = \mathcal{R}\left((F^*F)^{1/4}\right),$$

where $\left\{\sigma_j,\ \psi_j,\ \tilde{\psi}_j\right\}$ *is a singular system of* F.

Particularly, on the assumption that k^2 is not a Dirichlet eigenvalue for $-\Delta$ in D, it is easy to see that Φ_0^∞ is in the range of G if and only if $z \in D$. Consequently, if we solve the operator equation (10.3.40) with noisy far-field data u_δ^∞ by using the Tikhonov regularization with the regularization parameter chosen by the Morozov discrepancy principle, then the norm of regularized solution $\|g_z^\delta\|$ converges as the noise level $\delta \to 0$ if and only if $z \in D$. Comparing (10.3.37) with (10.3.40), these results can be explained in the sense that the operator F itself is too much more smoothing compared with $(F^*F)^{1/4}$ due to the fact $\Phi_0^\infty(\cdot, z) \notin F\left(L^2(\mathbb{S})\right)$ for $z \in D$.

On the other hand, the above arguments also yield a characterization of D, that is,

$$D = \left\{ z \in \mathbb{R}^3 : \sum_{j=1}^{\infty} \frac{|\rho_j^z|^2}{\sigma_j} < \infty \right\} = \left\{ z \in \mathbb{R}^3 : \Phi_0^\infty(\cdot, z) \in \mathcal{R}\left((F^*F)^{1/4}\right) \right\}, \tag{10.3.43}$$

where ρ_j^z are the expansion coefficients of $\Phi_0^\infty(\cdot, z)$ with respect to $\{\psi_j,\ j = 1, 2, \cdots\}$, i.e., $\rho_j^z = \langle \Phi_0^\infty(\cdot, z), \psi_j \rangle_{L^2}$. Moreover, there exists a constant $c > 1$ such that

$$\frac{1}{c^2}\|\Phi(\cdot, z)\|^2_{H^{\frac{1}{2}}(\partial D)} \le \sum_{j=1}^{\infty} \frac{|\rho_j^z|^2}{\sigma_j} \le c^2 \|\Phi(\cdot, z)\|_{H^{\frac{1}{2}}(\partial D)} \tag{10.3.44}$$

for all $z \in D$, which describes how the value of the series blows up when $z \to \partial D$. Indeed, we can easily show that $\|\Phi(\cdot, z)\|_{H^{\frac{1}{2}}(\partial D)}$ behaves as $\frac{1}{d(z, \partial D)}$ in $\mathbb{R}^3$ and $\ln|d(z, \partial D)|$ in $\mathbb{R}^2$, where $d(z, \partial D)$ denotes the distance of $z \in D$ from the boundary ∂D.

To visualize the scatterer in practice, the most convenient approach is via Picard's criterion for the solvability of linear ill-posed operator equations in terms of a singular system of F.

For the case of inverse medium scattering, the corresponding results are given in [36]. In [40], Kirsch provides the version of the factorization method for Maxwell' equations. The extension of the factorization method to the cases of limited far-field data and absorbing media has been carried out in [39], where a constrained optimization problem has to be solved. On the other hand, the factorization method can be viewed as an extension of the MUSIC algorithm from signal processing. The relation between them was first studied by Cheney [15], and further investigated by Kirsch [38]. More investigations on the factorization method and its applications can be found in [3, 4, 26, 27].

It should be noted that the behavior of the indicator function of the linear sampling method or factorization method with respect to the parameter is opposite to that of the probe method or singular sources method since it is bounded inside the obstacle D and becomes large when approaching to D and remains unbounded outside D.

The probe method, the singular sources method, the linear sampling method and the factorization method share the advantage that no knowledge about the boundary condition of the unknown scatterer is needed. Moreover, these methods are still valid in the limited aperture case, where the far field data is given only on an open subset of $\mathbb{S}$. The principal disadvantage of these methods lies in the fact that they all need to know the far-field patterns for incident plane waves of all or many directions. However, in practice, such a large number of input data is usually unavailable. Thus, the methods requiring only one or a few far-field patterns are of great importance. Of course, in this case, we cannot expect to reconstruct all properties of a scatterer if we do not have adequate *a priori* information, but we can detect some special information about the unknown scatterer, which is sufficient for practical applications in some cases.

10.3.5 Range test method

We firstly introduce the one-wave range test for *model problem I* with Dirichlet boundary condition. The basic idea is to test whether or not the far-field pattern is in the range of the single-layer potential operator [71, 66]. Then we can determine the convex scattering support of a scatterer with unknown physical properties from the far-field pattern for only one incident wave. The convex scattering support is a subset of the unknown scatterer, from which we can obtain some information about its location and rough shape.

Choose a test domain G such that the homogeneous interior Dirichlet problem for G has only the trivial solution. Define the single-layer potential $S_G\phi$ by

$$(S_G\phi)(x) := \int_{\partial G} \Phi(x,y)\phi(y)ds(y), \quad x \in \mathbb{R}^3, \tag{10.3.45}$$

then its far-field pattern is given by

$$(S_G^\infty\phi)(\hat{x}) = \frac{1}{4\pi}\int_{\partial G} e^{-ik\hat{x}\cdot y}\phi(y)ds(y), \quad \hat{x} \in \mathbb{S}. \tag{10.3.46}$$

For given far-field pattern u^∞, we establish the following far-field integral equation

$$(S_G^\infty\phi)(\hat{x}) = u^\infty. \tag{10.3.47}$$

As pointed out previously, the solvability of (10.3.47) is closely related to the question of whether or not the scattered wave u^s can be analytically extended into $\mathbb{R}^3 \setminus \overline{G}$. Precisely, the far-field equation (10.3.47) is solvable if and only if u^s can be analytically extended into $\mathbb{R}^3 \setminus \overline{G}$. This means the solvability of (10.3.47) can be used as a criterion for the analytical extensibility of u^s into $\mathbb{R}^3 \setminus \overline{G}$.

We employ the classical Tikhonov regularization and define the approximate solution to (10.3.47) as

$$\phi_\alpha := \left(\alpha I + S_G^{\infty,*} S_G^{\infty}\right)^{-1} S_G^{\infty,*} u^\infty, \quad \alpha > 0 \tag{10.3.48}$$

where $S_G^{\infty,*}$ denote the adjoint of S_G^{∞}.

By the standard arguments on Tikhonov regularization, it follows that

- If (10.3.47) does have a solution, then the norm $\|\phi_\alpha\|_{L^2(\partial G)}$ of (10.3.48) will be bounded for $\alpha \to 0$ and converges towards the norm of the true solution;
- If (10.3.47) does not have a solution, then the norm $\|\phi_\alpha\|_{L^2(\partial G)}$ will blow up as $\alpha \to 0$.

Thus, we can test the extensibility of the scattered field u^s by calculating the norm $\|\phi_\alpha\|$ for the test domain G. For sufficiently small fixed regularization parameter α and some appropriate cut-off constant C, if $\|\phi_\alpha\| \leq C$, then we think that (10.3.47) is solvable and therefore u^s can be analytically extended into $\mathbb{R}^3 \setminus \overline{G}$; if $\|\phi_\alpha\| > C$, then we conclude that (10.3.47) is unsolvable. For a set of test domains G_j, if the scattered field u^s can be analytically extended into $\mathbb{R}^3 \setminus \overline{G_j}$, we call G_j a positive test domain. Taking the intersection of all positive test domains, we obtain a subset of the unknown domain.

For the convergence and regularity of the range test, we refer to [71].

Here we point out that the range test method can be viewed as a generalization of the factorization method to the case of only one incident wave, since the range of the single-layer potential operator plays a key role for the blow-up property of the density function.

Since the boundary condition should be known *a priori* for the iterative methods and decomposition methods, while the far-field patterns for many incident waves are required for the probing method and linear sampling method, one of the advantages of range test is able to detect some information about the scatterer with unknown physical properties from the far-field pattern with only one incident wave. However, the range test in general cannot reconstruct the exact shape of D from the knowledge of the far-field pattern for one incident wave. To obtain the full reconstruction of the scatterer, the multi-wave range test is suggested by Potthast and Schulz [72].

Giving the far-field patterns for many incident plane waves, we firstly use the one-wave range test method for each wave to test some domain G. Assume that the domain G is positive for all $d \in \mathbb{S}$, then we can calculate $u^s(x,d)$ for $x \in \mathbb{R}^3 \setminus \overline{G}$ and $d \in \mathbb{S}$ via

$$u^s(x,d) = S_G\left((S_G^{\infty})^{-1} u^\infty(\hat{x},d)\right). \tag{10.3.49}$$

By using the mixed reciprocity relation (10.1.14), we obtain

$$\Phi^\infty(d,x) = \frac{1}{4\pi} S_G\left((S_G^{\infty})^{-1} u^\infty(\hat{x},-d)\right) \tag{10.3.50}$$

for $x \in \mathbb{R}^3 \setminus \left(\overline{G} \cup \overline{D}\right)$ and $d \in \mathbb{S}$.

Then we apply the one-wave range test to the following far-field equation corresponding to point source

$$(S_G^{\infty}\psi_x)(d) = \Phi^{\infty}(d,x). \tag{10.3.51}$$

Now we give the results of the extensibility of the scattered field $\Phi^s(\cdot,x)$ [72].

Theorem 10.3.12. *If $\overline{D} \subset G$, then the scattered field $\Phi^s(\cdot,x)$ can be analytically extended up to $\mathbb{R}^3 \setminus \overline{G}$ uniformly for all $x \in \mathbb{R}^3 \setminus \overline{G}$. If $D \not\subset G$, then the scattered field $\Phi^s(\cdot,x)$ cannot be analytically extended up to $\mathbb{R}^3 \setminus \overline{G}$ uniformly for all $x \in \mathbb{R}^3 \setminus \overline{G}$.*

This means that equation (10.3.51) is solvable for all $x \in \mathbb{R}^3 \setminus \overline{G}$ with uniformly bounded solution ψ_x only if $D \subset G$. With this characterization we can get a reconstruction of D by taking the intersection of all test domains G for which the supermum of the norm of ψ_x for some test points $x \in \mathbb{R}^3 \setminus \overline{G}$ is sufficiently small. In principle, the full shape of the unknown scatterer D can be reconstructed by properly choosing the set of test domains.

10.3.6 No response test method

The no response test method is firstly proposed by Luke and Potthast [56] to locate the support of a scatterer from a knowledge of the far-field pattern for only one incident wave. The main idea is to construct special incident waves which are small on some test domain and then to estimate the response to these waves. If the response is small, the unknown object is assumed to be a subset of the test domain. Especially, this method does not depend on the information about whether the scatterer is penetrable or impenetrable, nor does it depend on the physical properties of the scatterer.

We present the no response test method for *model problem I*, and then simply describe its multi-wave version.

Multiply (10.1.11) by $g \in L^2(\mathbb{S})$ and integrate $\mathbb{S}$, then we get

$$\begin{aligned} I(g) &:= \int_{\mathbb{S}} u^{\infty}(-\hat{x}) g(\hat{x}) ds(\hat{x}) \\ &= \frac{1}{4\pi} \int_{\partial D} \int_{\mathbb{S}} \left(u^s(y) \frac{\partial e^{iky\cdot d} g(d)}{\partial \nu(y)} - \frac{\partial u^s(y)}{\partial \nu(y)} e^{iky\cdot d} g(d) \right) ds(d) ds(y) \\ &= \frac{1}{4\pi} \int_{\partial D} \left(u^s \frac{\partial v_g}{\partial \nu} - \frac{\partial u^s}{\partial \nu} v_g \right) ds, \end{aligned} \tag{10.3.52}$$

where v_g is the Herglotz wave function.

Let v_g and its derivatives be small on some test domain G, then the above functional $I(g)$ should be small if $D \subset G$, while it will be arbitrarily large if $\overline{G} \subset \mathbb{R}^3 \setminus \overline{D}$. This idea is used in [56] to construct an approximation for an unknown scatterer D. It should be mentioned that the case where $D \cap \left(\mathbb{R}^3 \setminus \overline{G}\right) \neq \emptyset$ and $D \cap G \neq \emptyset$ remains open [56], but it was resolved by Potthast [69]. Now we give this further investigation as follows.

Let G be the admissible test domain, define I_ϵ for $\epsilon > 0$ by

$$I_\epsilon(G) := \sup\Big\{|I(g)| : g \in L^2(\mathbb{S}) \text{ such that } \|v_g\|_{C^1(\overline{G})} \le \epsilon\Big\}. \tag{10.3.53}$$

Then we construct the indicator function via

$$I_0(G) := \lim_{\epsilon\to 0} I_\epsilon(G). \tag{10.3.54}$$

Theorem 10.3.13. *Let G be an admissible test domain. We have $I_0(G) = 0$ if the field u^s can be analytically extended into $\mathbb{R}^3 \setminus G$. If u cannot be analytically extended into $\mathbb{R}^3\setminus\overline{G}$, then we have $I_\epsilon(G) = \infty$ for all $\epsilon > 0$ and hence $I_0(G) = \infty$.*

Based on this theorem, we can obtain some upper estimate for the set of singular points of u^s by taking the intersections of the sets G for all possible test domains G with $I_0(G) = 0$. In this way, some special information about the unknown scatterer D is detected. Of course, as pointed out for the range test method, we cannot hope to reconstruct the full shape of D by this one-wave method. However, it is possible when the far-field pattern $u^\infty(\hat{x}, d)$ is known for a large number of incident waves with different directions $d \in \mathbb{S}$.

Next we introduce two versions of the multi-wave no response test ([28]), which are proved to be equivalent with respect to the convergence properties.

Multiplying (10.1.11) by $f(\hat{x})g(d)$ and integrating $\mathbb{S}\times\mathbb{S}$, we have

$$\begin{aligned} I(f,g) &:= \int_{\mathbb{S}}\int_{\mathbb{S}} u^\infty(-\hat{x}, d)f(\hat{x})g(d)ds(\hat{x})ds(d) \\ &= \frac{1}{4\pi}\int_{\partial D}\left\{\frac{\partial v_f(y)}{\partial\nu(y)}v_g^s(y) - \frac{\partial v_g^s(y)}{\partial\nu(y)}v_f(y)\right\}ds(y). \end{aligned} \tag{10.3.55}$$

Define the indicator function for the multi-wave no response test by

$$I_1(G) := \lim_{\epsilon\to 0}\sup\big\{I(f,g) : \|v_f\|_{L^2(\partial G)} \le \epsilon,\ \|v_g\|_{L^2(\partial G)} \le \epsilon\big\}. \tag{10.3.56}$$

For the set $\mathcal{G}$ of non-vibrating domains G (see Definition 10.3.5), we calculate the indicator function $I_1(G)$ and take the intersection of all non-vibrating domains G with $I_1(G) = 0$

$$D_1 := \bigcap_{I_1(G)=0} G. \tag{10.3.57}$$

Now the following characterization of D from the far-field patterns can be given [28].

Theorem 10.3.14. *If $\overline{D} \subset G$ then $I_1(G) = 0$; If $\overline{D} \not\subset G$ then $I_1(G) = \infty$. Thus the unknown scatterer is given by $\overline{D} = D_1$.*

By combining the superposition principle and the range test, the second version of the no response test is also developed [28].

Firstly, we apply the range test to the far-field equation corresponding to the incident Herglotz wave:

$$\frac{1}{4\pi}\int_{\partial G} e^{-ik\hat{x}\cdot y}\phi(y)ds(y) = v_g^{\infty}, \quad \hat{x}\in\mathbb{S}, \tag{10.3.58}$$

where v_g^{∞} is the far-field pattern related to the incident Herglotz wave. The regularizing solution to (10.3.58) is

$$\phi_g^{\alpha} := \left(\alpha I + S_G^{\infty,*}S_G^{\infty}\right)^{-1} S_G^{\infty,*}v_g^{\infty}, \quad \alpha>0. \tag{10.3.59}$$

Secondly, we define the indicator function by

$$I_2(G) := \lim_{\epsilon\to 0}\sup\left\{\lim_{\alpha\to 0}\|\phi_g^{\alpha}\|_{L^2(\partial G)} : g\in L^2(\mathbb{S}) \text{ and } \|v_g\|_{L^2(\partial G)}\le\epsilon\right\}. \tag{10.3.60}$$

Finally, we calculate the indicator function $I_2(G)$ and take the intersection

$$D_2 := \bigcap_{G\in\mathcal{G}_1} G. \tag{10.3.61}$$

where $\mathcal{G}_1 := \{G\in\mathcal{G} : I_2(G) = 0\}$.

The convergence of this multi-wave no response test can be stated as [28]

Theorem 10.3.15. *If $\overline{D}\subset G$ then $I_2(G) = 0$; If $\overline{D}\not\subset G$ then $I_2(G) = \infty$. Thus, the scatterer is characterized by $\overline{D} = D_2$.*

All methods viewed in this section are based on the construction of some indicator functions depending on a parameter point. If the parameter varies in a certain way, then the behavior of these indicator functions changes drastically. By these special properties, the scatterer can be characterized, and hence some information about the unknown scatterer can be extracted.

10.4 Numerical schemes

In this section, we will focus on the numerical implementations of all inversion methods introduced in Section 10.3. In each method, the indicator function plays an important role, so the key to its numerical realization is to calculate the indicator function efficiently. For simplicity, we only consider the cases in two-dimensional space, in principle, we can generalize them to three-dimension with proper modifications.

As the basis of our work, we firstly consider how to compute the approximations for the point sources on ∂G efficiently, since in detecting the obstacle boundary, the approximated domain ∂G needs to be chosen for z approaching to ∂D along all directions. If we get this approximation always by solving the

minimum norm solution on ∂G, then the amount of computation will be quite large.

It has been shown in [54, 63] that if the approximation domain G is constructed from some fixed reference domain G_0 by rotation and translation, then the approximations of the point sources on ∂G can be constructed from the related minimum norm solutions defined in the fixed domain ∂G_0, using a simple function transform depending on the rotation matrix and translation vector. By this technique, the amount of computation for constructing the indicator function will be decreased dramatically.

For given fixed reference domain G_0 with $0 \notin G_0$ and smooth boundary ∂G_0, let G be a domain generated from G_0 by rotation and translation. We assume that

$$G = \mathbb{M}G_0 + z_0,$$

with an unit orthogonal matrix $\mathbb{M} = (m_{ij})_{2\times 2}$ and the translation vector z_0.

Consider two integral equations of the first kind

$$(Hg_0)(x) = \Phi(x, 0), \quad x \in \partial G_0 \tag{10.4.1}$$

and

$$(Hg)(x) = \Phi(x, z_0), \quad x \in \partial G. \tag{10.4.2}$$

It has been proven [63] that

Lemma 10.4.1. *Assume that $g_0(d)$ is the minimum norm solution of (10.4.1) with discrepancy $\epsilon > 0$. Then $g(d)$ defined by*

$$g(d) = g_0(\mathbb{M}^{-1}d)e^{-ikd\cdot z_0} \tag{10.4.3}$$

is the minimum norm solution of (10.4.2) with discrepancy $\epsilon > 0$.

This result means that we can determine the density function $g(d)$ using the minimum norm solution in the fixed domain G_0 from (10.4.3) such that

$$\|(Hg)(\cdot) - \Phi(\cdot, z_0)\|^2_{L^2(\partial G)} \le \epsilon^2. \tag{10.4.4}$$

Next, we need to generalize this result to the cases where Φ is replaced by its partial derivatives.

For vector-valued function $(\varphi_1, \varphi_2)^T \in L^2(\mathbb{S}) \times L^2(\mathbb{S}) := L^2(\mathbb{S} \times \mathbb{S})$, define

$$H[(\varphi_1, \varphi_2)^T](x) := ((H\varphi_1)(x), (H\varphi_2)(x))^T.$$

For functions $(f_1, f_2)^T \in L^2(\Gamma) \times L^2(\Gamma) := L^2(\Gamma \times \Gamma)$, we define the norm

$$\|(f_1, f_2)^T\|^2_{L^2(\Gamma\times\Gamma)} := \|f_1\|^2_{L^2(\Gamma)} + \|f_2\|^2_{L^2(\Gamma)},$$

where Γ may be $\mathbb{S}, \partial G_0$ or ∂G.

Theorem 10.4.2. *Assume that $f_0^j(d)$ with $j = 1, 2$ are the minimum norm solutions of*

$$(Hf_0^j)(x) = \Phi_{x_j}(x,0), \quad x \in \partial G_0 \tag{10.4.5}$$

with discrepancy $\epsilon > 0$. *Then the density function* $(f^1, f^2)^T$ *given by*

$$\begin{pmatrix} f^1(d) \\ f^2(d) \end{pmatrix} := \mathbb{M} \begin{pmatrix} f_0^1(\mathbb{M}^{-1}d) \\ f_0^2(\mathbb{M}^{-1}d) \end{pmatrix} e^{-ikd\cdot z_0} \tag{10.4.6}$$

satisfies that

$$\|H[(f^1, f^2)^T](\tilde{x}) - (\Phi_{\tilde{x}_1}, \Phi_{\tilde{x}_2})^T(\tilde{x}, z_0)\|^2_{L^2(\partial G)} \leq 2\epsilon^2. \tag{10.4.7}$$

Compared with Lemma 10.4.1, this result is its generalization in some weak sense, that is, we can not assert $(f^1(d), f^2(d))$ is the minimum norm solution of

$$H[(\varphi^1, \varphi^2)^T](\tilde{x}) = (\Phi_{\tilde{x}_1}, \Phi_{\tilde{x}_2})^T(\tilde{x}, z_0), \quad \tilde{x} \in \partial G \tag{10.4.8}$$

with discrepancy $\sqrt{2}\epsilon$. Fortunately, such a weak generalization is enough sometimes.

A small change of the assumptions on $f_0^j(d)$ can guarantee that $(f^1(d), f^2(d))$ is the minimum norm solution of (10.4.8). This is the following analogy to Lemma 10.4.1.

Theorem 10.4.3. *Assume that the vector-valued function* $(f_0^1(d), f_0^2(d)) \in L^2(\mathbb{S} \times \mathbb{S})$ *is the minimum norm solution of*

$$H[(\varphi^1, \varphi^2)^T](x) = (\Phi_{x_1}, \Phi_{x_2})^T(x, 0), \quad x \in \partial G_0 \tag{10.4.9}$$

with discrepancy $\epsilon > 0$. *Then the density function* $(f^1, f^2)^T$ *given by (10.4.6) is the minimum norm solution of equation (10.4.8) with discrepancy* $\epsilon > 0$.

Comparing Theorem 10.4.2 and Theorem 10.4.3, both of them can be used to construct the approximation of $\Phi_{x_j}(\tilde{x}, z_0)$ in ∂G from the minimum norm solution in the reference domain ∂G_0. However, in order to guarantee that (f^1, f^2) is the minimum norm solution, we need to solve the density (f_0^1, f_0^2) as the minimum norm solution from the coupled equation (10.4.9), which will lead to large amount of computation since H is defined in $L^2(\mathbb{S}) \times L^2(\mathbb{S})$ in this case. Therefore, from the numerical point of view, Theorem 10.4.2 is more suitable for the approximation.

For the second order derivative of the fundamental solution, we have the similar results.

Theorem 10.4.4. *Assume that* $h_0^j(d)$ *with* $j = 1, 2$ *are minimum norm solutions of*

$$(Hh_0^j)(x) = \Phi_{x_j x_2}(x, 0), \quad x \in \partial G_0 \tag{10.4.10}$$

with discrepancy $\epsilon > 0$. *Then the density function* $(h^1, h^2)^T$ *given by*

$$\begin{pmatrix} h^1(d) \\ h^2(d) \end{pmatrix} := \tag{10.4.11}$$

$$\mathbb{M}^2 \begin{pmatrix} h_0^1(\mathbb{M}^{-1}d) \\ h_0^2(\mathbb{M}^{-1}d) \end{pmatrix} e^{-ikd\cdot z_0} - \mathbb{M} \begin{pmatrix} k^2 m_{21} \\ 0 \end{pmatrix} g(\mathbb{M}^{-1}d) e^{-ikd\cdot z_0} \tag{10.4.12}$$

satisfies that

$$\|H[(h^1, h^2)^T](\cdot) - (\Phi_{\tilde{x}_1\tilde{x}_2}, \Phi_{\tilde{x}_2\tilde{x}_2})^T(\cdot, z_0)\|^2_{L^2(\partial G\times\partial G)} \le (2+k^2)\epsilon^2. \tag{10.4.13}$$

From the above results, we get the following

Corollary 10.4.5. *Assume that g_0, f_0^j, h_0^j are the minimum norm solutions of equations (10.4.1), (10.4.5) and (10.4.10) with discrepancy ϵ in the fixed reference domain ∂G_0, respectively. Then the density functions g, f^j, h^j constructed by (10.4.3), (10.4.6) and (10.4.11) respectively meet*

$$\begin{aligned} \|(Hg) - \Phi(\cdot, z_0)\| &\le \epsilon, \\ \|(Hf^j) - \Phi_{x_j}(\cdot, z_0)\| &\le \sqrt{2}\epsilon, \\ \|(Hh^j) - \Phi_{x_j x_2}(\cdot, z_0)\| &\le \sqrt{2+k^2}\epsilon \end{aligned} \tag{10.4.14}$$

for $j = 1, 2$, where the norm is in $L^2(\partial G)$ and $G = \mathbb{M}G_0 + z_0$.

Based on these preparations, we give the numerical scheme of each method as follows:

For simplicity, in the probe method, we take the near-field data, i.e., Dirichlet-to-Neumann map, simulated by solving the direct problem (10.3.2) as the input data. Without loss of generality, we complete the numerical implementation on the following assumption:

Assumption 10.4.6. *We assume that*

- *$0 \in D$ which determines the position of D approximately;*
- *diameter$(D) < 1$ which gives an upper bound on the size of D;*
- *$\partial D \in \mathcal{C}^2$ with all points can be approached by a suitable straight line outside D.*

Based on these assumptions, we can take the test domain as

$$\Omega = \{(x, y) : x^2 + y^2 < 1\} \subset \mathbb{R}^2 \tag{10.4.15}$$

satisfying $\overline{D} \subset \Omega$, and for every point $c(0) \in \partial\Omega$, the straight line needle c connecting $c(0)$ and $0 \in D$ has a joint point with ∂D. In this way, we can approximate all the points on ∂D by $c(t)$ for different straight line needles c with $c(0) \in \partial\Omega$. Then the algorithm can be stated as follows [13]:

Algorithm 10.4.7. *(Probe method)*

- *For given needle c and point $c(t) \notin \bar{D}$, construct the test domain $G(c, t)$ with $\mathcal{C}^2$ regular boundary such that $\{c(t') : 0 \le t' \le t\} \in \Omega \setminus \overline{G(c, t)}$.*

It is enough to construct $G(c_0, t)$ for a special needle with $c_0(0) = (0, 1)$ since $G(c, t)$ for other needles can be obtained by rotation.
The key to construction of $G(c_0, t)$ is that we should keep the continuity of the second-order derivative at the joint points with $\partial\Omega$ and have some cone-like shape nearing $\{c_0(t') : 0 \le t' \le t\} \in \Omega \setminus \overline{G(c_0, t)}$. To this end, we consider a curve described by

$$y = Ax^6 + Bx^4 + Cx^2 + D, \quad |x| \le x_0 < 1 \tag{10.4.16}$$

with $x_0, y > 0$, which touches $\partial\Omega$ at points $(\pm x_0, y_0)$ with $\mathcal{C}^2$ smoothness. We can solve A, B, D for different touching points $(\pm x_0, y_0)$ and constant C from the continuity of the curve at $(\pm x_0, y_0) \in \partial\Omega$. The shape of the cone can be changed by adjusting the parameter $C > 0$ and point $(x_0, y_0) \in \partial\Omega$.

- *Construct the Runge approximation function v_n, and then calculate $f_n(x) = v_n|_{\partial\Omega}$;*
  Here we give the scheme to determine f_n on $\partial\Omega$ by constructing the minimum norm solution $g_{\frac{1}{n}}(c(t), \xi)$ to

$$(Hg)(x) = \Phi(x, c(t)), \quad x \in \partial G(c, t). \tag{10.4.17}$$

with discrepancy $\frac{1}{n}$, where Hg is the Herglotz wave function with a density g. For any fixed $c(t) \notin \overline{G(c, t)}$, $g_{\frac{1}{n}}(c(t), \xi) := \phi_0(\xi)$ can be solved from the equations

$$\begin{cases} \|(H\phi_0)(\cdot) - \Phi(\cdot, c(t))\|_{L^2(\partial G(c,t))} = \frac{1}{n}, \\ \alpha\phi_0(\xi) + (H^*H\phi_0)(\xi) = (H^*\Phi)(\xi) \end{cases} \tag{10.4.18}$$

Thus, we can compute f_n by

$$f_n(x, c(t)) := (Hg_{\frac{1}{n}})(x) = \int_{\mathbb{S}} e^{ikx\cdot\xi} g_{\frac{1}{n}}(c(t), \xi) ds(\xi), \quad x \in \partial\Omega. \tag{10.4.19}$$

- *Simulate $\Lambda_{\partial D} f_n$, $\Lambda_0 f_n$ in the way given in [50] and compute the indicator function $I(t, c)$ approximately from (10.3.6) for n large enough;*
- *For some $c(t^*)$ such that $|\Re I(t^*, c)| = -\Re I(t^*, c)$ is large enough, we think $c(t^*)$ is near to ∂D. Usually, we determine $c(t)$ to clarify whether or not it is near to ∂D by suitably choosing some cut-off constant C.*
- *Rotate the needle c_0 by choosing different $c_0(0) \in \partial\Omega$ and repeat the above four steps, then the boundary of D can be expressed as*

$$\partial D = \{c(t^*) : \textit{all different needle } c\}.$$

Here, we want to point out that for the needle c generated from c_0 by rotation we can construct the corresponding domain $G(c, t)$ and minimum norm solution to (10.4.17) in a simple way.

Let $c = (\cos\tau, \sin(\tau))$ with $\tau \in [0, 2\pi)$, then it is easily shown that

$$\partial G(c,t) = R(\tau)\partial G(c_0,t), \tag{10.4.20}$$

where $R(\tau)$ is the unit orthogonal rotation matrix

$$R(\tau) = \begin{pmatrix} \cos\left(\tau - \frac{\pi}{2}\right) & -\sin\left(\tau - \frac{\pi}{2}\right) \\ \sin\left(\tau - \frac{\pi}{2}\right) & \cos\left(\tau - \frac{\pi}{2}\right) \end{pmatrix}. \tag{10.4.21}$$

From Lemma 10.4.1, we can see that

$$g_{\frac{1}{n}}(c,t) = g_{\frac{1}{n}}(c_0, (R(\tau))^{-1}\xi), \tag{10.4.22}$$

from which the amount of computation is weakened greatly. For more details, see [13].

Here, we would also like to mention that there are other numerical work on the probe method for single obstacle [24, 78] and for multiple obstacles [77, 79], where the input data is also the simulated Dirichlet-to-Neumann map. Obviously, all these numerical implementations avoid the reconstruction of the Dirichlet-to-Neumann map from the far field pattern. Whereas this procedure can be carried out by combining the classical potential theory with the mixed reciprocity principle.

Algorithm 10.4.8. *(Singular sources method)*

- *Let Ω be some test domain. For $z \in \Omega$, construct an approximate domain $G(z)$ by the similar approach as that in the probe method, where $c(t)$ is replaced by z.*
- *Find the minimum norm solution $g_{\frac{1}{n}}$ to point source equation*

$$(Hg)(\cdot) = \Phi(\cdot, z) \quad \text{on } \partial G(z) \tag{10.4.23}$$

  with discrepancy $\frac{1}{n}$ for n large enough, which can be completed efficiently using Lemma 10.4.1.
- *Calculate*

$$I_n(z) := \sqrt{8\pi k}e^{-i\pi/4} \int_{\mathbb{S}} \int_{\mathbb{S}} u^{\infty}(-\hat{x}, d) g_n(\hat{x}) g_n(d) ds(\hat{x}) ds(d) \tag{10.4.24}$$

  and take $I_n(z)$ as an approximation for $\Phi^s(z,z)$.
- *Find the unknown boundary ∂D as the set of points such that $|I(z)|$ is large enough. Choose a cut-off constant, if $|I(z)| > C$, then we think z is near to ∂D.*

Obviously, the key step in the above scheme is to calculate the indicator function $\Phi^s(z,z)$, where the point source approximations are essential. However, there is another approach to calculate the more general function $\Phi^s(x,z)$ for $x \in \mathbb{R}^2 \backslash G$, which is based on the potential method [72]. In fact, this procedure is also a key component of the multi-wave range test, which will be given in Algorithm 10.4.11.

As to the probe method for complex obstacles, our main task is also to approximately calculate the various indicator functions for dipoles or multipoles of

order two. To this end, we can adopt the similar scheme as that in Algorithm 10.4.8, where the point sources equation (10.4.23) is replaced by

$$(Hg)(x) = \frac{\partial}{\partial x_j}\Phi(x,z) \quad \text{on } \partial G(z)$$

or

$$(Hg)(x) = \frac{\partial}{\partial x_j}\frac{\partial}{\partial x_2}\Phi(x,z) \quad \text{on } \partial G(z)$$

for $j = 1, 2$. To decrease the amount of computation, the results in Theorem 10.4.2 and Theorem 10.4.4 are needed.

It is worth noticing that the accuracy of the minimum norm solution to (10.4.17) or (10.4.23) in practice is not very satisfactory since the right-hand side of them is almost a singular function. If the scatterer is of strong non-convexity, then the construction of approximation domains will be not easy. For the case of multiple obstacles, this difficulty is also encountered. In addition, the choice of cut-off constant C is difficult in practice. Usually, we choose it by trial and error, while it is not quite convincible.

Algorithm 10.4.9. *(Linear sampling method and factorization method)*

- *Choose a sampling grid $\mathcal{Z}$ covering the unknown scatterer D.*
- *For each $z \in \mathcal{Z}$, we solve the far-field equation (10.3.37) in the case of the linear sampling method and (10.3.40) in the case of the factorization method by using the classical Tikhonov regularization method with regularization parameter α.*
  Here, note that the $u^\infty(\hat{x}, d)$ is used as the kernel of the integral operator F, then the error in the input data leads to an error in the operator F. So, in this case, the regularization parameter α can be determined by the following generalized Morozov discrepancy principle:

$$\|Fg_z^\alpha - \Phi_0^\infty\| = \delta\|g_z^\alpha\| \quad or \quad \left\|(F^*F)^{\frac{1}{4}}\, g_z^\alpha - \Phi_0^\infty\right\| = \delta\,\|g_z^\alpha\|\,,$$

  where δ denotes the noisy level of u^∞. On the other hand, for the factorization method, we can calculate $\|g_z^\alpha\|_{L^2(\mathbb{S})}$ directly from the spectral data of F. In fact, let $\{\sigma_j,\ \psi_j,\ \tilde{\psi}_j\}$ be a singular system of F, then we have the representation

$$(F^*F)^{\frac{1}{4}}\, g_z = \sum_{j=1}^{\infty} \sqrt{\sigma_j}\langle g_z, \psi_j\rangle_{L^2}\psi_j.$$

  Hence, the regularized solution has the form of

$$g_z^\alpha = \sum_{j=1}^{\infty} \frac{\sqrt{\alpha}}{\alpha + \sigma_j}\rho_j^z\psi_j,$$

  and its norm is given by

$$\|g_z^\alpha\|^2 = \sum_{j=1}^{\infty} \frac{\alpha}{(\alpha+\sigma_j)^2} \left|\rho_j^z\right|^2$$

with $\rho_j^z = \langle \Phi_0^\infty(\cdot, z), \psi_j \rangle_{L^2}$.

- *Calculate the norm* $\|g_z\|_{L^2}$ *of the regularized solution to (10.3.37) or (10.3.40), and then choose a cut-off constant* C*, which yield a rough reconstruction* D_{rec} *of* D *via*

$$D_{rec} := \{z \in \mathcal{Z} : \|g_z\| \leq C\}. \tag{10.4.25}$$

From the above scheme, we see that both the linear sampling method and the factorization method share a noticeable advantage over the probing methods. They do not involve the approximation domains, and therefore the geometrical difficulties from the construction of these domains can be avoided. Particularly, these two methods can be easily used for the treatment of multiple obstacles.

Algorithm 10.4.10. *(One-wave range test)*

- *Choose a family of non-vibrating test domains* G_j *for* j *in some index set* $\mathcal{J}$.
- *For each test domain* G_j*, we test the extensibility of* u^s *into* $\mathbb{R}^2 \setminus G_j$ *by calculating the regularized solution*

$$\phi_\alpha^j := \left(\alpha I + S_{G_j}^{\infty,*} S_{G_j}^\infty\right)^{-1} S_{G_j}^\infty u^\infty \tag{10.4.26}$$

with small sufficiently regularization parameter α *to the far-field equation* (10.3.47)*. Choose a cut-off constant* C*, then we call* G_j *positive if*

$$\|\phi_\alpha^j\|_{L^2(\partial G_j)} < C.$$

Here, for some simple settings, we can efficiently calculate ϕ_α^j *for many test domains. Let* G_0 *be the reference domain and* G_j *be constructed from* G_0 *by translation*

$$G_j := G_0 + x_j \tag{10.4.27}$$

with vector x_j*, then we get*

$$\phi_\alpha^j := \left(\alpha I + S_{G_0}^{\infty,*} S_{G_0}^\infty\right)^{-1} S_{G_0}^{\infty,*} M_j^* u^\infty \tag{10.4.28}$$

with $M_j^* = e^{ik\hat{x}\cdot x_j}$*. For the details, we refer to [72].*

- *Take the intersection of the closure of all positive test domains* $D_{rec} := \bigcap_{G_j \text{ positive}} \overline{G_j}$*, which is a subset of the closure of the unknown scatterer* D.

Though we cannot reconstruct the exact shape by this one-wave range test, it is possible to detect some special information which can reflect basic properties of the unknown scatterer.

Algorithm 10.4.11. *(Multi-wave range test)*

- *Choose a set of non-vibrating test domains G_j such that the far-field equation is solvable for all incident directions, which can be carried out in principle by using the one-wave range test.*
- *For each direction $-d$, approximately solve the far-field equation (10.3.47) by using Tikhonov regularization method, and then calculate $u^s(x,-d)$ in terms of (10.3.45).*
- *Obtain $\Phi(d,x)$ from $u^s(x,-d)$ by using mixed reciprocity relation.*
- *Apply the one-wave range test to the far-field equation corresponding to point sources (10.3.51), its numerical scheme is the same as that in Algorithm 10.4.10.*

In principle, with a proper choice of the set of test domains we can reconstruct the full shape of the scatterer, but in practice we do not have explicit approach to choose these test domains G_j. Furthermore, for some complex situations such as D is of strong non-convexity, the choice of test domains might lead to quite complicated procedures.

Algorithm 10.4.12. *(No response test)*

- *Choose a set of test domains G_j for j in some index set $\mathcal{J}$ and a sufficiently small parameter ϵ.*
- *For each test domain G_j, construct the functions $v_g^{j,l}$ with $g_l \in L^2(\mathbb{S})$ for l in some index set $\mathcal{L}$ such that $\|v_g^{j,l}\| \leq \epsilon$.*
  Here, $v_g^{j,l}$ can be constructed by solving the point source equation as that in the probe method or singular sources method.
- *For each $l \in \mathcal{L}$, calculate $I^j(g_l)$ by (10.3.52) and take the supremum of $I^j(g_l)$ over all $l \in \mathcal{L}$*
$$I_\epsilon^j(G_j) := \sup_{l\in\mathcal{L}} I^j(g_l). \tag{10.4.29}$$
- *Choose a cut-off constant C. If $|I_\epsilon^j(G_j)| \leq C$, we call G_j positive. Then calculate the intersection of all positive test domains*
$$D_{rec} := \bigcap_{G_j \text{ positive}} \bigcap_{j\in\mathcal{J}} G_j. \tag{10.4.30}$$

Generally, the set D_{rec} is only an approximation for the set of singular points of the scattered field u^s, which is a subset of the closure of D. In fact, we cannot expect more from the knowledge of far-field pattern for only one incident if no enough *a priori* information about the unknown scatterer is provided [69, 70].

In conclusion, the calculation of indicator function is crucial to the numerical realization of each inversion method. Of course, the choice of approximate domains or test domains is also the key to the accuracy of numerical reconstruction. In addition, we cannot ignore the effect of cut-off constant C on numerical results. In practice, if C is chosen too large, the reconstruction is too small to exist. If C is chosen to be smaller, then the reconstruction may become larger.

At last, we would like to mention that the choice of regularization parameter α is also an interesting topic since all inversion methods involve the regularization to ill-posed integral equations.

Acknowledgements This work was supported by the National Natural Science Foundation of China under grant No. 10771033.

References

1. I. Akduman and R. Kress, Direct and inverse scattering problems for inhomogeneous impedance cylinders of arbitrary shape, *Radio Science*, **38**(3), 1055, 2003.
2. T. Arens, Why linear sampling method works, *Inverse Problems*, **20**, 167–173, 2004.
3. T. Arensa and N. Grinberg, A complete factorization method for scattering by periodic surfaces, *Computing*, **75**, 111–132, 2005.
4. T. Arens and A. Kirsch, The factorization method in inverse scattering from periodic structures, *Inverse Problems*, **19**: 1195–1221, 2003.
5. F. Cakoni and D. Colton, *Qualitative Methods in Inverse Scattering Theory*, Berlin: Springer-Verlag, 2006.
6. F. Cakoni, D. Colton and P. Monk, The direct and inverse scattering problems for partially coated obstacles, *Inverse Problems*, **17**(6), 1997–2015, 2001.
7. F. Cakoni, D. Colton and P. Monk, The determination of boundary coefficients from far field measurements (to appear).
8. F. Cakoni, M. Fares and H. Haddar, Analysis of two linear sampling methods applied to electromagnetic imaging of buried objects, *Inverse Problems*, **22**(3), 845–867, 2006.
9. F. Cakoni, D. Colton and E. Darrigrand, The inverse electromagnetic scattering problem for screens, *Inverse Problems*, **19**(3), 627–642, 2003.
10. F. Cakoni and D. Colton, The linear sampling method for cracks, *Inverse Problems*, **19**(2), 279–295, 2003.
11. F. Cakoni, D. Colton and H. Haddar, The linear sampling method for anisotropic media, *J. Comput. Appl. Math.*, **146**, 285–299, 2002.
12. J. Cheng, J. J. Liu and G. Nakamura, Recovery of the shape of an obstacle and the boundary impedence from the far-field pattern, *J. Math. of Kyoto Univ.*, **43**(1), 165–186, 2003.
13. J. Cheng, J. J. Liu and G. Nakamura, The numerical realization of probe method for the inverse scattering problems from the near-field data, *Inverse Problems*, **21**(3), 839–856, 2005.
14. J. Cheng, J. J. Liu, G. Nakamura and S. Z. Wang, Recovery of multiple obstacles by probe method, *Quart. Appl. Math.*, **67**(2), 221–247, 2009.
15. M. Cheney, The linear sampling method and the MUSIC algorithm, *Inverse Problems*, **17**, 591–595, 2001.
16. D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, New York: John Wiley Sons, 1983.
17. D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Berlin: Springer-Verlag, 1992.
18. D. Colton, J. Coyle and P. Monk, Recent developments in inverse acoustic scattering theory, *SIAM Review*, **42**(3), 369–414, 2000.
19. D. Colton and A. Kirsch, A simple method for solving inverse scattering problems in the resonance region, *Inverse Problems*, **12**, 383–393, 1996.
20. D. Colton, M. Piana and R. Potthast, A simple method using Morozov's discrepancy principle for solving inverse scattering problems, *Inverse Problems*, **13**, 1477–1493, 1997.

21. D. Colton and P. Monk, A linear sampling method for the detection of leukemia using microwaves, *SIAM J. Appl. Math.*, **58**, 926–941, 1998.
22. D. Colton and R. Kress, Using fundamental solutions in inverse scattering, *Inverse Problems*, **22**, 49–66, 2006.
23. D. Colton and P. Monk, Target identification of coated objects, *IEEE Trans. Ant. Prop.*, **54**(4), 1232–1242, 2006.
24. K. Erhard and R. Potthast, A numerical study of the probe method, *SIAM J. Sci. Comput.*, **28**(5), 1597–1612, 2006.
25. M. Fotouhi and M. Hesaaraki, The singular sources method for an inverse problem with mixed boundary conditions, *J. Math. Anal. Appl.*, **306**, 122–135, 2005.
26. N. Grinberg and A. Kirsch, The factorization method for obstacles with a priori separated sound-soft and sound-hard parts, *Math. and Computers in Simulation*, **66**, 267–279, 2004.
27. N. Grinberg, Obstacle visualization via the factorization method for the mixed boundary value problem, *Inverse Problems*, **18**, 1687–1704, 2002.
28. N. Honda, G. Nakamura, R. Potthast and M. Sini, The no-response approach and its relation to non-iterative methods for the inverse scattering, *Annali di Matematica*, **187**, 7–37, 2008.
29. M. Ikehata, The probe method and its applications, *Inverse Problems and Related Topics, Research Notes Maths*, **419**, Chapman Hall/CRC, Boca, Rotan, FL, 2000.
30. M. Ikehata, Reconstruction of the shape of the inclusion by boundary measurements, *Comm. PDEs.*, **23**, 1459–1474, 1998.
31. M. Ikehata, Reconstruction of obstacle from boundary measurements, *Wave Motion*, **30**, 205–223, 1999.
32. M. Ikehata, Reconstruction of an obstacle from the scattering amplitude at a fixed frequency, *Inverse Problems*, **14**, 949–954, 1998.
33. M. Ikehata, A new formulation of the probe method and related problems, *Inverse Problems*, **21**, 413–426, 2005.
34. V. Isakov, On uniqueness in the inverse transmission scattering, *Comm. PDEs.*, **15**, 1565–1587, 1990.
35. A. Kirsch, Characterization of the shape of a scattering obstacle using the spectral data of the far field operator, *Inverse Problems*, **14**, 1489–1512, 1998.
36. A. Kirsch, Factorization of the far-field operator for the inhomogeneous medium case and an application in inverse scattering theory, *Inverse Problems*, **15**, 413–429, 1999.
37. A. Kirsch and R. Kress, Uniqueness in inverse obstacle scattering, *Inverse Problems*, **9**, 285–299, 1993.
38. A. Kirsch, The MUSIC algorithm and the factorization method in inverse scattering theory for inhomogeneous media, *Inverse Problems*, **18**, 1025–1040, 2002.
39. A. Kirsch, New characterizations of solutions in inverse scattering theory, *Appl. Anal.*, **76**, 319–350, 2000.
40. A. Kirsch, The factorization method for Maxwell's equations, *Inverse Problems*, **20**, S117–S134, 2004.
41. A. Kirsch, The domain derivative and two applications in inverse scattering theory, *Inverse Problems*, **9**, 81–96, 1993.
42. R. Kress, *Linear Integral Equations*, Berlin, Heidelberg, New York: Spinger-Verlag, 1989.
43. R. Kress and W. Rundell, Inverse scattering for shape and impedance, *Inverse Problems*, **17**, 1075–1085, 2001.
44. R. Kress and P. Serranho, A hybrid method for two-dimensional crack reconstruction, *Inverse Problems*, **21**, 773–784, 2005.
45. R. Kress and P. Serranho, A hybrid method for sound-hard obstacle reconstruction, *J. Comput. Appl. Math.*, **204**, 418–427, 2007.
46. R. Kress, Newton's method for inverse obstacle scattering meets the method of least squares, *Inverse Problems*, **19**, S91–S104, 2003.

47. R. Kress and L. Paivarinta, On the far field in obstacle scattering, *SIAM J. Appl. Math.*, **59**(4), 1413–1426, 1999.
48. R. Kress and W. Rundell, Nonlinear integral equations and the iterative solution for an inverse boundary value problem, *Inverse Problems*, **21**, 1207–1223, 2005.
49. J. J. Liu, J. Cheng and G. Nakamura, Reconstruction and uniqueness of an inverse scattering problem with impedance boundary, *Science in China, Ser. A*, **45**(11), 1408–1419, 2002.
50. J. J. Liu, Determination of Dirichlet-to-Neumann map for a mixed boundary problem, *Applied Mathematics and Computation*, **161**(3), 843–864, 2005.
51. J. J. Liu, G. Nakamura and R. Potthast, A new approach and error analysis for reconstructing the scattered wave by the point soruce method, *J. Comput. Math.*, **25**(3), 113–130, 2007.
52. J. J. Liu, G. Nakamura and M. Sini, Reconstruction of the shape and surface impedance for acoustic scattering data for an arbitrary cylinder, *SIAM J. Appl. Math.*, **67**(4), 1124–1146, 2007.
53. J. J. Liu and M. Sini, On the accuracy of reconstruction of the obstacles from exterior measurements (submitted).
54. J. J. Liu and M. Sini, On the accuracy of the numerical detection of complex obstacles from far-field data using the probe method, *SIAM J. Sci. Comput.*, **31**(4), 2665–2687, 2009.
55. J. J. Liu and R. Potthast, On the duality of potential method and point source method in inverse scattering problems, *J. Integral Equations. Appl.*, **21**(2), 297–316, 2009.
56. D. R. Luke and R. Potthast, The no response test — a sampling method for inverse scattering problems, *SIAM J. Appl. Math.*, **63**(4), 1292–1312, 2003.
57. G. Nakamura, R. Potthast and M. Sini, Unification of the probe and singular sources methods for the inverse boundary value problem by the no-response test, *Comm. PDEs.*, **31**, 1505–1528, 2006.
58. G. Nakamura and M. Sini, Obstacle and boundary determination from scattering data, *SIAM J. Math. Anal.*, **39**(3), 819–837, 2007.
59. G. Nakamura, R. Potthast and M. Sini, A comparative study between some non-iterative methods for the inverse scattering, in: Inverse problems, multi-scale analysis and effective medium theory, *Contemp. Math.*, **408**, 249–265, 2006.
60. J. C. Nedelec, *Acoustic and Electromagnetic Equations*, New York: Spinger-Verlag, 2001.
61. R. Potthast, A fast new method to solve inverse scattering problems, *Inverse Problems*, **12**, 731–742, 1996.
62. R. Potthast, Stability estimates and reconstructions in inverse acoustic scattering using singular sources, *J. Comput. Appl. Math.*, **114**(2), 247–274, 2000.
63. R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, London: Chapman & Hall/CRC, 2001.
64. R. Potthast, Sampling and probe methods — an algorithmical view, *Computing*, **75**, 215–235, 2005.
65. R. Potthast, A survey on sampling and probe methods for inverse problems, *Inverse Problems*, **22**, R1–R47, 2006.
66. R. Potthast, From the Kirsch-Kress potential method via the range test to the singular sources method, *J. Phys.*: Conference Series **12**, 116–127, 2005.
67. R. Potthast and I. Stratis, The singular sources method for an inverse transmission problem, *Computing*, **75**, 237–255, 2005.
68. R. Potthast, A new non-iterative singular sources method for the reconstruction of piecewise constant media, *Numer. Math.*, **98**, 703–730, 2004.
69. R. Potthast, On the convergence of the no response test, *SIAM J. Math. Anal.*, **38**(6), 1808–1824, 2007.
70. R. Potthast, A set-handling approach for the no response test and related methods, *Mathematics and Computers in Simulation*, **66**, 281–295, 2004.

71. R. Potthast, J. Sylvester and S. Kusiak, A 'range test' for determining scatterers with unknown physical properties, *Inverse Problems*, **19**, 533–547, 2003.
72. R. Potthast and J. Schulz, A multiwave range test for obstacle reconstruction with unknown physical properties, *J. Comput. Appl. Maths.*, **205**, 53–71, 2007.
73. R. Potthast, Fréchet differentiability of boundary integral operators in inverse acoustic scattering, *Inverse Problems*, **10**, 431–447, 1994.
74. P. Serranho, A hybrid method for inverse scattering for shape and impedance, *Inverse Problems*, **22**, 663–680, 2006.
75. E. Sincich, Stable determination of the surface impedance of an obstacle by far-field measurements, *SIAM J. Math. Anal.*, **38**(2), 434–451, 2006.
76. A. Tacchino, J. Coyle and M. Piana, Numerical validation of the linear sampling method, *Inverse Problems*, **18**, 511–527, 2002.
77. H. B. Wang and J. J. Liu, Numerical realization of probe method for multiple obstacles, *Mathematica Numerica Sinica*, **29**(2), 189–202, 2007 (in Chinese).
78. M. Yuan and J. J. Liu, Numerical realization of probe method for 2-D inverse scattering, *Mathematica Numerica Sinica*, **28**(2), 189–200, 2006 (in Chinese).
79. S. Z. Wang, *Numerical Computation for Inverse Problems*, Sapporo, Dissertation of Hokkaidao University, 2006.

Chapter 11
Inverse Problems of Molecular Spectra Data Processing

Gulnara Kuramshina

Abstract. There are considered ill-posed inverse problems which arise in molecular spectroscopy data processing. The main of them, the general inverse problem of structural chemistry for free molecules is nonlinear inverse problem of molecular vibrations. The modern view on this problem joins the study of molecular force field and geometry restoration using all available experimental data (vibrational (IR and Raman) spectroscopy, electron diffraction, etc.), *a priori* constraints and *ab initio* quantum mechanical calculations. The second inverse problem of determination of intermolecular potential from the second virial coefficient is also closely connected with molecular structure investigation.

On the basis of formulation and formalization of possible obvious (and related to quantum mechanical results) model assumptions concerning the character of force fields which are widely used in vibrational spectroscopy we have constructed a principle for choosing a unique solution from the set of solutions in the framework of Tikhonov's regularization theory. The solution is chosen as the nearest to the given matrix of force constants which satisfy all *a priori* assumptions concerning the model characteristics of the solution.

Also there are presented stable numerical methods, based on Tikhonov's regularization method for computing the force fields of polyatomic molecules from experimental data as well as some numerical illustrations using real data.

11.1 Introduction

Many mathematical problems of science, technology and engineering are formulated in the form of an operator equation of the first kind with operator and

Gulnara Kuramshina
Department of Physical Chemistry, Faculty of Chemistry, Lomonosov Moscow State University, Moscow 119991, Russia.
e-mail: kuramshi@phys.chem.msu.ru

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

approximately known right-hand side. In many cases such problems belong to ill-posed ones and to solve such problems it is necessary to apply the special numerical methods based on regularizing algorithms. The theory of solving linear ill-posed problems greatly advanced at present [1] is not applicable in the case of nonlinear ill-posed problems. The special consideration of numerical methods for solving nonlinear ill-posed problems within the Tikhonov theory of regularization is presented in [2] and includes different variational approaches based on Tikhonov's functional, generalized principles for the choice of regularization parameter and formulation of very general conditions for their implementation.

Below we consider ill-posed inverse problems which arise in molecular spectroscopy data processing. The main of them, the general inverse problem of structural chemistry for free molecules is nonlinear inverse problem of molecular vibrations. The modern view on this problem joins the study of molecular force field and geometry restoration using all available experimental data (vibrational (IR and Raman) spectroscopy, electron diffraction, etc.), *a priori* constraints and *ab initio* quantum mechanical calculations. The second inverse problem described below is also closely connected with molecular structure investigation — a problem of determination of intermolecular potential from the second virial coefficient.

Formulation and formalization of possible obvious (and related to quantum mechanical results) model assumptions concerning the character of force fields which are widely used in vibrational spectroscopy were presented in our publications [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. On the basis of this formalization we have constructed a principle for choosing a unique solution from the set of solutions in the framework of Tikhonov's regularization theory. The solution is chosen as the nearest to the given matrix of force constants which satisfy all *a priori* assumptions concerning the model characteristics of the solution.

We also have investigated stable numerical methods, based on Tikhonov's regularization method, for computing the force fields of polyatomic molecules from experimental data and we provide some numerical illustrations using real data.

11.2 Inverse vibrational problem

A number of inverse problems arise in the data processing of experimental data obtained by means of infrared and Raman spectroscopy. The most important is the so-called inverse vibrational problem of determining parameters of the molecular force field from given experimental data (vibrational frequencies, isotope frequency shifts, Coriolis constants, centrifugal distortion constants, etc.). The accumulation of data on molecular constants is necessary for prediction of spectra and other properties of compounds not yet investigated and for development of physical models in a theory of molecular structure.

The idea about the force field arises when molecule is considered as a mechanical system of nuclei while all the interactions due to the electrons are included in an effective potential function $U(q_1, q_2, \ldots, q_n)$, where $q_1, q_2, \ldots, q_n$ denote $n = 3N - 6$ generalized coordinates describing mutual positions of N atomic nuclei of the molecule. Together with the nuclear masses, this function determines the most important properties of a molecule. As is well known (see, e.g., [3]), the equilibrium configuration of the molecule satisfies the relation

$$\frac{\partial U}{\partial q} = 0,$$

and if coordinates $q_1, q_2, \ldots, q_n$ are determined so that $q_1 = q_2 = \ldots = q_n = 0$ in the equilibrium configuration, the following expansion is valid:

$$U(q_1, \ldots, q_n) = U_0 + \frac{1}{2} \sum_{i,j=1}^{n} f_{ij} q_i q_j + O(\|q\|^3), \tag{11.2.1}$$

where U_0 is a certain constant, and the force constants

$$f_{ij} = \frac{\partial^2 U}{\partial q_i \partial q_j}, \qquad i, j = 1, \ldots, n$$

in the point $q_1 = q_2 = \ldots = q_n = 0$ constitute a positive definite matrix F determining all molecular characteristics connected with small vibrations. Mathematically, the concept of the force field may be obtained through the adiabatic theory of perturbations with the use of a small parameter related to the ratio of electron mass to the mass of nuclei, and it can be shown that in a certain approximation the nuclei may be treated as particles moving in the force field determined by the potential energy function (11.2.1).

There are two main purposes for using a molecular force field: (a) to check validity of various model assumptions commonly used by spectroscopists for approximation of the potential function; (b) to predict the vibrational properties of certain molecules (including those not yet observed) using fundamental properties of the force field such as its isotopic invariance and the approximate transferability of specific force constants in a series of related compounds.

The spectral frequencies ω_i are the main type of experimental data on molecular vibrations. They are connected with the matrix of force constants F by the eigenvalue equation

$$GFL = L\Lambda, \tag{11.2.2}$$

where $\Lambda = \text{diag}\ \{\omega_1^2, \ldots, \omega_n^2\}$ is a diagonal matrix consisting of the squares of the molecular normal vibrational frequencies, and G is the kinetic energy matrix in the momentum representation which depends only on nuclear masses and the equilibrium configuration (assumed to be known within specified limits of error). The matrix L characterizes the form of each normal vibration, i.e., the relative amplitudes of vibrations in terms of classical mechanics.

From (11.2.2), it is evident that (except for diatomic molecules) the $n(n+1)/2$ parameters of F cannot be found from n frequencies $\omega_1, \ldots \omega_n$ in the unique way. This has led both to attempts to use certain model assumptions concerning the structure of the matrix F, and to applying additional experimental data. Within the harmonic model, the molecular force field is independent of the nuclear masses, and hence in a case of m molecular isotopic species spectra we have, instead of (11.2.2), the next system

$$(G_i F)L_i = L_i \Lambda_i, \tag{11.2.3}$$

where the subscript $i = 1, 2, \ldots, m$ indicates the isotopomers. Usually, the introduction of isotopomers leads to a limited number of independent equations in system (11.2.3), thus leaving the inverse problem underdetermined. Important additional information on the molecular force is provided by Coriolis constants ζ which characterize the vibrational-rotational interaction in the molecule. They are connected with matrix F in terms of the eigenvectors L of the problem (11.2.2):

$$\zeta = \frac{1}{M^2} L^* A \tilde{M} A^* L, \tag{11.2.4}$$

where ζ is a matrix with vector elements consisting of the Coriolis constants, $\tilde{M}$ is a diagonal matrix consisting of the nuclear masses, M is the sum of nuclear masses of the molecule, and A is a matrix connecting the Cartesian displacements of atoms with coordinates q characterizing the equilibrium configuration of the molecule. In a similar manner we can write the dependencies of other measured values on the matrix F, such as the mean-square amplitudes of the vibrations (obtained from gas-phase electron diffraction [12]) which may be calculated from the eigenvalues and eigenvectors of (11.2.2).

Recent rapid progress in the extending of quantum mechanical calculations of vibrational spectra and harmonic force fields for moderate size molecules with inclusion of electron correlation at MP2 and DFT levels [15, 16] provides fundamental new possibilities for more accurate interpretation of experimental data, development of theoretical models of molecular structure and dynamics, and determination of molecular force fields based on the joint treatment of theoretical and experimental data. *Ab initio* quantum-mechanical calculations ensure obtaining information that has a clear physical meaning. When performed at a high theoretical level (taking account of electron correlation, etc.), such calculations are capable of satisfactorily reproducing patterns of experimental structural and vibrational spectroscopy data. However, in case of large molecular systems the best *ab initio* results significantly differ from experimental values and one has to use special empirical corrections [17].

11.3 The mathematical formulation of the inverse vibrational problem

Let us consider (11.2.2)–(11.2.4), or some part of them, depending on the available experimental data, as a single operator equation

$$AF = \Lambda, \tag{11.3.1}$$

where the nonlinear operator A maps the real symmetrical matrix F to the set of eigenvalues of (11.2.2) (or (11.2.3)), the Coriolis constants, ζ, (11.2.4), the mean square amplitudes, etc. This set of data may be represented as a vector in the finite-dimensional space $\mathbb{R}^l$, where l is a number of known experimental data. The matrix F is considered as a vector in the finite-dimensional space Z, consisting either of the elements of matrix F or of quantities of its parametrization. Note that (11.3.1) in general does not satisfy [3] any of the conditions of the well-posedness of the problem considered above.

1. *Solvability.*
   It is easy to see that system (11.2.3) (determined for different molecular isotopomers) is compatible only when the condition

   $$\det G_i / \det \Lambda_i = \text{const}, \qquad i = 1, 2, \ldots, m$$

   is satisfied (m is the number of isotopomers). This condition can be violated due to errors in experimental data Λ_i, and in experimental geometry used in G_i, or due to anharmonicity of the real vibrations ignored by the operator of (11.3.1). In these cases within harmonic model there is no matrix F reproducing the frequencies of all isotopomers within the limits of experimental errors.
2. *The uniqueness of the solution of the problem.*
   If we know only the vibrational frequencies of one isotopomer of the molecule, (11.3.1) reduces to the inverse eigenvalue problem (11.2.2); hence, when G is not singular, it follows that as solution of (11.3.1) we may take any matrix of the form

   $$F = G^{-1/2} C^* \Lambda C G^{-1/2}, \tag{11.3.2}$$

   where C is an arbitrary orthogonal matrix. To choose a definite solution it is necessary to use additional information or to take account of some model assumptions.
3. *The stability of the solution with respect to the perturbations of* Λ *and* A.
   An example of such instability may be easily constructed for a system of the form (11.2.3) in a case of molecular isotopomers.

Therefore, all three conditions of well-posedness are generally not satisfied for the considered problem and the inverse vibrational problem is an extreme example of an ill-posed problem. The unstability and non-uniqueness of the solution can lead to significant differences in the force field parameters of the same

molecule obtained by different methods and can lead to difficulties in comparing and transferring force fields in series of related compounds.

To elucidate any arbitrariness in the calculated force constants it is necessary to use stable solutions of the inverse vibrational problem which have some specific properties. In practical terms, chemists often employ certain model assumptions arising from the classic theory of chemical structure involving such concepts as monotone changes of the physico-chemical properties in a series of related molecules and the preservation of the properties of separate molecular fragments in various compounds (taking account of the nearest surrounding), all related to the general concept of transferability of the molecular force constants. All the necessary model assumptions may be taken into account by the choice of some given *a priori* matrix F^0 (see below): e.g. some off-diagonal elements of the matrix F may be taken to be equal to zero (the so-called valence force field) reflecting the assumption of insignificance of some intramolecular interactions, and/or it is possible to introduce some in-pair equalities of force constants for similar fragments, some elements of F may be taken from preliminary calculations, etc. Therefore, the inverse vibrational problem can be formulated in the following way [3, 4, 5, 6, 7, 8].

In the set of force constant matrices satisfying (11.3.1), we search for the matrix which is nearest in some given metric (normal) to *a priori* given matrix F^0 (the so-called *normal solution*). In the case of an inconsistent problem (within harmonic approximation, this may happen in the case of joint treatment of isotopomers spectra, or when we include an additional experimental data) it is possible to find the matrix F for which the distance from AF to Λ is minimal, i.e., to find the so-called *pseudosolution* of the problem. When the pseudosolution is also non-unique, we must proceed as in the case of non-uniqueness of the solution — i.e., from all possible pseudosolutions to choose the one which is nearest to the given force field matrix F^0 (the normal pseudosolution).

The general formulation of the inverse problem can be given in another way. Suppose, we are given (11.3.1) and the operator A put into correspondence to any symmetric, and positive definite matrix F, the set of experimental data (squares of molecular vibration frequencies and its isotopomers), and known molecular constants such as mean vibrational amplitudes, Coriolis constants, etc. The dimension of vector Λ is determined by the number of experimental data. Since the symmetric matrix F is determined by $n(n+1)/2$ elements we can consider the unknown force constants as a vector of dimension $n(n+1)/2$. Then the operator A acts from the Euclidean space $\mathbb{R}^{n(n+1)/2}$ into the Euclidean space $\mathbb{R}^l$. In these spaces we introduce the following norms:

$$\|F\| = \left(\sum_{i \le j=1}^{n} f_{ij}^2 \right)^{1/2}, \qquad \|\Lambda\| = \left(\sum_{k=1}^{l} \lambda_k^2 \rho_k \right)^{1/2},$$

where $\rho_k > 0$ are some positive weights; f_{ij} are the elements of matrix F; λ_k $(k = 1, \ldots, l)$ are the components of Λ.

The operator A is continuous for all considering problems. However, (11.3.1) may have non-unique solution, or no solution at all, due to the anharmonicity of vibrational frequencies. Suppose, we are given the matrix F^0 (vector of dimension $n(n+1)/2$). It is necessary to find the normal pseudosolution of (11.3.1), that is:

- find an element $F = \bar{F}_n$ for which $\min \|F - F^0\|$ is reached provided that $\|AF - \Lambda\| = \mu$, where $\mu = \inf_{F \in D} \|AF - \Lambda\|$ and D is the closed set of *a priori* constraints on the values of the force constants. If no constraints are imposed, then $D = \mathbb{R}^{n(n+1)/2}$.

The element F^0 should be specified from *a priori* considerations of the possible solutions including both the approximate quantum mechanical calculations and other ideas (for example, the transferability of the force constants among similar fragments in a series of related compounds).

Let us denote the solution (vector) to be found as $\bar{F}_n$, if this vector is non-unique the set of such vectors is denoted as $\{\bar{F}_n\}$. If (11.3.1) is solvable then $\mu = 0$, and it remains finding the solution of (11.3.1) which is the nearest to the given vector F^0. But we do not know the exact form of either the vector Λ or operator A (the matrix G or matrices G_i). We have only the vector Λ_δ, determined from experimental data such that $\|\Lambda_\delta - \Lambda\| \leq \delta$ (where $\delta > 0$ is the experimental error) and the operator A_h which approximates the operator A; $h \geq 0$ is a parameter characterizing the proximity of A_h to A. The uncertainty in operator A is related to errors of determination of the matrix G (or G_i) (errors of experimental data on the geometrical parameters). Therefore, we do not know exact forms of A and Λ, but only their approximations A_h and Λ_δ and errors (h, δ). It is necessary to find the vector $F_{h,\delta}$ approximating the exact solution $\bar{F}_n$. The difficulties in solving this problem are related to its ill-posed character.

The modern theory of solving ill-posed problems was founded by A. N. Tikhonov [1, 2] and developed by many authors. The main idea when designing stable methods for solving ill-posed problems is that ill-posed problems are generally underdetermined. To solve such problems it is necessary to use additional information and formulate the criteria for choosing approximate solutions. If such criteria are formulated and mathematically formalized, we can construct stable methods of solving ill-posed problems — the so-called *regularizing algorithms*.

11.4 Regularizing algorithms for solving the inverse vibrational problem

The inverse vibrational problem under investigation is nonlinear. Let us consider again (11.3.1) in the standard form

$$AF = \Lambda, \qquad F \in \mathbb{R}^{n(n+1)/2}, \quad \Lambda \in \mathbb{R}^l. \tag{11.4.1}$$

The existence of the normal (relatively to *a priori* given matrix F^0) pseudosolution $\bar{F}_n$ of the exact problem (11.4.1) may be guaranteed if the operator A includes the operator of the direct vibrational problem for a single molecule. The uniqueness of $\bar{F}_n$ cannot be guaranteed.

How can the error of the approximate operator A_h be estimated? The estimation $\|A - A_h\| \leq h$ is impossible because the nonlinear operators have no norm. It is obvious that for the operator of the inverse vibrational problem this error is related to errors of the G matrix. It is possible to obtain an estimate in the form [3, 6]

$$\|AF - A_hF\|\psi(h, F),$$

where ψ is a known continuous function, which approaches 0 as the errors of the equilibrium geometry of the molecule decrease to zero. Particularly, the estimates may be obtained in the form

$$\psi(h, F) = \psi(h, \|F\|), \qquad \psi(h, F) = h\|A_hF\|.$$

The error on the right-hand side of (11.4.1) is determined previously.

We arrive at the following formulation of the inverse problem.

Problem I. Suppose we are given (11.3.1) where $F \in D \subseteq Z$, $\Lambda \in U$, Z and U are finite-dimensional spaces, D is a closed set of *a priori* constraints of the problem, and A is a nonlinear operator continuous in D. It is required to find an approximate solution of (11.4.1) when instead of A and Λ, we are given their approximations A_h and Λ_δ such that $\|\Lambda - \Lambda_\delta\| \leq \delta$, $\|AF - A_hF\| \leq \psi(h, F)$ for $F \in D$; here $\psi(h, F)$ is a known continuous function, which approaches zero as $h \to 0$ uniformly for all $F \in D \cap \bar{S}(0, R)$, where $\bar{S}(0, R)$ is a closed ball with center at $F = 0$ and an arbitrary radius R. The error in specifying the operator A involves an error of determining the molecular equilibrium configuration from experiment. Note that Problem I does not satisfy any conditions of correctness.

We shall consider now the problem of constructing a normal pseudosolution of Problem I with exact right-hand side and operator.

Problem II. It is required to obtain

$$\bar{F}_n = \arg\min \|F - F^0\|, \qquad F : F \in D, \quad \|AF - \Lambda\| = \mu,$$

where $\mu = \inf \|AF - \Lambda\|$, $F \in D$.

The element $F^0 \in Z$ should be specified from *a priori* requirements on the solution, using both approximate quantum mechanical calculations and other ideas (for example, the transferability of the force constants for similar fragments of molecules).

In the case when a unique solution of (11.4.1) exists, it is clear that its normal pseudosolution is identical with the solution itself. Taking all the abovementioned into account we can formulate the following problem.

Problem III. Suppose we are given (11.4.1); it is required from the approximate data $\{A_h, \Lambda_\delta, h, \delta\}$ to obtain approximations $F_\eta \in D$ to the solution $\bar{F}_n$ of Problem II such that

$$F_\eta \to \bar{F}_n \quad \text{as} \quad \eta \to 0$$

i.e., the algorithm for finding F should be Tikhonov regularizing.

Now we shall consider the simplest formulation of Problem I.

Problem I′. The vibrational spectrum of a single molecule is known, and in (11.4.1) the operator A corresponds to the vector $F \in \mathbb{R}^{n(n+1)/2}$ which is made up of the elements of the symmetric matrix F of order n, the ordered set of eigenvalues of the matrix GF. We shall use the ordered set of squares of the molecular vibrational frequencies as the right-hand side $\Lambda \in \mathbb{R}^l$.

Problem II′. It is required to find the normal solution

$$\bar{F}_n = \arg\min \|F - F^0\|, \qquad F \in \{F : AF = \Lambda\}.$$

Problem I is always solvable, and, furthermore, solutions are nonunique (apart from the case when $n = 1$).

Since the operator A in (11.4.1) is completely defined by the specification of the matrix G, we shall estimate the deviation of the approximately specified operator A_h (corresponding to certain G_ξ) from the exact operator A (corresponding to G) by the error in specifying matrix G. We suppose that in a certain matrix norm $\|G - G_\xi\| \le \xi$.

In the space $\mathbb{R}^l$ of the right-hand sides we shall introduce the Euclidean norm with positive weights, while in $\mathbb{R}^{n(n+1)/2}$ we use the Euclidean norm. Suppose that instead of the accurate value of the right-hand side Λ we specify Λ_δ, such that $\|\Lambda - \Lambda_\delta\| \le \delta$.

The following theorems on the stability of Problems I′ and II′ hold [3, 6].

Theorem 11.4.1. *Problem I′ is stable in the Hausdorff metric with respect to the perturbations of the operator and the right-hand side.*

Here the Hausdorf distance (metrics), $\rho(A, B)$, is determined in the following way: for any closed set A and B in normed space

$$\rho(A, B) = \sup_{x \in A} \inf_{y \in B} \|x - y\| + \sup_{y \in A} \inf_{x \in B} \|x - y\|.$$

Theorem 11.4.2. *If Problem II′ has a unique solution, then it is stable to perturbations of the operator and the right-hand side.*

The proofs of these theorems are given in [3, 6].

The least-squares method which consists of minimizing $\|A_h F - \Lambda_\delta\|^2$ on a set of *a priori* constraints is the one most often encountered in data processing of molecular spectra. However, in the view of the ill-posed nature of the problem, this method cannot be applied directly to solving problems with approximate data. The method must be regularized by taking account of the possible inconsistency of the problem and also the non-uniqueness of its solutions. If we attempt to find the normal pseudosolutions (with respect to a certain F^0) we arrive at the formulation of the problem described above.

For finding a normal pseudosolution we have proposed a modification of the generalized discrepancy principle based on the possibility to estimate the error

of the operator in the form

$$\|AF - A_hF\| \le h\|A_hF\|, \qquad h < 1, \tag{11.4.2}$$

which corresponds to specification of the relative error AF and is a more convenient estimate for the problem considered than the monotone function ψ.

Suppose F_η^α is an extremum (possibly non-unique) of Tikhonov's functional

$$M^\alpha[F] = \|A_hF - \Lambda_\delta\|^2 + \alpha\|F - F^0\|^2 \tag{11.4.3}$$

in the set D. The existence of an extreme can be proved [7]. We shall introduce the function

$$\rho_\eta(\alpha) = \|A_hF_\eta^\alpha - \Lambda_\delta\| - \frac{1}{1-h}\left[\hat{\mu}_\eta + k(\delta + h\|\Lambda_\delta\|)\right],$$

where $k > 1$ is a constant and

$$\hat{\mu}_\eta = \inf_{F\in D}\left\{\|A_hF - \Lambda_\delta\| + \delta + h\|A_hF\|\right\}.$$

If the condition

$$\|A_hF^0 - \Lambda_\delta\| > \frac{1}{1-h}\left[\hat{\mu}_\eta + k(\delta + h\|\Lambda_\delta\|)\right] \tag{11.4.4}$$

is satisfied, then the equation $\rho_\eta(\alpha) = 0$ has a generalized solution $\alpha_\eta > 0$ (i.e., α_η is such that $\rho_\eta(\alpha) > 0$ when $\alpha > \alpha_\eta$; $\rho_\eta(\alpha) < 0$ when $\alpha < \alpha_\eta$). If α_η is a point of continuity of $\rho_\eta(\alpha)$, we have $\rho_\eta(\alpha_\eta) = 0$. This assertion follows from the monotonicity of $\rho_\eta(\alpha)$ and the limit relations (as $\alpha \to 0$ and $\alpha \to +\infty$).

We shall formulate the algorithm for finding approximations to the normal pseudosolutions of (11.4.1). If condition (11.4.4) is not satisfied, we take $F_\eta = F^0$ as an approximate solution; in the contrary case we find $\alpha_\eta > 0$ (the generalized solution of the equation $\rho_\eta(\alpha) = 0$), and assume $F_\eta = F^{\alpha_\eta}$. If the extreme of the functional (11.4.3) is non-unique, we choose the one for which

$$\|A_hF^0 - \Lambda_\delta\| \le \frac{1}{1-h}\left[\hat{\mu}_\eta + k(\delta + h\|\Lambda_\delta\|)\right]$$

(the possibility of making such a choice is shown in [6]).

Theorem 11.4.3. *The algorithm formulated above is Tikhonov regularizing.*

The proof is given in [6].

In a case when the estimate of the error of the operator cannot be written in the form (11.4.2), but the requirements of Problem I are satisfied we can use the following version of the generalized discrepancy method.

Problem IV. It is required to obtain $\inf\|F - F^0\|$,

$$F \in \mathbf{F}_\eta = \left\{F : F \in D, \|A_hF - \Lambda_\delta\| \le \delta + \psi(h, F) + \hat{\mu}(A_h, \Lambda_\delta)\right\},$$

where the estimate of the measure of incompatibility of the exact problem from above is

$$\hat{\mu}_\eta = \inf_{F \in D} \{\|A_h F - \Lambda_\delta\| + \psi(h, F) + \delta\}.$$

Lemma 11.4.4. *Suppose the conditions of Problem I are satisfied. Then* $\hat{\mu}_\eta \leq \mu$, *and* $\hat{\mu}_\eta \to \mu = \inf_{F \in D} \|AF - \Lambda\|$ *as* $\eta \to 0$.

Lemma 11.4.5. *Problem IV is solvable for any* $\Lambda_\delta \in U$ *such that* $\|\Lambda - \Lambda_\delta\| \leq \delta$ *and for a continuous operator* A *such that* $\|A_h F - AF\| < \psi(h, F)$.

The proof of this lemma relies on the fact that for all $\eta > 0$, the set $\mathbf{F}_\eta$ is nonempty (since $\bar{F}_n \in \mathbf{F}_\eta$), closed and bounded.

Theorem 11.4.6. *The algorithm defined by the extremal Problem IV is Tikhonov regularizing for Problem I.*

Details of proofs, estimates of the error of the operator and some details of the numerical realization of the methods are given in [3, 6].

11.5 Model of scaled molecular force field

In out strategy we can include results of quantum mechanical calculations in the statement of inverse problem and search for matrix F which is the nearest by the chosen Euclidean norm to the given *ab initio* F^0, the optimized solution is referred to as *Regularized Quantum Mechanical Force Field* (RQMFF) [3, 8]. In that statement we can use the Pulay model of scaled matrix [17] which is based on the representation of the force constant matrix as

$$F = BF^0B, \tag{11.5.1}$$

where B is a diagonal matrix of scale factors, and F^0 is an *a priori* given matrix. Such parameterization does not completely remove the ambiguity of the solution; however, this ambiguity may be resolved using ideas of regularization by searching for the scale matrix B closest to unit matrix or by searching for the matrix F closest to F^0, as shown in our previous studies [3, 13].

In the course of spectroscopic and structural research, introduction of the complete system of internal coordinates $\{q_i\}$ is the most tedious and time-consuming procedure, especially for the large molecular systems. From quantum chemistry we can obtain force constant matrix in Cartesian coordinates. The scaling procedure itself is formally given by Eq. (11.5.1); however, as we will see later, matrix B cannot be assumed diagonal any more. The main peculiar feature of the force matrices in Cartesian coordinates is in fact that they are not automatically independent of the molecular position and orientation. Physically meaningful force constant matrix should therefore satisfy a number of constraints.

Let the force matrix be represented as an array of 3 by 3 submatrices corresponding to each atom:

$$F = \begin{pmatrix} f_{(11)} & f_{(12)} & \cdots & f_{(1N)} \\ f_{(21)} & f_{(22)} & \cdots & f_{(2N)} \\ \vdots & \vdots & \ddots & \vdots \\ f_{(N1)} & f_{(N2)} & \cdots & f_{(NN)} \end{pmatrix},$$

where N is number of atoms in a molecule. Then the constraints are as follows [3]:

$$\sum_{i=1}^{N} f_{(ij)} = 0, \quad \sum_{i=1}^{N} V_i f_{(ij)} = 0, \qquad j = 1, 2, \ldots, N, \tag{11.5.2}$$

where

$$V_i = \begin{pmatrix} 0 & -R^0_{iz} & R^0_{iy} \\ R^0_{iz} & 0 & -R^0_{ix} \\ -R^0_{iy} & R^0_{ix} & 0 \end{pmatrix},$$

and $R^0_{ix}, R^0_{iy}, R^0_{iz}$ are Cartesian components of the i-th atom equilibrium position.

Equations (11.5.2) ensure that the force field represented by the matrix F does not contain terms related to displacement and rotation of a molecule as a whole. Imposing these constraints reduces the rank of matrix F to $3N - 6$ (or $3N - 5$ for linear molecules), thus leaving only vibrational degrees of freedom.

When the scaling procedure (11.5.1) is applied to the matrix F in Cartesian coordinates, we may assume [14] that *a priori* matrix F^0 satisfies the requirements (11.5.2). However, this does not necessarily mean that the scaled matrix also satisfies these requirements. To ensure that scaled matrix also contains only vibrational degrees of freedom, the scale matrix B should satisfy certain conditions. It can be shown that these conditions are as follows:

1. Matrix B should consist of the 3×3 unit submatrices E multiplied by some factors:

$$B = \begin{pmatrix} \beta_{11}E & \beta_{12}E & \ldots & \beta_{1N}E \\ \beta_{21}E & \beta_{22}E & \ldots & \beta_{2N}E \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{N1}E & \beta_{N2}E & \ldots & \beta_{NN}E \end{pmatrix}.$$

2. The factors β_{ij} $(i, j = 1, \ldots, N)$ should satisfy the conditions

$$\beta_{ij} = \beta_{ji};$$

$$\sum_{i=1}^{N} \beta_{1i} = \sum_{i=1}^{N} \beta_{2i} = \ldots = \sum_{i=1}^{N} \beta_{Ni} = S = \text{const}. \tag{11.5.3}$$

These conditions ensure that scaled matrix F will also correspond to only vibrational motion of a molecule. If there exist any extra constraints due to the symmetry or model assumptions, they should be used in addition to these constraints. In general, matrix B depends on $N(N-1)/2+1$ independent parameters, since all diagonal elements may be represented as

$$\beta_{ii} = S - \sum_{j \neq i} \beta_{ij}.$$

On this way we come to the formulation of inverse vibrational problem in a form (11.3.1) where a set of *a priori* constraints D on the molecular force field includes conditions (11.5.1) and (11.5.2).

This approach was tested for the series of molecules [14] and demonstrated very attractive advantages especially for the correction of molecular force fields of large scale molecules.

11.6 General inverse problem of structural chemistry

We consider the general problem of extracting internal molecular parameters (such as geometrical structure and properties of the intramolecular force field) from the available experimental data on infrared spectroscopy, electron diffraction analysis, microwave spectroscopy, etc., guided by the *ab initio* calculations. An implementation of such generalized approach is hindered by many well-known problems. Analysis of different experimental data is often carried out using different molecular models. For example, *ab initio* calculations and most spectroscopic studies use equilibrium geometry data, while the ED studies directly provide only thermally averaged values. Hence the results obtained for the same molecule using different experimental techniques may prove incompatible. Spectroscopic data is often insufficient to restore the complete force field, thus making it necessary to introduce model assumptions restricting the molecular force fields. As for electron diffraction data they are also often insufficient to determine all structural parameters, especially when a molecule possesses a set of similar interatomic distances; this also implies necessity of introducing external constraints on molecular geometry. Quantum mechanical calculations often lack accuracy to match the experimentally measured values. For example, an approach of scaling an *ab initio* force field is widely used to achieve a reasonable agreement between calculated and measured frequencies. Data on vibrations anharmonicity cannot be readily obtained from experimental data except for very small and simple molecules. This leads to different model evaluations (or to the usage of *ab initio* values) that may lack necessary precision.

We suggest the next scheme for the analysis of the molecules with relatively small atomic excursions taking account of anharmonicity and successfully deal with curvilinear motions which are referred to as a "small amplitude" approximation, to distinguish between this scheme and a more elaborate one that is capable of dealing with bending motions and internal rotation with fairly large amplitudes [12, 18].

A common molecular model is created that connects molecular parameters and experimentally measured values. Within this model, the molecular parameters to be defined are equilibrium geometry and force field parameters. All experimen-

tally measured values are calculated using the same set of parameters and the parameters are adjusted so as to fit the experimental evidence. It is important that all values should be determined from the same set of parameters. For example, the ED analysis will therefore obtain amplitudes compatible with spectroscopic evidence, and spectroscopy analysis will obtain frequencies compatible with the equilibrium geometry provided by the ED and microwave data.

Every time an experimental data is insufficient for the unique determination of some or all molecular parameters, we should employ some kind of external knowledge or experience. In accordance with the basics of regularization theory, we suggest to choose the solution that is in a certain sense nearest to some *a priori* chosen parameter set. This set may not necessarily conform to the experiment, but should be based on data complementary to the experiment. The external evidence may be derived from some general ideas (for example, molecular force field models, or data on similar molecular structures), or, preferably, be based on *ab initio* data. Within this approach, the results will tend to be as close to quantum-mechanical data as the experiment allows. From mathematical point of view, the algorithm should provide approximations to the solution that tend to an exact solution when the experimental data becomes more extensive and accurate.

These "soft" constraints may be combined with more rigid set of constraints implied on the solution to obtain the unique solution. For example, when there are close interatomic distances in a molecule, it is a common practice to determine only one of them from an ED experiment, fixing all differences between the distances at *ab initio* values. This may be called a "rigid" approach. Within a "soft" approach, it is possible to find a solution that will have the same properties unless it does not contradict experiment (and if it does, we shall find a solution which is the nearest to the one with that properties).

For simplicity we shall now limit the set of experimental data available by the vibrational frequencies (ω) and normalized electron scattering intensity (M). In a general case there may be more experimental data (rotational constants, etc.). All experimentally measured values depend on both equilibrium geometry (R) and force field (F) parameters. Harmonic vibrational frequencies can be obtained from the force constant matrix (the matrix of the second derivatives of the molecular potential at the point of equilibrium). To calculate anharmonic corrections to frequencies, we need cubic and quartic terms in the potential expansion. In many cases, however, anharmonic effects are relatively small and may be neglected during a normal coordinate analysis. The ED intensity shows only a moderate dependence on force field, while geometry changes are of much greater importance. However, the ED experiment is directly related to the thermally averaged interatomic distances rather than to their equilibrium values, and difference between these two kinds of distances depends on the anharmonic terms of the potential. Hence, calculation of ED intensity requires knowledge of the harmonic and anharmonic terms of the molecular potential (at least cubic terms).

Before formulating a problem of determining the molecular parameters, it makes sense to analyze what amount of data we are capable to find from a given

set of measured data. Indeed, if we introduce a very limited set of parameters (and thus create a very "rigid" model), we are likely to fail achieving a good fit between experimental and calculated data. On the other hand, if a model is very flexible (that is, contains too many adjustable parameters), we are likely to find a wide variety of solutions that all satisfy the experiment. Even if we employ the concept of a regularized solution, there must exist some kind of optimal parameter set that would correspond to the available experimental data. As for the force field determination, it is a common knowledge that (except for a limited set of small or very symmetrical molecules) we never have enough data to restore a complete force field. The ED data usually provides only a small additional data on force field, so as a rule we are in the situation when there exists a wide range of force fields compatible with spectroscopic experiment. Among the ways to reduce the ambiguity of the force fields, we could mention the following:

1. Introducing model assumptions based on general ideas of molecular structure (e.g. valence force field, etc.): these will result in neglecting some force constants, fixing the others, and/or introducing model potentials that would be allowed to generate force matrix depending on a small number of parameters.
2. Transferring some force field parameters from similar fragments in related molecules and assuming they are not likely to be significantly changed in a different environment.
3. Applying scale factors technique when all allowed force matrices to be obtained from *ab initio* values by the certain scaling procedure. Here these factors may be treated as force field parameters to be determined.

All of these ways may be formulated as the "soft" restrictions. Instead of introducing them, we may generate a force field F^0 that possesses the above properties and attempt to find solutions nearest to F^0. For the approach based on equilibrium configurations, the force field anharmonicity is of great importance because it defines difference between the equilibrium distances and thermally averaged ones used in electron diffraction analysis. Taking the mentioned limitations into account, we come to the following conclusions. Force field parameters — at most — should include a matrix of the quadratic terms in potential expansion. This set should be even more constrained by using *ab initio* data (as a stabilizer matrix F^0 or with the use of scaling scheme). Cubic (and quartic) terms cannot be determined from the experimental data under discussion and should be somehow evaluated. It is possible to introduce some anharmonicity models or — preferably — use *ab initio* data. To maintain consistence with quadratic potential, these terms may require adjustment when quadratic terms are changed during the fitting procedure.

Similar problems exist for the interatomic distances determined from the ED data. Though there are many successful results provided by this technique, it's evident that the accuracy of the obtained data may be insufficient in cases when a molecule possesses a number of bond lengths that are different in chemical nature but close to each other in their values. Here, again, we need to introduce model assumptions that, at best, are based on the *ab initio* calculations. Under

certain unfavorable conditions, the ED data may be insufficient even to define symmetry of the equilibrium configuration, which in this case should be obtained from alternative sources.

Taking all this into account, we come to the following formulation of the inverse problem. Let Λ be a set of available experimental data, and $A(R, F)$ be a procedure allowing of calculating this data from the set of molecular parameters R and F. We may suppose that Λ is a finite-dimensional vector from the normalized space $\mathbb{R}^m$; parameters (R, F) may also be chosen so as to constitute a vector from $\mathbb{R}^n$, then A is an operator acting from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Let (R^0, F^0) be an *a priori* given set of parameters (e.g. obtained from *ab initio* calculations), and we know that accuracy of experimental data Λ is such that it deviates from the "ideal" data not more than by a given constant δ. Let us also introduce a set of constraints D in $\mathbb{R}^n$ to which our solution should belong. We need to find an approximation $(R, F)_\delta$ to the exact parameters (R, F) such that:

1. the solution is compatible with the experimental data within the accuracy range $(R, F)_\delta \in Z_\delta$, where

$$Z_\delta = \{(R, F) \in D : \|A(R, F) - \Lambda_\delta\| \leq \delta\};$$

2. among all possible solutions, we choose the one most close to (R^0, F^0)

$$(R, F)_\delta = \arg\min_{(R,F)\in Z_\delta} \left\|(R, F) - (R^0, F^0)\right\|;$$

3. when accuracy increases, we get more accurate approximations to an exact solution $(R, F)_\delta \to (R, F)$ when $\delta \to 0$.

One of the possible implementation of the procedure is to obtain such approximations based on Tikhonov functional technique when we minimize

$$M^\alpha(R, F) = \|A(R, F) - \Lambda_\delta\|^2 + \alpha \left\|(R, F) - (R^0, F^0)\right\|^2 \tag{11.6.1}$$

on the set D, and regularization parameter α is chosen as a solution of the equation

$$\|A(R, F)_\alpha - \Lambda_\delta\| = \delta,$$

where $(R, F)_\alpha$ delivers a minimum to $M^\alpha(R, F)$.

Under certain conditions on the operator $A(R, F)$ the existence and uniqueness of the solution can be guaranteed. Eq. (11.6.1) shows that within this approach the deviations of solution from an *a priori* given set may be treated as a penalty. Obviously, when (R^0, F^0) itself is compatible with experimental data, no further adjustment is necessary. The formulation given above is very general; in the implementation that follows we shall assume that R is a set of independent equilibrium geometry parameters and F is a set of harmonic force constants (or the Pulay scale factors).

The presented approach is aimed at a simultaneous determination of the geometry and force field parameters of a molecule. It combines techniques previously

used in IR spectroscopy and ED data analysis. Particularly, it allows using more flexible force field models when fitting ED data, far beyond the usually employed scaling of the *ab initio* force field.

11.7 Intermolecular potential

The mechanical molecular model considers atoms as spheres and bonds as springs. Non-bonded atoms (greater than two bonds apart) interact through van der Waals attraction, steric repulsion, and electrostatic attraction/repulsion. These properties could be described mathematically when atoms are considered as spheres of characteristic radii. For a given spherical potential energy function $U(R)$ the virial expansion up to the second power of density is written as

$$\frac{p}{kT} = \rho + B(T)\rho^2,$$

where $B(T)$ is the second virial coefficient. It can be obtained using interaction potential function as

$$B(T) = -\frac{2\pi}{3kT}\int_0^\infty R^3 \frac{\partial U}{\partial R} \exp\left[-\frac{U(R)}{kT}\right] \mathrm{d}R, \tag{11.7.1}$$

or in equivalent form,

$$B(T) = 2\pi \int_0^\infty R^2 \left[1 - \exp\left(-\frac{U(R)}{kT}\right)\right] \mathrm{d}R. \tag{11.7.2}$$

So, the second virial coefficient performs the connection between the intermolecular potential and the thermodynamic relations. Various researchers used both equations to obtain the shape of potential function $U(R)$ from experimental data on virial coefficients. When we have a second virial coefficient measured at different temperatures, calculation of a potential energy function is a more complicated problem. It involves solving a first-order Fredholm integral equation, which is an example of typical ill-posed problem [1, 2]. In many cases, the potential function was not found directly; instead, a certain functional on potential $\Delta(U)$ was obtained following [19]. This technique was aimed at obtaining linear Fredholm's integral equation of the first kind with respect to $\Delta(U)$.

Equation (11.7.2) may be directly treated as nonlinear integral equation. If, for example, functions $B(T)$ and $U(R)$ belong to $L_2(0, \infty)$, this problem is apparently ill-posed and requires certain regularization technique. Previous investigations were based on fundamental properties of the potential function that was assumed to consist of repulsive term (monotonically decreasing function for small values of distance R) and attractive term (negative, monotonically increasing to zero function for large R). Well-known representatives of such functions are Lennard-Jones or Morse potentials.

Using restricted set of potential functions may, however, result in the well-posed problem for determining $U(R)$ from Eq. (11.7.2). Let us introduce the following set of constraints D [20]:

1. $U(R)$ is convex downwards for $R < R_p$, and convex upwards for $R > R_p$; obviously this form of potential generalizes all known empirical potential functions.
2. For $R > R_p$ (actually starting with even smaller distances), $U(R)$ is negative, and monotonically approaches zero as $R \to \infty$.

Here R_p is some unknown distance. It can be shown that mentioned constraints are sufficient to define a compact set in L_2 space; this guarantees convergence of approximate solutions of Eq. (11.7.2) to its exact solution of the problem when data errors of $B(T)$ tend to zero. Solving Eq. (11.7.2) may be implemented as finding minimum of the discrepancy functional [20]

$$\Phi[U] = \int_0^\infty \left[B(T) - \tilde{B}(T)\right]^2 dT \tag{11.7.3}$$

where $B(T)$ is calculated according to Eq. (11.7.2), and $\tilde{B}(T)$ are experimentally measured values. Minimization is subject to constraints D.

Technically, potential function $U(R)$ is represented on a discrete grid, and integration in (11.7.2) is restricted to the finite interval by introducing $R_{\min}$, $R_{\max}$ such that input of the omitted intervals was negligible. Discrepancy (11.7.3) was minimized for each value of the parameter R_p from a certain interval, and parameter yielding minimum discrepancy was chosen as the solution.

11.8 Examples of calculations

11.8.1 Calculation of methane intermolecular potential

The second virial coefficient data was simulated for the methane system using experimental data from [21]. A model solution was generated using the Lennard-Jones potential with parameter $\sigma = 4$. With this potential, the second virial coefficients were found for several temperatures, and inverse problem was solved by minimizing (11.7.3) on the set of constraints. Since all constraints are linear, the technique of conjugate gradients projection was applied.

Fig. 11.1 represents solutions of the inverse problem obtained for two cases:

1. When R_p parameter was close to its exact value (5.1 Å).
2. When R_p parameter was seriously in error (7.3 Å).

Note that the first solution (smooth curve with inflection point at $R = 5.1$ Å) practically coincides with the exact solution, because no additional noise was introduced into input data. However, the second solution (piecewise linear curve), obtained under assumption that inflection point is located at $R = 7.3$ Å, yields

practically the same discrepancy value (within 0.1 per cent of the relative error in the values of $B(T)$). The exact model data and simulated virial coefficients corresponding to both potentials (Fig. 11.1) are compared in Table 11.1.

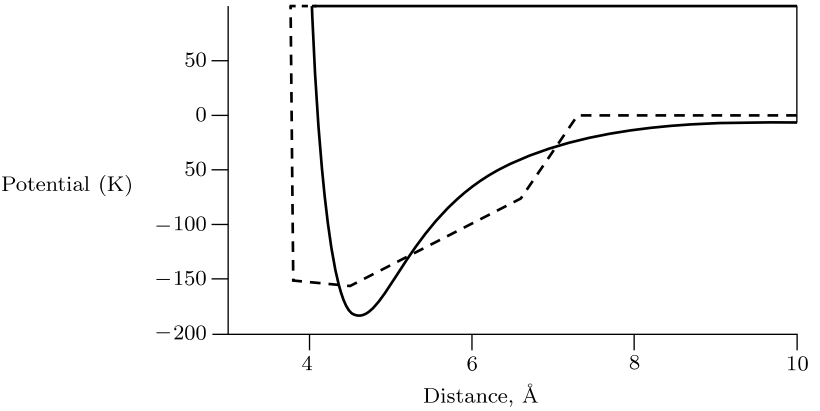


Fig. 11.1 Simulated potential energy function $U(R)$ with two initial inflection points (5.1 [solid line] and 7.3 Å [dash line]).

Table 11.1 Simulated $B(T)$ ($\mathrm{cm}^3 \cdot \mathrm{mol}^{-1}$) within different models.

$T(K)$	B(model)	$B(5.1)$	$B(7.3)$
232.095	−77.61	−77.55	−77.62
250.000	−67.73	−67.76	−67.73
270.420	−58.41	−58.47	−58.40
290.016	−50.96	−51.01	−50.95
309.718	−44.60	−44.62	−44.59
331.357	−38.64	−38.62	−38.65
350.366	−34.13	−34.05	−34.13

Evidently, the limited range and number of experimental points do not allow of accurately defining the properties of the potential function. Accurate solution of the problem requires introducing more temperature points and/or extra constraints (e.g. on solution smoothness).

11.8.2 Prediction of vibrational spectrum of fullerene C_{240}

Molecular design and molecular modeling for the large (nano or polymer, cluster etc.) molecular structures are based on the two major methods used to describe the interactions within a system — quantum mechanics and molecular mechanics. These results could be followed by the techniques that use the same energy models, including energy minimization, molecular dynamics, Monte Carlo simulations and conformational analysis. Both in molecular mechanics and molecular

dynamics the molecular potential energy (and its force field) is represented as the sum of contributions due to bond stretching, bond bending, energy changes accompanying internal rotation about single bonds (all describing the "internal" molecular vibrations, and this part is closely connected with empirical force field calculations), van der Waals attractions and repulsions between nonbonded atoms, and electrostatic interactions due to polar bonds.

For the calculations of vibrational spectra of the large size molecular systems such as polymers, nanostructures, biological systems, and clusters, the next scheme can be proposed:

1. preliminary quantum mechanical analysis of moderate size molecules (fragments of large molecular systems) chosen as key or model molecules;
2. joint treatment of *ab initio* and experimental data on vibrational spectra, structural electron diffraction (ED) and microwave (MW) data for model molecules within the regularization theory, determination the equilibrium geometry parameters and harmonic force constants;
3. determination of intermolecular potential parameters by means of stable numerical methods;
4. organization of a database on structural data, force field parameters and intermolecular potentials transferable in a series of related compounds;
5. synthesis (construction) of a large molecular system from separate fragments included in the database and calculation of its vibrational spectra and thermodynamical functions.

Here we demonstrate the application of this approach to the prediction of vibrational spectrum of C_{240} fullerene molecule (I_h symmetry) presented in Fig. 11.2. This is a hypothetical molecule of fullerene class not synthesized at present.

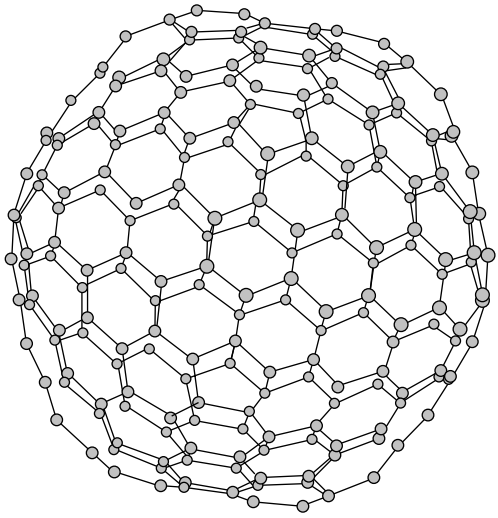


Fig. 11.2 Schematic illustration of C_{240} molecule.

The molecular geometry of C_{240} was optimized at the HF/STO-3G level. The equilibrium configuration of the icosahedral (I_h) symmetry molecule C_{240} is completely defined by five bond lengths shown in Fig. 11.3. Their optimized values are given in Table 11.2 in comparison with experimental data on corannulene $C_{20}H_{10}$ and fullerene C_{60}.

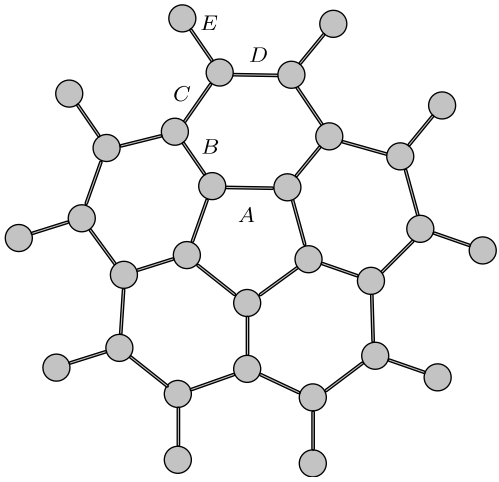


Fig. 11.3 Five types of bond lengths in C_{240}.

Table 11.2 Optimized geometry parameters of C_{240}.

	C_{240}	C_{60}	$C_{20}H_{10}$		
Bond (Å)	HF/STO-3G	X-Ray[a]	X-Ray[b]	ED[c]	B3LYP/6-31G*[c]
R(A)	1.4336	1.432	1.419	1.410	1.417
R(B)	1.3685	1.388	1.396	1.408	1.385
R(C)	1.4581			1.441	1.448
R(D)	1.4132		1.444	1.374	1.390
R(E)	1.4320				

[a] J. M. Hawkins et al. *Science*, **1991**, *252*, 312.
[b] J. C. Hanson and C. E. Nordman. *Acta Crystallogr.*, **1976**, *B32*, 1147.
[c] L. Hedberg et al. *J. Phys. Chem. A*, **2000**, *104*, 7689.

This geometry was used for the normal coordinate analysis of C_{240}. To perform the calculations for this molecule we introduced 1080 of internal coordinates consisting of 360 bond-stretching coordinates and 720 bond angles. Altogether 1080 redundant coordinates were introduced, with only 714 of them being independent. Internal coordinates were optimized automatically with the help of special utility in the SPECTRUM program package [3].

The 714 normal vibrations of C_{240} are distributed by irreducible representations as

$$7A_g + 5A_u + 16F_{1g} + 18F_{1u} + 17F_{2g} + 19F_{2u} + 24G_g + 24G_u + 31H_g + 29H_u$$

The list of 90 different force constants for C_{240} was extended by certain model assumptions of intraball forces on a base transferred from the regularized force constant matrices (RQMFF) of C_{60} and corannulene molecules. The symmetry properties allow one to reduce the complete force constant matrix of C_{240} into 10 blocks with orders varying from 6 to 48 in redundant symmetry coordinates. Symmetry coordinates were run by means of the SYMM program included into the SPECTRUM program package. The vibrational density plots (a distribution of calculated frequencies by a wavenumber scale) for fullerene C_{240} are presented in Fig. 11.4. There are two plots for vibrations with different inversion symmetry (g and u), one referring to the total number of frequencies active in the Raman spectrum, the other to frequencies active in the infrared absorption spectrum. These frequencies were used for the calculation of the thermodynamic functions of C_{240} in the 100–2000 K temperature region (Table 11.2).

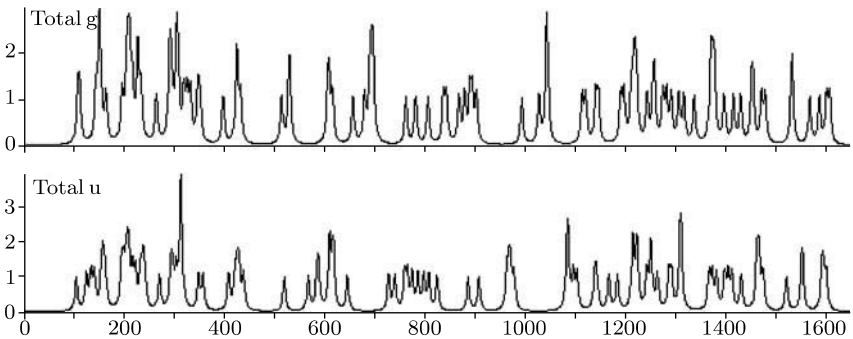


Fig. 11.4 Vibrational state density for C240 molecule in the infrared absorption region between $\omega = 100$ and 1650 cm^{-1}.

The package SPECTRUM allows the processing of several molecules simultaneously (each of them possessing a few isotopic species). This is a sensible approach when the model considerations require equivalence between certain force matrix elements. With this kind of constraint specified, these elements are held equivalent throughout the whole process of optimization. This option is of special value for verifying transferability properties of force constants. Additional features of the package include the following: all constraints are applied to matrices in internal coordinates; all the algorithms throughout the package allow internal coordinates to be redundant; redundancy conditions are taken into account automatically; and the regularization parameter is chosen in accordance with the generalized discrepancy principle. On a base of this package now we develop data base of nitrogen-containing compounds.

References

1. A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, Kluwer Academic Publishers, 1995.
2. A. N. Tikhonov, A. S. Leonov and A. G. Yagola, *Nonlinear Ill-Posed Problems*, Vol. 1, 2, Chapman and Hall, 1998.
3. A. G. Yagola, I. V. Kochikov, G. M. Kuramshina and Yu. A. Pentin, *Inverse Problems of Vibrational Spectroscopy*, VSP, Utrecht, Tokyo, 1999.
4. I. V. Kochikov, G. M. Kuramshina, Yu. A. Pentin and A. G. Yagola, Regularizing algorithm for solving inverse vibrational problem, *Dokl. Akad. Nauk SSSR*, **261**, 1104–1106, 1981.
5. I. V. Kochikov, G. M. Kuramshina, Yu. A. Pentin and A. G. Yagola, Calculation of polyatomic molecule force field by means of Tikhonov's method of regularization, *Dokl. Akad. Nauk SSSR*, **283**, 850–854, 1985.
6. I. V. Kochikov, G. M. Kuramshina and A. G. Yagola, Stable numerical methods of solving certain inverse problems of vibrational spectroscopy, *USSR Comput. Maths. Math. Phys.*, **27**, 33–40, 1987.
7. I. V. Kochikov, G. M. Kuramshina, Yu. A. Pentin and A. G. Yagola, Regularizing algorithms for molecular force field calculations, *J. Mol. Struct.: Theochem*, **272**, 13–33, 1992.
8. G. M. Kuramshina, F. Weinhold, I. V. Kochikov, A. G. Yagola and Yu. A. Pentin, Joint treatment of *ab initio* and experimental data in molecular force field calculations with Tikhonov's method of regularization, *J. Chem. Phys.*, **100**, 1414–1424, 1994.
9. G. M. Kuramshina and A. G. Yagola, *A priori* constraints in the force field calculations of polyatomic molecules, *J. Struct. Chem.*, **38**, 181–194, 1997.
10. G. M. Kuramshina and F. Weinhold, Constraints on the values of force constants for molecular force field models based on ab initio calculations, *J. Mol. Struct.*, **410–411**, 457–462, 1997.
11. I. V. Kochikov, G. M. Kuramshina and A. G. Yagola, Inverse problems of vibrational spectroscopy as nonlinear ill-posed problems, *Surveys on Mathematics for Industry*, **8**, 63–94, 1998.
12. I. V. Kochikov, Yu. D. Tarasov, G. M. Kuramshina, V. P. Spiridonov, A. G. Yagola and T. G. Strand, Regularizing algorithm for determination of equilibrium geometry and harmonic force field of free molecules from joint use of electron diffraction, vibrational spectroscopy and *ab initio* data with application to benzene, *J. Mol. Struct.*, **445**, 243–258, 1998.
13. I. V. Kochikov, G. M. Kuramshina, A. V. Stepanova and A. G. Yagola, Numerical aspects of the calculation of scaling factors from experimental data, *Numerical Methods and Programming*, **5**, 285–290, 2004.
14. I. V. Kochikov, G. M. Kuramshina and A. V. Stepanova, New approach for the correction of *ab initio* molecular force fields in cartesian coordinates, *Int. J. Quant. Chem.*, **109**, 28–33, 2009.
15. W. J. Hehre, L. Radom and P. V. R. Schlegel, *Ab Initio Molecular Orbital Theory*, Wiley, New York, 1985.
16. J. B. Foresman and A. E. Frish, *Exploring Chemistry with Electronic Structure Methods: A Guide to Using Gaussian*, Gaussian Inc., 1993.
17. G. Fogarashi and P. Pulay, *Ab initio* calculation of force fields and vibrational spectra, in: J. R. Durig, ed., *Vibrational Spectra and Structure*, **1**, 125–217, Elsevier, Amsterdam, 1985.
18. I. V. Kochikov, Yu. I. Tarasov, V. P. Spiridonov, G. M. Kuramshina, A. G. Yagola, A. S. Saakjan, M. V. Popik and S. Samdal, Extension of a regularizing algorithm for the determination of equilibrium geometry and force field of free molecules from joint use of electron diffraction, molecular spectroscopy and ab initio data on systems with large-amplitude oscillatory motion, *J. Mol. Struct.*, **485–486**, 421–443, 1999.

19. J. P. Braga and J. L. Neves, Long-range spherical potential energy function for second eficient decompositon into subspaces, *Phys. Chem. Chem. Phys.*, **3**, 4355–4358, 2001.
20. N. V. Anikeeva, I. V. Kochikov, G. M. Kuramshina and A. G. Yagola, Regularized algorithms of constructing intermolecular potential on a base of experimental data, *Numerical Methods and Programming*, **4**, 200–206, 2003.
21. G. Esper, W. Lemming, W. Beckermann and F. Kohler, Acoustic determination of ideal gas heat capacity and second virial coefficient of small hydrocarbons, *Fluid Phase Equilibria*, **105**, 173–179, 1995.

Chapter 12
Numerical Inversion Methods in Geoscience and Quantitative Remote Sensing

Yanfei Wang and Xiaowen Li

Abstract. To estimate structural parameters and spectral component signatures of Earth surface cover type, quantitative remote sensing seems to be an appropriate way to deal with these problems. Since the real physical system that couples the atmosphere with the land surface is very complicated and should be continuous, sometimes it requires comprehensive parameters to describe such a system, so any practical physical model can only be approximated by a model which includes only a limited number of the most important parameters that capture the major variation of the real system.

The pivot problem for quantitative remote sensing is the inversion. Inverse problems are typically ill-posed. The ill-posed nature is characterized by (C1) the solution may not exist; (C2) the dimension of the solution space may be infinite; (C3) the solution is not continuous with the variation of the observed signals. These issues exist for all quantitative remote sensing inverse problems. For example, when sampling is poor, i.e., there are very few observations, or directions are poorly located, the inversion process would be underdetermined, which leads to the large condition number of the normalized systems and the significant noise propagation. Hence (C2) and (C3) would be the chief difficulties for quantitative remote sensing inversion.

This chapter will address the theory and methods from the viewpoint that the quantitative remote sensing inverse problems can be represented by kernel-based operator equations.

Yanfei Wang
Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China.
e-mail: yfwang@mail.iggcas.ac.cn

Xiaowen Li
Research Center for Remote Sensing and GIS, Beijing Normal University, Beijing, 100875, China.
e-mail: lix@bnu.edu.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

12.1 Introduction

Both modeling and model-based inversion are important for geophysical problems and quantitative remote sensing. Hundreds of models related to vegetation and radiation have been established during past decades. The model-based inversion in solid geophysics and atmospheric science has been well understood. However, the model-based inverse problems for land surface received much attention from scientists only in recent years. Compared to modeling, model-based inversion is still in the stage of exploration. This is because intrinsic difficulties exist in the application of *a priori* information, inverse strategy and inverse algorithm. The appearance of hyperspectral and multiangular remote sensor enhanced the exploration means, and provided us more spectral and spatial dimension information than before. However, how to utilize these information to solve the problems faced in quantitative remote sensing to make remote sensing really enter the time of quantification is still an arduous and urgent task for remote sensing scientists. Remote sensing inversion for different scientific problems in different branches is paid more and more attention in recent years. In a series of international study projections, such as International Geosphere-Biosphere Programme (IGBP), World Climate Research Programme (WCRP) and NASA's Earth Observing System (EOS), remote sensing inversion has become a focal point of study.

Remote sensing inversions are usually optimization problems with different constraints. Therefore, how to incorporate the method developed in operation research field into remote sensing inversion field should be clarified. In quantitative remote sensing, since the real physical system that couples the atmosphere with the land surface is very complicated and should be continuous, sometimes it requires comprehensive parameters to describe such a system, so any practical physical model can only be approximated by a model which includes only a limited number of the most important parameters that capture the major variation of the real system. Generally speaking, a discrete forward model to describe such a system is in the form

$$y = h(\mathbf{C}, \mathbf{S}), \tag{12.1.1}$$

where y is single measurement, $\mathbf{C}$ is a vector of controllable measurement conditions such as wave band, viewing direction, time, sun position, polarization and so forth, $\mathbf{S}$ is a vector of state parameters of the system approximation, h is a function which relates $\mathbf{C}$ and $\mathbf{S}$, which is generally nonlinear and continuous.

With the ability of satellite sensors to acquire multiple bands, multiple viewing directions, and so on, while keeping $\mathbf{S}$ essentially the same, we obtain the following nonhomogeneous equation

$$\mathbf{D} = h(\mathbf{C}, \mathbf{S}) + \mathbf{n}, \tag{12.1.2}$$

where D is a vector in $\mathbb{R}^M$, which is an M dimensional measurement space with M values corresponding to M different measurement conditions, $\mathbf{n} \in \mathbb{R}^M$ is the vector of random noise with the same vector length M. Assume that there are m undetermined parameters which need to recover. Clearly, if $M = m$, (12.1.2) is a determined system, so it is not difficult to develop some suitable algorithms to

solve it. If more observations can be collected than the existing parameters in the model[25], i.e., $M > m$, the system (12.1.2) is overdetermined. In this situation, the traditional solution does not exist. We must define its solution in some other meaning, for example, the least squares error (LSE) solution. However Li [15] pointed out that "for physical models with about ten parameters (single band), it is questionable whether remote sensing inversion can be an overdetermined one in the foreseeable future." Therefore, the inversion problems in geosciences seem to be always underdetermined in some sense. Nevertheless, the underdetermined system in some cases can be always converted to an overdetermined one by utilizing multiangular remote sensing data or by accumulating some *a priori* knowledge [14].

Developed methods in literature for quantitative remote sensing inversion are mainly statistical methods with different variation from Bayesian inference. In this chapter, we analyze from algebraic point of view the solution theory and methods for quantitative remote sensing inverse problems.

12.2 Examples of quantitative remote sensing inverse problems: land surface parameter retrieval problem

As is well known, the anisotropy of the land surface can be best described by the bidirectional reflectance distribution function (BRDF). With the progress of the multiangular remote sensing, it seems that the BRDF models can be inverted to estimate structural parameters and spectral component signatures of Earth surface cover type [22], [21]. The state of the art of BRDF is the use of the linear kernel-driven models, mathematically described as the linear combination of the isotropic kernel, volume scattering kernel and geometric optics kernel. The information extraction on the terrestrial biosphere and other problems for retrieval of land surface albedos from satellite remote sensing have been considered by many authors in recent years, for instance, the survey papers on the kernel-based bidirectional reflectance distribution function (BRDF) models [17, 19, 18] and references therein. The computational stability is characterized by the algebraic operator spectrum of the kernel-matrix and the observation errors. Therefore, the retrieval of the model coefficients is of great importance for computation of the land surface albedos. Other than observation errors a limited or insufficient number of observations is one of the most severe obstacles for the estimation of BRDF. Therefore, it is very desirable to develop new techniques for the robust estimation of the BRDF model parameters due to the scarcity of the number of observations.

The linear kernel-based BRDF model can be described as follows [21]:

$$f_{iso} + k_{vol}(t_i, t_v, \phi) f_{vol} + k_{geo}(t_i, t_v, \phi) f_{geo} = r(t_i, t_v, \phi), \tag{12.2.1}$$

where r is the bidirectional reflectance; the kernels k_{vol} and k_{geo} are the so-called kernels, that is, known functions of illumination and of viewing geometry which describe volume and geometric scattering respectively; t_i and t_v are the zenith angle of the solar direction and the zenith angle of the view direction respectively; ϕ is the relative azimuth of the sun and view direction; and f_{iso}, f_{vol} and f_{geo} are three unknown parameters to be adjusted to fit observations. Theoretically, f_{iso}, f_{vol} and f_{geo} are closely related to the biomass such as leaf area index (LAI), Lambertian reflectance, sunlit crown reflectance, and viewing and solar angles. The vital task then is to retrieve appropriate values of the three parameters.

Generally speaking, the BRDF model includes kernels of many types. However, it was demonstrated that the combination of RossThick (k_{vol}) and LiSparse (k_{geo}) kernels had the best overall ability to fit BRDF measurements and to extrapolate BRDF and albedo [10, 32, 16, 20]. A suitable expression for the RossThick kernel k_{vol} was derived from [21]. It is reported that the LiTransit kernel k_{Transit}, instead of the kernel k_{geo}, is more robust and stable than LiSparse non-reciprocal kernel and the reciprocal LiSparse kernel k_{sparse} (LiSparseR) where the LiTransit kernel and the LiSparse kernel are related by

$$k_{\text{Transit}} = \begin{cases} k_{\text{sparse}}, & B \leq 2, \\ \frac{2}{B} k_{\text{sparse}}, & B > 2, \end{cases} \tag{12.2.2}$$

and B is given

$$B := B(t_i, t_v, \phi) = -O(t_i, t_v, \phi) + \sec t_i' + \sec t_v'$$

in [13]. More detailed explanation about O and t' in the definition of k_{Transit} can be found in [32].

To use the combined linear kernel model, a key issue is to numerically build the inverse model in a stable way. However, it is difficult to do in practical applications due to ill-posed nature of the inverse problem. So far, statistical methods have been developed for solving the inverse problem, which is based on expression of *a priori* knowledge of model parameter as a joint probability density $P_s(s_1, s2, \cdots, s_k)$; and the expression for *a priori* knowledge of the model accuracy and measurement noise should be a conditional joint probability density $P_D(d_{\text{obs}})$

$$P_D(d_{\text{obs}}) = P_D(d_{\text{obs}}|S)P_s(S).$$

By Bayesian inference formula, we have

$$P(s|d_{\text{obs}}) = P_D(d_{\text{obs}}|S)P_s(S)/P_D(d_{\text{obs}}),$$

where $P_D(d_{\text{obs}}) = \int_S P_D(d_{\text{obs}}|S)P_s(S)dV_s$, and dV_s is the volume differential elements in the parameter space. The $P_D(d_{\text{obs}}|S)$ can be interpreted as the prior knowledge of model prediction of d_{obs} giving parameters in parameter space, $P_s(S)$ as the prior knowledge of parameters, and $P_D(d_{\text{obs}})$ as prior knowledge of marginal density of observation. In [17, 18], the authors utilized the algebraic QR decomposition technique and also suggested using the singular value decom-

position for the inversion of the BRDF model. And optimal design scheme for the angular sampling is addressed. Later in [19], they compared several inversion techniques and uncertainty in albedo estimates from the SEVIRI/MSG observing system by using POLDER BRDF measurements. Though the solution method is algebraic, their description of the problem is still quite statistical.

12.3 Formulation of the forward and inverse problem

The inverse problems can be linear or nonlinear which is closely related with forward linear or nonlinear problems. A general framework of a forward model can be described by the following chart

Input x $\longrightarrow$ System model/Process K $\longrightarrow$ Output y.
Cause $x \longrightarrow$ System model/Process K $\longrightarrow$ Result y.
Model (model parameters) m $\longrightarrow$ Source dependent operator K
$\longrightarrow$ Observation (data) d.

Mathematically, the forward model can be written as

$$K(x) = y,\ K(m) = d \tag{12.3.1}$$

for nonlinear problems or

$$Kx = y,\ Km = d \tag{12.3.2}$$

for linear problems, where x, y, m and d are functions related with different variables in different applied sciences.

The inverse problems can be described by the following chart

Output y $\longrightarrow$ System model/Process K $\longrightarrow$ Input x.
Result y $\longrightarrow$ System model/Process K $\longrightarrow$ Cause x.
Observation (data) d $\longrightarrow$ Source dependent operator K
$\longrightarrow$ Model (model parameters) m.

Accordingly, the inverse model can be written mathematically as

$$x = K^{-1}(y),\ m = K^{-1}(d) \tag{12.3.3}$$

for nonlinear problems or

$$x = K^{-1}y,\ m = K^{-1}d \tag{12.3.4}$$

for linear problems, where x, y, m and d are functions defined as before.

It is clear that the kernel-based land surface parameter model (12.2.1) can be written as a linear forward model with

$$K = \left[1\ k_{vol}(t_i, t_v, \phi)\ k_{geo}(t_i, t_v, \phi)\right],$$

$$y = [r(t_i, t_v, \phi)],$$

and

$$x = [f_{iso}, f_{vol}, f_{geo}]^T,$$

where y and the components 1, $k_{vol}(t_i, t_v, \phi)$ and $k_{geo}(t_i, t_v, \phi)$ of K can be column vectors for multiangular data.

12.4 What causes ill-posedness

A problem is called ill-posed if the solution of unknown (to be recovered) parameters or functions does not exit, or is not unique, or it is not a continuous function of the data. This means even if a solution exists, it may severely unstable, i.e., a small perturbation of the data corresponds to a significant large perturbation of the solution.

For quantitative remote sensing, the basis is constructing the mathematical model which relates the remote sensing data/signal with the Earth surfaces. This model can be written implicitly as [9]

$$y_i = h(a_i, b_i, c_i, d_i, e_i, \epsilon), \tag{12.4.1}$$

where y_i records the radiation signal by remote sensor; a_i, b_i, c_i, d_i, e_i and ϵ represent the radiation source, atmospheric information, vegetation, ground and soil, sensor and unknown noise, respectively. The inverse problem is to recover one or several parameters of a_i, b_i, c_i, d_i and e_i. For example, giving data y and measurable parameters a_i, b_i, d_i and e_i to retrieve c_i is called the vegetation model inversion.

The BRDF physical model is a special case of remote sensing inverse problems. The ill-posedness arises because the linear kernel-based BRDF model is underdetermined, for instance, when there are very few observations or poor directional range. A single angular observation may lead to an underdetermined system whose solution set is infinite (the null space of the operator contains nonzero vectors) or which has no solution (the rank of the coefficient matrix is not equal to the rank of the augmented matrix). The lack of effective observation is not only a major obstacle of remote sensing inversion, but also an obstacle of most of the geophysical inverse problems. Another reason that leads to the ill-posedness is that error/noise propagation is significantly enlarged in computation due to bad algebraic spectrum distribution (see, e.g., [Wang et al., 2007]). Due to the ill-posedness of the inversion process, uncertainties in the model and in the reflectance measurements do not simply result in uncertainties as to the solution. More severely, the ill-posedness may lead to jumps in the solution space, i.e., the solution found may spread in the whole parameter space instead of being centered on the true solution (see [Xiao et al., 2003; Atzberger, 2004; Tikhonov and Arsenin, 1977; Wang, 2007; Wang et al., 2008]). To alleviate those difficulties, it

is necessary to impose *a priori* constraints on the unknown parameters and seek for a global optimized solution.

12.5 Tikhonov variational regularization

The Tikhonov regularization method is to solve a regularized minimization problem

$$J^{\alpha}(\mathbf{x}) := \|K\mathbf{x} - \mathbf{y}\|_2^2 + \alpha\|D^{1/2}\mathbf{x}\|_2^2 \longrightarrow \min . \tag{12.5.1}$$

In (12.5.1), α is the regularization parameter and D is a positively (semi-)definite operator. By a variational process, the minimizer of (12.5.1) satisfies

$$K^T K\mathbf{x} + \alpha D\mathbf{x} = K^T\mathbf{y_n}. \tag{12.5.2}$$

The operator D is a scale matrix which imposes smoothness constraint on the solution $\mathbf{x}$. The scale operator D and the regularization parameter α can be considered as some kind of *a priori* information, which will be discussed next.

12.5.1 Choices of the scale operator D

To regularize the ill-posed problem discussed in the previous subsection, the choice of the scale operator D has great impact upon the performance to the regularization. Note that the matrix D plays the role in imposing a smoothness constraint on the parameters and in improving the condition of the spectrum of the adjoint operator $K^T K$. Therefore, it should be positively definite or at least positively semi-definite. One may readily see that the identity may be a choice. However this choice does not fully employ the assumption about the continuity of the parameters.

We recall four classes of choosing the scale operator D developed in [31]. The first is the smoothness constrained optimization

$$\min J^{\alpha}(x) := \rho_F[Kx, y] + \alpha L(x), \tag{12.5.3}$$

where $\rho_F[Kx, y] = \frac{1}{2}\|Kx - y\|_{L_2}^2$, $L(x) = \frac{1}{2}\|x\|_{W^{1,2}}^2$. Assume that the variation of x is flat near the boundary of the integral interval $[a, b]$. In this case, the derivatives of x are zeros at the boundary of $[a, b]$. Let h_r be the step size of the grids in $[a, b]$, which could be equidistant or adaptive. Then after discretization of $L(x)$, D is a tridiagonal matrix in the form

$$D := D_1 = \begin{bmatrix} 1+\frac{1}{h_r^2} & -\frac{1}{h_r^2} & 0 & \cdots & 0 \\ -\frac{1}{h_r^2} & 1+\frac{2}{h_r^2} & -\frac{1}{h_r^2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{h_r^2} & 1+\frac{2}{h_r^2} & -\frac{1}{h_r^2} \\ 0 & \cdots & 0 & -\frac{1}{h_r^2} & 1+\frac{1}{h_r^2} \end{bmatrix}.$$

There are many kinds of techniques for choosing the scale matrix D appropriately. In Phillips-Twomey's formulation of regularization (see, e.g., [28]), the matrix D is created by the norm of the second differences, $\sum_{i=2}^{N-1}(x_{i-1}-2x_i+x_{i+1})^2$, which leads to the following form of matrix D

$$D := D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -2 & 5 & -4 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 & -4 & 5 & -2 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

However, the matrix D is badly conditioned and thus the solution to minimize the functional $J^\alpha[x]$ with D as the smooth constraint is observed to have some oscillations ([28]). Another option is the negative Laplacian (see, e.g., [26, 29]): $Lx := -\sum_{i=1}^{n} \frac{\partial^2 x}{\partial \tau_i^2}$, for which the scale matrix D for the discrete form of the negative Laplacian Lx is

$$D := D_3 = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix},$$

in which we assume the discretization step length to be 1. The scale matrix D_3 is positively semi-definite but not positively definite and hence the minimization problem may not work efficiently for severely ill-posed inverse problems. Another option of the scale matrix D is the identity, i.e., $D := D_4 = \mathrm{diag}(e)$, where e is the components of all ones, however this scale matrix is too conservative and may lead to over-regularization.

12.5.2 Regularization parameter selection methods

As noted above, the choice of the regularization parameter α is important to tackle the ill-posedness. *A priori* choice of the parameter α allows $0 < \alpha < 1$. However the *a priori* choice of the parameter does not reflect the degree of approximation that may lead to either over-estimate or under-estimate of the regularizer.

We will use the widely used discrepancy principle [23, 24, 24] to find an optimal regularization parameter. In fact, the optimal parameter α^* is a root of the nonlinear function

$$\Psi(\alpha) = \|K\mathbf{x}_\alpha - \mathbf{y_n}\|^2 - \delta^2, \tag{12.5.4}$$

where δ is the error level to specify the approximate degree of the observation to the true noiseless data, $\mathbf{x}_\alpha$ denotes the solution of the problem in equation (12.5.2) corresponding to the value α of the related parameter. Noting $\Psi(\alpha)$ is differentiable, fast algorithms for solving the optimal parameter α^* can be implemented. As in [31], the cubic convergent algorithm can be applied:

$$\alpha_{k+1} = \alpha_k - \frac{2\Psi(\alpha_k)}{\Psi'(\alpha_k) + (\Psi'(\alpha_k)^2 - 2\Psi(\alpha_k)\Psi''(\alpha_k))^{\frac{1}{2}}}. \tag{12.5.5}$$

In the above cubic convergent algorithm, the functions $\Psi'(\alpha)$ and $\Psi''(\alpha)$ have the following explicit expression:

$$\Psi'(\alpha) = -\alpha\beta'(\alpha),$$
$$\Psi''(\alpha) = -\beta'(\alpha) - 2\alpha\Big[\|\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{d}\alpha}\|^2 + (\mathbf{x}_\alpha, \frac{\mathrm{d}^2\mathbf{x}_\alpha}{\mathrm{d}\alpha^2})\Big],$$

where $\beta(\alpha) = \|\mathbf{x}_\alpha\|^2$, $\beta'(\alpha) = 2(\frac{\mathrm{d}\mathbf{x}_\alpha}{d\alpha}, \mathbf{x}_\alpha)$, and $\mathbf{x}_\alpha$, $\mathrm{d}\mathbf{x}_\alpha/\mathrm{d}\alpha$ and $\mathrm{d}^2\mathbf{x}_\alpha/\mathrm{d}\alpha^2$ can be obtained by solving the following equations:

$$(K^TK + \alpha D)\mathbf{x}_\alpha = K^T\mathbf{y_n}, \tag{12.5.6}$$

$$(K^TK + \alpha D)\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{d}\alpha} = -D\mathbf{x}_\alpha, \tag{12.5.7}$$

$$(K^TK + \alpha D)\frac{\mathrm{d}^2\mathbf{x}_\alpha}{\mathrm{d}\alpha^2} = -2D\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{d}\alpha}. \tag{12.5.8}$$

To solve the linear matrix-vector equations (12.5.6)–(12.5.8), we use the Cholesky (square root) decomposition method. A remarkable characteristic of the solution of (12.5.6)–(12.5.8) is that the Cholesky decomposition of the coefficient matrix $K^TK+\alpha D$ needs only once, then the three vectors $\mathbf{x}_\alpha$, $\mathrm{d}\mathbf{x}_\alpha/\mathrm{d}\alpha$, $\mathrm{d}^2\mathbf{x}_\alpha/\mathrm{d}\alpha^2$ can be obtained cheaply.

12.6 Solution methods

With the mathematical model relating the physical parameters, it is necessary to develop proper solution methods. It it proved that without robust solution methods, the retrieved parameters may be far away from the true solution or the solution process may be quite time-consuming. This is particularly important to numerical inversion and imaging, say, atmospheric remote sensing inversion and data assimilate problems.

We briefly review below some important methods developed in mathematics and applied sciences.

12.6.1 Gradient-type methods

In this subsection, we consider iterative gradient-type methods for solving linear operator equations

$$K\mathbf{x} = \mathbf{y}, \tag{12.6.1}$$

where K, $\mathbf{x}$ and $\mathbf{y}$ have the same meaning as before.

The functional $J(\mathbf{x})$ to be minimized is given by

$$J(\mathbf{x}) = \frac{1}{2}\|K\mathbf{x} - \mathbf{y}\|_{l^2}^2. \tag{12.6.2}$$

The gradient of $J(\mathbf{x})$ is given by

$$\text{grad}_{\mathbf{x}}[J(\mathbf{x})] = (K^T K)^{-1}\mathbf{x} - K^T\mathbf{y}.$$

At the kth iterative step, the gradient of $J(\mathbf{x}_k)$ can be expressed as $\text{grad}_k[J]$, which is evaluated by $\text{grad}_{\mathbf{x}_k}[J(\mathbf{x}_k)]$.

12.6.1.1 Bialy iteration

The Bialy iterative method for ill-posed linear inverse problems is in the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tau(y - K\mathbf{x}_k), \tag{12.6.3}$$

where $\tau \in (0, 2/\|A\|)$ and K is linear, bounded and nonnegative. The method can be considered as a special case of gradient method. One may readily see this method is very similar to the method of successive-approximations from standard textbook "numerical methods", where a very simple way to introduce the method is the following: consider the operator

$$T(\mathbf{x}) = \mathbf{x} + \tau(\mathbf{y} - K\mathbf{x}), \tag{12.6.4}$$

where τ is the so-called relaxation parameter. Any solution of (12.6.1) is equivalent to finding a fixed point of the operator T, i.e., solve $\mathbf{x}$ from

$$\mathbf{x} = T(\mathbf{x}). \tag{12.6.5}$$

Assuming that T is a contraction mapping, then by the method of successive approximations, we obtain the following iterative scheme

$$\mathbf{x}_{k+1} = T(\mathbf{x}_k), \tag{12.6.6}$$

i.e., iterative formula (12.6.3). Bialy algorithm converges if and only if $K\mathbf{x} = \mathbf{y}$ has a solution. The algorithm is straightforward to be given as follows:

Algorithm 12.6.1. *(Bialy iteration algorithm)*

Step 1 Input K and $\mathbf{y}$; Choose $\mathbf{x}_0$, $\tau \in (0, 2/\|K\|)$; Set $k := 0$.
Step 2 If the stopping rule is satisfied, STOP; Otherwise, GOTO Step 3.
Step 3 Iterates according to Bialy iterative formula (12.6.3).
Step 4 Set $k := k + 1$ and GOTO Step 2.

12.6.1.2 Landweber-Fridman iteration

Landweber and Fridman suggested rewriting the equation $Kx = y$ in the fix point form and iterating this equation, i.e., computing

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tau \cdot K^T(y - K\mathbf{x}_k), \tag{12.6.7}$$

where $\tau \in (0, 2/\|A^T A\|)$, K is linear, bounded. The algorithm is straightforward to be given as follows:

Algorithm 12.6.2. *(Landweber-Fridman iteration algorithm)*

Step 1 Input K and $\mathbf{y}$; Choose $\mathbf{x}_0$, $\tau \in (0, 2/\|A^T A\|)$; Set $k := 0$.
Step 2 If the stopping rule is satisfied, STOP; Otherwise, GOTO Step 3.
Step 3 Iterates according to Landweber-Fridman iterative formula (12.6.7).
Step 4 Set $k := k + 1$ and GOTO Step 2.

12.6.1.3 Steepest descent iteration

Steepest descent is, perhaps, the most famous and simple method in optimization. The formula is very similar to Landweber-Fridman iteration except that the stepsize is changed, i.e.,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tau_k \cdot K^T(y - K\mathbf{x}_k), \tag{12.6.8}$$

where τ_k is obtained by line search, i.e., $\tau_k = \text{argmin}_{\tau>0} J(\mathbf{x}_k - \tau g_k)$. The algorithm is straightforward to be given as follows:

Algorithm 12.6.3. *(Steepest descent iteration algorithm)*

Step 1 *Input K and $\mathbf{y}$; Choose $\mathbf{x}_0$; Compute τ_0 by the above line search formula; Set $k := 0$.*
Step 2 *If the stopping rule is satisfied, STOP; Otherwise, GOTO Step 3.*
Step 3 *Iterates according to the steepest descent iterative formula (12.6.7).*
Step 4 *Set $k := k + 1$ and GOTO Step 2.*

12.6.1.4 ν-method acceleration

This is a kind of accelerated Landweber-Fridman iterative method. The method is introduced by Brakhage[2], which belongs to a class of polynomial acceleration (semi-iterative method) to find the numerical solutions of linear algebraic systems. The iteration formula reads

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mu_k(\mathbf{x}_{k-1} - \mathbf{x}_{k-2}) + \omega_k A^*(y - A\mathbf{x}_{k-1}), \tag{12.6.9}$$

where $\mu_1 = 0,\ \omega_1 = \frac{4\nu+2}{4\nu+1}$; $\nu > 0$ is an *a priori* parameter which characterizes the smoothness of the solution, i.e., assuming that $x \in \text{Range}((A^*A)^\nu)$. For $k \geq 1$, we have the following iterative formulas:

$$\mu_k = \frac{(k-1)(2k-3)(2k+2\nu-1)}{(k+2\nu-1)(2k+4\nu-1)(2k+2\nu-3)},$$
$$\omega_k = 4\frac{((2k+2\nu-1)(k+\nu-1)}{(k+2\nu-1)(2k+4\nu-1)}.$$

If we set $z_0 = 0$, $\mathbf{x}_0 = 0$, $z_1 = \omega_1 y$ and $\mathbf{x}_1 = A^* z_1$, then the iteration formula can be simply refined as

$$\begin{aligned} z_k &= z_{k-1} + \mu_k(z_{k-1} - z_{k-2}) + \omega_k A^*(y - Az_{k-1}), \\ \mathbf{x}_k &= A^* z_k. \end{aligned} \tag{12.6.10}$$

Therefore, the algorithm can be given as follows:

Algorithm 12.6.4. *(ν-iteration algorithm)*

Step 1 *Input K and $\mathbf{y}$; Choose $\mathbf{x}_0 = 0$, $z_0 = 0$; Set $\nu > 0$; Compute ω_1.*
Step 2 *Compute $z_1 = \omega_1 y$; $\mathbf{x}_1 = A^* z_1$.*
Step 3 *Until convergence, do the iterations according to the iterative formula (12.6.10).*

12.6.1.5 Conjugate gradient method

The conjugate gradient (CG) method can be used for solving least-squares problem. The CG can be applied to the normal equation $K^T Kx = K^T y$ or to the

regularized normal equation $(K^TK+\alpha D^TD)x = K^Ty,\ \alpha > 0$. Here we re-outline the algorithm for solving the regularized normal equation as follows:

Algorithm 12.6.5. *(Regularized Conjugate Gradient algorithm)*

Step 1 *Input* K, y *and* $\mathbf{x}_0$*; Set* $\text{res}_0 := K^Ty - K^TK\mathbf{x}_0$*; Choose* $\alpha \in (0,1)$.

Step 2 *If* $\text{res}_0 \leq \epsilon$*, output* $\mathbf{x}_0$*, STOP; otherwise, set* $s_0 := \text{res}_0$, $\rho_0 := \text{res}_0^T\text{res}_0$ *and set* iter $:= 1$.

Step 3 *Compute next iteration point:*

$$\begin{aligned}
&\alpha_k := \rho_{k-1}/((Ks_{k-1})^T(Ks_{k-1}) + \alpha(Ds_{k-1})^T(Ds_{k-1})),\\
&\mathbf{x}_k := \mathbf{x}_{k-1} + \alpha_k s_{k-1},\\
&\text{res}_k := \text{res}_{k-1} - \alpha_k K^TKs_{k-1},\\
&\rho_k := \text{res}_k^T\text{res}_k,\\
&\beta_k := \rho_k/\rho_{k-1},\\
&s_k := \text{res}_k + \beta_k s_{k-1}.
\end{aligned}$$

Step 4 *If* $\text{res}_k \leq \epsilon$ *or* iter *exceeds the maximum iterative steps, output* $\mathbf{x}_k$*, STOP; otherwise, set* iter $:=$ iter $+ 1$*, GOTO Step 3.*

For the CG algorithm, in ideal situation, the iteration stops when $\text{res}_0 = 0$ or $\text{res}_k = 0$. But it is hard to maintain $\text{res}_0 = 0$ or $\text{res}_k = 0$ for noisy inverse problems. Therefore, we modify the algorithm that it should be stopped when $\text{res}_0 \leq \epsilon$ or $\text{res}_k \leq \epsilon$, for ϵ chosen in $(0,1)$.

12.6.1.6 Nonmonotone gradient method

Both the steepest descent and the Landweber-Fridman iteration methods are quite slow in convergence and are difficult to be used for practical problems[29]. Instead of using negative gradient in each iteration, the non-monotone gradient methods are developed recently (see, e.g., [7]). The famous one is the Barzilai-Borwein (BB) method. This method was first proposed for solving the unconstrained optimization problem [1]. A gradient method for solving the minimization of $J(\mathbf{x})$ calculates the next point from

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \nu_k \text{grad}[J]_k, \tag{12.6.11}$$

where ν_k is the steplength that depends on the method being used. The key point of Barzilai and Borwein's method is the two choices of the stepsize ν_k

$$\nu_k^{BB1} = \frac{s_{k-1}^Ts_{k-1}}{s_{k-1}^Tz_{k-1}} \tag{12.6.12}$$

and

$$\nu_k^{BB2} = \frac{s_{k-1}^Tz_{k-1}}{z_{k-1}^Tz_{k-1}}, \tag{12.6.13}$$

where $z_k = \text{grad}[J]_{k+1} - \text{grad}[J]_k$, $s_k = \mathbf{x}_{k+1} - \mathbf{x}_k$. The method is quite efficient for ill-posed convex quadratic programming problem [30].

The algorithm can be easily implanted by modifying iteration formula in Algorithm 12.6.3.

Algorithm 12.6.6. *(Nonmonotone algorithm)*

Step 1 *Initialization: Given initial point m_0, tolerance $\varrho > 0$ and set $k := 1$; Compute* $\text{grad}[J]_k$.
Step 2 *If the stopping rule is satisfied, STOP; Otherwise, set $s_k = -\text{grad}[J]_k$.*
Step 3 *Compute a BB step: ν_k by equation (12.6.12) or (12.6.13).*
Step 4 *Update the current iteration point by setting $\mathbf{x}_{k+1} = \mathbf{x}_k + \nu_k s_k$.*
Step 5 *Compute a new search direction $s_{k+1} = -\text{grad}[J]_{k+1}$ and go to* Step 6.
Step 6 *Loop: $k := k + 1$ and go to* Step 2.

For some of the gradient-type iterative methods, we do not provide the detailed stopping rule. Actually this can be realized by setting the maximum iteration numbers or by comparing the norm value of the difference of the current iteration point to the former iteration point or by evaluating the norm value of the gradient $\text{grad}[J]_k$. If these values are less than the preassigned tolerance, the iteration process should be terminated.

12.6.2 Newton-type methods

The conventional Tikhonov regularization method is equivalent to constrained l_2 minimization problem

$$\min_{\mathbf{x}} \ \|\mathbf{x}\|_2, \quad \text{s.t. } K\mathbf{x} = \mathbf{y}. \tag{12.6.14}$$

This reduces to solve an unconstrained optimization problem

$$\mathbf{x} = \text{argmin}_{\mathbf{x}} J^\alpha(\mathbf{x}), \quad J^\alpha(\mathbf{x}) = \frac{1}{2}\|K\mathbf{x} - \mathbf{y}\|_2^2 + \frac{\alpha}{2}\|x\|_2^2. \tag{12.6.15}$$

The gradient and Hessian of $J^\alpha(\mathbf{x})$ are given by

$$\text{grad}_{\mathbf{x}}[J^\alpha(\mathbf{x})] = (K^T K + \alpha I)^{-1}\mathbf{x} - K^T\mathbf{y} \text{ and } \text{Hess}_{\mathbf{x}}[J^\alpha(\mathbf{x})] = K^T K + \alpha I$$

respectively. Hence at the kth iterative step, the gradient and Hessian of $J^\alpha(\mathbf{x}_k)$ can be expressed as $\text{grad}_k[J^\alpha]$ and $\text{Hess}_k[J^\alpha]$, which are evaluated by $\text{grad}_{\mathbf{x}_k}[J^\alpha(\mathbf{x}_k)]$ and $\text{Hess}_{\mathbf{x}_k}[J^\alpha(\mathbf{x}_k)]$ respectively.

12.6.2.1 Gauss-Newton method

This is an extension of Newton method in one-dimensional space to higher dimensional space. The iteration formula reads

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\mathrm{Hess}_k[J^\alpha])^{-1}\mathrm{grad}_k[J^\alpha]. \tag{12.6.16}$$

The algorithm is straightforward to given as follows:

Algorithm 12.6.7. *(Gauss-Newton algorithm)*

Step 1 Choose x_0, $\epsilon > 0$*; Compute* $\mathrm{grad}_0[J^\alpha]$ *and* $\mathrm{Hess}_0[J^\alpha]$*; Set* $k := 0$.
Step 2 If $\|\mathrm{grad}_k[J^\alpha]\| \le \epsilon$*, STOP; Otherwise, GOTO Step 3.*
Step 3 Iterates according to formula (12.6.16).
Step 4 Set $k := k + 1$*, update* $\mathrm{grad}_k[J^\alpha]$ *and GOTO Step 2.*

12.6.2.2 Broyden method

For finite dimensional minimization problem, probably the most used approach is a so called secant method: at every step of an iterative process the Jacobian $R'(\mathbf{x}_k)$ is replaced with an approximation, obtained from $R(\mathbf{x}_{k+1})$ and $R(\mathbf{x}_k)$. The best known algorithm among the secant methods is the Broyden iterative process. The iteration formula of Broyden method reads

$$\mathbf{x}_{k+1} = \mathbf{x}_k - B_k^{-1}R(\mathbf{x}_k), \tag{12.6.17}$$

$$B_{k+1} = B_k + \frac{(s_k, \cdot)}{\|s_k\|^2}R(\mathbf{x}_{k+1}), \tag{12.6.18}$$

$$B_{k+1}^{-1} = B_k^{-1} - \frac{(s_k, B_k^{-1}\cdot)}{(s_k, B_k^{-1}(R(\mathbf{x}_{k+1}) - R(\mathbf{x}_k)))}B_k^{-1}R(\mathbf{x}_{k+1}), \tag{12.6.19}$$

where $R(\mathbf{x}_k) = K(\mathbf{x}_k) - y$, $s_k := \mathbf{x}_{k+1} - \mathbf{x}_k$. This method is also called secant method. The algorithm is straightforward to given as follows:

Algorithm 12.6.8. *(Broyden algorithm)*

Step 1 Choose $\mathbf{x}_0$, B_0, $\epsilon > 0$*, and a symmetric positive definite starting matrix* B_0*; Set* $k := 0$.
Step 2 If $\|\mathrm{grad}_k[J^\alpha]\| \le \epsilon$*, STOP; Otherwise, GOTO Step 3.*
Step 3 Iterates according to formulas (12.6.17)–(12.6.19).
Step 4 Set $k := k + 1$*, update* B_k *and* B_k^{-1} *and GOTO Step 2.*

12.6.2.3 BFGS method

The method BFGS comes from four mathematicians's work on unconstrained optimization. The iterative formula reads

$$\mathbf{x}_{k+1} = \mathbf{x}_k + B_k^{-1}\Delta g_k, \tag{12.6.20}$$

$$B_{k+1}^{-1} := H_{k+1} = (I - \frac{s_k \Delta g_k^T}{s_k^T \Delta g_k}) H_k (I - \frac{\Delta g_k s_k^T}{s_k^T \Delta g_k}) + \frac{s_k s_k^T}{s_k^T \Delta g_k}, \tag{12.6.21}$$

where $s_k := \mathbf{x}_{k+1} - \mathbf{x}_k$, $\Delta g_k = g_{k+1} - g_k$. The algorithm is straightforward to given as follows:

Algorithm 12.6.9. *(BFGS algorithm)*

Step 1 Choose $\mathbf{x}_0$, B_0, $\epsilon > 0$, *and a symmetric positive definite starting matrix* B_0*; Set* $k := 0$.
Step 2 If $\|\text{grad}_k[J^\alpha]\| \leq \epsilon$, *STOP; Otherwise, GOTO Step 3.*
Step 3 Iterates according to formulas (12.6.20)–(12.6.21).
Step 4 Set $k := k + 1$, *update* B_k *and* B_k^{-1} *and GOTO Step 2.*

12.6.2.4 L-BFGS method

This is a limited memory BFGS method, which is an efficient method for large scale scientific computing. The key of the method lies in that it does not need to store H_k in each iteration [30] The limited memory strategy avoids two of the bottlenecks specified above in the Newton algorithm when applied to big systems, the storage and the inversion of big matrices. Here although a quasi-Newton algorithm is used, the inverse of the Hessian matrix is never built up, but directly the product of the inverse of the Hessian by the gradient. Then, no Hessian diagonalization is required. What makes this method powerful is that in order to update this matrix product only information of last m steps is used. In this way only the geometry and gradient of the last m steps have to be stored. When a BFGS update formula is used this procedure is called L-BFGS. This method was developed by [18, 15]

$$H_{k+1} = \Big(I - \frac{s_k y_k^T}{s_k^T y_k}\Big) H_k \Big(I - \frac{y_k s_k^T}{s_k^T y_k}\Big) + \frac{s_k s_k^T}{s_k^T y_k}. \tag{12.6.22}$$

The source code can be obtained free of charge from the web.

Now we give a precise description of the L-BFGS method. We follow the description as is in the form [30]

$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T, \tag{12.6.23}$$

where $\rho_k = \frac{1}{s_k^T y_k}$, and

$$V_k = I - \rho_k y_k s_k^T.$$

Then, next iterative point $\mathbf{x}_{k+1}$ can be computed by

$$\begin{aligned} d_k &= -H_k g_k, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k d_k. \end{aligned} \tag{12.6.24}$$

Usually the L-BFGS method is implemented with a line search for the step length α_k to ensure a sufficient decrease of the misfit function. Assume that $\mathbf{x}_{k+1}^*$ is an approximate solution for problem (12.6.15) at $\mathbf{x}_k$. Convergence properties of the L-BFGS method are guaranteed if the steplength α_k in equation (12.6.24) satisfies the Wolfe line search conditions along $d_k = \mathbf{x}_{k+1}^* - \mathbf{x}_k$[13]

$$q(\mathbf{x}_k + \alpha_k d_k) \leq q(\mathbf{x}_k) + \gamma_1 \alpha_k g_k^T d_k, \tag{12.6.25}$$

$$|g(\mathbf{x}_k + \alpha_k d_k)^T d_k| \leq \gamma_2 |g(\mathbf{x}_k)^T d_k|, \tag{12.6.26}$$

where γ_1 and γ_2 are constants to be chosen *a priori*. The line search condition can ensure the iterates remaining in the feasible region.

We give the L-BFGS algorithm as follows:

Algorithm 12.6.10. *(L-BFGS algorithm)*

Step 1 *Choose* $\mathbf{x}_0$, m, $0 < \gamma_1 < \frac{1}{2}$, $\gamma_1 < \gamma_2 < 1$, *and a symmetric positive definite starting matrix* H_0; *Set* $k := 0$.

Step 2 *If the stopping rule satisfied, STOP; Otherwise, GOTO Step 3.*

Step 3 *Compute* d_k *and let* α_k *satisfies the above Wolfe conditions (12.6.25)–(12.6.26); Compute* $\mathbf{x}_{k+1}$ *by equation (12.6.24)*

Step 4 *Let* $\hat{m} = \min\{k, m-1\}$, *check if* $y_k^T s_k > 0$.
If NO: $H_{k+1} = I$ *(steepest descent step) and delete the pairs*

$$\{y_i, s_i\}_{i=k-\hat{m}}^{k};$$

If YES: Update H_0 $\hat{m}+1$ *times using the pairs* $\{y_i, s_i\}_{i=k-\hat{m}}^{k}$, *i.e., let*

$$\begin{aligned} H_{k+1} = {} & (V_k^T V_{k-1}^T \cdots V_{k-\hat{m}}^T) H_0 (V_{k-\hat{m}} \cdots V_{k-1} V_k) \\ & + \rho_{k-\hat{m}} (V_k^T V_{k-1}^T \cdots V_{k-\hat{m}+1}^T) s_{k-\hat{m}} s_{k-\hat{m}}^T (V_{k-\hat{m}+1} \cdots V_{k-1} V_k) \\ & + \rho_{k-\hat{m}+1} (V_k^T V_{k-1}^T \cdots V_{k-\hat{m}+2}^T) s_{k-\hat{m}+1} s_{k-\hat{m}+1}^T (V_{k-\hat{m}+2} \cdots V_{k-1} V_k) \\ & \vdots \\ & + \rho_k s_k s_k^T. \end{aligned}$$

Step 5 *Set* $k := k+1$ *and GOTO Step 2.*

In Step 1 of Algorithm 12.6.10, the initial guess for the Hessian H_0 is the identity matrix I. In the algorithm proposed by Liu and Nocedal (1989), the initial symmetric positive definite H_0 need to be scaled as follows:

$$H_0' = \frac{y_0^T s_0}{\|y_0\|^2} H_0 = \frac{y_0^T s_0}{\|y_0\|^2} I.$$

Then after one iteration is completed, all methods update H_0' instead of H_0. This scaling greatly improves the performances of the method.

In Step 4, the condition $y_k^T s_k > 0$ guarantees the positive definiteness of the L-BFGS matrix. However, this is not always the case[7]. If it is violated, a simple reparation is making a steepest descent step.

We note that in the L-BFGS method, the storing of the matrices H_k is unnecessary, instead, a prefixed number (say m) of vectors pairs $\{s_k, y_k\}$ that define them implicitly are stored. Therefore, during the first m iterations the L-BFGS and the BFGS methods are identical, but when $k > m$ only information from the m previous iterations is used to obtain H_k. The number m of BFGS corrections that must be kept can be specified by the users. Moreover, in the L-BFGS the product $H_k g_k$ which represents the search direction is obtained by means of a recursive formula involving g_k and the most recent vectors pairs $\{s_k, y_k\}$.

12.6.2.5 Trust region method

It is the best optimization method thus far for nonlinear programming problems. Its global convergence and regularity are proved recently[27]. In trust region method, one adjusts the trust region radius instead of adjusting the Levenberg-Marquardt parameter. The trust region method is indirectly solving the minimal least squares problem

$$\min_{x \in X} J(x) := \frac{1}{2}\|K(x) - y\|^2. \tag{12.6.27}$$

Instead, it takes a trial step from a subproblem at each iteration

$$\min_{\xi \in X} \psi_k(\xi) := (\mathrm{grad}_k[J], \xi) + \frac{1}{2}(\mathrm{Hess}_k[J]\xi, \xi), \tag{12.6.28}$$

$$s.\ t.\ \|\xi\| \leq \Delta_k, \tag{12.6.29}$$

where $\mathrm{grad}_k[J]$ is the gradient of J at the k-th iterate,

$$\mathrm{grad}[J(x)] = K'(x)^*(K(x) - y_\delta), \tag{12.6.30}$$

$Hess(J)_k$ is the Hessian of J at the k-th iterate,

$$\mathrm{Hess}[J(x)] = K'(x)^* K'(x) + K''(x)^*(K(x) - y), \tag{12.6.31}$$

and Δ_k is the trust region radius. The trust region subproblem (TRS) (12.6.28)–(12.6.29) is an approximation to the original optimization problem (12.6.27) with a trust region constraint which prevents the trial step becoming too large. A trust region algorithm generates a new point which lies in the trust region, and then decides whether it accepts the new point or rejects it. At each iteration, the trial step ξ_k is normally calculated by solving the "trust region subproblem" (12.6.28)–(12.6.29). Here $\Delta_k > 0$ is a trust region radius. Generally, a trust region algorithm uses

$$r_k = \frac{Ared_k}{Pred_k} \tag{12.6.32}$$

to decide whether the trial step ξ_k is acceptable and how the next trust region radius is chosen, where

$$Pred_k = \psi_k(0) - \psi_k(\xi_k) \tag{12.6.33}$$

is the predicted reduction in the approximate model, and

$$Ared_k = J(\mathbf{x}_k) - J(\mathbf{x}_k + \xi_k) \tag{12.6.34}$$

is the actual reduction in the objective functional.

Now we give the trust region algorithm for solving nonlinear ill-posed problems.

Algorithm 12.6.11. *(Trust region algorithm)*

STEP 1 *Choose parameters* $0 < \tau_3 < \tau_4 < 1 < \tau_1$, $0 \leq \tau_0 \leq \tau_2 < 1$, $\tau_2 > 0$ *and initial values* $\mathbf{x}_0$, $\Delta_0 > 0$*; Set* $k := 1$.

STEP 2 *If* **the stopping rule is satisfied** *then STOP; Else, solve (12.6.28)–(12.6.29) to give* ξ_k.

STEP 3 *Compute* r_k*;*

$$\mathbf{x}_{k+1} = \begin{cases} \mathbf{x}_k & if\ r_k \leq \tau_0, \\ \mathbf{x}_k + \xi_k & otherwise. \end{cases} \tag{12.6.35}$$

Choose Δ_{k+1} *that satisfies*

$$\Delta_{k+1} \in \begin{cases} [\tau_3\|\xi_k\|, \tau_4\Delta_k] & if\ r_k < \tau_2, \\ [\Delta_k, \tau_1\Delta_k] & otherwise. \end{cases} \tag{12.6.36}$$

STEP 4 *Evaluate* $\text{grad}[J]_k$ *and* $\text{Hess}[J]_k$*; k:=k+1; GOTO STEP 2.*

In STEP 2, the stopping rule is based on the discrepancy principle, i.e., *the iteration should be terminated at the first occurrence of the index k such that*

$$\|K(x_k) - y\| \leq \tau\delta, \tag{12.6.37}$$

where $\tau > 1$ *is the dominant parameter and can be chosen by users.*

The constants $\tau_i\,(i = 0, \cdots, 4)$ can be chosen by users. Typical values are $\tau_0 = 0$, $\tau_1 = 2$, $\tau_2 = \tau_3 = 0.25$, $\tau_4 = 0.5$. The parameter τ_0 is usually zero or a small positive constant. The advantage of using zero τ_0 is that a trial step is accepted whenever the objective function is reduced. When the objective function is not easy to compute, it seems that we should not throw away any "good" point that reduces the objective function (see [33] for details).

For linear inverse problems, the trust region method is straightforward. The nonlinear model (12.6.27) reduces to the quadratic model

$$\min_{\mathbf{x}\in X} J(\mathbf{x}) := \frac{1}{2}\|K\mathbf{x} - \mathbf{y}\|^2. \tag{12.6.38}$$

The gradient and Hessian of J can be written as

$$\text{grad}_{\mathbf{x}}[J(\mathbf{x})] = (K^T K)^{-1}\mathbf{x} - K^T\mathbf{y}$$

and

$$\text{Hess}_{\mathbf{x}}[J(\mathbf{x})] = K^T K$$

respectively. Note that (12.6.38) is the quadratic model, hence the ration r_k is always equal to 1. Therefore, if we use the trust region scheme directly, the trial step is always accepted. The regularization is straightforward if we stop the iteration step under proper termination criterion [29]. We note that the approximate accuracy is characterized by the discrepancy between the observation and the true data; therefore variations of the norm of the discrepancy may reflect the degree of approximation. Based on these considerations, we propose to accept or reject the trial step s_k at the k-th step by the ratio

$$\rho_k = \frac{J(\mathbf{x}_{k+1})}{J(\mathbf{x}_k)} = \frac{J(\mathbf{x}_k + \xi_k)}{J(\mathbf{x}_k)},$$

where $J(\mathbf{x}_{k+1})$ and $J(\mathbf{x}_k)$ are the reductions in norm of the discrepancy at k+1-th and k-th steps, respectively. Therefore, Algorithm 12.6.11 can be easily reformulated based on the above comments, we leave it for readers.

12.7 Numerical examples

Denote by M the number of measurements in the kernel-based models. Then the linear kernel-based BRDF model can be rewritten in the following matrix-vector form

$$K\mathbf{x} = \mathbf{y}, \tag{12.7.1}$$

where

$$K = \begin{bmatrix} 1 & k_{geo}(1) & k_{vol}(1) \\ 1 & k_{geo}(2) & k_{vol}(2) \\ \vdots & \vdots & \vdots \\ 1 & k_{geo}(M) & k_{vol}(M) \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} f_{iso} \\ f_{geo} \\ f_{vol} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_M \end{bmatrix},$$

in which, $k_{geo}(k)$ and $k_{vol}(k)$ represent the values of kernel functions $k_{geo}(t_i, t_v, \phi)$ and $k_{vol}(t_i, t_v, \phi)$ corresponding to the k-th measurement for $k = 1, 2, \cdots$; r_k represents the k-th observation for $k = 1, 2, \cdots$.

We use the combination of RossThick kernel and LiTransit kernel in the numerical tests. In practice, the coefficient matrix K cannot be determined accurately, and a perturbed version $\tilde{K}$ is obtained instead. Also instead of the true measurement $\mathbf{y}$, the observed measurement $\mathbf{y_n} = \mathbf{y} + \mathbf{n}$ is the addition of the true measurement $\mathbf{y}$ and the noise $\mathbf{n}$, which for simplicity is assumed to be additive Gaussian random noise. Therefore it suffices to solve the following operator equation with perturbation

$$\tilde{K}\mathbf{x} = \mathbf{y_n},$$

where $\tilde{K} := K + \delta B$ for some perturbation matrix B and δ denotes the noise level (upper bound) of $\mathbf{n}$ in (0,1). In our numerical simulation, we assume that B is a Gaussian random matrix, and also that

$$\|\mathbf{y_n} - \mathbf{y}\| \leq \delta < \|\mathbf{y_n}\|. \tag{12.7.2}$$

The above assumption of the noise can be interpreted as that the signal-to-noise ratio (SNR) should be greater than 1. We make such an assumption as we believe that observations (BRDF) are not trustable otherwise. It is clear that (12.7.1) is an underdetermined system if $M \leq 2$ and an overdetermined system if $M > 3$. Note that for satellite remote sensing, because of the restrictions in view and illumination geometries, $\tilde{K}^T\tilde{K}$ need not have bounded inverse [8, 25, 14]. We believe that the proposed regularization method can be employed to find an approximate solution $\mathbf{x}_\alpha^\dagger$ satisfying

$$\|\tilde{K}\mathbf{x}_\alpha^\dagger - \mathbf{y_n}\| \longrightarrow \min.$$

In this test, we choose the widely used 73 data sets referred to in [14]. Among the 73 sets of BRDF measurements, only 18 sets of field-measured BRDF data with detailed information about the experiments are known, including biophysical and instrumental information [10, 3, 4, 5, 6, 11, 12]. These data sets cover a large variety of vegetative cover types, and are fairly good representative of the natural and cultivated vegetation. Table 12.1 summarizes the basic properties of the data sets used in this chapter. For those selected field data, the observations are sufficient. If the kernel matrix $\tilde{K}$ is well-conditioned, the problem is well-posed, and the regularization is unnecessary, which can also be considered a regularization procedure with zero regularization parameter. Even for sufficient observations, most of the kernel matrices $\tilde{K}$ could be ill-conditioned, and hence the problem is discretely ill-posed. In that situation, one has to resort to regularization or preconditioning to make the proposed algorithm workable. Our proposed method is quite adaptive no matter whether the kernel matrix $\tilde{K}$ is ill-conditioned or well-conditioned.

With the nice field data, the numerical experiment can be done to test the robustness of the algorithm by using a limited number of observations, and then to compare the retrieval with the measurements. The ill-posed situations are generated by significantly reducing the number of observations from the field data in Table 12.1.

We choose one or two observations as limited number of observations and compare the retrieval results by different regularization methods [22, 20]. Comparison results are given in Tables 12.2 and 12.3. We do not list all of the results since the data sets can be used to generate an enormous number of such kinds of ill-posed situations. In the two tables, these methods are denoted by NTSVD, Tikh(a) ($\alpha = \delta^2$), Tikh(b) (D is in the form of D_1 and α is chosen by *a posteriori* method addressed in Section 12.5.1), Tikh(c) (D is in the form D_2 and α is

chosen by *a posteriori* method addressed in Section 12.5.1), Tikh(d) (D is in the form D_3 and α is chosen by *a posteriori* method addressed in Section 12.5.1) and Tikh(e) (D is in the form D_4 and α is chosen by *a posteriori* method addressed in Section 12.5.1). The true white sky albedo (WSA) is calculated from well-posed situations using AMBRALS, i.e., full observation data. It deserves pointing out that the standard operational algorithm used in AMBRALS does not work for such severely ill-posed situations. If we regard WSA > 1 or WSA < 0 as failed inversion, it is clear that our proposed method works for all of the cases. It follows from the experiments that our method (Tikh(b)) (proposed in [22]) works for a single observation, and performs better than the NTSVD (proposed in [20]) and the standard Tikhonov regularization with *a priori* choice of the regularization parameter (Tikh(a)).

Table 12.1 Data sets used in the simulation.

Data	**Cover Type**	**LAI**
ranson_soy.827	Soy	2.9
kimes.orchgrass	Orchard grass	1
Parabola.1994.asp-ifc2	Aspen	5.5

Table 12.2 Comparison of computational values of the WSAs from data sets in Table 12.1 for single observation and for two observations with the true WSAs values (multiangular observations) for VisRed band.

	Methods	Single Observation	Two Observations	True WSAs
ranson_soy.827	NTSVD	0.0449047	0.0442712	0.0405936
	Tikh(a)	0.0638937	0.0386419	
	Tikh(b)	0.0401528	0.0560726	
	Tikh(c)	0.0633967	0.0590594	
	Tikh(d)	-0.0009147	0.0539707	
	Tikh(e)	0.0476311	0.0583683	
kimes.orchgrass	NTSVD	0.1082957	0.1058740	0.0783379
	Tikh(a)	0.0397185	0.0860485	
	Tikh(b)	0.0753925	0.1214918	
	Tikh(c)	0.26211583	0.4365220	
	Tikh(d)	-0.0018020	0.0383555	
	Tikh(e)	0.1137684	0.1707774	
Parabola.1994.asp-ifc2	NTSVD	0.0364620	0.0389198	0.0227972
	Tikh(a)	0.0447834	-0.0040831	
	Tikh(b)	0.0262501	0.0102457	
	Tikh(c)	0.0798633	-0.0874783	
	Tikh(d)	-0.0006110	-0.0401510	
	Tikh(e)	0.0375009	0.0547068	

Table 12.3 Comparison of computational values of the WSAs from the data sets in Table 12.1 for single observation and for two observations with the true WSAs values (multiangular observations) for Nir band.

	Methods	Single Observation	Two Observations	True WSAs
ranson_soy.827	NTSVD	0.4469763	0.4348320	0.3653728
	Tikh(a)	0.6359822	0.4195730	
	Tikh(b)	0.3996775	0.5439493	
	Tikh(c)	0.6310461	0.9247240	
	Tikh(d)	-0.0091045	-0.0098136	
	Tikh(e)	0.4741162	0.6277249	
kimes.orchgrass	NTSVD	0.3890207	0.37216767	0.2963261
	Tikh(a)	0.2048903	0.2945934	
	Tikh(b)	0.2708260	0.4458619	
	Tikh(c)	0.9415755	1.8140732	
	Tikh(d)	-0.0064732	0.1927318	
	Tikh(e)	0.4086801	0.6015300	
Parabola.1994.asp-ifc2	NTSVD	0.5517209	0.5741842	0.4240376
	Tikh(a)	0.6776356	-0.0617838	
	Tikh(b)	0.3972022	0.2398577	
	Tikh(c)	1.2084479	-0.8953630	
	Tikh(d)	-0.0092437	-0.4071125	
	Tikh(e)	0.5674424	0.8185223	

Our algorithms to the Landsat Thematic Mapper (TM) data measured in Shunyi District of Beijing, China. The TM sensor is an advanced, multispectral scanning, Earth resources instrument designed to achieve higher image resolution, sharper spectral separation, improved geometric fidelity, and greater radiometric accuracy and resolution. Fig. 12.1 plots the reflectance for band 5 on May 17, 2001. The spatial resolution for the TM sensor on band 5 is 30 meters. The white-sky albedo (WSA) retrieved by Tikh(b) for band 5 of one observation on May 17, 2001 is plotted in Fig. 12.2 The retrieved results show that our algorithms work for satellite data with high spatial resolutions.

Experimental results of different data sets indicate that our proposed regularization method is feasible for ill-posed land surface parameter retrieval problems.

We want to emphasize that our method can generate smoothing data to help retrieval of parameters once sufficient observations are unavailable. As we have pointed out in [22, 20], we do not suggest discarding the useful history information (e.g., data that is not too old) and the multiangular data. Instead, we should fully employ such information if it is available. The key to why our algorithm outperforms previous algorithms is that our algorithm is adaptive, accurate and very stable, which solves kernel-based BRDF model of any order, which may be a supplement for BRDF/albedo retrieval product.

For the remote sensor MODIS, which can generate a product by using 16 days different observations data, this is not a strict restriction for MODIS, since it

Fig. 12.1 Reflectance for band 5 of Landsat Thematic Mapper Data (TM) on May 17, 2001.

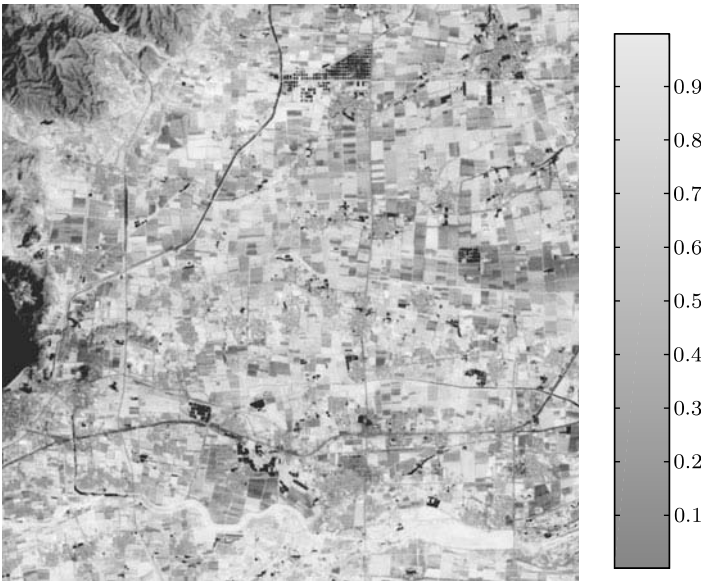


Fig. 12.2 White-sky albedo retrieved by proposed Tikhonov regularization method for band 5 of Landsat Thematic Mapper Data (TM) on May 17, 2001.

aims at global exploration. For other sensors, the period for their detection of the same area will be longer than 20 days or more. Therefore, for vegetation in the growing season, the reflectance and albedos will change significantly. Hence robust algorithms to estimate BRDF and albedos in such cases are highly desired. Our algorithm is a proper choice, since it can generate retrieval results which quite approximate the true values of different vegetation types of land surfaces by capturing just one time of observation.

Moreover, for some sensors with high spatial resolution, the quasi multiangular data is impossible to obtain. This is why there are not high resolution albedo products. But with our algorithm, we can achieve the results. This is urgently needed in real applications.

12.8 Conclusions

In this chapter, we investigate the regularization and optimization methods for the solution of the kernel-based remotely sensed inverse problems. We reformulate the problem in functional space by introducing the first kind operator equations, then the solution methods in l^1 and l^2 spaces are considered. The regularization strategy and optimization solution techniques are fully described. The equivalence between the Tikhonov regularization and Baysian statistical inference is established. The noise/error propagation for kernel-based model problems is deduced. We want to emphasize that there are different ways to impose *a priori* information [29]. For example, (P1) the unknowns $\mathbf{x}$ can be bounded. This method requires a good apriori upper bound for $\mathbf{x}$; (P2) applying different weights to the components of $\mathbf{x}$, then constructing the kernel model under the constraint of the weights; (P3) imposing historical information on $\mathbf{x}$ provided that such historical information exists; (P4) simplifying the physical model by solving a l^p norm problem, which means the unknowns $\mathbf{x}$ can be obtained at the l^p scale.

Acknowledgements This work was supported by the National Natural Science Foundation of China under grant No.10871191 and National "973" Key Basic Research Developments Program of China under grant numbers 2005CB422104 and 2007CB714400.

References

1. J. Barzilai and J. Borwein, Two-point step size gradient methods, *IMA Journal of Numerical Analysis*, **8**, 141–148, 1988.
2. H. Brakhage, On ill-posed problems and the method of conjugate gradients, in: H. W. Engl and C. W. Groetsch, eds., *Inverse and Ill-Posed Problems*, Academic Press, Boston, 165–175, 1987.
3. D. W. Deering, T. F. Eck and T. Grier, Shinnery oak bidirectional reflectance properties and canopy model inversion, *IEEE Trans. Geosci. Remote Sensing*, **30**(2), 339–348, 1992.

4. D. W. Deering, E. M. Middleton and T. F. Eck, Reflectance anisotropy for a spruce-hemlock forest canopy, *Remote Sens. of Environ.*, **47**, 242–260, 1994.
5. D. W. Deering, S. P. Shmad, T. F. Eck and B. P. Banerjee, Temporal attributes of the bidirectional reflectance for three forest canopies, *International Geoscience And Remote Sensing Symposium (IGARSS'95)*, 1239–1241, 1995.
6. T. F. Eck and D. W. Deering, Spectral bidirectional and hemispherical reflectance chacteristics of selected sites in the streletskay steppe, *Proceedings of the 1992 International Geoscience and Remote Sensing Symposium*, IEEE Geosci. and Remote Sens. Soc., New Jersey, 1053–1055, 1992.
7. R. Fletcher, On the Barzilai-Borwein method, *Numerical Analysis Report NA/207*, 2001.
8. F. Gao, A. H. Strahler, W. Lucht, Z. Xia and X. Li, Retrieving albedo in small sample size, *IEEE Int. Geosci. Remote Sens. Symp. Proc. 1998*, **5**, 2411–2413, 1998.
9. N. S. Geol, Models of vegetation canopy reflectance and their use in estimation of biophysical parameters from reflectance data, *Remote Sensing Reviews*, **4**, 1–212, 1988.
10. B. X. Hu, W. Lucht, X. W. Li and A. H. Strahler, Validation of kernel-driven semiempirical models for the surface bidirectional reflectance distribution function of land surfaces, *Remote Sensing of Environment*, **62**, 201–214, 1997.
11. D. S. Kimes, W. W. Newcomb and C. J. Tucker, Directional reflectance factor distributions for cover types of Northen Africa, *Remote Sensing of Environment*, **18**, 1–19, 1985.
12. D. S. Kimes, W. W. Newcomb and R. F. Nelson, Directional reflectance distributions of a hardwood and a pine forest canopy, *IEEE Transactions on Geoscience and Remote Sensing*, **24**, 281–293, 1986.
13. X. Li, F. Gao, Q. Liu, J. D. Wang and A. H. Strahler, Validation of a new GO kernel and inversion of land surface albedo by kernel-driven model (1), *Journal of Remote Sensing*, **4**(appl.), 1–7, 2000.
14. X. Li, F. Gao, J. Wang and A. H. Strahler, *A priori* knowledge accumulation and its application to linear BRDF model inversion, *J. Geophys. Res.*, **106**(D11), 11925–11935, 2001.
15. X. Li, J. Wang, B. Hu and A. H. Strahler, On unitilization of *a priori* knowledge in inversion of remote sensing models, *Science in China D*, **41**, 580–585, 1998.
16. X. Li, J. Wang and A. H. Strahler, Apparent reciprocal failure in BRDF of structured surfaces, *Progress of Natural Sciences*, **9**, 747–750, 1999.
17. O. Pokrovsky and J. L. Roujean, Land surface albedo retrieval via kernel-based BRDF modeling: I. statistical inversion method and model comparison, *Remote Sensing of Environment*, **84**, 100–119, 2002.
18. O. M. Pokrovsky and J. L. Roujean, Land surface albedo retrieval via kernel-based BRDF modeling: II. an optimal design scheme for the angular sampling, *Remote Sens. Environ.*, **84**, 120–142, 2003.
19. I. O. Pokrovsky, O. M. Pokrovsky and J. L. Roujean, Development of an operational procedure to estimate surface albedo from the SEVIRI/MSG observing system by using POLDER BRDF measurements: II. Comparison of several inversion techniques and uncertainty in albedo estimates, *Remote Sensing of Environment*, **87**(2–3), 215–242, 2003.
20. J. L. Privette, T. F. Eck and D. W. Deering, Estimating spectral albedo and nadir reflectance through inversion of simple bidirectional reflectance distribution models with AVHRR/MODIS-like data, *J. Geophys. Res.*, **102**, 29529–29542, 1997.
21. J. L. Roujean, M. Leroy and P. Y. Deschamps, A bidirectional reflectance model of the Earth's surface for the correction of remote sensing data, *J. Geophys. Res.*, **97**, 20455–20468, 1992.
22. A. H. Strahler, X. W. Li, S. Liang, J.-P. Muller, M. J. Barnsley and P. Lewis, *MODIS BRDF/Albedo Product: Algorithm Technical Basis Document*, NASA EOS-MODIS Doc., Vol. 2.1, 55pgs., 1994.

23. A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, New York, John Wiley and Sons, 1977.
24. A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, Dordrecht, Kluwer, 1995.
25. M. M. Verstraete, B. Pinty and R. B. Myneny, Potential and limitations of information extraction on the terrestrial biosphere from satellite remote sensing, *Remote Sensing Environment*, **58**, 201–214, 1996.
26. Y. F. Wang and Y. X. Yuan, A trust region algorithm for solving distributed parameter identification problem, *Journal of Computational Mathematics*, **21**(6), 759–772, 2003.
27. Y. F. Wang and Y. X. Yuan, Convergence and regularity of trust region methods for nonlinear ill-posed inverse problems, *Inverse Problems*, **21**, 821–838, 2005.
28. Y. F. Wang, S. F. Fan, X. Feng, G. J. Yan and Y. N. Guan, Regularized inversion method for retrieval of aerosol particle size distribution function in W space, *Applied Optics*, **45**(28), 7456–7467, 2006.
29. Y. F. Wang, *Computational Methods for Inverse Problems and Their Applications*, Beijing, Higher Education Press, 2007.
30. Y. F. Wang and S. Q. Ma, Projected Barzilai-Borwein methods for large scale nonnegative image restorations, *Inverse Problems in Science and Engineering*, **15**, 559–583, 2007.
31. Y. F. Wang, C. C. Yang and X. W. Li, A Regularizing Kernel-Based BRDF Model Inversion Method for Ill-posed Land Surface Parameter Retrieval Using Smoothness Constraint, *Journal of Geophysical Research*, **113**, D13101, doi:10.1029/2007JD009324, 2008.
32. W. Wanner, X. Li and A. H. Strahler, On the derivation of kernels for kernel-driven models of bidirectional reflectance, *J. Geophys. Res.*, **100**, 21077–21090, 1995.
Xiao T. Y. Xiao, S. G. Yu and Y. F. Wang, *Numerical Methods for the Solution of Inverse Problems*, Beijing, Science Press, 2003.
33. Y. X. Yuan, Nonlinear programming: trust region algorithms, in: S. T. Xiao and F. Wu, eds., *Proceedings of Chinese SIAM Annual Meeting*, Beijing, Tsinghua University Press, 83–97, 1994.

Chapter 13
Pseudo-Differential Operator and Inverse Scattering of Multidimensional Wave Equation

Hong Liu, Li He

Abstract. The infra-structural feature of multidimensional inverse scattering for wave equation is discussed in this chapter. Previous studies on several disciplines pointed out that the basic frame of multidimensional inverse scattering for wave equation is much similar to the one-dimensional (1D) case. For 1D wave equation inverse scattering problem, four procedures are included, i.e., time-depth conversion, Z transform, 1D spectral factorization and conversion of reflection and transmission coefficient to the coefficient of wave equation. In multidimensional or in the lateral velocity varying situation, the conceptions of 1D case should be replaced by image ray coordinate, one-way wave operator, multidimensional spectral factorization based on Witt production and the plane wave response of reflection and transmission operator. There are some important basic components of multidimensional inverse scattering problem, namely, effective one-way operator integral representation, differential form of wave equation in ray coordinate, wide application of Witt product and the modern development of multidimensional spectral factorization. The example of spectrum factorization shows that the energy is well focused, which may benefit the velocity analysis and the pickup of the reflection coefficients.

Hong Liu
Key Laboratory of Petroleum Geophysics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China.
e-mail: liuhong@mail.igcas.ac.cn

Li He
Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China.
e-mail: heli@mail.igcas.ac.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

13.1 Introduction

The 3D inverse scattering method is important in seismic propagation and imaging mute multiples and multi-scattering. For a long time, 1D wave inverse scattering methods show a good example to the solution of 3D problem. For 1D wave equation inverse problem, Schrodinger equation plays a key rule in time-depth convention. And, there are many other solving methods, such as Gilfend-Levitan equation method, Marchenko equation method, Gopinash-Sandhi equation method (Bube K P and Burridge R, 1983; Liu H, Li Y M and Wu R S, 1994; Song H B, 2002), Weiner-Hopf technique (Whittle factorization, Kolmogorov factorization) (Claerbout, 1985) and Wilson-Burg factorization. Spectral factorization, actually, is to express the reflected signal to the autocorrelation of prediction operators (auto-regression operator). The exponential form of prediction operators and exponential form of general transmission operators are related to stratigraphic filter (O'Doherty R F and Anstey N A, 1971) and seismic rock physics of fracture and porous medium.

Geophysicists have sought for solving inverse scattering problem of multidimensional wave equation or lateral velocity variation media since last two decades (Rose J H and DeFacio B, 1987; Weglein A, Foster D and Matson K, 2001). In recent years, some methods, such as multiples canceling (Matson K 2000, Verschuur D J and Berkhout A J, 2005; Berkhout A J and Verschuur D J, 2005) and multidimensional scattering methods (Prosser R T, 1969; Newton R G, 1982; Rose J H and DeFacio B, 1987; Cheney M and Rose J H, 1988; Gray S H, 1988; Rose J H, 1989; Chen X F, 1990) and decomposition in directional wave (de Hoop M V, 1996; Weglein A, Foster D and Matson K, 2001), especially in the study of reflection operator property and decomposition in directional wave (Fryer G J and Frazer L N, 1984, 1987; Wapenaar C P A and Grimbergen J L T, 1996; Song H B, 2002), require the infra-structural study in inverse scattering method for multidimensional wave equation. In this chapter, we review and integrate the previous researches and point out that the methods are much similar to 1D wave equation problem. In the inverse scattering method for multidimensional wave equation, time-depth conversion, Z transform, 1D spectral factorization and reflection transmission coefficient will be replaced by ray-coordinate (Larner K L, Hatton L and Gibson B S, 1981; Sava P and Fomel S, 2005; Iversen E and Tygel M, 2008; Shragge J C, 2008), one-way operator (implicit scheme such as finite-difference with LU decomposition on helix coordinator and multi-way splitting, explicit scheme generalized screen, separable kernel representation, time migration, residual migration, cascade migration and true amplitude migration) (Claerbout J F, 1985; Claerbout J F and Fomel S, 2002), multidimensional spectral factorization and the plane wave response of reflection operator.

13.2 Notations of operators and symbols

Though pseudo-differential operator theory has a very wide application, we concentrate symbol operation on Witt product (Qi M Y, Xu C J, Wang W K 2005, Liu H, Wang X M and Zeng R, 2007) in this chapter. Thus, the "pseudo-differential theory" means "symbol operation". Symbol of the operators is its plane wave response and its multiplication is similar to function multiplications. If the velocity does not vary horizontally, they are the same; otherwise, the multiplication of symbol should be added a series of derivative modification. And it is an important tool in structure preserving algorithm to keep the efficiency and accuracy in the lateral velocity varying case. So far, exponent mapping and logarithm mapping of the symbols have been found, from which we can derive Dix-like travel-time formula and the modification of the spectral factorization.

A symbol $\sigma(\hat{A})$ of the operator $\hat{A}$ is a function $a(x, y, k_x, k_y)$, which is the sine wave response under the pseudo-differential operator, i.e.,

$$a(x, y, k_x, k_y) \equiv \sigma(\hat{A}) = \exp(-ik_x x - ik_y y)(\hat{A}(x, y, D_x, D_y) \exp(ik_x x + ik_y y)), \tag{13.2.1}$$

where k_x, k_y are the wave numbers of the plane-wave, and

$$D_x = \frac{\partial}{i\partial x} = \frac{1}{i}\partial_x\,,\ D_y = \frac{\partial}{i\partial y} = \frac{1}{i}\partial_y. \tag{13.2.2}$$

The hat in $\hat{A}$ stands for operator, which will be used in the following context. Setting the symbol functions of $\hat{A}$ and $\hat{B}$ as $a = \sigma(\hat{A})$ and $b = \sigma(\hat{B})$, the multiplication of symbols can be written as

$$a\#b \equiv \sigma(\hat{A}\hat{B}) = \sum \frac{1}{n!}\left[\frac{\partial^a}{\partial k_x} D_x^b + \frac{\partial^a}{\partial k_y} D_y^b\right]^n ab, \tag{13.2.3}$$

where # is symbol notation, D_x^a means that D_x only operates on a and D_x^b means that D_x only operates on b. This is Witt product according to Leibniz rule, which plays a key role in exponent mapping's expression. Based on Witt product, we can develop symbols of operator, such as multiplication, division, power, exponential (Liu H, Wang X M and Zeng R, 2007), logarithm, vector normalization and matrix diagonalization.

The definition of communicator is

$$[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}, \tag{13.2.4}$$

$$[\hat{A}, \hat{B}, \hat{C}] = [\hat{A}, [\hat{B}, \hat{C}]], \tag{13.2.5}$$

$$ad_{\hat{A}}\hat{B} \equiv [\hat{A}, \hat{B}], \tag{13.2.6}$$

and the symbol of the communicator is defined as follows:

$$\{a, b\} = a\#b - b\#a = \sigma([\hat{A}, \hat{B}]) = \sigma(\hat{A}\hat{B} - \hat{B}\hat{A}). \tag{13.2.7}$$

In travel time and amplitude calculation of high frequency approximation, first order and zeroth order homogeneous terms of ω, k_x and k_y should be called the kinematic symbol (control travel time) and amplitude-controlling symbol (control amplitude), respectively. Letting comma in subscripts of the following formulas stand for derivative, we have

$$u_{,ix} = D_x u, \tag{13.2.8}$$

$$u_{,iy} = D_y u, \tag{13.2.9}$$

$$u_{,k_x} = \frac{\partial}{\partial k_x} u, \tag{13.2.10}$$

$$u_{,k_y} = \frac{\partial}{\partial k_y} u, \tag{13.2.11}$$

and the Newton's law for the elastic media can be written as

$$-\rho\omega^2 \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} = \begin{pmatrix} \partial_x \tau_{xx} + \partial_y \tau_{xy} + \partial_z \tau_{xz} \\ \partial_x \tau_{xy} + \partial_y \tau_{yy} + \partial_z \tau_{yz} \\ \partial_x \tau_{xz} + \partial_y \tau_{yz} + \partial_z \tau_{zz} \end{pmatrix} \tag{13.2.12}$$

where $\begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix}$ is displacement vector, ρ is density. And the Hooke's law can be written as

$$\begin{pmatrix} \tau_{xx} \\ \tau_{yy} \\ \tau_{zz} \\ \tau_{yz} \\ \tau_{zx} \\ \tau_{xy} \end{pmatrix} = M \begin{pmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \\ 2e_{yz} \\ 2e_{zx} \\ 2e_{xy} \end{pmatrix} = M \begin{pmatrix} \partial_{\mathrm{x}} & 0 & 0 \\ 0 & \partial_{\mathrm{y}} & 0 \\ 0 & 0 & \partial_{\mathrm{z}} \\ 0 & \partial_{\mathrm{z}} & \partial_{\mathrm{y}} \\ \partial_{\mathrm{z}} & 0 & \partial_{\mathrm{x}} \\ \partial_{\mathrm{y}} & \partial_{\mathrm{x}} & 0 \end{pmatrix} \begin{pmatrix} \mathrm{u_x} \\ \mathrm{u_y} \\ \mathrm{u_z} \end{pmatrix}, \tag{13.2.13}$$

where

$$M = \begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} \end{pmatrix}, \tag{13.2.14}$$

where C_{ij} is the elastic coefficient. Rather than using $b = \begin{pmatrix} u \\ \tau \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} \\ \begin{pmatrix} \tau_{xz} \\ \tau_{yz} \\ \tau_{zz} \end{pmatrix} \end{pmatrix}$

(Fryer and Frazer, 1984), we use $b = \begin{pmatrix} \begin{pmatrix} u_z \\ \tau_{xz} \\ \tau_{yz} \\ \tau_{zz} \\ u_y \\ u_y \end{pmatrix} \end{pmatrix}$ (Song, 1999), and we have

$$\partial_z b = \begin{pmatrix} T & C \\ S & T^T \end{pmatrix} b = \begin{pmatrix} S & T^T \\ T & C \end{pmatrix} \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} b = ANb, \tag{13.2.15}$$

where

$$N = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}, \tag{13.2.16}$$

$$A = \begin{pmatrix} S & T^T \\ T & C \end{pmatrix} \tag{13.2.17}$$

and

$$AW = NW\Lambda. \tag{13.2.18}$$

In the above expressions, T, C, S are 3×3 partitions and C and S are symmetric. These matrices may be obtained from expressions given by Woodhouse (1974) if due allowance is made for our choice of stress-displacement vector.

13.3 Description in symbol domain

The matrix A is 6×6 in general. In order to see how to solve equation (3.2.18), we consider the 2nd case. It was written as

$$\begin{pmatrix} a_{11}(x,k) & a_{12}(x,k) \\ a_{21}(x,k) & a_{22}(x,k) \end{pmatrix} \# \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix} = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix} \# \lambda_1(x,k), \tag{13.3.1}$$

where

$$\begin{aligned} a_{11}(x,k) &= \sigma(S), \\ a_{12}(x,k) &= \sigma\left(T^T\right), \\ a_{21}(x,k) &= \sigma(T), \\ a_{22}(x,k) &= \sigma(C). \end{aligned} \tag{13.3.2}$$

Let

$$\begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix} = \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(1)} + \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(0)} + \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(-1)} \tag{13.3.3}$$

and

$$\lambda_1(x,k) = \lambda_1^{(1)}(x,k) + \lambda_1^{(0)}(x,k) + \lambda_1^{(-1)}(x,k), \tag{13.3.4}$$

where the upper scripts (0),(1) and (-1) stand for the first zeroth and the minus first order homogeneous terms of ω, k_x, k_y, respectively. Through calculation we have

$$\begin{pmatrix} a_{11}(x,k)\ a_{12}(x,k) \\ a_{21}(x,k)\ a_{22}(x,k) \end{pmatrix} \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(1)} = \begin{pmatrix} 0\ I \\ I\ 0 \end{pmatrix} \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(1)} \lambda_1^{(1)}(x,k) \tag{13.3.5}$$

and

$$\begin{aligned} &\begin{pmatrix} a_{11}(x,k)\ a_{12}(x,k) \\ a_{21}(x,k)\ a_{22}(x,k) \end{pmatrix} \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(0)} \\ &\quad +\partial_k \begin{pmatrix} a_{11}(x,k)\ a_{12}(x,k) \\ a_{21}(x,k)\ a_{22}(x,k) \end{pmatrix} D_x \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(0)} \\ &= \begin{pmatrix} 0\ I \\ I\ 0 \end{pmatrix} \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(0)} \lambda_1^{(1)}(x,k) + \begin{pmatrix} 0\ I \\ I\ 0 \end{pmatrix} \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(1)} \lambda_1^{(0)}(x,k) \\ &\quad + \begin{pmatrix} 0\ I \\ I\ 0 \end{pmatrix} \partial_k \begin{pmatrix} w_{11}(x,k) \\ w_{12}(x,k) \end{pmatrix}^{(1)} D_x \lambda_1^{(1)}(x,k). \end{aligned} \tag{13.3.6}$$

Hence RT (radiation transfer) formula (Chandrasekhar S, 1960; Li X and Strahler A, 1992) can be derived. We know that

$$\begin{aligned} &a\#w_i = Nw_i\#\lambda_i, \\ &a\#w_j = Nw_j\#\lambda_j, \\ &w_j^T = \sigma(W_j^T) \end{aligned} \tag{13.3.7}$$

and

$$\begin{aligned} &w_i^T\#a\#w_j = w_i^T Nw_j\#\lambda_j, \\ &w_j^T\#a\#w_i = w_j^T Nw_i\#\lambda_i. \end{aligned} \tag{13.3.8}$$

It can be proved that

$$w_i^T Nw_j = \begin{cases} 0, & i \neq j, \\ q_i, & i = j \end{cases} \tag{13.3.9}$$

and

$$\begin{aligned} &q_i = w_i^T Nw_i, \\ &q_1 \geqslant 0, \\ &q_2 \leqslant 0. \end{aligned} \tag{13.3.10}$$

The sign in (13.3.10) stands for the flux direction: the positive refers to the up-going wave and the negative refers to the down-going wave. The following formulas can also be obtained straightforwardly

$$W = \left(w_1\#q_1^{-\frac{1}{2}}, w_2\#q_2^{-\frac{1}{2}} \right) \tag{13.3.11}$$

$$W^T NW = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} = \text{diag}\left(I \ -I \right), \tag{13.3.12}$$

$$\begin{aligned} \Lambda &= \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \\ AW &= NW\Lambda, \end{aligned} \tag{13.3.13}$$

$$F = (W)^{-1} (\partial_z W) = NW^T \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} (\partial_z W), \tag{13.3.14}$$

$$\begin{aligned} F &= \begin{pmatrix} F_1 & F_2^T \\ F_2 & F_4 \end{pmatrix}, \\ F_1 &= -F_1^T, \\ F_4 &= -F_4^T, \end{aligned} \tag{13.3.15}$$

$$\begin{aligned} AN &= NW\Lambda (NW)^{-1}, \\ b &= NW \begin{pmatrix} U \\ D \end{pmatrix}, \end{aligned} \tag{13.3.16}$$

$$\partial_z \begin{pmatrix} U \\ D \end{pmatrix} = (\Lambda + F) \begin{pmatrix} U \\ D \end{pmatrix} = \left(\begin{pmatrix} i\omega\Gamma & 0 \\ 0 & -i\omega\Gamma \end{pmatrix} + \begin{pmatrix} T & R \\ R & T \end{pmatrix} \right) \begin{pmatrix} U \\ D \end{pmatrix}, \tag{13.3.17}$$

$$\begin{aligned} &\text{tr}\, F = 0, \\ &\det(\exp F) = \text{constant}. \end{aligned} \tag{13.3.18}$$

where $\text{tr}\, F = \sum\limits_i F_{ii}$ and $\exp F = \sum\limits_{n=0}^{\infty} \frac{1}{n!} F^n$. When $\text{tr}\Lambda = \text{tr}\sigma (S + C) = 0$, it can be proved that

$$\partial_z \left(U^T (z, \omega) U (z, -\omega) + D^T (z, \omega) D (z, -\omega) \right) = 0, \tag{13.3.19}$$

which is the result of the energy flux normalization.

13.4 Lie algebra integral expressions

Applying the idea of structure preserving computation (Feng K, 1985; Celledoni E and Iserles A, 2001; Chen J B, Munthe-Kaas H and Qin M Z, 2002) to the integral of one-way wave differential equation (Liu H, Yuan J H and Chen J B, 2006) and prediction operator differential equation (Liu H, He L, Liu G F and Li B, 2008), we get the Lie algebra integral expressions one-way wave operator and prediction operator. The Lie algebra differential equation solution $\hat{\eta}$ indicates a mapping $\hat{f} \to \hat{\eta}(\hat{f})$ such that

$$\begin{aligned}\partial_\zeta \hat{\eta} &= \frac{ad_{\hat{\eta}}}{\exp(ad_{\hat{\eta}})-1}\hat{f} \\ &= \hat{f} + B_1 ad_{\hat{\eta}}\hat{f} + \frac{B_2}{2} ad_{\hat{\eta}} ad_{\hat{\eta}}\hat{f} + \frac{B_4}{24}\left(ad_{\hat{\eta}}\right)^4 \hat{f} + \frac{B_6}{720}\left(ad_{\hat{\eta}}\right)^6 \hat{f}\cdots \\ &= \hat{f} + B_1\left[\hat{\eta},\hat{f}\right] + \frac{B_2}{2}\left[\hat{\eta},\hat{\eta},\hat{f}\right] + \frac{B_4}{24}\left[\hat{\eta},\hat{\eta},\hat{\eta},\hat{\eta},\hat{f}\right] + \frac{B_6}{720}\overbrace{[\hat{\eta},\ldots,\hat{\eta}}^{6},\hat{f}]\cdots,\end{aligned} \tag{13.4.1}$$

hence

$$\partial_\zeta d(\zeta) = \hat{f}(\zeta)\, d(\zeta). \tag{13.4.2}$$

The solution of the above equation can be expressed as

$$d(\zeta) = \exp\left(\eta\left(\hat{f}\right)\right) d(0). \tag{13.4.3}$$

In equation (13.4.1), B_n is the Bernoulli number, i.e.,

$$\sum_{n=0}^{\infty} B_n \frac{t^n}{n!} = \frac{t}{e^t - 1} \tag{13.4.4}$$

and

$$B_0 = 1,\ \ B_1 = -\frac{1}{2},\ \ B_2 = \frac{1}{6},\ \ B_4 = -\frac{1}{30},\ \ B_6 = \frac{1}{42},\ \ B_5 = 0,\cdots \tag{13.4.5}$$

Using (13.4.1)–(13.4.3), we can prove that

$$\exp\left(\hat{\eta}(\hat{f})\right)^T = \exp\left(-\hat{\eta}(-\hat{f}^T)\right). \tag{13.4.6}$$

13.5 Wave equation on the ray coordinates

The concept of time migration is related with time domain, which actually is the non-orthogonal ray coordinates. The role of this non orthogonal coordinates in multidimensional cases is similar to the time-depth conversion in one-dimensional inversion. Larner (1981) and Hubral (1975) studied the transformation between velocity function in this coordinates and velocity function in depth domain. They cast light on the cause of anisotropy from layer dipping. Because the metric matrix of non-orthogonal ray coordinates is not diagonal (Iversen E and Tygel M, 2008), it is called Riemannian coordinates (Sava P and Fomel S, 2005) which gives the way to study velocity anisotropy in time domain. Curve and body-fitting coordinate simulation (Shragge J C, 2008) for complex maintaining area provide new plentiful experience.

The Helmholtz wave equation for 3D ray tracing coordinate can be written as

$$\frac{1}{\alpha J}\left[\frac{\partial}{\partial\zeta}\left(\frac{J}{\alpha}\frac{\partial W}{\partial\zeta}\right)+\frac{\partial}{\partial\xi}\left(G\frac{\alpha}{J}\frac{\partial W}{\partial\xi}-F\frac{\alpha}{J}\frac{\partial W}{\partial\eta}\right)+\frac{\partial}{\partial\eta}\left(E\frac{\alpha}{J}\frac{\partial W}{\partial\eta}-F\frac{\alpha}{J}\frac{\partial W}{\partial\xi}\right)\right]$$
$$=-\frac{\omega^2}{v^2}W. \tag{13.5.1}$$

In this equation, ω is circle frequency, v is wave velocity, W is wave field. The coefficients in this equation can be expressed by metric matrix with element g_{ij} and its conjugate matrix with element g^{ij}, defined by

$$[g_{ij}]=\begin{bmatrix}E & F & 0\\ F & G & 0\\ 0 & 0 & \alpha^2\end{bmatrix}, \tag{13.5.2}$$

$$\left[g^{ij}\right]=\begin{bmatrix}+G/J^2 & -F/J^2 & 0\\ -F/J^2 & +E/J^2 & 0\\ 0 & 0 & 1/\alpha^2\end{bmatrix}, \tag{13.5.3}$$

where $J^2=EG-F^2$, and the determinant of the matrix $|g|=\alpha^2J^2$. We refer (Sava P and Fomel S, 2005) to the geometry meaning. Equation (13.5.1) describes the two-way wave equation propagation in ray coordinates. The physical meaning of ζ is travel time along the ray, which is the reason why some ray tracing coordinate is called "time domain" in seismic migration. In the ray coordinate, the second order coefficient of ζ is equal to the one of ω^2, excluding the terms of $\frac{\partial}{\partial\zeta}\frac{\partial}{\partial\xi}$ and $\frac{\partial}{\partial\zeta}\frac{\partial}{\partial\eta}$. For the one-way equation, we should select the wave propagation direction and modify the acoustic wave equation. We introduce the following symbols

$$c_{\zeta\zeta}=\frac{1}{\alpha^2}\ ,c_{\xi\xi}=\frac{G}{J^2}\ ,c_{\eta\eta}=\frac{E}{J^2}\ ,c_{\xi\eta}=\frac{F}{J^2}\ ,c_{\zeta}=\frac{1}{\alpha J}\frac{\partial}{\partial\zeta}\left(\frac{J}{\alpha}\right), \tag{13.5.4}$$

$$c_{\xi}=\frac{1}{\alpha J}\left[\frac{\partial}{\partial\xi}\left(G\frac{\alpha}{J}\right)-\frac{\partial}{\partial\eta}\left(F\frac{\alpha}{J}\right)\right] \tag{13.5.5}$$

and

$$c_{\eta}=\frac{1}{\alpha J}\left[\frac{\partial}{\partial\eta}\left(E\frac{\alpha}{J}\right)-\frac{\partial}{\partial\xi}\left(F\frac{\alpha}{J}\right)\right]. \tag{13.5.6}$$

In the case of ray coordinates, wave equation can be transformed to the symbol domain equation

$$-\frac{1}{v^2}k_{\zeta}^2-c_{\xi\xi}k_{\xi}^2-c_{\eta\eta}k_{\eta}^2+ic_{\zeta}k_{\zeta}+ic_{\xi}k_{\xi}+ic_{\eta}k_{\eta}-c_{\xi\eta}k_{\xi}k_{\eta}=-\frac{\omega^2}{v^2}. \tag{13.5.7}$$

To get the wave number symbol k_{ζ} in the exploration direction, we solve symbol equation (13.5.7) and choose the minus or plus symbol corresponding to direction

$$k_\zeta = i\frac{c_\zeta}{2c_{\zeta\zeta}} \pm \sqrt{\left(\omega^2 - \left(\frac{c_\zeta}{2c_{\zeta\zeta}}\right)^2 - \sum_{j=\xi,\eta}\left(\frac{c_{jj}}{c_{\zeta\zeta}}k_j^2 - i\frac{c_j}{c_{\zeta\zeta}}k_j\right) - \frac{c_{\xi\eta}}{c_{\zeta\zeta}}k_\xi k_\eta\right)} \#. \tag{13.5.8}$$

In this equation, # stands for Witt product, with the meaning of square root of a symbol shown in [22], and the additive symbol shows that the wave propagates in the positive direction. To be simple, we express the coordinators (ξ, η, ζ) as (x, y, ζ). The general form of (13.5.8) is given by equations (13.3.1)–(13.3.6). The symbol of single square root is $k_\zeta = k_\zeta(x, y, \zeta, k_x, k_y, \omega)$. It can determine the one way wave operator uniquely.

13.6 Symbol expression of one-way wave operator equations

Using ray tracing coordinates, the wave field in the surface is denoted by $W(x, y, 0, \omega)$, and the wave field in the depth ζ is denoted by $W(x, y, \zeta, \omega)$. The extrapolation equation with the transmission term for wave field in frequency domain can be expressed as

$$\begin{aligned} \partial_\zeta U &= (\lambda_1 + F_1)\, U, \\ \partial_\zeta D &= (\lambda_2 + F_4)\, D, \end{aligned} \tag{13.6.1}$$

where U and D are up-going and down-going waves respectively. and in order to solve equation (13.6.1), we try to find the solution of the following equation.

$$\begin{aligned} &\partial_\zeta W(x, y, \zeta, \omega) = \left(-i\omega\hat{\Gamma} + \hat{T}\right) W(x, y, \zeta, \omega), \\ &\mathrm{tr}\hat{T} = 0. \end{aligned} \tag{13.6.2}$$

In a 2 by 2 matrix case, one has (by formulas (13.3.14), (13.3.15) and (13.3.18))

$$\partial_\zeta W(x, y, \zeta, \omega) = -i\omega\hat{\Gamma} W(x, y, \zeta, \omega). \tag{13.6.3}$$

This is the result of flux normalization.

Suppose the integral solution of the extrapolation equation is

$$W(\zeta, \omega) = \hat{G} W(\zeta = 0) = \exp\left(i\hat{A}\,(\zeta)\right) W(x, y, \zeta = 0, \omega), \tag{13.6.4}$$

where $\hat{G} = \exp(i\hat{A}(z))$ is the large step one-way operator, $\hat{A}(z)$ is the global integral operator of single square root operator $(\pm i\omega\hat{\Gamma} + \hat{T})$. We call it Lie algebra integral operator because it is the result of Lie algebra differential equation (4.1). Based on (4.1), we get:

$$i\hat{A}\,(\zeta) = \hat{\eta}\left(\pm i\omega\hat{\Gamma} + \hat{T}\right). \tag{13.6.5}$$

We use ϕ to express the complex phase function of $\hat{G}$ symbol functions, that is,

$$\begin{aligned}\exp\left(i\phi\left(x,y,\zeta,k_x,k_y,\omega\right)\right) &= \sigma\left(\hat{G}(\zeta)\right) = \sigma\left(\exp\left(i\hat{A}\left(\zeta\right)\right)\right)\\ &= \sigma\left(\exp\left(\hat{\eta}\left(-i\omega\hat{\Gamma}+\hat{T}\right)\right)\right).\end{aligned} \tag{13.6.6}$$

Thus equation (13.6.4) can be written as Fourier integral operator in the following way:

$$\begin{aligned}&W(x,y,z,\omega)\\ &= \frac{1}{\left(2\pi\right)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(i\phi\left(x,y,z,k_x,k_y,\omega\right)\right)w\left(k_x,k_y,z=0,\omega\right)\cdot\\ &\quad\exp\left(ik_x x+ik_y y\right)dk_x dk_y\\ &= \frac{1}{\left(2\pi\right)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(i\phi_1-\phi_0\right)w\left(k_x,k_y,z=0,\omega\right)\cdot\\ &\quad\exp\left(ik_x x+ik_y y\right)dk_x dk_y.\end{aligned} \tag{13.6.7}$$

This equation means that via complex phase function $\phi=\phi^{[1]}+i\phi^{[0]}$, the symbol of operator $\exp(iA(\zeta))=\exp(\hat{\eta}(-i\omega\hat{\Gamma}+\hat{T}))$ can be expressed as an exponent function. The one order homogeneous $\phi^{[1]}$ function determines the travel time, which can be solved from $\hat{\Gamma}$'s kinematic symbol. And the zero order homogeneous $\phi^{[0]}$ function determines the amplitude of the wave field, which can be gotten from one way wave equation operator $\hat{\Gamma}$ and transmission differential operator $\hat{T}$. The -1 order homogeneous function $\phi^{[-1]}$ shows the frequency-decreasing phase or dispersion. Equation (13.6.7) can be written as

$$\begin{aligned}&W(x,y,\zeta,\omega)=\\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\partial}{\partial\zeta}G\left(x,y,\zeta,x-x',y-y',\omega\right)W(x,y,0,\omega)dxdy\end{aligned} \tag{13.6.8}$$

and

$$\begin{aligned}&\frac{\partial}{\partial\zeta}G\left(x,y,\zeta,x-x',y-y',\omega\right)\\ &= \frac{1}{\left(2\pi\right)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(i\phi\left(x,y,\zeta,k_x,k_y,\omega\right)\right)\cdot\\ &\quad\exp\left(ik_x(x-x')+ik_y(y-y')\right)dk_x dk_y\\ &= \frac{1}{\left(2\pi\right)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(i\phi^{[1]}-\phi^{[0]}\right)\cdot\\ &\quad\exp\left(ik_x(x-x')+ik_y(y-y')\right)dk_x dk_y\\ &= M\left(x,y,\zeta,x-x',y-y',\omega\right)\exp\left(i\omega T\left(x,y,\zeta,x-x',y-y',\omega\right)\right).\end{aligned} \tag{13.6.9}$$

From the saddle point method in 2D cases, we get:

$$\begin{aligned}&\phi^{[1]}_{,p_x}\left(p_{x0},p_{y0}\right)+x-x'=0,\\ &\phi^{[1]}_{,p_y}\left(p_{x0},p_{y0}\right)+y-y'=0,\end{aligned} \tag{13.6.10}$$

$$T = \phi^{[1]}(p_{x0}, p_{y0}) + p_{x0}(x - x') + p_{y0}(y - y') \tag{13.6.11}$$

and

$$\begin{aligned} &M\left(x, y, \zeta, x - x', y - y', \omega\right) \\ &= \sqrt{\frac{2\pi}{\omega \left|b''(p_0)\right|}} \exp(-\phi^{[0]}(p_{x0}, p_{y0}) + \frac{i\pi}{4} \operatorname{sign} \phi^{[0]}(p_{x0}, p_{y0}), \end{aligned} \tag{13.6.12}$$

where $\phi^{[0]}(p_{x0}, p_{y0})$ is the term of transmission adjustment, which can be obtained from transmission operator term and zero order homogenous function of one-way wave equation operator. In formulas (13.6.10)–(13.6.12), $T(x, y, \zeta, x-x', y-y', \omega)$ and $M(x, y, \zeta, x - x', y - y', \omega)$ describe the travel time and amplitude of Green function respectively. Equation (13.6.12) is the expression of Green function in Kirchhoff integral formula.

13.7 Lie algebra expression of travel time

The expression of phase and amplitude of symbols need to be simplified for calculation. We can get concise expression of lateral velocity variation by adopting idea of Feng Kang's structure preserving calculation and combining the idea of time migration. The first step to simplify Lie algebra differential equation of one-way wave equation is realized by transferring it to symbol domain. The second step is to use BCH (Baker-Campbell-Hausdorff) formula (Cordes H O, 1995) for the exponential mapping from the result of last step. The third one is to transfer Lie algebra differential equation from the depth domain to ray tracing coordinate. Paraxonic approximation of travel time in the lateral velocity variation cases has non-symmetric term. Conventionally, this non-symmetric part has not been solved by numerical calculation of the differential wave equation(wave equation, eikonal equation or ray tracing equation) until the occurrence of the structure preserving algorithm. The structure preserving algorithm keeps discretization in wrong domain too early. Based on the study on Lie algebra integral and property of ray coordinates, it is found that the recursion method or Magnus method is one of the simple methods. That is, first get the low order Lie algebra integral, save them and their components, and solve the high order Lie algebra integrals and their components from the lower order ones by commutator and other operators. We have that

$$s = \sigma(\omega\hat{\Gamma} \pm i\hat{T}) = s^{[1]} + is^{[0]} + s^{[-1]} + is^{[-2]}, \tag{13.7.1}$$

where $s^{[1]}$ comes from $\hat{\Gamma}$ only, $s^{[0]}$ from $\hat{\Gamma}$ and $\hat{T}$. Kinematic symbol of one way wave equation operator has the following expression:

$$\begin{aligned} s^{[1]} &= \sigma_{\text{main}}(\omega\hat{\Gamma}) \\ &= \sqrt{\omega^2 - \frac{c_{\xi\xi}}{c_{\zeta\zeta}}k_x^2 - \frac{c_{\eta\eta}}{c_{\zeta\zeta}}k_y^2 - \frac{c_{\xi\eta}}{c_{\zeta\zeta}}k_x k_y} \\ &= \omega - \frac{1}{2\omega^2}\left(\frac{c_{\xi\xi}}{c_{\zeta\zeta}}k_x^2 + \frac{c_{\eta\eta}}{c_{\zeta\zeta}}k_y^2 + \frac{c_{\xi\eta}}{c_{\zeta\zeta}}k_x k_y\right) \\ &\quad + \frac{1}{8\omega^3}\left(\frac{c_{\xi\xi}}{c_{\zeta\zeta}}k_x^2 + \frac{c_{\eta\eta}}{c_{\zeta\zeta}}k_y^2 + \frac{c_{\xi\eta}}{c_{\zeta\zeta}}k_x k_y\right)^2 \cdots . \end{aligned} \tag{13.7.2}$$

This is the kinematic symbol of single square root operator in ray tracing coordinate. The Taylor extension of s only contains even part of k_x, k_y, without the third part of them. So equation (13.7.1) could be expressed as

$$\begin{aligned} s &= s^{[1]} + is^{[0]} \\ &= \left(s_0^{[1]} + s_2^{[1]} + s_4^{[1]} + s_6^{[1]} + \ldots\right) + i\left(s_0^{[0]} + s_1^{[0]} + s_2^{[0]} + s_3^{[0]} + \ldots\right). \end{aligned} \tag{13.7.3}$$

We use the superiors to express the whole homogeneous order, and inferior numerals to show the homogeneous order of k_x, k_y, and have that

$$s_m^{[n+m]} = \omega^n \sum_{i=0}^{m} s_{m-i,i}^{[n+m]} k_x^{m-i} k_y^i, \tag{13.7.4}$$

$$s_0^{[1]} = \omega, \tag{13.7.5}$$

$$s_1^{[1]} = s_3^{[1]} = s_5^{[1]} = 0, \tag{13.7.6}$$

$$s_2^{[1]} = -\frac{1}{2\omega}\left(\frac{c_{\xi\xi}}{c_{\zeta\zeta}}k_x^2 + \frac{c_{\eta\eta}}{c_{\zeta\zeta}}k_y^2 + \frac{c_{\xi\eta}}{c_{\zeta\zeta}}k_x k_y\right), \tag{13.7.7}$$

and

$$s_4^{[1]} = -\frac{1}{8\omega^3}\left(\frac{c_{\xi\xi}}{c_{\zeta\zeta}}k_x^2 + \frac{c_{\eta\eta}}{c_{\zeta\zeta}}k_y^2 + \frac{c_{\xi\eta}}{c_{\zeta\zeta}}k_x k_y\right)^2. \tag{13.7.8}$$

Thus, it is easy to get, from the property of ray coordinates, that for any symbol, the function f satisfies

$$\left\{f, s_0^{[1]}\right\} = 0. \tag{13.7.9}$$

In the depth coordinates, equation (13.7.9) is not correct because the zero order terms of k_x and k_y in one-way wave equation operator change horizontally, which makes the expression of Lie algebra complex and the multidimensional scattering predictive equation cannot be simplified through exchanging the orders of multiples and merging the exponent factors. So far, we have derived the odd term of paraxonic expression using velocity in time domain, root of mean square velocity and their lateral derivatives.

According to the recent results (Liu H, Yuan J H and Chen J B, 2006; Liu H, Wang X M and Zeng R, 2007), we get

$$ia = \sigma(i\hat{A}) = \hat{\eta}\left(\sigma\left(-i\omega\hat{\Gamma} + \hat{T}\right)\right). \tag{13.7.10}$$

Let $a \to \psi(a)$ represent exponential expression of operator symbol for exponent map (Celledoni E and Iserles A, 2001). Then

$$\exp(i\psi) = \sigma\left(\exp\left(i\hat{A}(\zeta)\right)\right) = \exp\left(\sigma\left(i\hat{A}(\zeta)\right)\#\right) = \exp(ia\#) \tag{13.7.11}$$

and

$$a = a^{[1]} + ia^{[0]} = \left(a_0^{[1]} + a_2^{[1]} + a_3^{[1]} + a_4^{[1]}\cdots\right) + i\left(a_0^{[0]} + a_1^{[0]} + a_2^{[0]} + a_3^{[0]}\cdots\right), \tag{13.7.12}$$

$$a_m^{[n+m]} = \omega^n \sum_{i=0}^{m} a_{m-i,i}^{[n+m]} k_x^{m-i} k_y^i, \tag{13.7.13}$$

$$a_m^{[n+m]} = \omega^n \sum_{i=0}^{m} a_{m-i,i}^{[n+m]} k_x^{m-i} k_y^i, \tag{13.7.14}$$

$$a_0^{[1]} = \omega\zeta, \tag{13.7.15}$$

$$a_1^{[1]} = 0, \tag{13.7.16}$$

$$a_2^{[1]} = \frac{1}{\omega}\left(a_{2,0}^{[1]} k_x^2 + a_{1,1}^{[1]} k_x k_y + a_{0,2}^{[1]} k_y^2\right), \tag{13.7.17}$$

$$a_3^{[1]} = \frac{1}{\omega^2}\left(a_{3,0}^{[1]} k_x^3 + a_{2,1}^{[1]} k_x^2 k_y + a_{1,2}^{[1]} k_x k_y^2 + a_{0,3}^{[1]} k_y^3\right). \tag{13.7.18}$$

From the derivative procedure, we can get the coefficient of quadratic terms of k_x and k_y in the global integral of square root operator on ray coordinator (time migration domain), which is only determined by quadratic terms of k_x and k_y in square root operator:

$$a_{2,0} = -\frac{1}{2}\int_0^\zeta \frac{c_{\xi\xi}}{c_{\zeta\zeta}} d\zeta \; a_{1,1} = -\frac{1}{2}\int_0^\zeta \frac{c_{\xi\eta}}{c_{\zeta\zeta}} d\zeta \; a_{0,2} = -\frac{1}{2}\int_0^\zeta \frac{c_{\eta\eta}}{c_{\zeta\zeta}} d\zeta\,. \tag{13.7.19}$$

This property is not be applicable in depth coordinator. It is, in fact, the strict theoretical base of Larner's cascaded migration (K. L. Larner, L. Hatton and B. S. Gibson, 1981). In equation (13.7.18),

$$i\left(a_3^{[1]} + ia_3^{[0]}\right) = -\frac{1}{2}\int_0^\zeta \left\{i\left(a_2^{[1]} + ia_2^{[0]}\right), i\left(s_2^{[1]} + is_2^{[0]}\right)\right\} d\zeta. \tag{13.7.20}$$

Equation (13.7.20) shows that although square root operator doesn't contain odd part, square root operator global integral (solution of Lie algebra differential equation) contains odd part of k_x and k_y, indicating that the solution of Lie algebra differential equation has considered lateral velocity variation. We have the following formulas:

$$\phi_{n,m} = \omega^n \sum_{i=0}^{m} \phi_{m-i,i} k_x^{m-i} k_y^i, \tag{13.7.21}$$

$$\begin{aligned}
\phi^{[1,1]} &= a^{[1]} \\
&= a_{0,0}\omega + \frac{1}{\omega}\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right) \\
&\quad + \frac{1}{\omega^2}\left(a_{3,0}k_x^3 + a_{2,1}k_x^2k_y + a_{1,2}k_xk_y^2 + a_{0,3}k_y^3\right) \\
&\quad + \frac{1}{\omega^3}\left(a_{4,0}k_x^4 + a_{3,1}k_x^3k_y + a_{2,2}k_x^2k_y^2 + a_{1,3}k_xk_y^3 + a_{0,4}k_y^4\right)\cdots,
\end{aligned} \tag{13.7.22}$$

$$\begin{aligned}
\phi^{[1,2]} = &\\
&\frac{1}{2}\begin{pmatrix}
\frac{1}{\omega^2}\left(2a_{2,0}k_x + a_{1,1}k_y\right)\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,x} \\
+\frac{1}{\omega^3}\left(3a_{3,0}k_x^2 + 2a_{2,1}k_xk_y + a_{1,2}k_y^2\right)\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,x} \\
+\frac{1}{\omega^3}\left(2a_{2,0}k_x + a_{1,1}k_y\right)\left(a_{3,0}k_x^3 + a_{2,1}k_x^2k_y + a_{1,2}k_xk_y^2 + a_{0,3}k_y^3\right)_{,x}
\end{pmatrix} \\
+&\frac{1}{2}\begin{pmatrix}
\frac{1}{\omega^2}\left(a_{1,1}k_x + 2a_{0,2}k_y\right)\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,y} \\
+\frac{1}{\omega^3}\left(a_{2,1}k_x^2 + 2a_{1,2}k_xk_y + 3a_{0,3}k_y^2\right)\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,y} \\
+\frac{1}{\omega^3}\left(a_{1,1}k_x + 2a_{0,2}k_y\right)\left(a_{3,0}k_x^3 + a_{2,1}k_x^2k_y + a_{1,2}k_xk_y^2 + a_{0,3}k_y^3\right)_{,y}
\end{pmatrix}
\end{aligned} \tag{13.7.23}$$

and

$$\begin{aligned}
\phi^{[1,3]} = &\frac{1}{6\omega^3}\begin{pmatrix}
2a_{2,0}\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,x} \\
+a_{1,1}\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,y}
\end{pmatrix}\cdot \\
&\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,x} \\
&+\frac{1}{6\omega^3}\begin{pmatrix}
2a_{1,1}\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,x} \\
+2a_{0,2}\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,y}
\end{pmatrix}\cdot \\
&\left(a_{2,0}k_x^2 + a_{1,1}k_xk_y + a_{0,2}k_y^2\right)_{,y},
\end{aligned} \tag{13.7.24}$$

where $\varphi^{[1,2]}$ contains a correction term of cubic term in Lie algebra integral to quadruplicate in phase, whereas $\varphi^{[1,1]}$ and $\varphi^{[1,3]}$ do not contain the correction term. Now, we can use stationary phase method to get Green's function expression from phase φ's expression. In the case of the layered media travel time in Green's function is "Dix formula". So, our derivation is to extend "Dix formula" to lateral variation condition, which can be called Pseudo-Dix formula, i.e.,

$$\begin{aligned}
&T^2(x, y, \zeta, x - x', y - y') \\
&= \phi_{0,0}^2 + \left(c_{2,0}\left(x - x'\right)^2 + c_{1,1}\left(x - x'\right)\left(y - y'\right) + c_{0,2}\left(y - y'\right)^2\right) \\
&+ \left(c_{3,0}\left(x - x'\right)^3 + c_{2,1}\left(x - x'\right)^2\left(y - y'\right) + c_{1,2}\left(x - x'\right)\left(y - y'\right)^2 + c_{0,3}\left(y - y'\right)^3\right),
\end{aligned} \tag{13.7.25}$$

$$c_{2,0}(x,y,\zeta) = -2\frac{\phi_{0,0}}{d}\phi_{0,2}, \tag{13.7.26}$$

$$c_{1,1}(x,y,\zeta) = 2\frac{\phi_{0,0}}{d}\phi_{1,1}, \tag{13.7.27}$$

$$c_{0,2}(x,y,\zeta) = -2\frac{\phi_{0,0}}{d}\phi_{2,0}, \tag{13.7.28}$$

$$c_{3,0}(x,y,\zeta) = -\frac{2\phi_{0,0}}{d^3}\left(8\phi_{0,2}^3\phi_{3,0} - 4\phi_{1,1}\phi_{0,2}^2\phi_{2,1} + 2\phi_{0,2}\phi_{1,1}^2\phi_{1,2} - \phi_{1,1}^3\phi_{0,3}\right), \tag{13.7.29}$$

$$c_{2,1}(x,y,\zeta) = -\frac{2\phi_{0,0}}{d^3}\begin{pmatrix} -12\phi_{0,2}^2\phi_{1,1}\phi_{3,0} + \left(4\phi_{1,1}^2\phi_{0,2} + 8\phi_{2,0}\phi_{0,2}^2\right)\phi_{2,1} \\ -\left(8\phi_{0,2}\phi_{1,1}\phi_{2,0} + \phi_{1,1}^3\right)\phi_{1,2} + 6\phi_{1,1}^2\phi_{2,0}\phi_{0,3} \end{pmatrix}, \tag{13.7.30}$$

$$c_{1,2}(x,y,\zeta) = -\frac{2\phi_{0,0}}{d^3}\begin{pmatrix} 12\phi_{0,2}\phi_{1,1}^2\phi_{3,0} - \left(\phi_{1,1}^3 + 8\phi_{2,0}\phi_{0,2}\phi_{1,1}\right)\phi_{2,1} + \\ +\left(8\phi_{0,2}\phi_{2,0}^2 + 4\phi_{1,1}^2\phi_{2,0}\right)\phi_{1,2} - 12\phi_{1,1}\phi_{2,0}^2\phi_{0,3} \end{pmatrix}, \tag{13.7.31}$$

$$c_{0,3}(x,y,\zeta) = -\frac{2\phi_{0,0}}{d^3}\left(-\phi_{1,1}^3\phi_{3,0} + 2\phi_{2,0}\phi_{1,1}^2\phi_{2,1} - 4\phi_{1,1}\phi_{2,0}^2\phi_{1,2} + 8\phi_{2,0}^3\phi_{0,3}\right), \tag{13.7.32}$$

and

$$d(x,y,\zeta) = \left(2\phi_{2,0}2\phi_{0,2} - \phi_{1,1}\phi_{1,1}\right). \tag{13.7.33}$$

In complex medium, formula (13.7.25) is not correct since it lost multi-value travel time. In this case, formulas (13.6.10)–(13.6.11) should be used. At present, we are doing application research on phase expression of single-square-root operator symbol and asymmetry travel time expression (Liu H and Li Y M, 2008). Asymmetry travel time expression can be applied to post-stack and pre-stack time migration through Kirchhoff integral method, the phase expression of single-square-root operator Symbol can be applied to post-stack and pre-stack time migration through PSPI method.

13.8 Lie algebra integral expression of prediction operator

From equation (13.3.17) we can get

$$\begin{aligned} &\partial_\zeta \begin{pmatrix} D(\zeta,-\omega)\ U(\zeta,\omega) \\ U(\zeta,-\omega)\ D(\zeta,\omega) \end{pmatrix} \\ &= \begin{pmatrix} -i\omega\hat{\Gamma}(\zeta,\omega) + \hat{T}(\zeta,\omega) & \hat{R}(\zeta,\omega) \\ \hat{R}(\zeta,\omega) & i\omega\hat{\Gamma}(\zeta,\omega) + \hat{T}(\zeta,\omega) \end{pmatrix} \begin{pmatrix} D(\zeta,-\omega)\ U(\zeta,\omega) \\ U(\zeta,-\omega)\ D(\zeta,\omega) \end{pmatrix}. \end{aligned} \tag{13.8.1}$$

Equation (13.8.1) is applicable to any function, so we can introduce an operator expression. Suppose that

$$\begin{pmatrix} D(\zeta,-\omega)\ U(\zeta,\omega) \\ U(\zeta,-\omega)\ D(\zeta,\omega) \end{pmatrix} = \begin{pmatrix} \hat{F}(\zeta,-\omega)\ \hat{G}(\zeta,\omega) \\ \hat{G}(\zeta,-\omega)\ \hat{F}(\zeta,\omega) \end{pmatrix} \begin{pmatrix} D(0,-\omega)\ U(0,\omega) \\ U(0,-\omega)\ D(0,\omega) \end{pmatrix}, \tag{13.8.2}$$

where $\begin{pmatrix} \hat{F}(\zeta,-\omega)\ \hat{G}(\zeta,\omega) \\ \hat{G}(\zeta,-\omega)\ \hat{F}(\zeta,\omega) \end{pmatrix}$ serves as the Green's function for up- and down-going wavefield, and

$$\begin{aligned} &D(\zeta,-\omega)+U(\zeta,\omega) \\ &=\hat{F}(\zeta,-\omega)(D(0,-\omega)+U(0,\omega))+\hat{G}(\zeta,\omega)(U(0,-\omega)+D(0,\omega)), \end{aligned} \tag{13.8.3}$$

$$\begin{pmatrix} \hat{F}(0,-\omega)\ \hat{G}(0,\omega) \\ \hat{G}(0,-\omega)\ \hat{F}(0,\omega) \end{pmatrix} = I, \tag{13.8.4}$$

then

$$\begin{aligned} &\partial_\zeta \begin{pmatrix} \hat{F}(\zeta,-\omega)\ \hat{G}(\zeta,\omega) \\ \hat{G}(\zeta,-\omega)\ \hat{F}(\zeta,\omega) \end{pmatrix} \\ &= \begin{pmatrix} -i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega) & \hat{R}(\zeta,\omega) \\ \hat{R}(\zeta,\omega) & i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega) \end{pmatrix} \begin{pmatrix} \hat{F}(\zeta,-\omega)\ \hat{G}(\zeta,\omega) \\ \hat{G}(\zeta,-\omega)\ \hat{F}(\zeta,\omega) \end{pmatrix}. \end{aligned} \tag{13.8.5}$$

Get the first line on both sides of equation (13.8.5) and minus each term, we obtain

$$\begin{aligned} &\partial_\zeta\left(\hat{F}(\zeta,-\omega)-\hat{G}(\zeta,\omega)\right)= \\ &\left(-i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega)\right)\hat{F}(\zeta,-\omega)+\hat{R}(\zeta,\omega)\hat{G}(\zeta,-\omega) \\ &-\left(-i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega)\right)\hat{G}(\zeta,\omega)-\hat{R}(\zeta,\omega)\hat{F}(\zeta,\omega). \end{aligned} \tag{13.8.6}$$

Let

$$\hat{B}(\zeta,\omega)=\hat{F}(\zeta,-\omega)-\hat{G}(\zeta,\omega) \tag{13.8.7}$$

and

$$\hat{C}(\zeta,\omega)=\hat{F}(\zeta,-\omega)+\hat{G}(\zeta,\omega), \tag{13.8.8}$$

equation (13.8.6) can be converted to

$$\partial_\zeta\hat{B}(\zeta,\omega)=\left(-i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega)\right)\hat{B}(\zeta,\omega)-\hat{R}(\zeta,\omega)\hat{B}(\zeta,-\omega), \tag{13.8.9}$$

where $\hat{B}(\zeta,\omega)$ is generalized transmission operator's inverse operator (will be explained later). Use one-way operator of Lie algebra integral expression equations (13.4.1),(13.6.5) and (13.6.6), and assume that

$$\hat{B}(\zeta,\omega)=\exp\left(\hat{\eta}\left(-i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega)\right)\right)\hat{A}(\zeta,\omega), \tag{13.8.10}$$

$$\hat{A}(0,\omega) = I, \tag{13.8.11}$$

and use Lie algebra integral expression of transposed operator equation (13.4.6) and transposed property of reflection, transmission and one-way operator equations (13.3.14)–(13.3.15), we get

$$\hat{B}^T(\zeta,-\omega) = \hat{A}^T(\zeta,-\omega)\exp\left(-\hat{\eta}\left(-i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega)\right)\right). \tag{13.8.12}$$

According to equation (13.8.10)and equation (13.8.12), we get

$$\hat{B}^T(\zeta,-\omega)\,\hat{B}(\zeta,\omega) = \hat{A}^T(\zeta,-\omega)\,\hat{A}(\zeta,\omega), \tag{13.8.13}$$

where $\hat{A}(\zeta,\omega)$ is prediction operator. Substituting equation (13.8.10) into equation (13.8.9), we can get the differential equation satisfied by prediction operator

$$\partial_\zeta \hat{A}(\zeta,\omega) = \hat{F}(\zeta,\omega)\,\hat{A}(\zeta,-\omega), \tag{13.8.14}$$

where

$$\begin{aligned}\hat{F}(\zeta,\omega) = -\exp\left(-\hat{\eta}\left(-i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega)\right)\right)\hat{R}(\zeta,\omega)\cdot\\ \exp\left(\hat{\eta}\left(i\omega\hat{\Gamma}(\zeta,\omega)+\hat{T}(\zeta,\omega)\right)\right).\end{aligned} \tag{13.8.15}$$

When ω and i in equation (13.3.17) change sign simultaneously, the $\hat{A}$ is invariant, that is

$$\hat{A}(\zeta,-\omega) = \hat{A}(\zeta,\omega)^*. \tag{13.8.16}$$

To distinguish complex numbers in Symbol, define that

$$\begin{aligned}\hat{A}_1(\zeta,\omega) &= \hat{A}(\zeta,\omega)+\hat{A}(\zeta,-\omega),\\ \hat{A}_2(\zeta,\omega) &= -i\left(\hat{A}(\zeta,\omega)+\hat{A}(\zeta,-\omega)\right).\end{aligned} \tag{13.8.17}$$

Then we can get prediction operator differential equation which is similar to equation (13.4.2)

$$\partial_\zeta\begin{pmatrix}\hat{A}_1(\zeta,\omega)\\ \hat{A}_2(\zeta,\omega)\end{pmatrix} = \begin{pmatrix}\mathrm{re}\left(\hat{F}\right) & \mathrm{im}\left(\hat{F}\right)\\ \mathrm{im}\left(\hat{F}\right) & -\mathrm{re}\left(\hat{F}\right)\end{pmatrix}\begin{pmatrix}\hat{A}_1(\zeta,\omega)\\ \hat{A}_2(\zeta,\omega)\end{pmatrix}. \tag{13.8.18}$$

From equations (13.4.1) and (13.4.3), the solution can be expressed as

$$\begin{pmatrix}\hat{A}_1(\zeta,\omega)\\ \hat{A}_2(\zeta,\omega)\end{pmatrix} = \exp\left(\hat{\eta}\begin{pmatrix}\mathrm{re}\left(\hat{F}\right) & \mathrm{im}\left(\hat{F}\right)\\ \mathrm{im}\left(\hat{F}\right) & -\mathrm{re}\left(\hat{F}\right)\end{pmatrix}\right)\begin{pmatrix}\hat{A}_1(0,\omega)\\ \hat{A}_2(0,\omega)\end{pmatrix} \tag{13.8.19}$$

and

$$\begin{pmatrix} \hat{A}_1(\zeta,-\omega) \\ \hat{A}_2(\zeta,-\omega) \end{pmatrix}^T = \begin{pmatrix} \hat{A}_1(0,-\omega) \\ \hat{A}_2(0,-\omega) \end{pmatrix}^T \exp\left(-\hat{\eta} \begin{pmatrix} -\operatorname{re}\left(\hat{F}\right) & \operatorname{im}\left(\hat{F}\right) \\ \operatorname{im}\left(\hat{F}\right) & \operatorname{re}\left(\hat{F}\right) \end{pmatrix} \right). \tag{13.8.20}$$

Equations (13.8.19) and (13.8.20) are called Lie algebra integral expression of prediction operator.

13.9 Spectral factorization expressions of reflection data

Spectral factorization expressions are using the autocorrelation of prediction operators to express reflection data. According to property of Earth surface reflection data (Claerbout J F, 1985; Frasier C W, 1970), we have that

$$\begin{pmatrix} D(0,-\omega)\ U(0,\omega) \\ U(0,-\omega)\ D(0,\omega) \end{pmatrix} = \begin{pmatrix} 1+\hat{R}(-\omega) & -\hat{R}(\omega) \\ -\hat{R}(-\omega) & 1+\hat{R}(\omega) \end{pmatrix} \exp\left(ik\left(x_H - x_{H,A}\right)\right), \tag{13.9.1}$$

$$\begin{pmatrix} D(\zeta,-\omega)\ U(\zeta,\omega) \\ U(\zeta,-\omega)\ D(\zeta,\omega) \end{pmatrix} = \begin{pmatrix} \hat{E}(\zeta,-\omega) & 0 \\ 0 & \hat{E}(\zeta,\omega) \end{pmatrix} \exp\left(ik\cdot\left(x_H - x_{H,A}\right)\right), \tag{13.9.2}$$

and

$$\begin{pmatrix} \hat{E}(\zeta,-\omega) & 0 \\ 0 & \hat{E}(\zeta,\omega) \end{pmatrix} = \begin{pmatrix} \hat{F}(\zeta,-\omega)\ \hat{G}(\zeta,\omega) \\ \hat{G}(\zeta,-\omega)\ \hat{F}(\zeta,\omega) \end{pmatrix} \begin{pmatrix} 1+\hat{R}(-\omega) & -\hat{R}(\omega) \\ -\hat{R}(-\omega) & 1+\hat{R}(\omega) \end{pmatrix}. \tag{13.9.3}$$

In the above equations, x_H and $x_{H,A}$ represent horizontal coordinates of receiver and shot. Comparing the first lines on both sides of equation (13.9.3), we get

$$\hat{E}(\zeta,-\omega) = \hat{F}(\zeta,-\omega)\left(1+\hat{R}(-\omega)\right) - \hat{G}(\zeta,\omega)\,\hat{R}(-\omega), \tag{13.9.4}$$

$$0 = -\hat{F}(\zeta,-\omega)\,\hat{R}(\omega) + \hat{G}(\zeta,\omega)\left(1+\hat{R}(\omega)\right). \tag{13.9.5}$$

Adding equation (13.9.4) to equation (13.9.5), we get

$$\hat{E}(\zeta,-\omega) = \hat{B}(\zeta,\omega)\left(1+\hat{R}(-\omega)+\hat{R}(\omega)\right). \tag{13.9.6}$$

By equations (13.3.18) and (13.3.19) or using reciprocity theorem of directional wave (Wapenaar C P A and Grimbergen J L T, 1996), we get the autocorrelation expression of reflection data by generalized transmission operator

$$\hat{E}^T(\zeta,\omega)\,\hat{E}(\zeta,-\omega) = 1+\hat{R}^T(-\omega)+\hat{R}(\omega) = 1+\hat{R}(-\omega)+\hat{R}(\omega). \tag{13.9.7}$$

Comparing equations (13.9.6) and (13.9.7), it can be gotten that

$$\hat{B}(\zeta,\omega)\hat{E}^T(\zeta,\omega)=I. \tag{13.9.8}$$

So

$$\hat{E}^T(\zeta,\omega)=\hat{B}(\zeta,\omega)^{-1}. \tag{13.9.9}$$

This is why $\hat{B}(\zeta,\omega)$ is called the inverse operator of generalized transmission operators. From equations (13.8.13), (13.9.7) and (13.9.9), one can get

$$\begin{aligned}&1+\hat{R}(-\omega)+\hat{R}(\omega)=\hat{B}^{-1}(\zeta,\omega)\hat{B}^T(\zeta,-\omega)^{-1}\\&=\hat{A}^{-1}(\zeta,\omega)\hat{A}^T(\zeta,-\omega)^{-1}.\end{aligned} \tag{13.9.10}$$

Equation (13.9.10) is the expression of reflection data via autocorrelation of prediction operators. Equation (13.8.19) gives an generalization to O'Doherty formula (O'Doherty R F and Anstey N A, 1971) in oblique incidence condition.

A laterally invariant medium with three strong reflecting boundaries and number of fine layers is considered to verify the schemes of the prediction and elimination of multiples using the inverse scattering theory. First, we verify the forward model: given the reflection coefficients, by equations (13.9.10) and (13.9.19), the wave field can be predicted. Second, the multiples elimination method is verified: given the wave field predicted above, by equations (13.9.19) and (13.8.14), the reflection coefficients are obtained, i.e., the multiples are removed. Fig. 13.1 (a) shows the reflection coefficients. Fig. 13.1 (b) is the prediction operator calculated by equation (13.9.10) giving the reflection coefficients shown in Fig. 13.1 (a). Substituting the prediction operator into Equation (13.9.19), the wave field can be easily obtained as shown in Fig. 13.2 (a). Fig. 13.2 (b) is the linear Radon transformation of the wave field shown in Fig. 13.2 (a). By 2D spectral factorization of the wave fields shown in Fig. 13.2 (a) and (b), in offset and ray parameter, respectively, the prediction operator can be obtained and shown in Fig. 13.3(a) and (b). By the prediction operator shown in Fig. 13.3 (b), the reflection coefficients shown in Fig. 13.4 (a) are calculated using the inverse scattering theory. And the linear Re-Radon transformation of the reflection coefficients is shown in Fig. 13.4 (b). Comparing Fig. 13.4 (b) with Fig. 13.2 (a) and Fig. 13.4 (a) with Fig. 13.3 (b), it can be easily found that all the multiples (surface multiples plus internal multiples) are removed simultaneously and automatically.

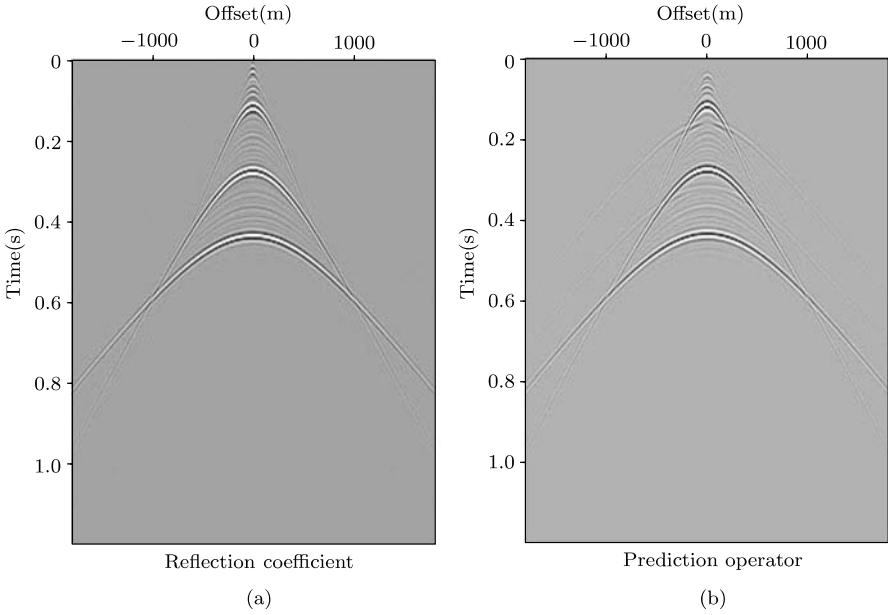


Fig. 13.1 (a) The reflection coefficients; (b) The prediction operator calculated by equation (13.9.10) giving the reflection coefficients.

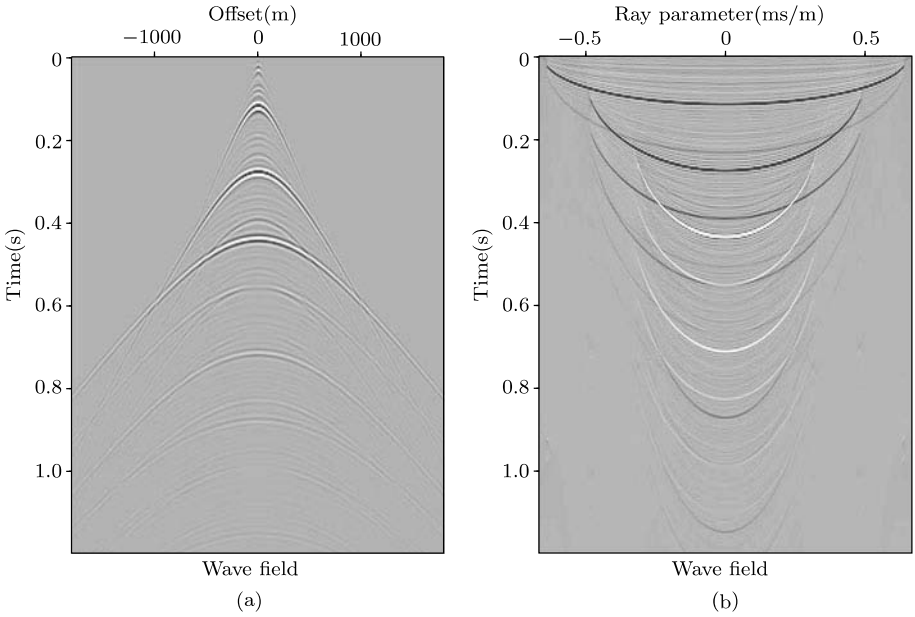


Fig. 13.2 (a) The wave field (primaries plus multiples) of a shot gather calculated by equation (13.9.19); (b) The linear Radon transformation of the wave field, i.e., the description of the wave field in ray parameter.

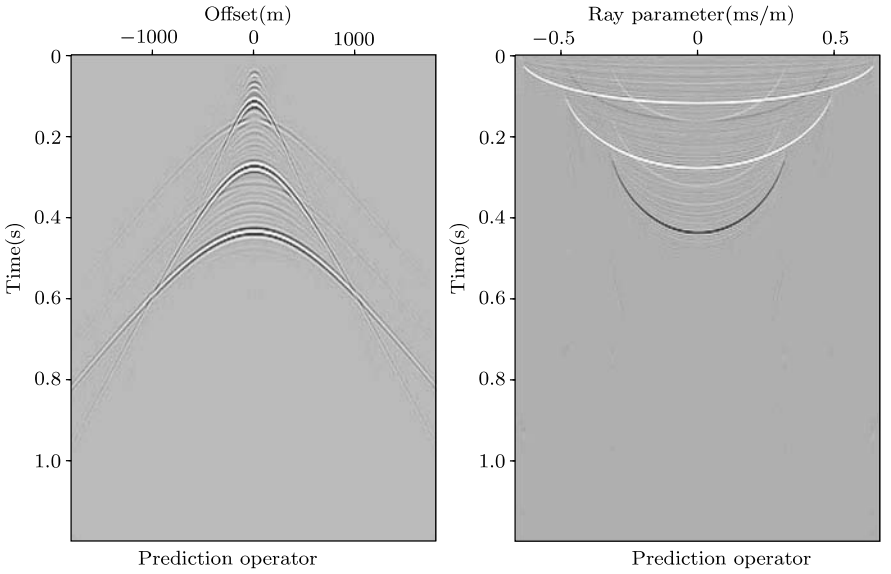


Fig. 13.3 (a) The prediction operator calculated by equation (13.9.19) and solved by a 2D spectral factorization of the wave field shown in Fig. 13.2 (a); (b) The radon transform of the prediction operator.

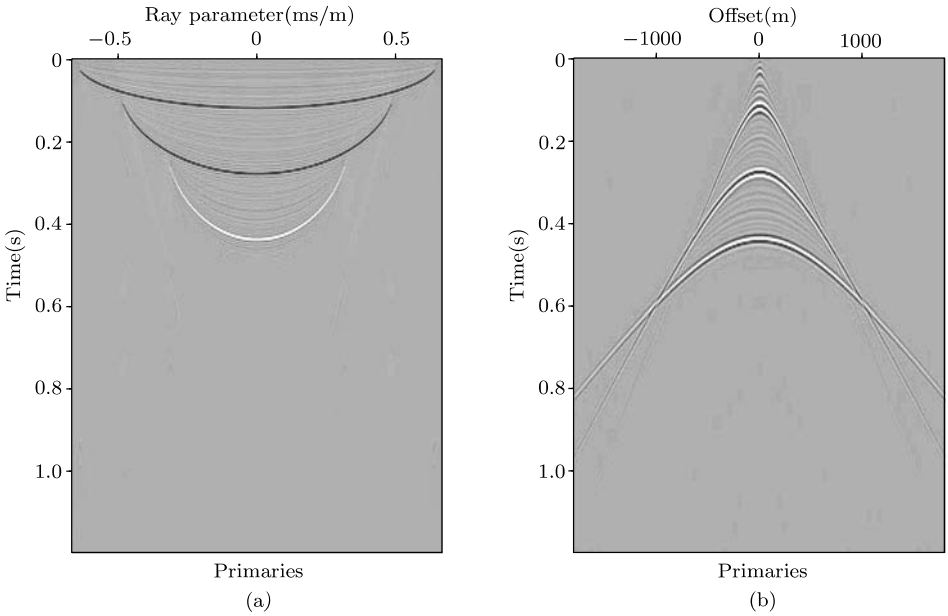


Fig. 13.4 (a) The primaries (reflection coefficients) calculated from the prediction operator shown in Fig. 13.3 (b); (b) The primaries of a shot gather Re-Radon transformed by the primaries shown in Fig. 13.4 (a).

13.10 Conclusions

In this chapter, the whole framework of the inverse scattering theory is presented. On the basis of the wave field decomposition and its flux conservation law,the pseudo-differential equation of the prediction operator is obtained. And by Lie Group method, it can be presented by the Lie integral expression which is the basis of the stratigraphic filter. And then it can bring a new method of the research of the absorption and the image of the full wave, and also increase the chance to reconstruct the real high frequency signal. The scattering wave field can be written as auto-correlation of prediction operator. And by spectral factorization of the wave field, the prediction operator can be obtained, which is the basis of the multiples elimination.

References

1. A. J. Berkhout and D. J. Verschuur, Removal of internal multiples with the common-focus-point (CFP) approach: part I — explanation of the theory, *Geophysics*, **70**(3), 45–60, 2005.
2. K. P. Bube and R. Burridge, The one-dimensional inverse problem of reflection seismology, *SIAM Review*, **25**(4), 497–559, 1983.
3. E. Celledoni and A. Iserles, Methods for the approximation of the matrix exponential in a Lie-algebraic setting, *IMA Journal of Numerical Analysis*, **21**(2), 463–488, 2001.
4. S. Chandrasekhar and N. Y. Dover, *Radiative Transferr*, New York: Dover, 1960.
5. J. B. Chen, H. Munthe-Kaas and M. Z. Qin, Square-conservative schemes for a class of evolution equations using Lie-group methods, *SIAM Journal on Numerical Analysis*, **39**(1), 2164–2178, 2002.
6. X. F. Chen, Seismogram synthesis for multi-layered media with irregular interfaces by global generalized reflection/transmission matrices method: I, theory of two-dimensional SH case, *Bulletin of the Seismological Society of America*, **80**(6), 1696–1724, 1990.
7. M. Cheney and J. H. Rose, Three-dimensional inverse scattering for wave equation: weak scattering approximations with error estimates, *Inverse Problems*, **4**(2), 435–447, 1988.
8. J. F. Claerbout and S. Fomel, *Image Estimation by Example, Geophysical Soundings Image Construction — Multidimensional Autoregression*, Palo Alto: Stanford University, 2002.
9. J. F. Claerbout, *Fundamentals of Geophysical Data Processing*, England: Blackwell Scientific Publications, 1985.
10. H. O. Cordes, *The Technique of Pseudodifferential Operators*, Cambridge University Press, Cambridge, 1995.
11. M. V. deHoop, Generalization of the Bremmer coupling series, *Journal of Mathematical Physics*, **37**(7), 3246–3282, 1996.
12. K. Feng, On difference schemes and symplectic geometry, in: K. Feng ed., *Symp. Diff. Geometry and Diff. Equations*, 42–58, Beijing: Science Press, 1985.
13. C. W. Frasier, Discrete time solution of plane p-sv waves in a plane layered medium, *Geophysics*, **35**(2), 197–219, 1970.
14. G. J. Fryer and L. N. Frazer, Seismic waves in stratified anisotropic media, *Geophys. J. R. astr. Soc.*, **78**, 691–710, 1984.

15. G. J. Fryer and L. N. Frazer, Seismic waves in stratified anisotropic media — II. Elastodynamic eigensolutions for some anisotropic systems, *Geophy. J. R. astr. Soc.*, **91**, 73–101, 1987.
16. S. H. Gray, Computational inverse scattering in multi-dimensions, *Inverse Problems*, **4**(1), 87–101, 1988.
17. P. Hubral, Time migration — some ray theoretical aspects, *Geophysi. Prosp.*, **25**, 738–745, 1975.
18. E. Iversen and M. Tygel, Image-ray tracing for joint 3D seismic velocity estimation and time-to-depth conversion, *Geophysics*, **73**(3), 99–114, 2008.
19. K. L. Larner, L. Hatton and B. S. Gibson, Depth migration of imaged time section, *Geophysics*, **46**(5), 734–750, 1981.
20. Li and Strahler, Geometric-optical bidirectional reflectance modeling of the discrete crown vegetation canopy: effect of crown shape and mutual shadowing, *IEEE Trans. Geosci. Remote Sens.*, **30**, 276–292, 1992.
21. G. F. Liu, H. Liu, B. Li and X. H. Meng, From the single square operator to the traveling time for prestack time migration earth science, *Journal of China University of Geosciences*, Research Report, IGG-CAS, 2009.
22. H. Liu, X. M. Wang and R. Zeng, Symbol description to integral solution of one-way wave operator, *Progress in Geophysics*, **22**(2), 463–471, 2007 (in Chinese).
23. H. Liu, J. H. Yuan, J.B. Chen, H. Shou and Y. M. Li, Theory and implementing scheme of large-step wavefield depth extrapolation, *Chinese Journal of Geophysics*, **49**(6), 1624–1641, 2006.
24. H. Liu, L. He, G. F. Liu and B. Li, *Derivation and Generalization of the Stratigraphic Filtering Formula via Lie Algebraic Integral*, the paper collection in memory of 80th anniversary of academician Liu Guangding, Beijing: Sciences Press, 2008.
25. H. Liu, Y. M. Li and R. S. Wu, 3D scattering operator and it's application in inverse scattering, *Progress in Geophysics*, **9**(2), 54–83, 1994 (in Chinese).
26. H. Liu and Y. M. Li, Review on the study of the reservoir geophysical method in hydrocarbon re-exploration, *Oil & Gas Geology*, 2008 (in Chinese).
27. H. Liu, J. Yuan and Y. Gou, Spectral factorizationof wavefield and operator in seismic inverse scattering, *Chinese Journal of Geophysics*, **50**, 231–240, 2007.
28. K. Matson, An overview of wavelet estimation using free-surface multiple remove, *The Leading Edge*, **19**(1), 50–55, 2000.
29. R. I. McLachlan and G. R. W. Quispel, Splitting methods, *Acta Numerica*, **11**(1), 341–434, 2002.
30. R. G. Newton, *Scattering Theory of Wave and Particles*, New York: Springer-Verlag, 1982.
31. R. F. O'Doherty and N. A. Anstey, Reflections on amplitudes, *Geophysical Prospecting*, **19**(3), 430–458, 1971.
32. R. T. Prosser, Formal solutions of inverse scattering problems, *Journal of Mathematical Physics*, **10**(8), 1819–1822, 1969.
33. M. Y. Qi, C. J. Xu and W. K. Wang, *Introduction of Modern Partial Differential Equation*, Wuhan: Wuhan University Press, 2005 (in Chinese).
34. J. H. Rose and B. DeFacio, Three-dimensional inverse scattering, in: I. W. Knowles, Y. Saito, ed., *Differential Equations and Mathematical Physics*, 46–54, New York: Springer-Verlag, 1987.
35. J. H. Rose, Elastic wave inverse scattering in nondestructive evaluation, *Pure and Applied Geophysics*, **131**(4),715–739, 1989.
36. P. Sava and S. Fomel, Riemannian wavefield extrapolation, *Geophysics*, **70**(3), 45–56, 2005.
37. J. C. Shragge, Riemannian wavefield extrapolation: Nonorthogonal coordinate systems, *Geophysics*, **73**(2), 11–21, 2008.
38. H. B. Song, *Inversion Research on Elastic Parameters of Stratified*, Beijing: Higher Education Press, 2002 (in Chinese).

39. D. J. Verschuur and A. J. Berkhout, Removal of internal multiples with the common-focus-point(CFP) approach: part 2-application strategies and data examples, *Geophysics*, **70**(3), 61–72, 2005.
40. C. P. A. Wapenaar and J. L. T. Grimbergen, Reciprocity theorem for one-way wavefield, *Geophysical Journal International*, **127**(1), 169–177, 1996.
41. A. Weglein, D. Foster and K. Matson, An inverse-scattering sub-series for predicting the spatial location of reflectors without the precise reference medium and wave velocity, *Expanded Abstracts of 71st Annual Internat SEG Mtg*, 2108–2111, 2001.

Chapter 14
Tikhonov Regularization for Gravitational Lensing Research

Boris Artamonov, Ekaterina Koptelova, Elena Shimanovskaya and Anatoly G. Yagola

Abstract. The gravitational lensing phenomenon can provide us with the information about luminous and dark matter in our Universe. But robust and effective tools are needed to extract that valuable information from observations. In this chapter, two inverse problems arising in gravitational lensing research are considered. Both problems are ill-posed, so regularization is needed. The first problem is the one of image reconstruction. Based on the Tikhonov regularization approach, several modifications of the regularizing algorithm, taking account of specific properties of gravitational lens systems, are proposed. The numerical results show that incorporation of all available *a priori* information allows reconstructing images of gravitational lens systems quite well. Furthermore, the algorithm decomposes the images into two parts – point sources and smooth background, that is needed for further investigations. The second problem concerns reconstruction of a distant quasar light distribution based on observations of microlensing events. In this case, a gravitational lens system as a gravitational telescope and the Tikhonov regularization method as a tool allow getting valuable information about the distant quasar unresolved by an instrument.

Boris Artamonov, Ekaterina Koptelova, Elena Shimanovskaya
Sternberg Astronomical Institute of Moscow State University, Universitetskiy prospekt, 19, Moscow, Russia.
e-mail: {artamon,koptelova,eshim}@sai.msu.ru

Anatoly G. Yagola
Department of Mathematics, Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia.
e-mail: yagola@inverse.phys.msu.ru

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

14.1 Introduction

There are various inverse problems almost in fields of sciences. And astronomy, as an observational science, is not an exception. Astronomers mostly deal with observational manifestation of space objects: images, spectra, lightcurves and so on. They have to solve various inverse problems from the image restoration to the reconstruction of object properties based on observations. Reconstructing images of space objects is one of the most important topics in astronomy, especially for the gravitational lens investigation, because distances between objects are almost equal to the resolution of a telescope.

According to the theory of general relativity developed by Albert Einstein, the presence of matter can curve spacetime and alter the trajectory of light [4]. That prediction was first verified during solar eclipse in 1919, when apparent positions of stars behind the solar limb had changed temporarily under the influence of the sun gravity [3]. Thus, if a massive body (a galaxy) is situated between a remote bright source (a quasar) and an observer, the image of the source is distorted. This phenomenon is called gravitational lensing. The deflecting body is called "a lens". A simple model of lensing is presented in Fig. 14.1 a). When the source and the deflector (lens) are perfectly aligned, an Einstein ring is seen by the observer (left figure). When the symmetry is broken, multiple images of the source can be observed (right figure).

Let θ denote an angle between a beam and an optic axis, and let β denote a position of an unlensed source. Then

$$\beta = \theta - \alpha(\theta), \tag{14.1.1}$$

where $\alpha(\theta)$ is a deflection angle. Lens equation (14.1.1) describes a mapping between the source plane and the lens plane, in general, it is irreversible. For fixed source position β there can be several solutions θ, which correspond to several source images. To solve the lens equation, a model of mass distribution of the deflector is needed. Positions in the lens plane, where the Jacobian of the lens equation is equal to zero, are called critical curves. Positions in the source plane corresponding to critical curves are called caustics. An example of critical curves and caustics for the isothermal sphere model of a deflector is presented in Fig. 14.1 b).

In dependance on the relative positions and distances between source, lens and observer, a whole zoo of lensed quasar images can be observed (Fig. 14.2). An image of a gravitational lens system can represent some complex piecewise continuous light distribution or it can consist of several intersecting images of point sources (images of the lensed object). In some cases, those images are superimposed on the image of the refracting body itself. Gravitational lensing can provide us with information about lensing galaxies, distant sources, dark matter and the structure of our Universe, e.g. the value of the Hubble constant. A more detailed introduction can be found in [19].

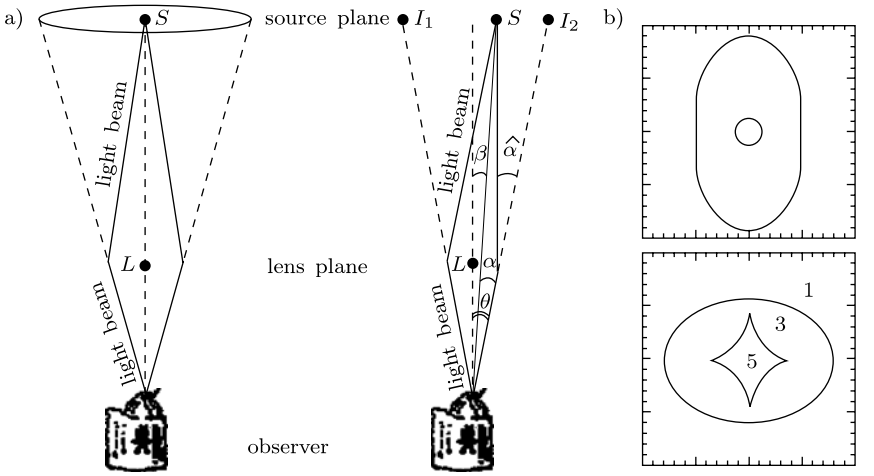
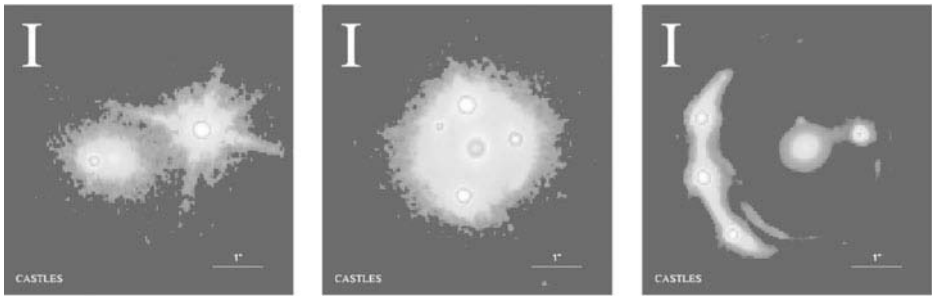


Fig. 14.1 a) A scheme of gravitational lensing: the Einstein ring, the double image, I_1 and I_2 are images of the same source S; b) critical curves (upper) and caustics (lower) of the elliptic lens, numbers correspond to 1, 3 and 5 images of a source.



UM673 SDSS0924+0219 RXJ1131-1231

Fig. 14.2 Images of some gravitational lens systems [8].

Investigation of gravitational lensing effect is closely related to various inverse problems. First of all, fast and robust algorithms for reconstructing images of gravitational lens systems acquired at the ultimate resolution are needed. A deconvolution algorithm for images composed of point sources and smooth background, that was developed for purpose of gravitational lens image reconstruction based on Tikhonov regularization approach, is described in the next chapter. That algorithm allows obtaining lightcurves of a lensed quasar components. Having those lightcurves and using a unique properties of gravitational lensing we can reconstruct an intrinsic bright distribution of a distant quasar. The problem of one-dimensional quasar profile reconstruction through the analysis of high magnification microlensing events in components of the quadruple gravitational

lens QSO2237+0305 is solved in the third chapter using Tikhonov regularization method.

14.2 Regularized deconvolution of images with point sources and smooth background

14.2.1 Formulation of the problem

The initial problem astronomers face is a need for processing images obtained with a telescope. An example of the gravitational lens image is presented in Fig. 14.3, in which A, B, C, and D are images of the same quasar Q2237+0305.

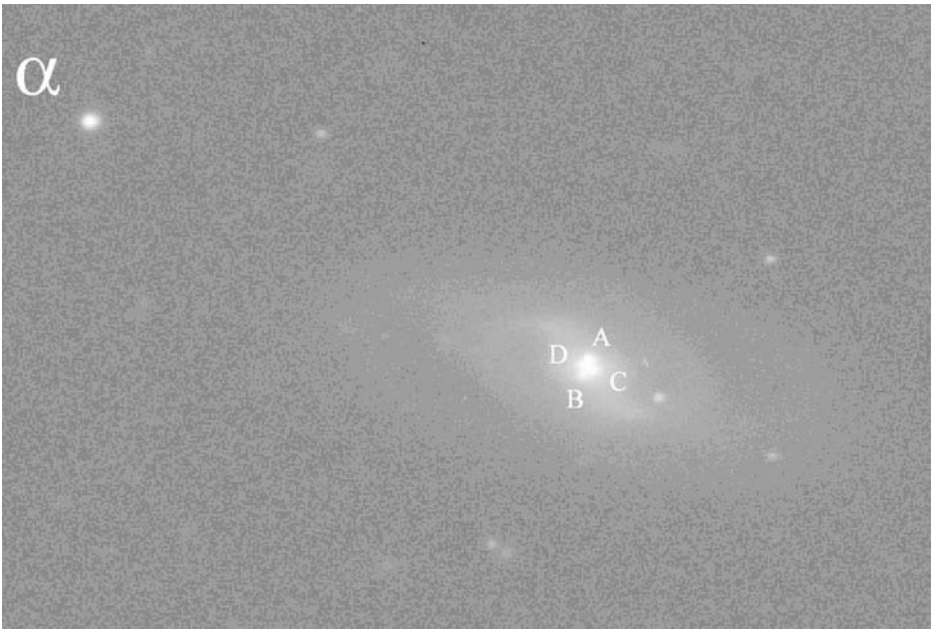


Fig. 14.3 An example of the CCD frame containing the image of the Q2237+0305 gravitational lens system (Einstein Cross).

The images obtained with ground-based telescopes are under corruption due to atmospheric perturbations and the telescope finite resolution. The simple model of the image formation can be represented as the Fredholm integral equation of the first kind:

$$A[z](x,y) = \iint\limits_B k(x-\xi, y-\eta) z(\xi,\eta)\, d\xi d\eta = u(x,y). \tag{14.2.1}$$

Here A is the convolution operator, $z(x, y)$ represents the unknown light distribution of the object, or sought solution, $k(x - \xi, y - \eta)$ is an instrument function which characterizes the distortion of the image, it is also called the point spread function (PSF), $u(x, y)$ represents the observed light distribution, B is the frame area, $B = [0, L] \times [0, L]$.

Let us assume the following smoothness properties of functions in (14.2.1): $u(x, y) \in U \equiv L_2[R^2]$, $k(x, y) \in L_1[R^2] \cap L_2[R^2]$. The convolution operator A is the linear operator which acts from some Hilbert space Z to the Hilbert space of second-power-integrable functions L_2. Suppose there is a unique solution of (14.2.1) $\bar{z}(x, y) \in Z$ for some function $\bar{u}(x, y) \in U$.

Instead of exact data, we have those approximations $u_\delta(x, y)$: $\|u_\delta - \bar{u}\|_{L_2} \leq \delta$, and $k_h(x, y)$: $\|A_h[z] - A[z]\|_{L_2} \leq \Psi(h, \|z\|), z \in Z(B)$. Let δ and $\Psi(h, \|z\|)$ be known and have standard properties [12]. Thus, having at our disposal the approximate data and those inaccuracies $A_h, u_\delta, h, \delta, \Psi$ we have to find the approximate solution of (14.2.1) that converges to the exact solution in the context of the norm of the function space Z.

14.2.1.1 Model of the CCD image

The continuous image of the object is the light energy density. It corresponds to the number of photons registered in every CCD (Charge-Coupled Device) cell. The energy in some CCD cell is:

$$g_{ij} = \int\int_{\Omega_{ij}} u(x, y)dxdy. \tag{14.2.2}$$

The statistical model of data in the ij-th pixel of CCD can be represented as a sum of the Poisson stochastic variable and the Normal stochastic variable:

$$U_{ij} \sim Poisson(g_{ij}) + Gauss(0, \sigma^2). \tag{14.2.3}$$

The first summand simulates a photon count, the second one simulates a background noise of the registration system. The realization of U_{ij} will be denoted by $\hat{u}_{ij}$. An array $\hat{u}$ that consists of $\hat{u}_{ij}$ is called a discrete image.

To increase the signal-to-noise ratio, frames can be averaged. The light energy uncertainty in every pixel for the averaged frame is calculated based on CCD properties:

$$\delta_{ij}^2 = \frac{\hat{u}_{ij}}{N \cdot g} + \frac{n^2}{N}, \tag{14.2.4}$$

where N is a number of frames to be averaged; $\hat{u}_{ij}$ the ij-th pixel count in the averaged frame; g the CCD gain factor; n the noise. Based on (14.2.4), the error estimate δ for $u_\delta(x, y)$ is:

$$\delta = \sqrt{\sum \delta_{ij}^2} = \sqrt{\sum (\frac{\hat{u}_{ij}}{N \cdot g} + \frac{n^2}{N})}. \tag{14.2.5}$$

If the ratio error of the kernel $\hat{k}$ is known and is equal to p, the operator uncertainty can be estimated in the following way:

$$h = \sqrt{\sum (p \cdot k_{ij})^2}. \tag{14.2.6}$$

14.2.1.2 Model of the kernel

The kernel of equation (14.2.1) is the primary characteristic of the imaging system. It is called an instrument function, or the Point Spread Function (PSF) – the response of an instrument to the point source. So it can be obtained through fitting a star in the frame with some model. For our images, the PSF is well modeled by Gauss function:

$$k(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\{-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\}, \tag{14.2.7}$$

$$x' = x \cdot \cos\varphi - y \cdot \sin\varphi, \tag{14.2.8}$$

$$y' = x \cdot \sin\varphi + y \cdot \cos\varphi. \tag{14.2.9}$$

Thus, we can fit the brightness distribution of a star in the frame with the following function:

$$K(x, y) = a_1 \cdot k(x - x_0, y - y_0) + a_2. \tag{14.2.10}$$

To find the unknown parameters $\{\sigma_x, \sigma_y, x_0, y_0, \varphi, a_1, a_2\}$, the least-squares method can be used. It corresponds to the search for the minimum of the following function:

$$\hat{\Phi}[\sigma_x, \sigma_y, x_0, y_0, \varphi, a_1, a_2] = \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} \frac{d^2}{\delta_{ij}^2} \cdot (\hat{K}_{ij}[\sigma_x, \sigma_y, x_0, y_0, \varphi, a_1, a_2] - \hat{S}_{ij})^2, \tag{14.2.11}$$

where $\hat{S}_{ij}$ is a star brightness distribution, δ_{ij} the intensity error in the pixel with coordinates (x_i, y_j), d a grid pitch.

Fig. 14.4 shows the results of fitting the star with the two-dimensional Gauss function.

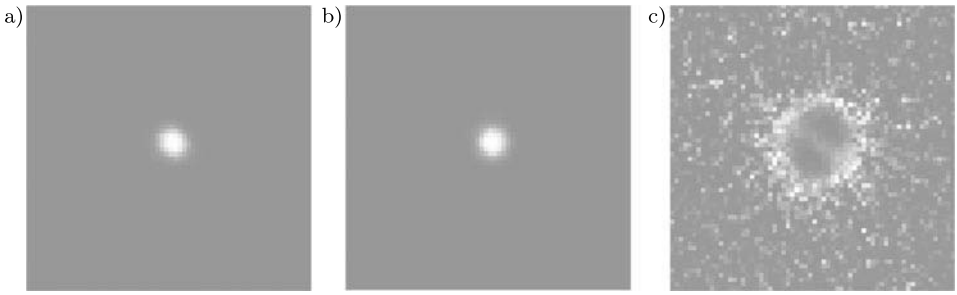


Fig. 14.4 Kernel modeling: a) a star from the frame; b) the model of the instrument function; c) simulation errors distribution.

14.2.2 Tikhonov regularization approach

The problem (14.2.1) belongs to the general class of ill-posed problems in the sense of Hadamar [7], i.e. it violates the conditions of the well-posedness for operator equations:

$$A[z] = u, \quad z \in Z, \quad u \in U, \tag{14.2.12}$$

where Z and U are normalized spaces. The problem (14.2.12) is called well-posed in the class of its admissible data if for any pair from the set of admissible data the solution: 1) exists, 2) is unique, 3) is stable, i.e. continuously depends on errors in A and u. If at least one of the requirements is not met, then the problem is called ill-posed.

To ensure the fulfillment of the first and second conditions, one can introduce a generalized solution concept. Let's introduce a class of pseudosolutions $Z^* \equiv \arg\inf_{z \in Z} \|A[z] - u\|_U$ and a function $\Omega[z]$, then one can choose Ω-optimal pseudosolution $\bar{z} = \arg\inf_{z \in Z^*} \Omega[z]$ as a generalized solution.

As a generalized solution, the so-called normal pseudosolution is often taken. It is the solution that minimizes discrepancy and is at the minimum distance from some fixed element of Z:

$$\tilde{z} = \arg \inf_{z \in Z^*} \|z - z^0\|. \tag{14.2.13}$$

It exists and is unique for any exact data of the problem (14.2.12) if A is a linear operator acting from the Hilbert space Z to a Hilbert space U: $A \in L(Z, U)$; $\bar{u} \in R(a) \oplus R(a)^{\perp}$ and $\tilde{z} = A^{+}[\bar{u}]$. Here $R(A)$ and $R(A)^{\perp}$ denote the range of the operator A and its orthogonal complement in U, and A^{+} stands for the operator pseudoinverse to A [12].

Academician Tikhonov in his fundamental work [16] introduced an idea of the regularizing algorithm as a way to find an approximate solution $z_\eta = R(A_h, u_\delta, \eta)$ that corresponds to the input data (A_h, u_δ, η) of (14.2.12), and has the convergence property $z_\eta \to \bar{z} = A^{+}[\bar{u}]$ as $\eta \to 0$. Since then this method has been extensively studied and widely adopted in many fields of sciences.

As an approximate solution of (14.2.12), the minimizer of the the smoothing function can be taken as:

$$z^{\alpha_\eta} = \arg \inf_{z \in Z} M^\alpha[z], \tag{14.2.14}$$

where

$$M^\alpha[z] = \|A_h[z] - u_\delta\|^2_{L_2} + \alpha \cdot \Omega[z] \tag{14.2.15}$$

is called a smoothing function.

The first term in (14.2.15) represents the squared discrepancy and is responsible for goodness of fit.

$\Omega[z]$ is called a stabilizer function. On the one hand, it is introduced for selection of a pseudosolution and has to contain information about the physics of the object, e.g. about the smoothness of the solution. On the other hand, it must have regularization property, i.e. ensuring the existence and stability of optimal pseudosolutions. A norm of the sought solution in some function space is often taken as a stabilizer function. The function space is chosen based on assumptions about smoothness of the sought solution.

α is a regularization parameter. The choice of α is crucial for solving ill-posed problems. It has to control the trade-off between the assumptions about smoothness of the sought solution and its consistency with the data. Generally, α should depend on the input data, their errors, and the method of approximation of the initial problem.

14.2.2.1 Discrepancy principle

One of the ways to coordinate the regularization parameter with the error of the input data is the discrepancy principle: if $\|u_\delta\| > 0$, then $\alpha > 0$ is a root of the following equation:

$$\rho(\alpha) \equiv \|A[z^\alpha_\delta] - u_\delta\|_U \simeq \delta. \tag{14.2.16}$$

The condition means that the discrepancy is equal to the noise level of the image.

Provided that the regularization parameter α is chosen according to this rule, z^α_δ can be considered as an approximate solution which tends to the exact solution in the context of the norm of the chosen function space as the error level of input data tends to zero.

If it is possible to solve the minimization problem for M^α with fixed $\alpha > 0$, then the root of (14.2.16) can be solved by means of one of the standard methods. The function $\rho(\alpha)$ is strictly monotonous, continuous and takes on all values from $[\delta, \|u_\delta\|_U)$, so (14.2.16) has only one root $\alpha > 0$. As a lower limit for α, the noise level δ can be taken. The upper limit for α was suggested in [5]:

$$\alpha^{**} = \frac{M^2\delta}{\|u_\delta\|_{L_2} - \delta}, \tag{14.2.17}$$

where

$$M = \int\int |k(x,y)|dxdy. \tag{14.2.18}$$

To find the root of (14.2.16), the following algorithm can be utilized. We specify the initial value of α so that $\rho(\alpha) > 0$, for example $\alpha^{(0)} = \alpha^{**}$. Then we solve the Tikhonov function minimization problem, compute the discrepancy value and check the equality (14.2.16) with the given accuracy. The next value $\alpha^{(s)}$ is chosen by formula $\alpha^{(s)} = \frac{\alpha^{(s-1)}}{10}$. The process is repeated until $\rho(\alpha^{(\tilde{s})}) \leq \delta$. This $\alpha^{(\tilde{s})}$ is taken as an approximate root of equation (14.2.16). This value can be fine tuned with the bisection method.

The discrepancy principle is a posteriori way of regularization parameter choice. It guarantees the convergence of approximate solutions $z^\alpha \to \bar{z}$ and a certain discrepancy value for the approximate solution. The regularizing algorithm with the discrepancy principle converges and is optimal [12].

14.2.2.2 Generalized discrepancy principle

If it is possible to estimate the uncertainty of the kernel, the generalized discrepancy principle can be adopted for the regularization parameter choice. Let us consider the function:

$$\rho_\eta(\alpha) \equiv \|A_h[z^\alpha] - u\|^2_{L_2} + (\delta + h\sqrt{\Omega[z^\alpha]})^2, \tag{14.2.19}$$

where $\eta = (h, \delta)$. The function $\rho_\eta(\alpha)$ is called the generalized discrepancy. If $\|u_\delta\|_U > \delta$, then $\rho_\eta(\alpha)$ is continuous, strictly monotone increasing and $\in (-\delta^2, \|u_\delta - \delta^2\|)$ for fixed η. Hence an equation

$$\rho_\eta(\alpha) = 0 \tag{14.2.20}$$

has a unique root $\alpha > 0$ that can be found by means of one of standard methods (the bisection method or the golden section method).

If z^α_η is an extremal of Tikhonov function and α_η is chosen according to the generalized discrepancy principle, then approximate solutions converge to the exact solution: given $\eta \to 0, lim_{\eta\to 0}\|z^\alpha - \bar{z}\|_Z = 0$.

14.2.3 *A priori* information

To obtain stable and physically valid results of the image reconstruction, the regularization ought to be based on a priori knowledge of the properties of the sought solution.

The first *a priori* assumption about sought solution is its nonnegativity:

$$z(x,y) \geq 0. \tag{14.2.21}$$

Various assumptions about the structure of the object under study can also be taken into account. Images of many gravitational lens systems consist of multiple overlapped quasar images superimposed on a background galaxy. So the image can be decomposed into two constituent parts – the sum of Q δ-functions and smooth background (galaxy):

$$z(x,y) = g(x,y) + \sum_{q=1}^{Q} a_q \delta(x - b_q, y - c_q), \tag{14.2.22}$$

where Q is the number of point sources with coordinates (b_q, c_q) and intensity a_q in the frame; $g(x,y)$ is the solution's component corresponding to a galaxy; δ represents Dirac function. We will look for the solution in the form (14.2.22).

Theorem 14.2.1. *If the kernel $K(u,v)$ satisfies $\hat{K}(\omega_1,\omega_2) \neq 0$ for all (ω_1,ω_2) (the hat corresponds to the Fourier transform), then the solution* (14.2.22) *is unique.*

Proof. It is sufficient to demonstrate that the homogeneous equation

$$k(x,y) * \{g(x,y) + \sum_{q=1}^{Q} a_q \delta(x - b_q, y - c_q)\} = 0 \tag{14.2.23}$$

has only trivial solution:

$$g(x,y) + \sum_{q=1}^{Q} a_q \delta(x - b_q, y - c_q) = 0, \tag{14.2.24}$$

$$g(x,y) = 0,$$

and

$$\sum_{q=1}^{Q} a_q \delta(x - b_q, y - c_q) = 0.$$

After Fourier transform of (14.2.23), we have:

$$\hat{k} \cdot \{\hat{g} + \sum_{q=1}^{Q} a_q \exp(-i\omega_1 b_q, -i\omega_2 c_q)\} = 0. \tag{14.2.25}$$

Hence

$$\hat{g} = \sum_{q=1}^{Q} a_q \exp(-i\omega_1 b_q, -i\omega_2 c_q). \tag{14.2.26}$$

The left part of (14.2.26) belongs to L_2, but right part doesn't. The contradiction proves the theorem. □

A prior knowledge about the smoothness of the unknown solution is embedded in the regularizing algorithm through the appropriate choice of the stabilizer function. In most cases it is the squared norm of the solution in some function space: $\Omega[z] = \|z\|_Z^2$. The choice of the stabilizer affects the order of convergence of approximate solutions.

14.2.3.1 Uniform regularization in the Sobolev space W_2^2

If *a priori* information about sought solution allows assuming the high-order smoothness of Z and choosing $Z \equiv W_2^2$, where W_2^2 is a set of L_2-functions having generalized derivatives of the second order which are second-power-integrable, the stabilizer can be written in the following form:

$$\Omega[z] = \|z\|_{W_2^2}^2 \equiv \iint\limits_B \left\{ z^2 + \left(\frac{\partial^2 z}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 z}{\partial y^2}\right)^2 \right\} dxdy. \tag{14.2.27}$$

The discrete representation of the Tikhonov function (14.2.15) is as follows:

$$\hat{M}^\alpha[\hat{z}] = \sum_{m=0}^{N_1-1}\sum_{n=0}^{N_2-1} \left\{ u_{mn} - \sum_{i=0}^{N_1-1}\sum_{j=0}^{N_2-1} k_{m-i,n-j}\left(\sum_{q=1}^{Q} a_q \delta_{i-b_q,j-c_q} + g_{ij}\right)\right\}^2$$
$$+\alpha \sum_{m=0}^{N_1-1}\sum_{n=0}^{N_2-1} \left\{ g_{mn}^2 + \left[\frac{\partial^2 g}{\partial x^2}(x_m, y_n)\right]^2 + 2\left[\frac{\partial^2 g}{\partial x \partial y}(x_m, y_n)\right]^2 + \left[\frac{\partial^2 g}{\partial y^2}(x_m, y_n)\right]^2 \right\} \tag{14.2.28}$$

When selecting α in accordance with the discrepancy principle, approximate solutions z^α tend to the exact solution of the problem as δ tends to zero in the context of the W_2^2 norm:

$$\|z^\alpha - z\|_{W_2^2} \to 0 \quad \text{as} \quad \delta \to 0.$$

According to the Sobolev's embedding theorem, $W_2^2[B]$ is embedded in $C[B]$ – the set of continuous functions on B. Thus, the convergence in the context of the W_2^2-norm means the convergence in the context of the norm of $C[B]$, i.e. regularized solutions converge to the exact solution uniformly:

$$\max_{(x,y)\in B} |z^\alpha(x,y) - z(x,y)| \to 0 \quad \text{as} \quad \delta \to 0.$$

The results of the W_2^2 regularization are presented in Fig. 14.5.

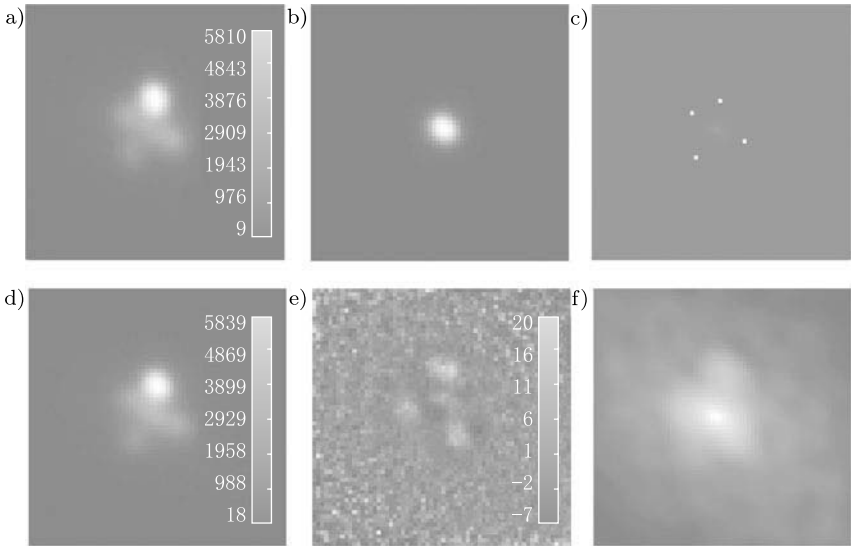


Fig. 14.5 Solution in the Sobolev space W_2^2 (frame size is 64×64): a) the image (data); b) the kernel; c) the solution; d) the solution convolved with the kernel (model); e) the residual $(\frac{data-model}{\sqrt{model}})$; f) the galaxy part of the solution in the logarithmic scale. $\alpha = 1.16 \cdot 10^{-4}$

14.2.3.2 Piecewise uniform regularization in the space of functions of bounded variation

Rapid intensity variations in the observed data can be processed by selection of the function representing the light distribution of the background galaxy from the appropriate function space. Let the smooth part of the solution belong to the class of functions with limited total variation defined by A. Leonov [11]. In that paper an approach to piecewise uniform regularization of two-dimensional ill-posed problems based on TV class of functions was developed.

Let us consider an arbitrary grid $S_{N_1N_2}$ introduced on B and define the total variation for a function z on B as follows:

$$V(z,B) = \sup_{S_{N_1N_2}} \left(\sum_{m=1}^{N_1-1} |z_{m+1,1} - z_{m,1}| + \sum_{n=1}^{N_2-1} |z_{1,n+1} - z_{1,n}| + \right.$$
$$\left. + \sum_{m=1}^{N_1-1} \sum_{n=1}^{N_2-1} |z_{m+1,n+1} - z_{m+1,n} - z_{m,n+1} + z_{m,n}|, \forall S_{N_1N_2} \right).$$

The function for which the total variation is a finite quantity is called bounded total variation function. It is continuous nearly everywhere with the exception, possibly, of the points of discontinuity positioned on the countable set of gridlines.

The regularizing algorithm with the proper choice of the regularization parameter and the stabilizer function

$$\Omega[z] \equiv \|z\|_{\nu[B]} = |z(0,0)| + V(z,B) \tag{14.2.29}$$

provides piecewise uniform convergence of approximate solutions [11].

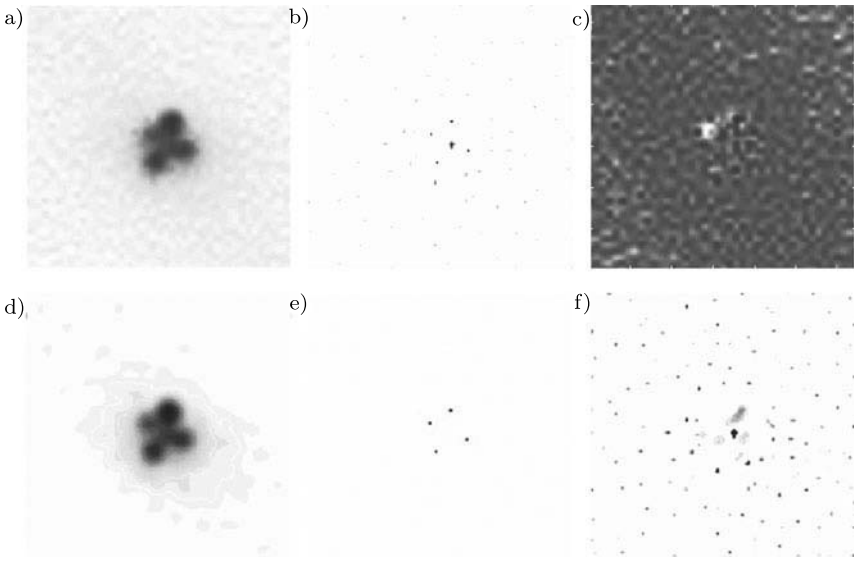


Fig. 14.6 The solution in the space of functions of bounded variation: a) the image; b) the solution; c) the residual; d) the solution convolved with the kernel finction; e) the quasar components part of the solution f) the galaxy part of the solution

The results of the image reconstruction are presented in Fig. 14.6. The solution part representing the lensing galaxy (Fig. 14.6 f)) is so discontinuous that the stabilizer for the image under consideration can not be used.

14.2.3.3 Closeness of the solution to some model

Additionally, one can penalize the unknown solution drastically different from the certain analytical model and construct the stabilizer in the following form:

$$\Omega[z] = \|g - g_{model}\|_G^2. \tag{14.2.30}$$

In this work, we assume that the light distribution in the central region of the galaxy is well modeled by generalized de Vaucouleurs profile (Sersic's model):

$$g_{model}(r) = I_0 \exp^{-b_n(\frac{r}{r_e})^{\frac{1}{n}}}, \tag{14.2.31}$$

where $b_n = 2n - 0.324$ for $1 \leq n \leq 4$.

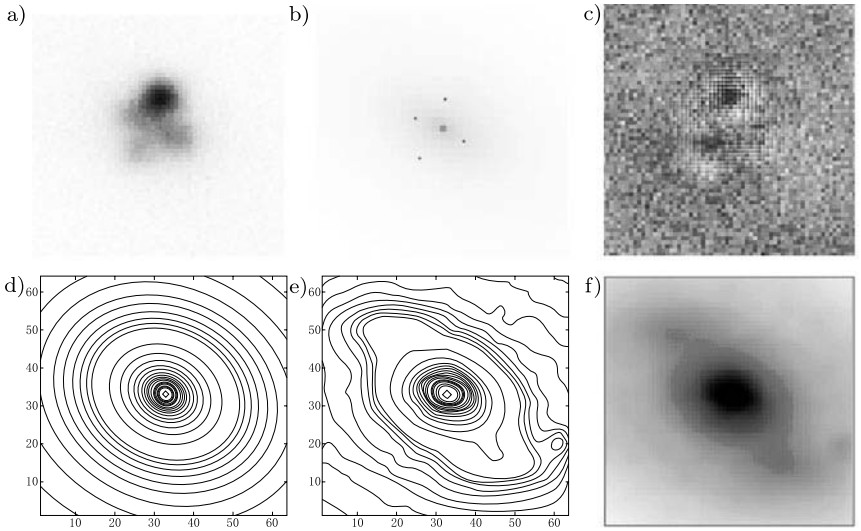


Fig. 14.7 The solution in the L_2 space of functions close to Sersic model: a) the image; b) the solution; c) the residual; d) the contour plot of the Sersic model; e) the contour plot of the galaxy part of the solution; f) the galaxy part of the solution. $\alpha = 5.8 \cdot 10^{-5}$.

Results of Q2237+0305 image reconstruction with the stabilizer function (14.2.30) where $G = L_2$ are presented in Fig. 14.7. The parameters of the Sersic model were calculated at the preliminary stage using the least-squares method. For the minimization of the smoothing function, the conjugate gradient method was used. Regularization with that stabilizer guarantees that approximate solutions converge to the exact solution in the context of the norm L_2, i.e. converges in mean (zero-order convergence):

$$\|z^\alpha - \bar{z}\|_{L_2} \to 0 \quad \text{as} \quad \delta \to 0.$$

Summary of *a priori* information used for quadruple gravitational lens image reconstruction: two-component representation of the solution – quasar components and a galaxy, the assumption about the smoothness of the galaxy part of the solution and its closeness to the Sersic model.

For the minimization of the Tikhonov function, the conjugate gradient method in combination with the enumeration of possible coordinates of the quasar components in the neighborhood of the initial points (b_{q_0}, c_{q_0}) was used.

The developed image reconstruction algorithm allows obtaining light distribution of lensing galaxy and coordinates and fluxes of quasar components. Furthermore, it allows getting light curves of quasar components during observation period (maybe years). The light curves of the Q2237+0305 "Einstein Cross" gravitational lens system based on Maidanak observations in 1995–2003 are presented in Fig. 14.8. And fortunately, thanks to unique properties of gravitational lensing phenomenon, the light curves can tell us more about the distant source $\cdots$

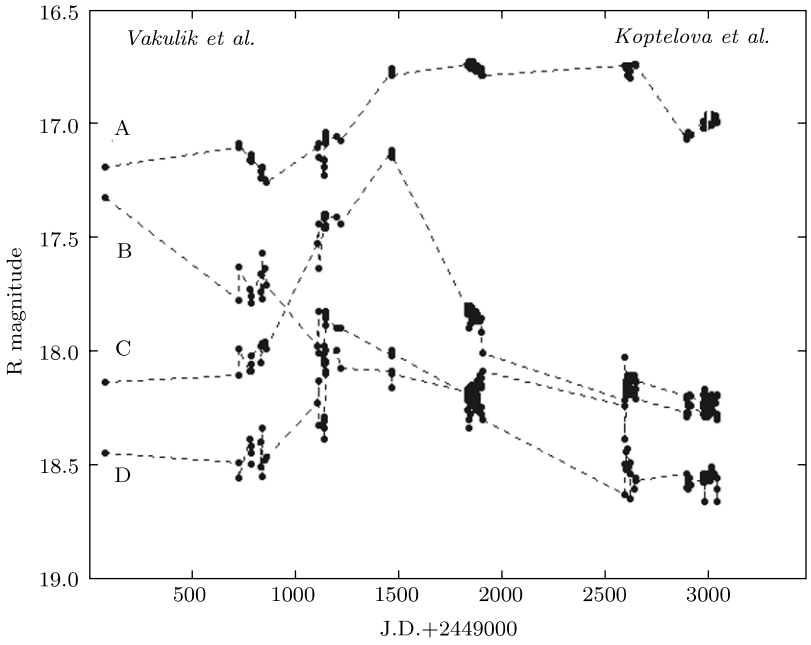


Fig. 14.8 Combined light curves of Q2237+0305 gravitational lens system "Einstein Cross" in the R band for 1995–2003 based on data from [17] and [9].

14.3 Application of the Tikhonov regularization approach to quasar profile reconstruction

14.3.1 Brief introduction to microlensing

As can be seen in Fig. 14.8, fluxes of quasar components vary. The cause may be the internal quasar variability or the microlensing variability. Microlensing fluctuations are produced by compact sub-structures of the lensing galaxy. The internal quasar variability manifests itself in all quasar components with some delay, whereas the microlensing variability is observed in certain quasar component. The microlensing variability can tell us about compact objects on the line of sight, quasar size, quasar brightness distribution, lens mass, and more. The microlensing with high amplification is of particular interest. It occurs when a source crosses a caustic – a map of the critical curve in the source plane (Fig. 14.1b)). The most probable type of the microlensing is connected with intersection of the fold caustic [18]. In this case, an additional couple of unresolved micro-images appears (or disappears) and the significant magnification of the flux of one of the quasar components can be observed.

The possibility of getting information on the size and possible multiple structure of the lensed source from the light curve was discussed in [2]. Since various approaches to reconstruct spatial structure of the quasar accretion disk have been proposed. In [14] analytical accretion disk model fitting method was applied to light curve of the image A of the Q2237+0305. In [6, 1, 13] the restoration of the one-dimensional profile from the high amplification light curve by means of a regularization method was proposed. Model-independent methods of observational data analysis can be used to test adequacy of accretion disk models employed in theoretical studies. Further we consider the algorithm for reconstruction of the accretion disk brightness profile based on the technique suggested in [6] using Tikhonov regularization approach [15].

14.3.2 Formulation of the problem

Assuming small angular size of the source in comparison with the curvature of the caustic, the caustic can be considered as a straight line. A scheme of the caustic crossing is presented in Fig. 14.9. Let the origin of cartesian coordinates match the source center, the x axes is perpendicular to a caustic. The magnification factor of a point source close to a caustic line along the x axis can be expressed as follows [2]:

$$A(x, x_c) = \frac{k}{\sqrt{x_c - x}} H(x_c - x), \tag{14.3.1}$$

where k is the caustic strength; $x_c = V_c(t - t_0)$ is the position of the caustic; V_c is the transverse velocity of the caustic, which is assumed to be constant; t_0 is the time when caustic crosses the center of the source (x=0); $H(x_c - x)$ is the Heaviside staircase function.

A one-dimensional profile of the source brightness distribution scanned by the caustic line along the x axis is:

$$P(x) = 2 \int_0^{\sqrt{R_s^2 - x^2}} I(x, y) dy, \tag{14.3.2}$$

where R_s is the radius of the source. Caustic crossing is accompanied by drastic magnification of the flux from one of the quasar images. The total magnification of an extended source with the brightness distribution $I(x, y)$ is obtained by summation over all micro-images and integration over the source. The observed light curve is supposed to be the convolution of the quasar brightness distribution and the magnification pattern:

$$\begin{aligned} F_{tot}(x) &= \int\int_S I(x, y) A(x, x_c) dx dy \\ &= k \int_x^{x_c} \frac{P(x)}{\sqrt{x_c - x}} dx = A(x, x_c) * P(x). \end{aligned} \tag{14.3.3}$$

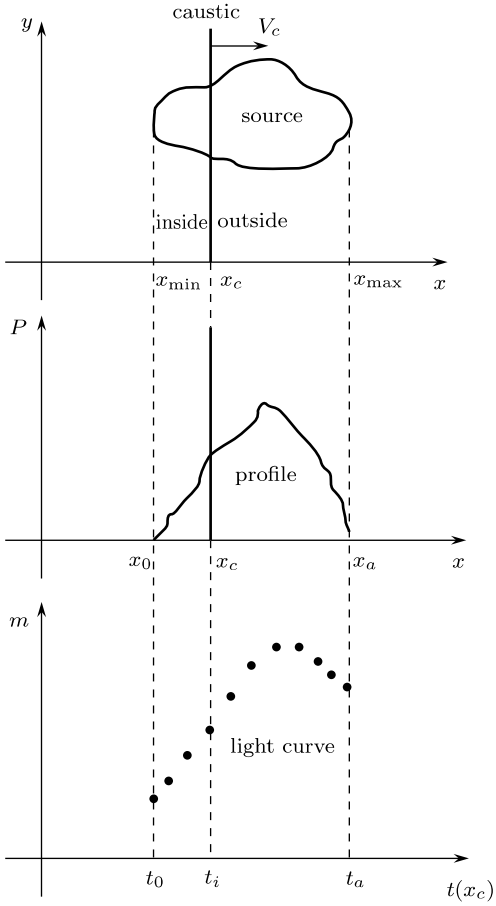


Fig. 14.9 A scheme of the caustic crossing.

The parameter of the caustic k can be estimated. The dependence of the caustic strength on the lens parameters was investigated in [20]. The uncertainty of this parameter leads to uncertainty in determination of the brightness of the source, i.e. we can determine the brightness distribution only to the constant factor. It means that we can restore only the shape of the brightness distribution of the source.

14.3.3 Implementation of the Tikhonov regularization approach

Introducing a regular grid x_j on the source plane, one can represent $P(x)$ as a continuous piecewise-linear function:

$$P(x) = P_{j-1} + \frac{P_j - P_{j-1}}{x_j - x_{j-1}}(x - x_{j-1}), \quad x_{j-1} \le x \le x_j. \tag{14.3.4}$$

It allows converting the integral convolution equation into the linear equations set [6]:

$$F_i = \sum_{j=1}^{m} K_{ij} P_j. \tag{14.3.5}$$

The dimension of the matrix K is $m \times n$ (where n is the number of observational points, m is the number of points of the regular grid on the source plane). The matrix K describes the connection between one-dimensional brightness distribution of the source and observational light curve. The direct inversion of this equation is impossible because the matrix K is singular. This is an ill-posed problem. The solution may be non-unique and unstable.

To solve this set of equations, the Tikhonov regularization approach described in Section 14.2.2 can be applied. The numerical representation of the smoothing function is:

$$M^{\alpha}[P] = \sum_{i=1}^{m} \frac{1}{\sigma_i^2}(F_i - \sum_{j=1}^{n} K_{ij} P_j)^2 + \alpha\Omega[P]. \tag{14.3.6}$$

Here the first term represents the squared discrepancy between the model and data, α is the regularization parameter, $\Omega[P]$ is a stabilizer function through which *a priori* information is introduced into the problem formulation. P^{α} that minimizes the Tikhonov function (14.3.6) will be taken as an approximate solution of (14.3.5). For co-ordination of the regularization parameter with the error of the input information we use the discrepancy principle – adoption of α satisfying:

$$\sum_{i=1}^{n}(\sum_{j=1}^{m} K_{ij} P_j^{\alpha} - F_i)^2 \approx \sigma_{tot}^2, \tag{14.3.7}$$

where $\sigma_{tot} = \sqrt{\sum \sigma_i^2}$ is the total error of observational data. Assuming that the source profile is a function that has square integrable second derivatives, one can choose the stabilizer in the following form:

$$\Omega[P] = \sum_{i=2}^{m-1} (P_{i+1} - 2P_i + P_{i-1})^2. \tag{14.3.8}$$

Further we describe an application of the above approach to the reconstruction of the Q2237+0305 quasar profile.

14.3.4 Numerical results of the Q2237 profile reconstruction

For the analysis of microlensing high magnification events with the aim of the quasar profile reconstruction, observations of Q2237+0305 gravitational lens system in V band by OGLE [21] and GLITP [22] collaborations were used. At the end of 1999 the brightness of the component A reached its maximum. The analysis of the light curve showed evidence for the caustic crossing between 1999 and 2000 years [23].

The search for the solution of the ill-posed problem was performed on the uniform grid with the number of points equal to 100 and step equal to 1 day. For minimization of the Tikhonov function (14.3.6), the conjugate gradient method was used. The results of the one-dimensional profile reconstruction are presented in Fig. 14.10. The best solution corresponds to the reduced $\overline{\chi}^2$ value of 1.213.

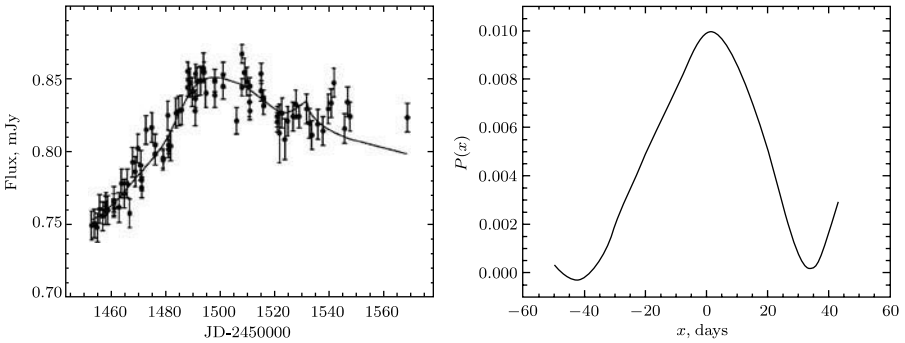


Fig. 14.10 Results of Q2237 profile reconstruction in V band, the A component microlensing event: a) the observational light curve and the light curve corresponding to the reconstructed 1D-profile (dash line); b) the reconstructed profile of the source in V band.

The reconstructed profile allows estimation of the source size. Moreover, it can be compared with theoretical models of an accretion disk to penetrate into the structure of the distant quasar. More detailed analysis of high-magnification microlensing events in Q2237+0305 with the presented approach can be found in [10].

14.4 Conclusions

In this chapter, two illustrative examples of the application of the Tikhonov regularization approach to practical problems of gravitational lensing research are presented. We have considered several modifications of the regularizing algorithm for processing images of gravitational lens systems and have found that our final approach allows successful reconstruction of images composed of point sources

and smooth background. We presented not only encouraging results of our final algorithm, but also several disappointing pictures we got during the evolution of our algorithm to show an importance of incorporating all available *a priori* information for solving ill-posed problems and getting the information you need. The second example shows that unique properties of gravitational lensing phenomenon and the Tikhonov regularization method allow getting information on the internal structure of distant quasars unresolved by a telescope.

Acknowledgements The work was supported by the Russian Foundation for Basic Research (grants 08-01-00160-a, 07-01-92103-NSFC, and 09-02-00244).

References

1. E. Agol and J. Krolik, Imaging a quasar accretion disk with microlensing, *The Astrophysical Journal*, **524**(1), 49–64.
2. K. Chang and S. Refsdal, Star disturbances in gravitational lens galaxies, *Astronomy and Astrophysics*, **132**(1), 168–178, 1984.
3. F. W. Dyson, A. S. Eddington and C. R. Davidson, A determination of the deflection of light by the sun's gravitational field from observations made at the total eclipse of May 29, 1919. *Mem. R. Astron. Soc.*, **62**, 291, 1920.
4. A. Einstein, Uber den Einflub der Schwerkraft auf die Ausbreitung des Lichtes (On the Influence of Gravity on the Propagation of Light), *Annalen der Physik*, **35**, 898, 1911.
5. A. V. Goncharskiy, A. S. Leonov and A. G. Yagola, On the solution of two-dimesional first kind Fredholm integral equations with kernel depending on the arguments difference, *Zhurnal vychislitelnoy matematiki i matematicheskoy fiziki*, **11**(5), 1971 (in Russian).
6. B. Grieger, R. Kayser and T. Schramm, The deconvolution of the quasar structure from microlensing light curves, *Astronomy and Astrophysics*, **252**(2), 508–512, 1991.
7. J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale Univ. Press, New Haven, 1923.
8. C. S. Kochanek, E. E. Falco, C. Impey, J. Lehar, B. McLeod and H.-W. Rix, CfA-Arizona Space Telescope LEns Survey, http://www.cfa.harvard.edu/glensdata/.
9. E. Koptelova, E. Shimanovskaya, B. Artamonov, M. Sazhin, A. Yagola, V. Bruevich, O. Burkhonov, Image reconstruction technique and optical monitoring of the QSO2237+0305 from Maidanak Observatory in 2002–2003, *Monthly Notices of the Royal Astronomical Society*, **356**(1), 323–330, 2005.
10. E. Koptelova, E. Shimanovskaya, B. Artamonov and A. Yagola, Analysis of the Q2237+0305 light-curve variability with regularization technique, *Monthly Notices of the Royal Astronomical Society*, **381**(4), 1655–1662, 2007.
11. A. S. Leonov, Numerical piecewise-uniform regularization for two-dimensional ill-posed problems, *Inverse Problems*, **15**, 1165–1176, 1999.
12. A. S. Leonov, A. N. Tikhonov and A. G. Yagola, *Nonlinear Ill-Posed Problems*, Chapman & Hall, London, 1998.
13. S. Mineshige and A. Yonehara, Gravitational microlens mapping of a quasar accretion disk, *Publ. of the Astronomical Society of Japan*, **51**, 497–504, 1999.
14. V. N. Shalyapin, L. J. Goicoechea, D. Alcalde, E. Mediavilla, J. A. Muñoz and R. Gil-Merino, The nature and size of the optical continuum source in QSO 2237+0305, *The Astrophysical Journal*, **579**(1), 127–135, 2002.

15. A. N. Tikhonov, et al., *Numerical Methods for the Solution of Ill-Posed-Problems*, Kluwer Academic Press, Dordrecht, 1995.
16. A. N. Tikhonov, On the solution of ill-posed problems and on the regularization method, *Doklady akademii nauk SSSR*, **151**(3), 1963 (in Russian).
17. V. G. Vakulik, R. E. Schild, V. N Dudinov, A. A. Minakov, S. N. Nuritdinov, V. S. Tsvetkova, A. P. Zheleznyak, V. V. Konichek, I. Ye. Sinelnikov, O. A. Burkhonov, B. P. Artamonov and V. V. Bruevich, Color effects associated with the 1999 microlensing brightness peaks in gravitationally lensed quasar Q2237+0305, *Astronomy and Astrophysics*, **420**, 447–457, 2004.
18. J. Wambsganss, H. J. Witt and P. Schneider, Gravitational microlensing — Powerful combination of ray-shooting and parametric representation of caustics, *Astronomy and Astrophysics*, **258**(2), 591–599, 1992.
19. J. Wambsganss, Gravitational lensing in astronomy, *Living Rev. Relativity*, **1**, http://www.livingreviews.org/lrr-1998-12, 1998.
20. H. J. Witt, R. Kayser and S. Refsdal, Microlensing predictions for the Einstein cross 2237+0305, *Astronomy and Astrophysics*, **268**(2), 501–510, 1993.
21. http://www.astrouw.edu.pl/∼ogle/ogle3/huchra.html.
22. http://www.iac.es/proyect/gravs lens/GLITP/.
23. J. S. B. Wyithe, R. L. Webster, E. L. Turner, MNRAS, Interpretation of the OGLE Q2237+0305 microlensing light curve (1997—1999), *Monthly Notices of the Royal Astronomical Society*, **318**(4), 1120–1130, 2000.

Index