Inverse and Ill-Posed Problems Series

Well-Posed, Ill-Posed, and Intermediate Problems with Applications

Yu. P. Petrov and V. S. Sizikov

VSP

Leiden • Boston

Well-Posed, Ill-Posed, and Intermediate Problems with Applications

Yu. P. Petrov and V.S. Sizikov

Well-Posed, Ill-Posed, and Intermediate Problems with Applications

Also available in the Inverse and Ill-Posed Problems Series:

Direct Methods of Solving Multidimensional
Inverse Hyperbolic Problems
S.I. Kabanikhin, A.D. Satybaev and M.A. Shishlenin
Characterisation of Bio-Particles from
Light Scattering - *V.P. Maltsev and K.A. Semyanov*
Carleman Estimates for Coefficient Inverse
Problems and Numerical Applications
M.V. Klibanov and A.A Timonov
Counterexamples in Optimal Control Theory
S.Ya. Serovaiskii
Inverse Problems of Mathematical Physics
M.M. Lavrentiev, A.V. Avdeev, M.M. Lavrentiev, Jr., and V.I. Priimenko
Ill-Posed Boundary-Value Problems - *S.E. Temirbolat*
Linear Sobolev Type Equations and Degenerate
Semigroups of Operators
G.A. Sviridyuk and V.E. Fedorov
Ill-Posed and Non-Classical Problems of
Mathematical Physics and Analysis
Editors: M.M. Lavrent'ev and S.I. Kabanikhin
Forward and Inverse Problems for Hyperbolic,
Elliptic and Mixed Type Equations
A.G. Megrabov
Nonclassical Linear Volterra Equations of the
First Kind - *A.S. Apartsyn*
Poorly Visible Media in X-ray Tomography
D.S. Anikonov, V.G. Nazarov, and I.V. Prokhorov
Dynamical Inverse Problems of Distributed
Systems - *V.I. Maksimov*
Theory of Linear Ill-Posed Problems and its
Applications
V.K. Ivanov, V.V. Vasin and V.P. Tanana
Ill-Posed Internal Boundary Value Problems for
the Biharmonic Equation - *M.A. Atakhodzhaev*
Investigation Methods for Inverse Problems
V.G. Romanov
Operator Theory. Nonclassical Problems
S.G. Pyatkov
Inverse Problems for Partial Differential
Equations - *Yu.Ya. Belov*
Method of Spectral Mappings in the Inverse
Problem Theory - *V. Yurko*
Theory of Linear Optimization - *I.I. Eremin*
Integral Geometry and Inverse Problems for
Kinetic Equations - *A.Kh. Amirov*
Computer Modelling in Tomography and
Ill-Posed Problems
M.M. Lavrent'ev, S.M. Zerkal and O.E. Trofimov
An Introduction to Identification Problems via
Functional Analysis - *A. Lorenzi*
Coefficient Inverse Problems for Parabolic Type
Equations and Their Application - *P.G. Danilaev*
Inverse Problems for Kinetic and Other
Evolution Equations - *Yu.E. Anikonov*
Inverse Problems of Wave Processes
A.S. Blagoveshchenskii
Uniqueness Problems for Degenerating
Equations and Nonclassical Problems
S.P. Shishatskii, A. Asanov and E.R. Atamanov
Uniqueness Questions in Reconstruction of
Multidimensional Tomography-Type
Projection Data
V.P. Golubyatnikov
Monte Carlo Method for Solving Inverse
Problems of Radiation Transfer - *V.S. Antyufeev*
Introduction to the Theory of Inverse Problems
A.L. Bukhgeim
Identification Problems of Wave Phenomena
- Theory and Numerics
S.I. Kabanikhin and A. Lorenzi
Inverse Problems of Electromagnetic Geophysical
Fields - *P.S. Martyshko*
Composite Type Equations and Inverse Problems
A.I. Kozhanov
Inverse Problems of Vibrational Spectroscopy
A.G. Yagola, I.V. Kochikov, G.M. Kuramshina and Yu.A. Pentin
Elements of the Theory of Inverse Problems
A.M. Denisov
Volterra Equations and Inverse Problems
A.L. Bughgeim
Small Parameter Method in Multidimensional
Inverse Problems - *A.S. Barashkov*
Regularization, Uniqueness and Existence of
Volterra Equations of the First Kind - *A. Asanov*
Methods for Solution of Nonlinear Operator
Equations - *V.P. Tanana*
Inverse and Ill-Posed Sources Problems
Yu.E. Anikonov, B.A. Bubnov and G.N. Erokhin
Methods for Solving Operator Equations
V.P. Tanana
Nonclassical and Inverse Problems for Pseudoparabolic Equations
A. Asanov and E.R. Atamanov
Formulas in Inverse and Ill-Posed Problems
Yu.E. Anikonov
Inverse Logarithmic Potential Problem
V.G. Cherednichenko
Multidimensional Inverse and Ill-Posed Problems
for Differential Equations - *Yu.E. Anikonov*
Ill-Posed Problems with A Priori Information
V.V. Vasin and A.L. Ageev
Integral Geometry of Tensor Fields
V.A. Sharafutdinov
Inverse Problems for Maxwell's Equations
V.G. Romanov and S.I. Kabanikhin

ISBN: 90 6764 432 3

A C.I.P. record for this book is available from the Library of Congress

The notion of well- and ill-posed problems, and also that of problems intermediate between well- and ill-posed ones, is described. Examples of such mathematical problems (systems of linear algebraic equations, systems of ordinary differential equations, partial differential equations, integral equations, and also examples of practical problems arising in control theory, image processing and tomography, are given. It is shown that classically equivalent transformations, when applied to well-posed equations, may yield ill-posed equations, and *visa versa.* The notion of transformations equivalent in the broadened sense is introduced. The stable Tikhonov regularization method and the solution-on-compact method are described. Solution results for some numerical examples are presented. The present book can be regarded both as a tutorial for advanced students and as a monograph.

For students, masters, post-graduate students, lecturers and scientific researches in pure and applied mathematics.

Recommended by EMA (educational-methodical association) on mathematical education as a training aid on the following branches of science: “Higher Mathematics”, “Applied Mathematics” and by EMA on optical and instrument-making education as a training aid on the following branches of science: “Control Theory”, “Image Processing”, “Foundations of Tomography”.

Approved by the Chair “Modeling of Electromechanical and Computer Systems”, Faculty of Applied Mathematics and Control Processes, Saint-Petersburg State University and by the Chair “Measuring Technologies and Computer Tomography”, Faculty of Precise Mechanics and Technologies, Saint Petersburg State University of Information Technologies, Mechanics and Optics.

Referees:

Doctor of Physical and Mathematical Sciences, Professor,
Head of Chair “Higher and Applied Mathematics”,
Penza State University
I. A. Boikov

Doctor of Technical Sciences, Professor,
Head of Chair “Electrical Engineering”,
Baltic State Technical University, Saint-Petersburg
P. A. Galaydin

Preface

This book is intended as a tutorial for students and post-graduate students, as a guide for lecturers, and as a monograph for scientific researches specializing in pure and applied mathematics and dealing with algebraic, differential, integral, and operator equations. More specifically, the book treats of one of key problems in applied mathematics, i.e., how to investigate into, and provide for, the solution stability in solving equations with due allowance for inaccuracies in set initial data, parameters and coefficients of a mathematical model for an object under study, instrumental function, initial conditions, etc., and also with allowance for miscalculations, including roundoff errors.

Until recently, all problems in mathematics, physics and engineering were believed to fall into two classes, — *well-posed problems*, wherein small inaccuracies in initial data give rise to small solution accuracies, and *ill-posed problems*, wherein small inaccuracies in initial data may result in arbitrarily large solution inaccuracies. For ill-posed problems to be solved, special (stable, regular) methods are required; some of these methods are considered in the present book. It is expedient, however, to additionally introduce a *third class of problems, intermediate ones between well- and ill-posed problems.* These are problems that change their property of being well- or ill-posed on equivalent transformations of governing equations, and also problems that display the property of being either well- or ill-posed depending on the type of the functional space used. From the standpoint of applications, problems that change their property of being well-or ill-posed on equivalent transformations should be paid special attention since their solution by classical methods may give rise to crude errors in calculations, resulting in accidents and crashes, and even in catastrophes of ships, control systems, etc.

The book consists of two complementary parts.

In the first part, general properties of all the three classes of mathematical, physical and engineering problems are considered, together with the approaches used to solve them.

In the second part, several stable methods for solving inverse ill-posed problems are described, illustrated with numerical examples.

The authors are grateful to Prof. I. A. Boikov and Prof. P. A. Galaydin for reading the manuscript and valuable comments. The authors would also like to thank I. A. Belov for his assistance in designing the russian variant of the book in the form of structure and for the computations that were needed to prepare Figures 5.2–5.5, 5.7, and 5.8.

The first part of the book was written by Doctor of Technical Sciences, Professor Yu. P. Petrov (Saint-Petersburg State University, e-mail: petrov1930@mail.ru) and the second part was written by Doctor of Technical Sciences, Professor V. S. Sizikov (Saint Petersburg State University of Information Technologies, Mechanics and Optics; e-mail: sizikov2000@mail.ru).

All questions and comments concerning the present monograph are welcome.

Contents

Part I
Three classes of problems in mathematics,
physics, and engineering

Part II

Stable methods for solving inverse problems

Part I

Three classes of problems in mathematics, physics, and engineering

Chapter 1.

Simplest ill-posed problems

1.1. STATEMENT OF THE PROBLEM. EXAMPLES

The necessity in studying ill-posed problems stems from one of the main problems in applied mathematics, gaining reliable computing results with due allowance for errors that inevitably occur in setting coefficients and parameters of a mathematical model used to perform computations.

Indeed, coefficients in a mathematical model, equations, or a set of equations used to perform computations are obtained, as a rule, from measurements; for this reason, they are accurate only to some limited accuracy. Moreover, parameters of an actual process or a technical object under simulation are never perfectly time-independent quantities; instead, they undergo uncontrollable changes, or display variations, whose exact value is usually unknown.

That is why we will differentiate between the nominal values of coefficients (to be denoted as $a_{i\,\mathrm{n}}$) and their real, "true" values (to be denoted as $a_{i\,\mathrm{t}}$). The nominal values are the values to be fed into the computer and used in all computations, whereas the real, "true" values $a_{i\,\mathrm{t}}$ are never known. What we can is just to claim that these unknown values are confined between certain limits and therefore obey some inequalities:

$$(1-\varepsilon)a_{i\,\mathrm{n}} \le a_{i\,\mathrm{t}} \le a_{i\,\mathrm{n}}(1+\varepsilon). \tag{1.1.1}$$

Here ε are numbers small compared to unity. Thus, the exact values of coefficients are never known, and only estimates for the coefficients are available.

We will call the products $\pm\varepsilon \cdot a_{i\,\mathrm{n}}$ the *nominal-coefficient variations*, or errors in nominal coefficients.

Since all computations are always conducted with nominal coefficients (while the real, "true" coefficients, as stated above, remain unknown), it is necessary to check how coefficient variations (and also variations of parameters, initial and boundary conditions, etc.) affect the computing accuracy.

There exist problems in which solution errors are the same order of magnitude as errors in setting coefficients (coefficient errors); this case is simplest to treat. However, there are problems wherein solution errors are greater than coefficient errors. Here, care must be taken in order to properly evaluate the computation accuracy, and efforts, made, to invent a computation procedure capable of minimizing the errors.

Finally, there are problems where even very small, practically unavoidable errors in setting coefficients, parameters, or initial and boundary conditions give rise to appreciable solution errors.

Such problems are called *ill-posed* ones (a more accurate definition will be given later); they are especially hard to solve. Nevertheless, such problems are often encountered in practice, and methods enabling their adequate solution need to be put to scrutiny.

Let us begin with several simple examples.

Consider a *system of two equations* for two variables x and y:

$$\begin{cases} a_{11}x + a_{12}y = b_1, \\ a_{21}x + a_{22}y = b_2. \end{cases} \tag{1.1.2}$$

In the plane (x, y), either of the equations defines a straight line; the solution of (1.1.2), i. e., the values of x and y making these equations identities are the coordinates of the intersection point of these straight lines. The solutions x and y can be expressed in terms of determinants by the well-known Cramer formulas:

$$x = \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix} \Bigg/ \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \qquad y = \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix} \Bigg/ \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}. \tag{1.1.3}$$

For instance, for the system

$$\begin{cases} x + 2y = 3, \\ 2x + y = 3 \end{cases} \tag{1.1.4}$$

we have $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} = -3$, $\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix} = \begin{vmatrix} 3 & 2 \\ 3 & 1 \end{vmatrix} = -3$, and $\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix} = \begin{vmatrix} 1 & 3 \\ 2 & 3 \end{vmatrix} = -3$. Hence, the solution here is $x = 1$ and $y = 1$.

Let us examine now how variation of the coefficient at x in the first equation of (1.1.4) affects the solution error.

The solution for the system

$$\begin{cases} (1+\varepsilon)x + 2y = 3, \\ \qquad 2x + y = 3 \end{cases} \tag{1.1.5}$$

is $x = 3/(3-\varepsilon)$, $y = (3-3\varepsilon)/(3-\varepsilon)$.

If $|\varepsilon| \le 0.01$, then

$$\begin{aligned} 0.996 \le x \le 1.0034, \\ 0.986 \le y \le 1.0034. \end{aligned} \tag{1.1.6}$$

Thus, we see that, here, a small error in the coefficient result in a small error in the solution. Nevertheless, there exist systems whose relative solution errors are much greater than relative coefficient errors. Such systems are often called "*ill-conditioned*" systems (an exact definition will be given below).

For example, for the system

$$\begin{cases} 1.1x + y = 1.1, \\ (1+\varepsilon)x + y = 1 \end{cases} \tag{1.1.7}$$

the determinant is

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 1.1 & 1 \\ 1+\varepsilon & 1 \end{vmatrix} = 0.1 - \varepsilon.$$

The solution of (1.1.7) is therefore $x = 1/(1-10\varepsilon)$, $y = -11\varepsilon/(1-10\varepsilon)$, and we obtain:

if $|\varepsilon| \le 0.001$, then $0.99 \le x \le 1.01$,

if $|\varepsilon| \le 0.01$, then $0.909 \le x \le 1.11$,

if $|\varepsilon| \le 0.1$, then $0.5 \le x \le \infty$.

Here, the solution error is greater than the coefficient error. Moreover, the solution error rapidly grows in value with increasing ε and may be arbitrarily large if $|\varepsilon| \le 0.1$.

With (1.1.2), everything is quite clear: solution errors are large if the determinant

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \tag{1.1.8}$$

is small.

A small value of (1.1.8) points to the fact that the straight lines intersect at a small angle, and even a small coefficient variation, changing this angle, gives rise to substantial changes in the coordinates of the intersection point.

Consider now the limiting case where, at nominal values of coefficients, the determinant (1.1.8) becomes zero. Here, even small coefficient variations may give rise to large, or even dramatically large, changes in the solution.

Consider the system

$$\begin{cases} x + y = b_1, \\ x + y = 1. \end{cases} \tag{1.1.9}$$

For this system, we have

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 0.$$

If $b_1 \neq 1$, then the system has no solutions (indeed, the straight lines here are parallel). If $b_1 = 1$, then there are many solutions: in this case, the straight lines are coincident, and any pair of numbers $x = 1 - y$ represents a solution.

Consider now the system

$$\begin{cases} (1 + \varepsilon)x + y = b_1, \\ x + y = 1, \end{cases} \tag{1.1.10}$$

i. e., the question how variation of a_{11} affects the solution of (1.1.9). Subtracting the second equation from the first one, we obtain

$$x = \frac{b_1 - 1}{\varepsilon}, \qquad y = 1 - \frac{b_1 - 1}{\varepsilon}. \tag{1.1.11}$$

If b_1 $\neq$1, then for any (even infinitesimal) ε the solution does exist, although it entirely depends on the unknown error ε. From the practical point of view, this solution, although existing for any $\varepsilon \neq 0$, is meaningless. If $b_1 = 1$, then, as previously, we subtract the second equation in (1.1.10) from the first one and obtain: $\varepsilon x = 0$; it follows from here that the pair

$$x = 0, \qquad y = 1 \tag{1.1.12}$$

also represents a solution. This solution is valid for all values of ε (including arbitrarily small ε) except for the single value $\varepsilon = 0$, to which an infinite set of solutions corresponds (for $\varepsilon = 0$, any pair $x = 1 - y$ is a solution). Whether the solution is unique or there are many solutions depends on the unknown value of ε.

The above consideration shows the solutions of (1.1.2) with vanishing determinant (1.1.8) to be of no practical significance.

For degenerate systems, Moore and Penrose invented the pseudo-inverse matrix method (Sizikov, 2001, p. 189); with this method, a unique (but generally unstable) solution, called the normal solution, can be obtained.

1.2. DEFINITIONS

With the above simplest example, we are going now to formulate a definition of well- and ill-posed problems.

There exists a definition given in one of most authoritative manual on ill-posed problems, A. N. Tikhonov and V. Ya. Arsenin, “Solution of Ill-Posed Problems”, Wiley, NY, 1977. (The first, second, and third editions of this book in Russian were issued in 1974, in 1979, and in 1986, respectively.) According to this book, a problem is *ill-posed* if at least one of the following three conditions is violated:

1. The solution exists.
2. The solution is unique.
3. The solution is stable, namely, arbitrarily small variations of coefficients, parameters, initial or boundary conditions give rise to arbitrarily small solution changes.

It follows from definition (1.1.1) of coefficient or parameter variations that, provided that the nominal value of the coefficient, i. e., the value of $a_{i\,\mathrm{n}}$, is zero, then its variation is also zero; in other words, a zero quantity is not to be varied, it always remains zero. This means that, when analyzing the question whether a problem of interest belongs or not to the category of well-posed problems, we will consider only relative, and not absolute, coefficient variations.

We exclude from the present consideration the cases in which the nominal value of at least one coefficient is zero while the varied value of this coefficient, even arbitrarily small, has become nonzero. This restriction is necessary; otherwise, very broad range of problems will fall into the category of ill-posed problems. For instance, the equation $a_1x + a_0 = 0$ can be considered as the equation $0 \cdot x^2 + a_1x + a_0 = 0$, i. e., as an equation of type $a_2x^2 + a_1x + a_0 = 0$ with $a_2 = 0$. If, on variation, the coefficient a_2 becomes nonzero, equal to ε, then already for arbitrarily small values of ε two roots of this equation, generally not close to each other, will emerge, and, in this situation, we will be forced to class the problem on finding roots of the polynomial $a_1x + a_0 = 0$, and also the problem on finding roots of an arbitrary polynomial

$$a_nx^n + a_{n-1}x^{n-1} + \cdots + a_0 = 0$$

to ill-posed problems, which policy seems to be unpractical.

Having abandoned the possibility of varying zero quantities, we expelled from the category of ill-posed problems a multitude of the so-called "singular-disturbed" systems of differential equations, i. e., systems whose parameters at higher-order derivatives are small, like in equations of type

$$\varepsilon\ddot{x} + a_1\dot{x} + a_0 = 0$$

and similar equations. These problems, although interesting, are not ill-posed ones. Even if ε is arbitrarily small, this quantity (of course, conventionally) is "by an infinitely large factor" greater than the exact zero. That is why the transition from exact zero to $\varepsilon \neq 0$ should not be regarded as variation of the zero-valued coefficient, or its small variation.

Below, it will become clear that, even with zero variations abandoned, the range of ill-posed problems all the same remains very broad.

1.3. EXAMPLES AND APPROACHES TO SOLVING ILL-POSED PROBLEMS

A very broad set of minimax, or extremum, problems falls into the category of ill-posed problems.

Example. Find the minimum length of a fence enclosing (without a gap) a parcel of land of area s.

At first glance, the problem is rather simple to solve: it is well known that a minimum fence length is achieved if the parcel of land is shaped as a circle having the perimeter $p = 2\pi R$ and area $s = \pi R^2$ (here, R is the radius of the circle). Eliminating R, we obtain: $p_{\min} = \sqrt{4\pi}\sqrt{s}$.

Yet, exact measurement of s is impossible. If the true area is greater than the nominal one, s_n, even by a small quantity Δs, then the fence of length $p_{\min} = \sqrt{4\pi}\sqrt{s_n}$ will remain open-ended, and the problem, unsolved. We see that the problem under consideration, as well as any other problem about minimizing elements, is an ill-posed one. There being an arbitrarily small error in a condition, the solution vanishes.

This simple example can be conveniently used to explain the general approach to solving ill-posed problems: an ill-posed problem, of no practical meaning, needs to be replaced by a well-posed problem close to it.

Here, the well-posed problem is as follows: let the area s_n be known to some accuracy Δs; then, which additional length Δp is to be added to the minimum length $p_{\min}$ so that to make the enclosing problem always solvable? From the condition $p_{\min} + \Delta p = \sqrt{4\pi}\sqrt{s_n + \Delta s}$, we readily obtain that $\Delta p = \sqrt{4\pi}\,(\sqrt{s_n + \Delta s} - \sqrt{s_n})$.

Since the error Δs is variable, equal to $\Delta_i s$, then, for any $\Delta_i s$, the above formula gives a sequence of quite meaningful well-posed problems. In the limit $\Delta s \to 0$, we obtain the limiting solution of no practical significance for the initial ill-posed problem.

This simple example illustrates the simplest case of the so-called regularization: we replace the initial ill-posed problem of no practical significance with a sequence of well-posed problems involving some parameter whose limiting value generates the initial ill-posed problem. In the above simplest case, the regularization parameter is Δs. Below, other regularization methods for this parameter will be considered.

As in the above example, the property of being ill-posed arises in any extremum problem or, more exactly, in any problem on determining an element on which some function, functional, etc. attains a minimum. Infinitesimal changes in the set condition will result in that the minimal element found (in the above example, the fence length $p_{\min} = \sqrt{4\pi}\,\sqrt{s_n}$) will become insufficient to satisfy it. The problem will lose sense, which proves it to be an ill-posed one.

Yet, since for all problems about an element providing an extremum the reasons for the property of being ill-posed were rather easy to understand, from time immemorial people used to solve such problems so that to have some reserve in the solution, for instance, not specifying that, say, in the problem about the enclosing fence one has to take, as the target value, not the nominal area s_n, but the maximum possible value of this area, i.e., $s = s_n + \Delta s$, where Δs is the maximum possible error in measuring or setting the value of s. The approach discussed being rather simple, at all times extremum problems were successfully treated without isolating them in a special class of ill-posed problems. The class of ill-posed problems was first identified by Jacques Hadamard, famous French mathematician (1865–1963), in 1902 Hadamard gave an example of a heat conduction problem in which an arbitrarily small error in setting boundary condition resulted in a large solution error. Yet, the Hadamard example used rather a complex mathematical model involving partial differential equations. Examples treated later in the well-known course of mathematics (Tikhonov and Arsenin, 1977) and in (Lavrent'ev, 1981; Ivanov, Vasin, and Tanana, 2002; Tikhonov, Leonov, and Yagola, 1997) were also rather complex.

That it why some of students, engineers and scientists — all those who deal with calculations — often share the opinion that ill-posed problems is a matter that belongs to the field of "very evolved" mathematics, something very difficult to understand and rarely met.

They are certainly wrong. Ill-posed problems are encountered very frequently, and care must be taken to adequately take their properties, and difficulties they cause, into account. A historical review of ill-posed problems can be found in (Petrov, 2001).

As mentioned above, numerous extremum problems in which an element providing an extremum is to be found, possess the property of being ill-posed.

Another example of ill-posed problems is given by problems on finding roots of polynomials in those cases where these polynomials have multiple roots, but only real solutions are physically meaningful.

Consider a simplest *second-degree polynomial*:

$$x^2 + 2x + 1.$$

Every scholar knows that this polynomial has the double root $x_1 = x_2 = -1$ yielded by the following elementary formula:

$$x_{1,2} = -b/2 \pm \sqrt{b^2/4 - c} = -1 \pm \sqrt{1 - 1} = -1.$$

Yet, if the coefficient at the last term is not unity exactly, i. e., if this coefficient equals $1+\varepsilon$, which always may be the case because all coefficients are known to some limited accuracy, then the real solution for arbitrarily small $\varepsilon > 0$ vanishes at once, and we have

$$x_{1,2} = -1 \pm \sqrt{1 - (1 + \varepsilon)} = -1 \pm \sqrt{-\varepsilon}.$$

If our concern is only with real solutions, then already for arbitrarily small $\varepsilon > 0$ the solution vanishes. The problem on finding real-valued multiple roots is therefore an ill-posed problem.

Many *boundary-value problems for ordinary differential equations* also possess the property of being ill-posed. Consider the equation

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} - x = 0 \tag{1.3.1}$$

with the boundary conditions $x(t = 0) = 0$, $x(t = a) = b$. The general solution of (1.3.1) is

$$x = c_1 \sin t + c_2 \cos t,$$

where c_1 and c_2 are constants of integration (indeed, it is an easy matter to check that each of the functions $c_1 \sin t$ and $c_2 \cos t$ satisfy equation (1.3.1); a linear combination of these functions therefore also satisfies this equation). From the first boundary condition it follows that $c_2 = 0$. The second boundary condition, $x(t = a) = b$, yields $c_1 = b/\sin a$. Thus, the set boundary conditions are satisfied with the solution

$$x_1(t) = b \sin t / \sin a.$$

Yet, if the boundary condition is set at the point $t = a + \varepsilon$ instead of the point $t = a$ (which is quite possible because small errors in boundary conditions readily occur), then the solution becomes

$$x_2(t) = b \sin t / \sin{(a + \varepsilon)}.$$

If, for instance, $a = \pi - \varepsilon$, then the absolute difference between $x_1(t)$ and $x_2(t)$ may be arbitrarily large even for arbitrarily small ε.

It should be noted that, unlike the case of the boundary-value problem, the Cauchy problem for one differential equation of arbitrary order presents a well-posed problem. Yet, very recently new interesting phenomena concerning the alteration between the property of being well-posed and the property of being ill-posed have been discovered in systems of differential equations; this matter will be discussed in more detail in Chapter 2.

Approaches to solving ill-posed problem will be illustrated below with the example of the problem on determining the position of a ship from bearing data. If the shore is well within sight, then the navigator measures the angles between the North direction and the directions towards two lighthouses, A and B, which are available on the map. Then, the navigator plots these

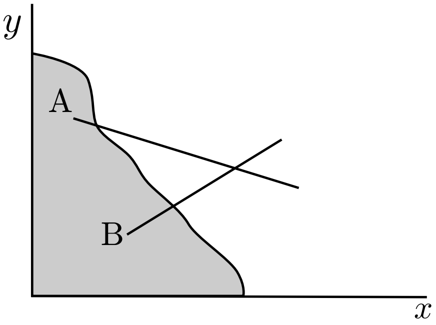

Figure 1.1

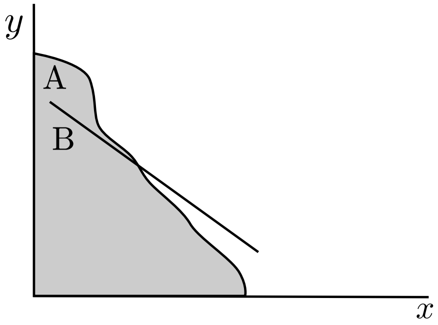

Figure 1.2

two directions (bearings) on the map with a pencil. The position of the ship then can be identified as the intersection point of the two bearings (Figure 1.1). Thus, here the navigator graphically solves the system of two linear equations similar to equations (1.1.2).

Of course, the angles are measured accurate to some precision, and errors in plotting lines on the map are either unavoidable; hence, the location problem can be solved only approximately, although the accuracy suits the navigator.

If two lighthouses are situated close together, then the directions towards them are almost identical, and the lines drawn with the pencil will intersect each other at an acute angle; they may even merge together within the precision typical of some particular maps (see Figure 1.2).

Consider the latter case, i. e., the case where the navigator has to solve an ill-posed problem, for instance, the problem on solving the system of equations

$$\begin{cases} x + y = 1, \\ x + y = 1, \end{cases} \tag{1.3.2}$$

whose determinant is zero; in fact, this problem therefore reduces to one equation,

$$x + y = 1.$$

Of course, the exact position of the ship cannot be identified based on the two coincident straight lines (defined by the equations $x + y = 1$ and $x + y = 1$), it may be any on the line $x + y = 1$.

Yet, this way or another, the ill-posed problem can be regularized.

Consider one (often recommended) regularization method: not any solution of (1.3.2) is to be found, but the solution that has the lowest norm.

(This method, yielding the solution with the minimum norm, called the *normal solution*, is considered in more detail by Sizikov (2001, pp. 179, 189). One of possible norms here is the square root of sum of squares of x and y. We are to solve the following problem: find x and y such that to minimize the square of norm

$$F = x^2 + y^2$$

providing that x and y are interrelated by equations (1.3.2), or (which is the same) by the equation $x + y = 1$.

Taking into account that $x = 1 - y$, we bring F to the form $F = (1 - y)^2 + y^2$ and, examining the derivative $\partial F/\partial y$, obtain that F is minimal at the point $x = 0.5$, $y = 0.5$.

Thus, with $F = x^2 + y^2$ we have passed from the ill-posed problem to a well-posed one (it is an easy matter to show that, with small coefficient variations in (1.3.2), the solution $x = 0.5$, $y = 0.5$ will change insignificantly).

Yet, for the practical problem of interest, that on determining the ship position, the solution $x = 0.5$, $y = 0.5$ is formal. This point has nothing in common with the actual position of the ship. If a method at hand fails to provide the solution, a more appropriate method needs to be invented.

One of such methods is invoking additional information. Let there exist a possibility to evaluate the distances to the coinciding lighthouses; these data provide for additional information. Then, we plot each of the distances on the map with a pair of compasses and obtain, on the straight line $x + y = 1$, the true position of the ship.

Yet another method employing additional information consists in determining, if possible, the bearing to a third lighthouse located far apart from the first two lighthouses. The point where the new bearing intersects the previous straight line $x + y = 1$ gives the sought position of the ship (Figure 1.1).

Thus, in the cases where the location problem proves to be an ill-posed one, the use of additional information offers a good regularization method.

Regularization should be regarded an approach to solving ill-posed problems, when one replaces the ill-posed problem of interest with a close well-posed problem, by way of using additional information, introducing a norm and searching for the solution for which the norm is minimal, introducing a sequence of well-posed problems approaching, in the limit, the ill-posed problem, etc.

There is a multitude of approaches to solving ill-posed problems; further example will be given below.

Furthermore, one can use general theorems. For instance, it long has been known that the problem on finding roots of a polynomial of arbitrary degree

$$a_n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0$$

presents a well-posed problem in the complex field. Below, we will see that, for instance, such common a problem as the solution of systems of ordinary differential equations is generally an ill-posed problem.

1.4. ILL-POSED PROBLEMS OF SYNTHESIS FOR OPTIMUM CONTROL SYSTEMS

One of most remarkable encounter with ill-posed problems had been in the 1970-ths, when synthesis methods for optimum control systems were under study. Shortly before (see, e. g., Larin, Naumenko, and Suntsev, 1971; Letov, 1960, 1969), a theory was developed and algorithms were proposed permitting the synthesis of optimum control systems; these algorithms made it possible to substantially improve performance characteristics of many industrial objects and transport facilities acted upon by incompletely known disturbing forces.

Under consideration were control objects governed by the following differential equations with constant coefficients:

$$A(D)x = B(D)u + \varphi(t). \tag{1.4.1}$$

In (1.4.1), $A(D) = a_nD^n + a_{n-1}D^{n-1} + \ldots + a_0$ and $B(D) = b_mD^m + b_{m-1}D^{m-1} + \ldots + b_0$ are some polynomials of differentiation operator $D = \mathrm{d}/\,\mathrm{d}t$; x is the quantity under control (or, more exactly, deviation of this quantity from its optimum value); u is the control action; and $\varphi(t)$ is the disturbing action, an incompletely known stationary random function of time, about which, as a rule, nothing is known except for its power spectral density (traditionally called, for short, the spectrum). The spectrum of a random function can be easily deduced from observations over the disturbing actions; this spectrum is then to be subsequently fitted with some even fractional rational function

$$S = \frac{a_p\omega^{2p} + a_{p-1}\omega^{2p-2} + \ldots + a_0}{b_q\omega^{2q} + b_{q-1}\omega^{2q-2} + \ldots + b_0}, \tag{1.4.2}$$

where ω is a variable having the units of frequency, i. e., 1/s.

The control quality criterion was the functional

$$J = \lim_{T\to\infty} \frac{1}{T} \int_0^T (\lambda_0 x^2 + u^2)\,\mathrm{d}t = \lambda_0 \langle x^2 \rangle + \langle u^2 \rangle, \tag{1.4.3}$$

in which the term $\langle x^2 \rangle$ reflected the control accuracy, and the term $\langle u^2 \rangle$ indirectly took into account inevitable restrictions on the control action magnitude. The nonnegative number λ_0 was a Lagrange multiplier; it depended on control restrictions and was calculated by rules given by Larin, Naumenko, and Suntsev (1971); Letov (1960, 1969), Petrov (1977).

As early as the 1950ths, it was proved that there exists an easily realizable linear control

$$W_1(D)x = W_2(D)u, \tag{1.4.4}$$

(where $W_1(D)$ and $W_2(D)$ are some polynomials of differentiation operator $D = \mathrm{d}/\,\mathrm{d}t$) providing for stability of the closed system (i.e., ensuring the solution stability for system (1.4.1)–(1.4.4)) and minimizing the quality criterion (1.4.3). In this manner, the problem on optimization was reduced to searching for the optimum operators $W_1(D)$ and $W_2(D)$ and their subsequent realization with the help of technical devices called controllers.

Example. For a Kazbek tanker making its sea route under rudder control (16 thousand ton displacement, 14 knot velocity), the mathematical model is

$$(T_1^2 D^2 + T_2 D)x = u + \varphi(t), \tag{1.4.5}$$

(i. e., $A(D) = T_1^2 D^2 + T_2 D$; $B(D) = 1$), where T_1 and T_2 are some time constants, expressed in seconds (for Kazbek tankers, $T_1 = 26.3$ s and $T_2 = 17.3$ s); x is the deviation from the preset course of the tanker, expressed in degrees; u is the control action, i.e., the deviation of the rudder from the diametric plane, expressed in degrees; and $\varphi(t)$ is the disturbing action influencing the vessel, the moment of forces due to wind and sea roughness, measured in rudder degrees, that produces an equivalent moment of force.

For a Kazbek tanker, the loss in velocity was minimized by the following easily realizable controller of type (1.4.4):

$$u = -(2.5 + 43.6D)x \tag{1.4.6}$$

(i. e., $W_1(D) = -(2.5 + 43.6D)$ and $W_2(D) = 1$).

Controller (1.4.6) is made up by an amplifier and a differentiator, connected in parallel; this controller was therefore easy to realize.

For other control objects, controllers of type (1.4.4) could also be constructed; with these controllers, performance characteristics of these objects could be substantially improved compared to objects equipped with previously used controllers (before the advent of the optimum control theory).

The coefficients in the optimum polynomials $W_1(D)$ and $W_2(D)$ of controllers (1.4.4) were calculated by rather evolved algorithms. By way of illustration, we give here a description of a simplest algorithm suiting the case of $B(D) = 1$. This algorithm comprises the following steps (Petrov, 1977):

1. Following the substitution of $j\omega$ with the variable s in spectrum (1.4.2) ($j\omega = s$), the spectrum can be factorized:

$$S_\varphi(s) = S_1(s)S_1(-s), \tag{1.4.7}$$

i. e., represented as the product of two symmetric multipliers, $S_1(s)$ and $S_2(s)$; the first multiplier includes all multipliers of type $s - a_i$, where a_i are roots whose real parts are negative, and the second one, all multipliers of the same form but with positive real parts of the roots.

Since spectrum (1.4.2) is an even function, its factorization (i. e., decomposition into symmetric multipliers) is always possible, but requires finding roots of the numerator and denominator of (1.4.2), which, of course, implies using computers.

2. Second, the even polynomial is to be factorized:

$$A(s)A(-s) + \lambda_0 = G(s)G(-s). \tag{1.4.8}$$

3. Third, decomposition into fractions (separation) is to be performed:

$$\frac{A(-s)}{G(-s)}\,S_1(s) = M_0 + M_+ + M_-, \tag{1.4.9}$$

where M_0 is an integer polynomial, and M_+ and M_- are proper fractions with poles that lie respectively in the left and right half-planes of the complex variable s. Of course, to perform this operation, a computer is necessary and, except for the case of simplest polynomials $A(s)$, the result cannot be expressed explicitly.

4. The fourth step is easy to perform: using M_0 and M_+, we construct the following equality for the polynomials $W_1(s)$ and $W_2(s)$ of the optimum controller (1.4.4):

$$\frac{W_1(s)}{W_2(s)} = A(s) - \frac{G(s)S_1(s)}{M_0 + M_+}. \tag{1.4.10}$$

Despite the complexity of the computational algorithm, easy realization of optimal controllers and, which is more important, considerable improvement in performance characteristics of diverse objects gained with the optimal control, were the factors that have made have the optimum control theory widely used; first monographs (Letov, 1969; Zubov, 1974) devoted to this theory and its various applications were issued, and soon afterwards this theory was introduced into lecture courses delivered at technical institutes and universities, including courses on control theory, automatic control theory, and theory of automatic systems.

Afterwards, unexpected events occurred: several wrecks (crashes) happened, the reason for which was that some optimal systems constructed in full compliance with the recommendations of the recently devised optimum control theory proved to be unstable against small deviations of control-object parameters or controller parameters from their nominal values.

In the 1960ths, such wrecks pointed to imperfection of the optimum control theory and its practical application, all the more so that, for long, the reason for these wrecks remained obscure. Indeed, at that time, it was quite clear that optimal systems had little sensitivity to parameter variations (because the main, linear component of the functional increment vanishes at the extremum curve); for this reason, these systems were believed to possess the property of parametric stability (it should be noted here that "parametrically stable" systems are systems that retain stability under parameter variations). That is why the lack of parametric stability looked a very strange observation, all the more so that not all systems displayed that property. Some optimal systems were parametrically stable, some not, and, for a long time, all attempts to find a regularity here were a failure.

Initially, it was the algorithm that was thrown doubt upon: parametric instability was hypothesized to result from some deficiencies of the control design algorithm; that is why active search for other algorithms capable of providing parametric stability was launched. In 1965–77, several monographs (Letov, 1969; Zubov, 1974) were issued where new synthesis algorithms were proposed; yet, every time these new algorithms were found to exhibit the same drawbacks concerning parametric stability.

The state of things remained unaltered until the year 1973, when a breakthrough was made. Early in this year, P. V. Nadezhdin (1973) found that yet another synthesis algorithm, long before that reported by three Ukrainian mathematicians (Larin, Naumenko, and Suntsev, 1971), fails to provide parametric stability; P. V. Nadezhdin expressed the opinion that this failure was a consequence of some algorithm deficiencies, and also the hope that these deficiencies could be eliminated in the future.

Later in 1973, it was shown (Petrov, 1973) that there exists no algorithm capable of guaranteeing parametric stability since the quality control minimum for many control objects lies on the verge of stability, and none of algorithms can alter the situation (Petrov, 1973). This statement is easy to prove since for several spectra of disturbing actions for the first time mathematical models of optimal controllers were obtained in finite, analytical form, and not in algorithmic form. This finding at once has made everything clear.

In particular, for the Rakhmanin–Firsov spectrum

$$S_\varphi = \langle \varphi^2 \rangle \, \frac{\alpha^2 + \beta^2}{(\alpha^2 + \beta^2 + \omega^2)^2 - 4\beta^2\omega^2} \tag{1.4.11}$$

widely used to optimize seasgoing crafts and marine control systems (where α and β are sea roughness controlled parameters) and control objects of type (1.4.1) with $B(D) = 1$, Petrov (1973) obtained the following optimum-controller formula:

$$u_{\text{opt}} = [A(D) - G(D)/(a + bD)]x. \tag{1.4.12}$$

The polynomial $G(D)$ in (1.4.12) is defined by formula (1.4.8); this polynomial is therefore of the same degree n as the polynomial $A(D)$, the higher term in $G(D)$ is $a_n D^n$, and the coefficients a and b can be expressed in terms of the real and imaginary parts of the complex number

$$A(\alpha - j\beta)/G(\alpha - j\beta) = K_1 + jK_2, \tag{1.4.13}$$

by the formulas $a = K_1 + \alpha K_2/\beta$ and $b = K_2/\beta$.

If control-object parameters equal the calculated, nominal parameters exactly, then, having closed the control object (1.4.1) with the optimal controller (1.4.13), we obtain the following mathematical model for the closed system:

$$G(D)x = (a + bD)\varphi(t). \tag{1.4.14}$$

It follows from formula (1.4.14) that, here, the closed system is stable since the polynomial $G(D)$ is a Hurwitz one. Yet, if control-object coefficients differ from their nominal values even by arbitrarily small numbers ε_i, i. e., if the mathematical model is

$$[(a_n + \varepsilon_n)D^n + (a_{n-1} + \varepsilon_{n-1})D^{n-1} + \dots a_0 + \varepsilon_0]x = u + \varphi, \tag{1.4.15}$$

then the equation for the closed system becomes

$$[\varepsilon_n D^n + \varepsilon_{n-1} D^{n-1} + \dots \varepsilon_0 + G(D)/(a + bD)]x = \varphi(t), \tag{1.4.16}$$

on multiplication by $(a + bD)$, considering the fact that the higher term in $G(D)$ is $a_n D^n$, instead of (1.4.16) we obtain:

$$[\varepsilon_n bD^{n+1} + \varepsilon_n aD^n + a_n D^n + \ldots]x = (a + bD)\varphi(t). \tag{1.4.17}$$

Here, the ellipsis stands for the terms whose power exponents are lower than n. From (1.4.17), it immediately becomes clear that even for arbitrarily small ε_n the closed system may become unstable, since the necessary condition for stability will be violated whenever the product $\varepsilon_n b$ and the terms in $G(D)$ are of opposite sign. Since ε_n may be of either sign, system (1.4.17) is parametrically unstable.

This means that for control objects of type (1.4.1) with $B(D) = 1$ acted upon by the disturbing action with spectrum (1.4.11) the design of a controller minimizing the quality criterion (1.4.3) presents an ill-posed problem: at nominal values of parameters, the closed system is stable, and criterion (1.4.3) attains its minimum value, but stability vanishes already with an arbitrarily small variation of a_n; for the unstable system, functional (1.4.3) has no finite value at all (i. e., it “tends to infinity”).

It is the property of being ill-posed, demonstrated by the problem of synthesizing optimal controllers for several spectra of disturbing actions and having remained unnoticed in due time, that resulted in several unforeseen wrecks and, which is more important, undermined the trust to the optimum control theory.

After the problem was proved to be ill-posed and the reasons for that were disclosed in monograph by Petrov (1973), it was found an easy matter to synthesize parametrically stable controllers and system close to optimal ones. The first synthesis method providing for parametric stability was proposed as early as 1973 in (Petrov, 1973). This method proved not to be a very good one, and the studies were resumed. In 1977, a better method was invented (Petrov, 1977), and a simple criterion was formulated enabling easy discrimination between the cases in which the synthesis of an optimal controller for control objects of type (1.4.1) presented a well- or ill-posed problem: if

$$p \geq m + q - 1, \tag{1.4.18}$$

then the synthesis is a well-posed problem, otherwise, the problem is ill-posed. In (1.4.18), m is the degree of the polynomial $B(D)$ in the mathematical model (1.4.1), and p and q are half-degrees of the numerator and denominator in the analytical approximation (1.4.2) of the spectrum $S(\omega)$ of the disturbing action $\varphi(t)$. Later, it became common practice to call formula (1.4.18) “the Petrov criterion”, or ”the Petrov inequality”.

If the Petrov criterion is not fulfilled, then the optimum controller synthesis presents an ill-posed, although easy to regularize, problem. The following example illustrates the regularization procedure.

Consider the simple control object

$$4Dx = (D+1)u + \varphi(t) \tag{1.4.19}$$

with the quality criterion

$$J = 9\langle x^2 \rangle + \langle u^2 \rangle \tag{1.4.20}$$

acted upon by a disturbing force with the spectrum

$$S_\varphi(\omega) = \frac{2}{\pi}\,\frac{1}{1+\omega^2}. \tag{1.4.21}$$

Next, we pose the following problem: it is required to find a mathematical model for the controller that minimizes criterion (1.4.20). Using the above-described synthesis algorithm (this algorithm is described in more detail in (Petrov, 1977, 1987), we readily obtain the sought mathematical model for the optimal controller:

$$(3D-5)u = 12(D+4)x. \tag{1.4.22}$$

On closing control object (1.4.19) with controller (1.4.22), we readily obtain that the closed system is stable; then, calculation of the quality criterion (1.4.20) in the system closed with the optimum controller yields the minimum possible value of (1.4.20): $J_{\min} = 0.436$.

Yet, the problem in question is an ill-posed one. Indeed, for the control object (1.4.19) and spectrum (1.4.21), we have: $m = 1$, $p = 0$, and $q = 1$; the latter means that Petrov inequality (1.4.18) is not fulfilled, and the system is parametrically unstable.

To check this, it suffices to take into account inevitable small deviations from calculated values in the coefficients of the real control object. If we close with the controller (1.4.22) the control object (1.4.19), in which possible coefficient variations had been taken into account and the control object therefore acquired the form

$$4(1+\varepsilon)Dx = (D+1)u + \varphi(t) \tag{1.4.23}$$

(for simplicity, here we vary just one coefficient), then the characteristic polynomial of the closed system will acquire the form

$$\Delta = -3\varepsilon D^2 + (20+5\varepsilon)D + 12. \tag{1.4.24}$$

With $\varepsilon = 0$, the only negative root of this polynomial is $D_1 = -0.6$, but already for an arbitrarily small $\varepsilon > 0$ the polynomial (1.4.24) has two roots, of which the second has a positive real part, and the closed system loses stability even for arbitrarily small $\varepsilon > 0$. If $\varepsilon \leq 0$, then stability is retained.

Had we had closed with the controller (1.4.22) the real control object (1.4.23), then, since the sign of ε is unpredictable, then the closed system may be either unstable (for $\varepsilon > 0$) or stable (for $\varepsilon \leq 0$). If $\varepsilon > 0$, then the system under design is unstable and is to be rejected. Even worse the situation will be if $\varepsilon \leq 0$. In this case, the constructed system will be able to successfully pass control tests, and it will be operated for rather a long time until the inevitable small drift of system parameters will give rise to system instability, which may result in an emergency situation.

Of course, for such simple a control object as (1.4.19), the possibility of a situation in which the system may lose stability is still easy to analyze. Yet, similar phenomena were observed with much more complicated systems. It is therefore not surprising that the application of the optimum control theory to practical cases was postponed for many years.

Meanwhile, the theory of ill-posed problems enables solution to the problem. It is important to find a simple criterion making it possible to discriminate between well- and ill-posed problems. As soon as the necessary criterion (1.4.18) was found and reported by Petrov (1977), all other things became in the following way: to provide for parametric stability, the analytical approximation of the spectrum, used to synthesize the controller, must be chosen such that the Petrov inequality be fulfilled.

In the particular case of the control object (1.4.19), it suffices to choose the following approximation for the disturbing-action spectrum:

$$S_\varphi(\omega) = \frac{2}{\pi}\,\frac{1 + k^2\omega^2}{1 + \omega^2}. \qquad (1.4.25)$$

For this approximation, we have $p = 1$, $q = 1$, and inequality (1.4.18) is therefore fulfilled. Thus, the optimal controller design for $k \neq 0$ presents a well-posed problem. Using the same synthesis algorithm, we obtain the following mathematical model for the optimal controller:

$$[(3 - 11k)D - (5 + 3k)]u = 12[(1 + 3k)D + 4]x. \qquad (1.4.26)$$

Then, the characteristic polynomial of the closed system is

$$\Delta = (20k - 3\varepsilon + 11\varepsilon k)D^2 + (20 + 12k + 5\varepsilon + 3\varepsilon k)D + 12. \qquad (1.4.27)$$

If the coefficient k is finite and ε is an arbitrarily small number, then the closed system is stable and, as could be expected from the Petrov inequality, the synthesis presents a well-posed problem.

As for finite, and not infinitesimal, variations ε, the problem is more difficult: for very small k, stability may be lacking. It follows from formula (1.4.27) that, for instance, with k=0.01 stability will be retained only if $\varepsilon < 0.069$, whereas already with $k = 0.1$ stability will be retained if $\varepsilon < 1.05$, i. e., almost for arbitrary variations.

Simultaneously, the coefficient k in approximation (1.4.25) is not to be chosen too large. The point here is as follows: if the disturbing-action spectrum is approximated with formula (1.4.21) and the controller is calculated by formula (1.4.25), then the magnitude of (1.4.20) will inevitably increase by a value that increases with the coefficient k. For instance, with $k = 0$, we have $J = 0.4336$; with $k = 0.1$, $J = 0.4374$ (increase by 0.88 %); and with $k = 0.3$, the value of J increases to $J = 0.4481$. The increase in the quality criterion is a kind of price to be paid for guaranteed parametric stability.

By introducing the parameter k into formula (1.4.25), we regularize the initial ill-posed problem, transforming it in a sequence of ill-posed problems whose solution in the limit $k \to 0$ tends to the solution of the initial ill-posed problem.

Note that, if $k \neq 0$ but is very small, then the synthesis is not an ill-posed problem, but remains an ill-conditioned problem: if, for instance, k=0.01, then already for $\varepsilon > 0.069$ the system loses stability, which in practice may result in a wreck. Yet, with $k = 0.1$ stability will be guaranteed.

The suggestion to make the optimal-controller synthesis a well-posed problem and the closed system parametrically stable by introducing into the analytical approximation of the spectrum multipliers of type $1 + k^2\omega^2$ was advanced and substantiated by Petrov (1977). This suggestion has enabled synthesis of controllers, appropriate for practical use, that have substantially improved performance characteristics of various control objects. The confidence to the optimum control theory was restored, and this theory gained a widespread practical application, providing a considerable economical effect. Practical applications of the theory to the design of optimal controllers were reflected, for instance, by Abdullaev and Petrov (1985), Zubov (1974), Larin, Naumenko, and Suntsev (1971), Letov (1960, 1969), Petrov (1973, 1977, 1987). We would like to mention here just one example: to diminish the vessel rolling, controlled rudders are widely used. The passage to the optimal rudder control enabled suppression of vessel rolling by a factor of 2.7 compared to the previously used control law (Abdullaev and Petrov, 1985,

p. 231). The optimal controller here is rather simple; its mathematical model is

$$(0.035D + 0.85)u = -(4.15D^3 + 10.91D^2 + 6.15D + 3.35)x, \qquad (1.4.28)$$

and the controller is easy to realize.

Equally easy to realize are controllers offering an effective control over seagoing crafts (Abdullaev and Petrov, 1985, p. 224) and over exciting regulators for synchronous machines in power engineering intended for maintaining stable voltage under load variations (Abdullaev and Petrov, 1985, pp. 209–215), etc.

All these controllers guarantee parametric instability of closed systems. Note that already after 1977, when the publication by Petrov (1977) was issued, another regularization method has gained widespread utility, i. e., instead of the quality control (1.4.3), it became a common practice to synthesize controllers that minimize the functional

$$J = \langle u^2 \rangle + k_0 \langle x^2 \rangle + k_1 \langle \dot{x}^2 \rangle + \ldots + k_n \langle (x^{(n)})^2 \rangle, \qquad (1.4.29)$$

where the additional terms $k_1 \langle \dot{x}^2 \rangle + \ldots + k_n \langle (x^{(n)})^2 \rangle$ are introduced just to overcome the property of the problem to be an ill-posed one. The synthesis algorithm for the controller that minimizes the functional (1.4.29) is not much more complex than that for (1.4.3), and parametric stability is guaranteed, although with a greater sacrifice in the quality criterion compared to the technique based on transformation of analytical approximation of the spectrum.

Yet, these are particularities. The main thing is that the investigation into factors resulting in parametric instability and unveiling the property of being ill-posed for a number of optimization problems enabled the use of optimal controllers making it possible to substantially improve performance characteristics of many industrial objects and transport facilities, thus providing a considerable economical effect.

1.5. ILL-POSED PROBLEMS ON FINDING EIGENVALUES FOR SYSTEMS OF LINEAR HOMOGENEOUS EQUATIONS

As is known, many important physical and technical problems require calculating the magnitude of some parameter (traditionally denoted as λ), for which a system of linear homogeneous equations with the parameter has nonzero solutions. By way of example, consider the system

$$\begin{cases} (\lambda^2 - 2\lambda)x_1 + (1 - 3\lambda)x_2 = 0, \\ \lambda x_1 - 3x_2 = 0. \end{cases} \tag{1.5.1}$$

Since system (1.5.1) is homogeneous, then, of course, this system has the trivial solution $x_1 = x_2 = 0$. Yet, for some values of λ, the system may also have nonzero solutions. For instance, system (1.5.1) has nonzero solutions if $\lambda = 0$. Then, this system assumes the form

$$\begin{cases} 0 \cdot x_1 + x_2 = 0, \\ 0 \cdot x_1 - 3x_2 = 0, \end{cases} \tag{1.5.2}$$

and any pair of numbers in which $x_2 = 0$ and x_1 is an arbitrary number will make equations (1.5.1) with $\lambda = 0$ identities.

The values of λ for which a system of linear homogeneous equations involving a parameter has nonzero solutions are called the *eigenvalues*. Finding the eigenvalues is an important step in solving systems of linear differential equations with constant coefficients; this problem is therefore often encountered in engineering calculations.

By way of example, consider the following system of equations for the variables y_1 and y_2:

$$\begin{cases} \ddot{y}_1 - 2\dot{y}_1 = 3\dot{y}_2 - y_2, \\ \dot{y}_1 = 3y_2. \end{cases} \tag{1.5.3}$$

We assume that the solutions of (1.5.3) are functions

$$\begin{aligned} y_1 &= x_1 \mathrm{e}^{\lambda t}, \\ y_2 &= x_2 \mathrm{e}^{\lambda t}, \end{aligned} \tag{1.5.4}$$

where, in the case of interest, x_1 and x_2 are the constants of integration (i. e., constant numbers). After insertion of (1.5.4) into (1.5.3), we obtain

$$\begin{cases} \mathrm{e}^{\lambda t}[(\lambda^2 - 2\lambda)x_1 + (1 - 3\lambda)x_2] = 0, \\ \mathrm{e}^{\lambda t}[\lambda x_1 - 3x_2] = 0. \end{cases} \tag{1.5.5}$$

Having cancelled out the nonzero function $\mathrm{e}^{\lambda t}$, we arrive at system (1.5.1). We see that nonzero solutions (1.5.4) exist for those values of λ, for which system (1.5.1) has nonzero solutions, i. e., for eigenvalues. It is these, and only these, values of λ that can be the exponents in the solutions of the system of linear differential equations with constant coefficients.

That is why finding eigenvalues of a homogeneous linear system of algebraic equations is a necessary stage in solving systems of linear differential equations with constant coefficients, — this statement is valid not only for the system under consideration, but also for any of the systems of interest.

In turn, the eigenvalues are roots of the polynomial-matrix determinant; the polynomial matrix for system (1.5.1) is

$$\begin{pmatrix} \lambda^2 - 2\lambda & 1 - 3\lambda \\ \lambda & -3 \end{pmatrix} \tag{1.5.6}$$

(recall that a *polynomial matrix* is a matrix whose elements are polynomials). The determinant of (1.5.6),

$$\Delta = \begin{vmatrix} \lambda^2 - 2\lambda & 1 - 3\lambda \\ \lambda & -3 \end{vmatrix} = 5\lambda, \tag{1.5.7}$$

has a single root, which is zero. Generally, for a system of differential equations

$$\begin{cases} A_1(D)x_1 + A_2(D)x_2 = 0, \\ A_3(D)x_1 + A_4(D)x_2 = 0, \end{cases} \tag{1.5.8}$$

where $D = \mathrm{d}/\mathrm{d}t$ is the differentiation operator, and $A_1(D)$, $A_2(D)$, $A_3(D)$ and $A_4(D)$ are polynomials of some degrees, whose eigenvalues are the roots of the determinant

$$\Delta = \begin{vmatrix} A_1(\lambda) & A_2(\lambda) \\ A_3(\lambda) & A_4(\lambda) \end{vmatrix}. \tag{1.5.9}$$

The same is valid for an arbitrary n-th order system of linear differential equations. The eigenvalues are the roots of the polynomial determinant (i. e., of a determinant whose elements are polynomials). For instance, the eigenvalues for the system

$$\begin{cases} A_{11}(D)x_1 + \ldots A_{1n}(D)x_n = 0, \\ \qquad \ldots\ldots\ldots \\ A_{n1}(D)x_1 + \ldots A_{nn}(D)x_n = 0 \end{cases} \tag{1.5.10}$$

are roots of the determinant

$$\Delta = \begin{vmatrix} A_{11}(\lambda) & \ldots & A_{1n}(\lambda) \\ \ldots & \ldots & \ldots \\ A_{n1}(\lambda) & \ldots & A_{nn}(\lambda) \end{vmatrix}. \tag{1.5.11}$$

The degrees of the operator polynomials in (1.5.10) depend on the problem under consideration. For instance, in the commonly encountered problem on finding frequencies of small-amplitude oscillations of mechanical or electrical systems, the equations for the oscillations are constructed based on second-kind Lagrange equations; hence, the polynomials $A_{ij}(\lambda)$ here are quadratic polynomials.

The problems on finding the roots of polynomial determinants similar to (1.5.11) can be either well- or ill-posed problem. For instance, calculation of (1.5.7) is an ill-posed problem. Indeed, consider variation of just one coefficient and calculate, instead of (1.5.7), the determinant

$$\Delta = \begin{vmatrix} \lambda^2 - 2\lambda & 1 - 3\lambda \\ \lambda & -3(1+\varepsilon) \end{vmatrix} = -3\varepsilon\lambda^2 + 6\varepsilon\lambda + 5\lambda. \tag{1.5.12}$$

Let us check, first of all, that for an arbitrarily small ε determinant (1.5.12) has not one, but two roots, $\lambda_1 = 0$ and $\lambda_2 = 2 + 5/(3\varepsilon)$, and, as $\varepsilon \to 0$, the second root by no means tends to the first one and vanishes if $\varepsilon = 0$ exactly.

For the polynomial-matrix determinants, the reason for the property of being ill-posed is quite clear: this property arises wherever, with nominal values of coefficients, the terms with the highest degree of λ cancel. It is clear that, even with arbitrarily small values of coefficients, we have no such cancellation; as a result, under small parameter variations, the degree of the polynomial of λ in the determinant undergoes changes; as a result, another polynomial root emerges.

Nevertheless, the fact that calculation of determinants of some polynomial matrices presents an ill-posed problem means that some of even more important and commonly encountered problems, problems on solving systems of ordinary differential equations, are also ill-posed problems. The property of being ill-posed may emerge even in solving the simplest class of differential equations often met in applications, namely, linear equations with constant coefficients.

1.6. SOLUTION OF SYSTEMS OF DIFFERENTIAL EQUATIONS. DO SOLUTIONS ALWAYS DEPEND ON PARAMETERS CONTINUOUSLY?

Since coefficients entering differential equations of a mathematical model of a natural process or a technical system under study are gained from tests and measurements, they are always known only to some limited accuracy; for this reason, the continuous dependence of solutions on coefficients and parameters represents a necessary (although not sufficient!) condition for practical appropriateness of calculation results. If there is no continuous dependence of solutions on parameters, then we cannot guarantee that inevitably occurring small deviations of true values of coefficients from their nominal values will not result in large errors in the solutions.

In such cases, even an experimental check guarantees nothing — with no continuous dependence of solutions on parameters, in an immediate vicinity of experimentally tried values there may be values of coefficients and parameters for which solutions may behave quite differently.

That is why the most important theorem in the theory of differential equations that forms a basis for all practical applications of the theory, is the well-known theorem about continuous dependence of solutions of ordinary differential equations on parameters. This theorem is discussed in all sufficiently detailed courses on differential equations, together with its proof, for one n-th order equation and for a system of n equations written in the normal Cauchy form, i. e., in the form

$$\begin{cases} \dot{x}_1 = f_1(x_1, \dots, x_n), \\ \qquad \cdots\cdots\cdots \\ \dot{x}_n = f_n(x_1, \dots, x_n). \end{cases} \tag{1.6.1}$$

For the theorem to be valid, it suffices that Lipschitz conditions be fulfilled for the right-hand parts of the equations. In practice, these conditions are almost always satisfied.

Since almost any system of equations unresolved with respect to higher-order derivatives can be brought to normal form, it is common practice to interpret this theorem loosely, assuming it to hold for all systems. As a matter of fact, this assumption is erroneous.

Consider system (1.5.3) of differential equations with a parameter m,

$$\begin{cases} \ddot{y}_1 = 2y_1 + 3\dot{y}_2 - y_2, \\ \dot{y}_1 = my_2 \end{cases} \tag{1.6.2}$$

and analyze how the solutions depend on the parameter m.

The characteristic polynomial of system (1.6.2) is given by the determinant

$$\Delta = \begin{vmatrix} \lambda^2 - 2\lambda & 1 - 3\lambda \\ \lambda & -m \end{vmatrix} = (3 - m)\lambda^2 + (2m - 1)\lambda; \tag{1.6.3}$$

this polynomial with $m \neq 3$ has two roots, $\lambda_1 = 0$ and $\lambda_2 = -(2m-1)/(3-m)$, whereas with $m = 3$, only one root, $\lambda_1 = 0$. Evidently, the value of $m = 3$ is singular. Suppose that $m = 3(1+\varepsilon)$; we are going to analyze now how the solutions depend on ε. The general solution of (1.6.2) is

$$y_1(t) = C_1 \mathrm{e}^{\lambda_1 t} + C_2 \mathrm{e}^{\lambda_2 t} = C_1 + C_2 \mathrm{e}^{(2+5/(3\varepsilon))t}. \tag{1.6.4}$$

The solution $y_2(t)$ is of the same form. The constants of integration in (1.6.4) can be found from boundary conditions. Let the initial conditions be such that $C_1 = 0$ and $C_2 = 1$. Then, let us ask ourselves, what will be the magnitude of $y_1(t)$, say, at the time t=1 as a function of ε. If ε is a small negative number, then the exponent $2+5/(3\varepsilon)$ is a large (in absolute value) negative number; it follows from here that the solution $y_1(t)$ will be close to zero. If, alternatively, ε is a small positive number, then the exponent $2+5/(3\varepsilon)$ is large and the solution $y_1(t)$ at $t = 1$ will be a very large number; this number will be the larger the smaller is ε. For $\varepsilon \to 0$ (i. e., for $m \to 3$), we have: $y_1(t) \to \infty$.

Thus, at the point $m = 3$ the solution $y_1(t)$ as a function of m suffers discontinuity, continuity being violated.

Next, consider practical consequences of this discontinuity. Let the nominal value of m in (1.6.2) be 2.999, and the error in setting this parameter (defined, for instance, by measurement error), be equal to two tenths. Calculating, for instance, the magnitude of $y_1(t)$ at $t = 1$ (or at any other time t) with the nominal value of m, $m = 2.999$, we readily obtain that$y_1(t) = 0$ to the fourth decimal place after the decimal point. Yet, with regard for the accuracy in setting m in (1.6.2), the true value of m can be 3.001; then, the true value of $y_1(t)$ will be large, and equally large will be the calculation error, pregnant with bad consequences, wrecks, and even catastrophes. The latter is caused by the fact that the problem of finding the solution of (1.6.2) with $m = 3$ is an ill-posed problem.

Crude errors in calculations will also arise in solving all system of equations whose characteristic polynomial changes its degree at some critical value of an involved parameter, resulting in cancellation of coefficients at the higher degree. If the magnitude of the parameter is not equal to the critical value exactly, but is close to it, then there will be no change of degree in the characteristic polynomial, but the coefficient at the highest degree will be small, i. e., the polynomial will acquire the form

$$\Delta = \varepsilon\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0, \tag{1.6.5}$$

where ε is a number much smaller than the coefficients $a_{n-1}, a_{n-2}, \ldots, a_0$. Here, polynomial (1.6.5) will have one large (in absolute value) root

$\lambda_n \approx -a_{n-1}/\varepsilon$, while the other $n-1$ roots will be close to the roots of (1.6.5) with omitted first term. This statement is easy to prove by dividing polynomial (1.6.5) by the binomial

$$\varepsilon\lambda/a_{n-1} + 1. \tag{1.6.6}$$

In residue, we obtain a polynomial close (for small values of ε) to polynomial (1.6.5) with omitted first term. The sign of the larger root $\lambda_n \approx -a_{n-1}/\varepsilon$ depends on the sign of ε; the root therefore changes its sign to the opposite when ε passes through zero. This means that in the case of $\varepsilon < 0$ a rapidly growing term will arise in the solution, while with $\varepsilon = 0$ the solution as a function of ε suffers discontinuity. All these features were previously displayed by system (1.6.2).

Thus, we arrive at the following conclusion: if at some critical value of some parameter in the characteristic polynomial of a system of differential equation of interest cancellation of coefficients at the higher term occurs, then at this value of the parameter the solution of the system presents an ill-posed problem: already under arbitrarily small variation of the parameter, the solution at any finite values of the argument t may suffer arbitrarily large changes. The reason for the property of being ill-posed here is discontinuity of solutions on the parameter.

If the magnitude of the parameter does not exactly equal the critical value, but, instead, differs from it by a small (but finite) number, then the solution of the system is not an ill-posed but ill-conditioned problem: small (but now only finite) variations of the parameter may change the solution by an arbitrarily large factor.

Note that solving ill-conditioned problem in this case is more risky than solving ill-posed problem: the property of being ill-posed will be evident from the large change of the degree of the characteristic polynomial under parameter variations, whereas the degree of the characteristic polynomial and the order of the system in an ill-conditioned problem both remain unchanged under parameter variations; for this reason, a crude error here is more probable.

Of course, there is a simple method making it possible to discriminate between ill-posed and ill-conditioned problems, i. e., to check that the solution remains almost unchanged for slightly changed values of parameters. Yet, for high-order systems wherein the total number of coefficients is large, this simple method becomes too labor-consuming.

All the aforesaid suggest a simple criterion that permits easy identification of suspicious systems that may be ill-conditioned: the highest term of the characteristic polynomial of such systems (or the highest term and several terms of lesser degree) is substantially smaller in value than other terms. This can be (but not necessarily!) indicative of the fact that the higher term of the characteristic polynomial has emerged as a small difference between large terms in the initial system, and small variations of these terms may give rise to large (relative) changes in the highest term or may make the term to change its sign for the opposite. Such a situation was already exemplified with system (1.6.2).

Interestingly, ill-posed problems, which, as was stated above, can be considered as a limiting case of ill-conditioned problems, are easier to identify: in an ill-posed problem for a system of differential equations, perfect cancellation of highest terms in the characteristic polynomial occurs at the nominal value of the coefficient, thus making the degree of the polynomial lower. The polynomial depression offers a simple indication for the property of being ill-posed. For instance, the characteristic polynomial of (1.6.2), given by the determinant (1.6.3), generally must be of the second degree, and the fact that, with $m = 3$, the degree of this polynomial diminishes presents a clear indication that the problem with $m = 3$ is an ill-posed one. The situation with ill-conditioned systems is more evolved: the degree of the characteristic polynomial for such systems remains unchanged, and the smallness of the coefficient at the highest term serves just an indication that this problem may be an ill-conditioned one with respect to coefficient variations.

Consider now the illusory contradiction between the well-known theorem about continuous dependence of solutions on parameters for one differential equation of arbitrary order and the lack of continuous dependence for some systems of differential equations. This illusory contradiction seems to result from the fact that a system of equations can be reduced to just one equation by equivalent transformations not affecting the solutions; that is why it seems, at first glance, that, once the theorem about continuous dependence of solutions on parameters holds for the single equation, then it must be also valid for the equivalent system. This is not the case.

System (1.6.2) with $m = 3$ has the solution $y_1 = C_1$ and can be equivalently transformed into the equation

$$(3 - m)\ddot{y}_1 + (2m - 1)\dot{y}_1 = 0. \qquad (1.6.7)$$

With $m = 3$, this equation reduces to

$$5\dot{y}_1 = 0; \qquad (1.6.8)$$

the latter equation has the same solution $y_1 = C_1$. The solution of (1.6.8) depends on the only coefficient $2m - 1$, which equals 5 if $m = 3$; on the other hand, this solution continuously depends on m. Yet, no immediate conclusions can be drawn from here as to continuous dependence of solutions of system (1.6.2) on m. We have already checked that solutions of (1.6.2) as functions of m suffer discontinuity at the point $m = 3$; formula (1.6.7) reveals the reason for this discontinuity: the coefficient at the term with the higher derivative vanishes if $m = 3$ and, hence, the term with the higher derivative, whose coefficient depends on m, vanishes if $m = 3$. Thus, it does not follow from the theorem about continuous dependence of solution of one differential equation on a parameter that a similar theorem holds for all systems of differential equations.

An analogous conclusion also applies to systems of differential equations written in Cauchy form, i. e. in form (1.6.1). For system written in normal form, the theorem about continuous dependence of solutions on parameters is valid, but this gives no grounds to assert that this theorem also holds for systems of equations written not in normal form, even if these systems can be brought to normal form by equivalent transformations (equivalent transformations and their properties will be discussed in more detail in Chapter 2).

Yet, the authors of many manuals on differential equations give the proof of the theorem about continuous dependence of solutions on parameters for systems in normal Cauchy form and for one equation, not mentioning the fact that for system written not in normal form continuous dependence may be lacking.

As a result, many graduate students of higher education institutions and even many graduate students of physico-mathematical and mathematico-mechanical departments of universities adhere to the opinion that the property of continuous dependence on coefficients is possessed by all systems of differential equations; this erroneous opinion can be the reason for crude errors in practical calculations.

That is why all specialists are to be forewarned that there exist systems of differential equations whose solutions do not continuously depend on parameters.

Let us give several examples of such systems:

1. The system

$$(D^3 + 4D^2 + 2D + 4)x_1 = (D^2 - D + 2)x_2, \tag{1.6.9}$$

$$(D^2 - 6D + 6)x_1 = (D - 25)x_2 \tag{1.6.10}$$

was discussed by Petrov and Petrov (1999). It is an easy matter to check that the coefficients of this system are such that to correspond to the point where the solutions as functions of the coefficient at Dx_1 in (1.6.10) and as functions of the coefficient at D^2x_2 in (1.6.9) suffer discontinuity.

2. The system

$$[TD^3 + (2+2T)D^2 + (4+T)D + s]x_1 = (D^2 + 2D + 1)x_2, \tag{1.6.11}$$

$$(D+1)x_2 = (D^2 + 4D + 5)x_1, \tag{1.6.12}$$

previously examined in (Petrov, 1994; Petrov and Petrov, 1999), describes processes in a system controlling the rotational speed of a direct-current drive. In (1.6.11), (1.6.12), x_1 is the deviation of the rotational speed from the optimum value and x_2 is the deviation of the motor armature current from the nominal value, i. e., the control action. Equation (1.6.12) presents the mathematical model for the controller optimal for disturbing actions with the spectrum

$$S_\varphi = 1/(1+\omega^2)^2. \tag{1.6.13}$$

The parameter T is the mechanical time constant of the electric drive and, since the time t in (1.6.11), (1.6.12) is expressed in fractions of the nominal value of the mechanical constant, the nominal value here is $T = 1$. One can check that the same value $T = 1$ defines the point where the solution $x_1(t)$ as a function of T suffers discontinuity. Indeed, the characteristic polynomial of (1.6.11), (1.6.12) is

$$\Delta = (1-T)\lambda^4 + (4-3T)\lambda^3 + (8-3T)\lambda^2 + (8-T)\lambda + 3, \tag{1.6.14}$$

and arbitrarily small deviations of T from the nominal value $T = 1$ give rise to a fourth, exponentially increasing term; this term grows in value the rapider the smaller is the difference $T - 1$.

For values of T as close to $T = 1$ as one likes, the growth rate of this term may become arbitrarily high. That is why the data calculated by the mathematical model (1.6.11), (1.6.13) for values of T close to $T = 1$ are not reliable. The solutions of (1.6.11), (1.6.12) may have nothing in common with the actual behavior of the electric drive. This example clarifies the difference between the variation of some differential-equation coefficient and the variation of some parameter of a technical apparatus or physical process. The variation of T, the mechanical time constant of the electric drive, may give rise to simultaneous variations of several coefficients in the system of differential equation that have different amplitudes (in (1.6.11), the coefficients suffering variations are the coefficients at D^3, D^2, and D); this circumstance needs to be taken into account in calculations.

The example of (1.6.11), (1.6.12) and many other similar examples given by Petrov (1987, 1994), Petrov and Petrov (1999) show that the loss of continuous dependence of solutions on parameters very often occurs in systems with the so-called noncanonic equations, i. e., with equations where the order of the derivatives in the right-hand part is higher than in the left-hand part; precisely such a situation is exemplified by equation (1.6.12). Noncanonic equations and sets of such equations were examined by V. A. Steklov (1927). In several last decades, these equations, although frequently met in applications and having many interesting properties, have been paid undeservedly little attention.

Now, it is easy a matter to establish a simple criterion making it possible to discriminate between well- and ill-posed problems in solving systems of linear differential equations with constant coefficients. To do this, it suffices to construct the "degree matrix", i. e., the matrix whose each next element represents the highest term of the operator polynomial of (1.5.10) and examine whether or not the highest terms in the determinant of the matrix will cancel.

The "degree matrix" for system (1.6.2) is

$$\begin{pmatrix} D^2 & -3D \\ D & -m \end{pmatrix}; \tag{1.6.15}$$

the terms with D^2 in the determinant of this matrix will cancel at $m = 3$. With $m = 3$, the solution of the system is an ill-posed problem.

The "degree matrix" for system (1.6.11), (1.6.12) is

$$\begin{pmatrix} TD^3 & -D^2 \\ D^2 & -D \end{pmatrix}, \tag{1.6.16}$$

and the terms with D^4 in the determinant of this matrix will cancel at the nominal value of T, the "mechanical time constant of the electric driver, i. e., at $T = 1$.

Thus, it is not difficult now to introduce a very simple criterion permitting easy discrimination between well- and ill-posed problems in the very important field, the solution of systems of linear differential equations with constant coefficients (the same criterion can easily be extended so that to cover many nonlinear problems). It is much more difficult a task to find a criterion enabling discrimination between ill-posed and ill-conditioned problems. The smallness of the highest term in the characteristic polynomial of the system of differential equations suggests nothing but suspicion that this problem may be ill-conditioned. In one of manuals, considered as an example is the system $\dot{x} = Ax$ with the matrix

$$A = \begin{pmatrix} 0 & 3.99 & 0 & 0 & 0 & 0 & 0 \\ -2575 & -127.4 & 2710 & 0 & 0 & 0 & 0 \\ 0 & 0 & -300 & 4200 & 520.8 & -520.8 & -692.7 \\ 0 & 0 & 0 & 0 & 15.8 & -15.8 & -21 \\ 0 & 0 & 0 & 0 & 0 & 0 & -21.3 \\ 0 & 10 & 0 & 0 & 0 & -500 & 0 \\ 24 & 0 & 0 & 0 & 0 & 0 & -200 \end{pmatrix};$$

the characteristic polynomial here is

$$\Delta = \lambda^7 + 1.127 \cdot 10^3\lambda^6 + 4.427\lambda^5 + 9.389 \cdot 10^7\lambda^4 + 1.181 \cdot 10^{10}\lambda^3 \\ + 7.838 \cdot 10^{11}\lambda^2 + 1.326 \cdot 10^{13}\lambda + 1.84 \cdot 10^{14}.$$

The first term in the polynomial is much smaller than all other terms; it therefore can be a small difference of large numbers that changes its sign for the opposite under small (but finite!) variations of the matrix coefficient; yet, this suspicion needs to be verified.

1.7. CONCLUSIONS

Based on the material discussed in Chapter 1, the following conclusions can be formulated:

1. Apart from the well-known ill-posed problems in the field of mathematical physics and partial differential and integral equations, there are many simpler yet not less important ill-posed problems among algebraic equations, differential equations, extremum problems, etc.

2. Each unforeseen encounter with an ill-posed problem may result in crude errors in computation, thus leading to wrecks and even catastrophes. This matter was discussed in more detail in § 1.4; many other examples and illustrative facts concerning this point can be found in (Petrov and Petrov, 1999).

3. To avoid errors, prior to solving any problem it is recommended to check if the problem is a well- or ill-posed one. Presently, such a check is not always performed, the opposite being highly desirable. Of course,

the direct check (performed by repetitive calculations with parameter variations) is too labor-consuming and cumbersome a procedure; for this reason, simple criteria permitting easy discrimination between well- and ill-posed problems can be of use here. For the control design problem, the simple inequality (1.4.18), called the Petrov criterion, serves such a criterion.

For systems of linear differential equations with constant coefficients, a simple criterion stems from the "degree matrices" that were considered in § 1.6.

Generally, operator closeness (or openness), inverse-operator finiteness (or infiniteness), etc. can serve as such criteria.

Chapter 2.

Problems intermediate between well- and ill-posed problems

2.1. THE THIRD CLASS OF PROBLEMS IN MATHEMATICS, PHYSICS AND ENGINEERING, AND ITS SIGNIFICANCE

Since 1902, when the class of ill-posed problems was discovered by Jacues Hadamard, the famous French mathematician, over a period of the first nine decades in the XX century all problems previously met in mathematics, physics and engineering were thought of as falling just into two classes of problems, the well-known class of well-posed problems and the class of ill-posed problems, the investigation into which, commenced in 1902, were with many important contributions due to Russian scientists (see Lavrent'ev, 1981; Ivanov, Vasin, and Tanana, 2002; Tikhonov and Arsenin, 1977; Verlan' and Sizikov, 1986; Sizikov, 2001; Lavrent'ev, Romanov, and Shishatskii, 1997; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995; Morozov, 1984; Bakushinsky and Goncharsky, 1994).

In the course of studies conducted at the Chair "Modeling of Electromechanical and Computer Systems", Faculty of Applied Mathematics and Control Processes, St.-Petersburg State University (SPbSU), a third class of problems, namely, the class of problems alternating, during their solution, between well- and ill-posed problems, was discovered. Such problems were isolated in a special, third class of problems, intermediate between well- and ill-posed ones (Petrov, 1998).

This third class of problems is important because its existence necessitates quite a different approach to the matter of discrimination between well- and ill-posed problems. After the year 1902, when first works by J. Hadamard were published, and up to the last decade of the XX century, it was supposed sufficient to check just once whether a problem of interest was a well- or ill-posed one, prior to turning to its solution. As a matter of fact, such a check is generally insufficient because there exist problems whose property of being well or ill posed alters under equivalent transformations applied to them in the course of their solution. That is why, to guarantee the reliability of calculation data, one has to perform the check repeatedly, after each transformation used, this check, of course, being a very labor-consuming procedure. Alternatively, equivalent transformations, traditionally used to treat the majority of problems, must be analyzed. The latter is just what we are going to discuss. We will see that it makes sense to subdivide all equivalent transformations into transformations equivalent in the classical sense and transformations equivalent in the widened sense. We will show that only transformations equivalent in the widened sense never alter well-posedness; such transformations can therefore be safely used to analyze all issues related to the matter of well- or ill-posedness, including parametric stability of control systems.

2.2. TRANSFORMATIONS EQUIVALENT IN THE CLASSICAL SENSE

Equivalent transformations, also called equipotent transformations, are taught at secondary schools; the theory of these transformations was accomplished as early as the XVIII century. Among simplest equivalent transformations are the addition of one and the same number to the left- and right-hand sides of an equation; the transposition of terms from one side of an equation to the other side with simultaneous change of sign; the multiplication of all terms in an equation by a nonzero number; the substitution, or replacement, of a term with an equivalent expression, etc.

Definition. *Equivalent (equipotent) systems of equations* are systems such that their solution sets coincide (Mathematical Encyclopedia, Vol. 4, p. 800, Sovetskaya Entsiklopediya, Moscow 1984). The Mathematical Encyclopedic Dictionary (1995) gives the same definition to equivalent (equipotent) systems of equations (see p. 511).

Equivalent transformations are such transformations the initial system and the transformed system for which are equivalent.

It is immediately seen from this definition that equivalence or nonequivalence of systems depends on the particular problem under consideration and on the manifold on which its solutions are sought for.

For instance, if we seek the values of x which satisfy the system

$$\begin{cases} 3x + y = 4, \\ x + y = 2, \end{cases} \tag{2.2.1}$$

(the values of y being of no interest), then the equation $2x = 2$, obtained by subtracting the second equation from the first one, is equivalent to system (2.2.1). Alternatively, if the values of both variables x and y that satisfy system (2.2.1) are of interest, then the equation $2x = 2$ is no longer equivalent to system (2.2.1).

The equations $x^2 - 1 = 0$ and $x^4 - 1 = 0$ are equivalent in the real field (either equation has the roots $x_1 = 1$ and $x_2 = -1$) and nonequivalent in the complex field because the equation $x^4 - 1 = 0$ has additional complex roots $x_3 = j$ and $x_4 = -j$.

In solving equations or systems of equations, it is necessary to observe if the transformations used were equivalent. In many transformations, in the transformed system there arise extra-solutions not satisfying the initial system; in other cases, there may be a loss of solutions. Of course, such transformations are not equivalent; yet (in contrast to obviously erroneous transformations that make the whole set of solutions entirely different), these transformations can be isolated in a special class of incompletely equivalent transformations whose equivalence can be restored by omitting extra-roots or by adding lost roots.

For instance, the equation $x - 2 = 0$ and the equation $x^2 - 3x + 2 = 0$, obtained by multiplication of the former equation by the binomial $x-1$, are of course nonequivalent: the first equation has one root, $x_1 = 2$, whereas the second, two roots, $x_1 = 2$ and $x_2 = 1$. Yet, it can be argued that after the omission of the extra-root $x{=}1$ (and only after this omission!) the equation $x^2-3x+2 = 0$ becomes equivalent to $x-2 = 0$. The use of such incompletely equivalent transformations (of course, after omission of extra-roots) can be of help in solving many practical problems.

In addition to the simplest equivalent transformation, consider termwise differentiation of an equation and its multiplication by a polynomial of the differentiation operator; these operations were examined by L. Euler in the XVIII century.

The equation

$$\dot{x} + x = 0, \tag{2.2.2}$$

with the initial condition $x(0) = 0$ has a single solution $x = 0$; this solution can be obtained from the general solution of (2.2.2), $x = C_1 \mathrm{e}^{-t}$, by determining the constant C_1 from the initial condition. Indeed, in the case under consideration we obtain $C_1 = 0$.

Termwise differentiation of (2.2.2) yields the second-order equation

$$\ddot{x} + \dot{x} = 0, \tag{2.2.3}$$

whose general solution,

$$x(t) = C_1 \mathrm{e}^{-t} + C_2, \tag{2.2.4}$$

contains two constants of integration; for a particular solution to be found, two initial conditions are necessary. Another condition must be added to the already available initial condition $x(0) = 0$. This second initial condition, for instance, a condition imposed on the derivative $\dot{x}$ at $t = 0$, is to be added, of course, not arbitrarily, but observing a certain strict rule. This rule is as follows: we calculate the derivative $\dot{x}$ at $t = 0$ from (2.2.2) and use the value thus obtained as a second initial condition for termwise differentiated equation (2.2.3). Since the solution of (2.2.2), $x = 0$, was already found, it follows from this solution that $\dot{x}(0) = 0$. With this second initial condition added, we can find the desired particular solution of (2.2.3) from the general solution of (2.2.3), that is, from (2.2.4): $x(t) = 0$; this solution coincides with the solution of (2.2.2). Thus, with properly chosen additional initial conditions, termwise differentiation presents an equivalent transformation.

Note that this statement is often contested by advancing objections based on indistinct understanding of the term “general solution of a differential equation”. They often say: “after termwise differentiation, equation (2.2.2), transformed into equation (2.2.3), assumes another general solution, namely, solution (2.2.4); hence, equations (2.2.2) and (2.2.3) are not equivalent”. Yet, the term “general solution”, whose coming into use dates back to the XVIII century, seems to be inadequate to the situation. In order to refer to a family of solutions depending on some constants of integration, it seems more rational to use, instead of the term “general solution”, the term “family of solutions”. When we determine constants of integration from some initial conditions, we obtain the desired particular solution.

It is not surprising now that, above, the family of possible solutions has changed after the termwise differentiation; yet, with a proper choice of additional initial conditions, the initial particular solution was restored.

As is well known, termwise differentiation is used in many integration methods for differential equations.

2.3. DISCOVERED PARADOXES

Despite the seeming simplicity of equivalent transformations taught at the secondary school, in the XX century some difficulties concerning these transformations were revealed; it was not soon that these difficulties got an adequate explanation.

The whole story takes its beginning from the control theory and, more specifically, from the synthesis method of control systems proposed by A. M. Letov in the 1960-ths; this method soon became widely known as "analytical design of optimal controllers" or simply "analytical design" (Letov, 1960, 1969). Note that, mathematically, a "controller" is an arbitrary relation between the control action, or the control, u and some controlled variables x_i. Relations (1.4.6), (1.4.12), (1.4.22), (1.4.26) exemplify such relations; these relations are to be subsequently embodied in technical devices also called "controllers".

A. M. Letov found that, for rather widely encountered linear control objects mathematically modeled with systems of linear differential equations written in the normal form

$$\begin{cases} \dot{x}_1 = a_{11}x_1 + \ldots + a_{1n}x_n + b_1 u, \\ \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dot{x}_n = a_{n1}x_1 + \ldots + a_{nn}x_n + b_n u \end{cases} \tag{2.3.1}$$

(where x_i are some controlled variables and u is the control), one can construct very simple a controller of form

$$u = -k_1 x_1 - \ldots - k_n x_n, \tag{2.3.2}$$

which (with properly chosen coefficients $k_1, k_2, \ldots, k_n$) is capable of ensuring both stability of the closed system and its parametric stability, simultaneously minimizing some quadratic form of the controlled variables and the control, for instance, the form

$$J = \int_0^\infty (u^2 + c_1 x_1^2 + \cdots + c_n x_n^2)\,\mathrm{d}t. \tag{2.3.3}$$

A. M. Letov developed a technique for calculating the amplification coefficients k_i in (2.3.2) that minimizes quadratic form (2.3.3) for arbitrary c_i, and also provides for stability of the closed system. A. M. Letov called this technique "analytical design" (Letov, 1960, 1969).

Since the minimum of quadratic forms similar to (2.3.3) often well reflected actual requirements to the operating quality of control objects and provided for acceptable transient processes, and, in addition, controllers of form (2.3.2) were easy to implement, the "analytical design" soon became very popular. Controller of form (2.3.2), intended for optimization of various control objects, rapidly came into use, and many studies devoted to various aspects of the "analytical design" were reported (more than 200 papers published only over the period of 1960–1968).

In practical applications of the technique, an interesting paradox has very soon emerged. In real control objects, very often not all controlled variables $x_1, x_2, \ldots, x_n$ in (2.3.1) could be directly measured and used in controller (2.3.2). The simplest and most natural way out of this situation was elimination of non-measurable variables out of (2.3.1)–(2.3.2) with the help of equivalent transformations. Following this elimination, only measurable variables and their derivatives were contained in the control-object equation (2.3.1) and in the controller equation (2.3.2). Since the transformations used were equivalent and, on these transformations, the solutions $x_i(t)$ suffered no changes, both stability of the closed system and its parametric stability were expected retained. These expectations did not prove true: stability was retained, transient processes were retained, but parametric stability often was not retained. Systems designed and constructed strictly following all "analytical design" recommendations often failed to preserve stability under small, inevitably occurring deviations of control-object or controller parameters from their design values, causing accidents. The latter caused the loss of interest to "analytical design", which soon came to naught, the more so that the causes for the loss of parametric stability remained obscure for a long time (see, for instance, Nadezhdin (1973), in which the matter of parametric stability was discussed, and also the discussion in "Avtomatika i Telemekhanika", the most authoritative journal in the field of control, initiated by this article). In spite of the heated debates, which, unfortunately, had not been then, in 1973, brought to completion, all subsequent attempts to gain a better insight into the problem and the reasons for alteration of parametric stability on equivalent transformations remained a failure.

2.4. TRANSFORMATIONS EQUIVALENT IN THE WIDENED SENSE

The problem of possible alteration of parametric stability upon equivalent transformation was substantially clarified by Petrov (1987), who posed a simple question: what, in fact, we mean by asserting that "a control system is parametrically stable or, which is the same, the problem of stability prediction for the given system presents a well-posed problem"? After all, this statement refers to the solutions of the system rather than to the system itself. This statement concerns the vicinity of the system, contending that other systems close (but not identical) to the initial system in the space of parameters has solutions that all are stable. When we say that the solution of the equation $\dot{x} + x = 0$ is parametrically stable, and the problem of stability prediction for this equation is a well-posed problem, this statement is equivalent to the statement that all solutions of the family of equations $(1 + \varepsilon_1)\dot{x} + (1 + \varepsilon_2)x = 0$, which, with small ε_1 and ε_2, represent a vicinity of the equation $\dot{x} + x = 0$, are stable. Then, provided that in the vicinity of the original system there is at least one system whose solutions are unstable, then the original system turns out to be parametrically unstable, and the problem of stability prediction for this system, presenting an ill-posed problem.

It immediately follows from here that equivalent transformations, which, by definition, do not affect the solutions of the system, by no means are bound to preserve all properties of the vicinity of the transformed system; hence, such transformations are not at all bound to preserve parametric stability of the transformed system and the well-posedness of the problem under study either. They may preserve, — and very often do preserve — the well-posedness, but may not preserve it. This means that the examples of non-conservation of parametric stability upon equivalent transformations met with surprise and hot debates by engineers and researchers in 1970-ths were, as a matter of fact, none of something extraordinary and paradoxial. Such situations may occur.

An example is given below.

The system of two differential equations

$$\begin{cases} (D^3 + 4D^2 + 5D + 2)x_1 = (D^2 + 2D + 1)x_2, \\ \qquad (D^2 + 4D + 5)x_1 = (D + 1)x_2, \end{cases} \tag{2.4.1}$$

where $D = \mathrm{d}/\mathrm{d}t$ is the differentiation operator, describes processes in a system devised to control the rotation frequency of a d. c. drive (x_1 is the rotation frequency and x_2 is the control, i. e., the motor-armature current). The characteristic polynomial of this system is

$$\Delta = \lambda^3 + 5\lambda^2 + 7\lambda + 3, \tag{2.4.2}$$

the roots are $\lambda_1 = -3$, $\lambda_2 = \lambda_3 = -1$, and the system is stable.

We introduce new variables x_3 and x_4, defined by the equalities

$$\begin{cases} x_3 = \dot{x}_1 + 2x_1 - x_2, \\ x_4 = \dot{x}_3, \end{cases} \tag{2.4.3}$$

and, as it should be, transform system (2.4.1) into the following system of four differential equations of the first order with respect to x_1, x_2, x_3, x_4:

$$\begin{cases} \dot{x}_1 = -2x_1 + x_2 + x_3, \\ \dot{x}_3 = x_4, \\ \dot{x}_4 = -x_3 - 2x_4, \\ \;\;0 = x_1 + x_2 + 2x_3 + x_4. \end{cases} \tag{2.4.4}$$

Here, the last differential equation is degenerate, yielding a fixed relation between the variables x_1, x_2, x_3, and x_4. The passage from (2.4.1) to the last equation in (2.4.4) will be considered in more detail in Section 2.5.

The characteristic polynomial of (2.4.4), polynomial (2.4.2), is the same as the characteristic polynomial of (2.4.1); like (2.4.1), system (2.4.4) is stable; and the families of solutions of both systems are

$$x_1(t) = C_1 \mathrm{e}^{-3t} + (C_2 t + C_3)\mathrm{e}^{-t}. \tag{2.4.5}$$

Systems (2.4.1) and (2.4.4) are therefore equivalent systems. Simultaneously, system (2.4.4) is parametrically stable, whereas system (2.4.1) is parametrically unstable: with arbitrarily small variations of some coefficients (the coefficient at D^2x_2 in the right-hand side of the first equation in (2.4.1), for instance) system (2.4.1) may lose stability.

Many other examples similar to systems (2.4.1)–(2.4.4) can be found in Petrov (1998), Petrov and Petrov (1999). these examples suggest introduction of a new mathematical notion: additionally to the previously used classical notion of equivalent transformations that do not change the solutions of transformed systems (in what follows, such transformations will be called transformations equivalent in the classical sense), we introduce the notion of transformations equivalent in the widened sense.

All transformations that

1) first, are equivalent in the classical sense, and

2) second, preserve the well-posedness of the problem under consideration,

will be called *transformations equivalent in the widened sense*.

The majority of equivalent transformations are equivalent both in the classical sense and in the widened sense; that is why the necessity in the new notion escaped notice for a long time. Yet, examples adduced by

Petrov (1987, 1994), Petrov and Petrov (1999) proved the existence of simple though rare transformations equivalent in the classical but not in the widened sense.

The notion of transformations equivalent in the widened sense (Petrov, 1994) is of significance because both the traditional procedures for analyzing stability of equations or systems of equations to parameter variations and the advanced methods (Petrov, 2001) using, for instance, the Kharitonov theorem (Kharitonov, 1978), are based on the examination of characteristic polynomials of closed systems. Meanwhile, already the example of (2.4.1) and (2.4.4), in which two systems having one and the same characteristic polynomial differed in parametric stability provides an indication that traditional methods for investigation into stability are obviously incomplete and by no means are capable of yielding reliable predictions. It does not suffice just to examine the characteristic polynomial; additionally, it must be checked which transformations were used to obtain this polynomial from the initial equations of the system under study. If the transformations used were classically equivalent, then there are no problems. Yet, if these transformations were equivalent in the classical but not in the widened sense, then the examination of characteristic polynomials will be of no help in avoiding possible miscalculations, that can reason of accidents in control systems (for more detail, see Petrov and Petrov, 1999).

The studies carried out by Petrov (1998), Petrov and Petrov (1999) showed that negativeness of real parts of roots of the characteristic polynomial of a system, negativeness of the real parts of eigenvalues of the coefficient matrix (for a system written in the normal Cauchy form), and the existence of Lyapunov function for a system under study do not give a conclusive answer to the question about parametric stability.

The simplest way to ensuring calculation reliability is to check whether the transformations used in calculations were equivalent in the widened sense. Here, however, a theory of such transformations is required, which is still far from being fully developed.

2.5. PROBLEMS INTERMEDIATE BETWEEN WELL- AND ILL-POSED PROBLEMS

From the existence of transformations that alter the property of a problem to be well or ill posed (i. e., from the existence of transformations equivalent in the classical but not in the widened sense), possible existence of problems that cannot be classed to well- or ill-posed problems, follows. It therefore makes sense to isolate such problems in a separate, *third class of problems.*

An example is the above-considered problem of stability prediction for the mathematical model of a d.c. drive. If the mathematical model is written in the normal Cauchy form, then the problem of stability prediction is a well-posed problem. Alternatively, if the mathematical model is written in the equivalent (in the classical sense) form of (2.4.1), then the same problem is an ill-posed one.

Numerous examples of problems belonging to the third class can be encountered in the cases where, in solving a problem, chains of transformations are used.

Consider the following system of linear homogeneous equations with some parameter λ:

$$\begin{cases} (1-\lambda)x_1 + x_2 + 2x_3 = 0, \\ x_1 + (1-\lambda)x_2 + 3x_3 = 0, \\ x_1 + x_2 = 0 \end{cases} \tag{2.5.1}$$

and pose a problem similar to one of the problems treated in Chapter 1: it is required to find the eigenvalues λ, i. e., the values of λ for which system (2.5.1) has nonzero solutions x_1, x_2, x_3.

Note that the parameter λ does not enter the third equation in (2.5.1). Such systems, i. e., systems whose one or several equations do not contain λ, are often met in mechanics. The equations that do not contain λ refer to holonomic (i. e., not containing derivatives) constraints between variables.

As it was argued in Chapter 1, the eigenvalues coincide with the roots of the determinant of (2.5.1):

$$\Delta = \begin{vmatrix} 1-\lambda & 1 & 2 \\ 1 & 1-\lambda & 3 \\ 1 & 1 & 0 \end{vmatrix}. \tag{2.5.2}$$

Calculation of (2.5.2) by expanding this determinant by the last-row minors yields $\Delta = -5\lambda$; hence, system (2.5.1) has a single eigenvalue $\lambda_1 = 0$.

The eigenvalue problem for (2.5.1) is a well-posed problem. This statement can be proved, in a most simple yet rather cumbersome way, by composing the varied determinant Δ_v in which to each of the nonzero coefficients of (2.5.1), its variation is added. With the coefficient variations, determinant (2.5.2) becomes

$$\Delta_v = \begin{vmatrix} 1+\varepsilon_1-(1+\varepsilon_2)\lambda & 1+\varepsilon_3 & 2(1+\varepsilon_4) \\ 1+\varepsilon_5 & 1+\varepsilon_6-(1+\varepsilon_7)\lambda & 3(1+\varepsilon_8) \\ 1+\varepsilon_9 & 1+\varepsilon_{10} & 0 \end{vmatrix}. \tag{2.5.3}$$

Calculation of (2.5.3) by expanding this determinant by the last-row minors yields (after omission of the small products $\varepsilon_i\varepsilon_j$)

$$\Delta_v = \varepsilon_3 + \varepsilon_5 - \varepsilon_1 - \varepsilon_6 + (5 + \varepsilon_2 + 2\varepsilon_4 + \varepsilon_7 + \varepsilon_9 + \varepsilon_{10})\lambda. \qquad (2.5.4)$$

From here, we find the eigenvalue

$$\lambda_1 = \frac{\varepsilon_1 - \varepsilon_3 - \varepsilon_5 + \varepsilon_6}{5 + \varepsilon_2 + 2\varepsilon_4 + \varepsilon_7 + \varepsilon_9 + \varepsilon_{10}}. \qquad (2.5.5)$$

We see that, with small ε_i, the eigenvalue changes weakly, and the eigenvalue problem is a well-posed problem.

Yet, as it is well known, the minor expansion of determinants is normally used to calculate determinants of order not higher than four-five since, with increase in the determinant order, the volume of calculations sharply increases. For this reason, for systems of linear homogeneous equations with a large number of equations, it has become common practice to reduce the total number of equations by elimination of variables. Let us illustrate this process with the example of the simple system (2.5.1). For the variable x_1 to be eliminated, we first take the first and second equations in (2.5.1), multiply these equations respectively by -1 and $1 - \lambda$, and add them together. The terms with x_1 cancel, yielding the equation

$$(\lambda^2 - 2\lambda)x_2 + (1 - 3\lambda)x_3 = 0 \qquad (2.5.6)$$

that contains only x_2 and x_3. Then, we take the second and the third equations of the system, multiply the second equation by -1 and add the resulting equation to the third equation. Again, the terms with x_1 cancel, and we obtain the equation

$$\lambda x_2 - 3x_3 = 0 \qquad (2.5.7)$$

that contains only x_2 and x_3. On the whole, on elimination of x_1 we obtain a system of two equations, (2.5.6) and (2.5.7), for two variables, whose determinant,

$$\Delta = \begin{vmatrix} \lambda^2 - 2\lambda & 1 - 3\lambda \\ \lambda & -3 \end{vmatrix} = 5\lambda, \qquad (2.5.8)$$

has depressed order compared to that of (2.5.1), but, like determinant (2.5.2), has the same unique root $\lambda_1 = 0$.

System (2.5.6)–(2.5.7) is equivalent to system (2.5.1) in the classical sense with respect to the eigenvalue problem and has the same eigenvalue $\lambda_1 = 0$ as system (2.5.1) does.

Yet, the eigenvalue problem for system (2.5.6), (2.5.7) is an ill-posed problem. System (2.5.6), (2.5.7) has been already dealt with in Chapter 1 (see system (1.5.3) with determinant (1.5.7)); then, we saw that if, for instance, the coefficient at x_3 in (2.5.7) will change by an arbitrarily small value to become $-3(1+\varepsilon)$, then system (2.5.6), (2.5.7) will happen to have two eigenvalues, $\lambda_1 = 0$, $\lambda_2 = 2 + 5/(3\varepsilon)$, and the second eigenvalue will differ significantly from the first one even with an arbitrarily small ε, vanishing only at $\varepsilon = 0$.

Thus, with respect to the eigenvalue problem, system (2.5.6), (2.5.7) is equivalent to system (2.5.1) in the classical sense and nonequivalent to it in the widened sense.

Such systems, equivalent in the classical but not in the widened sense are often encountered on elimination of variables out of systems of linear homogeneous equations with a parameter, in solution of eigenvalue problems for parameters, and in calculation of frequencies of small-amplitude oscillations of various mechanical and electrical systems.

Consider now practical effects from this situation. If all coefficients of a system are small integer numbers (like in the case of system (2.5.1)), then ill-posedness of the problem for the system is of no danger, provided that the order of this system is not too high. Yet, repeated multiplications of coefficients, necessary for elimination of variables, lead to accumulation of roundoff errors. Indeed, if a computer performs calculations accurate to eight digits and coefficients have eight significant digits, then multiplication of two such coefficients will yield a number with sixteen significant digits. On rounding to eight significant digits, an inevitable error occurs. As a result, already during the calculations the new coefficients acquire some variations ε_i independent of errors in setting the initial coefficients. If the transformed system (say, after elimination of x_1) is an ill-posed one, then even arbitrarily small errors may cause incorrect results.

For instance, in solving the eigenvalue problem for a system analogous to system (2.5.1) but with coefficients in the form of decimal fractions with the use of the whole digit capacity of the computer, when the eigenvalues are calculated by sequential elimination of variables starting from x_1, we may happen to obtain a false second eigenvalue whose magnitude will depend on the roundoff error.

Analogous false eigenvalues may also arise in the case of more complex systems; e. g., if, in the course of transformations of a particular mathematical model, transformations equivalent in the classical but not in the widened sense were used.

Thus, yet another possible reason for miscalculations is unveiled, namely, the alteration of well-posedness of a problem on transformations used in treating the problem. With the reason known and adequately understood, the correct result can be obtained. What is dangerous is ignorance, or an unexpected encounter with an erroneous result the reasons for which remain obscure. That is why at least brief information concerning the existence of the third class of problems in mathematics, physics and engineering, i. e., the existence of problems intermediate between well- and ill-posed ones, must be included in curricula. A more detailed consideration of these points is given by Petrov and Petrov (1999).

The aforesaid shows that one can judge well- or ill-posedness of a problem, considering the following triad:

1. Mathematical model.
2. Problem posed.
3. Solution method used to solve the problem.

One and the same problem for a mathematical model, or for a system of equations, may be well posed, whereas for another model, ill posed. This is rather clear. It is less obvious that for one and the same mathematical model, for one and the same problem considered in this model, one solution method may make the problem well posed, whereas another method may lead to ill-posedness.

An example is as follows: for system (2.5.1), finding eigenvalues through calculation of determinant (2.5.2) by minor expansion presents a well-posed problem, whereas the same problem using sequential elimination of variables is an ill-posed one.

That is why, when investigating into well- or ill-posedness, it is more adequate to consider the whole triad instead of just the problem of interest taken alone.

As mentioned above, there exist transformations equivalent in the widened sense and transformations equivalent in the classical but not in the widened sense, i. e., transformations that do not change the solutions of the transformed system, but alter correctness of a problem.

How do we discriminate between such transformations? The discrimination should be made considering triads. With transformations considered alone, it is hardly possible to find transformations that always, in all triads, will be equivalent in the widened sense. Even most simple transformations may alter correctness of a problem.

An example is as follows: we have already transformed the second equation in (2.4.1) to new variables x_1, x_2, x_3, x_4, where the variables x_3 and x_4 are defined by (2.4.3). Consider this transformation in more detail. Below, we will use only transformations of two types, namely, transpositions of terms from the left-hand side of an equation to its right-hand side with simultaneous change of sign and partition of terms (when, for instance, instead of $4Dx_1$ we write $2Dx_1 + 2Dx_1$). On such transformations, the second equation in (2.4.1) becomes

$$\big[(D^2 + 2D)x_1 - Dx_2\big] + \big[(2D + 4)x_1 - 2x_2\big] + x_2 + x_1 = 0.$$

Now, with (2.4.3), in the first square bracket we recognize the variable x_4, in the second bracket the term $2x_3$, and the whole second equation in (2.4.1) transforms into the last equation of (2.4.4). Yet, we saw that the transformation of (2.4.1) into (2.4.4) alters correctness. Thus, even simple transposition of terms from the left-hand side of an equation to its right-hand side with simultaneous change of sign may prove equivalent in the classical sense but not in the widened sense. Everything depends on the properties of a particular triad.

The properties of transformations equivalent in the widened sense are more complex than the simple properties, and rules, of classically equivalent transformations widely taught in the secondary school.

If some transformation (for instance, multiplication of all terms by a nonzero number) is equivalent in the classical sense, then this transformation will be equivalent always, in all problems, and for all equations. One and the same transformation can be equivalent in the widened sense for one system and not equivalent for another system. Equivalence in the widened sense can be established considering the already mentioned triad, namely, 1) the mathematical model; 2) the problem under consideration; and 3) the method used to solve the problem.

Statements about equivalence in the widened sense can only sound like the following statement: for such and such problem and such and such mathematical model such and such transformation is (or is not) equivalent in the widened sense. Examples will be given below.

2.6. APPLICATIONS TO CONTROL SYSTEMS AND SOME OTHER OBJECTS DESCRIBED BY DIFFERENTIAL EQUATIONS

Unveiling of the discussed differences between transformations equivalent in the classical sense and transformations equivalent in the widened sense allows a more adequate approach to the matter of analyzing parametric stability of systems of differential equations, to that of estimating stability margins in variation of parameters, and to other points that arise in the control theory. Finally, making distinction between the two types of transformations allows one to avoid significant errors in calculations.

Example. It is well known (see, for instance, Petrov, 1987) that the equations governing one of the numerous control objects, a d.c. drive of an actuating mechanism with the spectrum of drag-torque oscillations of form

$$S_\varphi = 1/(1+\omega^2)^2, \tag{2.6.1}$$

can be written as the following system of differential equations:

$$\begin{cases} \dot{x}_1 = -2x_1 + x_2 + x_3, \\ \dot{x}_3 = x_4, \\ \dot{x}_4 = -x_3 - 2x_4, \end{cases} \tag{2.6.2}$$

where x_1 is the deviation of the rotation frequency from its nominal value, x_2 is the control (deviation of the motor-armature current from its design value), x_3 is the deviation of the drag torque from its rated value, and x_4 is the derivative of the latter deviation; the independent variable is the time t. The last two equations in (2.6.2) are the equations of the so-called "shaping filter" that cuts off, out of the "white noise", the disturbing-action spectrum (2.6.1). On the whole, equations (2.6.2) coincide with the already considered first three equations in (2.4.4).

Provided that the control action is shaped, as a function of x_1, x_3, and x_4, by the law

$$x_2 = -x_1 - 2x_3 - x_4, \tag{2.6.3}$$

with applied feedback (2.6.3) then the block diagram of the control system will look as in Figure 2.1.

If the variables x_3 and x_4, as it is often the case, cannot be directly measured, then they can be eliminated out of (2.6.3) with the help of classically equivalent transformations. On the elimination, equation (2.6.3) acquires the form

$$(D^2 + 4D + 5)x_1 = (D+1)x_2. \tag{2.6.4}$$

Consider now the block diagram of the control system in which the control action is shaped according to equation (2.6.4), based on measurements

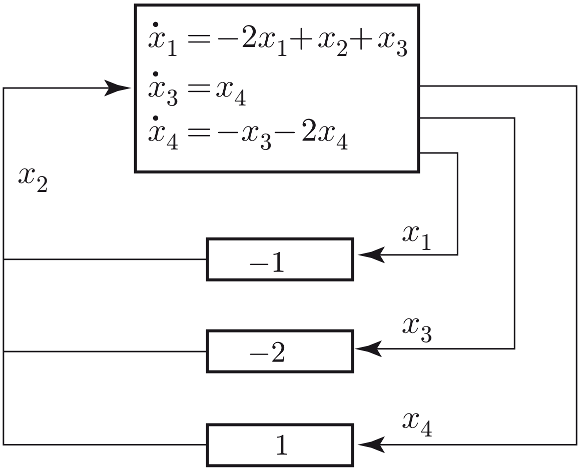

Figure 2.1. Block diagram of control object (2.6.2) with applied feedback (2.6.3)

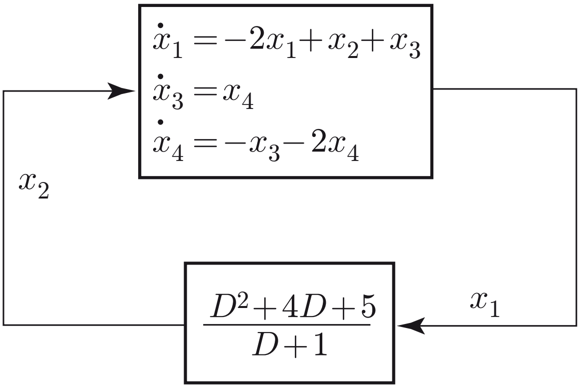

Figure 2.2. Block diagram of control object (2.6.2) with applied feedback (2.6.4)

of one variable, x_1, but the transient processes (solutions $x_1(t)$, $x_2(t)$, $x_3(t)$, and $x_4(t)$) remain unchanged. The new block diagram is shown in Figure 2.2.

A fleeting glance cast at the block diagrams of Figures 2.1 and 2.2 reveals at once that systems (2.6.2), (2.6.3) and (2.6.2)–(2.6.4) are not identical. These systems are equivalent in the classical sense; yet, they are not identical (this fact is immediately evident from Figures 2.1 and 2.2) and not equivalent in the widened sense. They have the same solutions but, simultaneously, possess different parametric stability. It is the introduction of the new mathematical notion, that of equivalence in the widened sense, that explains why on transformations equivalent in the classical but not in the widenee sense, some subtle yet important properties of control systems, such as parametric stability, suffer alteration.

That is why, when examining such properties, one cannot use, without additional checks, traditional calculation algorithms and habitual equivalent transformations of equations. Additionally, it is necessary to check that the transformations used were not only classically equivalent but also equivalent in the widened sense (or, in other words, to check correctness after each transformation). Using traditional methods without performing such checks leads to serious miscalculations that, as was mentioned above, may cause unexpected accidents in control systems (Petrov and Petrov, 1999).

Whether the calculation results will be erroneous or not, depends on which of available mathematical models equivalent to each other in the classical sense better reflects properties of the actual object. If the real relations between the control and the controlled variables in the electric-drive control system are more adequately described by system (2.6.2)–(2.6.4), then calculations by system (2.6.2), (2.6.3) will yield improper results concerning parametric stability of the real control system, although both systems, (2.6.2), (2.6.3) and (2.6.2)–(2.6.4), are equivalent in the classical sense, and either of them can be used to analyze stability (not parametric stability but just stability according to Lyapunov).

Note one more circumstance that often causes misunderstanding: when we perform an equivalent transformation of some system of equations, we arrive at a new system with some new coefficients. For instance, elimination of x_2 out of the system

$$a_{11}x_1 + a_{12}x_2 = b_1,$$
$$a_{21}x_1 + a_{22}x_2 = b_2,$$

yields the equation $(a_{12}a_{21} - a_{11}a_{22})x_1 = a_{12}b_2 - a_{22}b_1$. Instead of the four coefficients in the initial system, we now have just two coefficients remained.

Is the addition of arbitrarily small variations to the coefficients in the transformed system and subsequent examination of the impact of these variations on the solution capable of yielding the right answer to the question about the influence of coefficient variations in the initial system on the solutions?

The answer is easy to obtain: the coefficients in the transformed system are functions of the coefficients of the initial system; for instance, the first coefficient in the transformed system, a_{1tr}, is a certain function $k_1(a_{11}, a_{12}, \dots)$

of the coefficients of the initial system, i. e., $a_{1tr} = k_1(a_{11}, a_{12}, \dots)$; analogously, $a_{2tr} = k_2(a_{11}, a_{12}, \dots)$, etc. Here, two cases are possible: at the nominal values of the coefficients in the initial system these functions

1) are continuous,

2) have discontinuity.

In the first case, to arbitrarily small variations of coefficients in the transformed system, equally small variations of coefficients in the starting system correspond. For this reason, the check of stability can be performed considering any of the two systems, — of course, if the transformations used were equivalent in the widened sense. Otherwise, as we already showed, miscalculations are possible.

In the second case, to arbitrarily small variations of the coefficients of the initial system, finite, and even large, changes of coefficients in the transformed system may correspond. Here, one immediately sees that the problem for the initial system was an ill-posed problem.

Example. The system of equations

$$\begin{aligned} \ddot{x}_1 &= -\dot{x}_2 - x_2, \\ \dot{x}_2 &= -m\ddot{x}_1 + \mathrm{e}^{-t} \end{aligned}$$

with the parameter m and the zero initial conditions $x_1(0) = \dot{x}_1(0) = x_2(0) = \dot{x}_2(0) = 0$ can be transformed, on introduction of a new variable $x_3 = \dot{x}_1$, to the normal Cauchy form:

$$\begin{aligned} \dot{x}_1 &= x_3, \\ \dot{x}_2 &= \frac{m}{1-m}\,x_2 + \frac{1}{1-m}\,\mathrm{e}^{-t}, \\ \dot{x}_3 &= -\frac{1}{1-m}\,x_2 - \frac{1}{1-m}\,\mathrm{e}^{-t}. \end{aligned}$$

It is seen that at $m = 1$ the continuous dependence of the solution on the parameter is lost. This means that in the initial system with $m = 1$ there is no continuous dependence of solutions on m either.

This is a simple case that is easy to identify. Previously, more “tricky” examples were considered, when in a system reduced to the normal Cauchy form the solutions continuously depended on a parameter, whereas in the initial (non-transformed) system there was no continuous dependence of solutions on the parameter.

The introduction of the notion of transformations equivalent in the widened sense closes the discussion about admissibility of reduction of identical multipliers in transfer functions of control systems. Let, for instance, the transfer function be

$$\bar{u} = \frac{s+1}{(s+1)(s+2)}\,\bar{x}, \tag{2.6.5}$$

where the symbol s denotes the Laplace transform. Is reduction by $s+1$ admissible? On the one hand, this reduction is an equivalent transformation, it does not change the solutions $u(t)$ and facilitates computations. On the other hand, the use of such reductions in analyzing preservation of stability against parameter variations sometimes led to erroneous results. It was speculated that such reductions in transfer functions, although facilitating calculations, were inadmissible.

After the notion of transformations equivalent in the widened sense was introduced, everything has become clear: a reduction of the numerator and the denominator in the expression for the transfer function by equal multipliers is a transformation equivalent in the classical sense but by no means necessarily equivalent in the widened sense. If we are interested just in the calculation of the solution, i. e., in the calculation of the function $u(t)$ for nominal values of coefficients, then such a reduction is admissible and simplifies calculations. If, alternatively, we are interested in parametric stability, then such a reduction may prove inadmissible.

The same applies to any objects modeled with systems of ordinary differential equations of various orders. Provided that we are interested only in the solutions at nominal values of coefficients, we may use classically equivalent transformations and bring the systems of equations under study to the normal Cauchy form since this transformation makes it possible to use standard software, composed so that to treat systems written in the normal form. If, alternatively, we want to trace how small errors in experimentally determined coefficients affect the behavior of solutions, then, here, the use of transformations equivalent in the classical but not in the widened sense may yield incorrect results. For instance, in system (2.6.2)–(2.6.4), upon arbitrarily small deviations of the coefficients at D^2x_1 and Dx_2 from their nominal values, the behavior of the solutions will display significant changes, whereas in system (2.6.2), (2.6.3), written in the normal form and classically equivalent to system (2.6.2)–(2.6.4), small deviations of any of the coefficients from its nominal values weakly affect the solutions.

In order to avoid errors, the study of the dependence of solutions on almost inevitably occurring errors in setting coefficients must be performed using the initial (starting) model obtained directly from the physical and mathematical laws; if, for the sake of convenience, transformations were used to simplify the mathematical model, it is necessary to observe that the transformations were equivalent not only in the classical but also in the widened sense.

For instance, if we write the system whose block diagram is shown in Figure 2.2 and whose initial equations are equations (2.6.2)–(2.6.4) in the normal form, namely in the form of equations (2.6.2)–(2.6.3), then the calculation results — because of errors in set coefficients — may be erroneous. In calculations by the normal form, system (2.6.3), (2.6.4) is parametrically stable, but, in fact, parametric stability is lacking.

Note now the importance of making distinction between ill-posed and ill-conditioned problems. Recall that ill-posed problems are problems in which finite, and even large, changes of solutions result from arbitrarily small variations of coefficients, parameters, initial or boundary conditions, whereas ill-conditioned problems are problems in which finite (and often large) changes of solutions result from small but finite changes of coefficients, parameters, etc. The definition of "ill-conditioned" problems is somewhat obscure since the question of which precisely variations of coefficients and solutions should be regarded "small" and which "substantial" and "large" depends on specific features of the particular problem, on the measurement accuracy provided by particular measuring devices, on the requirements to the precision in obtaining the solutions, etc.

In contrast, ill-posed problems admit a strict definition (see Chapter 1); they are easier to identify compared to ill-conditioned problems.

An example is the system that consists of equation (2.6.4) and the equation

$$(D^3 + 4D^2 + 5D + 2)x_1 = (kD^2 + 2D + 1)x_2. \qquad (2.6.6)$$

With $k = 1$, calculation of solutions of system (2.6.4)–(2.6.6) and check for its stability are ill-posed problems: on an arbitrarily small deviation of k from unity, i. e., from $k = 1$, the solutions $x_1(t)$ and $x_2(t)$ for any $t > 0$ may be changed by any large values; with $k = 1$, system (2.6.4)–(2.6.6) is asymptotically stable, whereas with $k = 1 + \varepsilon$ it loses stability already at an arbitrarily small ε provided that $\varepsilon < 0$.

Yet, all these troubles can be foreseen since on elimination of, say, the variable x_2 out of system (2.6.4)–(2.6.6) this system transforms into the equation

$$[(k-1)D^4 + (4k-3)D^3 + 5kD^2 + 7D + 3]x_1 = 0 \tag{2.6.7}$$

that reduces the order at $k = 1$. This reduction gives a signal about possible ill-posedness of both the problem of finding solutions and the problem of checking stability of these solutions at $k = 1$.

Consider now the same system (2.6.4)–(2.6.6) at $k = 1.001$. On elimination of x_2, the system transforms, as it should be, into the fourth-order equation

$$(0.001D^4 + 1.004D^3 + 5.005D^2 + 7D + 3)x_1 = 0. \tag{2.6.8}$$

This equation shows no depression; hence, there is no "alarm signal" either. Indeed, system (2.6.4)–(2.6.6) at $k = 1.001$ is both stable and parametrically stable. The problem of checking stability of system (2.6.4)–(2.6.6) at $k = 1.001$ is a well-posed problem: on an arbitrarily small deviation of k from $k = 1.001$, the solutions of (2.6.4)–(2.6.6) will change by arbitrarily small values and remain stable. Yet, the stability margins with respect to variation of k in system (2.6.4)–(2.6.6) with $k = 1.001$ are very narrow: with the value of k having changed only by 0.11 %, system (2.6.4)–(2.6.6) loses stability. Since, normally, it is hard to guarantee beforehand that in the course of exploitation of the control system the variations of k will never happen to exceed 0.11 %, then, from the practical point of view, system (2.6.4)–(2.6.6) at $k = 1.001$ should be regarded as parametrically unstable or, more precisely, having inadmissibly narrow stability margins with respect to variations of k. Simultaneously, in this case there will be no warning alarm in the form of depression of the equation after elimination of x_2 or x_1. Moreover, calculating the roots of the characteristic polynomial of (2.6.4)–(2.6.6) at $k = 1.001$, we obtain that these roots (found accurate to the first significant digit) are $\lambda_1 = -3$, $\lambda_2 = \lambda_3 = -1$, and $\lambda_4 = -1000$. We see that all roots of the characteristic polynomial fall in the left complex half-plane and lie there far to the left from the imaginary axis; that is why, according to traditional checking methods, system (2.6.4)–(2.6.6) would be recognized stable at $k = 1.001$ and displaying broad stability margins. In fact, this is not the case, but the traditional examination of the characteristic polynomial gives no alarm signal. This is related to that fact that system (2.6.4)–(2.6.6) is not an ill-posed one at $k = 1.001$, being instead an ill-conditioned problem. Yet, such systems are more difficult to examine and test compared to ill-posed systems; hence, other methods are to be used here to investigate into their stability.

At the same time, isolation of ill-posed problems is also expedient in solving ill-conditioned problems. If at some particular values of coefficients or parameters a problem was found to be ill posed, then at coefficients and parameters close to the critical values the problem very often turns out ill-conditioned. The closeness to ill-posed problems enables identification of ill-conditioned problems.

2.7. APPLICATIONS TO PRACTICAL COMPUTATIONS

Making distinction between transformations equivalent in the classical sense and transformations equivalent in the widened sense makes it possible to unveil one of possible reasons for miscalculations. This reason consists in that, as we perform transformations of an initial model, we may unconsciously replace the initially well-posed model with an ill-posed one, and then any arbitrarily small roundoff error in calculations will lead to an erroneous result.

Consider the generalized eigenvalue problem for matrices and, by way of example, the system

$$\left.\begin{aligned}(a_{11}-\lambda)x_1+a_{12}x_2+\dots\dots\dots\dots\dots\dots+a_{1n}x_n&=0,\\ a_{21}x_1+(a_{22}-\lambda)x_2+\dots\dots\dots\dots\dots\dots+a_{2n}x_n&=0,\\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\\ a_{n-1,1}x_1+\dots\dots+(a_{n-1,n-1}-\lambda)x_{n-1}+a_{n-1,n}x_n&=0,\\ a_{n1}x_1+a_{n2}x_2+\dots\dots\dots\dots\dots\dots\dots\dots+a_{nn}x_n&=0,\end{aligned}\right\}\qquad(2.7.1)$$

i. e., a system of linear homogeneous equations with some parameter λ in which all equations, except for the last one, contain λ. It is required to find the values of λ at which system (2.7.1) has nonzero solutions.

Systems analogous to system (2.7.1), with various numbers of equations not containing λ are encountered in problems on finding the natural oscillation frequencies of various mechanical and electromechanical systems, in control problems, etc. (The equations that do not contain λ reflect holonomic constraints; such constraints are often met in practice).

Suppose that we are going to calculate the eigenvalues using sequential elimination of variables out of system (2.7.1) by termwise multiplications and additions (note that, as a matter of fact, many methods yielding eigenvalues are known; we do not compare the various methods and their advantages and disadvantages; instead, we focus on specific features that may be met when using one of the available methods).

On elimination of $n-2$ variables, prior to proceeding with the final elimination step, we arrive at a system of two equations:

$$\begin{aligned} A_1 x_{n-1} + A_2 x_n &= 0, \\ A_3 x_{n-1} + A_4 x_n &= 0, \end{aligned} \tag{2.7.2}$$

where $A_1, \ldots, A_4$ are polynomials of λ. Based on the Kramer formulas, we readily establish that these polynomials are given by the following determinants of the $n-1$ order:

$$A_1 = \begin{vmatrix} a_{11}-\lambda & \ldots & a_{1,n-1} \\ \ldots & \ldots & \ldots \\ a_{n-1,1} & \ldots & a_{n-1,n-1}-\lambda \end{vmatrix}, \tag{2.7.3}$$

$$A_2 = \begin{vmatrix} a_{11}-\lambda & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{n-1,1} & \ldots & a_{n-1,n} \end{vmatrix}, \tag{2.7.4}$$

$$A_3 = \begin{vmatrix} a_{21} & \ldots & a_{2,n-1} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{n,n-1} \end{vmatrix}, \tag{2.7.5}$$

$$A_4 = \begin{vmatrix} a_{21} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{vmatrix}. \tag{2.7.6}$$

Above, the first determinant is made up by the coefficients standing in the first $n-1$ equations of system (2.7.1) at its first $n-1$ variables from x_1 to x_{n-1}. Determinant (2.7.4) is determinant (2.7.3) in which the last column is substituted with the column of the coefficients standing in (2.7.1) at the variable x_n.

In a similar manner, determinant (2.7.5) is made up by the coefficients standing in system (2.7.1) starting from its second row, at the variables from x_2 to x_{n-1}, and determinant (2.7.6) differs from determinant (2.7.5) in that the last column in it is replaced with the column of the coefficients at the variable x_n.

It is seen from system (2.7.2) that the eigenvalues of (2.7.1) are contained among the roots of the polynomial

$$\Delta = A_1 A_4 - A_2 A_3. \tag{2.7.7}$$

We decompose determinants (2.7.3)–(2.7.6) by minors; then, it is easy to write out the terms with higher degrees of λ. We have:

$$\begin{aligned} A_1 &= (-1)^{n-1}\lambda^{n-1} + \dots, \\ A_2 &= (-1)^{n-2}a_{n-1,n}\lambda^{n-2} + \dots, \\ A_3 &= (-1)^{n-2}a_{n1}\lambda^{n-2} + \dots, \\ A_4 &= (-1)^{n-3}(a_{n-1,n}a_{n1} - a_{n-1,n}a_{nn})\lambda^{n-3} + \dots. \end{aligned} \tag{2.7.8}$$

On substitution of polynomials (2.7.8) into (2.7.7), we obtain

$$\Delta = -a_{n-1,n}a_{nn}\lambda^{2n-4} + (a_{n-1,n}a_{n1} - a_{n-1,n}a_{n1})\lambda^{2n-4} + \dots, \tag{2.7.9}$$

where the dots denote lower-degree terms.

To further proceed with the consideration, we retain the most interesting case of $a_{n-1,n}a_{nn} = 0$, i.e., the case in which one of the coefficients $a_{n-1,n}$ and a_{nn}, or both of these coefficients, are zero. Because of rounoff errors, the products $a_{n-1,n}a_{n,1}$ may differ from one another (indeed, the first of these products has come into polynomial (2.7.9) from the polynomials A_2 and A_3, and the second, from the polynomial A_4), the roundoff errors may have different impacts on these products and therefore the difference between the products may be not zero but a small number ε dependent on unpredicteable roundoff errors. This means that the eigenvalue problem for system (2.7.1) at any $n > 2$ may prove ill posed: with arbitrarily small roundoff errors, among the eigenvalues there may be a large extra-value, of order $1/\varepsilon$, whose sign will be unpredictable. There will be no such extra-value only if ε is zero exactly, $\varepsilon = 0$.

Note that, in this case, provided that all calculations are absolutely accurate, the measurement errors or variations of initial coefficients in (2.7.1) will not affect the eigenvalues.

Indeed, if the coefficient $a_{n-1,n}$ acquires some variation ε_1, and the coefficient a_{n1}, some variation ε_2, but there are no roundoff errors, then the difference $a_{n-1,n}(1+\varepsilon_1)a_{n,1}(1+\varepsilon_2) - a_{n-1,n}(1+\varepsilon_1)a_{n,1}(1+\varepsilon_2)$ in (2.7.9) will remain zero irrespective of ε_1 and ε_2. Here, the reason for the loss of well-posedness are roundoff errors.

Note that the classical eigenvalue problem for the matrix is now reduced to such a system of form (2.7.1), in which the parameter λ enters all the equations of the system, i.e., system (2.7.1) transforms into

$$\left.\begin{array}{c} (a_{11} - \lambda)x_1 + \dots\dots\dots + a_{1n}x_n = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{n1}x_1 + \dots\dots\dots + (a_{nn} - \lambda)x_n = 0. \end{array}\right\} \tag{2.7.10}$$

On passing from (2.7.1) to (2.7.10), the polynomials A_1, A_2, and A_3 retain their values, whereas the polynomial A_4 becomes

$$\bar{A}_4 = (-1)^{n-1} a_{n-1,1} \lambda^{n-2} + \ldots .$$

As a result, we obtain $\Delta = a_{n-1,n} \lambda^{2n-3} \ldots$, where dots stand for lower-degree terms; for this reason, small roundoff errors will no longer result in a loss of well-posedness. In manual calculations it is more convenient to start calculating eigenvalues from the equations that do not contain λ and, using them, eliminate part of variables; on the elimination, we arrive at the classical eigenvalue problem for a system with a lesser number of equations but, now, the parameter λ will enter each of the equations and, hence, subsequent eliminations of variables will not alter well-posedness.

In computations performed on programmable computing facilities, of primary significance is program unification; as a result, in a system similar to system (2.7.1) the computer may start eliminating the variables in the order of subscripts and, hence, it may happen to encounter a loss of well-posedness, the phenomenon not met in the epoch of manual calculus.

This circumstance points to the necessity of thorough inspection of computational algorithms when switching to computer-assisted calculations. Subtle things insignificant in manual calculations may give rise to serious errors in computer-assisted calculations.

Turning back to the problem of making distinction between transformations equivalent in the classical sense and transformations equivalent in the widened sense, we may say that classically equivalent transformations (termwise multiplications and additions, for instance) used in sequential elimination of variables in classical eigenvalue problems for matrices are also equivalent in the widened sense. The same transformations used in sequential elimination of variables in other mathematical models, for instance, in the generalized eigenvalue problem (in which some of equations do not contain λ) will no longer be equivalent in the widened sense; such transformations therefore may alter the well-posedness of the particular problem under consideration.

This once again underlines that, in an analysis of correctness and its possible alteration in the course of solution, a triad, i. e., the problem under consideration, the mathematical model, and the solution method, must be considered.

The problems that arise in various variants of the generalized eigenvalue problem were considered in more detail by Petrov and Petrov (1999). In the same monograph, the method of “degree matrices” is considered that makes it possible to identify third-class problems, i. e., problems whose correctness can be altered by transformations used to solve them.

If the reason for the alteration of correctness is known, it is not difficult to overcome it and avoid errors. What is dangerous is ignorance, or unexpected encounter with the unknown phenomenon. The following possibility must be therefore taken into account: in the course of sequential classically equivalent transformations of a mathematical model used in calculations, the correctness of the problem may alter. This being the case, even very small a roundoff error may lead to erroneous results.

Consider one more example that refers to calculation of functions with known Laplace transform (representation, image). Given the representation

$$\bar{f}(s) = \frac{s+b}{(s+a)(s+b)}, \tag{2.7.11}$$

it is required to calculate the original, i. e., the function $f(t)$. According to the well-known operational-calculus recommendations, we decompose representation (2.7.11) into simple fractions:

$$\frac{s+b}{(s+a)(s+b)} = \frac{A}{s+a} + \frac{B}{s+b} = \frac{As+Ab+Bs+Ba}{(s+a)(s+b)}. \tag{2.7.12}$$

The numerators here should be identical, and this gives us two equations from which two unknown numbers A and B can be found:

$$\begin{aligned} A + B &= 1, \\ Ab + Ba &= b. \end{aligned} \tag{2.7.13}$$

From these equations, we readily obtain that $B = 0$ and $A = 1$. Hence,

$$\bar{f}(s) = \frac{1}{s+a}, \qquad f(t) = \mathrm{e}^{-at}. \tag{2.7.14}$$

The calculations can be facilitated by reducing both the numerator and the denominator in (2.7.11) by $s + b$ (or, which is the same, by multiplying both of them by the nonzero number $1/(s+b)$). This readily yields (2.7.14). The decomposition into simple fractions (2.7.12) has once again confirmed that the reduction of both the numerator and the denominator in the Laplace representation by equal multipliers is a transformation equivalent in the classical sense. Is this transformation equivalent in the widened sense? Prior to answering this question, let us refine the meaning of the statement "both the numerator and the denominator of (2.7.11) contain equal multipliers $s + b$". The numbers b entering these multipliers were obtained (in the final analysis) from an experiment or from a measurement, and only several first decimal digits of b are known with absolute confidence. Let the value of b be measured with a good measuring instrument accurate to the fifth decimal digit, and let all four decimal digits both in the numerator and in the denominator be identical. Despite the fact that all the four digits are identical, we can only state that representation (2.7.11) is equal to

$$\bar{f}(s) = \frac{s + b_0 + \varepsilon_1}{(s+a)(s + b_0 + \varepsilon_2)}, \tag{2.7.15}$$

where ε_1 and ε_2 are some unknown numbers as to which we only know that $|\varepsilon_1/b_0| < 10^{-4}$ and $|\varepsilon_2/b_0| < 10^{-4}$. Decomposing representation (2.7.15) into simple fractions

$$\frac{s + b_0 + \varepsilon_1}{(s+a)(s + b_0 + \varepsilon_1)} = \frac{A}{s+a} + \frac{B}{s + b_0 + \varepsilon_1} \tag{2.7.16}$$

and calculating A and B with allowance for the unknowns ε_1 and ε_2, we see that

$$A = \frac{b_0 - a + \varepsilon_1}{b_0 - a + \varepsilon_2}, \qquad B = \frac{\varepsilon_2 - \varepsilon_1}{b_0 - a + \varepsilon_2}, \tag{2.7.17}$$

i. e., as a matter of fact, the number B is not necessarily zero and the true original $f_1(t)$ may be not $f(t) = \mathrm{e}^{-at}$ but

$$f_1(t) = \mathrm{e}^{-at} + \frac{\varepsilon_2 - \varepsilon_1}{b_0 - a + \varepsilon_2}\mathrm{e}^{-at} + \frac{\varepsilon_2 - \varepsilon_1}{b_0 - a + \varepsilon_2}\mathrm{e}^{-(b_0+\varepsilon_2)t}. \tag{2.7.18}$$

Let, for instance, $a = 1$, $b_0 = -2$, and $|\varepsilon_2 - \varepsilon_1| \le 10^{-6}$; then, at $t = 1$ the function $f_1(t)$ and $f(t)$ will differ within 10^{-5}, but already with $t = 20$ the functions $f_1(t)$ and $f(t)$ may happen to differ by many times. If, alternatively, $a > 0$ and $b > 0$, then the difference between $f_1(t)$ and $f(t)$ remains small for all values of t.

The reduction of the numerator and denominator of Laplace representations gives an example of a transformation equivalent in the classical but not in the widened sense. Reduction by equal multipliers in representations is often used to facilitate calculations; this transformation, however, may lead to erroneous results. Guides on control theory abound in warnings that care must be taken in performing reduction of identical multipliers in Laplace transformations of transfer functions. Again, we see that these warnings are quite rational, and now we are fully aware of the true reason for them.

2.8. CONCLUSIONS FROM CHAPTERS 1 AND 2

In Chapters 1 and 2, we considered some of ill-posed problems that often fall beyond the field of attention of researchers. Very often researchers deal with much more complex ill-posed problems that require solution of partial differential equations, integral equations, etc. The complexity of such problems often hampers unveiling and identification of fundamental points.

In Chapters 1 and 2, we saw that many simpler ill-posed problems arise in diverse fields of applications and, to avoid errors, an engineer must be familiar at least with the simplest properties of ill-posed problems and approaches to their solution.

It is important that ill-posed problems present a particular, limiting case of a certain broader but less distinctly defined class of ill-conditioned problems, problems in which substantial solution changes originate from small (but finite) variations of coefficients, parameters, initial or boundary conditions, etc. In Chapters 1 and 2, we saw that (in contrast to ill-conditioned problems) rather simple identification means can be proposed for ill-posed problems, and this circumstance greatly facilitates solution of these problems. Around ill-posed problems, ill-conditioned problems, more difficult to identify, lie. After isolation of ill-posed problems, ill-conditioned problems are simpler to handle.

Another distinctive feature of our approach to the matter under discussion is isolation, in the special, third class, of problems that cannot be classed to well- or to ill-posed problems since such problems may suffer alteration of their correctness under equivalent transformations used to treat them.

Of most significance for all researchers and engineers performing calculations is the recent discovery of systems of differential equations showing no continuous dependence of their solutions on coefficients and parameters. Until recently, the majority of engineers and scientific workers, when using solutions of differential equations in their calculations, took no care to observe if the solutions were correct, converting the equations to the normal Cauchy form in order to subsequently use standard computer programs and believing that reliability of such calculations was guaranteed by the theorem about continuous dependence of solutions on parameters. Now we see that this may be not the case and there are situations where the use of traditional research and calculation methods results in miscalculations leading to serious accidents. In order to avoid errors, simple additional checks described in Chapter 2 (and, in more detail, in the third edition of the monograph by Petrov and Petrov (1999)) must be applied.

Chapter 2 shows that the assignment of certain problems to the third class of mathematical, physical and engineering problems and the investigation into this class of problems makes it possible to avoid serious miscalculations.

Special cases, i. e., special systems of differential equations application of ordinary, traditional solution methods to which may result in serious miscalculations, are encountered not very frequently; that is why the existence of such cases has escaped notice until recently.

Yet, every unexpected encounter with such special systems may result in miscalculations and serious accidents in control systems (which was many times the case, see Petrov and Petrov (1999), pp. 21–23, 107–108). That is why such cases and such systems must be paid very concentrated attention, the more so that such special systems are rather easy to identify.

The same applies to many computations and computational algorithms that use chains of equivalent transformations. Here again, an unexpected encounter with problems belonging to the third class of problems, problems intermediate between well- and ill-posed ones, may result (and do result!) in serious miscalculations and accidents. The examples were presented above. That is why the third-class problems must be paid very concentrated attention.

Finally, we would like to note that the investigation into third-class problems, the investigation into transformations of mathematical models equivalent not only in the classical but also in the widened sense, and the investigation into transformations that alter correctness of posed problems, — all these investigations are still at their initial stage and, in this research field, much is still to be done.

Chapter 3.

Change of sensitivity to measurement errors under integral transformations used in modeling of ships and marine control systems

3.1. APPLICATION OF INTEGRAL TRANSFORMATIONS TO PRACTICAL PROBLEMS

In Chapters 1 and 2, the impact of measurement errors on the solution accuracy of various engineering problems was considered. Primary attention was paid to the important case of ill-posed problems, in which even arbitrarily small errors drastically affected solutions. We saw that there exists a class of problems intermediate between well- and ill-posed problems, whose correctness and, hence, sensitivity to errors may alter under equivalent transformations used in treating the problems. Such problems are especially hard to solve, and an unexpected encounter with such a problem may result in serious miscalculations.

In the present chapter, we will consider the change of sensitivity to measurement errors under integral transformations. Particular attention will be paid to the well-known Fourier cosine transformation,

$$S_x(\omega) = \frac{2}{\pi} \int_0^\infty K_x(\tau)\, \cos\omega\tau\, \mathrm{d}\tau, \tag{3.1.1}$$

which transforms a function of time $K(\tau)$ into a function of frequency $S(\omega)$. Transformation (3.1.1) is widely, although not widely enough, used in synthesis of marine control systems. In modeling of ships and marine systems, engineers cannot do without this transformation.

Indeed, disturbing actions exerted on ships and marine systems originate from wind, sea roughness, and vibrations, and the magnitude of these factors can never be reliably predicted.

Consider a simplest problem, namely, that of predicting ship roll for a ship oriented, as seamen say, “with the log to the wave”, i. e., the wave crests in the situation of interest are parallel to the center plane of the ship. The rolling equation has the form

$$(T_1^2D^2 + T_2D + 1)\theta = \varphi(t), \tag{3.1.2}$$

where T_1 and T_2 are some time constants, measured in seconds; $D = \mathrm{d}/\mathrm{d}t$ is the differentiation operator; θ is the roll angle, measured in degrees; and $\varphi(t)$ is the disturbing action, or the wave sloping angle, also measured in degrees. The time constant T_1 depends on the moment of inertia of the ship hull with respect to the longitudinal axis, and the constant T_2 reflects the damping action of water.

The constants T_1 and T_2 can be determined experimentally by recording natural oscillations of the ship on calm sea, i. e., at $\varphi(t) = 0$. If the initial (at $\varphi(t) = 0$) roll angle of the ship is θ_0, then further time evolution of the angle will be

$$\begin{aligned}\theta(t) = \mathrm{e}^{-(T_2/(2T_1^2))t}\Big[&c_1\cos\Big(\sqrt{1-\Big(\frac{T_2}{4T_1}\Big)^2}\,\frac{t}{T_1}\Big)\\ &+ c_2\sin\Big(\sqrt{1-\Big(\frac{T_2}{4T_1}\Big)^2}\,\frac{t}{T_1}\Big)\Big],\end{aligned} \tag{3.1.3}$$

where c_1 and c_2 are integration constants to be determined from initial conditions. Provided that $\theta = \theta_0$ and $\dot{\theta} = 0$ at $t = 0$, we have $c_1 = \theta_0$ and $c_2 = 0$. Since, usually, $T_2 \ll T_1$, then $\sqrt{1 - (T_2/(4T_1))^2} \approx 1$, and the natural frequency of the ship will be close to $1/T_1$. The ratio $T_2/(2T_1)$ characterizes the damping rate of natural oscillations.

Given the disturbing action $\varphi(t)$, equation (2), not difficult to solve, yields the function $\theta(t)$. If, for instance, the disturbing action is a harmonic function of time, $\varphi(t) = A\sin\beta t$, then, following the rapid decay of free oscillations, only forced oscillations will remain, representing a harmonic function of frequency β, whose amplitude, which depends on β and ship characteristics, and also on the constants T_1 and T_2, is given by the simple formula

$$\theta(t) = \frac{A}{\sqrt{(1 - T_1^2\beta^2)^2 + T_2^2\beta^2}} \sin(\beta t + \psi_0). \tag{3.1.4}$$

Yet, actual disturbing actions experienced by ships and marine systems are not harmonic oscillations: instead, they are random, poorly predictable functions. In this situation, direct solution of (3.1.2) gives nothing: the solution $\theta(t)$ will also be a random, poorly predictable function of no use for the modeling (a typical example of the roll angle θ of a ship on rough sea versus time is given in Figure 3.8 below). By solving the differential equation (3.1.2) with random right-hand side $\varphi(t)$, we do not solve the basic problem of calculating the ship roll: we cannot predict whether the ship will sink, or whether the roll angle will reach a limiting value for ship sinking.

In such a situation, the very approach to the problem needs to be modified; one should search for the mean square of the solution:

$$\langle\theta^2\rangle = \lim_{T\to\infty} \int_0^T \theta^2 \, \mathrm{d}t, \tag{3.1.5}$$

and for the root-mean-square value of the solution:

$$\sigma_\theta = \sqrt{\langle\theta^2\rangle}, \tag{3.1.6}$$

rather than for the solution $\theta(t)$ itself.

Since the distribution of disturbing actions almost always obeys the normal law, then, according to the well-known rule of “three root-mean-squares”, the maximum magnitude of $\theta(t)$ is thrice its root-mean-square value.

If, for instance, the critical roll angle for a ship, leading to its turning upside down, is $\theta_{cr} = 40^\circ$, and $\sigma_\theta = 12^\circ$, then we can be sure that $\theta_{\max} \le 3 \cdot 12^\circ = 36^\circ$, and the ship will not sink.

From here, there arises a question which for a long time, up to the second half of the XX century, remained unanswered: which characteristics of random disturbing actions need to be known in order the mean squares of the solutions of linear differential equations (and, in particular, equation (3.1.2)) with right-hand sides containing random functions (stochastic processes) $\varphi(t)$, could be found? The required characteristic is the spectral power density (or, briefly, the spectrum) of the random function $\varphi(t)$.

There holds the following important relation (a simplistic derivation of this relation will be given below): provided that the variables $x(t)$ and $\varphi(t)$

are related by a differential equation

$$A(D)x = B(D)\varphi(t), \tag{3.1.7}$$

where $A(D)$ and $B(D)$ are arbitrary-degree polynomials of the differentiation operator, i. e., provided that

$$A(D) = a_n D^n + a_{n-1} D^{n-1} + \ldots + a_0, \tag{3.1.8}$$

$$B(D) = b_m D^m + b_{m-1} D^{m-1} + \ldots + b_0, \tag{3.1.9}$$

then the spectra of $x(t)$ and $\varphi(t)$ (here and below, all spectra will be denoted with the subscribed character S, e. g., S_φ, S_x, etc.) are interrelated by the simple formula

$$S_x = |B(j\omega)/A(j\omega)|^2 S_\varphi. \tag{3.1.10}$$

Note that the mean square of a variable can be calculated from the spectrum of this variable:

$$\langle x^2 \rangle = \int_0^\infty S_x(\omega)\,\mathrm{d}\omega; \tag{3.1.11}$$

hence, we have a simple rule for calculating the mean squares of solutions of (3.1.7): first, the differentiation operator D in (3.1.8) and (3.1.9) is to be replaced with the number $j\omega$; second, the function

$$B(j\omega)/A(j\omega) \tag{3.1.12}$$

is to be calculated; third, the square of module of this function is to be multiplied by S_φ and, finally, integral (3.1.11) is to be calculated. Thus, provided that the spectrum S_φ is known, then subsequent calculations present no difficulty; as it is well known, the spectrum S_φ itself is the Fourier cosine transform of the correlation function $K_\varphi(\tau)$ of the process $\varphi(t)$:

$$S_\varphi(\omega) = \frac{2}{\pi}\int_0^\infty K_\varphi(\tau)\cos\omega\tau\,\mathrm{d}\tau. \tag{3.1.13}$$

Thus, the spectrum S_φ is a function of the variable ω (frequency), having dimension of 1/time, which can be calculated by transforming the correlation function by formula (3.1.13).

In turn, the correlation function can be calculated based on its definition

$$K_\varphi(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\varphi(t)\varphi(t+\tau)\,\mathrm{d}t. \tag{3.1.14}$$

Formula (3.1.14) shows that, for the function $K_\varphi(\tau)$ to be calculated, it suffices to multiply the values of $\varphi(t)$ separated by a time interval τ and, then, perform averaging of the products obtained.

Formulas (3.1.10), (3.1.11), (3.1.13), and (3.1.14) form the basis for modeling ships and marine systems whose disturbing action are stochastic processes originating from wind and sea roughness, for modeling aviation systems acted upon by random factors brought about by atmospheric turbulence, and for synthesizing many other systems; i. e, speaking generally, for modeling all systems acted upon by random factors.

Of course, integral (3.1.14) is actually calculated not over an infinitely broad interval $-T \le t \le T$, where $T \to \infty$, but over a finite interval, and the values of $\varphi(t)$ are measured with inevitable errors; both factors give rise to small deviations of the true correlation function from the correlation function used in (3.1.13).

There arises a question: to which extent small uncertainties in the correlation function affect the accuracy of the integral transform, i. e., the spectrum S_φ? To run a few steps forward, we say: small, and even arbitrarily small uncertainties in $K_\varphi(\tau)$ may strongly affect the spectrum $S_\varphi(\omega)$. Reconstruction of the spectrum $S_\varphi(\omega)$ from the known function $K_\varphi(\tau)$ by formula (3.1.13) is generally an ill-posed problem. For this problem to be successfully solved, additional information concerning the properties of correlation functions and spectra is required in addition to formulas (3.1.13) and (3.1.14), given in manuals. Below, the properties of correlation functions and spectra will be analyzed, required for the ill-posed problem on finding the disturbing-action spectrum to be successfully solved based on integral transformation (3.1.13). Most frequently occurring errors will also be considered.

3.2. PROPERTIES OF CORRELATION FUNCTIONS

First of all, note that the correlation function is calculated for the processes $\varphi(t)$ whose mean value, i. e., the integral

$$\langle \varphi(t) \rangle = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \varphi(t)\, \mathrm{d}t, \tag{3.2.1}$$

is zero. If $\langle \varphi \rangle \ne 0$, then the correlation function is to be calculated for the deviation of the process $\varphi(t)$ from its mean value, i. e., for the process $\varphi_1(t) = \varphi(t) - \langle \varphi \rangle$.

The properties of the correlation function that immediately follow from formula (3.1.14) are as follows:

1. The magnitude of $K_\varphi(t = 0)$, i. e., the magnitude of the correlation function at $\tau = 0$ is the mean square of the process $\varphi(t)$, i. e.,

$$K_\varphi(0) = \langle\varphi^2\rangle. \tag{3.2.2}$$

That is why, instead of correlation function (3.1.14) they often consider the normalized correlation function k_φ (to discriminate between this function and the ordinary correlation function, they denote the normalized function with the small letter) equal to the quotient of $K_\varphi(\tau)$ and $K_\varphi(0)$:

$$k_\varphi(\tau) = K_\varphi(\tau)/K_\varphi(0). \tag{3.2.3}$$

The normalized correlation function at $\tau = 0$ is always unity. The ordinary correlation function is equal to the normalized function multiplied by the mean square, i. e., $K_\varphi(\tau) = \langle\varphi^2\rangle k_\varphi(\tau)$.

2. The correlation function is an even function, i. e.,

$$K_\varphi(\tau) = K_\varphi(-\tau). \tag{3.2.4}$$

Indeed,

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \varphi(t)\varphi(t+\tau)\,\mathrm{d}t = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \varphi(t-\tau)\varphi(t)\,\mathrm{d}t,$$

which yields formula (3.2.4).

In view of (3.2.4), they often write the correlation function only for positive arguments, i. e., for $\tau \geq 0$. It should always be kept in mind that, in fact, this function is also defined for $\tau < 0$: it extends to the range of $\tau < 0$ symmetrically, in line with formula (3.2.4).

3. The obvious inequality

$$[\varphi(t) - \varphi(t+\tau)]^2 \geq 0,$$

yields

$$\varphi^2(t) + \varphi^2(t+\tau) \geq 2\varphi(t)\varphi(t+\tau). \tag{3.2.5}$$

We perform time averaging of both sides of (3.2.5) and obtain that

$$K_\varphi(0) \geq K_\varphi(\tau). \tag{3.2.6}$$

Thus, the magnitude of the correlation function at any τ cannot exceed the magnitude of this function at $\tau = 0$.

4. A fourth important property is the following formula for the correlation function of the derivative $\dot{\varphi}(t)$ of the process $\varphi(t)$:

$$K_{\dot{\varphi}}(\tau) = -\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\, K_{\varphi}(\tau). \tag{3.2.7}$$

This formula can be obtained from the basic formula (3.1.14) by differentiation under the integration sign with subsequent integration by parts. The correlation function of the derivative $\dot{\varphi}$ is equal to the second derivative (taken with the opposite sign) of the ordinary correlation function $K_{\varphi}(\tau)$ of the process $\varphi(t)$.

Examples. a) Suppose that $\varphi(t) = A\sin(\beta t + \psi_0)$; then, in the formula

$$K_{\varphi}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} A^2 \sin(\beta t + \psi_0)\sin(\beta t + \beta\tau + \psi_0)\,\mathrm{d}t$$

we substitute, according to the well-known trigonometric formulas, the product of the sines with the cosine of the half-difference and that of the half-sum. We take the fact into account that the integral of the half-sum of the cosines vanishes as $\tau \to \infty$, and obtain that

$$K_{\varphi}(\tau) = (A^2/2)\cos(\beta\tau). \tag{3.2.8}$$

We see that the correlation function of a harmonic oscillation $A\sin(\beta t + \psi_0)$ depends on its amplitude and frequency and does not depend on the phase.

b) If the process $\varphi(t)$ is a periodic one that can be expanded into a Fourier series,

$$\varphi(t) = \sum_{n=1}^{\infty} A_n \sin(n\beta t + \psi_n), \tag{3.2.9}$$

then, in view of the fact that the integrals of products of harmonics with different subscript numbers vanish, we use the same trigonometric transformation as in the derivation of (3.2.8) and obtain:

$$K_{\varphi}(\tau) = \frac{1}{2}\sum_{n=1}^{\infty} A_n^2 \cos(n\beta\tau). \tag{3.2.10}$$

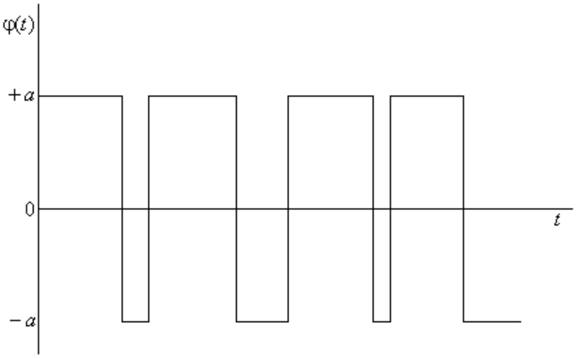

Figure 3.1

Formula (3.2.10) shows that the correlation function does not depend on the phase shift between individual harmonics of the periodic process. That is why various processes whose individual harmonics have identical amplitudes but different phases have identical correlation functions. The latter implies that the correlation function is not a unique characteristic of the process $\varphi(t)$. Different processes may have identical correlation functions.

c) Consider now a stochastic process in which the quantity $\varphi(t)$ randomly but with equal probabilities assumes the values $+a$ and $-a$ (see Figure 3.1).

The function $\varphi(t)$ looks as the output voltage of a telegraph apparatus; that is why this process is called "telegraph signal". The mean value of the "telegraph signal" is zero. Find now the correlation function. The probability of sign inversion obeys the Poisson law with a parameter μ, i. e., the probability of the situation that, during a time τ, no sign inversion will occur, is $P_0 = \mathrm{e}^{-\mu\tau}$, and the probability of the situation that, for the same time, there will be precisely n sign inversions, is

$$P_n = \frac{(\mu\tau)^n}{n!}\,\mathrm{e}^{-\mu|\tau|}.$$

The product $\varphi(t)\varphi(t+\tau)$ that enters the correlation function is either a^2 or $-a^2$, depending on whether the number of sign inversions of the function $\varphi(t)$ during the time τ is even or odd. But this means that

$$K_\varphi(\tau) = a^2(P_{\text{even}} - P_{\text{odd}}),$$

where P_{even} and P_{odd} are the probability of the situations in which the number of sign inversions during the time τ is an even or odd number, respectively.

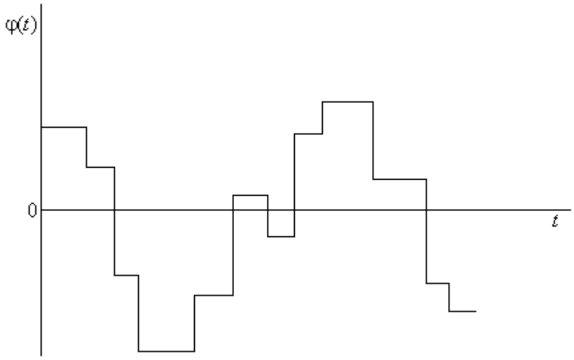

Figure 3.2

Since

$$P_{\text{even}} = \Big[\sum_{k=0}^{\infty} \frac{(\mu\tau)^{2k}}{(2k)!}\Big] \mathrm{e}^{-\mu|\tau|}, \qquad P_{\text{odd}} = \Big[\sum_{k=0}^{\infty} \frac{(\mu\tau)^{2k+1}}{(2k+1)!}\Big] \mathrm{e}^{-\mu|\tau|},$$

then, summing the series, we obtain:

$$K_\varphi(\tau) = a^2 \mathrm{e}^{-2\mu|\tau|}, \tag{3.2.11}$$

i. e., the correlation function decays exponentially, vanishing with increasing τ. The greater is μ, i. e., the more frequently the function $\varphi(t)$ changes its sign, the more rapidly the correlation function decays.

d) Consider now a piecewise constant process that proceeds by coming from one value φ_i to another at random times and in which the values φ_i obey the Poisson law with a parameter μ; in this case, the probability of the situation in which the function $\varphi(t)$ remains constant over the time interval τ is $\mathrm{e}^{-\mu\tau}$. This process is illustrated by Figure 3.2.

The product $\varphi(t)\varphi(t+\tau)$ will assume different values depending on whether or not during the time τ the process $\varphi(t)$ passes from one value, φ_k, to another value, φ_i.

In the first case

$$\varphi(t)\,\varphi(t+\tau) = \varphi_k\varphi_k = \varphi_k^2,$$

whereas in the second case

$$\varphi(t)\varphi(t+\tau) = \varphi_k\varphi_i.$$

The probability of the first case is $\mathrm{e}^{-\mu\tau}$, and the probability of the second case is $1 - \mathrm{e}^{-\mu\tau}$. Hence,

$$K_\varphi(\tau) = \langle\varphi_k\rangle \mathrm{e}^{-\mu\tau} + \langle\varphi_k\varphi_i\rangle(1 - \mathrm{e}^{-\mu|\tau|}).$$

Yet, if φ_k and φ_i are independent and obey one and the same distribution law, then $\langle \varphi_k \varphi_i \rangle = 0$ and, finally,

$$K_\varphi(\tau) = \langle \varphi^2 \rangle \mathrm{e}^{-\mu|\tau|}. \tag{3.2.12}$$

Here again, we see that rather dissimilar processes have similar correlation functions. This once again confirms the fact that different processes may have similar and even identical correlation functions. Processes that have identical correlation functions, function (3.2.12), for instance, are called "realizations" or, more precisely, realization of the process with a given correlation function.

3.3. PROPERTIES OF SPECTRA

The spectral power density of a process $\varphi(t)$, for short called the spectrum, is the integral Fourier transform of its correlation function, known from the course of mathematical analysis. In analysis, the integral Fourier transformation is the transformation

$$\int_{-\infty}^{+\infty} f(t)\mathrm{e}^{-j\omega t}\,\mathrm{d}t = \int_{-\infty}^{\infty} f(t)(\cos(\omega t) - j\sin(\omega t))\,\mathrm{d}t = F(j\omega), \tag{3.3.1}$$

which transforms a function of time $f(t)$ into a generally complex-valued function $F(j\omega)$ of frequency ω. Since the correlation function is an even function, we are interested in the Fourier transformation of even functions for which

$$\int_{-\infty}^{+\infty} f(t)\sin(\omega t)\,\mathrm{d}t = 0$$

and, therefore, Fourier transformation (3.3.1) turns into the cosine transformation

$$\int_{-\infty}^{+\infty} f(t)\cos(\omega t)\,\mathrm{d}t = F_c(\omega), \tag{3.3.2}$$

which transforms an even function of time $f(t)$ into a real-valued (and also even) function of frequency $F_c(\omega)$. As it is known from the course of analysis, the inverse Fourier transformation

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} F_c(\omega)\cos(\omega t)\,\mathrm{d}\omega = f(t) \tag{3.3.3}$$

restores the initial function of time $f(t)$. Taking the evenness of the functions $f(t)$ and $F_c(\omega)$ into account, one can substitute integration from $-\infty$

to $+\infty$ with integration over a narrower interval, from 0 to $+\infty$; this yields the following pair of formulas for the direct and inverse cosine Fourier transformations:

$$F_c(\omega) = \int_0^{+\infty} f(t)\cos(\omega t)\,\mathrm{d}t, \tag{3.3.4}$$

$$f(t) = \frac{2}{\pi}\int_0^{+\infty} F_c(\omega)\cos(\omega t)\,\mathrm{d}\omega. \tag{3.3.5}$$

Inverse cosine transformation (3.3.5) differs from direct cosine transformation (3.3.4) by the factor $2/\pi$. Yet, this factor may be conventionally included in the direct cosine transformation so that to be eliminated from the inverse transformation. It is generally agreed to use the latter form of the cosine transformation of correlation functions in modeling ships and marine control systems. The following important formulas hold:

$$S_\varphi(\omega) = \frac{2}{\pi}\int_0^{+\infty} K_\varphi(\tau)\cos(\omega\tau)\,\mathrm{d}\tau, \tag{3.3.6}$$

$$K_\varphi(\tau) = \int_0^{+\infty} S_\varphi(\omega)\cos(\omega\tau)\,\mathrm{d}\omega. \tag{3.3.7}$$

The function $S_\varphi(\omega)$, defined by (3.3.6), is the "spectral power density" or, for short, the "spectrum" of the process $\varphi(t)$.

Example: for the correlation function $K_\varphi = \mathrm{e}^{-\alpha\tau}$, we calculate the integral

$$\frac{2}{\pi}\int_0^{\infty} \mathrm{e}^{-\alpha\tau}\cos(\omega\tau)\,\mathrm{d}\tau = \frac{2}{\pi}\,\frac{\alpha}{\alpha^2+\omega^2}, \tag{3.3.8}$$

and see that, in this case, the spectrum is $S_\varphi(\omega) = \dfrac{2}{\pi}\,\dfrac{\alpha}{\alpha^2+\omega^2}$.

Calculation of the integral

$$\int_0^{\infty} \frac{2}{\pi}\,\frac{\alpha}{\alpha^2+\omega^2}\cos(\omega\tau)\,\mathrm{d}\omega = \mathrm{e}^{-\alpha\tau}, \tag{3.3.9}$$

shows that inverse transformation (3.3.7) indeed restores the initial correlation function.

Of interest is the calculation of the spectrum of the correlation function $K_\varphi = \mathrm{e}^{-\alpha\tau}$ in the limiting case of $\alpha \to 0$, when $K_\varphi(\tau) \to 1$ and the process $\varphi(t)$ also transforms in a constant, $\varphi(t) = 1$.

The limit of spectrum (3.3.8) for $\alpha \to 0$ yields very peculiar a function that is zero at all $\omega \neq 0$ and infinity at $\omega = 0$.

Simultaneously, since formula (3.3.9) is valid for all α, including $\alpha \to 0$, then the integral of this function, which assumes a single nonzero value, equals unity.

This peculiar function, denoted as $\delta(\omega)$, is called the *Dirac δ-function.* Thus, the spectrum of a constant quantity $\varphi(t) = A$ is the delta-function

$$S_{\varphi=A} = A^2\delta(\omega).$$

Delta-functions, first introduced by Dirac, the English physicist in 1928, are widely used in applications. These functions can be considered as limits of ordinary functions. For instance, the function $\varphi(t)$ that is zero in the interval $-\infty < t < 0$ and $\varphi(t) = \alpha \mathrm{e}^{-\alpha\tau}$ in the interval $0 \le t < \infty$ turns in the limit of $\alpha \to 0$ into the delta-function: $\lim_{\alpha\to 0} \alpha \mathrm{e}^{-\alpha t} = \delta(t)$.

Consider the main properties of the function $S_\varphi(\omega)$ defined by formulas (3.3.6) and (3.3.7).

1. Formula (3.3.7) is valid for all τ, including $\tau = 0$. Substitution of $\tau = 0$ into (3.3.7) yields

$$K_\varphi(0) = \int_0^\infty S_\varphi(\omega)\,\mathrm{d}\omega. \tag{3.3.10}$$

From here, in view of (3.2.2), we have

$$\langle\varphi^2\rangle = \int_0^\infty S_\varphi(\omega)\,\mathrm{d}\omega. \tag{3.3.11}$$

Thus, the integral of $S_\varphi(\omega)$ gives the mean square of the process $\varphi(t)$. If we take into account the fact that the mean square is often called the "power" of the process $\varphi(t)$, the meaning of the full name of the function $S_\varphi(\omega)$, the "spectral power density" of the process $\varphi(t)$, now becomes clear. This function shows the distribution of power over frequencies. In short, the function $S_\varphi(\omega)$ is normally called just "spectrum".

2. From basic formula (3.3.6), it immediately follows that if $K_\varphi(\tau) = K_1(\tau)+K_2(\tau)$, then $S_\varphi(\omega) = S_1(\omega)+S_2(\omega)$, i. e., the spectrum of the sum of two correlation functions is equal to the sum of the spectra of the addends.

3. It follows from the same formula that the transformation is unambiguous. To each correlation function, its own spectrum corresponds. With the spectrum known, we can calculate the correlation function, and vice versa.

4. As it is known from mathematical analysis, for the general Fourier transformation (3.3.1) the following formula holds:

$$\int_{-\infty}^{+\infty} \frac{\mathrm{d}f}{\mathrm{d}t}\,\mathrm{e}^{-j\omega t}\,\mathrm{d}t = j\omega F(j\omega), \tag{3.3.12}$$

i. e., the transform of the derivative $\mathrm{d}f/\mathrm{d}t$ is equal to the Fourier transform of the function $f(t)$ multiplied by the number $j\omega$. In view of (3.2.7), for the derivative of the correlation function $\dot{\varphi}(t)$ of a process $\varphi(t)$ we obtain:

$$S_{\dot{\varphi}}(\omega) = |j\omega|^2 S_\varphi(\omega) = \omega^2 S_\varphi(\omega). \tag{3.3.13}$$

Thus, the spectrum of the derivative $\dot{\varphi}(t)$ is equal to the spectrum of the process $\varphi(t)$ multiplied by the square of the modulus of the number $j\omega$ or, which is the same, by the number ω^2.

Using formula (3.3.13), we can find the relation between the spectrum $S_x(\omega)$ of the solution $x(t)$ of (3.1.7) and the spectrum $S_\varphi(\omega)$ of the right-hand side of this differential equation. This relation is given by the already presented formula (3.1.10). How, the derivation of this formula is also given.

With formula (3.3.11) taken into account, we can also find the mean square of the solution $x(t)$ of (3.1.7):

$$\langle x^2 \rangle = \int_0^\infty \left| \frac{B(j\omega)}{A(j\omega)} \right|^2 S_\varphi(\omega)\,\mathrm{d}\omega. \tag{3.3.14}$$

Formula (3.3.14) opens up a simple way to calculating the mean squares for all those various engineering systems whose disturbing actions are stochastic, incompletely predictable processes. For this to be done, it suffices, using the measured current values of the disturbing actions $\varphi(t)$ and formula (3.1.14), to calculate the correlation function $K_\varphi(\tau)$ and to convert this function into the spectrum $S_\varphi(\omega)$ by formula (3.3.6), and then it suffices for any differential equation (3.1.7) to replace the differentiation operator $D = \mathrm{d}/\mathrm{d}t$ in $A(D)$ and $B(D)$ with the number $j\omega$, find the squares of the moduli $|B(j\omega)|^2$ and $|A(j\omega)|^2$, and calculate integral (3.3.14).

The spectra of disturbing actions and formula (3.3.14) presently form the basis for the synthesis of all marine engineering objects (because the disturbing actions exerted on ships and marine systems by wind and rough sea are stochastic, incompletely predictable processes) and aviation systems (because the disturbing actions brought about by atmospheric turbulence are also stochastic processes), and many other systems and engineering apparatus.

Nonetheless, many important aspects of integral transformations have not been adequately covered in the literature. The next section discusses the important matter of correctness of transformations in more detail.

3.4. CORRECTNESS OF INTEGRAL TRANSFORMATIONS

In calculating the correlation function of a process $\varphi(t)$ by formula (3.1.14), we must take into account the fact that inevitable small errors in the measured values of $\varphi(t)$ and $\varphi(t+\tau)$, and also the replacement of the theoretically infinite interval of integration in formula (3.1.14) with a real, finite interval, results in inevitable miscalculations of the correlation function. There arises a question, to which extent these miscalculations will affect the final result, i. e., the calculated mean square of the solution $x(t)$ of the differential equation

$$A(D)x = B(D)\varphi. \tag{3.4.1}$$

Does calculation of the mean square of the solution through the integral Fourier transformation of the process $\varphi(t)$ presents a well-posed problem?

By way of example, consider the simplest equation

$$x = D\varphi \tag{3.4.2}$$

for the case of $K_\varphi(\tau) = \mathrm{e}^{-\alpha|\tau|}$. Since, here,

$$S_\varphi(\omega) = \frac{2}{\pi}\int_0^\infty \mathrm{e}^{-\alpha\tau}\cos(\omega\tau)\,\mathrm{d}\tau = \frac{2}{\pi}\,\frac{\alpha}{\alpha^2+\omega^2}, \tag{3.4.3}$$

then the spectrum of $x(t)$ is

$$S_x(\omega) = \frac{2}{\pi}\,\frac{\alpha\omega^2}{\alpha^2+\omega^2}. \tag{3.4.4}$$

We calculate the mean square of $x(t)$ by formula (3.3.11) and see that, here, the mean square is infinite since the improper integral

$$\langle x^2\rangle = \int_0^\infty \frac{2}{\pi}\,\frac{\alpha\omega^2}{\alpha^2+\omega^2}\,\mathrm{d}\omega$$

diverges.

Consider now the same problem on calculation of the mean square of $x(t)$ for the process $\varphi(t)$ with the correlation function

$$K_\varphi(\tau) = \mathrm{e}^{-\alpha|\tau|} + \alpha\tau\mathrm{e}^{-\beta|\tau|}. \tag{3.4.5}$$

By calculating the integral

$$\begin{aligned} S_\varphi(\omega) &= \frac{2}{\pi}\int_0^\infty (\mathrm{e}^{-\alpha\tau} + \alpha\tau\mathrm{e}^{-\beta\tau})\cos(\omega\tau)\,\mathrm{d}\tau \\ &= \frac{2\alpha}{\pi}\,\frac{\beta^4+\alpha^2\beta^2+(3\beta^2-\alpha^2)\omega^2}{(\alpha^2+\omega^2)(\beta^2+\omega^2)^2}, \end{aligned} \tag{3.4.6}$$

we can also calculate the mean square of $x(t)$ by formula

$$\langle x^2 \rangle = \int_0^\infty \frac{2\alpha\omega^2}{\pi} \frac{\beta^4 + \alpha^2\beta^2 + (3\beta^2 - \alpha^2)\omega^2}{(\alpha^2 + \omega^2)(\beta^2 + \omega^2)^2} \, \mathrm{d}\omega. \tag{3.4.7}$$

We notice at once that integral (3.4.7) is finite for all β, because with increasing frequency ω the integrand vanishes as $1/\omega^2$, and such integrals converge.

Simultaneously, at small values of α/β (i. e., at high β) the correlation function (3.4.5) differs as little as one likes from $K_\varphi(\tau) = \mathrm{e}^{-\alpha|\tau|}$.

Indeed, the maximum difference of these correlation functions,

$$(\mathrm{e}^{-\alpha|\tau|} + \alpha\tau\mathrm{e}^{-\beta|\tau|}) - \mathrm{e}^{-\alpha\tau} = \alpha\tau\mathrm{e}^{-\beta\tau}, \tag{3.4.8}$$

is attained at $\tau = 1/\beta$ to be

$$\Delta_{\max} = \alpha/(\mathrm{e}\beta) = 0.368\alpha/\beta. \tag{3.4.9}$$

As $\alpha/\beta \to 0$, the difference between the correlation function $K_\varphi(\tau) = \mathrm{e}^{-\alpha|\tau|}$ and (3.4.5) can be arbitrarily small (at the expense of large β), while the difference between the mean squares of the solution $x(t)$ of differential equation (3.4.2) is of fundamental nature: at $K_\varphi(\tau) = \mathrm{e}^{-\alpha|\tau|}$ the mean square is infinite, whereas at $K_\varphi(\tau) = \mathrm{e}^{-\alpha|\tau|} + \alpha\tau\mathrm{e}^{-\beta|\tau|}$ the mean square is finite for any α and β. (Note that the difference between the correlation functions, equal to $\alpha\tau\mathrm{e}^{-\beta\tau}$, for $\alpha/\beta \to 0$ is small also in the mean-square metric since the integral

$$\int_0^\infty (\alpha\tau\mathrm{e}^{-\beta\tau})^2 \, \mathrm{d}\tau \tag{3.4.10}$$

is also arbitrarily small for $\alpha/\beta \to 0$).

Thus, we arrive at an important conclusion: calculation of the mean squares of solutions of differential equations through the integral cosine Fourier transformation may appear an ill-posed problem. Inevitable small errors in the calculated correlation function (and even arbitrarily small errors) may result in fundamental differences in the mean squares.

With this example, one of the approaches to ill-posed problems can be figured out. Improved accuracy in calculating the correlation function gives nothing: the fundamental differences in the value of the mean square will be retained with arbitrarily small errors. Yet, one can use additional information about real processes whose spectra are of interest, and this information will allow him to adequately approach the ill-posed problem, i. e., to choose an analytical formula for the correlation function that will provide a best fit to experimental data distorted by inevitable measurement errors.

Actual processes display a finite rate of change and the mean square of the rate $\dot{\varphi}(t)$ of a real process $\varphi(t)$, i. e., the integral

$$\langle\dot{\varphi}\rangle = \int_0^\infty S_{\dot{\varphi}}(\omega)\,\mathrm{d}\omega = \int_0^\infty \omega^2 S_\varphi\,\mathrm{d}\omega \tag{3.4.11}$$

must be finite. Yet, we have already established that this integral is equal to the second derivative of the correlation function at the point $\tau = 0$, i. e.,

$$\int_0^\infty \omega^2 S_\varphi(\omega)\,\mathrm{d}\omega = -\frac{\mathrm{d}^2}{\mathrm{d}\tau}\,K_\varphi(0). \tag{3.4.12}$$

But the second derivative at $\tau = 0$ will be finite only if the first derivative at $\tau = 0$ is continuous. We turn to the widely used correlation function

$$K_\varphi(\tau) = \mathrm{e}^{-\alpha|\tau|}, \tag{3.4.13}$$

and see at once that at the left of the point $\tau = 0$ the derivative $\mathrm{d}K_\varphi/\mathrm{d}\tau$ is positive, equal to $\alpha\mathrm{e}^{\alpha\tau}$, and, as $\tau \to 0$, this derivative tends to $+\alpha$, while at the right of the point $\tau = 0$ the derivative $\mathrm{d}K_\varphi/\mathrm{d}\tau = -\alpha\mathrm{e}^{-\alpha\tau}$ is negative, tending, with $\tau \to 0$, to $-\alpha$. Hence, the first derivative at $\tau = 0$ is discontinuous. It follows from here that at $\tau = 0$ the second derivative of (3.4.13) is infinitely high, and it is therefore not surprising now that the mean square of the derivative $\dot{\varphi}(t)$ of the process $\varphi(t)$ whose correlation function is function (3.4.13) turned out to be infinitely large.

Yet, since finiteness or infiniteness of the mean square of the derivative $\dot{\varphi}(t)$ depends only on the behavior of the second derivative of the correlation function at the single point $\tau = 0$, then finiteness of the mean square can be restored by an arbitrarily small additive to function (3.4.13): it just suffices to provide for identical values of $\mathrm{d}K_\varphi/\mathrm{d}\tau$ at the right and at the left of the point $\tau = 0$.

We saw that correlation function (3.4.5), which for small values of α/β differs arbitrarily little from (3.4.13), refers to a process $\varphi(t)$ whose mean squared rate is finite, because, as $|\tau| \to 0$, the derivative of (3.4.5):

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\,(\mathrm{e}^{-\alpha|\tau|} + \alpha\tau\mathrm{e}^{-\beta|\tau|}) = -\alpha\beta|\tau|\mathrm{e}^{-\beta|\tau|}, \tag{3.4.14}$$

tends to one and the same nonzero value both at the left and at the right of the point $\tau = 0$.

Thus, analytical approximation (3.4.5) of the correlation function adequately reflects the physical essence of the real process, whose rate cannot be infinite. This additional information (about finiteness of the rate) permits a right choice of the analytical approximation for the correlation function in the ill-posed problem on reconstruction of the function from the measured values of $\varphi(t)$ inevitably distorted by small measurement errors.

As for correlation function (3.4.13), this function cannot adequately reflect such an important characteristics of almost all actual processes as finiteness of their rate. (We recall that the correlation function of type (3.4.13) was written for the processes shown in Figures 3.1 and 3.2, i. e., for the processes with “jumps” represented in the graphs by vertical segments, at which the derivative $\dot{\varphi}(t)$ is infinite. In actual processes, the derivative can be large, and even very large, but never infinite. That is why the processes in Figures 3.1 and 3.2, and other processes whose correlation function is given by (3.4.13) are mathematically idealized processes.)

In spite of the aforesaid, correlation function (3.4.13) is widely used in practice. The reasons for that will be discussed in Section 3.5.

3.5. PROBLEMS LOW SENSITIVE TO ERRORS IN THE SPECTRUM

The majority of practical problems on calculation of mean squared values of characteristics of dynamic systems acted upon by disturbing actions representing stochastic processes are weakly sensitive to errors in disturbing-action spectra that lie in the high-frequency portion of the spectrum, or in the region of large values of the variable ω in the spectra $S_\varphi(\omega)$.

We have already mentioned that, provided that variables $x(t)$ and $\varphi(t)$ are interrelated by differential equation (3.1.7) in which the polynomials of the differentiation operator $D = \mathrm{d}/\mathrm{d}t$ have the form (3.1.8) and (3.1.9), the spectra S_x and S_φ are interrelated by formula (3.1.10), in which the quantity $B(j\omega)/A(j\omega)$ is called the “*frequency characteristic*” of the dynamic system or, which is the same, the frequency characteristic of mathematical model (3.1.7). Since for the majority of system the degree of $A(D)$ is greater than the degree of $B(D)$, then the squared modulus of the frequency characteristic

$$|B(j\omega)/A(j\omega)|^2, \tag{3.5.1}$$

vanishes at high ω, and even large deviations between the spectra S_1 and S_2 of the processes $\varphi_1(t)$ and $\varphi_2(t)$ in the high-frequency region result in very small differences between the mean squares $\langle x_1^2\rangle$ and $\langle x_2^2\rangle$ of the solutions $x_1(t)$ and $x_2(t)$.

This circumstance has been already used in Section 1.4 of Chapter 1, where, to provide for well-posedness of a control optimization problem with distorted control-object parameters, we substituted the disturbing-action spectrum

$$S_1(\omega) = (2/\pi)/(1+\omega^2) \tag{3.5.2}$$

(answering the experimental data to the largest possible degree) with the spectrum

$$S_2(\omega) = (2/\pi)(1+0.01\omega^2)/(1+\omega^2), \tag{3.5.3}$$

with which the Petrov criterion for well-posedness of the optimization problem under consideration is satisfied. In spite of the fact that the difference between spectra (3.5.2) and (3.5.3) at high ω is large (since at $\omega = 10$ we have $S_2(10) = 2S_1(10)$, and at $\omega = 20$ even $S_2(20) = 5S_1(20)$), the difference between the mean squares of the controlled process was only 0.88 % (formula (3.3.14)).

That is why instead of correlation function (3.4.5), to which spectrum (3.4.6) corresponds, they often use function (3.4.13), to which spectrum (3.4.4) corresponds. The difference between spectra (3.4.4) and (3.4.6) is large only at high frequencies, and spectrum (3.4.4) is much simpler.

Moreover, they often replace spectrum (3.4.4) with even simpler spectrum $S_\varphi(\omega) = \text{const}$, i. e. with a frequency-independent spectrum. A process with such a spectrum is called "*white noise*". Of course, "white noise" is pure mathematic idealization, and none of actual processes can be "white noise" just because the mean square of the noise is infinite. Nonetheless, replacement of an actual spectrum with a simple constant value is widely used. The admissibility of such a replacement can be conveniently explained with the following example. Consider the motion of a "Kazbek" ship (16 thousand ton displacement, 14 knot velocity) under steering control, acted upon by wind and rough sea induced disturbing actions. The mathematical model for the ship is the equation

$$(690D^2 + 60.8D + 2.5)\theta = \varphi(t), \tag{3.5.4}$$

where θ is the angle of deviation of the ship from the set course and $\varphi(t)$ is the moment of the wind and rough sea induced disturbing force. The spectrum of the disturbing forces can be fitted with the well-known Rakhmanin–Firsov formula:

$$S_\varphi(\omega) = \langle\varphi^2\rangle \, \frac{4\alpha}{\pi} \, \frac{\alpha^2+\beta^2}{(\alpha^2+\beta^2+\omega^2)^2 - 4\beta^2\omega^2}, \tag{3.5.5}$$

where the parameters α and β depend on the wave strength. For waves of moderate strength often encountered values are $\beta = 1/\text{sec}$ and $\alpha = 0.21\beta$.

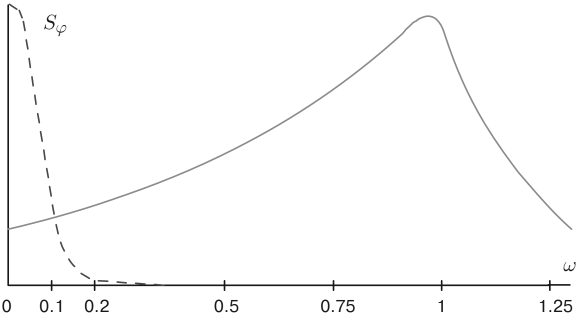

Figure 3.3

Spectrum (3.5.5) for $\beta = 1/\mathrm{sec}$ and $\alpha = 0.21\beta$ is plotted in Figure 3.3 (solid curve); the dashed curve in the same figure shows the squared modulus of the frequency characteristics of the “Kazbek” ship,

$$1/|690(j\omega)^2 + 60.8j\omega + 2.5|^2. \tag{3.5.6}$$

We see that the squared modulus of the frequency characteristic differs noticeably from zero only in the frequency band from $\omega = 0$ to $\omega = 0.15$; in this band, spectrum (3.5.5) differs from $S_\varphi(\omega) = \mathrm{const}$ by not more than 4 %. That is why the difference in the root-mean-square value of the deviation angle of the “Kazbek” ship from the set course with spectrum (3.5.5) replaced with a simpler spectrum $S_\varphi(\omega) = \mathrm{const}$ will remain within 1–1.5 % (for ships with a greater displacement, i. e., with a displacement greater than 16 thousand tons, the difference will be even smaller). That is why in calculating the control rule for steering, in calculating the ship deviation from the course, and in similar calculations it is common practice to substitute the actual disturbing action with a “white noise”. Such substitution simplifies both the calculations and realization of steering and, simultaneously, guarantees a sufficient accuracy. Detailed calculations of optimum steering control systems for various analytical approximations of disturbing-action spectra can be found in the monograph by Petrov (1973, pp. 132–147). These calculations have lent support to the admissibility of replacement of spectrum (3.5.5) with $S_\varphi(\omega) = \mathrm{const}$. This is related to the fact that such a dynamic system as a ship under steering control is weakly sensitive to spectrum variations outside the frequency band substantial for the ship, and this band is very narrow, contained in the interval $0 \le \omega \le 0.1$ or $0 \le \omega \le 0.15$.

In ship-roll calculations, the situation is quite different. For the roll, the natural frequency of the ship and the frequency of maximum disturbing forces are comparable. For this reason, the most significant portion of the spectrum appears to be a vicinity of the maximum, whereas the portion of the spectrum in the vicinity of $\omega = 0$ is insignificant. That is why spectrum (3.5.5) with the correlation function

$$K_\varphi(\tau) = \langle\varphi^2\rangle \mathrm{e}^{-\alpha|\tau|}(\cos(\beta\tau) + (\alpha/\beta)\sin(\beta|\tau|)) \tag{3.5.7}$$

and a modification of spectrum (3.5.5),

$$S_\varphi(\omega) = \langle\varphi^2\rangle \frac{2\alpha}{\pi} \frac{\alpha^2+\beta^2+\omega^2}{(\alpha^2+\beta^2+\omega^2)^2 - 4\beta^2\omega^2}, \tag{3.5.8}$$

with the correlation function

$$K_\varphi(\tau) = \langle\varphi^2\rangle \mathrm{e}^{-\alpha|\tau|}\cos(\beta\tau) \tag{3.5.9}$$

can both be successfully used in roll calculations.

It is readily seen that spectrum (3.5.5) and correlation function (3.5.7) refer to a stochastic process $\varphi(t)$ with finite mean square of the derivative $\dot\varphi(t)$, whereas spectrum (3.5.8) and correlation function (3.5.9), to a more idealized process $\varphi(t)$ for which the mean square of the derivative $\dot\varphi(t)$ has no finite value since the integral

$$\langle\dot\varphi\rangle = \int_0^\infty \omega^2 S_\varphi(\omega)\,\mathrm{d}\omega$$

for spectrum (3.5.8) diverges (for spectrum (3.5.5), this integral is finite since, as $\omega \to \infty$, spectrum (3.5.5) vanishes more rapidly).

Both spectra, (3.5.5) and (3.5.8), are called the *Rakhmanin–Firsov spectra*; either of them can be successfully used in roll calculation despite the fact that at frequencies close to $\omega = 0$ the ordinates of (3.5.5) at equal $\langle\varphi^2\rangle$ are twice the ordinates of (3.5.8). In roll calculations, low frequencies are insignificant, and both spectra are equally applicable.

The situation has changed with the necessity to have, for drilling and exploration works, ships capable of preserving, in spite of the disturbing action of wind and rough sea, their almost exact position in the sea with the help of dynamic stabilization systems. Calculations of such systems have shown that, for them, both high and low frequencies are of significance. The Rakhmanin–Firsov spectra (3.5.5) and (3.5.8), both in use till the 1980ths, were soon recognized to yield wrong proportions between the maxima of the spectrum and the magnitude of the spectrum at $\omega = 0$, thus resulting in overestimation of the power required for normal operation of dynamic-stabilization systems. Even worse results were obtained with the widely known Derbishire, Neumann, Bretshneider, Roll, and Fisher spectra, available in the majority of manuals. All these spectra can be expressed by the general Barling formula:

$$S_\varphi(\omega) = A\omega^{-k}\mathrm{e}^{-B\omega^{-n}} \tag{3.5.10}$$

with coefficients A, B, k, and n slightly differing from each other. All the spectra based on formula (3.5.10) have one feature in common: as $\omega \to 0$, their ordinates tend to zero and, therefore, $S_\varphi(\omega = 0) = 0$. In ship-roll calculations, these spectra all were used quite successfully, but when applied to synthesis of dynamic-stabilization systems, wherein low frequencies are of significance, there soon emerged a sharp difference between the predicted and real behavior of the systems.

The reason consists in that inevitable errors in the measured values of the process $\varphi(t)$, and also errors that has arisen in treating the values, i. e., in calculating the correlation function and the spectrum $S_\varphi(\omega)$, have a most pronounced impact on the results at small values of ω in a certain vicinity of $\omega = 0$. Calculation of $S_\varphi(\omega)$ in the vicinity of $\omega = 0$ turns out to be an ill-posed problem (or a problem close to an ill-posed problem); it is not therefore surprising that different experiments performed by different researchers yielded different values for $S_\varphi(0)$, and a proposition was advanced to apply an approach traditionally practiced with ill-posed problems, namely, to invoke, in addition to the measured data, additional information about the spectrum $S_\varphi(\omega)$ drawn from physical considerations.

The train of thoughts was as follows: for marine dynamic-stabilization systems, the disturbing action is the wave sloping angle $\varphi(t)$, whereas the real sea roughness presents a complex combination of various waves widely ranging in their wavelengths and frequencies. It was believed that very long "elementary waves" refer to the values of $S_\varphi(\omega)$ at low frequencies ω, but such waves are unable to replenish their energy from wind and, therefore, have to be expected to rapidly decay, leading to $S_\varphi(\omega) = 0$ for low ω's. Even if measurements gave values $S_\varphi(0) \neq 0$, these values, believed to result from measurement errors, were replaced with $S_\varphi(0) = 0$. Unfortunately, all these considerations from the very beginning were erroneous. As a matter of fact, the ordinates of the spectrum $S_\varphi(\omega)$ at low frequencies are not bound at all to answer real "long waves". In fact, these ordinates reflect the non-uniformity of $\varphi(t_i)$ over arbitrarily long time intervals. If we consider a particular realization of wave sloping angles over time intervals $t_i \leq t \leq t_{i+1}$, we will see that positive values will prevail in some intervals and negative values will prevail in other intervals, and this fluctuations, although small, will be observed over time intervals of arbitrary length irrespective of how large will be the difference $t_{i+1}-t_i$ (it is these fluctuations that results, in the first place, in the ship crab to be compensated with dynamic-stabilization systems). In the spectrum $S_\varphi(\omega)$ these fluctuations will result in that at low frequencies, including the point $\omega = 0$, we will have $S_\varphi(\omega) \neq 0$. Note that if $S_\varphi(0) \neq 0$, then it does not mean at all (as it is generally believed) that the process $\varphi(t)$ contains a constant component: the spectrum of a "telegraph signal" with the mean value equal to zero, whose correlation function is $K_\varphi(\tau) = a^2 \mathrm{e}^{-2\mu\tau}$, is

$$S_\varphi(\omega) = a^2 \, \frac{8\mu}{\pi(4\mu^2+\omega^2)}, \qquad S_\varphi(0) = \frac{a^2}{\pi}\,\frac{2}{\mu}.$$

We see that, although the use of additional information remains a most powerful approach to ill-posed (or close to ill-posed) problems, care must be taken to choose this additional information soundly. An error in choosing additional information results in an erroneous solution of the main problem, which was initially the case in calculating dynamic-stabilization systems. A breakthrough occurred after 1985, when it was proposed to introduce, in order to describe the disturbing actions exerted on drill and oil exploring ships, a new correlation function with three, instead of two, parameters,

$$K_\varphi(\tau) = \langle\varphi^2\rangle \mathrm{e}^{-\alpha|\tau|}(\cos(\beta\tau) + \gamma\sin(\beta|\tau|)), \qquad (3.5.11)$$

with the spectrum

$$S_\varphi(\omega) = \langle\varphi^2\rangle\,\frac{2\alpha}{\pi}\,\frac{\alpha(\alpha^2+\beta^2+\omega^2)+\gamma\beta(\alpha^2+\beta^2+\omega^2)}{(\alpha^2+\beta^2+\omega^2)^2-4\beta^2\omega^2}. \qquad (3.5.12)$$

The magnitude of the spectrum at $\omega = 0$,

$$S_\varphi(0) = \langle\varphi^2\rangle\,\frac{2}{\pi}\,\frac{\alpha+\gamma\beta}{(\alpha^2+\beta^2)}, \qquad (3.5.13)$$

depends on γ. Practical calculations performed for sea roughness of various intensities shows that, normally, the coefficient γ is negative and its modulus does not exceed 0.1; the magnitude of γ has therefore almost no effect on the spectrum near the maximum of the spectrum, which circumstance greatly facilitates calculations. Earlier, using spectra with three parameters was often avoided since it was hard to simultaneously choose such values of three parameters with which experimental data could be accurately reproduced. Specific features of spectrum (3.5.12) allows one to choose the values of the parameters not simultaneously but in succession: the parameter β is the frequency at which the spectrum attains its highest, the ratio α/β for sea roughness varies in a narrow range, $0.1 < \alpha/\beta < 0.25$, and can be found from the damping rate of the correlation function by traditional methods, as in the previous case of the Rakhmanin–Firsov spectrum. With chosen values of α and β, it is required to calculate, as accurately as it can be done, the value of $S_\varphi(0)$, and then calculate γ by formula (3.5.13).

Three-parametric spectrum (3.5.12), proposed and substantiated by Yu. P. Petrov and V. V. Chervyakov in their monograph "Drill-ship stabilization systems" (the first edition of this monograph was issued in 1985 and the second, additional edition, in 1997 at the St.-Petersburg Technical University), improved the reliability of calculations of dynamic-stabilization systems of core and drill ships.

3.6. DIFFERENTIATION OF DISTORTED FUNCTIONS

Since there are no perfectly precise measuring instruments capable of yielding measurement results with zero error, any continuously measured quantity always suffers inevitable distortions.

Suppose that we intend to measure a function $f_1(t)$; instead, at the output of the meter we have a function

$$f_2(t) = f_1(t) + \varphi(t), \tag{3.6.1}$$

where $\varphi(t)$ is an unknown and, as a rule, random process. This process cannot be made equal to zero. Yet, it is necessary (and possible) to strive for small values of $\varphi(t)$ or, more precisely, for small ratios between the random component $\varphi(t)$ and the function $f_2(t)$, i. e., one should make his best to achieve a small value of the fraction $\varphi(t)/f_2(t)$, which will automatically yield a good approximation of the measured quantity $f_2(t)$ to the unknown true function $f_1(t)$.

New difficulties arise in the differentiation problem if not only the function $f_1(t)$ itself, but also its derivative $\mathrm{d}f_1/\,\mathrm{d}t$ is of interest, whereas the meter gives us just the sum $f_1(t) + \varphi(t)$.

The differentiation problem, or the problem of reconstruction of the true derivative of the function $f_1(t)$ from distorted measurement results, often arises in the theory and practice of automatic control and is of primary significance in this field.

Since even a function whose absolute magnitude is small may have a large derivative, then differentiation of a distorted function may often be an ill-posed problem. However small the difference is between the measured value $f_2(t)$ and the true function $f_1(t)$, the difference between their derivatives

$$\frac{\mathrm{d}f_2(t)}{\mathrm{d}t} - \frac{\mathrm{d}f_1(t)}{\mathrm{d}t} = \frac{\mathrm{d}\varphi(t)}{\mathrm{d}t} \tag{3.6.2}$$

may be large. Even more frequently this problem turns out to be an ill-conditioned problem, when to a small (but finite) difference of the functions, $f_2(t) - f_1(t)$, a more appreciable difference of their derivatives corresponds, resulting from that the random component $\varphi(t)$ very often, as a rule, has a higher variation frequency.

Now, let us pose a problem that requires finding an optimum operator that best solves the differentiation problem for the distorted function. We will find an optimum linear operator comprising various combinations of the differentiation operator $D = \mathrm{d}/\mathrm{d}t$ and integration operator $D^{-1} = \int_0^t \mathrm{d}t$, or, in other words, it is required to find an operator of the form $B(D)/A(D)$, where $A(D)$ and $B(D)$ are some polynomials of the differentiation operator $D = \mathrm{d}/\mathrm{d}t$.

We denote the sought operator as L. It follows from linearity of L that

$$L[x_1(t) + x_2(t)] = L[x_1(t)] + L[x_2(t)], \tag{3.6.3}$$

i. e., the operator L applied to the sum of functions x_1+x_2 yields a result that equals the sum of the results obtained by application of the operator to each of the term individually. Note that the differentiation operator $D = \mathrm{d}/\mathrm{d}t$ provides an example of a linear operator. Formula (3.6.3) expresses the well-known rule: the derivative of the sum of two functions is equal to the sum of their derivatives.

So, suppose that the sum of a measured function $x(t)$ and an interfering random function $\varphi(t)$ is acted upon by the sought linear operator (and is also fed to the input of the device that realizes this operator); we seek the operator L that transforms the sum $x+\varphi$ so that to simultaneously minimize the difference between the transformed sum $L(x + \varphi)$ and the derivative $\dot{x}$, i. e., the difference

$$L(x + \varphi) - \frac{\mathrm{d}}{\mathrm{d}t}\,x = y. \tag{3.6.4}$$

It follows from linearity of L that

$$y = L(x) + L(\varphi) - \frac{\mathrm{d}}{\mathrm{d}t}\,x = \Big(L - \frac{\mathrm{d}}{\mathrm{d}t}\Big)(x) + L(\varphi). \tag{3.6.5}$$

Let us apply now the integral cosine Fourier transformation to all terms in (3.6.5). After the transformation, the functions y, x, and φ will be substituted with their spectra $S_y(\omega)$, $S_x(\omega)$, and $S_\varphi(\omega)$, and the operators, with squared moduli of their Fourier transforms, or squared moduli of the frequency characteristics of these operators. We obtain:

$$S_y(\omega) = |L_\omega - j\omega|^2 S_x(\omega) + |L_\omega|^2 S_\varphi(\omega). \tag{3.6.6}$$

In (3.6.6), the subscribed character L_ω denotes the cosine Fourier transform of L, a function of ω (recall that the cosine transform of the differentiation operator $\mathrm{d}/\mathrm{d}t$ is $j\omega$). It follows from (3.6.6) that

$$\langle y^2\rangle = \int_0^\infty S_y(\omega)\,\mathrm{d}\omega = \int_0^\infty (|L_\omega - j\omega|^2 S_x + |L_\omega|^2 S_\varphi)\,\mathrm{d}\omega. \tag{3.6.7}$$

The minimum of $\langle y^2\rangle$ is provided by a function L_ω that minimizes the integrand. It is this fact, by the way, that makes the passage to the Fourier transform a useful operation: it is a difficult task to directly find the operator L which minimizes difference (3.6.4) since methods for finding optimum operators are still poorly developed. At the same time, it is quite an easy matter to find the function L_ω that minimizes integral (3.6.7): it suffices to invoke the well-known variational-calculus methods. With these methods applied, we obtain that

$$L_{\omega,\mathrm{opt}} = j\omega S_x(\omega)/(S_x(\omega) + S_\varphi(\omega)), \tag{3.6.8}$$

and the squared modulus of the optimum operator is

$$|L_{\omega,\mathrm{opt}}|^2 = \omega^2 (S_x)^2/(S_x + S_\varphi)^2.$$

We substitute (3.6.8) into (3.6.7) and find the least possible root mean square value of the differentiation error:

$$\langle y^2\rangle = \int_0^\infty \omega^2\,\frac{S_x S_\varphi}{S_x + S_\varphi}\,\mathrm{d}\omega. \tag{3.6.9}$$

The limiting case of $S_\varphi \to 0$ refers to the case of no random component present in the function under consideration and $S_\varphi(\omega) = 0$. In this limiting case, as it could be expected,

$$L_{\omega,\mathrm{opt}} = j\omega, \qquad L_{\mathrm{opt}} = \frac{d}{dt}; \tag{3.6.10}$$

i. e., with no random component involved, one can differentiate the function $x(t)$ in the ordinary manner since, in this case, the problem presents no difficulties. With a random component involved, the situation, of cause, becomes more complicated. In addition, exact realization of operator (3.6.8) is not always possible, and one has to restrict himself to some approximation to this operator; yet, the use of additional information about the spectrum of the function to be differentiated and the spectrum of its random component enables substantial reduction of the differentiation error.

Example. Consider the problem of differentiation of a function $x(t)$ whose mean square is $\langle x^2\rangle = 1$, the correlation function is $K_\varphi(\tau) = \mathrm{e}^{-\alpha\tau}(1+\alpha\tau)$, and the spectrum is

$$S_x = (4/\pi)\alpha^3/(\alpha^2+\omega^2)^2. \tag{3.6.11}$$

The measured function $x(t)$ involves a random component $\varphi(t)$ whose mean square is $\langle\varphi^2\rangle$ and the spectrum is

$$S_\varphi = \langle\varphi^2\rangle(2/\pi)\beta/(\beta^2+\omega^2), \tag{3.6.12}$$

i. e., in fact, we have to differentiate the sum $x(t)+\varphi(t)$. As it was noted previously, spectrum (3.6.12) refers to a non-differentiable process $\varphi(t)$. Note that, if the rate of change of real physical processes is finite and these processes are therefore differentiable, then the measurement error can vary stepwise as in Figure 3.2 and can therefore have the spectrum given by formula (3.6.12). The involvement of a non-differentiable random component makes differentiation of the signal $x(t) + \varphi(t)$ that contains a random component $\varphi(t)$, an ill-posed problem.

Indeed, we calculate the mean squared derivative of the signal $y = x+\varphi$ (i. e., the result of transformation of the signal $y = x + \varphi$ by the simple differentiation operator $D = \mathrm{d}/\mathrm{d}t$) and obtain:

$$\langle y^2\rangle = \int_0^\infty \frac{4}{\pi}\,\frac{\alpha^3\omega^2}{(\alpha^2+\omega^2)^2}\,\mathrm{d}\omega + \int_0^\infty \langle\varphi^2\rangle\,\frac{2}{\pi}\,\frac{\beta\omega^2}{\beta^2+\omega^2}\,\mathrm{d}\omega. \tag{3.6.13}$$

It is seen at once from formula (3.6.13) that the first improper integral converges and is equal to α^2, whereas the second integral diverges at any, even very small, mean square $\langle\varphi^2\rangle$ of the random component $\varphi(t)$. The differentiation error for any small random component $\varphi(t)$ with spectrum (3.6.12) turns out to be infinitely large.

To approach this ill-posed problem, one can employ additional information contained in the spectra of the useful signal $x(t)$ and random component $\varphi(t)$ and use, instead of the differentiation operator $\mathrm{d}/\mathrm{d}t$, the operator whose Fourier transform has the form (3.6.8) or an operator close to the

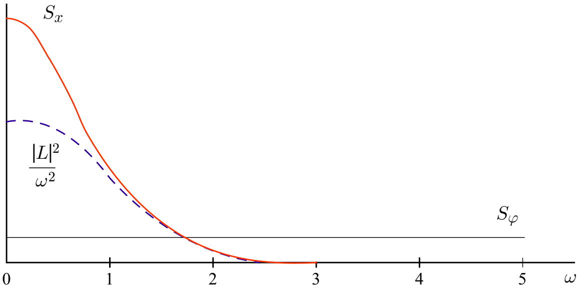

Figure 3.4

latter. It is easy to check (by formula (3.6.9)) that, in this case, the differentiation error will be most frequently small.

If the exact spectra of the useful signal $x(t)$ and the random component $\varphi(t)$ are hard to obtain, then a simple rule can be used: if a predominant part in the spectrum of the random component is played by higher frequencies than in the useful signal, then, instead of the differentiation operator $\mathrm{d}/\mathrm{d}t$, one should use an operator for which the squared Fourier-transform modulus decreases with increasing frequency. The lesser the squared modulus of the operator differs from the optimum one given by (3.6.8), the lesser the differentiation error will be.

By way of illustration, Figure 3.4 shows the spectrum of useful signal (3.6.11) for $\alpha = 1$ and the spectrum of random component (3.6.12) for $\langle\varphi^2\rangle = 1$ and $\beta = 10$; the dashed curve in this figure is the ratio of the squared modulus of optimum operator (3.6.8) to that of the frequency characteristics of the ideal-differentiation operator $\mathrm{d}/\mathrm{d}t$. The spectrum of the random component decays more slowly with increasing ω than that of the useful signal, and the squared modulus of the frequency characteristic of the optimum filter rapidly decreases with increasing ω.

A similar consideration allows one to solve a more simple optimal-filtration problem, wherein it is required to extract, in a best possible manner, a signal $x(t)$ from meter indications $x(t) + \varphi(t)$ distorted by a random component $\varphi(t)$. We assume that the mean squares $\langle x^2\rangle$ and $\langle\varphi^2\rangle$ of the signal $x(t)$ and the error $\varphi(t)$ are known, together with their spectra $S_x(\omega)$ and $S_\varphi(\omega)$. It is required to find a linear operator L such that to minimize the difference between the yield of the operator $L(x + \varphi)$ and the sought signal $x(t)$:

$$L(x + \varphi) - x = y. \tag{3.6.14}$$

It follows from linearity of L that

$$y = L(x) - L(\varphi) - x = (L - 1)(x) + L(\varphi). \tag{3.6.15}$$

Apply now the integral Fourier transformation to all terms in (3.6.15). On this transformation, the functions $y(t)$, $x(t)$, and $\varphi(t)$ will pass into their spectra $S_y(\omega)$, $S_x(\omega)$, and $S_\varphi(\omega)$, and the operators L and $L - 1$ will be replaced with squared moduli of their Fourier transforms, which are functions of ω and will be denoted as L_ω and $L_\omega - 1$:

$$S_y(\omega) = |L_\omega - 1|^2 S_x(\omega) - |L_\omega|^2 S_\varphi(\omega). \tag{3.6.16}$$

It follows from (3.6.16) that

$$\langle y^2 \rangle = \int_0^\infty S_y(\omega)\,\mathrm{d}\omega = \int_0^\infty [|L_\omega - j\omega|^2 S_x + |L_\omega|^2 S_\varphi]\,\mathrm{d}\omega. \tag{3.6.17}$$

To find the optimum operator minimizing the mean square of $y(t)$, it suffices to find the function L_ω that minimizes integral (3.6.17). We take the derivative with respect to L_ω, equate it with zero, and obtain:

$$L_{\omega,\mathrm{opt}} = S_x/(S_x + S_\varphi). \tag{3.6.18}$$

Then, we insert (3.6.18) into (3.6.17) to find the least possible filtration error

$$\langle y^2 \rangle_{\min} = \int_0^\infty \frac{S_x S_\varphi}{S_x + S_\varphi}\,\mathrm{d}\omega. \tag{3.6.19}$$

Example. Let a useful signal $x(t)$ whose correlation function is $K_x = \mathrm{e}^{-\tau}$ and spectrum is

$$S_x = 1/(1 + \omega^2), \tag{3.6.20}$$

be fed to the input of a control system. Admixed to the signal are measurement errors, noise, etc., presenting a "white noise" with the spectrum

$$S_\varphi = 1 \tag{3.6.21}$$

(i. e., with infinite mean square; eventually, this means that the signal is heavily distorted, the power of the random component is high, and its spectrum can be approximated with simple dependence (3.6.21)).

Here, the optimum operator L_ω is

$$L_{\omega,\mathrm{opt}} = 1/(2 + \omega^2), \tag{3.6.22}$$

and its squared modulus is

$$|L_\omega|^2 = 1/(4 + 4\omega^2 + \omega^4). \tag{3.6.23}$$

The mean squared error is

$$\langle y^2 \rangle = \int_0^\infty \frac{(1+\omega^2)^{-1}}{(1+\omega^2)^{-1}+1}\,\mathrm{d}\omega = \int_0^\infty \frac{1}{2+\omega^2}\,\mathrm{d}\omega = \frac{\sqrt{2}\,\pi}{4}. \tag{3.6.24}$$

Filtration can be realized with a dynamic system whose differential equation is

$$(D^2 + \sqrt{8}\,D + 2)y = x + \varphi. \tag{3.6.25}$$

Since the spectrum of the random component is uniform throughout the whole range of frequencies ω, and the spectrum of the useful signal decreases with increasing frequency, it follows from here that the squared modulus of the frequency characteristic of the filter can be expected to decrease with increasing ω, which fact is reflected by formula (3.6.23). If exact realization of the dynamic system with squared modulus of frequency characteristic (3.6.8) or (3.6.18) is difficult to perform, then a simpler dynamic system can be used, such that its frequency characteristic can be approximated with formulas (3.6.8) or (3.6.18).

For instance, instead of dynamic system (3.6.25) one can use a simpler dynamic system

$$(\sqrt{2} + D)y = x + \varphi, \tag{3.6.26}$$

for which

$$\langle y^2 \rangle = \int_0^\infty S_y(\omega)\,\mathrm{d}\omega = \int_0^\infty \frac{S_x + S_\varphi}{2+\omega^2}\,\mathrm{d}\omega = \frac{\pi}{2}. \tag{3.6.27}$$

The filtration error of (3.6.26) is greater than that of (3.6.25), but still remains quite acceptable.

3.7. PROGNOSTICATION

In this section, we will consider a problem that cannot be solved exactly in principle: it is required to predict the future values of a stochastic process $\varphi(t)$ based on observations of this process in the past and present.

The problem can be formulated as follows: we monitor the process $\varphi(t)$ over a certain time interval; these observations prove the process to be stationary and allow us to calculate the correlation function $K_\varphi(\tau)$ and the spectrum $S_\varphi(\omega)$ of the process. Note that the correlation function can be calculated with sufficient accuracy in rather short a time: provided that the correlation function has the form (3.2.12) with $\mu = 1/\text{sec}$, for instance, then the measurements can be completed in a 7-second interval since, with $\tau \geq 7$ seconds, the magnitude of correlation function (3.2.12) will be lesser than one thousandth fraction of its maximum value at $\tau = 0$; that is why in ordinary engineering problems observations over time intervals longer than 7 seconds makes no sense. If $\mu \neq 1/\text{sec}$, then it suffices to perform observations over the time interval $0 \leq t \leq 7/\mu$.

So, let by the time $t = 0$ we have obtained the correlation function and the spectrum $S_\varphi(\omega)$ of the process $\varphi(t)$, and are able to measure $\varphi(0)$ and $\dot{\varphi}(0)$, but have no other additional information about the process. It is required to predict, as reliably as possible, the future values of the process $\varphi(t)$, or its values at $t > 0$.

This problem was first posed and solved in 1943 by Norbert Wiener (1894–1964), the outstanding American mathematician. World War II was escalating, and Wiener advanced a method for predicting the position of an enemy airplane to make the flak fire more efficient.

The prediction method proposed by N. Wiener was complex. It is described in the well-known book "Cybernetics" by Wiener (see Russian edition issued in 1958, pp. 81–123); in 1943, when this method was first published in a secret yellow-cover booklet, this booklet became widely known among design engineers as "yellow danger". The engineers felt that the Wiener method was capable of raising gunfire efficiency, but, simultaneously, it was almost impracticable. Many hundred thousand hours, such expensive in wartime, were spent in vein in attempts to comprehend the "yellow danger". Some people even proposed to parachute the book over Germany so that German engineers had also to spend precious time on attempts to understand N. Wiener.

We will give below a simpler, although not rigorous, derivation based on the calculation of the spectrum of a sequence of impulse functions, i. e., such functions whose mean square (power) is zero because

$$\langle \varphi^2 \rangle = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \varphi^2 \, \mathrm{d}t \to 0, \tag{3.7.1}$$

but the integral

$$E_\varphi = \lim_{T\to\infty} \int_{-T}^{T} \varphi^2 \, \mathrm{d}t \tag{3.7.2}$$

exists and has a finite magnitude. The terms "power" and "energy" are used because, with $\varphi(t)$ being electric current, the square $\varphi^2(t)$ presents the instant power of the current released in the resistance $r = 1$ Ohm, mean square (3.7.1) is the mean power of the current, and integral (3.7.2) is the energy of the current released in the resistance 1 Ohm during the time $-\infty < t < \infty$ (from here, in particular, the full name of the spectrum $S_\varphi(\omega)$, spectral power density, originates; recently, a new term "spectrum" has come into use). By analogy with the ordinary correlation function, we introduce the correlation function $r_\varphi(\tau)$ defined by the equality

$$r_\varphi = \int_0^\infty \varphi(t)\varphi(t+\tau)\,\mathrm{d}t = \int_{-\infty}^0 \varphi(t)\varphi(t-\tau)\,\mathrm{d}t, \tag{3.7.3}$$

and the spectral energy density $E_\varphi(\omega)$ defined as the cosine Fourier transform of the correlation function:

$$E_\varphi(\omega) = \frac{2}{\pi}\int_0^\infty r_\varphi(\tau)\cos(\omega\tau)\,\mathrm{d}\tau. \tag{3.7.4}$$

The inverse cosine transformation

$$\int_0^\infty E_\varphi(\omega)\cos(\omega\tau)\,\mathrm{d}\omega = r_\varphi(\tau) \tag{3.7.5}$$

reconstructs the correlation function $r_\varphi(\tau)$; we put $\tau = 0$ in formula (3.7.5) and obtain

$$r_\varphi(0) = \int_0^\infty \varphi^2\,\mathrm{d}t = \int_0^\infty E_\varphi(\omega)\,\mathrm{d}\omega = E_\varphi, \tag{3.7.6}$$

i. e., the energy of the impulse function $\varphi(t)$ is equal to the integral of the spectral energy density.

Example. Consider an exponential pulse, i. e.,

$$\varphi(t) = \begin{cases} \alpha\mathrm{e}^{-\alpha t}, & 0 \le t < \infty, \\ 0, & -\infty < t < 0. \end{cases} \tag{3.7.7}$$

For this function, we have

$$r_\varphi = \alpha^2 \int_0^\infty \mathrm{e}^{-\alpha t}\mathrm{e}^{-\alpha(t+\tau)}\,\mathrm{d}t = \frac{\alpha}{2}\,\mathrm{e}^{-\alpha|\tau|}. \tag{3.7.8}$$

We apply the cosine transformation and obtain

$$E_\varphi(\omega) = \frac{2}{\pi}\int_0^\infty \frac{1}{2\alpha}\,\mathrm{e}^{-\alpha\tau}\cos(\omega\tau)\,\mathrm{d}\tau = \frac{1}{\pi}\,\frac{1}{\alpha^2+\omega^2}. \tag{3.7.9}$$

We see that the spectral energy density of the exponential function differs from the spectral power density of the exponential correlation function just by a constant factor.

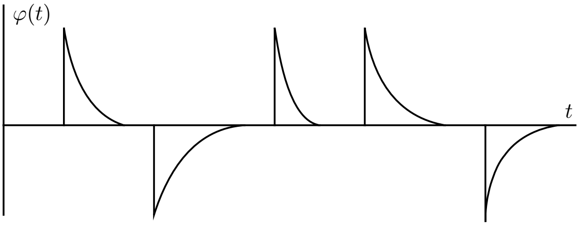

Figure 3.5

Consider now a random sequence of exponential pulses of both signs, $\pm\alpha e^{-\alpha t}$, randomly spaced in time according to the Poisson law. In a unit time interval, we have, in the mean, μ pulses. Such a pulse sequence is shown in Figure 3.5.

Let the alternation of the signs be such that the mean magnitude of the pulse sequence is zero, and the pulses follows at a very low repetition frequency, so that each pulse has enough time to completely decay by the time the next pulse arrives. Since the spectral energy density of a pulse $\pm\alpha e^{-\alpha t}$ is

$$E_\varphi(\omega) = \frac{1}{\pi}\,\frac{1}{\alpha^2+\omega^2} \tag{3.7.10}$$

and, in the mean, μ pulses arrives during unit time, then the spectral density (spectrum) of the pulse sequence is

$$S_\varphi(\omega) = \frac{\mu\alpha}{2}\,\frac{2}{\pi}\,\frac{1}{\alpha^2+\omega^2}, \tag{3.7.11}$$

i. e., it differs from the processes with spectrum (3.3.8), more than once considered previously, just by a constant factor. We mentioned already that the correlation functions and the spectrum are not exhaustive characteristics of a process $\varphi(t)$. There may exist different processes with one and the same spectrum (these processes are called realizations of a process with the spectrum $S_\varphi(\omega)$). We saw that the exponential function $K_\varphi = e^{-\alpha\tau}$ and spectrum (3.3.8) are displayed by the realization having the form of a "telegraph signal" (Figure 3.1), and also by the realization in the form of the function shown in Figure 3.2.

Figure 3.5 depicts another realization of a stochastic process with an exponential correlation function, namely, the realization in the form of a random sequence of exponential functions.

In the limit of $\alpha \to \infty$, each pulse $\alpha e^{-\alpha\tau}$ passes into a δ-function, the random sequence of the pulses transforms into a random sequence of δ-func-

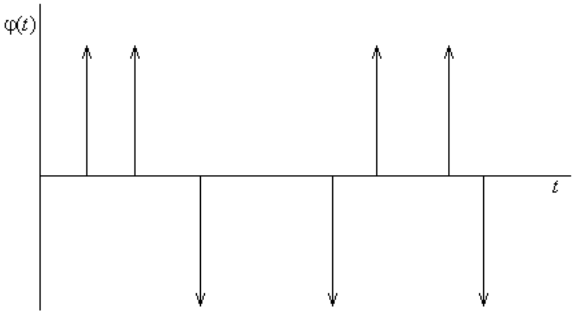

Figure 3.6

tions, and spectrum (3.7.11) passes into a constant quantity

$$S_\varphi(\omega) = \mu/\pi. \tag{3.7.12}$$

As was mentioned above, a process with a frequency-independent spectrum is called “white noise”. We see that one of the realizations of the “white noise” is a random sequence of δ-functions of different signs. A schematic of this sequence is shown in Figure 3.6. Since the amplitude of the δ-function is infinite, this function is conventionally depicted as a “pointed arrow” whose tip symbolizes the fact that this function has infinite amplitude.

Now, we are ready to answer the basic question: for the random process $\varphi(t)$ observed till the moment $t = 0$, the spectrum and the correlation function $K_\varphi = \langle\varphi^2\rangle \mathrm{e}^{-\alpha\tau}$ are calculated. At $t = 0$, the measurements give $\varphi(t) = \varphi(0)$. Which most plausible statement can be made about the values of β at $t > 0$ based on the available information about the spectrum? The answer is clear: the best prediction for the future values of $\varphi(t)$ is the function $\varphi(0)\mathrm{e}^{-\alpha\tau}$.

At least for one realization of the process $\varphi(t)$ with $K_\varphi = \mathrm{e}^{-\alpha\tau}$, namely, for the realization in the form of a random sequence of pulses shown in Figure 3.5, this prediction is an exact prediction, and for all other realizations, an approximate prediction.

The approximate nature of the prediction $\varphi_{\text{pred}} = \varphi(0)\mathrm{e}^{-\alpha\tau}$ is evident for the realization of $\varphi(t)$ in the form of a “telegraph signal”, whose correlation function is also exponential. The dashed curve in Figure 3.7 shows the best prediction for $\varphi(t)$ for $t \geq 2$ (here, the magnitude of the process $\varphi(t)$ was last measured at $t = 2$, and the curve shows the values predicted for $t \geq 2$).

The prediction accuracy rapidly decreases with increasing time t (or, more exactly, with increasing product αt). For $t = (0.7 \div 0.8)\alpha^{-1}$, the

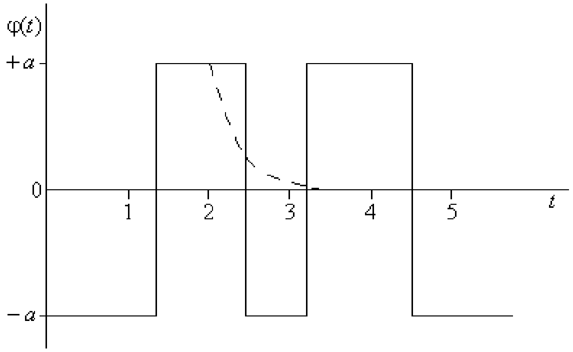

Figure 3.7

prediction accuracy is still quite acceptable, and then it gradually decreases. This behavior is easy to understand since the correlation function and the spectrum give us only limited information about the process $\varphi(t)$.

If we are interested in a more accurate prediction, then deep penetration into the mechanisms behind the process $\varphi(t)$ is necessary. Nonetheless, the correlation function is easy to calculate and provides a good prediction over short prediction intervals.

Let us turn now to a more complex problem of predicting future values of a process $\varphi(t)$ whose spectrum is

$$S_\varphi(\omega) = A/(\omega^4 + M\omega^2 + N), \tag{3.7.13}$$

where $N \geq M^2/4$ (if $N < M^2/4$, then spectrum (3.7.13) can be decomposed in a sum of spectra of form (3.3.8); the problem for such spectra was solved previously). If $N \geq M^2/4$, then spectrum (3.7.13) can be transformed into

$$S_\varphi = A/((\omega^2 + \alpha^2 + \beta^2)^2 - 4\beta^2\omega^2), \tag{3.7.14}$$

where

$$\alpha^2 = M/4 + \sqrt{N}/2, \qquad \beta^2 = \sqrt{N}/2 - M/4.$$

Let the realization of the process $x(t)$ in the form of a random sequence of δ-functions is fed to the input of a dynamic system whose differential equation is

$$(D^2 + 2\alpha D + \beta^2 - \alpha^2)\,\varphi = x(t). \tag{3.7.15}$$

Such a realization (as it was shown previously) has a frequency-independent spectrum; hence, the spectrum of $\varphi(t)$ at the output of system (3.7.15) will have the form (3.7.14), and the realization of the process with spectrum (3.7.14) in this case will be a random sequence of decaying pulses of form

$$\varphi_i(t) = \pm A_i \mathrm{e}^{-\alpha t} \sin(\beta t). \tag{3.7.16}$$

Each pulse is generated by a δ-function in the sequence. (Here, we consider $\varphi(t)$-functions so widely spaced in time that the processes generated by one δ-function have enough time to decay unless a next δ-function arrives at the input of the dynamic system). The coefficients A_i at the pulses depend on the magnitude and sign of the coefficients at the δ-functions in the sequence.

Consider now the spectrum

$$S_\varphi(\omega) = B\omega^2/(\omega^4 + M\omega^2 + N). \tag{3.7.17}$$

The process with spectrum (3.7.17) is a derivative of the process with spectrum (3.7.14). Hence, one of the realizations of this process may be a random sequence of pulses of form

$$\varphi_i(t) = \pm A_i \mathrm{e}^{-\alpha t}(\beta\cos(\beta t) - \alpha\sin(\beta t)) \tag{3.7.18}$$

(each of the pulses is the derivative of (3.7.16)).

With formulas (3.7.16) and (3.7.18), we can find one of the realizations of the process whose spectrum has the form

$$S_\varphi(\omega) = (A + B\omega^2)/(\omega^4 + M\omega^2 + N). \tag{3.7.19}$$

Such a realization will be a random sequence of decaying pulses

$$\varphi_i(t) = \mathrm{e}^{-\alpha t}(c_1\cos(\beta t) + c_2\sin(\beta t)), \tag{3.7.20}$$

the first one starting at the point $t = 0$; the constants c_1 and c_2 depend on and in spectrum (3.7.19).

Now, we can readily solve the problem of best prediction of the process $\varphi(t)$ with spectrum (3.7.19) from the values of $\varphi(t)$ and $\dot\varphi(t)$ at $t = 0$. The best prediction will be function (3.7.20) with constants c_1 and c_2 found based on the measured values of $\varphi(0)$ and $\dot\varphi(0)$. It follows from (3.7.20) that

$$c_1 = \varphi(0), \qquad c_2 = (\dot\varphi(0) + \alpha\varphi(0))/\beta. \tag{3.7.21}$$

With allowance for (3.7.21), the optimum prediction of the future (for $t \geq 0$) values of the random process with spectrum (3.7.19) acquires the form

$$\varphi(t \geq 0) = \mathrm{e}^{-\alpha t}\Big[\varphi(0)\cos(\beta t) + \frac{\dot\varphi(0) + \alpha\varphi(0)}{\beta}\sin(\beta t)\Big]. \tag{3.7.22}$$

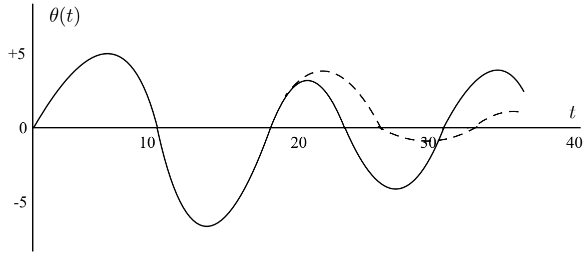

Figure 3.8

For realization (3.7.20), prediction (3.7.22) will be the exact one (unless the next pulse arrives at the input of the dynamic system); for all other realizations of the random process with spectrum (3.7.19) prediction (3.7.22) from the very beginning will be an approximate, although yielding an acceptable accuracy for the time $t = (0.2 \div 0.3)\alpha^{-1}$, prediction.

For seagoing ships whose rolling spectrum can be described with formula (3.7.14), and the typical values of α and β for which are $\beta = 0.4 \div 1$ sec^{-1} and $\alpha = 0.08 \div 0.2$ sec^{-1}, one can expect that the roll angles will be predicted rather accurately over times $t = 3 \div 5$ sec.

Admittedly, even such a prediction is a good one since realizations of the random process with spectrum (3.7.14), even realizations of form (3.7.16), may differ greatly from each other. A simplest case is the one in which pulses follow at such a low repetition frequency that they do not overlap. Otherwise, if the pulses overlap, then the realization of the process $\varphi(t)$ may be rather queer; in these conditions prediction (3.7.22), which provides for the various realizations a good accuracy over the times $t = (0.2 \div 0.3)\alpha^{-1}$, should be acknowledged as a breakthrough of the theory of random processes.

The solid curve in Figure 3.8 is a typical realization of the random process with spectrum (3.7.19), ship rolling on rough sea. Shown is the roll angle θ, expressed in degrees, as a function of time t, and the dashed curve shows the optimum prediction for $t \geq 20$. At $t = 20$ sec, the value of $\theta(20)$ and the derivative $\dot{\theta}(20)$ were last measured, and then the prediction follows.

Note that, having solved the prediction problem for the processes with spectra (3.7.11) and (3.7.19), we, in fact, have it solved in the general statement, for any fractional rational spectrum

$$S_\varphi(\omega) = \frac{a_p\omega^{2p} + a_{p-1}\omega^{2p-2} + \ldots + a_0}{b_q\omega^{2q} + b_{q-1}\omega^{2q-2} + \ldots + b_0}, \tag{3.7.23}$$

where $p < q$. Since any spectrum is an even function, and spectrum (3.7.23) is a function of ω^2, we can introduce the notation $\omega^2 = x$ and write the denominator as

$$b_q x^q + b_{q-1} x^{q-1} + \ldots + b_0. \tag{3.7.24}$$

Afterwards, as it is well known, fractional rational spectrum (3.7.23) can be decomposed into the sum of spectra:

$$S_\varphi(\omega) = \sum_{i=1}^{n} \frac{c_i}{\alpha_i^2 + \omega^2} + \sum_{i=1}^{n} \frac{A_i + B_i\omega^2}{N_i + M_i\omega^2 + \omega^4}. \tag{3.7.25}$$

Each of the spectra of form

$$S_i = c_i/(\alpha_i^2 + \omega^2) \tag{3.7.26}$$

refers to one of the real roots $x_i = \omega_i^2$ of (3.7.24), and each of the spectra

$$S_i = (A_i + B_i\omega^2)/(\omega^4 + M_i\omega^2 + N_i), \tag{3.7.27}$$

to each pair of complex-conjugate roots. The constants c_i, A_i, and B_i in (3.7.25) can be found by the ordinary rules of indefinite multipliers.

The problem of prognostication based on information about the correlation function (or, which is the same, about the spectrum of the random process) presents an interesting example of a problem in which the solution error depends not only on the measurement and calculation errors, but also on fundamental factors. Even if, supposedly, we have precisely calculated the correlation function of a process $\varphi(t)$, this calculation will not allow an exact prediction of the future values of $\varphi(t)$ just because the information contained in the correlation function is insufficient for the exact prediction to be made.

To make a more accurate prediction, one should use an approach that has already proved useful in solving ill-posed problems — one has to invoke some additional information about the process under study; this information will allow more accurate prognostication.

By way of example, consider weather forecasting or, more precisely, prediction of the mean temperature for tomorrow and a few days ahead. The temperature or, more precisely, the deviation of the current temperature from the known smoothly varying many-year temperature is a stationary random process whose correlation function (more exactly, the values of the correlation function for $\tau_1 = 1$ day, $\tau_2 = 2$ days, and so on) is easy to calculate. These values, as it can be easily checked, often fall onto roughly an exponential curve.

Calculations performed for the Moscow Oblast' yielded $K_\varphi(\tau = 1 \text{ day}) = 0.65$; $K_\varphi(\tau = 2 \text{ days}) = 0.42$, etc.

From here, a simplest prognostication rule follows: if the today's temperature, for instance, is 4 degrees greater that the many-year temperature, then, very probably, tomorrow this temperature will be 2.6 degrees greater than the many year's temperature, the next day, 1.7 degree greater, etc. You can check it for yourself that this very simple prognostication rule yields a better result than just a guess about the tomorrow's temperature. Nonetheless, this simple forecast is far less accurate compared to the forecasts delivered by present-day meteorological services based on versatile information about atmospheric processes, cyclone trajectories, etc.

The general rule is as follows: if you happened to encounter an ill-posed problem or a problem close to an ill-posed problem, you should try to find additional information. This information may prove (and often does prove) to be of much help.

Bibliography to Part I

Basic literature

Abdullaev N. D. and Petrov Yu. P. (1985). *Theory and Methods of Optimal-Controller Design.* Energoatomizdat, Leningrad (in Russian).

Andronov A. A., Vitt A. A., and Khaikin S. E. (1981). *Vibration Theory.* Nauka, Moscow (in Russian).

Danilevich Ya. B. and Petrov Yu. P. (2000). On the necessity of broadening the notion of equivalence of mathematical models. *Dokl. Ros. Akad. Nauk*, 371 (4), 473–475 (in Russian).

Gaiduk A. R. (1997). Stability of linear systems. *Avtomatika i Telemekhanika*, 3, 153–160 (in Russian).

Ivanov V. K., Vasin V. V., and Tanana V. P. (2002). *Theory of Linear Ill-Posed Problems and its Applications.* VSP, Zeist.

Kharitonov V. L. (1978). Asymptotic stability of the equilibrium position of a family of linear differential equations. *Diff. Uravneniya*, 11 (in Russian).

Larin V. B., Naumenko K. I., and Suntsev V. N. (1971). *Spectral Synthesis Methods for Linear Feedback Systems.* Naukova Dumka, Kiev (in Russian).

Lavrent'ev M. M. (1981). *Ill-Posed Problems for Differential Equations.* Novosibirsk State University, Novosibirsk (in Russian).

Letov A. M. (1960). Analytical controller design. *Avtomatika i Telemekhanika*, 4, 5, 6; 4 (1961) (in Russian).

Letov A. M. (1969). *Flight Dynamics and Control.* Nauka, Moscow (in Russian).

Nadezhdin P. V. (1973). Loss of coarseness in elementary transformations of differential control-system equations. *Avtomatika i Telemekhanika*, 1, 185–187 (in Russian).

Petrov Yu. P. (1973). *Optimization of Cotrolled Systems under Wind and Sea Roughness.* Sudostrienie, Leningrad (in Russian).

Petrov Yu. P. (1977). *Variation Methods in the Optimal-Control Theory.* Energiya, Leningrad (in Russian).

Petrov Yu. P. (1987). *Synthesis of Optimal Control Systems with Partially Known Disturbing Forces.* Leningrad State University, Leningrad (in Russian).

Petrov Yu. P. (1994). Stability of linear control systems under parameter variations. *Avtomatika i Telemekhanika*, 11, 186–189 (in Russian).

Petrov Yu. P. (1998). *The Third Class of Problems of Physics and Technics, Intermediate between Well- and Ill-Posed Ones.* St.-Petersburg State University, St.-Petersburg (in Russian).

Petrov Yu. P. (2001). *Lectures on the History of Applied Mathematics.* St.-Petersburg State University, St.-Petersburg (in Russian).

Petrov Yu. P. and Petrov L. Yu. (1999). *Unexpected in Mathematics and Its Relation to Recent Accidents and Catastrophes.* St.-Petersburg State University, St.-Petersburg, 1st edition; 3rd edition (2002) (in Russian).

Sergeev V. O. (1999). *Ill-Posed Problems and Methods for Their Solution.* St.-Petersburg State University, St.-Petersburg (in Russian).

Tikhonov A. N. and Arsenin V. Ya. (1977). *Solution of Ill-Posed Problems.* Wiley, NY.

Tikhonov A. N., Leonov A. S., and Yagola A. G. (1997). *Nonlinear Ill-Posed Problems.* CRC Press UK, London.

Zubov V. I. (1974). *Mathematical Methods for Investigating into Automatic Control Systems.* Mashinostroenie, Leningrad (in Russian).

Supplementary reading

Bakushinsky A. and Goncharsky A. (1994). *Ill-Posed Problems: Theory and Applications.* Kluwer, Dordrecht.

Engl H. (1980). *Analyse und Numerik Schlecht Gesteller Probleme.* J. Kepler Univ., Linz.

Engl H. W., Hanke M., and Neubauer A. (1996). *Regularization of Inverse Problems.* Kluwer, Dordrecht.

Galaktionov M. A. (1999). On the hidden nature of the coarseness property of linear optimal-control systems. *Izv. VUZov. Electromekhanika*, 4, 48–50 (in Russian).

Galaktionov M. A. (2001). *Structural Instability in Linear Optimal-Control Systems.* St.-Petersburg State University, St.-Petersburg (in Russian).

Groetsch C. W. (1984). *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind.* Pitman, Boston.

Lavrent'ev M. M., Romanov V. G., and Shishatskii S. P. (1997). *Ill-Posed Problems of Mathematical Physics and Analysis.* AMS, Providence.

Morozov V. A. (1984). *Methods for Solving Incorrectly Posed Problems.* Springer-Verlag, NY.

Petrov Yu. P. (2000). *New Chapters of Control Theory.* St.-Petersburg State University, St.-Petersburg (in Russian).

Petrov Yu. P. (2002). *Control, Stability, and Optimization (Popular Science Sketches).* St.-Petersburg State University, St.-Petersburg (in Russian).

Petrov Yu. P. and Frolenkov D. B. (2000). Alteration of well-posedness in transformation of equations. *Vestnik Sankt-Peterb. Gos. Universiteta. Ser. 1*, 1, 52–57 (in Russian).

Sizikov V. S. (2001). *Mathematical Methods for Processing the Results of Measurements.* Politekhnika, St.-Petersburg (in Russian). Electronic version: Sizikov V. S. *Stable Methods for Processing the Results of Measurements.* `http://de.ifmo.ru/--books/SIZIKOV.PDF` or `http://dsp-book.narod.ru/SIZIKOV.pdf` .

Steklov V. A. (1927). *Foundations of the Theory of Integration of Ordinary Differential Equations.* GIZ, Moscow–Leningrad (in Russian).

Tararykin S. V. and Tyutikov V. V. (2002). Robust modal control of dynamic systems. *Avtomatika i Telemekhanika*, 5, 41–55 (in Russian).

Tikhonov A. N., Goncharsky A. V., Stepanov V. V., and Yagola A. G. (1995). *Mathematical Methods for the Solution of Ill-Posed Problems.* Kluwer, Dordrecht.

Verlan' A. F. and Sizikov V. S. (1986). *Integral Equations: Methods, Algorithms, Programs.* Naukova Dumka, Kiev (in Russian).

Wilkinson J. H. (1993). *The Algebraic Eigenvalue Problem.* Oxford Univ. Press, Oxford.

Part II

Stable methods for solving inverse problems

Chapter 4.

Regular methods for solving ill-posed problems

Before turning to the matter of correctness of various problems and equations, and before considering the solution methods for them, it is strongly recommended first to master the fundamentals of functional analysis and recall some facts from linear algebra.

4.1. ELEMENTS OF FUNCTIONAL ANALYSIS

Recall the basic notions of functional analysis (Verlan' and Sizikov, 1986; Kantorovic and Akilov, 1964; Kolmogorov and Fomin, 1981).

4.1.1. Some designations and definitions from the set theory

$a \in A$ or $a \notin A$ — the element a belongs or does not belong to the set A; $A \subset B$ — all elements of the set A are contained in the set B or, in other words, the set A is a *subset* of B; $A \subseteq B$ — the same, but, in addition, it may happen that $A = B$; $\varnothing$ — *null* (*empty*) *set*; $A \cup B$ — *sum of sets* — the set that consists of elements belonging either to A or B, or to both of the sets; $A \cap B$ — *product of sets* — the set that consists of elements belonging both to A and B; *countable set* is a set whose elements can be put in correspondence to natural numbers, or enumerated.

4.1.2. Topological spaces

A system of sets τ of a set X is called a *topology* in X if: 1) the set X itself and the empty set $\varnothing$ both belong to τ, 2) the sum $\bigcup_k G_k$ of any number and the product $\bigcap_{k=1}^n G_k$ of a finite number of sets from τ belong to τ. The set X with a topology τ defined on this set is called a *topological space* (X, τ) or, in short, T.

Sets that belong to τ are called *open sets.* The elements of a topological space are called *points. A vicinity* of a point $x \in T$ is any open set $G < T$ that contains the point x; a point $x \in T$ is called a *touch point of a set* $M \subset T$ if any vicinity of x contains at least one point from M; x is a *limiting point of a set* M if each vicinity of x contains at least one point from M different than x. The totality of all touch points of a set M is called the *closure of the set* M and is denoted as $\bar{M}$.

A set M of a topological space T is called an *everywhere dense set* if its closure is T, i. e. $\bar{M} = T$. A topological space with a countable, everywhere dense set is called a *separable space.*

4.1.3. Metric spaces

A particular case of topological spaces are metric spaces. A *metric space* $R = (X, \rho)$ or, in short, X, is a set X of elements (numbers, functions, etc.) such that, to each pair of members $x, y \in X$, a nonnegative real number $\rho(x, y)$ corresponds, called the *distance* between x and y; this number obeys the following *axioms*:

1) $\rho(x, y) \geq 0$ and, in addition, $\rho(x, y) = 0$ iff $x = y$;

2) $\rho(x, y) = \rho(y, x)$ (*symmetry axiom*);

3) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (*triangle axiom*).

Let x_n, $n = 1, 2, 3, \ldots$ be some infinite sequence of elements. An element x_0 is called the *limit of the sequence* x_n if $\rho(x_n, x_0) \to 0$ as $n \to \infty$. The notation is $\lim_{n\to\infty} x_n = x_0$. In the latter case, for any $\varepsilon > 0$ there exists a number N such that

$$\rho(x_m, x_n) \leq \varepsilon \qquad \text{for} \quad m, n \geq N; \tag{4.1.1}$$

the sequence x_n is called a *fundamental sequence.*

Generally speaking, the opposite statement, i. e., the statement that the sequence x_n has a limit does not follow from (4.1.1). With an additional condition stating that the existence of a limit of the sequence x_n follows from (4.1.1) (i. e., that any fundamental sequence converges), such a metric space is called a *full space.*

A set U of elements X is called a *compact set* if any infinite sequence of elements x_n from U contains a converging sequence. The closure of a compact set is called a *compact.*

Given two metric spaces X and Y, a mapping $y = Ax$ that puts certain elements $y \in Y$ in correspondence to elements $x \in X$ is called an *operator A* from X to Y. This mapping is denoted as $A : X \to Y$. A particular case of operators are *functionals*. Functionals are operators if Y is a space of numbers with the distance $\rho(y_1, y_2) = |y_1 - y_2|$, where $y_1, y_2 \in Y$.

4.1.4. Linear spaces

A *linear, or vector, space* L is a nonzero set of elements $x, y, z, \ldots$ that satisfies the following *conditions*:

I. To any two elements $x, y \in L$, a third element $z = x + y \in L$, called the *sum*, is put into one-to-one correspondence; in addition,

1) $x + y = y + x$ (*commutativity*),

2) $x + (y + z) = (x + y) + z$ (*associativity*),

3) there exists an element $0 \in L$ such that $x + 0 = x$ (existence of *zero*),

4) for each element $x \in L$, there exists such an element $-x$ such that $x + (-x) = 0$ (existence of the *opposite element*).

II. To each element $x \in L$ and to each number α, an element $\alpha x \in L$ (*product* of the element x and a scalar α) is put into correspondence; in addition,

5) $\alpha(\beta x) = (\alpha\beta)x$,

6) $1 \cdot x = x$,

7) $(\alpha + \beta)x = \alpha x + \beta x$,

8) $\alpha(x + y) = \alpha x + \alpha y$.

Elements $x, y, \ldots, w$ of a linear space L are called *linearly dependent elements* if there exist numbers $\alpha, \beta, \ldots, \lambda$, not all equal to zero, such that $\alpha x + \beta y + \cdots + \lambda w = 0$. Otherwise, these elements are called *linearly independent elements.*

4.1.5. Normalized spaces

A *normalized space* is a linear space L where the notion of norm is defined. The norm of an element $x \in L$ is denoted with the symbol $\|x\|$. The *norm* $\|x\|$ in L is a real nonnegative number such that

1) $\|x\| \geq 0$ and, in addition, $\|x\| = 0$ only if $x = 0$ (*non-degeneracy*),

2) $\|\alpha x\| = |\alpha| \cdot \|x\|$ (*homogeneity*),

3) $\|x + y\| \leq \|x\| + \|y\|$ (*triangle inequality*).

A normalized space where the notion of *distance* $\rho(x, y) = \|x - y\|$ is introduced becomes a *metric space.*

Let the set A that consists of elements $x, y, \ldots, w$ of a linear normalized space L be introduced. Then, the set C is called the *linear span of a set A* if its elements are linear combinations $\alpha x + \beta y + \cdots + \lambda w$, where $\alpha, \beta, \ldots, \lambda$ are arbitrary numbers.

4.1.6. Banach spaces

A normalized full space is called a *Banach space* (B-space, space of type B, space B).

Examples of linear normalized spaces that are Banach spaces, and designation of the norm in them.

1. Space $\mathbb{R}^1$ of real numbers:

$$\|x\| = |x|.$$

2. Real n-dimensional space $\mathbb{R}^n$ of numbers $x = (x_1, x_2, \ldots, x_n)$:

$$\|x\| = \Big\{ \sum_{k=1}^{n} x_k^2 \Big\}^{1/2}. \tag{4.1.2}$$

3. Space l_2 of the sequences of square-summed numbers $x = (x_1, x_2, \ldots, x_k, \ldots)$:

$$\|x\| = \Big\{ \sum_{k=1}^{\infty} x_k^2 \Big\}^{1/2}.$$

4. Space $L_2[a,b]$ of square-integrated functions $y(x)$:

$$\|y\| = \Big\{ \int_a^b |y(x)|^2 \,\mathrm{d}x \Big\}^{1/2} < \infty. \tag{4.1.3}$$

5. Space $L_p[a,b]$ $(p \geq 1)$:

$$\|y\| = \Big\{ \int_a^b |y(x)|^p \,\mathrm{d}x \Big\}^{1/p} < \infty.$$

6. Space $L_1[a,b]$ of modulus-integrated functions $y(x)$:

$$\|y\| = \int_a^b |y(x)| \,\mathrm{d}x < \infty. \tag{4.1.4}$$

7. Space $C[a,b]$ of functions continuous over a segment $[a,b]$:

$$\|y\| = \max_{a \leq x \leq b} |y(x)|. \tag{4.1.5}$$

8. Space $C^{(n)}[a,b]$ $(n \geq 1)$:

$$\|y\| = \max_{a \leq x \leq b} \sum_{k=0}^{n} \Big| \frac{\mathrm{d}^k y(x)}{\mathrm{d}x^k} \Big|.$$

9. Sobolev space $W_p^l[a,b]$ of l times continuously differentiable functions $y(x)$:

$$\|y\| = \Big\{ \sum_{k=0}^{l} \int_a^b \Big| \frac{\mathrm{d}^k y(x)}{\mathrm{d}x^k} \Big|^p \,\mathrm{d}x \Big\}^{1/p}.$$

10. Sobolev space $W_2^1[a,b]$, representing a particular case of space 9, namely, the space of functions $y(x)$ that have square-integrated derivative:

$$\|y\| = \Big\{ \int_a^b y^2(x) \,\mathrm{d}x + \int_a^b y'^2(x) \,\mathrm{d}x \Big\}^{1/2}.$$

4.1.7. Spaces with scalar product (Euclidean space, Hilbert space, etc.)

Euclidean space. A real linear space E is called an *Euclidean space* if a real number (x, y), *scalar product*, is put in correspondence to each pair of its elements x and y; the real number (x, y) is chosen such that to obey the following axioms:

1) $(x, y) = (y, x)$ (*symmetry*),

2) $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$ (*additivity*),

3) $(\lambda x, y) = \lambda(x, y)$ (*homogeneity*),

4) $(x, x) \geq 0$, and $(x, x) = 0$ only if $x = 0$.

Euclidean space with the norm defined as

$$\|y\| = \sqrt{(y, y)}$$

turns into a *normalized space.*

Let E be a space with scalar product. A system of nonzero elements $x_1, x_2, \ldots, x_m \in E$ is called an *orthogonal system* if $(x_k, x_l) = 0$ for $k \neq l$. If the system $x_1, x_2, \ldots, x_m$ is such that

$$(x_k, x_l) = \begin{cases} 0, & k \neq l, \\ 1, & k = l, \end{cases}$$

then this system is called orthonormalized system.

Hilbert space. A space H with scalar product is called a *Hilbert space* if this space is full in the norm generated by scalar product. An alternative definition is as follows: a *Hilbert space* is a Banach space where the norm is defined as the scalar product $\|y\| = \sqrt{(y, y)}$ or $\|x - y\| = \sqrt{(x - y, x - y)}$.

Spaces 1–4 and 10 are Hilbert spaces.

A set $A \subseteq H$ is called a *linearly independent space* if each finite collection of its elements is a linearly independent one. A set $A \subseteq H$ is called the *basis* of the Hilbert space H if this set is a linearly independent set and the closure of the linear span A coincides with H. A basis is called an *orthogonal basis* if the system of its elements is an orthogonal system, and an *orthonormalized basis* if the system of its elements is an orthonormalized system.

4.1.8. Operators in Hilbert space

Given some sets Y and F, suppose that a subset $D \subseteq Y$ is allotted in Y. Provided that to each element $y \in D$ a certain element $f \in F$ is put into correspondence, then the *operator* $f = Ay$ is defined. In this case, the set D is called the *domain of the operator* A; for this set, the designation $D(A)$ or $\operatorname{Im} A$ is adopted. The set $R \equiv R(A) = \{f \in F \mid f = Ay,\ y \in D\}$ is called the *value area of the operator* A. The action of the operator A is designated as $A : Y \to F$.

An operator A that acts from a Banach space B_1 to a Banach space B_2 is called a *linear operator* if for $y_k \in D(A)$ and for arbitrary numbers c_k, $k = 1, \ldots, m$, the equality $A(c_1 y_1 + \cdots + c_m y_m) = c_1 A y_1 + \cdots + c_m A y_m$ holds. The operator A is called a *bounded operator* if $\|Ay\| \le g\|y\|$ for any $y \in D(A)$, where g is a finite number. An example of a linear bounded operator is the linear integral operator $A : L_2[a,b] \to L_2[a,b]$ that generates the function

$$Ay = \int_a^b K(x,s)y(s)\,\mathrm{d}s \equiv f(x), \qquad a \le x \le b.$$

The norm of the operator is defined as

$$\|A\| = \sup_{y \in D(A)} \frac{\|Ay\|}{\|y\|} = \sup_{\|y\|=1} \|Ay\|. \tag{4.1.6}$$

An operator B that acts from a Hilbert space H into H is called the *conjugate operator* with respect to the operator $A : H \to H$ if $(Ax,y)_H = (x,By)_H$ for any $x,y \in H$; this operator is denoted as A^*. For instance, the conjugate operator for the integral operator A is the operator A^* that generates the function

$$A^* y = \int_a^b K^*(x,s)y(s)\,\mathrm{d}s = \int_a^b \overline{K(s,x)}y(s)\,\mathrm{d}s,$$

where the bar is used to denote complex conjugation.

A *self-conjugate (Hermitean) operator* is a linear bounded operator A: $H \to H$ such that for any $x, y \in H$ the equality $(Ax,y)_H = (x,Ay)_H$ holds. The *unit operator* is the operator E (or I) such that $Ey = y$ for any $y \in D(A)$. An *unitary operator* is an operator U such that $U^*U = E$. A self-conjugate operator A is called a *positively defined (positive) operator* if $(Ay,y) > 0$ for any elements $y \in D(A)$. An example of a positively defined operator is the operator A^*A. The *inverse operator* with respect to A (this operator is denoted as A^{-1}) is the operator defined on $R(A)$ and putting in correspondence to each element $f \in R(A)$ a certain element $y \in D(A)$. The latter is denoted as $y = A^{-1}f$. From linearity of A, it follows that the operator A^{-1} is also linear. Yet, generally speaking, boundedness of A does not imply that the operator A^{-1} is also a bounded operator (the latter case is typical, first of all, for ill-posed problems).

Operators A and B are called *distributive* (*permutable*) *operators* if $AB = BA$.

A linear operator $A : Y \to F$ is called *continuous* at a point $y_0 \in Y$ if $Ay \to Ay_0$ as $y \to y_0$. For linear operators, continuousness and boundedness are equivalent notions.

The linear operator A is called *quite continuous* if it transforms any norm-bounded set in a compact set.

Let X be a linear space and A be a linear operator acting from X into X with a domain of operator $D(A)$. A number λ is called the *eigenvalue of the operator* A if there exists a vector $x \neq 0$, $x \in D(A)$ such that $Ax = \lambda x$. In this case, the vector x is called the *eigenvector* (*eigenfunction*) *of the operator* A which corresponds to the given eigenvalue λ. If A is a linear self-conjugate operator, then all λ are real, and $\|A\| = |\lambda(A)|_{\max}$, $\|A^{-1}\| = 1/|\lambda(A)|_{\min}$. If the operator A is positive, then all λ are real and nonnegative, and $\|A\| = \lambda(A)_{\max}$, $\|A^{-1}\| = 1/\lambda(A)_{\min}$. If A is an arbitrary (linear) operator, then all λ are, generally speaking, complex numbers, and $\|A\| = \sqrt{\lambda(A^*A)_{\max}}$, $\|A^{-1}\| = \sqrt{1/\lambda(A^*A)_{\min}}$. The totality of all eigenvalues of an operator is called the *spectrum of the operator*.

4.2. SOME FACTS FROM LINEAR ALGEBRA

Recall some facts from linear algebra (Bronshtein and Semendyaev, 1986, pp. 156–161; Verlan' and Sizikov, 1986, pp. 504–509; Sizikov, 2001, pp. 142–147; Voevodin, 1980; Gantmacher, 1959; Kurosh, 1975; Wilkinson, 1993).

4.2.1. Vector and matrix

A *rectangular* $m \times n$*-matrix* is a set of (generally speaking, complex) numbers arranged in a rectangular table that has m rows and n columns:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \equiv (a_{ij}), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

If $m > 1$ and $n = 1$, then A is a *column vector*, and if $m = 1$ and $n > 1$, then A is a *row vector*. If $m = n$, then the matrix A is called a *square matrix.* If $a_{ij} = 0$ for $i \neq j$, then a square matrix A is called a *diagonal* matrix. If $a_{ij} = 0$ for $i \neq j$ and $a_{ij} = 1$ for $i = j$, then the square matrix A is called the *unit matrix* and is denoted as E or I. If all $a_{ij} = 0$, then the matrix A is called the *zero matrix.*

A matrix $A^\top = (b_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, composed of the elements $b_{ij} = a_{ji}$ is called the *transposed matrix* with respect to the $m \times n$-matrix A. In particular, transposition transforms a column vector in a row vector, and vice versa. The matrix $\bar{A} = (\bar{a}_{ij})$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, is called the *complex conjugate* matrix for the matrix A. The matrix $A^* = \bar{A}^\top$ is called the *conjugate* (or *Hermitian-conjugate*), or *conjugate and transposed* matrix for the matrix A. In this case, $(A^*)^* = A$. If A is a real matrix, then $A^* = A^\top$.

The *determinant* of a square matrix A is denoted as $|A|$ or $\det(A)$. A *minor* of order k of a matrix A is the determinant of the k-th order composed of any part of A with preserved arrangement of the elements a_{ij}. The *rank* $r = \operatorname{rang}(A)$ of a matrix A is the maximum order $k_{\max}$ of the nonzero minors of A.

A *banded* $m \times n$-matrix $A = (a_{ij})$ is a matrix such that $a_{ij} = 0$ for $|i - j| \geq \Delta$ and, in addition, $\Delta < \max(m, n)$. A *quasi-diagonal matrix* A is a square $n \times n$-matrix that has square cells of order $< n$ at its principal diagonal, while all other elements of the matrix are zero. The banded and quasi-diagonal (or *sparse*) matrices are opposed by *dense* matrices, or by matrices with comparatively small number of zero elements.

A square matrix A is called *degenerate* (or *singular*) if $|A| = 0$; otherwise, the matrix A is *nongenerate.*

A square $n \times n$-matrix A^{-1} is called the *inverse matrix* to a square $n \times n$-matrix A if $A^{-1}A = E$. A necessary and sufficient condition for the existence of A^{-1} is nongeneracy of A.

A square $n \times n$-matrix $A = (a_{ij})$ is called *left*, or *lower* (or, correspondingly, *right*, or *upper*) *triangular* matrix if $a_{ij} = 0$ for $j > i$ (for $j < i$). For a triangular $n \times n$-matrix A, we have: $|A| = \prod_{i=1}^{n} a_{ii}$ (the determinant of the matrix is equal to the product of the diagonal elements of the matrix).

The sum of the diagonal elements, $a_{11} + \dots + a_{nn}$, is called the *spur* of the square $n \times n$-matrix $A = (a_{ij})$ and is denoted as $\operatorname{Sp} A$ or $\operatorname{Tr} A$.

A square $n \times n$-matrix $A = (a_{ij})$ is called a *symmetric* matrix if $a_{ij} = a_{ji}$, $i, j = 1, \dots, n$. A real symmetric $n \times n$-matrix $A = (a_{ij})$ is called a *positively defined* (*positive*) *matrix* if $\sum_{i,j=1}^{n} a_{ij} x_i x_j > 0$ for all real x_i and $\sum_{i,j=1}^{n} a_{ij} x_i x_j = 0$ only if $x_1 = x_2 = \dots = x_n = 0$. Examples of positively defined matrices are the matrices E, $A^\top A$, $AA^\top$, A^*A, and AA^* (A is a rectangular matrix).

A square matrix A is called a *normal matrix* if $A^*A = AA^*$. A squared complex matrix A is called a *Hermitean* (*self-conjugate*) matrix if $A = A^*$. A real Hermitean matrix is symmetric. A *unitary* matrix is a complex matrix A such that $A^*A = E$ or $AA^* = E$. A real unitary matrix is called an *orthogonal* matrix ($A^\top A = E$ or $AA^\top = E$).

Any arbitrary nongenerate square matrix A can be represented as $A = LU$, where L and U are respectively a lower and an upper triangular matrix. If, additionally, A is a banded matrix, then L and U are triangular banded matrices.

A positively defined matrix A can be represented as $A = LL^\top$, where L is a lower triangular matrix, in particular, a matrix with positive diagonal elements (*Cholesky scheme*) or as $A = LDL^\top$, where L is a lower triangular matrix with unit diagonal, and D is a diagonal matrix with positive elements. If, additionally, the positively defined matrix A is a banded matrix, then this matrix it can be represented as $A = LL^\top$, where L is a lower triangular banded matrix.

4.2.2. Eigenvalues and singular numbers of matrices

The roots λ_i, $i = 1, \dots, n$, of the *characteristic equation*

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0 \tag{4.2.1}$$

are called the *eigenvalues* of the square $n \times n$-matrix $A = (a_{ij})$. The quantities $1/\lambda_i$, $i = 1, \dots, n$, are called *characteristic numbers*. The *singular numbers* of an $m \times n$-matrix A are real nonnegative numbers $\mu_i(A) = \sqrt{\lambda_i(A^*A)}$,

$i = 1, \ldots, n$, normally arranged in non-increasing order (if the matrix A^*A is degenerate and has a rank $r < n$): $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r > \mu_{r+1} = \cdots = \mu_n = 0$ or (if the matrix A^*A is non-degenerate): $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n > 0$. The following equality holds: $|\det(A)| = \mu_1\mu_2\cdots\mu_n$.

The eigenvalues λ_i of an arbitrary (square) matrix A are, generally speaking, complex numbers. If the matrix A is symmetric, then all λ_i are real numbers (of either sign, in principle). If the matrix A is positive and symmetric, then all λ_i are real and nonnegative. All singular numbers μ_i of an arbitrary matrix are real and nonnegative. In addition, in all cases λ_i and μ_i may have order amounting to n. The set of eigenvalues of a matrix is called its *spectrum*.

If A is a rectangular $m \times n$-matrix, then, instead of eigenvalues, they usually use singular numbers $\mu_i(A)$. If the rank of the matrix $r < n$, then the matrix A is *degenerate*, or *singular*; in this case, the determinant $\det(A) \equiv |A| = 0$ and the inverse matrix A^{-1} (for $m = n$) or $(A^*A)^{-1}$ (in the general case) does not exist. If $r = n$, then the matrix A is nondegenerate, its determinant is $|A| = \sqrt{|A^*A|} = \mu_1\mu_2\cdots\mu_n$, and the inverse, or reciprocal, matrix A^{-1} (for $m = n$) or $(A^*A)^{-1}$ (in the general case) exists. If $m = n$,

$$A^{-1} = \frac{1}{|A|}\begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix},$$

where A_{ij} are the algebraic adjuncts.

4.2.3. Norms of vectors and matrices

Normally, it is common practice to first introduce the notion of the *norm of vector*, $\|x\|$, and, then, another, subordinate, notion, that of the *norm of matrix*, $\|A\|$, generally defined as (cp. (4.1.6))

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

Given below are some specific norms of a vector or a matrix.

Let $x = (x_1, x_2, \ldots, x_n)$ be a row vector and $x^\top = (x_1, x_2, \ldots, x_n)^\top$ be a column vector in the real n-dimensional space $\mathbb{R}^n$. In this case, the *Euclidean* (or *Schurean*) *norm of the vector* x is (cp. (4.1.2))

$$\|x\| = \Big(\sum_{k=1}^{n} x_k^2\Big)^{1/2} = \sqrt{xx^\top}. \tag{4.2.2}$$

The Euclidean norm of a real square $n \times n$-matrix A generated by norm (4.2.2) is

$$\|A\| = \Big(\sum_{i,j=1}^{n} |a_{ij}|^2\Big)^{1/2} = \mu(A)_{\max}.$$

If x is a complex vector and A is a complex matrix, then norm (4.2.2) is

$$\|x\| = \Big(\sum_{k=1}^{n} |x_k|^2\Big)^{1/2} = \sqrt{xx^*};$$

in the latter case, the norms $\|x\|$ and $\|A\|$ are called *Hermitean* norms.

If A is a positively defined symmetric matrix, then

$$\|A\| = \mu(A)_{\max} = \lambda(A)_{\max}, \qquad \|A^{-1}\| = 1/\mu(A)_{\min} = 1/\lambda(A)_{\min}.$$

If A is a symmetric matrix, then

$$\|A\| = \mu(A)_{\max} = |\lambda(A)|_{\max}, \qquad \|A^{-1}\| = 1/\mu(A)_{\min} = 1/|\lambda(A)|_{\min}.$$

If A is an arbitrary (for instance, a rectangular) matrix, then

$$\|A\| = \mu(A)_{\max} = \sqrt{\lambda(A^*A)_{\max}}, \qquad \|A^{-1}\| = 1/\mu(A)_{\min} = 1/\sqrt{\lambda(A^*A)_{\min}}.$$

They also use the *"octahedral" norm of vector* (see (4.1.4))

$$\|x\| = \sum_{k=1}^{n} |x_k|$$

and the related "octahedral" norm of $m \times n$-matrix

$$\|A\| = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|,$$

and also the *"cubic" norm of vector* (see (4.1.5))

$$\|x\| = \max_{1\le k\le n} |x_k|$$

and the related "cubic" norm of $m \times n$-matrix

$$\|A\| = \max_{1\le j\le m} \sum_{j=1}^{n} |a_{ij}|,$$

also called *maximum-norms.*

Sometimes, they first introduce the notion of the norm of matrix, and derive from it the notion of the norm of vector.

4.2.4. Multiplication of matrices and vectors

The multiplication of two rectangular matrices is

$$\underset{m\times l}{Z} = \underset{m\times n}{X} \times \underset{n\times l}{Y}$$

or, in more detail,

$$z_{ik} = \sum_{j=1}^{n} x_{ij} y_{jk}, \qquad i = 1, \dots, m, \quad k = 1, \dots, l. \tag{4.2.3}$$

Multiplication of a matrix by a vector (a particular case of multiplication of matrices) is

$$\underset{m\times 1}{f} = \underset{m\times n}{A} \times \underset{n\times 1}{y}$$

or, in more detail,

$$f_i = \sum_{j=1}^{n} a_{ij} y_j, \qquad i = 1, \dots, m. \tag{4.2.4}$$

4.2.5. SLAE, condition number, LSM, and pseudo-inverse matrix

Consider a *system of linear algebraic equations* (*SLAE*):

$$Ay = f \tag{4.2.5}$$

or

$$\sum_{j=1}^{n} a_{ij} y_j = f_i, \qquad i = 1, \dots, m, \tag{4.2.6}$$

where A is an $m \times n$-matrix, y is an n-column vector, and f is an m-column vector. If $m > n$ (or, more precisely, the number of linear independent rows in (4.2.5) is greater than n), then the SLAE is called an *overdetermined* system. If $m = n$ (and $r = m$), then the SLAE is called a *determined* system. If, alternatively, $m < n$, then the SLAE is called an *underdetermined* system.

To analyze an SLAE, we can also use another, more stringent criterion. To this end, along with the *rank* $r = \operatorname{rang}(A)$ of the matrix A (the maximum order of the nonzero minors of A), we introduce the number $\rho = \operatorname{rang}(A \mid f)$, the *rank of extended matrix.* Then, if $\rho > r$, then the SLAE has no solutions and presents an overdetermined system. If $\rho = r$, then in the case of $\rho = n$ the SLAE has a single solution and presents a determined system; if $\rho < n$, then the SLAE has many solution and presents an underdetermined system.

The solution of a determined SLAE is symbolically denoted as $y = A^{-1}f$, where A^{-1} is the inverse matrix; in practice, to solve an SLAE, the Gauss method (or other methods) are used, which invoke the product decomposition $A = LU$; in the latter case, the problem is reduced to solving two SLAEs with triangular matrices $Lx = f$ and $Uy = x$. If the matrix A is positive, then most efficient methods are the Cholesky method and the Kraut method, or the "square root", method.

If, instead of the exact f and A, we have an approximate $\tilde{f}$ and $\tilde{A}$ such that $\|\tilde{f} - f\| \le \delta$ and $||\tilde{A} - A|| \le \xi$ (δ and ξ are the errors in setting the right-hand side and the matrix or, more precisely, their upper bounds), then a frequently used estimate of the relative solution error is

$$\frac{\|\delta y\|}{\|y\|} \le \operatorname{cond}(A)\Big(\frac{\delta}{\|f\|} + \frac{\xi}{\|A\|}\Big), \tag{4.2.7}$$

where $\operatorname{cond}(A) = \operatorname{cond}(A^{-1}) = \|A\| \cdot \|A^{-1}\| = \mu(A)_{\max}/\mu(A)_{\min} \ge 1$ is the *condition number* of A. If $\operatorname{cond}(A)$ is a relatively small number (normally, $< 10^3$), then the matrix A (and the SLAE) is called a *well-conditioned* matrix (and system). Otherwise, if $\operatorname{cond}(A)$ is a relatively large number (normally, $> 10^4$), then the matrix A (and the SLAE) is called an *ill-conditioned* matrix (and system). Note that smallness (compared to unity) of the determinant $|A|$ is not, generally speaking, a criterion for ill-conditionality.

The solution of an overdetermined SLAE (pseudo-solution) can be found by the Gauss *least-squares method* (*LSM*); in this method, instead of (4.2.5), one solves the so-called *normal SLAE*

$$A^*Ay = A^*f \tag{4.2.8}$$

with the square positively defined $n \times n$-matrix A^*A. SLAE (4.2.8) results from the condition

$$\|Ay - f\|^2 = \min_y. \tag{4.2.9}$$

The quantity $\|Ay - f\|$ or $\|Ay - f||^2$ is called the *discrepancy* (*residual*) between the left-hand and right-hand sides of (4.2.5). A solution y that satisfies condition (4.2.9) or, in other words, minimizes the discrepancy, is called the *pseudo-solution*. For more detail see Section 4.6.

An underdetermined SLAE has many solutions. The only (normal) solution of such a system can be found by the Moore–Penrose *pseudo-inverse matrix method*:

$$y = A^+f,$$

where A^+ is the *pseudo-inverse* $n \times m$*-matrix*. For more detail see Section 4.6.

4.3. BASIC TYPES OF EQUATIONS AND TRANSFORMATIONS

Listed below are the main types of equations (or system of equations) that will be considered in subsequent sections of Part II.

4.3.1. Systems of linear algebraic equations

A system of m linear algebraic equations (SLAE) (Bronshtein and Semendyaev, 1986, p. 161; Sizikov, 2001, p. 140; Voevodin, 1980; Gantmacher, 1959; Kurosh, 1975; Wilkinson, 1993) for n unknowns can be written in the form of (4.2.5) or (4.2.6), where A is an $m \times n$-matrix, y is the unknown $n \times 1$-column vector, and f is the right-hand side (a set $m \times 1$-column vector), or, in more detail,

$$\left.\begin{aligned} a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n &= f_1, \\ a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n &= f_2, \\ &\vdots \\ a_{m1}y_1 + a_{m2}y_2 + \cdots + a_{mn}y_n &= f_m. \end{aligned}\right\} \tag{4.3.1}$$

Some propositions about SLAEs were given in Section 4.2. It should be noted here that most hard to solve numerically are degenerate and ill-conditioned SLAEs (Tikhonov and Arsenin, 1977).

Below, several definitions of degenerate SLAEs are given.

1. Suppose that $m = n$, i.e., the matrix of the system is a square matrix. In this case, the SLAE is called a *degenerate* SLAE if its determinant is zero, i.e., $|A| = 0$.

2. Suppose that, generally speaking, $m \neq n$. Then, the SLAE is called a *degenerate* SLAE if its $m \times n$-matrix A has at least one zero singular number μ. In the latter case, the condition number of A is infinity, i.e., $\operatorname{cond}(A) = \mu(A)_{\max}/\mu(A)_{\min} = \infty$. This definition also applies to the case of $m = n$, i.e., it is a general definition.

The notion of an ill-conditioned SLAE is defined less stringently. An SLAE is an *ill-conditioned SLAE* if the condition number $\operatorname{cond}(A)$ of its matrix A is a sufficiently large, although finite, number. The words "sufficiently large" have different meanings for different practical problems. Relation (4.2.7) shows that if, for instance, the relative error in setting the initial data (the right-hand side f and the matrix A) is of order 10^{-4} and it is required to find the solution y with a relative error of 10^{-2}, and the condition number cond (A) is of order 10^4, then the required error of y is unachievable, the condition number should be regarded as a "sufficiently large" number and the SLAE, an ill-conditioned SLAE. At the same time, if the relative error in setting the initial data is of order 10^{-6}, then the value cond $(A) \sim 10^4$ is sufficient for a relative solution error of order 10^{-2} be achieved; in this case, the condition number can be regarded a "small" number and the SLAE, a well-conditioned SLAE.

Note the following. If the calculations are to be carried out accurate to some finite accuracy, then it is often impossible to establish whether the given SLAE is a degenerate SLAE (i. e., $\mu(A)_{\min} = 0$) or an ill-conditioned SLAE (i. e., $\mu(A)_{\min} > 0$). That is why degenerate and ill-conditioned SLAEs are normally considered uniformly and treated with unified (stable) methods. For more detail concerning these methods see Section 4.7.

4.3.2. Systems of ordinary differential equations

Consider a *system of ordinary differential equations* (Bronshtein and Semendyaev, 1986, p. 306; Pontryagin, 1982) written in the normal Cauchy form

$$\frac{dy_i(t)}{dt} = f_i(t, y_1, \ldots, y_n), \qquad i = 1, \ldots, n, \tag{4.3.2}$$

with some initial conditions (Cauchy problem)

$$y_i(t^0) = y_i^0. \tag{4.3.3}$$

For such a system, the Cauchy theorem about solution existence and uniqueness holds (Bronshtein and Semendyaev, 1986, p. 306).

Cauchy theorem. *Suppose that the following conditions are fulfilled:*

1) the functions $f_i(t, y_1, \ldots, y_n)$ are continuous and bounded (i. e., $|f_i| \le A$) in a closed domain $G = \{|t - t^0| \le a,\ |y_i - y_i^0| \le b,\ i = 1, \ldots, n\}$;

2) in the domain G, the Lipschitz condition is fulfilled:

$$|f_i(t, \tilde{y}_1, \ldots, \tilde{y}_n) - f_i(t, y_1, \ldots, y_n)| \le L \sum_{k=1}^{n} |\tilde{y}_k - y_k|, \qquad i = 1, \ldots, n,$$

where L is the Lipschitz constant. Then, system (4.3.2) *with initial conditions* (4.3.3) *has a unique solution if $|t - t^0| \le \alpha$, where $\alpha = \min(a, b/A)$.*

Remark. The Lipschitz condition is fulfilled if, in G, the functions f_i have bounded partial derivatives with respect to y_k, i. e., then

$$\left|\frac{\partial}{\partial y_k}\, f_i(t, y_1, \dots, y_n)\right| \le M, \qquad i, k = 1, \dots, n. \tag{4.3.4}$$

Consider now the matter of *solution stability* for system (4.3.2) (Bronshtein and Semendyaev, 1986, pp. 325–326). System (4.3.2) can be represented as

$$\frac{dy_i(t)}{dt} = \sum_{j=1}^{n} a_{ij} y_j(t) + \varphi_i(t; y_1, \dots, y_n), \qquad i = 1, \dots, n, \tag{4.3.5}$$

with $a_{ij} = (\partial f_i/\partial y_j)(t; 0, \dots, 0) = \mathrm{const}$, $i, j = 1, \dots, n$, i. e., a_{ij} do not depend on t. The system

$$\frac{dy_i(t)}{dt} = \sum_{j=1}^{n} a_{ij} y_j(t), \qquad i = 1, \dots, n, \tag{4.3.6}$$

is called a system *linearized* with respect to (4.3.5).

We call $y_i = y_i(t; t^0, y_1^0, \dots, y_n^0)$, $i = 1, \dots, n$, *the undisturbed solution of the system* (solution of (4.3.2) with initial data (4.3.3)), and $\tilde{y}_i = y_i(t; t^0, \tilde{y}_1^0, \dots, \tilde{y}_n^0)$, $i = 1, \dots, n$, the *disturbed solution of the system* (solution of (4.3.2) with disturbed initial data $\tilde{y}_1^0, \dots, \tilde{y}_n^0$). Next, we call the difference $\bar{y}_i = \tilde{y}_i - y_i$ between the disturbed and undisturbed solutions the *trivial solution.*

A *sufficient condition for stability* of the trivial solution is given by the following theorem.

Theorem. *Let*

1) all roots of the characteristic equation of linearized system (4.3.6)*, i. e., all roots λ of* (4.2.1)*, or of the equation* $\det(a_{ij} - \lambda\delta_{ij}) = 0$*, have negative real parts:* $\operatorname{Re}\lambda_i < 0$*, $i = 1, \dots, n$;*

2) all functions $\varphi_i(t; y_1, \dots, y_n)$ satisfy the condition

$$|\varphi_i(t; y_1, \dots, y_n)| \le M\Big\{\sum_{i=1}^{n} y_i^2\Big\}^{1/2+\alpha}, \qquad i = 1, \dots, n,$$

where $M = \mathrm{const}$ *and* $\alpha > 0$*. Then, the trivial solution of* (4.3.2) *is a Lyapunov-stable solution. Otherwise, if at least one root λ of the characteristic equation has a positive real part (and the functions φ_i satisfy condition 2), then the trivial solution of* (4.3.2) *is unstable.*

Below, we will also consider a system of ordinary differential equations with a control function (Tikhonov and Arsenin, 1977):

$$\frac{dy(t)}{dt} = f(t, y, u), \tag{4.3.7}$$

where $y(t) = \{y_1(t), \ldots, y_n(t)\}$ is the unknown vector-function considered over an interval $t_0 \le t \le T$, $f(t) = \{f_1(t), \ldots, f_n(t)\}$ is the right-hand side (vector-function), and $u(t) = \{u_1(t), \ldots, u_m(t)\}$ is a control vector-function. The *initial conditions* are

$$y(t_0) = y_0, \tag{4.3.8}$$

where y_0 is a given vector.

Further consideration of systems of ordinary differential equations will be given in Sections 4.4 and 4.7.

4.3.3. Partial differential equations

Among the variety of partial differential equations (Bronshtein and Semendyaev, 1986, p. 340; Ivanov, Vasin, and Tanana, 2002; Lavrent'ev, Romanov, and Shishatskii, 1997; Tikhonov and Arsenin, 1977; Lavrent'ev, 1981; Tichonov and Samarskij, 1963), consider most typical equations, namely, the Laplace equation and the heat conduction equation.

In the two-dimensional case, the (elliptic) *Laplace equation* is of the form

$$\Delta u(x, y) \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0; \tag{4.3.9}$$

here, we consider the *boundary conditions* (*Cauchy problem*):

$$u(x, 0) = f(x), \qquad \frac{\partial u}{\partial y}\bigg|_{y=0} = \varphi(x), \tag{4.3.10}$$

where $f(x)$ and $\varphi(x)$ are some given functions.

Consider also the one-dimensional (parabolic) *heat conduction equation.* We will consider the following four statements of the problem.

1-st statement of the problem (*with forward time*). It is required to solve the *heat conduction equation*

$$\frac{\partial u(x, t)}{\partial t} = \frac{\partial^2 u(x, t)}{\partial x^2}, \qquad 0 \le x \le l, \tag{4.3.11}$$

with the initial condition $u(x,0) = \varphi(x)$ and the boundary conditions $u(0,t) = \psi_1(t)$ and $u(l,t) = \psi_2(t)$. The function $u(x,t)$ gives the temperature distribution along a rod at the time t provided that the initial (at $t = 0$) distribution was $\varphi(x)$ and the ends of the rod are kept under the temperature regimes $\psi_1(t)$ and $\psi_2(t)$. The equation is to be solved from the time $t = 0$ towards increasing t; this statement is therefore called the *direct problem* or the *problem with forward time.*

2-nd statement of the problem (*with reverse time*). In the inverse problem, it is required, by solving the problem towards decreasing t, to find the initial temperature distribution $u(x,0) = \varphi(x)$ from the temperature distribution along the rod known at some time $t_* > 0$, i. e., from the known function $u(x,t_*)$. Here, the boundary conditions $\psi_1(t)$ and $\psi_2(t)$ are not set.

3-rd statement of the problem. Here, to be treated is the heat conduction problem posed as a Cauchy problem: it is required to solve the heat conduction equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \qquad -\infty < x < \infty, \quad 0 \le t \le T, \tag{4.3.12}$$

with the initial conditions $u(x,0) = \varphi(x)$ and $u(x,T) = \chi(x)$, where $\varphi(x)$ and $\chi(x)$ are some given functions.

4-th statement of the problem. Like the third statement, this statement considers a Cauchy problem: it is required to solve heat conduction equation (4.3.12), but with a single initial condition $u(x,0) = \varphi(x)$.

Solution of partial differential equations will be analyzed in Section 4.4; stable solution methods will be considered in Section 4.7.

4.3.4. Integral equations

Below, only several types of integral equations (Verlan' and Sizikov, 1986; Ivanov, Vasin, and Tanana, 2002; Lavrent'ev, Romanov, and Shishatskii, 1997; Sizikov, 2001; Tikhonov and Arsenin, 1977; Tikhonov, Arsenin, and Timonov, 1987; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) will be considered; namely, equations of the first kind, whose solution problem, as a rule, lacks stability.

The *one-dimensional Fredholm integral equation of the first kind* is

$$\int_a^b K(x,s)y(s)\,\mathrm{d}s = f(x), \qquad c \le x \le d, \tag{4.3.13}$$

where $K(x,s)$ is the kernel (some given function), $f(x)$ is the right-hand side (some measured function), and $y(s)$ is the unknown function. The types of spaces to which the functions K, f and y belong will be indicated in what follows. The kernel $K(x,s)$ in (4.3.13) is a Fredholm kernel.

Definition. A kernel $K(x,s)$ is a *Fredholm kernel* if $K(x,s)$ is a function continuous in the rectangle $\Pi = \{a \le s \le b,\ c \le x \le d\}$, a closed function (i. e., the homogeneous equation (equation (4.3.13) with $f(x) \equiv 0$) has only zero solution), and a problem satisfying the Hilbert–Schmidt condition

$$\int_a^b \int_c^d |K(x,s)|^2 \, \mathrm{d}x \, \mathrm{d}s < \infty. \tag{4.3.14}$$

In practical problems, $y(s)$ is normally the input process or the signal to be reconstructed, $f(x)$ is the measured output process, and $K(x,s)$ is the instrumental function of the meter.

One-dimensional first-kind Volterra integral equation:

$$\int_a^x K(x,s)y(s)\,\mathrm{d}s = f(x), \qquad a \le x \le b, \tag{4.3.15}$$

where the functions K, f and y have the same meaning as in (4.3.13).

One-dimensional first-kind convolution-type Fredholm integral equation:

$$\int_{-\infty}^{\infty} K(x-s)y(s)\,\mathrm{d}s = f(x), \qquad -\infty < x < \infty. \tag{4.3.16}$$

Two-dimensional first-kind convolution-type Fredholm integral equation:

$$\begin{gathered}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(x_1-s_1, x_2-s_2)y(s_1,s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2 = f(x_1,x_2), \\ -\infty < x_1, x_2 < \infty.\end{gathered} \tag{4.3.17}$$

The integral equations will be further considered in Sections 4.4–4.8 and in Chapter 5.

4.3.5. Operator equations

All the above-indicated equations and their systems (4.3.1), (4.3.2), (4.3.7), (4.3.9), (4.3.11)–(4.3.13), (4.3.15)–(4.3.17) can be represented as a general *first-kind operator equation*:

$$Ay = f, \qquad y \in Y, \quad f \in F, \tag{4.3.18}$$

where A is some given linear operator (algebraic, differential, integral, etc.), f is the right-hand side of the equation (an element belonging to some space F), y is the sought solution (an element belonging to some space Y), and Y and F are some Hilbert spaces, for instance, the spaces L_2 and C (see (4.1.3) and (4.1.5)). The apparatus of operator equations will be extensively used in Sections 4.4 and 4.7.

4.3.6. Fourier, Hartley, and Laplace transformations

Consider some *integral transformations* (Bronshtein and Semendyaev, 1986; Verlan' and Sizikov, 1986; Sizikov, 2001; Bracewell, 1986; Vasil'ev and Gurov, 1998; Rabiner and Gold, 1975).

Continuous Fourier transformation. Given some piecewise-continuous function (an initial process) $y(t)$, $-\infty < t < \infty$, where t is time, a linear coordinate, an angular coordinate, etc. (if t is time, then $y(t)$ is a time-dependent process). Then, the integral

$$Y(\omega) = \int_{-\infty}^{\infty} y(t)\mathrm{e}^{i\omega t}\,\mathrm{d}t, \qquad -\infty < \omega < \infty, \tag{4.3.19}$$

is called the *one-dimensional direct continuous Fourier transform* (CFT) or, for short, the *Fourier transform* (FT), *Fourier image*, *image according to Fourier*, *spectrum*, etc. The function $y(t)$ is called the *inverse Fourier transform* (IFT), or the original. The variable ω is called the *Fourier frequency.* The function $y(t)$ is either a real or complex function, and $Y(\omega)$ is generally a complex function that results from the following operations.

We apply the *Euler formula*

$$\mathrm{e}^{i\varphi} = \cos\varphi + i\sin\varphi \tag{4.3.20}$$

to obtain

$$\mathrm{e}^{i\omega t} = \cos\omega t + i\sin\omega t. \tag{4.3.21}$$

If $y(t)$ is a real function, then expression (4.3.19) can be written as

$$Y(\omega) = \int_{-\infty}^{\infty} y(t)\cos\omega t\,\mathrm{d}t + i\int_{-\infty}^{\infty} y(t)\sin\omega t\,\mathrm{d}t$$

or

$$Y(\omega) = \operatorname{Re} Y(\omega) + i\operatorname{Im} Y(\omega), \tag{4.3.22}$$

where

$$\operatorname{Re} Y(\omega) = \int_{-\infty}^{\infty} y(t)\cos\omega t\,\mathrm{d}t, \tag{4.3.23}$$

$$\operatorname{Im} Y(\omega) = \int_{-\infty}^{\infty} y(t)\sin\omega t\,\mathrm{d}t. \tag{4.3.24}$$

Relation (4.3.23) is called the *Fourier cosine transformation,* and relation (4.3.24), the *Fourier sine transformation.* They also often use the modulus-squared Fourier transform (FT)

$$|Y(\omega)|^2 = \operatorname{Re}^2 Y(\omega) + \operatorname{Im}^2 Y(\omega),$$

called the *power spectrum,* and the FT modulus

$$|Y(\omega)| = \sqrt{\operatorname{Re}^2 Y(\omega) + \operatorname{Im}^2 Y(\omega)}\,,$$

called the *intensity spectrum.*

From (4.3.19), the *inverse Fourier transformation* can be obtained:

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega)\mathrm{e}^{-i\omega t}\,\mathrm{d}\omega, \qquad -\infty < t < \infty. \tag{4.3.25}$$

Given a two-dimensional function (*initial process*) $y(t_1, t_2)$, one can apply the *two-dimensional direct continuous Fourier transformation* to the function:

$$\begin{gathered} Y(\omega_1,\omega_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y(t_1,t_2)\mathrm{e}^{i(\omega_1 t_1 + \omega_2 t_2)}\,\mathrm{d}t_1\,\mathrm{d}t_2, \\ -\infty < \omega_1,\omega_2 < \infty; \end{gathered} \tag{4.3.26}$$

then, the inversion formula analogous to (4.3.25) yields

$$\begin{gathered} y(t_1,t_2) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y(\omega_1,\omega_2)\mathrm{e}^{-i(\omega_1 t_1 + \omega_2 t_2)}\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2, \\ -\infty < t_1, t_2 < \infty, \end{gathered} \tag{4.3.27}$$

the two-dimensional inverse continuous Fourier transformation.

Continuous Hartley transformation. Consider also the *one-dimensional direct continuous Hartley transformation*

$$Y_H(\omega) = \int_{-\infty}^{\infty} y(t)\,\mathrm{cas}\,(\omega t)\,\mathrm{d}t, \qquad -\infty < \omega < \infty,$$

where the (real) function is defined as

$$\mathrm{cas}\,x = \cos x + \sin x$$

or

$$\mathrm{cas}\,\omega t = \cos \omega t + \sin \omega t.$$

The *inverse continuous Hartley transformation* is

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Y_H(\omega)\,\mathrm{cas}\,\omega t\,\mathrm{d}\omega, \qquad -\infty < t < \infty.$$

A distinctive feature of the Hartley transformation (HT) is the fact that both the direct Hartley transform $Y_H(\omega)$ and the inverse Hartley transform $y(t)$ are real functions.

Continuous Laplace transformation. The *one-sided direct Laplace transformation* is the integral transformation

$$\Psi(p) = \int_0^{\infty} \varphi(x)\mathrm{e}^{-px}\,\mathrm{d}x, \tag{4.3.28}$$

where $p = \lambda + i\sigma$ is a complex variable; $\varphi(x)$ is some function of a real variable x (normally time), called the *original*; and $\Psi(p)$ is the *image* of $\varphi(x)$, which is often expressed as $\varphi(x) \to \Psi(p)$ or $\Psi(p) = L[\varphi(x)]$.

The *double-sided Laplace transformation* differs from (4.3.28) by that the lower integration limit in it is set equal to $-\infty$. Yet, it is one-sided transformation (4.3.28) that is normally meant under the term "*Laplace transformation*".

The original $\varphi(x)$ must obey the following conditions:

a) $\varphi(x)$ is a piecewise continuous function;

b) $\varphi(x) = 0$ for $x < 0$;

c) $|\varphi(x)| < M\mathrm{e}^{cx}$ for $x > 0$, where $M > 0$ and $c \geq 0$ are some constants, and, in addition, if $|\varphi(x)| \leq |\varphi(0)|$, then $c = 0$.

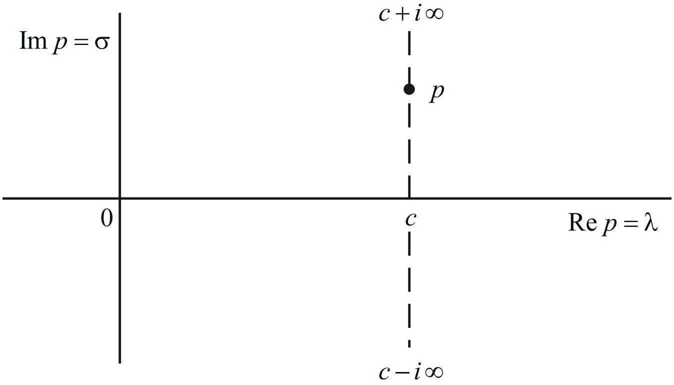
$c + i\infty$
$\operatorname{Im} p = \sigma$
p
0
c
$\operatorname{Re} p = \lambda$
$c - i\infty$

Figure 4.1.

Then, the *inverse Laplace transformation* takes place, which yields the following expression for the original:

$$\varphi(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Psi(p)\mathrm{e}^{px}\,\mathrm{d}p. \tag{4.3.29}$$

The dashed line in Figure 4.1 shows the straight line along which the integration in (4.3.29) is to be performed if the function $\Psi(p)$ has no singular points; otherwise, the path of integration in (4.3.29) lies on the right of all singular points of $\Psi(p)$.

Note that, strictly speaking, the integral Fourier, Hartley, Laplace (and also Mellin, Hilbert, Hankel, etc. (Verlan' and Sizikov, 1986; Korn and Korn, 1961)) transformations cannot be considered as integral equations with respect to their originals. For instance, expression (4.3.19) written as

$$\int_{-\infty}^{\infty} \mathrm{e}^{i\omega t} y(t)\,\mathrm{d}t = Y(\omega), \qquad -\infty < \omega < \infty, \tag{4.3.30}$$

cannot be considered as the first-kind Fredholm integral equation for the function $y(t)$ with the kernel $\mathrm{e}^{i\omega t}$ and the right-hand side $Y(\omega)$. This is related with the fact that the definition of the Fredholm integral equation implies fulfillment of the Hilbert–Schmidt condition (4.3.14) for the kernel, whereas for the kernel $\mathrm{e}^{i\omega t}$, which enters relation (4.3.30), this condition is violated because

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\mathrm{e}^{i\omega t}|^2\,\mathrm{d}\omega\,\mathrm{d}t = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{d}\omega\,\mathrm{d}t = \infty.$$

4.4. WELL- AND ILL-POSEDNESS ACCORDING TO HADAMARD

4.4.1. Definitions of well- and ill-posedness

In 1902, J. Hadamard introduced the notions of *well-* and *ill-posedness* (see also Verlan' and Sizikov (1986), Ivanov, Vasin, and Tanana (2002), Lavrent'ev, Romanov, and Shishatskii (1997), Sizikov (2001), Tikhonov and Arsenin (1977), Morozov (1984)). Consider these definitions of these notions with the example of the first-kind operator equation

$$Ay = f, \qquad y \in Y, \quad f \in F, \tag{4.4.1}$$

where y is the sought solution, f is a given right-hand side, Y and F are some Hilbert spaces (the spaces W_2^1 and L_2, for instance), and A is a given continuous operator (linear, nonlinear, algebraic, differential, integral, etc.).

Definition 4.4.1. The problem for solving equation (4.4.1) presents a *correct*, or *well-posed*, problem according to Hadamard if the following three conditions are fulfilled:

1) the solution exists;

2) the solution is unique;

3) the solution is stable.

If at least one of these conditions is violated, then the problem is an *incorrect*, or *ill-posed, problem.*

A more strict definition of well-posedness according to Hadamard is as follows (Ivanov, Vasin, and Tanana, 2002).

Definition 4.4.2. The problem for solving equation (4.4.1) presents a *well-posed problem according to Hadamard* if:

1) for any $f \in F$, there exists an element $y \in Y$ such that $Ay = f$, i. e., the operator domain is $R(A) = F$ (*solution existence*);

2) the solution y is uniquely determined by the element f, i. e., the inverse operator A^{-1} exists (*solution uniqueness*);

3) the function y depends on f continuously, i. e., the inverse operator A^{-1} is a continuous operator (*solution stability*).

The definition of well-posedness according to Hadamard is also called the *classical definition of well-posedness.*

In addition, there exists another definition of ill-posedness given by Fichera (Ivanov, Vasin, and Tanana, 2002); this definition further develops the definition of well-posedness according to Hadamard, placing emphasis on the properties of the operator A.

Definition 4.4.3. The problem for solving equation (4.4.1) presents a *well-posed problem according to Fichera* if the operator A is normally solvable (i. e., its operator domain is closed).

In Section 4.7, the definition of well-posedness according to Tikhonov will be given.

Hadamard advanced a statement that ill-posed problems are physically meaningless problems; in other words, if an equation that describes some applied problem is an ill-posed equation, then the problem presents an artificial (unreal) problem or this problem is described in an inadequate manner (for instance, some constraints on the solution making the problem correct are not taken into account). Hadamard gave several examples of ill-posed problems, for instance, the Cauchy problem for the Laplace equation (see Section 4.3). This problem, however, has many applications in astronomy, geophysics, cosmonautics, etc., i. e., presents a physically meaningful problem.

Moreover, many applied problems (in signal and image processing, tomography, spectroscopy, control theory, etc.) are ill-posed problems, which fact was widely recognized in several last decades. It should therefore be concluded that the statement advanced by Hadamard was invalid and has slowed down the development of many fields of research in pure and applied mathematics.

4.4.2. Examples of ill-posed problems

Below, several examples of equations or system of equations whose solution presents an ill-posed problem, are given.

Example 4.4.1. Consider the following overdetermined system of linear algebraic equations (SLAE) (Sizikov, 2001, p. 179):

$$\begin{aligned} 2y_1 - 3y_2 &= -4, \\ -y_1 + 2y_2 &= 3, \\ y_1 + 4y_2 &= 15. \end{aligned} \tag{4.4.2}$$

This SLAE is indeed an overdetermined system having no solution because the rank of the extended matrix here is $\rho = \text{rang}\,(A|f) = 3$, and the rank of A is $r = \text{rang}\,(A) = 2$, i. e., $\rho > r$ (see Section 4.2). In addition, the number of independent rows in (4.4.2) is 3, which is greater than n, that equals 2.

The fact that this SLAE has no solutions y_1, y_2 can be proved immediately. Indeed, with only the first two equations taken into consideration we obtain the solution $y_1 = 1$, $y_2 = 2$; with the second and the third equations, we obtain $y_1 = y_2 = 3$; if, alternatively, we take into account only the first and the third equation, then $y_1 = 2.635$, $y_2 = 3.09$, i. e., there is no unique solution and, here, the first condition for well-posedness according to Hadamard is violated.

Example 4.4.2. Consider the following underdetermined SLAE (Sizikov, 2001, p. 179):

$$2y_1 - 3y_2 = -4. \tag{4.4.3}$$

For this SLAE, $\rho = r = 1 < n = 2$; this SLAE is therefore an underdetermined system that has many solutions (see Section 4.2). Indeed, this SLAE has many solutions; for instance, 1) $y_1 = 1$, $y_2 = 2$; 2) $y_1 = 2$, $y_2 = 8/3$; 3) $y_1 = 0$, $y_2 = 4/3$, etc. are all solutions of the system. Thus, the solution of the SLAE is non-unique, and the second condition for well-posedness according to Hadamard is violated.

Example 4.4.3. Consider the determined SLAE (Sizikov, 2001, p. 179):

$$\begin{aligned} 2y_1 - 3y_2 &= 3, \\ -1.33y_1 + 2y_2 &= -1.99. \end{aligned} \tag{4.4.4}$$

For this SLAE, $\rho = r = n = 2$ and, in addition, $m = n = 2$ (and $r = m = 2$), this system is therefore indeed a determined system.

The solution of (4.4.4) exists and is unique: $y_1 = 3, \quad y_2 = 1$. Yet, with somewhat modified right-hand sides, which yields the SLAE

$$\begin{aligned} 2y_1 - 3y_2 &= 3.01, \\ -1.33y_1 + 2y_2 &= -2, \end{aligned} \tag{4.4.5}$$

i. e., with introduced relative errors $\|\delta f\|/\|f\| < 0.5\,\%$, we obtain a new, notably different solution: $y_1 = 2$ (relative error $|\delta y_1|/|y_1| \approx 33\,\%$), $y_2 = 0.33$

(relative error $\approx 67\,\%$), i. e., the relative solution error here is two orders of magnitude greater than the relative error in setting the right-hand side. These estimates can also be made considering the condition number cond (A) (see Section 4.2). To do this, we write the characteristic equation

$$\begin{vmatrix} 2-\lambda & -3 \\ -1.33 & 2-\lambda \end{vmatrix} = 0; \tag{4.4.6}$$

this equation yields $\lambda_1(A) = 3.9975$ and $\lambda_2(A) = 0.0025016$, i. e., the eigenvalues of A are real and positive numbers. Hence, the matrix A of (4.4.4) is a positively defined yet nonsymmetric matrix. That is why one has to calculate the condition number cond (A) with the help of the matrix A^*A, a positively defined symmetric matrix. We use (4.2.3) and obtain:

$$A^*A = \begin{pmatrix} 5.7689 & -8.66 \\ -8.66 & 13 \end{pmatrix}.$$

The eigenvalues of this matrix are

$$\lambda_1(A^*A) = 18.769, \qquad \lambda_2(A^*A) = 0.53996 \cdot 10^{-5}.$$

The singular numbers of A are

$$\mu_1(A) = \sqrt{\lambda_1(A^*A)} = 4.3323, \qquad \mu_2(A) = \sqrt{\lambda_2(A^*A)} = 0.0023237.$$

The condition number of A is

$$\operatorname{cond}(A) = \frac{\mu(A)_{\max}}{\mu(A)_{\min}} = 1.8644 \cdot 10^3, \tag{4.4.7}$$

and, according to (4.2.7) (at $\xi = 0$), we have:

$$\frac{\|\delta y\|}{\|y\|} \le \operatorname{cond}(A) \frac{\|\delta f\|}{\|f\|}. \tag{4.4.8}$$

We see that the solution of (4.4.4) is moderately unstable, that is, the relative solution error here is two orders of magnitude greater than the relative error of f.

In practice, many SLAEs have greater condition numbers, i. e., they are unstable in a greater extent. For such SLAEs, it makes sense to use the term “ill-conditioned SLAEs”. If, alternatively, cond $(A) = \infty$ (the determinant is zero), then such a SLAE is degenerate and the solution error may appear arbitrarily large.

Example 4.4.4. System of ordinary differential equations (ODE). In Section 4.3, the Cauchy problem for a system of ODEs was considered. We write this problem in the form (see (4.3.2)):

$$\frac{dy(t)}{dt} = f(t, y), \tag{4.4.9}$$

where $y(t) = \{y_1(t), \ldots, y_n(t)\}$ is the unknown vector function considered over the interval $t_0 \le t \le T$ and $f(t) = \{f_1(t), \ldots, f_n(t)\}$ is the right-hand side of the equation (some vector function). The initial conditions are

$$y(t_0) = y_0, \tag{4.4.10}$$

where y_0 is a given vector.

In Section 4.3, existence, uniqueness and stability conditions for system (4.4.9) were formulated. Namely, provided that the right-hand sides f_i of (4.4.9) are continuous and bounded functions that satisfy the Lipschitz condition in a certain closed region, then system (4.4.9) has a unique solution. Next, provided that the real parts of all roots of the characteristic equation are negative, then system (4.4.9) is stable, which means that small changes in initial conditions never result in large and arbitrarily large changes in the solution.

Consider now (as in Section 4.3) a system of ordinary differential equations with control function (see (4.3.7)):

$$\frac{dy(t)}{dt} = f(t, y, u), \tag{4.4.11}$$

where $u(t) = \{u_1(t), \ldots, u_m(t)\}$ is a control vector function. The initial conditions are set in the form (4.3.8).

We assume that the function $u(t)$, as well as the function $y(t)$, is the unknown function; in this case, we obtain an optimum-control problem.

Consider a particular example of optimum-control problem, the *vertical motion of a rocket* of variable mass launched vertically upward so that to reach a maximum height. This motion is governed by the following system of ODEs (Tikhonov and Arsenin, 1977):

$$\begin{gathered} \frac{dv(t)}{dt} = \frac{au(t) - cv^2(t)}{m(t)} - g, \\ \frac{dm(t)}{dt} = -u(t) \end{gathered} \tag{4.4.12}$$

with the initial conditions $m(0) = m_0$ and $v(0) = 0$. Here, $m(t)$ is the variable mass of the rocket; $v(t)$ is the velocity of the rocket; $u(t)$ is the control function, the consumed fuel mass as a function of time; and a, c, and g are some constants.

It is required to find an optimum control function $u_{\text{opt}}(t)$ such that the rocket could reach the maximum height ($H = \max$).

Yet, this problem is an ill-posed problem (Tikhonov and Arsenin, 1977): to small variations of H, arbitrarily large variations of $u(t)$ correspond, i. e., the third condition of well-posedness according to Hadamard is violated here. In Section 4.7 we will consider a regularization method for solution of ODE systems making the above problem stable.

Example 4.4.5. The *Cauchy problem for the Laplace equation.* In Section 4.3, we considered the *two-dimensional Laplace equation* (elliptic-type partial differential equation)

$$\Delta u(x,y) \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{4.4.13}$$

together with the boundary conditions (Cauchy problem)

$$u(x,0) = f(x), \qquad \left.\frac{\partial u}{\partial y}\right|_{y=0} = \varphi(x), \tag{4.4.14}$$

where $f(x)$ and $\varphi(x)$ are some given functions.

In a number of studies (Lavrent'ev, Romanov, Shishatskii (1997) and others) it was shown that, with some minimum constraints imposed on the class of functions among which the solution $u(x,y)$ is sought, this solution exists and is unique, i. e., the first and second conditions for well-posedness according to Hadamard are fulfilled here. Yet, the solution is unstable. Let us show that.

With $f(x) = f_1(x) = 0$ and $\varphi(x) = \varphi_1(x) = a \sin \omega x$, the solution of the Cauchy problem is the function

$$u(x,y) = u_1(x,y) = (a/\omega) \sin \omega x \operatorname{sh} \omega y,$$

where $\operatorname{sh} x = (\mathrm{e}^x - \mathrm{e}^{-x})/2$ is the hyperbolic sine.

Alternatively, if $f(x) = f_2(x) = \varphi(x) = \varphi_2(x) = 0$, then the solution of the Cauchy problem is the function $u(x,y) = u_2(x,y) = 0$. For any functional spaces and $\varepsilon > 0$, $c > 0$, and $y > 0$, one can choose a and ω such that the equality

$$\|f_1 - f_2\| = 0$$

and inequalities

$$\|\varphi_1 - \varphi_2\| = \|a \sin \omega x\| < \varepsilon, \qquad \|u_1 - u_2\| = \|(a/\omega) \sin \omega x \operatorname{sh} \omega y\| > c,$$

will be fulfilled, i. e., small variations of the boundary conditions $f(x)$ and $\varphi(x)$ will result in arbitrarily large variations of $u(x, y)$.

Thus, the solution of the Cauchy problem for the Laplace equation is unstable (the third condition of well-posedness according to Hadamard is violated) and, hence, the problem is incorrect.

In Section 4.7, it was shown how a regular (stable) solution of the Cauchy problem for the Laplace equation can be constructed.

Example 4.4.6. The Cauchy problem for heat conduction equation (parabolic-type partial differential equation). In Section 4.3, four variants of this problem were considered (Ivanov, Vasin, and Tanana, 2002).

In the first variant (direct problem, or forward-time problem), it is required to solve the *heat conduction equation*

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \qquad 0 \le x \le l, \quad t > 0, \tag{4.4.15}$$

with the initial condition $u(x, 0) = \varphi(x)$ and the boundary conditions $u(0, t) = \psi_1(t)$ and $u(l, t) = \psi_2(t)$.

The sought function $u(x, t)$ gives the temperature distribution along a rod at the time t provided that, initially (at $t = 0$), this distribution was $\varphi(x)$ and the ends of the rod were kept under the temperature regimes $\psi_1(t)$ and $\psi_2(t)$.

The equation is to be solved from the time $t = 0$ towards increasing t. This is the classical heat conduction problem. This problem is a well-posed problem: the solution continuously depends on the initial data $\varphi(x)$.

In the second variant of the problem (in the inverse problem, or in the reverse-time problem) it is required to reconstruct the initial temperature distribution $u(x, 0) = \varphi(x)$ along the rod from the temperature distribution known at some time $t_* > 0$, i. e., to reconstruct, by solving problem towards decreasing t, the initial temperature distribution from some given function $u(x, t_*)$. In this case, the boundary conditions $\psi_1(t)$ and $\psi_2(t)$ are not set. This is an unstable problem.

In the third variant of the problem, it is required to solve the heat conduction equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \qquad -\infty < x < \infty, \quad 0 \le t \le T, \tag{4.4.16}$$

with the initial conditions $u(x, 0) = \varphi(x)$ and $u(x, T) = \chi(x)$, where $\varphi(x)$ and $\chi(x)$ are some given functions. Generally speaking, this problem has no solution (i. e., the first condition for well-posedness according to Hadamard is violated here). Yet, it is possible to find, using just the initial condition $u(x, T) = \chi(x)$ and proceeding from the time T to the past $(t < T)$, the function $\varphi(x) = u(x, 0)$, instead of setting this function. Yet, this problem (analogous to the second variant) is unstable (i. e., the third condition of well-posedness according to Hadamard is violated here).

In the fourth variant of the problem, it is required to solve equation (4.4.15) or (4.4.16) with just one boundary condition $u(x, 0) = \varphi(x)$. This problem is an ill-posed problem; the incorrectness here results from data insufficiency.

In Section 4.7, some stable methods for solving the heat conduction equation will be discussed.

4.4.3. Correctness or incorrectness as dependent on the particular type of space

The problem for solving an equation can be a well- or ill-posed problem depending on the particular types of the functional spaces in which the solution is sought and the initial conditions are set. Moreover, different types of equations display different "degree of ill-posedness". Let us illustrate this with several examples.

Example 4.4.7. *Differentiation of an approximate function* $f(x)$ (Tikhonov and Arsenin, 1977). Let $y(x)$ be the derivative of a function $f(x)$, i. e. $y(x) = f'(x)$.

A distorted function, for instance, the function $\tilde{f}(x) = f(x) + A \sin \omega x$ differs from the function $f(x)$ in the metric C by a finite value

$$\|\tilde{f} - f\|_C = \max_{a \le x \le b} |\tilde{f}(x) - f(x)| = \max_{a \le x \le b} |A \sin \omega x| = |A|$$

for any ω.

Yet, in the metric C the derivative $\tilde{y}(x) = \tilde{f}'(x) = f'(x) + A\omega\cos\omega x$ differs from $y(x) = f'(x)$ by

$$\|\tilde{y} - y\|_C = \max_{a\le x\le b} |A\omega\cos\omega x| = |A\omega|;$$

the latter quantity may appear arbitrarily large at sufficiently large values of $|\omega|$. Thus, the differentiation problem generally does not possess the property of stability in the space C, i. e., it presents an ill-posed problem (the third condition of well-posedness according to Hadamard is violated here).

Consider the same problem on the other pair of spaces, namely, as previously, we consider the difference between $\tilde{y}(x)$ and $y(x)$ in the space C, i. e.,

$$\|\tilde{y} - y\|_C = \max_{a\le x\le b} |\tilde{y}(x) - y(x)|,$$

whereas the difference between $\tilde{f}(x)$ and $f(x)$ in the space $C^{(1)}$ is

$$\|\tilde{f} - f\|_{C^{(1)}} = \max_{a\le x\le b}[|\tilde{f}(x) - f(x)| + |\tilde{f}'(x) - f'(x)|].$$

In the latter case, $\|\tilde{y} - y\|_C = |A\omega|$ and

$$\|\tilde{f} - f\|_{C^{(1)}} = \max_{a\le x\le b}[|A\sin\omega x| + |A\omega\cos\omega x|] = |A| + |A\omega|.$$

Now, we cannot say that at the finite norm $||\tilde{f}-f||_{C^{(1)}}$ the norm $\|\tilde{y}-y\|_C$ can be arbitrarily large; we can just say that, with finite $|A|$ and $|\omega|$, the difference $\|\tilde{y} - y\|_C$ will also be finite. In other words, the differentiation problem in the pair of spaces $(C, C^{(1)})$ (with $y(x) \in C$ and $f(x) \in C^{(1)}$) is a well-posed problem. The latter results from the fact that, having used the space $C^{(1)}$, we have imposed harder constraints on the function $f(x)$ compared to the case of the space C.

Remark. In the latter example, to imitate the error in the function $f(x)$, we used the deterministic function $A\sin\omega x$. In practical problems, the error is normally a random process, and a deeper analysis is necessary. Yet, the qualitative conclusion that the differentiation problem can be correct or incorrect depending on the particular pair of spaces on which this problem is considered, remains valid.

Example 4.4.8. *Solution of the first-kind Fredholm integral equation*

$$\int_a^b K(x,s)y(s)\,\mathrm{d}s = f(x), \qquad c \le x \le d, \tag{4.4.17}$$

and the first-kind Volterra integral equation

$$\int_a^x K(x,s)y(s)\,\mathrm{d}s = f(x), \qquad a \le x \le b. \tag{4.4.18}$$

Similarly to Example 4.4.7, it can be shown (Verlan' and Sizikov, 1986; Apartsyn, 2003) that the solution of the first-kind Fredholm equation is unstable in any "rational" pair of functional spaces: (C, L_2), (L_2, L_2), (C, C), $(C, C^{(1)})$, (W_2^1, L_2), etc., i. e., the third condition (and also the first and second conditions) of well-posedness according to Hadamard is violated here.

As for the first-kind Volterra equation, this equation is correct on the pair of spaces $(C, C^{(1)})$ (where $y(s) \in C$ and $f(x) \in C^{(1)}$) and incorrect on the pair (C, C) (where $y(s) \in C$, $f(x) \in C$) (Apartsyn, 2003). In this sense, the first-kind Volterra equation occupies an intermediate position, for instance, between the first-kind Fredholm equation and the second-kind Volterra equation

$$y(x) + \int_a^x K(x,s)y(s)\,ds = f(x), \qquad a \le x \le b,$$

whose solution presents a well-posed problem under rather weak constraints imposed on to the kernel $K(x,s)$ and on the function $f(x)$ (Verlan' and Sizikov, 1986, p. 20).

Remark. Notice that the conditions for well- or ill-posedness of the differentiation problem and for the solution of the first-kind Volterra equation are identical conditions. The latter is related to the fact that the problem on determination of the n-th order derivative $y(x)$ of a function $f(x)$ can be reduced to solving the following first kind Volterra integral equation (Tikhonov and Arsenin, 1977)

$$\int_a^x \frac{(x-s)^{n-1}}{(n-1)!}\,y(s)\,\mathrm{d}s = f(x), \qquad x \ge a,$$

with the additional condition $f(a) = f'(a) = \cdots = f^{(n-1)}(a) = 0$.

Example 4.4.9. This example is an example taken from *the problem on reconstruction of blurred images* (see Section 5.1). The starting equation of the problem is the *first-kind Volterra integral equation*

$$\int_{x}^{x+\Delta} \frac{1}{\Delta} w(\xi)\,\mathrm{d}\xi = g(x), \qquad -\infty < x < \infty, \tag{4.4.19}$$

where $g(x) \equiv g_y(x)$ is the intensity distribution on an image blurred by a uniform rectilinear shift, $w(\xi) \equiv w_y(\xi)$ is the sought intensity distribution on the true (non-blurred) image, Δ is the shift magnitude, and the axes x and y are directed along and normal to the blurring shift. In (4.4.19), the quantity y is a parameter (at each y, one has to solve one-dimensional equation (4.4.19)).

In terms of Apartsyn (2003), this equation is the *non-classical Volterra equation of the first kind* (because both limits here are variable). Yet, as it was shown by Apartsyn (2003) (see also Verlan' and Sizikov, 1986), the problem for solving non-classical equation, as well as the problem for solving equation (4.4.18), is a well-posed problem on the pair of spaces $(C, C^{(1)})$, i. e., when $w(\xi) \in C$ and $g(x) \in C^{(1)}$, and an ill-posed problem on the pair (C, C), when $w(\xi) \in C$ and $g(x) \in C$.

Next, equation (4.4.19) can be transformed in an integral equation of another type, namely, in the first-kind (convolution-type) Fredholm integral equation:

$$\int_{-\infty}^{\infty} k(x-\xi) w(\xi)\,\mathrm{d}\xi = g(x), \qquad -\infty < x < \infty, \tag{4.4.20}$$

where

$$k(x) = \begin{cases} 1/\Delta, & -\Delta \le x \le 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4.21}$$

The first-kind Fredholm equation (4.4.20) is incorrect in any pair of spaces, in the pair $(C, C^{(1)})$ for instance. It should be noted here that Fredholm equation (4.4.20) was obtained from Volterra equation (4.4.19) by applying some equivalent transformations (see Section 2.2). Thus, this example confirms the conclusion made in Chapters 1 and 2 that an equivalent transformations can transform a correct equation in an incorrect equation, and vice versa. Yet, in solving either of the equations, (4.4.19) or (4.4.20), by the regularization method (see Section 4.7) both equations yield a stable solution, i. e., regularization makes it possible to avoid detrimental effects that may be induced by equivalent transformations (Danilevich and Petrov, 2000; Petrov, 1998).

Example 4.4.10. *Determination of the unit impulse response* $g(t,\tau)$ of a transducer by given input $x(\tau)$ and output $f(t)$ signals bound up in Volterra integral equation of the first kind (Boikov and Krivulin, 2000)

$$\int_0^t g(t,\tau)\,x(\tau)\,\mathrm{d}\tau = f(t), \qquad 0 \le t \le T. \tag{4.4.22}$$

This problem is the problem for determining the kernel $g(t,\tau)$ of the integral equation using the given functions $x(\tau)$ and $f(t)$. For solvability of the problem, we consider the functions x and f as the functions of two variables, i. e., $x = x(p,\tau)$ and $f = f(p,t)$, where p is an additional parameter (for instance, the effective width of the function $x(\tau)$), and rewrite (4.4.22) in the form

$$\int_0^t g(t,\tau)\,x(p,\tau)\,\mathrm{d}\tau = f(p,t), \qquad 0 \le t \le T, \quad a \le p \le b.$$

For every fixed t, the latter equation can be reduced to one-dimensional first-kind Fredholm equation on the function $g_t(\tau)$:

$$\int_0^t x(p,\tau)\,g_t(\tau)\,\mathrm{d}\tau = f_t(p), \qquad a \le p \le b, \tag{4.4.23}$$

where $x(p,\tau)$ is a new kernel.

The main feature of the example is reduction of first-kind Volterra equation (4.4.22) to first-kind Fredholm equation (4.4.23), the “degrees of instability” of which, as it was noted above, are various.

Remark. An opinion can be expressed that it is possible to choose an appropriate pair of functional spaces for solving a particular equation so that to make the problem for solving this equation a well-posed problem (Lavrent’ev, Romanov, and Shishatskii, 1997, p. 8). For instance, one can solve the first-kind Volterra integral equation not on the pair of spaces (C, C), where the problem is incorrect, but on the pair $(C, C^{(1)})$, where the problem is correct. As it was shown above, the same is valid for the differentiation problem. In a similar manner, one can choose an appropriate pair of spaces for solving partial differential equations to make their solution a well-posed problem. Nonetheless, first, there exist equations whose solution presents an ill-posed problem in any pair of spaces (an example is the first-kind Fredholm integral equations, for instance) and, second, the choice of the pair of spaces should not be made arbitrarily but, instead, it should be matched with specific features of the measurement procedure for the function $f(x)$ with due allowance for the experimental error $\delta f(x)$. If one supposes that $f(x) \in C$, then this means that we are going to use the estimated maximum measurement error

$$\delta \equiv \|\delta f(x)\|_C = \max_{c \le x \le d} |\delta f(x)|,$$

which can be readily obtained from the experiment. In a similar manner, the error δ can be estimated in the case of $f(x) \in L_2$: in this case, the error δ is the mean-root square error

$$\delta \equiv \|\delta f(x)\|_{L_2} = \Big(\int_c^d |f(x)|^2 \, \mathrm{d}x \Big)^{1/2}.$$

Yet, it is practically impossible to obtain an estimate of δ in the case of $f(x) \in C^{(1)}$, where

$$\delta \equiv \|\delta f(x)\|_{C^{(1)}} = \max_{c \le x \le d} \{|\delta f(x)| + |\delta f'(x)|\}.$$

That is why we will be unable to use the pair of spaces $(C, C^{(1)})$, and will be forced to use instead the less desirable pair (C, C) or (C, L_2). Note that the value of δ is used in regularization methods intended for solving ill-posed problems (see Section 4.7).

Below, we will show that: 1) if no solution exists (the first condition for well-posedness according to Hadamard is violated), then some modification of the Gauss least-mean-square method should be used (in this case, a pseudo-solution will be obtained (see Section 4.6)); 2) if the solution is not unique (the second condition for well-posedness according to Hadamard is violated), then some modification of the Moore–Penrose pseudo-matrix method should be used (here, the normal solution will be obtained (see Section 4.6)); 3) if the solution is unstable (the third condition for well-posedness according to Hadamard is violated), then it will be necessary to use a stable (regular) method such as the Tikhonov regularization method (in this case, the regular solution will be obtained (see Section 4.7)).

4.5. CLASSICAL METHODS FOR SOLVING FREDHOLM INTEGRAL EQUATIONS OF THE FIRST KIND

Prior to discussing the stable (regular) methods for solving ill-posed problems, let us dwell on some classical methods; to make the narration more convincing, we will treat the most unstable equation, namely, the first-kind Fredholm integral equation.

4.5.1. Quadrature method (Sizikov 2001, pp. 180–182)

Consider the *Fredholm integral equation of the first kind*

$$\int_a^b K(x,s)y(s)\,\mathrm{d}s = f(x), \qquad c \le x \le d, \tag{4.5.1}$$

where $K(x,s)$ is the kernel, $y(s)$ is the unknown function, $f(x)$ is the right-hand side, $[a,b]$ is the domain of s, and $[c,d]$ is the domain of x.

Suppose that, instead of $f(x)$, a right-hand side $\tilde{f}(x)$ distorted by measurement errors is available, and, instead of $K(x,s)$ the inaccurate $\tilde{K}(x,s)$ is known; then, under $y(s)$, a function $\tilde{y}(s)$ should be meant, representing the solution distorted by errors that result from the right-hand side and kernel inacuracies (and also from the errors brought about by the solution algorithm). Hence, instead of (4.5.1), we must write:

$$\int_a^b \tilde{K}(x,s)\tilde{y}(s)\,\mathrm{d}s = \tilde{f}(x), \qquad c \le x \le d; \tag{4.5.2}$$

yet, to simplify notation, instead of (4.5.2) we will further use equation (4.5.1).

The essence of the quadrature method is as follows (Sizikov, 2001, pp. 180–182).

1) We introduce a uniform node grid with steps $\Delta s = h_1 = \mathrm{const}$ and $\Delta x = h_2 = \mathrm{const}$ along s and x, respectively. Then, the number of nodes along s will be $n = (b-a)/h_1 + 1$ and the number of nodes along x, $m = (d-c)/h_2 + 1$.

2) We replace the integral in (4.5.1) with a finite sum constructed by some quadrature formula, for instance, by the trapezoid formula:

$$\int_a^b K(x,s)y(s)\,\mathrm{d}s \approx \sum_{j=1}^{n} p_j K(x,s_j)y(s_j),$$

where

$$p_j = \begin{cases} 0.5h_1, & j = 1 \quad \text{or} \quad j = n, \\ h_1, & \text{otherwise}, \end{cases}$$

$$s_j = a + (j-1)h_1.$$

3) We take into account the discretization along x:

$$x_i = c + (i-1)h_2,$$

and finally obtain a discrete analog of (4.5.1):

$$\sum_{j=1}^{n} A_{ij} y_j = f_i, \qquad i = 1, \ldots, m, \tag{4.5.3}$$

where $A_{ij} = p_j K(x_i, s_j)$ are elements of an $m \times n$ matrix A, $y_j = y(s_j)$, and $f_i = f(x_i)$.

As a result, we have a system of m linear algebraic equations (4.5.3) for n unknowns y_j. By solving this system, we can find the solution of (4.5.1) in discrete numerical form.

Generally, the matrix A of system (4.5.3) is a rectangular matrix. If $m = n$, then the matrix A is a square matrix and SLAE (4.5.3) can be solved by the Cramer rule, Gauss methods, etc. If $m > n$, then SLAE (4.5.3) should be solved by the Gauss least-squares method (see Section 4.6; in this case we will obtain a pseudo-solution), and if $m < n$, then the Moore – Penrose pseudo-inverse method is to be used (see Section 4.6; we will obtain the normal solution). Thus, the first two conditions for well-posedness according to Hadamard will be formally fulfilled.

Yet, all these solutions are very unstable, i. e., here the third condition for well-posedness according to Hadamard will be violated. This instability results from the fact that the inverse operator of equation (4.5.1) is infinite. The latter is manifested in that the minimum singular number $\mu_{\min}$ of the integral operator of (4.5.1) is zero, and the condition number is cond $= \infty$. If, alternatively, the integral operator is approximated with an algebraic operator with finite m and n, then $\mu_{\min}$ may happen to somewhat differ from zero, but, nonetheless, the solution of (4.5.3) will again be very unstable.

Figure 4.2 show the solution of an example (given below in Section 4.7) obtained by the quadrature method according to (4.5.3) with $m = n = 137$. The exact solution $y(s)$ is the sum of five Gaussians. The almost vertical dashed curves show the approximate (discrete) solution y_j, $j = 1, \ldots, n$, of (4.5.3).

We see that the solution y_j emerges as the so-called modulated "saw" of large amplitude (Tikhonov and Arsenin, 1977; Glasko, Mudretsova, and Strakhov, 1987, p. 91) that has nothing in common with the exact solution. Meanwhile, if, to check the solution, to insert the "saw" into (4.5.3), then we obtain the right- and left-hand sides of (4.5.3) to be identical within three-five digits provided that the calculations were performed on a computer with ordinary accuracy (to seven accurate digits), or within 6–10 digits if the calculations are performed with double accuracy (to 14 accurate digits).

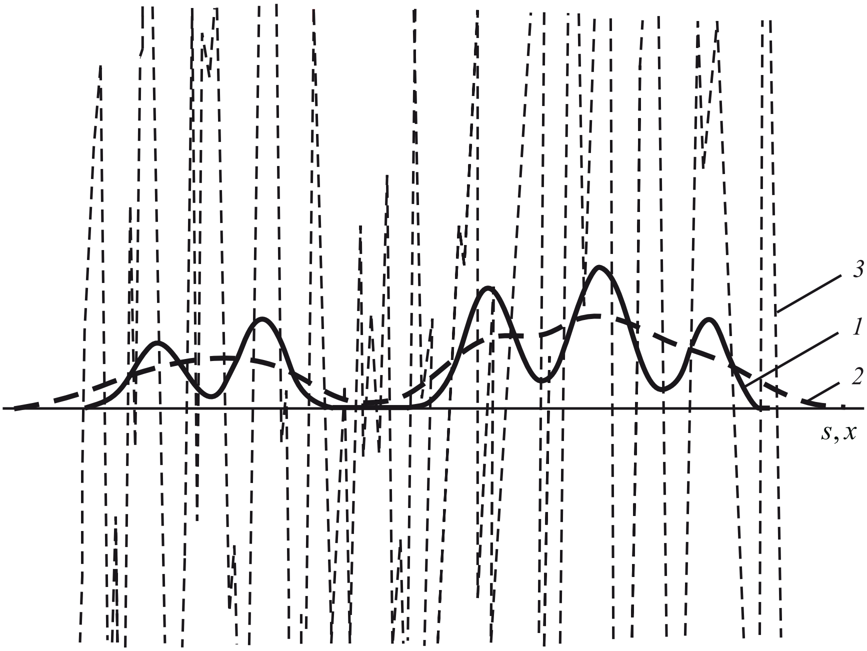

Figure 4.2. Instable solution of the Fredholm integral equation of the first kind obtained by the quadrature method:
1 — exact solution $y(s)$, *2* — right-hand side $f(x)$, *3* — approximate solution y_j

Note that the shape of the "saw" depends on the particular method used to solve the SLAE, on the computer program, etc.

From the aforesaid, the following *conclusions* can be drawn:

1) The quadrature method for solving the first-kind Fredholm integral equation is rather an unstable method (i. e., the third condition for well-posedness according to Hadamard is violated here). As applied to the Volterra integral equation of the first kind, the quadrature method turns out to be more stable (Verlan' and Sizikov, 1986, pp. 111–114, 120–123).

2) The classical definition of the exact solution $\bar{y}$ as a solution for which the discrepancy between the two sides of the equation is zero,

$$\|A\bar{y} - f\| = 0, \tag{4.5.4}$$

is generally inappropriate for ill-posed problems. This is related with the fact that: 1) in the case where no solution exists (see Example 4.4.1), there is no such $\bar{y}$ that make equality (4.5.4) an identity; 2) in the case where the solution is non-unique (see Example 4.4.2), there are many $\bar{y}$ satisfying equality (4.5.4); 3) in the case of instability (see Example 4.4.3 and Figure 4.2), criterion (4.5.4) yields an unstable solution.

4.5.2. Fourier transformation method for convolution-type equations

(Verlan' and Sizikov, 1986, pp. 256–259; Sizikov, 2001, pp. 182–184)

Consider the *Fredholm integral equation of the first kind of convolution type*

$$\int_{-\infty}^{\infty} K(x-s)y(s)\,\mathrm{d}s = f(x), \qquad -\infty < x < \infty. \tag{4.5.5}$$

This equation has a solution that can be represented analytically as the *inverse Fourier transform* (IFT):

$$y(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega)\mathrm{e}^{-i\omega s}\,\mathrm{d}\omega. \tag{4.5.6}$$

Here, the spectrum, or FT, of the solution is

$$Y(\omega) = F(\omega)/\lambda(\omega), \tag{4.5.7}$$

where

$$F(\omega) = \int_{-\infty}^{\infty} f(x)\mathrm{e}^{i\omega x}\,\mathrm{d}x, \tag{4.5.8}$$

$$\lambda(\omega) = \int_{-\infty}^{\infty} K(x)\mathrm{e}^{i\omega x}\,\mathrm{d}x \tag{4.5.9}$$

are the spectra, or FTs, of $f(x)$ and $K(x)$.

Yet, like the solution obtained by the quadrature method, solution (4.5.6) is also very unstable; this fact can be explained as follows. Normally, the kernel $K(x)$ is set with the help of a smooth analytical formula whose spectrum $\lambda(\omega)$ rapidly decays with increasing $|\omega|$ so that $\lim_{\omega\to\infty}\lambda(\omega) = 0$. The function $f(x)$ is normally set as a tabulated sequence of experimental noise-distorted data; i.e., instead of the exact function $f(x)$ we have $\tilde{f}(x) = f(x) + \delta f(x)$, where $\delta f(x)$ is some error whose spectrum normally tends, as $|\omega| \to \infty$, to some yet small constant, a "white-noise" level. For this reason, $\lim_{\omega\to\infty} F(\omega)/\lambda(\omega) = \lim_{\omega\to\infty} Y(\omega) = \infty$, and integral (4.5.6) diverges. In other words, instability of the FT method emerges owing to the very strong response of high Fourier harmonics to arbitrarily small measurement-induced errors of $f(x)$.

Alternatively, if the calculations are performed by finite quadrature formulas, then, instead of the continuous Fourier transforms (CFT) $F(\omega)$, $\lambda(\omega)$ and $Y(\omega)$, their discrete Fourier transforms (DFT) are used (Sizikov, 2001; Bracewell, 1986; Vasil'ev and Gurov, 1998; Rabiner and Gold, 1975). In the latter case, the maximum Fourier frequency is finite: $\omega_{\max} = 2\pi/h$, where $h = \Delta x = \Delta s$ is the discretization step. As a result, the solution instability decreases, but still remains large.

Solution of several examples shows that the FT method yields a more stable solution compared to the quadrature method; the latter can be attributed to the fact that, first, the FT method yields analytical solution (4.5.6) and, second, in the numerical realization of the FT method, high Fourier frequencies turn out to be automatically cut out from the spectrum because of the discretization step h is finite.

4.6. GAUSS LEAST-SQUARES METHOD AND MOORE–PENROSE INVERSE-MATRIX METHOD

Consider, with the examples of a system of linear algebraic equations (SLAE) and an integral equation, the Gauss least-squares method (LSM) (Bronshtein and Semendyaev, 1986, p. 521; Sizikov, 2001, pp. 186–188), and also the Moore–Penrose pseudo-inverse matrix method (Verlan' and Sizikov, 1986, p. 508; Sizikov, 2001, p. 189; Voevodin, 1980; Gantmacher, 1959).

4.6.1. Overdetermined SLAE

Consider a system of m linear algebraic equations (SLAE) for n unknowns such that $m > n$ and $\operatorname{rang}(A|f) > \operatorname{rang}(A)$ or, which is equivalent, such a system in which the number of independent rows is greater than n, i. e., consider an overdetermined SLAE (system (4.4.2), for instance):

$$Ay = f, \tag{4.6.1}$$

where A is an $m \times n$-matrix; y is the unknown $n \times 1$ column vector; and f is the set right-hand side, an $m \times 1$ column vector. Such an SLAE has no solution; in other words, there is no $\bar{y}$ such for which the discrepancy turns into zero:

$$\|A\bar{y} - f\| = 0. \tag{4.6.2}$$

4.6.2. Pseudo-solution and normal SLAE

In the Gauss least-squares method (LSM), instead of (4.6.2), the following condition is introduced:

$$\|Ay - f\| = \min_{y}. \tag{4.6.3}$$

Definition. *Pseudo-solution* of (4.6.1) is a solution y that satisfies condition (4.6.3), i. e., the solution that minimizes the discrepancy $\|Ay - f\|$.

Thus, in the LSM the condition of zero discrepancy is replaced with the condition of minimum discrepancy, and, instead of the exact solution $\bar{y}$, a pseudo-solution y is considered. Normally, the norm $\|Ay - f\|$ is considered in the space $\mathbb{R}^n$ (see (4.2.2)), and the method is therefore called the *least-squares method.*

Remark. If $\|Ay - f\| = 0$, then the pseudo-solution y coincides with the exact solution $\bar{y}$, i. e., the pseudo-solution presents a generalization of the notion of exact solution.

We write condition (4.6.3) as

$$\|Ay - f\|^2 = \min_{y}. \tag{4.6.4}$$

Let us derive a new SLAE from condition (4.6.4). Minimization of (4.6.4) implies that the variation, or derivative, of the discrepancy with respect to y (also called the Frechet derivative (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) is zero: $2A^*(Ay - f) = 0$ or $A^*Ay - A^*f = 0$. As a result, we obtain:

$$A^*Ay = A^*f. \tag{4.6.5}$$

Thus, instead of the overdetermined SLAE (4.6.1) we have a new SLAE (4.6.5), called the *normal SLAE.*

We write (4.6.5) as

$$By = u, \tag{4.6.6}$$

where

$$B = A^*A, \tag{4.6.7}$$

$$u = A^*f \tag{4.6.8}$$

or, provided that the matrix A is a real matrix,

$$B = A^\top A, \tag{4.6.9}$$

$$u = A^\top f. \tag{4.6.10}$$

We use rule (4.2.3) of multiplication of matrices and rule (4.2.4) of multiplication of a matrix by a vector, and write detailed expressions for the elements of the new matrix B and for the elements of the new right-hand side u of normal SLAE (4.6.6) in the case of real A:

$$B_{ij} = \sum_{k=1}^{m} A_{ik}^\top A_{kj} = \sum_{k=1}^{m} A_{ki} A_{kj}, \tag{4.6.11}$$

$$u_i = \sum_{k=1}^{m} A_{ik}^\top f_k = \sum_{k=1}^{m} A_{ki} f_k. \tag{4.6.12}$$

Formally, the solution of (4.6.5) or (4.6.6) is

$$y = (A^* A)^{-1} A^* f \tag{4.6.13}$$

or

$$y = B^{-1} u. \tag{4.6.14}$$

In practice, this solution can be found by the Cramer rule, Gauss method, Cholesky method, or other algorithms.

4.6.3. LSM as applied to integral equations

If we apply the Gauss LSM to the first-kind Fredholm equation (4.5.1), we obtain the following new integral equation (cp. (4.6.6)):

$$\int_a^b B(t,s) y(s)\,\mathrm{d}s = u(t), \qquad a \le t \le b, \tag{4.6.15}$$

where (cp. (4.6.11) and (4.6.12))

$$B(t,s) = B(s,t) = \int_c^d K(x,t) K(x,s)\,\mathrm{d}x, \tag{4.6.16}$$

$$u(t) = \int_c^d K(x,t) f(x)\,\mathrm{d}x. \tag{4.6.17}$$

The *main specific features of the Gauss LSM* are as follows:

1) The matrix B of the new (normal) SLAE (4.6.6) is an $n \times n$ square matrix, i. e., to be solved is a new system of n equations for n unknowns; if $\det(B) \neq 0$, then SLAE (4.6.6) has a unique solution.

2) The matrix B of the new SLAE (4.6.6), and also the kernel $B(t, s)$ of the new integral equation (4.6.15) both are symmetrical and positively defined elements. That is why the most appropriate strategy here is to solve SLAE (4.6.6) by special methods (Cholesky method, Kraut method, etc.), and not by the general methods (Cramer method, Gauss method, etc.). The same applies to the solution of the SLAE obtained in solving the new integral equation (4.6.15) by the quadrature method.

3) Provided that the solutions of the starting SLAE (4.6.1) and integral equation (4.5.1) are both unstable, then the solution of the new SLAE (4.6.6) and the solution of the integral equation (4.6.15) are also unstable, i. e., the LSM does not solve the problem of instability.

Example. To illustrate the LSM, consider Example 4.4.1 (cp. (4.4.2)). The matrix of (4.4.2) is

$$A = \begin{pmatrix} 2 & -3 \\ -1 & 2 \\ 1 & 4 \end{pmatrix}. \tag{4.6.18}$$

Using formulas (4.6.11) and (4.6.12) we obtain

$$B = \begin{pmatrix} 6 & -4 \\ -4 & 29 \end{pmatrix}, \tag{4.6.19}$$

$$u = \begin{pmatrix} 4 \\ 78 \end{pmatrix}, \tag{4.6.20}$$

i. e., the new (normal) SLAE is

$$\left.\begin{aligned} 6y_1 - 4y_2 &= 4, \\ -4y_1 + 29y_2 &= 78. \end{aligned}\right\} \tag{4.6.21}$$

The solution of this SLAE is $y_1 = 2.71$ and $y_2 = 3.06$. To this solution, the (minimum possible) discrepancy $\|Ay - f\| = 0.3993 \approx 0.4$ corresponds (according to (4.2.2)). By solving the characteristic equation for the matrix $B = A^*A$

$$\begin{vmatrix} 6-\lambda & -4 \\ -4 & 29-\lambda \end{vmatrix} = 0,$$

we find the eigenvalues of B: $\lambda_1(B) = 29.675$ and $\lambda_2(B) = 5.325$. We see that λ_1 and λ_2 are real and nonnegative numbers, as it should be for the symmetric positively defined matrix $B = A^*A$.

Consider now the Moore–Penrose pseudo-inverse matrix method (PIMM). As with the Gauss LSM, consider first the PIMM as applied to solving SLAEs (Verlan' and Sizikov, 1986, p. 508; Sizikov, 2001, pp. 189–191; Voevodin, 1980; Gantmacher, 1959).

4.6.4. Underdetermined SLAE

Consider an SLAE

$$Ay = f, \tag{4.6.22}$$

where A is an $m \times n$-matrix; y is the unknown n-vector; and f is the right-hand side of the equation, an m-vector such that $m < n$, or $\rho < n$, where $\rho = \operatorname{rang}(A|f)$ is the rank of the extended matrix. Such an SLAE is called an *underdetermined SLAE.*

An underdetermined SLAE has many solutions $\bar{y}$; hence, the second condition for well-posedness according to Hadamard is violated here. For instance, SLAE (4.4.3) has the following solutions: 1) $\bar{y}_1 = \{1, 2\}^\top$, 2) $\bar{y}_2 = \{2, 8/3\}^\top$, 3) $\bar{y}_3 = \{0, 4/3\}^\top$, etc. Each of these solutions makes equality (4.6.2) an identity.

4.6.5. Normal solution and pseudo-inverse matrix

The choice of a single solution from the whole set of solutions of the underdetermined SLAE can be made by means of the Moore–Penrose *pseudo-inverse matrix method* (PIMM) (1930).

Definition. The *normal solution* is the solution with the minimum norm among the whole set of solutions; in other words, this solution satisfies the condition

$$\|y\| = \min_y \tag{4.6.23}$$

or

$$\|y\|^2 = \min_y. \tag{4.6.24}$$

The normal solution is the smoothest solution among all solutions.

According to PIMM, the normal solution is to be chosen from the set of all solutions of the underdetermined SLAE. It is proved by Gantmacher (1959) that the normal solution exists and is unique; it can be found by the formula

$$y = A^{+} f, \tag{4.6.25}$$

where A^{+} is the *Moore–Penrose pseudo-inverse $n \times m$-matrix.* The matrix A^{+} is defined by the relation

$$AA^{+}A = A \tag{4.6.26}$$

or by the asymptotic formula

$$A^{+} = \lim_{\alpha \to 0} (\alpha E + A^{*}A)^{-1} A^{*}. \tag{4.6.27}$$

There exists a unique pseudo-inverse matrix A^{+} defined by relation (4.6.26) or (4.6.27).

Yet, the matrix A^{+} is inconvenient to find by relation (4.6.26) or (4.6.27). In practice, the matrix A^{+} can be most easily found by the formula (Verlan' and Sizikov, 1986, p. 508; Gantmacher, 1959)

$$A^{+} = C^{*}(CC^{*})^{-1}(B^{*}B)^{-1}B^{*}, \tag{4.6.28}$$

where $A = BC$ is a *skeleton* (ambiguous) *product decomposition* of the matrix A, in which B is some non-degenerate $m \times r$-matrix; C is some $r \times n$-matrix equal to $C = B^{-1}A$; and $r = r_A = r_B = r_C$, where r_A, r_B, and r_C are the ranks of the matrices A, B, and C, respectively. Different skeleton product decompositions of A yield one and the same pseudo-inverse matrix A^{+}.

In the case of a square nondegenerate matrix A we have $A^{+} = A^{-1}$, and in the case of an overdetermined system, $A^{+} = (A^{*}A)^{-1}A^{*}$ (cp. (4.6.13)), i. e., notation (4.6.25) is common for overdetermined, determined, and underdetermined SLAEs. Moreover, solution (4.6.25), which can be appropriately written in the form $y^{+} = A^{+}f$ (where y^{+} is the normal solution), provides for $\|Ay^{+} - f\| = 0$, i. e., this solution is a pseudo-solution (cp. (4.6.3)). Here, among all pseudo-solutions (numerous in the case of an overdetermined SLAE) the solution y^{+}, as the normal solution, has the least norm. In other words, the normal solution is simultaneously a pseudo-solution. That is why the normal solution is also called the *normal pseudo-solution* (the smoothest solution among the set of all pseudo-solutions). The normal pseudo-solution exists and is unique.

Example. By way of example, consider the underdetermined SLAE (4.4.3). In Section 4.4, several solutions of this SLAE were given: 1) $y_1 = 1$, $y_2 = 2$, the norm of the solution (defined by (4.2.2)) is $\|y\| = \sqrt{1^2 + 2^2} = 2.24$; 2) $y_1 = 2$, $y_2 = 8/3$, $\|y\| = 3.34$; 3) $y_1 = 0$, $y_2 = 4/3$, $\|y\| = 1.33$.

Let us find the normal solution of (4.4.3) using formulas (4.6.25) and (4.6.28). In this example, $m = 1$, $n = 2$, the matrix $A = (2 \;\; -3)$, and the rank $r_A = 1$. We use the skeleton product decomposition $A = BC$, where B is some 1×1-matrix (i. e., a scalar) and C is some 1×2-matrix (i. e., a row vector); in this case, $r = r_A = r_B = r_C = 1$. We set $B = b$, where $b \neq 0$ is some number. Then, $C = B^{-1}A = (2/b \;\; -3/b)$,

$$C^\top = \begin{pmatrix} 2/b \\ -3/b \end{pmatrix}, \qquad CC^\top = (13/b^2), \qquad (CC^\top)^{-1} = (b^2/13),$$

$$C^\top(CC^\top)^{-1} = \begin{pmatrix} 2b/13 \\ -3b/13 \end{pmatrix}, \qquad (B^\top B)^{-1}B^\top = (1/b).$$

Formula (4.6.28) yields for any b:

$$A^+ = C^\top(CC^\top)^{-1}(B^\top B)^{-1}B^\top = \begin{pmatrix} 2/13 \\ -3/13 \end{pmatrix}.$$

Since the right-hand side is $f = (-4)$, then, according to (4.6.25), the normal solution is

$$y = A^+ f = \begin{pmatrix} -8/13 \\ 12/13 \end{pmatrix}.$$

Thus, the normal solution of (4.4.3) is $y_1 = -8/13 \approx -0.615$ and $y_2 = 12/13 \approx 0.923$, and the norm of this solution is $\|y\| \approx 1.113 = \min$. The check of the solution is $y_2 = (2y_1 + 4)/3 = 12/13$.

Next, we have

$$A = (2 \;\; -3), \qquad A^* = \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \qquad A^*A = \begin{pmatrix} 4 & -6 \\ -6 & 9 \end{pmatrix}.$$

The characteristic equation

$$\begin{vmatrix} 4-\lambda & -6 \\ -6 & 9-\lambda \end{vmatrix} = 0$$

yields the roots $\lambda_1(A^*A) = \lambda(A^*A)_{\max} = 13$ and $\lambda_2(A^*A) = \lambda(A^*A)_{\min} = 0$. Hence, the singular numbers here are

$$\mu(A)_{\max} = \sqrt{13}, \qquad \mu(A)_{\min} = 0,$$

and the condition number and the determinant are

$$\operatorname{cond}(A) = \mu(A)_{\max}/\mu(A)_{\min} = \infty, \qquad \det(A^*A) = 0,$$

i. e., the matrix A^*A is degenerate and the inverse matrix $(A^*A)^{-1}$ is nonexistent: its norm is $\|(A^*A)^{-1}\| = 1/\mu(A)_{\min} = \infty$. In other words, SLAE (4.4.3) is an unstable SLAE (not only the second, but also the third condition for well-posedness according to Hadamard is violated here).

4.6.6. PIMM as applied to other equations

The above pseudo-inverse matrix method can also be employed if (4.6.22) is an arbitrary equation, for instance, the first-kind integral Fredholm equation (4.5.1), differential equations (4.4.9), (4.4.11), (4.4.13) and (4.4.15), operator equation (4.4.1), etc. In this case, to be chosen as the solution of a particular (integral, differential, operator, etc.) equation is the *normal solution y* that satisfies condition (4.6.23) and can be found by formula (4.6.25), where A^+ is the *pseudo-inverse operator* defined by (4.6.27) or (4.6.28). For more detail see Section 4.7.

4.6.7. General conclusion

From this section, the following conclusion can be drawn: with the so-called *normal pseudo-solution* (the smoothest pseudo-solution among the whole set of pseudo-solutions), the first two conditions of well-posedness according to Hadamard (solution existence and solution uniqueness) will be satisfied.

Yet, neither LSM nor PIMM solves the problem of solution instability, i. e., generally speaking, with these methods the third condition for well-posedness according to Hadamard is not fulfilled. How the problem of stability can be solved will be discussed in Sections 4.7 and 4.8.

4.7. TIKHONOV REGULARIZATION METHOD

In this section, we will consider one of most powerful method for solving ill-posed problems, namely, the Tikhonov regularization method

(Bakushinsky and Goncharsky, 1994; Verlan' and Sizikov, 1986; Ivanov, Vasin, and Tanana, 2002; Lavrent'ev, Romanov, and Shishatskii, 1997; Sizikov, 2001; Tikhonov and Arsenin, 1977; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995; Vasilenko, 1979; Voskoboynikov, Preobrazhenski, and Sedel'nikov, 1984; Lavrent'ev, 1981; Morozov, 1984; Tikhonov, Leonov, and Yagola, 1997), and show how this method can be applied to solving various equations. The Tikhonov regularization method presents further development of the Gauss least-squares method (LSM), which gives a pseudo-solution, and the Moore–Penrose pseudo-inverse matrix method (PIMM), which gives the normal solution.

4.7.1. Essence of the method

To begin with, consider the Tikhonov regularization method as applied to the first-kind operator equation

$$Ay = f, \qquad y \in Y, \quad f \in F, \tag{4.7.1}$$

where Y and F are some metric spaces, $A : Y \to F$ is some linear operator that acts from Y into F, $f \in F$ is a given element (right-hand side of the equation), and $y \in Y$ is the sought element (solution). Let, instead of the exact f and A, known are their approximations $\tilde{f}$ and $\tilde{A}$ such that

$$\|\tilde{f} - f\| \le \delta,$$
$$\|\tilde{A} - A\| \le \xi,$$

where $\delta > 0$ and $\xi \ge 0$ are the errors in setting f and A, or, more precisely, their upper bounds. Thus, instead of (4.7.1), it is required to solve the equation

$$\tilde{A}\tilde{y} = \tilde{f}, \qquad \tilde{y} \in Y, \quad \tilde{f} \in F. \tag{4.7.2}$$

Remark. If under operator equation (4.7.1) an integral equation is meant, then the error δ in setting the right-hand side f of the integral equation arises from the measurement error of f, and the error ξ of the operator A, from the uncertainty of the kernel $K(x,s)$ and from the error brought about by the numerical solution algorithm, for instance, by the error brought about by the quadrature method. If, alternatively, we deal with SLAE, then δ is the error in setting the right-hand side vector, and ξ is the error in setting the elements of the matrix A. If, alternatively, we deal with homogeneous differential equations (Laplace equation, heat conduction equation, etc.), then $\delta = 0$, and ξ is the error in setting the boundary conditions plus the uncertainty brought about the numerical solution algorithm, for instance, that of the grid method.

To simplify notation, we will subsequently use notation (4.7.1), bearing in mind that, in fact, we consider equation (4.7.2).

In Section 4.4, the classical (proposed by Hadamard) definition of well-posedness was given. A. N. Tikhonov gave a new definition of well-posedness, later called by M. M. Lavrent'ev the *well-posedness according to Tikhonov.*

Definition. The problem for solving equation (4.7.1) is called a *conditionally well-posed*, or *Tikhonov-well-posed problem* if (Ivanov, Vasin, and Tanana, 2002; Lavrent'ev, Romanov, and Shishatskii, 1997; Tikhonov and Arsenin, 1977):

1) it is known *a priori* that the solution y exists and belongs to a certain set (correctness set) $M \subset Y$: $y \in M$;

2) the solution is unique in the class of functions belonging to M, i. e., the operator A is invertible on the set M;

3) the solution y depends on f continuously or, in other words, arbitrarily small variations of f that do not expel the solution y beyond the limits of M result in arbitrarily small variations of y, i. e., the inverse operator A^{-1} is a continuous operator.

The difference between the conditional correctness (according to Tikhonov) and the classical correctness (according to Hadamard) consists in the introduction of the *correctness set* that substantially narrows the class of possible solutions.

The most typical example of correctness set is a *compact* (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) (examples of compacts are the set of functions monotonically bounded from above and below, the set of bounded convex functions, etc.). Yet, in the majority of applied problems the class of possible solutions Y is not a compact (i. e., it is impossible to impose, based on physical considerations, any hard constraints on the solution) and, in addition, variations of the right-hand side f of equation (4.7.1) resulting from its errors can expel f beyond the limits of the set AY, the image of the set Y mapped by the operator A. Such problems are called *essentially ill-posed problems* (Tikhonov and Arsenin, 1977).

A. N. Tikhonov developed (Tikhonov and Arsenin, 1977) a fundamentally new approach yielding stable solutions of essentially ill-posed problems.

This approach is based on the notion of *regularizing operator* (RO), or *regularizing algorithm* (RA).

In the Tikhonov regularization method, two conditions are set, the condition of minimum discrepancy of type (4.6.4), as in the Gauss LSM, and the condition of minimum solution norm of type (4.6.24), as in the Moore–Penrose PIMM. As a consequence, we arrive at a conditional-minimization problem to be normally solved by the *Lagrange method of undetermined multipliers*; namely, to be satisfied is the minimization condition for the *smoothing functional*:

$$\|Ay - f\|^2 + \alpha\|y\|^2 = \min_y, \tag{4.7.3}$$

where $\alpha > 0$ is the *regularization parameter* that plays role of a Lagrange undeterminate multiplier (remarks concerning the choice of α will be given below). Condition (4.7.3) yields the *Euler–Tikhonov equation* (cp. (4.6.5)) (Tikhonov and Arsenin, 1977)

$$(\alpha E + A^*A)y_\alpha = A^*f, \tag{4.7.4}$$

where E is the unit operator ($Ey = y$). Thus, instead of first-kind equation (4.7.1) we have second-kind equation (4.7.4).

4.7.2. Brief analysis of the method

Let us consider condition (4.7.3) and equation (4.7.4).

If $\alpha = 0$, then the Tikhonov regularization method passes into the Gauss LSM with the minimum discrepancy $\|Ay - f\|^2$ and very unstable solution. As the value of α increases, the solution y_α becomes smoother and more stable, i. e., the solution norm $\|y_\alpha\|^2$ decreases, although the discrepancy increases. The truth is somewhere in the middle; i. e., at some moderate α the solution y_α will display some moderate smoothness and some moderate discrepancy.

If $\delta, \xi \to 0$, then $\alpha \to 0$ and

$$y_\alpha = \lim_{\alpha \to 0} (\alpha E + A^*A)^{-1}A^*f \equiv A^+f \tag{4.7.5}$$

(cp. (4.6.27)); i. e., the solution y_α passes into the normal pseudo-solution. Thus, the Tikhonov regularization method presents a generalization of the Gauss least-squares method and the Moore–Penrose pseudo-inverse operator method.

The Tikhonov regularization method is a stable method, i. e., with it, the third condition for well-posedness according to Hadamard is fulfilled; this stability can be mathematically explained as follows. The operator A^*A in (4.7.4) is a symmetric positively defined operator, and all eigenvalues of this operator are therefore real and nonnegative numbers, $\lambda_i(A^*A) \geq 0$, and, in addition, $\lambda(A^*A)_{\min} = 0$. The involvement of the addend αE in (4.7.4) makes all $\lambda_i(A^*A)$ greater by α and, hence, $\lambda(\alpha E + A^*A)_{\min} = \alpha$. As a consequence, the operator $\alpha E + A^*A$ becomes an invertible operator, the norm of the inverse operator $\|(\alpha E + A^*A)^{-1}\| = 1/\alpha \neq \infty$, and the problem becomes a stable problem.

The solution of (4.7.4) is

$$y_\alpha = (\alpha E + A^*A)^{-1}A^*f. \tag{4.7.6}$$

A matter of importance here is the choice of the regularization parameter α. Several ways for choosing α in the Tikhonov regularization method were developed. Consider two of these ways.

The first one is the discrepancy principle (Verlan' and Sizikov, 1986; Tikhonov and Arsenin, 1977; Morozov, 1984). In this wasy, the value of α is to be chosen from the condition (at $\xi = 0$)

$$\|Ay_\alpha - f\| = \delta. \tag{4.7.7}$$

If $\|f\| \geq \delta$, then there exists a unique solution of (4.7.7) with respect to α. With $\xi \neq 0$, the discrepancy principle passes into the *generalized discrepancy principle* (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995). Here, it is conventional practice to subdivide all ways for choosing α that use the errors δ and ξ into two types of ways, namely, into *a priori* and *a posteriori* ways (Tikhonov, Leonov, and Yagola, 1997). Both the discrepancy principle and the generalized discrepancy principle are *a posteriori* ways.

The second one is the fitting (inspection) way (Sizikov, 2001; Belov and Sizikov, 2001; Sizikov, Akhmadulin, and Nikolaev, 2002; Sizikov and Belov, 2000). In this way, to be found are solutions y_α for several values of α in the “reasonable” interval; afterwards, the final choice of α is to be made based not on mathematical criteria but on physiological perception criteria. This way closely resembles the TV contrast tuning procedure. Indeed, as the value of α decreases, then the solution y_α becomes more unstable (the picture contrast increases, provided that under the picture the function y_α is meant); and vice versa, as the value of α increases, then the solution y_α becomes more smooth (the picture contrast decreases). In spite of its simplicity, the fitting way proved to be a highly efficient method in solving such problems as the reconstruction of blurred and defocused images (see Sections 5.1 and 5.2), obtaining x-ray tomograms (see Section 5.3), etc.

Other ways for choosing the value of the regularization parameter α include the way of the quasi-optimal (quasi-best) value, the ratio way, the way of independent realizations, the cross-validation method, the modeling way, etc. (Verlan' and Sizikov, 1986; Tikhonov and Arsenin, 1977; Voskoboynikov, Preobrazhenski, and Sedel'nikov, 1984).

Below, we will consider the Tikhonov regularization method as applied to stable solution of various equations (integral equations, SLAE, differential equations, etc.).

Remark. Apart from the Tikhonov regularization method, the following *stable methods* were developed for solving first-kind equations (integral, algebraic, and differential equations): the regularization methods (Lavrent'ev, Bakushinsky, Denisov), the iterative-regularization methods (Fridman, Morozov, Vainikko, Bakushinsky, Emelin, Krasnosel'skii), the local-regularization methods (Arsenin, Voiskoboinikov, Sizikov), the Ivanov quasisolution method, the Tikhonov solution-on-the-compact method, the Morozov descriptive regularization method, etc. (Verlan' and Sizikov, 1986; Ivanov, Vasin, and Tanana, 2002; Lavrent'ev, Romanov, and Shishatskii, 1997; Tikhonov, Arsenin, and Timonov, 1987; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995; Voskoboynikov, Preobrazhenski and Sedel'nikov, 1984; Lattes and Lions, 1969; Morozov, 1984). These are *deterministic regularization methods.* One has also developed the *statistical regularization methods* (Turchin, Khalfin, Lavrent'ev, Zhukovskii), the *optimal filtration methods* (Wiener, Kalman), the *suboptimal filtration methods* (Brunner, Sizikov), etc. (Verlan' and Sizikov, 1986; Voskoboynikov, Preobrazhenski, and Sedel'nikov, 1984; Brunner and Sizikov, 1998). Of these methods, in the present monograph only the Tikhonov regularization method will be considered.

4.7.3. Solving the Fredholm integral equation of the first kind

Consider the *Fredholm integral equation of the first kind*

$$Ay \equiv \int_a^b K(x,s)y(s)\,\mathrm{d}s = f(x), \qquad c \le x \le d. \tag{4.7.8}$$

The Tikhonov regularization method as applied to equation (4.7.8) yields the second-kind Fredholm integral equation (cp. (4.6.15) and (4.7.4)):

$$\alpha y_\alpha(t) + \int_a^b R(t,s) y_\alpha(s)\,\mathrm{d}s = F(t), \qquad a \le t \le b, \tag{4.7.9}$$

with some new (symmetric and positively defined) kernel (cp. (4.6.16))

$$R(t,s) = R(s,t) = \int_c^d K(x,t)K(x,s)\,\mathrm{d}x \tag{4.7.10}$$

and with some new right-hand side (cp. (4.6.17))

$$F(t) = \int_c^d K(x,t) f(x)\,\mathrm{d}x. \tag{4.7.11}$$

Numerical algorithm. Consider the numerical solution of integral equation (4.7.9). We will dwell here on the *quadrature method*, one of most powerful solution algorithm for this equation (Verlan' and Sizikov, 1986, pp. 249–251; Sizikov, 2001, p. 195; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995).

Suppose that the right-hand side $f(x)$ is given in tabulated form on the following, generally speaking, non-uniform x-grid of nodal points:

$$c = x_1 < x_2 < x_3 < \cdots < x_l = d, \tag{4.7.12}$$

and the solution $y_\alpha(s)$ is sought on another non-uniform s-grid of nodal points coincident with the t-grid:

$$a = s_1 = t_1 < s_2 = t_2 < s_3 = t_3 < \cdots < s_n = t_n = b, \tag{4.7.13}$$

and, in addition, $l \lesseqgtr n$.

Then, we apply some quadrature formula, in particular, the trapezoid formula, to the integral in (4.7.9) and obtain:

$$\alpha y_k + \sum_{j=1}^{n} r_j R_{kj} y_j = F_k, \qquad k = 1,\ldots,n, \tag{4.7.14}$$

where $y_k \equiv y_\alpha(t_k)$, $y_j \equiv y_\alpha(s_j)$, $R_{kj} \equiv R(t_k, s_j)$, and $F_k \equiv F(t_k)$. In a similar manner, we approximate the integrals in (4.7.10) and (4.7.11) with finite sums by the quadrature formula and obtain:

$$R_{kj} = R_{jk} = \sum_{i=1}^{l} p_i K_{ik} K_{ij}, \qquad k,j = 1,\ldots,n, \tag{4.7.15}$$

$$F_k = \sum_{i=1}^{l} p_i K_{ik} f_i, \qquad k = 1, \ldots, n, \tag{4.7.16}$$

where $K_{ik} \equiv K(x_i, t_k)$, $K_{ij} \equiv K(x_i, s_j)$, and $f_i \equiv f(x_i)$. In (4.7.14)–(4.7.16), r_j and p_i are the quadrature coefficients.

Notation (4.7.14) presents a SLAE with respect to y_j, $j = 1, \ldots, n$. Details concerning the numerical algorithm can be found in Verlan' and Sizikov (1986, pp. 249–251).

Programs. In Verlan' and Sizikov (1986, pp. 371–379), the TIKH1, TIKH2, TIKH3, TIKH4 and TIKH5 programs, and in Tikhonov, Goncharsky, Stepanov, and Yagola (1995), the PTIMR and PTIZR programs written in the FORTRAN computer language and implementing the Tikhonov regularization method according to formulas of type (4.7.12)–(4.7.16) with different choices of α, are presented.

Numerical example. Let us give some results of solution of integral equation (4.7.8) by the Tikhonov regularization method. Consider a model example (Sizikov, 2001, pp. 198, 199), in which the exact solution was set as a superposition of five gaussians:

$$\begin{aligned} y(s) = 6.5\mathrm{e}^{-[(s+0.66)/0.085]^2} + 9\mathrm{e}^{-[(s+0.41)/0.075]^2} \\ + 12\mathrm{e}^{-[(s-0.14)/0.084]^2} + 14\mathrm{e}^{-[(s-0.41)/0.095]^2} + 9\mathrm{e}^{-[(s-0.67)/0.065]^2}, \end{aligned}$$

$a = -0.85$, $b = 0.85$, $c = -1$, $d = 1$, the kernel

$$K(x, s) = \sqrt{q/\pi}\, \mathrm{e}^{-q(x-s)^2/(1+x^2)},$$

where the exact value of q is $q = 59.924$. The discretization steps are $\Delta x = \Delta s = \text{const} = 0.0125$, and the numbers of nodes are $l = 161$ and $n = 137$.

First, the *direct problem* was solved. The values of f_i, $i = 1, \ldots, l$, were calculated by the trapezoid quadrature formula

$$f_i \approx \sum_{j=1}^{N} p_j K(x_i, s_j) y(s_j)$$

with a step much lesser than $\Delta s = 0.0125$, i. e., with $N \gg n$. Then, the RNDAN program (random-number generator) (Sizikov, 2001, p. 152) was used to add to the values of f_i the errors δf_i distributed by the normal law with zero average and with the root-mean-square deviation equal to

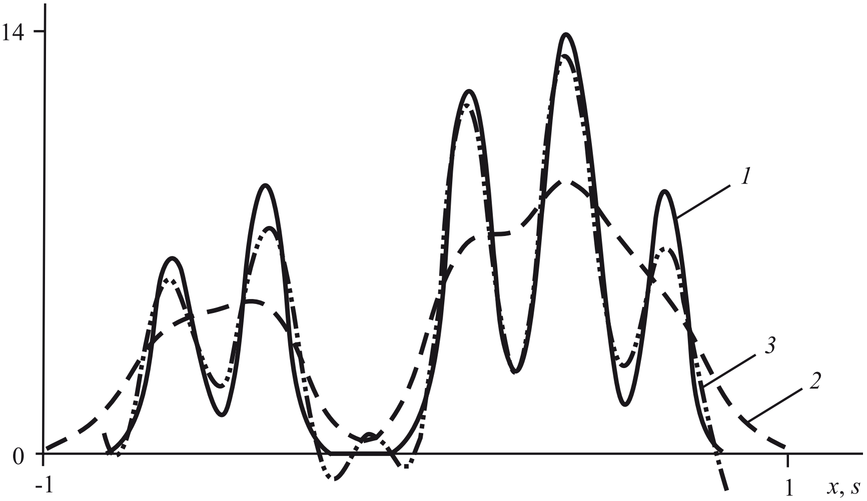

Figure 4.3. Solution of the Fredholm integral equation of the first kind obtained by the Tikhonov regularization method: *1* — $y(s)$, *2* — $f(x)$, *3* — $y_\alpha(s)$

$\delta = 0.0513$; the latter value corresponds to the relative right-hand side error of $\delta_{\text{rel}} \approx 1\,\%$.

Afterwards, the *inverse problem* was solved. The exact value of q was replaced with an approximate value $\tilde{q} = 60$; this replacement corresponds to $\xi_{\text{rel}} \approx 1\,\%$. To solve the example, the TIKH2 and TIKH3 programs were used. The value of the regularization parameter α was chosen by the modeling way (by solving several close model examples) with the help of the TIKH2 program: $\alpha = 10^{-3.5}$. Then, the TIKH3 program was used to find the solution $y_\alpha(s)$. The exact solution $y(s)$, the exact right-hand side $f(x)$, and the regularized solution $y_\alpha(s)$ are shown in Figure 4.3 (the solution obtained for this example without regularization is shown in Figure 4.2).

This example is typical, for instance, for the *hydroacoustic problem* (reduction of extended signals (Sizikov, 2001, pp. 114–115)). The function $y(s)$ is the sound-field intensity dependent on the direction s; the function $K(x, s)$ is the directional characteristic of an antenna (by power) used to measure the field; and $\tilde{f}(x)$ is the measurement result (indicator process). In the problem, it is required to mathematically reconstruct the input field $y(s)$ from the measured function $\tilde{f}(x)$ and from the known (with some error) function $\tilde{K}(x, s)$.

Here, the unknown field $y(s)$ displays considerable fluctuations. Yet, the measured function $f(x)$, first, shows almost no fluctuations due to the finite width of the directional characteristic $K(x,s)$ and, second, this function is distorted with measurement errors.

Figure 4.3 shows that, with the Tikhonov regularization method, the field $y(s)$ at the antenna input could be reconstructed rather accurately. In other words, the meter (antenna) connected to a computing device (computer, specially designed processor, etc.) with a program implementing the Tikhonov regularization method, offers a new measuring means that enables a better resolving power.

This example can also be interpreted as an example from the *inverse spectroscopy problem* (Sizikov, 2001, p. 79), assuming that the variable x, and also the variable s, is the frequency (or wavelength); the function $y(s)$ is the true distribution of the intensity over the spectrum; $K(x,s)$ is the frequency characteristic (FC) of the spectrometer; and $f(x)$ is the measured spectrum. In the case under consideration, it is required to mathematically reconstruct the true spectrum $y(s)$ from the measured spectrum $\tilde{f}(x)$ and the known FC, thus improving the resolving power of the spectrometer.

Among many other problems, the above (hydroacoustic and spectroscopy) problems are problems that illustrate the so-called *Rayleigh reduction problem* (Sizikov, 2001, p. 111).

Consider an important particular case of equation (4.7.8), the one- and two-dimensional first-kind convolution-type Fredholm integral equations.

Tikhonov regularization method as applied to first-kind convolution-type integral equations. Consider the *one-dimensional first-kind convolution-type Fredholm integral equation*:

$$Ay \equiv \int_{-\infty}^{\infty} K(x-s)y(s)\,\mathrm{d}s = f(x), \qquad -\infty < x < \infty. \tag{4.7.17}$$

In Section 4.5, we saw how this equation can be solved by the Fourier transformation method and showed that the solution thus obtained is unstable, this instability resulting from the strong sensitivity of high Fourier harmonics to errors in setting the right-hand side $f(x)$ of the equation.

The Tikhonov regularization method offers a stable algorithm for solving equation (4.7.17). According to this method, the stable solution of the equation is to be found from the *minimum condition* for the *smoothing functional* (cp. (4.7.3))

$$\int_{-\infty}^{\infty} [Ay - f(x)]^2\,\mathrm{d}x + \alpha \int_{-\infty}^{\infty} M(\omega)|Y(\omega)|^2\,\mathrm{d}\omega = \min_{y}, \tag{4.7.18}$$

where

$$M(\omega) = |\omega|^{2q} \tag{4.7.19}$$

is the *q-order regularizer*, where $q > 0$ is some set regularization order, for instance, $q = 1$.

Condition (4.7.18) yields the following expression for the *regularized solution* (cp. (4.5.6)):

$$y_\alpha(s) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\lambda(-\omega)F(\omega)}{L(\omega) + \alpha M(\omega)} \, \mathrm{e}^{-i\omega s} \, \mathrm{d}\omega, \tag{4.7.20}$$

where

$$\lambda(\omega) = \int_{-\infty}^{\infty} K(x) \mathrm{e}^{i\omega x} \, \mathrm{d}x, \tag{4.7.21}$$

$$F(\omega) = \int_{-\infty}^{\infty} f(x) \mathrm{e}^{i\omega x} \, \mathrm{d}x \tag{4.7.22}$$

are the spectra, or Fourier transforms, of the kernel $K(x)$ and the right-hand side $f(x)$ and

$$L(\omega) = |\lambda(\omega)|^2 = \lambda(\omega)\lambda(-\omega) = \mathrm{Re}^2\, \lambda(\omega) + \mathrm{Im}^2\, \lambda(\omega). \tag{4.7.23}$$

Compare the classical solution (4.5.6) and the regularized solution (4.7.20). In (4.7.20), the subintegral function vanishes as $|\omega| \to \infty$ due to the term $\alpha M(\omega)$, which suppresses the response of high Fourier harmonics to the error in setting the right-hand side $f(x)$; the larger the values of α and q, the more pronounced this suppression is. The larger q, the stronger is the suppression of high harmonics in the solution in comparison with low harmonics; the parameter α determines the global suppression: an increase in α suppresses all harmonics. That is why, considering q, one has to apply the following *rule*: if the sought solution contains a fluctuating component (and, hence, the solution spectrum contains a high-frequency component), then the value of q is to be reduced, for instance, to $q = 1$; otherwise, if the solution is smooth (and its spectrum contains no high-frequency harmonics), then the value of q is to be increased, for instance, to $q = 2$. As for α, the ways in which this parameter can be chosen are the same as for equation (4.7.9) (the discrepancy principle, the fitting way, etc.).

Numerical algorithm for finding the solution $y_\alpha(s)$ can be obtained by replacing integrals (4.7.20)–(4.7.22) with finite sums (according to the rectangle formula, trapezoid formula, etc.). As a result, continuous Fourier transformations (CFT) will be replaced by discrete Fourier transformations (DFT) and even by fast Fourier transformations (FFT) (Sizikov, 2001, pp. 166–170; Rabiner and Gold, 1975).

The following *programs* for solving equation (4.7.17) by the Tikhonov regularization method were developed: PTIKR (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995), CONV1, CONV2, CONV3, CONV4, CONV5 (Verlan' and Sizikov, 1986, pp. 379–388), etc.

Numerical example. Consider solution results for the following *example* (Sizikov, 2001, pp. 201, 202; Brunner and Sizikov, 1998) (close to the example of Tikhonov, Goncharsky, Stepanov, and Yagola (1995)):

$$\int_0^1 k(t-\tau)y(\tau)\,\mathrm{d}\tau = f(t), \qquad 0 \le t \le 2, \tag{4.7.24}$$

where the (difference) kernel is

$$k(t) = \mathrm{e}^{-80(t-0.5)^2},$$

the exact solution is

$$y(\tau) = \Big\{0.45\exp\Big[-\Big(\frac{\tau-0.29}{0.18}\Big)^2\Big] + \exp\Big[-\Big(\frac{\tau-0.71}{0.16}\Big)^2\Big]\Big\}\sqrt{1-\Big(\frac{\tau-0.5}{0.5}\Big)^2},$$

and the local supports (carriers) (intervals of, generally speaking, nonzero values of the functions) are $\operatorname{supp} k(t) \subseteq [0,1]$, $\operatorname{supp} f(t) \subseteq [0,2]$, and $\operatorname{supp} y(\tau) \subseteq [0,1]$. Although the ranges of τ and t in equation (4.7.24) are finite, this equation is a convolution-type equation since outside the τ- and t-intervals (local supports) the functions $k(t)$, $f(t)$ and $y(\tau)$ are zero, i. e., as a matter of fact, the ranges of τ and t are infinite.

According to the discretization technique for the problem of (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995), instead of (4.7.24) we solved the equation

$$\int_{-0.5}^{1.5} k(t-\tau)y(\tau)\,\mathrm{d}\tau = f(t), \qquad 0 \le t \le 2,$$

with new supports of identical lengths: $\operatorname{supp} k(t) \subseteq [-0.5, 1.5]$, $\operatorname{supp} f(t) \subseteq [0,2]$, and $\operatorname{supp} y(\tau) \subseteq [-0.5, 1.5]$. The number of discretization nodes was set equal to $N = 64$. The discretization step was $h = 2/N$. The grid nodes along k and y were $t_j = -0.5 + h(j-1/2)$, $j = 1,\dots,N$, and the grid nodes along f were $t_i = h(i-1/2)$, $i = 1,\dots,N$.

Instead of the exact function $f(t)$, we used the function $\tilde f(t) = f(t) + \delta f(t)$ distorted with a random error $\delta f(t)$ distributed according to the normal law with zero average and with the root-mean-square deviation $\delta = 0.0164$; the latter value corresponds to the relative error of $\delta_{\mathrm{rel}} \approx 0.1\max_{t\in[0,2]} f(t)$ in setting the right-hand side of the equation.

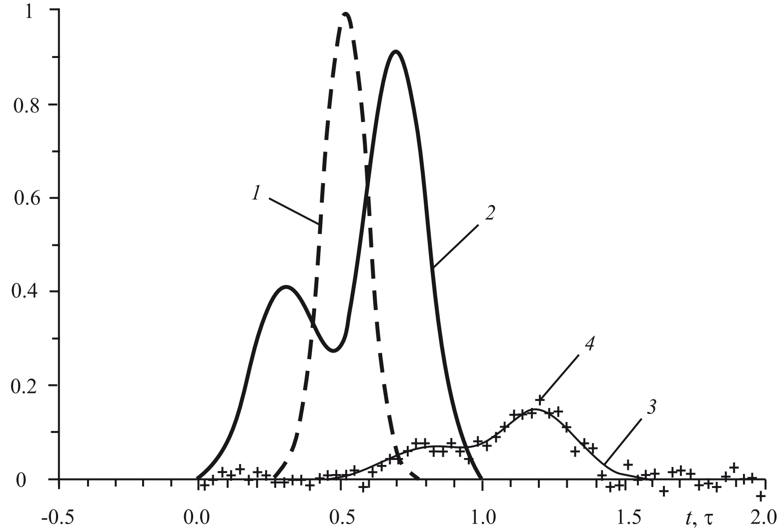

Figure 4.4. Numerical example: the first-kind convolution-type Fredholm integral equation:
1 — $k(t)$, *2* — $y(\tau)$, *3* — $f(t)$, *4* — $\tilde f(t)$

The exact functions $k(t)$, $y(\tau)$ and $f(t)$, and also the function $\tilde f(t)$, are shown in Figure 4.4.

Shown in Figure 4.5 are the exact solution $y(\tau)$ and the solution results obtained for $\operatorname{supp} k(t) \subseteq [-0.5, 1.5]$, $\operatorname{supp} f(t) \subseteq [0, 2]$, and $\operatorname{supp} y(\tau) \subseteq [-0.5, 1.5]$, namely, the too smooth yet very stable solution $y_{\alpha_1}(\tau)$ obtained by the Tikhonov regularization method with the value of α chosen by the discrepancy principle ($\alpha = \alpha_1 = 4\cdot 10^{-4}$) and the solution $y_{\alpha_2}(\tau)$ obtained by the Tikhonov regularization method with $\alpha = \alpha_2 \ll \alpha_1$, that is, $\alpha_2 = 10^{-5}$; in the latter solution, the two maxima are better resolved, although this solution is more unstable, displaying false alternating fluctuations, manifestations of the so-called Gibbs effect, at the ends of the interval.

Like several other examples (see Verlan' and Sizikov (1986, p. 283), Tikhonov, Goncharsky, Stepanov, and Yagola (1995)), the latter example shows that if the error in setting the initial data is large ($\approx 10\,\%$), then the discrepancy principle (and also the generalized discrepancy principle) yields the overestimated value of α, and one has to use additional ways, the fitting way, for instance.

As for the false fluctuations in the solution $y_{\alpha_2}(\tau)$ emerging at the edges of the interval $\tau \in [-0.5, 1.5]$, these fluctuations can be reduced or even completely eliminated, for instance, by shortening the lengths of the supports. Figure 4.6 shows the solution results obtained for the same example

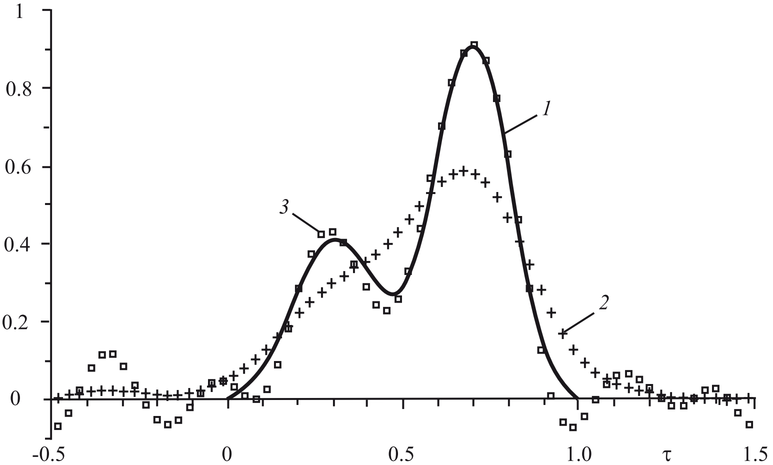

Figure 4.5. Solution of the first-kind convolution-type Fredholm integral equation obtained by the Tikhonov regularization method: *1* — $y(\tau)$, *2* — $y_{\alpha_1}(\tau)$, *3* — $y_{\alpha_2}(\tau)$

with shortened (identical) support lengths, namely, for $\operatorname{supp} k(t) \subseteq [0,1]$, $\operatorname{supp} f(t) \subseteq [0.5, 1.5]$, and $\operatorname{supp} y(\tau) \subseteq [0,1]$ (the designations here are the same as in Figure 4.5).

Next, consider the two-dimensional first-kind convolution-type Fredholm integral equation:

$$Ay \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(x_1 - s_1, x_2 - s_2) y(s_1, s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2 = f(x_1, x_2), \qquad -\infty < x_1, x_2 < \infty. \tag{4.7.25}$$

Let us solve this equation by the Tikhonov regularization method. We introduce the *minimum condition* for the *smoothing functional* (cp. (4.7.18))

$$\|Ay - f\|^2 + \alpha \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M(\omega_1, \omega_2)|Y(\omega_1, \omega_2)|^2\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2 = \min_y .$$

This condition yields the *regularized solution* (cp. (4.7.20))

$$y_\alpha(s_1, s_2) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\lambda(-\omega_1, -\omega_2) F(\omega_1, \omega_2)}{L(\omega_1, \omega_2) + \alpha M(\omega_1, \omega_2)}\,\mathrm{e}^{-i(\omega_1 s_1 + \omega_2 s_2)}\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2, \tag{4.7.26}$$

where (cp. (4.7.21)–(4.7.23))

$$F(\omega_1, \omega_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)\mathrm{e}^{i(\omega_1 x_1 + \omega_2 x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2, \tag{4.7.27}$$

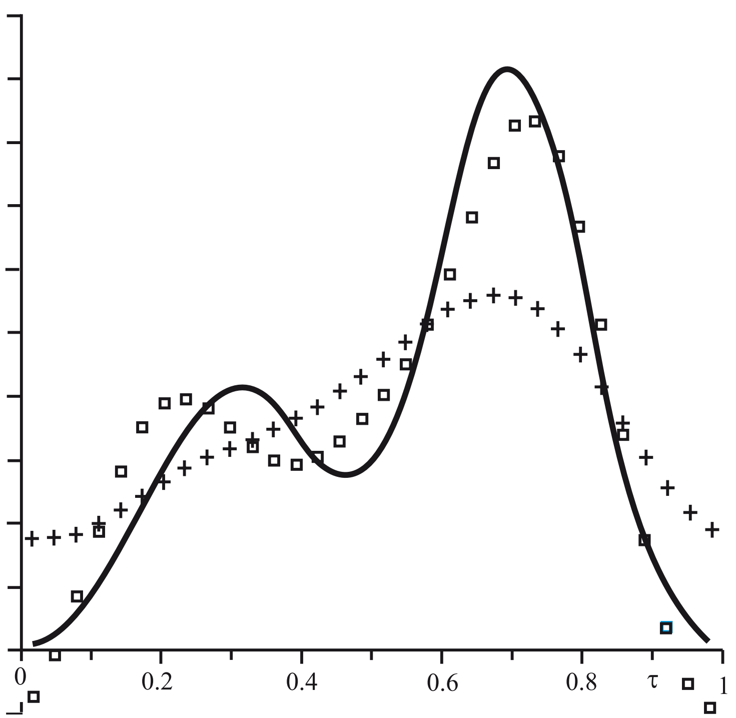

Figure 4.6.

$$\lambda(\omega_1,\omega_2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K(x_1,x_2)\mathrm{e}^{i(\omega_1x_1+\omega_2x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2, \qquad (4.7.28)$$

$$\begin{aligned}L(\omega_1,\omega_2)=|\lambda(\omega_1,\omega_2)|^2&=\lambda(\omega_1,\omega_2)\lambda(-\omega_1,-\omega_2)\\&=\mathrm{Re}^2\,\lambda(\omega_1,\omega_2)+\mathrm{Im}^2\,\lambda(\omega_1,\omega_2).\end{aligned} \qquad (4.7.29)$$

The regularizer $M(\omega_1,\omega_2)$ can be chosen, for instance, in the form (cp. (4.7.19)):

$$M(\omega_1,\omega_2)=(\omega_1^2+\omega_2^2)^2. \qquad (4.7.30)$$

In the practical realization of the two-dimensional problem, the values of x_1, s_1, x_2, s_2 are to be set discrete within some finite limits, and two-dimensional CFTs (4.7.26)–(4.7.28) are to be replaced with two-dimensional DFTs or FFTs.

The regularization parameter α is to be chosen in the same manner as in the case of one-dimensional equations (4.7.8) and (4.7.17) (by the discrepancy principle, generalized discrepancy principle, fitting way, etc.).

In Tikhonov, Goncharsky, Stepanov, and Yagola (1995), the PTITR program, written in the FORTRAN language, was reported which solves the two-dimensional first-kind convolution-type Fredholm integral equation (4.7.25) by the Tikhonov regularization method.

In Section 5.2, we will consider the problem of reconstruction of defocused images described by the two-dimensional first-kind convolution-type Fredholm integral equation. To solve this solution, the Tikhonov regularization method will be used.

4.7.4. Solution of degenerate and ill-conditioned SLAEs

In Sections 4.2–4.4, we already discussed degenerate and ill-posed SLAEs.

We write the system of linear algebraic equations (SLAE) in the form

$$Ay = f, \tag{4.7.31}$$

or, in more detail,

$$\sum_{j=1}^{n} a_{ij} y_j = f_i, \qquad i = 1, \ldots, m, \tag{4.7.32}$$

where A is an $m \times n$-matrix, y is an n-column vector, and f is an m-column vector.

This system may be unsolvable, i. e., having no solution (when being an overdetermined system), uniquely solvable (when being a determined system), or degenerate (having infinite number of solutions; in this case, this system is called an underdetermined system). Recall some definitions that were previously given in Sections 4.2–4.4.

The *rank* $r = \operatorname{rang}(A)$ of the matrix A is the maximum order of the nonzero minors of the matrix. The *rank of extended matrix* is the quantity $\rho = \operatorname{rang}(A \mid f)$. An SLAE is called an *overdetermined SLAE* if $m > n$, or, more exactly, the number of linearly independent rows in (4.7.32) is greater than n or $\rho > r$. An SLAE is called a *determined SLAE* if $m = n = r$ or $\rho = r$. An SLAE is called an *underdetermined SLAE* if $m < n$ or $r < n$ or $\rho < n$. An overdetermined SLAE has no solutions, a determined SLAE has a single solution, and an underdetermined SLAE has many solutions.

An SLAE is called a *degenerate SLAE* if the $m \times n$-matrix A of this SLAE has at least one zero singular number μ. In this case, the condition number is $\operatorname{cond}(A) = \mu(A)_{\max}/\mu(A)_{\min} = \infty$. If $m = n$, i. e., the matrix of the system is a square matrix, then the SLAE is called a *degenerate matrix* provided that the determinant $|A| = 0$. An SLAE is called an ill-conditioned SLAE if the condition number $\operatorname{cond}(A)$ of its matrix A is a sufficiently large yet finite number.

A *pseudo-solution* of SLAE (4.7.31) is a vector y that minimizes the discrepancy $\|Ay - f\|$. System (4.7.31) may have more than one solution. The *normal* solution of SLAE (4.7.31) is the pseudo-solution with the minimum norm $\|y\|$. For any system of form (4.7.31), the normal solution exists and is unique.

Yet, the normal solution may be unstable, and its degree of instability increases with increasing condition number

$$\operatorname{cond}(A) = \mu(A)_{\max}/\mu(A)_{\min}.$$

The essence of the Tikhonov regularization method for SLAEs consists in finding a stable approximation to the normal solution of the SLAE.

We introduce the minimum condition for the *smoothing functional* (cp. (4.7.3)):

$$\|Ay - f\|^2 + \alpha\|y\|^2 = \min_y, \tag{4.7.33}$$

where $\alpha > 0$ is the regularization parameter. Condition (4.7.33) yields a new SLAE (cp. (4.7.4))

$$(\alpha E + A^*A)y_\alpha = A^*f, \tag{4.7.34}$$

where E is the unit matrix. Alternatively, system (4.7.34) can be written as (cp. (4.6.6)):

$$(\alpha E + B)y_\alpha = u, \tag{4.7.35}$$

where (cp. (4.6.7) and (4.6.8))

$$B = A^*A, \tag{4.7.36}$$

$$u = A^*f \tag{4.7.37}$$

is the new matrix B (a symmetric positively defined matrix) and the new right-hand side u, or, provided that the matrix A is a real matrix, as

$$B = A^\top A, \tag{4.7.38}$$

$$u = A^\top f. \tag{4.7.39}$$

In more detail, SLAE (4.7.35) can be written as

$$\alpha y_i + \sum_{j=1}^{n} B_{ij} y_j = u_i, \qquad i = 1, \dots, n, \tag{4.7.40}$$

where, provided that A is a real matrix (cp. (4.6.11) and (4.6.12)),

$$B_{ij} = \sum_{k=1}^{m} A_{ki} A_{kj}, \qquad i, j = 1, \dots, n, \tag{4.7.41}$$

$$u_i = \sum_{k=1}^{m} A_{ki} f_k, \qquad i = 1, \dots, n. \tag{4.7.42}$$

The regularized solution of SLAE (4.7.34) or (4.7.35) is

$$y_\alpha = (\alpha E + A^* A)^{-1} A^* f \tag{4.7.43}$$

or

$$y_\alpha = (\alpha E + B)^{-1} u. \tag{4.7.44}$$

In practice, this solution can be found numerically by solving SLAE (4.7.40) with due regard for the fact that the matrix $\alpha E + B$ is a symmetric positively defined matrix; here, the Kraut, Cholesky, Voevodin or other methods can be used.

In this case, the regularization parameter α can be chosen by one of the following ways: discrepancy principle, general discrepancy principle, fitting way, etc.

4.7.5. Solving systems of ordinary differential equations with control function

Consider (like in Sections 4.3 and 4.4) the *system of ordinary differential equations (ODE) with control function* (Tikhonov and Arsenin, 1977):

$$\frac{dy(t)}{dt} = f(t, y, u), \qquad t_0 \le t \le T, \tag{4.7.45}$$

where $y(t) = \{y_1(t), \dots, y_n(t)\}$ is the unknown function, $f(t) = \{f_1(t), \dots, f_n(t)\}$ is the right-hand side, and $u(t) = \{u_1(t), \dots, u_m(t)\}$ is the control function (all the above functions are vector functions). The initial conditions are

$$y(t_0) = y_0,$$

where y_0 is a given vector.

The function $u(t)$ can be considered as a set function with which the control will not be optimal; alternatively, this function can be considered to be the sought function that minimizes some functional $F[u]$ (with possible constraints imposed on the function $u(t)$); in the latter case, the control will be optimal.

Consider a particular *example* of the Cauchy problem for a system of ODEs with optimum control, namely, the *vertical motion of a rocket* of variable mass in a uniform atmosphere launched so that a maximum height be reached. The motion of the rocket is governed by the following system of ODE (Tikhonov and Arsenin, 1977):

$$\left.\begin{aligned} \frac{dv(t)}{dt} &= \frac{au(t) - cv^2(t)}{m(t)} - g, \\ \frac{dm(t)}{dt} &= -u(t) \end{aligned}\right\} \tag{4.7.46}$$

with the initial conditions $m(0) = m_0$ and $v(0) = 0$. Here, $m(t)$ is the variable mass of the rocket with the fuel, $\mu \le m(t) \le m_0$, where μ is the net (without fuel) mass of the rocket; $v(t)$ is the velocity of the rocket; $u(t)$ is the control function, the time-dependent fuel consumption; and a, c, and g are some constants (a is the gas outflow velocity with respect to the rocket, and c is the drag coefficient due to air, and g is the free-fall acceleration).

The maximum height to be reached by the rocket is $H = H[v(u)] = \int_0^T v(t)\,\mathrm{d}t$; here, we assume that $v(T) = 0$.

It is required to find the optimum control function $u_{\mathrm{opt}}(t)$, a function such that $H = \max$. Usually, this problem is solved as follows. They introduce the so-called *objective functional*, normally in the form

$$F[u] = H[v(u)] - \tilde{H}, \tag{4.7.47}$$

where $\tilde{H}$ is some estimate of the sought maximum height. The problem on finding $u_{\mathrm{opt}}(t)$ is to be solved based on the minimum condition for the objective functional:

$$F[u_{\mathrm{opt}}(t)] = \min_{u(t)}. \tag{4.7.48}$$

Yet, this problem is ill-posed: small variations of F result in arbitrarily large variations of $u(t)$; i. e., the third condition for well-posedness according to Hadamard is violated here.

Consider a regularization procedure for the system of ODEs that makes the problem stable. According to the regularization method, to be minimized is the *smoothing functional*

$$\Phi_\alpha[u(t)] = F[u(t)] + \alpha\Omega[u(t)], \tag{4.7.49}$$

where the *stabilizing functional* $\Omega[u(t)]$ can be chosen, for instance, in the form

$$\Omega[u(t)] = \|u\|^2_{L_2} = \int_0^T u^2(t)\,\mathrm{d}t \tag{4.7.50}$$

or

$$\Omega[u(t)] = \|u\|^2_{W_2^2} = \int_0^T u''^2(t)\,\mathrm{d}t, \tag{4.7.51}$$

and $\alpha > 0$ is the regularization parameter.

Here, the functional $\Phi_\alpha[u(t)]$ is to be minimized numerically, by the gradient method, Newton method, chord method, etc., whereas the parameter α can be chosen, for instance, by the discrepancy principle. A numerical example is given by Tikhonov and Arsenin (1977).

4.7.6. Solving partial differential equations

Consider first the two-dimensional (elliptic) *Laplace equation* previously discussed in Sections 4.3 and 4.4:

$$\Delta u(x,y) \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{4.7.52}$$

with the following boundary conditions (Cauchy problem):

$$u(x,0) = f(x), \qquad \left.\frac{\partial u}{\partial y}\right|_{y=0} = \varphi(x), \tag{4.7.53}$$

where $f(x)$ and $\varphi(x)$ are given functions.

It is known (Lavrent'ev, Romanov, and Shishatskii, 1997) that under some rather loose requirements imposed on the solution $u(x,y)$ this solution exists and is unique, i. e., the first and second condition of well-posedness according to Hadamard are fulfilled here. Yet, as it was shown in Section 4.4, the solution is unstable.

Consider a one of possible construction algorithm for the regular (stable) solution of the Cauchy problem for the Laplace equation, namely, *reduction to an integral equation.*

We can reduce the Laplace equation (4.7.52) to the first-kind Fredholm integral equation with the Poisson kernel (Lavrent'ev, Romanov, and Shishatskii, 1997; Glasko, Mudretsova, and Strakhov, 1987):

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{y}{y^2 + (x-\xi)^2}\, u(\xi, y)\,\mathrm{d}\xi = u(x,0), \qquad -\infty < x < \infty. \tag{4.7.54}$$

In this equation, the quantity y is a parameter. That is why at each fixes value of y the equation is the one-dimensional first-kind convolution-type Fredholm integral equation. The latter equation can be effectively solved by the Fourier transformation method with regularization according to Tikhonov (see (4.7.17)–(4.7.23)).

In (Verlan' and Sizikov, 1986, pp. 30, 146, 152–157, 230) the matter of reduction of various (ordinary and partial) differential equations and systems of such equations with various boundary conditions to corresponding integral equations is considered in details.

Note that equation (4.7.54) is widely used in the inverse gravimetry problem, namely, in the problem of analytic continuation of the measured gravitational potential u from the Earth surface ($y = 0$) into depth (to the depth y), or in the recalculation problem for the gravitational potential.

Consider now the one-dimensional (parabolic) *heat conduction equation* discussed in Sections 4.3 and 4.4. Recall that, previously, four statements of the problem were considered.

In the first variant of the problem (direct problem, or the problem with forward time), it is required to solve the equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \qquad 0 \le x \le l, \quad t > 0, \tag{4.7.55}$$

with the initial condition $u(x,0) = \varphi(x)$ and the boundary conditions $u(0,t) = \psi_1(t)$ and $u(l,t) = \psi_2(t)$. The direct (classical) problem is a well-posed problem: the solution exists, is unique, and continuously depends on the initial data $\varphi(x)$.

In the second variant of the problem (inverse problem, or reverse-time problem) it is required to find the initial temperature distribution $u(x,0) = \varphi(x)$ from the given function $u(x,t_*)$, known at some time $t_* > 0$, by solving the problem towards decreasing t. Here, the initial conditions $\psi_1(t)$ and $\psi_2(t)$ are not set.

In the third variant of the problem it is required to solve the equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \qquad -\infty < x < \infty, \quad 0 \le t \le T, \tag{4.7.56}$$

with the initial conditions $u(x,0) = \varphi(x)$ and $u(x,T) = \chi(x)$, where $\varphi(x)$ and $\chi(x)$ are some given functions.

In the fourth variant of the problem it is required to solve equation (4.7.55) or (4.7.56), but with one initial condition: $u(x,0) = \varphi(x)$.

As it was noted in Section 4.4, variants 2–4 are incorrect. Consider two stable methods for solving variants 2–4 of the problem on solving the heat conduction equation.

The first method. The second and the third variants of the problem can be solved by the *method of quasi reversibility* (Lattes and Lions, 1969). According to this method, instead of (4.7.56), to be solved the *regularized equation* (Ivanov, Vasin, and Tanana, 2002; Tikhonov and Arsenin, 1977):

$$\frac{\partial u_\alpha(x,t)}{\partial t} = \frac{\partial^2 u_\alpha(x,t)}{\partial x^2} + \alpha\,\frac{\partial^4 u_\alpha(x,t)}{\partial x^4}, \qquad -\infty < x < \infty, \qquad 0 \le t \le T, \tag{4.7.57}$$

where $\alpha > 0$ is the regularization parameter. In this case, we set the only initial condition $u_\alpha(x,T) = \chi(x)$ and, proceeding from $t = T$ to $t = 0$ (towards decreasing t), find, in quite a stable manner, the function $u_\alpha(x,0) = \varphi_\alpha(x)$.

Afterwards, to check the solution, we solve the well-posed Cauchy problem for equation (4.7.57) with the initial condition $u_\alpha(x,0) = \varphi_\alpha(x)$ and, proceeding from $t = 0$ to $t = T$ (towards increasing t), we find $u_\alpha(x,t)$ and, in particular, $u_\alpha(x,T) = \chi_\alpha(x)$. As it was shown by Lattes and Lions (1969), average convergence takes place here: $\chi_\alpha(x) \to \chi(x)$ as $\alpha \to 0$.

The second method. This is the *integral-equation method* (Lavrent'ev, Romanov, and Shishatskii, 1997). According to this method, the (differential) heat conduction equation can be reduced to the integral equation

$$\frac{1}{2\sqrt{\pi t}}\int_{-\infty}^{\infty} \exp\Big[-\frac{(x-\xi)^2}{4t}\Big]u(\xi,t)\,\mathrm{d}\xi = \varphi(x), \qquad -\infty < x < \infty. \tag{4.7.58}$$

Here $\varphi(x) = u(x,0)$ is the initial condition. Equation (4.7.58) is the first-kind convolution-type Fredholm integral equation with the parameter t. By solving this equation at each fixed value of t by the one-dimensional Fourier transformation method and by the Tikhonov regularization method (see above), we obtain a stable solution $u_\alpha(\xi,t)$, where $\alpha > 0$ is the regularization parameter.

4.8. SOLUTION-ON-THE-COMPACT METHOD

4.8.1. The essence of the method

The possibility of constructing stable solutions of ill-posed problems is based on the use of *a priori* (additional) information about the solution. In Section 4.7, we discussed the case of essentially ill-posed problems, whose solution can be obtained by the methods (e. g., by the Tikhonov regularization method, etc.) in which *qualitative information* about the solution, namely, information about smoothness of the solution, is used. For instance, in the Tikhonov regularization method the solution smoothness is controlled by the regularization parameter α.

In this section we will consider cases in which *quantitative information* about the solution is available, that is, information that allows one to narrow the class of possible solutions to a compact set (compact) (for the definition of compact see Section 4.1). For the first time, this approach to solving ill-posed problem was advanced by A. N. Tikhonov (1943).

Consider the equation

$$Ay = f, \qquad y \in Y, \quad f \in F, \tag{4.8.1}$$

where Y and F are some metric spaces and A is a continuous operator. Let the problem for solving equation (4.8.1) be a Tikhonov-well-posed problem (see Section 4.7), for which in the space Y we isolate a certain subspace $M \subseteq Y$ representing the correctness set of the problem (i. e., the set on which a unique stable solution of the problem exists). We assume that the set M is a compact (*examples of compacts* will be given below), and the elements f and A are set exactly. In this case, the following theorem is valid (Tikhonov, 1943) (see also Verlan' and Sizikov, 1986, p. 302; Lavrent'ev, Romanov, and Shishatskii, 1997; Tikhonov and Arsenin, 1977).

Theorem 4.8.1 [Tikhonov's theorem]. *If a mapping $M \to AM$ of a set $M \subseteq Y$ onto a set $AM \subseteq F$ (image of M) is a continuous one-to-one mapping and the set M is a compact, then the inverse mapping $AM \to M$ is also a continuous mapping, i. e., the operator inverse to A is also a continuous operator on the set $AM \subseteq F$.*

Next, let all the conditions of Theorem 4.8.1 be fulfilled and, concerning the exact solution of (4.8.1), we know that $y \in M$, but, instead of f, known is a function $\tilde{f}$ such that $\rho_F(\tilde{f}, f) \le \delta$ and, in addition, $\tilde{f} \in AM$. Next, consider the set

$$Y_\delta^M = \{\tilde{y} \mid \tilde{y} \in M,\ \rho_F(A\tilde{y}, \tilde{f}) \le \delta\}.$$

Then, the inverse mapping will also be continuous, which fact can be formulated in terms of ε, δ as the following theorem (Verlan' and Sizikov, 1986, p. 302):

Theorem 4.8.2. *For any $\varepsilon > 0$, there exists $\delta_0(\varepsilon) > 0$ such that $\rho_Y(\tilde{y}, y) < \varepsilon$ (where y is the exact solution) for all $\tilde{y} \in Y_\delta^M$ as soon as $\delta < \delta_0(\varepsilon)$.*

It follows from Theorems 4.8.1 and 4.8.2 that, first, the solution on compact is a stable solution and, second, in the case of $\tilde f \in AM$ any element $\tilde y \in Y_\delta^M$ can be used as an approximate solution for the ill-posed problem. Here, $\lim_{\delta\to 0} \rho_Y(\tilde y, y) = 0$.

By way of illustration, consider the first-kind Fredholm integral equation

$$\int_a^b K(x,s)y(s)\,\mathrm{d}s = \tilde f(x), \qquad c \le x \le d, \tag{4.8.2}$$

where $\tilde f(x) \in L_2$. Consider the case of approximate right-hand side and exact kernel. The solution-on-the-compact method as applied to equation (4.8.2) consists in *minimizing* the *discrepancy functional* (cp. (4.6.4))

$$\Phi(y) \equiv \|Ay - \tilde f\|^2_{L_2} = \int_c^d \Big[\int\limits_a^b K(x,s)y(s)\,\mathrm{d}s - \tilde f(x)\Big]^2 \mathrm{d}x \tag{4.8.3}$$

on the functions $\tilde y \in Y_\delta^M$, where M is some compact defined, based on physical reasoning, with the help of some constraints in the form of equalities or inequalities (see below).

Let us make the aforesaid clear by giving some examples of compacts (Verlan' and Sizikov, 1986, pp. 303–305; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) for integral equation (4.8.2).

4.8.2. Solution on the set of monotonic bounded functions

This set presents a *first example of compacts.* Suppose that it is known *a priori* (from physical considerations) that the exact solution $\bar y$ of an ill-posed problem is a *monotonic* (for definiteness, *non-increasing*) *function bounded* from below and above respectively by some constants C_1 and C_2. Thus, concerning the sought exact solution we assume that $\bar y \in M \in L_2$, where M is the set of non-increasing, bounded above and below functions. We designate the set of non-increasing functions bounded from above and below as $Y{\downarrow}_{C_1}^{C_2}$ (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995), i. e.

$$Y{\downarrow}_{C_1}^{C_2} = \{y \mid y'(s) \le 0,\ C_1 \le y(s) \le C_2,\ a \le s \le b\}. \tag{4.8.4}$$

Notation (4.8.4) refers to the set of functions whose first derivative is a non-positive function bounded from above and below in the interval $[a,b]$. In other words, the constraints in the first example of compact are imposed on the function and on the first derivative of this function.

In this case, the *practical algorithm* that can be used to solve the problem is as follows.

We discretize equation (4.8.2) using, generally speaking, a non-uniform x-grid of nodal points

$$c = x_1 < x_2 < x_3 < \cdots < x_l = d \tag{4.8.5}$$

and a non-uniform s-grid of nodal points

$$a = s_1 < s_2 < s_3 < \cdots < s_n = b, \tag{4.8.6}$$

and also, for instance, the trapezoid formula.

As a result, the *problem of conditional minimization of functional* (4.8.3) acquires the form

$$\Phi(y) \equiv \sum_{i=1}^{l} p_i \Big[\sum_{j=1}^{n} r_j K_{ij} y_j - \tilde{f}_i \Big]^2 = \min_y, \tag{4.8.7}$$

with some *constraints* added in the form of inequalities (Verlan' and Sizikov, 1986, p. 303; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995):

$$\begin{gathered} y_{j+1} - y_j \le 0, \qquad j = 1, 2, \ldots, n-1, \\ y_1 - C_2 \le 0, \qquad C_1 - y_n \le 0. \end{gathered} \tag{4.8.8}$$

Here $K_{ij} \equiv K(x_i, s_j)$, $y_j \equiv y(s_j)$, $\tilde{f}_i \equiv \tilde{f}(x_i)$, and p_i and r_j are the quadrature coefficients (Verlan' and Sizikov, 1986, p. 250).

Here, as it was stated by Tikhonov, Goncharsky, Stepanov, and Yagola (1995), it is not necessary to seek the exact minimum of the discrepancy functional (4.8.7). It just suffices to make the following condition fulfilled:

$$\Phi(y) \equiv \sum_{i=1}^{l} p_i \Big[\sum_{j=1}^{n} r_j K_{ij} y_j - \tilde{f}_i \Big]^2 \le \delta^2, \tag{4.8.9}$$

where δ is the error in setting the right-hand side: $\|\tilde{f} - \bar{f}\|_{L_2} \le \delta$.

In the case of the discretization problem, the set $Y{\downarrow}_{C_1}^{C_2}$ (4.8.4) passes into the set (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995)

$$\begin{aligned} M{\downarrow}_{C_1}^{C_2} = \{ y \mid y \in \mathbb{R}^n,\ y_{j+1} - y_j \le 0,\ j = 1, 2, \ldots, n-1, \\ y_1 - C_2 \le 0,\ C_1 - y_n \le 0 \}. \end{aligned} \tag{4.8.10}$$

Thus, the problem of numerical solution in the class of monotonically non-increasing bounded upper and below functions consists in minimizing the discrepancy functional $\Phi(y)$ (cp. (4.8.7)) or in drawing the magnitude of this functional to δ^2 or lower (cp. (4.8.9)) with due regard for (4.8.8). Such a problem, called nonlinear (or, alternatively, linear, quadratic, etc.) programming, can be solved by various conditional-minimization methods: by the method of conjugate-gradient projections, by the method of conditional gradient, etc.

4.8.3. Solution on the set of monotonic bounded convex functions

This set presents a *second example of compacts.* Let the sought solution $\bar{y}$ belong to the set of *monotonically non-increasing convex functions bounded* from below and above by some constants C_1 and C_2. We designate this set as $\overset{\frown}{Y}\downarrow_{C_1}^{C_2}$ (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995), i. e.,

$$\overset{\frown}{Y}\downarrow_{C_1}^{C_2} = \{y \mid y'(s) \le 0,\ y''(s) \le 0,\ C_1 \le y(s) \le C_2,\ a \le s \le b\}. \tag{4.8.11}$$

Notation (4.8.11) means that, here, constraints are imposed on the function and on the first and second derivatives of the function.

In the case of discretized problem (see (4.8.5)–(4.8.7) and (4.8.9)), to be added are *constraints* in the form of inequalities (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995):

$$\begin{gathered} y_{j+1} - y_j \le 0, \qquad j = 1, 2, \ldots, n-1, \\ y_{j-1} - 2y_j + y_{j+1} \le 0, \qquad j = 2, 3, \ldots, n-1, \\ y_1 - C_2 \le 0, \qquad C_1 - y_n \le 0. \end{gathered} \tag{4.8.12}$$

In the case of discretized problem the set $\overset{\frown}{Y}\downarrow_{C_1}^{C_2}$ (4.8.11) passes into the set $\overset{\frown}{M}\downarrow_{C_1}^{C_2}$ (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) defined by constraints (4.8.12).

4.8.4. Solution on the set of bounded convex functions

This set presents a *third example of compacts.* Let the exact solution $\bar{y}$ belong to the set of *convex functions bounded* from below and above by some constants C_1 and C_2. We designate this set as $\overset{\frown}{Y}{}_{C_1}^{C_2}$ (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995), i. e.,

$$\overset{\frown}{Y}{}_{C_1}^{C_2} = \{y \mid y''(s) \le 0,\ C_1 \le y(s) \le C_2,\ a \le s \le b\}. \tag{4.8.13}$$

Notation (4.8.13) means that, here, *constraints* on the function and on the second derivative of the function are imposed.

In the case of discretized problem (see (4.8.5)–(4.8.7) and (4.8.9)), to be added are *constraints* in the form of inequalities (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995):

$$\begin{gathered} y_{j-1} - 2y_j + y_{j+1} \leq 0, \qquad j = 2, 3, \ldots, n-1, \\ y_1 - C_2 \leq 0, \qquad C_1 - y_n \leq 0. \end{gathered} \tag{4.8.14}$$

In the case of discretized problem, the set $\widehat{Y}{}_{C_1}^{C_2}$ (4.8.13) passes into the set $\widehat{M}{}_{C_1}^{C_2}$ (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) defined by constraints (4.8.14).

4.8.5. Solution on the set of bounded-variation functions

This set presents a *fourth example of compacts.* Let the sought exact solution $\bar{y}$ belong to the set of, generally speaking, fluctuating functions whose total variation is bounded from above by some constant C. We designate this set as V_C (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995).

In the case of discretized problem, the variation is the sum (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) $|y_2-y_1|+|y_3-y_2|+\cdots+|y_n-y_{n-1}|$. That is why, to solve the problem, relations (4.8.5)–(4.8.7) and (4.8.9) are to be used, supplemented with a *constraint* in the form of inequality

$$|y_2 - y_1| + |y_3 - y_2| + \cdots + |y_n - y_{n-1}| \leq C. \tag{4.8.15}$$

4.8.6. Numerical examples

To realize the above variants of the solution on the compact, one can use the Fortran computer programs PTIGR, PTIGR1, PTIGR2, etc. (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995). With these programs, several examples were solved.

We solved the first-kind Fredholm integral equation (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995)

$$\int_a^b K(x,s)y(s)\,\mathrm{d}s = f(x), \qquad c \leq x \leq d, \tag{4.8.16}$$

where $a = 0$, $b = 1$, $c = -1$, $d = 2$, and the kernel is

$$K(x,s) = \frac{1}{1+100(x-s)^2}. \tag{4.8.17}$$

Initially, the exact solution was set in the form (*first example*)

$$\bar{y}(s) = 1 - s^2, \qquad a \le s \le b. \tag{4.8.18}$$

Figure 4.7 shows the exact functions $y(s)$, $K(x,0)$ and $f(x)$ from the first example. A characteristic feature of the first example is that the solution $y(s)$ is a monotonic bounded convex function if $s \in [a,b]$, i. e., the solution, in principle, can be sought on the sets $M{\downarrow}_{C_1}^{C_2}$, $\overset{\frown}{M}{\downarrow}_{C_1}^{C_2}$ or $\overset{\frown}{M}{}_{C_1}^{C_2}$ depending on available information.

The problem was discretized as follows: the total number of nodal points along x was set equal to $m = 61$ (the step was $\Delta x = (d-c)/(m-1) = 0.05$), and the total number of nodal points along s was taken equal to $n = 41$ (the step along s was $\Delta s = (b-c)/(n-1) = 0.025$).

With the help of the RNDAN random-number generator (Sizikov, 2001, p. 152), errors uniformly distributed over the interval $[-\delta_0, \delta_0]$ were added to the right-hand side $f(x)$, where $\delta_0 = 0.01 \max_{c \le x \le d} f(x) = 0.2325 \cdot 10^{-2}$. This yields (Sizikov, 2001, p. 148) $\delta^2 = (d-c)\delta_0^2/3 = 7.21 \cdot 10^{-6}$.

Figure 4.8 shows the exact solution $\bar{y}(s)$ and the approximate solution $\tilde{y}(s)$ found, with the use of (4.8.9), on the set $M{\downarrow}_{C_1}^{C_2}$ of monotonically non-increasing bounded functions (cp. (4.8.10)). Here, $C_1 = 0$ and $C_2 = 1$. The initial approximation was set as $y_0(s) = 0$. A total of 800 iterations were performed to reach the value of the discrepancy functional $\Phi(y)$ equal to $7.30 \cdot 10^{-6}$, i. e., practically, δ^2.

Figures 4.9 and 4.10 show the solution found on the set $\overset{\frown}{M}{\downarrow}_{C_1}^{C_2}$ of monotonically non-increasing bounded convex functions and the solution found on the set $\overset{\frown}{M}{}_{C_1}^{C_2}$ of bounded convex functions. Next, the same example (4.8.16)–(4.8.18) was solved with a greater error: $\delta_0 = 0.03 \max\limits_{c \le x \le d} f(x) = 0.6975 \cdot 10^{-2}$, $\delta^2 = 6.49 \cdot 10^{-5}$. In either case, a total of 800 iterations were performed. The zero initial approximation was used: $y_0(s) = 0$.

Figures 4.8–4.10 demonstrate that, first, the use of additional conditions in the form of some constraints imposed on the solution so that to place it into a compact (if such constraints follow from the essence of the problem) makes the solution stable. Second, a comparison of results shows that the solution reconstructing accuracy depends on the particular type of constraints imposed on the solution ((4.8.8), (4.8.12), (4.8.14), etc.).

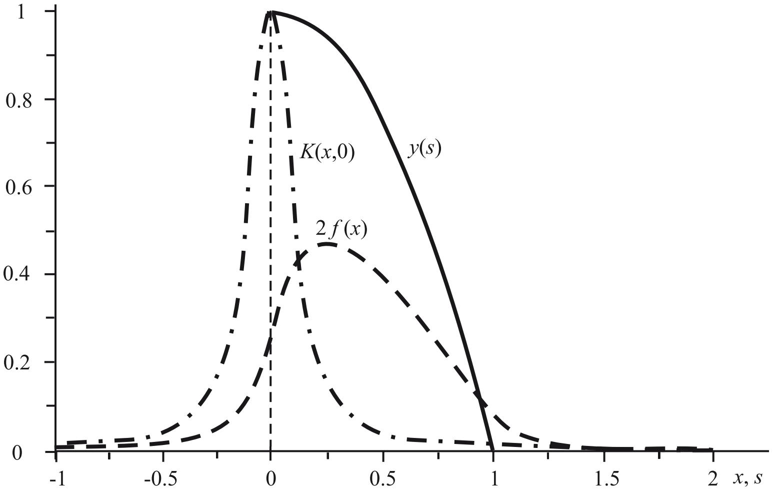

Figure 4.7. First example: exact functions $y(s)$, $K(x,0)$, and $f(x)$

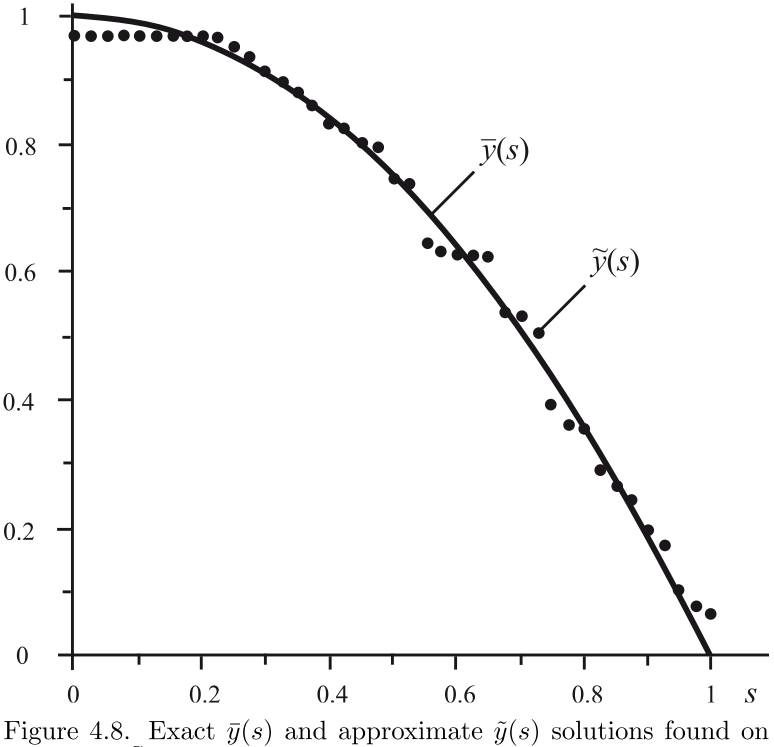

Figure 4.8. Exact $\bar{y}(s)$ and approximate $\tilde{y}(s)$ solutions found on the set $M{\downarrow}_{C_1}^{C_2}$ of monotonically non-increasing bounded functions

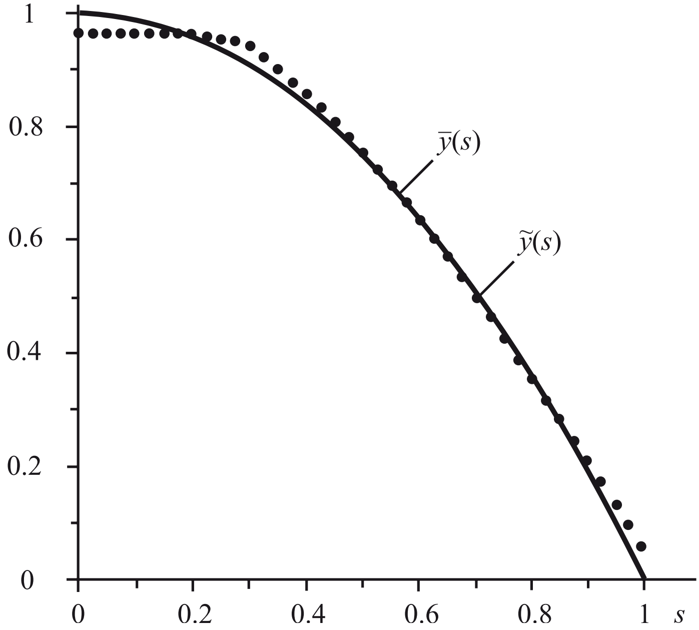

Figure 4.9. Exact $\bar{y}(s)$ and approximate $\tilde{y}(s)$ solutions found on the set $\widehat{M}{\downarrow}_{C_1}^{C_2}$ of monotonically non-increasing bounded convex functions

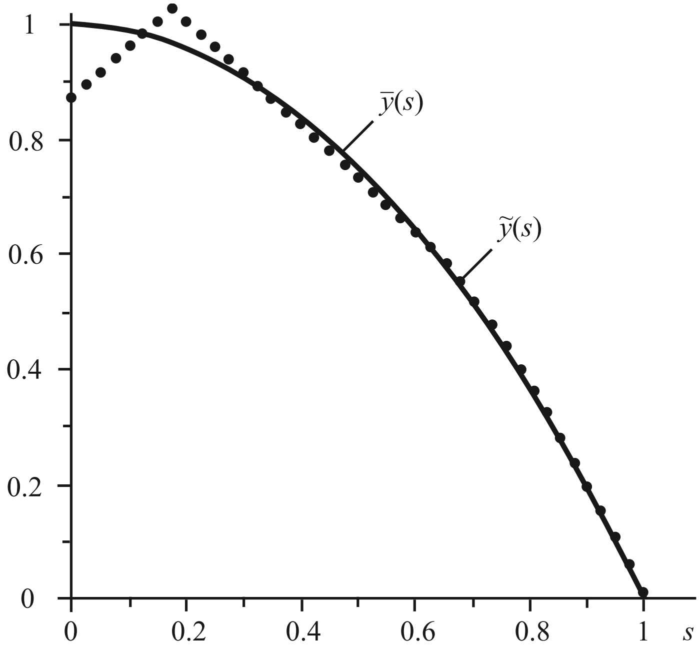

Figure 4.10. Exact solution $\bar{y}(s)$ and approximate solution $\tilde{y}(s)$ found on the set $\widehat{M}_{C_1}^{C_2}$ of bounded convex functions

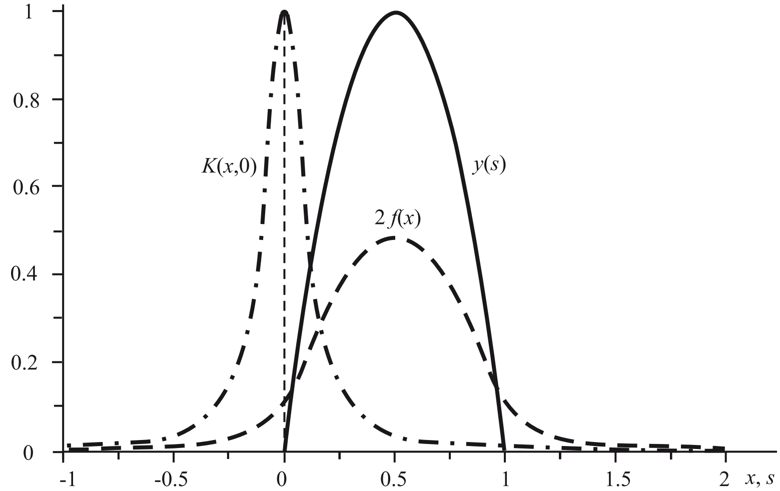

Figure 4.11. Second example. Exact solution $y(s)$, right-hand side $f(x)$, and kernel $K(x,0)$

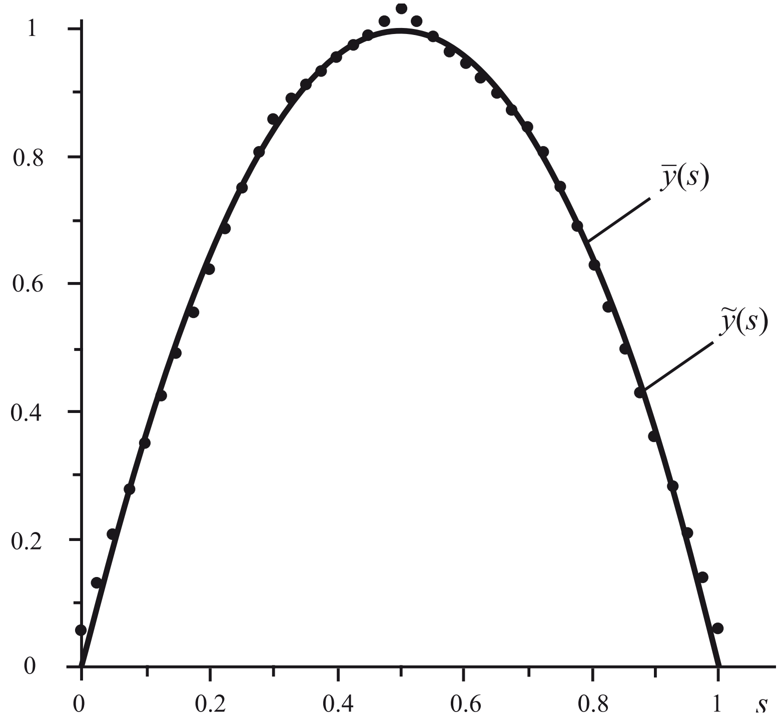

Figure 4.12. Exact solution $\bar{y}(s)$ and approximate solution $\tilde{y}(s)$ found on the set $\widehat{M}_{C_1}^{C_2}$ of bounded convex functions

We also solved a *second example*, in which the exact solution was $\bar{y}(s) = 4s(1-s)$, $s \in [0,1]$ (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995). Figure 4.11 shows the exact solution, the right-hand side, and the kernel from the second example. In this example, $\delta_0 = 0.01 \max f(x) = 0.2457 \cdot 10^{-2}$ and $\delta^2 = 8.05 \cdot 10^{-6}$. A characteristic feature here is that the desired solution $y(s)$ for $s \in [0,1]$ is a bounded convex function.

Figure 4.12 shows the exact solution and the approximate solution from the second example found on the set $\overset{\frown}{M}{}_{C_1}^{C_2}$ of bounded convex functions. A total of 800 iterations were performed to reach the final value of $\Phi(y)$ equal to $9.99 \cdot 10^{-6}$.

Solutions obtained by solving a numerical example on the set of functions with bounded variation V_C are given by Tikhonov, Goncharsky, Stepanov, and Yagola (1995).

These examples show that the use of additional information about the solution allows one to make finding the solution a well-posed problem.

Chapter 5.

Inverse problems in image reconstruction and tomography

In this chapter, we will consider some applied ill-posed inverse problems. Statements of the problems are given, and their mathematical descriptions are presented. Stable solutions of the problems are obtained by the Tikhonov regularization method. Numerical examples are discussed.

5.1. RECONSTRUCTION OF BLURRED IMAGES

In Sections 5.1 and 5.2, we will consider an inverse problem that arises in optics, namely, reconstruction (restoration) of distorted images, namely, reconstruction of blurred (smeared) or defocused images (Bakushinsky and Goncharsky, 1994; Bates and McDonnell, 1986; Sizikov, 2001).

Under the term *image*, we mean a photograph of a person, a text, or a natural object (including photos taken from the outer space); a TV- or cine-picture; a telescopic image of a cosmic object; etc. Yet, for definiteness, in all cases under this term we will understand a *photograph.*

We assume that the image of interest was already given a preliminary treatment so that to remove scratches from it, to adjust brightness and contrast range, etc.; in other words, all non-mathematical operations with the image are assumed accomplished. Here, we consider the most complex problem, that is, mathematical reconstruction of blurred images whose distortion was a result of camera or object motion (shift, displacement).

5.1.1. Statement of the problem

Consider the problem with the example of a blurred (or shift- or displacement-distorted) photograph (Bakushinsky and Goncharsky, 1994; Bates and McDonnell, 1986; Sizikov, 2001; Sizikov and Belov, 2000; Tikhonov, Goncharskii, and Stepanov, 1987). Let an object of interest (assumed to be flat due to its remoteness) and the photographic film in the camera be both parallel to the aperture of the thin lens of the camera. The object and the film are located on different sides of the lens at distances f_1 and f_2 from it. Then,

$$\frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{f}, \tag{5.1.1}$$

where f is the focal distance of the lens and $f_1 \geq f$ (see Figure 5.1). As a result, on the photographic film we will obtain an inverted image of the object.

We introduce a rectangular coordinate system $\xi' O' \eta'$ in the plane of the object and a rectangular coordinate system $\xi O \eta$ in the plane of the film. We take some point $A'(\xi', \eta')$ on the object emitting intensity $w'(\xi', \eta')$. The rays emanating out of this point and having passed through the lens intersect each other at a certain point $A(\xi, \eta)$ in the plane of the film. It follows from similarity of the triangles $A'CO'$ and ACO that

$$\frac{\overrightarrow{O'A'}}{f_1} = \frac{\overrightarrow{OA}}{f_2}.$$

In projections, we have:

$$\frac{\xi'}{f_1} = -\frac{\xi}{f_2}, \qquad \frac{\eta'}{f_1} = -\frac{\eta}{f_2}. \tag{5.1.2}$$

As a result, the point $A(\xi, \eta)$ on the film (this point is the image of the point $A'(\xi', \eta')$ on the object) has the same brightness w and coordinates ξ and η (*direct problem*):

$$w(\xi, \eta) = w'(\xi', \eta'), \qquad \xi = -\frac{\xi'}{q}, \qquad \eta = -\frac{\eta'}{q}, \tag{5.1.3}$$

where $q = f_1/f_2$. The distance f_2 can be found as

$$f_2 = \Big(\frac{1}{f} - \frac{1}{f_1}\Big)^{-1}. \tag{5.1.4}$$

Thus, to each point A' on the object, a certain point A on the film corresponds, with the same intensity $w(\xi, \eta) = w'(\xi', \eta')$, but with inverted coordinates additionally reduced by the factor $q = f_1/f_2$ (see (5.1.3)).

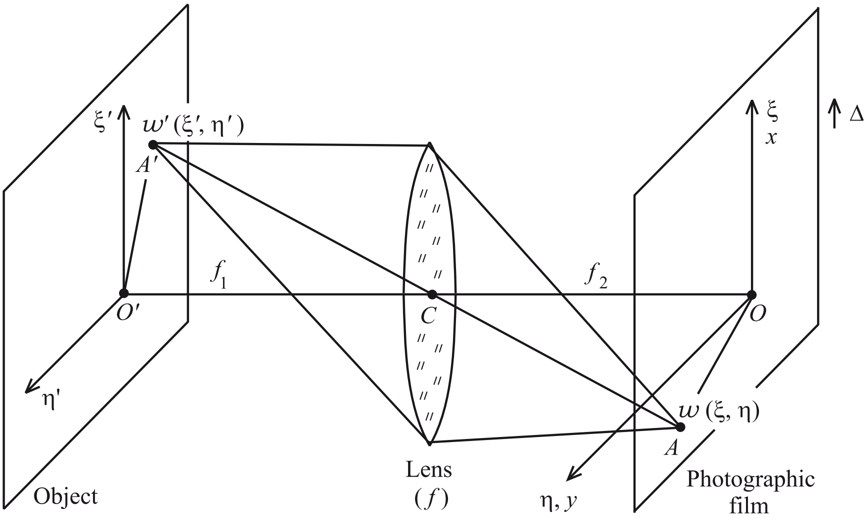

Figure 5.1. Diagram illustrating the manner in which a blurred image can be obtained

Example. Let $f_1 = 7$ m and $f = 5$ cm, then $f_2 = 5.03$ cm (according to (5.1.4)) and $q = 139.2$, i. e., the inverted image has 139.2 times smaller dimensions than the object.

From the photograph, one can reconstruct (restore) the object (*inverse problem*):

$$w'(\xi', \eta') = w(\xi, \eta), \qquad \xi' = -q\xi, \qquad \eta' = -q\eta. \tag{5.1.5}$$

In this case,

$$f_1 = \left(\frac{1}{f} - \frac{1}{f_2}\right)^{-1}. \tag{5.1.6}$$

Next, we assume that, during the exposure time τ, the camera has been given a rectilinear and uniform *shift* (displacement) with a velocity $v =$ const along the ξ axis, i. e., has been displaced by the distance $\Delta = v\tau$, or, alternatively, the object (a fast target) has traversed the distance $-q\Delta$. As a result, the image on the film will be blurred along ξ (see Figure 5.2a below).

5.1.2. Derivation of an integral equation

Describe the problem of image blurring mathematically. In addition to the stationary coordinate system $\xi O \eta$, we introduce a coordinate system xOy

attached to the moving film, which was initially (at $\tau = 0$) coincident with the system $\xi O \eta$ (see Figure 5.1). During the time τ, a continuous set of points A with abscissas from $\xi = x$ to $\xi = x + \Delta$ and various intensities $w(\xi, y)$ will be projected onto some point (x, y) of the film; in other words, the resultant brightness (denoted as g) of the point (x, y) can be found as the sum (or, more precisely, integral) of intensities $w(\xi, y)$, $\xi \in [x, x + \Delta]$ (for more detail, see Sizikov, 2001, p. 66):

$$g(x, y) = \frac{1}{\Delta} \int_x^{x+\Delta} w(\xi, y)\, \mathrm{d}\xi. \tag{5.1.7}$$

We write (5.1.7) in a different way:

$$\frac{1}{\Delta} \int_x^{x+\Delta} w(\xi, y)\, \mathrm{d}\xi = g(x, y). \tag{5.1.8}$$

Relation (5.1.8) is the master relation in the problem of reconstruction of blurred images. In this relation, $g(x, y)$ is the distribution of brightness over the rectangular coordinates x and y on the film (in the blurred image) (the x axis is directed along the blurring direction); Δ is the blurring distance, assumed known; and $w(\xi, y)$ is the distribution of the true (undistorted) intensity on the film (the intensity which would be registered by stationary film, i. e., in the case of $\Delta = 0$).

Relation (5.1.8) is one-dimensional Volterra equation of the first kind for the function $w(\xi, y)$ at each fixed value of y, representing a parameter; in other words, relation (5.1.8) represents a *set of one-dimensional integral equations.*

This equation was briefly considered in Section 4.4 (Example 4.4.9). According to the terminology of Apartsyn (2003), this equation is *non-classical Volterra equation of the first kind* (because both integration limits in it are variable quantities). Solution of this equation is a well-posed problem on the pair of spaces $(C, C^{(1)})$, i. e., when $w(\xi, y) \in C$ and $g(x, y) \in C^{(1)}$, and an ill-posed problem on the pair of spaces (C, C), when $w(\xi, y) \in C$ and $g(x, y) \in C$.

Note that in a number of works (Bakushinsky and Goncharsky, 1994; Tikhonov, Goncharskii, and Stepanov, 1987; and others) more complex statements of this problem were considered, including the cases of non-uniform and/or non-rectilinear shift of the camera or object, non-parallel object and film planes, etc.

5.1.3. Solving integral equation by the Fourier transformation method and by the Tikhonov regularization method

Equation (5.1.8) can be written as convolution type Fredholm integral equation of the first kind (Bakushinsky and Goncharsky, 1994; Sizikov, 2001; Sizikov and Belov, 2000):

$$\int_{-\infty}^{\infty} k(x-\xi)w(\xi,y)\,\mathrm{d}\xi = g(x,y), \qquad -\infty < x,y < \infty, \tag{5.1.9}$$

where

$$k(x) = \begin{cases} 1/\Delta, & x \in [-\Delta, 0], \\ 0, & x \notin [-\Delta, 0]. \end{cases} \tag{5.1.10}$$

Solving equation (5.1.9) is an ill-posed problem. To solve this equation, we use the Fourier transformation method and the Tikhonov regularization method (Bakushinsky and Goncharsky, 1994; Verlan' and Sizikov, 1986; Sizikov, 2001; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) (see Section 4.7). The *regularized solution* has the form (cp. (4.7.20)):

$$w_\alpha(\xi,y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_\alpha(\omega,y)\mathrm{e}^{-i\omega\xi}\,\mathrm{d}\omega, \tag{5.1.11}$$

where the regularized Fourier-spectrum of the solution is

$$W_\alpha(\omega,y) = \frac{K(-\omega)G(\omega,y)}{L(\omega)+\alpha M(\omega)}. \tag{5.1.12}$$

Here (cp. (4.7.21), (4.7.22))

$$K(\omega) = \int_{-\infty}^{\infty} k(x)\mathrm{e}^{i\omega x}\,\mathrm{d}x = \frac{\sin(\omega\Delta)}{\omega\Delta} + \frac{\cos(\omega\Delta)-1}{\omega\Delta}\,i, \tag{5.1.13}$$

and

$$G(\omega,y) = \int_{-\infty}^{\infty} g(x,y)\mathrm{e}^{i\omega x}\,\mathrm{d}x \tag{5.1.14}$$

are respectively the Fourier spectra of the kernel $k(x)$ and the right-hand side $g(x,y)$ of (5.1.9), the regularizer $M(\omega)$ can be chosen, for instance, in the form $M(\omega)=\omega^2$, $\alpha > 0$ is the regularization parameter, and

$$L(\omega) = |K(\omega)|^2 = K(\omega)K(-\omega) = \mathrm{Re}^2\,K(\omega) + \mathrm{Im}^2\,K(\omega).$$

As was stated in Section 4.7, there exist a number of ways by which the value of α can be chosen, for instance, the discrepancy principle or the generalized discrepancy principle (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995; Morozov, 1984). Yet, as tests showed, in the case of the image reconstruction problem the most efficient way is the *fitting way* (Sizikov, 2001, p. 71; Sizikov and Belov, 2000). In this way, the function $w_\alpha(\xi, y)$ is to be calculated for several values of α by formula (5.1.11) (with the use of (5.1.10), (5.1.12)–(5.1.14)). Then, graphs of $w_\alpha(\xi, y)$ should be inspected to choose the value of α ensuring the best reconstruction quality of the image from the standpoint of physiological (and not mathematical) perception criteria. This method is analogous to the TV contrast adjustment procedure (in the case under consideration, the parameter α is the reciprocal to the contrast).

Note that the blurring distance Δ is unknown *a priori*; it can be found by the fitting way or from the length of strokes in the distorted image. As for the blurring direction (along which the axis x is directed), this direction can be found from the direction of strokes in the distorted image (see Figures 5.2, 5.4, and 5.5 below).

Thus, with properly chosen direction of x on the distorted image (along the shift) and with properly chosen blurring distance Δ, one can reconstruct, in a stable way, the brightness $w_\alpha(x, y)$ on undistorted photograph from the brightness $g(x, y)$ on the distorted photograph by solving equation (5.1.9) (or, more precisely, set of equations) using formulas (5.1.10)–(5.1.14) with the value of α chosen, for instance, by the fitting way. In many cases the distorted image contains valuable yet directly non-decipherable information: a photograph of a very important person, a historic building, an important text, a terrestrial object whose photograph was taken from the outer space, a fast target, etc. This information can be extracted from the photo (with the help of a computer) only mathematically.

5.1.4. Numerical results

The *program package* IMAGE (under Windows 98 and higher) in language Visual C++ has been worked out for solving problem of reconstructing blurred (and also defocused, see Section 5.2) images by the Tikhonov regularization method and by the Fourier transform method according to formulas (5.1.9)–(5.1.14) with inspecting α (and also Δ) and outputing processing results onto a display. Using this program package, both the *direct problem* (modeling of intensity $g(x, y)$ on a distorted image according to (5.1.6)) and the *inverse problem* (reconstruction of the undistorted intensity $w_\alpha(\xi,\, y)$ on an image according to (5.1.11)) can be solved.

Calculation of one-dimensional inverse and direct Fourier transforms (see (5.1.11), (5.1.14)) is performed in the form of discrete Fourier transformation (DFT). The DFT is realized both as fast Fourier transform (FFT)

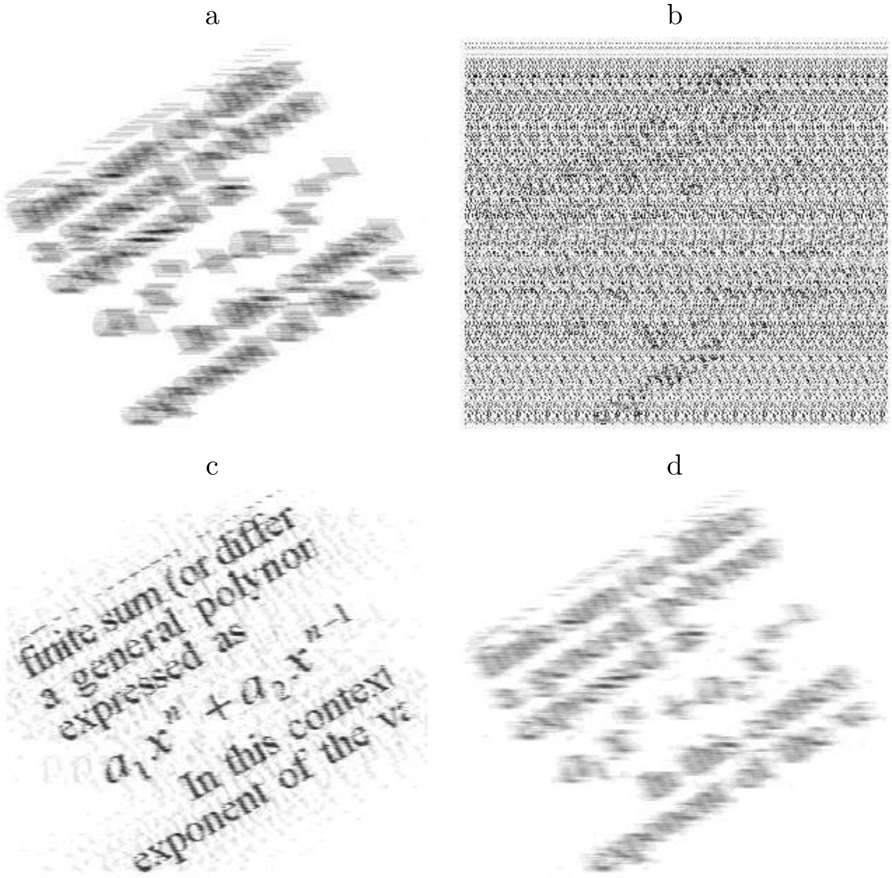

Figure 5.2. Smeared (a) ($\Delta = 20$) and reconstructed (b–d) gray images: $\alpha = 0$ (b), $2.5 \cdot 10^{-3}$ (c), 0.1 (d)

and ordinary Fourier transform. FFT is used if rapid image reconstruction is required, for instance, that of a shot of a fast object. Ordinary DFT can be used in the cases where there is no need in high processing rate, for instance, in reconstruction of texts, old photos, etc.

In processing black-and-white images, the *gray color* (a mixture of the red, green and blue colors taken in identical proportions) is normally used to widen the brightness range; in processing colored images, separate processing in the three colors with subsequent superposition of obtained images is used. Further details can be found in Sizikov and Belov (2000).

Figure 5.2a shows a gray text fragment blurred at an angle to the lines (blurring distance $\Delta = 20$ pixels). The blurring direction can be identified and the blurring distance Δ, estimated, by inspecting strokes in the photo. Figures 5.2b,c, and d show the regularized solution $w_\alpha(\xi, y)$ found with $\alpha = 0$ (i. e., without regularization), with the best-adjusted $\alpha = 2.5 \cdot 10^{-3}$, and with a certain estimated value of α, namely, with $\alpha = 0.1$, respectively. Here, a total of 256 discrete values of y were used, and 256 discrete readings of g along x at each fixed y were taken; when the total number of sampled values was lesser than 256, missing values were substituted with zeros. In other words, FFT was used, although ordinary DFT would have been more efficient in this case.

Figure 5.3 shows a reconstructed image of a blurred stationary object (building with a grid), and Figures 5.4 and 5.5 show reconstructed images of a fast air object (airplane) and a fast ground object (passenger car). All the three images are colored pictures; they were therefore treated separately in the three colors with subsequent superposition of obtained images.

Figures 5.3–5.5 present black-and-white reproductions of the colored images. In these examples, the total number of discrete readings along the axes x and ω was taken equal to 700 pixels. The total number of discrete readings along the y-axis was 500 pixels. In this case, to the left and to the right of the images 100 zero-pixels were added to suppress the Gibbs effect in the reconstructed images. As a result, the total number of readings along x was 900 pixels.

The results presented in Figures 5.2–5.5 are indicative of the following. The Fourier transformation method without regularization as applied to equation (5.1.9) yields very unstable a solution (Figure 5.2b). On the contrary, the Tikhonov regularization method with a properly chosen value of α yields rather good results: the text becomes readable (Figure 5.2c) and the images of the stationary object (Figure 5.3b) and both moving objects (Figures 5.4b and 5.5b) become distinctly seen to finest features (men, grid and clocks in Figure 5.3b, plane markings in Figure 5.4b, the passenger's face in Figure 5.5b), although still contain errors (noise in Figure 5.2c and the Gibbs effect on the background of Figure 5.4b). Experience gained in reconstruction of many images shows that the value of α can be conveniently and adequately chosen by the fitting way. As for determination of blurring direction and the value of Δ, in some cases (like in Figure 5.2a) this can be done by inspecting strokes, whereas in other cases (like in Figures 5.3a, 5.4a, and 5.5a), by inspecting the image blur at the left and right edges of the blurred image.

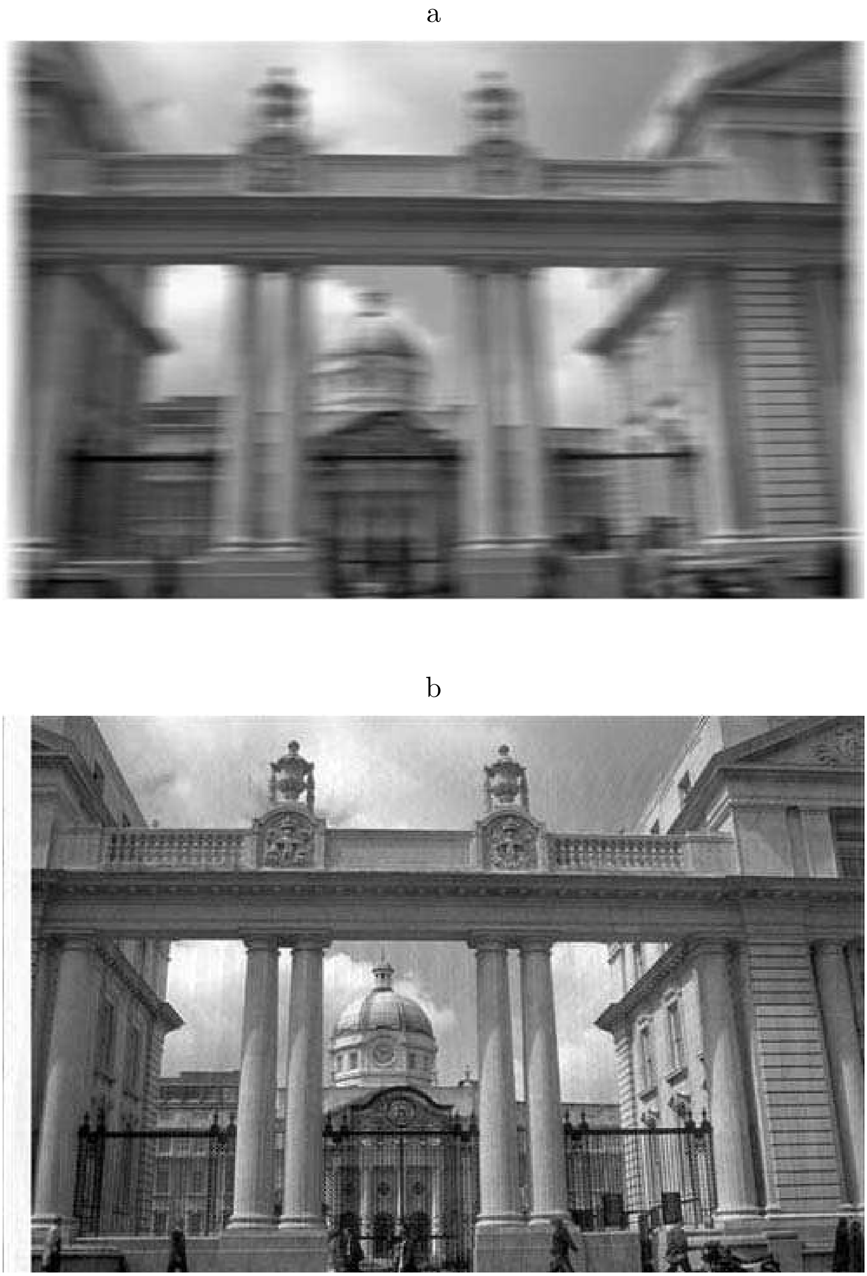

Figure 5.3. Smeared (a) and reconstructed (b) images of a stationary object

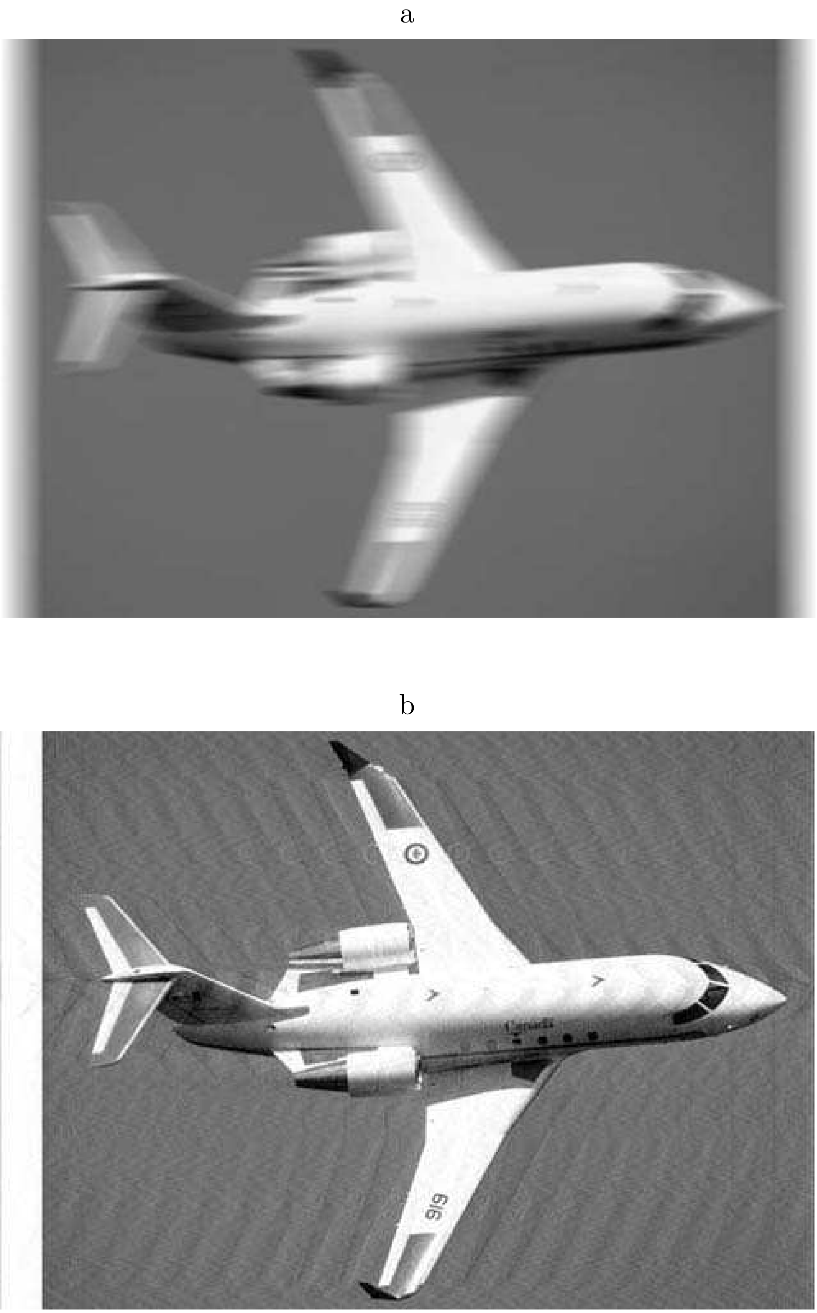

Figure 5.4. Smeared (a) and reconstructed (b) images of a fast-flying air object

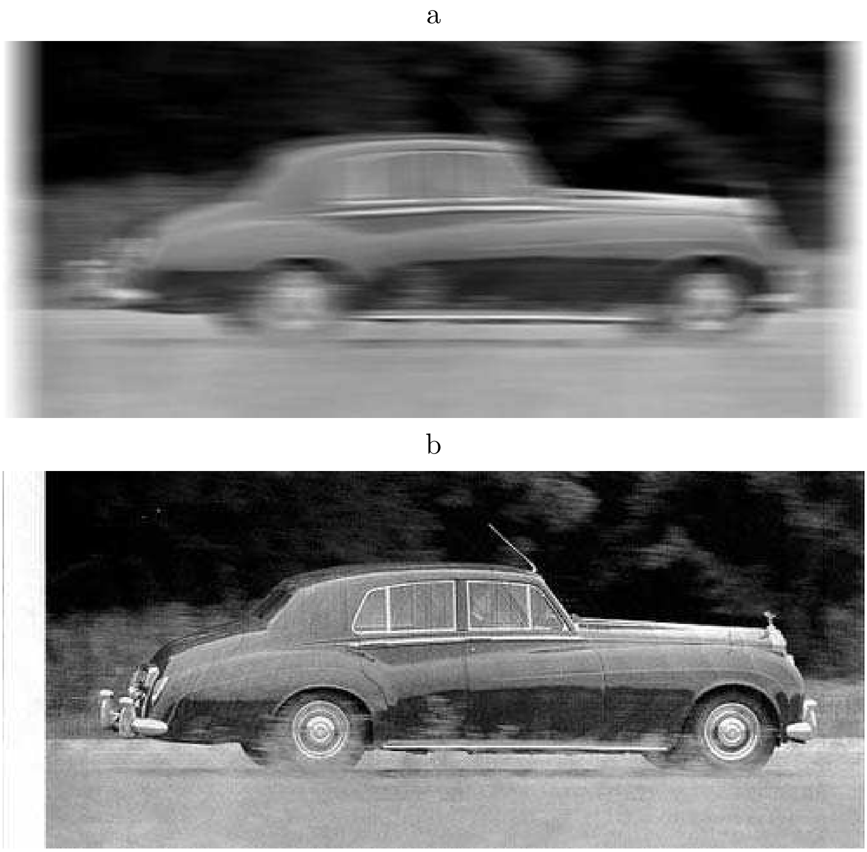

Figure 5.5. Smeared (a) and reconstructed (b) images of a fast-moving ground object

5.2. RECONSTRUCTION OF DEFOCUSED IMAGES

Consider another problem that arises in optics, namely, reconstruction (restoration) of defocused images (photos of a man, a text, a cosmic object, etc.) (Bakushinsky and Goncharsky, 1994; Bates and McDonnell, 1986; Sizikov, 2001; Vasilenko, 1979; Vasil'ev and Gurov, 1998; Sizikov and Belov, 2000; Tikhonov, Goncharskii, and Stepanov, 1987; Chaisson, 1992). This problem will be considered with the example of a defocused photograph.

5.2.1. Statement of the problem

This problem has much in common with the previous problem (reconstruction of blurred images); yet, there are substantial differences.

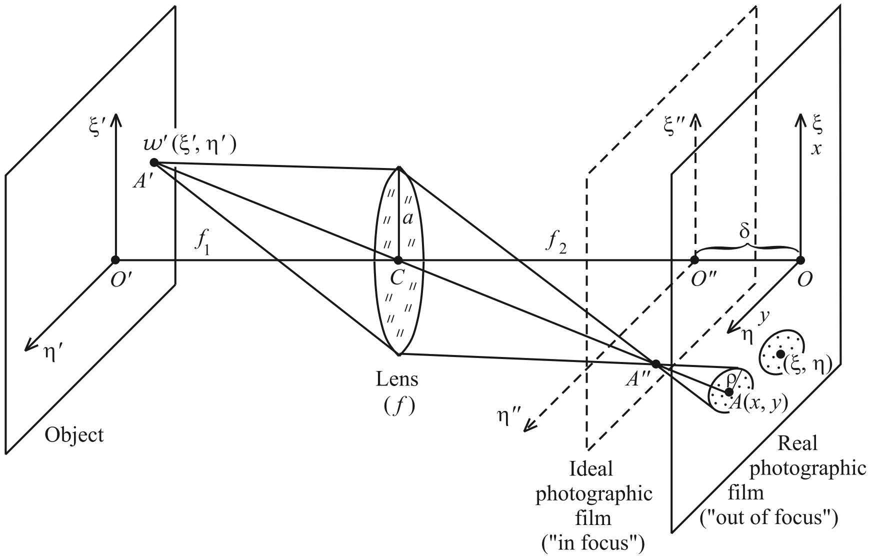

Figure 5.6. Schematic illustrating the geometry of the problem

Let the object of interest (assumed flat) and the photographic film be both parallel to the thin lens and lie on different sides of it at distances f_1 and $f_2 + \delta$, respectively, where δ is the image focusing error (see Figure 5.6).

Like in the previous problem, relation (5.1.1) holds here, where f is the lens focal distance.

We introduce a rectangular coordinate system $\xi' O' \eta'$ in the plane of the object, a coordinate system $\xi'' O'' \eta''$ on the "ideal" photographic film located at the focal point of the lens ($\delta = 0$), and a coordinate system $\xi O \eta$, and also a coincident system xOy, on the real photographic film displaced from the focal point ($\delta \neq 0$). We denote the intensity emitted from some point $A'(\xi', \eta')$ on the object as $w'(\xi', \eta')$. The point A' will also be mapped into the point A'' on the "ideal" photographic film; the latter point has the same brightness $w''(\xi'', \eta'') = w'(\xi', y')$ and the coordinates $\xi'' = -\xi'/q$ and $\eta'' = -\eta'/q$, where $q = f_1/f_2$ (cp. (5.1.2)), and the distance f_2 can be found by formula (5.1.3).

On the real photographic film, the point A' will be mapped not into a point, but in a diffraction circle of radius

$$\rho = a\delta/f_2 \tag{5.2.1}$$

centered at the point $A(x, y)$, where a is the aperture radius of the lens, and

$$x = -\frac{f_2 + \delta}{f_1}\,\xi', \qquad y = -\frac{f_2 + \delta}{f_1}\,\eta' \tag{5.2.2}$$

(cp. (5.1.2)).

5.2.2. Derivation of the master equation

Let us give mathematical formulation to the image-defocusing problem. In addition to the diffraction circle centered at the point $A(x, y)$, consider the circle centered at the point (ξ, η) (see Figire 5.6). The radii of these (and other) circles are identical, equal to ρ (cp. (5.2.1)), and the areas of such circles are $S = \pi\rho^2$. As a result, the intensity $w(\xi, \eta)$ emitted by the point (ξ, η) will be "blurred" over a circle, whose radius is ρ and area is $S = \pi\rho^2$, with density $w(\xi, \eta)/\pi\rho^2$, in the first approximation assumed uniform over the diffraction circle.

The intensity at the point $A(x, y)$ will be the result of summation (integration) of partial intensities over all circles covering the point $A(x, y)$. The condition of coverage of the point $A(x, y)$ by a circle of radius ρ centered at the point (ξ, η) is

$$\sqrt{(x - \xi)^2 + (y - \eta)^2} \le \rho. \tag{5.2.3}$$

As a result, the intensity at the point $A(x, y)$ will be

$$g(x, y) = \iint\limits_{\sqrt{(x-\xi)^2+(y-\eta)^2} \le \rho} \frac{w(\xi, \eta)}{\pi\rho^2}\,\mathrm{d}\xi\,\mathrm{d}\eta. \tag{5.2.4}$$

Relation (5.2.4) is the master relation in the problem of reconstruction of defocused images. We write this relation as the equation

$$\iint\limits_{\sqrt{(x-\xi)^2+(y-\eta)^2} \le \rho} \frac{w(\xi, \eta)}{\pi\rho^2}\,\mathrm{d}\xi\,\mathrm{d}\eta = g(x, y). \tag{5.2.5}$$

Equation (5.2.5) is the *master equation* in the problem of reconstruction of defocused images. In this equation, $g(x, y)$ is the measured function, and $w(\xi, \eta)$ is the unknown function.

5.2.3. Bringing the master equation to the standard form

Equation (5.2.5) is a first-kind integral equation for the function $w(\xi,\eta)$. Yet, this equation is written in non-standard form; the latter makes this equation hard to solve. Let us bring this equation to the standard form. We write (5.2.5) as (Sizikov, 2001, p. 75):

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(x-\xi, y-\eta)w(\xi,\eta)\,\mathrm{d}\xi\,\mathrm{d}\eta = g(x,\,y), \qquad -\infty < x,y < \infty, \tag{5.2.6}$$

where

$$k(x-\xi,y-\eta) = \begin{cases} 1/(\pi\rho^2), & \sqrt{(x-\xi)^2+(y-\eta)^2} \le \rho, \\ 0, & \sqrt{(x-\xi)^2+(y-\eta)^2} > \rho, \end{cases} \tag{5.2.7}$$

or

$$k(x,y) = \begin{cases} 1/(\pi\rho^2), & \sqrt{x^2+y^2} \le \rho, \\ 0, & \sqrt{x^2+y^2} > \rho. \end{cases} \tag{5.2.8}$$

Relation (5.2.6) is the two-dimensional convolution-type Fredholm integral equation of the first kind. In this equation, $g(x,y)$ is the brightness on the defocused photograph, $w(\xi,y)$ is the unknown brightness on the undistorted photograph (on the photo taken with $\delta = 0$), and $k(x,y)$ is the kernel of the integral equation, called the *point scattering function* (in the present case, the scattering function of the point (x,y)).

The radius ρ can be found using formula (5.2.1), where a and f_2 are known quantities, and the value of δ can be obtained by the fitting way. Yet, it is more rational to adjust the radius ρ itself without using formula (5.2.1). Figure 5.7a exemplifies a defocused image (Sizikov, 2001, p. 75; Sizikov and Belov, 2000); it is seen from this figure that the radius ρ can be estimated as the radius of diffraction rings on the defocused image.

Note that the problem on reconstruction of defocused images in the case of non-parallel object and film planes was considered by Tikhonov, Goncharskii, and Stepanov (1987).

After solving equation (5.2.6), one can reconstruct the initial image in the object plane (*inverse problem*, cp. (5.1.3)):

$$w'(\xi',\eta') = w(\xi,\eta), \qquad \xi' = -\frac{f_1}{f_2+\delta}\,\xi, \qquad \eta' = -\frac{f_1}{f_2+\delta}\,\eta.$$

5.2.4. Solution of the equation by the method of two-dimensional Fourier transformation and by the Tikhonov regularization method

Like the two-dimensional first-kind convolution type Fredholm integral equation, equation (5.2.6) can be solved by the two-dimensional Fourier transform (FT) method (the inverse filtration method), analogously to one-dimensional equation (4.5.5) or (5.1.9). Yet, solution of (5.2.6), and also solution of (5.1.9), presents an ill-posed problem (Bakushinsky and Goncharsky, 1994; Verlan' and Sizikov, 1986; Vasilenko, 1979; Voskoboynikov, Preobrazhenski, and Sedel'nikov, 1984). The latter is related with the fact that the function $g(x,y)$ is measured with errors, which fact leads to arbitrarily large errors in the solution $w(\xi,y)$.

To solve the problem in a stable manner, we will use, in addition to two-dimensional FT, the Tikhonov regularization method. Then, the solution of (5.2.6) obtained by these methods can be written as (cp. (4.7.26)–(4.7.30))

$$w_\alpha(\xi,\eta) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_\alpha(\omega_1,\omega_2)\mathrm{e}^{-i(\omega_1\xi+\omega_2\eta)}\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2, \tag{5.2.9}$$

where

$$W_\alpha(\omega_1,,\omega_2) = \frac{K(-\omega_1,-\omega_2)G(\omega_1,\omega_2)}{L(\omega_1,\omega_2)+\alpha M(\omega_1,\omega_2)}. \tag{5.2.10}$$

In (5.2.10),

$$K(\omega_1,\omega_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(x,\ y)\mathrm{e}^{i(\omega_1 x+\omega_2 y)}\,\mathrm{d}x\,\mathrm{d}y, \tag{5.2.11}$$

$$G(\omega_1,\omega_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\mathrm{e}^{i(\omega_1 x+\omega_2 y)}\,\mathrm{d}x\,\mathrm{d}y, \tag{5.2.12}$$

$$L(\omega_1,\omega_2) = |K(\omega_1,\omega_2)|^2 = K(\omega_1,\omega_2)K(-\omega_1,-\omega_2), \tag{5.2.13}$$

$$M(\omega_1,\omega_2) = (\omega_1^2+\omega_2^2)^{2p}, \tag{5.2.14}$$

$\alpha > 0$ is the regularization parameter, and $p = 1,2,3,\dots$ is the regularization order.

Normally, the value of α can be chosen by the generalized discrepancy principle (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995). Yet, for the problem of reconstruction of defocused images (like for the problem of reconstruction of blurred images, see Section 5.1), the *fitting way* is more efficient (Sizikov, 2001, p. 77).

5.2.5. Numerical results

We developed the IMAGE program package written in the Visual C++ language and intended for reconstruction of blurred (see Section 5.1) and defocused images. In the problem of reconstruction of defocused images, both the direct problem (modeling of defocused images) and the inverse problem (reconstruction of initial images from defocused images) can be solved.

In the inverse problem, equation (5.2.6) is solved by the method of two-dimensional FT and by the Tikhonov regularization method according to (5.2.9)–(5.2.14). The two-dimensional continuous FTs (5.2.9) and (5.2.12) are calculated as two-dimensional discrete FTs, which in turn are realized, by the FTFTC program (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995), as fast FTs. The value of ρ (diffraction-circle radius) is set as estimated from the defocused image; if necessary, this radius can be subsequently refined in an adjustment procedure. The regularization parameter α can be chosen by the fitting way with subsequent graphical display of solutions $w_\alpha(\xi,\eta)$.

Here, like in the problem of reconstruction of blurred images, to obtain a wider brightness range in processing black-and-white images, the gray color (a mixture of the red, green, and blue colors taken in equal proportions) is used, and in processing of colored images, separate treatment in the three colors is used with subsequent superposition of the three images.

Figure 5.7 presents solution results for a model example. Figure 5.7a shows a defocused image of a gray text fragment (direct problem). A total of 256×256 discrete readings were taken. Figure 5.7b shows the reconstructed image (inverse problem) obtained with $\alpha = 0$, i. e., without regularization; this image presents an unstable solution. Figure 5.7c shows the reconstructed image obtained with $\alpha = 5 \cdot 10^{-4}$, the best value of α chosen by inspection of displayed images. Figure 5.7d shows the reconstructed image obtained with $\alpha = 5 \cdot 10^{-2}$, an overestimated value of α. In all these cases, $p = 1$. This way for choosing α is analogous to the TV-picture contrast adjustment procedure based on physiological rather than mathematical criteria.

Figure 5.8 shows results obtained in reconstruction of a defocused colored image.

Thus, with the help of stable relations (5.2.9)–(5.2.14) and with properly chosen values of ρ and α, one can reconstruct the undistorted image with intensity $w_\alpha(\xi,\eta)$ from a defocused image with intensity $g(x,y)$.

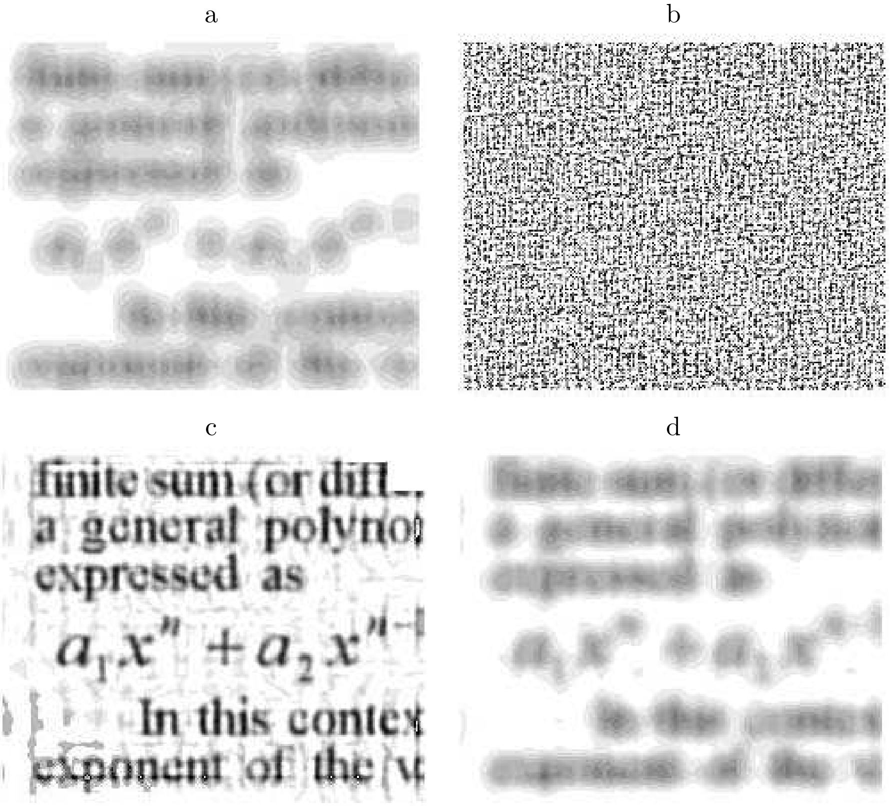

Figure 5.7. Defocused ($\rho = 10$ pix) (a) and reconstructed (b–d) text fragment in the gray color:
$\alpha = 0$ (b), $5 \cdot 10^{-4}$ (c), $5 \cdot 10^{-2}$ (d)

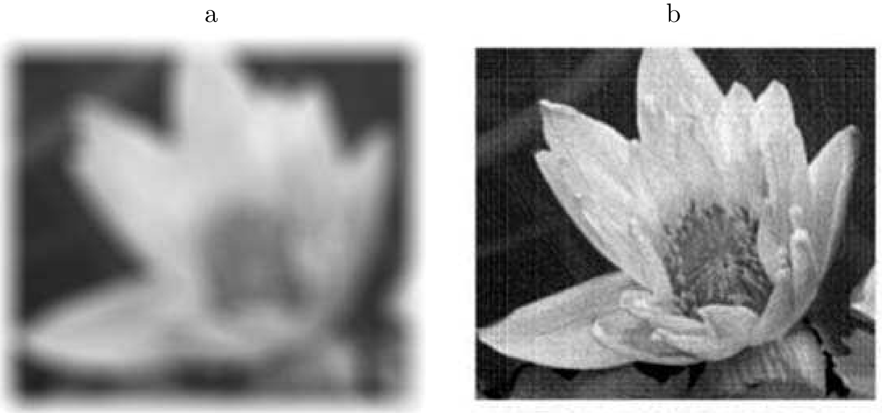

Figure 5.8. Defocused (a) and reconstructed (b) color image

Examples of images of extraterrestial objects (planets, nebuli, galaxies, etc.) reconstructed from defocused images of these objects obtained on the Hubble space telescope (HST) are given by Chaisson (1992). The defocusing was caused by an error of curvature of the HST mirror when its making. The computer-aided mathematical treatment of the photos made it possible to reconstruct the undistorted images of the objects, i. e., to compensate the metering error. Yet, in this case, in addition to the two-dimensional Fourier transformation method, instead of some modification of the Tikhonov regularization method, a method like the Fourier-spectrum truncation procedure and the method of smoothing windows were used (Verlan' and Sizikov, 1986, pp. 259, 260).

5.3. X-RAY TOMOGRAPHY PROBLEMS

In this section, we will consider the mathematical aspect of x-ray computer tomography (CT) (Sizikov, 2001, pp. 17–33; Tikhonov, Arsenin, and Timonov, 1987; Webb, 1988, V. 1; Natterer, 1986; Tikhonov, Arsenin, Rubashov, and Timonov, 1984; Troitskii, 1989), namely, some problems in it related with an ill-posed integral equation. Note that there exist nuclear magnetic resonance (NMR-) tomography (Sizikov, 2001, pp. 33–62; Webb, 1988, V. 2; Afanas'ev, Studentsov, Khorev, et al, 1979; Galaydin, Ivanov, and Marusina, 2004; Montgomery, 1969; Brunner and Sizikov, 1998); ultrasonic tomography (Webb, 1988, V. 2); positron-emission tomography, etc.

The word "tomography" takes its origin from the Greek words $\tau o \mu \eta$ (cross-section) and $\gamma \rho \alpha \varphi \omega$ (to write), i.e., this word means "to write in cross-sections". The *essence* of all types of tomographies is the same: from integral characteristics taken from some cross-section of a body, it is required to extract local characteristics, namely, the distribution of substance density (more exactly, of x-ray absorption factor) $c(x, y)$ over this cross-section, where x and y are the coordinates in this cross-section; and, then, with densities $c_z(x, y)$ known in some set of cross-sections, where z is the coordinate normal to the cross-sections, to find (reconstruct) the volume density $c(x, y, z)$. Different tomographies deal with entirely different types of integral data: CT deals with the intensity measured by radiation detectors; NMR-tomography deals with echo-signals; etc. The mathematical descriptions of tomographic problems also differ: in CT, the problem is described by the Radon (or Fredholm) integral equation; in NMR-tomography, by two-dimensional Fourier transformation; etc. Nonetheless, in all cases one and the same *final goal* is pursued: it is required to obtain cross-sectional

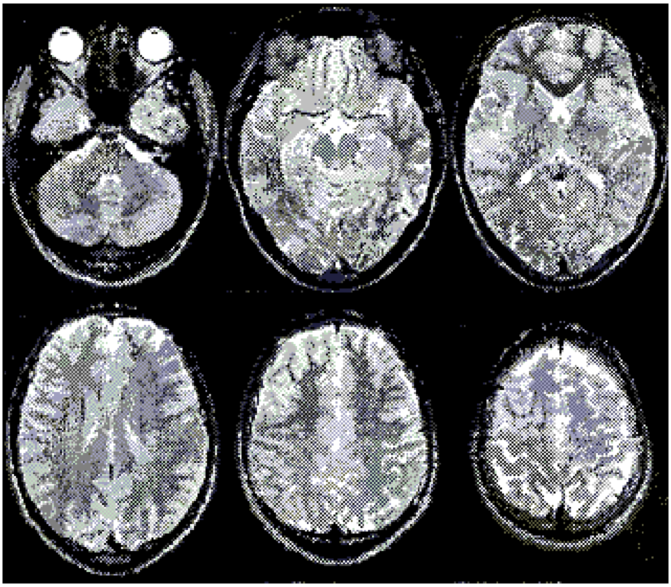

Figure 5.9. Tomograms of six horizontal sections of human brain

distributions of the density $c(x, y)$. That is why, for instance, x-ray and NMR tomograms (representations of $c(x, y)$ in transparent films) closely resemble each other, although these tomograms have entirely different physics and mathematics behind them and require different instrumentation to be obtained. Figure 5.9 shows typical NMR tomograms.

5.3.1. Experimental procedure

Figure 5.10 shows a parallel scan scheme in x-ray tomograph intermediate between tomographs of the first and second generation (by now, five generations of x-ray tomographs are known to exist (Sizikov, 2001, pp. 24–26; Webb, 1988, V. 1)).

On a frame, an array of x-ray tubes-sources is mounted; each of these x-ray tubes emits a well-collimated x-beam. All the beams are parallel to each other. The beams pass a cross-section of the object of interest (brain, for instance), and radiation detectors measure their attenuated intensities. Subsequently, the frame as a whole, with the sources and the detectors installed on it, is rotated through an angle θ around the object to repeat the measurements.

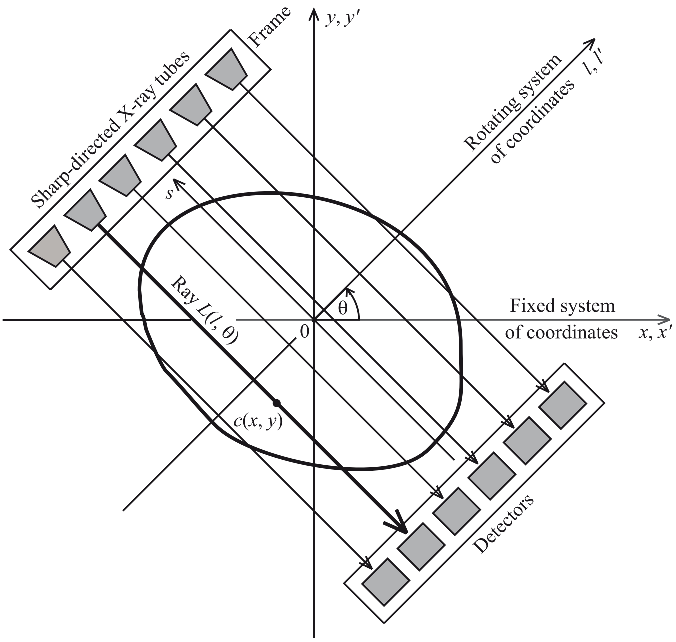
Frame
y, y'
Sharp-directed X-ray tubes
Rotating system of coordinates
l, l'
s
Ray $L(l, \theta)$
θ
0
Fixed system of coordinates
x, x'
$c(x, y)$
Detectors

Figure 5.10. Diagram illustrating x-ray tomographic examination of a body (parallel scan scheme)

5.3.2. Beer law

According to the Bouguer–Lambert–Beer law (Sizikov, 2001, p. 18; Webb, 1988, V. 1), the intensity of an x-beam incident on the detector is

$$I(l,\theta) = I_0 \exp\Big[-\int_{L(l,\theta)} c(x,y)\,\mathrm{d}s\Big], \tag{5.3.1}$$

where l is the coordinate of the detector, θ is the frame angle, I_0 is the intensity emitted by the tube-source, $c(x,y)$ is the substance density (more exactly, the x-ray absorption factor) on the path of the ray $L(l,\theta)$, which represents a straight line whose equation is

$$x\cos\theta + y\sin\theta = l. \tag{5.3.2}$$

The integration in (5.3.1) is to be performed along the ray $L(l,\theta)$. The integral $\int_{L(l,\theta)} c(x,y)\,\mathrm{d}s$ is the *mass of substance in the ray.* The greater the integral, the lesser is the transmitted intensity $I(l,\theta)$.

5.3.3. Radon transformation and equation

An alternative form of (5.3.1) is

$$I(l,\theta)/I_0 = \exp\left[-q(l,\theta)\right], \tag{5.3.3}$$

where

$$q(l,\theta) = \int_{L(l,\theta)} c(x,y)\,\mathrm{d}s. \tag{5.3.4}$$

Definition. Expression (5.3.4), in which $L(l,\theta)$ is some beam passing through the substance, $c(x,y)$ is the density of the substance in the beam, and the coordinate s is directed along the beam, is called the *Radon transform* (1917).

We find the logarithm of (5.3.3) and obtain:

$$q(l,\theta) = -\ln\left[I(l,\theta)/I_0\right]. \tag{5.3.5}$$

Definition. The function $q(l,\theta)$ is called *absorption* or *shadow*.

The absorption $q(l,\theta)$ assumes values ranging from 0, when $I(l,\theta) = I_0$, i. e., the medium is transparent, to ∞, when $I(l,\theta) = 0$, i. e., the medium is opaque.

Definition. The ratio $I(l,\theta)/I_0$ is called *transparency*.

The transparency $I(l,\theta)/I_0$ assumes values ranging from 0 (the medium is opaque) to 1 (the medium is transparent).

We may write (5.3.4) as

$$\int_{L(l,\theta)} c(x,y)\,\mathrm{d}s = q(l,\theta). \tag{5.3.6}$$

Relation (5.3.6) is called the *Radon equation* (or *shadow equation*). This equation is the master equation in CT. Here, the two-dimensional function $q(l,\theta)$ is defined by (5.3.5), where $I(l,\theta)$ is a measured quantity, and the two-dimensional function $c(x,y)$ is the function to be found. In view of this, relation (5.3.6) can be considered as an *integral equation* whose solution makes it possible to determine the function $c(x,y)$ from the measured right-hand side $q(l,\theta)$. In principle, by solving (5.3.6), one can find the density distribution $c(x,y)$ in some cross-section of the object (a human brain, for instance) from the measured intensity $I(l,\theta)$ and, hence, the absorption $q(l,\theta)$. Such a problem is called *reconstruction of an x-ray image.*

For the first time, this problem was considered by Radon in 1917.

One of the solutions of (5.3.6) (solution by means of *convolution and back projection method* with use of Fourier transformation) has the form (Bracewell and Riddle, 1967; Ramachandran and Lakshminarayanan, 1971; Webb, 1988, V. 1; Troitskii, 1989, p. 33)

$$c(x,y)=\frac{1}{\pi}\int_0^{\pi}\mathrm{d}\theta\int_{-\infty}^{\infty}q(l,\theta)p(x\cos\theta+y\sin\theta-l)\,\mathrm{d}l, \tag{5.3.7}$$

where

$$p(t)=\pi\int_{-\infty}^{\infty}|\omega|\mathrm{e}^{i\omega t}\,\mathrm{d}\omega \tag{5.3.8}$$

is the so-called *impulse response of a filter* with the frequency characteristic $\pi|\omega|$. Let us briefly examine solution (5.3.7)–(5.3.8). We use the Euler formula $\mathrm{e}^{i\omega t}=\cos\omega t+i\sin\omega t$ and take the fact into account that $\int_{-\infty}^{\infty}|\omega|\cos\omega t\,\mathrm{d}\omega=2\int_0^{\infty}\omega\cos\omega t\,\mathrm{d}\omega$ and $\int_{-\infty}^{\infty}|\omega|\sin\omega t\,\mathrm{d}\omega=0$ (here, respectively integrals of an even and odd function are considered). Then, we obtain:

$$p(t)=\pi\int_{-\infty}^{\infty}|\omega|\cos\omega t\,\mathrm{d}\omega+i\pi\int_{-\infty}^{\infty}|\omega|\sin\omega t\,\mathrm{d}\omega=2\pi\int_0^{\infty}\omega\cos\omega t\,\mathrm{d}\omega.$$

Whatever $t\in(-\infty,\infty)$, the latter integral never converges; as a consequence, whatever $x,y\in(-\infty,\infty)$, expression (5.3.7) also never converges. In other words, solution (5.3.7)–(5.3.8) is a solution that diverges even if the function $q(l,\theta)$ is known exactly. A converging and stable variant of this solution will be presented below.

Another solution of (5.3.6) is (Tikhonov, Arsenin, and Timonov, 1987, p. 40)

$$c(x,y)=-\frac{1}{2\pi^2}\int_0^{\pi}\mathrm{d}\theta\int_{-\infty}^{\infty}\frac{\partial q(l,\theta)}{\partial l}\,\frac{\mathrm{d}l}{l-(x\cos\theta+y\sin\theta)}. \tag{5.3.9}$$

Solution (5.3.9) is an unstable solution; the latter fact is related with the necessity of numerical calculation of the derivative $\partial q(l,\theta)/\partial l$ of the function $q(l,\theta)$ that normally contains a noise-induced random component. In addition, the integral in (5.3.9) is singular since the denominator $l-(x\cos\theta+y\sin\theta)$ may turn into zero.

Thus, the problem on solution of equation (5.3.6) is an ill-posed problem.

5.3.4. List of stable methods for solving the Radon equation

One of the first stable procedures for solving the Radon equation was realized by G. Hounsfield in the first commercial tomographic system, the CT-1010 brain scanner of the firm EMI. Another method for solving this equation is the Shepp–Logan method (filter) (Natterer, 1986; Webb, 1988, V. 1). These and similar methods (in which stability is achieved by truncation of the Fourier spectrum, by using smoothing windows, etc.) are sometimes called "intuitive regularization". These methods, although possessing some stability, are not stable enough (see Figures 5.11 and 5.12 below). Better results can be obtained by the Tikhonov regularization method or by even more precise a method, the Arsenin local regularization method, realized in the first Soviet x-ray computer tomograph CPT-1000 (Tikhonov, Arsenin, Rubashov, and Timonov, 1984). Consider the use of Tikhonov regularization method for solving the problem on reconstruction of x-ray images.

5.3.5. Stable solution of the problem by the Tikhonov regularization method

First of all, dwell on solution (5.3.7). The regularized variant of this solution is (Troitskii, 1989, p. 33):

$$c_\alpha(x,y) = \frac{1}{\pi}\int_0^\pi \mathrm{d}\theta \int_{-\infty}^{\infty} \tilde{q}(l,\theta)p_\alpha(x\cos\theta + y\sin\theta - l)\,\mathrm{d}l, \tag{5.3.10}$$

where $\tilde{q}(l,\theta)$ is the function distorted with some errors and

$$p_\alpha(t) = \pi\int_{-\infty}^{\infty} |\omega| W_\alpha(|\omega|)\mathrm{e}^{i\omega t}\,\mathrm{d}\omega. \tag{5.3.11}$$

In (5.3.11) (cp. (4.7.19), (4.7.20))

$$W_\alpha(|\omega|) = \frac{H^2(|\omega|)}{H^2(|\omega|) + \alpha\omega^{2p}}, \qquad H(|\omega|) = \frac{1}{\pi|\omega|},$$

$\alpha > 0$ is the regularization parameter, and $p \geq 0$ is the regularization order. Even simpler a form of $W_\alpha(|\omega|)$ is (Sizikov and Shchekotin, 2004)

$$W_\alpha(|\omega|) = \frac{1}{1 + \alpha\omega^{2p}}. \tag{5.3.12}$$

Results of computer modelling according to (5.3.10)–(5.3.12) are given in Sizikov and Shchekotin (2004).

Consider now another variant of Tikhonov regularization. The Radon equation (5.3.6) can be brought to a more habitual type of equation, namely, to *two-dimensional convolution-type Fredholm integral equation of the first kind* (Tikhonov, Arsenin, and Timonov, 1987):

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{c(x',y')\,\mathrm{d}x'\,\mathrm{d}y'}{\sqrt{(x-x')^2+(y-y')^2}}=S(x,y),\qquad -\infty<x,y<\infty,\tag{5.3.13}$$

where

$$S(x,y)=\frac{1}{\pi}\int_0^{\pi}q(x\cos\theta+y\sin\theta,\theta)\,\mathrm{d}\theta.\tag{5.3.14}$$

Equation (5.3.13) has a standard form admitting a well-established solution algorithm (Verlan' and Sizikov, 1986; Tikhonov and Arsenin, 1977; Tikhonov, Goncharsky, Stepanov, and Yagola, 1995). In this equation, the kernel is

$$K(x,y)=\frac{1}{\sqrt{x^2+y^2}},\tag{5.3.15}$$

the function to be found is $c(x,y)$, and the right-hand side $S(x,y)$ can be calculated numerically from known function $q(l,\theta)$ by formula (5.3.14).

In Section 4.7, the Tikhonov regularization method for solving the two-dimensional first-kind convolution-type Fredholm integral equation was described. According to this method, the regularized solution of (5.3.13) has the form (cp. (4.7.26)–(4.7.30)):

$$c_\alpha(x,y)=\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{c}_\alpha(\omega_1,\omega_2)\mathrm{e}^{-i(\omega_1x+\omega_2y)}\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2,\tag{5.3.16}$$

where the regularized-solution spectrum is

$$\hat{c}_\alpha(\omega_1,\omega_2)=\frac{\hat{K}(-\omega_1,-\omega_2)\hat{S}(\omega_1,\omega_2)}{L(\omega_1,\omega_2)+\alpha M(\omega_1,\omega_2)}.\tag{5.3.17}$$

Here, the spectrum of the right-hand side,

$$\hat{S}(\omega_1,\omega_2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}S(x,y)\mathrm{e}^{i(\omega_1x+\omega_2y)}\,\mathrm{d}x\,\mathrm{d}y,\tag{5.3.18}$$

is to be calculated numerically, whereas the spectrum of the kernel can be found analytically:

$$\hat{K}(\omega_1,\omega_2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K(x,y)\mathrm{e}^{i(\omega_1x+\omega_2y)}\,\mathrm{d}x\,\mathrm{d}y=\frac{2\pi}{\sqrt{\omega_1^2+\omega_2^2}}=\frac{2\pi}{\omega}.\tag{5.3.19}$$

Here

$$L(\omega_1, \omega_2) = |\hat{K}(\omega_1, \omega_2)|^2 = \hat{K}(\omega_1, \omega_2)\hat{K}(-\omega_1, -\omega_2) = \frac{4\pi^2}{\omega^2}, \tag{5.3.20}$$

where $\omega = \sqrt{\omega_1^2 + \omega_2^2}$, $\alpha > 0$ is the regularization parameter. By Tikhonov, Arsenin, and Timonov (1987, p. 68) the following form of the regularizer $M(\omega_1, \omega_2)$ was adopted:

$$M(\omega_1, \omega_2) = 4\pi^2[(\omega_1^2 + \omega_2^2)^2 + 1] = 4\pi^2(\omega^4 + 1). \tag{5.3.21}$$

As a result, for the regularized spectrum of the solution we obtain the expression (Tikhonov, Arsenin, and Timonov, 1987, p. 68)

$$\hat{c}_\alpha(\omega_1, \omega_2) = \frac{1}{2\pi} \frac{\omega}{1 + \alpha\omega^2(\omega^4 + 1)} \hat{S}(\omega_1, \omega_2). \tag{5.3.22}$$

For expression (5.3.22) the following features are typical. First, $\hat{c}_\alpha(\omega_1, \omega_2) \to 0$ as $\omega \to \infty$, and integral (5.3.16) converges. Second, expression (5.3.22) provides for more accurate suppression of high frequencies ω compared to methods in which truncating high frequencies or smoothing windows are used (Verlan' and Sizikov, 1986, pp. 259, 260). The point here is that, on the one hand, high frequencies more strongly respond to errors and should therefore be suppressed and, second, high frequencies are necessary for more accurate representation of close features in tomograms; for this reason, a moderate suppression should be achieved. The Tikhonov regularization method satisfies both of these criteria (in a proportion governed by the value of α).

5.3.6. Numerical illustrations

Below, we give some examples of reconstruction of x-ray images using the various procedures. Figure 5.11 shows the solution of a model example (Tikhonov, Arsenin, Rubashov, and Timonov, 1984).

Figure 5.11 shows the solutions obtained by the Shepp–Logan method (this method belongs to "intuitive-regularization" methods), and by the Arsenin local regularization method (Tikhonov, Arsenin, and Timonov, 1987; Tikhonov, Arsenin, Rubashov, and Timonov, 1984) (this method ensures even a better accuracy than the Tikhonov regularization method). The solution obtained by the Shepp–Logan procedure is seen to represent very unstable a solution, whereas the local regularization method yields a high-accuracy solution.

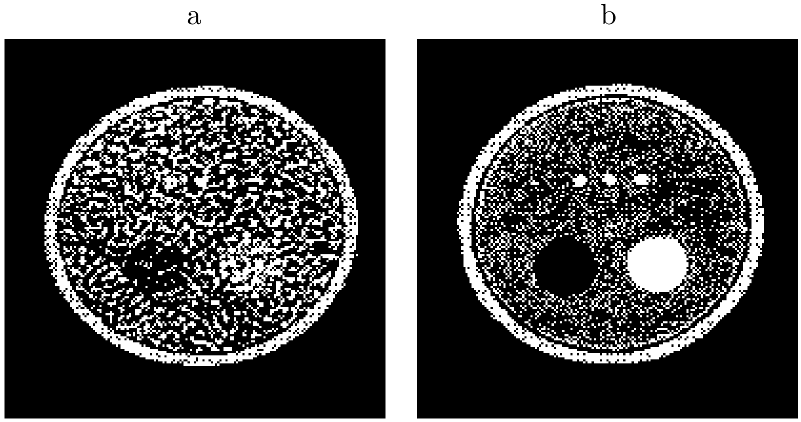

Figure 5.11. Reconstruction of an x-ray image (the model example) by the Shepp–Logan filter (a) and by the Arsenin local-regularization method (b)

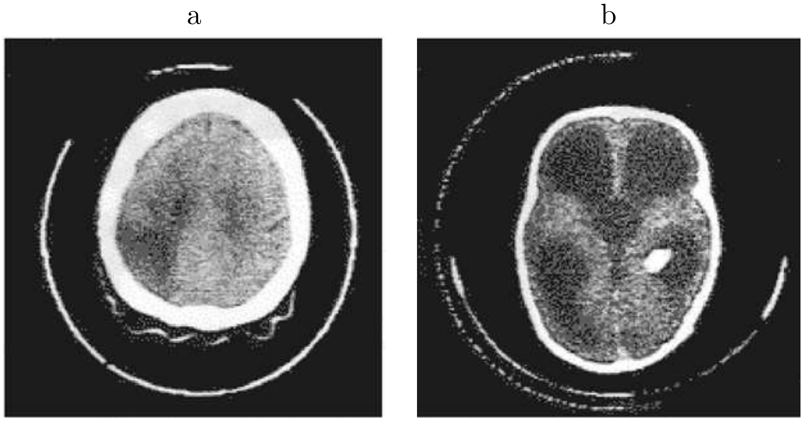

Figure 5.12. Tomograms of a brain cross-section obtained by the EMI procedure (a) and by the local regularization method (b)

Figure 5.12 shows tomograms (Tikhonov, Arsenin, Rubashov, and Timonov, 1984) of one and the same section of human brain obtained by the EMI procedure (CT-1010 tomograph) and by the local-regularization method (CPT-1000 tomograph). A specific feature here is that the substance density varies in this cross-section only within $\approx 0.5\,\%$; the EMI procedure (see Figure 5.12a) fails to resolve these variations, whereas the local-regularization method (see Figure 5.12b) successfully resolves them.

Additional examples can be found in Tikhonov, Arsenin, and Timonov (1987), Webb (1988, V. 1), Natterer (1986), etc.

5.3.7. About algorithms and programs

In practice, the above-described methods for reconstructing x-ray images are realized in the form of numerical algorithms and computer programs. The above-considered equations are integral equations of the convolution type whose solutions can be expressed in terms of Fourier transforms (with regularization). That is why a considerable portion of calculations reduces to calculating continuous Fourier transforms (CFT), (5.3.8), (5.3.11), (5.3.16), (5.3.18), (5.3.19), etc., direct or inverse, one- or two-dimensional. In turn, in practical calculations CFTs are replaced with discrete Fourier transforms (DFTs) realized normally as fast Fourier transforms (FFT).

For computer realization of FFT, many standard programs were developed, written in general-purpose programming languages (Fortran, C, Matlab, MathCAD, etc.). Among such programs are the FFT (Rabiner and Gold, 1975) and FTF1C (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) programs, intended for calculating one-dimensional FFTs, and the FTFTC (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) program, intended for calculating two-dimensional FFTs and written in the Fortran language.

In addition, programs for solving one- and two-dimensional first-kind convolution-type Fredholm integral equation by the Fourier transformation and regularization methods were developed. These are the PTIKR program (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995), the CONV1 program, etc. (Verlan' and Sizikov, 1986, pp. 379–388) (all these programs are intended for solving one-dimensional equations); and the PTITR program (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995) (this program is intended for solving two-dimensional equations); etc.

Real tomographs are normally supplied with software written not in general-purpose languages but, for instance, in assembler; sometimes such tomographs are provided with FFT co-processors.

In conclusion, it should be noted that there are also the algebraic reconstruction method, the iterative reconstruction method, etc. (Webb, 1988, V. 1; Natterer, 1986). Yet, we do not consider these methods.

5.4. MAGNETIC-FIELD SYNTHESIS IN AN NMR TOMOGRAPH

Among the numerous problems that arise in nuclear-magnetic-resonance tomography (NMR-tomography) (Sizikov, 2001, pp. 33–62; Tikhonov, Arsenin, and Timonov, 1987; Webb, 1988, V. 2; Afanas'ev, Studentsov, Khorev, et al, 1979; Galaydin, Ivanov, and Marusina, 2004; Lugansky, 1987; Lukhvich and Churilo, 1987; Montgomery, 1969; Cho, Jones, and Singh, 1993), there is an ill-posed problem that requires a regular method to be applied for its stable solution, namely, the problem of magnetic-field synthesis in an NMR tomograph (Sizikov, 2001; Afanas'ev, Studentsov, Khorev, et al, 1979; Galaydin, Ivanov, and Marusina, 2004; Lugansky, 1987; Lukhvich and Churilo, 1987; Sizikov, Akhmadulin, and Nikolaev, 2002; Adamiak, 1978).

5.4.1. Statement of the problem

In NMR-tomography, the magnetic field inside the tomograph must be made highly uniform: the intensity H of the basic (static) magnetic field in the working volume of the tomograph must vary only within $\Delta H/H \approx 10^{-5} \div 10^{-6}$ (Webb, 1988, V. 2; Galaydin, Ivanov, and Marusina, 2004; Lukhvich and Churilo, 1987; Cho, Jones, and Singh, 1993). The high uniformity of the magnetic field makes it possible to solve, with a high resolving power, the main problem of NMR-tomography, the problem of reconstruction of NMR images (Sizikov, 2001, pp. 45–51; Webb, 1988, V. 2; Cho, Jones, and Singh, 1993, pp. 270–274).

As a rule, synthesis of a highly uniform magnetic field is achieved by introducing Helmholtz solenoidal correcting coils of different orders (Galaydin, Ivanov, and Marusina, 2004). This engineering problem is very difficult to solve. Consider another approach to the problem of synthesis of highly uniform magnetic fields in NMR-tomographs. This approach is based on calculating a distribution law of the electric current over the tomograph coil winding enabling a highly uniform magnetic field inside the coil. This problem can be a *direct*, or *analysis*, *problem* (calculation of the magnetic field from known current distribution) or *inverse*, or *synthesis*, *problem* (calculation of the current distribution from a given magnetic field). The inverse problem is much more complex a problem than the direct one (Afanas'ev, Studentsov, Khorev, et al, 1979, p. 17).

Consider the problem of *magnetic-field synthesis* inside a tomographic coil (Sizikov, 2001, p. 55–61; Afanas'ev, Studentsov, Khorev, et al, 1979, p. 17; Lukhvich and Churilo, 1987; Sizikov, Akhmadulin, and Nikolaev, 2002; Adamiak, 1978) and, by way of example, the simplest variant of this problem, – magnetic field synthesis on the axis of a cylindrical coil with infinitely thin winding.

K. Adamiak (1978) was one of the first researchers who formulated the magnetic-field synthesis problem with allowance for its ill-posedness; he solved the problem by the Tikhonov regularization method. This problem was also treated by L. Lugansky (1987) and some other researchers. Yet, several inaccuracies were committed in Adamiak (1978); in particular, in the choice of α in the Tikhonov regularization method, K. Adamiak used the *minimum-discrepancy principle*, which gave out a very small value of $\alpha \approx 10^{-13}$; this, in fact, meant that no regularization was used; as a result, an unstable distribution of current was obtained. In Sizikov, Akhmadulin, and Nikolaev (2002), some of the statements and idea proposed by Adamiak (1978) were reformulated.

So, consider the following *inverse NMR-tomograhy problem*: given the magnetic-field intensity (strength) $H(z)$ on the axis of the coil, it is required to find the distribution of the current density $J(a)$ along the infinitely thin winding of a cylindrical coil. This is the *problem of integral synthesis of magnetic field* on the axis of an NMR-tomograph coil (Lukhvich and Churilo, 1987). Here, a is the coordinate along the coil winding and z is the coordinate along the axis of the coil. Consider the case of $H(z) = \text{const}$.

Figure 5.13 shows a cylindrical coil with infinitely thin winding, where $a \in [-l, l]$ is the distance along the coil winding; $z \in [-l, l]$ is the distance along the coil axis, reckoned from its center; l is the half-length of the coil; $H(z) = H = \text{const}$, $z \in [-l, l]$ is the given magnetic-field strength on the axis of the coil; $J(a)$, $a \in [-l, l]$ is the desired distribution of current along the coil winding; and R is the coil radius.

According to Montgomery (1969), in the case of $J(a) \neq \text{const}$ the coil is called *solenoid with variable current density*, or *solenoid with non-uniform current distribution.* A non-uniform distribution of the electric current can

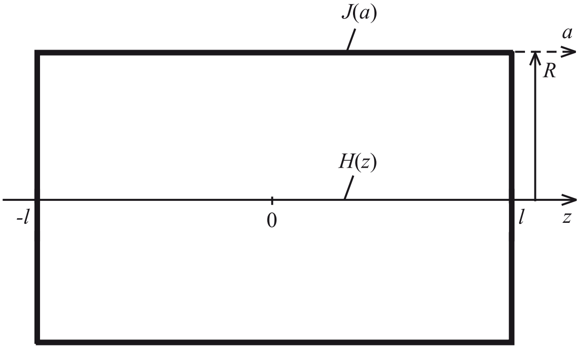

Figure 5.13. Cylindrical coil with infinitely thin winding

be achieved by dividing the solenoid into several individual sections. Other methods (use of insulated turns with individual current supply organized to each of the turns from a single source or use of a solenoid wound from a wire whose individual turn have different resistances) were proposed by Sizikov, Akhmadulin, and Nikolaev (2002).

5.4.2. Derivation of the integral equation for the synthesis problem

The magnetic-field strength on the axis of a thin turn (circular current) is (Afanas'ev, Studentsov, Khorev, et al, 1979, p. 20; Galaydin, Ivanov, and Marusina, 2004; Druzhkin, 1964, p. 327; Montgomery, 1969; Frish and Timoreva, 1952, pp. 330–333; Cho, Jones, and Singh, 1993) (see Figure 5.14)

$$H(z) = k\,\frac{JR^2}{\sqrt{[R^2+(z-a)^2]^3}}, \qquad -\infty < z < \infty,$$

where k is the proportionality factor, J is the electric current in the turn, R is the radius of the turn, z is the coordinate of the point at which we calculate the field H, and a is the z-coordinate of the center of the turn (the axis z is normal to the plane of the coil). To simplify further notation, we assume that $k = 1$. Then,

$$H(z) = \frac{JR^2}{\sqrt{[R^2+(z-a)^2]^3}}, \qquad -\infty < z < \infty. \tag{5.4.1}$$

Next, we assume that we have an infinite set of turns wound on a cylinder of radius R and half-length l with a current density $J = J(a)$ (see Figure 5.13). In this case, performing integration over all turns producing field strength (5.4.1), we obtain the following expression for the integral

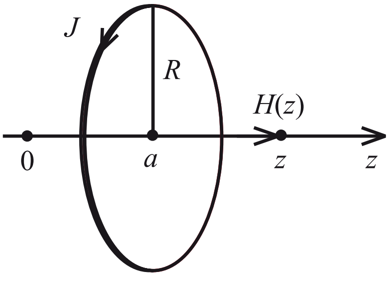

Figure 5.14. Field strength at the axis of the thin turn

magnetic-field strength at a point with coordinate z at the axis of the cylinder with infinitely thin winding (see Figure 5.13):

$$H(z) = \int_{-l}^{l} \frac{J(a)R^2\,\mathrm{d}a}{\sqrt{[R^2 + (z-a)^2]^3}}, \qquad -\infty < z < \infty. \tag{5.4.2}$$

Calculation of the field $H(z)$ from a given current $J(a)$ by formula (5.4.2) is the *direct problem.* It is worth noting that formula (5.4.2) is valid for all points both inside ($|z| \le l$) and outside ($|z| \ge l$) the cylinder. Let us briefly analyze the direct problem.

If $J(a) = J = \text{const}$, then integration (5.4.2) yields:

$$H(z) = \Big[\frac{z+l}{\sqrt{R^2 + (z+l)^2}} - \frac{z-l}{\sqrt{R^2 + (z-l)^2}}\Big] J, \qquad -\infty < z < \infty. \tag{5.4.3}$$

Particular cases of (5.4.3) are as follows: the field at the center of the coil is $H(0) = 2J/\sqrt{1 + (R/l)^2}$, and the field at the edges of the coil is $H(l) = J/\sqrt{1 + (R/2l)^2}$. If $R = l$, then $H(0) = 2J/\sqrt{2}$ and $H(l) = 2J/\sqrt{5}$; i. e., the field strength at the center of the coil is greater than the field strength at its edges by the factor $\sqrt{5/2} \approx 1.58$. If $z \to \infty$, then $H(z) \to 2JR^2l/z^3$ (Frish and Timoreva, 1952, p. 333); i. e., the field $H(z)$ outside the coil vanishes as $\sim z^{-3}$. This asymptotics holds both in the case of a thin coil (see (5.4.1)) and in the case of a cylindrical coil with $J(a) \neq \text{const}$.

If $l \gg R$ (infinitely long coil), then $H(0) = 2J$ and $H(l) = J$. We see that, here, the field strength at the center of the coil is two times greater than at its edges.

If, for instance, $l = 10R$, i. e., the length of the coil is twenty times greater than its radius, then $H(0) = 1.9901J$ and $H(R) = 1.9898J$; hence, $|H(R) - H(0)|\,/H(0) \approx 1.5 \cdot 10^{-4}$. We see that, in principle, using a long coil, one can achieve an magnetic field non-uniformity of order 10^{-4} in the region $|z| < R$; such a coil, however, is hard to fabricate.

The above brief analysis of the direct problem shows that in the case of $J(a) = \text{const}$ the field $H(z)$ decreases from the center of the coil towards its edges, whereas outside the coil $H(z) \sim |z|^{-3}$ as $|z| \to \infty$.

Let us consider now the *inverse problem.* We write (5.4.2) as

$$\int_{-l}^{l} \frac{R^2}{\sqrt{[R^2 + (z-a)^2]^3}}\, J(a)\,\mathrm{d}a = H(z), \qquad -l \le z \le l. \tag{5.4.4}$$

Relation (5.4.4) is the *first-kind Fredholm integral equation*, where $H(z)$ is the set right-hand side (magnetic-field strength at the axis of the coil)

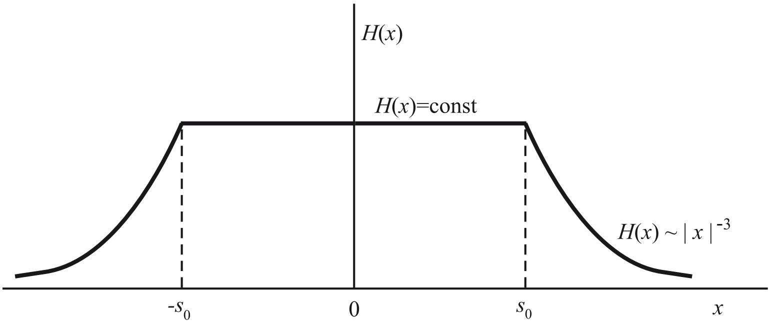

Figure 5.15. Magnetic-field strength $H(x)$ inside and outside the coil

($H(z) = H =$ const, for instance), and $J(a)$ is the function to be found (distribution of current along the infinitely thin winding of the cylindrical tomographic coil). The problem on solving equation (5.4.4) is an ill-posed problem (see Section 4.4).

We use the dimensionless variables $s = a/R$, $x = z/R$, and $s_0 = l/R$. Then, we can rewrite equation (5.4.4) as follows:

$$\int_{-s_0}^{s_0} K(x,s)J(s)\,\mathrm{d}s = H(x), \qquad -s_0 \le x \le s_0, \tag{5.4.5}$$

where

$$K(x,s) = 1/\sqrt{[1+(x-s)^2]^3} \tag{5.4.6}$$

is the kernel of the integral equation.

The solution of (5.4.5) and its realization make it possible to obtain a polarizing field on the axis of the coil with a given law of field strength ($H(x) = H =$ const, for instance).

5.4.3. Solution of the equation by the regularization method with constraints

Consider now how equation (5.4.5) can be solved.

Perform first a *qualitative analysis.* Note that notation (5.4.5), (5.4.6) is valid not only for $x \in [-s_0, s_0]$, but also for $x \notin [-s_0, s_0]$.

Let $H(x) =$ const for $x \in [-s_0, s_0]$. Then, the field $H(x) \neq$ const at $x \notin [-s_0, s_0]$ and monotonically vanishes with increasing $|x|$; as $|x| \to \infty$, the field follows the asymptotic $H(x) \sim |x|^{-3}$ (see Figure 5.15).

We see that the function $H(x)$ has discontinuous derivatives $H'(x)$, $H''(x)$, ... at $x = \pm s_0$. It follows from here that the desired function $J(s)$

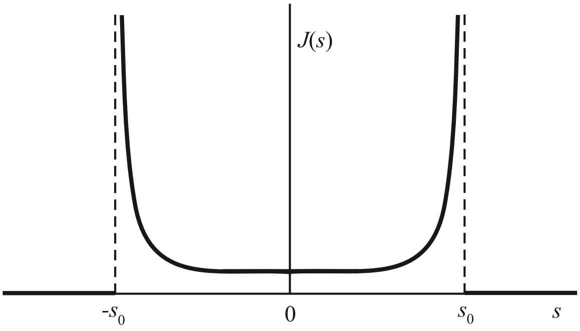

Figure 5.16. Function $J(s)$ (schematically)

is singular at $s = -s_0 + 0$ and $s = s_0 - 0$, and for $|s| > s_0$ the current $J(s) = 0$, in compliance with the physical statement of the problem. Besides, if $H(-x) = H(x)$, then $J(-s) = J(s)$.

Thus, the function $J(s)$ is a singular and symmetric function, i. e., a function schematically shown in Figure 5.16.

Consider the *numerical solution algorithm.* In principle, one can try to seek for an analytical solution of equation (5.4.5). Yet, no analytical solution can be found here. Consider therefore the matter of numerical solution of this equation. Taking into account incorrectness of (5.4.5), let us use the Tikhonov regularization method (see Section 4.7).

Adamiak (1978) also used the Tikhonov regularization method. The numerical solution of (5.4.5) obtained him showed that the regularized solution $J_\alpha(s)$ obtained with small values of α displays alternating large-amplitude fluctuations. At the same time, it can be expected that the regularization parameter should indeed be very small (of order $10^{-5} \div 10^{-10}$) because the function $H(x)$ is known exactly and the problem involves only algorithm-induced errors. Besides, the regularized solution $J_\alpha(s)$ is expected to vary in a broad range of values (because the exact solution is singular). In addition, the solution $J_\alpha(s)$ must monotonically increase from the center of the coil ($s = 0$) towards its edges ($s = \pm\, s_0$).

For the regularized solution $J_\alpha(s)$ to possess the above-described properties, namely, to be non-negative and monotonically increasing from the center of the coil towards its edges, it must be sought for by the *Tikhonov regularization method with constraints imposed on solution* (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995).

The *constraints* consist in that the solution $J_\alpha(s)$ is sought on the set of non-negative monotonically non-increasing functions (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995). Yet, only left part of the solution $J_\alpha(s)$ (for $s \in (-s_0, 0])$ satisfies this condition. Therefore, we modify equation (5.4.5) taking into account the symmetry of the solution: $J(-s) = J(s)$.

Modification of equation (5.4.5). We write (5.4.5) as

$$H(x) = \int_{-s_0}^{0} \frac{J(s)\,\mathrm{d}s}{\sqrt{[1+(x-s)^2]^3}} + \int_{0}^{s_0} \frac{J(s)\,\mathrm{d}s}{\sqrt{[1+(x-s)^2]^3}}$$
$$= \int_{-s_0}^{0} \frac{J(s)\,\mathrm{d}s}{\sqrt{[1+(x-s)^2]^3}} + \int_{-s_0}^{0} \frac{J(-s)\,\mathrm{d}s}{\sqrt{[1+(x+s)^2]^3}}\,.$$

Since $J(s) = J(-s)$, i. e., the current distribution is symmetric, then $H(x) = H(-x)$, i. e., the magnetic field $H(x)$ is also symmetric. The converse is also valid: if $H(x) = H(-x)$, then $J(s) = J(-s)$.

As a result, we can write equation (5.4.5) as follows:

$$\int_{-s_0}^{0} R(x,s)J(s)\,\mathrm{d}s = H(x), \qquad -s_0 \le x \le 0, \tag{5.4.7}$$

where the new kernel is

$$R(x,s) = \frac{1}{\sqrt{[1+(x+s)^2]^3}} + \frac{1}{\sqrt{[1+(x-s)^2]^3}}. \tag{5.4.8}$$

So, we seek the solution $J(s)$ in the left half-space by solving equation (5.4.7), i. e., for $s \in (-s_0, 0]$, and assume that in the right half-space $J(s) = J(-s)$, $s \in [0, s_0)$.

We will additionally take the fact into account that the function $J(s)$ is a non-negative function monotonically non-increasing over the interval $s \in [-s_0, 0]$; such a function is schematically shown in Figure 5.17.

To obtain the numerical solution $J_\alpha(s)$ by the Tikhonov regularization method on the set of non-negative monotonically non-increasing functions, one can employ the PTIPR Fortran program (Tikhonov, Goncharsky, Stepanov, and Yagola, 1995). In this program, a smoothing functional is minimized by the conjugate gradient projection method on the set of non-negative monotonically non-increasing functions at each value of α ($\alpha > 0$).

Example. Using the PTIPR program, we solved the following *example* (similar to the examples considered by Adamiak (1978); Sizikov (2001, p. 57); Sizikov, Akhmadulin, and Nikolaev (2002)): $s_0 = 0.5$, discretization

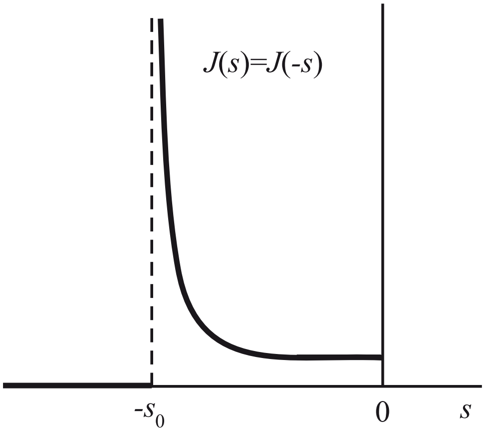

Figure 5.17.

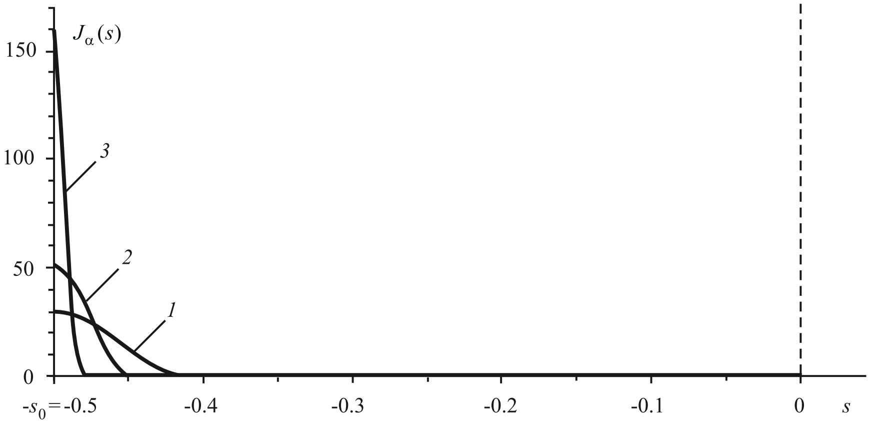

Figure 5.18. Regularized solutions $J_\alpha(s)$ with constraints: *1* — $\alpha = 10^{-4}$; *2* — $\alpha = 1.1 \cdot 10^{-5}$; *3* — $\alpha = 10^{-6}$

step $h = \Delta s = \Delta x = 0.00625$, number of discretization steps in the left half-space $n = s_0/h = 80$, total number of turns in the coil $N = 2n + 1 = 161$, field strength $H(x) = H = \text{const} = 1$, $-s_0 \leq x \leq s_0$.

The problem was solved for several values of α. Figure 5.18 shows the solutions $J_\alpha(s)$ obtained with $\alpha = 10^{-4}$ (curve *1*), $\alpha = 1.1 \cdot 10^{-5}$ (curve *2*), and $\alpha = 10^{-6}$ (curve *3*).

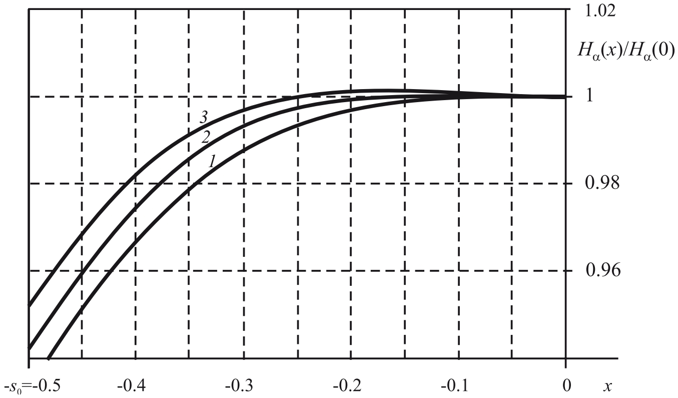

Figure 5.19. Regularized magnetic-field strength $H_\alpha(x)$: *1* — $\alpha = 10^{-4}$; *2* — $\alpha = 1.1 \cdot 10^{-5}$; *3* – $\alpha = 10^{-6}$

Figure 5.19 shows the magnetic field strength distributions calculated from the found solutions $J_\alpha(s)$ by the formula (cp. (5.4.7))

$$H_\alpha(x) = \int_{-s_0}^{0} R(x,s) J_\alpha(s)\,\mathrm{d}s, \qquad -s_0 \le x \le 0, \tag{5.4.9}$$

with $\alpha = 10^{-4}$ (curve *1*), $\alpha = 1.1 \cdot 10^{-5}$ (curve *2*), and $\alpha = 10^{-6}$ (curve *3*).

The integral in (5.4.7) was represented and the integral in (5.4.9) calculated by the trapezoid formula with a constant step h.

Figure 5.18 shows that with decreasing α, the ratio $J_\alpha(-s_0)/J_\alpha(0)$ increases and, as $\alpha \to 0$, the solution $J_\alpha(s)$ passes into a singular function. Figure 5.19 shows that, for some moderate (optimal) value of α (in this example $\alpha_{\text{opt}} \approx 1.1 \cdot 10^{-5}$), the relative field non-uniformity

$$\Delta H_{\alpha\,\text{rel}}(x) = \frac{|H_\alpha(x) - H_\alpha(0)|}{H_\alpha(0)} = \left|\frac{H_\alpha(x)}{H_\alpha(0)} - 1\right|$$

is equal $\approx 10^{-4} \div 10^{-5}$ for $|x| \in [0, 0.15] = [0, 0.3]s_0$.

From the above consideration, the following *conclusions* can be drawn: 1) although the use of the Tikhonov regularization method without constraints imposed on the solution stabilizes solution, the solution $J_\alpha(s)$ obtained with small values of α displays fluctuations whose amplitude increases as the parameter α decreases (see Adamiak (1978) and Sizikov (2001, p. 58)); 2) in using the Tikhonov regularization method with constraints imposed on solution, the fluctuations in the solution $J_\alpha(s)$ vanish even at very small values of α (see Figure 5.18); 3) as it is seen from Figure 5.18, as $\alpha \to 0$, the solution $J_\alpha(s)$ passes into a singular function resembling the δ-function; from the practical point of view this means that the coil must be made in the form of a solenoid with uniform current distribution except for several end turns in which the current increases towards the edges of the coil.

5.4.4. Solution obtained by the approach to the δ-function

The latter example was also solved as follows (Sizikov, Akhmadulin, and Nikolaev, 2002). The desired function $J(s)$ was set equal to

$$J(s) = \begin{cases} J_0, & s = -s_0, \\ J, & s \neq -s_0, \end{cases} \tag{5.4.10}$$

(approach to the δ-function), where $J = \text{const} = 1$, and J_0 was a varied parameter. From the engineering point of view, function (5.4.10) is equivalent to a combination of a solenoidal coil with the current J and a pair of thin turns (Helmholtz rings) with the current J_0 located at the edges of the coil (the solenoid and the turns are the same radius).

The integral in (5.4.7) was approximated by the trapezoid formula and the magnetic-field strength $H(x)$ was calculated as

$$H(x) = h \sum_{j=0}^{n} p_j R(x, s_j) J(s_j), \tag{5.4.11}$$

where

$$s_j = -s_0 + hj, \qquad p_j = \begin{cases} 0.5, & j = 0 \text{ or } j = n, \\ 1, & 0 < j < n. \end{cases}$$

Figure 5.20 shows the predicted field strengthes $H(x)$ for $J_0 = 110$, 120, and 130 (curves *1*, *2*, and *3*, respectively), $s_0 = 0.5$, $h = 0.00625$, $n = 80$.

We see that, at a certain value of J_0, the average field uniformity (curve *2* in Figure 5.20 for $|x| \le 0.4 s_0$, i. e., in the working volume) is of order 10^{-4}. The greater the number N of turns, the better is the field uniformity.

The following *conclusion* can be drawn. With the electric-current distribution $J(s)$ made uniform along the tomographic coil except for its edge turns (i. e., with the distribution of type (5.4.10)), a more uniform magnetic-field strength $H(x)$ can be achieved at the axis of the coil. Although more simple than the previous method (Tikhonov regularization method with constraints imposed on solution), the method under consideration (approach to

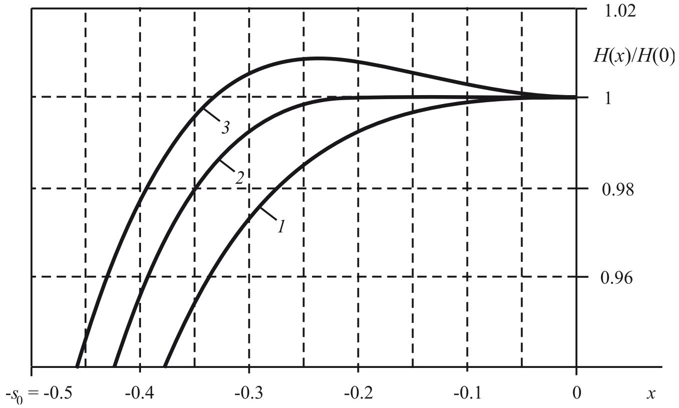

Figure 5.20. Field strengthes $H(x)$ calculated by formula (5.4.11):
1 — $J_0 = 110$; *2* — $J_0 = 120$; *3* — $J_0 = 130$

the δ-function) yields better results (provides for a more uniform magnetic field). This example shows how important it is to use additional information in solving ill-posed problems and to construct, on the basis of this information, an adequate algorithm. Although both methods use information about the solution, the second method admits a more adequate mathematical description of the problem and, as a consequence, yields a more accurate solution.

The second method is not a novelty. As a matter of fact, is was previously used by L. Lugansky (1987) and others. Yet, the purpose of the present narration was demonstration of various variants of stable methods for solving ill-posed problem rather than gaining most accurate practical results.

Bibliography to Part II

Basic literature

Bakushinsky A. and Goncharsky A. (1994). *Ill-Posed Problems: Theory and Applications.* Kluwer, Dordrecht.

Bates R. H. T. and McDonnell M. J. (1986). *Image Restoration and Reconstruction.* Clarendon Press, Oxford.

Bronshtein I. N. and Semendyaev K. A. (1986). *Reference Book on Mathematics for Engineers and Students of Engineering Schools.* Nauka, Moscow, 13th edition (in Russian).

Ivanov V. K., Vasin V. V., and Tanana V. P. (2002). *Theory of Linear Ill-Posed Problems and its Applications.* VSP, Zeist.

Lavrent'ev M. M., Romanov V. G., and Shishatskii S. P. (1997). *Ill-Posed Problems of Mathematical Physics and Analysis.* AMS, Providence.

Sizikov V. S. (2001). *Mathematical Methods for Processing the Results of Measurements.* Politekhnika, St.-Petersburg (in Russian). Electronic version: Sizikov V. S. *Stable Methods for Processing the Results of Measurements.* `http://de.ifmo.ru/--books/SIZIKOV.PDF` or `http://dsp-book.narod.ru/SIZIKOV.pdf` .

Tikhonov A. N. and Arsenin V. Ya. (1977). *Solution of Ill-Posed Problems.* Wiley, NY.

Tikhonov A. N., Arsenin V. Ya., and Timonov A. A. (1987). *Mathematical Problems of Computer Tomography.* Nauka, Moscow (in Russian).

Tikhonov A. N., Goncharsky A. V., Stepanov V. V., and Yagola A. G. (1995). *Mathematical Methods for the Solution of Ill-Posed Problems.* Kluwer, Dordrecht.

Verlan' A. F. and Sizikov V. S. (1986). *Integral Equations: Methods, Algorithms, Programs.* Naukova Dumka, Kiev (in Russian).

Webb S. (Ed.). (1988). *The Physics of Medical Imaging.* Vol. 1, 2. Hilger, Bristol.

Supplementary reading

Adamiak K. (1978). Method of the magnetic field synthesis on the axis of cylinder solenoid. *Appl. Phys.*, 16, 417–423.

Afanas'ev Yu. V., Studentsov N. V., Khorev V. N., et al. (1979). *Measuring Tools for Magnetic-Field Characteristics.* Energiya, Leningrad (in Russian).

Apartsyn A. S. (2003). *Nonclassical Linear Volterra Equations of the First Kind.* VSP, Zeist.

Belov I. A. and Sizikov V. S. (2001). *Software for Reconstruction of Blurred and Defocused Images.* Registration certificate of a computer program No. 2001610743. 18.06.2001 (in Russian).

Boikov I. V. and Krivulin N. P. (2000). Determination of dynamic characteristics of transducers with distributed parameters. *Izmeritel'naya tekhnika*, 9, 3–7 (in Russian).

Bracewell R. N. (1986). *The Hartley Transform.* Oxford Univ. Press, NY.

Bracewell R. N. and Riddle A. C. (1967). Inversion of fan-beam scans in radio astronomy. *Astrophys. J.*, 150, 427–434.

Brunner H. and Sizikov V. (1998). On a suboptimal filtration method for solving convolution-type integral equations of the first kind. *J. Math. Anal. Appl.*, 226 (2), 292–308.

Chaisson E. J. (1992). Early results from the Hubble space telescope. *Sci. American*, 266 (6), 6–14.

Cho Z. H., Jones J. P., and Singh M. (1993). *Foundations of Medical Imaging.* Wiley, New York.

Danilevich Ya. B. and Petrov Yu. P. (2000). On the necessity of extending the notion of equivalence of mathematical models. *Dokl. Akad. Nauk*, 371 (4), 473–475 (in Russian).

Druzhkin L. A. (1964). *Problems in Field Theory.* MIRGE, Moscow (in Russian).

Frish S. E. and Timoreva A. V. (1952). *A Course of General Physics.* V. 2. GITTL, Moscow–Leningrad (in Russian).

Galaydin P. A., Ivanov V. A., and Marusina M. Ya. (2004). *Calculation and Design of Electromagnetic Systems of Magneto-Resonance Tomographs.* Students Guide. SPbSU ITMO, St.-Petersburg (in Russian).

Gantmacher F. R. (1959). *The Theory of Matrices.* AMS Chelsea Publ., NY.

Glasko V. B., Mudretsova E. A., and Strakhov V. N. (1987). Inverse problems in gravimetry and magnetometry. In: *Ill-Posed Problems in Natural Science.* A. N. Tikhonov and A. V. Goncharsky (Eds). Moscow State University, Moscow, 89–102 (in Russian).

Kantorovic L. and Akilov V. (1964). *Functional Analysis in Normed Spaces.* Pergamon Press, Oxford.

Kolmogorov A. N. and Fomin S. V. (1981). *Elements of Function Theory and Functional Analysis.* Nauka, Moscow, 5th edition (in Russian).

Korn G. A. and Korn T. M. (1961). *Mathematical Handbook for Scientists and Engineers.* McGraw-Hill, NY.

Kurosh A. G. (1975). *A Course of Higher Algebra.* Nauka, Moscow, 11th edition (in Russian).

Lattes R. and Lions J. L. (1969). *The Method of Quasi Reversibility.* American Elsevier, NY.

Lavrent'ev M. M. (1981). *Ill-Posed Problems for Differential Equations.* Novisibirsk State University, Novosibirsk (in Russian).

Lugansky L. B. (1987). Optimal coils for producing uniform magnetic fields. *J. Phys. E: Sci. Instrum.*, 20, 277–285.

Lukhvich A. A. and Churilo V. R. (1987). Sources of polarizing magnetic fields and its gradients for NMR-tomography: a review. *Prib. Tekhn. Exper.* 5, 172–173 (in Russian).

Montgomery D. B. (1969). *Solenoid Magnet Design. The Magnetic and Mechanical Aspects of Resistive and Superconducting Systems.* Wiley, NY.

Morozov V. A. (1984). *Methods for Solving Incorrectly Posed Problems.* Springer-Verlag, NY.

Natterer F. (1986). *The Mathematics of Computerized Tomography.* Wiley, Chichester.

Petrov Yu. P. (1998). *Third Class of Problems in Physics and Engineering, Intermediate between Well- and Ill-Posed Ones.* St.-Petersburg State University, St.-Petersburg (in Russian).

Pontryagin L. S. (1982). *Ordinary Differential Equations.* 5th edition. Nauka, Moscow (in Russian).

Rabiner L. R. and Gold B. (1975). *Theory and Application of Digital Signal Processing.* Englewood Cliffs, Prentice-Hall.

Ramachandran G. N. and Lakshminarayanan A. V. (1971). Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. *Proc. Nat. Acad. Sci. US*, 68, 2236–2240.

Sizikov V. S., Akhmadulin R. I., and Nikolaev D. B. (2002). Synthesis of a magnetic field along the axis of a NMR-tomograph coil. *Izv. VUZov. Priborostroenie*, 45 (1), 52–57 (in Russian).

Sizikov V. S. and Belov I. A. (2000). Reconstruction of smeared and out-of-focus images by regularization. *J. Optical Technology*, 67 (4), 60–63.

Sizikov V. S. and Shchekotin D. S. (2004). On combination of high resolution and safety in CT-diagnosis. In: *Proc. 4th Int. Conf. IEHS'2004.* Solnitsev R. (Ed). SPUAI, St.-Petersburg, 131–134.

Tikhonov A. N. (1943). On stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39 (5), 195–198 (in Russian).

Tikhonov A. N., Arsenin V. Ya., Rubashov I. B., and Timonov A. A. (1984). First soviet computer tomograph. *Priroda*, 4, 11–21 (in Russian).

Tikhonov A. N., Goncharskii A. V., and Stepanov V. V. (1987). Ill-posed problems in photo-image processing. In: *Ill-Posed Problems in Natural Science.* A. N. Tikhonov and A. V. Goncharskii (Eds). Moscow State University, Moscow, 185–195 (in Russian).

Tikhonov A. N., Leonov A. S., and Yagola A. G. (1997). *Nonlinear Ill-Posed Problems.* CRC Press UK, London.

Tichonov A. N. and Samarskij A. A. (1963). *Equations of Mathematical Physics.* 2nd ed. Pergamon Press, NY.

Troitskii I. N. (1989). *Statistical Tomography Theory.* Radio i Svyaz', Moscow (in Russian).

Vasilenko G. I. (1979). *Signal Restoration Theory.* Sovetskoe Radio, Moscow (in Russian).

Vasil'ev V. N. and Gurov I. N. (1998). *Computer-Assisted Signal Processing as Applied to Interferometric Systems.* BKhV, St.-Petersburg (in Russian).

Voevodin V. V. (1980). *Linear Algebra.* Nauka, Moscow (in Russian).

Voskoboynikov Yu. E., Preobrazhenski N. G., and Sedel'nikov A. I. (1984). *Mathematical Treatment of Experiment in Molecular Gas Dynamics.* Nauka, Novosibirsk (in Russian).

Wilkinson J. H. (1993). *The Algebraic Eigenvalue Problem.* Oxford Univ. Press, Oxford.

Index