**V. A. Morozov and A. I. Grebennikov**

# METHODS FOR SOLUTION OF ILL-POSED PROBLEMS

*ALGORITHMIC ASPECTS*

**Monograph**

**Electronic publication**

**Moscow, 2011**

# Contents


**1 BASIC CONCEPTS OF REGULARIZATION THEORY 4**

1.1 Variational methods of regularization . . . . . . . . . . . . . . . 4
1.2 Special regularizing algorithms and their properties . . . . . . . 17
1.3 On the problem of correction of linear inequalities . . . . . . . 31
1.4 Shift regularization and its application for solving singular systems of linear algebraic equations . . . . . . . . . . . . . . . . . . 41
1.5 Methods for solving singular integral equations with Hilbert kernels . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62

**2 METHODS OF DIGITAL IMAGE RESTORATION AND ILL-POSED PROBLEMS OF AUTOMATIC CONTROL 76**

2.1 Computerized systems of image processing . . . . . . . . . . . . 76
2.2 Fast algorithms of digital image recovery . . . . . . . . . . . . . 81
2.3 Fast algorithms of signal recovery . . . . . . . . . . . . . . . . . 91
2.4 A collection of fast Fourier transform algorithms for discrete series 108
2.5 Fast and accelerated solution of special linear systems . . . . . 117
2.6 Stable design of pulse transition functions . . . . . . . . . . . . 129
2.7 Fast algorithms for solving the Wiener–Hopf equations . . . . . 153
2.8 Approximate solution of Wiener–Hopf discrete equations and estimation of accuracy . . . . . . . . . . . . . . . . . . . . . . . 162

**3 THE SPLINE APPROXIMATION METHOD AND ITS APPLICATIONS 172**

3.1 The collocation method for the first-kind linear operator equations 172
3.2 The sequential smoothing method for the first-kind linear operator equations . . . . . . . . . . . . . . . . . . . . . . . . . . . 178
3.3 Explicit spline approximation of functions and experimental data processing algorithms . . . . . . . . . . . . . . . . . . . . . 185
3.4 Algorithms for solving linear integral equations with singularities in kernels . . . . . . . . . . . . . . . . . . . . . . . . . . . 200

3.5 Collections of FORTRAN subroutines and the solution of some applied problems . . . . . . . . . . . . . . . . . . . . . . . . . 208

**4 THE METHOD OF DESCRIPTIVE REGULARIZATION 218**

4.1 The method of descriptive regularization and the quality of approximate solutions . . . . . . . . . . . . . . . . . . . . . . . . . 218

4.2 Peculiarities of numerical implementations of the descriptive regularization method . . . . . . . . . . . . . . . . . . . . . . . 226

4.3 Algorithms of piecewise monotone and convex approximation . 242

4.4 Convergence of monotone and convex solutions of integral equations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 262

4.5 The descriptive regularization method and the stable solution of linear systems . . . . . . . . . . . . . . . . . . . . . . . . . . 275

4.6 Chebyshev's solutions to linear systems . . . . . . . . . . . . . 286

Bibliography . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 293

# Chapter 1

# BASIC CONCEPTS OF REGULARIZATION THEORY

## 1.1 Variational methods of regularization

**1.** It is well known that the problem of solving the operator equation

$$Au = f, \tag{1.1.1}$$

where $A : U \to F$ (here $U$ and $F$ are complete metric spaces) and $A$ is an operator with a nonempty definition domain $D_A \subseteq U$, is said to be a classically (according to Hadamard) well-posed problem when the following conditions are satisfied:

a) solvability: $\forall f \in F \Rightarrow \exists u \in D_A : Au = f$;

b) uniqueness: $\forall u_1, u_2 \in D_A$, $Au_1 = Au_2 \Rightarrow u_1 = u_2$;

c) stability: if $\forall u, \tilde{u} \in D_A$: $Au = f$, $A\tilde{u} = \widetilde{f}$, then $\tilde{u} \to u$ in $U$ as $\tilde{f} \to f$ in $F$.

However, a number of practically important problems do not satisfy the above conditions of Hadamard's well-posedness. Such problems are, for example, the problems that lead to the Fredholm linear integral equations of the first kind

$$Au \equiv \int\limits_a^b k(x,\xi)u(\xi)\,d\xi = f(x), \quad c \le x \le d, \tag{1.1.2}$$

where $U = L_2[a,b]$, $F = L_2[c,d]$, $k \in L_2[\Pi]$, and $\Pi = [c,d] \times [a,b]$.

This remark requires a refinement of the concept of well-posed problems. As a result, the concept of conditionally well-posed problems was developed (well-posed problems according to A. N. Tikhonov [132]): the problem (1.1.1) is said to be well posed according to Tikhonov if it is possible to specify a nonempty class $M \subseteq D_A$ such that the following conditions are satisfied:

a$'$) $f \in N = A[M] \Rightarrow \exists u \in M: \ Au = f$;

b$'$) $M$ is a uniqueness set: $\forall u_1, u_2 \in M$: $Au_1 = Au_2 \Rightarrow u_1 = u_2$;

c$'$) $M$ is a set of stability: $\forall f, \tilde{f} \in N$: $Au = f$, $A\tilde{u} = \tilde{f} \Rightarrow \tilde{u} \to u$ if $\tilde{f} \to f$ in $F$.

If the uniqueness set $M$ is a compactum and $A$ is a continuous operator, then it is said that $M$ is a set of well-posedness [254], i.e., the problem (1.1.1) is well posed according to A. N. Tikhonov.

This approach was repeatedly generalized by many authors in various directions. For example, the operator $A$ may not be continuous and $M$ may not be a set of uniqueness; moreover, $M$ may not be a compactum [115], [172].

The existence of conditionally well-posed problems provides a principal opportunity to study classically ill-posed problems. However, the question of efficient methods for constructing the approximate solutions to such problems remains open. A difficulty consists in the fact that input data for the problem (1.1.1) (i.e., the elements $\tilde{f}$) may not belong to the set $N$ (even the establishment of the fact that $\tilde{f} \in N$, as a rule, is nontrivial); even if $\tilde{f} \in N$, the problem of practical construction of a solution to equation (1.1.1) with such data remains open.

Thus, the problem is to give a meaning to the "solution" of equation (1.1.1) and to those data $\tilde{f}$ that do not belong to $N$ (ideally, to any $\tilde{f} \in F$). An objective base for this purpose is *a priori* selection of a class $M \in D_A$ of feasible solutions to (1.1.1). On this basis, V. K. Ivanov [110] has introduced the concept of a quasisolution to equation (1.1.1) as an element $u_f \in M$ for which

$$\rho_F\left(Au_f, f\right) = \inf_{v \in M} \rho_F\left(Av, f\right) = \mu_f. \tag{1.1.3}$$

Let $U_f$ be the set of all quasisolutions (1.1.3) corresponding to the right-hand side $f \in F$. If $f \in N = A\left[M\right]$, then $Au_f = f$, i.e., $u_f$ is the classical solution to (1.1.1).

**Theorem 1.1.1** *Let the following conditions be fulfilled:*

1) *the problem* (1.1.1) *is well posed according to Tikhonov;*

2) $U_g$ *is nonempty for any* $g \in F$.

*If* $u_f \in M: \ Au_f = f$, *then* $\lim\limits_{\tilde{f} \to f} \sup\limits_{u \in U_{\tilde{f}}} \rho_U(u, u_f) = 0$.

**Remark 1** . *The solvability condition* 2) *can be eliminated if the* $\varepsilon$-*quasisolutions* $u_f^\varepsilon \in M: \ \rho_F(Au_f^\varepsilon, f) \leq \mu_f + \varepsilon, \ \varepsilon \to 0+$, *are used as qua-*

*sisolutions. Then, $U_{\tilde{f}}$ should be replaced by the set $U_{\tilde{f}}^{\varepsilon}$ of all $u_{\tilde{f}}^{\varepsilon}$, which is always nonempty.*

**Theorem 1.1.2** *Let the problem* (1.1.1) *be well posed according to Tikhonov. If $U$ is a Banach space, $F$ is a Hilbert space, $A$ is a linear operator, and the set $N = A[M]$ is closed and convex, then*

1) $\forall f \in F \;\; \exists u_f \in M$;
2) *the element $u_f$ is uniquely defined by a given element $f$;*
3) *if $\tilde{f} \to f$, then $u_{\tilde{f}} \to u_f$, i.e., the stability of quasisolutions takes place.*

This theorem is of fundamental importance, since under its hypotheses the conditions of well-posedness according to Hadamard are fulfilled if the solution to (1.1.1) is understood as a quasisolution (for the same input data). Note also that *a priori*, Theorems 1.1.1 and 1.1.2 do not require the assumption of compactness of the set $M$, which sometimes is considered as a necessary assumption when solving an ill-posed problem according to Hadamard. In some cases, informal separation of well-posedness sets for the problem (1.1.1) is rather delicate and requires special research. Therefore, $M = D_A$ is often considered as a natural domain of definition for the operator $A$. In this case, the set $U_f$ (if it is nonempty) defines the set of the so-called pseudosolutions to (1.1.1) or solutions to (1.1.1) in the least-squares sense [321]. Nonuniqueness of pseudosolutions is allowed. Then, for a given "trial" element $u^* \in U$ ($u^* = 0$ is often chosen) it is natural to choose $u_f^* \in U_f$ in such a way that

$$\rho_U(u_f^*, u^*) = \inf_{u \in U_f} \rho_U\left(u, u^*\right) = \mu^*. \tag{1.1.4}$$

The set $U_f^*$ of all such $u_f^*$ is called the set of normal pseudosolutions. If $Au_f^* = f$, then $u_f^*$ is said to be the normal solution to (1.1.1). This concept was introduced by A. N. Tikhonov and plays an important role in the theory of regularization of ill-posed problems. An economic interpretation of the normal solution was also given by A. N. Tikhonov, which allows one to consider these objects in mathematics informally.

The following theorem is valid.

**Theorem 1.1.3** *Let $U$, $F$ be Hilbert spaces and $A$ be a linear closed operator on a linear set $D_A$. Then:*

*a)* $\exists$ *a unique* $u_f^* \in U_f \neq \emptyset$;
*b)* $U_f \neq \emptyset \Leftrightarrow Pf \in Q_A = A(D_A)$, *$P$ is an orthoprojector onto $Q_A$;*
*c)* $U_f \neq \emptyset \; \forall f \in F \Leftrightarrow Q_A = \overline{Q}_A$, *i.e., $Q_A$ is closed. In this case,*

$$\lim_{\tilde{f} \to f} \left| u_{\tilde{f}}^* - u_f^* \right|_A = 0, \quad \left|\cdot\right|_A^2 \equiv \left\|\cdot\right\|_U^2 + \left\|A\cdot\right\|_F^2 .$$

If $Q_A = \overline{Q}_A$, the operator $A$ is said to be normally solvable. The normal solvability of $A$ restores, as it follows from Theorem 1.1.3, the well-posedness of (1.1.1) in the sense of pseudosolutions. Note that in general this condition does not provide neither the classical well-posedness nor the well-posedness in the sense of quasisolutions to (1.1.1).

**2.** Let us put the question: what is meant by the term "well-posed problem"? First of all, it seems necessary to clarify the term "posed problem", since it is obvious that the concept of "well-posedness" of a problem originates from the concept "problem".

As a rule, a mathematical formulation of a problem is preceded by the development and construction (at a qualitative level at first) of a mathematical model for the physical process (phenomenon, medium) under study. On the basis of particular physical laws, this model should adequately describe a quantitative interrelation between an unknown state of the medium and its known manifestation.

The main features of a mathematical formulation of a problem are:

1) a clear separation of the class of original data $D$ (these data may be elements or operators of some spaces, the spaces themselves or their subsets, and etc.);

2) the choice of a space $U$ of feasible solutions to a problem; these solutions should characterize the medium's state;

3) the fixation of a rule (an operator, an algorithm) $R$ for the construction of a solution defined for some original data $d \in D$, so that $R(d) = U_d \subseteq U$.

It is allowed that the solution set $U_d$ may consist of more than one element. The choice of the rule $R$ is often dictated (and rather rigorously) by the two reasons: a practical significance of a solution to a given problem and the possibility of simple realization of this solution. If a problem $R$ is posed, the question of the validity of its posedness (is it right or wrong) is not important in itself, since this problem should adequately reflect the physical reality under study and the usefulness of its solution is dictated by the intrinsic necessity of the mathematics evolution. Unfortunately, the terms "right" and "wrong" are the synonyms of the terms "correct" and "incorrect" included into the Russian language; this fact causes some misunderstanding (especially, among nonspecialists) in interpretation of the concept of correctness for a mathematical problem. In essence, this concept reflects only one, but a very important property — the stability of the problem with respect to variations of input data.

With this in mind, we shall say that a problem $R$ is $A$-correct if

1) $R$ is defined for all $d \in D$;

2) $U_d = R(d),\ U_{\tilde{d}} = R(\tilde{d}) \Rightarrow U_{\tilde{d}} \to U_d$ when $\tilde{d} \to d$.

Here the symbol "$\to$" indicates a well-defined approximation in the space of data (in the space of results). For example, if $U$ is metric, the symbol "$\to$" can be understood as a synonym of the $\beta$-convergence of $U_{\tilde{d}}$ to $U_d$, i.e., in the sense of the relation

$$\beta(U_{\tilde{d}}, U_d) = \sup_{u \in U_{\tilde{d}}} \rho_U(u, U_d) \to 0, \quad \tilde{d} \to d. \tag{1.1.5}$$

A problem $R$ is said to be $T$-correct if it is possible to specify a nonempty subclass of data $\overline{D} \subset D$ on which the problem is $A$-correct.

Basically, these definitions belong to A. N. Tikhonov [254] and substantially expand the meaning of the concept of correctness of a problem. For example, under the hypotheses of Theorem 1.1.3 the problem (1.1.1), as mentioned above, is $A$-correct in the sense of pseudosolutions if $D = \{f \in F\}$; in general, however, it is not $A$-correct if $D = \{A, f\}$, i.e., under the assumption of variations of the operator (this fact can easily be checked by the example of singular systems of linear algebraic equations).

At the same time, this problem will be $T$-correct if we define the data $\overline{D} = \{A, f\}$ with those operators $A$ that possess the same domain of definition and the same kernel. This fact was first established by V. A. Morozov [164]. The problem of $T$-correctness for (1.1.1) was studied by V. I. Meleshko [147], P. A. Wedin [333], G. M. Vajnikko and A. Ju. Veretennikov [43].

We will also consider the problem of choice for the rule $R$ in the further discussion; it will be shown that sometimes it is easier to solve this problem with approximate data than with exact ones.

Having discussed the concept of problem formulation and its correctness from rather general positions, we turn now to the concept of approximate solutions to a problem, i.e., solutions to a problem $R$ with approximate data $\tilde{d}$. If this problem is $A$-correct, then this concept can be formulated in a simple way: the value $R$ on $\tilde{d}$ (i.e., the set $U_{\tilde{d}} = R(\tilde{d})$) may be taken as an approximate solution. In the case of $T$-correctness, this question becomes nontrivial. Recall, however, what we have done when solving the problem (1.1.1) in a similar case: we have changed the rule of solution choice. This remark leads us to the following definition of a regularizing algorithm.

Let a parameter family of algorithms $R_\alpha$, $\alpha > 0$, be given in such a way that

1) $\forall \alpha > 0$ the algorithm $R_\alpha$ is defined for any $d \in D$;

2) if $U_d = R(d)$ is the exact solution, $\tilde{d}$ are approximate data, and $U^\alpha_{\tilde{d}} = R_\alpha(\tilde{d})$, then there exists a dependence $\alpha = \alpha(\tilde{d}, d) \to 0+$ such that $U^\alpha_{\tilde{d}} \to U_d$ as $\tilde{d} \to d$.

In the case of the above choice of the regularization parameter $\alpha$, the algorithm $R_\alpha$ is said to be regularizing and the corresponding method of construction of the set $U_{\tilde{d}}^{\alpha}$ is called the method of regularization [155, 255, 265].

As can be seen, a distinctive feature of the regularization method is the necessary transition from the rule $R$ to another rule $\tilde{R}$ of definition of solutions with preservation of stability for approximations.

The theory of regularizing algorithms includes

a) methods of their construction;

b) a theory of their estimation and their choice with optimal characteristics;

c) methods of their software implementation.

Our further discussion will follow the above scheme.

**3.** The basic ideas of the regularization theory are considered here under the assumption that the spaces $F$ and $U$ are Banach, whereas the operator $A$ and the set $D_A$ are linear. We say that an operator $\tilde{A}$ (of the same type as $A$) approximates $A$ if

$$\left\|\tilde{A}u - Au\right\|_F \leq h\Omega(u), \tag{1.1.6}$$

where $h \to 0+$ characterizes an accuracy of approximation, the functional $\Omega(u) = \|Lu\|_G$, $L$ is a linear operator defined on $D_L \subseteq U$ and acting into a Banach space $G$. The inequality (1.1.6) is assumed to be valid for $u \in D_A \cap D_L \neq \emptyset$.

Let a nonempty convex set $D \subset D_A \cap D_L$ be given and $f \in A(D)$, so that the set $\overline{U}_f = \{u_f \in D: \ Au_f = f\} \neq \emptyset$. An element $u_f^* \in \overline{U}_f$ such that

$$\Omega(u_f^*) = \Omega^* \equiv \inf_{u \in \overline{U}_f} \Omega(u) \tag{1.1.7}$$

is called the $\Omega$-minimum solution (the $L$-pseudosolution) to equation (1.1.1). Let $U_f^*$ be the set of all $u_f^*$. The mapping $f \to U_f^*$ is said to be $\Omega$-minimal inverse ($L$-pseudoinverse) with respect to $A$ and is denoted by $A_\Omega^-$ (or $A_L^-$). Obviously, its definition domain is $D_{A_\Omega^-} \subset Q = A(D)$. If $A_\Omega^-$ (or $A_L^-$) is a (single-valued) operator, it is denoted by $A_\Omega^+$ (or $A_L^+$).

In what follows, the rule $R$ defining our problem is identified with $A_\Omega^-$. Note that the choice of this rule is completely specified by the approximation condition (1.1.6). This fact is very important for practical applications of regularizing algorithms.

The residual is denoted by $\rho(u) \equiv \|Au - f\|_F$.

**Definition 1** . *A sequence $u_n \in D$ is said to be $\Omega$-regular if*

$$\lim_{n\to\infty} \rho(u_n) = 0, \quad \overline{\lim_{n\to\infty}}\, \Omega(u_n) \leq \Omega^*. \tag{1.1.8}$$

*A sequence $u_n \in D$ is said to be weakly $\Omega$-regular if*

$$\lim_{n\to\infty} \rho(u_n) = 0, \quad \Omega(u_n) \leq K < +\infty, \tag{1.1.9}$$

*where $K$ is a constant.*

**4.** It is not difficult to see that the regularity conditions already define some topology of convergence for the sequence $u_n$. Therefore, it is reasonable to consider some basic ways of construction of such sequences. The most natural of them is the following method.

*The method of $\varepsilon$-optimum residual.* We have ($\|f - \tilde{f}\|_F \le \delta$)

$$\tilde{\rho}(u) \equiv \left\| \tilde{A}u - \tilde{f} \right\|_F \le h\Omega(u) + \delta + \rho(u) \quad \forall u \in D_A \cap D_L;$$

therefore,

$$\tilde{\rho}(u_f) \le h\Omega(u_f) + \delta \quad \forall u_f \in \overline{U}_f.$$

Let us consider the set $U_\sigma = \{u \in D : \tilde{\rho}(u) \le h\Omega(u) + \delta\}$ ($\sigma \equiv (\delta, h)$) of formal approximations of solutions. Obviously, $U_\sigma \ne \emptyset$. As approximate solutions, it is quite reasonable to choose those elements from $U$ that minimize the residual $\tilde{\rho}(u)$ (the classical residual principle). However, in the case of an unstable problem, the direct application of this principle, obviously, is not successful. Moreover, its application does not guarantee even the $\Omega$-boundedness of approximate solutions (for example, for singular systems of linear algebraic equations). Therefore, it is required to formulate a new residual principle. This can be made in a very simple way if we draw our attention to the fact that in the estimate for $\tilde{\rho}(u)$, $u \in U_\sigma$, the functional $\Omega(u)$ is the only free parameter. That is why we require that this parameter should be minimal. This requirement guarantees that the residual $\tilde{\rho}(u)$ is certainly small on the resulting approximations (as in the classical case) and, at least, the functional $\Omega$ is bounded, which may potentially restore the stability of approximations. Such a principle of selection of approximate solutions without any theoretical substantiation has been proposed in [325] and completely investigated in the works of V. K. Ivanov [112], A. N. Tikhonov [259], V. A. Morozov [153], A. G. Yagola and A. V. Goncharsky [78], including the presence of perturbations in $A$ (see also [188, 189]).

Let

$$U_\sigma^\varepsilon = \left\{ u \in U_\sigma : \Omega(u) \le \tilde{\Omega}^* + \varepsilon,\ \tilde{\Omega}^* = \inf_{u \in U_\sigma} \Omega(u),\ \varepsilon \to 0+ \right\}.$$

Then, obviously,

$$\rho(u) \le 2h(\Omega^* + \varepsilon) + 2\delta,\ \Omega(u) \le \Omega^* + \varepsilon \quad \forall u \in U_\sigma^*. \tag{1.1.10}$$

*The method of $\varepsilon$-quasisolutions.* Let $R > \Omega^*$ be given *a priori*; then, the set $M_R = \{u \in D : \Omega(u) \le R\} \ne \emptyset$. Based on the same reasoning as before, we define the approximate solutions as follows:

$$U^{\varepsilon}_{\sigma,R} = \left\{ u \in M_R : \tilde{\rho}(u) \le \tilde{\rho}^* + \varepsilon, \quad \tilde{\rho}^* = \inf_{u \in M_R} \tilde{\rho}(u), \quad \varepsilon \to 0+ \right\}.$$

This selection principle for solutions (the principle of quasisolutions) is almost classical. It has been proposed and substantiated by V. K. Ivanov [110] for a wide class of operator equations and was used in [299] to solve integral equations of the first kind numerically.

We have

$$\rho(u) \le 2hR + 2\delta + \varepsilon, \quad \Omega(u) \le R \quad \forall u \in U^{\varepsilon}_{\sigma,R}. \tag{1.1.11}$$

Note that the method of $\varepsilon$-quasisolutions does not require any specifications of accuracy characteristics for the input data $\delta$.

*The method of $\varepsilon$-regularization.* Let us define the regularizing functional

$$\tilde{\Phi}_\alpha[u] = \tilde{\rho}^2(u) + \alpha \Omega^2(u),$$

where $\alpha > 0$ is a parameter of regularization. Let $\tilde{\Phi}^*_\alpha = \inf_{u \in D} \tilde{\Phi}_\alpha[u]$.

The set of $\varepsilon$-regularized solutions is

$$U^{\varepsilon}_{\sigma,\alpha} = \left\{ u \in D : \tilde{\Phi}_\alpha[u] \le \tilde{\Phi}^*_\alpha + \varepsilon, \quad \varepsilon \to 0+ \right\}.$$

It is easy to show that

$$\begin{aligned} \rho(u) &\le \left( (h\Omega^* + \sigma)^2 + \alpha \Omega^{*2} + \varepsilon \right)^{1/2} + h\Omega(u) + \delta, \\ \Omega(u) &\le \left( \Omega^{*2} + \varepsilon/\alpha + (h\Omega^* + \delta)^2 / \alpha \right)^{1/2}. \end{aligned} \tag{1.1.12}$$

The method of regularization was first proposed and substantiated by A. N. Tikhonov. It has become the subject of further studies by Russian and foreign scientists (see, for example, the survey of V. A. Morozov [167]). Now, this method is one of the most powerful instruments for solving ill-posed problems . From the works devoted to the general problems of regularization, we note the interesting theoretical works of V. A. Vinokurov [50], which also are of interest in practice.

The methods of optimal residual, quasisolutions, and regularization belong to the class of basic variational regularizing algorithms and form a basis for constructing other regularizing algorithms.

The conditions that should be imposed on the tendency of $\sigma \equiv \sigma_n$ and $\varepsilon \equiv \varepsilon_n$ to zero easily follow from (1.1.10)–(1.1.12). These conditions provide the possibility of obtaining (weakly) regular sequences.

**5.** Some assumptions on properties of the operators $A$ and $L$ are necessary to study the convergence of regular sequences. Namely we assume that, along with an original topology, a certain (probably, another) topology $\tau$ is given in $U$ such that from any sequence $u_n \in D : Au_n \xrightarrow{F} v$ (the sequence $Lu_n$ is bounded as $n \to \infty$), it is possible to select the subsequence

$$\begin{aligned} \{u_{n_k}\} &\subseteq \{u_n\} : u_{n_k} \xrightarrow{\tau} u_0 \in D, \\ v &= Au_0, \;\; \Omega(u_0) \leq \varliminf_{k\to\infty} \Omega(u_{n_k}). \end{aligned}$$

This condition is called the condition of $\tau$-complementarity for the operators $A$ and $L$.

**Theorem 1.1.4** *Let the operators $A$ and $L$ be $\tau$-complementary on $D$. Then, $U_f^* \neq \emptyset$ $\forall f \in Q$, i.e., $D_{A_L^-} = Q$. If $G$ is a Hilbert space, then*

$$u_f^* \in U_f^* \Leftrightarrow \left(Lu_f^*, L\left(\omega - u_f^*\right)\right)_G \geq 0 \quad \forall \omega \in \overline{U}_f.$$

*If the spaces $F$ and $G$ are strictly normalized and for any*

$$u_1, u_2 \in D, \quad Au_1 = Au_2, \quad Lu_1 = Lu_2 \Rightarrow u_1 = u_2,$$

*then $A_L^- = A_L^+$, i.e., $U_f^* = \left\{u_f^*\right\}$.*

Some other important properties of $A_L^+$ (in particular, its continuity, i.e., the $T$-correctness of the problem under consideration) have been established by V. A. Morozov [175] (see also [314]).

A sequence $u_n \in D$ is said to be $\tau$-convergent to a set $\overline{U}$ if (perhaps, except for a finite number of its elements) all elements of this sequence are contained in any $\varepsilon$-neighborhood of the set $\overline{U}$ (in the topology $\tau$). This fact can be written down as

$$u_n \xrightarrow{\tau} \overline{U}.$$

**Theorem 1.1.5** *Let $A$ and $L$ be $\tau$-complementary on $D$ and a sequence $u_n \in D$, $n \to \infty$, be given. If $\{u_n\}$ is weakly regular, then*

$$u_n \xrightarrow{\tau} \overline{U}_{f,R} = \left\{u \in \overline{U}_f : \Omega(u) \leq R\right\}, \quad Au_n \to f, \quad n \to \infty, \qquad (1.1.13)$$

*where $R > \Omega^*$ is a constant; if $\{u_n\}$ is regular, then*

$$u_n \xrightarrow{\tau} U_f^*, \quad \Omega(u_n) \to \Omega^*, \quad Au_n \to f, \quad n \to \infty. \qquad (1.1.14)$$

**Corollary.** *Let $U_f^* = \{u_f^*\}$, the spaces $U$ and $G$ satisfy the conditions of Efimov–Stechkin (i.e., they are reflexive and the strong convergence follows from the weak convergence of elements and their norms). Let $\tau$ mean the weak convergence in $U$. If $u_n \in D$ is regular, then from the weak convergence of $u_n$ to $u_f^*$ we have*

$$\left|u_n - u_f^*\right|^2 \equiv \left\|A(u_n - u_f^*)\right\|_F^2 + \left\|L(u_n - u_f^*\right\|_G^2 \to 0, \quad n \to \infty. \qquad (1.1.15)$$

For the above situation, moreover, from the condition of $\tau$-complementarity it follows the existence of a constant $\gamma > 0$ such that

$$\|z - v\|_U \le \gamma\,|z - v| \quad \forall z,\, v \in D, \qquad (1.1.16)$$

i.e., in reality we have $u_n \to u_f^*$ in $U$.

**Remark 2** . *The regularity conditions are also necessary for convergence of the sequence $u_n \in D$ in the sense mentioned above. At the same time, $\overline{U}_f \ne \emptyset$ and $f \in F \Leftrightarrow \exists u_n \in D$ is weakly regular. We assume here that $A$ and $L$ are $\tau$-complementary on $D$.*

**6.** An important field of application for $L$-pseudosolutions is the theory of splines. In this case, the operator $A$ is regarded as $A \equiv A^{(n)} = (l_1, l_2, \ldots, l_n)$, where $l_i = l_i^{(n)}$ are some linear functionals defined on $D_L$. Then we assume that $f = f^{(n)} = (l_1(\hat{u}), l_2(\hat{u}), \ldots, l_n(\hat{u}))$, where $\hat{u} \in D_L$ is some fixed element. Let $u_f^* \equiv s_n$; the element $s_n$ is called the $L$-spline interpolating the element $\hat{u}$. Determination of conditions for which the splines approximate the original element $\hat{u}$ with arbitrary accuracy is a very important problem in the theory of splines.

Let us assume that $\hat{u} \in D$ and $D$ is a convex subset of $D_L$, the operator $L$ is closed on $D$, i.e., for any $u_n \in D,\ u_n \to u,\ Lu_n \to g,\ n \to \infty \Rightarrow u \in D,\ Lu = g$. The original space $U$ is regarded as the space of Efimov–Stechkin with a strictly convex norm, i.e., $U$ is the $E$-space [112].

**Theorem 1.1.6** *Let the above conditions be valid and let there exist $\chi_i = \chi_i^{(n)} > 0$ such that*

$$\left|\|u\|_U^2 - \sum_{i=1}^{n} \chi_i l_i^2\,(u)\right| \le r_n^2\,\|Lu\|_G^2\,, \quad u \in D,$$

*where $r_n \to 0$. Then, for all sufficiently large $n$ there exists a unique interpolating spline $s_n \in D$ and $\|s_n - \hat{u}\|_L \to 0$ as $n \to \infty$, where $\|v\|_L \equiv |v|\ \forall v \in D_L$.*

This theorem allows us to substantiate the convergence of splines under very weak original assumptions and has wide applications in the theory of interpolation of discrete data and for the substantiation of various approximate collocation-type methods.

**7.** The following theorem demonstrates the property of universality for the method of regularization.

**Theorem 1.1.7** *Let a regularized element $\tilde{u}_\alpha$ be uniquely defined from the condition $\tilde{u}_\alpha \in D$, $\tilde{\Phi}^*_\alpha = \tilde{\Phi}_\alpha[\tilde{u}_\alpha]$ for all $\alpha > 0$. If there exists $\alpha : 0 < \alpha < \infty$ for which the equality*

$$\tilde{\rho}(\tilde{u}_\alpha) = h\Omega(\tilde{u}_\alpha) + \delta \tag{1.1.17}$$

*holds, then $\Omega(\tilde{u}_\alpha) = \tilde{\Omega}^*$, i.e., the optimal residual method is solvable; if $\alpha$ satisfies the relation*

$$\Omega(\tilde{u}_\alpha) = R, \tag{1.1.18}$$

*then $\tilde{\rho}(\tilde{u}_\alpha) = \tilde{\rho}^*_R$, i.e., the method of quasisolutions is solvable.*

The conditions for the choice of regularization parameter from the criteria (1.1.17) and (1.1.18) are fundamental in the theory of regularizing algorithms. Their remarkable feature consists in the fact that the regularized solutions are chosen from some limiting relations specified by reasons of accuracy or by *a priori* considerations of a class of feasible solutions. To some extent, this predetermines the fact of optimality of the corresponding algorithms. We will discuss this fact below.

The operators $A$ and $L$ are said to be strongly $\tau$-complementary on $D$ if from any sequence $u_n \in D : \ |u_n| \le K = \text{const}, \ n = 1, 2, \dots,$ it is possible to select a subsequence $\{u_{n_k}\}$ which is $\tau$-convergent to some element $u_0 \in D$ and

$$\begin{aligned} \|Au_0 - g\|_F &\le \varliminf_{k\to\infty} \|Au_{n_k} - g\|_F, \quad g \in F, \\ \Omega(u_0) &\le \varliminf_{k\to\infty} \Omega(u_{n_k}). \end{aligned}$$

**Theorem 1.1.8** *If the operators $\tilde{A}$ and $L$ are strongly $\tau$-complementary on $D$, then the set of regularized solutions $\tilde{U}_\alpha = \{u \in D : \tilde{\Phi}_\alpha[u] = \tilde{\Phi}^*_\alpha\} \ne \emptyset$ and the following relation holds:*

$$\tilde{U}_\beta \xrightarrow{\tau} \tilde{U}_\alpha, \quad \sup_{v\in\tilde{U}_\beta} \inf_{z\in\tilde{U}_\alpha} |z - v| \to 0, \quad \beta \to \alpha.$$

**Remark 3** . *Let $F$, $U$, and $G$ be the spaces of Efimov–Stechkin, $\tilde{A}$ and $L$ be linear operators, the topology $\tau$ be a weak topology in $U$, the operators $A$ and $L$ satisfy the relation* (1.1.16), *and for any sequence $u_n \in D : u_n \to u_0, \ \tilde{A}u_n \to f, \ Lu_n \to g, \ n \to \infty$ we have: $u_0 \in D, \ \tilde{A}u_0 = f, \ Lu_0 = g$. Then the operators $\tilde{A}$ and $L$ are strongly $\tau$-complementary on $D$.*

Further we assume (see also [189]) that the problem of regularization is uniquely solvable, i.e., $\tilde{U}_\alpha = \{\tilde{u}_\alpha\}$. Then the following functions are defined:

$$\tilde{\rho}(\alpha) \equiv \tilde{\rho}(\tilde{u}_\alpha), \quad \tilde{\gamma}(\alpha) \equiv \|L\tilde{u}_\alpha\|_G\,, \; \tilde{\varphi}^2(\alpha) \equiv \tilde{\Phi}^*_\alpha;$$

these functions are called the criterial functions. Let us consider their main properties under the assumption that $F$, $U$, and $G$ are Hilbert spaces, whereas the operators $\tilde{A}$ and $L$ are linear. Let

$$\begin{aligned}
\tilde{\mu} &= \inf_{z\in D} \tilde{\rho}(z), \\
\tilde{\nu} &= \begin{cases} \inf\limits_{z\in \tilde{U}_{\tilde{f}}} \Omega(z) \;\text{ if }\; \tilde{U}_{\tilde{f}} = \left\{ z \in D : \inf\limits_{v\in D} \tilde{\rho}(v) = \tilde{\rho}(z) \right\} \neq \emptyset; \\ +\infty \;\text{ if }\; \tilde{U}_{\tilde{f}} = \emptyset; \end{cases} \\
\tilde{\eta} &= \begin{cases} \inf\limits_{z\in N_0} \tilde{\rho}(z) \;\text{ if }\; N_0 = \left\{ z \in D : \nu_0 \equiv \inf\limits_{v\in D} \Omega(v) = \Omega(z) \right\} \neq \emptyset; \\ +\infty \;\text{ if }\; N_0 = \emptyset. \end{cases}
\end{aligned}$$

**Lemma 1.1.1** *The following limiting relations are valid:*

$$\begin{aligned}
\lim_{\alpha\to 0+} \tilde{\rho}(\alpha) = \lim_{\alpha\to 0+} \sqrt{\tilde{\varphi}(\alpha)} = \tilde{\mu}; \quad \lim_{\alpha\to 0+} \tilde{\varphi}(\alpha) = \tilde{\nu}; \\
\lim_{\alpha\to \infty} \tilde{\rho}(\alpha) = \lim_{\alpha\to \infty} \sqrt{\tilde{\varphi}(\alpha)} = \tilde{\eta}; \quad \lim_{\alpha\to \infty} \tilde{\gamma}(\alpha) = \nu_0.
\end{aligned}$$

**Lemma 1.1.2** *For $\alpha \in (0, +\infty)$, the functions $\rho(\alpha)$ and $\sqrt{\tilde{\varphi}(\alpha)}$ are continuous, monotonically increase, and their values belong to the interval $(\tilde{\mu}, \tilde{\eta})$; the function $\tilde{\gamma}(\alpha)$ is continuous, convex downwards, monotonically decreases, and its values belong to the interval $(\nu_0, \tilde{\nu})$.*

It is easy to see that $\tilde{\mu} < h\Omega^* + \delta$. Assuming that $\theta \in D$, we obtain $\nu_0 = 0,\; N_0 = \{\theta\}$ (these conditions are fulfilled, as a rule, in practical problems). Thus, $\tilde{\eta} = \left\|\tilde{f}\right\|_F$ and for $h,\; \delta \ll \|f\|_F$ the interval $(\tilde{\mu}, \tilde{\eta})$ is, obviously, nontrivial. Similarly, it can be shown that the interval $(0, \tilde{\nu})$ is nontrivial.

The above reasoning allows us to formulate some conditions of solvability for the criteria (1.1.17) and (1.1.18) of the choice of regularization parameter. Note that the function $\tilde{\rho}(\alpha)$ may change its curvature; this fact makes it difficult to construct an efficient numerical algorithm for the choice of regularization parameter. However, putting $\alpha = 1/\lambda$ and defining $\tilde{\rho}_0(\lambda) \equiv \tilde{\rho}(1/\lambda)$, it is easy to establish the downward convexity of $\tilde{\rho}_0(\lambda)$. Together with the condition of differentiability of this function, the above fact allows us to construct fast-convergent Newton-type algorithms [172] similar to those proposed earlier for the construction of smoothing splines [326].

If $\Delta = 2(hR + \delta)^{2(1-s)},\ R > \Omega^*,\ 0 \le s < 1$, then the choice of the regularization parameter $\alpha_\Delta$ from the condition (called the smoothing functional principle [152])

$$\tilde{\varphi}^{1/2}(\alpha) = \Delta \tag{1.1.19}$$

is possible for sufficiently small $h$ and $\delta$ and leads to regular ($s > 0$) or weakly regular solutions even for nonlinear problems [319].

The need for knowledge of a value of $h$ complicates a procedure of the choice of regularization parameter. Let, however,

$$\tilde{\rho}(\tilde{u}) = \sum_{i=1}^{n} \left( l_i(\tilde{u}) - \tilde{f}_i \right)^2,$$

where $\tilde{f}_i = l_i(\tilde{u}),\ \tilde{u} \in U : \|\hat{u} - \tilde{u}\|_U \le \delta$, i.e., the problem of construction of regularized (smoothing) splines $\tilde{s}_{\alpha,n}$ is considered. If the closure condition

$$\|z\|_U^2 = \sum_{i=1}^{\infty} l_i^2(z) \quad \forall z \in U$$

is valid, then the regularization parameter can be chosen from the condition

$$\tilde{\rho}(\tilde{s}_{\alpha,n}) = \delta; \tag{1.1.20}$$

in addition we have $|\tilde{s}_{\alpha,n} - \hat{u}|_L \to 0$ as $n \to \infty,\ \delta \to 0$. An explicit specification of $\tilde{u}$ is optional. All the systems of functionals defined as Fourier coefficients with respect to complete orthonormal systems of elements have the above properties.

Under some conditions, the problem of choosing the regularization parameter becomes not so important. For example, if the operator $A^+$ is defined on all $Q$ and for any sequence $f_n \in Q$ convergent to $f \in Q$ the elements $u^*_{f_n}$ form a regular sequence, then the sequence of regularized solutions $\hat{u}_\alpha$ corresponding to the data $\tilde{A} \equiv A$ and $\tilde{f} = f_n$ is regular for any way of tendency of $\alpha$ to zero: $\alpha \to 0+$. In this scheme, it is even possible to admit perturbations of the operators $A$ but of very special character. The indicated problems are discussed in [158], [164] in more detail.

The following theorem plays an important part in numerical construction of regularized solutions.

**Theorem 1.1.9** *Let the hypotheses of Theorem* 1.1.8 *be valid. Then,* $\tilde{U}_\alpha \subset D^- = A^-(Q)$, *i.e., in a regularized problem the set* $D$ *can be replaced by* $D^-$.

Often, the structure of $D^-$ happens to be well known and substantially simpler than the structure of the original feasible set $D$. This allows us to make easier the search for regularized solutions, especially if $D^-$ has a finite-dimensional structure (as, for example, in the theory of splines). Similar facts take place for the methods of residual and quasisolutions.

## 1.2 Special regularizing algorithms and their properties

**1.** Let us consider some other methods for choosing the regularization parameter and some other ways for constructing the regularizing algorithms.

*Quasioptimal regularization.* This approach assumes that there exists a "sheaf" of regularized solutions and is based on the determination of the quantities

$$\alpha_{\mathrm{qopt}} = \arg\min_{\alpha>0} \left\| \alpha \frac{d\tilde{u}_\alpha}{d\alpha} \right\|.$$

If $\alpha = \alpha_j = \tau\alpha_{j-1}$ and $\tau \approx 1$, then

$$\alpha \frac{d\tilde{u}_\alpha}{d\alpha}\bigg|_{\alpha=\alpha_j} \approx \tau \frac{\tilde{u}_{\alpha_j} - \tilde{u}_{\alpha_{j-1}}}{\tau - 1}.$$

This means that $\alpha_{\mathrm{qopt}} \approx \arg\min_j \|\tilde{u}_{\alpha_j} - \tilde{u}_{\alpha_{j-1}}\|$. Usually, the behavior of $\tilde{u}_\alpha$ is studied on several grids, whereas the choice of $\alpha_{\mathrm{qopt}}$ is governed by the values (or by the norms) of the residual on the regularized solution (the "fork" technique) [156]. This method was proposed by A. N. Tikhonov and V. B. Glasko in [270], was verified in many numerical experiments, and was proved to be reliable and efficient in practice.

The same idea underlies the choice of values of $\tilde{\alpha}_{\mathrm{qopt}} = \arg\min_{\alpha>0} \left\| \alpha \frac{d\tilde{f}_\alpha}{d\alpha} \right\|$, where the norm is stronger than the norm of the space $F$ (for example, if $F = L_2$, then $\|\cdot\| = \|\cdot\|_C$; here, obviously, it is additionally required that the operators $\tilde{A}$ would act in $C$). This algorithm was proposed by V. A. Morozov [184] and has been tested on many model problems by V. V. Badeva.

The theoretical substantiation of the above methods for the choice of regularizing parameter is given in [264].

*Regularization with a control right-hand side.* Let us assume that in addition to $\tilde{f}$ another element $\tilde{f}_1 \in F$ is given (or is modeled on a computer) with the same accuracy. Then, we can choose $\hat{\alpha} = \arg\min \|\tilde{A}\tilde{u}_\alpha - \tilde{f}_1\|$. This method was proposed by I. I. Kochetov [130], was tested on practical problems, and was proved to be efficient.

*Adaptive regularization.* Let the stabilizing functional be represented in the form $\Omega(u - u^*)$. One can easily see that in the case $u^* = u^*_f$ and in the absence of errors ($\sigma = 0$) the regularized solutions satisfy the identity $\tilde{u}_\alpha \equiv u_f^*$ $\forall\alpha > 0$ (i.e., they are stationary), whereas the stability domain of solutions in $\alpha$ is very extensive. It is natural to assume that this property of regularized solutions remains valid for $u^* \cong u_f^*$. However, it is extremely difficult to choose a sufficiently good trial solution *a priori*. Fortunately, there exists a possibility to achieve the desired aim on the basis of the above property of regularized solutions themselves. This can be done by the method of adaptive regularization. Let us call $\tilde{u}_\alpha^{(1)} \equiv \tilde{u}_\alpha$ the primary family of regularized solutions. If we replace $u^*$ by $\tilde{u}_\alpha^{(1)}$ and apply the regularization process once again, then we obtain the secondary family $\tilde{u}_\alpha^{(2)}$ of regularized solutions, etc. If the set $D$ is linear and all the spaces are Hilbert ones, then $\tilde{u}_\alpha^{(2)} = \tilde{u}_\alpha^{(1)} - \alpha\,\dfrac{d\tilde{u}_\alpha^{(1)}}{d\alpha}$. V. A. Morozov proved in [157] that the family $\tilde{u}_\alpha^{(2)}$ defines a regularizing algorithm. A simple discrete expression on the grid $\alpha_j = \tau\alpha_{j-1}$ can be written down for $\tilde{u}_\alpha^{(2)}$.

Note that the choice of quasioptimal values for the regularization parameter is equivalent to the minimization of the second term in the representation of the family $\tilde{u}_\alpha^{(2)}$ or, what is the same, to the minimization of deviation between the secondary family of regularized solutions and the primary one. This method is intended to expand the natural stability domain of regularized solutions.

*Iterative regularization.* If the regularization parameter $\alpha$ is fixed, then the iterative regularization coincides with the adaptive regularization. In this case, we obtain a sequence of elements $\tilde{u}_p \equiv \tilde{u}_\alpha^{(p)}$, $p = 1, 2, \dots$. If the original data for the problem are exact, then $\tilde{u}_p \to u_f^*$ (in the above sense) as $p \to \infty$. If these data are given approximately, then the consistent tendency $p \to \infty$ is required as $\sigma \to 0$. The problem of convergence is considered in [132], [145], [155], [313]. I. V. Emelin and M. A. Krasnoselsky have substantiated the choice of the parameter $p$ in the residual $\tilde{\rho}(\tilde{u}_p)$ and have studied others properties of iterative regularization [96]. The iterative regularization method was generalized to the solution of inconsistent operator equations [151], [146], [67]. It is attractive to consider both the parameters ($\alpha$ and $p$) as regularizing [152]; however, no meaningful results in this direction of research are known to us.

A generalization of this method is known for the case of the continuous parameter $p$ in the form of a stabilization technique for an evolutionary equation [152].

*Descriptive regularization.* This method is one of the modifications of the method of quasisolutions and uses the qualitative *a priori* information on the sought-for solution. As a rule, the subjective basis for this information is an intuitive consideration of the simplicity for the structure of the sought-for solution, whereas the objective basis is *a priori* known general idea on the behavior of the physical process under study [76]. This is connected with the researcher's wish to reproduce the most important and essential features in the mathematical model.

Some of them may be induced by purely aesthetic reasons when selecting the solutions. At an intuitive level, indeed, the behavior of a function is reasonably well represented by a possible arrangement of its characteristic points (the extremum points and the points of change of curvature) and by the domain of its values. It is assumed that on the whole the function behaves naturally (i.e., it is single-valued, has no points of return, and is smooth enough (can be drawn with a stroke of the pen)). Such a class of simple-structure functions is efficiently defined by the specification of domains of its monotonicity and (or) convexity. It is possible to vary not only the arrangement of these domains but their number, too. The minimization of the residual $\tilde{\rho}(u)$ on the so-defined set $D$ is the essence of the method of descriptive regularization. It has been proved that the monotonicity of a function ensures the uniform stability of regularizing solutions [77], whereas the use of information on its curvature ensures the stability of these solutions in $C^1$ (see [239] and Chapter IV of this book).

The numerical testing of this method on a number of model problems has shown the high efficiency and quality of the approximate solutions obtained. The application of special computational tricks substantially reduces the computer time and storage [183], [75]. Several standard subroutines implementing this method were developed at the Research Computing Center of Moscow State University.

The method of descriptive regularization may efficiently be used in interactive mode of operation, i.e., in the "trial-and-error" mode. A new type of a computer-aided "solver" appears; its basic elements are: a computer carrying out the most routine part of work, a display (a communication device) enabling the user to analyze the resulting information and to accept/reject the solutions, and a human operator forming at each step a new model for testing on the basis of data obtained at the previous steps.

The concepts of the method of descriptive regularization are of interest, since, in particular, they can easily be generalized to the case of nonlinear inverse problems.

*The pointwise residual method.* Let $F$ be a functional space and let $A(D) = Q \subseteq C \subseteq F$, $\tilde{f} \in C : |f - \tilde{f}| \leq \delta(x)$, where $\delta(x) \in C$. This method (in its continuous form) consists in finding $\tilde{u}_\epsilon \in D$ such that

$$\Omega(\tilde{u}_\epsilon) \leq \inf_u \Omega(u) + \epsilon, \quad u \in D : |Au - \tilde{f}| \leq \delta(x), \quad \epsilon \to 0+\,. \qquad (1.2.21)$$

If $\epsilon \to 0+$, $\|\delta(x)\|_F \to 0$, then $\tilde{u}_\epsilon$ satisfies the conditions of regularity. The method (in its discrete form) consists in the minimization of the functional $\Omega$ on the set $u \in D : \left| Au|_{x=x_i} - \tilde{f}(x_i) \right| \leq \delta(x_i)$, $i = 1, 2, \ldots, n$, i.e., in the solution of a finite-dimensional problem of mathematical programming. If $Au \equiv u$, then the method can easily be interpreted geometrically: to find a function of the most simple structure ($\Omega \to \min$) falling within the given limits (a given "corridor"). In the case when $\Omega^2(u) = \int\limits_a^b u''^2 dx$, this means that it is necessary to determine a function with the least mean-square curvature. This approach corresponds to the most widespread intuitive notions of "beauty" (acceptability, rationality) of the sought-for approximations. The method was proposed by V. A. Morozov [154] and does not require the complicated programming.

*The method of statistical regularization.* If we set $\alpha = \delta^2/R^2$, $R \geq \Omega^*$ in the Tikhonov regularization method, we come to the minimization of the functional (assuming $h = 0$)

$$\|Au - \tilde{f}\|_F^2/\delta^2 + \|Lu\|_G^2/R^2, \quad u \in D. \tag{1.2.22}$$

It can be shown that a weakly regular approximation is obtained in this way [319].

The method of statistical regularization coincides in form with (1.2.22) if we assume *a priori* that $\tilde{f} - f = \xi \in N(0; \delta^2 E_m)$, $Lu \in N(0; R^2 E_n)$, where $E_m$ and $E_n$ are the identity matrices of order $m$ and $n$, and if we apply the Bayes strategy for the choice of solutions. The idea of this method completely coincides with the above-stated concepts.

The specific feature of the method consists in the fact that a statistically regularized problem is, as a rule, stable in the mathematical sense, i.e., is well posed (for example, if $L = E$). In our opinion, some authors [287] make a fetish of this rather simple fact when declaring that the method of statistical regularization is the only correct and physically justified method and when opposing it to the method of determinate regularization. The groundlessness of this approach is quite obvious, although it has found some followers among the researchers in the field of applied physics, who can be unaware of the essence of the problem.

Serious studies on this method was carried out by M. M. Lavrentjev and V. G. Vasiljev [134], V. Ya. Arsenin [18], E. L. Zhukovskiy [99], [100], A. P. Petrov [220], V. A. Morozov [162], M. V. Muravjeva [203] (see also [304], [330], etc.). The results of [203] are substantially developed in [295] .

*Regularization of variational inequalities.* In the case when $F$ and $G$ are Hilbert spaces, the minimization of $\Phi_\alpha[u]$, $u \in D$, is equivalent to the solution of Euler's regularized variational inequality

$$(\tilde{A}\tilde{u}_\alpha - \tilde{f}, \tilde{A}(\omega - \tilde{u}_\alpha))_F + \alpha(L\tilde{u}_\alpha, L(\omega - \tilde{u}_\alpha))_G \geq 0 \quad \forall \omega \in D.$$

If $D$ is linear, this inequality becomes the Euler identity

$$(\tilde{A}\tilde{u}_\alpha - \tilde{f}, \tilde{A}\omega)_F + \alpha(L\tilde{u}_\alpha, L\omega)_G = 0 \quad \forall\omega \in D,$$

from which (under certain conditions) the Euler equation can easily be derived:

$$\alpha L^* L\tilde{u}_\alpha + \tilde{A}^*\tilde{A}\tilde{u}_\alpha = \tilde{A}^*\tilde{f},$$

where $*$ indicates the operation of conjugation.

The above relations often serve as a source for the construction of regularized finite-dimensional models suitable for software implementations. The method of regularized variational inequalities has been studied by A. B. Bakushinsky [28], [29] and by other researchers. In the book [138] the method is discussed for the case of unstable problems of control.

*Simplified regularization.* If $F = U$, $F$ and $G$ are Hilbert spaces, and $A^* = A \geq 0$, then instead of the functional $\tilde{\rho}^2(z)$ we can consider the functional $(\tilde{A}z, z)_U - 2(z, \tilde{f})_U$; as a result, we come to the method of simplified regularization. The corresponding algorithm was first considered by A. N. Tikhonov [255] and M. M. Lavrentjev [132]. Later on, this algorithm was studied by many authors, basically for the case of consistent ($\mu_f = 0$) equations. V. A. Morozov proposed a modified version of the algorithm for inconsistent equations [172]. The elements

$$\tilde{\omega}_\alpha = \tilde{u}_\alpha + \alpha\frac{d\tilde{u}_\alpha}{d\alpha}, \quad \alpha\tilde{u}_\alpha + \tilde{A}\tilde{u}_\alpha = \tilde{f}$$

are taken as approximate solutions.

The advantage of this regularization algorithm consists in the fact that it does not contain the Gauss symmetrization, which is common in the classical methods of regularization. Due to this circumstance, it becomes possible to obtain a more exact solution to a problem at less cost. The simplified regularization algorithm is also more reasonable for unbounded operators $A$. A comparative analysis of numerical characteristics for the complete and simplified regularization algorithms in the case of inconsistent linear problems has been done by V. A. Morozov and S. F. Gilyazov [190].

In general, the regularization of inconsistent problems has its specific peculiarities. The books of V. A. Morozov [172] and S. F. Gilyazov [68] are almost completely devoted to the fundamental study of these questions.

*Shift regularization.* This method is similar to the simplified regularization applied to the solutions of singular equations, but does not require that the condition $A = A^* > 0$ be satisfied. I. K. Lifanov [140] applied this method to the systems of linear algebraic equations obtained by discretization of singular equations with the discrete vortex method in problems of aerodynamics. The general scheme of regularization by shift and the main results of studies on this method are described in Section 1.4.

*The method of self-regularization.* This method is applied to the problems reduced to the solution of integral equations (1.1.2) with $\delta$-type kernels. When (1.1.2) are approximated, the grid step (which should be agreed with the accuracy of input data) is taken as a regularizing parameter. This method coincides in form with the usual quadrature method for the solution of (1.1.2). The method was developed by A. N. Tikhonov, V. I. Dmitriev, E. V. Zakharov [273] and was repeatedly tested when solving the practical diffraction problems [92]. The applicability of the method (when the right-hand sides and the kernels are given exactly) was proved in [60] for periodic equations (1.1.2) with kernels possessing a logarithmic singularity [119].

*The spline approximation method.* This method was proposed by A. I. Grebennikov for a class of ill-posed problems similar (in the sense of instability) to the differentiation problem. This class of problems includes the integral equations of the first kind of the Abel, Volterra, or Fredholm type with a logarithmic singularity in kernels. It is assumed that the right-hand side is discrete and is given approximately by a set of functionals $l_i(f)$, $i = 1, \dots, n$. In essence, the method consists in the construction of a special smoothing of the right-hand side $s_n$. The smoothing procedure is repeated several times, but does not require the solution of linear systems, since it is based on explicit local approximating splines; hence, its software implementation is very efficient. At the next step, the inversion of the operator $A$ is performed. If $A^{-1}$ is calculated by known formulas, then $A^{-1}s_n$ is taken as a solution. Otherwise, an approximate solution is sought in the form of a spline with unknown coefficients whose values are computed from the collocation conditions for the smoothed right-hand side $s_n$. The regularization takes place without restrictions imposed on the interrelation between the error level $\delta$ and $n$ if a certain dependence of the number of smoothings on $n$ and on $\delta$ exists. This fact is an advantage of the spline approximation method compared to the self-regularization method and to the other methods of stepwise regularization. The above scheme can be applied for the regularization of direct methods (discrete vortices [33], [34], self-regularization, etc.). This method is thoroughly studied in Chapter 3 of this book.

*The method of quasiinversion.* The essence of this method consists in the replacement of the original ill-posed problem by a well-posed problem which is, in some sense, close to the original one. Usually, the solutions to the second problem approximate the sought-for solution, i.e., this method defines a regularizing algorithm. However, the resulting approximations are not always exact enough. This observation has been made as early as in [318] whose author has formulated the problem of stabilization of the method for evolutionary problems. In our opinion, the solution proposed by this author is too complicated. For ill-posed problems, the method "modifies" the basic equation, while the evolutionary property of the problem is preserved. The method was proposed in [136] and has been developed by many authors. We regard the approach suggested by B. G. Karasik [122] as particularly promising. It is based on the idea similar to that on which the method of quasiinversion is based; in this method, however, the regularization consists in the replacement of initial conditions by mixed ones with preservation of the basic equation. This method exhibits good accuracy characteristics and is, in essence, a version of the regularization method [185], [302], [298].

We could discuss here many regularizing algorithms developed for solving special classes of problems, for example, the lexicographic problems of optimization, the game problems, or the linear programming problems [281], but the consideration of these algorithms is beyond the scope of this book.

**2.** The above discussion shows that at our disposal we have a rather large spectrum of regularizing algorithms. How do we choose the best one? This question has no direct answer, since the very concept of the "best" algorithm requires clarification. The natural criteria for the choice of the "best" algorithm are the accuracy characteristics and the amount of computer time and storage needed to carry out the computations. Certain technological factors such as programming peculiarities, modularity of an algorithm, availability of efficient software implementations of these modules (all these factors specify a final quality of performance of the algorithm as a whole) are also of importance.

Let us first analyze the accuracy characteristics of algorithms. Assume that a convex class of feasible solutions $M \subseteq D$ is given in $U$ *a priori.* This class is regarded to be centrally symmetric ($-z \in M$ if $z \in M$). It is required to restore the solution (having the following information on it: $z \in M,\ \tilde{A}z \approx \tilde{f}$) as accurately as possible.

Let $\omega_B(z, v) = \|Bz - Bv\|_V,\ z, v \in D_B \supset D$, where $B$ is a given operator (to be estimated) with its values in the normed space $V$. Often, $B$ is a linear functional or the identity operator.

Let $T$ be any (not necessarily linear) operator $T(\tilde{f}, \tilde{A}, \sigma) = U_\sigma \subset M$ called a restoration operator and let $U_{f,M} = \{z \in M : Az = f,\ f \in N = A[M]\}$. We characterize the accuracy of restoration by the quantity

$$\omega_B(\sigma, M, T) = \sup_{\xi, \delta A} \sup_{v \in U_\sigma} \inf_{z \in U_{f,M}} \omega_B(z, v),$$

$$f \in N, \quad \xi = \tilde{f} - f, \quad \|\xi\|_F \leq \delta.$$

The feasible errors $\delta A = \tilde{A} - A$ are defined here by the approximation condition (1.1.6).

The optimal restoration operator $T_0$ is defined by

$$\omega_B(\sigma, M, T_0) = \inf_T \omega_B(\sigma, M, T) = \omega_B^0(\sigma, M), \tag{1.2.23}$$

while the quasioptimal operator $\tilde{T}_0$ satisfies the relation

$$\omega_B(\sigma, M, \tilde{T}_0) \asymp \omega_B^0(\sigma, M). \tag{1.2.24}$$

The function $\omega_B(\delta; M) \equiv \sup_z \|Bz\|_V, \ z \in M : \|Az\|_F \leq \delta$ is called the estimator. The following remarkable theorem is valid [172], [249], [144], [195].

**Theorem 1.2.10** *For any restoration algorithm $T$ its error satisfies the inequality $\omega_B(\sigma, M, T) \geq \omega_B(\delta; M)$.*

Is this bound achievable (at least, by an order of magnitude)? The answer is positive. Indeed, let $M = M_R = \{z \in D : \Omega(z) \leq R\}$ and $R > \Omega^*$. Let us define the functional

$$\|\tilde{A}z - \tilde{f}\|_F^2/(\delta + hR)^2 + \Omega^2(z)/R^2, \quad z \in D,$$

and assume that $\tilde{z}_\epsilon \in D$ is an $\epsilon$-minimizing element for it. It is easy to show that for $\epsilon \ll 1$, $h \asymp \delta$ the algorithm $\tilde{T}$ (which calculates $\tilde{z}_\epsilon$) is quasioptimal. The same is true for the method of $\epsilon$-quasisolutions for $M = M_R$. Moreover, the following theorem is valid.

**Theorem 1.2.11** *Any weakly regular method defines a quasioptimal algorithm.*

The importance of the estimator $\omega_B(\delta, M)$ follows from the above theorems. The answer to the question as to under what conditions we have $\omega_B(\delta, M) \to 0, \ \delta \to 0$ can be readily obtained from the following theorem.

**Theorem 1.2.12** *Let $\omega_B(\delta; R) \equiv \omega_B(\delta, M_R)$. Then, $\omega_B(\delta; R) \to 0$ as $\delta \to 0$ if and only if for $\forall z_n \in D$ from $\|Az_n\|_F \to 0, \ \Omega(z_n) \leq K = \mathrm{const}$ it follows that $\|Bz_n\|_V \to 0$ as $n \to \infty$.*

If $\omega_B(\delta, R) \to 0$, $\delta \to 0$, then the operator $B$ is said to be estimable. In the case $N_A = \{z \in D : Az = 0\} \neq \{0\}$, the operator $B = E$ is not estimable, since $\omega_B(\delta; R) \equiv R$. This case is realized, for example, for a singular system of linear algebraic equations, which certainly is the most widely-known model of ill-posed problems. However, if we consider the equation $AA^*v = f$ (instead of $Az = f$) and put $B = A^*$, then the operator $B$ becomes estimable. Moreover, as was shown by V. A. Morozov [177], the following estimate is valid:

$$\|u_f^* - \tilde{A}^*\tilde{v}\| \leq O(h + \delta),$$

where $\tilde{v} : \|\tilde{v}\| \approx \inf_v \|v\|,\ v : \|\tilde{A}\tilde{A}^*v - \tilde{f}\|_F \leq 2h\|v\| + \delta.$

We see that here the rank of the matrix $A$ may not necessarily be known, as opposed to the common practice in the classical approach to the solution of singular systems of equations.

The evaluation of the function $\omega_B(\delta; R)$ is not always simple [177], [114]. If we put

$$\hat{\omega}_B(\delta; R) \equiv \sup_z \|Bz\|_V, \quad z \in D : \|Az\|_F^2/\delta^2 + \Omega^2(z)/R^2 \leq 2,$$

then $\omega_B(\delta; R) \leq \hat{\omega}_B(\delta; R) \leq \sqrt{2}\omega_B(\delta; R)$. Sometimes, an explicit expression for $\hat{\omega}_B(\delta; R)$ can be calculated easier, as for $\omega_B\,(\delta, R)$ [116]. There are known some other simplification techniques for the evaluation of the function $\omega_B(\delta; R)$ (see, for example, [177]).

**3.** We now proceed to the discussion on constructing optimal recovery algorithms. The general solution of this problem has not been found yet, therefore we will consider only some special cases.

The case $\sigma = 0$. Let $\delta = h = 0$.

**Theorem 1.2.13** *Let the hypotheses of Theorem* 1.1.4 *be fulfilled and, hence, let $A^-$ be defined for all $f \in A[M_R]$. If $\Omega(z - v) \leq R\ \forall z \in U_f, v \in \overline{U}_f$, then $T_0 = A^-$, i.e. the mapping $A^-$ is optimal.*

It seems that the condition $\sigma = 0$ makes the above optimization problem unattractive for researchers. However, this impression is wrong. For example, under the conditions of Subsection 6 of Section 1.1 the equality $A^- = A^+$ defines the spline interpolation operator $S_n$, which, according to Theorem 1.2.13, is optimal [174] in the above sense. Note that the functionals $l_i(z)$ are fixed here, but it is possible to refine the optimization problem, choosing $l_i$ in the best way. This problem in the one- and two-dimensional cases was considered by A. I. Grebennikov [82], [84] on the basis of the technique for the evaluation of integrals on optimal grid nodes $\{x_i\}$ developed by A. N. Tikhonov and S. S. Gaisaryan [269]. In some cases, as was shown by V. A. Morozov [163], the problem is reduced to the choice of optimal quadrature formulas.

In the case $\sigma \neq 0$, the optimization problem is simplified in a natural way. Namely, let a family $T_\alpha$ of regularizing algorithms be given. It is said that the algorithm $T_{\alpha_{\rm opt}}$ is optimal on this family if

$$\omega_B(\sigma, M; T_{\alpha_{\rm opt}}) = \inf_{\alpha} \omega_B(\sigma, M; T_\alpha).$$

**Theorem 1.2.14** *For the families of regularizing algorithms induced by Tikhonov's regularization method, the inequalities*

$$\omega_B(\delta; R) \leq \inf_{\alpha>0} \omega_B(\delta, M_R; T_\alpha) \leq K\omega_B(\delta; R), \quad K < +\infty,$$

*hold if* $h \asymp \delta$.

The problem of finding an optimal algorithm in the class of Tikhonov's regularizing algorithms at least for $h = 0$ has a positive solution. Namely, let $F = U$ be a Hilbert space, let $L = E$, and let $A = A^*$ be completely continuous, invertible and positive ($A > 0$), $B = A^{1-t}$, $0 \leq t \leq 1$. Let us consider the simplified regularization algorithm (for $h = 0$)

$$\alpha z + Az = \tilde{f}. \tag{1.2.25}$$

The approximations obtained by this algorithm are denoted by $\omega_\alpha$.

**Theorem 1.2.15** *The simplified regularization algorithm is optimal when* $\alpha = \dfrac{\delta}{R}\dfrac{t}{1-t}$; *in addition we have* $\omega_B(\delta; R) = \delta^{1-t}R^t$.

The optimality of Tikhonov's regularizing algorithms was established by V. N. Strakhov for a wider class of problems, but with rather burdensome restrictions [249]. Several regularization algorithms quasioptimal with respect to accuracy were studied in [250].

The problem of determining $\omega_\alpha(t) = B\tilde{\omega}_\alpha$ can easily be reduced to some regularized (under certain initial-boundary conditions) evolutionary problems [302]. Hence, for this class of problems the optimality of regularization algorithms is also shown. In [298], another optimal algorithm for the solution of ill-posed problems of the inverse heat-conduction type is suggested; in our opinion, this algorithm seems to be more complicated for practical realization.

The optimization problem for quadrature formulas with exact input data is studied in S. M. Nikol'sky's book [211]. An optimal algorithm for restoration of functions from discrete approximate data was constructed by A. G. Marchuk and K. Yu. Osipenko [144]. The method developed by them is a generalization of the N. S. Bakhvalov method [32] based on the approach to the optimization of problems with exact discrete data suggested by S. A. Smolyak [245].

The optimality of Tikhonov's regularizing algorithms for $h = 0$ was proved in [316] for the case when all the considered spaces are Hilbert ones. For the case $h > 0$, the solution of optimization problem (1.2.23) is unknown even in special cases. However, V. A. Morozov and A. I. Grebennikov [195] have succeeded in proving the optimality of Tikhonov's regularizing algorithms for $h > 0$ on a class of nonlinear feasible algorithms $T$ (compare with [327], [329]) on the basis of another definition of proximity measure (using a "relative" accuracy) for the values of the operator $B$. Some local-optimization problems for regularizing algorithms are also considered in [181], [195].

Obviously, the problem of the choice of an optimal regularization algorithm can be treated as the problem of optimal approximation of the operator $B$. Along with the above-considered approach (1.2.23), (1.2.24), the problem of the best approximation of $B$ by linear bounded operators from some *a priori* given class (stated by S. B. Stechkin [247]) became widely known.

Connections between these approaches and other points of interest are discussed in the paper of V. V. Arestov [6]. Quasioptimal regularization algorithms for the problems of calculation of values of linear operators were also studied by V. V. Vasin [149] and by V. P. Tanana [215]; the same problems for nonlinear operators were studied by A. I. Grebennikov [83]. In this work, some regularization algorithms optimal in a certain sense are also constructed (see [181], [317], [332]). The algorithms of optimal image recovery have been considered by Yu. P. Pytjev [229].

Other approaches to the problem of optimization of algorithms for the solution of unstable problems are also known. Let $U_\sigma$ be, as before, a set of formal approximate solutions, i.e., $U_\sigma = \{z \in D : \tilde{\rho}(z) \le h\Omega(z) + \delta\}$. Then, the set $U_{\sigma,M} = U_\sigma \bigcap M$ contains all the solutions which are feasible from the standpoint of *a priori* input information at hand. Sometimes $U_{\sigma,M}$ is called the uncertainty set of a solution. It is natural to choose one of the remarkable points of the set $U_{\sigma,M}$ (namely, its Chebyshev center) as an "optimal solution". Assume that the problem of determining the Chebyshev center is solvable. Then, for $h = 0, M = M_R$ the method of quasisolutions or the method of (optimal) residual (depending on whether $R$ or $\delta$ is given) is optimal in the sense of (1.2.23). The pointwise residual method also has this property. The problem of optimization of the solutions to operator equations was considered in the above formulation by V. V. Ivanov [108] and A. G. Marchuk [143] (see also [6]). Sometimes, the uncertainty set of a solution is given in the form $z \in D : \tilde{\Phi}_\alpha[z] \le \Delta^2$ (it is assumed that this set is not empty). Under the above conditions for solvability of a regularized problem, the Chebyshev center of this set is the regularized solution $\tilde{z}_\alpha$. If a value of $\Delta$ is given *a priori*, then the regularization parameter is chosen, obviously, according to the smoothing functional principle (1.1.19). This fact of optimality of regularized solutions for interpolating and smoothing splines was established by M. Golomb and

Y. F. Weinberger [307], P. M. Anselone and J. Loran [294]. Note that in this approach the "reasonable" choice of $\Delta$ should be made in the form $\Delta \approx \delta + h$. The assignment of $R$ is optional.

All the above problems are to be considered as containing random errors. The locally optimal algorithms of regularization with noisy right-hand sides of equation (1.1.1) have been considered by V. Ya. Arsenin [18] and V. A. Morozov [162]. It is assumed that $\tilde{f} = f + \xi$, where $\xi$ is a generalized stochastic process. Then, the regularization parameter is chosen from the condition for minimum of the mathematical expectation $\|\tilde{u}_\alpha - u_f^*\|^2$. As a result, we come to the algorithms of the ridge-estimation type. Naturally, the optimum solution depends on $u_f^*$, which should be considered as an unfavorable factor. Various realizations of the algorithms optimal in this sense are suggested in [162], [20]. Several statistical procedures for the choice of regularization parameter (not based on the quantitative information on the accuracy of input data) are developed in [331] (see also [282], [305]).

**4.** We assume that the approximation condition (1.1.6) is valid. However, the choice of $\tilde{A}$ substantially influences both the accuracy and a numerical algorithm for the solution of a regularized problem. It follows from the theoretical results that if the operator $\tilde{A}$ is of a finite rank (in this case, we put $\tilde{A} = A_n$, assuming that the rank of $A_n$ is less than or equal to $n$), then the regularized solutions are finite-dimensional. Hence, in this case we do not need to follow those authors who consider an additional approximation of $\tilde{A}$ as necessary. Obviously, the amount of computations needed to solve the problem depends on $n$. Then, it is natural to choose approximations such that $h = h_n$ be as small as possible for a given $n$.

Let $U$ and $F$ be Hilbert spaces and let the operator $A$ (which is supposed to be, for simplicity, invertible) be completely continuous. Let $\{\varphi_i\}$ ($\{\psi_i\}$) be the eigenfunctions of the operator $A^*A$ ($AA^*$) orthonormal in $U$ ($F$) and let $\lambda_i > 0$ be the common eigenvalues of both these operators enumerated with regard to their multiplicity in decreasing order: $\lambda_{i+1} \leq \lambda_i, \; \lambda_i \to 0+$.

**Theorem 1.2.16** *The following relation is valid:*

$$\inf_{A_n:\ rank\ A_n \leq n} \|A - A_n\| = \|A - \overline{A}_n\| = \lambda_{n+1}.$$

*Here* $\overline{A}_n z = \sum_{i=1}^{n} \lambda_i z_i \psi_i, \; z_i \equiv (z, \varphi_i)_U$, *is the partial sum of the Schmidt series expansion.*

So, if we put $\tilde{A} = \overline{A}_n$, then $h_n \equiv \lambda_{n+1} \to 0+$. The finite-dimensional analogue of this theorem is widely used in the processing of noisy images.

Theorem 1.2.16 has been established by D. E. Allakhverdiev in [5]. In [182], the attention is drawn to its importance for solving ill-posed problems.

One more way of constructing the optimal approximations consists in application of splines. The spline $s_n$ is uniquely defined by a finite set of values of the functionals $u_i = l_i(\hat{u})$; the values of the operator $A$ on $s_n$ (for such a way of defining the information about the element $\hat{u}$) are calculated in the best (in the sense of Theorem 1.2.13) way. If the condition (1.1.17) holds, then (assuming $A_n \equiv AS_n$, where $S_n$ is a spline interpolation operator) we come to an estimate of the type (1.1.6) with $h = h_n \approx r_n$, i.e., this estimate is an order of magnitude optimal. Since the operator $A_n$ is obviously of finite rank, Theorem 1.1.9 is applicable. The formulation of the above-mentioned problems is discussed in greater detail in [174] and [182] (see also [188], [84]).

**5.** Now we consider some general aspects of software implementation of regularizing algorithms. A significant experience has been already accumulated in this direction of research; it is impossible to completely consider this experience in the form of a brief outline. Note only that everyone who starts the development of any numerical algorithm must take into account the specific features of the problem under consideration, since these features substantially influence the performance of an algorithm.

The experience in the numerical solution of unstable problems has revealed, already at the first stage, some computing peculiarities specific for regularizing algorithms. The main aim of a numerical experiment at that stage of development of regularization methods was to prove that ill-posed (unstable) problems with approximate data can be solved numerically. First of all, these specific peculiarities consist in the parametric character of the solution; hence, the necessity arises to solve the problems of the same type repeatedly. Some users became therefore skeptical with respect to the possibilities of computerizing this class of problems. However, it is known that the necessity is the mother of invention. The question has arisen whether it is possible to use these specific peculiarities of the regularization method for simplification of computing processes. And a positive answer was found. Namely, it has appeared reasonable to perform algebraic transformations of original data with the purpose of reducing the original problem to a simpler one, not requiring substantial computing cost. V. V. Voevodin [54] was the first who drew the attention to this possibility in the regularization method. This approach was used as a basis for a number of subroutines included into the Numerical Analysis Library in FORTRAN developed at the Research Computing Center of Moscow State University [56]. The use of these procedures requires no more than $O(N^3)$ arithmetic operations and $O(N^2)$ words of storage, where $N$ is the algebraic dimension of a problem. We see that these parameters correspond to those for the solution of stable algebraic problems. The above approach was further developed in [169], [179], etc.

Let a regularized solution $\tilde{u}_\alpha$ be given (not necessarily precisely). Suppose it is required to determine $\tilde{u}_{\tau\alpha}$, $0 < \tau \leq 1$. It is known [172] that the regularized solutions are uniformly close. Hence, choosing some minimization method $m$ and taking $\tilde{u}_\alpha$ as an initial approximation, we can obtain $\tilde{u}_{\tau\alpha}$ according to the rule $\tilde{u}_{\tau\alpha} = m(\tilde{u}_\alpha)$ with an insignificant number of additional operations to refine $\tilde{u}_\alpha$. This method is especially efficient for nonlinear problems (with the possible linearization of the problem in a neighborhood of $\tilde{u}_\alpha$) and for the problems where the operator is given algorithmically. This is typical for the inverse problems expressed in terms of partial differential equations; the main aim of solving such problems is the determination (identification) of unknown boundary (initial) regimes or the coefficients of equations when some additional characteristics of the sought-for solution are known.

The problems of nuclear spectroscopy, gravimetry, analytic continuation, electrodynamics, statistics, processing of images, etc. are often reduced to the solution of convolution-type Fredholm integral equations of the first kind (i.e., with kernels dependent on the difference of arguments) on infinite intervals. Here the algorithms of fast (discrete) Fourier transform are promising, since they allow one to solve a problem with no more than $O(N \log N)$ operations and $O(N)$ words of storage, i.e., they are sufficiently efficient [12], [76] in combination with a computer-aided choice of regularization parameter on the basis of various criteria.

In the automatic control and regulation, the main problem is the stable numerical solution of Wiener–Hopf type equations. Special regularizing algorithms developed for this class of problems allow one to solve them with no more than $O(N \log N)$ operations and $O(N)$ words of storage [184], [199].

The application of the descriptive regularization method requires the development of efficient numerical algorithms for projection on specific sets. Further optimization of the method is possible due to the choice of economic ways of calculation of the gradient for the minimized functional and due to the choice of the conjugate gradient projection method as a basic one. This choice should be done because of the well-known property of finiteness of the conjugate gradient method for singular problems, so that its application in the descriptive regularization method requires (this was confirmed by computing experiments) about $r \approx 10$ iterations, which coincides, as a rule, with the practical rank of a problem. These and other features of the method are considered in [183] and in Chapter 4 of this book.

In recent years, an increased interest is observed in numerical algorithms for solving the inverse heat-conduction problem. This is explained by new applications of this model describing various physical processes in crystallography, atmosphere, physics, in fixing the images, etc. General algorithms applied to the solution of the above problem, especially in the two-dimensional case, come across substantial difficulties, which requires the development of special regularizing algorithms. Such developments are presented, for example, in [185], [302], [298].

In summary, one may come to the conclusion that the solutions to ill-posed problems can be obtained with the same efficiency as the solutions to similar well-posed problems. This conclusion is confirmed both by the profound theoretical studies and by computing practice, not leaving a possibility to be skeptical in any form as to the practical importance of regularization methods. We paid most of our attention to linear models of ill-posed problems. This is explained not only by the fact that the more profound results have been obtained for the linear models than for the nonlinear ill-posed problems, but also by the A. N. Tikhonov result [256], [257] from which it follows that the basic techniques of regularization of linear ill-posed problems can be extended without essential changes to the regularization of nonlinear ill-posed problems. In the sequel, these results were developed both for the solution of nonlinear operator equations [319] and for the general problem of minimization of functionals [267], [268]. An additional material on this topic can be found in [148]–[150], [62].

## 1.3 On the problem of correction of linear inequalities

**1.** Many applied problems can be reduced to searching for an arbitrary solution or a solution with given properties for the following linear functional inequality:

$$\overline{f} - \overline{A}u \in K, \quad u \in D. \tag{1.3.26}$$

Here $u$ and $\overline{f}$ are elements of separable Hilbert spaces $U$ and $F$, $K$ is a closed convex cone, $D$ is a closed convex set of *a priori* constraints, $0 \in D$. The operator $\overline{A} : U \to F$ is an operator from the Hilbert space $\hat{U}$ of linear bounded operators with the absolute norm

$$\|\overline{A}\| = N(\overline{A}) = \left[ \sum_{i,j} \left|(\psi_i,\ \overline{A}\varphi_j)\right|^2 \right]^{1/2},$$

where $\{\psi_i\}$ and $\{\varphi_j\}$ are arbitrary orthogonal bases of the spaces $F$ and $U$, i.e., $\overline{A}$ is a kernel operator. We will assume that the inequality (1.3.26) is solvable in $u$.

Let us consider the complexity functional $\Omega(u) \equiv \|u - u^*\|$, where $u^*$ is a given element. This functional is included in the formulation of the problem. Without loss of generality, it is assumed below that $u^* = 0$.

**Definition 2** *An element $u$ minimizing the functional $\Omega(u)$ on the set of all solutions to the inequality* (1.3.26) *is said to be the normal solution of this inequality.*

The existence and uniqueness of the normal solution follow from the convexity of the set of solutions to the inequality (1.3.26) and its property of being closed.

**2.** Let the inequality

$$\widetilde{f} - \widetilde{A}u \in K, \quad u \in D, \tag{1.3.27}$$

with approximate original data be given instead of inequality (1.3.26). Here $\widetilde{f} \in \widetilde{F}$, $\widetilde{A} \in \hat{U}$ is a linear bounded operator, and $\widetilde{\sum} = (\widetilde{F}, \hat{U})$ is a class of linear inequalities equivalent with respect to accuracy (the inequalities (1.3.26), (1.3.27) belong to this class).

In general, the inequality (1.3.27) may have properties significantly different from the properties of (1.3.26), for example, it may be unsolvable. Thus, it is necessary to correct the inequality (1.3.27), i.e., to construct (on the basis of given data $(\widetilde{f}, \widetilde{A})$) a pair $(f, A) \in \widetilde{\sum}$ such that the inequality corresponding to it has given properties. In this book, we consider a correction method; this method is a natural generalization of the approach proposed in [260], [262]. A number of other ways of correction are studied in [97].

**3.** In what follows it is assumed that a vector $\sigma = (\delta, h)$ of error levels in the original data of inequality (1.3.27) is given, i.e.,

$$\left\|\widetilde{f} - \overline{f}\right\| \leq \delta, \quad \left\|\widetilde{A} - \overline{A}\right\| \leq h, \quad \delta > 0, \quad h > 0.$$

Then, the class $\widetilde{\sum}$ may be described as a set of inequalities $(f, A)$ equivalent with respect to accuracy:

$$f - Au \in K, \quad u \in D, \tag{1.3.28}$$

where

$$f \in \widetilde{F} = \left\{f \in F : \left\|f - \widetilde{f}\right\| \leq \delta\right\}, \ A \in \hat{U} = \left\{A : \left\|\widetilde{A} - A\right\| \leq h\right\}.$$

Here $A$ is a linear bounded operator.

**Definition 3** *An element $u_\sigma \in D$ is said to be a stable approximation to the normal solution $\overline{u}$ of* (1.3.26) *if*

$$\|u_\sigma - \overline{u}\| \to 0 \quad as \quad \sigma \to 0.$$

Now we formulate the main problem being solved in this section.

*The correction problem.* It is required to construct a solvable inequality of the type (1.3.26) with a defining pair $(f_\sigma, A_\sigma) \in \widetilde{\sum}$ and its normal solution $u_\sigma$ in such a way that the element $u_\sigma$ be a stable approximation to $\overline{u}$.

**Definition 4** *An element $u \in D$ is said to be a feasible solution of the class $\widetilde{\sum}$ if there is a defining pair $(f, A) \in \widetilde{\sum}$ such that $f - Au \in K$.*

Let $\widetilde{Z}$ be the set of all feasible solutions of the class $\widetilde{\sum}$. It is easy to see that the set $\widetilde{Z}$ may be introduced as follows:

$$\widetilde{Z} = \left\{ u \in D : \rho(f - Au, K) = 0,\ (f, A) \in \widetilde{\sum} \right\},$$

where $\rho(f - Au, K) = \|f - Au - P_K(f - Au)\|$ and $P_K$ is an operator of projection onto the set $K$.

**Definition 5** *An element minimizing the complexity functional $\Omega(u)$ on the set $\widetilde{Z}$ is called the normal solution of the class $\widetilde{\sum}$.*

According to Definition 5, the problem of finding the normal solution of the class $\widetilde{\sum}$ is formulated as the following variational problem.

*Problem* 1. To find an element $z_\sigma \in \widetilde{Z}$ such that $\|z_\sigma\| = \widetilde{\eta} = \inf \|z\|$ with $z \in \widetilde{Z}$.

**Lemma 1.3.3** *The existence of a feasible solution $z \in D$ to the problem $f - Az \in K$, $(f, A) \in \widetilde{\sum}$ is equivalent to the solvability of the inequality $\rho(\widetilde{f} - Az, K) \le \delta$, $A \in \hat{U}$.*

**Proof.** Let $z \in D$ be a feasible solution to the problem (1.3.28), i.e., there is a pair $(\hat{f}, \hat{A})$ such that $\rho(\hat{f} - \hat{A}z, K) = 0$. Then we have

$$\rho(\widetilde{f} - \widetilde{A}z, K) \le \left\| \widetilde{f} - \widetilde{A}z - P_K(\hat{f} - \hat{A}z) \right\| \le \delta,$$

where $\hat{A} \in \hat{U}$.

Now we assume that for a given $z \in D$ the inequality

$$\rho(\widetilde{f} - Az, K) \le \delta, \quad A \in \hat{U},$$

holds and that $\hat{A} \in \hat{U}$ is such that $\rho(\widetilde{f} - \hat{A}z, K) \leq \delta$. Denoting

$$C = \hat{A}z + K = \left\{ y \in F : y = \hat{A}z + k,\ k \in K \right\},$$

we come to the relation

$$\rho(\widetilde{f} - \hat{A}z, K) = \rho(\widetilde{f}, C) \leq \delta.$$

Since the set $C$ is closed and convex, the element $\hat{f} = P_C \widetilde{f}$ is uniquely defined. Then, $\left\| \hat{f} - \widetilde{f} \right\| \leq \delta$ and $\hat{f} - \hat{A}z \in K$ by definition of the element $\hat{f}$.

It follows from Lemma 1.3.3 that Problem 1 is equivalent to the following problem.

*Problem* 2. To find an element $z_\sigma \in \widetilde{Z}$ such that

$$\|z_\sigma\| = \tilde{\eta} = \inf \|z\|,\ z \in \widetilde{Z} = \left\{ z \in D : \rho(\widetilde{f} - Az, K) \leq \delta,\ A \in \hat{U} \right\}.$$

**4.** Let us consider the functional inequality with respect to the operator $A$:

$$f - Au \in K \tag{1.3.29}$$

with given $u \in U,\ u \neq 0,\ f \in F$.

**Lemma 1.3.4** *Among the operators with finite absolute norm, there exists a unique operator $A$ with minimal norm. This operator is a solution to the inequality* (1.3.29) *and has the form*

$$\widehat{A} = \frac{(f - \widehat{r})(u, \cdot)}{\|u\|^2},$$

*or*

$$\widehat{a}_{ij} = (\psi_i, \widehat{A}\varphi_j) = \frac{(\psi_i, f - \widehat{r})(u, \varphi_j)}{\|u\|^2}, \quad \left\| \widehat{A} \right\| = \frac{\|f - \widehat{r}\|}{\|u\|}, \tag{1.3.30}$$

*where $\widehat{r} = P_K f$.*

**Proof.** First we note that the operator $A$ defined by (1.3.30) satisfies the inequality (1.3.29). In addition, it is easy to show that the set of solutions to (1.3.29) in the Hilbert space of operators with finite absolute norm is closed and convex. Thus, there is a unique operator with minimal norm which is a the solution to (1.3.29). Now we show that the sought-for operator is expressed by (1.3.30).

Let $\widehat{A}$ be a solution to (1.3.29) and be minimal with respect to norm. Then, $f - \widetilde{A}u = \widehat{r}$, where $\widehat{r} \in K$. Let us consider the functional equation with respect to the operator $A$

$$f - Au = \widehat{r}. \tag{1.3.31}$$

It follows from [262] that equation (1.3.31) has a unique solution $\widehat{A}_1$ with minimal norm. It can easily be seen that $\widehat{A}_1$ coincides with $\widehat{A}$. It is shown in [262] that the operator $\widehat{A}_1$ is given by the formula

$$\widehat{A}_1 = \frac{(f-\widehat{r})(u,\cdot)}{\|u\|^2}, \qquad \left\|\widehat{A}_1\right\| = \frac{\|f-\widehat{r}\|}{\|u\|}.$$

Taking into account that

$$a(\widehat{r}) = \left\|\widehat{A}_1\right\| = \left\|\widehat{A}\right\| = \inf a(r) = \frac{\|f-\widehat{r}\|}{\|u\|}, \quad r \in K,$$

we obtain $\widehat{r} = P_K f$, which proves the lemma.

In what follows we assume that the normal solution of the class $\widetilde{\sum}$ differs from zero of the space $U$. This assumption is certainly valid under the condition

$$\rho(\widetilde{f}, K) > \delta. \tag{1.3.32}$$

**Lemma 1.3.5** *If the normal solution $z_\sigma$ of the class $\widetilde{\sum}$ exists and if the pair $(f_\sigma, A_\sigma)$ realizes it, then $\left\|A_\sigma - \widetilde{A}\right\| = h$ and $\rho(\widetilde{f} - Az_\sigma, K) = \delta$, $\left\|f_\sigma - \widetilde{f}\right\| = \delta$.*

The proof of this lemma can be carried out in the same way as in [262], [263]. It is necessary only to take into account the continuity of the functional $\widetilde{\varphi}(z) \equiv \rho(\widetilde{f} - A_\sigma z, K)$ and the basic properties of the set $D$.

It follows from Lemma 1.3.5 that the problem of determining the normal solution of the class $\widetilde{\sum}$ is reduced to the following problem.

*Problem* 3. To find an element $z_\sigma \in \widetilde{Z}_1$ such that

$$\begin{aligned} \|z_\sigma\| = \widetilde{\eta} = \inf \|z\| \quad \text{for} \quad z \in \widetilde{Z}_1 = \\ = \left\{ z \in D : \rho(\widetilde{f} - Az, K) = \delta,\ \left\|A - \widetilde{A}\right\| = h \right\}. \end{aligned}$$

**5.** Let us now formulate the following auxiliary variational problem.

*Problem* 4. To find an operator $\widehat{A}$ realizing

$$\inf \rho(\widetilde{f} - Az, K), \quad A \in \widehat{U},$$

where $\widetilde{f}$ and $z$ are given.

**Lemma 1.3.6** *Let elements $z \in U$, $z \neq 0$, $y \in F$, $y \neq 0$ and a class of linear operators $S = \{\Delta : \|\Delta\| \leq h\}$ be given. There is a unique operator $\overline{\Delta} \in S$ for which the inequality*

$$\alpha y - \overline{\Delta} z \in K$$

*holds for some maximal value of the parameter $\alpha$. This operator is given by the formula*

$$\overline{\Delta} = h\frac{(y - P_K y)(z,\cdot)}{\|y - P_K y\|\,\|z\|}, \quad \|\overline{\Delta}\| = h, \quad \alpha = h\frac{\|z\|}{\|y - P_K y\|}.$$

The proof is carried out in a similar way as in [260], [262].

Let $d(z) \equiv \widetilde{f} - Az - P_K(\widetilde{f} - Az)$ and $\widetilde{d}(z) \equiv \widetilde{f} - \widetilde{A}z - P_K(\widetilde{f} - \widetilde{A}z)$. In addition, we put $\Delta = A - \widetilde{A}$, $\widetilde{\nu}(z) \equiv \left\|\widetilde{d}(z)\right\| - h\,\|z\| = \rho(\widetilde{f} - \widetilde{A}z, K) - h\,\|z\|$.

**Theorem 1.3.17** *If $\widetilde{\nu}(z) \le 0$ or $\widetilde{d}(z) = \theta h\,\|z\|$, $0 \le \theta \le 1$, then the inequality*

$$\widetilde{f} - Az \in K, \quad A \in \hat{U}, \tag{1.3.33}$$

*is solvable and its solution $\widehat{A}$ of least deviation from $\widetilde{A}$ in norm is equal to $\widehat{A} = \widetilde{A} + \widehat{\Delta}$, where*

$$\widehat{\Delta} = \frac{\widetilde{d}(z)(z,\cdot)}{\|z\|^2} = \theta h\frac{\widetilde{d}(z)(z,\cdot)}{\left\|\widetilde{d}(z)\right\|\,\|z\|}, \quad \left\|\widehat{\Delta}\right\| = \theta h.$$

**Proof**. Reducing the inequality (1.3.33) to the form $\widetilde{f} - \widetilde{A}z = \Delta z \in K$ and applying Lemma 1.3.4, we come to the conclusion of the theorem.

**Corollary**. *The normal solution $z_\sigma$ of the class $\widetilde{\sum}$ satisfies the inequality $\widetilde{\nu}(z_\sigma) > 0$.*

Indeed, if $\widetilde{\nu}(z_\sigma) \le 0$, then $z_\sigma$ satisfies the inequality $\widetilde{f} - \widehat{A}z_\sigma \in K$, $\widehat{A} \in \hat{U}$, but for the normal solution, according to Lemma 1.3.5, $\rho(\widetilde{f} - \widehat{A}z_\sigma, K) = \delta > 0$.

For Problem 4 we have

$$\begin{aligned}\|d(z)\| &= \left\|\widetilde{f} - \widetilde{A}z - \Delta z - P_K(\widetilde{f} - Az)\right\| \ge \\ &\ge \left\|\widetilde{f} - \widetilde{A}z - P_K(\widetilde{f} - \widetilde{A}z)\right\| - \|\Delta z\| \ge \left\|\widetilde{d}(z)\right\| - h\,\|z\|\end{aligned}$$

for $A \in \widehat{U}$. Thus,

$$\left\|\widetilde{d}(z)\right\| - h\,\|z\| \le \inf\|d(z)\|\,, \quad A \in \hat{U}. \tag{1.3.34}$$

**Theorem 1.3.18** *If $\widetilde{\nu}(z) > 0$, i.e., $\left\|\widetilde{d}(z)\right\| > h\,\|z\|$, then the infimum in Problem 4 is realized for the only operator $\widehat{A} = \widetilde{A} + \widehat{\Delta}$, where the operator $\widehat{\Delta}$ is defined in Lemma 1.3.6 for $y = \widetilde{f} - \widetilde{A}z$, i.e.,*

$$\widehat{\Delta} = h\,\frac{\widetilde{d}(z)(z,\cdot)}{\left\|\widetilde{d}(z)\right\|\,\|z\|}, \quad \left\|\widehat{\Delta}\right\| = h,$$

*and the following equality holds:*

$$\|d(z)\| = \rho(\widetilde{f} - \widehat{A}z, K) = \left\|\widetilde{d}(z)\right\| - h\,\|z\|\,. \tag{1.3.35}$$

**Proof.** Let us show that the operator $\widehat{A}$ satisfies the equality (1.3.35), i.e., $\widehat{A}$ is a solution to Problem 4. Taking Lemma 1.3.6 into account, we obtain

$$\begin{gathered}
\rho(\widetilde{f} - \widehat{A}z, K) = \left\|\widetilde{f} - \widehat{A}z - P_K(\widetilde{f} - \widehat{A}z)\right\| \leq \\
\leq \left\|\widetilde{f} - \widehat{A}z - P_K(\alpha(\widetilde{f} - \widetilde{A}z) - \widehat{\Delta}z) - (1-\alpha)P_K(\widetilde{f} - \widetilde{A}z)\right\| \leq \\
\leq \left\|\alpha(\widetilde{f} - \widetilde{A}z) - \widehat{\Delta}z - P_K(\alpha(\widetilde{f} - \widetilde{A}z) - \widehat{\Delta}z)\right\| + \\
+ (1-\alpha)\left\|\widetilde{f} - \widetilde{A}z - P_K(\widetilde{f} - \widetilde{A}z)\right\| = (1-\alpha)\left\|\widetilde{d}(z)\right\|,
\end{gathered}$$

where $\alpha$ is specified in Lemma 1.3.6. Hence,

$$\rho(\widetilde{f} - \widehat{A}z, K) \leq (1-\alpha)\left\|\widetilde{d}(z)\right\| = \left\|\widetilde{d}(z)\right\| - h\,\|z\|\,.$$

Taking the relation (1.3.34) into account, we come to the equality (1.3.35); hence, $\widehat{A}$ is a solution to Problem 4.

Now we show that Problem 4 has a unique solution. Note first that if the operator $A_1 = \widetilde{A} + \Delta_1$ is a solution to Problem 4, then $\|\Delta_1\| = h$. Indeed, it is easy to see that the following equalities hold:

$$\begin{aligned}
\rho(\widetilde{f} - \widehat{A}z, K) &= \left\|\widetilde{d}(z)\right\| - \left\|\widehat{\Delta}z\right\| = \left\|\widetilde{d}(z)\right\| - h\,\|z\|\,, \\
\rho(\widetilde{f} - A_1 z, K) &= \left\|\widetilde{d}(z)\right\| - \|\Delta_1 z\| = \left\|\widetilde{d}(z)\right\| - h\,\|z\|\,.
\end{aligned}$$

It follows from here that $\left\|\widehat{\Delta}z\right\| = \|\Delta_1 z\| = h\,\|z\|$ and $\|\Delta_1\| = h$. Let us assume that $A_1 \neq \hat{A}$. The operator $A_2 = \frac{1}{2}(A_1 + \widehat{A})$ is also a solution to Problem 4. Indeed, we have

$$\begin{aligned}
\rho(\widetilde{f} - A_2 z, K) &= \left\|\frac{1}{2}(\widetilde{f} - A_1 z) + \frac{1}{2}(\widetilde{f} - \widehat{A}z) - P_K(\widetilde{f} - A_2 z)\right\| \leq \\
&\leq \left\|\frac{1}{2}(\widetilde{f} - A_1 z - P_K(\widetilde{f} - A_1 z)) + \frac{1}{2}(\widetilde{f} - \widehat{A}z - P_K(\widetilde{f} - \widehat{A}z))\right\| \\
&\leq \frac{1}{2}\rho(\widetilde{f} - A_1 z, K) + \frac{1}{2}\rho(\widetilde{f} - \widehat{A}z, K) = \rho(\widetilde{f} - \widehat{A}z, K).
\end{aligned}$$

Hence, $\rho(\widetilde{f} - A_2 z, K) = \rho(\widetilde{f} - \widehat{A}z, K)$. Thus, for $\Delta_2 = \frac{1}{2}(\Delta_1 + \widehat{\Delta})$ we have $\|\Delta_2 z\| = \left\|\widehat{\Delta}z\right\| = h\,\|z\|\,,\ \|\Delta_2\| = h$. Taking into account that $\Delta_1 \neq \widehat{\Delta}$, we obtain $\|\Delta_2\| < h$. This contradiction proves the theorem.

**6.** In the further discussion, the following variational problems will play an important part.

*Problem* 5. To find an element $\widetilde{z} \in Z_2$ such that $\|\widetilde{z}\| = \widehat{\eta} = \inf \|z\|$ for $z \in \widetilde{Z}_2 \equiv \{z \in D : \widetilde{\nu}(z) \leq \delta\}$.

*Problem* 6. To find an element $\widetilde{z} \in \widetilde{Z}_3$ such that $\|\widetilde{z}\| = \widetilde{\eta} = \inf \|z\|$ for $z \in \widetilde{Z}_3 \equiv \{z \in D : \widetilde{\nu}(z) = \delta\}$.

**Lemma 1.3.7** *Problems* 5 *and* 6 *are equivalent under the condition* (1.3.32).

The proof can be carried out according to the known scheme [265]; it is necessary only to take into account that $D$ is a closed convex set and $0 \in D$.

**Lemma 1.3.8** *Problems* 6 *and* 3 *are equivalent under the condition* (1.3.32). *Moreover, if* $z_\sigma$ *is a solution to Problem* 3 *and* $(f_\sigma, A_\sigma) \in \widetilde{\sum}$ *is a pair realizing it, then the operator* $A_\sigma$ *coincides with the operator* $\hat{A}$ (*the solution to Problem* 4) *for* $z = z_\sigma$ *and* $f_\sigma = A_\sigma z_\sigma + P_k(\widetilde{f} - A_\sigma z_\sigma)$.

The proof can be carried out with consideration of Lemmas 1.3.5 and 1.3.6 actually in the same way as in [260]. Some modification of the reciprocity principle [261] in which $D \neq U$ should be used.

It follows from Lemma 1.3.8 that the problem of finding a normal solution of the class $\widetilde{\sum}$ can be reduced to the solution of the variational Problem 6. In addition, a functional inequality of the type (1.3.27) is defined; its normal solution coincides with the solution to Problem 6.

**7.** Let us show now that under certain conditions, Problem 6 solves the correction problem formulated in Subsection 3. Moreover, we formulate a sufficient condition for the unique solvability of Problem 6 and propose a method for its solution.

Let us introduce the smoothing parametric functional

$$F^\alpha[z] \equiv \left\|\widetilde{d}(z)\right\|^2 + \alpha \|z\|^2 ,$$

where $\alpha > 0$ is the regularization parameter [265]. Let us consider the following variational problem.

*Problem* 7. To find an element $\tilde{z}^\alpha \in D$ such that

$$F^\alpha[\widetilde{z}^\alpha] = \inf F^\alpha[z], \quad z \in D.$$

**Lemma 1.3.9** *Problem* 7 *is uniquely solvable for any* $\alpha > 0$.

**Proof.** The conclusion of the lemma follows from the fact that the smoothing functional is continuous and strongly convex, whereas the set $D$ is closed and convex [48].

Let us define the following auxiliary functions of regularization parameter $\alpha > 0$:

$$\widetilde{\gamma}(\alpha) \equiv \|\widetilde{z}^{\alpha}\|, \quad \widetilde{\beta}(\alpha) \equiv \left\|\widetilde{d}(\widetilde{z}^{\alpha})\right\|, \quad \widetilde{r}(\alpha) \equiv \widetilde{\beta}(\alpha) - h\widetilde{\gamma}(\alpha). \tag{1.3.36}$$

**Lemma 1.3.10** *The functions* (1.3.36) *have the following properties:*

*1) all of them are continuous for $\alpha > 0$;*

*2) the function $\widetilde{\gamma}(\alpha)$ is monotonically nonincreasing, whereas the functions $\widetilde{\beta}(\alpha)$ and $\widetilde{r}(\alpha)$ are monotonically nondecreasing;*

*3) the limiting relations hold:*

$$\lim_{\alpha\to+\infty} \widetilde{r}(\alpha) = \lim_{\alpha\to+\infty} \widetilde{\beta}(\alpha) = \left\|\widetilde{d}(0)\right\| = \rho(\widetilde{f}, K).$$

The proof is similar to the proof of the corresponding statements in [172], [272], [125].

**Corollary** 1. *For solvability of the equation*

$$\widetilde{r}(\alpha) = \delta, \tag{1.3.37}$$

*it is necessary and sufficient that there exists $\overline{\alpha} > 0$ such that $\widetilde{r}(\overline{\alpha}) < \delta$.*

**Corollary** 2. *If $\alpha_1$ and $\alpha_2$ are roots of equation* (1.3.37), *then $\widetilde{z}^{\alpha_1} = \widetilde{z}^{\alpha_2}$.*

**Theorem 1.3.19** *If $\alpha^*$ is a root of equation* (1.3.37), *then the element $\widetilde{z}^{\alpha^*}$ is a stable approximation to the normal solution $\overline{z}$ of inequality* (1.3.26).

**Proof.** Let us put $z_\sigma \equiv \widetilde{z}^{\alpha^*}$ and take a sequence $\sigma_k = (\delta_k, h_k)$ such that $\left\|\widetilde{f}_k - \overline{f}\right\| = \delta_k$, $\left\|\widetilde{A}_k - \overline{A}\right\| = h_k$, $\delta_k \to 0$, $h_k \to 0$, $k \to \infty$.

Taking into account the definition of the element $z_{\sigma_k}$, we have the following estimate:

$$\rho^2(\widetilde{f}_k - \widetilde{A}_k z_{\sigma_k}, K) + \alpha_k^* \|z_{\sigma_k}\|^2 \le \rho^2(\widetilde{f}_k - \widetilde{A}_k \overline{z}, K) + \alpha_k^* \|\overline{z}\|^2 .$$

Using the relations

$$\begin{aligned} \rho(\widetilde{f}_k - \widetilde{A}_k z_{\sigma_k}, K) &= \delta_k + h_k \|z_{\sigma_k}\|, \\ \rho(\widetilde{f}_k - \widetilde{A}_k \overline{z}, K) &\le \delta_k + h_k \|\overline{z}\|, \end{aligned}$$

we obtain

$$\begin{gathered} h_k(\|z_{\sigma_k}\| - \|\overline{z}\|)\,(2\delta_k + h_k \|z_{\sigma_k}\| + h_k \|\overline{z}\|) \le \\ \le \alpha_k^*(\|\overline{z}\| - \|z_{\sigma_k}\|)(\|\overline{z}\| + \|z_{\sigma_k}\|). \end{gathered}$$

Thus,

$$\|z_{\sigma_k}\| \leq \|\overline{z}\|, \quad k = 1, 2, \dots . \tag{1.3.38}$$

It follows from here that from the sequence $z_{\sigma_k}$ it is possible to select a weakly convergent subsequence. We assume without loss of generality that $z_{\sigma_k} \overset{weakly}{\to} \widetilde{z}$. Note that $\widetilde{z} \in D$, since the set $D$ is weakly closed. Proceeding to the limit as $k \to \infty$ in the inequality

$$\rho(\overline{f} - \overline{A} z_{\sigma_k}, K) \leq 2(\sigma_k + h_k \|\overline{z}\|),$$

we obtain $\rho(\overline{f} - \overline{A}\widetilde{z}, K) = 0$, i.e., $\widetilde{z}$ is a solution to the inequality (1.3.26). Here the weak lower semicontinuity of the functional $\varphi(z) = \rho(\overline{f} - \overline{A}z, K)$ was used. Then, it follows from (1.3.38) that $\lim_{k\to\infty} \|z_{\sigma_k}\| = \|\widetilde{z}\|$. Since in a Hilbert space from the weak convergence and from the convergence of norms it follows the strong convergence, we obtain $\lim_{k\to\infty} z_{\sigma_k} = \widetilde{z}$.

**Theorem 1.3.20** *If equation* (1.3.37) *is solvable, then Problem* 6 *is equivalent to Problem* 7*, where* $\alpha = \alpha^*$ *is a root of equation* (1.3.37).

**Proof.** Let us put $z_\sigma \equiv \widetilde{z}^{\alpha^*}$ and show that $z_\sigma$ is a solution to Problem 6. Since $\alpha^*$ is a root of equation (1.3.37), then $z_\sigma \in \widetilde{Z}_3$. Assume, on the contrary, that for some element $\widetilde{z} \in \widetilde{Z}_3$ the inequality $\|\widetilde{z}\| < \|z_\sigma\|$ holds. Using the inequality $F^{\alpha^*}[z_\sigma] \leq F^{\alpha^*}[\widetilde{z}]$ (as in Theorem 1.3.19), we obtain $\|z_\sigma\| < \|\widetilde{z}\|$. As a result, we come to the contradiction with the above assumption. Hence, Problem 6 is solvable and $z_\sigma$ is its solution.

Let us show now that Problem 6 is uniquely solvable under the hypotheses of Theorem 1.3.20. Taking into account that $\alpha^* > 0$ and $F^{\alpha^*}[z_\sigma] \leq F^{\alpha^*}[z]$ for any $z \in D$, we have

$$\begin{aligned} \|z_\sigma\|^2 &= \lambda^* \left[\rho^2(\widetilde{f} - \widetilde{A}z_\sigma, K) - (\delta + h\,\|z_\sigma\|^2)\right] + \|z_\sigma\|^2 \leq \\ &\leq \lambda^* \left[\rho^2(\widetilde{f} - \widetilde{A}z, K) - (\delta + h\,\|z_\sigma\|)^2\right] + \|z\|^2 , \end{aligned}$$

where $\lambda^* = 1/\alpha^*$. It follows from here that the point $(z_\sigma, \lambda^*)$ is a saddle point for the Lagrange function of the following variational problem.

*Problem* 8. To find an element $\widehat{z} \in \widetilde{Z}_4$ such that

$$\|\widehat{z}\| = \inf \|z\|, \quad z \in \widetilde{Z}_4 = \left\{ z \in D : \rho(\widetilde{f} - \widetilde{A}z, K) \leq \delta + h\,\|z_\sigma\| \right\}.$$

By the Kuhn–Tucker theorem [48], [115], the element $z_\sigma$ is a solution to Problem 8. Note that Problem 8 has a unique solution, since the set $\widetilde{Z}_4$ is closed and convex. Therefore, this solution is $\widetilde{z}$. Taking into account that for any solution $\widetilde{z}$ to Problem 6 we have $\widetilde{z} \in \widetilde{Z}_3$ and $\|\widetilde{z}\| = \|z_\sigma\|$, we obtain $\widetilde{z} \in \widetilde{Z}_4$; then $\widetilde{z} = z_\sigma$.

**Corollary**. *The solvability of equation* (1.3.37) *is a sufficient condition for the unique solvability of Problem* 7.

**Remark 1.** *An example, where Problem* 7 *is uniquely solvable, but equation* (1.3.37) *is unsolvable for* $\alpha > 0$, *is considered in* [115] (*see also* [187], [263]).

**Theorem 1.3.21 (main)** *Let equation* (1.3.37) *be solvable and* $z_\sigma = \widetilde{z}^{\alpha^*}$ *be a solution to Problem* 7, *where* $\alpha^*$ *is a root of equation* (1.3.37). *Then, a pair* $(f_\sigma, A_\sigma) \in \widetilde{\sum}$ *defined by the formulas*

$$A_\sigma = \widetilde{A} + \Delta_\sigma, \quad \Delta_\sigma = h\,\frac{\widetilde{d}(z_\sigma)(z_\sigma, \cdot)}{\left\|\widetilde{d}(z_\sigma)\right\| \|z_\sigma\|}, \quad f_\sigma = A_\sigma z_\sigma + P_K(\widetilde{f} - A_\sigma z_\sigma),$$

*solves the correction problem and* $z_\sigma$ *is the normal solution to the sought-for linear functional inequality.*

**Proof.** The conclusions of this theorem directly follow from Lemmas 1.3.5–1.3.9 and Theorem 1.3.20.

**Remark 2.** *It is possible to use a gradient-type method for the numerical solution of a convex variational problem, since the smoothing functional* $F^\alpha[z]$ *is Frechét differentiable* [48]. *In order to find a root of equation* (1.3.37), *it is possible, for example, to use a special modification of the secant method* [272]. *If the problem of projecting onto the cone* $K$ *can be solved explicitly, then the formulas of Theorem* 1.3.30 *have a constructive character.*

## 1.4 Shift regularization and its application for solving singular systems of linear algebraic equations

Let a linear bounded operator $A = B - K$ be given in a Hilbert space $H$. Here $B$ is a boundedly invertible operator and $K$ is a completely continuous operator. It is necessary to solve the equation

$$Au = f, \tag{1.4.39}$$

where the element $f \in H$ is given. In order to study this problem, we use the shift regularization method [279], which was successfully applied to the numerical solution of singular integral equations in the theory of elasticity and in aerodynamics [140], [34], to the approximate calculation of bifurcation points of nonlinear operator equations [243], [274], to the linear and nonlinear spline approximations [176] as well as to the solution of systems of linear algebraic equations [200].

**1.** It follows from the above representation for the operator $A$ that the kernel of $A$ and the kernel of the adjoint operator $A^*$ are finite-dimensional and their dimensions are $\dim\ker A = \dim\ker A^* = k$ [142]. Let $e_1, e_2, \ldots, e_k$ and $g_1, g_2, \ldots, g_k$ be orthonormal bases in $\ker A$ and $\ker A^*$, respectively. Along with equation (1.4.39), we consider the equation

$$Au + Gu = f, \tag{1.4.40}$$

where the operator $G$ is of the form

$$Gu \equiv \sum_{i=1}^{k} (e_i, u) g_i, \quad u \in H.$$

Hence, the following lemma is valid.

**Lemma 1.4.11** *There exists an operator $(A+G)^{-1}$ bounded on $H$, i.e., the operator $A+G$ is boundedly invertible.*

**Proof.** It is sufficient to show that the kernels of the operators $A+G$ and $A^*+G^*$ consist only of a zero element. Then, the boundedness of the invertible operator follows from normal solvability of $A+G$.

Let for some $u_0 \in H$ the following relation is valid:

$$Au_0 + Gu_0 = 0. \tag{1.4.41}$$

Let us show that $u_0 = 0$. Indeed, since $Gu_0 \in \ker A^*$, $Au_0 \in (\ker A^*)^{\perp}$, the equalities $Au_0 = 0$, $Gu_0 = 0$ follow from (1.4.41). The representations $u_0 = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_k e_k$ and $\alpha_i = (e_i, u_0) = 0$, $i = 1, \ldots, k$, follow from the first and second equalities, i.e., $u_0 = 0$.

The relation $\ker(A^* + G^*) = \{0\}$ follows from [212]. However, it is possible to prove it directly. Indeed, from the obvious equality

$$\left( \sum_{i=1}^{k} (e_i, x)\, g_i, y \right) = \left( \sum_{i=1}^{k} (g_i, y)\, e_i, x \right), \quad x, y \in H,$$

it follows that

$$G^* y = \sum_{i=1}^{k} (g_i, y)\, e_i.$$

As above, it is possible to show that if $A^*x_0 + G^*x_0 = 0$, then $x_0 = 0$. The lemma is proved.

The pseudoinverse operator for the operator $A$ is denoted by $A^+$.

**Theorem 1.4.22** *Let equation* (1.4.39) *be solvable. Then, the unique solution to equation* (1.4.40) *coincides with the normal pseudosolution to equation* (1.4.39), *i.e.,* $(A+G)^{-1}f = A^+f$.

**Proof.** If $x_0$ is the normal solution to equation (1.4.39), then $x_0 \perp \ker A$, i.e., $(e_i, x_0) = 0$, $i = 1, 2, \ldots, k$, and, hence, $Gx_0 = 0$. Therefore, $x_0$ is a solution to equation (1.4.40).

Let now $x_0$ be a solution to equation (1.4.40):

$$Ax_0 + Gx_0 = f. \tag{1.4.42}$$

Since $Gx_0 \in \ker A^*$, then the equalities $Ax_0 = f$, $Gx_0 = 0$ follow from (1.4.42). The first of these equalities means that $x_0$ is a solution, whereas the second one means that $x_0$ is the normal solution to (1.4.39). The theorem is proved.

**2.** Let $H_n$, $n = 1, 2, \ldots,$ be a given sequence of Hilbert spaces and let $\{\varphi_n\}$ and $\{\psi_n\}$ be families of uniformly bounded linear operators such that $\varphi_n : H \to H_n$, $\psi_n : H_n \to H$. In addition, it is assumed that the following approximation conditions are fulfilled:

a) $\varphi_n\psi_n y = y \quad \forall y \in H_n$;

b) $\lim_{n\to\infty} \|\psi_n\varphi_n x - x\| = 0 \quad \forall x \in H$;

c) $\lim_{n\to\infty} \|\varphi_n x\| = \|x\| \quad \forall x \in H$.

It is said that a family of linear operators $C_n : H_n \to H_n$ is stable (respectively, pseudostable) if there are positive numbers $n_0$ and $M < \infty$ such that for $n \geq n_0$ the operators $C_n^{-1}$ (respectively, $C_n^+$) exist and $\left\|C_n^{-1}\right\| \leq M$ (respectively, $\|C_n^+\| \leq M$).

**Remark.** *If the spaces* $H_n$ *are finite-dimensional, the existence of the operators* $C_n^+$ *is certainly ensured for any* $n$ [158].

**Lemma 1.4.12** *Let condition b) be fulfilled and* $A_n \equiv \varphi_n A\psi_n$. *Then, the following relation is valid:*

$$\lim_{n\to\infty} \|\varphi_n Ax - A_n\varphi_n x\| = 0 \quad \forall x \in H.$$

**Proof.** Indeed,

$$\begin{aligned}\|\varphi_n Ax - A_n\varphi_n x\| &= \|\varphi_n A(x - \psi_n\varphi_n x)\| \leq \\ &\leq \|\varphi_n A\|\,\|x - \psi_n\varphi_n x\| \to 0, \quad n \to \infty.\end{aligned}$$

The lemma is proved.

Below we use the following analogue of the Lax theorem for nonuniquely solvable operators [93].

**Theorem 1.4.23** *In order that the family of linear operators $C_n$ be pseudostable, it is necessary and sufficient that for all $x_n \perp \ker C_n$ from $\|C_n x_n\| \to 0$, $n \to \infty$, it follows that $\|x_n\| \to 0$, $n \to \infty$.*

**3.** Let us study the question of stability and pseudostability for families of approximating operators. Let $e_{in} \equiv \varphi_n e_i$, $g_{in} \equiv \varphi_n g_i$, $i = 1, \ldots, k$, and let $\overline{e}_{in}$ and $\overline{g}_{in}$, $i = 1, \ldots, k$, be some bases in $\ker A_n$ and $\ker A_n^*$, respectively. Let us put $B_n \equiv \varphi_n B \psi_n$ and

$$G_n x \equiv \sum_{i=1}^{k} (e_{in}, x) g_{in}, \quad \overline{G} x \equiv \sum_{i=1}^{k} (\overline{e}_{in}, x) \overline{g}_{in}, \quad x \in H_n. \tag{1.4.43}$$

**Lemma 1.4.13** *If the family $\{B_n\}$ is pseudostable, $\ker B_n = \{0\}$, and the following relations are valid:*

$$\|e_{in} - \overline{e}_{in}\| \to 0, \quad \|g_{in} - \overline{g}_{in}\| \to 0, \quad n \to \infty, \quad i = 1, \ldots, k, \tag{1.4.44}$$

*where $\{\overline{e}_{in}\}$ and $\{\overline{g}_{in}\}$ are orthonormal, then the family $\{A_n + G_n\}$ is stable, whereas the family $\{A_n\}$ is pseudostable.*

**Proof.** Let us prove the stability of the family $\{A_n + G_n\}$. Let us assume the contrary. Then, there exists a sequence of normalized elements $x_n$ such that

$$\|(A_n + G_n) x_n\| \to 0, \quad n \to \infty. \tag{1.4.45}$$

We show that the validity of the relation

$$\lim_{n \to \infty} \left\| \left(G_n - \overline{G}_n\right) x \right\| = 0 \ \ \forall x \in H_n \tag{1.4.46}$$

follows from (1.4.44). Indeed, writing down the equality

$$\begin{aligned} \left(G_n - \overline{G}_n\right) x \ = \ & \sum_{i=1}^{k} (e_{in} - \overline{e}_{in}, x)(g_{in} - \overline{g}_{in}) + \\ & + \sum_{i=1}^{k} (\overline{e}_{in}, x)(g_{in} - \overline{g}_{in}) + \sum_{i=1}^{k} (e_{in} - \overline{e}_{in}, x) \overline{g}_{in} \end{aligned}$$

and estimating each term in norm, we obtain

$$\begin{aligned}\left\|\left(G_n-\overline{G}_n\right)x\right\| &\le \sum_{i=1}^{k}\Big[\|e_{in}-\overline{e}_{in}\|\left(\|g_{in}-\overline{g}_{in}\|+1\right)+\\ &\qquad +\|g_{in}-\overline{g}_{in}\|\Big]\,\|x\|\to 0,\quad n\to\infty.\end{aligned}$$

It follows from (1.4.45) and (1.4.46) that

$$\begin{aligned}\left\|\left(A_n+\overline{G}_n\right)x_n\right\| &\le \|(A_n+G_n)\,x_n\|+\\ &\quad +\left\|\left(G_n-\overline{G}_n\right)x_n\right\|\to 0,\quad n\to\infty.\end{aligned}\tag{1.4.47}$$

In its turn, from here it follows the relations

$$\|A_nx_n\|\to 0,\quad (\overline{e}_{in},x_n)\to 0,\quad n\to\infty,\quad i=1,\ldots,k.\tag{1.4.48}$$

Putting $x_n=y_n+z_n$, where $y_n\in\ker A_n$, $z_n\in(\ker A_n)^{\perp}$, from (1.4.48) we come to the validity of the limiting relations

$$\|A_nz_n\|\to 0,\quad (\overline{e}_{in},y_n)\to 0,\quad n\to\infty,\quad i=1,\ldots,k.\tag{1.4.49}$$

Since $y_n\in\ker A_n$, there exist numbers $\alpha_{in}$ such that $y_n=\sum_{i=1}^{k}\alpha_{in}\overline{e}_{in}$. From (1.4.49) it follows that $\alpha_{in}=(\overline{e}_{in},y_n)\to 0$, $n\to\infty$. Thus, $\|y_n\|\to 0$, $n\to\infty$, and, hence,

$$\|z_n\|\to 1,\quad n\to\infty.\tag{1.4.50}$$

Then, the sequences of elements $\{\psi_nz_n\}$ are bounded. Using the complete continuity of the operator $K$, it is possible to assume, without loss of generality, that there exists an element $z_0\in H$ such that

$$K\psi_nz_n\to z_0,\quad n\to\infty.\tag{1.4.51}$$

From (1.4.49), (1.4.51) and condition b) it follows that

$$\begin{aligned}\left\|B_n\left(z_n-\varphi_nB^{-1}z_0\right)\right\| &\le \|B_nz_n-\varphi_nz_0\|+\|\varphi_nz_0-\varphi_nK\psi_nz_n\|+\\ &\quad +\|\varphi_nB\|\left\|\left(\psi_n\varphi_n-I\right)B^{-1}z_0\right\|\to 0,\ \ n\to\infty.\end{aligned}$$

Therefore, by virtue of pseudostability of the family $\{B_n\}$ and by the equality $\ker B_n=\{0\}$ we have

$$\left\|z_n-\varphi_nB^{-1}z_0\right\|\to 0,\quad n\to\infty.\tag{1.4.52}$$

Further we obtain

$$\begin{aligned}\left\|\varphi_n AB^{-1}z_0\right\| &\le \|A_n z_n\| + \left\|A_n\left(\varphi_n B^{-1}z_0 - z_n\right)\right\| + \\ &\quad + \|\varphi_n A\| \left\|\left(I-\psi_n\varphi_n\right)B^{-1}z_0\right\| \to 0, \quad n\to\infty.\end{aligned}$$

From here and from condition c) it follows the equality $AB^{-1}z_0 = 0$, i.e., $B^{-1}z_0 \in \ker A$. Let $B^{-1}z_0 = \alpha_1 e_1 + \alpha_2 e_2 + \dots + \alpha_k e_k$. Having applied the operator $\varphi_n$, we obtain

$$\varphi_n B^{-1}z_0 = \sum_{i=1}^{k} \alpha_i e_{in}, \quad n = 1,2,\dots . \tag{1.4.53}$$

From (1.4.44), (1.4.52) and (1.4.53) it follows that

$$\left\| z_n - \sum_{i=1}^{k} \alpha_i \overline{e}_{in} \right\| \to 0, \quad n\to\infty. \tag{1.4.54}$$

By virtue of the inclusion $\sum_{i=1}^{k} \alpha_i \overline{e}_{in} \in \ker A_n,\ z_n \in (\ker A_n)^{\perp}$, the relation $\|z_n\| \to 0$, $n\to\infty$, follows from (1.4.54). This contradicts (1.4.50). The contradiction obtained confirms the stability of the family $\{A_n + G_n\}$. Hence, the family $\{A_n\}$ is pseudostable. The lemma is proved.

**Remark.** *The condition* $\ker B_n = \{0\}$ *in the formulation of Lemma* 1.4.13 *is essential. For the example considered below, all the hypotheses of Lemma* 1.4.13 *(except for this condition) are fulfilled. However, the stability of the family* $\{A_n + G_n\}$ *does not follow from the pseudostability of the family* $\{B_n\}$.

**Example.** Let $H = l_2$, $B = \operatorname{diag}(\mathcal{B}_1, \mathcal{B}_2, \dots \mathcal{B}_n, \dots)$, and $K = \operatorname{diag}(\mathcal{K}_1, \mathcal{K}_2, \dots \mathcal{K}_n, \dots)$, where

$$\mathcal{B}_m = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{K}_m = \begin{bmatrix} -1/m & 0 \\ 0 & 1/m \end{bmatrix}, \quad m = 1,2,\dots .$$

Let $H_n \equiv R^{2n+1}$, where $R^m$ is the $m$-dimensional Euclidean space. The operators $\varphi_n$ and $\psi_n$ are defined as follows:

$$\begin{aligned}\varphi_n x &= \left(\xi_1, \xi_2, \dots, \xi_{2n+1}\right), \ x = (\xi_1, \xi_2, \dots) \in l_2, \\ \psi_n y &= \left(\eta_1, \eta_2, \dots, \eta_{2n+1}, 0, 0, \dots\right), \ y = \left(\eta_1, \dots, \eta_{2n+1}\right) \in R^{2n+1}.\end{aligned}$$

It is easy to see that the approximation conditions a)–c) are fulfilled. The approximating operators $A_n$ and $B_n$ are of the form

$$\begin{aligned}A_n &\equiv \operatorname{diag}\left(\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n, \frac{1}{n+1}\right), \\ B_n &\equiv \operatorname{diag}\left(\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{K}_n, 0\right),\end{aligned}$$

where $\mathcal{A}_m = \mathcal{B}_m - \mathcal{K}_m,\ m = 1, \ldots, n$. From here we obtain $B_n^+ = B_n$,

$$A_n^+ = \operatorname{diag}\left(\mathcal{A}_1^+, \mathcal{A}_2^{-1}, \mathcal{A}_3^{-1}, \ldots, \mathcal{A}_n^{-1}, n+1\right),$$

where

$$\mathcal{A}_1^+ = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}, \quad \mathcal{A}_m^{-1} = \frac{1}{m^2-1}\begin{bmatrix} -m & m^2 \\ m^2 & -m \end{bmatrix}, \quad m = 2, \ldots, n.$$

Hence, $\|B_n^+\| = 1$, i.e., the family $\{B_n^+\}$ is pseudostable, whereas the family $\{A_n\}$ is not pseudostable, since $\|A_n^+\| = n + 1 \to \infty$ as $n \to \infty$.

In some sense, the following lemma is the converse of Lemma 1.4.13.

**Lemma 1.4.14** *Let a sequence $\{\varphi_n^* \varphi_n\}$ be strongly convergent to the operator $\Phi$ and*

$$\begin{aligned} &\lim\nolimits_{n\to\infty} \|e_{in} - \bar{e}_{in}\| = 0,\ \ i = 1, \ldots, k; \\ &\inf\nolimits_{n=1,2,\ldots} \|\bar{e}_{in} - \bar{e}_{jn}\| > 0, \quad i \neq j. \end{aligned} \tag{1.4.55}$$

*Then, from the pseudostability of the family $\{A_n\}$ it follows the pseudostability of the family $\{B_n\}$.*

**Proof.** First we show the validity of the equality $\Phi = I$. Indeed, since $\forall x \in H$ we have $(\varphi_n^*\varphi_n x, x) \to (x, x)$, $(\varphi_n^*\varphi_n x, x) \to (\Phi x, x)$ as $n \to \infty$, the equalities $(\Phi x, x) = (x, x)$, $\Phi^* = \Phi$ are valid. From here it follows that $\Phi = I$.

Now we establish the pseudostability of the family $\{B_n\}$. Assuming the contrary, we assume the existence of a sequence of normalized elements $x_n \perp \ker B_n$ such that $\|B_n x_n\| \to 0$, $n \to \infty$. By virtue of complete continuity of the operator $K$, it is also possible to assume without loss of generality that $K\psi_n x_n \to x_0 \in H,\ n \to \infty$. Hence,

$$\begin{aligned} \|\varphi_n (A\psi_n x_n + x_0)\| \ \le\ & \|B_n x_n\| + \\ & + \|\varphi_n (x_0 - K\psi_n x_n)\| \to 0, \quad n \to \infty. \end{aligned} \tag{1.4.56}$$

Since the sequence $\{\varphi_n^*\varphi_n\}$ strongly converges to the unit operator, $\forall y \in H$ we have $(y, A\psi_n x_n + x_0) \to 0$ as $n \to \infty$.

Thus, the sequence $\{A\psi_n x_n + x_0\}$ weakly converges to zero. From here we conclude that $x_0$ belongs to the weak closure of the image of the operator $A$. According to the hypotheses of the lemma, however, this image is closed and, hence, coincides with its weak closure. Thus, the element $x_0$ belongs to the image of the operator $A$. Further,

$$\begin{aligned} &\varphi_n x_0 - \varphi_n A\psi_n\varphi_n A^+\psi_n\varphi_n x_0 = \\ &\quad = \varphi_n A\psi_n\varphi_n A^+ \left(I - \psi_n\varphi_n\right) x_0 + \varphi_n A \left(I - \psi_n\varphi_n\right) A^+ x_0; \end{aligned}$$

hence, $\|\varphi_n x_0 - \varphi_n A\psi_n\varphi_n A^+\psi_n\varphi_n x_0\| \to 0$ as $n \to \infty$. From here and from (1.4.56) it follows that $\|A_n (x_n + \varphi_n A^+\psi_n\varphi_n x_0)\| \to 0$ as $n \to \infty$. Then, using the pseudostability of the family $\{A_n\}$, we come to the relation

$$\left\|x_n + \varphi_n A^+ x_0 + y_n\right\| \to 0, \quad n \to \infty, \tag{1.4.57}$$

where $y_n \in \ker A_n,\ x_n + \varphi_n A^+ x_0 + y_n \in (\ker A_n)^\perp$.

Let $y_n = \sum_{i=1}^k \alpha_{in}\overline{e}_{in}$. Note that since $x_n$ are normalized and the operators $\varphi_n$ are uniformly bounded, from (1.4.57) we conclude that the sequence $\{y_n\}$ is bounded. Then, from (1.4.55) it follows that $|\alpha_{in}| \le \alpha = \text{const} < \infty$, $i = 1, \dots, k$; $n = 1, 2, \dots$ . Without loss of generality, it is possible to assume that $\alpha_{in} \to \alpha_i$ as $n \to \infty$, $i = 1, \dots, k$.

Let $y_0 = \sum_{i=1}^k \alpha_i e_i$. Then, $\|\varphi_n y_0 - y_n\| \to 0$, $n \to \infty$. From here and from (1.4.57) we obtain

$$\left\|x_n - \varphi_n \left(A^+ x_0 + y_0\right)\right\| \to 0, \quad n \to \infty. \tag{1.4.58}$$

From the relation $\|B_n x_n\| \to 0$ as $n \to \infty$ and condition c) we have $B\left(A^+ x_0 + y_0\right) = 0$. However, the operator $B$ is boundedly invertible; therefore, $A^+ x_0 + y_0 = 0$ and $\varphi_n\left(A^+ x_0 + y_0\right) = 0$. Then, from (1.4.58) we obtain $\|x_n\| \to 0$, $n \to \infty$. This contradicts the fact that the elements $\{x_n\}$ are normalized. The lemma is proved.

Let us prove the following lemma.

**Lemma 1.4.15** *Let* $\overline{e}_{in}$ *and* $\overline{g}_{in}$, $i = 1, \dots, k$, *be orthonormal bases in* $\ker A_n$ *and* $\ker A_n^*$, *respectively, and the relations* (1.4.44) *take place. Then, for the stability of the family* $\{A_n + G_n\}$ *it is necessary and sufficient that the family* $\{A_n\}$ *be pseudostable.*

**Proof.** *Necessity.* Let the family $\{A_n + G_n\}$ be stable. We show the pseudostability of the family $\{A_n\}$. If we assume the contrary, then there exists a sequence of normalized elements $z_n \in (\ker A_n)^\perp$ such that $\|A_n z_n\| \to 0$, $n \to \infty$. Since

$$\begin{aligned} \|G_n x_n\| &= \left\|\sum_{i=1}^k (e_{in}, z_n)\, g_{in}\right\| = \left\|\sum_{i=1}^k (e_{in} - \overline{e}_{in}, z_n)\, g_{in}\right\| \le \\ &\le \sum_{i=1}^k \|e_{in} - \overline{e}_{in}\|\, \|g_{in}\| \to 0, \quad n \to \infty, \end{aligned}$$

then

$$\|(A_n + G_n)\, z_n\| \le \|A_n z_n\| + \|G_n z_n\| \to 0, \quad n \to \infty.$$

From the stability of the family $\{A_n + G_n\}$ it follows that $\|z_n\| \to 0$ as $n \to \infty$. This contradicts the choice of the elements $z_n$. The necessity is proved.

*Sufficiency.* Let the family $\{A_n\}$ be pseudostable. We show the stability of the family $\{A_n + G_n\}$. If we assume the contrary, then there exists a sequence of normalized elements $x_n$ such that $\|(A_n + G_n)\, x_n\| \to 0,\ n \to \infty$. Let $x_n = y_n + z_n$, where $y_n \in \ker A_n,\ \ z_n \in (\ker A_n)^{\perp}$. From (1.4.44) we establish the validity of the relations $\|A_n z_n\| \to 0,\ \left\|\overline{G}_n y_n\right\| \to 0,\ \ n \to \infty$. Since $y_n \in \ker A_n$, the representation $y_n = \sum_{i=1}^{k} \alpha_{in} \overline{e}_{in}$ is valid. Therefore, $\left\|\overline{G}_n y_n\right\|^2 = \sum_{i=1}^{k} \alpha_{in}^2 \to 0,\ \ n \to \infty$. It follows from here that $\|y_n\| \to 0,\ n \to \infty$. By Theorem 1.4.23, from the relations $\|A_n z_n\| \to 0,\ n \to \infty,\ z_n \in (\ker A_n)^{\perp}$ and from the pseudostability of the family $\{A_n\}$ it follows that $\|z_n\| \to 0,\ n \to \infty$.

Thus, $\|x_n\| \le \|y_n\| + \|z_n\| \to 0,\ n \to \infty$. This is in contradiction with the fact that the elements $x_n$ are normalized, which proves the lemma.

**4.** Let us study the question of convergence and its rate for approximate solutions and the question of influence of the approximate character of input data.

**Theorem 1.4.24** *Let the hypotheses of Lemma* 1.4.13 *be fulfilled. Then, the equality*

$$\lim_{n\to\infty} \left\|\varphi_n x_f - (A_n + G_n)^{-1} \varphi_n f\right\| = 0 \quad \forall f \in H$$

*is valid, where* $x_f \equiv (A + G)^{-1} f$.

**Proof.** Since

$$\left\|\varphi_n x_f - (A_n + G_n)^{-1} \varphi_n f\right\| \le \\ \le \left\|(A_n + G_n)^{-1}\right\| \ \|(A_n + G_n)\, \varphi_n x_f - \varphi_n (A + G)\, x_f\|,$$

then by Lemma 1.4.12 it is sufficient to show the validity of the relation

$$\|(G_n \varphi_n - \varphi_n G)\, x_f\| \to 0, \quad n \to \infty. \tag{1.4.59}$$

Let us consider the following equality:

$$(G_n \varphi_n - \varphi_n G)\, x_f = \sum_{i=1}^{k} (e_{in}, \varphi_n x_f)\, g_{in} - \sum_{i=1}^{k} (e_i, x_f)\, \varphi_n g_i = \\ = \sum_{i=1}^{k} (\varphi_n^* \varphi_n e_i - e_i, x_f)\, g_{in}. \tag{1.4.60}$$

By virtue of condition c), the sequence of the operators $\{\varphi_n^* \varphi_n\}$ weakly converges to the unit operator; therefore, the validity of relation (1.4.59) follows from (1.4.60). The theorem is proved.

**Corollary.** *If equation* (1.4.39) *is solvable, then the equality* $\lim_{n\to\infty} \|\varphi_n A^+ f - A_n^+ \varphi_n f\| = 0$ *is valid when the hypotheses of Lemma* 1.4.13 *are fulfilled.*

The validity of this statement follows from Theorems 1.4.22 and 1.4.24.

Similar statements are also valid when equation (1.4.39) is not consistent. In this case, instead of equation (1.4.39) it is necessary to consider the solvable equation $Au = f_0$, where $f_0 = f - \sum_{i=1}^{k} (g_i, f)\, g_i$, whose normal solution coincides with the normal pseudosolution to equation (1.4.39).

Note further that condition b) means the strong convergence of the sequence $\{\psi_n \varphi_n\}$ to the unit operator, whereas condition c) means the weak convergence of the sequence $\{\varphi_n^* \varphi_n\}$ to the unit operator. In order to estimate the convergence rate, we assume that

$$\|\psi_n \varphi_n x - x\| \leq a_n(x), \tag{1.4.61}$$

$$|(\varphi_n^* \varphi_n e_i - e_i, x)| \leq b_n(x),\ i = 1, \ldots, k, \tag{1.4.62}$$

where $a_n(x) \geq 0$, $b_n(x) \geq 0$ are given, whereas $\lim_{n\to\infty} a_n(x) = 0$ and $\lim_{n\to\infty} b_n(x) = 0$.

**Lemma 1.4.16** *The following estimates are valid:*

$$\|(\varphi_n A - A_n \varphi_n)\, x\| \leq \|\varphi_n A\|\ a_n(x);$$
$$\|(\varphi_n G - G_n \varphi_n)\, x\| \leq \sqrt{k}\, b_n(x) \quad \forall x \in H.$$

The proof of this lemma is obvious.

**Theorem 1.4.25** *Let the family* $\{A_n + G_n\}$ *be stable. Then,* $\forall f \in H$ *the following estimate is valid:*

$$\left\| \varphi_n x_f - (A_n + G_n)^{-1} \varphi_n f \right\| \leq m_n \left( \|\varphi_n A\|\, a_n(x_f) + \sqrt{k} b_n(x_f) \right),$$

*where* $\left\| (A_n + G_n)^{-1} \right\| \equiv m_n \leq m < \infty$.

**Proof.** The proof follows from the equality

$$\varphi_n x_f - (A_n + G_n)^{-1} \varphi_n f = (A_n + G_n)^{-1} \left[ (A_n + G_n)\, \varphi_n - \varphi_n\, (A + G) \right] x_f$$

and Lemma 1.4.16.

The following theorem can be proved similarly.

**Theorem 1.4.26** *Let the family* $\{A_n\}$ *be pseudostable and equation* (1.4.39) *be solvable. Then, the estimate*

$$\left\|\varphi_n A^+ f - A_n^+ \varphi_n f\right\| \le m_n \left(\|\varphi_n A\|\, a_n(x_f) + \sqrt{k} b_n(x_f)\right)$$

*is valid.*

**Lemma 1.4.17** *Let* $\left\|(A_n + G_n)^{-1}\right\| = m_n \le m < \infty$, *the operators* $\widetilde{A}_n$ *and* $\widetilde{G}_n$ *act in* $H_n$ *and satisfy the inequalities* $\|A_n - \widetilde{A}_n\| \le \eta_n$, $\|G_n - \widetilde{G}_n\| \le \delta_n$, $\eta_n$, $\delta_n \ge 0$. *If* $m_n(\eta_n + \delta_n) = q_n \le q < 1$, *then the operator* $\widetilde{A}_n + \widetilde{G}_n$ *is invertible and the following equality holds:*

$$\begin{aligned}\left(\widetilde{A}_n + \widetilde{G}_n\right)^{-1} &= (A_n + G_n)^{-1} \{I_n - [(A_n + G_n) - \\ &\quad - \left(\widetilde{A}_n + \widetilde{G}_n\right)\Big] (A_n + G_n)^{-1}\Big\}^{-1},\end{aligned} \tag{1.4.63}$$

*where* $I_n$ *is the unit operator in* $H_n$; *in addition, the estimate*

$$\left\|\left(\widetilde{A}_n + \widetilde{G}_n\right)^{-1}\right\| \le \frac{m}{1-q} \tag{1.4.64}$$

*is valid.*

**Proof.** Putting $D_n = A_n + G_n$ and $\widetilde{D}_n = \widetilde{A}_n + \widetilde{G}_n$, we obtain

$$\widetilde{D}_n = \left[I_n \left(D_n - \widetilde{D}_n\right) D_n^{-1}\right] D_n.$$

Since

$$\left\|\left(D_n - \widetilde{D}_n\right) D_n^{-1}\right\| \le \left\|D_n - \widetilde{D}_n\right\| \left\|D_n^{-1}\right\| \le (\eta_n + \delta_n)\, m_n \le q < 1,$$

the operator $I_n - \left(D_n - \widetilde{D}_n\right) D_n^{-1}$ is invertible. Hence, the operator $\widetilde{D}_n$ is also invertible; moreover, the equality (1.4.63) and the inequality (1.4.64) are valid. The lemma is proved.

**Theorem 1.4.27** *Let the hypotheses of Lemma* (1.4.17) *and the relations*

$$m_n(\eta_n + \delta_n) = q_n \le q < 1, \quad \left\|f_n - \widetilde{f}_n\right\| \le \varepsilon_n \|f_n\|$$

*be fulfilled. Then, the following estimate is valid:*

$$\left\|\widetilde{D}_n^{-1} \widetilde{f}_n - D_n^{-1} f_n\right\| \le \frac{m_n(\varepsilon_n + q_n)}{1-q} \|f_n\|.$$

**Proof.** Let us write down the following obvious inequality:

$$\left\|\widetilde{D}_n^{-1}\widetilde{f}_n - D_n^{-1}f_n\right\| \leq \left\|\widetilde{D}_n^{-1}\right\| \ \left\|\widetilde{f}_n - f_n\right\| + \left\|\widetilde{D}_n^{-1} - D_n^{-1}\right\| \ \|f_n\| . \quad (1.4.65)$$

Now we estimate the second term in the right-hand side of (1.4.65). Since

$$\widetilde{D}_n^{-1} - D_n^{-1} = \widetilde{D}^{-1}\left(D_n - \widetilde{D}_n\right) D_n^{-1},$$

the estimate

$$\left\|D_n^{-1} - \widetilde{D}_n^{-1}\right\| \leq \left\|\widetilde{D}n^{-1}\right\| \left\|D_n - \widetilde{D}_n\right\| \left\|D_n^{-1}\right\| \leq \frac{m_n^2\left(\eta_n + \delta_n\right)}{1 - m_n\left(\eta_n + \delta_n\right)} \leq \frac{mq_n}{1-q}$$

holds by Lemma 1.4.17. Therefore, the estimate being proved follows from (1.4.65). The theorem is proved.

Note that some stable methods for calculation of operator kernels are discussed in [1], [117] in the context of the above approach.

**5.** Let us consider some necessary and sufficient conditions for regularizability of singular systems of linear algebraic equations by the shift method.

Let $A, B : C^n \to C^n$ be matrix operators and $C^n$ be the $n$-dimensional complex Euclidean space. We consider a system of linear algebraic equations

$$Au = f, \quad (1.4.66)$$

where $f \in C^m$ is a given vector and $u \in C^m$ is the sought-for vector. Along with (1.4.66), we consider the regularizing family of equations

$$(A + \lambda B)\, u = f, \quad (1.4.67)$$

where $\lambda \in C = C^1$ is the regularization parameter.

It is well known that if the matrix $A$ is invertible, then there is a positive number $\rho_1$ such that for any $\lambda \in C,\ 0 \leq |\lambda| \leq \rho_1$, the matrix $A + \lambda B$ is also invertible. If $u_0$ is a solution to (1.4.66) and $u_\lambda$ is a solution to (1.4.67), then the following estimate holds:

$$\|u_\lambda - u_0\| = O(\lambda). \quad (1.4.68)$$

If the matrix $A$ is not invertible, then, in general, the estimate (1.4.68) does not take place.

The case when the matrix $A$ is not invertible and, instead of it, an approximation $\widetilde{A}$ is given is of importance in research. In this case, the problem (1.4.66) is ill posed.

If $A = A^* > 0$, then instead of data $(A, B, \lambda, f)$ it is possible to take $(A, E, \alpha, f)$; then, (1.4.67) becomes a simplified regularization according to

A. N. Tikhonov. If $A = A^*$, it is possible to take $(A, E, i\alpha, f)$, i.e., we come to the regularization according to A. B. Bakushinskiy [28]. If

$$Bu \equiv \sum_{m=1}^{k} (e_m, u)\, g_m,$$

where $e_m$ and $g_m$, $m = 1, 2, \dots, k$, are orthonormal bases in $\ker A$ and $\ker A^*$, respectively, then (1.4.67) is transformed to the regularizator considered above.

The *pencil* of matrices $A + \lambda B$ is said to be *singular* if $\varphi(\lambda) \equiv \det(A + \lambda B) \equiv 0$ for any $\lambda \in C$. If there exists $\lambda_0 \in C$ such that $\varphi(\lambda_0) \neq 0$, the pencil $A + \lambda B$ is said to be *regular.*

It follows from the definition of a regular pencil that there exists $\rho_0 > 0$ such that for any $\lambda \in \Lambda_0 = \{\lambda \in C : 0 < |\lambda| < \rho_0\}$ the matrix $A + \lambda B$ is invertible. Indeed, if $\lambda_1, \dots, \lambda_k$ are all the roots of the equation $\varphi(\lambda) = 0$, then $\rho_0$ may be defined as follows:

$$\rho_0 = \begin{cases} \min\{|\lambda_i| : \lambda_i \neq 0,\ \varphi(\lambda_i) = 0,\ i = 1, 2, \dots, k\}; \\ > 0 \quad if \quad \lambda_i = 0,\ i = 1, 2, \dots, k. \end{cases}$$

**Theorem 1.4.28** *Let the pencil of* $A + \lambda B$ *be regular and* $0 < \rho < \rho_0$. *Then the equality*

$$(A + \lambda B)^{-1} = \sum_{m=-\infty}^{\infty} \lambda^m A_m \tag{1.4.69}$$

*holds, where* $A_m = \frac{1}{2\pi i} \int\limits_{|\mu|=\rho} \mu^{-m-1} (A + \mu B)^{-1}\, d\mu,\ m = 0, \pm 1, \dots$ .

**Proof.** It is easy to see that the elements $c_{ij}(\lambda)$ of the matrix $(A + \lambda B)^{-1}$ on the set $\Lambda_0$ are analytic functions. Therefore, the matrix $(A + \lambda B)^{-1}$ can be expanded in Loran's series on this set. Let us show that this expansion is the right-hand side of equality (1.4.69). Indeed, multiplying the right-hand side of (1.4.69) by $A + \lambda B$, we obtain

$$\begin{aligned} &(A + \lambda B) \sum_{m=-\infty}^{\infty} \lambda^m A_m = \sum_{m=-\infty}^{\infty} \lambda^m (A A_m + B A_{m-1}) = \\ &= \sum_{m=-\infty}^{\infty} \lambda^m \frac{1}{2\pi i} \left[ \int\limits_{|\mu|=\rho} \mu^{-m-1} A (A + \mu B)^{-1}\, d\mu + \right. \\ &\left. + \int\limits_{|\mu|=\rho} \mu^{-m} B (A + \mu B)^{-1}\, d\mu \right] = \sum_{m=-\infty}^{\infty} \lambda^m \frac{1}{2\pi i} \int\limits_{|\mu|=\rho} \mu^{-m-1}\, d\mu. \end{aligned} \tag{1.4.70}$$

Since

$$\frac{1}{2\pi i}\int_{|\mu|=\rho}\mu^{-m-1}d\mu = \begin{cases} 0 & if \quad m \neq 0, \\ 1 & if \quad m = 0, \end{cases}$$

it follows from (1.4.70) that

$$(A+\lambda B)\sum_{m=-\infty}^{\infty}\lambda^m A_m = E.$$

The theorem is proved.

Let $r$ be the multiplicity of the zero root for the equation $\varphi(\lambda) = 0$, i.e., $\varphi(0) = \varphi^{'}(0) = \cdots = \varphi^{(r-1)}(0) = 0$, $\varphi^{(r)}(0) \neq 0$; by $q_{ij}$ we denote the multiplicity of the zero root for the equation $\varphi_{ij}(\lambda) = 0$, where $\varphi_{ij}(\lambda)$ is the algebraic cofactor for the element $a_{ij} + \lambda b_{ij}$ of the matrix $A + \lambda B$. Let $C(\lambda) \equiv \left\{\varphi_{ij}(\lambda)\right\}$, $1 \leq i, j \leq n$.

**Theorem 1.4.29** *Let the hypotheses of Theorem* 1.4.28 *be fulfilled and*

$$q = \min\left\{|q_{ij}| : 1 \leq i,\, j \leq n\right\}.$$

*Then, in* (1.4.69) $A_m = 0$ *for all* $m < -r + q$ *and* $A_{-r+q} \neq 0$.

**Proof.** We have

$$(A+\lambda B)^{-1} = \varphi^{-1}(\lambda)C(\lambda). \tag{1.4.71}$$

It follows from the definition of the numbers $r$ and $q$ that

$$\varphi(\lambda) = \lambda^r\varphi_1(\lambda), \quad \varphi_1(0) \neq 0, \quad C(\lambda) = \lambda^q C_1(\lambda), \quad C_1(0) \neq 0.$$

Therefore, it follows from (1.4.71) that for $\lambda \neq 0$, $\lambda \to 0$,

$$(A+\lambda B)^{-1} = \lambda^{-r+q}\varphi_1^{-1}(\lambda)C_1(\lambda), \quad \varphi_1^{-1}(0)C_1(0) \neq 0.$$

Since the elements of the matrix $\varphi_1^{-1}(\lambda)C_1(\lambda)$ are analytic functions on the set $\Lambda_0 \cup \{0\}$, the expansion of the matrix $\varphi_1^{-1}(\lambda)C_1(\lambda)$ in Loran's series does not contain negative degrees of $\lambda$ and begins with $\lambda^0$ with a nonzero coefficient. The theorem is proved.

Let $A = [a_1, \ldots, a_n]$ and $B = [b_1, \ldots, b_n]$, where $a_j = [a_{1j}, \ldots, a_{nj}]^T$ and $b_j = [b_{1j}, \ldots, b_{nj}]^T$. We put

$$c_m = \sum_{m_1=0}^{m}\sum_{1\leq j_1<\cdots<j_{m_1}<n}\det\left[a_1 \ldots a_{j_1-1}b_{j1}a_{j_1+1}\ldots a_{jm_1-1}b_{jm_1}a_{jm_1+1}\ldots a_n\right].$$

Then, obviously, $\varphi(\lambda) = c_0 + \lambda c_1 + \lambda^2 c_2 + \cdots + \lambda^n c_n$. If $r$ is a multiplicity of the zero root of the polynomial $\varphi(\lambda)$, then $c_0 = \cdots = c_{r-1} = 0$, $c_r \neq 0$. The elements $q_{ij}$ are similarly determined in terms of the sum of determinants.

**6.** In this subsection we use the following notation: $A^+$ is the pseudoinverse matrix for a matrix $A$; $Q = E - A^+A$ and $P = E - AA^+$ are orthogonal projectors onto $\ker A$ and $\ker A^*$, respectively; $A^*$ is the adjoint matrix for the matrix $A$; $R(A)$ is the range of values for the operator $A$.

**Theorem 1.4.30** *The following conditions are equivalent:*
*the pencil* $A + \lambda B$ *is regular and the equality*

$$R(BQ) \cap R(A) = \{0\} \tag{1.4.72}$$

*holds*;
*the matrix equalities*

$$Q = F^+F, \quad F = PBQ \tag{1.4.73}$$

*are valid.*

**Proof.** 1.4.72 $\Rightarrow$ 1.4.73. Note that if $P_1$ and $P_2$ are orthogonal projectors acting in the same space, then in order that $P_1 = P_2$ it is necessary and sufficient that $\ker P_1 = \ker P_2$. Since $\ker F^+F = \ker F = \ker PBQ \supset \ker Q$, it is sufficient to show that $\ker PBQ \subset \ker Q$. Let $z \in \ker F$, i.e., $PBQz = 0$. Hence, $BQz \in \ker P = R(A)$. From here and from (1.4.72) it follows that $BQz = 0$. Since $Q$ is an orthogonal projector onto $\ker A$, we have $AQz = 0$. Thus, $(A + \lambda B)\,Qz = 0$. Since the pencil $A + \lambda B$ is regular, we get $Qz = 0$. The implication 1.4.72 $\Rightarrow$ 1.4.73 is proved.

1.4.73 $\Rightarrow$ 1.4.72. It follows from (1.4.72) that

$$\ker PBQ = \ker Q. \tag{1.4.74}$$

Let $y \in R(BQ) \cap R(A)$; then, there exists $z \in C^n$ such that $y = BQz \in R(A) = \ker P$ and $Py = PBQz = 0$. By virtue of (1.4.74) we obtain the equality $Qz = 0$. Therefore, $y = BQz = 0$. The implication (1.4.73) $\Rightarrow$ (1.4.72) is proved. Now we show that the pencil $A + \lambda B$ is regular. Let us assume the contrary: for all $\lambda \in C$ the matrix $A + \lambda B$ is singular. Then, the equation $(A + \lambda B)\,x = 0$ has a nontrivial solution $x(\lambda)$. Let us expand $x(\lambda)$ in $\lambda$:

$$x(\lambda) = \lambda^{-m}x_{-m} + \lambda^{-m+1}x_{-m+1} + \cdots, \quad x_{-m} \neq 0.$$

From this expansion and from the identity $(A + \lambda B)\,x(\lambda) = 0$ we obtain

$$Ax_{-m} = 0, \tag{1.4.75}$$
$$Ax_{-m+1} + Bx_{-m} = 0. \tag{1.4.76}$$

It follows from (1.4.75) that $x_{-m} = Qx_{-m}$. From (1.4.76) it follows that $PAx_{-m+1} + PBx_{-m} = PBQx_{-m} = Fx_{-m} = 0$. Using the equality (1.4.73), from the relation $Fx_{-m} = 0$ we come to the equality $Qx_{-m} = 0$, which contradicts the assumption that $x_{-m} \neq 0$. The theorem is proved.

In the proof of the next theorem we will use the following geometric fact: if $E_1$ and $E_2$ are subspaces of $C^n$ and $E_1 \cap E_2 = \{0\}$, $\dim E_1 + \dim E_2 = n$, then ${}^\perp E_1 \cap^\perp E_2 = \{0\}$ and ${}^\perp E_1 \oplus^\perp E_2 = C^n$, where ${}^\perp E_i$ is an orthogonal complement to the subspace $E_i$, $i = 1, 2$.

**Theorem 1.4.31** *For the equality* (1.4.73) *to be valid, it is necessary and sufficient that*

$$P = FF^+, \quad F = PBQ. \tag{1.4.77}$$

**Proof.** *Necessity* (1.4.73) $\Rightarrow$ (1.4.77). Since $P$ and $FF^+$ are orthogonal projectors acting in $C^n$, it is sufficient to show the validity of the inclusion $\ker QB^*P \subset \ker P$, because the inverse inclusion is obvious. This inclusion takes place if $\ker QB^* \cap \ker A^* = \{0\}$. This equality is equivalent to the equality ${}^\perp R(BQ) \cap^\perp R(A) = \{0\}$. Let $\dim\ker A = k$. Then, $\dim R(A) = n - k$. Let us show that $\dim R(BQ) = k$. Let this be not the case, i.e., $\dim R(BQ) < k$. Then, there exists a nonzero vector $z \in \ker A$ such that $Bz = 0$, i.e., $(A + \lambda B)\, z = 0$ for any $\lambda \in C$. Further, taking into account the regularity of the pencil $A+\lambda B$, we obtain $z = 0$. This contradiction shows that $\dim R(BQ) = k$. Since $R(BQ) \cap R(A) = \{0\}$ and $\dim R(A) + \dim R(BQ) = n$, by virtue of the above fact we have ${}^\perp R(BQ) \cap^\perp R(A) = \{0\}$. The necessity is proved.

*Sufficiency.* It follows from (1.4.77) that

$$\ker QB^*P = \ker P. \tag{1.4.78}$$

Let us show the validity of the equality

$$\ker QB^* \cap \ker A^* = \{0\}\,. \tag{1.4.79}$$

Let $z \in \ker QB^* \cap \ker A^*$. Then, $Pz = z$ and $QB^*z = 0$; therefore, $QB^*Pz = 0$. From this equality and from (1.4.78) it follows that $Pz = z = 0$. Now, using (1.4.79), we obtain the equality (1.4.72). The theorem is proved.

**Lemma 1.4.18** *For any matrices $A$ and $B$ the following equalities are valid:*

$$QF^+ = F^+P = F^+, \quad F = PBQ, \tag{1.4.80}$$

$$AT = -AA^+ = P - E, \quad T = -\left(E - F^+B\right)A^+. \tag{1.4.81}$$

*If* (1.4.77) *holds, then*

$$PBT = 0. \tag{1.4.82}$$

**Proof.** Indeed, from the equalities $F^{+} = F^{*}(FF^{*})^{+} = (F^{*}F)^{+} F^{*}$, $QF^{*} = F^{*}$, $F^{*}P = F^{*}$ it follows the validity of (1.4.80). For (1.4.81) we have $AT = -A(E - QF^{+}B)A^{+} = -AA^{+}$; for (1.4.82) we have

$$PBT = -PB(E - F^{+}B)A^{+} = -PB(E - QF^{+}B)A^{+} =$$
$$= (-PB + PBQF^{+}B)A^{+} = (-PB + FF^{+}B)A^{+} = (-PB + PB)A^{+} = 0.$$

The lemma is proved.

**Theorem 1.4.32** *Let the equality* (1.4.72) *be valid. Then, the series*

$$\begin{gathered}\lambda^{-1}F^{+} + \sum_{m=0}^{\infty} \lambda^{m}(TB)^{m} T(BF^{+} - E), \\ 0 < |\lambda| < \|TB\|^{-1},\end{gathered} \tag{1.4.83}$$

*is convergent and its sum is the inverse of the matrix $A + \lambda B$.*

**Proof.** Within the above parameter range, the convergence of (1.4.83) is obvious. Let us show the validity of the second part of this theorem. We have

$$\begin{gathered}(A + \lambda B)\left[\lambda^{-1}F^{+} + \sum_{m=0}^{\infty} \lambda^{m}(TB)^{m} T(BF^{+} - E)\right] = \\ = \lambda^{-1}AF^{+} + [AT(BF^{+} - E) + BF^{+}] + \\ + \sum_{m=0}^{\infty} \lambda^{m}\left[A(TB)^{m} T + B(TB)^{m-1} T\right](BF^{+} - E).\end{gathered} \tag{1.4.84}$$

From (1.4.80) it follows that the coefficient $AF^{+}$ at $\lambda^{-1}$ is equal to zero. Taking (1.4.81) and (1.4.77) into account, we obtain

$$\begin{aligned}AT(BF^{+} - E) + BF^{+} &= (P - E)(BF^{+} - E) + BF^{+} = \\ &= FF^{+} - P + E = E.\end{aligned}$$

Using (1.4.82), we get

$$A(TB)T(BF^{+} - E) + BT(BF^{+} - E) =$$
$$= (-AA^{+} + E)BT(BF^{+} - E) = PBT(BF^{+} - E) = 0,$$
$$A(TB)^{m} T + B(TB)^{m-1} T = (AT + E)B(TB)^{m-1} = PBT(BT)^{m-1} = 0.$$

The theorem is proved.

**Theorem 1.4.33** *Let the equality*

$$(A+\lambda B)^{-1} = \lambda^{-1}A_{-1} + A_0 + \lambda A_1 + \cdots \tag{1.4.85}$$

*hold, where $A_m$, $m=-1,0,1,\dots,$ are some matrices. Then,*

$$QQ_1 = 0, \tag{1.4.86}$$

*where $Q_1 = E - F^+F$.*

**Proof.** Note that $Q$ and $Q_1$ are orthogonal projectors onto $\ker A$ and $\ker F$, respectively, and $QQ_1 = Q_1Q$. Therefore, $QQ_1$ is also an orthogonal projector [22] onto $\ker A \cap \ker F$.

Let $z \in \ker A \cap \ker F$. Then, $Az = 0$, $Qz = z$, and $PBz = 0$. Therefore, $(A+\lambda B)\, z \equiv \lambda A^+Bz$ for any $\lambda \in C$, and for $\lambda \in \Lambda_0$ we have

$$z \equiv \lambda\,(A+\lambda B)^{-1}\,AA^+Bz.$$

From (1.4.85) it follows that

$$z = A_{-1}AA^+Bz.$$

But $A_{-1}A = 0$; therefore, $z = 0$ and $\ker A \cap \ker F = \{0\}$. Hence, $QQ_1 = 0$. The theorem is proved.

**Corollary 1.** *The validity of* (1.4.72) *in Theorem* 1.4.30 *follows from* (1.4.85).

Indeed, from (1.4.85) we obtain the relation (1.4.86) equivalent to (1.4.72). Further it is necessary to use Theorem 1.4.30.

**Theorem 1.4.34** *Let the condition* (1.4.72) *be valid. Then, for any $f \in R(A)$ the following relations hold:*

$$\lim_{\lambda\to 0}\,(A+\lambda B)^{-1}\,f = u(f), \quad Au(f) = f, \quad u(f) \equiv A^+f + F^+BA^+f.$$

**Proof.** By Theorem 1.4.32 we have

$$(A+\lambda B)^{-1}\,f = \left[\lambda^{-1}F^+ + \sum_{m=0}^{\infty}\lambda^m\,(TB)^m\,T\,(BF^+ - E)\right]AA^+f =$$
$$= -TAA^+f - \sum_{m=1}^{\infty}\lambda^m\,(TB)^m\,TAA^+f.$$

Since $TAA^+f = A^+f + QF^+BA^+F$, we have $A\,(-TAA^+f) = AA^+f = f$ and

$$\sum_{m=1}^{\infty}\lambda^m\,(TB)^m\,TAA^+f \;\to 0 \quad \text{as} \quad \lambda \to 0.$$

The theorem is proved.

**Theorem 1.4.35** *Let for all* $f \in R(A)$ *the relation* $(A + \lambda B)^{-1} f \to u_f$, $\lambda \to 0$, *be valid. Then,* $Au_f = f$ *and the equality* (1.4.85) *takes place.*

**Proof.** It follows from the hypothesis of this theorem that $(A + \lambda B)^{-1} f = u_f + u_\lambda$, where $u_\lambda \to 0$ as $\lambda \to 0$. Therefore, $f = (A + \lambda B)(u_f + u_\lambda) = Au_f + \lambda Bu_f + (A + \lambda B) u_\lambda \to Au_f$ as $\lambda \to 0$. Hence, $Au_f = f$.

Now we show the validity of (1.4.85). Let

$$(A + \lambda B)^{-1} = \lambda^{-k} A_{-k} + \cdots + \lambda^{-1} A_{-1} + A_0 + \lambda A_1 + \cdots . \tag{1.4.87}$$

Let us show that $A_{-k} = \cdots = A_{-2} = 0$. Indeed, from (1.4.87) and from the hypothesis of this theorem we obtain

$$A_{-k} A = \cdots = A_{-1} A = 0. \tag{1.4.88}$$

On the other hand, it follows from (1.4.87) and from the identity

$$E = (A + \lambda B)^{-1} (A + \lambda B)$$

that $A_{-k}A = 0$, $A_{-m+1}A + A_{-m}B = 0$, $m = 2, 3, \ldots, k$. From these equalities and from (1.4.88) we have $A_{-k}B = \cdots = A_{-2}B = 0$. Therefore, $A_{-m}(A + \lambda B) = 0$ for $m = 2, 3, \ldots, k$. Using the regularity of the pencil $A + \lambda B$, we obtain $A_{-k} = \cdots = A_{-2} = 0$. The theorem is proved.

**Corollary 2.** *It follows from the above theorem that the inequality*

$$\left\| (A + \lambda B)^{-1} A \right\| \le \text{const}$$

*implies the fulfilment of the condition* (1.4.72).

The results obtained above may be summed up as follows.

**Theorem 1.4.36** *In order that the solutions to equation* (1.4.67) *be convergent for any* $f \in R(A)$ *to some solution of equation* (1.4.66) *as* $\lambda \to 0$, *it is necessary and sufficient that any of the following conditions be fulfilled:* 1) *condition* (1.4.72)*;* 2) *equality* (1.4.73)*;* 3) *equality* (1.4.77)*;* 4) $\left\| (A + \lambda B)^{-1} \right\| = O(\lambda^{-1})$*;* 5) $\left\| (A + \lambda B)^{-1} A \right\| \le \text{const}$.

In conclusion we point out some properties of the expansion for the matrix $A + \lambda B$ in Loran's series:

1) if $A_m = 0$ for $m < -k$, $A_{-k} \ne 0$, then $A_{-k}A = 0$ and $A_{k+1}A \ne 0$;

2) if $f \perp R(A)$ and $\left\| (A + \lambda B)^{-1} f \right\| \le \text{const}$, then $f = 0$.

**7.** If in addition to the regularity of the pencil $A+\lambda B$ there is some extra information on the matrix $B$, then it is possible to obtain stronger statements.

For example, if $R(B) = \ker A^*$ and $\ker B = R(A^*)$, then the matrix $A+\lambda B$ is invertible for all nonzero $\lambda \in C$ and the following equality holds:

$$(A+\lambda B)^{-1} = A^+ + \frac{1}{\lambda} B^+. \tag{1.4.89}$$

Indeed,

$$\begin{aligned}\left(A^+ + \frac{1}{\lambda} B^+\right)(A+\lambda B) &= A^+A + \lambda A^+B + \frac{1}{\lambda} B^+A + B^+B = \\ &= A^+A + \lambda\,(A^*A)^+ A^*B + \frac{1}{\lambda}(B^*B)^+ B^*A + B^+B = \\ &= A^+A + B^+B.\end{aligned} \tag{1.4.90}$$

Since $B^+B$ is an orthogonal projector onto ${}^\perp\ker B = \ker A$, we have $B^+B = E - A^+A$. It follows from (1.4.90) that the matrix $A+\lambda B$ is invertible and the equality (1.4.89) is valid.

Let $\dim\ker A = k$ and $e_m$ and $g_m$, $m = 1, 2, \ldots, k$, be orthonormal bases in $\ker A$ and $\ker A^*$, respectively. Let

$$B = B_0 = \sum_{m=1}^{k} g_m e_m^*. \tag{1.4.91}$$

Then,

$$B_0 u \equiv \sum_{m=1}^{k} g_m e_m^* u \equiv \sum_{m=1}^{k} (e_m, u)\, g_m.$$

It follows from this example that the matrix

$$(A+\lambda B_0) \equiv A + \lambda \sum_{m=1}^{k} g_m e_m^* \tag{1.4.92}$$

is invertible for all $\lambda \neq 0$.

**Theorem 1.4.37** *Let the relation* (1.4.91) *be valid. Then,*

$$(A+\lambda B_0)^{-1} = A^+ + \frac{1}{\lambda}\sum_{m=1}^{k} e_m g_m^*. \tag{1.4.93}$$

**Proof.** We have

$$\left(A^{+}+\frac{1}{\lambda}\sum_{m=1}^{k}e_m g_m^*\right)\left(A+\lambda\sum_{m=1}^{k}g_m e_m^*\right)=A^{+}A+\sum_{m=1}^{k}e_m e_m^*. \quad (1.4.94)$$

Now we show that

$$\sum_{m=1}^{k}e_m e_m^*=E-A^{+}A. \quad (1.4.95)$$

Let $z\in\ker A$. Then,

$$z=\sum_{j=1}^{k}\alpha_j e_j,\quad \sum_{m=1}^{k}e_m e_m^* z=\sum_{m=1}^{k}\sum_{j=1}^{k}\alpha_j\left(e_m,e_j\right)e_m=\sum_{m=1}^{k}\alpha_m e_m=z.$$

Suppose $z\in^{\perp}\ker A=R(A^*)$. Then, $z=A^*y$ and

$$\sum_{m=1}^{k}e_m e_m^* z=\sum_{m=1}^{k}\left(e_m,z\right)e_m=\sum_{m=1}^{k}\left(Ae_m,y\right)e_m=0.$$

So, the equality (1.4.95) is valid; therefore, from (1.4.94) it follows that the matrix (1.4.92) is invertible for $\lambda\neq 0$ and the equality (1.4.93) holds. The theorem is proved.

**Corollary 3.** *If* $R(B)=\ker A^*$ *and* $\ker B=R(A^*)$, *then for* $f\in R(A)$ *the solution to* (1.4.67) *is the normal solution to* (1.4.66)*:* $(A+\lambda B)^{-1}f\equiv A^{+}f$, $\lambda\neq 0$.

The following more general theorem is valid.

**Theorem 1.4.38** *In order that the solution to* (1.4.67) *be convergent for all* $f\in R(A)$ *to the normal solution of* (1.4.66)*, it is necessary and sufficient that the matrix* $F^{+}BA^{+}A$ *be the zero matrix.*

**8.** Let us assume that instead of exact data $A$, $B$, and $f$ their approximation $\widetilde{A}$, $\widetilde{B}$, and $\widetilde{f}$ are given in such a way that

$$\max\left(\left\|\widetilde{A}-A\right\|,\left\|\widetilde{B}-B\right\|,\left\|\widetilde{f}-f\right\|\right)\leq\varepsilon,\quad \varepsilon\geq 0. \quad (1.4.96)$$

**Lemma 1.4.19** *Let the condition* (1.4.72) *and the inequalities* (1.4.96) *be valid and*

$$\varepsilon\left(1+|\lambda|\right)\left\|(A+\lambda B)^{-1}\right\|<1.$$

*Then, the matrix* $\widetilde{C}_\lambda\equiv\widetilde{A}+\lambda\widetilde{B}$ *is invertible and the following estimate is fulfilled:*

$$\left\|\widetilde{C}_\lambda^{-1}\right\|=O(\lambda^{-1}). \quad (1.4.97)$$

**Proof.** Let $C_\lambda = A + \lambda B$; then, $\widetilde{C}_\lambda = C_\lambda \left(E - C_\lambda^{-1}\left(C_\lambda - \widetilde{C}_\lambda\right)\right)$. Since $\left\|C_\lambda^{-1}\left(C_\lambda - \widetilde{C}_\lambda\right)\right\| \le \|C_\lambda^{-1}\| \left(\left\|\widetilde{A} - A\right\| + |\lambda| \left\|\widetilde{B} - B\right\|\right) < 1$, from the Banach theorem on close matrices it follows that the matrix $\widetilde{C}_\lambda$ is invertible. Further, from the inequality $\left\|E - C_\lambda^{-1}\left(C_\lambda - \widetilde{C}_\lambda\right)\right\| \le \text{const}$ for $\lambda \in \Lambda_0$ the relation (1.4.97) follows. The lemma is proved.

**Theorem 1.4.39** *Let the hypotheses of Lemma* 1.4.19 *be fulfilled,* $f \in R(A)$ *and* $\varepsilon/\lambda \to 0$ *as* $\varepsilon, \lambda \to 0$. *Then, the relation* $\widetilde{C}_\lambda^{-1} f \to u(f)$ *as* $\varepsilon, \lambda \to 0$ *is valid, where* $u(f) \equiv A^+ f + F^+ BA^+ f$ *is a solution to equation* (1.4.66).

**Proof.** We have

$$\begin{aligned}\left\|\widetilde{C}_\lambda^{-1}\widetilde{f} - u(f)\right\| \le{}& \left\|\widetilde{C}_\lambda^{-1}\left(\widetilde{f} - f\right)\right\| + \\ &+ \left\|\left(\widetilde{C}_\lambda^{-1} - C_\lambda^{-1}\right) f\right\| + \left\|C_\lambda^{-1} f - u(f)\right\|.\end{aligned}$$

For the first and third terms the following estimates hold:

$$\left\|\widetilde{C}_\lambda^{-1}\left(\widetilde{f} - f\right)\right\| = O\left(\varepsilon/\lambda\right), \ \left\|C_\lambda^{-1} f - u(f)\right\| = O\left(\lambda\right).$$

Since

$$\left\|C_\lambda - \widetilde{C}_\lambda\right\| = O\left(\varepsilon\right), \ \left\|C_\lambda^{-1} f\right\| \le \text{const}$$

and

$$\left(\widetilde{C}_\lambda^{-1} - C_\lambda^{-1}\right) f = \widetilde{C}_\lambda^{-1}\left(C_\lambda - \widetilde{C}_\lambda\right) C_\lambda^{-1} f,$$

by Lemma 1.4.19 we obtain

$$\left\|\left(\widetilde{C}_\lambda^{-1} - C_\lambda^{-1}\right) f\right\| = O\left(\varepsilon/\lambda\right).$$

The theorem is proved.

## 1.5 Methods for solving singular integral equations with Hilbert kernels

**1.** Here we consider singular integral equations with Hilbert kernels written in the form of the operator equation

$$Ax(t) = f(t), \quad 0 \le t \le 2\pi, \tag{1.5.98}$$

where the operator $A$ is of the form

$$Ax = Tx + K_1 x + K_2 x. \tag{1.5.99}$$

Here $T$ is the Hilbert integral operator

$$Tx(t) \equiv \frac{1}{2\pi} \int_0^{2\pi} \cot \frac{t-s}{2}\, x(s)\, ds \tag{1.5.100}$$

and $K_1$ and $K_2$ are the regular parts of the operator $A$:

$$K_1 x(t) \equiv \frac{1}{2\pi} \int_0^{2\pi} k_1\,(t-s)\, x(s)\, ds, \tag{1.5.101}$$

$$K_2 x(t) \equiv \frac{1}{2\pi} \int_0^{2\pi} k_2\,(t+s)\, x(s)\, ds, \tag{1.5.102}$$

where

$$k_m(t) = a_m \cos t + b_m \sin t + c_m \tag{1.5.103}$$

and the coefficients $a_m$, $b_m$, and $c_m$, $m = 1, 2$, are complex.

Equation (1.5.98) with the operator $A$ defined by the equality (1.5.99) arises in the theory of elasticity [34], aerodynamics [33], [36], and the theory of cracks [224].

Equations (1.5.98) are studied in [205] and [35] for the case when one or both the terms $K_1$ and $K_2$ are omitted in the representation (1.5.99) of the operator $A$.

In this subsection we study the questions of solvability and, in particular, of unique solvability for equation (1.5.98) in the space $H_\alpha$ of $2\pi$-periodic Hölder functions with exponent $\alpha$: $0 < \alpha < 1$. The answer to these question we seek in the form of conditions imposed on $a_m$, $b_m$, and $c_m$, $m = 1, 2$, which are the coefficients of functions (1.5.103).

Along with the space $H_\alpha$ we consider the space $L_2$ of square integrable $2\pi$-periodic functions measurable according to Lebesgue.

The following lemma is valid.

**Lemma 1.5.20** *Let the function* $f(t) \equiv \sum_{k=-\infty}^{\infty} f_k e^{ikt}$ *belong to the space* $H_\alpha$. *Then, for any integer* $k_0$ *the function* $g(t) = \sum_{k=k_0}^{\infty} f_k e^{ikt}$ *also belongs to the space* $H_\alpha$.

**Proof.** Let us consider the two cases: 1) $k_0 = 0$ and 2) $k_0 \neq 0$.

*Case* 1. Putting $\varsigma = e^{it}$ and $\varphi(\varsigma) \equiv f(t)$, we have

$$\varphi(\varsigma) = \sum_{k=-\infty}^{\infty} f_k \left(e^{it}\right)^k \equiv \sum_{k=-\infty}^{\infty} f_k \varsigma^k. \tag{1.5.104}$$

For $|z| < 1$ and $|\varsigma| = 1$ the equality

$$\frac{1}{\varsigma - z} = \varsigma^{-1} \sum_{m=0}^{\infty} \frac{z^m}{\varsigma^m} \tag{1.5.105}$$

holds. Substituting (1.5.104) and (1.5.105) into the Cauchy integral

$$\Phi\left(z\right) = \frac{1}{2\pi i} \int\limits_{|\varsigma|=1} \frac{\varphi(\varsigma) d\varsigma}{\varsigma - z}, \quad |z| < 1,$$

we obtain the equality

$$\begin{aligned} \Phi\left(z\right) &= \frac{1}{2\pi i} \int\limits_{|\varsigma|=1} \sum_{k=-\infty}^{\infty} f_k \varsigma^k \varsigma^{-1} \sum_{m=0}^{\infty} \frac{z^m}{\varsigma^m}\, d\varsigma = \\ &= \sum_{k=-\infty}^{\infty} f_k \sum_{m=0}^{\infty} z^m \frac{1}{2\pi i} \int\limits_{|\varsigma|=1} \varsigma^{k-m-1}\, d\varsigma. \end{aligned} \tag{1.5.106}$$

Since

$$\frac{1}{2\pi i} \int\limits_{|\varsigma|=1} \varsigma^{k-m-1} d\varsigma = \begin{cases} 1 & if \quad k = m, \\ 0 & if \quad k \neq m, \end{cases}$$

from (1.5.106) we come to the equality

$$\Phi\left(z\right) = \sum_{k=0}^{\infty} f_k z^k, \quad |z| < 1.$$

By the theorem of Plemel–Privalov [205], the function $\Phi(z)$ is continuously extendable as $|z| \to 1$ with preservation of Hölder properties. Hence, the function $\Phi\left(\varsigma\right)$, $|\varsigma| = 1$, is a Hölder one and belongs to $H_\alpha$.

*Case* 2. Let $k_0 \neq 0$. For definiteness, we assume that $k_0 < 0$. Then, the function $g(t)$ can be represented in the form

$$g(t) = \sum_{k=k_0}^{-1} f_k e^{ikt} + \sum_{k=0}^{\infty} f_k e^{ikt}. \tag{1.5.107}$$

According to the first part of the proof, the second sum in the right-hand side of (1.5.107) belongs to $H_\alpha$, whereas the first one is Lipschitzian. Therefore, $g(t) \in H_\alpha$. The lemma is proved.

Information on the spectrum of $A$ plays an important part in the study of the question of solvability of equation (1.5.98). This information is given in the lemma below.

**Lemma 1.5.21** *The following equalities are valid:* $\sigma_{L_2}(A) = \sigma_{H_\alpha}(A) = \sigma$. *Here* $\sigma_{L_2}(A)$ *and* $\sigma_{H_\alpha}(A)$ *are the spectra of the operator* $A$ *in the spaces* $L_2$ *and* $H_\alpha$, *respectively, and*

$$\sigma = \left\{ \pm i,\ c_1 + c_2,\ \frac{1}{2}\left[a_1 \pm \sqrt{a_2^2 + b_2^2 - (b_1+2)^2}\right]\right\}.$$

**Proof.** Let us find $\sigma_{L_2}(A)$. To accomplish this, we consider the parametric family of equations

$$Ax(t) - \lambda x(t) = f(t) \tag{1.5.108}$$

dependent on the complex number $\lambda$. Show that if $\lambda \notin \sigma$, then for any function $f(t) \in L_2$, equation (1.5.108) has a unique solution from the space $L_2$. Let

$$x(t) = \sum_{k=-\infty}^{\infty} x_k e^{ikt}. \tag{1.5.109}$$

The following equalities are valid:

$$\begin{aligned}
Tx(t) &= -i \sum_{k=-\infty}^{\infty} x_k \, sign\, k e^{ikt}, \\
K_1 x(t) &= \frac{1}{2}(a_1 + ib_1)\, x_{-1} e^{-it} + c_1 x_0 + \frac{1}{2}(a_1 - ib_1)\, x_1 e^{-it}, \\
K_2 x(t) &= \frac{1}{2}(a_2 + ib_2)\, x_1 e^{-it} + c_2 x_0 + \frac{1}{2}(a_2 - ib_2)\, x_{-1} e^{it}.
\end{aligned}$$

Thus, by virtue of (1.5.99) we have

$$\begin{aligned}
Ax(t) = i &\sum_{k=-\infty}^{-2} x_k e^{ikt} - i \sum_{k=2}^{\infty} x_k e^{ikt} + \\
&+ \left[\frac{1}{2}(a_1 + ib_1 + i2)\, x_{-1} + \frac{1}{2}(a_2 + ib_2)\, x_1\right] e^{-it} + \\
&+ (c_1 + c_2)\, x_0 + \left[\frac{1}{2}(a_2 - ib_2)\, x_{-1} + \frac{1}{2}(a_1 - ib_1 - i2)\right] e^{it}.
\end{aligned} \tag{1.5.110}$$

Let us expand the right-hand side $f(t)$ of equation (1.5.108) in the Fourier series:

$$f(t) = \sum_{k=-\infty}^{\infty} f_k e^{ikt}. \tag{1.5.111}$$

A solution $x(t)$ to equation (1.5.108) is sought in the form (1.5.107). Then, by virtue of (1.5.110) and (1.5.111) we obtain the infinite system for determining the unknown coefficients of the function $x(t)$:

$$\begin{aligned}
&(i-\lambda)\, x_k = f_k, \quad k \le -2,\\
&\frac{1}{2}\,(a_1 + ib_1 + i2 - 2\lambda)\, x_{-1} + \frac{1}{2}\,(a_2 + ib_2)\, x_1 = f_{-1},\\
&(c_1 + c_2 - \lambda)\, x_0 = f_0,\\
&\frac{1}{2}\,(a_2 - ib_2)\, x_{-1} + \frac{1}{2}\,(a_1 - ib_1 - i2 - 2\lambda)\, x_1 = f_1,\\
&(-i-\lambda)\, x_k = f_k, \quad k \ge 2.
\end{aligned} \tag{1.5.112}$$

If $\lambda \notin \sigma$, then the system (1.5.112) is uniquely solvable and the solution has the form

$$\begin{aligned}
x_k &= \frac{1}{i-\lambda}\, f_k, \quad k \le -2,\\
x_{-1} &= \frac{2}{\Delta(\lambda)}\left[(a_1 - ib_1 - i2 - 2\lambda)\, f_{-1} - (a_2 + ib_2)\, f_1\right],\\
x_0 &= \frac{1}{c_1 + c_2 - \lambda}\, f_0,\\
x_1 &= \frac{2}{\Delta(\lambda)}\left[(a_2 - ib_2)\, f_{-1} + (a_1 + ib_1 + i2 - 2\lambda)\, f_1\right],\\
x_k &= -\frac{1}{i+\lambda}\, f_k, \quad k \ge 2,
\end{aligned} \tag{1.5.113}$$

where

$$\Delta(\lambda) \equiv (2\lambda - a_1)^2 + (b_1 + 2)^2 - a_2 - b_2^2.$$

From (1.5.109) and (1.5.113) it follows the equality

$$\begin{aligned}
x(t) = {}& \frac{1}{i-\lambda} \sum_{k=-\infty}^{-2} f_k e^{ikt} - \frac{1}{i+\lambda} \sum_{k=2}^{\infty} f_k e^{ikt} +\\
&+ \frac{2}{\Delta(\lambda)}\left[(a_1 - ib_1 - i2 - 2\lambda)\, f_{-1} - (a_2 + ib_2)\, f_1\right] e^{-it} +\\
&+ \frac{1}{c_1 + c_2 - \lambda}\, f_0 + \frac{2}{\Delta(\lambda)}\left[-\,(a_2 - ib_2)\, f_{-1} + \right.\\
&\left. + \,(a_1 + ib_1 + i2 - 2\lambda)\, f_1\right] e^{it},
\end{aligned} \tag{1.5.114}$$

which shows the validity of the inclusion $x(t) \in L_2$ if $f(t) \in L_2$.

Further, if $\lambda \notin \sigma$ and $f(t) \in H_\alpha$, by Lemma 1.5.20 the terms

$$\frac{1}{i-\lambda}\sum_{k=-\infty}^{-2} f_k e^{ikt}, \quad -\frac{1}{i+\lambda}\sum_{k=2}^{\infty} f_k e^{ikt}$$

in the right-hand side of (1.5.114) belong to $H_\alpha$, whereas the three middle terms are the Lipschitzian functions. Hence, $x(t) \in H_\alpha$.

To complete the proof of Lemma 1.5.21, it remains to note that if $\lambda \in \sigma$, then by virtue of system (1.5.112) we have the inclusions $\lambda \in \sigma_{L_2}(A)$ and $\lambda \in \sigma_{H_\alpha}(A)$. The lemma 1.5.21 is proved.

The following theorem is valid.

**Theorem 1.5.40** *In order that equation* (1.5.98) *be uniquely solvable for all* $f(t) \in H_\alpha$, *it is necessary and sufficient that the relations*

$$c_1 + c_2 \neq 0, \quad a_1^2 + (b_1+2)^2 - a_2^2 - b_2^2 \neq 0 \tag{1.5.115}$$

*be fulfilled, where* $a_m$, $b_m$, *and* $c_m$, $m = 1, 2$, *are the coefficients of the functions* (1.5.103).

**Proof.** *Necessity.* Let equation (1.5.98) be uniquely solvable for all $f(t) \in H_\alpha$. Then, the system (1.5.112) is uniquely solvable for $\lambda = 0$. This means that $0 \notin \sigma$. The validity of (1.5.115) follows from Lemma 1.5.21.

*Sufficiency.* Let the relations (1.5.115) take place. From here and from Lemma 1.5.21 it follows that $0 \notin \sigma$. Hence, the system (1.5.112) is uniquely solvable for $\lambda = 0$ and for all $f(t) \in H_\alpha$. The theorem is proved.

If one or both inequalities (1.5.115) are violated in the hypotheses of Theorem 1.5.40, for the solvability or unique solvability of equation (1.5.98) it is necessary to require the fulfillment of extra conditions both for the right-hand side and for the sought-for solution. The enumeration of all allowable cases is possible, but too cumbersome. Let us consider some of them.

**Assertion 1.5.1** *Let*

$$c_1 + c_2 = 0, \quad a_1^2 + (b_1+2)^2 - a_2^2 - b_2^2 \neq 0. \tag{1.5.116}$$

*The necessary and sufficient condition for the solvability of equation* (1.5.98) *is the equality*

$$\int_0^{2\pi} f(t)\, dt = 0; \tag{1.5.117}$$

*if the condition* (1.5.117) *is valid, then for each complex number* $d_0$ *there exists only a one solution* $x(t)$ *to equation* (1.5.98) *satisfying the condition*

$$\int_0^{2\pi} x(t)\,dt = d_0. \tag{1.5.118}$$

**Assertion 1.5.2** *If*

$$c_1 + c_2 \neq 0, \quad a_1^2 + (b_1+2)^2 - a_2^2 - b_2^2 = 0, \tag{1.5.119}$$

*then the necessary and sufficient conditions for the solvability of equation* (1.5.98) *are the equalities*

$$(a_2 - ib_2)\int_0^{2\pi} f(t)e^{it}\,dt - (a_1 + ib_1 + i2)\int_0^{2\pi} f(t)e^{-it}\,dt = 0, \tag{1.5.120}$$

$$(a_1 - ib_1 - i2)\int_0^{2\pi} f(t)e^{it}\,dt - (a_2 + ib_2)\int_0^{2\pi} f(t)e^{-it}\,dt = 0. \tag{1.5.121}$$

**2.** Now we consider the problem of numerical solution of the singular integral equation (1.5.98) with the singular operator $A$ defined by the equality (1.5.99); this operator acts in $H_\alpha$. For this purpose, we replace equation (1.5.98) by the system of linear algebraic equations

$$A_N x_N = f_N, \tag{1.5.122}$$

where $A_N : C^N \to C^N$, $x_N$ and $f_N$ are the sought-for and given $N$-vectors, respectively. The matrix $A_N$ and the vector $f_N$ are defined below.

Let us define the matrix $A_N$. To accomplish this, we consider the following two uniform grids for each $N \geq 2$ on the segment $[0, 2\pi]$:

$$s_m^{(N)} = s_0^{(N)} + \frac{2\pi}{N}m, \quad t_m^{(N)} = s_m^{(N)} + \frac{\pi}{N}, \quad m = 1, 2, \ldots, N. \tag{1.5.123}$$

Here $s_0^{(N)} \in [-2\pi/N, -\pi/N]$ is an arbitrary point. Put

$$A_N \equiv T_N + K_{1N} + K_{2N},$$

where $T_N$ and $K_{1N}$ are circulant matrices and $K_{2N}$ is a percirculant matrix [58]; the first rows of these matrices respectively are

$$\begin{aligned}
&\left[\frac{1}{N}\cot\frac{\xi_0^{(N)}}{2}, \frac{1}{N}\cot\frac{\xi_1^{(N)}}{2}, \dots, \frac{1}{N}\cot\frac{\xi_{N-1}^{(N)}}{2}\right],\\
&\left[\frac{1}{N}k_1\left(\xi_0^{(N)}\right), \frac{1}{N}k_1\left(\xi_1^{(N)}\right), \dots, \frac{1}{N}k_1\left(\xi_{N-1}^{(N)}\right)\right],\\
&\left[\frac{1}{N}k_2\left(\eta_0^{(N)}\right), \frac{1}{N}k_2\left(\eta_1^{(N)}\right), \dots, \frac{1}{N}k_2\left(\eta_{N-1}^{(N)}\right)\right],
\end{aligned} \tag{1.5.124}$$

where $\xi_m^{(N)} = t_1^{(N)} - s_{m+1}^{(N)}$, $\eta_m^{(N)} = t_1^{(N)} + s_{m+1}^{(N)}$, $m = 0, 1, \dots, N-1$, and the functions $k_1(t)$ and $k_2(t)$ are given by the equalities (1.5.103).

Let us consider the operators $\varphi_N$ and $\psi_N$ that associate each function $x(t)$ from $H_\alpha$ with the following grid functions on the grids (1.5.123):

$$\varphi_N x(t) = \left[x\left(s_1^{(N)}\right), x\left(s_2^{(N)}\right), \dots, x\left(s_N^{(N)}\right)\right]^T, \tag{1.5.125}$$

$$\psi_N x(t) = \left[x\left(t_1^{(N)}\right), x\left(t_2^{(N)}\right), \dots, x\left(t_N^{(N)}\right)\right]^T. \tag{1.5.126}$$

In what follows, the right-hand side of (1.5.122) is defined by the equality

$$f_N \equiv \psi_N f(t), \quad f(t) \in H_\alpha. \tag{1.5.127}$$

It is proved in [35] that if $x(t) \in H_\alpha$, then the estimate

$$\left|(Tx)\left(t_m^{(N)}\right) - (T_N x_N)_m\right| = O\left(N^{-\alpha}\ln N\right) \tag{1.5.128}$$

is valid. Here $T$ is the singular operator (1.5.100), $T_N$ is a circulant matrix with its first row (1.5.124), and $(T_N x_N)_m$ is the $m$th component of the vector

$$x_N = \varphi_N x(t). \tag{1.5.129}$$

For the regular parts (1.5.101) and (1.5.102) of the operator (1.5.99), the estimate

$$\begin{gathered}\left|(K_n x)\left(t_m^{(N)}\right) - (K_{nN} x_N)_m\right| = O(N^{-\alpha}),\\ n = 1, 2; \quad m = 1, \dots, N\end{gathered} \tag{1.5.130}$$

is valid for any $x(t) \in H_\alpha$.

The approximation estimate

$$\|(A_N\varphi_N - \psi_N A)\,x(t)\|_{L_2^{(N)}} = O(N^{-\alpha}\ln N), \quad x(t) \in H_\alpha, \tag{1.5.131}$$

follows from (1.5.125)–(1.5.130), where

$$\|y\|_{L_2^{(N)}} \equiv \sqrt{\frac{1}{N}\sum_{m=1}^{N}|y_m|^2}, \quad y = [y_1, y_2, \dots, y_N],$$

is a discrete analogue of the norm in $L_2$.

**Theorem 1.5.41** *Let the relations* (1.5.115) *be fulfilled. Then, the system* (1.5.122) *is uniquely solvable for all the right-hand sides* $f_n$ *defined by the equality* (1.5.127) *and the following estimate is valid:*

$$\left\| \left(A_N^{-1}\psi_N - \varphi_N A^{-1}\right) f(t) \right\|_{L_2^{(N)}} = O(N^{-\alpha}\ln N). \tag{1.5.132}$$

*Here* $\varphi_N$ *and* $\psi_N$ *are the operators* (1.5.125) *and* (1.5.126).

**Proof.** Let us put

$$B_N = F_N^* A_N F_N, \tag{1.5.133}$$

where $F_N$ is the matrix of the discrete Fourier transform. By virtue of the properties of circulant and percirculant matrices, we have

$$F_N^* K_{1N} F_N = \operatorname{diag}\left[c_1, (a_1 + ib_1)\, s,\, 0, \dots, 0, (a_1 - ib_1)\,\overline{s}\right], \tag{1.5.134}$$

$$F_N^* K_{2N} F_N = \operatorname{pd}\left[c_2, (a_2 + ib_2)\, c, 0, \dots, 0, (a_2 - ib_2)\,\overline{c}\right], \tag{1.5.135}$$

where $s = \dfrac{1}{2}\, e^{-i\pi/N}$, $c = \dfrac{1}{2}\, e^{i\left(2\Delta_0^{(N)} + \pi/N\right)}$, and

$$\operatorname{pd}(a_0, a_1, \dots, a_N) = \begin{bmatrix} a_0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & a_1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_N & \cdots & 0 \end{bmatrix}.$$

Therefore, we have the equality

$$B_N = \begin{bmatrix} c_1 + c_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & (a_1 + ib_1 + i2)s & 0 & \cdots & 0 & (a_2 + ib_2)c \\ 0 & 0 & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{N-2} & 0 \\ 0 & (a_2 - ib_2)\overline{c} & 0 & \cdots & 0 & (a_1 - ib_1 - i2)\overline{s} \end{bmatrix},$$

where $\lambda_m = e^{i\alpha_m}$ and $\alpha_m = \pi/2 - m\pi/N$, $m = 2, \ldots, N-2$. From this equality we obtain the estimate

$$\left\| B_N^{-1} \right\| \le M, \tag{1.5.136}$$

where

$$M = \max\left\{1, |c_1 + c_2|^{-1}, |\omega_1|^{-1/2}, |\omega_2|^{-1/2}\right\},$$

$$\omega_{1,2} = \frac{1}{8}\left(|a_1|^2 + |b_1 + 2|^2 + |a_2|^2 + |b_2|^2\right) \pm \frac{1}{4}\sqrt{\omega},$$

$$\omega = 2\left(|a_1|^2 + |b_1 + 2|^2\right)\left(|a_2|^2 + |b_2|^2\right) + 2\left[a_1\left(\overline{b}_1 + 2\right) - \overline{a}_1\left(b_1 + 2\right)\right] \times$$
$$\times\left(a_2\overline{b}_2 - \overline{a}_2 b_2\right) + \left[a_1^2 + (b_1 + 2)^2\right]\left(\overline{a}_2^2 + \overline{b}_2^2\right) + \left[\overline{a}_1^2 + \left(\overline{b}_1 + 2\right)^2\right]\left(a_2^2 + b_2^2\right).$$

Since $\|F_N\| = \|F_N^*\| = 1$, then from (1.5.133) and (1.5.136) it follows the estimate

$$\left\|A_N^{-1}\right\| = \left\|F_N B_N^{-1} F_N^*\right\| \le \|F_N\| \left\|B_N^{-1}\right\| \|F_N^*\| \le M.$$

Passing to norm in the equality

$$\left(A_N^{-1}\psi_N - \varphi_N A^{-1}\right) f(t) = A_N^{-1}\left(\psi_N A - A_N \varphi_N\right) A^{-1} f(t),$$

from (1.5.131) we come to the validity of estimate (1.5.132). The theorem is proved.

Let the relations (1.5.116) take place. By virtue of Assertion 1.5.1, the relations (1.5.117) and (1.5.118) are necessary and sufficient conditions for the unique solvability of equation (1.5.98). For definiteness, we assume that in (1.5.118) $d_0 = 0$, i.e., we consider the system

$$Ax(t) = f(t), \quad \frac{1}{2\pi}\int\limits_0^{2\pi} x(t)\,dt = 0. \tag{1.5.137}$$

Let us rewrite the system (1.5.137) in the form of the single operator equation:

$$Bx(t) = g(t), \quad B : H_\alpha \to H_\alpha \times C. \tag{1.5.138}$$

Here

$$Bx(t) = \begin{bmatrix} Ax(t) \\ \dfrac{1}{2\pi}\displaystyle\int\limits_0^{2\pi} x(t)\,dt \end{bmatrix}, \quad g(t) = \begin{bmatrix} f(t) \\ 0 \end{bmatrix}. \tag{1.5.139}$$

Note that by virtue of Assertion 1.5.1, the operator $B$ acts from $H_\alpha^0$ to $H_\alpha^0 \times \{0\}$ and is boundedly invertible. Here $H_\alpha^0 \subset H_\alpha$ is the subspace of functions with zero mean value.

Along with (1.5.125) and (1.5.126) we consider the grid operators

$$\overline{\varphi}_N x(t) = \begin{bmatrix} \varphi_N x(t) \\ 0 \end{bmatrix}, \quad \overline{\varphi}_N : H_\alpha \to C^{N+1}, \tag{1.5.140}$$

$$\overline{\psi}_N \begin{bmatrix} x(t) \\ c \end{bmatrix} = \begin{bmatrix} \psi_N x(t) \\ c \end{bmatrix}, \quad \overline{\psi}_N : H_\alpha \times C \to C^{N+1}. \tag{1.5.141}$$

We replace equation (1.5.138) by the system of linear algebraic equations

$$B_N \overline{x}_N = \overline{f}_N, \tag{1.5.142}$$

where

$$B_N = \begin{bmatrix} A_N & 0 \\ a_N^* & 0 \end{bmatrix}, \quad \overline{x}_N = \begin{bmatrix} x_N \\ x_N^0 \end{bmatrix},$$

$$a_N^* = \left[\frac{1}{\sqrt{N}}, \cdots, \frac{1}{\sqrt{N}}\right] \in C^N, \quad \overline{f}_N = \overline{\psi}_N g(t).$$

Now we show that the operator $B$ defined by equality (1.5.139) and the matrix $B_N$ from (1.5.143) are related by

$$\left\| \left(B_N \overline{\varphi}_N - \overline{\psi}_N B\right) x(t) \right\|_{L_2^{(N+1)}} = O\left(N^{-\alpha} \ln N\right) \tag{1.5.143}$$

for all $x(t) \in H_\alpha^0$. Indeed, since

$$\left(B_N \overline{\varphi}_N - \overline{\psi}_N B\right) x(t) = \begin{bmatrix} \left(A_N \varphi_N - \psi_N A\right) x(t) \\ a_N^* \varphi_N x(t) \end{bmatrix},$$

then, passing to norm, we obtain the estimate

$$\begin{aligned} &\left\| \left(B_N \overline{\varphi}_N - \overline{\psi}_N B\right) x(t) \right\|_{L_2^{(N+1)}} = \\ &= \sqrt{\frac{1}{N+1} \left[ \sum_{k=1}^N \left(\psi_N A x(t) - A_N \varphi_N x(t)\right)_k^2 + \left(a_N^* \varphi_N x(t)\right)^2 \right]} \le \\ &\le \sqrt{\frac{1}{N} \sum_{k=1}^N \left(\psi_N A x(t) - A_N \varphi_N x(t)\right)_k^2} + \frac{1}{N} \left| \sum_{k=1}^N x\left(s_k^{(N)}\right) \right| = \\ &= \left\| \left(\psi_N A - A_N \varphi_N\right) x(t) \right\|_{L_2^{(N)}} + \frac{1}{N} \left| \sum_{k=1}^N x\left(s_k^{(N)}\right) \right|. \end{aligned}$$

The estimate (1.5.131) is valid for the first term of the right-hand side of the last inequality, whereas for the second term the following estimate holds:

$$\frac{1}{N}\left|\sum_{k=1}^{N} x\left(s_k^{(N)}\right)\right| = O\left(N^{-\alpha}\right),$$

since the mean value of the function $x(t)$ is equal to zero. Hence, the validity of (1.5.143) is proved.

**3.** The linear system (1.5.142) is overdetermined and, in general, is inconsistent. In order to eliminate this inconsistency and overdetermination for the case $A \equiv T$ (i.e., $k_1 \equiv k_2 \equiv 0$, where $T$, $K_1$, and $K_2$ are the operators (1.5.100)–(1.5.102)), an additional unknown called the regularizing factor is introduced in [35].

The shift regularization discussed in Section 1.4 is used to solve the linear system (1.5.142).

By direct check it is possible to ensure that the vectors

$$\begin{aligned} \overline{e}_n &= [0,\ldots,0,1]^T \in C^{N+1}, \\ \overline{g}_n &= \left[\frac{1}{\sqrt{N}},\cdots,\frac{1}{\sqrt{N}},0\right]^T \in C^{N+1} \end{aligned} \tag{1.5.144}$$

satisfy the equality $B_N\overline{e}_N = B_N^*\overline{g}_N$.

The linear system (1.5.142) is replaced by the equation

$$D_N\overline{x}_N = h_N, \tag{1.5.145}$$

where $D_N \equiv B_N + \overline{g}_N\overline{e}_N^*$ and $h_N \equiv \overline{f}_N - \overline{g}_N\overline{g}_N^*\overline{f}_N$.

**Theorem 1.5.42** *Let $g(t) \in H_\alpha^0 \times \{0\}$. Then the following estimate is valid:*

$$\left\|D_N^{-1}h_N - \overline{\varphi}_N B^{-1}g(t)\right\|_{L_2^{(N+1)}} = O(N^{-\alpha}\ln N). \tag{1.5.146}$$

**Proof.** Note that the vector $h_N$ belongs to the image of the matrix $B_N$, whereas the matrix $D_N$ is invertible. The unique solution to (1.5.145) coincides with the normal pseudosolution to (1.5.142), i.e.,

$$D_N^{-1}h_N = B_N^+\overline{f}_N. \tag{1.5.147}$$

To solve the system (1.5.145), we may use the method of fast Fourier transform. Let us put

$$\Phi_N = \begin{bmatrix} F_N & 0 \\ 0 & 1 \end{bmatrix}, \tag{1.5.148}$$

where $F_N$ is the Fourier matrix. Along with (1.5.145), we also consider the system

$$\begin{aligned} &\Phi_N^* D_N \Phi_N \overline{y}_N = \Phi_N^* h_N, \\ &\overline{x}_N = \Phi_N \overline{y}_N. \end{aligned} \tag{1.5.149}$$

From the first equation of (1.5.149) we find $\overline{y}_N$. Since

$$\Phi_N^* D_N \Phi_N = \begin{bmatrix} F_N^* A_N F_N & F_N^* a_N \\ a_N^* F_N & 0 \end{bmatrix}, \tag{1.5.150}$$

$$F_N^* a_N = [1, 0, \ldots, 0]^T, \quad a_N^* F_N = [1, 0, \ldots, 0], \tag{1.5.151}$$

from (1.5.150) and (1.5.151) it follows the equality

$$\Phi_N^* D_N \Phi_N =$$

$$= \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & (a_1 + ib_1 + i2)\, s & 0 & \cdots & 0 & (a_2 + ib_2)\, c & 0 \\ 0 & 0 & \lambda_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{N-2} & 0 & 0 \\ 0 & (a_2 - ib_2)\, \overline{c} & 0 & \cdots & 0 & (a_1 - ib_1 - i2)\, \overline{s} & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}. \tag{1.5.152}$$

The following equality holds for the right-hand side of the first equation in (1.5.145):

$$\Phi_N^* h_N = [0, (F_N^* f_N)_2, (F_N^* f_N)_3, \ldots, (F_N^* f_N)_N, 0]^T. \tag{1.5.153}$$

Therefore, from (1.5.152) and (1.5.153) we obtain

$$y_{k,N} = \overline{\lambda}_{k-1} (F_N^* f_N)_k, \quad k = 3, 4, \ldots, N-1; \quad y_{1,N} = y_{0,N} = 0.$$

The remaining unknowns $y_{2,N}$ and $y_{N,N}$ can be found from the second-order system

$$\begin{aligned} &[(a_1 + ib_1 + i2)\, s]\, y_{2,N} + [(a_2 + ib_2)\, c]\, y_{N,N} = (F_N^* f_N)_2, \\ &[(a_2 - ib_2)\, \overline{c}]\, y_{2,N} + [(a_1 - ib_1 - i2)\, \overline{s}]\, y_{N,N} = (F_N^* f_N)_N. \end{aligned} \tag{1.5.154}$$

It is easy to show that the matrix of (1.5.154) is not singular.

If the linear system (1.5.154) is solved, then we can find $\overline{x}_N$ by substituting the vector $\overline{y}_N$ into the second equation of (1.5.149). Note that the solution $\overline{x}_N$ to the linear system (1.5.149) is also a solution to (1.5.145). By virtue of this and the equality (1.5.147), the above-obtained vector $\overline{x}_N$ is the normal pseudosolution to the linear system (1.5.142). As above, it is possible to prove the estimate

$$\left\|D_N^{-1}\right\| \leq M < \infty \tag{1.5.155}$$

for all $N$. Passing to norm in the equality

$$D_N^{-1} h_N - \varphi_N B^{-1} g(t) = D_N^{-1}\left[\overline{\psi}_N B - B_N \overline{\varphi}_N\right] B^{-1} g(t) - D_N^{-1} \overline{g}_N \overline{g}_N^{*} \overline{\psi}_N g(t),$$

from (1.5.143), (1.5.155) and from the relation

$$\left\|\overline{g}_N \overline{g}_N^{*} \overline{\psi}_N g(t)\right\|_{L_2^{(N+1)}} = O(N^{-\alpha})$$

we come to the validity of the estimate (1.5.146). The theorem is proved.

# Chapter 2

# METHODS OF DIGITAL IMAGE RESTORATION AND ILL-POSED PROBLEMS OF AUTOMATIC CONTROL

## 2.1 Computerized systems of image processing

**1.** The digital processing of images is applied in many areas of science and technology, such as space research, crystallography, optics, roentgenoscopy, archival photodocument renewal, etc. In a broad sense, the image processing is understood as certain transformations of signals that are essentially two-dimensional in their nature. The main aim of digital image processing consists in restoration (reconstruction, recovery) of the true (whenever possible) source image with minimum distortion by means of special mathematical methods. A comprehensive survey of existing methods for digital image restoration is given in [213], [214].

The basic mathematical model of the image restoration problem is a two-dimensional convolution-type integral equation of the first kind, i.e., an integral equation with a kernel depending on the difference of arguments [225]. The equations of this type describe many other problems arising in various areas of research (for example, restoration of two-dimensional distributions of brightness across a radio-astronomical source, resolution of multiplets in complicated two-dimensional spectra, etc.). Therefore, the development of methods for solving the two-dimensional convolution-type equations is of considerable interest in modern research.

The solution of integral equations with difference kernels is an ill-posed problem; therefore, the regularization method [293] is of fundamental importance for the construction of noise-resistant approximations. In particular, an analysis of optimal and quasioptimal recovery algorithms for noisy images is given in [227]. Optimal regularizing algorithms and the theory of accuracy for approximate solutions to two-dimensional convolution-type integral equations are developed in [228]. Numerical methods for solving such equations are discussed, for example, in [228]–[230], where regularizing algorithms for special classes of two-dimensional convolution-type equations are constructed in the case when a kernel admits the separation of variables or possesses the circular symmetry.

The problems of quality improvement of smeared and out-of-focus images (these problems can be reduced to solving equations with special-type kernels [14], [16]) are important particular cases of processing the images distorted by a linear homogeneous system.

For computerized image processing, as is known, an image is coded with scanners, i.e., is divided into $N_1 \times N_2$ small elements (squares, pixels) and is represented in the form of a matrix $A\,(N_1, N_2)$ whose elements are optical densities of the corresponding image elements. Scanners of different kind exist for technical realization of this procedure. To express the brightness of image elements (pixels), the $2^B$ quantified levels are introduced, where $B$ is the number of bits per element; hence, a coded image is represented by $N_1 \times N_2 \times B$ bits.

The specific features of image recovery problems are: 1) instability (these problems belong to the class of ill-posed problems); 2) the necessity to process large amounts of digital data in core memory or in external memory devices (magnetic tapes or disks).

**2.** A subroutine library of a computerized system for image processing (LCSIP) was developed at the Numerical Analysis Section of Moscow University Research Computing Center. The aim of this library is to provide the users with software for the solution of basic digital image recovery problems and for the development and application of stable fast algorithms intended for the processing of large amounts of information with and without external storage.

The first edition of the LCSIP library contains more than thirty subroutines grouped in accordance with the following topics:

1. Fast transforms. Analysis and synthesis of series.
2. Integral equations.

3. Elementary operations on images.

4. Restoration of images distorted by a linear homogeneous system.

5. Digital filtration.

6. Service subroutines.

Modernization, development and periodic issuing of new extended editions of the Library is carried out according to the progress in the field of digital image processing methods.

The work with the LCSIP library assumes that the following conditions are imposed on the images to be processed:

1) brightness of images is digitized into 256 levels, i.e., $B = 8$ bits (one byte) per pixel, the elements of the image matrix $A(N_1, N_2)$ are integers from 0 up to 255;

2) when kept in external storage, an image is given in the form of a file consisting of $N_1$ records $N_2$ bytes each, i.e., each record contains one line of the image;

3) the number $N_2$ of elements in a line of the image is no more than $\left(2^{15} - 1\right)$.

4) the number $N_1$ of lines in the image is no more than $\left(2^{15} - 1\right)$.

The LCSIP library is supplied with a manual which contains an introduction, an analysis and illustrations of some numerical results, a catalog of subroutines, and the unified descriptions of all the subroutines with brief discussions of underlying methods. The algorithms implemented for the Library are described in detail in [7]–[16].

All the library subroutines are written in FORTRAN-IV. The subroutines operating with external storage are oriented to the use of the IBM data format and of the image record format compatible with the image processing device "Optronics".

The subroutines from the library sections "Fast transforms. Analysis and synthesis of series" and "Integral equations" are written in standard FORTRAN and are included into the Numerical Analysis Library developed at the Research Computing Center of Moscow State University for BESM-6, IBM-360, and "Elbrus" computer systems. These subroutines are used to solve various research problems in many scientific centers of Russia (USSR) and other countries.

The LCSIP library is widely used to solve the problems of reconstruction of archival photodocuments [201].

**3.** In what follows, a brief description of the LCSIP library sections is given.

The section "*Fast transforms. Analysis and synthesis of series*" contains high-performance subroutines realizing various versions of algorithms for the fast Fourier transform (FFT) and for the fast Walsh transform (FWT). In particular, these subroutines compute: the discrete Fourier transform (DFT) for a complex series segment whose length is equal to a power of two or is even, the DFT for a real series and its Fourier coefficients in the system of even or odd trigonometric functions, the DFT for a complex series with a prescribed number of nonzero terms, a required number of DFT coefficients for a complex series of even length, and the two-dimensional DFT for a complex matrix.

Some FFT-based algorithms for computation of various special convolutions (one-dimensional and two-dimensional, both periodic and aperiodic) as well as algorithms for the correlation and spectral analysis of stationary processes are also included into this section. For example, it is possible to construct smoothed approximations of autospectra and mutual spectra by the method of modified periodograms or by averaging in frequencies and to find the estimates of autocorrelations, mutual correlations, mutual amplitude spectra, the functions of phase differences, and the coherence functions [10], [11].

Various methods of image processing are, to some extent, based on these algorithms (in particular, the methods of statistical processing of stationary stochastic processes when their realizations are given). The FFT and FWT algorithms can also be used to code the images in order to reduce their redundancy.

The section "*Integral equations*" contains several very efficient subroutines for numerical solution of one-dimensional and two-dimensional convolution-type integral equations in infinite domains. Their algorithms are based on the regularization method proposed by A. N. Tikhonov and on the FFT algorithm. Two options are possible for construction of regularized approximations: one of them is based on a fixed value of regularization parameter, while the other one is based on its value chosen according to the residual criterion. The quality of a regularized solution is estimated by the four criterial functions which can be used for the choice of regularization parameter: the residual, the solution norm, the regularizing functional, and the "sensitivity" function. The algorithms for solution of integral equations with difference kernels whose Fourier transforms are given analytically are also implemented.

The section "*Elementary operations on images*" contains a set of subroutines implementing the most necessary operations on images recorded in coded form on a disk or a tape, such as the separation of a fragment of an image, insertion of a fragment in an image, etc.

The section "*Restoration of images distorted by a linear homogeneous system*" contains subroutines for reconstruction of smeared and defocused images with and without the use of external storage. It is assumed that a distorted image is given in core memory in the form of a matrix of real numbers or is recorded on a disk or a tape in byte form. The smearing and defocusing parameters are assumed to be known. The algorithms for restoration of images are based on the solution of one-dimensional and two-dimensional convolution-type integral equations by A. N. Tikhonov's regularization method with application of FFT; the subroutines from the first three sections of the Library are used.

Due to the use of FFT, the above-discussed algorithms for image recovery are highly efficient in the sense of their speed, the saving of storage space, and the reduction of round-off errors. The necessary amount of core memory without the use of external storage is $\sim 2N_1N_2+4N_2$ for the smeared image recovery and $\sim 4N_1N_2$ for the defocused image recovery; the number of arithmetic operations is proportional to $N_1\,(N_2\log_2 N_2)$ and $N_1N_2\,(\log_2 N_1+\log_2 N_2)$, respectively, where $N_1$ is the number of lines and $N_2$ is the number of elements in a line of a distorted image.

The section "*Digital filtration*" contains a number of subroutines for one-dimensional linear filtration of processes in a frequency domain on the basis of FFT and in a time domain by recursive or nonrecursive filters. These algorithms are widely used in the digital image processing (in particular, for reconstruction of noisy images with periodic noises [15]).

The section "*Service subroutines*" contains auxiliary subroutines used by subroutines from other sections of the LCSIP library. For example, the subroutines for transformation of an image recorded on a disk or a tape from its byte form to its numerical form or from its numerical form to its byte form, the subroutines for recording the resulting image on a disk or a tape, the subroutines for transposition of a matrix of real numbers given on a disk, the subroutine for evaluation of coefficients of recursive or nonrecursive filters when they are given analytically on some segment of their frequency domain.

**4.** The efficiency of the algorithms considered above was confirmed by numerical experiments on various model problems of spectral and correlation analysis for the solution of convolution-type integral equations, for modeling of filters, digital filtering and restoration of distorted images. A detailed analysis of experiments and illustrations of typical numerical results are given in [15], [16].

Note that the subroutines for image recovery in core memory can process the matrices of maximum dimension $64\times128$ on a BESM-6 computer and of maximum dimension $128 \times 128$ on an IBM-22 computer for smeared images and, respectively, $64 \times 64$ and $64 \times 128$ for defocused images. The computing time for restoration of defocused images of size $64 \times 64$ with a one value of regularization parameter and with computation of criterial functions is about 15 seconds on a BESM-6 computer and $\sim 45$ seconds on an IBM-22 computer. The time for restoration of smeared images of size $512 \times 512$ with external storage on an IBM-22 computer is $\sim 40$ minutes.

Our numerical experiments have shown that the above fast regularizing algorithms provide the efficient solution of the restoration problems for smeared or defocused images with optimal values of regularization parameter (generally speaking, these values are not equal to zero). These algorithms are stable with respect to approximation errors, round-off errors, noises of quantization, and additive random disturbances of a distorted image.

## 2.2 Fast algorithms of digital image recovery

**1.** The aim of this section is to describe a high-performance computing algorithm for the solution of two-dimensional convolution-type integral equations with random noise in their kernels and right-hand sides.

A regularizing algorithm for the solution of this problem was constructed. This algorithm is based on Tikhonov's one-parametric regularization method and on the two-dimensional discrete Fourier transform [196]. The algorithm of two-dimensional fast Fourier transform (FFT) [196] is used to optimally compute the direct and inverse Fourier transforms. Due to FFT, our numerical algorithm for the solution of convolution-type equations is efficient in the sense of speed of operation, the saving of storage space, and the reduction of round-off errors; the total number of arithmetic operations is proportional to $N_1N_2(\log N_1 + \log N_2)$, whereas the amount of storage is proportional to $N_1N_2$, where $N_1$ and $N_2$ are the grid point numbers along the coordinate axes for each of the two variables (the original data are given at these points). The algorithm we developed is intended for processing of large amounts of information, which is typical for image reconstruction problems.

A similar algorithm for the solution of one-dimensional equations with difference kernels is considered in [8], [13]. The subroutines implementing the above algorithms are written in standard FORTRAN.

Numerical experiments based on a number of model problems were performed to check the efficiency of the algorithms. Our analysis of numerical results we obtained is illustrated below by tables and figures.

Let us consider the following two-dimensional convolution-type integral equation of the first kind on the plane $R^2 = (-\infty, +\infty) \times (-\infty, +\infty)$:

$$Af \equiv \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} k(x-\xi, y-\eta) f(\xi, \eta)\, d\xi\, d\eta = g(x,y), \qquad (2.2.1)$$

where $-\infty < x, y < +\infty$, $k \in L_1(R^2)$ is a kernel dependent on the difference of arguments and $f \in L_2(R^2)$ is the unique exact solution corresponding to an exact right-hand side $g \in L_2(R^2)$. By the multiplication theorem for convolution, the sought-for solution $f$ can be represented in the form

$$f(\xi, \eta) = \Phi^{-1} F = \frac{1}{(2\pi)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{G(\lambda, \omega)}{K(\lambda, \omega)} e^{i(\lambda\xi + \omega\eta)} d\lambda\, d\omega,$$
$$i = \sqrt{-1}, \quad -\infty < \xi, \eta < +\infty,$$

where $\Phi$ is the Fourier–Plancherel operator [40], $F$, $K$, and $G$ are the two-dimensional Fourier transforms (Fourier-images) for the functions $f$, $k$, and $g$ defined, for example in the case of $f$, by the formula

$$F(\lambda, \omega) = \Phi f \equiv \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(\xi, \eta) e^{-i(\lambda\xi + \omega\eta)}\, d\xi\, d\eta, \qquad (2.2.2)$$
$$-\infty < \lambda, \omega < +\infty.$$

It is assumed that the kernel $k$ and the right-hand side $g$ in (2.2.1) are given approximately as follows:

$$\widetilde{k} = k + n_k \in L_1, \quad \widetilde{g} = g + n_g \in L_2.$$

Here $n_k$ and $n_g$ are additive noises (disturbances) from some allowable classes $\{n_k\}$ and $\{n_g\}$. It is well known that the problem of solving convolution-type integral equations of the first kind in the presence of errors in the operators and in the right-hand sides is ill-posed. For the construction of noiseproof approximations, we apply the Tikhonov method of one-parametric regularization, which reduces the problem to the minimization of the following smoothing functional in $\widetilde{f}$:

$$M^\alpha\left[\widetilde{f}; \widetilde{g}\right] \equiv \left\| \widetilde{A}\widetilde{f} - \widetilde{g} \right\|_{L_2}^2 + \alpha \Omega\left[\widetilde{f}\right]. \qquad (2.2.3)$$

Here

$$\left\| \widetilde{A}\widetilde{f} - \widetilde{g} \right\|_{L_2}^2 \equiv \frac{1}{(2\pi)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left| \widetilde{K}(\lambda, \omega) \widetilde{F}(\lambda, \omega) - \widetilde{G}(\lambda, \omega) \right|^2 d\lambda\, d\omega$$

is the square of the residual, $\alpha > 0$ is the regularization parameter,

$$\Omega\left[\widetilde{f}\right] \equiv \frac{1}{(2\pi)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[1 + \left(\lambda^2 + \omega^2\right)^p\right] \left| \widetilde{F}(\lambda, \omega) \right|^2 d\lambda\, d\omega$$

is a stabilizing functional of order $p \geq 0$ ($p$ is not supposed to be natural), $\widetilde{K}$, $\widetilde{F}$, and $\widetilde{G}$ are the Fourier images of the functions $\widetilde{k}$, $\widetilde{f}$, and $\widetilde{g}$ defined similarly to (2.2.2).

**Remark 4** *For a natural value of $p$, the application of the regularizator $R(\lambda,\omega) \equiv 1 + \left(\lambda^2+\omega^2\right)^p$ in the frequency domain is equivalent to the construction of a solution in the space $W_2^p\left[R^2\right]$ and the following relation is valid:*

$$\Omega\left[\widetilde{f}\right] \equiv \left\|\widetilde{f}\right\|^2_{W_2^p} = \left\|\widetilde{f}\right\|^2_{L_2} + \sum_{p_1=0}^{p} \frac{p!}{p_1!(p-p_1)!} \left\| \frac{\partial^p \widetilde{f}}{\partial \xi^{p_1} \partial \eta^{(p-p_1)}} \right\|^2_{L_2}.$$

The functional (2.2.3) for $\alpha$ fixed is minimized by the function

$$\widetilde{f}^{\alpha}(\xi,\eta) = \frac{1}{(2\pi)^2} \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} \frac{\widetilde{K}^*(\lambda,\omega)\,\widetilde{G}(\lambda,\omega)\,e^{i(\lambda\xi+\omega\eta)}\,d\lambda\,d\omega}{\left|\widetilde{K}(\lambda,\omega)\right|^2 + \alpha\left[1+\left(\lambda^2+\omega^2\right)^p\right]}, \qquad -\infty < \xi,\, \eta < +\infty, \tag{2.2.4}$$

where $^*$ indicates complex conjugation; this function is adopted as a regularized solution to the problem. The parameter $\alpha$ may be chosen in accordance with some well-known criteria, for example, according to the residual criterion or may be put to be equal to its quasioptimal value. Thus, the problem of solving the convolution-type equation (2.2.1) with an approximate kernel and an approximate right-hand side is reduced to the triple calculation of the Fourier transform (including the inverse Fourier transform) and to the application of the regularization method in the frequency domain with an appropriate choice of the regularization parameter $\alpha$.

**2.** Let us write down a discrete analogue of the problem. It is assumed that $k(x,y) \to 0$, $f(\xi,\eta) \to 0$, $g(x,y) \to 0$ as $x, y, \xi, \eta \to \pm\infty$ and that the functions $\widetilde{k}(x,y)$, $\widetilde{g}(x,y)$ are given on a half-segment $T_1$ in the first variable and on a half-segment $T_2$ in the second variable (i.e., $x \in [-T_1/2, T_1/2]$ and $y \in [-T_2/2, T_2/2]$) at the nodes of the uniform grids in $x$ and $y$:

$$\begin{aligned} x_{s_1} &= s_1 d_1, \quad s_1 = \overline{-N_1/2,\ N_1/2-1}, \\ y_{s_2} &= s_2 d_2, \quad s_2 = \overline{-N_2/2,\ N_2/2-1}. \end{aligned} \tag{2.2.5}$$

Here $N_1$ is an even number of nodes, $d_1 = T_1/N_1$ is the grid step in $x$, $N_2$ is an even number of nodes, $d_2 = T_2/N_2$ is the grid step in $y$, $T_1$ and $T_2$ are the sufficiently large parameters of the problem. Thus, the convolution-type equation (2.2.1) with the discrete kernel $\widetilde{k}_{s_1,s_2} = \widetilde{k}(x_{s_1}, y_{s_2})$ and the discrete right-hand side $\widetilde{g}_{s_1,s_2} = \widetilde{g}(x_{s_1}, y_{s_2})$ should be solved.

The regularized solutions $\widetilde{f}^{\alpha}_{s_1,s_2} = \widetilde{f}^{\alpha}\left(\xi_{r_1}, \eta_{r_2}\right)$ are constructed on the same grid (2.2.5) at the nodes

$$\xi_{r_1} = r_1 d_1, \quad r_1 = \overline{-N_1/2,\ N_1/2-1}, \quad \eta_{r_2} = r_2 d_2, \quad r_2 = \overline{-N_2/2,\ N_2/2-1}.$$

Approximating the integrals in the Fourier transforms $\widetilde{K}$, $\widetilde{G}$, and $\widetilde{F}$ in the equalities (2.2.1), (2.2.3), (2.2.4) by the rectangular formula, we obtain the discrete analogues of the Fourier transforms for

a) the kernel:

$$\widetilde{K}\left(\lambda_{m_1}, \omega_{m_2}\right) \equiv \widetilde{K}_{m_1,m_2} = d_1 d_2 \sum_{s_1=-N_1/2}^{N_1/2-1} \sum_{s_2=-N_2/2}^{N_2/2-1} \widetilde{k}_{s_1,s_2} e^{-2\pi i (s_1 m_1/N_1 + s_2 m_2/N_2)},$$

where

$$\begin{aligned} \lambda_{m_1} &= m_1 \Delta\lambda, \quad m_1 = \overline{-N_1/2,\ N_1/2-1}, \quad \Delta\lambda = 2\pi/N_1 d_1, \\ \lambda_{m_2} &= m_2 \Delta\omega, \quad m_2 = \overline{-N_2/2,\ N_2/2-1}, \quad \Delta\omega = 2\pi/N_2 d_2; \end{aligned}$$

b) the convolution equation:

$$d_1 d_2 \sum_{r_1=-N_1/2}^{N_1/2-1} \sum_{r_2=-N_2/2}^{N_2/2-1} k_{s_1-r_1,s_2-r_2} f_{r_1,r_2} = g_{s_1,s_2},$$
$$s_1 = \overline{-N_1/2,\ N_1/2-1}, \quad s_2 = \overline{-N_2/2,\ N_2/2-1};$$

c) the smoothing functional:

$$\frac{1}{N_1 N_2 d_1 d_2} \sum_{m_1=-N_1/2}^{N_1/2-1} \sum_{m_2=-N_2/2}^{N_2/2-1} \left( \left| \widetilde{K}_{m_1,m_2} \widetilde{F}_{m_1,m_2} - \widetilde{G}_{m_1,m_2} \right|^2 + \right.$$
$$\left. + \alpha R_{m_1,m_2} \left| \widetilde{F}_{m_1,m_2} \right|^2 \right),$$

where $R_{m_1,m_2} = 1 + \left(\lambda_{m_1}^2 + \omega_{m_2}^2\right)^p$;

d) the regularized solution minimizing this functional:

$$\widetilde{f}^{\alpha}_{s_1,s_2} = \frac{1}{N_1 N_2 d_1 d_2} \sum_{m_1=-N_1/2}^{N_1/2-1} \sum_{m_2=-N_2/2}^{N_2/2-1} \frac{\widetilde{K}^{*}_{m_1,m_2} \widetilde{G}_{m_1,m_2}}{\left( \left| \widetilde{K}_{m_1,m_2} \right|^2 + \alpha R_{m_1,m_2} \right)} \times$$
$$\times e^{2\pi i (r_1 m_1/N_1 + r_2 m_2/N_2)},$$
$$r_1 = \overline{-N_1/2,\ N_1/2-1}, \quad r_2 = \overline{-N_2/2,\ N_2/2-1}.$$

It is assumed that for $p = 0$, $m_1 = 0$, and $m_2 = 0$ the relations $\left(\lambda_{m_1}^2 + \omega_{m_2}^2\right)^p = 1$ and $R_{m_1,m_2} = 2$ hold.

**3**. Our numerical algorithm for solving the convolution-type equations (2.2.1) with discrete kernels and discrete right-hand sides by the regularization method is based on the application of the two-dimensional discrete Fourier transform (DFT).

**Definition 6** *By the discrete* (*two-dimensional*) *Fourier transform for a function* $u_{s_1,s_2}$, $s_1 = \overline{0, N_1-1}$, $s_2 = \overline{0, N_2-1}$, *we mean the function*

$$U_{m_1,m_2} = \sum_{s_1=0}^{N_1-1} \sum_{s_2=0}^{N_2-1} u_{s_1,s_2} e^{\pm 2\pi i (s_1 m_1/N_1 + s_2 m_2/N_2)}, \tag{2.2.6}$$

*where the sign "$-$" in the exponent corresponds to the direct two-dimensional DFT, whereas the sign "$+$" corresponds to the inverse two-dimensional DFT.*

For efficient calculation of sums of the form (2.2.6), the algorithm of two-dimensional fast Fourier transform (FFT) [16] can be used.

Let $\widehat{K}_{m_1,m_2}$, $\widehat{G}_{m_1,m_2}$, and $\widehat{F}^{\alpha}_{m_1,m_2}$ be the DFTs of the functions

$$\begin{aligned} \widehat{k}_{s_1,s_2} &= \widetilde{k}_{s_1-N_1/2,\, s_2-N_2/2}, \quad \widehat{g}_{s_1,s_2} = \widetilde{g}_{s_1-N_1/2,\, s_2-N_2/2}, \\ \widehat{f}^{\alpha}_{s_1,s_2} &= \widetilde{f}^{\alpha}_{s_1-N_1/2,\, s_2-N_2/2}, \quad s_1 = \overline{0, N_1-1}, \quad s_2 = \overline{0, N_2-1}. \end{aligned}$$

It is easy to check that the formula for the above regularized solution $\widehat{f}^{\alpha}_{r_1,r_2}$ can be rewritten in the form

$$\begin{aligned} \widehat{f}^{\alpha}_{r_1,r_2} &= \frac{1}{N_1 N_2} \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} (-1)^{m_1+m_2} F^{\alpha}_{m_1,m_2} e^{2\pi i (r_1 m_1/N_1 + r_2 m_2/N_2)}, \\ r_1 &= \overline{-N_1/2, N_1/2-1}, \quad r_2 = \overline{-N_2/2, N_2/2-1}, \end{aligned} \tag{2.2.7}$$

where

$$\begin{aligned} (-1)^{m_1+m_2} F^{\alpha}_{m_1,m_2} &\equiv \widehat{F}^{\alpha}_{m_1,m_2}, \quad F^{\alpha}_{m_1,m_2} = \frac{1}{d_1 d_2} \frac{\widehat{K}^*_{m_1,m_2} \widehat{G}_{m_1,m_2}}{\left|\widehat{K}_{m_1,m_2}\right|^2 + \dfrac{\alpha}{d_1^2 d_2^2} \widehat{R}_{m_1,m_2}}, \\ m_1 &= \overline{0, N_1-1}, \quad m_2 = \overline{0, N_2-1}; \end{aligned}$$

$$\widehat{R}_{m_1,m_2} = \begin{cases} R_{m_1,m_2}, & m_1 = \overline{0, N_1/2}, \qquad\qquad m_2 = \overline{0, N_2/2}, \\ R_{N_1-m_1,m_2}, & m_1 = \overline{N_1/2+1, N_1-1}, \ m_2 = \overline{0, N_2/2}, \\ R_{m_1,N_2-m_2}, & m_1 = \overline{0, N_1/2}, \qquad\qquad m_2 = \overline{N_2/2+1, N_2-1}, \\ R_{N_1-m_1,N_2-m_2}, & m_i = \overline{N_i/2+1, N_i-1}, \quad i = 1, 2; \end{cases}$$

$\widehat{R}_{m_1,m_2}$ is the periodic continuation of the function $R_{m_1,m_2}$ with period $N_1$ in the first coordinate and with period $N_2$ in the second coordinate.

For the software implementation of the algorithm, thus, the original problem is reduced to its canonical form by calculation of FFT for the kernel $\widehat{K}_{m_1,m_2}$ and for the right-hand side $\widehat{G}_{m_1,m_2}$; then, according to the formula (2.2.7), the inverse FFT for the function $\widehat{F}^{\alpha}_{m_1,m_2}$ is taken.

The quality of the regularized solution $\widehat{f}^{\alpha}_{r_1,r_2}$ for a given $\alpha$ is estimated by the following four criterial functions:

1) the residual

$$\rho(\alpha) = \left\| \widetilde{A}\widetilde{f}^{\alpha} - \widetilde{g} \right\|_{L_2} \approx \left( \frac{d_1 d_2}{N_1 N_2} \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} \left| d_1 d_2 \widehat{K}_{m_1,m_2} F^{\alpha}_{m_1,m_2} - \widehat{G}_{m_1,m_2} \right|^2 \right)^{1/2},$$

2) the $\Omega$-norm of the solution

$$\gamma(\alpha) = \Omega^{1/2}\left[\widetilde{f}^{\alpha}\right] \approx \left( \frac{d_1 d_2}{N_1 N_2} \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} \widehat{R}_{m_1,m_2} \left| F^{\alpha}_{m_1,m_2} \right|^2 \right)^{1/2},$$

3) the regularizing functional

$$\varphi(\alpha) = \left(\rho^2(\alpha) + \alpha\gamma^2(\alpha)\right)^{1/2},$$

4) the "sensitivity" function

$$\tau(\alpha) = \alpha\Omega^{1/2}\left[\frac{d\widetilde{f}^{\alpha}}{d\alpha}\right] \approx$$

$$\approx \left[ \frac{1}{N_1 N_2 d_1^3 d_2^3} \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} \frac{\widehat{R}^3_{m_1,m_2} \left| F^{\alpha}_{m_1,m_2} \right|^2}{\left( \left| \widehat{K}_{m_1,m_2} \right|^2 + \frac{\alpha}{d_1^2 d_2^2} \widehat{R}_{m_1,m_2} \right)^2} \right]^{1/2}.$$

These criterial functions can be used to choose the regularization parameter $\alpha$. The calculation of all the criterial functions is performed in the frequency domain (in terms of DFT-components), which requires $\sim N_1 N_2$ operations and considerably simplifies the problem of choosing $\alpha$. The use of inverse DFT in the formula (2.2.7) is reasonable only for a chosen value of $\alpha$.

The algorithm allows us to reduce the problem to its canonical form only once and to use the DFTs obtained for a kernel and a right-hand side in the further calculations. This is important for the efficiency of the algorithm (for example, when the same integral equation is solved repeatedly or when the regularization parameter is chosen, as well as for the repeated solution of an equation with the same kernel but with different right-hand sides or in the case when the Fourier transform of a kernel is known analytically).

**4.** As previously mentioned, an image can be described by a real bounded nonnegative function $f(\xi, \eta)$ of two space variables $\xi$, $\eta$; this function presents the intensity or brightness of the image at each point of the plane $(\xi,\ \eta)$. No optical reproduction system can ensure the perfect quality of images; in reality, therefore, distorted noisy images are obtained. It is necessary to perform a certain correction of an image to improve the quality of its digital representation. The equation (2.2.1) is a basic mathematical model for the problem of recovering images distorted by a linear system which is space-invariant with respect to shift (a homogeneous system). Here $f(\xi, \eta)$ is the original ideal image, $g(x, y)$ is an image of $f$ at outlet from the system, $k(x-\xi, y-\eta)$ is the pulse characteristic of the system or the dispersion function of a point; this function corresponds to the response of the system at a point $(x, y)$ to the two-dimensional $\delta$-function (the point light source) at the point $(\xi, \eta)$.

In any real system, a fuzzy image $g$ is additionally distorted by various noises during transmission or registration; these noises may be considered as additive random noises with known statistical characteristics. In practice, a distorted image $\widetilde{g} = g + n$ is usually given. For example, if $\widetilde{g}$ is an image recorded on a photographic film, then $n$ is the film grain as a noise.

The basic problem of image recovery can be reduced to the construction of an image $\widetilde{f}$ which is maximally close (for example, in the mean-square sense) to the ideal image $f$ under the assumption that the noisy image $\widetilde{g}$, the dispersion function $k$, and *a priori* statistical data on the noise $n$ are given [201].

Important particular problems of image processing are the problems of quality improvement of defocused and smeared images. In the case of two-dimensional defocusing of optical objectives, the dispersion function of a point has the Gaussian form

$$k\,(x-\xi, y-\eta) = \frac{1}{2\pi\beta^2}\, e^{-\left(1/2\beta^2\right)\left[(x-\xi)^2+(y-\eta)^2\right]}, \quad \beta > 0, \tag{2.2.8}$$

or is given as a constant inside a circle of radius $r > 0$:

$$k\,(x-\xi, y-\eta) = \begin{cases} 1/\pi r^2, & (x-\xi)^2 + (y-\eta)^2 \le r^2, \\ \quad 0, & (x-\xi)^2 + (y-\eta)^2 > r^2. \end{cases} \tag{2.2.9}$$

In the case of one-dimensional smearing, the loss of image sharpness occurs in the direction of a one coordinate (for example, as a result of rectilinear motion of the image). The dispersion function admits the separation of variables and is of the form

$$k\,(x-\xi, y-\eta) = k_1\,(x-\xi)\,\delta\,(y-\eta)\,. \tag{2.2.10}$$

In the particular case of uniform rectilinear motion of an image, the function $k_1(x-\xi)$ becomes constant in some strip:

$$k_1(x-\xi) = \begin{cases} 1/(b-a), & a \le x-\xi \le b, \\ 0, & x-\xi < a,\ x-\xi > b. \end{cases} \tag{2.2.11}$$

In the case of recovery of smeared images, the basic equation (2.2.1) is equivalent to the equation

$$\int_{-\infty}^{\infty} k_1(x-\xi)\, f(\xi, y)\, d\xi = g(x,y), \quad -\infty < y < +\infty,$$

i.e., leads to the repeated solution of one-dimensional convolution-type integral equations.

Thus, the above fast regularizing algorithm can be used to solve the problems of recovering the quality of defocused and smeared images; these problems are reduced to the solution of two-dimensional convolution-type integral equations (2.2.1) with kernels of the form (2.2.8)–(2.2.10).

**5.** In order to check the efficiency of the above algorithm for the solution of two-dimensional convolution-type equations (2.2.1) with approximate kernels and right-hand sides, many numerical experiments based on a number of model problems were carried out.

The subroutine which implements the above algorithm uses the FFT subroutine requiring that the grid point numbers $N_1$ and $N_2$ be natural powers of 2; otherwise, the definition of the original matrices should be completed at the grid nodes (2.2.5) by an appropriate number of zeros.

The values of the regularization parameter $\alpha$ are taken on the grid

$$\{\alpha\} = \left\{\alpha : \alpha_i = \theta\alpha_{i-1},\ i = \overline{1,q},\ \alpha_{q+1} = 0\right\},$$

where $\alpha_0$ is the initial value of $\alpha$, $0 < \theta < 1$ is the grid step, and $q$ is the number of grid steps in $\alpha$. Among the family of regularized solutions $\left\{\widetilde{f}^{\alpha}\right\}$, $\alpha \in \{\alpha\}$, the optimal approximation $\widetilde{f}^{\text{opt}} = \widetilde{f}^{\alpha_{\text{opt}}}$ was chosen by searching for the optimal (in the sense defined below) value of the parameter $\alpha = \alpha_{\text{opt}} \ge 0$.

*Model problem* 1. Let us consider a two-dimensional convolution-type integral equation (2.2.1) with the Gaussian kernel $k(x-\xi, y-\eta) = e^{-\left[(x-\xi)^2+(y-\eta)^2\right]}$ of the form (2.2.8). The exact solution $f$ and the exact right-hand side $g$ of the equation are as follows:

$$f(\xi,\eta) = e^{-\left(\xi^2+\eta^2\right)},\ g(x,y) = (\pi/2)\, e^{-\left(x^2+y^2\right)/2}.$$

As an optimal solution, we choose the approximation $\widetilde{f}^{\text{opt}}$ which is the best in the sense of minimum in $\alpha \in \{\alpha\}$ for the uniform error $\gamma_C(\alpha)$ or for the mean-square error $\gamma_{L_2}(\alpha)$:

$$\begin{aligned}
\gamma_C(\alpha) &= \max_{r_1,r_2} \left| f_{r_1,r_2} - \widetilde{f}^{\alpha}_{r_1,r_2} \right|, \\
\gamma_{L_2}(\alpha) &= \left( (d_1 d_2) \sum_{r_1=-N_1/2}^{N_1/2-1} \sum_{r_2=-N_2/2}^{N_2/2-1} \left| f_{r_1,r_2} - \widetilde{f}^{\alpha}_{r_1,r_2} \right|^2 \right)^{1/2},
\end{aligned}$$

where $f_{r_1,r_2} \equiv f\left(\xi_{r_1}, \eta_{r_2}\right)$ are the discrete values of the known model solution at the grid nodes (2.2.5).

The efficiency of our algorithm was checked for additive random noises $n_k$ and $n_g$ of the kernel $k$ and the right-hand side $g$ of equation (2.2.1); these noises are uniformly distributed in the squares $[-\sigma_k, \sigma_k] \times [-\sigma_k, \sigma_k]$ and $[-\sigma_g, \sigma_g] \times [-\sigma_g, \sigma_g]$. Our computations were carried out for various values of the parameters specifying the original discrete equation (the grid point numbers $N_1$ and $N_2$ and the lengths of the realization segments $T_1$ and $T_2$), the regularizator order $p$, and the amplitudes $\sigma_k$, $\sigma_g$ of noises in the kernel and in the right-hand side of the equation. Some typical numerical results are given in Tables 2.2.1 and 2.2.2.

**Table 2.2.1.** Dependence of the optimal value of $\alpha$ and the optimal errors on the lengths of the realization segments $T_1$, $T_2$, on the numbers $N_1$, $N_2$, and on the regularizator order $p$ (without perturbations in the kernel and in the right-hand side)

<table>
<tr><th>$p$</th><th>$N_1$</th><th>$N_2$</th><th>$T_1$</th><th>$T_2$</th><th>$\alpha_{\text{opt}}$</th><th>$\gamma_C(\alpha_{\text{opt}})$</th><th>$\gamma_{L_2}(\alpha_{\text{opt}})$</th></tr>
<tr><td rowspan="4">1</td><td rowspan="4">64</td><td rowspan="4">64</td><td>4</td><td>4</td><td>1e-2</td><td>2e-1</td><td>2e-1</td></tr>
<tr><td>10</td><td>10</td><td>1e-8</td><td>3e-4</td><td>8e-4</td></tr>
<tr><td>20</td><td>20</td><td>1e-13</td><td>6e-7</td><td>9e-7</td></tr>
<tr><td>50</td><td>50</td><td>1e-15</td><td>1e-2</td><td>2e-2</td></tr>
<tr><td rowspan="4">1</td><td>32</td><td>32</td><td rowspan="4">10</td><td rowspan="4">10</td><td>1e-8</td><td>4e-4</td><td>1e-3</td></tr>
<tr><td>64</td><td>64</td><td>1e-8</td><td>3e-4</td><td>8e-4</td></tr>
<tr><td>32</td><td>128</td><td>1e-8</td><td>4e-4</td><td>9e-4</td></tr>
<tr><td>16</td><td>256</td><td>1e-8</td><td>7e-4</td><td>1e-3</td></tr>
<tr><td>0</td><td rowspan="5">64</td><td rowspan="5">64</td><td rowspan="5">10</td><td rowspan="5">10</td><td>1e-7</td><td>4e-4</td><td>1e-3</td></tr>
<tr><td>0.5</td><td>1e-7</td><td>3e-4</td><td>6e-4</td></tr>
<tr><td>1</td><td>1e-8</td><td>3e-4</td><td>8e-4</td></tr>
<tr><td>2</td><td>1e-9</td><td>4e-4</td><td>5e-4</td></tr>
<tr><td>5</td><td>1e-14</td><td>3e-4</td><td>7e-4</td></tr>
</table>

It follows from Table 2.2.1 that the optimal accuracy and the optimal value $\alpha_{\text{opt}}$ of regularization parameter essentially depend on the lengths of the realization segments $T_1$ and $T_2$. For small values of these quantities, an error due to replacement of infinite limits of integration by finite ones has a noticeable effect, whereas for their very large values the main effect is caused by a too large step of discretization. The dependence of $\alpha_{\text{opt}}$ and of an error in the solution on the grid point numbers $N_1$ and $N_2$ (for some average lengths of the realization segments) is considerably less, which can be explained by the sufficient smoothness of the functions under consideration. For this model problem, the optimal accuracy of resulting approximations is weakly dependent on $p$, although $\alpha_{\text{opt}}$ substantially decreases with increasing $p$.

**Table 2.2.2.** Dependence of the optimal value of $\alpha$ and the optimal errors on perturbations in the kernel and in the right-hand side

| | $\sigma_k \cdot 100\%$ | $\sigma_g \cdot 100\%$ | $\alpha_{\text{opt}}$ | $\gamma_C(\alpha_{\text{opt}})$ | $\gamma_{L_2}(\alpha_{\text{opt}})$ |
|---|---|---|---|---|---|
| | | 0 | 1e-8 | 3e-4 | 8e-4 |
| | 0 | 0.1% | 1e-5 | 8e-3 | 2e-2 |
| $N_1$=64 | | 1% | 1e-4 | 2e-2 | 7e-2 |
| $N_2$=64 | | 5% | 1e-3 | 5e-2 | 1e-1 |
| $T_1$=10 | 0.1% | | 1e-7 | 7e-4 | 2e-2 |
| $T_2$=10 | 1% | 0 | 1e-6 | 4e-3 | 1e-2 |
| $p$=1 | 5% | | 1e-7 | 2e-2 | 6e-2 |
| | 0.1% | 0.1% | 1e-5 | 1e-8 | 5e-2 |
| | 1% | 1% | 1e-4 | 1e-8 | 8e-2 |
| | 0.1% | 10% | 1e-3 | 1e-14 | 3e-1 |

Table 2.2.2 shows that when the perturbations of the kernel and the right-hand side in the integral equation increase, the values of $\alpha_{\text{opt}}$ and the errors of the optimal solution also increase; the optimal solution has a satisfactory accuracy even for sufficiently large noises in the kernel and in the right-hand side. The lesser dependence of the optimal accuracy on perturbations of the kernel compared to perturbations of the right-hand side is somewhat unexpected.

*Model problem* 2. The problem of recovering the quality of noisy smeared and defocused images was solved on the basis of the above fast regularizing algorithm. The images were described by two-dimensional functions with their values in the range $[0, 1]$. The brightness of all images was digitized at 256 levels, i.e., 8 bits per readout. The convolution-type equations (2.2.1) with the kernel (2.2.9) for defocused images and with the kernels (2.2.10), (2.2.11) for smeared images were considered. The original test image $f$ was the figure "**5**" given discretely by a $64 \times 64$-matrix. The image was two-tone. Each point inside the image was associated with 1, each point outside the image was associated with 0. The one-dimensional smearing was modeled by the averaging of each element of the original image over a horizontal segment of length $l$ ($l$ is a parameter of smearing) with its beginning at the point under consideration, whereas the two-dimensional defocusing was modeled by the averaging of each element of the original image over a circle of radius $l$ ($l$ is the parameter of defocusing) with its center at a given point.

The distorted image $\widetilde{g}$ to be recovered was chosen in the form $\widetilde{g} = g + n + n_1$, where $g$ is the fuzzy image obtained as a result of smearing or defocusing of the original image, $n$ is the modeled normally distributed centralized random noise with its amplitude $\sigma$, and $n_1$ is the noise of quantization; if a value of $\widetilde{g}$ does not belong to the range $[0, 1]$, then we put $\widetilde{g} = 0$ when $\widetilde{g} < 0$ and $\widetilde{g} = 1$ when $\widetilde{g} > 1$.

Our numerical experiments were carried out for various values of the smearing and defocusing parameter $l$ at various levels of the noise $\sigma$. The regularizator order $p$ was fixed ($p = 0$). The values of the resulting solution $\widetilde{f}^{\alpha}$ were reduced to the segment $[0, 1]$ by a linear transformation. The optimal value $\alpha_{\text{opt}}$ on the grid $\alpha \in \{\alpha\}$ was visually chosen by comparison of quality of the recovered images $\widetilde{f}^{\alpha}$ with the original image. Some typical numerical results are illustrated in the photos that are presented in Figures 2.2.1–2.2.5.

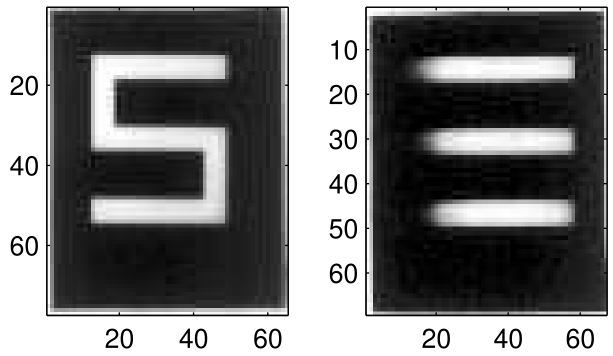


Figure 2.2.1: The original image: the first photo; the smeared image ($l = 41$, $\sigma = 0$): the second photo

The case $\alpha = 0$ corresponds to the image recovery without regularization. The photos were obtained with an instrumental complex for image processing at the Research Center of Technical Documentation (Russian Federation).

## 2.3 Fast algorithms of signal recovery

**1.** A broad spectrum of problems arising in applications can be reduced to solving convolution-type integral equations of the first kind (for example, the inverse problems of automatic control, optics, nuclear spectroscopy, gravimetry, spectrometry, analytical continuation, and many others).

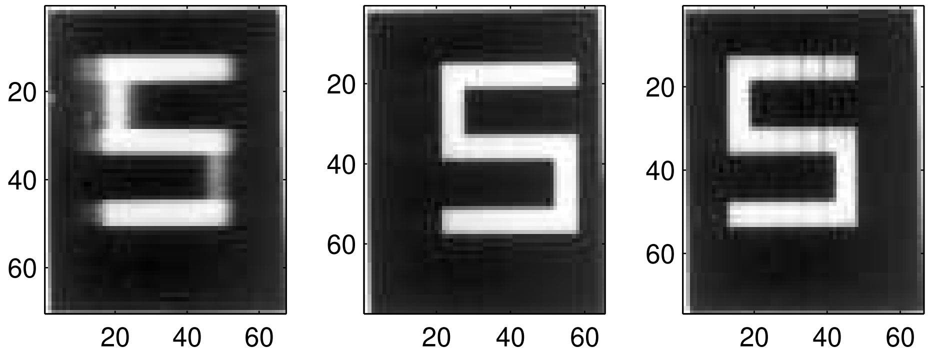


Figure 2.2.2: Recovery of smeared images with $l = 41$, $\sigma = 0$ for various values of $\alpha$ ($\alpha = 10^{-2}$: the first photo; $\alpha_{\text{opt}} = 4 \cdot 10^{-5}$: the second photo; $\alpha = 0$: the third photo)

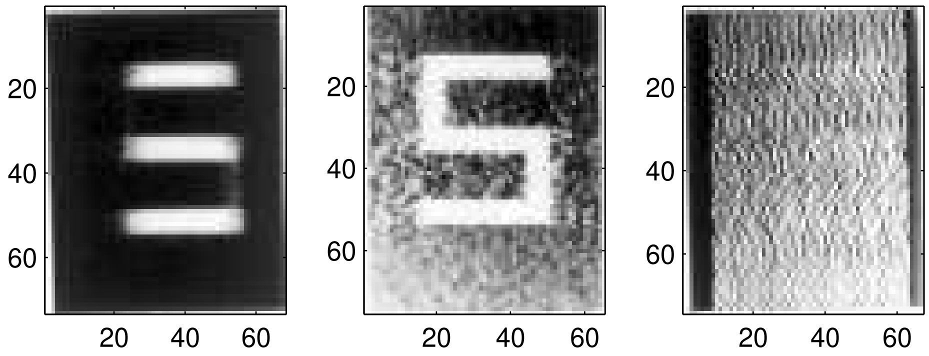


Figure 2.2.3: Recovery of smeared images with $l = 41$, $\sigma = 0.04$ for various values of $\alpha$ ($\alpha = 10^{-2}$: the first photo; $\alpha_{\text{opt}} = 10^{-3}$: the second photo; $\alpha = 0$: the third photo)

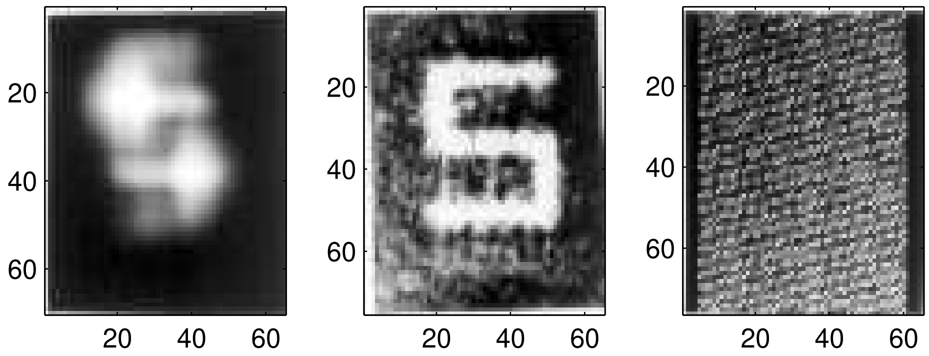


Figure 2.2.4: The first photo: the defocused image with $l = 10$, $\sigma = 0$; the second and third photos: the recovered defocused images with $l = 10$, $\sigma = 0.04$ for $\alpha_{\text{opt}} = 10^{-3}$ and $\alpha = 0$

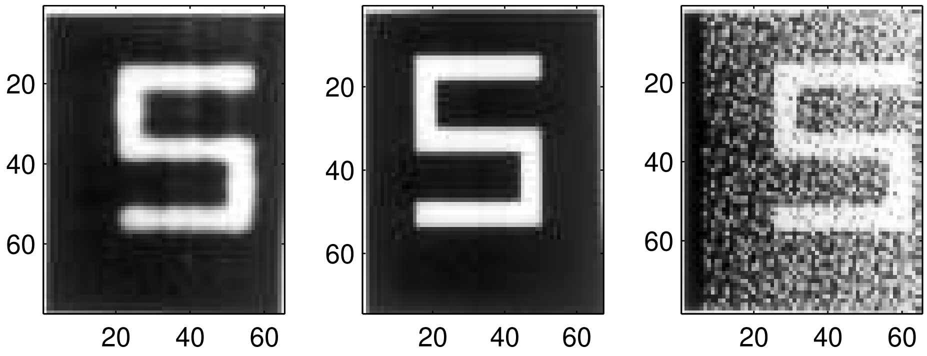


Figure 2.2.5: Recovery of defocused images with $l = 10$, $\sigma = 0.0$; the first photo: $\alpha = 10^{-3}$; the second photo: $\alpha_{\text{opt}} = 10^{-5}$; the third photo: $\alpha = 0$

Let us discuss some methods for solving one-dimensional convolution-type integral equations of the first kind with random noises from the standpoint of development of fast numerical algorithms. Taking into account the specific property of these equations (dependence of their kernels on the difference of arguments), we construct an efficient regularizing algorithm based on the use of the regularization method and the method of fast (discrete) Fourier transform [196]. The regularization parameter is chosen according to the residual criterion [157] and according to the computing scheme proposed in [194]. We also analyze numerical results obtained on the basis of various model problems.

Let us consider the following one-dimensional convolution-type integral equation of the first kind in the space $L_2\,(-\infty, +\infty)$:

$$Ax \equiv \int_{-\infty}^{+\infty} k\,(t-\tau)\,x(\tau)\,d\tau = y(t), \quad -\infty < t < +\infty. \tag{2.3.12}$$

Here the kernel $k(t)$ belongs to $L_1\,(-\infty, +\infty)$. It is assumed that for some exact right-hand side $y(t) \in L_2$ there exists a unique exact solution $x(t) \in L_2$. By the multiplication theorem for convolution, this solution can be represented in the form

$$x(t) = F^{-1}X = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{Y(\lambda)}{K(\lambda)}\, e^{i\lambda t}\, d\lambda, \quad -\infty < t < +\infty,$$

where $F$ is the Fourier–Plancherel operator, whereas $K(\lambda)$, $X(\lambda)$, and $Y(\lambda)$ are the Fourier transforms (the Fourier images) of the functions $k$, $x$, and $y$, respectively.

It is further assumed that the kernel $k$ and the right-hand side $y$ of equation (2.3.12) are given approximately as

$$\widetilde{k}(t) = k(t) + n_k(t) \in L_1, \quad \widetilde{y}(t) = y(t) + n_y(t) \in L_2,$$

where $n_k$ and $n_y$ are additive noises from some allowable classes $\{n_k\}$ and $\{n_y\}$. It is well known [265] that the problem of solving convolution-type integral equations with errors in their operators and right-hand sides is ill posed. In order to construct stable approximations, we apply the regularization method [255], which reduces the problem to the minimization of the following smoothing functional in $\tilde{x}$:

$$M^{\alpha}\left[\widetilde{x};\widetilde{y}\right] \equiv \left\|\widetilde{A}\widetilde{x} - \widetilde{y}\right\|_{L_2}^{2} + \alpha\Omega\left[\widetilde{x}\right]. \tag{2.3.13}$$

Here $\left\|\widetilde{A}\widetilde{x} - \widetilde{y}\right\|_{L_2}^{2} \equiv \dfrac{1}{2\pi}\displaystyle\int_{-\infty}^{+\infty}\left|\widetilde{K}(\lambda)\,\widetilde{X}(\lambda) - \widetilde{Y}(\lambda)\right|^{2} d\lambda$ is the square of the residual, $\alpha > 0$ is the regularization parameter, $\Omega\left[\widetilde{x}\right] \equiv \dfrac{1}{2\pi}\displaystyle\int_{-\infty}^{+\infty}|\lambda|^{2p}\left|\widetilde{X}(\lambda)\right|^{2} d\lambda$ is the stabilizing functional of order $p \geq 0$ ($p$ is not assumed to be integer), $\widetilde{K}(\lambda)$, $\widetilde{X}(\lambda)$, and $\widetilde{Y}(\lambda)$ are the Fourier images of the functions $\widetilde{k}$, $\widetilde{x}$, and $\widetilde{y}$, respectively.

The function

$$\widetilde{x}^{\alpha}(t) \equiv \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{\widetilde{K}^{*}(\lambda)\,\widetilde{Y}(\lambda)}{\left|\widetilde{K}(\lambda)\right|^{2} + \alpha\,|\lambda|^{2p}}\,e^{i\lambda t}\,d\lambda \tag{2.3.14}$$

(here $^*$ indicates complex conjugation) minimizes the functional (2.3.13) and is adopted as a regularized solution to equation (2.3.12). An appropriate value of the parameter $\alpha$ can be chosen according to any well-known criteria (for example, according to the residual criterion [157]) or can be taken as quasioptimal. One of the efficient numerical algorithms for choosing the regularization parameter is considered below.

Let us describe a discrete analogue of the problem. It is assumed that $k(t), x(t), y(t) \to 0$ as $t \to \pm\infty$ and that the functions $\widetilde{k}(t)$ and $\widetilde{y}(t)$ are given on a finite interval $[-T/2, T/2]$ on the uniform grid in $t$:

$$t_s = s\Delta t, \quad s = \overline{-N/2,\ N/2-1}. \tag{2.3.15}$$

Here $T$ is the sufficiently large parameter of the problem, $N$ is an even number of nodes, and $\Delta t = T/N$ is the grid step. Our numerical algorithm for solving convolution-type equations with discrete kernels and right-hand sides by the regularization method is based on the use of the discrete Fourier transform (DFT) of time series.

Let us represent the original problem in its canonical form by applying the DFT to the functions $\widetilde{k}_s = \widetilde{k}\,(t_s)$ and $\widetilde{y}_s = \widetilde{y}\,(t_s)$:

$$\widetilde{K}_m = \sum_{s=-N/2}^{N/2-1} \widetilde{k}_s e^{-i\lambda_m t_s}, \quad \widetilde{Y}_m = \sum_{s=-N/2}^{N/2-1} \widetilde{y}_s e^{-i\lambda_m t_s}.$$

Here $\lambda_m = m\Delta\lambda$, $m = \overline{-N/2, N/2-1}$, $\Delta\lambda = 2\pi/N\Delta t$, and $i = \sqrt{-1}$. The DFTs $X_m$ and $\widetilde{X}^\alpha_m$ of the series $x_s = x(t_s)$ and $\widetilde{x}^\alpha_s = \widetilde{x}^\alpha(t_s)$ are defined similarly. Further, approximating the integrals in the equalities (2.3.13), (2.3.14) and in the Fourier transforms by the rectangular formula, we obtain the discrete analogues for the smoothing functional and the regularized solution minimizing it:

$$\widetilde{x}^\alpha_s = \frac{1}{N\Delta t} \sum_{m=-N/2}^{N/2-1} \left[ \widetilde{K}^*_m \widetilde{Y}_m \,/\, \left( \left|\widetilde{K}_m\right|^2 + \frac{\alpha}{(\Delta t)^2} \left|\lambda_m\right|^{2p} \right) e^{2\pi i m s/N} \right]. \tag{2.3.16}$$

Here $\lambda_m = (2\pi/N\Delta t)\, m$ and $s = \overline{-N/2, N/2-1}$.

Thus, the problem of solving convolution-type equations with approximately given kernels and right-hand sides is reduced to the triple usage of one-dimensional DFT and the application of the regularization method in spectral domains with an appropriate choice of the regularization parameter $\alpha$.

**2.** Let us construct Newton's iterative process of searching for a root of the residual equation. It is assumed that additive noises in the kernel of equation (2.3.12) are absent and that the right-hand side is given approximately with an error not exceeding a known estimate $\delta$: $\|\tilde{y} - y\|_{L_2} \le \delta$. Algorithmic ways for the choice of the regularization parameter $\alpha = \alpha(\delta)$ that ensure the convergence of appropriate regularized solutions $\tilde{x}^{\alpha(\delta)}$ to the sought-for exact solution $x$ as $\delta \to 0$ are of importance for the practical application of the regularization method. As a criterion of such a choice in the formula (2.3.16), we use the residual principle [157] according to which the residual value on the constructed regularized solutions should be agreed with the accuracy of original data of the problem under consideration. Namely, we choose an appropriate value of $\alpha$ as a root of the equation

$$\rho(\alpha) = \delta, \quad \rho(\alpha) \equiv \|A\tilde{x}^\alpha - \tilde{y}\|_{L_2}. \tag{2.3.17}$$

The residual $\rho_N(\alpha)$ on the discrete regularized solution (2.3.16) is defined by the formula

$$\rho_N(\alpha) = \left( \frac{\Delta t}{N} \sum_{m=-N/2}^{N/2-1} \left| \Delta t K_m \tilde{X}_m^\alpha - \tilde{Y}_m \right|^2 \right)^{1/2} = \\ = \left[ \frac{\Delta t}{N} \sum_{m=-N/2}^{N/2-1} \left| \tilde{Y}_m \right|^2 \Big/ \left( \frac{(\Delta t)^2}{\alpha} \left| K_m \right|^2 / \left| \lambda_m \right|^{2p} + 1 \right)^2 \right]^{1/2}, \tag{2.3.18}$$

where $K_m$, $\tilde{X}_m^\alpha$, and $\tilde{Y}_m$ are the DFTs of the series $k_s$, $\tilde{x}_s^\alpha$, and $\tilde{y}_s$.

**Lemma 2.3.22** *If $\alpha \in (0, \infty)$, then the residual function $\rho_N(\alpha)$ is continuous, increases in $\alpha$ strictly monotonically, and has the following limit values:*

$$\rho_N(0) = \lim_{\alpha \to 0+} \rho_N(\alpha) = \left( \frac{\Delta t}{N} \sum_{m: K_m = 0} \left| \tilde{Y}_m \right|^2 \right)^{1/2}, \\ \rho_N(\infty) = \lim_{\alpha \to \infty} \rho_N(\alpha) = \left( \frac{\Delta t}{N} \sum_{m: \lambda_m \neq 0} \left| \tilde{Y}_m \right|^2 \right)^{1/2}, \quad p > 0. \tag{2.3.19}$$

Note that the following relations hold:

$$\rho_N(0) \le \rho_N(\infty) \le \left( \frac{\Delta t}{N} \sum_{m=-N/2}^{N/2-1} \left| \tilde{Y}_m \right|^2 \right)^{1/2} = \left\| \tilde{y}_s \right\|_N, \\ \lim_{N \to +\infty} \rho_N(\infty) / \left\| \tilde{y}_s \right\|_N = 1.$$

It follows from Lemma 2.3.22 that if a given residual level $\delta$ belongs to the range of values of the function $\rho_N(\alpha)$, then equation (2.3.17) is uniquely solvable. Let $\alpha_H$ be its root. The above properties of the residual do not allow us to construct an efficient numerical algorithm for the choice of the regularization parameter, since the function $\rho_N(\alpha)$ may change its curvature. In what follows we discuss a modification of equation (2.3.17) that ensures the unconditional applicability of rapidly convergent Newton-type algorithms for determining the roots of equations. The computing scheme we propose here is similar to the scheme given in [194].

Let us define the function

$$R(v) \equiv \rho_N(1/v), \quad 0 < v < +\infty.$$

Taking into account the relations (2.3.18) and (2.3.19) for the descrete residual, it is easy to see that the following lemma is valid.

**Lemma 2.3.23** *When $v \in (0, +\infty)$, the function $R(v)$ is continuous, decreases in $v$ strictly monotonically, and is convex downwards; its values belong to the interval $(\rho_N(0), \rho_N(\infty))$.*

Further we apply the change of variable $\alpha = 1/v$ in equation (2.3.17), reduce it to the form $R(v) = \delta$, and search for the root $v = v_H$ of the equivalent equation

$$f(v) = 0, \quad f(v) \equiv R^{-1}(v) - \delta^{-1} \tag{2.3.20}$$

by Newton's method. Then, the sought-for value of the regularization parameter is determined from the equality $\alpha_H = 1/v_H$.

From Lemma 2.3.23 and the condition of solvability for equation (2.3.17) we come to the following theorem.

**Theorem 2.3.43** *Newton's iterative method*

$$v_n = v_{n-1} + \Delta v, \quad n = 1, 2, \ldots, \tag{2.3.21}$$

*where*

$$\Delta v = \frac{R\left(v_{n-1}\right)}{R'\left(v_{n-1}\right)}\left[1 - R\left(v_{n-1}\right)/\delta\right], \tag{2.3.22}$$

$$R'(v) = -\frac{1}{R(v)} \frac{(\Delta t)^3}{N} \times$$
$$\times \sum_{m=-N/2}^{N/2-1} |K_m|^2 \left|\tilde{Y}_m\right|^2 |\lambda_m|^{-2p} \left(v\,(\triangle t)^2 |K_m|^2 / (|\lambda_m|^{2p} + 1\right)^{-3},$$

*converges to a root of equation* (2.3.20) *for any initial guess* $v_0 \geq 0$.

The above algorithm of solving the above one-dimensional integral equations with difference kernels on the infinite number axis $(-\infty, +\infty)$ was implemented in FORTRAN by M. V. Kolos. A collection of subroutines was developed and included into the Numerical Analysis Library of Moscow University Research Computing Center. These subroutines enable one to construct regularized approximations of equation (2.3.12) by the formula (2.3.16) both with a given value of $\alpha$ and with an automatic choice of $\alpha$ according to the residual criterion (2.3.17).

Let us discuss some features of software implementation of this algorithm. As an initial guess for the root $\alpha_H$ in the iterative process (2.3.21), the value $v_0 = 0$ is taken (i.e., $\alpha_0 = +\infty$); the convrgence to a root occurs from the left at each increment step $\Delta v > 0$.

It follows from the formulas (2.3.18), (2.3.22) that for $v=0$ the functions $R\,(v)$ and $R'\,(v)$ take the form ($p>0$)

$$\begin{aligned} R(0) &= \left(\frac{\Delta t}{N}\sum_{m\neq 0}\left|\tilde{Y}_m\right|^2\right)^{1/2}, \\ R'\,(0) &= -\left(\frac{(\Delta t)^3}{N}\sum_{m\neq 0}|K_m|^2\left|\tilde{Y}_m\right|^2\Big/|\lambda_m|^{2p}\right)\Big/\,R(0). \end{aligned}$$

It is assumed that as a measure of an error in the right-hand side $y$, the relative residual level $\varepsilon=\delta/\left\|\tilde{y}_s\right\|_N$, $0<\varepsilon<1$, is taken.

If the relative residual level $\varepsilon$ is such that $\rho_N(\infty)/\left\|\tilde{y}_s\right\|_N\le\varepsilon<1$, then the following approximation to the limiting solution $x^\infty=\lim_{\alpha\to+\infty}\tilde{x}^\alpha\equiv 0$ is constructed:

$$\tilde{x}_s^\alpha\equiv\frac{1}{\Delta t N}\frac{\tilde{Y}_0}{K_0},\quad s=\overline{-N/2,\,N/2-1},\quad p>0.$$

If $\varepsilon$ is too small and cannot be achieved, then (if it is possible) a nonregularized solution $\tilde{x}_s^\alpha$ is constructed for $\alpha=0$.

The above subroutines provide computations in various modes, including: 1) the solution of an integral equation with reducing the problem to its canonical form (the general case); 2) the solution of an integral equation provided that the problem is already reduced to its canonical form; 3) the solution of an integral equation under the assumption that the DFT of its kernel is already calculated. These modes are reasonable to use, for example, when the same integral equation is solved repeatedly, or when an equation is solved with the same kernel but with another right-hand side, or in the case when the analytic Fourier transform of the kernel $K(\lambda)$ is known in advance.

The quality of a constructed regularized solution $\tilde{x}_s^\alpha$ with a given value of $\alpha$ (or with $\alpha$ chosen according to the residual criterion) is estimated by the following criterial functions calculated approximately:

1) the residual $\rho(\alpha)\equiv\left\|A\tilde{x}^\alpha-\tilde{y}\right\|_{L_2}$;

2) the $\Omega$-norm $\gamma(\alpha)\equiv\Omega^{1/2}\left[\tilde{x}^\alpha\right]$ of a solution;

3) the regularizing functional $\varphi(\alpha)\equiv\left(\rho^2(\alpha)+\alpha\gamma^2(\alpha)\right)^{1/2}$;

4) the sensitivity function $\tau(\alpha)\equiv\alpha\hat{\Omega}^{1/2}\left[\dfrac{d\tilde{x}^\alpha}{d\alpha}\right]$, where $\hat{\Omega}\,[x]\equiv\Omega\,[x]$ for $p=1/2$ (in order to calculate a quasioptimal value of the regularization parameter $\alpha$).

The functions $R\,(v)$ and $R'\,(v)$ in the iterative process (2.3.21) and all the criterial functions $\rho(\alpha)$, $\gamma(\alpha)$, and $\varphi(\alpha)$ are evaluated in the spectral domain $\lambda$ in terms of DFT- components of the kernel $K_m$ and the right-hand side $\tilde{Y}_m$.

For the optimal computation of the direct and inverse DFTs, the well-known algorithm of fast Fourier transform (see Section 2.4) is used. Due to this, the implemented numerical algorithm solving the problem by the formula (2.3.16) with an automatic choice of regularization parameter is efficient in the sense of computing cost, the saving of storage space, and the reduction of round-off errors. The total number of arithmetic operations is proportional to $N \log_2 N$, whereas the amount of storage is proportional to $N$, where $N$ is the number of points where the original data are given. The implemented subroutines allow one to process the series of maximal length 2048, using only the core memory. Naturally, this cannot be done if a convolution-type integral equation is solved by reducing it to a system of linear algebraic equations. Note also that the gain in time of this algorithm compared to the direct algebraic methods is, for example, about 60 times for a system of order $N = 64$.

**3.** In order to check the efficiency of the above algorithm, many numerical experiments were carried out on the basis of model problems with known exact solutions. Various performance capabilities of the algorithm for solving convolution-type equations with inexactly given kernels and right-hand sides by the formula (2.3.16) were studied. The abobe-considered parameters were varied in rather a wide range. Let us discuss some peculiarities characterizing our numerical results.

1) The approximation errors $\tilde{x}_s^\alpha$ were estimated in the uniform and root-mean-square grid norms, i.e., the following characteristics of accuracy were calculated:

$$\gamma_C(\alpha) = \max_s |x_s - \tilde{x}_s^\alpha| , \quad \gamma_{L_2}(\alpha) = \left( \frac{1}{N} \sum_{s=-N/2}^{N/2-1} |x_s - \tilde{x}_s^\alpha|^2 \right)^{1/2} .$$

Here $x_s = x(t_s)$ are discrete values of the known model solution at the grid nodes (2.3.15).

2) The appropriate values of the regularization parameter $\alpha$ were chosen in accordance with the residual criterion ($\alpha_H$) or were given on the grid

$$\{\alpha\} = \{\alpha : \alpha_i = \Delta_\alpha \alpha_{i-1}, \quad i = 1, 2, \ldots, q_\alpha, \quad q_{\alpha+1} = 0\} ,$$

where $\alpha_0$ is an initial value of $\alpha$, $0 < \Delta_\alpha < 1$ is a “grid” step, and $q_\alpha$ is the number of grid steps in $\alpha$. Among the regularized solutions $\{\tilde{x}_s^\alpha\}$, $\alpha \in \{\alpha\}$, an optimal approximation $\tilde{x}_s^{\text{opt}} = \tilde{x}_s^{\alpha_{\text{opt}}}$ was chosen by search for an optimal value $\alpha = \alpha_{\text{opt}}$ minimizing one of the deviations $\gamma_C(\alpha)$ or $\gamma_{L_2}(\alpha)$ in $\alpha \in \{\alpha\}$: $\gamma_C(\alpha_{\text{opt}}) \equiv \min_\alpha \gamma_C(\alpha)$ (the best uniform approximation) or $\gamma_{L_2}(\alpha_{\text{opt}}) \equiv \min_\alpha \gamma_{L_2}(\alpha)$ (the best mean-square approximation).

Note that usually the value of $\alpha_{\text{opt}}$ was first found on a sufficiently coarse grid in $\alpha$ (with a large initial value $\alpha_0$ and with the step $\Delta_\alpha = 0.1$) and, then, this value was refined for a given $\alpha_0$ in a neighborhood of $\alpha_{\rm opt}$ and with increasing the step in $\alpha$ up to $\Delta_\alpha = 0.9$.

3) Our computations were carried out for various values of the parameters characterizing the original discrete equation and the regularizator in use; those parameters were:

— the grid point number $N = 2^\gamma$, $\gamma = 4, 5, \ldots, 11$;

— the length of the realization segment $T = 2, 4, 6, 8, 10, 20$;

— the regularizator order $p = 0, 1, 2$.

All possible combinations of the above values were considered.

4) The stability of the numerical method with respect to errors of input data was checked for additive uniformly distributed random perturbations $n_k$ and $n_y$ of the kernel $k$ and the right-hand side $y$ in equation (2.3.12); these perturbations were chosen in the form $n_k = \sigma_k\xi$ and $n_y = \sigma_y\xi$, where $\xi$ is an uniformly distributed centered random noise on the segment $[-1, 1]$, $\sigma_k$ and $\sigma_y$ are the amplitudes of the noises $n_k$ and $n_y$ taking the values: 0.001, 0.01, 0.025, 0.05, 0.1, 0.15. For the model problems under consideration, the perturbations $n_k$ and $n_y$ were about $0.1 - 15\,\%$ of the maximal value of input data.

5) The relative residual level $\varepsilon$ was prescribed by the following three ways.

a) The values of $\varepsilon$ were taken on the grid

$$\{\varepsilon\} = \{\varepsilon : \varepsilon_i = \Delta_\varepsilon \varepsilon_{i-1}, \quad i = 1, 2, \ldots, q_\varepsilon\},$$

where $\varepsilon_0 = 1$ is an initial value of $\varepsilon$, $0 < \Delta_\varepsilon < 1$, $\Delta_\varepsilon$ is the grid step, and $q_\varepsilon$ is the number of grid steps in $\varepsilon$. For each $\varepsilon \in \{\varepsilon\}$, a value $\alpha_H = \alpha_H(\varepsilon)$ was specified according to the residual criterion and an appropriate approximation $\tilde{x}_s^{\alpha_H(\varepsilon)}$ was constructed. A family of the regularized solutions $\left\{\tilde{x}_s^{\alpha_H(\varepsilon)}\right\}$, $\varepsilon \in \{\varepsilon\}$, was thus obtained. This family contains an optimal approximation $\tilde{x}_s^{\alpha_H(\varepsilon_{\rm opt})} = \tilde{x}_s^{\rm opt}$ with some $\varepsilon = \varepsilon_{\rm opt}$ for which $\alpha_H(\varepsilon_{\rm opt}) = \alpha_{\rm opt}$ on the set $\{\alpha_H(\varepsilon_{\rm opt})\}$.

b) The error $\delta_1 = \|y_s - \tilde{y}_s\|_N$ of the right-hand side was calculated; the residual level $\varepsilon$ was prescribed in the form $\varepsilon_1 = \delta_1 / \|\tilde{y}_s\|_N$.

c) The total error was calculated for the case of an exact kernel:

$$\delta_2 = \|\Delta t(Cx)_s - \tilde{y}_s\|_N \le \|\Delta t(Cx)_s - y_s\|_N + \delta_1.$$

Here $\|\Delta t(Cx)_s - y_s\|_N$ is an error of approximation for the integral in equation (2.3.12), $C$ is the circulant matrix defined by its column $k_0$, $k_1$, $\ldots$, $k_{N/2-1}$, $k_{-N/2}$, $k_{-N/2+1}$, $\ldots$, $k_{-1}$, and $\delta_1$ is an error of the right-hand side; the residual level $\varepsilon$ was taken in the form $\varepsilon_2 = \delta_2 / \|\tilde{y}_s\|_N$.

**4. Model problems**. Let us discuss some numerical results obtained for a number of model problems. Many other numerical experiments are described in [9].

*Model problem* 1. A convolution-type integral equation with Gaussian kernel $k(t) = e^{-t^2}$ was considered on the infinite number axis $(-\infty, +\infty)$. Such equations arise, for example, in the inverse problems of optics and nuclear spectroscopy. The exact solution $x(t)$ and the exact right-hand side $y(t)$ were also prescribed in exponential form:

$$x(t) = e^{-t^2}, \quad y(t) = \left(\frac{\pi}{2}\, e^{-t^2}\right)^{1/2}.$$

*Model problem* 2. A convolution-type integral equation with Poisson's kernel $k(t) = 1/\left(1+t^2\right)$ was considered on the finite segment $[-1, 1]$. Such equations with the above kernels arise, for example, in the inverse problems of gravimetry, in the problems of analytical continuation of harmonic functions on a horizontal layer, and in the Cauchy problems for the Laplace equations in a half-plane. The exact solution $x(t)$ and the exact right-hand side $y(t)$ were prescribed on the segment $[-1, 1]$ in the form

$$\begin{gathered} x(t) = 1 - t^2, \\ y(t) = \left(2 - t^2\right)\left[\arctan\left(1-t\right) + \arctan\left(1+t\right)\right] - 2- \\ -t \ln\left[\left(1 + (1-t)^2\right) \Big/ \left(1 + (1+t)^2\right)\right]. \end{gathered}$$

Note that the length $T$ of the realization segment in this model example was fixed ($T = 2$).

Some typical numerical results are represented graphically in Figures 2.3.6–2.3.11. These figures illustrate: 1) the exact solutions of the above model problems; 2) the optimal regularized solutions $\tilde{x}_s^{\rm opt}$ for $\alpha = \alpha_{\rm opt} > 0$ and the solutions $\tilde{x}_s^{\alpha}$ for $\alpha = 0$ or for $\alpha$ close to zero in the case of approximation errors caused by the replacement of the original problem with its discrete analogue ($\sigma_k = 0, \sigma_y = 0$); 3) the optimal regularized solutions $\tilde{x}_s^{\rm opt}$ and the uniform errors $\gamma_C\left(\alpha\right)$ for various noises in the kernel and in the right-hand side of the equation.

Some other numerical results are represented in Tables 2.3.1–2.3.5. These tables illustrate: 1) how the optimal values of $\alpha_{\rm opt}$, the optimal errors $\gamma_C\left(\alpha_{\rm opt}\right)$ and $\gamma_{L_2}\left(\alpha_{\rm opt}\right)$, and the number $k$ of steps for the iterative process when searching for $\alpha = \alpha_H$ depend on the length $T$ of the realization segment, on the grid point number $N$, on the regularizator order $p$, and on the noise levels $\sigma_k$ and $\sigma_y$; 2) how the values of $\alpha = \alpha_H$ chosen according to the residual criterion, the errors $\gamma_C\left(\alpha_H\right)$ and $\gamma_{L_2}\left(\alpha_H\right)$ of the corresponding regularized solutions, and the number $k$ of iterations depend on the way of choosing the residual

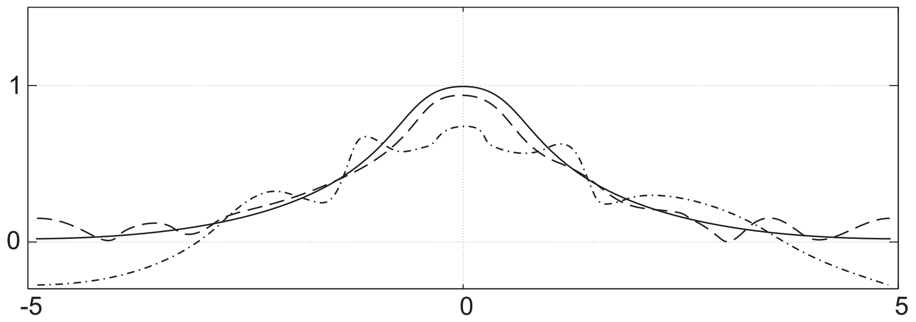


Figure 2.3.6: The solution of model problem 1 without perturbations of the kernel and the right-hand side $\big(T = 10,\ N = 64,\ R = \alpha\lambda^4, \sigma_k = 0,\ \sigma_y = 0\big)$: the solid curve — the exact solution; the dashed curve — the solution $\tilde{x}_s^{\rm opt}$ for $\alpha = \alpha_{\rm opt} = 4.5 \cdot 10^{-9}$; the dot-and-dashed curve — the solution $\tilde{x}_s^{\alpha}$ for $\alpha = 10^{-17}$

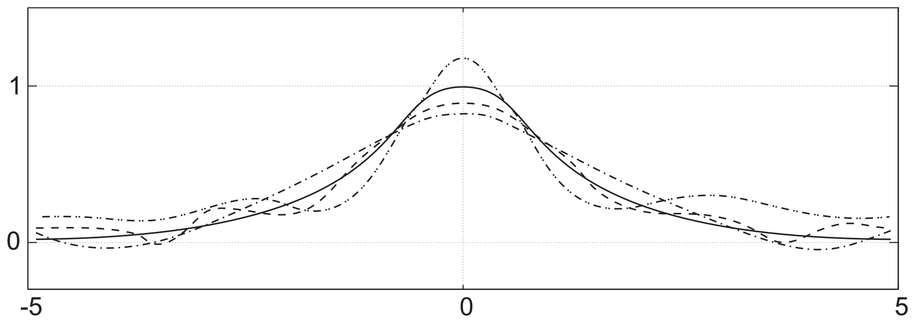


Figure 2.3.7: The solution of model problem 1 with perturbations of the kernel and the right-hand side $\big(T = 10,\ N = 64,\ R = \alpha\lambda^4,\ \sigma_k = 0,\ \sigma_y > 0\big)$: the solid curve — the exact solution; the dashed curve — the solution $\tilde{x}_s^{\rm opt}$ for $\alpha = \alpha_{\rm opt} = 2.3{\cdot}10^{-4}$, $\sigma_y = 0.01$; the dot-and-dashed curve — the solution $\tilde{x}_s^{\rm opt}$ for $\alpha = \alpha_{\rm opt} = 4.1 \cdot 10^{-3}$, $\sigma_y = 0.05$; the dot-dot-and-dashed curve — the solution $\tilde{x}_s^{\rm opt}$ for $\alpha = \alpha_{\rm opt} = 6.7{\cdot}10^{-3}$, $\sigma_y = 0.1$

level $\varepsilon$ $(\varepsilon = \varepsilon_{\rm opt}, \varepsilon_1, \varepsilon_2)$ and on the values of $\varepsilon$ on the grid $\{\varepsilon\}$ for various noises in the right-hand side of the equation.

An analysis of numerical results we obtained allows us to come to the following conclusions.

**5.** General conclusions.

1) It follows from our numerical experiments based on various model problems that the efficient numerical solution of the one-dimensional convolution-type integral equation (2.3.12) with an approximately given kernel and an approximately given right-hand side can be obtained by joint application of A. N. Tikhonov's methods of one-parametric regularization and of the fast discrete Fourier transform.

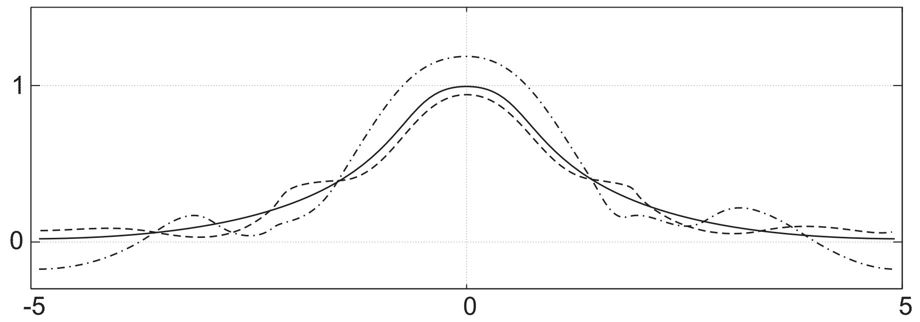


Figure 2.3.8: The solution of model problem 1 with perturbations of the kernel and the right-hand side $\big(T = 10,\ N = 64,\ R = \alpha\lambda^4,\ \sigma_k > 0,\ \sigma_y > 0\big)$: the solid curve — the exact solution; the dashed curve — the solution $\tilde{x}_s^{\rm opt}$ for $\alpha = \alpha_{\rm opt} = 2.5 \cdot 10^{-4}$, $\sigma_k = 0.025$, $\sigma_y = 0.01$; the dot-and-dashed curve — the solution $\tilde{x}_s^{\rm opt}$ for $\alpha = \alpha_{\rm opt} = 6.8 \cdot 10^{-3}$, $\sigma_k = 0.025$, $\sigma_y = 0.1$

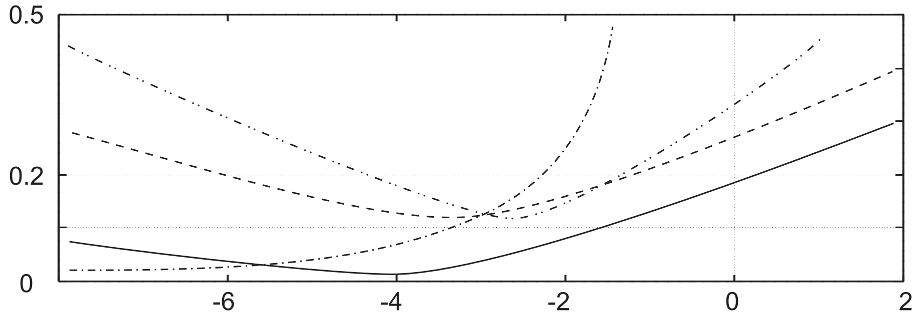


Figure 2.3.9: The dependence of the uniform error $\gamma_c\,(\alpha)$ on $\alpha$ for perturbations of the kernel and the right-hand side of model problem 1: the solid curve — for $\sigma_k = 0$, $\sigma_y = 0,01$; the dashed curve — for $\sigma_k = 0$, $\sigma_y = 0.1$; the dot-and-dashed curve — for $\sigma_k = 0.05$, $\sigma_y = 0$; the dot-dot-and-dashed curve — for $\sigma_k = 0.01$, $\sigma_y = 0.1$

2) The regularization method substantially influences the accuracy of the approximations obtained. The optimal (in the sense of minimal absolute or mean-square errors) solution is obtained for some optimal value $\alpha = \alpha_{\rm opt}$ of regularization parameter (generally, not equal to zero) and has rather high accuracy.

3) The solution $\tilde{x}_s^0$ without regularization in the model problem 2 is unsatisfactory and has a very large error. In the model problem 1, it is not always possible to construct numerically such a solution (since the amplitudes of eigenvalues $|K_m|$, $m = 0, 1, \ldots, N/2 - 1$, of equation (2.3.12) decrease to zero with increasing $m$; as a result, the corresponding subroutine terminates its run because of emergency stop). The only exception is the case of a strongly perturbed kernel (5–15 %) and an exact right-hand side when $\alpha_{\rm opt} = 0$ and

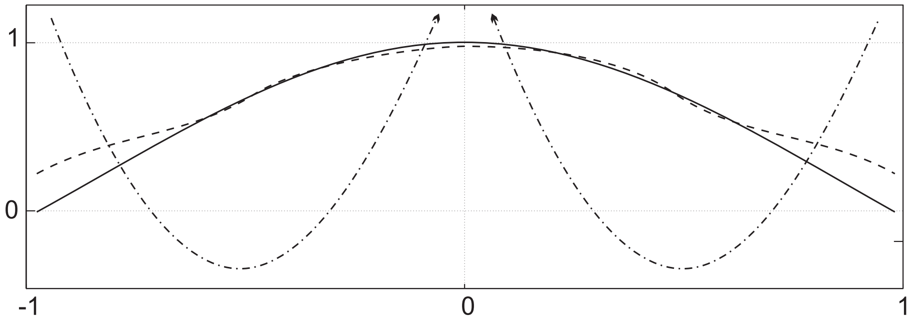


Figure 2.3.10: The solution of model problem 2 without perturbations of the kernel and the right-hand side $\big(T = 2,\ N = 64,\ R = \alpha\lambda^4,\ \sigma_k = 0,\ \sigma_y = 0\big)$: the solid curve — the exact solution; the dashed curve — the solution $\tilde{x}_s^{\mathrm{opt}}$ for $\alpha = \alpha_{\mathrm{opt}} = 1.5 \cdot 10^{-3}$; the dot-and-dashed curve — the solution $\tilde{x}_s^{\alpha}$ for $\alpha = 0$

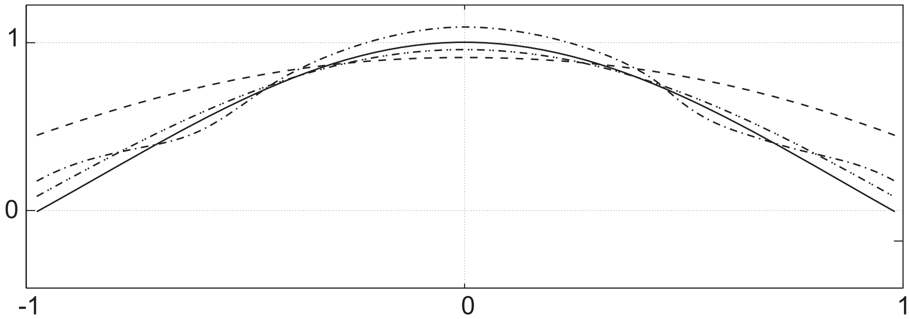


Figure 2.3.11: The solution of model problem 2 with perturbations of the kernel and the right-hand side $\big(T = 2,\ N = 64,\ R = \alpha\lambda^4,\ \sigma_k > 0,\ \sigma_y > 0\big)$: the solid curve — the exact solution; the dashed curve — the solution $\tilde{x}_s^{\mathrm{opt}}$ for $\alpha = \alpha_{\mathrm{opt}} = 1.5 \cdot 10^{-3}$, $\sigma_k = 0$, $\sigma_y = 0.01$; the dot-and-dashed curve — the solution $\tilde{x}_s^{\mathrm{opt}}$ for $\alpha = \alpha_{\mathrm{opt}} = 1.6 \cdot 10^{-3}$, $\sigma_k = 0.01$, $\sigma_y = 0.05$; the dot-dot-and-dashed curve — the solution $\tilde{x}_s^{\mathrm{opt}}$ for $\alpha = \alpha_{\mathrm{opt}} = 2.0 \cdot 10^{-3}$, $\sigma_k = 0.025$, $\sigma_y = 0.1$

the solution without regularization has an acceptable accuracy.

4) The algorithm we developed is stable with respect to errors in approximations, round-off errors, and additive perturbations in kernels and in right-hand sides. Naturally, the accuracy of resulting approximations is deteriorated when noises grow, but the algorithm allows one to construct a quite satisfactory solution in the sense of accuracy even when there are significant errors in kernels and in right-hand sides. In the model problem 1 for $\sigma_k \sim 10 - 15\,\%$ and $\sigma_y \sim 1 - 3\,\%$, for example, the absolute error in the resulting approximations is 3–4 %. A similar situation is observed in the model problem 2 for $\sigma_k \sim 10 - 15\,\%$ and $\sigma_y \sim 10 - 15\,\%$.

5) When the length $T$ increases (with a coordinated choice of the grid point

**Table 2.3.1. Model problem 1**. The dependence of optimal values of $\varepsilon$, $\alpha$ and of optimal errors on the length $T$ of the realization segment, on the grid point number, and on the regularizator order $p$ (without perturbations of the kernel and the right-hand side)

| $p$ | $N$ | $T$ | $\varepsilon_{\text{opt}}$ | $\alpha_{\text{opt}}$ | $k$ | $\gamma_C(\alpha_{\text{opt}})$ | $\gamma_{L_2}(\alpha_{\text{opt}})$ |
|---|---|---|---|---|---|---|---|
| 2 | 64 | 4 | $10^{-1}$ | $4.4 \cdot 10^{-2}$ | 2 | $1.6 \cdot 10^{-1}$ | $7.4 \cdot 10^{-2}$ |
| | | 6 | $10^{-2}$ | $1.0 \cdot 10^{-3}$ | 5 | $5.9 \cdot 10^{-2}$ | $2.9 \cdot 10^{-2}$ |
| | | 10 | $10^{-6}$ | $4.5 \cdot 10^{-9}$ | 5 | $3.2 \cdot 10^{-4}$ | $1.5 \cdot 10^{-4}$ |
| 2 | 64 | 10 | $10^{-6}$ | $4.5 \cdot 10^{-9}$ | 5 | $3.2 \cdot 10^{-4}$ | $1.5 \cdot 10^{-4}$ |
| | 128 | 10 | $10^{-6}$ | $4.5 \cdot 10^{-9}$ | 5 | $3.3 \cdot 10^{-4}$ | $1.5 \cdot 10^{-4}$ |
| | 256 | 10 | $10^{-6}$ | $5.0 \cdot 10^{-9}$ | 5 | $3.3 \cdot 10^{-4}$ | $1.6 \cdot 10^{-4}$ |
| 0 | 64 | 10 | $10^{-6}$ | $1.0 \cdot 10^{-6}$ | 4 | $1.7 \cdot 10^{-4}$ | $8.7 \cdot 10^{-5}$ |
| 1 | | 10 | $10^{-6}$ | $2.1 \cdot 10^{-7}$ | 4 | $2.5 \cdot 10^{-4}$ | $9.2 \cdot 10^{-5}$ |
| 2 | | 10 | $10^{-6}$ | $4.5 \cdot 10^{-9}$ | 5 | $3.2 \cdot 10^{-4}$ | $1.5 \cdot 10^{-4}$ |

**Table 2.3.2. Model problem 1.** The dependence of optimal values of $\varepsilon, \alpha$ and of optimal errors on perturbations of the kernel and the right-hand side

| | $\sigma_k$ | $\sigma_y$ | $\varepsilon_{\text{opt}}$ | $\alpha_{\text{opt}}$ | $k$ | $\gamma_C(\alpha_{\text{opt}})$ | $\gamma_{L_2}(\alpha_{\text{opt}})$ |
|---|---|---|---|---|---|---|---|
| $T = 10$, $N = 64$, $p = 2$ | 0 | 0 | $10^{-6}$ | $4.5 \cdot 10^{-2}$ | 5 | $3.2 \cdot 10^{-4}$ | $1.5 \cdot 10^{-4}$ |
| | 0 | 0.01 | $10^{-2}$ | $2.3 \cdot 10^{-4}$ | 7 | $3.3 \cdot 10^{-2}$ | $1.3 \cdot 10^{-2}$ |
| | 0 | 0.05 | $5 \cdot 10^{-2}$ | $4.1 \cdot 10^{-3}$ | 6 | $7.9 \cdot 10^{-2}$ | $3.1 \cdot 10^{-2}$ |
| | 0 | 0.1 | $10^{-1}$ | $6.7 \cdot 10^{-3}$ | 7 | $1.1 \cdot 10^{-1}$ | $4.3 \cdot 10^{-2}$ |
| | 0.01 | 0 | $10^{-6}$ | $1.0 \cdot 10^{-8}$ | 4 | $5.6 \cdot 10^{-3}$ | $2.3 \cdot 10^{-3}$ |
| | 0.05 | 0 | $10^{-7}$ | 0 | 2 | $2.1 \cdot 10^{-2}$ | $1.0 \cdot 10^{-2}$ |
| | 0.1 | 0 | $10^{-7}$ | 0 | 2 | $3.8 \cdot 10^{-2}$ | $1.9 \cdot 10^{-2}$ |
| | 0.01 | 0.01 | $10^{-2}$ | $2.4 \cdot 10^{-4}$ | 7 | $3.6 \cdot 10^{-2}$ | $1.4 \cdot 10^{-2}$ |
| | 0.01 | 0.1 | $10^{-1}$ | $6.8 \cdot 10^{-3}$ | 7 | $1.2 \cdot 10^{-1}$ | $4.4 \cdot 10^{-2}$ |
| | 0.025 | 0.01 | $10^{-2}$ | $2.5 \cdot 10^{-4}$ | 7 | $4.1 \cdot 10^{-2}$ | $1.5 \cdot 10^{-2}$ |
| | 0.025 | 0.1 | $10^{-1}$ | $6.8 \cdot 10^{-3}$ | 7 | $1.3 \cdot 10^{-1}$ | $4.5 \cdot 10^{-2}$ |

number $N$), the approximation of a problem is improved and the accuracy of corresponding optimal approximations grows.

6) We have not succeeded to establish a noticeable influence of the regularizator order on accuracy of the resulting approximations for the model problems under consideration. The best accuracy is reached for the regularizator of zeroth order, but this improvement is not significant.

7) The uniform and mean-square errors $\gamma_C(\alpha)$ and $\gamma_{L_2}(\alpha)$ monotonically

**Table 2.3.3. Model problem 1.** The dependence of $\alpha$ chosen according to the residual criterion and of errors in the regularized solution on a way of specification of the residual level

| | $\sigma_y$ | $\varepsilon$ | $\alpha_H$ | $k$ | $\gamma_C(\alpha_H)$ | $\gamma_{L_2}(\alpha_H)$ |
|---|---|---|---|---|---|---|
| $T = 10$ $N = 64$ $p = 2$ $\sigma_k = 0$ | 0.001 | $\varepsilon_1 = \varepsilon_2 = 1.02 \cdot 10^{-3}$ | $1.1 \cdot 10^{-5}$ | 7 | $9.7 \cdot 10^{-3}$ | $4.0 \cdot 10^{-3}$ |
| | | $\varepsilon_{\text{opt}} = 10^{-3}$ | $1.4 \cdot 10^{-5}$ | 7 | $1.2 \cdot 10^{-2}$ | $5.2 \cdot 10^{-3}$ |
| | 0.001 | $\varepsilon_1 = \varepsilon_2 = 1.02 \cdot 10^{-2}$ | $2.4 \cdot 10^{-4}$ | 7 | $3.7 \cdot 10^{-2}$ | $1.5 \cdot 10^{-2}$ |
| | | $\varepsilon_{\text{opt}} = 10^{-2}$ | $2.3 \cdot 10^{-4}$ | 7 | $3.3 \cdot 10^{-2}$ | $1.3 \cdot 10^{-2}$ |
| | 0.025 | $\varepsilon_1 = \varepsilon_2 = 2.6 \cdot 10^{-2}$ | $9.9 \cdot 10^{-4}$ | 7 | $6.1 \cdot 10^{-2}$ | $2.5 \cdot 10^{-2}$ |
| | | $\varepsilon_{\text{opt}} = 3.0 \cdot 10^{-2}$ | $5.9 \cdot 10^{-4}$ | 6 | $4.2 \cdot 10^{-2}$ | $2.2 \cdot 10^{-2}$ |
| | 0.05 | $\varepsilon_1 = \varepsilon_2 = 5.1 \cdot 10^{-2}$ | $2.7 \cdot 10^{-3}$ | 7 | $8.4 \cdot 10^{-2}$ | $3.5 \cdot 10^{-2}$ |
| | | $\varepsilon_{\text{opt}} = 5.0 \cdot 10^{-2}$ | $4.1 \cdot 10^{-3}$ | 6 | $7.9 \cdot 10^{-2}$ | $3.1 \cdot 10^{-2}$ |
| | | $\varepsilon_1 = \varepsilon_2 = 1.02 \cdot 10^{-1}$ | $8.6 \cdot 10^{-3}$ | 7 | $1.2 \cdot 10^{-1}$ | $5.0 \cdot 10^{-2}$ |
| | | $\varepsilon_{\text{opt}} = 10^{-1}$ | $6.7 \cdot 10^{-3}$ | 7 | $1.1 \cdot 10^{-1}$ | $4.3 \cdot 10^{-2}$ |

**Table 2.3.4. Model problem 2.** The dependence of $\alpha$ chosen according to the residual criterion and of errors in the regularized solution on the residual level $\varepsilon$

| | $\sigma_y$ | $\varepsilon$ | $\alpha_H$ | $k$ | $\gamma_C(\alpha_H)$ | $\gamma_{L_2}(\alpha_H)$ |
|---|---|---|---|---|---|---|
| $T = 10$ $N = 64$ $p = 2$ $\sigma_k = 0$ | 0.001 | 1 | $1.1 \cdot 10^{18}$ | 0 | $6.3 \cdot 10^{-1}$ | $3.0 \cdot 10^{-1}$ |
| | | $10^{-1}$ | $1.5 \cdot 10^{-3}$ | 2 | $4.1 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1}$ |
| | | $10^{-2}$ | $2.7 \cdot 10^{-8}$ | 6 | $2.5 \cdot 10^{0}$ | $1.2 \cdot 10^{0}$ |
| | | $10^{-3}$ | $2.8 \cdot 10^{-14}$ | 10 | $4.8 \cdot 10^{2}$ | $2.4 \cdot 10^{2}$ |
| | | $10^{-4}$ | $1.2 \cdot 10^{-16}$ | 6 | $1.0 \cdot 10^{4}$ | $4.9 \cdot 10^{3}$ |
| | | $10^{-5}$ | 0 | 4 | $1.2 \cdot 10^{4}$ | $6.1 \cdot 10^{3}$ |
| | 0.05 | 1 | $1.0 \cdot 10^{18}$ | 0 | $6.3 \cdot 10^{-1}$ | $3.0 \cdot 10^{-1}$ |
| | | $10^{-1}$ | $1.5 \cdot 10^{-3}$ | 2 | $2.4 \cdot 10^{-1}$ | $1.7 \cdot 10^{-1}$ |
| | | $10^{-2}$ | $2.7 \cdot 10^{-15}$ | 10 | $7.9 \cdot 10^{-1}$ | $3.5 \cdot 10^{1}$ |
| | | $10^{-3}$ | $1.7 \cdot 10^{-17}$ | 6 | $1.6 \cdot 10^{2}$ | $6.1 \cdot 10^{1}$ |
| | | $10^{-4}$ | 0 | 4 | $1.7 \cdot 10^{2}$ | $6.4 \cdot 10^{1}$ |

decrease down to some positive minimal value when $\alpha$ decreases from $\alpha_0$ to $\alpha_{\text{opt}}$ $(\alpha_{\text{opt}} \le \alpha \le \alpha_0)$; these errors monotonically increase with further decrease in $\alpha$ $(0 \le \alpha \le \alpha_{\text{opt}})$.

8) Newton's iterative process (2.3.21) for finding a root of equation (2.3.17) converges rather rapidly to the sought-for value of the regularization parameter $\alpha = \alpha_H$; the number of iterations in all the model problems under consideration does not exceed 10.

9) When the residual level $\varepsilon$ on the grid $\{\varepsilon\}$ decreases, the found values of the regularization parameter $\alpha = \alpha_H(\varepsilon)$ and the corresponding residuals

**Table 2.3.5. Model problem 2.** The dependence of $\alpha$ chosen according to the residual criterion and of the errors in the regularized solution on a way of specification of the residual level $\varepsilon$.

| | $\sigma_y$ | $\varepsilon$ | $\alpha_H$ | $k$ | $\gamma_C(\alpha_H)$ | $\gamma_{L_2}(\alpha_H)$ |
|---|---|---|---|---|---|---|
| | | $\varepsilon_1 = 5.5 \cdot 10^{-4}$ | $2.2 \cdot 10^{-15}$ | 9 | $2.8 \cdot 10^{1}$ | $3.9 \cdot 10^{0}$ |
| | 0.001 | $\varepsilon_2 = 9.7 \cdot 10^{-2}$ | $1.3 \cdot 10^{-3}$ | 2 | $3.9 \cdot 10^{-1}$ | $1.5 \cdot 10^{-1}$ |
| | | $\varepsilon_{\text{opt}} = 10^{-1}$ | $1.5 \cdot 10^{-3}$ | 2 | $4.1 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1}$ |
| | | $\varepsilon_1 = 5.5 \cdot 10^{-3}$ | $1.6 \cdot 10^{-10}$ | 9 | $5.5 \cdot 10^{0}$ | $1.8 \cdot 10^{0}$ |
| $T = 2$ | 0.01 | $\varepsilon_2 = 9.7 \cdot 10^{-2}$ | $1.3 \cdot 10^{-3}$ | 2 | $3.9 \cdot 10^{-1}$ | $1.5 \cdot 10^{-1}$ |
| $N = 64$ | | $\varepsilon_{\text{opt}} = 10^{-1}$ | $1.5 \cdot 10^{-3}$ | 2 | $4.2 \cdot 10^{-1}$ | $1.7 \cdot 10^{-1}$ |
| $p = 2$ | | $\varepsilon_1 = 2.7 \cdot 10^{-2}$ | $5.8 \cdot 10^{-8}$ | 8 | $1.8 \cdot 10^{0}$ | $8.9 \cdot 10^{-1}$ |
| $\sigma_k = 0$ | 0.05 | $\varepsilon_2 = 9.8 \cdot 10^{-2}$ | $1.3 \cdot 10^{-3}$ | 2 | $4.0 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1}$ |
| | | $\varepsilon_{\text{opt}} = 10^{-1}$ | $1.5 \cdot 10^{-3}$ | 2 | $4.2 \cdot 10^{-1}$ | $1.7 \cdot 10^{-1}$ |
| | | $\varepsilon_1 = 5.5 \cdot 10^{-2}$ | $1.2 \cdot 10^{-5}$ | 8 | $4.6 \cdot 10^{-1}$ | $2.8 \cdot 10^{-1}$ |
| | 0.1 | $\varepsilon_2 = 1.07 \cdot 10^{-3}$ | $1.4 \cdot 10^{-3}$ | 2 | $4.1 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1}$ |
| | | $\varepsilon_{\text{opt}} = 10^{-1}$ | $1.6 \cdot 10^{-3}$ | 2 | $4.3 \cdot 10^{-1}$ | $1.8 \cdot 10^{-1}$ |

$\rho(\alpha_H(\varepsilon))$ monotonically decrease.

10) If $\varepsilon$ is given in the form $\varepsilon_1 = \dfrac{\|y_s - \tilde{y}_s\|_N}{\|\tilde{y}_s\|_N}$, then the value of $\alpha = \alpha_H(\varepsilon_1)$ is close to $\alpha_{\text{opt}}$ only in some cases. More often this condition results in an underestimated value of $\alpha$ compared to $\alpha_{\text{opt}}$ (in other words, the value of $\delta_1$ is underestimated). To obtain more exact results, it is necessary to add the value of approximation error to $\delta_1$, i.e., when errors in a kernel are absent, it is necessary to choose $\varepsilon$ in the form $\dfrac{\|\Delta t(Cx)_s - \tilde{y}_s\|_N}{\|\tilde{y}_s\|_N}$.

Our numerical results show that the value of $\varepsilon_2$ is always close to the optimal value $\varepsilon_{\text{opt}}$ and, hence, leads to the optimal approximation $\tilde{x}_s^{\alpha_H(\varepsilon)}$ for $\alpha_{\text{opt}} = \alpha_H(\varepsilon_2)$.

In the model problem 1, the errors of approximation in the integral is almost equal to zero, $\delta \approx \delta_1$ for any perturbations of the right-hand side of the equation, and the relations $\varepsilon_1 \approx \varepsilon_2 \approx \varepsilon_{\text{opt}}$ hold; hence, $\alpha_H(\varepsilon_1) \approx \alpha_H(\varepsilon_2) \approx \alpha_{\text{opt}}$.

In the model problem 2, the errors of approximation in the integral make a main contribution to the total error $\delta_2$, $\delta_1 \ll \delta_2$, of any perturbations in the right-hand side of the equation; the relations $\varepsilon_1 \ll \varepsilon_2 \approx \varepsilon_{\text{opt}}$ are valid and, hence, $\alpha_H(\varepsilon_1) < \alpha_H(\varepsilon_2) \approx \alpha_{\text{opt}}$.

Thus, the above analysis of our numerical results allows us to come to the conclusion that it is possible to use the above-proposed algorithm for solving the particular applied problems that can be reduced to one-dimensional convolution-type integral equations of the first kind.

## 2.4 A collection of fast Fourier transform algorithms for discrete series

**1.** A collection of subroutines intended for computing the discrete (inverse discrete) Fourier transform on the basis of the fast Fourier transform method was developed at the Moscow University Research Computing Center. Recall that the discrete Fourier transform of a complex series $A_k$, $k = 0, 1, \ldots, N-1$, is the series

$$X_r = \sum_{k=0}^{N-1} A_k e^{-\frac{2\pi i}{N} kr}, \quad r = 0, 1, \ldots, N-1. \tag{2.4.23}$$

The coefficients of the original series $A_k$ are expressed in terms of the coefficients of the Fourier transform $X_r$ as follows:

$$A_k = \frac{1}{N} \sum_{r=0}^{N-1} X_r e^{\frac{2\pi i}{N} kr}, \quad k = 0, 1, \ldots, N-1.$$

In practice, the following somewhat different definition of Fourier transform is used: the expression $X_r/N$ (or $X_r/\sqrt{N}$) is taken as the coefficients of the discrete Fourier transform. In the sequel, the series (2.4.23) is called the direct discrete Fourier transform (DFT), whereas the series

$$A_k = \sum_{r=0}^{N-1} X_r e^{\frac{2\pi i}{N} kr}, \quad k = 0, 1, \ldots, N-1, \tag{2.4.24}$$

is called the inverse discrete Fourier transform (IDFT). In the context of this definition, the IDFT applied to the direct DFT of a series $A_k$ results in the series $\hat{A}_k$ differing from the original series only by the factor $N$: $\hat{A}_k = NA_k$. The real Fourier transform is defined in a similar way. The Fourier coefficients $X_j$, $j = 0, 1, \ldots, 2N-1$, of a real series are defined as follows:

$$\begin{aligned} A_k &= \frac{1}{N} \sum_{j=0}^{2N-1} X_j \cos\left(\frac{\pi}{N} kj\right), \quad k = 0, 1, \ldots, N, \\ B_k &= \frac{1}{N} \sum_{j=0}^{2N-1} X_j \sin\left(\frac{\pi}{N} kj\right), \quad k = 1, \ldots, N-1. \end{aligned} \tag{2.4.25}$$

Knowing the real Fourier coefficients, it is possible to restore the original series by the formula ($j = 0, 1, \dots, 2N-1$)

$$X_j = \frac{A_0}{2} + (-1)^j \frac{A_N}{2} + \sum_{k=1}^{N-1} \left( A_k \cos\left(\frac{\pi}{N} kj\right) + B_k \sin\left(\frac{\pi}{N} kj\right)\right). \quad (2.4.26)$$

Direct calculations by the formulas (2.4.23) and (2.4.24) require $N^2$ complex multiplications and additions and, for a sufficiently large $N$, leads to a considerable increase in computing cost.

At present, several similar algorithms allowing to substantially reduce the computing time are known. These algorithms are said to be the algorithms of fast Fourier transform (FFT). The FFT algorithms may be divided into three groups, depending on whether the length $N$ of a series is prime, composite, or an integer power of 2.

For $N$ equal to an integer power of 2, there are known the time and frequency decimation algorithms [69]. These algorithms have many versions, depending on whether the data are available in core memory before, during, or after computations. The algorithms allowing to compute recurrently (i.e., using the same storage in which input data are recorded) require rearrangement of input or output data. This rearrangement is known as binary inversion, since data are recorded in the binary-inverse order (for example, $A_6 = A_{(110)_2}$ is recorded in the position $(011)_2 = 3$). The elimination of binary inversion from an algorithm requires an additional storage for computations (a working array of the same length as an original series has). Our comparison carried out for various binary inversion algorithms [324] shows that the time spent for rearrangement of data is the same for the best of them, comprises not so large $(5-10\,\%)$ part of computing cost of the whole FFT, and decreases with increasing $N$.

All versions of FFT with binary inversion are roughly equivalent. In the collection under consideration, the Cooley–Tukey time decimation algorithm with binary inversion of input data is implemented.

The computation of FFT by the fast Fourier transform method requires $O(N \log_2 N)$ complex multiplications and additions.

The basic modifications of the FFT algorithm can be generalized for the case when $N$ is composite [306]. If $N = n_1 n_2 \dots n_k$, the FFT algorithm requires $O(N(n_1 + n_2 + \dots + n_k))$ operations.

If $N$ is a prime number, then the DFT is reduced to circular convolution with the aid of some transformations of an original series. This convolution can be performed as a part of circular convolution for series of a greater length $N'$. Putting $N' = 2^M$ and using the convolution theorem, it is possible to use the FFT for series of length $N'$ in order to obtain the DFT for series of length $N$.

In this case, the number of operations required to compute the DFT for a series of length $N$ is proportional to $N \log_2 N$, but the coefficient of proportionality is greater than for the DFT of a series whose length is an integer power of 2 [231].

All the subroutines of this collection are written in FORTRAN, almost all of them do not use any specific features of FORTRAN language for a particular computer. Therefore, this collection can directly be used on any computer with a FORTRAN compiler.

**2.** In all the above subroutines, complex arrays are represented in the form of two real arrays in which the real and imaginary parts of complex numbers are recorded. This allows us to avoid the complex arithmetic and, then, to increase (to some extent) the speed of operation of the subroutines (in addition, their use for FORTRAN dialects without complex arithmetic is possible).

The accuracy of results depends basically on the accuracy of computation of $\cos a_n$ and $\sin a_n$, $a_n = 2\pi n/N$. The recursive computation of cos and sin is used in all subroutines of the collection. This allows us to increase the speed of operation with preservation of sufficiently high accuracy. If it is necessary, the accuracy of computations can be increased. To accomplish this, it is recommended to compute $\cos a_n$ and $\sin a_n$ directly. Our numerical experiments show that the use of double precision for recursive (or direct) computations of the above values does not cause a noticeable increase in accuracy, but leads to a considerable increase in time of operation of the subroutines.

The first part of the collection contains the subroutines for the discrete complex (one- and two-dimensional) and real (one-dimensional) Fourier transforms of series whose length is an integer power of 2. The subroutines for computing the Fourier coefficients of series of length $N = KL$, where $L = 2^M$, are also included into the collection. These subroutines allow one to obtain:

1) all the $N$ Fourier coefficients;

2) only $L$ Fourier coefficients;

3) all the $N$ Fourier coefficients of sparse series, i.e., of series for which most of terms are equal to zero.

These subroutines can be used, for example, to separate harmonic components in narrow ranges of frequencies and to restore functions at a large number of points. The collection also contains the subroutines for the Fourier transform of complex series whose length is a prime or composite number as well as the subroutines for harmonic analysis of even and odd real series.

Special tests are developed for all the subroutines included into the first part of the collection. For each subroutine (except for the real and two-dimensional Fourier transforms), the Fourier coefficients (divided by 32) for a complex series (32 points) are calculated. Then, the inverse Fourier transform is applied to the direct one. As a result, the original series should be obtained. An arithmetic progression is taken as an original series.

The test for the real Fourier transform differs from the above one only in that the real Fourier coefficients are printed when the direct transform is applied. The test for the two-dimensional Fourier transform is constructed in a similar way.

**3. Description of the subroutines.**

### Subroutine FFT

*Purpose.* This subroutine is intended for computing the discrete or inverse discrete Fourier transform of a complex series of length $L = 2^M$ ($M$ is integer) by the fast Fourier transform method. The subroutine allows one to compute the Fourier transform for a series $X$ of length $L$ allocated in an array $A$ of length $N \geq L$ in the following way:

$$X_r = A(IN + rk), \quad r = 0, 1, \ldots, L-1,$$

where $IN$ and $k$ are given integers ($N = rL$).

*Method.* The subroutine implements the Cooley–Tukey algorithm with binary inversion of input data. The direct or inverse discrete Fourier transform of a complex series $X$ of length $L$ is the calculation of the sum

$$A_r = \sum_{k=1}^{L-1} X_k W^{kr}, \quad r = 0, 1, \ldots, L-1, \tag{2.4.27}$$

where $W = e^{-\frac{2\pi i}{N}}$ for the direct and $W = e^{\frac{2\pi i}{N}}$ for the inverse Fourier transform.

The indices $k$ and $r$ can be represented in the form

$$k = k_{M-1}2^{M-1} + \cdots + k_1 2 + k_0, \quad r = r_{M-1}2^{M-1} + \cdots + r_1 2 + r_0, \tag{2.4.28}$$

where $k_\nu$ and $r_\nu$ are equal to 0 or 1 ($\nu = 0, 1, \ldots, M-1$).

Substituting (2.4.28) into the formula (2.4.27) and grouping the terms with respect to the indices $k_\nu$, we obtain

$$A_r = \sum_{k_0=0}^{1} \sum_{k_1=0}^{1} \ldots \sum_{k_{M-1}=0}^{1} X_k W^{rk_{M-1}2^{M-1}} \ldots W^{rk_0 2^0}.$$

Taking into account that

$$W^{rk_{M-J}2^{M-J}} = W^{(r_{J-1}2^{J-1}+\cdots+r_0)k_{M-J}2^{M-J}},$$

we come to the formula

$$A_{r_{M-1}2^{M-1}+\dots+r_0} = \sum_{k_0=0}^{1} W^{(r_{M-1}2^{M-1}+\dots+r_0)k_0 2^0} \times \\ \times \left[ \sum_{k_1=0}^{1} W^{(r_{M-2}2^{M-2}+\dots+r_0)k_1 2^1} \left[ \cdots \right.\right. \\ \left.\left.\left. \cdots \left[ \sum_{k_{M-1}=0}^{1} W^{r_0 k_{M-1} 2^{M-1}} X_{k_{M-1}\ 2^{M-1}\cdot+\dots+k_0} \right]\right]\right]. \tag{2.4.29}$$

The computations by the formula (2.4.29) are performed recurrently and the intermediate and final results are recorded in the same two arrays, where the original series was allocated.

To obtain the coefficients $A_r$ in the order of increasing $r$, it is necessary that the original series be represented in the binary-inverse order. Finally, the formulas take the following form: $A_0(r) = X(I(r))$, where $I(r)$ is the binary inversion of $r$, $r = 0, 1, \dots, L-1$,

$$A_J(2^J k + r) = A_{J-1}(2^J k + r) + W^{r2^{M-J}} A_{J-1}(2^J k + r + 2^{J-1}),$$
$$A_J(2^J k + r + 2^{J-1}) = A_{J-1}(2^J k + r) - W^{r2^{M-J}} A_{J-1}(2^J k + r + 2^{J-1}),$$
$$J = 1, \dots, M, \quad r = 0, 1, \dots, 2^{J-1} - 1, \quad k = 0, 1, \dots, 2^{M-J} - 1.$$

**Subroutine FFTKL**

*Purpose.* This subroutine is intended for computing the discrete or inverse discrete Fourier transform of a complex series of length $N$ by the accelerated Fourier transform method. The length $N$ is assumed to be $N = KL$, where $L = 2^J$, $K$ and $J$ are integer. This restriction is weaker than the restriction imposed on the length of a series in the subroutine FFT.

*Method.* Let the length $N$ of a complex series $A$ can be factored in the form $N = KL$, where $K$ is integer and $L$ is an integer power of 2. Let us consider the problem of computing the inverse discrete Fourier transform of a series $A$:

$$X_r = \sum_{m=0}^{N-1} e^{\frac{2\pi i}{N} rm} A_m, \quad r = 0, 1, \dots, N-1. \tag{2.4.30}$$

Let us represent the indices $m$ and $r$ in the form

$$m = lK + k, \quad l = 0, 1, \dots, L-1, \quad k = 0, 1, \dots, K-1;$$
$$r = jL + n, \quad j = 0, 1, \dots, K-1, \quad n = 0, 1, \dots, L-1.$$

Let $B_l^{(k)} = A_{lK+k}$. Then, the formula (2.4.30) can be rewritten as

$$X_{jL+n} = \sum_{k=0}^{K-1} e^{\frac{2\pi i}{N}(jL+n)k} \left( \sum_{l=0}^{L-1} B_l^{(k)} e^{\frac{2\pi i}{L} ln} \right),$$
$$n = 0, 1, \ldots, L-1, \quad j = 0, 1, \ldots, K-1.$$

When $k$ is fixed, the expression in the round brackets is the inverse discrete Fourier transform of a series of length $L$. For its calculation, the fast Fourier transform algorithm is used:

$$Y_n^{(k)} = \sum_{l=0}^{L-1} B_l^{(k)} e^{\frac{2\pi i}{L} ln}, \quad n = 0, 1, \ldots, L-1.$$

The vectors $Y_n^{(k)}$ are recorded into the storage occupied by the original series $A$. Then, for all possible $n$ and $j$ the following values are calculated:

$$X_{jL+n} = \sum_{k=0}^{K-1} e^{\frac{2\pi i}{N}(jL+n)k} \; Y_n^{(k)},$$
$$n = 0, 1, \ldots, L-1, \quad j = 0, 1, \ldots, K-1.$$

In the case of direct Fourier transform, the computing scheme is similar with the only exception that the sign in the exponent is changed.

### Subroutine FFTL

*Purpose.* This subroutine is intended for computing the $L$ ($L = 2^J$, $J$ is integer) values of the discrete or inverse discrete Fourier transform for a complex series of length $N = KL$ with the use of the fast Fourier transform algorithm. The coefficients to be computed are stored in one of the $K$ segments of length $L$: $X_{ML}$, $X_{ML+1}, \ldots, X_{ML+(L-1)}$, $0 \le M \le K-1$, ($M$ is integer; the series is divided into these segments). The subroutine can be used, for example, to separate the harmonic components in some frequency range.

*Method.* Let the length $N$ of a complex series $A$ can be factored in the form $N = KL$, where $K$ is integer and $L$ is an integer power of 2. For the inverse discrete Fourier transform of a series, we consider the problem of computing the coefficients with the numbers $ML$, $ML+1, \ldots, ML+(L-1)$, where $M$ is integer,

$$X_r = \sum_{m=0}^{N-1} A_m e^{\frac{2\pi i}{N} rm}, \quad r = ML, ML+1, \ldots, ML+(L-1). \qquad (2.4.31)$$

Let us represent the indices $m$ and $r$ in the form

$$\begin{aligned} m &= lK + k, \quad l = 0, 1, \ldots, L-1, \quad k = 0, 1, \ldots, K-1, \\ r &= ML + n, \quad n = 0, 1, \ldots, L-1. \end{aligned}$$

Let $B_l^{(k)} = A_{lK+k}$. Then, the formula (2.4.31) can be rewritten as

$$X_{ML+n} = \sum_{k=0}^{K-1} e^{\frac{2\pi i}{N}(ML+n)k} \left( \sum_{l=0}^{L-1} B_l^{(k)} e^{\frac{2\pi i}{N} ln} \right), \quad n = 0, 1, \ldots, L-1.$$

When $k$ is fixed, the expression in the round brackets is the inverse discrete Fourier transform of a series of length $L$. For its computation, the fast Fourier transform algorithm is used:

$$Y_n^{(k)} = \sum_{l=0}^{L-1} B_l^{(k)} e^{\frac{2\pi i}{L} ln}, \quad n = 0, 1, \ldots, L-1.$$

The vectors $Y^{(k)}$ are recorded into the storage occupied by the original series $A$.

Then, for each $n$ the following coefficients are computed:

$$X_{ML+n} = \sum_{k=0}^{K-1} e^{\frac{2\pi i}{N}(ML+n)} Y_n^{(k)}, \quad n = 0, 1, \ldots, L-1.$$

In the case of direct Fourier transform, the computing scheme is the same with the only exception that the sign in the exponent is changed.

**Subroutine FFTLN**

*Purpose.* This subroutine is intended for computing the $N$ values of the discrete or inverse discrete Fourier transform of a complex series $A$ of length $N$ whose number of nonzero terms is equal to or less than $L$. $N$ should be equal to $KL$, where $K$ is integer, $L = 2^J$, and $J$ is integer. All the nonzero terms of the series should be stored in one of the $K = N/L$ segments of length $L$ into which the series is divided: $A_{ML}$, $A_{ML+1}$, ..., $A_{ML+L-1}$, $0 \le M \le K-1$, $M$ is integer.

The subroutine can be used to compute the $N$ values of a function, using the $L$ values of the Fourier coefficients.

*Method.* Let us consider the problem of computing the inverse discrete Fourier transform of a complex series $A$ of length $N$:

$$X_r = \sum_{n=0}^{N-1} e^{\frac{2\pi i}{N} rn} A_n, \quad r = 0, 1, \ldots, N-1. \tag{2.4.32}$$

Since only the $L$ successive terms of the series $A$ may be different from zero, starting from the number $ML$, then the formula (2.4.32) can be rewritten as

$$X_r = \sum_{n=ML}^{ML+L-1} e^{\frac{2\pi i}{N} rn} A_n, \quad r = 0, 1, \ldots, N-1. \tag{2.4.33}$$

Let us represent the indices $n$ and $r$ in the form

$$\begin{aligned} n &= ML + j, \quad j = 0, 1, \ldots, L-1, \quad r = LK + k, \\ l &= 0, 1, \ldots, L-1, \quad k = 0, 1, \ldots, K-1, K = N/L. \end{aligned}$$

Let $B_j^{(k)} = A_{ML+j} e^{\frac{2\pi i}{N}(ML+j)k}$. Then, the formula (2.4.33) can be rewritten as follows:

$$X_{lK+k} = \sum_{j=0}^{L-1} e^{\frac{2\pi i}{N} jl} B_j^{(k)}. \tag{2.4.34}$$

When $B_j^{(k)}$ are computed, the $j$th components of all vectors $B^{(k)}$ are computed consistently (at first $B_1^{(k)}$, $k = 0, 1, \ldots, K-1$, then $B_2^{(k)}$, etc.). When $k$ is fixed, the formula (2.4.34) represents the Fourier transform of the vector $B^{(k)}$. The fast Fourier transform method is used to implement the formula (2.4.34).

The formulas for the direct Fourier transform differ only by the sign in the exponent.

### Subroutine FFT2

*Purpose.* This subroutine is intended for computing the two-dimensional discrete or inverse discrete Fourier transform of a complex matrix of size $N_1 \times N_2$ such that $N_1 = 2^{J_1}$ and $N_2 = 2^{J_2}$, where $J_1$ and $J_2$ are integer.

*Method.* Let us consider the direct or inverse Fourier transform of a complex $N_1 \times N_2$-matrix $A$:

$$X_{r_1,r_2} = \sum_{k_1=0}^{N_1-1} \sum_{k_2=0}^{N_2-1} A_{k_1,k_2} e^{\pm 2\pi i \left(\frac{k_1 r_1}{N_1} + \frac{k_2 r_2}{N_2}\right)};$$
$$r_1 = 0, 1, \ldots, N_1 - 1; \quad r_2 = 0, 1, \ldots, N_2 - 1,$$

where the signs "minus" and "plus" in the exponent are used for the direct and inverse two-dimensional discrete Fourier transforms, respectively.

The formula for the inverse two-dimensional Fourier transform can be rewritten in the form

$$X_{r_1,r_2} = \sum_{k_1=0}^{N_1-1} e^{\frac{2\pi i}{N_1} k_1 r_1} \left( \sum_{k_2=0}^{N_2-1} e^{\frac{2\pi i}{N_2} k_2 r_2} A_{k_1,k_2} \right),$$
$$r_1 = 0, 1, \ldots, N_1 - 1, \quad r_2 = 0, 1, \ldots, N_2 - 1.$$

Thus, the computation of the two-dimensional Fourier transform is reduced to the successive application of one-dimensional Fourier transforms for which the fast Fourier transform method is used.

The direct two-dimensional Fourier transform is computed in a similar way.

### Subroutine RFFT

*Purpose.* This subroutine is intended for computing the coefficients of the real Fourier transform or for computing the values of a function when its real Fourier coefficients are known. The length $N_2$ of a real series (to be transformed or computed) should be an integer power of 2.

*Method.* Let us consider the problem of computing the Fourier coefficients of a real series: $X_0,\ldots,X_{N_2-1}$. First we find the discrete Fourier transform of the complex series $X_{2j}+iX_{2j+1}$:

$$Y_k = \frac{1}{N}\sum_{j=0}^{N-1}\left(X_{2j}+iX_{2j+1}\right)e^{-\frac{2\pi i}{N}jk}, \quad k=0,1,2,\ldots,N-1; \quad N=N_2/2.$$

The further computations are performed by the following formulas:

$$T_k = e^{i\left(\frac{\pi}{2}-\frac{\pi}{N}k\right)}\left(Y_k-\overline{Y}_{N-k}\right),$$
$$U_k = \left(\overline{Y_k}+Y_{N-k}\right), \quad k=1,2,\ldots,N/2-1.$$

The coefficients $A_k$ and $B_k$ of the real Fourier transform are computed by the formulas:

$$\begin{aligned}
&A_k = \frac{1}{2}\operatorname{Re}\left(U_k-\overline{T_k}\right), && B_k = \frac{1}{2}\operatorname{Im}\left(U_k-\overline{T_k}\right),\\
&A_{N-k} = \frac{1}{2}\operatorname{Re}\left(\overline{U_k}+T_k\right), && B_{N-k} = \frac{1}{2}\operatorname{Im}\left(\overline{U_k}+T_k\right),\\
&k=1,2,\ldots,N/2-1; &&\\
&A_{N/2} = \operatorname{Re}Y_{N/2}, && B_{N/2} = \operatorname{Re}Y_{N/2},\\
&A_0 = \operatorname{Re}Y_0+\operatorname{Im}Y_0, && B_0 = 0,\\
&A_N = \operatorname{Re}Y_0-\operatorname{Im}Y_0, && B_N = 0.
\end{aligned}$$

The following algorithm is used to compute the values of a function on the basis of its real Fourier coefficients $A_l$ and $B_l$. At first stage, the computations are carried out by the formulas ($k = 1, 2, \ldots, N/2 - 1$)

$$
\begin{aligned}
&Z_l = A_l + iB_l, && l = 0, 1, \ldots, N-1;\\
&T_k = e^{i(\pi/2 - \pi k/N)}\left(Z_k - \overline{Z}_{N-k}\right), && U_k = \overline{Z_k} + Z_{N-k},\\
&Y_k = \frac{1}{2}\left(U_k - \overline{T_k}\right), && Y_{N-k} = \left(\overline{U_k} + T_k\right),\\
&Y_{N/2} = Z_{N/2}, && Y_0 = \frac{1}{2}\left(A_0 + A_N\right) + \frac{1}{2}\, i\left(A_0 - A_N\right).
\end{aligned}
$$

At the second stage, the inverse discrete Fourier transform of the complex series $Y_k$ is computed:

$$
X_{2j} + iX_{2j+1} = \sum_{k=0}^{N-1} Y_k e^{\frac{2\pi i}{N} jk}, \quad j = 0, 1, \ldots, N-1.
$$

The fast Fourier transform algorithm is used in the above scheme. The real and imaginary parts of the resulting series are the even and odd terms of the sought-for series, respectively.

## 2.5 Fast and accelerated solution of special linear systems

**1.** For mathematical description of some physical systems, it is necessary to solve the following minimization problem:

$$
\widehat{\varpi}_{\text{opt}} = \arg \min_{\widehat{\varpi} \in R_{2n}} e^2(\widehat{\varpi}), \tag{2.5.35}
$$

where $\widehat{\varpi} = \left(\widehat{\varpi}^1, \widehat{\varpi}^2\right)'$, $\widehat{\varpi}^i \in R_n$, $i = 1, 2$, is the composite vector and the prime indicates the operation of transposition. The objective functional specifying the quality of the system is defined by the formula

$$
\begin{aligned}
e^2(\widehat{\varpi}) \;=\; & (A_{11}\widehat{\varpi}^1, \widehat{\varpi}^1) + 2(A_{12}\widehat{\varpi}^1, \widehat{\varpi}^2) + \\
& + (A_{22}\widehat{\varpi}^2, \widehat{\varpi}^2) - 2(\widehat{f^1}, \widehat{\varpi}^1) - 2(\widehat{f^2}, \widehat{\varpi}^2) + e_0^2.
\end{aligned}
$$

Here the scalar product $(u, v)$ is defined as $\sum_{i=1}^n u_i v_i$, $u, v \in R_n$, $A_{ij}$ are square matrices of order $n$ such that $A_{ii} = A_{ii}'$, $i = 1, 2$, $A_{12} = A_{21}'$, $\widehat{f} = (\widehat{f^1}, \widehat{f^2})'$ is a given vector, and $e_0^2$ is a constant.

Let us introduce the block matrix of order $2n$:

$$
A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.
$$

Then,

$$e^2(\widehat{\varpi}) = [A\widehat{\varpi}, \widehat{\varpi}] - 2\left[\widehat{f}, \widehat{\varpi}\right] + e_0^2,$$

where the symbol $[\cdot, \cdot]$ indicates the scalar product in $R_{2n}$.

Under the above conditions, the matrix $A = A' \geq 0$. For a given $\widehat{f}$, there exists an optimal vector $\widehat{\varpi}_{\text{opt}}$ (perhaps, not unique). This vector can be obtained as a solution to the linear system

$$A\widehat{\varpi} = \widehat{f}. \tag{2.5.36}$$

Indeed, the symmetry of $A$ is obvious. Let us reduce it to the diagonal form $D$ by an orthogonal transformation $Q$: $A = Q'DQ$.

If we assume that there exists $d_{ii} < 0$, then by diagonalizing the form $e^2(\widehat{\varpi})$ it is easy to construct a vector $\widehat{\varpi}$ for which this form takes a negative value, i.e., we come to the contradiction. The existence of the vector $\widehat{\varpi}_{\text{opt}}$ can be established in a similar way. The transition to the system (2.5.36) is trivial.

**Remark.** *If the system* (2.5.36) *is solvable, then any its solution may be taken as* $\widehat{\varpi}_{\text{opt}}$. *At the same time, the system* (2.5.36) *is uniquely solvable for any* $\hat{f}$ *if and only if the matrix* $A$ *is positive definite:* $A > 0$.

In the case under consideration, the solvability of (2.5.36) is exclusively a consequence of nonnegativity of the functional $e^2(\hat{\varpi})$ even when $A \geq 0$. In addition, the equality

$$e^2(\widehat{\varpi}_{\text{opt}}) = e_0^2 - \left[\widehat{f}, \widehat{\varpi}_{\text{opt}}\right]$$

holds and the following condition of orthogonality is satisfied with necessity:

$$\left[\widehat{f} - A\widehat{\varpi}_{\text{opt}}, \widehat{\varpi}_{\text{opt}}\right] = 0.$$

When the system (2.5.36) is solved numerically, the following difficulties may arise:

1) the matrix $A$ may be singular or near singular (ill conditioned);

2) the order of (2.5.36) may be large (especially, when $m \geq 2$, where $m$ is the number of blocks in $A$).

It is well known that it is required $\sim (mn)^3$ operations and $\sim (mn)^2$ words of storage to solve the system (2.5.36) with $m$ blocks even in the case $A > 0$; these parameters cannot be improved [57]. In the sequel, we consider the system (2.5.36) with special matrices. This allows us to construct substantially more economical algorithms in the sense of speed of operation and in the sense of computer storage required.

Let us introduce some necessary definitions.

The matrix

$$T = \begin{pmatrix} t_0 & t_1 & t_2 & \dots & t_{n-2} & t_{n-1} \\ t_{-1} & t_0 & t_1 & \dots & t_{n-3} & t_{n-2} \\ t_{-2} & t_{-1} & t_0 & \dots & t_{n-4} & t_{n-3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ t_{-(n-2)} & t_{-(n-3)} & t_{-(n-4)} & \dots & t_0 & t_1 \\ t_{-(n-1)} & t_{-(n-2)} & t_{-(n-3)} & \dots & t_{-1} & t_0 \end{pmatrix}$$

defined by its $2n-1$ elements $t_i$, $i = -(n-1), \dots, 0, 1, \dots, (n-1)$, is said to be a *Toeplitz* matrix.

The matrix $T$ is called a *block-Toeplitz* matrix if $t_i$, $i = -(n-1), \dots, n-1$, are square matrices of order $m$.

A matrix $A$ is referred to as a *t-type* matrix if it consists of $m^2$ blocks of order $n$ and each block is a Toeplitz matrix.

In what follows we assume that all matrices $A$ are $t$-type with $m \ll n$; this condition defines the class of problems under consideration. Without loss of generality, moreover, we assume that almost everywhere $m = 2$.

Our fast algorithm for solving (2.5.36) is based on the following assertion.

**Assertion**. *Any t-type matrix A can be reduced to a block-Toeplitz matrix T by permutation of rows and the corresponding columns.*

Indeed, let the block $A_{11}$ of $A$ be defined by the elements ($m = 2$)

$$a_{-(n-1)}, a_{-(n-2)}, \dots, a_0, \dots, a_{n-1},$$

the block $A_{12}$ be defined by the elements

$$c_{-(n-1)}, \dots, c_0, \dots, c_{n-1},$$

and the block $A_{22}$ be defined by the elements

$$b_{-(n-1)}, b_{-(n-2)}, \dots, b_0, b_1, \dots, b_{n-1}.$$

Let us form the block-Toeplitz matrix $T$ with the elements

$$t_0 = \begin{pmatrix} a_0 & c_0 \\ c_0 & b_0 \end{pmatrix}, \quad t_1 = \begin{pmatrix} a_1 & c_1 \\ c_{-1} & b_1 \end{pmatrix}, \dots,$$

$$t_i = \begin{pmatrix} a_i & c_i \\ c_{-i} & b_i \end{pmatrix}, \quad t_{n-1} = \begin{pmatrix} a_{n-1} & c_{n-1} \\ c_{-(n-1)} & b_{n-1} \end{pmatrix}, \quad t_{-i} = t_i'.$$

It is easy to see that the transformation $A \to T$ is defined by the permutation

$$\Pi = \begin{pmatrix} 1 & 2 & 3 & . & . & . & n & n+1 & . & . & . & 2n-1 & 2n \\ 1 & 3 & 5 & . & . & . & 2n-1 & 2 & . & . & . & 2n-2 & 2n \end{pmatrix},$$

i.e., the $(2i-1)$th row (column) of $T$ is replaced by the $i$th row (column) of $A$ and the $2i$th row (column) of $T$ is replaced by the $(n+i)$th row (column) of $A$ for $i = 1, 2, \ldots, n$.

Obviously, $T' = T \geq 0$; if $A > 0$, then $T > 0$.

Let us introduce a permutation matrix $P$ such that $PAP = T$. Then, the system (2.5.36) is reduced to the equivalent system with the block-Toeplitz matrix $T$:

$$T\hat{y} = \hat{g}. \tag{2.5.37}$$

Here $\hat{y} = P\hat{\varpi}$ and $\hat{g} = P\hat{f}$.

The algorithm proposed in [59] can be used to solve the system (2.5.37) efficiently. In order to find $\hat{y}$, in general it takes only $\sim m^3n^2$ arithmetic operations and $\sim m^2n$ words of storage. The vector $\hat{\varpi}$ is obtained by the inverse permutation $\Pi^{-1}$ applied to the vector $\hat{y}$ in $\sim mn$ operations.

The above computing process is realizable if the matrix under consideration has a certain reserve of positive definiteness. As a rule, this is not the case for a large $n$. Then, the system (2.5.36) can be solved by the simplified Tikhonov method of regularization [257]. Let us consider the linear system

$$A_\alpha \hat{\varpi} = \hat{f} \tag{2.5.38}$$

with the matrix $A_\alpha \equiv A + \alpha L$, where $\alpha > 0$ is the regularization parameter and $L$ is a symmetric nonnegative matrix such that $A+L > 0$. Then, obviously, $A_\alpha > 0$ for any $\alpha > 0$.

If the matrix $L$ is of the same structure as $A$, then the fast algorithm discussed in [59] can be used to solve the regularized system (2.5.38). In general, it is not the case and the efficient solution of (2.5.38) becomes impossible.

A matrix $\tilde{T}$ is called a *block-quasi-Toeplitz* matrix if $\tilde{T} = T + T_0$, where $T$ is a block-Toeplitz matrix and the columns with their numbers $i_1, i_2, \ldots, i_s$ $(s \ll n)$ in the matrix $T_0$ may not be equal to zero only.

Let us consider the nonsingular system

$$\tilde{T}\hat{y} = \hat{g}$$

with a block-quasi-Toeplitz matrix $\tilde{T}$. It is assumed that the matrix $T$ satisfies the conditions for which the fast algorithm becomes applicable. This system can be rewritten in the form

$$T\hat{y} + \sum_{k=1}^{s} y_{i_k}\tilde{t}_{i_k} = \hat{g},$$

where $\tilde{t}_{i_k}$ are the column vectors of the matrix $T_0$. The fast algorithm can be used to find the solutions $u_k$ to the linear systems

$$Tu_k = \tilde{t}_{i_k}, \quad k = 1, 2, \ldots, s+1,$$

where $\tilde{t}_{i_{s+1}} = \hat{g}$. Then, it is obvious that

$$\hat{y} + \sum_{k=1}^{s} u_k y_{i_k} = u_{s+1}. \tag{2.5.39}$$

Let $e_i$ be the elements of the Cartesian basis in $R_{2n}$ ($m = 2$). Multiplying the equation (2.5.39) scalarly by $e_{i_r}$, $r = 1, 2, \ldots, s$, we obtain the system of equations

$$y_{i_r} + \sum_{k=1}^{s} y_{i_k} \left[u_k, e_{i_r}\right] = \left[u_{s+1}, e_{i_r}\right], \quad r = 1, 2, \ldots, s,$$

from which $y_{i_k}$ are uniquely determined; using (2.5.39), then, we can find the vector $\hat{y}$.

The above scheme can be applied to the regularized system (2.5.38), in particular, if the matrix $L$ has the form ($m = 2$)

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix},$$

where each block is of order $n$ and $L_i$ are quasi-Toeplitz matrices.

**2.** Let us consider a gradient-type accelerated algorithm for minimization of the functional $e^2(\hat{\varpi})$. The computation of the gradient

$$\operatorname{grad} e^2(\hat{\varpi}) = 2(A\hat{\varpi} - \hat{f}\ )$$

is the main operation at each step of the algorithm and its efficiency depends on the time it takes to perform the operation

$$A\hat{\varpi} = \begin{pmatrix} A_{11}\hat{\varpi}^1 + A_{12}\hat{\varpi}^2 \\ A_{21}\hat{\varpi}^1 + A_{22}\hat{\varpi}^2 \end{pmatrix},$$

where $A_{ij}$ are Toeplitz matrices.

Now we turn to the algorithm of accelerated calculation of the product $Tx$, where $T$ is a Toeplitz matrix and $x \in R_N$. Note that for the direct calculation of $Tx$ it is easy to save storage, but the required time is proportional to $N^2$.

It is well known that the Toeplitz matrix

$$C = \begin{pmatrix} c_0 & c_1 & c_2 & \cdot & \cdot & \cdot & c_{N-2} & c_{N-1} \\ c_{N-1} & c_0 & c_1 & \cdot & \cdot & \cdot & c_{N-3} & c_{N-2} \\ c_{N-2} & c_{N-1} & c_0 & \cdot & \cdot & \cdot & c_{N-4} & c_{N-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_1 & c_2 & c_3 & \cdot & \cdot & \cdot & c_{N-1} & c_0 \end{pmatrix}$$

defined by the vector $c = (c_0, c_1, \dots, c_{N-1})$ is said to be *circulant*.

The computation of the vector $Cy$, $y \in R_N$, can be performed as follows:

1) to apply the discrete Fourier transform [196] to the vectors $c$ and $y$;

2) to compute the componentwise product of the images of $c$ and $y$;

3) to apply the inverse discrete Fourier transform to the resulting vector.

When $N = 2^r$ ($r$ is a natural number), all the above stages require $\sim N \log_2 N$ arithmetic operations if the DFT is calculated on the basis of the FFT algorithm [196] (in the case of direct calculations, it takes $N^2$ operations).

Now we return to the calculation of the vector $Tx$, $x \in R_N$. Let us put $c_i = t_i$, $i = 0, 1, 2, \dots, n-1$, $c_{N-i} = t_{-i}$, $i = 1, 2, \dots, n-1$. Let $N = 2^r$, where $r$ is the least natural number for which $N \geq 2n - 1$. The other $c_i = 0$. Let the circulant matrix $C$ be defined by the vector $c$ and the vector $y = (x, 0)'$. Then, when $z = Cy$, it is easy to see that the first $n$ components of this vector define the sought-for vector $Tx$. The computation of $z$ can be performed, according to the above accelerated algorithm, in $\sim n \log_2 n$ operations.

Thus, the following assertion is valid.

**Assertion.** *For any $t$-type matrix $A$ the expression $A\hat{\varpi}$ can be computed in $\sim m^2 n \log_2 n$ operations.*

**Remark 1.** *If the system* (2.5.36) *is solved by the conjugate gradient method, then it is sufficient to perform $\sim rm^2 n \log_2 n$ operations, where $r = \operatorname{rank} A \leq mn$. In the case of regularized system* (2.5.38), *the matrix $A_\alpha$ is always of full rank. The accelerated algorithm is also applicable if $L$ is a $t$-type matrix or slightly differs from it* (*for example, can be reduced to the block-quasi-Toeplitz form*).

**Remark 2.** *A similar situation takes place if the matrix $L$ is sparse, i.e., if the number of its nonzero elements is small, which usually is the case in practice.*

**Remark 3.** *In order to accelerate the process of solving a regularized system, the following technique* [159] *can be used. Let a solution $\hat{\varpi}_{\alpha_1}$ to the system* (2.5.38) *be known for some $\alpha_1$; it is required to find a solution $\hat{\varpi}_{\alpha_2}$ for $\alpha_2 \approx \alpha_1$. It is established in* [159] *that the regularized solutions are "uniformly" close with respect to the regularization parameter. For the above methods, therefore, it is natural to take $\hat{\varpi}_{\alpha_1}$ as an initial approximation when searching for $\hat{\varpi}_{\alpha_2}$. In particular, if $L > 0$, then $\hat{\varpi}_{\alpha_1} \equiv 0$ for $\alpha_1 = +\infty$. The successive application of this technique to the solution of the system* (2.5.38) *together with the accelerated algorithm for computing gradients is very efficient in practice.*

Let us consider the regularized functional

$$e_\alpha^2(\hat{\varpi}) \equiv [A\hat{\varpi}, \hat{\varpi}] - 2\left[\hat{f}, \hat{\varpi}\right] + e_0^2 + \alpha\,[L\hat{\varpi}, \hat{\varpi}]\,,$$

where $0 < L = L'$. Let $L$ be represented in the form $L = K'K$, where $K$ is the upper triangular matrix obtained, for example, by the Cholesky method. Putting $K\hat{\varpi} = \hat{y}$, we obtain

$$e_\alpha^2(\hat{\varpi}) = [B\hat{y}, \hat{y}] - 2\,[\hat{g}, \hat{y}] + e_0^2 + \alpha\,[\hat{y}, \hat{y}] \equiv E_\alpha^2\,(\hat{y})\,,$$

where $B = \left(K^{-1}\right)' AK^{-1}$ and $\hat{g} = \left(K^{-1}\right)' \hat{f}$.

It is easy to see that the problem of minimization of the functional $e_\alpha^2(\hat{\varpi})$ in $\hat{\varpi}$ is equivalent to the problem of finding

$$\hat{y}_\alpha = \arg\min_{\hat{y}} E_\alpha^2\,[\hat{y}]$$

and computing $\hat{\varpi}_\alpha = K^{-1}\hat{y}_\alpha$. If we assume that the matrix $L$ is of the form

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix},$$

where $L_1$ and $L_2$ are band matrices of moderate width, then the matrix $K$ is represented in the form

$$K = \begin{pmatrix} K_1 & 0 \\ 0 & K_2 \end{pmatrix},$$

where $K_i$ are upper band triangular matrices. Then, the number of operation for decomposition of $L$ and for solving the system $K\hat{x} = \hat{z}$ is insignificant. In order to compute

$$\operatorname{grad} E_\alpha^2\,(\hat{y}) \equiv B\hat{y} - \hat{g} + \alpha\hat{y},$$

it is possible to use the above reasoning (which removes the necessity of computing $K^{-1}$) or to apply the algorithm for accelerated multiplication of $t$-type matrices by a vector, which leads also to $\sim m^2 n \log_2 n$ operations in total.

The fast computation of gradients can be very useful in other situations (in particular, for the descriptive regularization method [191]–[193] when the functional $e_\alpha^2(\hat{\varpi})$ is minimized on a set of piecewise-monotone or piecewise-convex vector functions).

**3.** A number of physical problems lead to the necessity of numerical solution of Wiener–Hopf integral equations of the first kind with symmetric nonnegative definite kernels of the form

$$Au \equiv \int_0^\infty K(x-t)u(t)\,dt = f(x), \quad x \geq 0, \tag{2.5.40}$$

where $K(x)$, $f(x)$ are given functions and $u(x)$ is the sought-for function.

In practice, the right-hand side and the kernel in equation (2.5.40) are given, as a rule, discretely and, hence, approximately. Taking into account the instability of a solution to equation (2.5.40) with respect to small perturbations in input data, it is impossible to search for this solution in the classical sense, i.e., according to the rule $u = A^{-1}f$, where $A^{-1}$ is the inverse of the operator $A$ in (2.5.40). In order to overcome this difficulty, the simplified Tikhonov method of regularization is proposed in [206], [199] (in particular, for the problem of designing optimal linear systems for automatic control); in the classical case, this problem can be reduced to the solution of equation (2.5.40) with a symmetric nonnegative definite kernel.

According to [199] and to the approach discussed at the beginning of this section, for numerical solution of the regularized problem we come to the necessity of solving three linear systems with the same symmetric Toeplitz positive definite matrix.

It is well known that the solution of a linear system with a Toeplitz matrix requires $\sim N^2$ operations, where $N$ is the order of the matrix. The algorithm described in [196] and modified for the case of symmetric matrices is efficient for solution of such systems. This algorithm reduces the number of operation and computing cost in half. Moreover, using a specific property of the right-hand sides in the other two systems and using partially the results of computations obtained during the solution of the first system, it is possible to construct the sought-for solutions in $\sim N$ operations only. Thus, the solution of the regularized problem is obtained practically with the same efficiency as the single solution of a linear system with a symmetric Toeplitz matrix.

The integral equation (2.5.40) with a smooth nonnegative definite symmetric kernel $K$ is approximately solved by Tikhonov's method of simplified one-parametric regularization of first order by minimization of the following smoothing functional in $u$:

$$\Phi_\alpha[u] \equiv \int_0^\infty \int_0^\infty K(x-t)u(x)u(t)\,dx\,dt - 2\int_0^\infty f(x)u(x)\,dx + \alpha \int_0^\infty u'^{\,2}(x)\,dx.$$

Here $\alpha > 0$ is the regularization parameter and $\int_0^\infty u'^{\,2}(x)\,dx$ is the stabilizing functional.

Assuming that $K(t) \to 0$, $u(t) \to 0$, $u'(t) \to 0$, $f(t) \to 0$ (rapidly enough) as $t \to \infty$ and that the functions $K(t)$ and $f(t)$ are given on a finite segment $[0, T]$ (perhaps, discretely), we come to the following representation of the smoothing functional:

$$\Phi_\alpha[u] \approx \int_0^T \int_0^T K(x-t)u(x)u(t)\,dx\,dt - 2\int_0^T f(x)u(x)\,dx + \alpha \int_0^T u'^{\,2}(x)\,dx.$$

For discretization of the first and second terms in this representation, the trapezoidal quadrature formula on the uniform grid $0 = x_1 < x_2 < \ldots < x_N = T$, $x_{i+1} - x_i = h$, is used; the forward differences

$$\frac{du(x_i)}{dx} \approx (u(x_{i+1}) - u(x_i))/h, \quad i = 1, 2, \ldots, N-1,$$

are used to approximate the third term. It is assumed that $u'(T) = 0$. Then, the discrete analogue of the above representation is of the form

$$h^2(\hat{D}\hat{K}\hat{D}\hat{u}, \hat{u}) - 2h(\hat{D}\hat{f}, \hat{u}) + \alpha(\hat{C}\hat{u}, \hat{u}). \tag{2.5.41}$$

Here $\hat{D} = \operatorname{diag}\{d_{11}, \ldots, d_{NN}\}$, $d_{11} = d_{NN} = 1/2$, $d_{ii} = 1$, $i = 2, \ldots, N-1$; $\hat{K}$ is the symmetric Toeplitz matrix defined by the elements of the first raw $K_{1i} = K(x_i)$; the vectors $\hat{u} = (u_1, u_2, \ldots, u_N)'$, $\hat{f} = (f_1, \ldots, f_N)'$, $f_i = f(x_i)$, $i = 1, 2, \ldots, N$; $\hat{C}$ is a symmetric tridiagonal quasi-Toeplitz matrix (this matrix approximates the stabilizing functional and, hence, is nonnegative definite).

The problem of minimizing the functional (2.5.41) in $\hat{u}$ is equivalent to the solution of the linear system

$$h^2\hat{D}\hat{K}\hat{D}\hat{u} + \alpha\hat{C}\hat{u} = h\hat{D}\hat{f}. \tag{2.5.42}$$

Putting $\hat{D}\hat{u} = \hat{y}$, we obtain

$$h^2\hat{K}\hat{y} + \alpha\hat{D}^{-1}\hat{C}\hat{D}^{-1}\hat{y} = h\hat{f}, \tag{2.5.43}$$

where $\hat{D}^{-1}$ is the inverse of $\hat{D}$. It is easy to show that $\hat{D}^{-1}\hat{C}\hat{D}^{-1} = \hat{C}' + \hat{C}''$, where $\hat{C}'$ is the symmetric Toeplitz (tridiagonal) matrix defined by the two elements $c'_{11} = 2/h$ and $c'_{12} = -1/h$, whereas the elements of $\hat{C}''$ are equal to zero, except for

$$c''_{22} = c''_{N-1,N-1} = -1/(2h).$$

The solution of equation (2.5.43) can be reduced to the solution of the following equation:

$$\hat{T}\hat{y} = \frac{\alpha}{2h}y_2\hat{e}_2 + \frac{\alpha}{2h}y_{N-1}\hat{e}_{N-1} + h\hat{f}. \tag{2.5.44}$$

Here $\hat{T} = h^2\hat{K} + \alpha\hat{C}'$ is a symmetric Toeplitz positive definite matrix,

$$\hat{e}_2 = (0, 1, 0, \ldots, 0)', \quad \hat{e}_{N-1} = (0, \ldots, 0, 1, 0)',$$

$y_2$ and $y_{N-1}$ are the second and penultimate elements of the vector $\hat{y}$, respectively.

Let us solve the following auxiliary equations:

$$\hat{T}\hat{z} = h\hat{f}, \tag{2.5.45}$$

$$\hat{T}\hat{z}' = \frac{\alpha}{2h}\,\hat{e}_2, \tag{2.5.46}$$

$$\hat{T}\hat{z}'' = \frac{\alpha}{2h}\,\hat{e}_{N-1}. \tag{2.5.47}$$

Then, from (2.5.44) we obtain

$$\hat{y} = y_2\hat{z}' + y_{N-1}\hat{z}'' + \hat{z}; \tag{2.5.48}$$

hence, $y_2$ and $y_{N-1}$ are the solution to the following system:

$$y_2 = y_2 z_2' + y_{N-1} z_2'' + z_2, \quad y_{N-1} = y_2 z_{N-1}' + y_{N-1} z_{N-1}'' + z_{N-1}.$$

From (2.5.48) we obtain $\hat{y}$ and, then, $\hat{u} = \hat{D}^{-1}\hat{y}$.

Thus, the problem consists in the efficient numerical solution of the linear systems (2.5.45)–(2.5.47) with the same symmetric Toeplitz positive definite matrix.

**4.** Now we describe our algorithm for the solution of linear systems $Ax = b$ with a symmetric Toeplitz matrix $A$. Let $A_k$, $k = 0, 1, \ldots, N-1$, be truncated matrices of order $k+1$ such that

$$A_{N-1} \equiv A, \quad x = (x_0, x_1, \ldots, x_{N-1})', \quad b = (b_0, b_1, \ldots, b_{N-1})'.$$

If $\alpha_k = (\alpha_{0,k}, \ldots, \alpha_{k,k})'$ and $\beta_k = (\beta_{0,k}, \ldots, \beta_{k,k})'$ are the first and last columns of the matrix $A_k^{-1}$, respectively, then it is clear that for $k = 0$ these columns coincide and contain the one element $a_0^{-1}$. It is easy to see that $\alpha_{i,k} = \beta_{k-i+1,k}$, $i = \overline{0,k}$, $k = \overline{0, N-1}$. It is assumed that $\alpha_{k-1}$ is known. We search for $\alpha_k$ in the following form:

$$\begin{aligned}\alpha_k \;\; = \;\; & u_k(\alpha_{0,k}, \alpha_{1,k-1}, \ldots, \alpha_{k-1,k-1})' + \\ & + v_k(0, \alpha_{k-1,k-1}, \alpha_{k-2,k-1}, \ldots, \alpha_{0,k-1})',\end{aligned}$$

where $u_k$ and $v_k$ are some scalars.

Since $\alpha_k$ is the first column of $A_k^{-1}$, then $A_k\alpha_k = (1, 0, \ldots, 0)'$. Denoting $F_k = \sum\limits_{i=0}^{k} a_i\alpha_{k-i,k-1}$, from here we obtain

$$u_k + F_k v_k = 1, \quad F_k u_k + v_k = 0,$$

i.e., $u_k = (1 - F_k^2)^{-1}$ and $v_k = -F_k u_k$.

Thus, for the matrices $A_0^{-1}, \ldots, A_{N-1}^{-1}$ it is possible to determine the first and last columns without computing the elements of other columns.

In the sequel, using the last column of $A_k^{-1}$, we can find recurrently the solution $y_k$ to the truncated system $A_k y_k = d_k$, where $d_k = (b_0, b_1, \ldots, b_k)'$, $k = \overline{0, N-1}$. Then we obtain $x = y_{N-1}$. See [59] for more details.

Note that we followed the algorithm of [59] with the only distinction: in general, the last column is obtained in the same amount of operations as it takes for the first column ($\sim N^2$), whereas in our case (due to the symmetry of the original matrix) the last column is obtained automatically, i.e., without additional operations.

As in the case of the algorithm given in [59], our algorithm requires $\sim N^2$ operations. The time spent for solution of a linear system with a Toeplitz matrix of order $N = 1000$ on a BESM-6 computer is equal 1 minute and 57 seconds without taking its symmetry into account and is only half as large when the symmetry is taken into account.

Note that the solutions to equations (2.5.46) and (2.5.47) are related by the equality $z'_i = z''_{N-i+1}$, $i = \overline{1, N}$. Moreover, our analysis of the above algorithm shows that for the solution of equation (2.5.47) we need to perform only the last two steps of the recurrent procedure.

Thus, the essence of our approach consists in the following. First we solve equation (2.5.45). Then, using the computed last columns of the matrices $A^{-1}_{N-2}$ and $A^{-1}_{N-1}$, we find the solutions to equation (2.5.47) and, consequently, to equation (2.5.46). The number of operations is only $\sim N$.

**5.** The above-described numerical algorithm was used to solve some problems in the theory of identification of models. Its rather high efficiency (in speed of operation and in the amount of storage required) was shown.

*General characteristics of the subroutine for numerical solution of the problem*:

1) the subroutine implements the above algorithm for construction of discrete regularized solutions $\hat{u}^{\alpha_i}$, $i = \overline{1, M}$;

2) the values of the regularization parameter $\alpha$ are taken on the grid $\alpha_{i+1} = \tau\alpha_i$, where $\tau$ is the grid step, $\alpha_0$ is an initial approximation in $\alpha$, $i = 0, 1, \dots$; $M$ is the number of grid steps in $\alpha$;

3) errors are estimated in the root-mean-square grid norms for each approximate solution $\hat{u}^{\alpha_i}$; namely, the accuracy characteristic $N1(\alpha) \approx \|u^\alpha - u_{\text{ex}}\|_{L_2}$, where $u_{\text{ex}}$ is the known model solution, is evaluated;

4) a quasioptimal approximation $\hat{u}^{\alpha_{\text{qopt}}}$ is chosen among the regularized solutions $\{\hat{u}^{\alpha_i}\}$ by searching for the value $\alpha = \alpha_{\text{qopt}}$ that minimizes the deviation $N2(\alpha_i) \approx \left\|u^{\alpha_i} - u^{\alpha_i - 1}\right\|_{L_2}$ in $\alpha_i$;

5) input data are perturbed to study the stability of approximate solutions.

The following two model problems were considered. The model problem I is specified by the following data:

$$\begin{aligned} K(t) &= \exp(-|t|), \quad |t| < +\infty, \\ f(t) &= (2/5)\exp(-t)(\cos t + 2\sin t), \quad t \geq 0. \end{aligned}$$

The exact solution: $u_{\mathrm{ex}}(t) = \exp(-t)\cos t$, $t \geq 0$.

The model problem II is specified by the following data:

$$\begin{aligned} K(t) &= (2/3)\exp(-|t|/2) + 1/6\exp(-|t|), \quad |t| < +\infty, \\ f(t) &= (4/3)\exp(-|t|/2) + t/6\exp(-|t|) - 29/36\exp(-|t|), \quad t \geq 0. \end{aligned}$$

The exact solution: $u_{\mathrm{ex}}(t) = \exp(-t)$, $t \geq 0$.

Our computations were performed with $N = 200$ and $T = 10$.

When the input data were exact (there was an approximation error only) and the regularization was absent ($\alpha = 0$), the results obtained were satisfactory. However, small perturbations of data (for example, $\Delta K \approx \dfrac{1}{400}\sin(200t)$, $\Delta f \approx \dfrac{1}{200}\sin(200t)$) cause the strong numerical instability of the approximations obtained, whereas this phenomenon is not observed when the appropriate regularized solutions are chosen (see Tables 2.5.1 and 2.5.2) and the quite qualitative results ($N1(\alpha) \approx 0.03$, $\hat{u}_{\alpha_7} = \hat{u}_{\alpha_{\mathrm{opt}}}$) are obtained.

It takes 3.21 seconds to compute a one regularized solution for $N = 200$ on a BESM-6 computer. Note that for large ($\alpha = \alpha_1$) and for small ($\alpha = \alpha_9$) values of regularization parameter (see Tables 2.5.1 and 2.5.2) the solutions obtained are not adequate in accuracy.

**Table 2.5.1.** Problem I (perturbed input data)

| t | 0.0 | 0.05 | 0.5 | | |
|---|---|---|---|---|---|
| $u_{\mathrm{ex}}(t)$ | 1.0 | 0.5722 | $-0.0128$ | | |
| $\hat{u}_0(t)$ | 10.8049 | $-8.2529$ | $-5.7702$ | N1(0)=20.185 | |
| $\hat{u}_{\alpha_1}(t)$ | 0.4287 | 0.3885 | $-0.0264$ | $\mathrm{N1}(\alpha_1) = 0.302$ | $\mathrm{N2}(\alpha_1) = 20.1944$ |
| $\hat{u}_{\alpha_7}(t)$ | 0.9007 | 0.5873 | $-0.0070$ | $\mathrm{N1}(\alpha_7) = 0.0328$ | $\mathrm{N2}(\alpha_7) = 0.0185$ |
| $\hat{u}_{\alpha_9}(t)$ | 0.9017 | 0.6212 | 0.0404 | $\mathrm{N1}(\alpha_9) = 0.1572$ | $\mathrm{N2}(\alpha_9) = 0.1037$ |

**Table 2.5.2.** Problem II (perturbed input data)

| t | 0.0 | 0.05 | 0.5 | | |
|---|---|---|---|---|---|
| $u_{\mathrm{ex}}(t)$ | 1.0 | 0.6362 | 0.0189 | | |
| $\hat{u}_0(t)$ | 5.5567 | $-4.9126$ | 4.9179 | N1(0)=14.224 | |
| $\hat{u}_{\alpha_1}(t)$ | 0.5122 | 0.4864 | 0.0180 | $\mathrm{N1}(\alpha_1) = 0.2471$ | $\mathrm{N2}(\alpha_1) = 14.2298$ |
| $\hat{u}_{\alpha_7}(t)$ | 0.8872 | 0.6581 | 0.0247 | $\mathrm{N1}(\alpha_7) = 0.0371$ | $\mathrm{N2}(\alpha_7) = 0.0192$ |
| $\hat{u}_{\alpha_9}(t)$ | 0.8810 | 0.6869 | 0.0826 | $\mathrm{N1}(\alpha_9) = 0.1597$ | $\mathrm{N2}(\alpha_9) = 0.1053$ |

## 2.6 Stable design of pulse transition functions

**1.** Many problems in synthesis of control systems that are optimal in the mean-square sense and are subject to random actions can be reduced to solving the systems of Wiener–Hopf-type integral equations [288]. The following problems belong to this class: multidimensional filtration; compensation of information in the case of excess of data transmission channels; identification of objects with technological feedback; synthesis of systems (optimal in the mean-square sense) with different degrees of structural restrictions imposed on their information and control components [310]; synthesis of optimal systems of indirect distributing control; problems for random fields with discrete and distributed control, etc.

The absence of satisfactory methods, algorithms, and software implementations for solving the systems of Wiener–Hopf-type integral equations makes serious difficulties for the solution of the above problems. A critical analysis [219] of the literature devoted to the methods for solving the Wiener–Hopf integral equations shows that the most suitable method is the method of regularization, since it imposes the least restrictions on input data and gives stable numerical solutions.

The Wiener–Hopf type integral equations belong to the class of convolution-type integral equations of the first kind with consideration of the physical realizability condition. Numerical methods for solving convolution-type integral equations are discussed in [270], [248] and in a number of other works. Nevertheless, the questions devoted to the stable solution of Wiener–Hopf integral equations (especially for multidimensional systems of automatic control and regulation) are poorly studied.

An algorithm for numerical synthesis of a class of multidimensional automatic control systems is proposed in [286], where the exact solution is found in the frequency range of an optimal filter being synthesized under the condition that this solution is physically realizable on a set of exact original data. However, the question of influence of errors in input data was not studied.

In the sequel, we propose a numerical algorithm for synthesis of such systems based on the method of regularization and the theory of designing $\Omega$-optimal automatic control systems. In particular, we discuss an algorithm for solving multidimensional Wiener–Hopf-type integral equations (this algorithm ensures the numerical stability of the solutions obtained).

Let us consider a linear stationary system characterized in the general case by $m$ inputs and by one output. Let $z_1, z_2, \dots, z_m$ and $y$ be given stationary and stationary-connected random processes at the input and at the output of the system, respectively. By the symbol "$*$" we denote the convolution operation. The discrepancy functional of the system characterizing its quality (error variance in the response of the system) is defined as follows:

$$e^2(\varpi) = e^2(\varpi; y) \equiv M \left| \sum_{i=1}^{n} z_i * \varpi_i - y \right|^2 .$$

Here $M$ is the symbol of mathematical expectation and $\varpi = (\varpi_1, \dots, \varpi_n)$ is a vector pulse transition function (PTF) of the system.

The main problem in the design of optimal systems of automatic control (SAC) consists in minimizing the discrepancy functional (the error of a system), i.e., in searching for the function

$$\varpi_{\text{opt}} = \arg \min_{\varpi \in W} e^2(\varpi). \tag{2.6.49}$$

Here $W$ is a functional class of allowable PTFs given *a priori*. It is often required that a PTF should satisfy the physical realizability condition for a system; then,

$$W = \{\varpi : \varpi(t) = 0,\ t < 0\} .$$

Let us introduce the scalar product

$$(u, v) = \int_0^\infty u(t)v(t)\, dt, \quad u, v \in L_2(0, \infty).$$

Putting

$$\begin{aligned} k_{i,j}(\tau) &\equiv M\left[z_i(t) z_j(t+\tau)\right], \quad i, j = 1, 2, \dots, m, \\ f_i(\tau) &\equiv M\left[y(t) z_i(t+\tau)\right], \quad i = 1, 2, \dots, m, \\ e_0^2 &\equiv M\left[y^2(t)\right], \end{aligned}$$

we can specify the operators $K_{ij} : K_{ij}\varpi_j = k_{ij} * \varpi_j$. It is easy to see that $K_{ij}^* = K_{ji}$, where the symbol "$*$" indicates the conjugation operation.

Then, the discrepancy functional can be represented in the form

$$e^2(\varpi) = [K\varpi, \varpi] - 2[f, \varpi] + e_0^2,$$

where

$$\begin{aligned} [u, v] &\equiv (u_1, v_1) + \cdots + (u_m, v_m) \ \ \forall u, v \in W, \ \ f = (f_1, \dots, f_m)', \\ K\varpi &\equiv \begin{pmatrix} k_{11} * & k_{12} * & . & . & . & k_{1m} * \\ . & . & . & . & . & . \\ k_{m1} * & k_{m2} * & . & . & . & k_{mm} * \end{pmatrix} \begin{pmatrix} \varpi_1 \\ \varpi_2 \\ . \\ . \\ . \\ \varpi_m \end{pmatrix}. \end{aligned}$$

Now we consider the operator equation

$$K\varpi = f \tag{2.6.50}$$

and denote the set of its solutions by $W_0$; the solution set for the problem (2.6.49) is denoted by $W_{\text{opt}}$.

**Theorem 2.6.44** *The sets $W_0$ and $W_{\text{opt}}$ coincide, i.e., the problems* (2.6.49) *and* (2.6.50) *are equivalent* [25].

Note, however, that the sets $W_0$ and $W_{\text{opt}}$ may be empty (simultaneously). It is a very important fact used in the further discussion to construct a stable method for solving the above problem.

Let $m = 1$. Then,

$$K\varpi \equiv \int_0^\infty k(t-\tau)\varpi(\tau)\,d\tau.$$

Let $\varpi = \varpi_\varepsilon(t) = 0$, $t \geq \varepsilon > 0$, and $\varpi_\varepsilon(t) = a > 0$, $0 \leq t \leq \varepsilon$. If the function $k(t)$ is bounded, then

$$|K\varpi_\varepsilon| \leq \max|k(t)|\,a\varepsilon \to 0, \quad \varepsilon \to 0,$$

where $\max|\varpi_\varepsilon| = a$.

The above means that the right-hand side of (2.6.50) can be measured with arbitrary accuracy in the uniform sense, but in an approximate solution we can obtain an arbitrary large error in $L_2$, i.e., the problem of designing the optimal SACs is unstable. It is required to develop a special approach to the solution of such problems. Such an approach is considered in the further discussion.

**2.** Let us come back to the problem of designing the optimal SACs formulated in Subsection 1. One of the peculiarities of the problem under consideration is its possible unsolvability. Indeed, if the correlation functions $k_{ij}$ are differentiable and the functions $f_i$ are only continuous, then the equation $K\varpi = f$ is obviously unsolvable. Thus, the set $W_{\text{opt}}$ may be empty even if

$$e^2_{\text{opt}} \equiv \inf_{\varpi} e^2(\varpi) = 0,$$

i.e., the system is invariant.

To avoid this unpleasant emptiness, we introduce the following definition [25].

**Definition.** *The function $\varpi^\varepsilon_{\text{opt}}$ is called the $\varepsilon$-optimal PTF for some $\varepsilon > 0$ if $e(\varpi^\varepsilon_{\text{opt}}) \leq e_{\text{opt}} + \varepsilon$.*

Obviously, the set $W_{\text{opt}}^{\varepsilon}$ of all $\varepsilon$-optimal PTFs corresponding to a given $\varepsilon$ is always nonempty and contains $W_{\text{opt}}$ if it is nonempty. Therefore, it is natural to replace the requirement of finding an optimal PTF by the weaker (and always feasible) requirement of constructing $\varepsilon$-optimal PTFs. Note that this approach is typical in practice, but the formalization of these operations allows one (at least, qualitatively) to estimate their consequences.

It is natural to consider the above system as adequate if the condition $e_{\text{opt}} \ll e_0$ is fulfilled. Obviously, any inadequate system is mathematically inconsistent, i.e., the mathematical model under consideration cannot correspond to a real physical system. The level of adequacy can be estimated by $\mu e_0$, where $0 < \mu \ll 1$ is a numerical parameter characterizing a measure of adequacy. When $\mu \approx 0$, a system is adequate, when $\mu \gg 0$ it is inadequate.

A measure of adequacy can be given by various ways, and we will not pay attention here to details. It is important to note that at a level appropriate for practical requirements, such a choice of measure is always possible. In the sequel, the systems under study are assumed to be adequate.

Thus, the first stage for the solution of the problem under consideration is the construction of such a system for which $e(\varpi) \approx e_\mu = \mu e_0$. If $e_\mu > e_{\text{opt}}$ (this is supposed in what follows), then there always exists at least one function $\varpi \in W$ such that $e(\varpi) \le e_\mu$.

*A system with the quality characteristic $e_\mu$ is said to be $\mu$-adequate (or $\mu$-feasible). How to choose the most reasonable system among all $\mu$-adequate ($\mu$ is fixed) systems?*

We base on some principles used to select solutions. On some subset $D \subset W$ (for simplicity, $D$ is assumed to be linear) we introduce a positive definite quadratic functional $\Omega^2(\varpi) \ge \gamma^2 |\varpi|^2$, $\varpi \in W$, where $|\varpi|^2 \equiv [\varpi, \varpi]$; this functional has the meaning of cost for the realization of a PTF (or any other similar meaning).

**Definition.** *A PTF $\varpi_\mu$ is said to be $\Omega$-optimal if it is $\mu$-feasible and realizes a minimum of the functional $\Omega(\varpi - \varpi^*)$ among all $\mu$-feasible PTFs, i.e.,*

$$\Omega(\varpi_\mu - \varpi^*) = \min_{\varpi \in D : e(\varpi) \le e_\mu} \Omega(\varpi - \varpi^*), \tag{2.6.51}$$

*where $\varpi^*$ is a trial PTF corresponding to the available a priori knowledge about the sought-for PTF.*

Thus, the problem of designing the $\Omega$-optimal SACs is completely formulated.

What is an advantage of this formulation compared to the classical one? In our opinion, the advantage consists in its flexibility, in better adequacy to practical requirements, and in taking account of a number of additional factors, which are inevitable when such systems are studied: for example, possible errors in input data (correlation and mutually correlation functions) whose presence makes the classical formulation for the problem of constructing the optimal SACs hardly reasonable (because of its property of being ill-posed); more often, the classical formulation is used only due to tradition. In the new formulation, the problem of designing $\Omega$-optimal systems is stable in small variations of input data $d = \{k_{ij}, f\}$ for $\mu$ fixed.

Note that if $\varpi_\mu$ is an $\Omega$-optimal PTF, then $e(\varpi_\mu) = e_\mu$, i.e., in the problem (2.6.51) the functional $\Omega\,(\varpi - \varpi^*)$ can be minimized on the set $\varpi \in D$, $e(\varpi) = e_\mu$. In other words, an $\Omega$-optimal PTF satisfies the *principle of guaranteed quality* of a SAC.

The above approach to the design of $\Omega$-optimal SACs completely corresponds to one of the basic methods for solving ill-posed problems: to the residual (discrepancy) method [172].

The problem (2.6.51) is convenient from the standpoint of interpretation of $\Omega$-optimal solutions to SACs; however, it is not quite convenient for algorithmization. In this connection, it is reasonable to formulate the following definition.

**Definition**. *A vector function $\varpi^\alpha \in D$ is called the regularized pulse transition function* (*RPTF*) *if*

$$\varpi^\alpha = \arg\min_{\varpi \in D} \left\{ e^2(\varpi) + \alpha\Omega^2(\varpi - \varpi^*) \right\}, \tag{2.6.52}$$

*where $\alpha > 0$ is the regularization parameter.*

We will establish some properties of RPTFs and their relation to $\Omega$-optimal PTFs and to $\varepsilon$-optimal PTFs.

**Theorem 2.6.45** *The following relation takes place:*

$$\lim_{\alpha \to 0+} e\,(\varpi^\alpha) = e_{\text{opt}},$$

*i.e., for a sufficiently small $\alpha > 0$, a RPTF is $\varepsilon$-optimal with any $\varepsilon > 0$. If, moreover, in $D$ there exists at least a one optimal PTF, then*

$$\Omega\,(\varpi^\alpha - \hat{\varpi}_{\text{opt}}) \to 0, \quad \alpha \to 0+,$$

*where $\hat{\varpi}_{\text{opt}}$ is the only $\Omega$-optimal PTF satisfying the optimality condition: $e\,(\hat{\varpi}_{\text{opt}}) = e_{\text{opt}}$. The existence of $\hat{\varpi}_{\text{opt}}$ is guaranteed.*

Note that the operator $K$ introduced previously in (2.6.50) is nonnegative, i.e., $[K\varpi, \varpi] \geq 0$. If $K$ is positive, i.e., $K\varpi = 0 \iff \varpi = 0$, then equation (2.6.50) is uniquely solvable, i.e., the set $W_{\text{opt}} = \{\varpi_{\text{opt}}\}$ is one-element. Then $\hat{\varpi}_{\text{opt}} = \varpi_{\text{opt}}$ if $\varpi_{\text{opt}} \in D$, i.e., the functional $\Omega$ is defined on $\varpi_{\text{opt}}$.

Theorem 2.6.45 shows that RPTFs approximate optimal PTFs in a certain sense and can be used as a tool to find optimal PTFs approximately.

An important property of RPTFs is expressed by the following theorem.

**Theorem 2.6.46** *Regularized PTFs are uniquely defined for any input data and for any* $\alpha > 0$.

Let us put $\rho(\alpha) \equiv e(\varpi^{\alpha})$, $\alpha > 0$. It is natural to take the function $\rho(\alpha)$ as a *discrepancy function of a regularized system* (error variance for a response of this system). It turns out that $\rho(\alpha)$ is a strictly increasing function of $\alpha$ and

$$\lim_{\alpha \to 0+} \rho(\alpha) = e_{\text{opt}}, \quad \lim_{\alpha \to +\infty} \rho(\alpha) = e_0,$$

so that $e_{\text{opt}} < \rho(\alpha) < e_0$ for $0 < \alpha < +\infty$. This implies the following

**Theorem 2.6.47** *Let* $e_\mu$, $e_{\text{opt}} < e_\mu < e_0$, *be a given level. Then, there exists a unique value* $\alpha_\mu$ *of regularization parameter such that*

$$\rho(\alpha_\mu) = e_\mu \tag{2.6.53}$$

*and* $\varpi^{\alpha_\mu} = \varpi_\mu$, *i.e., the corresponding RPTF is an* $\Omega$*-optimal PTF.*

Thus, the design of an $\Omega$-optimal system is divided into the following two stages.

I. The construction of a regularized PTF of a system for any $\alpha > 0$.

II. The choice of the regularization parameter $\alpha$ from the condition (2.6.53), i.e., from the principle of guaranteed quality of the system.

It follows from here that the concept of $\Omega$-optimal SACs is substantially wider than the classical concept of optimal SACs and imposes substantially lesser restrictions on input data. Thus, $\Omega$-optimal solutions can be found even in the case when optimal solutions do not exist in the classical sense.

The above-considered questions are the application of the general regularization theory discussed in Chapter 1 and in [188], [189] to the design of optimal SACs.

**3.** The solution of the scalar equation (2.6.53) is important for constructing the $\Omega$-optimal systems. In what follows we consider a number of methods for the choice of appropriate values of the regularization parameter.

*The method of direct choice* is very simple in its idea. Let $\alpha_0 > \alpha_1 > \ldots > \alpha_M > 0$ be a grid in $\alpha$ (here $\alpha_0$ is sufficiently large and $\alpha_M$ is sufficiently small). Then, we calculate (perhaps, approximately) the regularized solutions $\varpi^i = \varpi^{\alpha_i}$, $i = 0, 1, \ldots, M$ (several methods for their calculation are considered below) and the values $\rho_i = \rho(\alpha_i)$. Next we find $i_0$ such that $\rho_{i_0} \le e_\mu \le \rho_{i_0+1}$.

Therefore, $\alpha_{i_0} \leq \alpha_\mu \leq \alpha_{i_0+1}$. If the grid in $\alpha$ is sufficiently dense, then it is possible to put $\varpi_\mu \approx \varpi^{\alpha_{i_0}}$. Otherwise, the process of division of the segment $[\alpha_{i_0}, \alpha_{i_0+1}]$ is applied (for example, bisection). If the sought-for $i_0$ is not found (on the original grid), then we take new values of $\alpha_i$ on the basis of the results obtained and the fact that $\rho(\alpha)$ monotonically increases. Usually, this process rapidly terminates, since the behavior of $\rho(\alpha)$ is very simple. It is possible to imagine a semi-automatic system realizing the selection of values of the regularization parameter and performing all necessary routine work: construction of $\varpi^i$ and computation of $\rho_i$. If this system is supplied by a display, the necessary trial value of $\alpha$ can easily be found even by an unskilled user.

*The method of model problems* consists in the formulation of a problem similar to the problem under consideration, but with a known exact solution. Errors in input data are modeled on a computer in such a way that these errors are similar to those in the original problem in the sense of their nature and effect. Since the model solution is known, the best values $\alpha_H$ of regularization parameter can be determined from the condition of the best approximation of the solution to the model problems by the regularized solutions. The original problem is solved only with $\alpha = \alpha_H$. The success of this approach depends on the correct modeling of the original problem; in its turn, this depends on some intuitive (*a priori*) knowledge about the problem under study (we consider the solution of a problem as a process and not as a single action).

*The method for the choice of quasioptimal values* $\alpha_{\text{qopt}}$ [265] is also simple in its computer implementation. This method is based on some intuitive knowledge about the existence of a "sheaf" of regularized solutions, i.e., about the points of their condensation:

$$\alpha_{\text{qopt}} = \arg\inf_\alpha \left\| \alpha \frac{d\varpi^\alpha}{d\alpha} \right\|.$$

If a geometric grid in $\alpha$: $\alpha_0, \alpha_1 = \tau\alpha_0, \ldots, \alpha_{i+1=}\tau\alpha_i, \ldots, \alpha_M$ is given (here $\tau \leq 1$ and $\alpha_0$ is sufficiently large), then

$$\alpha \frac{d\varpi^\alpha}{d\alpha}\bigg|_{\alpha=\alpha_i} \approx \alpha_i \frac{\varpi^i - \varpi^{i-1}}{\alpha_i - \alpha_{i-1}} = \tau \frac{\varpi^i - \varpi^{i-1}}{\tau - 1};$$

hence,

$$\alpha_{\text{qopt}} \approx \arg\min_i \left\| \varpi^i - \varpi^{i-1} \right\|.$$

In order to refine $\alpha_{\text{qopt}}$, several grids in $\alpha$ are used. The method was tested on many model and practical problems; however, the method is not substantiated. It was noted that this method does not always provide for necessary results, especially when the errors in input data are large. It is reasonable to combine this method with other methods, for example, with the method of model solutions or the residual (discrepancy) method.

*The modified method for the choice of quasioptimal values of the regularization parameter* (also not substantiated) is based on certain intuitive knowledge. Let $f^\alpha$ be an output of a regularized system, i.e., $f^\alpha = K\varpi^\alpha$. Then, the method consists in the direct choice of the values

$$\hat{\alpha} = \arg\inf_{\alpha} \left\| \alpha \frac{df^\alpha}{d\alpha} \right\|_C$$

or in the choice of these values on a geometric grid:

$$\hat{\alpha} \approx \arg\min_{i} \left\| f^i - f^{i-1} \right\|_C .$$

Usually, $f^\alpha \in C$, whereas a given right-hand side $f$ may belong to $L_2$ and may contain a high-frequency component.

The modified method chooses the parameter corresponding to the "sheaf" of regularized (i.e., sufficiently smooth) right-hand sides in $C$.

*The automated choice of the regularization parameter* according to the principle of guaranteed quality of a SAC. The process of solution of the regularized problem (2.6.52) is equivalent to the solution of Euler's equation

$$\alpha L(\varpi - \varpi^*) + K\varpi = f, \tag{2.6.54}$$

where $L$ is a linear positive definite operator uniquely defined by the identity $\Omega^2(\varpi) \equiv [L\varpi, \varpi]$. In (2.6.54) it is convenient to use the Lagrangian parameter $\lambda = 1/\alpha$, i.e., to consider the equation

$$L(\varpi - \varpi^*) + \lambda K\varpi = \lambda f \tag{2.6.55}$$

with the solution $\varpi(\lambda) \equiv \varpi^\alpha$, $\alpha = 1/\lambda$. Let us put $\hat{\rho}(\lambda) \equiv e(\varpi(\lambda))$ and consider the scalar equation

$$\hat{\rho}(\lambda) = e_\mu. \tag{2.6.56}$$

If $\lambda_\mu$ is its solution, then $\alpha_\mu = 1/\lambda_\mu$ is a solution to equation (2.6.53). The transition to the parameter $\lambda$ is connected with the fact that the function $\hat{\rho}(\lambda)$ *strictly decreases* and is *convex downwards*, i.e., equation (2.6.56) can always be solved by Newton's method with the initial approximation $\lambda_0 = 0$:

$$\lambda_{k+1} = \lambda_k - (\hat{\rho}(\lambda_k) - e_\mu)/\hat{\rho}'(\lambda_k), \quad k = 0, 1, \ldots;$$

it is well known that the convergence rate of Newton's method is quadratic. Since

$$\hat{\rho}^2(\lambda) = [K\varpi(\lambda), \varpi(\lambda)] - 2[f, \varpi(\lambda)] + e_0^2,$$

we have

$$\hat{\rho}'(\lambda)\hat{\rho}(\lambda) = \left[K\varpi(\lambda) - f, \varpi'(\lambda)\right] = -\left[L(\varpi(\lambda) - \varpi^*), \varpi'(\lambda)\right]/\lambda,$$

where, by virtue of (2.6.55), the function $\varpi'(\lambda)$ is a solution to the equation

$$(L + \lambda K)\varpi'(\lambda) = f - K\varpi(\lambda).$$

Note that $\varpi(0) = 0$ (with $\varpi^* = 0$), whereas $\varpi'(0)$ can be determined from the last equation, i.e., Newton's process is completely defined. The difficulty caused by the choice of an initial approximation $\lambda_0$ (this is typical for Newton's method) is thus absent.

The sequence of approximations $\lambda_k$ strictly increases, rapidly converges (5–7 iterations), and does not require additional operations. The method is theoretically substantiated, is practically tested on a great number of problems, and demonstrates good performance. The method is developed in [172]–[194].

*The choice of the regularization parameter according to a control realization of the system output* consists in the choice of yet another output realization $\tilde{y}$ measured with the same accuracy as $y$. Then, the suitable values of the parameter can be found from the condition

$$\bar{\alpha} = \arg\inf_{\alpha} e(\varpi^{\alpha}; \tilde{y}),$$

where $\varpi^{\alpha}$ is defined in (2.6.52). Note that if we know an approximate level of errors in $y$ and its nature, we can simulate $\tilde{y}$ on a computer. The method is applicable if the errors in other data are negligibly small in comparison with errors in $y$.

*The parameter choice by a method of quasisolutions* [115] consists in the choice of a control value $R$ bounding the values of the stabilizing functional $\gamma(\alpha) \equiv \Omega(\varpi^{\alpha} - \varpi^*)$ on the realized solutions. The regularization parameter $\alpha$ is chosen as a solution to the scalar equation $\gamma(\alpha) = R$; this equation can be solved by a trial-and-error method or by Newton's process with an appropriate choice of an initial approximation. The following property of the function $\gamma(\alpha)$ is used: *this function strictly monotonically decreases and is convex downwards.* The process of computations is similar to the process described above, but with the following distinction. Generally speaking, it is impossible to put $\alpha_0 = 0$, since the solution to the regularized problem may not be defined or may be so large that its software implementation is not possible. Therefore, it is required to find $\alpha_0$, $\gamma(\alpha_0) \geq R$, experimentally and, then, to construct Newton's approximations [194]:

$$\alpha_{k+1} = \alpha_k - (\gamma(\alpha_k) - R)/\gamma'(\alpha_k), \quad k = 0, 1, \ldots .$$

Since

$$\gamma'(\alpha)\gamma(\alpha) = \left[L\left(\varpi^{\alpha} - \varpi^{*}\right), \frac{d\varpi^{\alpha}}{d\alpha}\right],$$

it is sufficient to determine $\dfrac{d\varpi^{\alpha}}{d\alpha}$ from the equation

$$(\alpha L + K)\,\frac{d\varpi^{\alpha}}{d\alpha} = -\alpha L\left(\varpi^{\alpha} - \varpi^{*}\right).$$

The meaning of the above way for the choice of the regularization parameter consists in the choice of a regularized solution of a certain "cost" $R$. The method is equivalent to the solution of the variational problem

$$\min_{\varpi : \Omega(\varpi - \varpi^{*}) \leq R} e(\varpi);$$

this problem is easily interpreted from the physical standpoint.

*The method of a smoothing functional* [152], [172]. It is well known that the functional $e_{\alpha}^{2}(\varpi) \equiv e^{2}(\varpi) + \alpha\Omega^{2}(\varpi)$, $\varpi \in D$, is said to be smoothing [265]. Let us put $\varphi(\alpha) \equiv e_{\alpha}^{2}(\varpi^{\alpha})$. The criterion for the parameter choice consists in the solution of the scalar equation $\varphi(\alpha) \equiv e_{\mu}^{2}$ that can be solved by the above techniques (see also [194]). The functional $e_{\alpha}^{2}(\varpi)$ characterizes the discrepancy at the output of a regularized system.

*The adaptive regularization.* If $\varpi^{*}$ coincides with the unique optimal PTF $\varpi_{\mathrm{opt}}$, then $\varpi^{\alpha} \equiv \varpi^{*}$, $\alpha > 0$, i.e., the regularized PTFs are stationary. From here it follows the recommendation: to choose a trial solution $\varpi^{*} \approx \varpi_{\mathrm{opt}}$. In this case, the range of suitable values of the regularization parameter is substantially expanded, i.e., the errors in the determination of the regularization parameter do not lead to significant errors in a solution. It is possible to use the following approach. Let $\varpi^{*} = 0$, i.e., there is no information on the exact solution. The family $\varpi^{\alpha}$ is called the *primary regularized family of PTFs.* Let us denote this family by $\varpi_{1}^{\alpha}$. Now we replace $\varpi^{*}$ by the function $\varpi_{1}^{\alpha}$ and solve the problem of minimizing the smoothing functional once again. The new regularized family we call the secondary one and denote it by $\varpi_{2}^{\alpha}$, etc. For the secondary regularized family $\varpi_{2}^{\alpha}$ the following relation holds:

$$\varpi_{2}^{\alpha} = \varpi_{1}^{\alpha} - \alpha\,\frac{d\varpi_{1}^{\alpha}}{d\alpha}.$$

Similar representations can be written inductively for the $p$th regularized solution:

$$\varpi_{p}^{\alpha} = \varpi_{p-1}^{\alpha} - \alpha\,\frac{d\varpi_{p-1}^{\alpha}}{d\alpha}.$$

It turns out that each $p$th regularized family may be considered as an approximate solution to the original problem with an appropriate choice of the regularization parameter. The above regularization process is said to be adaptive. Its essence consists in the following: the range of suitable values of the regularization parameter for the $p$th regularized solution is extended due to the reasons discussed above.

The secondary regularized family of solutions can be obtained if we take into account that $\dfrac{d\varpi_1^\alpha}{d\alpha}$ satisfies the equation

$$(\alpha L + K)\,\frac{d\varpi_1^\alpha}{d\alpha} = -L\varpi_1^\alpha.$$

Similarly we sequentially compute $\varpi_p^\alpha$ for any $p \geq 2$. If a problem is solved on a geometric grid, then it is possible to use only the family $\varpi_1^\alpha$ and, then, to put

$$\varpi_2^j \simeq \varpi_1^j - \tau\,\frac{\varpi_1^j - \varpi_1^{j-1}}{\tau - 1} = \frac{\tau\varpi_1^{j-1} - \varpi_1^j}{\tau - 1},$$

where $\varpi_p^j \equiv \varpi_p^{\alpha_j}$ and the grid step $\tau < 1$.

Similar representations can be written for $\varpi_p^\alpha$ [172].

Let us consider the relation between the secondary regularized family of solutions and the choice of quasioptimal values of the regularization parameter. It is easy to see that the norm of the second term in the representation of the secondary family is minimized. It is not difficult to transfer this principle on families of higher order.

*The iterative regularization.* Let a value of the regularization parameter $\alpha$ be fixed in the smoothing functional $e_\alpha^2\,(\varpi)$. For convenience we denote $e_\alpha^2\,(\varpi;\varpi^*) \equiv e_\alpha^2\,(\varpi)$; thus, we explicitly indicate the dependence of the stabilizing functional on the trial function $\varpi^*$. Let $\varpi^* = \varpi_0$ be an initial approximation to the sought-for PTF (for example, $\varpi^* = 0$). Then we solve the problem of minimizing the functional

$$e_\alpha^2\,[\varpi;\varpi_0]\,, \quad \varpi \in D.$$

The solution to this problem we denote by $\varpi_1$ and substitute $\varpi_1$ for $\varpi_0$. Further we solve a similar problem and obtain $\varpi_2$, etc. As a result, we come to a sequence of the functions $\varpi_0, \varpi_1, \ldots, \varpi_n$ that converges (in the functional $\Omega$) to an optimal PTF $\varpi_{\text{opt}}$ as $n \to \infty$ (with any fixed $\alpha > 0$) if the original data of the problem are exact. In the case of noisy data, the coordination of the necessary iteration number $n_{\text{opt}}$ with levels of errors in input data is required. Such an iterative process of regularization was first considered and studied in [155]; in the sequel, these results were conceptually repeated in some later works.

If the regularization parameter is assumed to be variable, then, obviously, we come to the adaptive regularization. The iterative regularization method can easily be formulated in operator form: the functions $\varpi_k$ satisfy the operator equation

$$(\alpha L + K)\varpi_k = \alpha L \varpi_{k-1} + f, \quad k = 1, 2, \ldots . \tag{2.6.57}$$

*The iterative determination of regularized solutions* [265], [159]. A regularized problem should often be solved repeatedly for various values of the regularization parameter $\alpha$. Let $\alpha_0 > \alpha_1 > \ldots > \alpha_M$ be a grid in $\alpha$ and $\alpha_0$ be sufficiently large. Then, $\varpi^{\alpha_0} \approx \varpi^*$ (this fact follows from the general results of the regularization theory: $\varpi^\alpha \to \varpi^*$, $\alpha \to \infty$ [172]). Therefore, the problem of minimizing the quadratic functional $e^2_{\alpha_0}(\varpi)$, $\varpi \in D$, can be solved by a minimization method with $\varpi^*$ as an initial approximation. Having determined $\varpi^{\alpha_0}$, it is reasonable to choose $\varpi = \varpi^{\alpha_0}$ as an initial approximation for the problem of minimizing $e^2_{\alpha_1}(\varpi)$, $\varpi \in D$. If $\varpi^{\alpha_i}$ is already obtained, then $\varpi^{\alpha_i}$ is chosen as an initial approximation for the minimization of $e^2_{\alpha_{i+1}}(\varpi)$, $\varpi \in D$. As a minimization method, it is possible to take the gradient method, the method of conjugate gradients, etc. For some reasons, we prefer the method of conjugate gradients.

The further savings in storage and computing time are possible for the problem of designing $\Omega$-optimal SACs if we use the above fast algorithms for computing the gradients of finite-dimensional functionals approximating $e^2_\alpha(\varpi)$ (see [25] and the previous section).

**4.** *SACs with noise.* When we have considered the problem of designing the $\Omega$-optimal SACs, we did not explicitly indicate the aspects associated with inaccuracy in input data. In order to make the problem more specific, we discuss this case in this subsection. It is assumed that the input processes $z_i$ and the output process $y$ contain stationary noises $\Delta z_i$ and $n$ in such a way that

$$z_i = \bar{z}_i + \Delta z_i, \quad y = \bar{y} + n,$$

where $M\Delta z_i = Mn = 0$, whereas $\bar{z}_i$, $i = 1, \ldots, n$, and $\bar{y}$ are exact data. Then,

$$e^2(\varpi) = \bar{e}^2(\varpi) + M\left|\Delta z * \varpi - n\right|^2,$$

where $\bar{e}^2(\varpi)$ is the discrepancy functional for an ideal system. Using the well-known inequalities, we obtain

$$e(\varpi) \le \bar{e}(\varpi) + [K_{\Delta z}\varpi, \varpi]^{1/2} + \sigma_n,$$

where $\sigma_n^2 \equiv Mn^2$ and $K_{\Delta z}$ is an operator which is similar to $K$ and is induced by noises at the input.

Let

$$[K_{\Delta z}\varpi, \varpi] \le \sigma^2_{\Delta z}\Omega^2(\varpi).$$

If the processes $\Delta z_i$ do not mutually correlate, i.e., $K_{\Delta z} = \mathrm{diag}\left\{\sigma_1^2, \ldots, \sigma_m^2\right\}$, then

$$[K_{\Delta z}\varpi, \varpi] \le \sigma^2 \sum_{i=1}^{m} \int\limits_0^{\infty} \varpi_i^2\, dt \le \frac{\sigma^2}{\gamma^2}\, \Omega^2(\varpi),$$

i.e., $\sigma_{\Delta z}^2 = \sigma^2/\gamma^2$ and $\sigma^2 = \max \sigma_i^2$. In this case,

$$e(\varpi) \le \bar{e}(\varpi) + \sigma\Omega(\varpi) + \sigma_n, \bar{e}(\varpi) \le e(\varpi) + \sigma\Omega(\varpi) + \sigma_n,$$

where the second inequality is obtained similarly. These inequalities relate the values of discrepancy at the output of the real and "ideal" systems.

Let us assume that the "ideal" system is invariant, i.e., $\inf \overline{e}(\varpi) = 0$ and there exists a unique optimal PTF $\varpi_{\mathrm{opt}}$ such that $\bar{e}(\varpi_{\mathrm{opt}}) = 0$. Then we obtain the following estimate

$$e(\varpi_{\mathrm{opt}}) \le \sigma\Omega(\varpi_{\mathrm{opt}}) + \sigma_n.$$

Now we introduce a class of feasible PTFs of the real system $\varpi \in D$, $e(\varpi) \le \sigma\Omega(\varpi) + \sigma_n$. It follows from the previous discussion that $\varpi_{\mathrm{opt}}$ is a feasible PTF. However, the set of all feasible PTFs is too wide, since its diameter, in general, tends to infinity. How can we reasonably (from the physical and mathematical standpoints) narrow this set? It is natural to use the concept of $\Omega$-optimal PTFs analyzed earlier. Then, we come to the following minimization problem:

$$\min \Omega(\varpi), \quad \varpi \in D : e(\varpi) \le \sigma\Omega(\varpi) + \sigma_n. \tag{2.6.58}$$

Its solution $\tilde{\varpi}_{\mathrm{opt}}$ is also called an $\Omega$-optimal PTF.

Note that here the condition of $\Omega$-optimality was imposed by the course of study. Hence, the "cost" functional $\Omega(\varpi)$ of the system was obtained quite naturally on the basis of analysis of the system's properties. This is an essential methodological advantage of the above method for designing the $\Omega$-optimal SACs; this advantage follows from the method of optimal residual (discrepancy) [172], which is one of the general methods of regularization.

How to implement the solution of the variational problem (2.6.58)? We consider the regularized PTFs $\varpi^{\alpha}$ and the scalar equation, which defines the choice of the regularization parameter $\alpha$:

$$\rho(\alpha) = \sigma\gamma(\alpha) + \sigma_n. \tag{2.6.59}$$

The following theorem is valid.

**Theorem 2.6.48** *For sufficiently small $\sigma$ and $\sigma_n$, equation* (2.6.59) *has a unique solution $\tilde{\alpha}$. The function $\varpi^{\tilde{\alpha}} = \tilde{\varpi}_{\mathrm{opt}}$ is a solution to the problem* (2.6.58) *and $\Omega(\tilde{\varpi}_{\mathrm{opt}} - \varpi_{\mathrm{opt}}) \to 0$ as $\sigma$, $\sigma_n \to 0$.*

Thus, the solution to (2.6.58) can be obtained on the basis of the regularization method, using the choice criterion (2.6.59). The numerical aspects are considered in [172].

Note that for the solution $\tilde{\varpi}_{\text{opt}}$ to the problem (2.6.58) we have the estimate $\Omega\left(\tilde{\varpi}_{\text{opt}}\right) \leq \Omega\left(\varpi_{\text{opt}}\right)$, i.e., the "cost" of the $\Omega$-optimal solution is always less than the "cost" of the optimal solution. Then, for an equation satisfying the criterion (2.6.59) we have

$$e\left(\tilde{\varpi}_{\text{opt}}\right) \leq \sigma\Omega\left(\varpi_{\text{opt}}\right) + \sigma_n,$$

i.e., $\tilde{\varpi}_{\text{opt}}$ has the characteristics similar (and close) to those of $\varpi_{\text{opt}}$. If the estimate $\Omega\left(\varpi_{\text{opt}}\right) \leq R$ is known, then we may put $e_\mu = \sigma R + \sigma_n$, which simplifies the algorithm.

**5.** Let us consider another method for designing the SACs under noise conditions; this method is applicable for noninvariant systems. The inequality

$$\bar{e}\left(\varpi\right) \leq e\left(\varpi\right) + \sigma\Omega(\varpi) + \sigma_n, \quad \varpi \in D,$$

was previously obtained. The right-hand side of this inequality may be interpreted as a measure of discrepancy of some weighted nonideal system, which is fictitious in a certain sense. Then, a smoothing functional can be defined by putting $\tilde{e}\left(\varpi\right) \equiv e\left(\varpi\right) + \sigma\Omega(\varpi)$; hence, the acceptable PTFs $\tilde{\varpi}$ can be chosen from the condition for a minimum of this functional.

Note that the problem of choosing the regularization parameter does not arise for the above approach which is very similar to the method of regularization. The basic idea of this method for designing SACs is the *minimax principle* for the residual (discrepancy) functional developed in [141], [180].

Let $\tilde{\varpi}$ be a solution to the problem under study. Let us consider its properties. For invariant systems we have

$$\begin{aligned} e\left(\tilde{\varpi}\right) &\leq \tilde{e}\left(\varpi_{\text{opt}}\right) = e\left(\varpi_{\text{opt}}\right) + \sigma\Omega\left(\varpi_{\text{opt}}\right) \leq \\ &\leq 2\sigma\Omega\left(\varpi_{\text{opt}}\right) + \sigma_n \leq 2\left(\sigma\Omega\left(\varpi_{\text{opt}}\right) + \sigma_n\right). \end{aligned}$$

From here it follows the relation for "cost":

$$\Omega\left(\tilde{\varpi}\right) \leq 2\Omega\left(\varpi_{\text{opt}}\right) + \sigma_n/\sigma.$$

If $\sigma_n/\sigma \ll \Omega\left(\varpi_{\text{opt}}\right)$, then the "cost" $\tilde{\varpi}$ of a PTF is only three times greater than the "cost" of an optimal PTF, whereas the discrepancy is two times greater. If such "losses" are acceptable, then $\tilde{\varpi}$ can be adopted as an approximation to the sought-for $\varpi_{\text{opt}}$.

Note that it is possible to search for $\tilde{\varpi}$ as a solution to the operator equation

$$\left(\alpha L + K\right)\tilde{\varpi}_\alpha = f, \quad \alpha > 0,$$

defining $\alpha$ by the condition $\tilde{e}\,(\tilde{\varpi}_\alpha) = \min\limits_{\alpha}$.

**6.** Now we try to answer the question: what is the principal distinctions between the methods for solving stable and unstable problems? We already saw that at a level of formulation of a problem these distinctions are insignificant (in our interpretation of this question).

Obviously, the essence of the question consists in the rules of obtaining a solution. If a problem is stable, such a rule is the same for exact and approximate data, i.e., it is unique for any data. Another situation takes place for unstable problems. In this case, it is impossible to use the same rule to construct a solution for inexact input data (because of instability of the problem). It remains only to change this rule and to make it more flexible.

In our opinion, this is the fundamental idea of the regularization method proposed by A. N. Tikhonov whom the general definition of regularizing algorithms for solving unstable problems belongs to. Without going into details, we refer the reader to [265]. Note that a distinguished feature of regularizing algorithms is their dependence on accuracy of input data, so that in the case of absence of such errors a regularized algorithm becomes the original rule, i.e., they are asymptotically similar. This is similar to the case when instead of operation on an unknown value of some limit we would operate on a completely specified member of a sequence defining this limit and on the characteristics of accuracy in the representation of the limit by this member.

Such an approach to solving mathematical problems requires more careful consideration of accuracy in input data. As a matter of fact, for the rigorous solution of an unstable problem it is necessary to specify the original data together with the characteristics of their accuracy. This circumstance changes the character of understanding of input data in comparison with the classical case.

The variety of regularization methods expands our possibilities in solving unstable problems because of taking account of their specific character and the specific character of input data; this allows one to develop efficient numerical algorithms for classes of problems.

There exist the two kinds of methods for solving unstable problems: regular and weakly regular [172]. The regular methods have the following peculiarity: asymptotically, the "cost" of an approximate solution is equal to the "cost" of the $\Omega$-optimal (normal according to A. N. Tikhonov) solution. The "cost" of a weakly regular solution is equal to the "cost" of the $\Omega$-optimal solution only by an order of magnitude. It is easy to establish the regularity or the weak regularity of the above stable methods for solving the problem of designing $\Omega$-optimal SACs on the basis of their characteristics.

**7.** Let us consider the numerical aspects of designing $\Omega$-optimal SACs. It is possible to point out the two approaches to this problem on the basis of

1) approximation of Euler's equation and

2) approximation of a smoothing functional.

The second approach seems to be more preferable, since it allows one to preserve the main properties (self-conjugacy and positive definiteness) of discrete approximations and to apply special methods for solving the resulting discrete problems (see the previous section).

Note that for any function $\varpi \in D$ the following condition is necessarily fulfilled: $\varpi(t) \to 0,\ t \to +\infty$. Therefore, it is natural to choose a sufficiently large $T > 0$ and to approximate a system with infinite memory by a system with finite memory. According to this remark, it is assumed that $D = \{\varpi : \varpi(t) = 0,\ t \geq T\}$. Then, in the functional $e_\alpha^2\,(\varpi) \equiv e^2\,(\varpi) + \alpha\Omega^2(\varpi)$ the time integration is performed only over the finite interval from 0 to $T$. For simplicity we assume that $m = 2$ and the functional is of the form

$$\Omega^2(\widehat{\varpi}) \equiv \int\limits_0^T \left(\widehat{\varpi}_1'^{\,2} + \widehat{\varpi}_2'^{\,2}\right) dt,$$

i.e., the first-order regularization [265] is used. The following grid of nodes $\{t_i\}$ is given on the segment $[0, T]$: $0 = t_1 < t_2 < \ldots < t_N < t_{N+1} = T$.

The first approximation method is *the method of basis functions.* It consists in using the basis $\varpi_1(t), \ldots, \varpi_N(t)$ with a local support associated with the grid $\{t_i\}$, i.e., the function $\varpi_i(t)$ is not zero only in some neighborhood of the node $t_i$. For example, the $B$-splines [296] have such a locality property. The elementary piecewise linear $B$-splines are the functions

$$\varpi_1(t) = \begin{cases} \dfrac{t_2 - t}{\Delta t_1} & \text{for} \quad 0 \leq t \leq t_2; \\ 0 & \text{for} \quad t \geq t_2; \end{cases}$$

$$\varpi_i(t) = \begin{cases} \dfrac{t - t_{i-1}}{\Delta t_{i-1}} & \text{for} \quad t_{i-1} \leq t \leq t_i; \\ \dfrac{t_{i+1} - t}{\Delta t_i} & \text{for} \quad t_i \leq t \leq t_{i+1}; \\ 0 & \text{for} \quad t \leq t_{i-1},\ t \geq t_{i+1}. \end{cases}$$

Here $i = 1, 2, \ldots, N$ and $\Delta t_i = t_{i+1} - t_i$.

In $e_\alpha^2\,(\varpi)$ we put $\widehat{\varpi}_s(t) \approx \sum_{j=1}^N \xi_j^s \varpi_j(t)$, where $\xi_j^s$ are the sought-for values. If we put $\xi_s = (\xi_1^s, \ldots, \xi_N^s)$ and $\xi = (\xi_1, \xi_2)'$, then by minimizing $e_\alpha^2$ in $\xi$, we obtain the linear system

$$\left(\alpha\hat{L} + \hat{K}\right)\xi = \hat{f}, \tag{2.6.60}$$

where

$$\hat{K} = \begin{pmatrix} K^{11} & K^{12} \\ K^{21} & K^{22} \end{pmatrix}, \ \hat{L} = \begin{pmatrix} L^{11} & 0 \\ 0 & L^{22} \end{pmatrix}, \ \hat{f} = \begin{pmatrix} f^1 \\ f^2 \end{pmatrix}.$$

The elements of the matrices $K^{sp}$ and the components of the vectors $f^s$, $s, p = 1, 2$, are defined by the formulas $(i, j = 1, 2, \dots, N)$:

$$K_{ij}^{sp} = \int_0^T \int_0^T k_{sp}(t-\tau)\varpi_i(t)\varpi_j(\tau)\,dt\,d\tau, \quad f_i^s = \int_0^T f_s(t)\varpi_i(t)\,dt.$$

It is easy to see that in the case of $B$-splines the coefficients $\xi_j^s$ have a clear geometric meaning.

An advantage of the above approximation method is the fact that the matrix of system (2.6.60) is symmetric and positive definite for all $\alpha > 0$, i.e., for solving this system we can always use the Cholesky method [55]. Its disadvantage is the necessity of careful computing the corresponding integrals, which may require a significant computing cost. Note that the grid $\{t_i\}$ may not be uniform. This fact can be used to improve the approximation of a problem with increasing $N$ if the qualitative behavior of the sought-for solution is known *a priori*.

The second approximation method is *the quadrature method*. For any smooth function $u(t)$ we write down the quadrature formula

$$\int_0^T u(t)\,dt \approx \sum_{i=1}^{N} \sigma_i u(t_i),$$

where $\sigma_i = \sigma_i^N$ are the quadrature weights. Using this formula, we approximate $e^2(\varpi)$ by a quadratic form of the variables $\xi = (\xi_1, \xi_2)'$, $\xi_j^s = \varpi_s(t_j)$, $s = 1, 2$. The functional $\Omega^2(\varpi)$ is approximated using the difference ratios. Minimizing the thus-obtained quadratic form in $\xi$, we come to the following system of linear algebraic equations:

$$\left(\alpha\hat{L} + \hat{K}_\sigma\right)\xi = \hat{f}_\sigma. \tag{2.6.61}$$

Here

$$\hat{K}_\sigma = \begin{pmatrix} K_\sigma^{11} & K_\sigma^{12} \\ K_\sigma^{21} & K_\sigma^{22} \end{pmatrix}, \ \hat{L} = \begin{pmatrix} L^{11} & 0 \\ 0 & L^{22} \end{pmatrix}, \ \hat{f}_\sigma = \begin{pmatrix} f_\sigma^1 \\ f_\sigma^2 \end{pmatrix}$$

and the elements of the blocks are as follows $(s = 1, 2,\ i, j = 1, 2, \dots, N)$:

$$\begin{aligned} (K_\sigma^{sp})_{ij} &= \sigma_i k_{sp}\left(\Delta t\,(i-j)\right)\sigma_j, \\ L^{11} &= L^{22} = \frac{1}{\Delta t} L_0, \qquad (f_\sigma^s)_i = \sigma_i f_s(t_i). \end{aligned}$$

In the above formulas it is assumed that the grid in $t$ is uniform: $\Delta t_i \equiv \Delta t$. The matrix $L_0$ is tridiagonal:

$$L_0 = \begin{pmatrix} 1 & -1 & 0 & . & . & . & 0 & 0 \\ -1 & 2 & -1 & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 2 & -1 \\ 0 & 0 & 0 & . & . & . & -1 & 2 \end{pmatrix}.$$

The matrix $\alpha\hat{L} + \hat{K}_\sigma$ is symmetric and positive definite. Let

$$\hat{K}_0 = \begin{pmatrix} K_0^{11} & K_0^{12} \\ K_0^{21} & K_0^{22} \end{pmatrix}, \quad \hat{\sigma} = \begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}, \quad \hat{f} = \begin{pmatrix} f^1 \\ f^2 \end{pmatrix},$$

where $K_0^{sp} = \{k_{sp}(\Delta t\,(i-j))\}$, $f^s = \{f_s(t_i)\}$, and $\sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_N)$; then, it is possible to rewrite (2.6.61) in the form

$$\left(\hat{\sigma}\hat{K}_0\hat{\sigma} + \alpha\hat{L}\right)\xi = \hat{\sigma}\hat{f}.$$

Putting $\xi_{\hat{\sigma}} = \hat{\sigma}\xi$ and $L_{\hat{\sigma}} = \hat{\sigma}^{-1}\hat{L}\hat{\sigma}^{-1}$, we obtain

$$\left(\hat{K}_0 + \alpha L_{\hat{\sigma}}\right)\xi_{\hat{\sigma}} = \hat{f}, \quad \xi = \hat{\sigma}^{-1}\xi_{\hat{\sigma}}. \tag{2.6.62}$$

The system (2.6.62) is more convenient for computation.

*Algorithms for solving regularized systems of linear algebraic equations.* For definiteness, we consider the system (2.6.62). The matrix $L_{\hat{\sigma}}$ is symmetric and positive definite, whereas the matrix $\hat{K}_0$ is symmetric and nonnegative definite. Let $L_\sigma = \sigma^{-1}L_0\sigma^{-1}$. We decompose the matrix $L_\sigma$ in the form $L_\sigma = S'S$, where $S$ is the upper triangular matrix and $S'$ is its transpose. Then, $L_{\hat{\sigma}} = \hat{S}'\hat{S}$, where $\hat{S} = \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}$. This decomposition can be performed by the Cholesky method [55]. Let $\hat{S}\xi_{\hat{\sigma}} = \hat{y}$. Then, from (2.6.62) we come to the system

$$\left(\hat{A} + \alpha\hat{E}\right)\hat{y} = \hat{g},$$

where $\hat{g}$ is easily obtained as the solution to the system $\hat{S}'\hat{g} = \hat{f}$ with the triangular matrix and the matrix $\hat{A}$ is defined as the solution to the matrix equation $\hat{S}'\hat{A}\hat{S} = \hat{K}_0$. Thus, it is natural to use the block structures of the matrices $\hat{S}$ and $\hat{K}_0$. This substantially reduces the number of arithmetic operations, since instead of solution of the system of order $2N$ it is sufficient to solve the system of order $N$ several times. The matrix $\hat{E}$ has the form $\hat{E} = \begin{pmatrix} E & 0 \\ 0 & E \end{pmatrix}$, where $E$ is the unit matrix of order $N$. Note that the inversion of $\hat{S}$ (or $S$) is not required.

The matrix $\hat{A}$ is symmetric and nonnegative definite. By an orthogonal similarity transformation $Q$, it can be reduced to a symmetric tridiagonal matrix: $\hat{T} = Q'\hat{A}Q$ [55]. Putting $Q'\hat{y} = \hat{z}$, we come to the canonical regularized system

$$\left(\hat{T} + \alpha\hat{E}\right)\hat{z} = \hat{\varphi}, \quad \hat{\varphi} = Q'\hat{g}, \tag{2.6.63}$$

with the positive definite tridiagonal matrix; the solution to (2.6.63) can easily be obtained by the sweep method or by the Cholesky method.

The above algorithm is very efficient for the repeated solution of a regularized problem, especially if we take into account that the criterial functions defining the regularization parameter are easily calculated in terms of the solutions to the system (2.6.63). Its essential disadvantage is that it almost completely ignores the specific character of equation (2.6.62) enabling to construct a more efficient algorithm; therefore, it can be recommended only for small $N$ and $m$ in the case of a nonuniform grid $\{t_i\}$. If the problem (2.6.61) is solved for $k$ values of the regularization parameter, then the total number of arithmetic operations is of order $m^3N^3 + km^2N^2$, where the first term takes into account the preliminary work on the above decomposition of the matrices.

What is a specific character of the system (2.6.62)? It turns out that each of the blocks $K_0^{sp}$ is a Toeplitz matrix. Such matrices were previously called the $t$-type matrices [25]. Any $t$-type matrix can be represented in the form of a block-Toeplitz matrix of order $N$ by permutation of rows and column; each block of this matrix has the order $m$ (in our case, $m = 2$). For linear systems with block-Toeplitz matrices there exist more efficient algorithms in the sense of computing time and storage: the number of operations $\sim m^3N^2$ and the amount of storage $\sim m^2N$, which is by an order of magnitude (in $N$) less than in the case of the general approach considered above. The addition of the regularizing term $\alpha\hat{L}$ to the matrix $\hat{K}_0$ changes the properties of $\hat{K}_0$, but rather slightly; therefore, the main conclusions and estimates remain valid. The method is described in [25] and in Section 2.5 in greater detail.

The above specific character of a regularized system can efficiently be used to implement the above-considered algorithm for sequential refinement of regularized solutions on the basis of gradient-type methods. The main operation in this case is the multiplication of the matrix $\alpha L_\sigma + \hat{K}_0$ by a vector, i.e., the computation of $\left(\hat{K}_0 + \alpha L_\sigma\right)\xi$. The direct computation of this vector requires $\sim m^2N^2$ operations and $\sim mN$ words of storage.

The above-developed algorithms for fast multiplication of a Toeplitz matrix by a vector require only $\sim N\log_2 N$ operations with approximately the same storage [58]. This fact allows one to compute a required gradient in $\sim m^2 N \log_2 N$ operations. Obviously, this leads to substantial savings in computing cost due to the use of FFT (fast Fourier transform) algorithms.

The fast computation of gradients may be rather useful in other situations, in particular, for application of the descriptive regularization method when the functional $e_\alpha^2(\varpi)$ is minimized on a set of piecewise-monotone or piecewise-convex functions $\varpi$ as well as for software implementation of the iterative regularization method.

**8.** The above discussion demonstrates that for solving the problem of designing the $\Omega$-optimal SACs it is possible to develop specific numerical methods that take into account the general properties of regularized problems (positive definiteness and symmetry of their matrices) and the other deeper properties, enabling to develop highly efficient algorithms. Note that the form of regularization of problems treated in some works (see, for example, [39], [219], [234]) does not allow one to take into account these properties and, moreover, leads to algorithms that are difficult to interpret at the physical standpoint. In addition, this approach requires extra (not necessary) cost for symmetrization of the operator $K$ (equal to the cost for solving the whole problem); moreover, the condition number of the resulting algebraic problem is substantially greater. Our approach is free from the above disadvantages caused by the formal application of the main ideas of Tikhonov's regularization method. We also draw the reader's attention to the work [109], where some numerical methods for solving similar problems are developed (without explicit application of the regularization method).

Let us describe some numerical results. For an approbation of the above technique of computing the $\Omega$-optimal SACs, the following three model problems with known solutions were considered.

*Model problem* I. The autocorrelation function of an input signal $z$ is

$$k_{zz}(t) = \frac{2}{3}\exp\left(-\frac{1}{2}|t|\right) + \frac{1}{6}\exp\left(-|t|\right);$$

the mutual correlation function of the input signal $z$ and of the output signal $y$ is

$$f = k_{zy}(t) = \frac{4}{3}\exp\left(-\frac{1}{2}t\right) + \frac{1}{6}\exp(-t) - \frac{29}{36}\exp(-t).$$

The exact solution is $\bar{\varpi}(t) = \exp(-t)$, $t \geq 0$, and, obviously, corresponds to the first-order aperiodic link.

*Model problem* II. The autocorrelation function of an input signal $z$ is

$$k_{zz} = \exp\left(-|t|\right);$$

the mutual correlation function of the input signal $z$ and of the output signal $y$ is

$$f = k_{zy} = \frac{2}{5}\,\exp(-t)\,(\cos t + 2\sin t)\,.$$

The exact solution is $\bar{\varpi}(t) = \exp(-t)\cos t$, $t \geq 0$. The same problem was earlier considered in [39].

*Model problem* III. A linear stationary SAC with two inputs is considered. This problem is associated with synthesis of information SACs with indirect action. The initial data are taken from [234]:

$$k_{z_1z_1}(t) = \frac{2}{3}\,\exp\left(-\frac{1}{2}\,|t|\right) + \frac{1}{6}\,\exp(-\,|t|)\,,$$

$$k_{z_2z_2}(t) = \frac{3}{8}\,\exp(-\,|t|) + \frac{3}{8}\,\exp\left(-\frac{1}{3}\,|t|\right),$$

$$k_{z_1z_2}(t) = \frac{1}{3}\,\exp(-t) + \frac{9}{20}\,\exp\left(-\frac{1}{3}\,t\right), \quad t \geq 0,$$

$$k_{z_2z_1}(t) = \frac{1}{4}\,\exp(-t) + \frac{8}{15}\,\exp\left(-\frac{1}{2}\,t\right), \quad t \geq 0,$$

$$f_1 = k_{z_1y}(t) = \frac{5}{12}\,\exp(-t) + \frac{t}{2}\,\exp(-t) - \frac{4}{3}\,\exp(-t/2) + \frac{27}{20}\,\exp(-t/3)\,,$$

$$f_2 = k_{z_2y}(t) = \frac{3}{8}\,\exp(-t) + \frac{t}{2}\,\exp(-t) - \frac{16}{15}\,\exp(-t/2) + \frac{9}{8}\,\exp(-t/3)\,.$$

The exact solution is $\bar{\varpi}(t) = \left(-e^{-t}, 2e^{-t}\right)$, $t \geq 0$.

The stabilizing functional for Problems I and II is $\Omega^2(\varpi) \equiv \int\limits_0^T \varpi'^{\,2}\,dt$ and for problem III is $\Omega^2(\varpi) \equiv \int\limits_0^T \left(p_1\varpi_1'^{\,2} + p_2\varpi_2'^{\,2}\right)dt$, where $p_1$, $p_2 > 0$ are the weighting coefficients.

For the better balancing of the regularization process it is reasonable to take

$$p_1 \approx \frac{1}{\int\limits_0^T \bar{\varpi}_1'^{\,2}\,dt}, \quad p_2 \approx \frac{1}{\int\limits_0^T \bar{\varpi}_2'^{\,2}\,dt}.$$

From the properties of the exact solution it follows that $p_1 = p_2/4$. Since $\Omega^2(\varpi)$ is specified with an accuracy up to a constant, it is possible to put $p_1 = 1/4$ and $p_2 = 1$.

It seems that such a choice of weighting factors corresponds to their "optimal" choice. In our practical applications of the method, we first took

$p_1 = p_2 = 1$ and then, after computing the approximate PTFs $\tilde{\varpi}_1$ and $\tilde{\varpi}_2$, we refined the values of $p_1$ and $p_2$ on the basis of replacement of the exact characteristics by their approximations, which led to the substantial refinement of numerical results. This approach is applicable in the general case for a problem with $m$ inputs if it is impossible to take the weights $p_i$ on the basis of other (for example, *a priori*) considerations devoted to the influence of system channels.

The aim of our computations was not only an approbation of the regularization method, but also an improvement of practical algorithms for the choice of regularization parameter, since these algorithms predetermine the stable solution of problems. To accomplish this, the following characteristics were approximately computed:

1) $\|\varpi^\alpha - \bar{\varpi}\|$ for the determination of the optimal values $\alpha_{\text{opt}}$ of regularization parameter;

2) $\left\|\alpha \dfrac{d\varpi^\alpha}{d\alpha}\right\|$ for the determination of quasioptimal values $\alpha_{\text{qopt}}$ of regularization parameter;

3) $\left\|\alpha \dfrac{df^\alpha}{d\alpha}\right\|$ for the determination of quasioptimal values $\hat{\alpha}_{\text{qopt}}$ of regularization parameter by the above modified method.

In order to establish the stability of approximate solutions, errors were introduced in the correlation functions defining the input data of the problems; these errors were of various intensity from $1\,\%$ to $5\,\%$.

The decay time for the correlation functions was $T = 6.7$ for all the model problems. The number of partition points on the segment $[0, T]$ was $N = 80$ for Problems I and II and $N = 40$ (so that $2 \times N = 80$) for Problem III. These values of $N$ were chosen from the following reasoning: the error caused by the "cut" of correlation functions and the error caused by the quadrature formulas on $[0, T]$ should be equal approximately (the violation of this rule leads to the unpractical use of computer resources or to an inadequately accurate approximation of the original problem).

For Problems I and II, the choice of such values for the parameters $T$ and $N$ was well acceptable, while for Problem III an increase in $T$ and $N$ was essentially required (this was due to its greater complexity and the slower decay of the correlation functions and the solution, but was limited by the resources of the computer in use because of the chosen numerical method for solving the regularized systems of linear algebraic equations: the Cholesky method, which essentially does not use, as mentioned above, the specific character of the problems under consideration). These circumstances, certainly, have caused a lower accuracy of approximations in Problem III, which is necessary to take into account when analyzing the numerical results.

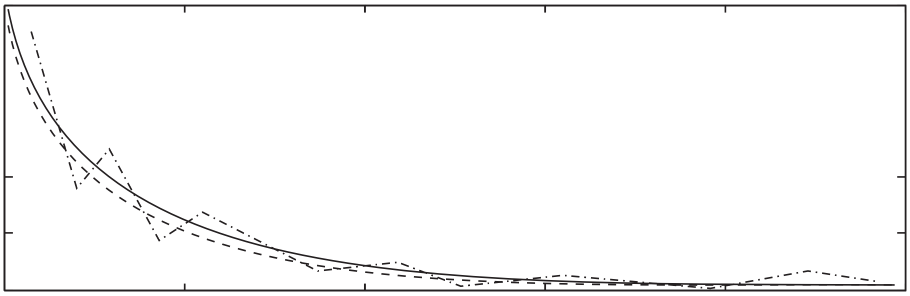

Figure 2.6.12: Influence of the choice of a quadrature formula in the absence of noise: the solid curve — the exact solution; the dashed curve — $\alpha = 0$, the trapezoidal formula; the dot-and-dashed curve — $\alpha = 0$, the Simpson formula

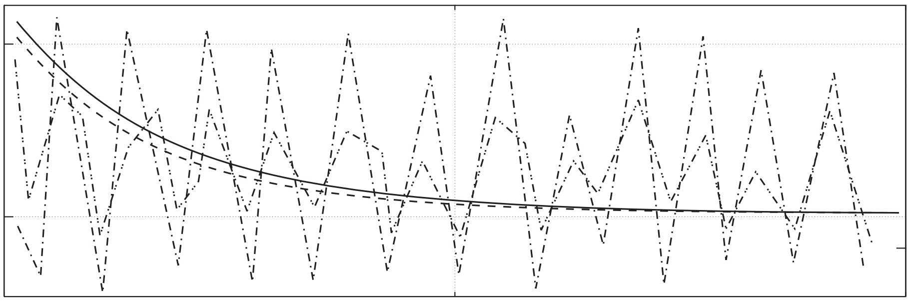

Figure 2.6.13: Effect of regularization in the presence of noise: the solid curve — the exact solution; the dashed curve — $\alpha = \alpha_{\text{opt}} = 10^{-5}$; the dot-and-dashed curve — $\alpha = 3 \cdot 10^{-6}$; the dot-dot-dashed curve — $\alpha = 7 \cdot 10^{-8}$

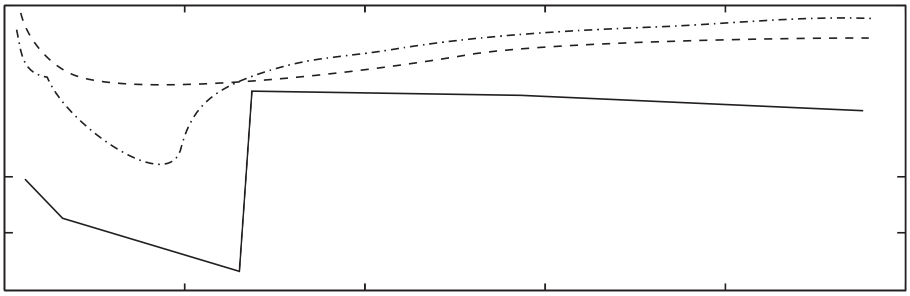

Figure 2.6.14: Dependence of criterial functions on the regularization parameter $\alpha$: the solid curve — $\left\| \alpha \, \dfrac{df^\alpha}{d\alpha} \right\|$; the dashed curve — $\left\| \alpha \, \dfrac{d\varpi^\alpha}{d\alpha} \right\|$; the dot-dashed curve — $\|\varpi^\alpha - \bar{\varpi}\|$

Our computations were basically performed with the use of the trapezoidal quadrature formula, since the use of the more exact Simpson formula led to the worse results, especially in the presence of errors in input data, which

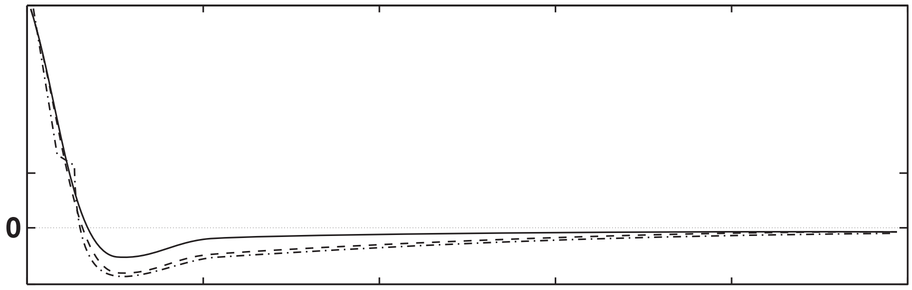


Figure 2.6.15: Influence of the choice of a quadrature formula in the absence of noise: the solid curve — the exact solution; the dashed curve — $\alpha = 0$, the trapezoidal formula; the dot-dashed curve — $\alpha = 0$, the Simpson formula

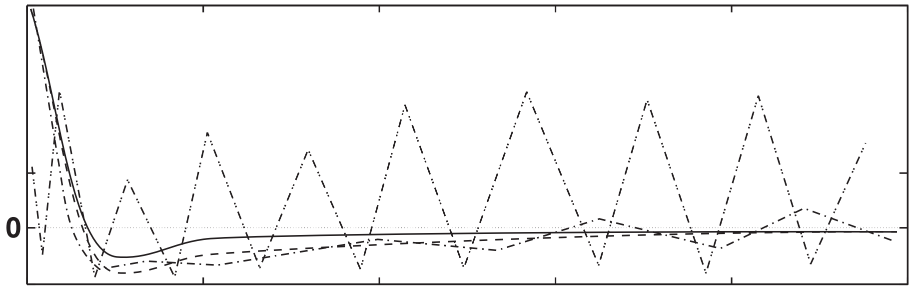


Figure 2.6.16: Effect of regularization in the presence of noise: the solid curve — the exact solution; the dashed curve — $\alpha = \alpha_{\text{opt}} = 9 \cdot 10^{-6}$; the dot-dashed curve — $\alpha = 3 \cdot 10^{-6}$; the dot-dot-dashed curve — $\alpha = 2 \cdot 10^{-8}$

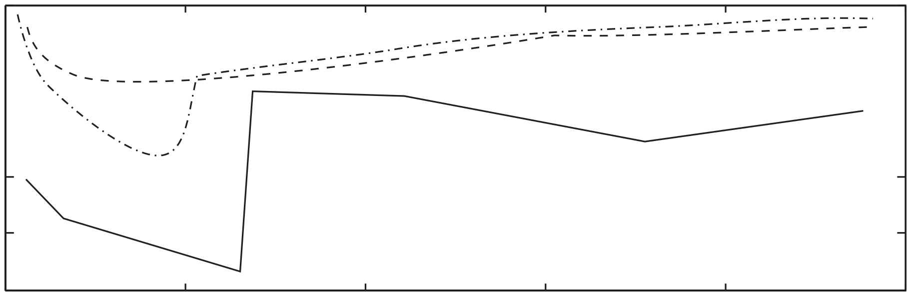

Figure 2.6.17: Dependence of criterial functions on the regularization parameter $\alpha$: the solid curve — $\left\| \alpha \dfrac{df^\alpha}{d\alpha} \right\|$; the dashed curve — $\left\| \alpha \dfrac{d\varpi^\alpha}{d\alpha} \right\|$; the dot-dashed curve — $\|\varpi^\alpha - \bar{\varpi}\|$

should be expected.

The numerical results obtained are represented in Figures 2.6.12–2.6.14 for Problem I, in Figures 2.6.15–2.6.17 for Problem II, and in Figure 2.6.18 for Problem III. It can be seen that for exact data the satisfactory results

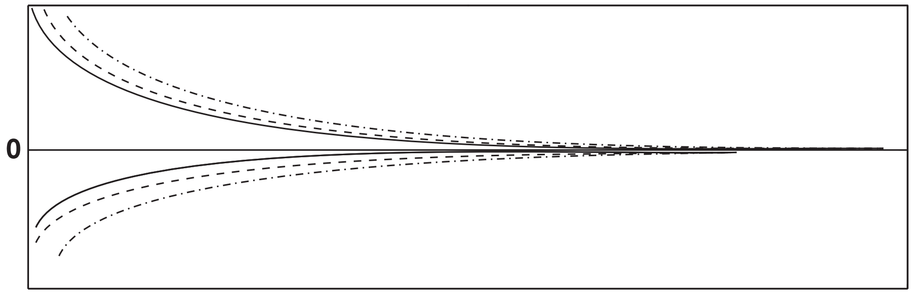


Figure 2.6.18: Influence of the choice of a quadrature formula in the absence of noise in Problem III: the solid curve — the exact solution; the dashed curve — $\alpha = 0$, the trapezoidal formula; the dot-dashed curve — $\alpha = 0$, the Simpson formula

are also obtained with $\alpha = 0$, i.e., in the absence of regularization. This is explained by the "good" $\delta$-shaped character of the autocorrelation functions of input processes. However, small perturbations in data result in the strong numerical instability of the approximations obtained; this instability manifests itself in substantial oscillations for $\alpha \approx 0$, whereas this phenomenon, naturally, disappears when the regularizing solutions are chosen appropriately and we obtain the more qualitative results.

Figures 2.6.14 and 2.6.17 show that an objective choice of the regularization parameter is possible. Similar results are observed for Problem III. The pleasant surprise was the satisfactory performance of the method for the choice of $\hat{\alpha}_{\mathrm{qopt}}$ whose values were in good agreement with optimal values of the regularization parameter. This method was tested for the first time and, therefore, requires the further experimental tests.

All the above numerical results were obtained by V. V. Badeva with a BESM-6 computer installed at the Moscow University Research Computing Center. The more complete description of numerical results is given in [23].

## 2.7 Fast algorithms for solving the Wiener–Hopf equations

**1.** The Wiener–Hopf integral equations form a special class of linear equations. They describe many physical problems; among them are the Milne classical problem connected with the process of neutron transport (or radiative transport) through a substance, the problem of diffraction of electromagnetic waves on a flat screen, various problems of the theory of control, and many others.

The Wiener–Hopf integral equation has the form

$$Au = au(t) + \int_0^\infty k(t-s)u(s)\,ds = f(t), \quad 0 \le t < +\infty, \tag{2.7.64}$$

where $a$ is a constant, $k(t)$ $(-\infty < t < +\infty)$ is an absolutely Lebesgue integrable function, $f(t)$ is a given function, and $u(s)$ is the sought-for function (the functions $f$ and $u$ belong to some Banach spaces $F$ and $U$, respectively; these spaces are described below).

When $a = 0$, equation (2.7.64) is called the Wiener–Hopf integral equation of the first kind; when $a \ne 0$, it is the Wiener–Hopf integral equation of the second kind.

In this section, some theoretical questions relating to the Wiener–Hopf integral equations are discussed, an algorithm for solving the Wiener–Hopf integral equations of the second kind with the use of Fourier transform is described, and the corresponding discrete algorithm for their approximate solution with the use of fast Fourier transform is given. This algorithm allows one to obtain an approximate solution in $O(N \log_2 N)$ operations, where $N$ is the number of grid points on which a solution is sought for. The solutions to model problems and an analysis of numerical results are given. The discussion given below follows [206].

Let us consider the questions connected with the well-posedness for the formulation of the problem of solving the Wiener–Hopf equations. Let $L_p$, $1 \le p < \infty$, be the space of all measurable complex functions $f(t)$ given on the real axis $(-\infty, \infty)$; the $p$th power of $f$ is integrable; the norm of $f$ is defined as

$$\|f\| = \left( \int_{-\infty}^{\infty} |f(t)|^p \, dt \right)^{1/p}.$$

Let $M$ be the space of all functions measurable almost everywhere and bounded on the real axis. The norm of $f \in M$ is defined as

$$\|f\| = \sup_{-\infty<t<+\infty} \operatorname{ess} |f(t)| \, .$$

Let $M^c$ and $M^u$ be the subspaces of all continuous and uniformly continuous functions from $M$, respectively.

Let $C$ be the space of all continuous functions on the real axis $(-\infty, \infty)$ such that there exist the finite limits $f(+\infty) = f(-\infty) = \lim\limits_{t\to\pm\infty} f(t)$. By $C^0$ we denote the space of all functions $f \in C$ such that $f(+\infty) = f(-\infty) = 0$.

The norm in $C$ and $C^0$ is defined as

$$\|f\| = \max_{-\infty \le t \le +\infty} |f(t)|\,.$$

All the above spaces are Banach spaces and $C^0 \subset C \subset M^u \subset M^c \subset M$.

Let $E$ be any of the spaces $L_p$ $(1 \le p < \infty)$, $C^0$, $C$, $M^u$, $M^c$, or $M$ and let $E_+$ be the corresponding space of functions on the positive semiaxis $[0, +\infty)$. The space $E_+$ is also a Banach space.

As is known [40], for any function $f \in L_1$ its Fourier transform

$$(\mathcal{F}f)(\varpi) = \int\limits_{-\infty}^{+\infty} f(t)\exp(i\varpi t)\,dt, \quad -\infty < \varpi < \infty,$$

is a function from $C^0$.

The following function is closely connected with equation (2.7.64):

$$A(\varpi) \equiv a + \int\limits_{-\infty}^{+\infty} k(t)\exp(i\varpi t)\,dt, \quad -\infty < \varpi < \infty.$$

This function is said to be the *symbol of the operator* $A$, the number $\kappa = \dfrac{1}{2\pi}\,[\arg A(\varpi)]_{\varpi=-\infty}^{\infty}$ is called the *index of the symbol* $A(\varpi)$.

Let $D_A \subset U$ be the definition domain of the operator $A$, $R_A \subset F$ be the domain of values for the operator $A$, and $\dim X$ be the dimension of a linear (finite-dimensional or infinite-dimensional) space $X$. The set $\ker A = \{u \in D_A,\ Au = 0\}$ is referred to as the kernel of the operator $A$. The factor-space $\operatorname{coker} A = F/\bar{R}_A$ is called the cokernel of the operator $A$. From here on, the bar over a symbol indicates the operation of closure.

**Definition 1.** *The operator $A$ is said to be normally solvable if $R_A = \bar{R}_A$.*

**Definition 2.** *The operator $A : U \to F$ is said to be* (*continuously*) *invertible if there exists an operator $A^{-1} \in \mathcal{L}(F, U)$ with $R_{A^{-1}} \subset D_A$ such that the following relations hold:*

$$A^{-1}Ax = x, \quad x \in D_A; \quad AA^{-1}y = y, \quad y \in F.$$

Here $A : U \to F$ indicates a linear operator from $U$ into $F$ and $\mathcal{L}(F, U)$ indicates the set of linear continuous (and, hence, bounded) operators $B$ from $F$ into $U$ and $D_B = F$.

**Definition 3.** *The operator $A : U \to F$ is said to be left* (*right*) *invertible if there exists an operator $A^{-1} \in \mathcal{L}(F, U)$ with $R_{A^{-1}} \subset D_A$ such that the following relations hold:*

$$A^{-1}Ax \;=\; x, \quad x \in D_A \quad (AA^{-1}y = y, \;\; y \in F).$$

The theorems on solvability of equation (2.7.64) are formulated as follows (see [224]).

**Theorem 2.7.49** *Let $k(t) \in L_1$. In order that the operator $A$ defined by the formula* (2.7.64) *be at least one-sidedly invertible, it is necessary and sufficient that its symbol be vanished nowhere, i.e., $A(\omega) \neq 0$, $-\infty \leq \omega \leq +\infty$. Under this condition, the following formulas take place:*

$$\dim \ker A = \max(-\kappa, 0), \quad \dim \operatorname{coker} A = \max(\kappa, 0).$$

*If $A(\omega) = 0$ for some $\omega$, $-\infty \leq \omega \leq +\infty$, then the operator $A$ is not normally solvable.*

**Theorem 2.7.50** *Let $A(\omega) \neq 0$, $-\infty \leq \omega \leq +\infty$. Then, if $\kappa \leq 0$, equation* (2.7.64) *is solvable in $E_+$ for any $f \in E_+$.*

**Theorem 2.7.51** *Let $A$ be a linear continuous operator bijectively mapping $D_A$ onto $R_A$. Then, for the continuity of the operator $A^{-1}$ it is necessary and sufficient that $R_A = \bar{R}_A$.*

From Theorems 2.7.49–2.7.51 we easily obtain the following assertion.

**Theorem 2.7.52** *In order that the problem of solving equation* (2.7.64) *be well posed according to Hadamard on the pair of spaces $(E_+, E_+)$, it is necessary and sufficient that the symbol of the operator $A$ be vanished nowhere, i.e., $A(\omega) \neq 0$, $-\infty \leq \omega \leq +\infty$, and the index of the symbol of $A(\omega)$ be equal to zero.*

Since the symbol $A(\omega)$ of the Wiener–Hopf equation of the first $(a = 0)$ kind belongs to $C^0$, then from Theorems 2.7.49 and 2.7.50 we come to the following assertion.

**Theorem 2.7.53** *The problem of solving the Wiener–Hopf equation of the first kind is ill posed according to Hadamard on the pair of spaces $(E_+, E_+)$.*

Using Theorem 124 from [253], we obtain the validity of the following theorem.

**Theorem 2.7.54** *If in equation* (2.7.64) *the kernel $k(x)$ is an even function bounded and continuous with its first derivative for $-\infty < x < +\infty$, the function $k(x)$ monotonically decreases to zero as $x \to +\infty$ and is convex downwards on $[0, +\infty]$, and $a > 0$, then the symbol $A(\omega) > 0$, $-\infty \leq \omega \leq +\infty$, and the index of the symbol of $A(\omega)$ is equal to zero, i.e., the problem of solving equation* (2.7.64) *is well posed according to Hadamard on the pair of spaces $(E_+, E_+)$.*

**2.** Let us consider the application of Fourier transform to the solution of the Wiener–Hopf equations of the second kind. Following [131], we discuss an algorithm for analytic solution of the Wiener–Hopf equation of the second kind:

$$Au \equiv u(x) + \int_0^\infty k(x-t)u(t)\,dt = f(x), \tag{2.7.65}$$

where $u(x)$, $f(x) \in L_1(0,\infty)$, and $k(x) \in L_1(-\infty,\infty)$. It is assumed that the conditions of well-posedness according to Hadamard for the problem (2.7.65) are valid, i.e.,

$$\kappa = 0, \quad A(\varpi) \neq 0, \quad -\infty \leq \varpi \leq +\infty.$$

The space $L_1(0,\infty)$ ($L_1(-\infty,0)$) can be considered as the subspace $L^+$ ($L^-$) of the space $L_1(-\infty,\infty)$ consisting of the functions $f(t)$ equal identically to zero for $t<0$ ($t>0$). On the subspace $L^+$ ($L^-$) the projector $P_+$ ($P_-$) is naturally defined:

$$P_+x(t) = x(t), \quad t \geq 0; \quad P_+x(t) = 0, \quad t<0;$$
$$(P_-x(t) = x(t), \quad t<0; \quad P_-x(t) = 0, \quad t \geq 0)\,.$$

Now we turn to equation (2.7.65). This equation can be rewritten in operator form as follows:

$$u(x) + Ku(x) = f(x) + P_-Ku(x), \tag{2.7.66}$$

where $K$ is the integral operator in (2.7.65) acting from $L^+$ into $L_1$.

It is well known that the Fourier transform of convolution is equal to the product of the Fourier transforms of each multiplier [253] (the theorem of multiplication). After application of Fourier transform, therefore, equation (2.7.66) takes the form

$$U(\varpi) + K(\varpi)U(\varpi) = F(\varpi) + Z(\varpi), \tag{2.7.67}$$

where $Z(\varpi)$ is the Fourier transform of the function $P_-Ku(x)$. Here $U(\varpi)$, $K(\varpi)$, and $F(\varpi)$ are the Fourier images for the functions $u$, $k$, and $f$, respectively.

Let us rewrite equation (2.7.67) in the form

$$A(\varpi)U(\varpi) = F(\varpi) + Z(\varpi), \tag{2.7.68}$$

where $A(\varpi)$ is the symbol of the operator $A$.

It is known [224] that if $\kappa = 0$ and $A(\varpi) \neq 0$, $-\infty \leq \varpi \leq +\infty$, then the following representation is valid:

$$A^{-1}(\varpi) = A_-(\varpi)A_+(\varpi). \tag{2.7.69}$$

Here

$$A_-(\varpi) = 1 + \int\limits_{-\infty}^{0} \alpha_1(t)\, \exp(i\varpi t)\, dt, \quad A_+(\varpi) = 1 + \int\limits_{0}^{\infty} \alpha_2(t)\, \exp(i\varpi t)\, dt.$$

Similarly,

$$A_-^{-1}(\varpi) = 1 + \int\limits_{-\infty}^{0} \beta_1(t)\, \exp(i\varpi t)\, dt, \quad A_+^{-1}(\varpi) = 1 + \int\limits_{0}^{\infty} \beta_2(t)\, \exp(i\varpi t)\, dt,$$

where the functions $\alpha_1(t)$, $\beta_1(t) \in L^-$ and the functions $\alpha_2(t)$, $\beta_2(t) \in L^+$. The representation (2.7.69) is called the *factorization of* $A(\varpi)$.

Applying the factorization (2.7.69) to $A^{-1}(\varpi)$, from the relation (2.7.68) we obtain

$$A_+^{-1}(\varpi) U(\varpi) = A_-(\varpi) F(\varpi) + A_-(\varpi) Z(\varpi). \qquad (2.7.70)$$

The operator $P_+$ $(P_-)$ generates the projective operator $\Pi_+$ $(\Pi_-)$ in the Fourier images; this operator associates the Fourier image of a function $f$ with the Fourier image of the function $P_+ f$ $(P_- f)$:

$$\Pi_+ F(\varpi) = \mathcal{F} P_+ \mathcal{F}^{-1} F(\varpi) \quad (\Pi_- F(\varpi) = \mathcal{F} P_- \mathcal{F}^{-1} F(\varpi)).$$

Here $\mathcal{F}^{-1} F(\varpi) = f(x)$ and $F(\varpi) = \mathcal{F} f(x)$.

Therefore, it is possible to write down the following representation:

$$F(\varpi) = \Pi_+ F(\varpi) + \Pi_- F(\varpi) = F_+(\varpi) + F_-(\varpi), \qquad (2.7.71)$$

where $F_+(\varpi)$ $(F_-(\varpi))$ is the Fourier image of a function from $L^+$ $(L^-)$. The representation (2.7.71) is called the *decomposition of the function* $F(\varpi)$.

Using this decomposition, from (2.7.70) we obtain the equality

$$A_+^{-1}(\varpi) U(\varpi) = \Pi_+ \left[ A_-(\varpi) F(\varpi) \right] + \Pi_- \left[ A_-(\varpi) F(\varpi) \right] + A_-(\varpi) Z(\varpi).$$

The following assertion [253] is also valid: the product of the Fourier images of a function from $L$ $(L_+, L_-)$ is the Fourier image of a function from $L$ $(L_+, L_-)$.

Having applied the operator $\Pi_+$ to both sides of the last equality and taking into account the above properties, we obtain

$$A_+^{-1}(\varpi) U(\varpi) = \Pi_+ \left[ A_-(\varpi) F(\varpi) \right].$$

From here we come to the final result

$$U(\varpi) = A_+(\varpi) \Pi_+ \left[ A_-(\varpi) F(\varpi) \right].$$

Under certain conditions, for the calculation of $A_+(\varpi)$ it is possible to use the formula [101]

$$A_+(\varpi) = \exp(\mathcal{F}P_+\mathcal{F}^{-1}\ln A^{-1}(\varpi)),$$

whereas for the calculation of $\mathcal{F}^{-1}F(\varpi)$ we can use the inverse Fourier transform [253]:

$$\mathcal{F}^{-1}F(\varpi) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} F(\varpi)\exp(-i\varpi t)\,d\varpi.$$

Thus, it is required to apply the Fourier transform (or the inverse to it) seven times to calculate $u(x)$. Indeed,

1) $F(\varpi) = \mathcal{F}f(x)$;

2) $K(\varpi) = \mathcal{F}k(x)$, $\quad A(\varpi) = 1 + K(\varpi)$;

3) $g(x) \equiv \mathcal{F}^{-1}\ln A^{-1}(\varpi)$;

4) $G_+(\varpi) = \mathcal{F}P_+g(x)$, $\quad A_+(\varpi) = \exp\left(G_+(\varpi)\right)$

5) $A_-(\varpi) = A^{-1}(\varpi)/A_+(\varpi)$, $\quad c(x) \equiv \mathcal{F}^{-1}\left[A_-(\varpi)F(\varpi)\right]$;

6) $C_+(\varpi) = \mathcal{F}P_+c(x)$, $\quad U(\varpi) = A_+(\varpi)C_+(\varpi)$;

7) $u(x) = \mathcal{F}^{-1}U(\varpi)$.

**3.** The algorithm described above was used to solve equation (2.7.65) with a symmetric kernel $k(t) = k(-t)$ [206]. The symmetry of the kernel allows us to simplify the general algorithm and, then, to reduce computing cost in the process of approximate solution.

Let us describe this algorithm in brief. Since $k(t)$ is an even function, we have

$$K(\varpi) = 2\int\limits_0^{+\infty} k(t)\cos\varpi t\,dt.$$

Hence, $\ln A^{-1}(\varpi)$ is also an even function and, therefore,

$$g(t) \equiv \mathcal{F}^{-1}\ln A^{-1}(\varpi) = \frac{1}{\pi}\int\limits_0^{\infty}\ln A^{-1}(\varpi)\cos\varpi t\,d\varpi.$$

Further,

$$\mathcal{F}P_+\mathcal{F}^{-1}\ln A^{-1}(\varpi) = \int\limits_0^{\infty} g(t)\exp(i\varpi t)\,dt \equiv G_+(\varpi).$$

It is easy to show the validity of the following lemma.

**Lemma 2.7.24** *Let*

$$G_+(\varpi) = \int\limits_0^\infty g(t)\exp(i\varpi t)\,dt, \quad g(t) = \frac{1}{2\pi}\int\limits_{-\infty}^\infty G_+(\varpi)\exp(-i\varpi t)\,d\varpi.$$

*Then, in order that $g(x)$ be a real function, it is necessary and sufficient that $G_+(\varpi) = \overline{G_+(-\varpi)}$. If $G_+(\varpi) = \overline{G_+(-\varpi)}$, then*

$$g(t) = \frac{1}{\pi}\,\mathrm{Re}\left[\int\limits_0^\infty \overline{G_+(\varpi)}\exp(i\varpi t)\,d\varpi\right], \quad t \geq 0.$$

Since $A_+(\varpi) = \exp G_+(\varpi)$, it follows from this lemma that $A_+(\varpi) = \overline{A_+(-\varpi)}$. With the aid of simple transformations and taking into account that $g(x)$ is an even function, it is possible to show that $A_+(\varpi) = \overline{A_-(\varpi)}$. Therefore,

$$A_-(\varpi)F(\varpi) = \overline{A_+(\varpi)}F(\varpi) \equiv C(\varpi).$$

By the above lemma we conclude that $C(\varpi) = \overline{C(-\varpi)}$ and, hence,

$$\mathcal{F}^{-1}C(\varpi) = \frac{1}{\pi}\,\mathrm{Re}\left[\int\limits_0^\infty \overline{C(\varpi)}\exp(i\varpi t)\,d\varpi\right] \equiv c(t).$$

Further,

$$\Pi_+\left[A_-(\varpi)F(\varpi)\right] = C_+(\varpi) \equiv \int\limits_0^\infty c(t)\exp(i\varpi t)\,dt.$$

Since $U(\varpi) = A_+(\varpi)C_+(\varpi)$, then by the above lemma we finally come to the following result:

$$u(t) = \frac{1}{\pi}\,\mathrm{Re}\left[\int\limits_0^\infty \overline{U(\varpi)}\exp(i\varpi t)\,d\varpi\right].$$

It can be seen from the above formulas that all of them require the knowledge of values of functions on the positive semiaxis; this property is the source of savings mentioned previously.

The numerical realization of this algorithm is based on the discrete fast Fourier transform (FFT). As is known, the FFT requires $\sim N\log_2 N$ operations [196], where $N$ is the number of readings.

In order to evaluate the integrals in the above-described algorithm, the infinite intervals of integration are replaced by sufficiently large finite intervals, where the integrals are approximated by the trapezoidal formula on a uniform grid.

On the half-interval $[0, T)$, in the variable $t$ we take a uniform grid whose grid point number $N$ is equal to an integer power of 2 and whose step is $\Delta t = T/N$, $t_r = r\Delta t$, $r = \overline{0, N-1}$. In the variable $\varpi$, we take a uniform grid with the grid point number equal to $N$ and with the step $\Delta\varpi = \pi/T$, $\varpi_k = k\Delta\varpi$, $k = \overline{0, N-1}$.

To be specific, we consider the following formulas for $K(\varpi)$ and $u(t)$:

$$\begin{aligned} K(\varpi_j) &\approx 2\mathrm{Re}\left[\int_0^T k(t)\exp(i\varpi_j t)\,d\varpi\right] \approx \\ &\approx \frac{2T}{N}\mathrm{Re}\sum_{r=0}^{N-1} k_r \exp\left(\frac{\pi i}{N} jr\right), \quad j = \overline{0, N-1}. \end{aligned}$$

Here

$$k_r = k(t_r), \quad r = \overline{1, N-2}, \quad k_0 = \frac{1}{2}k(t_0), \quad k_{N-1} = \frac{1}{2}k(t_{N-1});$$

$$u(t_r) \approx \frac{1}{\pi}\left[\int_0^{\pi N/T} \overline{U(\varpi)}\exp(i\varpi t_r)\,d\varpi\right] \approx$$

$$\approx \frac{1}{T}\mathrm{Re}\sum_{j=0}^{N-1} U_j \exp\left(\frac{\pi i}{N} jr\right), \quad r = \overline{0, N-1},$$

where

$$U_j = \overline{U(\varpi_j)}, \quad j = \overline{1, N-2}, \quad U_0 = \frac{1}{2}\overline{U(\varpi_0)}, \quad U_{N-1} = \frac{1}{2}\overline{U(\varpi_{N-1})}.$$

The above numerical algorithm was implemented as a subroutine in FORTRAN with taking into account the symmetry of the kernel (2.7.65). The following two model problems were solved with this subroutine:

$$1)\ k(t) = 4\exp(-|t|), \quad -\infty < t < +\infty, \quad f(t) = \exp(-t), \quad 0 \le t < +\infty,$$

the exact solution is $u_{\mathrm{ex}} = 0.5\exp(-3t)$, $0 \le t < +\infty$;

$$\begin{aligned} 2)\quad & k(t) = \exp(-|t|), && -\infty < t < +\infty, \\ & f(t) = \frac{1}{5}\exp(-t)(7\cos t + 4\sin t), && 0 \le t < +\infty, \end{aligned}$$

the exact solution is $u_{\rm ex} = \exp(-t)\cos t$, $0 \le t < +\infty$.

The quite satisfactory numerical results (see Tables 2.7.1 and 2.7.2) were obtained. The deviations of the approximate solutions from the exact ones in the discrete norms in $L_2$ and $C$ for $N = 512$ and $T = 8$ were equal to $\sim 4\cdot 10^{-5}$ and $\sim 10^{-4}$ for the first problem and to $\sim 2\cdot 10^{-4}$ and $\sim 2\cdot 10^{-3}$ for the second problem. The computing time (without compilation) on a BESM-6 computer was 4.78 seconds for the first problem and 5.10 seconds for the second problem. As mentioned in Section 2.5, the solution of a similar problem reduced to a system of linear algebraic equations with a symmetric Toeplitz matrix of order 1000 requires 1 minute 27 seconds [206]; hence, saving of computing time is more than eight times. Note also that the required storage is proportional to $N$, i.e., our subroutine is designed for processing large amounts of data.

**Table 2.7.1. Problem 1**

| | $t_1$ | $t_{10}$ | $t_{20}$ | $t_{30}$ | $t_{100}$ | $t_{200}$ | $t_{300}$ |
|---|---|---|---|---|---|---|---|
| $u_{\rm ex}(t)$ | 0.50000 | 0.33667 | 0.21695 | 0.13980 | 0.00416 | 0.00005 | 0.00000 |
| $u_{\rm ap}(t)$ | 0.49986 | 0.33662 | 0.21692 | 0.13979 | 0.00416 | 0.00005 | 0.00000 |

**Table 2.7.2. Problem 2**

| | $t_1$ | $t_{10}$ | $t_{20}$ | $t_{30}$ | $t_{100}$ | $t_{200}$ | $t_{300}$ |
|---|---|---|---|---|---|---|---|
| $u_{\rm ex}(t)$ | 1.00000 | 0.86888 | 0.72792 | 0.59578 | -0.00524 | -0.04667 | -0.00200 |
| $u_{\rm ap}(t)$ | 0.99836 | 0.86886 | 0.72791 | 0.59577 | -0.00524 | -0.04667 | -0.00200 |

In Tables 2.7.1 and 2.7.2, $t = (j-1)\,\Delta t$, $j = \overline{1, N}$, $u_{\rm ex}(t)$ is the exact solution, and $u_{\rm ap}(t)$ is the approximate solution obtained by our numerical algorithm.

## 2.8 Approximate solution of Wiener–Hopf discrete equations and estimation of accuracy

**1.** In many problems of mathematical physics the discrete Wiener–Hopf equation

$$\sum_{j=1}^{\infty} a_{k-j} x_j = b_k, \quad k = 0, 1, \ldots, \tag{2.8.72}$$

has a wide application. It is assumed here that $a = \{a_j\}_{-\infty}^{\infty} \in l_1$ and $b = \{b_j\}_0^{\infty}$ are given vectors and $x = \{x_j\}_0^{\infty}$ is the sought-for vector in the space $l_p$, $1 \le p < \infty$.

A fast algorithm for approximate solution of discrete Wiener-Hopf equations with symmetric real matrices is discussed in [199]. Here we consider the general case when the matrix in (2.8.72) is nonsymmetric and its elements are complex numbers. An algorithm with $O(N \log_2 N)$ operations, where $N$ is the order of the reduced (truncated) system (2.8.72), is proposed and substantiated and some estimates for the convergence rate of an approximate solution to the exact one are obtained.

Let $l_p$, $1 \le p < \infty$, be the space of all sequences of complex numbers $a = \{a_j\}_{-\infty}^{\infty}$ with the norm

$$\|a\|_p = \left( \sum_{n=-\infty}^{\infty} |a_n|^p \right)^{1/p} < +\infty$$

and $c^0$ be the space of all sequences $a = \{a_j\}_{-\infty}^{\infty}$ converging to zero. The space $c^0$ after introducing the norm

$$\|a\| = \sup_n |a_n|$$

becomes a Banach space.

In the sequel, by $E$ we mean any of the above spaces $\big(l_p,\ 1 \le p < \infty;\ c^0\big)$; by $E_+$ and $E_-$ we respectively denote the subspaces of the sequences

$$a^+ = \{\dots, 0, a_0, a_1, \dots\} \quad \text{and} \quad a^- = \{\dots, a_{-2}, a_{-1}, 0, \dots\}.$$

Let $P^+$ be the projector from the space $E$ onto the space $E_+$:

$$P^+ a = a^+, \quad a \in E, \quad a^+ \in E_+, \quad a_j^+ = a_j, \quad j = 0, 1, \dots,$$

and $P^- = I - P^+$ be the projector onto $E_-$, where $I$ is the unit operator in $E$.

Let $m \ge 0$ be integer. By $E^m$ we denote the space of all sequences of complex numbers $a = \{a_j\}_{-\infty}^{\infty}$ such that

$$\left\{ n^k a_n \right\}_{-\infty}^{\infty} \in E, \quad k = 0, 1, \dots, m.$$

The introduction of the norm

$$\|a\|_{E^m} = \sum_{k=0}^{m} \left\| \left\{ n^k a_n \right\}_{-\infty}^{\infty} \right\|_E$$

causes the space $E^m$ to become a Banach space.

By $W^m$, $0 \le m < \infty$, $W^0 \equiv W$ we denote the Banach algebra of all functions $a(z)$ such that they are defined on the unit circle $|z| = 1$ and their Fourier coefficients form a sequence $a$ from the space $l_1^m$ with the norm

$$\|a(z)\|_{W^m} = \|a\|_{l_1^m}.$$

Let $E_N$ be the subspace $E$ of vectors $a^{(N)} \in E$ such that $a^{(N)} = \left\{\ldots, 0, a_{-N}^{(N)}, \ldots, a_{N-1}^{(N)}, 0, \ldots\right\}$ and $E_N^+$ ($E_N^-$) be the subspace of the space $E_N$ of vectors $a_+^{(N)} \in E_N$ ($a_-^{(N)} \in E_N$) such that

$$\left(a_+^{(N)}\right)_{-k} = 0 \quad \left(\left(a_-^{(N)}\right)_{k-1} = 0\right), \quad k = 1, 2, \ldots, N.$$

By $P_N$ we denote the projector from $E$ onto $E_N$:

$$P_N a = a^{(N)}, \quad a \in E, \quad a^{(N)} \in E_N, \quad a_j^{(N)} = a_j, \quad j = \overline{-N, N-1}.$$

By the multiplication of vectors $a$ and $b$ we mean their component-wise multiplication $ab = \{a_j b_j\}_{-\infty}^{\infty}$.

The discrete Fourier transform [196] of a vector $a^{(N)} \in E_N$ in $E_N$ is defined as the vector $A^{(N)} = \Im_N a^{(N)}$, where

$$A_k^{(N)} \equiv \sum_{n=-N}^{N-1} a_n^{(N)} \exp\left(\frac{\pi i}{N} kn\right), \quad k = \overline{-N, N-1}, \quad i = \sqrt{-1}.$$

The inverse transform $a^{(N)} = \Im_N^{-1} A^{(N)}$ is defined by the relations

$$a_n^{(N)} \equiv \frac{1}{2N} \sum_{k=-N}^{N-1} A_k^{(N)} \exp\left(-\frac{\pi i}{N} kn\right), \quad n = \overline{-N, N-1}.$$

Let $f$ be a function of complex variable. By $f(a)$ we mean the vector $f(a) \equiv \{f(a_j)\}_{-\infty}^{\infty}$.

Let $G_N$ be the following operator acting from $W$ onto $E_N$:

$$\{G_N a(z)\}_k = a\left[\exp\left(\frac{\pi i}{N} k\right)\right], \quad |z| = 1, \quad -N \le k \le N-1.$$

Note that

$$\left\|\Im_N^{-1} G_N a(z)\right\|_1 \le \|a(z)\|_W, \quad |z| = 1. \tag{2.8.73}$$

The convolution of two vectors $a \in l_1$ and $b \in E$ is the vector $c$ with its components

$$c_k = \sum_{j=-\infty}^{\infty} a_{k-j} b_j, \quad k = 0, \pm 1, \pm 2, \ldots .$$

The cyclic convolution of two vectors $a^{(N)}$ and $b^{(N)} \in E$ is called the vector $c^{(N)} \in E_N$ with its components

$$c_k^{(N)} = \sum_{n=-N}^{N-1} a_{k-n}^{(N)} b_n^{(N)}, \quad k = \overline{-N, N-1},$$

where $a_{-N-j}^{(N)} = a_{N-j}^{(N)}$ and $a_{N+j-1}^{(N)} = a_{-N+j-1}^{(N)}$, $j = 1, 2, \ldots, N$. The above convolutions are denoted by $c \equiv a * b$ and $c^{(N)} \equiv a^{(N)} \otimes b^{(N)}$.

Direct calculations show the validity of the multiplication theorem:

$$\Im_N \left( a^{(N)} \otimes b^{(N)} \right) = \Im_N a^{(N)} \cdot \Im_N b^{(N)}.$$

It is well known [224] that if $c = a * b$, where $a \in l_1$, $b \in E^m$, $0 \le m < \infty$, then $c \in E^m$ and

$$\|c\|_{E^m} \le \|a\|_{l_1^m} \|b\|_{E^m} . \tag{2.8.74}$$

For any vector $a \in l_1^m$ the function

$$a(z) = \sum_{j=-\infty}^{\infty} a_j z^j, \quad |z| = 1,$$

is called the symbol of $a$ and the number

$$\kappa_a = \operatorname{ind} a(z) = \frac{1}{2\pi} \left[ \arg a(z) \right]_{|z|=1}$$

is called its index.

Recall the well-known theorem on convergence of the reduction method for the discrete Wiener–Hopf equation [224].

**Theorem 2.8.55** *Let sequences $a$ and $b$ satisfy the conditions: $a \in l_1^m$, $b \in P^+ E^m$, $a(z) \ne 0$ $(|z| = 1)$, $\kappa_a = 0$. Then, it is possible to indicate a sufficiently large natural number $N_0$ such that the reduced system*

$$\sum_{k=0}^{N-1} a_{j-k} x_k = b_j, \quad j = 0, 1, \ldots, N-1, \tag{2.8.75}$$

*has a unique solution $x_+^{(N)} \in E_N^+$ for $N \ge N_0$ and the sequence $\left\{ x_+^{(N)} \right\}_1^\infty$ converges to $x^+$ in the norm of the space $E^m$, where $x^+$ is a unique solution to equation* (2.8.72). *Moreover, the following convergence rate estimate is valid:*

$$\left\| x^+ - x_+^{(N)} \right\|_E = o(N^{-m}).$$

**2.** In the sequel, we establish some properties of the convolution operation and the discrete Fourier transform; based on these properties and on Theorem 2.8.55, we substantiate an algorithm for approximate solution of equation (2.8.72).

**Lemma 2.8.25** *If a sequence* $\left\{a^{(N)}\right\}_1^\infty$ *converges in* $E^m$, $0 \le m < \infty$, *to a vector* $a$, *then*

$$\left\|a^{(2N)} - P_N a^{(2N)}\right\|_E = o(N^{-m}),$$

*where* $a^{(2N)} \in E_{2N}$.

The proof directly follows from the inequality

$$\begin{aligned}\left\|a^{(2N)} - P_N a^{(2N)}\right\|_{E^m} &\le \left\|a^{(2N)} - a\right\|_{E^m} + \|a - P_N a\|_{E^m} + \\ &+ \|P_N\| \left\|a - a^{(2N)}\right\|_{E^m}.\end{aligned}$$

**Lemma 2.8.26** *Let* $\left\{a^{(N)}\right\}_1^\infty$ *converge in* $l_1^m$ *to a vector* $a$ *and a sequence* $\left\{x^{(N)}\right\}_1^\infty$ *converge in* $E^m$ *to a vector* $x$. *Then, the sequence* $\left\{y^{(N)}\right\}_1^\infty$, *where* $y^{(N)} = a^{(N)} \otimes x^{(N)}$, *converges in* $E^m$ *to the vector* $y = a * x$ *and* $\left\|a^{(N)} * x^{(N)} - a^{(N)} \otimes x^{(N)}\right\|_E = o(N^{-m})$.

**Proof**. Using the inequality (2.8.74), we obtain

$$\begin{aligned}&\left\|y - z^{(2N)}\right\|_{E^m} = \left\|a * x - a^{(N)} * x^{(N)}\right\|_{E^m} \le \\ &\quad \le \|a\|_{l_1^m} \left\|x - x^{(N)}\right\|_{E^m} + \left\|x^{(N)}\right\|_{E^m} \left\|a - a^{(N)}\right\|_{l_1^m},\end{aligned} \tag{2.8.76}$$

where $z^{(2N)} = a^{(N)} * x^{(N)}$.

Further, let $Q_N = I - P_N$. Using Lemma 2.8.25, from (2.8.76) we get

$$\begin{aligned}\left\|Q_N\left(a^{(N)} * x^{(N)}\right)\right\|_{E^m} &\to 0, \quad N \to \infty, \\ \left\|Q_N\left(a^{(N)} * x^{(N)}\right)\right\|_{E} &= o(N^{-m}).\end{aligned}$$

It is easy to check that

$$z^{(2N)} - y^{(N)} = Q_N z^{(2N)} - V^{-2N} P^+ Q_N z^{(2N)} - V^{2N} P^- Q_N z^{(2N)},$$

where $V$ is the shift operator: $V^k \left\{a_n\right\}_{-\infty}^\infty = \left\{a_{n+k}\right\}_{-\infty}^\infty$, $V^0 = 1$.

Obviously,

$$\left\|V^{-2N} P^+ Q_N z^{(2N)}\right\|_{E^m} \le \left\|P^+ Q_N z^{(2N)}\right\|_{E^m}.$$

From here it follows that

$$\left\| y^{(N)} - z^{(2N)} \right\|_{E^m} \leq 3 \left\| Q_N z^{(2N)} \right\|_{E^m} \to 0, \quad N \to \infty, \tag{2.8.77}$$

and

$$\left\| a^{(N)} * x^{(N)} - a^{(N)} \otimes x^{(N)} \right\|_E = o(N^{-m}).$$

This completes the proof.

With the aid of the inequalities (2.8.74) and (2.8.77), we obtain the following inequality used in the sequel:

$$\left\| a^{(N)} \otimes x^{(N)} \right\|_{E^m} \leq 4 \left\| a^{(N)} \right\|_{l_1^m} \left\| x^{(N)} \right\|_{E^m}. \tag{2.8.78}$$

**Lemma 2.8.27** *If a sequence* $\left\{ a_+^{(N)} \right\}_1^\infty$ $\left( \left\{ a_-^{(N)} \right\}_1^\infty \right)$ *converges to a vector* $a^+ \in P^+ l_1^m$ $(a^- \in P^- l_1^m)$ *and a sequence* $\left\{ x_+^{(N)} \right\}_1^\infty$ $\left( \left\{ x_-^{(N)} \right\}_1^\infty \right)$ *converges in* $E^m$ *to the vector* $x^+$ $(x^-)$, *then*

$$\left\| P^- \left( a_+^{(N)} \otimes x_+^{(N)} \right) \right\|_{E^m} \to 0, \quad N \to \infty$$

$$\left( \left\| P^+ \left( a_-^{(N)} \otimes x_-^{(N)} \right) \right\|_{E^m} \to 0, \quad N \to \infty \right)$$

*and*

$$\left\| P^- \left( a_+^{(N)} \otimes x_+^{(N)} \right) \right\|_E = o\left( N^{-m} \right), \quad \left( \left\| P^+ \left( a_-^{(N)} \otimes x_-^{(N)} \right) \right\|_E = o\left( N^{-m} \right) \right).$$

**Proof**. The substantiation is performed for $P^- \left( a_+^{(N)} \otimes x_+^{(N)} \right)$ (the proof is similar for $P^+ \left( a_-^{(N)} \otimes x_-^{(N)} \right)$).

Since $P^- \left( a_+^{(N)} * x_+^{(N)} \right) = 0$, then

$$P^- \left( a_+^{(N)} \otimes x_+^{(N)} \right) = P^- \left( a_+^{(N)} \otimes x_+^{(N)} \right) - P^- \left( a_+^{(N)} * x_+^{(N)} \right).$$

Hence,

$$\left\| P^- \left( a_+^{(N)} \otimes x_+^{(N)} \right) \right\|_{E^m} \leq \left\| P^- \right\| \; \left\| a_+^{(N)} \otimes x_+^{(N)} - a_+^{(N)} * x_+^{(N)} \right\|_{E^m}.$$

Now it is sufficient to use Lemma 2.8.26 in order to obtain the proof of Lemma 2.8.27.

**3.** Let us consider some properties of the discrete Fourier transform.

**Lemma 2.8.28** *Let a vector $a \in l_1^m$ be given, its symbol $a(z) \neq 0$, $|z| = 1$, and its index $\kappa_a = 0$. Then, for a sufficiently large natural $N$ there exists a sequence $b \in l_1^m$ such that*

$$\ln\left(\Im_N P_N a\right) = \Im_N P_N b + \Im_N \varepsilon^{(N)}, \tag{2.8.79}$$

*where $\varepsilon^{(N)} \in E_N$ and $\left\|\varepsilon^{(N)}\right\|_1 = 0\left(N^{-m}\right)$.*

**Proof**. By $\ln a(z)$ for $|z| = 1$ we mean any branch of this function under the condition of its continuity on $\Gamma : |z| = 1$.

Since $a(z) \in W^m$, $a(z) \neq 0$, and $\kappa_a = \operatorname{ind} a(z) = 0$, then by the theorem from [289] we come to the conclusion that $\ln a(z) = b(z) \in W^m$, i.e.,

$$\ln\left(\sum_{n=-\infty}^{\infty} a_n z^n\right) = \sum_{n=-\infty}^{\infty} b_n z^n, \quad b \in l_1^m, \quad |z| = 1. \tag{2.8.80}$$

From the hypotheses of this lemma it follows that there exist an absolute constant $c > 0$ and a natural number $N_0$ such that

$$|a_N(z)| \geq c, \quad a_N(z) = \sum_{n=-N}^{N-1} a_n z^n, \quad |z| = 1, \quad N \geq N_0.$$

Let us rewrite (2.8.73) in the form

$$\ln a_n(z) + \ln\left[1 + \frac{a(z) - a_N(z)}{a_N(z)}\right] = b_N(z) + b(z) - b_N(z).$$

Applying the operator $G_N$ to the last equality, we obtain

$$\ln \Im_N P_N a + G_N \ln\left[1 + \frac{a(z) - a_N(z)}{a_N(z)}\right] = \Im_N P_N b + G_N\left[b(z) - b_N(z)\right].$$

Using the fact that in the Banach algebra, $x^{-1}$ is a continuous function of $x$ on the set of invertible elements [65], we conclude that $\left\|\left[a_N(z)\right]^{-1}\right\| \to \left\|\left[a(z)\right]^{-1}\right\|$. Since $\|a(z) - a_N(z)\| = o(N^{-m})$, with the aid of (2.8.73) and the inequality (2.8.74) it is easy to show that

$$\left\|\Im_N^{-1} G_N \ln\left[1 + \frac{a(z) - a_N(z)}{a_N(z)}\right]\right\|_1 = o(N^{-m}).$$

Since $\|b(z) - b_N(z)\|_W = o(N^{-m})$, from (2.8.73) we come to the validity of formula (2.8.79).

The similar reasoning can be used to prove the following two lemmas.

**Lemma 2.8.29** *Let a vector $a^+ \in P^+ l_1^m$. Then, there exists a vector $b^+ \in P^+ l_1^m$ such that* $\exp(\Im_N P_N a^+) = \Im_N P_N b^+ + \Im_N \varepsilon^{(N)}$*, where $\varepsilon^{(N)} \in E_N$ and $\left\|\varepsilon^{(N)}\right\|_1 = o(N^{-m})$.*

**Lemma 2.8.30** *Let a vector $a^- \in P^- l_1^m$. Then, there exists a vector $b^- \in P^- l_1^m$ such that* $\exp(\Im_N P_N a^-) = 1 + \Im_N P_N b^- + \Im_N \varepsilon^{(N)}$*, where $\varepsilon^{(N)} \in E_N$ and $\left\|\varepsilon^{(N)}\right\|_1 = o(N^{-m})$.*

With the aid of Lemmas 2.8.28–2.8.30, it is easy to show that the following lemma is valid.

**Lemma 2.8.31** *Let a vector $a$ be given such that $a \in l_1^m : a(z) \neq 0,\ |z| = 1,\ \kappa_a = 0$. Then, for a sufficiently large natural $N$ there exist vectors $b^+ \in P^+ l_1^m$, $c^+ \in P^+ l_1^m$, and $b^- \in P^- l_1^m$ such that the representation*

$$\Im_N P_N a = \Im_N \tilde{a}_+^{(N)} \left(1 + \Im_N \tilde{b}_-^{(N)}\right)^{-1} \tag{2.8.81}$$

*is valid, where*

$$\Im_N \tilde{a}_+^{(N)} = \exp\left[\Im_N P^+ \Im_N^{-1} \ln(\Im_N P_N a)\right] = \Im_N P_N b^+ + \Im_N \varepsilon_1^{(N)},$$

$$\left[\Im_N \tilde{a}_+^{(N)}\right]^{-1} = \exp\left[-\Im_N P^+ \Im_N^{-1} \ln(\Im_N P_N a)\right] = \Im_N P_N c^+ + \Im_N \varepsilon_2^{(N)},$$

$$\left(1 + \Im_N \tilde{b}_-^{(N)}\right)^{-1} = \exp\left[-\Im_N P^- \Im_-^{-1} \ln(\Im_N P_N a)\right] = 1 + \Im_N P_N b^- + \Im_N \varepsilon_3^{(N)}.$$

*Here the vectors $\varepsilon_i^{(N)} \in E_N$, $\left\|\varepsilon_i^{(N)}\right\|_1 = o(N^{-m})$, $i = 1, 2, 3$.*

**4.** Now we describe an algorithm for solving equation (2.8.72). Let $a \in l_1^m$ be a vector, its symbol $a(z) \neq 0$, $|z| = 1$, and its index $\kappa_a = 0$. Let us consider the equation

$$a^{(N)} \otimes x_+^{(N)} = b^{(N)} + y_-^{(N)}. \tag{2.8.82}$$

Here $b_+^{(N)} = P_N b^+$ and $a^{(N)} = P_N a$ are given, $x_+^{(N)}$ and $y_-^{(N)}$ are the sought-for vectors. It is easy to see that $x_+^{(N)}$ is a solution to equation (2.8.75).

By Theorem 2.8.55 and Lemma 2.8.26 we can conclude that the sequence $\left\{y_-^{(N)}\right\}_1^\infty$ converges in norm to some vector $y$ in the space $E^m$. Applying the discrete Fourier transform to (2.8.82) and denoting the images of the vectors $a^{(N)}$, $x_+^{(N)}$, $b_+^{(N)}$, $y_-^{(N)}$ by $A^{(N)}$, $X_+^{(N)}$, $B_+^{(N)}$, $Y_-^{(N)}$, respectively, we obtain

$$A^{(N)} X_+^{(N)} = B_+^{(N)} + Y_-^{(N)}. \tag{2.8.83}$$

If the vector $A^{(N)}$ is represented in the form (2.8.81), then by Lemma 2.8.31 the equation (2.8.83) can be rewritten in the form

$$\Im_N \tilde{a}_+^{(N)} X_+^{(N)} = \left[1 + \Im_N \tilde{b}_-^{(N)}\right] B_+^{(N)} + \left[1 + \Im_N \tilde{b}_-^{(N)}\right] Y_-^{(N)}.$$

Let us apply the inverse Fourier transform and the projector $P^+$ to both sides of this equality followed by the application of the direct Fourier transform. Then, we get

$$\Im_N \tilde{a}_+^{(N)} X_+^{(N)} = \Im_N P^+ \Im_N^{-1} \left(\left[1 + \Im_N \tilde{b}_-^{(N)}\right] B_+^{(N)}\right) + \Im_N \varepsilon^{(N)}. \qquad (2.8.84)$$

Since $x_+^{(N)}$, $b_+^{(N)}$, $y_-^{(N)}$ converge in $E^m$ and $\Im_N \tilde{a}_+^{(N)}$, $1 + \Im_N \tilde{b}_-^{(N)}$ have the representations given in Lemma 2.8.31, from Lemma 2.8.27 we obtain $\left\|\varepsilon^{(N)}\right\|_E = o(N^{-m})$.

Let us rewrite (2.8.84) in the form

$$X_+^{(N)} = \left[\Im_N \tilde{a}_+^{(N)}\right]^{-1} \Im_N P^+ \Im_N^{-1} \left(\left[1 + \Im_N \tilde{b}_-^{(N)}\right] B_+^{(N)}\right) + \Im_N \varepsilon_1^{(N)}, \quad (2.8.85)$$

where $\varepsilon_1^{(N)} = \Im_N^{-1} \left(\left[\Im_N \tilde{a}_+^{(N)}\right]^{-1} \Im_N \varepsilon^{(N)}\right)$. Since $\left\|\varepsilon^{(N)}\right\|_E = o(N^{-m})$, on the basis of Lemma 2.8.31 and the estimate (2.8.78) we conclude that $\left\|\varepsilon_1^{(N)}\right\|_E = o(N^{-m})$. Applying the inverse discrete Fourier transform to equality (2.8.85), we obtain $x_+^{(N)}$, which is a solution to equation (2.8.75).

As an approximate solution to equation (2.8.75), we adopt

$$\tilde{x}_+^{(N)} = \Im_N^{-1} \left\{\left[\Im_N \tilde{a}_+^{(N)}\right]^{-1} \Im_N P^+ \Im_N^{-1} \left(\left[1 + \Im_N \tilde{b}_-^{(N)}\right] B_+^{(N)}\right)\right\}. \qquad (2.8.86)$$

Thus, the following main theorem is proved.

**Theorem 2.8.56** *Let in equation* (2.8.72) *a vector* $a \in l_1^m : a(z) \neq 0,\ |z| = 1$, $\kappa_a = \operatorname{ind} a(z) = 0$, *and a vector* $b \in E^m$. *Then, for a sufficiently large* $N$ *the approximate solution* (2.8.86) *can be found in* $O(N \log_2 N)$ *operations and the following accuracy estimate is valid:* $\left\|x_+^{(N)} - \tilde{x}_+^{(N)}\right\|_E = o(N^{-m})$.

Note in conclusion that a number of numerical experiments were performed to study the efficiency of software implementation of the above algorithm for the case of the Wiener–Hopf integral equation of the second kind (see [199]):

$$u(t) + \int_0^\infty K(t-s) u(s)\, ds = f(t), \quad 0 \le t < \infty.$$

This numerical algorithm is implemented in FORTRAN. The following model problems were solved:

1) $K(t) = 4\exp(-|t|)$, $f(t) = \exp(-t)$, the exact solution is $u_{\mathrm{ex}}(t) = 0.5\exp(-3t)$, $0 \le t < \infty$.

2) $K(t) = \exp(-|t|)$, $f(t) = 0.2\exp(-t)(7\cos t + 4\sin t)$, the exact solution is $u_{\mathrm{ex}}(t) = \exp(-t)\cos t$, $0 \le t < \infty$.

The Wiener–Hopf integral equation was replaced by the following Wiener–Hopf discrete equation:

$$u(t_j) + \triangle t \sum_{i=1}^{\infty} K(t_j - t_i)\, u(t_j) = f(t_j), \quad t_j = j\triangle t, \quad j = 0, 1, \dots .$$

**Table 2.8.1. Problem 1**

| | $t_0$ | $t_{10}$ | $t_{20}$ | $t_{30}$ | $t_{40}$ | $t_{50}$ | $t_{60}$ |
|---|---|---|---|---|---|---|---|
| $u_{\mathrm{ex}}(t)$ | 0.5000 | 0.2951 | 0.1642 | 0.0914 | 0.0509 | 0.0283 | 0.0158 |
| $u_{\mathrm{ap}}(t)$ | 0.4998 | 0.2950 | 0.1642 | 0.0914 | 0.0509 | 0.0283 | 0.0158 |

**Table 2.8.2. Problem 2**

| | $t_0$ | $t_{10}$ | $t_{20}$ | $t_{30}$ | $t_{40}$ | $t_{50}$ | $t_{60}$ |
|---|---|---|---|---|---|---|---|
| $u_{\mathrm{ex}}(t)$ | 1.0000 | 0.8259 | 0.6430 | 0.4789 | 0.3378 | 0.2212 | 0.1284 |
| $u_{\mathrm{ap}}(t)$ | 0.9988 | 0.8259 | 0.6430 | 0.4789 | 0.3378 | 0.2212 | 0.1284 |

Quite satisfactory numerical results (see Table 2.8.1 for the first problem and Table 2.8.2 for the second one) were obtained. The deviation of the approximate solution $u_{\mathrm{ap}}$ from the exact one $u_{\mathrm{ex}}$ in the discrete norms $L_2$ and $C$ for $N = 512$ and $\triangle t = 0.02$ is approximately equal to $4 \cdot 10^{-5}$ and $10^{-4}$ for the first problem and to $2 \cdot 10^{-5}$ and $2 \cdot 10^{-3}$ for the second one. The computing time on a BESM-6 computer (without compilation) was equal to 4.78 seconds for the first problem and 5.10 seconds for the second one.

Note that it takes 1 minute and 27 seconds to solve a similar problem reduced to a linear system with a symmetric Toeplitz matrix of order 1000 (see [200]); hence, saving of computing time is more than eight times. Note also that the amount of storage is proportional to $N$, i.e., our algorithm is intended for processing large amounts of data.

# Chapter 3

# THE SPLINE APPROXIMATION METHOD AND ITS APPLICATIONS

## 3.1 The collocation method for the first-kind linear operator equations

**1. Formulation of the collocation method.** Let $G$ be a Hilbert space, $X$, $Y$, and $F$ be Banach spaces, $Y \subset X \subseteq F \subset G$. A linear operator $A$ acting uniquely and completely continuously from $G$ into $G$ is given. Let linear independent bounded functionals $l_i$, $i = 1, \ldots, N$, be given in $F$. We consider the problem of approximate solution of the operator equation

$$Au = f, \quad f \in F, \tag{3.1.1}$$

when $l_i\,(f)$, $i = 1, \ldots, N$, are known. An approximation to the exact solution is estimated in the norm of $X$. It is assumed that there exists the inverse operator $A^{-1}$ bounded from $Y$ into $X$, but unbounded from $F$ into $X$; therefore, the problem (3.1.1) is unstable.

In $G$ we choose $N$ linearly independent elements $\psi_i$, $i = 1, \ldots, N$, whose linear envelope forms a subspace $H_N \subset G$. An approximate solution to equation (3.1.1) is sought in the form

$$g_N = \sum_{i=1}^{N} c_i \psi_i, \tag{3.1.2}$$

where $c_i$ are unknown coefficients.

By the collocation method we mean a method for finding an approximate solution to equation (3.1.1) in the form (3.1.2) in which the coefficients $c_i$ are determined from the conditions

$$l_i(A^\tau g_N) = l_i(f), \quad i = 1, \dots, N, \tag{3.1.3}$$

where the linear operator $A^\tau$ is a chosen approximation of the operator $A$, $A^\tau : X \to Y$. For example, if $A$ is an integral operator, then $A^\tau$ is the operator with an approximate kernel. Below, in the process of substantiation of the collocation method we impose on the operator $A^\tau$ some additional approximation conditions. Since the operator $A^\tau$ and the functionals $l_i$ are linear, the conditions (3.1.3) lead to the system of $N$ linear algebraic equations

$$A_N \hat{c} = \hat{f}, \tag{3.1.4}$$

where $\hat{c} = (c_1, \dots, c_N)$ is the sought-for vector, $A_N = (a_{ij})$, $a_{ij} = l_i(A^\tau \psi_j)$, $i, j = 1, \dots, N$, is the matrix of the system, and $\hat{f} = (l_1(f), \dots, l_N(f))^T$ is the right-hand side.

The components of the vector $\hat{c}$ (i.e., the components of the solution to (3.1.4)) are denoted by $\bar{c}_i$; the corresponding approximate solution to (3.1.1) is represented in the form $\bar{g}_{N,\tau} = \sum_{i=1}^N \bar{c}_i \psi_i$.

Let us mention the important case when $H_N$ is a subspace of splines; in order to define these splines, a Hilbert space $U \subset G$ and a linear operator $L : U \to G$ should be chosen. It is assumed that $l_i$, $i = 1, \dots, N$, are defined on $U$.

An element $v_N \in U_N$ is said to be the interpolating spline for a given $v \in F$ if $v_N$ is a solution to the problem

$$\|Lv_N\|_G = \inf_{g \in U(v)} \|Lg\|_G, \quad U(v) = \{g \in U : \mathcal{L}g = \mathcal{L}v\}, \tag{3.1.5}$$

where $\mathcal{L}g \equiv \{l_1(g), \dots, l_N(g)\}$, $g \in U$. It is assumed that the conditions guaranteeing the existence and the uniqueness of the interpolating spline (see [172]) are fulfilled.

Let us introduce the pseudoscalar product

$$(u, v)_{G_L} \equiv (Lu, Lv)_G$$

and the basic scalar product of the spline theory

$$(u, v)_N \equiv \sum_{i=1}^N l_i(u) l_i(v) + (u, v)_{G_L}, \quad u, v \in U.$$

It is shown in [172] that the set of all splines forms an $N$-dimensional linear space $U_N \subset U$; this space is complete with respect to the norm $\|u\|_N \equiv (u,u)_N^{1/2}$ and, hence, is a Hilbert space. The spline $v_N$ interpolating an element $v$ has an extremal property, which can be written down in the form of the condition for orthogonality [172] of the element $(v_N - v)$ to any element $r_N \in U_N$:

$$(r_N, v_N - v)_N = 0. \tag{3.1.6}$$

For the case of interpolation, this condition can be rewritten as

$$(r_N, v_N - v)_{G_L} = 0. \tag{3.1.7}$$

The basis in the space $U_N$ is denoted by $\varphi_1, \dots, \varphi_N$. Let $P_N$ be the spline interpolation operator and $B = P_N A^\tau P_N$. Then, the system (3.1.3) can be rewritten in operator form as follows:

$$Bg_N = P_N f, \qquad g_N \in U_N.$$

The existence and uniqueness of the solution obtained by the collocation method is equivalent to the existence of the operator $B^{-1}$.

**2. Substantiation of the collocation method. The existence and uniqueness.** The substantiation of the collocation method consists in: 1) the study of conditions imposed on the operators $A^\tau$ and $L$, on the functionals $l_i$, $i = 1, \dots, N$, and on the elements $\psi_i$, $i = 1, \dots, N$, for which the system (3.1.4) is uniquely solvable and 2) in finding *a priori* estimates for approximation errors.

It is further assumed that the element $A^\tau g_N \in U$ for $g_N \in H_N$ and $f \in U$. Let $A_N$ be the matrix with the elements $a_{ij} = \left(A^\tau \psi_j, \varphi_i\right)_N$ and $\left((f,\varphi_1)_N, \dots, (f,\varphi_N)_N\right)^T$ be the column vector $\hat{f}_N$. It is assumed that the operator $A^\tau$ is positive and self-adjoint in $U_N$. In this case, the system of linear algebraic equations for finding the coefficients $\{c_i\}$ of the approximate solution (3.1.2) by the Galerkin method with respect to the scalar product $(\cdot,\cdot)_N$ has the following form:

$$A_N \hat{c} = \hat{f}_N. \tag{3.1.8}$$

Here $\hat{c} = (c_1, \dots, c_N)^T$.

**Theorem 3.1.57** *Under the above assumptions, any solution to the system* (3.1.4) *obtained by the collocation method is also a solution to the system* (3.1.8) *obtained by the Galerkin method.*

**Proof.** Choosing the element $v = A^\tau g_N$, we can rewrite the condition (3.1.6) in the form

$$(A^\tau g_N, r_N)_N = (f, r_N)_N$$

if the coefficients $\{c_i\}$ in the representation of $g_N$ satisfy the collocation conditions (3.1.3), i.e., they are the components of the solution to the system (3.1.4). As $r_N$ we choose the elements of the basis $\varphi_1, \dots, \varphi_N$ and obtain the linear system (3.1.8) of the Galerkin method, which proves the theorem.

In the above case, thus, it is sufficient to study the system (3.1.8) in order to establish the existence and uniqueness of the solution obtained by the collocation method.

Let us consider the spline collocation method for the case when $\psi_i = \varphi_i$, $i = 1, \dots, N$, i.e., $H_N = U_N$.

**Theorem 3.1.58** *Let the operator $A^\tau$ satisfy the conditions*

$$1) \quad (LA^\tau g, Lv)_G = (A^\tau Lg, Lv)_G \quad \textit{for all} \quad g, v \in U_N; \qquad (3.1.9)$$

$$2) \quad A^\tau = (A^\tau)^* > 0 \quad \textit{in} \quad G; \qquad (3.1.10)$$

$$3) \quad \textit{if} \quad g \in \ker L \cap U_N \quad \textit{and} \quad P_N A^\tau g = 0, \quad \textit{then} \quad g \equiv 0. \qquad (3.1.11)$$

*Then, the solution obtained by the spline collocation method exists and is unique.*

**Proof.** Let us consider the case $f \equiv 0$. Let $\bar{g}_{N,\tau} \in U_N$ and $\bar{g}_{N,\tau}$ be the solution obtained by the collocation method. From the relation (3.1.7) for $v = A^\tau \bar{g}_{N,\tau}$, $r_N = \bar{g}_{N,\tau}$, and $v_N = 0$ we obtain

$$(A^\tau \bar{g}_{N,\tau}, \bar{g}_{N,\tau})_{G_L} = 0. \qquad (3.1.12)$$

From the relations (3.1.9) and (3.1.12) it follows that

$$(A^\tau L\bar{g}_{N,\tau}, L\bar{g}_{N,\tau})_G = 0.$$

Then, from the condition (3.1.10) we conclude that $\bar{g}_{N,\tau} \in \ker L$. From the condition (3.1.11) it follows that $\bar{g}_{N,\tau} \equiv 0$. By virtue of linear independence of the elements $\{\varphi_j\}$ this means that $c_j = 0$, $j = 1, \dots, N$.

Thus, the linear system (3.1.4) has only the trivial solution when its right-hand side is equal to zero; this is possible if and only if $\det A_N \neq 0$. Hence, the system (3.1.4) is uniquely solvable for any right-hand side. The theorem is proved.

Let the operator $A$ be represented in the form $A = A_0 + A_1$. Let us introduce the following notation: $A_1^\tau$ is a linear operator from $X$ into $Y$ (this operator approximates the operator $A_1$), $A^\tau = A_0 + A_1^\tau$, $B = P_N A^\tau P_N$,

$B_0 = P_N A_0 P_N$, $B_1 = P_N A_1^\tau P_N$, $B_1^0 = P_N A_1 P_N$, $B = B_0 + B_1$, and $B^0 = B_0 + B_1^0$.

Let $ZU_N$ be a set $U_N$ with the norm of a space $Z$ and $E$ be the unit operator.

**Theorem 3.1.59** *Let the following conditions be fulfilled:*

1) *there exist the operators $A^{-1}$ and $A_0^{-1}$ bounded from $Y$ into $X$,*

2) *the operator $A_1$ is completely continuous from $X$ into $Y$,*

3) *there exists the operator $B_0^{-1}$ bounded from $YU_N$ into $XU_N$ uniformly in $N$,*

4) $\|A_1 - A_1^\tau\|_{X\to Y} \to 0$ *as* $\tau \to 0$,

5) $\|P_N g - g\|_Z \to 0$ *as* $N \to \infty$ *for any fixed* $g \in Z$, $Z = X$ *or* $Z = Y$.

*Then, the operator $B^{-1}$ exists for a sufficiently large $N$ and a sufficiently small $\tau$.*

**Proof.** Let us represent the operator $B^0$ in the form

$$B^0 = B_0 \left(E + B_0^{-1} B_1^0\right).$$

Using the idea of compact approximation, we show that there exists the operator $\left(E + B_0^{-1} B_1^0\right)^{-1}$ for a sufficiently large $N$. To accomplish this, it is sufficient to show that the operator $T_N = B_0^{-1} B_1^0$ is completely continuous from $X$ into $X$ and the sequence $T_N$ is compactly convergent to the operator $T = A_0^{-1} A_1$ with respect to the connecting operators $P_N$.

The complete continuity of $T_N$ follows from the fact that $B_0^{-1}$ is bounded uniformly in $N$ from $YU_N$ into $XU_N$ and the operator $B_1^0$ can be represented by the product of the operators $P_N$ and $A_1$, where $P_N$ is bounded uniformly in $N$ from $X$ into $X$ and from $Y$ into $Y$ and $A_1$ is completely continuous from $X$ into $Y$.

The compact convergence $T_N$ to $A_0^{-1} A_1$ with respect to $P_N$ follows from the fact that by virtue of the complete continuity $A_1$ from $X$ into $Y$ and the boundedness of $P_N$, the sequence $\{g_N\}$ being bounded in $X$ goes into the sequence $\{T_N g_N\}$ compact in $Y$ and from the fact that for any $g \in X$ the following relation is fulfilled:

$$\|T_N P_N g - P_N T g\|_X \to 0, \quad N \to \infty.$$

Now we conclude that for a sufficiently large $N$ there exists the operator $\left(E + B_0^{-1} B_1^0\right)^{-1}$ bounded uniformly in $N$ from $XU_N$ into $YU_N$. Hence, for the same $N$ there exists the operator $\left(B^0\right)^{-1}$ bounded from $YU_N$ into $XU_N$

uniformly in $N$. According to the conditions 4) and 5) of this theorem, the following relation takes place:

$$\left\|B - B^0\right\|_{XU_N \to YU_N} \to 0, \quad \tau \to 0;$$

therefore, the operator $B^{-1}$ exists for a sufficiently large $N$ and a sufficiently small $\tau$. The theorem is proved.

**3. An approximation error estimate for the collocation method.** Let $\bar{u}$ be the exact solution to equation (3.1.1) with the operator $A$ and the right-hand side $f$, $\hat{g}_N$ be the spline interpolating the element $\bar{u}$, and $\bar{g}_N$ be the solution obtained by the collocation method for $A^\tau = A$.

**Theorem 3.1.60** *Let the following conditions be fulfilled:*

1) $A = A^* > 0$ *in* $G$;
2) $LAg = ALg$ *for all* $g \in U$;
3) *there exists a constant* $c \geq 0$ *and a quantity* $\rho_N > 0$ *such that* $\|g\|_Y \leq c\,\|\mathcal{L}g\|_{R_N} + \rho_N\,\|Lg\|_G$ *for all* $g \in U$;
4) $\bar{u} \in U$.

*Then, the following estimate is valid:*

$$\|\bar{u} - \bar{g}_N\|_X \leq \gamma\rho_N\,\|\bar{u} - \hat{g}_N\|_{G_L}^{1/2}\,\|A\left(\bar{u} - \hat{g}_N\right)\|_{G_L}^{1/2}, \tag{3.1.13}$$

*where* $\gamma = \left\|A^{-1}\right\|_{Y \to X}\|A\|_{G \to G}^{1/2}$.

**Proof.** The orthogonality conditions for the spline $\bar{g}_N$ determined from the collocation conditions mean that under the above assumptions, $\bar{g}_N$ is a solution to the extremum problem

$$\left(A\left(\bar{u} - \bar{g}_N\right), \bar{u} - \bar{g}_N\right)_{G_L} = \min_{g \in U_N}\left(A\left(\bar{u} - g\right), \bar{u} - g\right)_G.$$

Hence,

$$\left(A\left(\bar{u} - \bar{g}_N\right), \bar{u} - \bar{g}_N\right)_{G_L} \leq \left(AL\left(\bar{u} - \hat{g}_N\right), L\left(\bar{u} - \hat{g}_N\right)\right)_G.$$

The inequalities

$$\begin{gathered}
\|\bar{u} - \bar{g}_N\|_X \leq \left\|A^{-1}\right\|_{Y \to X}\|A\left(\bar{u} - \bar{g}_N\right)\|_Y \leq \\
\leq \left\|A^{-1}\right\|_{Y \to X}\rho_N\,\|AL\left(\bar{u} - \bar{g}_N\right)\|_G \leq \\
\leq \rho_N\left\|A^{-1}\right\|_{X \to Y}\|A\|_{G \to G}^{1/2}\left(A\left(\bar{u} - \bar{g}_N\right), \bar{u} - \bar{g}_N\right)_{G_L}^{1/2} \leq \\
\leq \rho_N\,\|A\|_{G \to G}^{1/2}\left\|A^{-1}\right\|_{Y \to X}\left(AL\left(\bar{u} - \hat{g}_N\right), L\left(\bar{u} - \hat{g}_N\right)\right)_G^{1/2}
\end{gathered}$$

are fulfilled; the estimate (3.1.13) follows from these inequalities. The theorem is proved.

It is assumed that the operator $A$ and the right-hand side $f$ contain errors. The exact operator is denoted by $\bar{A}$ and the exact right-hand side is denoted by $\bar{f}$. An approximation of the right-hand side used instead of $f$ in the collocation method we denote by $\varphi$. Let $\tilde{g}_N$ be a solution to the problem

$$\|\tilde{g}_N - \bar{u}_0\|_{Z_1} = \inf_{g \in U_N} \|g - \bar{u}_0\|_{Z_1},$$

where $\bar{u}_0$ is the exact solution to the equation $\bar{A}u = \bar{f}$.

**Theorem 3.1.61** *Let the operator $B^{-1}$ exist and the spaces $Z_1$ and $Z_2$ be such that the quantities*

$$\nu_{N,\tau} = \left\|B^{-1}\right\|_{Z_2 U_N \to Z_1 H_N}, \quad \mu_{N,\tau} = \nu_{N,\tau} \left\|P_N A^\tau\right\|_{Z_1 \to Z_2 U_N}$$

*are finite. Then the following estimate is valid:*

$$\begin{aligned} \|\bar{u}_0 - \bar{g}_{N,\tau}\|_{Z_1} \leq \left(1 + \mu_{N,\tau}\right) \|\bar{u}_0 - \tilde{g}_N\|_{Z_1} + \\ + \nu_{N,\tau} \left\{ \left\|P_N \left(A^\tau - \bar{A}\right) \bar{u}_0\right\|_{Z_2 U_N} + \left\|P_N \left(\bar{f} - \varphi\right)\right\|_{Z_2 U_N} \right\}. \end{aligned} \tag{3.1.14}$$

From Theorem 3.1.61 it follows that if

$$\begin{aligned} \mu_{N,\tau} \|\bar{u}_0 - \tilde{g}_N\|_{Z_1} + \nu_{N,\tau} \Big\{ \left\|P_N \left(A^\tau - \bar{A}\right) \bar{u}_0\right\|_{Z_2 U_N} + \\ + \left\|P_N \left(\bar{f} - \varphi\right)\right\|_{Z_2 U_N} \Big\} \to 0 \end{aligned} \tag{3.1.15}$$

as $N \to \infty$, $\tau \to 0$, then the collocation method converges in $Z_1$. In general, when $\bar{u}$ is not sufficiently "smooth" or the operator $\bar{A}$ and the right-hand side $\bar{f}$ are not adequately approximated, the relation (3.1.15) may not be fulfilled; therefore, the collocation method may not be convergent. The reduction of "smoothness" for $\bar{u}$ or the reduction of quality in the approximation of the operator and the right-hand side may be caused by errors in $A$ and $f$. In this case, the regularization of the collocation method is necessary, i.e., a special choice of the approximations $A^\tau$ and $\varphi$ is required. In Section 3.2 the regularization by the sequential smoothing method is considered.

## 3.2 The sequential smoothing method for the first-kind linear operator equations

**1. Formulation of the method.** Let the Banach spaces $X$, $Y$, and $F$ be such that $X \subset Y \subseteq F$. It is assumed that for equation (3.1.1) there exists an inverse operator $A^{-1}$ unbounded from $F$ into $X$, but bounded from $Y$ into $X$. The right-hand side $f$ is assumed to be represented in the form $f = \bar{f}+\delta$, where $\bar{f} \in Y$ is an exact right-hand side, $\delta \in F$ is an error, and $\|\delta\|_F \leq \delta_0$. Thus, the problem (3.1.1) may be ill posed (unstable) [172]. From the estimate (3.1.14) of Theorem 3.1.61 it follows that in general the spline collocation method is not stable. Here we propose an approach to the regularization of the collocation method for the problem (3.1.1) on the basis of construction of $\varphi$ by a special smoothing of the right-hand side $f$.

Let $\mathcal{L}$ be an operator from $F$ into $R_N$:

$$\mathcal{L}g \equiv (l_1(g), \ldots, l_N(g))^T, \quad g \in F,$$

where $l_i$, $i = 1, \ldots, N$, are given functionals. Let $P$ be a known operator such that for any $g \in F$ this operator associates an element $\mathcal{L}g$ with an approximating spline $P\mathcal{L}g \in U_N$. Here $U_N$ is the space of splines (the linear envelope of elements $\varphi_i$, $i = 1, \ldots, N$) introduced in Section 3.1.

Let $V = \mathcal{L}P$. Further we assume that $\|P\|_{R_N \to FU_N} < C = \text{const}$, $C$ does not depend on $N$, and the spectral norm $\|V\| \leq 1$. Let $V^K$ be the product of $K$ operators $V$, where $K > 0$ is integer.

The calculation of the element $y_K$ by the formula

$$y_K = PV^{K-1}\mathcal{L}f, \quad y_0 = f, \tag{3.2.16}$$

is said to be the sequential smoothing method for an elements $f \in F$.

The spline approximation method we propose for approximate solution of the problem (3.1.1) consists in the combination of the collocation method and the sequential smoothing method, i.e., in the calculation of the element

$$u_N^K = B^{(-1)}y_K, \tag{3.2.17}$$

where $B^{(-1)} = (P_N A P_N)^{-1} \equiv B^{-1}$ or $B^{(-1)} = P\mathcal{L}A^{-1}$ if $A^{-1}$ is simple for computation. The number $K$ of smoothings is chosen depending on an error in input data and on the discretization parameter $N$. Thus, $\varphi = y_K$.

Let $\bar{u}$ be the exact solution to equation (3.1.1) with an exact right-hand side $f = \bar{f}$. Let $\delta_0 \geq \|\delta\|_F$ and

$$R(N, \delta_0, K, f) \equiv \left\| \bar{u} - B^{(-1)}y_K \right\|_X .$$

The method $B^{(-1)}y_K$ is said to be convergent to $X$ for a given $f$ : $\left\|f - \bar{f}\right\|_F \leq \delta_0$ if the number $K = K(N, \delta_0)$ of smoothings depends on $N$ and $\delta_0$ in such a way that

$$\lim_{N\to\infty, \delta_0\to 0} R(N, \delta_0, K(N, \delta_0), f) = 0. \tag{3.2.18}$$

The method $B^{(-1)}y_K$ is said to be regularizing in $X$ if the number of smoothings depends on $N$, $\delta_0$, and $K = K(N, \delta_0)$ in such a way that

$$\lim_{N\to\infty,\delta_0\to 0} \sup_{f:\left\|f-\bar{f}\right\|_F\le\delta_0} R\,(N, \delta_0, K(N, \delta_0), f) = 0. \tag{3.2.19}$$

If (3.2.18) or (3.2.19) is valid only for some dependence on $N$ and $\delta_0$, we call the method $B^{(-1)}y_K$ dependently convergent or dependently regularizing in $X$, respectively; the double limit is reduced to the single one.

Note also that the difference methods of numerical differentiation and the method of self-regularization for the first-kind integral equations with a logarithmic singularity in kernels are dependently regularizing for $X = F = C\,[a, b]$.

Let $\delta^0 \ge 0$ be a given error estimate such that

$$\left(\frac{1}{N}\sum_{i=1}^{N} l_i^2(\delta)\right)^{1/2} \le \delta^0,$$

and $\delta^0 \to 0$ as $\delta_0 \to 0$.

Further we assume that the method (3.2.17) converges as $N \to \infty$ when input data are exact, i.e.,

$$R_1(N) \equiv \left\|\bar{u} - B^{-1}P\mathcal{L}\bar{f}\right\|_X \to 0, \quad N \to \infty. \tag{3.2.20}$$

In what follows we assume that the eigenvalues of the matrix $V$ are real and satisfy the conditions

$$\begin{gathered} 1 = \lambda_1 = \lambda_2 = \cdots = \lambda_m > \lambda_{m+1} > \ldots > \lambda_N > 0, \\ 1 - \lambda_i = \mu_i, \quad 0 < \mu_0 < \mu_i < 1, \\ \mu_i = O((i/N)^m), \quad i = m+1, \ldots, N. \end{gathered} \tag{3.2.21}$$

It is also assumed that the matrix $V$ has eigenvectors $g_1, \ldots, g_N$ orthonormal with respect to the Euclidean norm in the finite-dimensional space $R_N$ and that $m$ linearly independent eigenvectors $g_1, \ldots, g_m$ correspond to $\lambda_i = 1$, $i = 1, \ldots, m$. Let $G_m$ be the linear envelope of the first $m$ eigenvectors $g_1, \ldots, g_m$ of the operator $V$. Any vector $v \in R_N$ can uniquely be represented in the form

$$v = \sum_{i=1}^{N} \hat{v}_i g_i.$$

By $v^{(m)} = \mathrm{Pr}_m\, v$ we denote the projection of $v$ onto $G_m$. The expansions of the elements $\mathcal{L}f$, $\mathcal{L}\bar{f}$, and $\mathcal{L}\delta$ in the basis $\{g_i\}_{i=1}^N$ are written down as follows:

$$\mathcal{L}f = \sum_{i=1}^N \hat{f}_i g_i, \quad \mathcal{L}\bar{f} = \sum_{i=1}^N \hat{\bar{f}}_i g_i, \quad \mathcal{L}\delta = \sum_{i=1}^N \hat{\delta}_i g_i.$$

Further we assume that the operator $B^{(-1)}$ acts in such a way that

$$\left\| B^{(-1)} P v \right\|_X^2 = O\left( \frac{1}{N} \sum_{i=1}^N i^{2q} \hat{v}_i^2 \right), \quad v \in R_N, \tag{3.2.22}$$

where $q \geq 0$ is an integer number.

**2. The choice of the number of smoothings by the residual principle**. The residual (discrepancy) principle is a widespread way for the choice of the regularization parameter [172]. Here we formulate similar residual principles for the choice of the number $K$ of smoothings in the sequential smoothing method.

Let $v = (v_1, \ldots, v_N) \in R_N$. We denote the Euclidian norm by $\|v\| = \left( \sum_{i=1}^N v_i^2 \right)^{1/2}$ and the uniform norm by $\|v\|_1 = \max_{1 \leq i \leq N} |v_i|$. It is assumed that one of the following error estimates is given: $\|\mathcal{L}\delta\| / \sqrt{N} \leq \delta^0$ or $\|\mathcal{L}\delta\|_1 \leq \delta^1$. Here $\delta^0$ is the error estimate in the root-mean-square norm and $\delta^1$ is the error estimate in the uniform norm. Depending on which estimate is given, the following two variants of formulation for the residual principle are possible.

*Variant 1.* Let us introduce the function

$$r_1(K, v) \equiv \frac{1}{N} \left\| v - V^K v \right\|^2, \tag{3.2.23}$$

where $K$ is a natural number.

The residual principle for the root-mean-square norm is called the choice of the minimal number $K_1 = K_1(N, \varepsilon)$ from all $K$ for which the inequality

$$r_1(K, \mathcal{L}f) \geq \varepsilon^2 \tag{3.2.24}$$

holds for a given number $\varepsilon > 0$.

*Variant 2.* Let us introduce the function

$$r_2(K, v) \equiv \left\| v - V^K v \right\|_1 \tag{3.2.25}$$

where $K$ is a natural number.

The residual principle for the uniform norm is called the choice of the maximal number $K_2$ from all $K$ for which the inequalities

$$r_2(K-1, \mathcal{L}f) \leq \varepsilon, \quad r_2(K, \mathcal{L}f) > \varepsilon \tag{3.2.26}$$

hold for a given number $\varepsilon > 0$.

In order to substantiate the possibility for choosing $K$ from the conditions (3.2.24) and (3.2.26), we study the behavior of the sequence $V^K v$, $v \in R_N$, and of the functions $r_1(K,v)$ and $r_2(K,v)$ in greater detail.

**Theorem 3.2.62** *Under the assumptions* (3.2.21), *for any fixed* $v \in R_N$ *the following relation is valid:*

$$\lim_{K\to\infty} V^K v = v^{(m)}.$$

**Proof.** Representing a vector $v$ in the form (3.2.22), we obtain

$$V^K v = \sum_{i=1}^{N} \hat{v}_i \lambda_i^K g_i = \sum_{i=1}^{m} \hat{v}_i g_i + \sum_{i=m+1}^{N} \lambda_i^K \hat{v}_i g_i. \tag{3.2.27}$$

In (3.2.27) we pass to the limit in $K$ and take into account that $\|\hat{v}_i\|$ and $\|g_i\|$ $(i = 1, \dots, N)$ are bounded uniformly in $K$. Since $\lambda_i < 1$, $i = m+1, \dots, N$, we obtain the conclusion of the theorem.

Let

$$r_\infty^1(v) = \left\| v - v^{(m)} \right\|^2, \quad r_\infty^2(v) = \left\| v - v^{(m)} \right\|_1.$$

**Theorem 3.2.63** *Under the above assumptions, the functions* $r_i(K,v)$, $i = 1,2$, *change from* $0$ *to* $r_\infty^i(v)$ *when* $K$ *changes from* $0$ *to* $\infty$; *the function* $r_1(K,v)$ *changes monotonically. When* $\mathcal{L}f \notin G_m$ *for all* $\varepsilon$, $0 < \varepsilon < r_\infty^i(\mathcal{L}f)$, *there exists a finite* $K_i$, $i = 1,2$, *satisfying the residual principle* (3.2.24) *or* (3.2.26), *respectively.*

**Proof.** The fact that the functions $r_i(K,v)$ vary in the above limits follows from their definition and Theorem 3.2.62. In order to prove that the function $r_1(K,v)$ is monotone, we expand the vector $v$ in the basis $\{g_i\}$. Then, we obtain the expression

$$r_1(K,v) = \frac{1}{N} \sum_{i=m+1}^{N} \hat{v}_i^2 \left(1 - \lambda_i^K\right)^2.$$

Since $0 < \lambda_i < 1$ for $i = m+1, \dots, N$, obviously $r_1(K,v) \le r_1(K+1,v)$ and the equality is possible only in the case when $v \in G_m$.

The existence of $K_1$ and $K_2$, which are the solutions to the problems (3.2.24) and (3.2.26), respectively, follows from the established limits for the change of $r_i(K,v)$ and the restrictions on $\varepsilon$.

The uniqueness of $K_1$ follows from the monotonicity of $r_1(K, \mathcal{L}f)$ and the uniqueness of $K_2$ follows from its definition. The theorem is proved.

The substantiation of regularizing and convergence properties for the spline approximation method when the number of smoothings is chosen by the residual principle is given in the next theorem.

**Theorem 3.2.64** *Let the following conditions be fulfilled:*

1) *there exist constants $\beta > 0$ and $c > 0$ independent of $N$ such that*

$$\left\| B^{-1} P v \right\|_X^2 \le c \left\{ \|v\|^2 / N + N^{\beta - 1} \|v\| \, \|(E - V)\, v\| \right\}, \qquad v \in R_N, \quad N > N_0; \tag{3.2.28}$$

2) $\bar{f} \in M_\beta = \left\{ g \in Y : N^{\beta - 1/2} \|(E - V)\, \mathcal{L} g\| = O(1), \; N \to \infty \right\};$ (3.2.29)

3) *a number $\varepsilon$ is chosen according to the rules:*

a) *for the root-mean-square residual in* (3.2.24)

$$\varepsilon = \varepsilon_1 = \delta^0; \tag{3.2.30}$$

b) *for the uniform residual in* (3.2.26)

$$\varepsilon = \hat{\varepsilon}_1 = \sqrt{N} \delta^1; \tag{3.2.31}$$

4) *for $\bar{f}$ the relation* (3.2.20) *holds.*

*Then, the spline approximation method for the problem* (3.2.16) *is regularizing in $X$ for the choice of the number of smoothings by the root-mean-square residual principle and is dependently regularizing in $X$ for the choice of the number of smoothings by the uniform residual principle if $\hat{\varepsilon}_1 \to 0$ as $N \to \infty$ and $\delta^1 \to 0$.*

**Proof.** First we consider the root-mean-square residual principle. Suppose $\delta \notin M_\beta$. Let us show that the element $\bar{v} = V^{K_1} \mathcal{L} f$ has the property

$$\|(E - V)\, \bar{v}\|^2 \le \left\|(E - V)\, \mathcal{L} \bar{f}\right\|^2. \tag{3.2.32}$$

Now we use the expansions of $\mathcal{L}f$, $\mathcal{L}\bar{f}$, and $\mathcal{L}\delta$ in the basis $\{g_i\}$. For the difference between the left-hand side and the right-hand side of the inequality (3.2.32) we obtain the relations

$$\begin{aligned} J &= \|(E - V)\, \bar{v}\|^2 - \left\|(E - V)\, \mathcal{L}\bar{f}\right\|^2 = \\ &= \sum_{i=m+1}^{N} (1 - \lambda_i)^2 \left( \hat{f}_i^2 \lambda_i^{2K_1} - \hat{\hat{f}}_i^2 \right) < - \sum_{i=m+1}^{N} \hat{f}_i^2 (1 - \lambda_i)^2 (1 - \lambda_i^{K_1})^2 + \\ &+ 2 \|(E - V)\, \mathcal{L} f\| \, \|(E - V)\, \mathcal{L}\delta\| - \|(E - V)\, \mathcal{L}\delta\|^2. \end{aligned}$$

Let us introduce quantities $\mu_1$ and $\mu_2 = \mu_2(K)$ such that

$$\begin{aligned} \mu_1^2 \|\mathcal{L}\delta\|^2 &= \sum_{i=m+1}^{N} (1-\lambda_i)^2 \hat{\delta}_i^2, \\ \mu_2^2(K) \left\|\left(E - V^K\right)\mathcal{L}f\right\|^2 &= \sum_{i=m+1}^{N} (1-\lambda_i)^2 (1-\lambda_i^K)^2 \hat{f}_i^2. \end{aligned}$$

By $p_i$ and $q_i(k)$ we denote the nonzero weighting coefficients at $(1-\lambda_i)^2$ in the expressions for $\mu_1^2$ and $\mu_2^2(K)$:

$$p_i = \hat{\delta}_i^2 / \|\mathcal{L}\delta\|^2, \quad q_i(K) = (1-\lambda_i^K)^2 \hat{f}_i^2 / \left\|\left(E - V^K\right)\mathcal{L}f\right\|^2.$$

In the case under consideration, from the conditions 1), 2), 4) of this theorem and the assumption (3.2.22) we conclude that the quantities $\left|\hat{f}_i/\hat{f}_{i+l}\right|$ and $\left|\hat{\delta}_i/\hat{\delta}_{i+l}\right|$ increase at the same rate in $l$ for those $i$, $l$ for which $\hat{\delta}_i, \hat{\delta}_{i+l}, \hat{f}_i, \hat{f}_{i+l} \neq 0$. Therefore, for all $K > 0$ and for all $i$, $j$ such that $p_i p_j q_i q_j \neq 0$ the quantity

$$\gamma_i = \frac{p_i q_i(K)}{p_j q_j(K)} \sim \frac{(1-\lambda_i^K)^2}{(1-\lambda_j^K)^2}, \quad N \to \infty,$$

increases with increasing $j$ for $i$ fixed. For those numbers $i$ for which $\hat{\delta}_i = 0$, the coefficients $q_i(K)$ grow when $i$ increases from $m+1$ to some $i = i_0$ and, then, decrease. For $i \leq i_0$, the order of magnitude for $q_i$ in $i$ is no less than that for $q_{N-i+1}$. Therefore, for all $K > 0$ the inequality $\mu_2(K) > \mu_1$ holds.

From the above relations for $J$, thus, we obtain the inequality

$$J \leq \mu_2^2(K) N \left\{-r_1(K_1, \mathcal{L}f) + \left(\delta^0\right)^2\right\} \tag{3.2.33}$$

for a sufficiently large $N$ and a fixed $\delta^0$. When (3.2.30) is fulfilled, from (3.2.33) it follows that $J < 0$ and the inequality (3.2.32) is valid. Hence, $P\bar{v} \in M_\beta$.

Now we consider the case $\delta \in M_\beta$. Since $\bar{f} \in M_\beta$, we have $f \in M_\beta$ and, by virtue of the property $\lambda_i \leq 1$, obtain the condition $P\bar{v} \in M_\beta$.

From the inequalities

$$\begin{aligned} &\left\|\bar{v} - \mathcal{L}\bar{f}\right\| / \sqrt{N} \leq \frac{1}{\sqrt{N}} \left\{\left\|\mathcal{L}f - V^{K_1 - 1}\mathcal{L}f\right\| + \right. \\ &\qquad + \left\|V^{K_1-1}\left(E - V\right)\mathcal{L}f\right\| + \left\|\mathcal{L}f - \mathcal{L}\bar{f}\right\| \leq 2\delta^0 + \left\|\left(E - V\right)\mathcal{L}f\right\| / \sqrt{N} \end{aligned}$$

and from the convergence of the method with $\delta_0 = 0$ and the condition (3.2.28) it follows the relation

$$\left\|\bar{u} - B^{-1} P V^{K_1} \mathcal{L}f\right\|_X \leq \left\|\bar{u} - B^{-1} P \mathcal{L}\bar{f}\right\|_X + \left\|B^{-1} P \left(\mathcal{L}\bar{f} - \bar{v}\right)\right\|_X \to 0$$

as $N \to \infty$ and $\delta_0 \to 0$, which means that the convergence is proved.

Now we pass to the case of uniform residual. Let us show that under the hypotheses of this theorem the inequality (3.2.32) is also fulfilled for $\bar{v} = V^{K_2}\mathcal{L}f$. The above relations for $J$ are valid. Using the condition $\mu_2(K) > \mu_1$ for $\delta \notin M_\beta$ and the inequality

$$r_2^2(K, \mathcal{L}f) < N r_1(K, \mathcal{L}f), \tag{3.2.34}$$

we conclude that for a sufficiently large $N$ and a fixed $\delta^1$ the inequality

$$J \leq -\mu_2^2 r_2^2(K_2, \mathcal{L}f) + N\left(\delta^1\right)^2 \mu_2$$

is fulfilled; if (3.2.31) holds, from this inequality it follows the validity of (3.2.32). The further reasoning repeats the above proof for the case of root-mean-square residual. The theorem is proved.

The computation speed of a software implementation for the sequential smoothing method substantially depends on the amount of operations required to compute the values of the operator $P$. In the next section we use a technique with low computing cost for constructing the operator $P$ on the basis of explicit approximations of functions by splines.

## 3.3 Explicit spline approximation of functions and experimental data processing algorithms

**1. The explicit spline approximation operator and its spectral properties.** Let the values of a $(b-a)$-periodic function $f(x)$, $f_i = f(x_i)$, $i = 1, \dots, N$, be given on a segment $[a, b]$ of the real axis at the points $\{x_i\}$: $a = x_1 < x_2 < \ldots < x_N = b$. Let $s_{i,2n-1}(x)$ be a local basis polynomial spline of degree $2n-1$ constructed at the points $x_i, \dots, x_{i+2n}$, $i = -2n+1, \dots, N-1$; the additional points $x_i$ for $j < 1$ and $j > N$ are chosen with taking account of periodicity of the function $f(x)$. The formulas for the local basis spline and some algorithms for their stable computation are given, for example, in [88], [84]. As the approximating operator $P$ introduced in Section 3.2, the operator of explicit spline approximation is chosen:

$$Pf \equiv S_{2n-1,p}(x, f) = \sum_{i=1}^{N} \left( \sum_{j=-p}^{p} c_{ij} f_{i+j} \right) s_{i-n,2n-1}(x). \tag{3.3.35}$$

Here the coefficients $c_{ij}$ are taken from the approximation condition irrespective of $f(x)$. In the simplest case for $p = 0$, we have $c_{ij} = 1$ for all $i$, $n$. In the case of cubic splines on a uniform grid with $p = 1$, the above formula takes the form

$$S_{3,p}(x, f) = \sum_{i=1}^{N} \left( -\frac{1}{6} f_{i-1} + \frac{4}{3} f_i - \frac{1}{6} f_{i+1} \right) s_{i-2,3}(x).$$

The corresponding formulas for splines of some other degrees are given in [84], [86]. Thus, in the method of Section 3.2 we put

$$\varphi_i = s_{i,2n-1}(x), \quad l_i(f) = \sum_{j=-p}^{p} c_{ij} f_{i+j}, \quad i = 1, \ldots, N.$$

In the case of a uniform grid $\{x_i\}$, the analysis of values $S_{2n-1,p}(x, f)$ at the point $x_i$ results in the following formula for the matrix $V$:

$$V = E + \sum_{l=p+1}^{n+p-1} d_l \Delta^{2l}.$$

Here $E$ is the unit matrix in $R_N$, $d_l$ are some constants, and the operator $\Delta^{2l}$ has the form $\Delta^{2l} = \operatorname{diag}(\Delta_1^{2l}, \ldots, \Delta_i^{2l}, \ldots, \Delta_N^{2l})$, where $\Delta_i^{2l}$ is the central finite difference of order $2l$ at the points $x_{i-l}, \ldots, x_{i+l}$. Further we assume that $a = 0$ and $b = 1$.

Let us find the eigenvalues and the eigenvectors of the matrix $V$, i.e., the nonzero solutions to the problem

$$\left( E + \sum_{l=p+l}^{n+p-1} d_l \Delta^{2l} \right) v = \lambda v, \tag{3.3.36}$$

where $v \neq 0$ and $v \in R_N$. Obviously, the eigenvalue $\lambda = 1$ corresponds only to those eigenvectors $v$ for which

$$(V - E)\, v = \sum_{l=p+l}^{n+p-1} d_l \Delta^{2l} v = 0. \tag{3.3.37}$$

Any vector $v = \mathcal{L}g$, where $g$ is a polynomial of degree not higher than $2p+1$, satisfies this equation. Since the approximating formula (3.3.35) is exact only for polynomials of degree not higher than $2p + 1$ and does not restore any other functions precisely, the equation (3.3.37) has no other solutions, except for the above ones. This means that $\lambda = 1$ is the $(2p + 2)$-fold eigenvalue of the operator $V$ and the $2p+2$ linearly independent eigenvectors correspond to it:

$$\bar{g}_1 = \begin{pmatrix} 1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}, \quad \bar{g}_2 = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{pmatrix}, \dots, \quad \bar{g}_{2p+2} = \begin{pmatrix} x_1^{2p+1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_N^{2p+1} \end{pmatrix}.$$

The other eigenvectors are sought in the form of the vector function $v = \sin \alpha x$, $\alpha \neq 0$. Substituting it into equation (3.3.36), we get

$$(1-\lambda)\sin \alpha x + \sum_{l=p+1}^{n+p-1} d_l \Delta^{2l} \sin \alpha x = 0.$$

Taking into account the equality

$$\Delta^{2l} \sin \alpha x = (-1)^l 2^l \sin \alpha x \, (1-\cos \alpha h)^l, \quad h = 1/(N-1),$$

we obtain

$$(1-\lambda)\sin \alpha x + \sin \alpha x \sum_{l=p+1}^{n+p-1} d_l \, (-2)^l \, (1-\cos \alpha h)^l = 0. \tag{3.3.38}$$

Since we search for the nontrivial solutions, we cancel $\sin \alpha x \neq 0$ out of this equation and come to the expression for $\lambda$:

$$\lambda = 1 + \sum_{l=p+l}^{n+p-1} d_l \, (-2)^l \, (1-\cos \alpha h)^l. \tag{3.3.39}$$

Now we find such values of the parameter $\alpha$ for which the function $\sin \alpha x$ would be periodic with period 1. To accomplish this, it is sufficient to solve the equation $\sin \alpha = 0$: $\alpha = \alpha_i = \pi i$, $i = 1, 2, \dots$. The linearly independent eigenvectors are obtained for $i = 1, 2, \dots, N-1$. For periodic functions we take into consideration only the $N$ eigenvectors

$$\bar{g}_1, \quad \bar{g}_i = \sin \alpha_{i-1} x, \quad i = 2, \dots, N.$$

From them we obtain the system $\{g_1, \dots, g_N\}$ orthonormal in $R_N$. Any vector $v = \mathcal{L}f$, where $f$ is a function periodic on $[0,1]$, can be represented uniquely in the form

$$v = \sum_{i=1}^{N} \hat{v}_i g_i. \tag{3.3.40}$$

In order to use the theoretical results discussed in Section 3.2, it is necessary to show that $0 < \lambda_i < 1$, $i = 2, \dots, N$. We check these properties for some specific $n$ and $p$.

Cubic splines, $n = 2$:

a) $p = 0: \ \lambda_i = 1 - \dfrac{1}{3}(1 - \cos(i-1)\pi h), \quad i = 2, \dots, N,$

$\lambda_2 = \dfrac{2}{3} + \dfrac{1}{3}\cos \pi h < 1, \quad \lambda_N = \dfrac{1}{3} > 0;$

b) $p = 1: \ \lambda_i = 1 - \dfrac{1}{9}(1 - \cos(i-1)\pi h)^2, \quad i = 2, \dots, N,$

$\lambda_2 = \dfrac{8}{9} + \dfrac{2}{9}\cos \pi h - \dfrac{1}{9}\cos^2 \pi h < 1, \quad \lambda_N = \dfrac{5}{9} > 0.$

The fifth-degree spline, $n = 3$:

$p = 0: \ \lambda_i = 1 - \dfrac{1}{2}[1 - \cos(i-1)\pi h] + \dfrac{1}{30}[1 - \cos(i-1)\pi h]^2, \ i = 2, \dots, N,$

$\lambda_2 = \dfrac{1}{2} + \dfrac{1}{2}\cos \pi h + \dfrac{1}{30}[1 - \cos \pi h]^2 < 1, \quad \lambda_N = \dfrac{2}{15} > 0.$

It is easy to check that the necessary properties are fulfilled for other values of $n$ and $p$.

**2. The sequential smoothing method for approximately-given functions of one and many variables.** Let us assume that $f_i = \bar{f}_i + \delta_i$, $i = 1, \dots, N$, are given, where $\bar{f}_i = \bar{f}(x_i)$ are the values of an exact function periodic on $[0, 1]$ and $\delta_i$ are errors. For a uniform grid, the sequential smoothing method consists in the evaluation of the function

$$F_{\bar{K}}(x) = P V^{\bar{K}} \hat{f}, \tag{3.3.41}$$

where $\hat{f} = (f_1, \dots, f_N)$, the additional values of $f_j$, $j < 1$, $j > N$, are chosen with taken account of periodicity of the exact function, the operator $P$ is defined by the formula (3.3.35), and the number $\bar{K}$ of smoothings is chosen by the residual principle.

Let us consider the numerical differentiation problem, which can be represented in the form of the operator equation (3.1.1) as follows:

$$Au \equiv \int_0^x u(t)\, dt = f(x) - f(0), \quad x \in [0, 1], \tag{3.3.42}$$

so that $A^{-1} = d/dx$ is the differentiation operator. We put $X = F = C[0, 1]$ and $Y = C^1[0, 1]$; then, the conditions (3.2.21) are valid for the above-written eigenvectors of the operator $V$. For the operator $B^{-1} = P\mathcal{L}A^{-1}$ the conditions (3.2.22) are fulfilled with $q = 1$. Let the estimate $\delta^1 = \max_{1 \le i \le N} |\delta_i|$ be known; then, it is also possible to estimate $\delta^0$:

$$\left(\frac{1}{N}\sum_{i=0}^{N}\delta_i^2\right)^{1/2} = \delta^0 \le \delta^1. \tag{3.3.43}$$

The substantiation of the residual principle is made more specific for the problem (3.3.42) in the next theorem.

**Theorem 3.3.65** *Let $\bar f \in W_2^{2r}[0,1]$, $r \ge 1$. Then, the sequential spline smoothing method on the basis of the formula* (3.3.35) *for $n \ge r$ and $p+1 = r$ is regularizing for the problem* (3.3.42) *in the space $W_2^{r-1}[0,1]$ when the number of smoothings is chosen according to the root-mean-square residual principle with $\varepsilon = \delta^1$ and is dependently regularizing in $W_2^{r-1}[0,1]$ when the number of smoothings is chosen according to the uniform residual principle with $\varepsilon = \hat\varepsilon_1 = \sqrt{N}\delta^1$ if $\hat\varepsilon_1 \to 0$ as $N \to \infty$ and $\delta^1 \to 0$.*

**Proof**. Let us check the fulfilment of the hypotheses of Theorem 3.2.64. Taking into account the above representation of the matrix $V$, we show that the conditions (3.2.28) and (3.2.31) are valid with $\beta = 2p+2$. The convergence condition with exact data follows from the results on convergence of the spline approximation method [84], [102]. Thus, all the hypotheses of Theorem 3.2.64 are fulfilled; hence the regularizing properties of the spline approximation method are valid. The theorem is proved.

Now we consider the case of a function $f(u) = f(u_1, \dots, u_M)$ of $M$ variables defined in the parallelepiped $\Omega_0 = \Pi_{i=1}^M \Omega_i$, where $\Omega_i = [0,1]$ are the unit segments. Along each axis $u_i$ we introduce a uniform grid $\{x_{ij}\}$, $i = 1, \dots, N_i$, with the step $h_i = x_{i,j+1} - x_{i,j}$, $N = \Pi_{i=1}^M N_i$. Let us use the explicit formulas for the spline approximation of functions of $M$ variables; each of these formulas is a superposition of one-dimensional formulas.

The operators $\mathcal{L}$, $P$, $V$ are defined in terms of the corresponding operators $\mathcal{L}_i$, $P_i$, $V_i$ for the $i$th variable by the formulas

$$V_i = \mathcal{L}_i P_i, \quad V = V_M \times \dots \times V_1, \quad P\mathcal{L}g = P_M \times \dots \times P_1 \mathcal{L}_M \times \dots \times \mathcal{L}_1 g,$$

where $g$ is a function defined in $\Omega_0$.

The sequential smoothing method consists in calculation of the element

$$F_{\bar K} = P V_M^{\bar K_M} V_{M-1}^{\bar K_{M-1}} \dots V_1^{\bar K_1} \mathcal{L}_M \times \dots \times \mathcal{L}_1 g, \tag{3.3.44}$$

where the number $\bar K = (\bar K_1, \dots, \bar K_M)$ of smoothings along different variables may be distinct. In the simplest case, $\bar K_i = \bar K$, $i = 1, \dots, M$, the number $\bar K$ of smoothings can be chosen according to the residual principle, like in the case of functions of one variable. Since the formula (3.3.44) is the superposition of the formulas for one variable, the substantiation of convergence for the spline approximation method (for example, for the problem of numerical differentiation) is reduced to the substantiation for the case of functions of one variable.

Note also that for aperiodic functions, when calculations at edges of the domain $\Omega_0$ are performed, an extrapolatory smoothing may be used for each $K$ by the least-squares method on the basis of $(n+p)$ points nearest to the boundary. The use of boundary conditions if they are given is possible.

**3. Experimental data processing algorithms.** Here by the processing of experimental data we mean the problems of smoothing, restoration, and numerical differentiation of functions whose values are known approximately at a finite number of points. Such problems often arise for the processing of experimental data in physics, chemistry, the X-ray structure analysis, etc. One of the widespread stable methods for solving the numerical differentiation problem is the method of regularized (according A. N. Tikhonov) smoothing polynomial splines [172], [44]. However, a software implementation of this method requires the solution of some linear systems; although these systems have a special structure, their solution requires a significant amount of operations, especially for functions of two and more variables. If, for example, the choice of the regularization parameter in Tikhonov's method is made numerically on the basis of the principle of residual or quasisolutions, then the above systems have to be solved repeatedly.

The spline approximation method under consideration does not require the solution of linear systems; this allows one to construct experimental data processing algorithms that are simpler in implementation and require a lower computing cost in comparison with the traditional spline smoothing method.

Let us formulate our algorithm for the case of functions of one variable. *Algorithm 1.*

1) $y_0(x) = \begin{cases} f(x) & \text{if the grid } \{x_i\} \text{ is uniform,} \\ S_{2n-1,p}(x,f) & \text{is the spline on } \{x_i\} \text{ if the grid is nonuniform;} \end{cases}$

2) if the original grid is nonuniform, then the uniform grid $\{\bar{x}_i\}$ with the step $h = 1/(N-1)$ is introduced (if $\{x_i\}$ is uniform, then $\{\bar{x}_i\} \equiv \{x_i\}$);

3) $y_k(x) = \bar{S}_{2n-1,p}(x, y_{k-1})$, $k = 1, \ldots, \bar{K}$, where the spline $\bar{S}_{2n-1,p}$ is constructed on the grid $\{\bar{x}_i\}$ and $\bar{K}$ is chosen according to the residual principle;

4) the sought-for value of the operator $dy_{\bar{K}}/dx$ is computed.

The amount of operations for the derivative obtained by the above algorithm at $N_0$ points of the segment $[0,1]$ is $O\left(N_0(n+p)\bar{K}\right)$; when $N_0 = o(N)$, this amount is less in magnitude than the amount of operations by traditional spline smoothing methods [44], [172], [326], whereas for $N_0 = O(N)$ these amounts are comparable.

Let us describe an algorithm for the sequential smoothing method in the case when a function of $M$ variables is given by its values $f_i = f(Q_i)$ at the points $Q_i = \left(u_1^i, \dots, u_M^i\right)$, not necessarily forming a rectangular grid in a domain $\Omega \subseteq \Omega_0$. To accomplish this, the irrational approximation $IR(u, f)$ of the function $f(u)$ is used:

$$\begin{aligned} IR(u,f) &= \sum_{j=1}^{N} r_j(u) f(Q_j) \Big/ \sum_{j=1}^{N} r_j(u), \\ r_j(u) &= \left( \sum_{i=1}^{M} \left(u_i - u_i^j\right)^2 \right)^{1/2}, \quad j = 1, \dots, N. \end{aligned} \tag{3.3.45}$$

The function $IR(u, f)$ interpolates $f(Q_i)$ at a given point, is exact for constants and, hence, has the approximating properties when $d = \max_{1 \le i \le N} \min_{j \ne i} \|P_i - P_j\|_{R_M} \to 0$, $N \to \infty$. If $\|f\|_{C[\Omega]} \le 1$, then $\|IR(u, f)\|_{C[\Omega]} \le 1$; this ensures that uniform errors in input data do not increase for the irrational approximation if $d \ll \delta^1$.

*Algorithm 2.*

1) $y_0(u) = \begin{cases} f(u) & \text{if } \Omega \text{ is a parallelepiped, the points } Q_i \text{ form} \\ & \text{a rectangular uniform grid;} \\ P\mathcal{L}f(u) & \text{if } \Omega \text{ is a parallelepiped, the points } Q_i \text{ form} \\ & \text{a rectangular nonuniform grid;} \\ IR(u,f) & \text{otherwise;} \end{cases}$

2) an auxiliary rectangular uniform grid $\bar{x}_{ij}$ is introduced in $\Omega_0$;

3) $y_k(u) = P\overline{\mathcal{L}} y_{k-1}(u)$, $k = 1, \dots, \bar{K}$, where the operator $P\overline{\mathcal{L}}$ corresponds to the formula of explicit spline approximation on the grid $\{\bar{x}_{ij}\}$, and $\bar{K}$ is chosen according to the residual principle under the assumptions similar to the case of one variable;

4) the sought-for value of the operator on $y_{\bar{K}}$ is computed.

The fact that the number $\bar{K}$ almost does not depend on $M$ is very important. If the operator to be computed is local (for example, the differentiation operator), then the amount of operations to find its values is proportional to $N + N_0(2n + 2p)^M \bar{K}$ (here $N_0$ is the number of points where these values are computed). If the traditional approach is used to construct the smoothing splines, then the solution of a linear system is required; to accomplish this by the sweep method (for example, in the case $M = 2$), it is necessary to perform about $O\left(N^{3/2}\right)$ arithmetic operations.

The sequential spline smoothing method for functions of two variables by bicubic splines was implemented in the subroutine SG20. The input data array was modeled by the table $f_{i,j} = \bar{f}(x_i, x_j) + (1 - 2W)\delta^1$, where $\{x_i\}_{i=1}^n$ is a uniform grid of $N$ points on $[-\pi, \pi]$, $\delta^1$ is a given level of uniform errors, and $W$ is a random quantity uniformly distributed on $[0, 1]$. The quality of smoothing was estimated by

$$\Delta^0 = \max_{1 \le i,j \le n} \Delta_{i,j}, \quad \Delta^0_{\mathrm{m}} = \frac{1}{N^2} \sum_{i,j=1}^{n} \Delta_{ij},$$

where

$$\Delta_{ij} = \left| \bar{f}(x_i, x_j) - y_{\bar{K}}(x_i, x_j) \right|, \quad \Delta^1 = \max_{1 \le i,j \le n} \triangle^1_{i,j}, \quad \Delta^1_{\mathrm{m}} = \frac{1}{n^2} \sum_{i,j=1}^{N} \Delta^1_{i,j},$$

$$\Delta^1_{ij} = \left| \frac{\partial}{\partial x} \bar{f}(x_i, x_j) - \frac{\partial}{\partial x} y_{\bar{K}}(x_i, x_j) \right|.$$

The above quantities are given in Table 3.3.1 in the case of the choice of $\bar{K}$ according to the uniform residual principle for $\varepsilon = 1.5\,\delta^1$, $N = 61$, and various $\delta^1$ and $\bar{f}(x, y)$. In addition, this table contains the number $\bar{K}$ of smoothings and the run time $t_{\bar{K}}$ of the subroutine SG20 on a BESM-6 computer (in seconds).

**Table 3.3.1.**

| | $\bar{f}(x,y) = e^{-x^2}\cos y + e^{-y^2}\cos x, \quad x, y \in [-\pi, \pi]$ | | | |
|---|---|---|---|---|
| $\delta^1$ | 0.05 | 0.1 | 0.2 | 0.4 |
| $\triangle^0$ | $5.48 \cdot 10^{-2}$ | $9.10 \cdot 10^{-2}$ | $9.20 \cdot 10^{-2}$ | $2.46 \cdot 10^{-1}$ |
| $\triangle^0_m$ | $1.14 \cdot 10^{-2}$ | $1.91 \cdot 10^{-2}$ | $2.56 \cdot 10^{-2}$ | $5.69 \cdot 10^{-2}$ |
| $\triangle^1$ | $2.35 \cdot 10^{-1}$ | $3.23 \cdot 10^{-1}$ | $3.85 \cdot 10^{-1}$ | $4.17 \cdot 10^{-1}$ |
| $\triangle^1_m$ | $6.15 \cdot 10^{-2}$ | $7.22 \cdot 10^{-2}$ | $8.55 \cdot 10^{-2}$ | $1.01 \cdot 10^{-1}$ |
| $\bar{K}$ | 3 | 5 | 7 | 20 |
| $t_{\bar{K}}$ | 3.5 | 4 | 5 | 9 |
| | $\bar{f}(x,y) = \sin(x+y) + \frac{1}{3}\sin(3x+3y), \quad x, y \in [-\pi, \pi]$ | | | |
| $\delta^1$ | 0.05 | 0.1 | 0.2 | 0.4 |
| $\triangle^0$ | $5.81 \cdot 10^{-2}$ | $9.23 \cdot 10^{-2}$ | $1.55 \cdot 10^{-1}$ | $2.67 \cdot 10^{-1}$ |
| $\triangle^0_m$ | $2.28 \cdot 10^{-2}$ | $3.37 \cdot 10^{-2}$ | $6.05 \cdot 10^{-2}$ | $1.23 \cdot 10^{-1}$ |
| $\triangle^1$ | $4.56 \cdot 10^{-1}$ | $5.62 \cdot 10^{-1}$ | $7.03 \cdot 10^{-1}$ | $3.62 \cdot 10^{-1}$ |
| $\triangle^1_m$ | $1.51 \cdot 10^{-1}$ | $1.75 \cdot 10^{-1}$ | $2.19 \cdot 10^{-1}$ | |
| $\bar{K}$ | 2 | 3 | 6 | 15 |
| $t_{\bar{K}}$ | 3 | 4 | 5 | 7 |

The graphs of the functions $\bar{f}$, $f$, $y_{\bar{K}}$, $\partial\bar{f}/\partial x$, $\partial y_0/\partial x$, and $\partial y_{\bar{K}}/\partial x$ are represented in Figures 3.3.1–3.3.6. for $\bar{f} = e^{-x^2}\cos y + e^{-y^2}\cos x$, $N = 61$,

$\delta^1 = 0.3$. It follows from these figures that the useful information is completely lost for differentiation without smoothing, whereas $\partial y_{\bar{K}}/\partial x$ is close to $\partial \bar{f}/\partial x$.

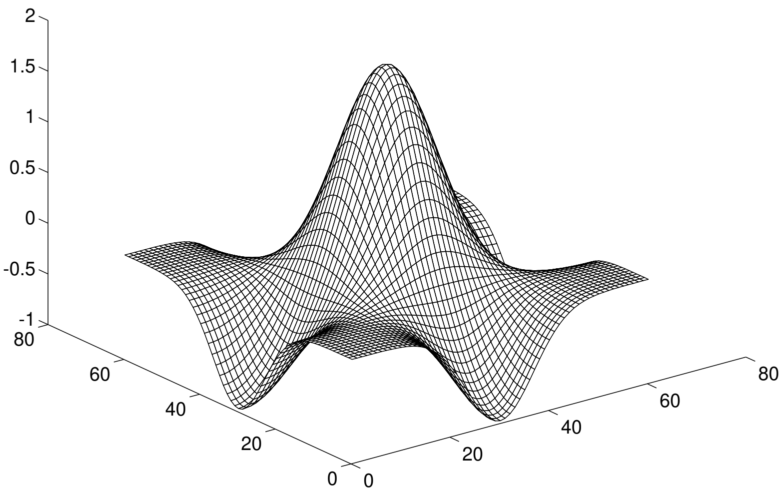


Figure 3.3.1: The exact function

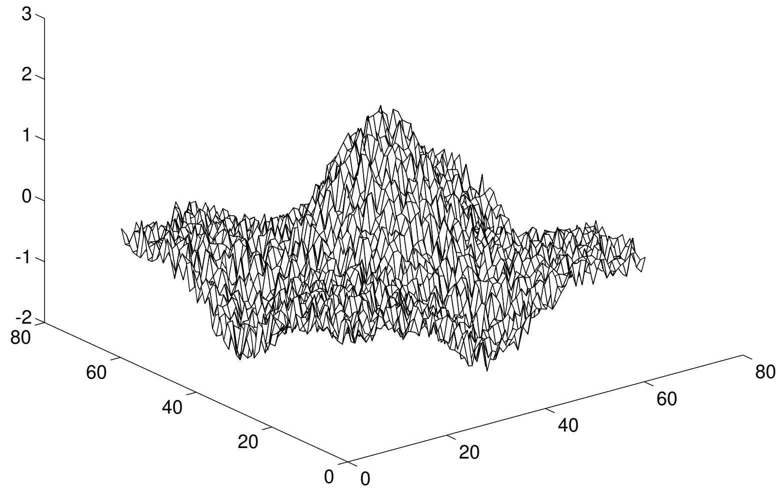


Figure 3.3.2: The noisy function

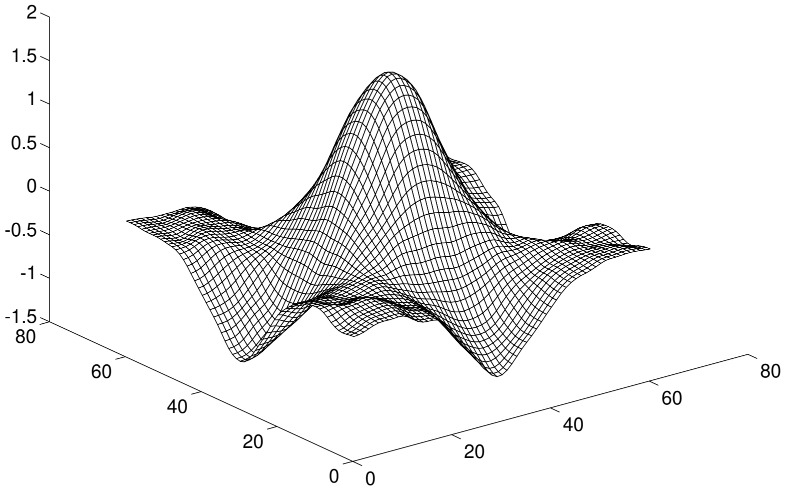


Figure 3.3.3: The restored function

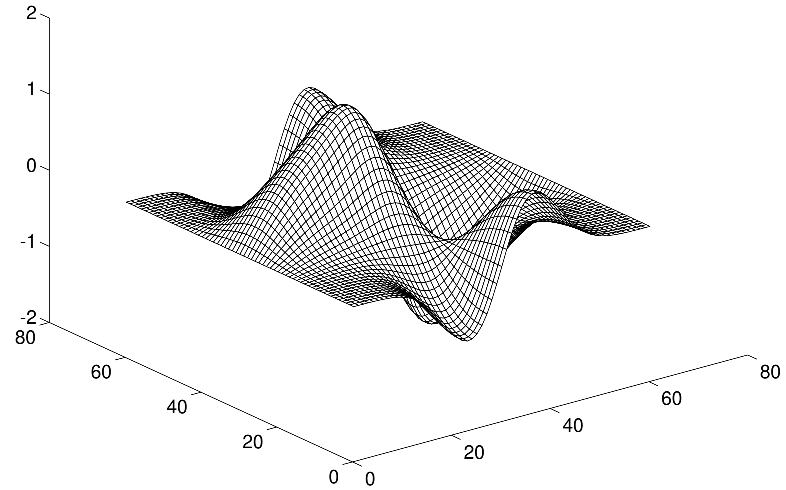


Figure 3.3.4: The derivative of the exact function

**4. Experimental data processing algorithms with additional a priori information.** The locality property of the spline approximation method allows one to easily take into account the additional qualitative *a*

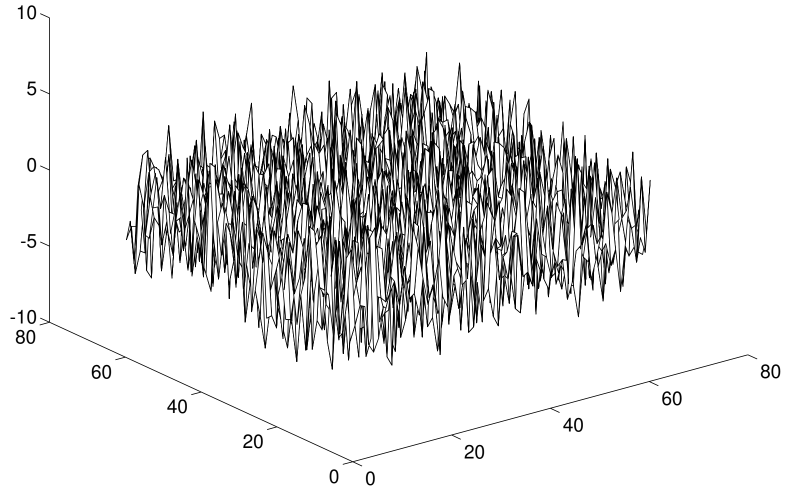


Figure 3.3.5: The derivative of the noisy function

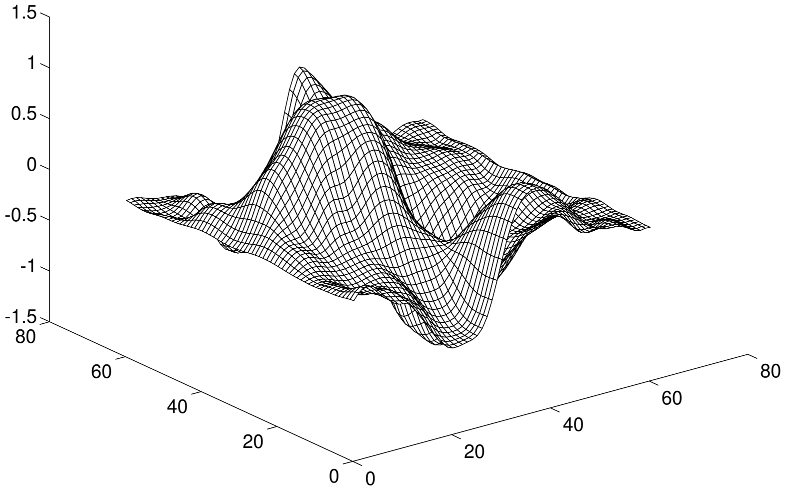


Figure 3.3.6: The restored derivative

*priori* information on an exact function $\bar{f}$ in the form of given segments of sign-definiteness, monotonicity, convexity, etc. Such qualitative *a priori* information is often known in physical, chemical, and other experiments. Its usage allows one to construct regularizing algorithms for wide classes of ill-posed problems by minimizing the residual on a class of functions satisfying some qualitative *a priori* restrictions [70]–[72], [75].

The computing cost required by these algorithms is high; this is quite reasonable for solving the integral equations, for example. However, for the problems of restoration and differentiation of experimental data, in our opinion, the methods based on correction of violations of necessary *a priori* properties because of random errors in input data are more natural and simpler in computations. Let us consider some algorithms based on correction of approximating splines constructed by the sequential smoothing method.

Let the following additional information be known under the assumptions of Subsection 3.3.2:

a) $\bar{f}(x) \le 0,\ \ x \in [0,1]$; it is required to restore a value of this function;
b) $\bar{f}'(x) \le 0,\ \ x \in [0,1]$; it is required to restore a value of this function;
c) $\bar{f}''(x) \le 0,\ \ x \in [0,1]$; it is required to restore the first derivative.

*Algorithm 3.*

1) $\varphi_{\bar{K}} = \begin{cases} y_{\bar{K}} & \text{for cases a) and b)}, \\ dy_{\bar{K}}/dx & \text{for case c)}, \end{cases}$

where $y_{\bar{K}}$ is computed by *Algorithm 1*;

2) in case a):

$$\varphi_{\text{cor}}(\bar{x}_{i+1}) = \begin{cases} \varphi_{\bar{K}}(\bar{x}_{i+1}) & \text{for } \ \varphi_{\bar{K}}(\bar{x}_{i+1}) \le 0, \\ 0 \quad \text{for } \ \varphi_{\bar{K}}(\bar{x}_{i+1}) > 0; \end{cases}$$

in cases b) and c):

$$\varphi_{\text{cor}}(\bar{x}_{i+1}) = \begin{cases} \varphi_{\bar{K}}(\bar{x}_{i+1}) & \text{for } \ \varphi_{\bar{K}}(\bar{x}_i) \ge \varphi_{\bar{K}}(\bar{x}_{i+1}), \\ \varphi_{\bar{K}}(\bar{x}_i) \ \ \text{for } \ \varphi_{\bar{K}}(\bar{x}_i) < \varphi_{\bar{K}}(\bar{x}_{i+1}); \end{cases}$$

3) $\varphi = PV^{K_0}\mathcal{L}\varphi_{\text{cor}}$; $K_0$ additional smoothings are performed.

The necessity of item 3) is caused by the fact that $\varphi_{\text{cor}}$ may have not a necessary smoothness. In practice, $K_0$ is chosen to be equal to 1, 2, or 3. By virtue of the *isogeometric properties* of explicit spline approximation [84], the function $\varphi$ has necessary geometrical properties (in particular, $\varphi$ is monotone). The substantiation of *Algorithm 3* is reduced to the above substantiation of item 1), since it is obvious that the corrections made in item 2) do not increase the uniform error in item 1).

It is easy to construct several algorithms implementing some combinations of conditions a), b), and c) on various subintervals of the segment $[0,1]$, as well as an algorithm computing the half-sum of smoothings and the values corrected from left to right and from right to left.

*Algorithm 3* was implemented in the FORTRAN subroutine SG2 for $n = 2$ and $p = 0$ on a uniform grid with the choice of $K = \bar{K}$ according to the residual principle in uniform norm for $\varepsilon = 1.5\,\delta^1$. The input data for model tests were taken in the form of noisy values of an exact function on a uniform grid. The functions $\bar{f}(x) = \cos x$, $x \in [0, \pi/2]$, and $\bar{f}(x) = \sin x$, $x \in [0, \pi]$, were chosen as model ones.

The results of restoration of the derivative with taking account of its monotonicity were estimated by $\Delta_0$, $\Delta_{0\,\mathrm{m}}$, $\Delta^1$, $\Delta^1_{\mathrm{m}}$, $\bar{\Delta}^1$, $\bar{\Delta}^1_{\mathrm{m}}$:

$$\begin{aligned}
\Delta^1 &= \max_{1 \le i \le N} \Delta_i, \quad \Delta^1_{\mathrm{m}} = \frac{1}{N} \sum_{i=1}^{N} \Delta_i, \\
\bar{\Delta}^1 &= \max_{1 \le i \le N} \bar{\Delta}_i, \quad \bar{\Delta}_{\mathrm{m}} = \frac{1}{N} \sum_{i=1}^{N} \bar{\Delta}_{0i}, \\
\Delta_0 &= \max_{1 \le i \le N} \Delta_{0i}, \quad \Delta_{0\,\mathrm{m}} = \frac{1}{N} \sum_{i=1}^{N} \Delta_{0i},
\end{aligned}$$

where $(i = 1, \dots, N)$

$$\begin{aligned}
\Delta_i &= \left| \varphi_{\bar{K}}(x_i) - \frac{d\bar{f}(x_i)}{dx} \right|, \quad \bar{\Delta}_i = \left| \varphi(x_i) - \frac{d\bar{f}(x_i)}{dx} \right|, \\
\Delta_{0i} &= \left| \frac{d\bar{f}(x_i)}{dx} - \frac{f_{i+1} - f_i}{h} \right|.
\end{aligned}$$

The numerical results for $\bar{f}\,(x) = \cos x$, $N = 20$ and for $\bar{f}\,(x) = \sin x$, $N = 40$ are given in Tables 3.3.2 and 3.3.3. Numerical experiments allow us to conclude that the usage of *Algorithm 3* with correction decreases errors of restoration of the derivative compared to *Algorithm 1*: the uniform and mean errors approximately decrease by 40 % and 25 %, respectively.

**Table 3.3.2.**

| $\delta_0 \backslash \Delta$ | $\bar{\Delta}^1$ | $\Delta^1$ | $\Delta_0$ | $\bar{\Delta}^1_{\mathrm{m}}$ | $\Delta^1_{\mathrm{m}}$ | $\Delta_{0\,\mathrm{m}}$ | $\bar{K}$ |
|---|---|---|---|---|---|---|---|
| 0.03 | 0.999 | 0.143 | 0.292 | 0.043 | 0.055 | 0.122 | 3 |
| 0.05 | 0.083 | 0.260 | 0.338 | 0.033 | 0.053 | 0.161 | 5 |
| 0.09 | 0.176 | 0.435 | 0.557 | 0.059 | 0.095 | 0.228 | 11 |

The subroutine SG2 was tested for the case of restoration of a function with consideration of its monotonicity. Our computations were performed for $\bar{f}\,(x) = \cos x$, $x \in [0, \pi/2]$. Taking account of monotonicity allowed us to improve the quality of approximation. The subroutine has the mode of operation for the case when the positiveness (negativeness) of the function under restoration is taken into account.

**Table 3.3.3.**

| $\delta_0\backslash\Delta$ | $\bar{\Delta}^1$ | $\Delta^1$ | $\Delta_0$ | $\bar{\Delta}^1_{\mathrm{m}}$ | $\Delta^1_{\mathrm{m}}$ | $\Delta_{0\,\mathrm{m}}$ | $\bar{K}$ |
|---|---|---|---|---|---|---|---|
| 0.03 | 0.185 | 0.189 | 0.286 | 0.043 | 0.044 | 0.103 | 3 |
| 0.05 | 0.186 | 0.278 | 0.505 | 0.074 | 0.081 | 0.168 | 4 |
| 0.09 | 0.241 | 0.273 | 0.869 | 0.093 | 0.105 | 0.345 | 5 |

A number of numerical experiments were performed to compare the subroutine SG2 with the standard subroutine IA20R from the Numerical Analysis Library of the Moscow University Research Computing Center. The standard subroutine allows one to construct the best root-mean-square approximation of a function of one variable on the set of piecewise convex functions with the bounded first derivative. For comparison, the model function $\bar{f}(x) = \cos x$, $x \in [0, \pi/2]$, was used on the uniform grid with the step $h = \pi/39$, $\delta^1 = 0.04$, 0.06, 0.08.

By some changes of parameters in the subroutine IA20R, we empirically achieved an approximate correspondence between the errors of restoration in the derivatives computed by IA20R and SG2. Our experiments allow us to conclude that the computing time spent by the subroutine SG2 is 4–6 times less than that of IA20R for the same level of accuracy in restoring the first derivative.

*Algorithm 2* for a nonuniform grid was compared with the algorithm of variational spline smoothing [44]. The function

$$f(x, y) = 3x^2 \sin(5y + x) + 4(y - x)\cos(3x - 2y), \quad (x, y) \in [0, 1] \times [0, 1]\,,$$

was taken as a test one.

Its image on the map of isolines (the isolines were drawn for integer values from $-2$ to 5 with pointwise computations on the grid $240 \times 175$) is represented in Figure 3.3.7 (left).

For comparison of algorithms, three series of experiments with $N$ original points ($N = 50$, 100, 200) were performed. The original points were generated with the aid of a random-number generator as uniformly distributed on a unit square with a restriction on proximity (not closer than 0.05 %). Errors introduced into the values of the function were modeled as normally distributed random numbers with zero mean and with mean-square deviation 0.5. The transition to a greater number of points was carried out by addition of new points to the existing ones. The map of isolines for the noisy function is represented for $N = 100$ in Figure 3.3.7 (right).

The results of numerical experiments are given in Table 3.3.4, 3.3.5 and in Figure 3.3.8, where the maps of isolines for the splines constructed by the

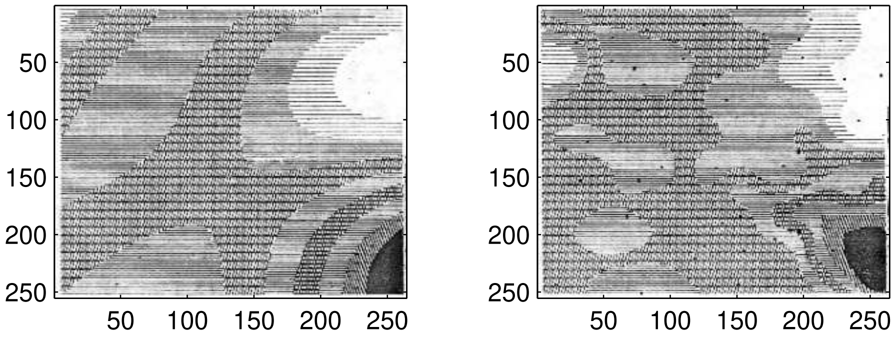


Figure 3.3.7: The maps of isolines for the exact (left) and noisy (right) functions

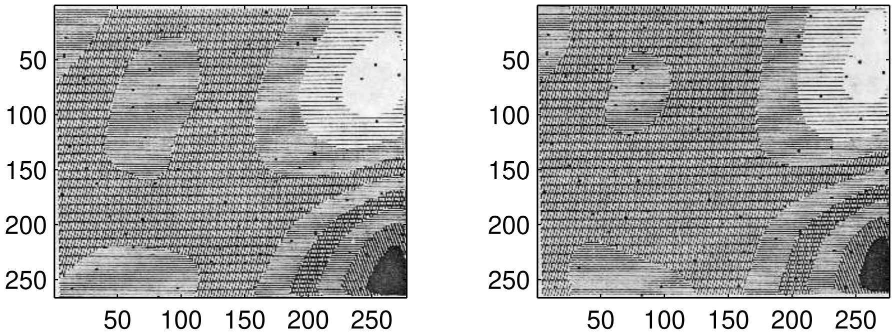


Figure 3.3.8: The maps of isolines for the splines constructed by the variational (left) and explicit (right) algorithms

**Table 3.3.4.** Restoration errors (mean square deviation)

| Number of points $N$ | Variational algorithm | Explicit algorithm |
|---|---|---|
| 50 | 0.309 | 0.336 |
| 100 | 0.299 | 0.310 |
| 200 | 0.249 | 0.280 |

**Table 3.3.5.** The computing time for approximation at the grid points (in seconds)

| Number of points $N$ | Var. algorithm | | | Explicit algorithm | | |
|---|---|---|---|---|---|---|
| $N$ | $T_f$ | $T_g$ | $T_V$ | $T_I$ | $T_S$ | $T_C$ |
| 50 | 12 | 30 | 42 | 25 | 7 | 32 |
| 100 | 60 | 60 | 120 | 50 | 6 | 56 |
| 200 | 525 | 115 | 640 | 100 | 7 | 107 |

variational and explicit algorithms are represented for $N = 100$. The quality of approximation is characterized by the estimation of mean square deviation for the restoration error of exact values of the test function (Table 3.3.4) and is illustrated by the maps of isolines of the approximating functions. The choice of the smoothing parameter in both the algorithms was performed according to the residual principle with $\delta^0 = 0.5$.

The computing time of approximation for the above two algorithms on the $20 \times 20$ grid measured on a personal computer IBM PC XT with a collection of FORTRAN subroutines is illustrated in Table 3.3.5. The time $T_V$ of the variational algorithm is presented as the sum of $T_f$ and $T_g$, where $T_f$ is the time spent for computation of the spline and $T_g$ is the time spent for computation of the values of the spline at the grid points. The time $T_C$ of the explicit algorithm is presented as the sum of $T_I$ and $T_S$, where $T_I$ is the time spent for the irrational interpolation and $T_S$ is the time spent for spline smoothing.

The above maps of isolines and the accuracy estimates in Table 3.3.4 allow us to conclude that both the algorithms demonstrate a satisfactory quality of restoration of a function $f(x, y)$ with inexact data, but the accuracy of the variational algorithm is slightly better. The algorithms differ widely in the amount of storage required; in this sense, the explicit algorithm has the main advantage. The variational algorithm requires to store a dense symmetric matrix of order $N+3$ and the original values of a function, i.e., $(N+3)(N+6)/2$ words of storage in the storage-saving variant and $(N+3)(N+4)$ in the time-saving variant. Thus, the amount of storage is proportional to $N^2$. The explicit algorithm requires to store the coordinates of the original points and the values of a function at these points as well as the values of the approximating function at the grid points, i.e., $3N + M \times M$ words of storage. Note that the number $M \times M$ of points usually is chosen of order $N$. Thus, the amount of storage is proportional to $N$.

The time characteristics of the two algorithms (Table 3.3.5) are similar for small $N$; for $N > 50$, the explicit algorithm has some advantage. Its advantage is doubled for $N = 100$ and becomes five times as much for $N = 200$.

## 3.4 Algorithms for solving linear integral equations with singularities in kernels

Let us consider the special case of equation (3.1.1) with the integral operator

$$Au \equiv \int_a^b K(x,t)u(t)\,dt \equiv \bar{A}_0 u + \bar{A}_1 u = f(x), \tag{3.4.46}$$

where

$$\bar{A}_0 u \equiv \int\limits_a^b \bar{K}_0\left(|x-t|\right) u(t) dt$$

is the main part of the operator $A$.

Further we consider the periodic case when the kernel is of the form

$$\bar{K}_0\left(|x-t|\right) = \ln\left(\frac{1}{|\sin(\pi(x-t)/(b-a))|}\right). \tag{3.4.47}$$

The operator $\bar{A}_1 u \equiv \int\limits_a^b \bar{K}_1(x,t)u(t)\,dt$ has a sufficiently smooth doubly periodic kernel $\bar{K}_1(x,t)$ and the functions $f(x)$, $u(t)$ are periodic with period $(b-a)$. We search for an approximate solution in the form

$$g_N(t) = \sum_{j=-n+1}^{N+n-1} c_j s_{j-n,2n-1}(t), \quad t \in [a,b], \tag{3.4.48}$$

where $c_{j+1} = c_{N+j}$, $j = -n, \ldots, n$.

The system (3.1.4) is of the form

$$\sum_{j=1}^{N-1} c_j (I_{ij} + J_{ij}) = f(x_i), \quad i = 1, \ldots, N-1, \tag{3.4.49}$$

where

$$I_{ij} = \int\limits_{x_{j-n}}^{x_{j+n}} K_0(x_i,t) s_{j-n,2n-1}(t)\,dt, \quad J_{ij} = \int\limits_{x_{j-n}}^{x_{j+n}} K_1(x_i,t) s_{j-n,2n-1}(t)\,dt,$$

$$\begin{aligned} K_0\left(x,t\right) &= \ln\frac{(b-a)^2}{|x-t|\left[(b-a)-|x-t|\right]}, \\ K_1\left(x,t\right) &= \bar{K}_1\left(x,t\right) + \bar{K}_0\left(x,t\right) - K_0\left(x,t\right). \end{aligned}$$

The quantities $I_{ij}$ can be calculated analytically and $J_{ij}$ can be found, in general, only approximately. Therefore, in order to perform realistic calculations, we pass from the system (3.4.49) to the system

$$\sum_{j=1}^{N-1} c_j (I_{ij} + \tilde{J}_{ij}) = f(x_i), \quad i = 1, \ldots, N-1, \tag{3.4.50}$$

where $\tilde{J}_{ij}$ is an approximation of $J_{ij}$. It is of prime importance the following fact: if a grid is uniform and the kernel $K(x,t)$ of the operator $A$ depends only on the absolute value of the difference of its arguments, then the matrix of system (3.4.50) is a symmetric Toeplitz matrix.

**1. Algorithms for equations whose right-hand sides and operators contain errors.** Let the right-hand side $f(x)$ and the kernel $K(x,t)$ be represented in the form

$$f(x) = \bar{f}(x) + \delta(x), \quad K(x,t) = \hat{K}(x,t) + \Delta(x,t),$$

where $\bar{f}$, $\hat{K}$ are exact functions and $\delta(x)$, $\Delta(x,t)$ are errors bounded only in uniform norm in such a way that $\|\delta(x)\|_{C[a,b]} \leq \delta_0$ and $\|\Delta(x,t)\|_{C[\Omega]} \leq \Delta$. Then, the collocation method should be applied to the equation with a presmoothed right-hand side and a presmoothed kernel.

In this case, when the right-hand side of system (3.4.50) is computed, it is necessary to use the function $y_{\bar{K}}(x)$ instead of $f(x)$ and the function $\breve{K}_1(x,t)$ instead of $K_1$ to find the matrix $\tilde{J}_{ij}$. The functions $y_{\bar{K}}$ and $\breve{K}_1$ are obtained by the sequential smoothing method from the functions $f$ and $K_1(x,t)$, respectively.

Let us discuss the following algorithm for solving the equations with logarithmic singularity in kernels.

1. First we compute the function $y_{k_1}$ on a grid $\{x_i\}$:

$$y_k = P\mathcal{L}y_{k-1}(x), \quad k = 1,\ldots,k_1. \tag{3.4.51}$$

Here the operator $P$ is defined by the formula (3.1.1), $y_0 = f$, and $k_1 = k_1(N,\delta_0)$ is chosen according to the root-mean-square residual principle.

2. Next we compute the function $\breve{K}_1(x,t)$ on the grid $\{x_i\} \times \{x_j\}$ by the formula (3.3.44) for the case of two variables:

$$\breve{K}_1(x,t) = P_2 V_2^{k_2} P_1 V_1^{k_2} \mathcal{L}_1 \mathcal{L}_2 K_1(x,t), \tag{3.4.52}$$

$$K_1(x,t) = K(x,t) - \bar{K}_0(x,t). \tag{3.4.53}$$

Here $k_2 = k_2(N,\Delta)$ is chosen according to the root-mean-square residual principle.

3. Finally, we solve the system

$$\sum_{j=1}^{N-1} c_j (I_{ij} + \breve{J}_{ij}) = y_{k_1}(x_i), \quad i = 1,\ldots,N-1, \tag{3.4.54}$$

where $I_{ij}$ are defined as in Subsection 3.4.1 and

$$\breve{J}_{ij} = \int\limits_{x_{j-n}}^{x_{j+n}} \breve{K}_1(x_i,t)\, s_{j-n,2n-1}(t)\, dt.$$

**2. Algorithms for solving some two-dimensional integral equations.** Let us consider the case of equation (3.1.1) with the integral operator

$$Au \equiv \int_a^b \int_c^d K(x,y,t,\tau)u(t,\tau)\,dt\,d\tau = f(x,y), \quad A = A_0 + A_1. \tag{3.4.55}$$

The operator $A_0$ is the main part of the operator $A$:

$$A_0 u = \int_a^b \int_c^d K_0(x,y,t,\tau)u(t,\tau)\,dt\,d\tau.$$

The operator

$$A_1 u = \int_a^b \int_c^d K_1(x,y,t,\tau)u(t,\tau)\,dt\,d\tau$$

has a sufficiently smooth kernel $K_1$. Let us introduce the grids $\{x_i\}$ and $\{y_j\}$:

$$\begin{aligned} x_{-2n+1} &< \ldots < x_0 < a = x_1 < x_2 < \ldots < x_{N_1} = \\ &= b < x_{N_1+1} < \ldots < x_{N_1+2n}, \\ y_{-2n+1} &< \ldots < y_0 < c = y_1 < y_2 < \ldots < y_{N_2} = \\ &= d < y_{N_2+1} < \ldots < y_{N_2+2n}. \end{aligned}$$

Let $s_{i,2n-1}(x)$ and $s_{j,2n-1}(y)$ be the local basis splines of degree $2n-1$ constructed using the points $x_i, \ldots, x_{i+2n}$ and $y_j, \ldots, y_{j+2n}$, respectively.

An approximate solution is sought in the form

$$u_N(t,\tau) = \sum_{i \in J_1} \sum_{j \in J_2} c_{ij} s_{i-n,2n-1}(t) s_{j-n,2n-1}(\tau).$$

It is assumed that the functions $f$, $K$, and $u$ are periodic on $\Omega = [a,b] \times [c,d]$. Then,

$$J_l = \{-n+1, \ldots, N+n-1\}, \quad l = 1,2;$$

in addition, the following conditions are fulfilled:

$$c_{i+1,j+1} = c_{N_1+i,N_2+j}, \quad i,j = -n, \ldots, n.$$

The system (3.1.4) has the form

$$\sum_{i \in \bar{J}_1} \sum_{j \in \bar{J}_2} c_{ij} \left( I_{ijlk} + J_{ijlk} \right) = \varphi_{l,k}, \quad l \in \bar{J}_1, \quad k \in \bar{J}_2, \tag{3.4.56}$$

where

$$
\begin{aligned}
\bar{J}_m &= \{1, \ldots, N_m - 1\}, \quad m = 1, 2; \quad \varphi_{l,k} = f(x_l, y_k), \\
I_{ijlk} &= \int\limits_{x_{i-n}}^{x_{i+n}} \int\limits_{y_{j-n}}^{y_{j+n}} K_0(x_l, y_k, t, \tau) s_{i-n,2n-1}(t) s_{j-n,2n-1}(\tau)\, dt\, d\tau, \\
J_{ijlk} &= \int\limits_{x_{i-n}}^{x_{i+n}} \int\limits_{y_{j-n}}^{y_{j+n}} \tilde{K}_1(x_l, y_k, t, \tau) s_{i-n,2n-1}(t) s_{j-n,2n-1}(\tau)\, dt\, d\tau
\end{aligned}
$$

and $\tilde{K}_1$ is an approximation of the kernel $K_1$. It is assumed that the functions $K_0$ and $\tilde{K}_1$ are chosen in such a way that $I_{ijlk}$ and $J_{ijlk}$ can be calculated analytically.

Note that if the kernel $K_0$ depends only on $|x - t|, |y - \tau|$ and $K_1 \equiv 0$, then for uniform grids the matrix of system (3.4.56) is a symmetric block-Toeplitz matrix.

**3. Substantiation of the algorithms for solving the above integral equations.** Substantiation of the algorithms consists in the following: we should prove that the matrices of the corresponding linear systems are not singular and that the approximate solutions converge to the exact one. The grid $\{x_i\}$ may be nonuniform, but we assume that this grid is regular, i.e., the following relation is valid uniformly as $h \to 0$: $h/\min h_i \le c$, $c = \text{const} > 0$. Here $h_i = x_{i+1} - x_i$ and $h = \max_{1 \le i \le N-1} h_i$.

**Theorem 3.4.66** *Let* $K_1(x,t) \equiv 0$ *and equation* (3.4.46) *be uniquely solvable; then, the solution to the system* (3.4.50) *exists and is unique.*

**Proof.** Let us show that if $K_1(x,t) \equiv 0$ and, hence, $A^\tau = A_0$, then the hypotheses of Theorem 3.1.58 are fulfilled. The operator $A_0^{-1}$ exists [205]. Having chosen $G = L_2[a,b]$, we see that the integral operator with the kernel (3.4.47) is positive and self-adjoint.

For our case, $U_N$ is the space of splines of degree $2n-1$, which are periodic with period $(b-a)$. The commutativity condition (3.1.9) for the operators $A_0$ and $L = \dfrac{d^n}{dx^n}$ can be proved with integration by parts. A periodic polynomial is only a constant; therefore, the condition (3.1.11) is also valid.

By Theorem 3.1.58, thus, the system (3.4.50) is not singular and the operator $B_0^{-1}$ exists. The theorem is proved.

**Theorem 3.4.67** *Let* $K_1(x,t) = 0$ *and the exact solution* $\bar{u}(t)$ *to equation* (3.4.46) *with the exact-right side hand be such that* $\bar{u}(t) \in W_2^r[a,b]$, $r \ge n$.

*Then, the following estimates are valid:*

$$\|\bar{u}(t)-\bar{g}_N(t)\|_{C^{0,\alpha}[a,b]} \le O\left(h^{\nu-1-\alpha}\right), \quad 0<\alpha<1; \tag{3.4.57}$$

$$\|\bar{u}(t)-\bar{g}_N(t)\|_{W_2^l[a,b]} \le O\left(h^{\nu-l-1/2}\right); \tag{3.4.58}$$

$$\left\|A_0^{1/2}\left(\bar{u}-\bar{g}_N\right)\right\|_{W_2^l[a,b]} \le O\left(h^{\nu-l}\left\|A_0^{1/2}\bar{u}\right\|_{W_2^{\nu}[a,b]}\right); \tag{3.4.59}$$
$$\nu=\min\{r,2n\}, \quad 0\le l\le \nu-1.$$

**Proof**. The hypotheses of Theorem 3.1.60 are fulfilled under the above assumptions. Indeed, the condition 1) of this theorem follows from the fact that the operator $A_0$ is positive and self-adjoint in $G=L_2[a,b]$. The commutativity of the operators $A_0$ and $L=\dfrac{d^n}{dx^n}$ is established with integration by parts. The condition 3) follows from the well-known estimates for the interpolation of functions [2]; if $Y=C^{1,\alpha}[a,b]$, $0<\alpha<1$, we have $\rho_N\le O(h^{n-3/2-\alpha})$, whereas if $Y=W_2^l[a,b]$, we have $\rho_N\le O(h^{n-l})$. The following estimates are valid [2]:

$$\|\bar{u}-\hat{g}_N\|_{G_L}\le O\left(h^{\nu-n}\right), \quad \|A_0\left(\bar{u}-\hat{g}_N\right)\|_{G_L}\le O\left(h^{\nu-n+1}\right).$$

The estimates (3.4.57) and (3.4.58) follow from the above ones. The estimates (3.4.59) in energy norm follow from the extremal property of the spline $\bar{g}_N(t)$ and the previous estimates. The theorem is proved.

**Corollary 3.4.1.** *The estimate* (3.4.59) *implies the boundedness of the operator $B_0^{-1}$ from the space $W_2^{l+1}U_N$ into the space $W_2^lU_N$ for all $l\ge 0$.*

Note that the estimates (3.4.57) and (3.4.58) are preserved if $(A^{\tau})^{-1}$ exists and $K_1(x,t)\ne 0$, but depends only on $|x-t|$.

If $K_1(x,t)\ne 0$, it is possible to study the solvability of the linear systems obtained by the collocation method with the aid of Theorem 3.1.59 and to analyze the error estimates with the aid of Theorem 3.1.61. In order to formulate the corresponding result, by $W_2^{l,q}[\Omega]$ we denote the space of functions of two variables $g(x,y)$ defined on $\Omega=[a,b]\times[a,b]$ such that $\partial^{l+q}g/\partial x^l\partial y^q$ is square integrable on $\Omega$ and by $W_2^lC[\Omega]$ we denote the space of functions such that $d^lg/dx^l$ is continuous in $y$ and is square integrable with respect to $x$ with the pseudonorm

$$\|g(x,y)\|_{W_2^lC[\Omega]}=\max_{a\le y\le b}\left(\int_a^b\left|\frac{d^lg(x,y)}{dx^l}\right|^2dx\right)^{1/2}.$$

**Theorem 3.4.68** *Let all the hypotheses of Theorems* 3.4.66 *and* 3.4.67 *be fulfilled with the exception of* $K_1(x,t) \equiv 0$ *and the following assumptions be valid:*

1) *the operator* $A$ *of equation* (3.4.46) *has the inverse bounded from* $W_2^{l+1}[a,b]$ *into* $W_2^l[a,b]$, $0 \le l \le r$;

2) $K_1(x,t)$, $\tilde{K}_1(x,t) \in W_2^{r+2,q}[\Omega]$, $q \ge 1$;

3) $\left\| K_1(x,t) - \tilde{K}_1(x,t) \right\|_{W_2^{l+1}C[\Omega]} = O\left(h^{\mu-l}\right)$, $\mu = \min\{r+2, q-1/2\}$, $0 < l < \mu$.

*Then, for a sufficiently large* $N$ *the solution to the system* (3.4.50) *exists and is unique and the following estimate is valid:*

$$\|\bar{u}(t) - \bar{g}_N(t)\|_{W_2^l[a,b]} \le O\left(h^{\nu-l}\right), \quad \nu = \min\{\mu, r, 2n\}, \quad 0 < l < \nu.$$

**Proof.** All the hypotheses of Theorem 3.1.59 are fulfilled with $X = W_2^m[a,b]$, $Y = W_2^{m+1}[a,b]$, $m \le n-1$, $\tau = 1/N$; hence, the solution to the system (3.4.50) exists and is unique for a sufficiently large $N$. The estimate in the conclusion of this theorem follows from the estimates of the best approximations by splines, whereas the estimates 1)—3) of this theorem follow from the estimates (3.1.14) of Theorem 3.1.61 for $Z_1 = W_2^l[a,b]$ and $Z_2 = W_2^{l+1}[a,b]$. The theorem is proved.

Let us consider an algorithm that takes into account errors in the kernel and the right-hand side. It is assumed that the estimates $\delta_0 \ge \|\delta(x)\|_{C[a,b]}$ and $\Delta \ge \|\Delta(x,t)\|_{C[a,b]}$ are known. Let $g_{N,\delta_0,\triangle}(t)$ be the spline with its coefficients determined from (3.4.54). The grid $\{x_i\}$ is supposed to be uniform.

**Theorem 3.4.69** *Let the following conditions be fulfilled:*

1) *the operator* $A$ *of equation* (3.4.46) *has the inverse bounded from* $W_2^{r+1}[a,b]$ *into* $W_2^r[a,b]$;

2) $\hat{K}(x,t) - \bar{K}_0(x,t) \in W_2^{2r+4}C[\Omega]$;

3) $\bar{f}(x) \in W_2^{2r+4}[a,b]$.

*Then, the system* (3.4.54) *is uniquely solvable for a sufficiently large* $N$ *and a small* $\Delta < \bar{\Delta}$ *and the algorithm* (3.4.51)–(3.4.54) *is regularizing in* $W_2^r[a,b]$, *i.e.,*

$$\lim_{N\to\infty,\Delta\to 0,\delta_0\to 0} \sup_{f:\|f-\tilde{f}\|_C \le \delta_0, K:\|K-\hat{K}\|_C \le \Delta} \|\bar{u}_0(t) - g_{N,\delta_0,\Delta}(t)\|_{W_2^r[a,b]} = 0.$$

The proof of this theorem is based on the regularizing properties of the spline approximation method for problems of approximation and numerical differentiation of functions of one and two variables and on Theorems 3.4.68 and 3.1.61.

Let us discuss the case of two-dimensional integral equations. For definiteness, we consider bicubic splines ($n = 2$), the operator $L$ is of the form

$$Lg = L_{x,y}g(x, y) = \frac{\partial^4 g(x, y)}{\partial x^2 \partial y^2},$$

and the functionals $l_i(g)$, $i = 1, \ldots, N_1 \times N_2$, are the values of $g(x, y)$ on the grid $\{x_l, x_k\}$. The space $U_N$ is the space of bicubic splines periodic on $\Omega$.

**Theorem 3.4.70** *Let $K_1 \equiv 0$ and the kernel $K_0$ of equation* (3.4.55) *satisfy the following conditions under the above assumptions:*

1) *the operator $A_0$ is not singular;*
2) *$A_0$ is positive in $L_2[\Omega]$;*
3) *$K_0(x, y, t, \tau) = K_0(|x - t|, |y - \tau|)$;*
4) *the function $L_{xy}A_0 g \in L_2[\Omega]$ for all $g \in U_N$.*

*Then the solution to the system* (3.4.56) *exists and is unique.*

**Proof.** Let us show that the hypotheses of Theorem 3.1.58 are fulfilled. In our case, the conditions (3.1.9) and (3.1.10) follow from the hypotheses 1) and 2) of this theorem if we put $G = L_2[\Omega]$. Only bilinear functions belong to the kernel of the operator $L$. Only a constant may be a bilinear function periodic in $\Omega$. Substituting $g = \text{const} = c_0$ into the conditions $l_i(A_0 c_0) = 0$, $i = 1, \ldots, N_1 \times N_2$, we obtain $c_0 = 0$.

Thus, the condition (3.1.11) is fulfilled. In view of the properties of the operator $L$, the verification of the condition (3.1.9) is reduced to the verification (in the space $U_N$) of commutativity of the operator $A_0$ and of the operator of single differentiation with respect to $x$ and (or) to $y$; this can be established with integration by parts with consideration of periodicity. The theorem is proved.

In the case when $K_1 \neq 0$ and the kernel $K_1$ is such that the operator $A_1$ is completely continuous from $L_2[\Omega]$ into $W_2^{1,1}[\Omega]$, the substantiation of the existence and uniqueness of the solution to the system (3.4.55) is established on the basis of Theorem 3.1.59 by analogy with the one-dimensional case for sufficiently large $N_1$ and $N_2$. Similarly, the error estimates for the spline collocation method are established on the basis of Theorems 3.1.60, 3.1.61 and the error estimates for interpolation of functions of two variables by bicubic splines.

Obviously, the scheme of application and substantiation for the sequential smoothing method can be generalized to the case when the right-hand side and the kernel $K_1$ are noisy.

## 3.5 Collections of FORTRAN subroutines and the solution of some applied problems

**1. A collection of subroutines for approximation and smoothing of functions by splines.** This collection of subroutines consists of several subroutines for uniform grids: SG1, SG2, SG20; these subroutines are independent of the others, since they implement the most simple and fast algorithms. The software implementation of algorithms for nonuniform grids is based on the modular approach: several subroutines were implemented for the one-dimensional case (MASSIV, INTERV, BVAL, and others). They are used as the main modules in the other subroutines for functions of one, two, and three variables. The descriptions of the main subroutines for approximation of functions of one and two variables are given in [84], [86]. The subroutines for the case of three variables are similar to those for the two-dimensional case.

Here we describe the subroutines implementing the algorithms for approximation of surfaces on the basis of finite sets of points, not forming regular grids, with the use of parametric spline approximation or irrational interpolation. The descriptions of sequential spline smoothing subroutines for functions of one and two variables on uniform rectangular grids are also given.

**SAP13**

*Purpose*

Computation of Cartesian coordinates of a curve and their derivatives with respect to a parameter by parametric cubic approximating splines constructed on the basis of a finite number of points.

**SAP22**

*Purpose*

Restoration of Cartesian coordinates of a surface on the basis of given Cartesian coordinates of a finite set of original surface points lying on space curves by bicubic parametric splines.

**IRIN21**

*Purpose*

This function subroutine implements the irrational interpolation method for functions whose values are given on irregular grids on a plane. Let the values of a functions $Z$ of two variables be given on an arbitrary grid: $Z(I) = Z(X(I), Y(I))$, $I = 1, \ldots, N$. An approximate value $Z(x, y)$ at any point of the plane $(X, Y)$ is computed by the Schepard-type irrational interpolation method [85] by the formulas (3.3.45).

**SG1**

*Purpose*

Smoothing of functions of one variable on the basis of its noisy values on a uniform grid given on a segment of the real axis.

**SG2**

*Purpose*

Restoration of a function or its derivative on the basis of its noisy values on a uniform grid with consideration of its sign-definiteness, monotonicity, and convexity by the sequential spline smoothing method with the choice of the number of smoothings according to the residual principle.

**SG20**

*Purpose*

This subroutine implements the sequential spline smoothing method for noisy functions of two variables by bicubic splines. Let approximate values of a function $F(X, Y)$ be given on a two-dimensional grid $\{X(I), Y(J)\}$, $I = 1, \ldots, N1$, $J = 1, \ldots, N2$; $F(I, J) = F(X(I), Y(J))$. A smoothing bicubic spline is constructed on the basis of these values by the sequential smoothing method.

The single smoothing is performed according to the explicit formula

$$S(I, J) = F(I, J) + \frac{1}{6}\,\Phi 1(I, J) + \frac{1}{6}\,\Phi 2(I, J) + \frac{1}{36}\,\Phi 3(I, J),$$

where

$$\begin{aligned}
\Phi 1(I, J) &= F(I-1, J) - 2F(I, J) + F(I+1, J), \\
\Phi 2(I, J) &= F(I, J-1) - 2F(I, J) + F(I, J+1), \\
\Phi 3(I, J) &= \Phi 2(I-1, J) - 2\Phi 2(I, J) + \Phi 2(I+1, J).
\end{aligned}$$

When the further smoothings are performed, the values of the smoothed function obtained after the previous smoothing are used in this formula instead of the function $F$. The process of smoothing is repeated $K$ times; $K$ is given by the user or is calculated according to the residual principle with the use of the error bound EPS if $K = 0$ on entry of this subroutine. The values of the smoothed function on the same grid are obtained on return.

**2. A collection of subroutines for solving integral equations with logarithmic singularities in kernels.** This collection has a modular structure and consists of auxiliary subroutines and the main subroutines RIU23 and RIU25 corresponding to the third- and fifth-degree splines.

**RIU23**

*Purpose*

Computation of the coefficients of a cubic spline from the collocation conditions on a uniform grid for an integral equation with a logarithmic singularity in its kernel.

**FP**

*Purpose*

Evaluation of the indefinite integral

$$\int \ln|x-t|\, t^k\, dt.$$

**FL**

*Purpose*

Evaluation of the integral

$$\int_{x_{j-2}}^{x_{j+2}} \tilde{K}_1(x,t) s_{j-2,3}(t)\, dt$$

by a Gauss quadrature formula.

**3. On solution of some applied problems.** Let us consider some problems connected with processing of observations on the basis of restoration of functions and their derivatives when approximate values of these functions are known in finite sets of points.

1) *The data processing problem for calculation of parameters of radio engineering devices.* In order to calculate the interrelation of parameters for electronic devices during their design, it is necessary to calculate the differential parameter dependence when the values of a function are known with errors for a finite discrete set of parameter values. Thus, it is necessary to solve the problem of numerical differentiation. To accomplish this, the sequential smoothing algorithm described in Section 3.3 may be used. The number of smoothings is chosen according to a given relative error characterizing experimental data. It is very important that the algorithm is strictly substantiated theoretically, is simple in software implementation, and has low computing cost; this allows one to use this algorithm in systems of automatic design. The program SG1 is the basis for the software implementation of the algorithm.

2) *The data processing problem for low-temperature experiments.* For the design of systems of cryogenic technology, of importance is the mathematical

and numerical modeling of thermophysical processes based on data of low-temperature experiments. In models and numerical calculations, as a rule, it is necessary to use the rate of change of physical fields whose values are measured with small time steps, but with considerable errors.

The important specific character of the problem is availability of additional *a priori* information on monotonicity of exact dependences. From physical considerations it follows that some of them have the form illustrated in Figure 3.5.9 (left).

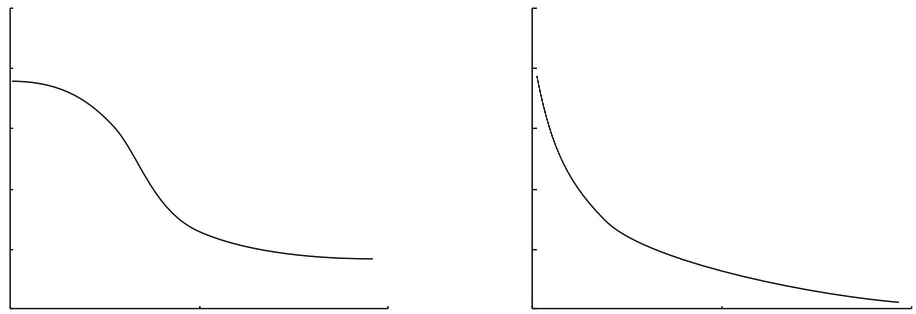

Figure 3.5.9: Dependences in some reactions

The monotonicity is disturbed by a random error whose level is known and is 10–20 %. The modified *Algorithm 3* discussed in Section 3.3 for sequential smoothings with consideration of *a priori* information was applied to restore the functions and their derivatives. It was very important that this algorithm is efficient in storage (it is proportional to the size of the original table of data) and in computing time, since the realization of the corresponding numerical model, as a rule, is carried out on a minicomputer. The basis for the software implementation of the algorithm is the subroutine SG2.

3) *The problem of calculating the rates of some chemical reactions.* Suppose it is known that a chemical reaction occurs with consumption of a substance. The concentration of this substance is measured at various time instants. The information on the concentration of the substance is obtained experimentally; therefore, it is given approximately. The shape of the curve $K(t)$ representing the exact time dependence of the concentration is known from chemical laws (see Figure 3.5.9, right). It is known that this curve is convex downwards and monotonically decreases. It is also known that at the initial instant the concentration of the substance is equal to unity: $K(t)\,|_{t=0} = 1$. The measurement errors for the concentration are about 4–5 %. It is required to compute an approximation for the derivative $dK(t)/dt$. The computations should be performed on a strongly nonuniform grid when $h_i = t_{i+1} - t_i$ considerably (by a factor of 10–20) increases with increasing $i$.

As a result of chemical experiments, another data processing problem may arise when during a chemical reaction the substance is formed and spent. In this case, the exact curve $K(t)$ has another shape. It is known that the curve is convex upwards, at the initial instant of time the precise concentration of the substance is given: $K(t)\,|_{t=0} = 0$. The measurement errors have the same level as in the previous problem. It is necessary to compute an approximation for the first derivative.

In order to solve this problem, a modification of *Algorithm 3* from Section 3.3 was applied. The algorithm was implemented as a collection of subroutines on the basis of the subroutine SG2.

4) *Some applied problems of biomechanics.* A number of applied problems in biomechanics are connected with determination of allowable (or optimal) loads (efforts) on a part of a human body or elements of a training device. Mathematically, the problem is reduced to finding the second derivative (accelerations) on the basis of data on the trajectory of center-of-mass motion of an object. The coordinates of points on the trajectory are determined by the stroboscope with some errors dependent on the measuring device and on some other factors. For the solution of these problems, the double numerical differentiation algorithms were developed on the basis of the sequential smoothing method. Let us describe the two particular problems for which these algorithms were used.

Problem 1. The process of skiing is considered. An array of equidistant-in-time points $\{t_i\}$, $i = 1, \ldots, 60$, with the step $\Delta t = 0.02$ is given. The center-of-mass heights $h_i = h(t_i)$ of a human forearm measured with a stroboscope with an error bounded by 0.05 are known at the time instants $t_i$ (the vertical projection is considered). An analysis of the calculated second derivative showed its qualitative agreement with the stages of movements made by a skier.

Problem 2. A person jumps off a chair down to the floor. An array of equidistant-in-time points $\{t_i\}$, $i = 1, \ldots, 60$, with the step $\Delta t = 0.02$ is given. The center-of-mass heights $h_i = h(t_i)$ of a human body measured with a stroboscope with an error bounded by 0.03 are known at the time instants $t_i$. Physically it is clear that at the instant of landing the acceleration sharply changes from a value close to its maximum to zero (the so-called shock motion). These peculiarities are graphically reflected on the behavior of the trajectory $h(t)$ in the form of the breakpoint corresponding to landing.

The specific character of the problem has resulted in the necessity to develop special algorithms for the sequential smoothing method on the basis of explicit spline approximation with the use of local basis splines of variable degree.

In the vicinity of jumps in the exact dependence, the local basis splines of zero degree are used, in the vicinity of breakpoints we use the first-degree splines, and on the segments of smoothness we use the local basis splines of degree $2n-1$, $n \geq 2$. The coordinates of the jump points and breakpoints are assumed to be known. The algorithm is implemented in the subroutine SGL on the basis of the sequential smoothing method with splines of variable degree. The results of solving the problem under consideration with our collection of subroutines completely correspond to the physical character of the above motion.

Let us describe some problems that lead to smoothing the functions of two variables.

5) *The problem of evaluating the thermodynamic functions of molecules.* This problem is connected with evaluation of thermodynamic functions of molecules with consideration of the contributions of internal rotation. It is also necessary to take into account the contribution of motion in a periodic potential field of a top.

In the problems of evaluating the thermodynamic functions, the data on contributions for arbitrary values of variables are obtained on the basis of approximation of functions of two variables with the aid of the Pitzer and Gwinn tables, where data are given with some errors bounded by a known quantity. For the solution of the problem, therefore, *Algorithm 2* from Section 3.3 realizing the sequential smoothing method on a rectangular grid is applied. In practice, it is sufficient to perform one or two smoothings. The algorithm was implemented as a collection of subroutines on the basis of the subroutine SG20.

6) *The problem of smoothing the electron density distribution function.* This function in a two-dimensional centrally symmetric structure is described by the formula [244]:

$$\bar{f}(x,y) = \bar{F}_{00} + 2\sum_{l=-\infty}^{+\infty}\sum_{m=1}^{+\infty}\bar{F}_{l,m}\cos 2\pi\,(lx+my) + 2\sum_{l=1}^{+\infty}\bar{F}_{l,0}\cos 2\pi lx,$$

where $\bar{F}_{l,m}$ are the exact values of structure amplitudes. As a rule, the chemical composition of a substance forming the structure is known. An analysis of the electron density distribution function is necessary to determine the mutual arrangement of atoms of various elements. The coordinates of the arrangement of atoms correspond to peaks of this function. The position of a peak is determined according to the level lines forming an electron density map.

In practice, instead of the exact structure amplitudes $\bar{F}_{l,m}$ their approximate values $F_{l,m}$ are known from experiments. Moreover, these values are known only for a finite set of indices $l$ and $m$. Hence, instead of the function

$\bar{f}(x, y)$ it is possible to evaluate only an approximate function $f(x, y)$ whose map of isolines has extra peaks and hollows with negative values. The problem of smoothing the function $f(x, y)$ consists in selection of main peaks on an electron density map and in suppression of extra peaks.

This problem was solved by the modified *Algorithm 2* of the sequential smoothing method described in Section 3.3. The values of $f(x, y)$ on a rectangular uniform grid were taken as original data. Our computations were performed with various grids for several values of $K$, where $K$ is the number of smoothings.

The subsequent visual analysis of approximate electron density maps for different variants allows one (with consideration of *a priori* information on the quantitative and qualitative compositions of the structure under study) to select a variant with the most clear positions of the main peaks.

The modification of the algorithm consisted in taking account of the fact that the exact electron density distribution function is positive; as a result, the nonphysical negative values on the calculated electron density map were eliminated. This modification is very simple algorithmically, since it is sufficient to replace negative values in the explicit formulas by zero at the stage of first smoothing.

The isogeometric properties of the explicit approximating formulas for functions of two variables guarantee the nonnegativity of splines obtained at the stages of next smoothings [84]. The periodicity of $\bar{f}(x, y)$ was also taken into account. The modified algorithm is implemented in the subroutine SG20E, which is similar to SG20. The algorithm was improved on the basis of several test problems.

A particular structure consisting of two atoms of iron and eight atoms of oxygen was taken from [244], where the table of approximate structure amplitudes for $q = 6$, $p = 7$ is given. We obtained the arrangement of atoms corresponding to that given in [244]; the resulting electron density map does not contain negative hollows. We emphasize that the computation of one variant was carried out very fast with the help of the subroutine SG20E on a BESM-6 computer. The computation of original values of $f(x, y)$ on the grid $30 \times 30$ takes 50 seconds, one smoothing takes about 1 second. Algorithms and subroutines of smoothing for similar three-dimensional problems were also created. The algorithms and subroutines we developed are of importance for a number of problems in the structural analysis.

7) *The problem of calculating the coal bed depth.* Original data are the depth values for a coal bed obtained by exploratory boring at points irregularly located in the field under exploration. The problem consisted in restoration of the coal bed surface and was solved with the help of *Algorithm 2* of Section 3.3

whose implementation was based on the subroutines IRIN21 and SG20. The irrational interpolation method is implemented in the subroutine IRIN21. This is the first stage of the algorithm. It should be noted that at this stage the approximating function has extra peaks, which deforms the map of isolines. The sequential smoothing method was applied for the reduction of distortions.

When the coal bed surface was restored, the maps of isolines were constructed for various numbers of smoothings. The map with sufficiently smooth isolines which are close to the given values at the original points was chosen.

8) *The geometric data processing problem for automatic design systems in mechanical engineering.* A number of applied problems that can be reduced to calculation of surfaces for products in mechanical engineering and for flying vehicles as well as the algorithms of their solution with the aid of splines are described in [84], [85]. More complex shapes of surfaces (arising, for example, in automobile industry and in the design of turbines) require the use of parametric splines for representation of such surfaces. Original data on a surface may be given by Cartesian coordinates of points lying on nonintersecting lines. The parametrization is performed, for example, over the total length of chords connecting the points on the lines and, in a similar way, in the direction orthogonal to the given lines [86]. This parametrization is used in the subroutines SAP13 and SAP22 suitable for restoration of surfaces of complex shape.

The above algorithms and subroutines for approximation and smoothing of functions were applied to the solution of some other problems: trajectory calculations, efficiency estimation of turbines, data processing in oceanology and geophysics. The algorithms and subroutines for approximation of functions of three variables were applied to the approximation of kernels and right-hand sides of integral equations describing some thermophysical processes [3].

The subroutines SG1, SG20 together with SAP13, SAP22, IRIN21 allow one to solve the experimental data processing problems for complex domains and irregular grids. The subroutine SAP22 can also be used for generation of grids in the finite element method or in the difference methods for solving boundary value problems. The above sequential smoothing algorithms can be used, for example, to solve the applied problems described in [61].

9) *Some problems of electromagnetic wave diffraction and an algorithm for their solution by splines.* Let us consider the plane problem of $E$-polarized electromagnetic field diffraction on ideally conducting cylindrical surfaces. This problem is described by the integral equation

$$Au \equiv \int_a^b K(x,t)u(t)\,dt = f(x), \tag{3.5.60}$$

where $u(t) = \dfrac{\varpi\mu}{2\pi} j(t)\sqrt{\varphi'^2(t) + \psi'^2(t)}$ is the sought-for function and

$$\begin{aligned} K(x,t) &= H_0^{(2)}\left[kR(x,t)\right], \\ R(x,t) &= \sqrt{[\varphi(x)-\varphi(t)]^2 + [\psi(x)-\psi(t)]^2}. \end{aligned}$$

Here $\{\varphi(t), \psi(t)\}$, $t \in [a,b]$, is a parametric representation of a smooth contour $\Gamma$ being the generator for a surface; $f(x) = E_0(x)$ is the external field on the surface; $j(t)$ is the density of the induced current; $\varpi$ is the frequency of the external field; $k$ is the wave number; $\mu$ is the magnetic permeability; $H_0^{(2)}$ is the Hankel function, $H_0^{(2)} = J_0 - iN_0$; $J_0(kR)$ is the Bessel function of zeroth order, $J_0(0) = 1$; $N_0(kR)$ is the Neumann function possessing a logarithmic singularity at zero:

$$N_0(kR)_{R\to 0} \sim \frac{2}{\pi}\ln(kR). \tag{3.5.61}$$

Equation (3.5.60) is written in complex variables; we pass to the equivalent system of equations in real variables:

$$u(t) = u_2(t) + iu_1(t), \quad E_0(x) = f_1(x) + if_2(x),$$

$$\begin{aligned} \int_a^b N_0(kR)u_1(t)\,dt + \int_a^b J_0(kR)u_2(t)\,dt &= f_1(x), \\ \int_a^b J_0(kR)u_1(t)\,dt - \int_a^b N_0(kR)u_2(t)\,dt &= f_2(x). \end{aligned} \tag{3.5.62}$$

Approximations to the functions $u_1(t)$ and $u_2(t)$ we seek in the form of the splines $S_1(t)$ and $S_2(t)$:

$$S_1(t) = \sum_{j=-n}^{N+n} c_j^1 s_{j-n,2n-1}(t), \quad S_2(t) = \sum_{j=-n}^{N+n} c_j^2 s_{j-n,2n-1}(t),$$

where $c_j^l = c_{j+N-1}^l$, $l = 1, 2$, $j = -n, \dots, n$. The undetermined coefficients $\left\{c_j^1\right\}$ and $\left\{c_j^2\right\}$ are found by the collocation method on the grid $\{x_i\}$, where the above splines are constructed.

In order to apply the general theory to the substantiation of the collocation method for the problem (3.5.62), we assume that $U$, $G$, $X = X^r$, and $Y = X^{r+1}$ are two-dimensional spaces of vector functions with the scalar product

$$(u,v) = (u_1,v_1)_{H_1} + (u_2,v_2)_{H_2}; \quad u = (u_1,u_2); \quad v = (v_1,v_2).$$

The system (3.5.62) can be represented in the form

$$\begin{aligned}
\mathbf{A}u &= \mathbf{A}_0 u + \mathbf{A}_1 u = f; \\
\mathbf{A}_0 u &= \begin{pmatrix} A_0 & 0 \\ 0 & A_0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \mathbf{A}_1 u = \begin{pmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.
\end{aligned}$$

The operator $\mathbf{A}_0$ is positive and self-adjoint in $G$, since the operator $A_0$ has these properties in $L_2[a, b]$. In the case of a closed contour, the subspace $U_N$ is the subspace of splines periodic on $[a, b]$. Let the operator $L = d^n/dx^n$; the commutativity of $\mathbf{A}_0$ with $L$ is reduced to the commutativity of $\mathbf{A}_0$ with the differentiation operator; this can be checked with integration by parts. The functionals $l_i(f) = f(x_i)$, $i = 1, \ldots, N$. The operator $\mathbf{A}_1$ is completely continuous from $X^r$ into $X^{r+1}$ if the contour $\Gamma$ is given by the functions $\varphi$ and $\psi$ possessing $(r+2)$-integrable derivatives, since in this case the operators $A_1$ and $A_2$ are completely continuous from $W_2^r[a, b]$ into $W_2^{r+1}[a, b]$. Thus, the hypotheses of Theorem 3.2.64 guaranteeing the existence and uniqueness of an approximate solution are fulfilled. The convergence of the approximate solution can be studied with the aid of the estimates of this theorem.

In conclusion we note that the above spline approximation method can be applied to special classes of ill-posed problems equivalent in their instability to the differentiation problem. Except for the problems discussed above, the Volterra- and Abel-type equations belong to the classes of such problems. The classes of problems we considered are similar in properties to the class of normally solvable ones [43], but do not coincide with them. A careful analysis of properties of the operator $A$ is necessary for the choice of an efficient method for solving the corresponding ill-posed problems.

# Chapter 4

# THE METHOD OF DESCRIPTIVE REGULARIZATION

## 4.1 The method of descriptive regularization and the quality of approximate solutions

**1.** At present, the mathematical theory of solving ill-posed problems is developed sufficiently well. The basis of this theory is the use of some *a priori* information on accuracy of input data and (or) on the sought-for solution. The character of this information may be of two kinds: quantitative or qualitative. The majority of the methods use, as a rule, the quantitative information on accuracy of input data and the rather general information on the smoothness of a solution (Tikhonov's regularization method, the residual method). A distinctive feature of Ivanov's method of quasisolutions is the possibility of using not only the quantitative information, but the qualitative information on the *a priori* knowledge of behavior of the sought-for solution as well. An objective base for the information of such a kind is, as a rule, some intuitive reasoning about the simplicity of structure of the sought-for solution as well as some general reasoning about the behavior of the physical process under study. The first kind of reasoning is connected with the natural desire of a researcher to introduce the most important and essential features into a mathematical model and may be caused by certain aesthetic reasons. The second kind of information arises, for example, in the following situation: when studying the distribution of brightness of a star, the absolutely obvious fact is the fall of intensity from the center of the star to its edges, if the star is single, and the presence of two maxima, if the star is double [76].

Here we describe a method for solving the Fredholm integral equations of the first kind on the basis of *a priori* knowledge concerning the arrangement of extrema and the points of inflection for the sought-for solution. This knowledge allows one to take into account the main qualitative regularities inherent to the exact solution.

Let us assume that the phenomenon under study is characterized quantitatively by a function $u = u(x)$, $a \le x \le b$. As a measure of its "simplicity", we may adopt some quantitative characteristics, such as a variation of $u(x)$, the mean-square value of its $k$th derivative, etc.; these characteristics are very often used to solve ill-posed problems. It is also well known that at an intuitive level the behavior of a function is quite efficiently simulated if we prescribe a possible arrangement of its characteristic points: the points of extrema and the points of curvature change. In addition, it is assumed that, on the whole, the function behaves naturally, i.e., it is single-valued, has no cuspidal points, is sufficiently smooth and, hence, may be drawn with a stroke of the pen. Such a class of simple functions may be given if we specify the segments of monotonicity and convexity.

The class of smooth functions with $L-1$ segments of monotonicity can be defined by the conditions

$$M = \left\{ u : (-1)^{l+i} u'(x) \le 0, \quad x_i \le x \le x_{i+1}, \quad i = 1, 2, \dots, L-1 \right\},$$

where $x_i$, $i = 1, 2, \dots, L$, are extrema of the function $u(x)$, $a = x_1 < x_2 < \dots < x_L = b$, and the parameter $l$ equal to 1 or 2 defines the character of monotonicity on the first segment. Thus, it is allowed that $M = M(x_2, \dots, x_{L-1}; l, L)$, i.e., the number of extrema may be changed, the segments where the function increases and decreases may alternate, and the arrangement of internal extrema may vary. A more detailed specification of this class of functions can be performed if we select the segments where the function's curvature is of constant sign. As a result, we come to the class

$$V = \left\{ u : (-1)^{v+j} u''(x) \le 0, \quad \bar{x}_j \le x \le \bar{x}_{j+1}, \quad j = 1, 2, \dots, N-1 \right\},$$

where $\bar{x}_j$, $j = 2, 3, \dots, N-1$, are the points of inflection of the function $u(x)$, $a = \bar{x}_1 < \dots < \bar{x}_N = b$, the parameter $v$ equal to 1 or 2 defines the sign of curvature on the first segment, and $N-1$ is the number of segments where the curvature is of constant sign. It is allowed that $V = V(\bar{x}_2, \dots, \bar{x}_{N-1}; v, N)$.

The points $x_i$ and $\bar{x}_j$ should satisfy the following matching conditions. If $x_1 < \bar{x}_2 < x_2$, it is necessary that $x_2 < \bar{x}_3 < x_3, \dots, x_{L-2} < \bar{x}_{L-1} < x_{L-1}$, whereas on the last segment $[x_{L-1}, b]$ the curvature may change its sign (then, $x_{L-1} < \bar{x}_{L-1} < b$, $N = L+1$ ) or may preserve its sign (then, $\bar{x}_L = b$, $N = L$).

If $x_2 < \bar{x}_2 < x_3$, i.e., on the segment $[x_1, x_2]$ the curvature is constant, it is necessary that $x_3 < \bar{x}_3 < x_4, \ldots, x_{L-2} < \bar{x}_{L-2} < x_{L-1}$. If the curvature does not change, then on the last segment it should be $\bar{x}_{L-1} = x_L$, i.e., $N = L - 1$; otherwise, it should be $x_{L-1} < \bar{x}_{L-1} < x_L$ and, hence, $\bar{x}_L = b$, $N = L$.

These conditions correspond to intuitive reasoning about the simplicity of behavior of a function from the class under consideration and take place in many applied problems.

Let us consider the following Fredholm integral equation:

$$Ku \equiv \int_a^b k(x, y)\, u(y)\, dy = f(x), \quad c \le x \le d.$$

Here $f \in L_2[c, d]$, $k \in L_2(\Pi)$, and $\Pi = [c, d] \times [a, b]$. The methods for solving such equations with taking account of the conditions $M$ and (or) $V$ are called *the methods of descriptive regularization* [71]. Now we consider the following problem of mathematical programming: to find a function $u^*(y)$ which is an approximate solution to the problem

$$u^* = \arg\min_{u \in D} \|Ku - f\|_{L_2},$$

where $D = M$, $V$ or $M \cap V$. The kernel and the right-hand side of the equation being considered may be given approximately.

A peculiarity of the descriptive regularization method is the description of a feasible set of functions by selection of characteristic points (the points of extrema and the points where the curvature changes its sign) and the assumption on the simplicity of structure of the sought-for solution in the above sense.

It is known [80] that the conditions $M$ stabilize an approximate solution in uniform metric. Taking account of the convexity conditions, as shown below, provides the stabilization of first-order smoothness, i.e., under certain conditions, the uniform convergence of approximate solutions together with their derivatives takes place.

It is easy to see that the numerical discretization of a problem leads to a problem of quadratic programming with specific linear constraints. We pay our main attention to the construction of efficient numerical algorithms for solving a discrete problem with consideration of its specific character and to testing their performance on model problems. We are also interested in the quality of approximations defined by their nature. Note that the quality of approximations is defined not only by their accuracy, but basically by manifestation of main laws inherent to the exact solution and, in many respects, depends on subjective analysis of these approximations.

**2.** Now we formulate a discrete problem. Let $\varpi_x^h = \{x_i\}$, $i = \overline{1,m}$, be a grid on the segment $[c,d]$ and $\varpi_y^h = \{y_j\}$, $j = \overline{1,n}$, be a grid on the segment $[a,b]$. Putting $u(y_j) = u_j$, we can write

$$Ku \approx \sum_{j=1}^{n} q_j k(x, y_j) u_j = f(x),$$

where $q_j$ are quadrature weights (in the sequel, for the trapezoid formula). Putting here $x = x_i$, we come to the following collocation conditions:

$$A\hat{u} = \hat{f}. \tag{4.1.1}$$

Here $A$ is the $m \times n$-matrix with elements $a_{ij} = q_j k(x_i, y_j)$, $\hat{u} = (u_1, u_2, \ldots, u_n)'$, $\hat{f} = (f_1, \ldots, f_m)'$, and $f_i = f(x_i)$. As an objective functional we choose

$$\Phi[\hat{u}] \equiv \Phi\left[\hat{u}; \hat{f}\right] = \sum_{i=1}^{m} p_i \left(\sum_{j=1}^{n} a_{ij} u_j - f_i\right)^2 \equiv \left\| A\hat{u} - \hat{f} \right\|_P^2,$$

where $p_i > 0$ are some weighting coefficients.

Writing the condition of monotonicity and convexity for the discrete functions $\hat{u}$ and preserving the previous notation for them, we come to the following problem of mathematical programming:

$$\Phi[\hat{u}] \to \min_{\hat{u} \in D} \tag{4.1.2}$$

with specific constraints in the form of linear inequalities.

As a numerical algorithm for solving the problem (4.1.2), we choose the conjugate gradient projection method. Our choice is based on the well-known fact that for quadratic functionals without constraints this method (belonging to the class of quasi-Newton methods) converges to the sought-for solution after at least $r$ steps, where $r$ is the rank of the matrix $A$. In this case we have the situation similar to that when the "practical" rank of the problem substantially less than $n$, i.e., when $r$ is sufficiently small.

This conclusion is based on the following. In practice, any regularization method leads to a reduction of influence of high-frequency components in the solution expressed in some special basis; this is equivalent to the replacement of the original problem by a similar one which is singular and has a small rank. It is known from practice that singular problems may be solved numerically more efficiently than the ill-conditioned (or badly-"organized") problems of large dimension. Our numerical experiments described below are in good agreement with this observation, since the actual number of iterations is not large.

Another reason we are guided by is a substantial reduction in computing cost and an increase in accuracy of computations because of efficient implementation of the projection operation. This can be achieved for sets $D$ of the type $M$ or $V$ by application of special algorithms developed for these cases in [70] and for the set $D = M \cap V$ by sequential use of additional information [191].

Let us consider the formulas for the conjugate gradient projection method in the case of the problem (4.1.2).

Let $u_0 \in D$ be an initial approximation. The iterative process is constructed according to the following formulas:

$$\begin{aligned} u^{s+1} &= \mathcal{P}_D\left(u^s - \alpha_s g^s\right), \quad s = 0, 1, \dots, \quad g^0 \equiv \operatorname{grad}\Phi\left[u^0\right], \\ g^s &= \operatorname{grad}\Phi\left[u^s\right] - \beta_s g^{s-1}, \quad s = 1, 2, \dots, \\ \beta_s &= \frac{\left(\operatorname{grad}\Phi\left[u^s\right], \operatorname{grad}\Phi\left[u^{s-1}\right] - \operatorname{grad}\Phi\left[u^s\right]\right)}{\left\|\operatorname{grad}\Phi\left[u^{s-1}\right]\right\|^2}. \end{aligned} \tag{4.1.3}$$

Here $\mathcal{P}_D$ is the operator of projection onto the set $D$:

$$\left\|\mathcal{P}_D(z) - z\right\| = \min_{\varpi \in D} \|\varpi - z\|$$

and $\|\cdot\|$ is the root-mean-square norm in a finite-dimensional space. A value of the descent step $\alpha_s$ is chosen from the monotone decrease condition for the objective functional:

$$\Phi\left[u^s - \alpha_s g^s\right] \le \Phi\left[u^s\right]. \tag{4.1.4}$$

Using the formulas

$$\alpha_s^* = \frac{\left(\operatorname{grad}\Phi\left[u^s\right], g^s\right)}{2\left(A'PAg^s, g^s\right)}, \quad \Phi\left[u^s - \alpha_s^* g^s\right] = \min_{\alpha \ge 0} \Phi\left[u^s - \alpha g^s\right],$$

we put $\alpha_s = \alpha_s^*$ when the condition (4.1.4) is fulfilled and $\alpha_s = \alpha_s^*/2^k$ ($k = 1, 2, \dots$) otherwise.

When the iterative process (4.1.3) is constructed in a finite-dimensional analogue of the space $L_2$, the gradient of the objective functional takes the form

$$\operatorname{grad}\Phi[u] = 2\left(A'PAu - A'Pf\right), \tag{4.1.5}$$

where the prime indicates the transposition operation and $P = \operatorname{diag}\{p_1, \dots, p_m\}$.

If it is known *a priori* that the solution has, for example, the first-order smoothness, then an increase in accuracy and an increase of quality of the solution can be achieved by applying the conjugate gradient method in a finite-dimensional analogue of the space $W_2^1$. In this case, $\operatorname{grad}\Phi[u] = \left(\bar{\Phi}_1, \bar{\Phi}_2, \ldots, \bar{\Phi}_n\right)$, where $\bar{\Phi}_j$, $j = 1, 2, \ldots, n$, are determined from the linear system

$$\frac{\rho_j \bar{\Phi}_{j+1}}{\gamma_j} - \left(1 + \frac{\rho_j}{\gamma_j} + \frac{\rho_{j-1}}{\gamma_j}\right)\bar{\Phi}_j + \frac{\rho_{j-1}\bar{\Phi}_{j-1}}{\gamma_j} = -\Phi_j,$$
$$j = 1, 2, \ldots, n, \quad \bar{\Phi}_0 = \bar{\Phi}_{n+1} = 0, \quad \rho_0 = \rho_n = 0,$$

which can be solved by the sweep method. Here $\Phi_j$ are the components of the vector (4.1.5), $\gamma_j > 0$ and $\rho_j \geq 0$ are given numbers associated with normalization in a discrete analogue of the space $W_2^1$.

**3.** In order to check the efficiency of the above method, a FORTRAN subroutine was implemented. Several numerical experiments were conducted for a number of model problems with known exact solutions.

*Model problem 1*: to solve the problem (4.1.2) in the class of piecewise monotone functions ($D = M$), when the exact solution is $u_{\mathrm{ex}} = 1 - y^2$, $k(x,y) = 1/\left(1 + (x-y)^2\right)$, and

$$f(x) = \left(2 - x^2\right)\left(\arctan(1-x) + \arctan(1+x)\right) - 2 - x\ln\frac{1 + (1-x)^2}{1 + (1+x)^2}.$$

*Model problem 2*: to solve the problem (4.1.2) in the class $D = M$, when the exact solution is $u_{\mathrm{ex}} = \sin\frac{\pi}{2}y$ and $k(x,y) = (x-y)^2$.

The computations were performed on the uniform grids $\varpi_x^h \in [-2, 2]$ and $\varpi_y^h \in [-1, 1]$ with the steps $h_x = 0.1$ and $h_y = 0.05$, respectively. The function $u^0 \equiv 0$ was taken as an initial approximation. The explicit form of the projector onto the set of piecewise monotone functions described in [238] was used to construct the projections onto $M$. In order to check the stability of the method with respect to errors in input data, a number of numerical experiments were performed with uniformly distributed random perturbations of the kernel $k$ and the right-hand side $f$.

Some of numerical results are represented in Figures 4.1.1 and 4.1.2. It follows from these figures that the consideration of the constraints $M$ and the choice of a space metric (the parameters $\rho_j$, $j = \overline{1, n-1}$) allow one to obtain an approximate solution close to the exact one.

*Model problem 3*: to solve the problem (4.1.2) in the class of convex functions, when the exact solution is $u_{\mathrm{ex}} = \cos\frac{\pi}{2}y$, $k(x,y) = x - y$, $f(x) = 4x/\pi$, $x \in [-2, 2]$, $y \in [-1, 1]$, $h_x = 0.1$, and $h_y = 0.05$.

The solution to this problem is represented in Figure 4.1.3.

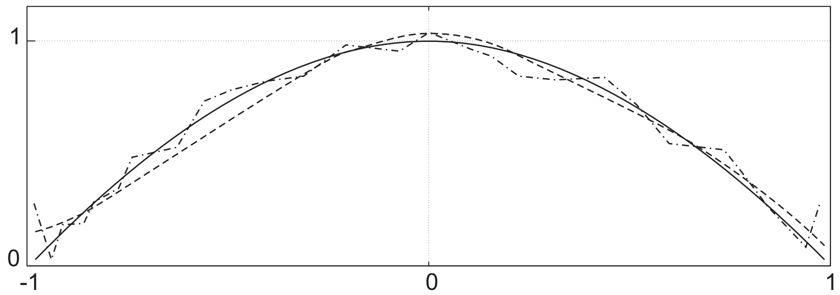


Figure 4.1.1: Numerical results for the model problem 1 with errors 5 % in the kernel and errors 3 % in the right-hand side; the solid curve — exact solution; the dot-dashed curve — the approximate solution with $\rho_i = 0$, $i = 1, \ldots, 40$, without taking account of constraints; the dashed curve — the approximate solution with $\rho_i = 1$, $i = 1, \ldots, 40$, with taking account of conditions $M$

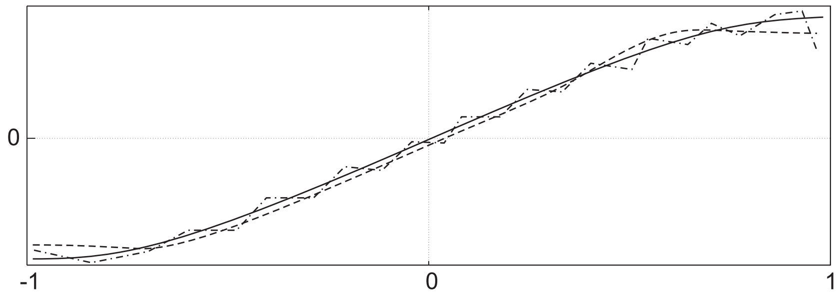


Figure 4.1.2: Numerical results for the model problem 2 with errors 5 % in the kernel and errors 2 % in the right-hand side; the solid curve — the exact solution; the dot-dashed curve — the approximate solution with $\rho_i = 0$, $i = 1, \ldots, 40$, without taking account of constraints; the dashed curve — the approximate solution with $\rho_i = 1.5$, $i = 1, \ldots, 38$, $\rho_i = 0.5$, $i = 39, 40$, with taking account of conditions M

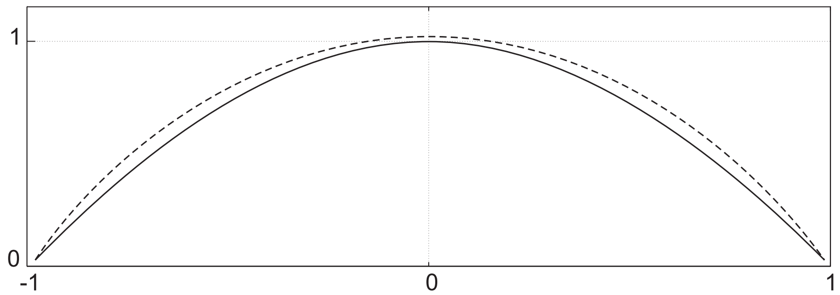


Figure 4.1.3: Numerical results for the model problem 3; the solid curve — the exact solution; the dashed curve — the approximate solution with $\rho_i = 0$, $i = 1, \ldots, 40$, with taking account of conditions $V$

The values of the functional $\Phi$ in the two next iterations coincide with the accuracy $10^{-7}$ for each of the above problems after 10 iterations (this corresponds to one minute of computing time on a BESM-6 computer together with compilation of the program).

If in the descriptive regularization method the result obtained after several iterations by the conjugate gradient method without any constraints is taken as an initial approximation, then we can achieve even more substantial reduction in computing time.

Our direct comparison of Figures 4.1.1, 4.1.2, and 4.1.3 shows that taking account of the $M$-type conditions only is insufficient for consideration of approximations as qualitative in full measure, although the accuracy of these approximations is quite satisfactory. The approximations in Figures 4.1.1 and 4.1.2 have a stepwise character (this fact is quite consistent with the conclusions obtained in [193]). At the same time, taking account of the conditions $V$ leads to the qualitative reconstruction of the exact solution (Figure 4.1.3). Probably, this conclusion is connected not only with the above particular examples, but has a sufficiently general nature.

Note that when considering the conditions $M$, the improvement of perception of approximations is also reached by transition to descent along the conjugate gradient in another space which takes into account the smoothness of the sought-for function. The same effect can be achieved if we take into account that the method of descriptive regularization, as a method of quasisolutions in general [115], is a limiting case of Tikhonov's regularization method

$$\|Ku - f\|^2 + \alpha \left\|u^{(k)}\right\|^2 \to \min_D,$$

when the regularization parameter $\alpha > 0$ is sufficiently small, but its influence as a stabilizing factor is still preserved.

Our analysis of computing cost (storage and computing time) necessary for solving a problem by the descriptive regularization method shows that this computing cost is comparable with that needed for minimization of finite-dimensional quadratic problems without constraints by the conjugate gradient method. Hence, this method is quite suitable for the use in practice.

The application of the descriptive regularization method is especially efficient in the case when the operator $K$ is unit, i.e., when the problem of approximation on the set of piecewise monotone functions is considered.

The main conclusions from the above results can be generalized without changes to the case of nonlinear integral equations; it follows from the above discussion that it is not necessary to prescribe an explicit form of the operator. This allows us to recommend the descriptive regularization method for a wide range of inverse problems in the one-dimensional case. This method may efficiently be applied with the use of a computer display as a trial method. Thus, we may speak of a new type of a "solver" whose main elements are a computer performing the most routine part of work, a display enabling to analyze and to make a decision, and a computer operator who forms a new model to be studied on the basis of data obtained.

## 4.2 Peculiarities of numerical implementations of the descriptive regularization method

**1.** An advantage of Tikhonov's regularization method is the simplicity of its implementation and the asymptotic (with respect to perturbations of input data) stability of approximate solutions. In practice, however, such properties are not always sufficient, as stated in the previous section, to judge the quality of approximate solutions and their conformity to the exact solution. The essence consists not only in the asymptotic convergence of the method, but also in the nature of this convergence.

The last means that the algorithm should preserve the main qualitative characteristics of the sought-for solution: monotonicity, convexity, the presence of extrema at certain points, etc.; these characteristics are usually known from *a priori* considerations during formulation of a problem. Taking account of these requirements results in the qualitative improvement of approximate solutions and, sometimes, stabilizes them to an extent that the algorithm of searching for a solution may be considered as regularizing.

This situation becomes apparent in the best way when the methods of mathematical programming are applied for solving ill-posed problems. Of prime importance is the description of major qualitative characteristics of the sought-for solution. For this reason, the above numerical methods of regularization are said to be descriptive. Note that, apparently, this approach was first recommended in [191] to solve integral equations of the first kind in the framework of the pointwise residual method and in [197] to solve the problem of construction of splines. In the sequel, the methods of descriptive regularization and the questions of their implementation were considered in various aspects in [192], [193], and [272].

Here we consider the application of the descriptive regularization method to solving the ill-posed problems for which *a priori* information on their solutions can be written down with the aid of linear inequalities. The problems of such a type arise in the case of numerical solution of the Fredholm integral equation of the first kind

$$A\,[x;u] \equiv \int_a^b k(x,y)u(y)\,dy = f(x), \quad c \le x \le d, \tag{4.2.6}$$

when *a priori* information on boundedness, monotonicity, and convexity of its solution on subsegments of the segment $[a, b]$ is known. Let $\varpi_{h_x} = \{x_j\}$, $j = 1, \ldots, M$, be a grid on the segment $[c, d]$ and $\varpi_{h_y} = \{y_j\}$, $j = 1, \ldots, N$, be a grid on the segment $[a, b]$. The integral in (4.2.6) can be approximated with a quadrature formula. Hence, we can pass to the consideration of a discrete analogue of equation (4.2.6):

$$\hat{A}\hat{u} = \hat{f}. \tag{4.2.7}$$

Here $\hat{A}$ is an $M \times N$ matrix, $\hat{u} \in R_N$, and $\hat{f} \in R_M$. Additional information on the qualitative behavior of an approximate solution $\hat{u} = (u_1, \ldots, u_N)'$ on the grid $\varpi_{h_y}$ can be represented by a system of linear constraints. For example, the monotonicity conditions on each individual subsegment of the segment $[a, b]$ take the form

$$\begin{aligned} (-1)^l(u_i - u_{i+1}) &\le 0, \quad i \in \overline{i_1, i_2 - 1}, \\ (-1)^{l+1}(u_i - u_{i+1}) &\le 0, \quad i \in \overline{i_2, i_3 - 1}, \\ (-1)^{l+L-2}(u_i - u_{i+1}) &\le 0, \quad i \in \overline{i_{L-1}, i_L - 1}, \end{aligned} \tag{4.2.8}$$

where $(L - 1)$ is the number of subsegments of monotonicity and $l > 0$ is an integer number characterizing the type of monotonicity (nondecrease or nonincrease) on the first subsegment. In matrix form, these conditions can be written down as $Su \le 0$, where $S$ is an $(N - 1) \times N$ two-diagonal matrix.

The conditions of sign-definiteness for the second derivative of the sought-for solution (more exactly, the conditions of its convexity or concavity) on the subsegments of the segment $[a, b]$ can be written down in the form

$$(-1)^m(du_i) \le 0, \quad i \in \overline{j_1, j_2 - 1}, \quad (-1)^{m+1}(du_i) \le 0, \quad i \in \overline{j_2,, j_3 - 1}, \quad \ldots,$$
$$(-1)^{m+k-2}(du_i) \le 0, \quad i \in \overline{j_{k-1},, j_k - 1},$$
$$(du_i) \equiv (u_{i+1} - u_i)/h_i - (u_i - u_{i-1})/h_{i-1}, \quad i \in \overline{2, N - 1},$$

where $k - 1$ is the number of subsegments of sign-definiteness for curvature, $m > 0$ is an integer number characterizing the sign of curvature on the first subsegment, and $h_i = y_{i+1} - y_i$ is the step of the grid $\varpi_{h_y}$. In matrix form, these conditions can be represented as $Du \le 0$, where $D$ is an $(N - 2) \times N$ tridiagonal matrix.

Additional *a priori* information on the solution (for example, on its symmetry, nonnegativity, etc.) can be represented in matrix form in a similar way.

A discrete problem in a more general case may be formulated as follows. Let the ill-conditioned linear system (4.2.7) approximate equation (4.2.6) with the constraints

$$S\hat{u} \le 0, \quad D\hat{u} \le 0, \quad \hat{u} \ge 0. \tag{4.2.9}$$

The $(N-1) \times N$-matrix $S$ is uniquely defined by the specification of integers $L > 0$ and of a set of the extreme points for the solution: $I_L = \{i_1, \ldots, i_L\}$, $i_1 = 1$, $i_L = N$. The $(N-2) \times N$-matrix $D$ is uniquely defined by the specification of the grid $\varpi_{h_y}$, of integers $K > 0$, and of a set of the points of inflection for the solution: $J_k = \{j_1, \ldots, j_K\}$, $j_1 = 1$, $j_K = N$. The conditions (4.2.9) specify the "natural" behavior of the sought-for discrete solution (without the cuspidal points).

As a regularizing algorithm, we consider the choice of an approximate solution to equation (4.2.6) as a grid function satisfying the restrictions (4.2.9) and the inequality

$$\Phi\left[\hat{u}; \tilde{f}\right] \le \delta^2, \tag{4.2.10}$$

where

$$\Phi\left[\hat{u}; \tilde{f}\right] \equiv \left(A\hat{u} - \tilde{f}, A\hat{u} - \tilde{f}\right)_P = \left(P\left(A\hat{u} - \tilde{f}\right), A\hat{u} - \tilde{f}\right),$$

$\tilde{f}$ is a grid $\delta$-approximation to the exact right-hand side, i.e., $\|f - \tilde{f}\|_P \le \delta$, $P = \operatorname{diag}(p_1, p_2, \ldots, p_M)$ is the matrix of weighting coefficients (in particular, these coefficients take into account the nonuniformity of the grid $\varpi_{h_x}$), and

$$(v, g)_P \equiv \sum_{i=1}^{M} p_i v_i g_i, \quad v, g \in R_M.$$

The problem of constructing such an approximate solution can be reduced to the following problem of quadratic programming with linear constraints:

$$\Phi\left[\hat{u}; \tilde{f}\right] \to \min_{\hat{u}}, \quad \hat{u}: \ S\hat{u} \le 0, \quad D\hat{u} \le 0, \quad \hat{u} \ge 0. \tag{4.2.11}$$

Solving this problem with the use of minimization methods (whose survey can be found, for example, in [48]), we construct a minimizing sequence until a discrete function satisfying (4.2.10) be defined.

In the previous section, the conjugate gradient projection method was used as an efficient method for minimizing the functional $\Phi$ on the set of monotone functions. Here we study an algorithm for solving the problem (4.2.11) on the basis of information about the qualitative behavior of the solution. Our subroutine implementing the conjugate gradient projection algorithm is described and some numerical results we obtained are discussed.

**2.** *A priori* information on the qualitative character of the sought-for solution to the problem (4.2.6) (its sign-definiteness as well as the sign-definiteness of its first and second derivatives) can be used sequentially as this character is refined. By the example of problem (4.2.11), below we consider one of the possible ways of such a usage of additional information.

At the first stage, the minimization problem

$$\Phi\left[\hat{u};\tilde{f}\right] \to \min_{\hat{u}} \tag{4.2.12}$$

is considered without taking account of constraints on the sought-for solution or only with taking account of its sign-definiteness.

For example, the gradient method

$$u^{s+1} = u^s - \alpha_s \operatorname{grad}\Phi\left[u^s; f\right], \quad s = 0, 1, \dots ,$$

may be used to solve the problem (4.2.12) approximately. If the minimization of the functional is carried out on the set of sign-definite functions, the account of these constraints is taken by projection. For example,

$$u^{s+1} = \mathcal{P}_{u\geq 0}\left(u^s - \alpha_s \operatorname{grad}\Phi\left[u^s; f\right]\right)$$

if $u$ is nonnegative.

At the second stage, the refinement of the approximate solution obtained is made by introducing the additional information on its character, i.e., the problem (4.2.11) is considered with taking account of all constraints. The iterative Pshenichny method [226] is applied as a method of solving. The solution obtained at the first stage is used to construct an initial approximation.

The Pshenichny method belongs to the class of feasible direction methods, i.e., all successive iterations satisfy the constraints and the direction of minimization is chosen in such a way that this direction would not cause the violation of constraints and be a direction along which the minimizing functional decreases.

In brief, the general scheme of the Pshenichny method consists in the following. Suppose we seek a solution to the finite-dimensional problem

$$\varphi(x) \to \min_x, \quad \left(a^i, x\right) \leq b_i, \quad i = 1, \dots, m, \quad x \in Q \subset R_n,$$

where $a^i \in R_n$, $b_i$ are given numbers, $\varphi$ is a quadratic function, and $Q \subset R_n$ is a convex set.

At a point $x^s \in Q$ we find $\varepsilon$-active constraints such that they become inequalities

$$I_{\varepsilon_s} = \left\{i : \left(a^i, x\right) \geq b_i - \varepsilon_s\right\}$$

with the accuracy $\varepsilon_s \geq 0$. Next we solve the following auxiliary problem of minimizing the quadratic form with simple constraints for the dual variables $\lambda_i$:

$$\left\|\operatorname{grad}\varphi\left(x^s\right) + \sum_{i\in I_{\varepsilon_s}} \lambda_i a^i\right\|^2 \to \min_\lambda, \quad \lambda_i \geq 0.$$

Its solution $\lambda_i^s$, $i \in I_{\varepsilon_s}$, is used to construct the new approximation

$$x^{s+1} = \mathcal{P}_Q\left(x^s - \alpha_s\left(\operatorname{grad}\varphi(x^s) + \sum_{i\in I_{\varepsilon_s}} \lambda_i a^i\right)\right).$$

Here $\mathcal{P}_Q$ is the operator of projection onto the set $Q$ and $\alpha_s > 0$ is the step of descent along the above-found direction (this step is chosen in such a way that $\varphi\left(x^{s+1}\right) \le \varphi(x^s)$ and all the constraints be valid).

The process is assumed to be complete if

$$0 \le \left(\operatorname{grad}\varphi(x^s),\ \operatorname{grad}\varphi(x^s) + \sum_{i\in I_{\varepsilon_s}} \lambda_i a^i\right) \le \varepsilon_s,$$

where $\varepsilon_s < \varepsilon$ and $\varepsilon > 0$ are given numbers. If the last condition is not fulfilled, then we put $\varepsilon_s := \varepsilon_s/2$.

Unlike the feasible direction method proposed in [106], where the auxiliary problem of choosing the descent direction is considered, here the dual problem is solved. In general, this method converges with a geometric progression rate.

Before making the general scheme of Pshenichny's algorithm more specific for the problem under consideration, we represent the problem (4.2.11) as follows:

$$\Phi(U) \to \min_U, \quad BU \le 0, \quad CU \le 0, \quad U \ge 0.$$

Here $\Phi(U) = \left(\tilde{A}U - \tilde{F}, \tilde{A}U - \tilde{F}\right)_{\tilde{P}}$, $U = \{\hat{u}, \hat{u}\} \in R_{2N}$, $\tilde{F} = \left\{\tilde{f}, \tilde{f}\right\} \in R_{2M}$.

The matrices $\tilde{A}$, $B$, $C$, $\tilde{P}$ have the dimensions $2M \times 2N$, $(2N-3) \times 2N$, $N \times 2N$, $2M \times 2M$, respectively, and are of the form

$$\tilde{A} = \begin{pmatrix} \mathrm{A} & 0 \\ 0 & \mathrm{A} \end{pmatrix},\ B = \begin{pmatrix} \mathrm{S} & 0 \\ 0 & 0 \end{pmatrix},\ C = \begin{pmatrix} \mathrm{I} & \mathrm{I} \end{pmatrix},\ \tilde{P} = \begin{pmatrix} \mathrm{P} & 0 \\ 0 & \mathrm{P} \end{pmatrix},$$

where $I$ is the unit matrix of order $N$.

In our case, the set of $\varepsilon$-active constraints becomes

$$I_\varepsilon = \left\{i \in \overline{1, 2N-3} : \left(b^i, U\right) \ge -\varepsilon\right\},$$

where $b^i \in R_{2N}$ is the $i$th row of the matrix $B$.

First we consider the dual problem. In order to find a feasible direction, the following minimization problem for dual variables (the Lagrange multipliers) is solved:

$$J(\lambda) \to \min_\lambda, \quad J(\lambda) \equiv \left\|\operatorname{grad}\Phi(U) + \sum_i b^i \lambda_i\right\|, \quad \lambda \ge 0, \tag{4.2.13}$$

where $\lambda = \{\lambda_1, \dots, \lambda_{2N-3}\}$ and $\lambda_i = 0$ for $i \in I_\varepsilon$.

The gradient of the functional $\Phi(U)$ has the form

$$\operatorname{grad} \Phi\left(U\right) \equiv \left\{\operatorname{grad} \Phi\left[\hat{u}; \hat{f}\right], \operatorname{grad} \Phi\left[\hat{u}; \hat{f}\right]\right\}.$$

A solution to the extremal problem (4.2.13) is determined from the condition

$$\operatorname{grad} J\left(\lambda\right) = 0, \quad \lambda \geq 0,$$

i.e., for $i \in \overline{1, 2N-3}$ and $j \in \overline{1, 2N-3}$ we have

$$\sum_j \left(b^i, b^j\right) \lambda_j = \left(\operatorname{grad} \Phi\left(U\right), b^i\right), \quad \lambda_j \geq 0. \tag{4.2.14}$$

The system (4.2.14) can be rewritten in matrix form as follows:

$$\begin{aligned} T\lambda &= v, \quad v = (v_1, \dots, v_{2N-3})', \\ v_i &= -\left(\operatorname{grad} \Phi\left(U\right), b^i\right), \ T = \begin{pmatrix} \mathrm{T}_1 & 0 \\ 0 & \mathrm{T}_2 \end{pmatrix}. \end{aligned}$$

Based on the form of the matrices $S$ and $D$, it is easy to establish that $T_1$ is a tridiagonal matrix of order $(N-1)$ and $T_2$ is a five-diagonal matrix of order $(N-2)$. Thus, the system (4.2.14) can be represented in the form of two independent systems whose numerical solutions may be found by the sweep method. The stability of this method follows from the positive definiteness of the corresponding matrices.

At the final stage of solving the dual problem, we put

$$\lambda = \begin{cases} 0, \ \lambda_i < 0, \\ \lambda_i, \ \lambda_i \geq 0, \end{cases} \qquad i = \overline{1, 2N-3},$$

and use the above-found Lagrange multipliers to construct the vector $\varpi = (\varpi_1, \varpi_2, \dots, \varpi_{2N})$, $\varpi = \sum_{i=1}^{2N-3} \lambda_i b^i$.

Let us consider the problem for the direct variables $U$.

One of the possible ways used to satisfy to the constraint $CU \leq 0$ is the requirement

$$C\varpi = 0, \quad \text{i.e.,} \ \ \varpi_i = \varpi_{N+i}, \ \ i \in \overline{1, N}. \tag{4.2.15}$$

This requirement is fulfilled if, for example, we put

$$\varpi_i := \varpi_{N+i} := \frac{\varpi_i + \varpi_{N+i}}{2}, \quad i \in \overline{1, N}.$$

Assuming that the equality (4.2.15) is fulfilled, we consider the construction of iterations in the space of the initial variables $\hat{u} \in R_N$.

Let $\hat{u}^s \in R_N$ be an approximation satisfying the constraints (4.2.9). The next approximation $\hat{u}^{s+1}$ is defined as

$$\hat{u}^{s+1} = \mathcal{P}_{\hat{u}^s \geq 0}\left(\hat{u}^s - \alpha_s\left(\operatorname{grad}\Phi\left[\hat{u}^s; \tilde{f}\right] + \varpi^s\right)\right),$$

where $\varpi^s = (\varpi_1^s, \ldots, \varpi_N^s)$ is the direction vector found as a result of solving the dual problem for the point $\hat{u}^s$.

In the case of a quadratic functional, the step $\alpha_s$ can be found in the explicit form:

$$\alpha_s = \frac{\left(\operatorname{grad}\Phi\left[\hat{u}^s; \tilde{f}\right], V^s\right)}{\left(\Phi''\left[\hat{u}^s; \tilde{f}\right] V^s, V^s\right)},$$

where

$$V^s \equiv \operatorname{grad}\Phi\left[\hat{u}^s; \tilde{f}\right] + \varpi^s, \quad \Phi''\left[\hat{u}^s; \tilde{f}\right] \equiv 2A'PA.$$

If the constraints (4.2.9) are not valid for $\hat{u}^{s+1}$ with an accuracy $\varepsilon_s$ or the functional does not decrease, then the step $\alpha_s$ is halved.

**3.** In order to study the performance of the above algorithm, N. L. Gol'dman implemented a FORTRAN subroutine and solved a number of model problems. Some of them were the model problems 1 and 2 from the previous section with the same input data and the problem (4.2.6)–(4.2.8) with the following model data:

$$\begin{aligned}
u_e(y) &= 1 - y, \;\; k(x,y) = 1/\left(1 + \psi^2 (x-y)^2\right), \;\; \psi = 10, \\
a &= 0, \;\; b = 1, \;\; c = -2, \;\; d = 2; \\
f(x) &= \frac{1}{\psi^2}\{\psi(1-x)(\arctan\psi(1-x) + \arctan\psi x) - \\
&\quad -\frac{1}{2}\ln\left[\left(1 + \psi^2(1-x)^2\right)/\left(1 + \psi^2 x^2\right)\right]\}.
\end{aligned}$$

The computations were performed on the uniform grids $\varpi_{h_x} = \{x_i\}$, $i = \overline{1,41}$, $\varpi_{h_y} = \{y_j\}$, $j = \overline{1,41}$, with the steps $h_x = 0.01$, $h_y = 0.025$ for the model problem (4.2.6)–(4.2.8) and with the steps $h_x = 0.1$, $h_y = 0.05$ for the model problems 1 and 2.

Some numerical results for the model problem (4.2.6)–(4.2.8) are represented in the Figure 4.2.4 (left). The functional being minimized has the same form as in the previous section.

The information on qualitative behavior of the solution was introduced sequentially. The function $\hat{u}^0 \equiv 0.5$ was taken as an initial approximation for the zeroth iteration of the gradient descent on the set of nonnegative functions. The value of the functional decreased from $\Phi\left[\hat{u}^0; \tilde{f}\right] = 2.9 \cdot 10^{-2}$ down to

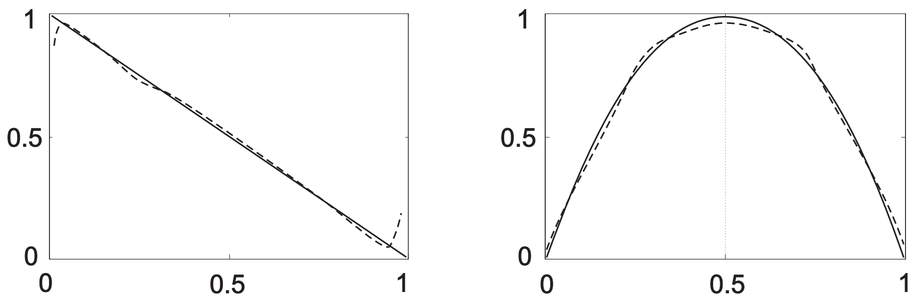


Figure 4.2.4: 1) the model problem (4.2.6)–(4.2.8); $M = 41$, $N = 41$; the solid curve — the exact solution; the dashed curve — $\hat{u}^{33}$ (the gradient method); 2) the model problem 1; $M = 41$, $N = 21$; the solid curve — the exact solution; the dashed curve — $\hat{u}^8$

$\Phi\left[\hat{u}^{33}; \tilde{f}\right] = 1.2 \cdot 10^{-6}$ after 33 iterations; the two next iterations coincided with the given accuracy $\varepsilon = 10^{-8}$. The resulting approximation $\hat{u}^{33}$ was projected then onto the set of decreasing convex functions and the projection obtained was taken as a zeroth iteration for Pshenichny's method. The residual for $\hat{u}^{33}$ was equal to $2.8 \cdot 10^{-5}$. After 12 iterations of Pshenichny's method, the residual decreased down to $1.6 \cdot 10^{-6}$; the corresponding approximation satisfied all constraints with a given accuracy ($\varepsilon = 10^{-4}$ for the condition of monotonicity, $\varepsilon = 10^{-2}$ for the condition of convexity) and was different from the exact solution on the grid $\varpi_{h_y}$ by no more than 2 %. The computing time on a BESM-6 computer was approximately equal to 1.5 minutes.

Some numerical results for the model problems 1 and 2 without account of additional information are represented in Figure 4.2.4 (right) and in Figure 4.2.5.

Similar numerical experiments with the sequential introduction of additional information on the solution were carried out for the model problem 1 with $x \in [-2, 2]$ and $y \in [-1, 1]$.

The approximate solution to the model problem 1 from the previous section obtained after 19 iterations of gradient descent in the class of nonnegative functions ($\Phi\left[\hat{u}^0; \tilde{f}\right] = 1.825$ with $\hat{u}^0 = 0.5$, $\Phi\left[\hat{u}^{19}; \tilde{f}\right] = 1.825 \cdot 10^{-5}$ after coincidence of two successive iterations with the given accuracy $\varepsilon = 10^{-7}$) is used to construct an initial approximation in Pshenichny's method by projection onto the set of convex functions.

The value of residual for this initial approximation was equal to $1.41 \cdot 10^{-5}$; when the given accuracy $\varepsilon = 10^{-9}$ was reached (at the third step of the iterative process), the functional became equal to $1.39 \cdot 10^{-5}$. Such a small decrease of the functional can be explained by the fact that its value on the constructed initial approximation is comparable with the error of approxima-

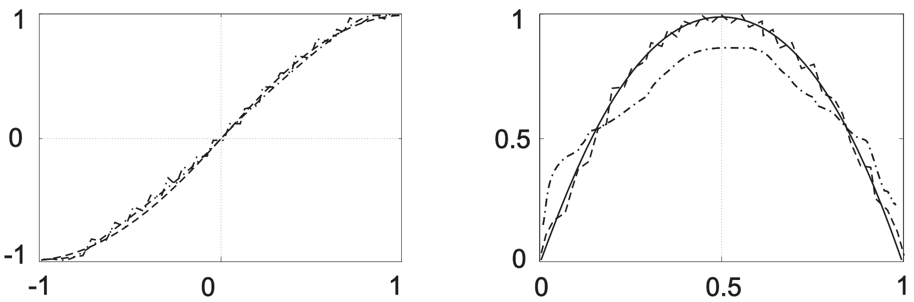


Figure 4.2.5: 1) the model problem 2; $M = 41$, $N = 41$; the solid curve — the exact solution; the dashed curve — $\hat{u}^7$; the dot-dashed curve — $\hat{u}^{13}$ with the error 5 % in the kernel and the right-hand side; 2) the model problem 1; $M = 41$, $N = 21$; the solid curve — the exact solution; the dashed curve — $\hat{u}^8$ with the error 5 % in the kernel and the right-hand side; the dot-dashed curve — $\hat{u}^4$ with the error 25 % in the kernel and the right-hand side

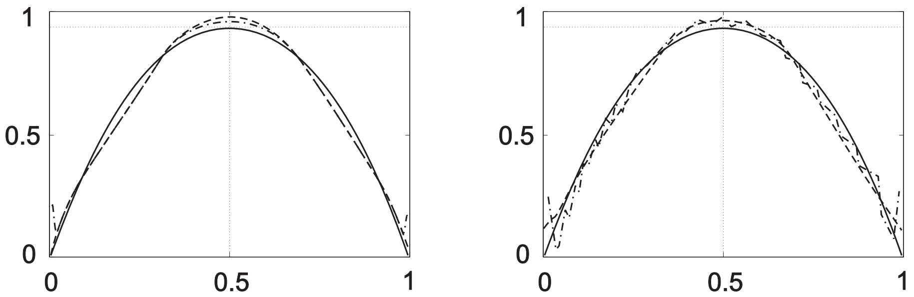


Figure 4.2.6: The model problem 1; $M = 41$, $N = 41$; the solid curve — the exact solution; 1) the dot–dashed curve — $\hat{u}^{19}$ (the gradient method); the dashed curve — $u^3$ (Pshenichny's method); 2) the error 1 % in the kernel and the right-hand side; the dot–dashed curve — $\hat{u}^3$ (the gradient method); the dashed curve — $u^3$ (Pshenichny's method)

tion of the integral in (4.2.6). The exact solution to the model problem 1 and the numerical results of successive stages are represented in Figure 4.2.6 (left). The computing time for the model problem 1 on a BESM-6 computer was approximately equal to 1 minute.

In order to study the stability of the above algorithm with respect to errors in input data, the model problem 1 was solved with a perturbed kernel $\tilde{k}$ and a perturbed right-hand side $\tilde{f}$ of equation (4.2.6). Some results of numerical experiments are represented in Figure 4.2.6 (right) for the error 1 % in the kernel and the right-hand side. The value of the objective functional after three steps of the gradient method was decreased from $\Phi\left[\hat{u}^0; \tilde{f}\right] = 1.828$ for the zeroth iteration $\hat{u}^0 \equiv 0.5$ to $\Phi\left[\hat{u}^3; \tilde{f}\right] = 3.92 \cdot 10^{-3}$ with the coincidence of two successive iterations with the given accuracy $\varepsilon = 10^{-4}$. The corresponding iteration $\hat{u}^3$ was used for Pshenichny's method to construct an initial approximation in the class of functions monotone on the segments $[-1, 0]$ and $[0, 1]$. After three steps of the iterative process, the coincidence of iterations with the accuracy $\varepsilon = 10^{-8}$ was reached. The resulting approximation satisfies the conditions of monotonicity on the segments $[-1, 0]$ and $[0, 1]$, whereas its deviation from the exact solution is approximately $5\,\%$ inside the grid $\varpi_{h_y}$ and is $12\,\%$ at the endpoints $a = -1$ and $b = 1$; the required computing time was a little over one minute.

In conclusion of our discussion of numerical results, we briefly consider the model problem 1 from the previous section.

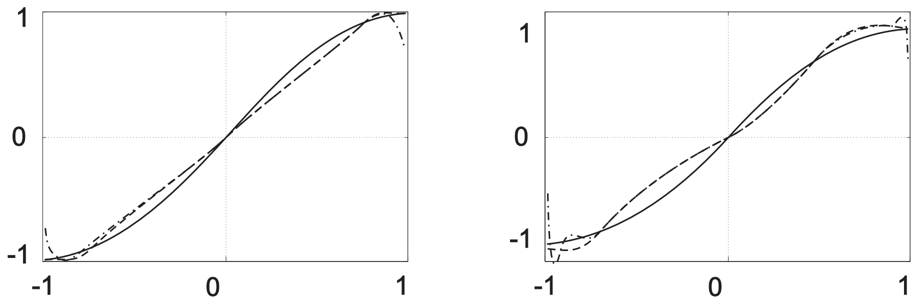


Figure 4.2.7: The model problem 1; $M = 41$, $N = 41$; the solid curve — the exact solution; 1) the dot–dashed curve — $\hat{u}^8$ (Pshenichny's method); 2) the error in the kernel is $3\,\%$ and in the right-hand side is $1\,\%$; the dot–dashed curve — $\hat{u}^{31}$ (the gradient method); the dashed curve — $\hat{u}^7$ (Pshenichny's method)

The exact solution to the model problem 2, the approximate solution obtained by the gradient method after coincidence of two successive iterations with the accuracy $10^{-7}$, and the approximation in the class of monotone functions with sign-definite curvature obtained after eight iterations by Pshenichny's method are represented in Figure 4.2.7 (left). The computing time was approximately equal to 1.5 minutes.

Our additional model computations demonstrated the stability of the above algorithm. Some numerical results for the model problem 2 in the class of monotone functions are given in Figure 4.2.7 (right) with the error $3\,\%$ in the kernel and with the error $1\,\%$ in the right-hand side.

The following conclusion may be drawn from our results of model computations: the algorithm of sequential use of information on the character of the sought-for solution allows one to considerably refine the approximate solution obtained without account of constraints. The computing time required by the algorithm is not large and is comparable with that required by the conjugate gradient projection method described in the previous section.

**4.** Let us consider some peculiarities of numerical implementation of the descriptive regularization method, enabling to use the resources of a computer (storage and time) more efficiently and to increase the performance of a computer substantially when this method is applied. Our reasoning is not always strict, since it is often based on an intuitive analysis.

Suppose it is required to solve the Fredholm integral equation of the first kind

$$Au \equiv \int_a^b k(x,\xi)\, u(\xi)\, d\xi = f(x), \quad c \le x \le d, \tag{4.2.16}$$

where the kernel $k(x,\xi)$ and the right-hand side $f(x)$ are at least continuous functions and may be given approximately.

Let us define the functional

$$\varphi(u) \equiv \int_c^d (Au - f)^2\, dx \tag{4.2.17}$$

and assume that the following *a priori* information on the behavior of the sought-for solution $u(\xi)$ is known: $u \in M$, or $u \in V$, or $u \in M \cap V$, where the sets $M$ and $V$ are defined in the same way as in Section 4.1. In addition we assume that the following information is available:

$$|u(\xi)| \le R(\xi), \quad |u'(\xi)| \le R_1(\xi).$$

Here $R(\xi)$ and $R_1(\xi)$ are given functions defining the sets $R$ and $R_1$ of the functions $u(\xi)$.

The definition domain $D$ of the functional (4.2.17) may be any one of the sets $M$, $V$, $R$, or $R_1$ or may be an intersection of two or more sets from this list.

The descriptive regularization method consists in the construction of an element $u_\varepsilon^* \in D$ such that

$$\varphi(u_\varepsilon^*) \le \inf_{u\in D} \varphi(u) + \varepsilon, \quad \varepsilon > 0. \tag{4.2.18}$$

The quantity $\varepsilon$ characterizes the accuracy of minimization of the objective functional and is given *a priori* or is determined during the computing process. The set $D$ may be known up to a distribution of extreme points, so that the further optimization in these parameters is possible. Here we consider the above parameters to be fixed and pay our main attention to the numerical schemes of solving the problem (4.2.18).

As a minimization method for (4.2.18), it is recommended the conjugate gradient projection method. Why this method is recommended? Let us clarify this choice. Let *a priori* knowledge about the sought-for solution be given in the form $u = Bv$, $\|v\| \le C$, and the solution to (4.2.16) with exact data belong to this set (here $B$ is an operator). Then, equation (4.2.16) is equivalent to the equation $ABv = f$ with the further computation $u = Bv$. Let $\lambda_i^A$ be the singular values of the operator $A$, whereas $\lambda_i^{AB}$ are the singular values of the operator $AB$. If the operator $B$ is completely continuous, then the tendency rate of $\lambda_i^{AB}$ to zero is substantially higher than that of $\lambda_i^A$ as $i \to \infty$, i.e., the problem $ABv = f$ is "more" singular and, hence, has a considerably smaller "practical" rank $r$. To solve the transformed problem by the conjugate gradient method, then, it is required no more than $r$ steps. Usually, $r$ is not large. Our experimental computations demonstrated that, as a rule, $r \le 10$ for a wide class of problems. This value of $r$ is substantially less than the order of linear systems approximating the problem (4.2.6). It is impossible to explicitly specify the operator $B$ in the descriptive regularization method (strictly speaking, there is no need in this specification); however, the computing effect of this method is similar to that described above (see Section 4.1).

Thus, it is possible to assume that the number of steps (iterations) in the descriptive regularization method is small. Where we can find an additional reserve for an increase of the algorithm efficiency?

Let us analyze each step of the method. It consists in the two stages: the computation of the gradient of the functional $\varphi$ and the computation of its projection onto the set $D$. At the second stage, certainly, it is possible to solve the corresponding problem of mathematical programming with constraints by a standard method. This approach appears to be inefficient, since the standard methods do not take into account the essential specific character of the problem. It is required to develop special methods of projection onto the set of piecewise-monotone (or piecewise-convex) functions with a minimum of operations. Such direct algorithms of projection were developed by M. K. Samarin and V. A. Malyshev.

Each projection operation, according to this approach, is performed no more than in $n^2$ arithmetic operations, where $n$ is the number of components in the projection vector; hence, the total number of arithmetic operations required for $r$ steps is of order $rn^2$.

Thus, the two reserves of increase of efficiency for the numerical algorithm based on the descriptive regularization method have been considered: due to the choice of an algorithm for minimization of the functional $\varphi$ with taking account of the specific character of the problem under study and due to an efficient implementation of the operation of projection onto the special set $D$.

The further increase in performance of the algorithm is possible due to the efficient computation of $\operatorname{grad}\varphi$. Here a specific character of the kernel in equation (4.2.16) plays the most important role. Let us consider the following cases of representation of this kernel:

a) general;

b) symmetric: $k(x, \xi) \equiv k(\xi, x)$;

c) difference: $k(x, \xi) \equiv k(\xi - x)$;

d) difference and symmetric, i.e., $k(t) = k(-t)$;

e) difference and $k(t)$ is a periodic function.

After a finite-dimensional approximation of (4.2.6), we come to the linear system

$$\hat{A}\hat{u} = \hat{f},$$

where the matrix $\hat{A}$ corresponds to the type of the kernel under consideration:

a) $\hat{A}$ is a rectangular matrix;

b) $\hat{A} = \hat{A}'$, where the prime indicates the operation of transposition;

c) $\hat{A}$ is a Toeplitz matrix [58];

d) $\hat{A}$ is a Toeplitz symmetric matrix;

e) $\hat{A}$ is a circulant matrix [58].

In order to store the matrix $\hat{A}$, it is required the following amount of storage: case a) — $n^2$, case b) — $n^2/2$, case c) — $2n$, case d), and e) — $n$.

If direct algorithms are used to multiply a matrix by a vector (this operation is the main one when $\operatorname{grad} \varphi(u)$ is computed), then in all the above cases it is required $O(n^2)$ arithmetic operations with significant savings in storage. However, the use of special algorithms based on the algorithms of fast Fourier transform (FFT) allows one in case e) to multiply a circulant matrix by a vector only in $\sim n \log_2 n$ operations, which is an optimal number of operations. The multiplication of a Toeplitz matrixes by a vector, as shown in [58], is also possible to perform in $\sim n \log_2 n$ operations with insignificant loss in storage. These algorithms of fast multiplication of Toeplitz and circulant matrices are included as basic subroutines into the Numerical Analysis Library of Moscow University Research Computing Center [38] and are used in the algorithms implementing the descriptive regularization method (these algorithms are also included into this Library).

Despite its obvious simplicity, the above reasoning allowed us to develop a collection of algorithms for special classes of integral equations of the first kind whose efficiency is similar to the efficiency of methods for solving the same type integral equations of the second kind.

It follows from the above discussion that the most difficult cases (from the standpoint of further reduction of storage and computing time) are cases a) and b). However, in our opinion, there are doubtless reserves of increase of performance for the algorithms in these cases also. Certainly, if this problem is considered only from the algebraic point of view, then such an increase seems to be questionable. However, if the kernel $k(x, \xi)$ is smooth, then with a high degree of accuracy we can write down the following expansion for a small $r$ [118]:

$$k(x, \xi) \approx \sum_{i=1}^{r} \varphi_i(x) \psi_i(\xi). \tag{4.2.19}$$

An algebraic analogue of this expansion is the truncated singular (one-rank) value decomposition of the matrix $\hat{A}$:

$$\hat{A} \approx \sum_{i=1}^{r} \hat{\varphi}_i \hat{\psi}_i', \quad r \ll \max(m, n). \tag{4.2.20}$$

When equation (4.2.16) is solved many times, the use of (4.2.20) with various right-hand sides substantially reduces the number of operations required for computing $\hat{A}\hat{u}$ (and $\hat{A}'\hat{v}$).

Unfortunately, the majority of kernels in integral equations connected, for example, with the experimental data processing formally have a large degree of smoothness, but are organized insufficiently "well". Namely, these kernels may have the weakly expressed $\delta$-shaped character. In this case, the representations (4.2.19) and (4.2.20) converge, in general, slowly. Hence, the useful effect of the above-proposed method is minimal. What can be recommended for this case?

The following approaches to the solution of the above problem may be applied. The essence of these approaches is rather simple. First, it can be recommended to partition the rectangle $\Pi = [c, d] \times [a, b]$ into a small number of subrectangles where the representations of the type (4.2.19) are suitable. As a result, the matrix $\hat{A}$ becomes partitioned into blocks in which the representations of the type (4.2.20) are appropriate. Of course, the following problem arises: to partition the rectangle $\Pi$ optimally in such a way that the number of subrectangles be minimal with a fixed level of accuracy for the approximation of a kernel. The authors assume that this always can be done efficiently on the basis of intuitive reasoning. It should be noted that in [84] a strict mathematical solution of a problem similar to the above one is given. The results of our discussion can also be useful for solving integral equations of the second kind by iterative methods.

The second approach is based on the usage of *a priori* information on the smoothness of the sought-for solution. Let $u = Bv$, where $B$ is a linear completely continuous integral operator. Then, it is quite reasonable to replace the operator $A$ by the operator $AB$ and to pass to the problem of finding the new sought-for function $v$. A smoother kernel corresponds to the operator $AB$; hence, it is possible to apply the representations (4.2.19) and (4.2.20) in the entire domain of its definition. The main difficulty of this approach is to obtain the required representations without explicit computation of the superposition of the kernels of the operators $A$ and $B$ (respectively, without multiplication of the approximating matrices $\hat{A}$ and $\hat{B}$), requiring a significant amount of operations compared to the further data processing. This difficulty seems to be unimportant if the kernels of the operators $A$ and $B$ are given in explicit form.

The third approach is based on the representation of the kernel $k(x,\xi)$ in the form $k(x,\xi) = k_0(x-\xi) + k_1(x,\xi)$, where $k_0(x-\xi)$ takes into account the $\delta$-shaped part of the kernel of the original equation and $k_1(x,\xi)$ is its smooth part. Then, for $k_1$ the representation (4.2.19) (the representation (4.2.20) in matrix form) is applicable; the specific character of the matrix approximation for the function $k_0$ is taken into account according to the above discussion.

In this approach, the total number of arithmetic operations for multiplication of a matrix by a vector is of order $n \log_2 n + rn$, where $r$ is the "practical" rank of the function $k_1$, and can be substantially less than $n^2$. There exist other ways of reducing the number of operations and the amount of storage required.

Now we briefly analyze the quality of the approximations obtained by the descriptive regularization method. As shown in [191], [192], the account of $M$-type conditions allows one to speak of a uniform approximation for the sought-for solution only, whereas the account of $V$-type conditions allows one to speak of a uniform approximations for the sought-for solution and its derivative [193]. Of prime importance is the character of these approximations. In the case $u \in M$, the approximations are certainly stepwise; even if the number of points is large and the threshold of steps is small, such approximations do not correspond to an intuitive idea about the behavior of the solution from the visual standpoint. Our numerical experiments demonstrate that the account of $V$-type conditions leads to numerical results that are satisfactory from the visual standpoint: the corresponding functions can be drawn with a stroke of the pen; the algorithm of solution is complicated insignificantly [193].

**5.** Let us consider some additional advantages of the descriptive regularization method. The specification of $D$ defined basically by the choice of "remarkable" points of the sought-for solution can be characterized as an implicit formation of a skeleton model for this solution, whereas the method itself can be characterized as a numerical reconstruction of this skeleton. The quality of this reconstruction may be characterized by $\varphi_D^* = \inf_{u \in D} \varphi(u)$. It is possible to select the set $D$ in such a way that a minimum of $\varphi_D^*$ be attained with variations of $D$. Mathematically, this problem is rather delicate and requires the development of special methods of solving. Our experience shows that from the physical reasoning (sometimes, after several steps of the conjugate gradient method) it is possible to specify a set $\tilde{D}$ close to an "optimal" set $D^*$. Therefore, choosing $D_1 \approx D^*$ and $D_2 \approx D^*$, analyzing the corresponding solutions $u_1$ and $u_2$ and the characteristics of their quality $\varphi_1^* = \varphi_{D_1}^*$ and $\varphi_2^* = \varphi_{D_2}^*$, sometimes it is not difficult to formulate some rules for the choice of a set $D_3$ closer to $D^*$. Usually, it is sufficient to make several trials. Even an unskilled user quickly copes with such a technique of solving the integral equation (4.2.16). This way of solution is rather efficient with the use of displays, graph plotters and other devices for geometric representation of discrete information. These ideas can easily be implemented within the framework of a specialized system of solving integral equations of the first kind by the descriptive regularization method in interactive mode.

The descriptive regularization method is applicable not only to the equations of the type (4.2.16), but also to more general ones (in particular, to systems of integral equations). For example, in (4.2.16) it is possible to consider $k(x, \xi)$ as a matrix function, whereas $u(\xi)$ and $f(x)$ can be supposed to be vector functions of some dimension $l$. Such equations are obtained, for example, in problems of designing the optimal linear systems of automatic control with $l$ inputs [25], [26], i.e., when optimal pulse transition functions are determined for stationary objects. In this case, the matrix $\hat{A}$ acquires the block-Toeplitz structure. The application of the descriptive regularization method does not require an explicit specification of a kernel in equation (4.2.16). Of importance is the possibility of efficient approximate computation of $Au$ and $A^*v$, where $A^*$ is the operator conjugate with respect to $A$. This situation takes place when the so-called inverse problems for partial differential equations are solved in the problems of restoration of boundary regimes (for example, the inverse heat conduction problem [3]).

The general formulation of such problems is the following. Let us consider the operator (differential) equation $Lz = g$ with the boundary conditions $l_0 z = u$ and $lz = f$, where $z$ and $u$ are the sought-for functions. Usually, the problem $Lz = g$, $l_0 z = u$ is well posed for a fixed $u$ and defines some operator $z = z(u)$. Then, the function $u$ can be found as a solution to the following operator equation of the type (4.2.16) with an implicitly given kernel: $Au \equiv lz(u) = f$.

The expression $A^*v$ can efficiently be found as a solution to some adjoint problem. These questions are considered in detail, for example, in [68]. The descriptive regularization method may also be applied for solving the inverse problems of the above-considered type in the case when the operators $A$, $l_0$, and $l$ are nonlinear if the sought-for function is one-dimensional. When this function is many-dimensional, the method requires its further development (in particular, the development of efficient numerical algorithms for projection onto special sets arising in this case). The problem of efficient computation of $Au$ and $A^*v$ becomes more complicated. These questions are still open (see [4]).

## 4.3 Algorithms of piecewise monotone and convex approximation

**1.** Let us describe a direct method of orthogonal projection of an element $f$ from an $N$-dimensional space $R_N$ with the norm $\|f\| = \left(\sum_{i=1}^N p_i f_i^2\right)^{1/2}$ (where $p_i > 0$, $i = 1, \dots, N$, are weighting coefficients) onto the set $M$ of elements from $R_N$ defined as

$$M = \{v \in R_N : \; m_k(v_{k+1} - v_k) \geq 0, \; k = 1, \dots, N-1\}, \tag{4.3.21}$$

where $m_k$, $k = 1, 2, \dots, N-1$, are the parameters characterizing only the signs of the corresponding inequalities and taking the values 1, $-1$, or 0. This method was developed by M. K. Samarin [237].

The problem of projecting a vector $f$ onto the set (4.3.21) may be interpreted as a mean-square approximation of a grid function $f$ on a set $M$ of grid functions $v$ with constraints like monotonicity.

The orthogonal projection $t \in M$ of an element $f \in R_N$ onto the set $M$ is a solution to the following problem of convex programming:

$$\|f - t\|^2 = \min_{v \in M} \|f - v\|^2 . \tag{4.3.22}$$

Since the set $M$ is convex and closed, the problem (4.3.22) certainly has a unique solution.

First of all we note that if the parameter $m_k$ is equal to zero, then the problem (4.3.22) is divided into two independent problems of the same type; without loss of generality, therefore, we may assume that $m_k \neq 0$, $k = 1, \dots, N-1$. According to the Kuhn–Tucker theorem [283], for an element $t \in R_N$ to be a solution to the problem (4.3.22), it is necessary and sufficient that there exist numbers $u_1, u_2, \dots, u_N$ such that

$$g_k(t) \equiv m_k(t_{k+1} - t_k) \geq 0, \tag{4.3.23}$$

$$u_k g_k(t) = 0, \tag{4.3.24}$$

$$u_k \geq 0, \tag{4.3.25}$$

$$\operatorname{grad}\sigma(t)=\sum_{k=1}^{N-1}u_k\operatorname{grad}g_k(t)\,,\tag{4.3.26}$$

where

$$\sigma(t)=\|f-t\|^2\,.$$

The condition (4.3.26) leads to the system of equations

$$\begin{aligned}&p_1(t_1-f_1)=-m_1u_1,\quad p_2(t_2-f_2)=m_1u_1-m_2u_2,\quad \dots,\\ &p_k(t_k-f_k)=m_{k-1}u_{k-1}-m_ku_k,\quad \dots\quad p_N(t_N-f_N)=m_Nu_N.\end{aligned}$$

Adding up the first $k$ equations of this system, we obtain the following expression for $u_k$:

$$m_ku_k=\sum_{i=1}^{k}p_i(f_i-t_i)\quad\text{or}\quad u_k=m_k\sum_{i=1}^{k}p_i(f_i-t_i)\,.\tag{4.3.27}$$

Let

$$P_l^n=\sum_{k=n+1}^{l}p_k,\quad F_l^n=\sum_{k=n+1}^{l}p_kf_k/P_l^n,\quad T_l^n=\sum_{k=n+1}^{l}p_kt_k/P_l^n,\tag{4.3.28}$$

where $l>n$. Then, the expression (4.3.27) can be rewritten in the form

$$u_k=m_k\left(F_k^0-T_k^0\right),\quad k=1,\dots,N-1.\tag{4.3.29}$$

Note that for $l_2>l_1>n$ the following relation takes place:

$$\operatorname{sign}\left(F_{l_2}^n-F_{l_1}^n\right)=\operatorname{sign}\left(F_{l_2}^{l_1}-F_{l_1}^n\right),\tag{4.3.30}$$

since by virtue of (4.3.28) we have

$$F_{l_2}^n-F_{l_1}^n=\left(F_{l_2}^{l_1}-F_{l_1}^n\right)\alpha,\quad \alpha=P_{l_2}^{l_1}\backslash P_{l_2}^n,\quad 0<\alpha<1.\tag{4.3.31}$$

It follows from (4.3.23)–(4.3.29) that the solution $t$ to the problem (4.3.22) is the piecewise constant function

$$t_k=t_{k_i},\quad k=k_{i-1}+1,\dots,k_i,\quad i=1,2,\dots;\quad k_0=0,\tag{4.3.32}$$

and

$$t_{k_i}=F_{k_i}^{k_{i-1}},\quad i=1,2,\dots;\quad k_0=0.\tag{4.3.33}$$

Indeed, if $g_1(t) = \cdots = g_{k-1}(t) = 0$, $g_k(t) > 0$, then from (4.3.23) it follows that $t_1 = t_2 = \cdots = t_{k_1}$, whereas from (4.3.24) and (4.3.29) it follows that $t_{k_1} = F_{k_1}^0$; since for $k > k_1$ we have

$$u_k = m_k \sum_{i=1}^{k} p_i (f_i - t_i) = m_k P_k^{k_1} \left( F_k^{k_1} - T_k^{k_1} \right),$$

then on the basis of the same reasoning we come to the validity of the relations (4.3.32) and (4.3.33).

Thus, in order to find the solution $t$ to the problem (4.3.22), it is sufficient to find the points of discontinuity on the right for the piecewise constant function $t$.

**Theorem 4.3.71** *For a point $k_{i+1}$ to be a right-hand discontinuity point for the piecewise constant function $t$ which is a solution to the problem* (4.3.22), *it is necessary and sufficient that the following conditions be fulfilled:*

$$1)\ m_k \left( F_k^{k_i} - F_{k_{i+1}}^{k_i} \right) \geq 0, \quad k = k_i + 1, \ldots, k_{i+1}; \tag{4.3.34}$$

2) *if $k_{i+1} < N$, then there exists a point $j_{i+1} > k_{i+1}$ such that*

$$m_{k_{i+1}} \left( F_{k_{i+1}}^{k_i} - F_{j_{i+1}}^{k_i} \right) < 0, \quad m_{k_{i+1}} = -m_{j_{i+1}} \quad \text{or} \quad j_{i+1} = N; \tag{4.3.35}$$

*moreover, for any $k$ such that $k_{i+1} < k < j_{i+1}$ and $m_k = -m_{k_{i+1}}$, the following relation takes place:*

$$m_{k_{i+1}} \left( F_{k_{i+1}}^{k_i} - F_k^{k_i} \right) < 0. \tag{4.3.36}$$

**Proof**. Since in the hypotheses of this theorem there are no points whose subscripts are less than $k_i$, without loss of generality we may assume that $i = 0$ (in this case, $k_0 = 0$).

*Necessity.* By virtue of (4.3.29), (4.3.32), and (4.3.33), the requirement (4.3.34) is a constituent of the necessary condition (4.3.25); hence, the condition (4.3.34) itself is also necessary.

The necessity of the second hypothesis of this theorem is proved by contradiction. Let the point $k_1$ be a right-hand discontinuity point for a solution $t$ to the problem (4.3.22) for which the conditions (4.3.35) and (4.3.36) are not valid. This means that there exists a point with a subscript $n > k_1$ such that

$$m_{k_1} \left( F_{k_1}^0 - F_n^0 \right) \geq 0, \ \ m_n = m_{k_1} \ \ (\text{or} \ \ n = N) \tag{4.3.37}$$

and that for any $k$, $k_1 \leq k < n$, such that $m_k = m_{k_1}$ the following inequality holds:

$$m_{k_1} \left( F_{k_1}^0 - F_k^0 \right) \geq 0. \tag{4.3.38}$$

Let us find the second discontinuity point $k_2$ for this case. If $k_2 \geq n$, then

$$\begin{aligned} m_n\left(F_n^{k_1} - F_{k_2}^{k_1}\right) &= m_n\left(F_n^{k_1} - F_{k_1}^{0}\right) + m_n\left(F_{k_1}^{0} - F_{k_2}^{k_1}\right) = \\ &= m_{k_1}\left(F_n^{k_1} - F_{k_1}^{0}\right) + m_{k_1}\left(F_{k_1}^{0} - F_{k_2}^{k_1}\right) \geq 0 \end{aligned} \tag{4.3.39}$$

by virtue of the above-proved condition (4.3.34); but then from (4.3.30) and (4.3.39) the inequality

$$m_{k_1}\left(F_{k_1}^{0} - F_{k_2}^{k_1}\right) \geq 0 \tag{4.3.40}$$

follows; this inequality contradicts the above assumption that $k_1$ is a right-hand discontinuity point for the function $t$, i.e., for this point the following condition should be valid:

$$m_{k_1}\left(F_{k_2}^{k_1} - F_{k_1}^{0}\right) > 0. \tag{4.3.41}$$

For a similar reason (the contradiction with the condition (4.3.41)) the situation when $k_2 < n$ and $m_{k_2} = -m_{k_1}$ is impossible, since in this case the condition (4.3.38) leads to the inequality (4.3.40).

Thus, it turns out that $k_2 < n$ and $m_{k_2} = m_{k_1}$. But then we have

$$m_{k_2}\left(F_{k_2}^{k_1} - F_{k}^{k_1}\right) = m_{k_2} m_{k_1}\left[m_{k_1}\left(F_{k_2}^{k_1} - F_{k_1}^{0}\right) - m_{k_1}\left(F_{k}^{k_1} - F_{k_1}^{0}\right)\right] > 0$$

for all $k$, $k_1 < k < n$, and such that $m_k = -m_{k_1}$ by virtue of the conditions (4.3.30), (4.3.38), and (4.3.41) and for $k = n$ by virtue of (4.3.37). For the point $k_2$, the conditions similar to the conditions (4.3.37) and (4.3.38) take place. This means that when the point with subscript $n$ is reached, the assumption that $k_2$ is a right-hand discontinuity point for the function $t$ becomes invalid, i.e., the conditions (4.3.37) and (4.3.38) make it impossible to construct a continuation of the solution $t$ such that the following condition be fulfilled:

$$m_{k_i}\left(F_{k_{i+1}}^{k_i} - F_{k_i}^{k_{i-1}}\right) > 0, \quad k_i < n \leq k_{i+1}.$$

*Sufficiency.* In order to prove the sufficiency of the conditions (4.3.35) and (4.3.36), it is necessary to show that the second right-hand discontinuity point $k_2$ for the function $t$ is necessarily chosen in such a way that the condition (4.3.41) would be valid, since only in this case all the conditions (4.3.23)–(4.3.26) will be fulfilled.

1) If $k_2 < j_1$ and $m_{k_2} = m_{k_1}$, then the validity of the condition (4.3.41) follows from (4.3.30) and (4.3.36).

2) Let $k_2 < j_1$ and $m_{k_2} = -m_{k_1}$. Then, for any $k > k_2$ we have

$$z_k = m_{k_2}\left(F_{k_2}^{k_1} - F_k^{k_1}\right) = \frac{1}{\alpha} m_{k_1}\left(F_k^0 - F_{k_1}^0\right) - m_{k_1}\left(F_{k_2}^{k_1} - F_{k_1}^0\right), \quad (4.3.42)$$

where $0 < \alpha < 1$; hence, from (4.3.36) it follows that if the condition (4.3.41) is not fulfilled, then for any $k$, $j_1 > k > k_2$, and $m_k = m_{k_1}$ the inequality $z_k \geq 0$ takes place, i.e., if for $k_2$ there exists a point $j_2$ mentioned in the hypotheses of this theorem, then this point is necessarily positioned to the right of the point $j_1$, i.e., $j_2 > j_1$. In this case, on the other hand, from (4.3.42) and (4.3.35) it follows that $z_{j_1} \geq 0$, i.e., the point $k_2$ certainly does not satisfy the necessary conditions (4.3.35), (4.3.36) and, therefore, cannot be a right-hand discontinuity point for the function $t$.

3) Finally, if $k_2 > j_1$, then

$$m_{j_1}\left(F_{j_1}^{k_1} - F_{k_2}^{k_1}\right) = m_{k_1}\left(F_{k_2}^{k_1} - F_{k_1}^0\right) - \frac{1}{\alpha} m_{k_1}\left(F_{j_1}^0 - F_{k_1}^0\right) \geq 0$$

by virtue of condition (4.3.34); hence, from the fulfilment of the requirement (4.3.35) it follows the validity of the condition (4.3.41). The proof of the theorem is completed.

The above projection method was used to solve Fredholm integral equations of the first kind by the descriptive regularization method and to solve a number of inverse problems [239]. A subroutine implementing this method was included into the Numerical Analysis Library of Moscow University Research Computing Center [37], [38].

**2.** Following [241], here we describe a direct algorithm for uniform approximation of a function $f$ given approximately by its values $f = (f_1, \ldots, f_N)$ on a discrete set of points $X = \{x_1, \ldots, x_N\}$, $x_1 < x_2 < \ldots < x_N$, by discrete monotone functions. The functional

$$\sigma(y) = \max_{1 \leq k \leq N} p_k \left|f_k - y_k\right| \quad (4.3.43)$$

is used as a measure of approximation. Here $p_k > 0$ are the weighting coefficients, $f \in R_N$ is a given function, $y \in R_N$ is an approximating element (a vector), and $R_N$ is an $N$-dimensional linear space.

This algorithm allows one to compute the best uniform approximation in the case of unit weighting coefficients in $O\left(N^2\right)$ arithmetic operations (in the general case, in $O\left(N^3\right)$ operations).

First we consider the process of solving the following auxiliary problem. Suppose it is required to find

$$\min_{y \in C} \sigma(y) = \sigma(t), \quad (4.3.44)$$

where $C = \{y \in R_N : y_1 = y_2 = \cdots = y_N\}$.

**Lemma 4.3.32** *If $t \in C$ is a solution to the problem* (4.3.44)*, then there exist $k_1$ and $k_2$ such that*

$$\sigma(t) = p_{k_1}(f_{k_1} - t_{k_1}) = p_{k_2}(t_{k_2} - f_{k_2}). \tag{4.3.45}$$

*In addition we have*

$$t_1 = t_2 = \cdots = t_N = (p_{k_1} f_{k_1} + p_{k_2} f_{k_2}) / (p_{k_1} + p_{k_2}). \tag{4.3.46}$$

**Proof**. The first conclusion of this lemma is a particular case of the general theorem on Chebyshev's alternance (see, for example, [63]). The relations (4.3.46) immediately follow from (4.3.45). The lemma is proved.

Let

$$\delta(k,j) = (f_k - f_j) / \left(\frac{1}{p_k} + \frac{1}{p_j}\right). \tag{4.3.47}$$

**Lemma 4.3.33** *In order that an element $t \in C$ satisfying the condition* (4.3.46) *be a solution to the problem* (4.3.44)*, it is necessary and sufficient that the following relation be fulfilled:*

$$|\delta(k_1, k_2)| = \max_{1 \le k,j \le N} |\delta(k,j)|. \tag{4.3.48}$$

**Proof.** *Necessity*. Estimating $|\delta(k,j)|$ from above, we obtain

$$|\delta(k,j)| \le \sigma(t) = |\delta(k_1,k_2)|.$$

The necessity of the condition (4.3.48) follows from this estimate.

*Sufficiency*. Let an element $z \in C$ satisfying the condition

$$z_1 = z_2 = \cdots = z_N = (p_m f_m + p_n f_n) / (p_m + p_n)$$

be a solution to the problem (4.3.43). Then,

$$\begin{aligned} |\delta(m,n)| &= \sigma(z) = \min_{y \in C} \max_{1 \le k \le N} p_k |f_k - y_k| \ge \\ &\ge \min_{y_1 \in R_1} \max \{p_{k_1} |f_{k_1} - y_1| \ , \ p_{k_2} |f_{k_2} - y_1|\} = |\delta(k_1,k_2)|. \end{aligned}$$

Since the problem of Chebyshev's interpolation

$$\min_{y_1 \in R_1} \max \{p_{k_1} |f_{k_1} - y_1|, p_{k_2} |f_{k_2} - y_1|\}$$

has a unique solution [124] defined by the formula (4.3.46), we have $z_1 = t_1$, $z = t$, i.e., the element $z \in C$ defined by the relations (4.3.48) and (4.3.46) is a solution to the problem (4.3.43). The proof of the lemma is completed.

**Lemma 4.3.34** *Let the weighting coefficients $p_k$, $k = 1, \ldots, N$, in* (4.3.43) *be equal to* 1 *and*

$$\delta_n = \max_{1 \le i,j \le n} \delta(i,j) = \delta(k_1, k_2).$$

*Then,*

$$\delta_{n+1} = \max_{1 \le i,j \le n+1} \delta(i,j) = \max\{\delta_n, \delta(k_1, n+1), \delta(k_2, n+1)\}. \quad (4.3.49)$$

**Proof.** It is obvious that

$$\delta_{n+1} = \max\left\{\delta_n, \max_{1 \le i \le n} \delta(i, n+1)\right\}. \quad (4.3.50)$$

Therefore, it is sufficient to prove the formula

$$\max_{1 \le i,j \le n} \delta(i,j) = \left(\max_{1 \le i \le n} f_i - \min_{1 \le j \le n} f_j\right)/2. \quad (4.3.51)$$

Since $p_k = 1$ for $k = 1, \ldots, N$, we have

$$\max_{1 \le i \le n} \delta(i, n+1) = \max\{\delta(k_1, n+1), \delta(k_2, n+1)\};$$

from here we easily obtain (4.3.28). The lemma is proved.

**Corollary.** *When $p_k = 1$, $k = 1, \ldots, N$, the problem* (4.3.43) *can be solved in $O(N^2)$ arithmetic operations.*

This corollary follows from Lemmas 4.3.33 and 4.3.34.

The next example shows that if not all weighting coefficients are equal to 1, then the formula (4.3.50) can be used only to compute the value of (4.3.48) recurrently. Indeed, let

$$N = 4, \quad p_1 = p_2 = p_4 = 1, \quad p_3 = 2, \quad f = (10, -10, 0, -90).$$

Then,

$$\max_{1 \le i,j \le 3} \delta(i,j) = \delta(1,2) = 10;$$

however,

$$\max_{1 \le i,j \le 4} \delta(i,j) = \delta(3,4) = 60 > \max\{\delta(1,4), \delta(2,4)\}.$$

This result contradicts the formula (4.3.49).

Now we formulate the problem of uniform approximation by monotone functions: to find

$$\min_{y \in M} \sigma(y) = \sigma(t), \quad (4.3.52)$$

where the functional $\sigma(y)$ is defined by the formula (4.3.43) and the set $M \in R_N$ is of the form

$$M = \{y \in R_N : y_{k+1} \geq y_k,\ k = 1, 2, \ldots, N-1\}.$$

Generally speaking, the problem (4.3.52) has many solutions. Let us construct one of the possible solutions to this problem by the following algorithm:

$$\begin{array}{l} a)\ t := 1;\ t_0 := -\infty; \\ b)\ \delta(n_k, m_k) := \max\limits_{L \leq m,n \leq k} \delta(m,n),\ \ k = L, L+1, \ldots, N; \\ c)\ F_k := (p_{n_k} f_{n_k} + p_{m_k} f_{m_k}) / (p_{n_k} + p_{m_k}),\ \ k = L, \ldots, N; \\ d)\ F_M := \min\limits_{L \leq k \leq n} F_k; \\ e)\ z := \max\{t_{L-1}, F_M\}; \\ g)\ t_k := z,\ k = L, L+1, \ldots, m_N; \\ h)\ \text{if } m_N = N, \text{ then the end of the algorithm;} \\ i)\ L := m_N + 1,\ \text{go to step b).} \end{array} \tag{4.3.53}$$

**Theorem 4.3.72** *Algorithm* (4.3.53) *yields a solution to the problem* (4.3.52).

**Proof**. Let a function $t$ be obtained by algorithm (4.3.53). By construction, the function $t$ is piecewise constant and does not decrease (its monotonicity is provided at step e)). Therefore, it is necessary to show that the function $t$ minimizes the functional $\sigma(y)$ on the set $M$.

Let

$$\begin{array}{rcl} t_1 & = & \cdots = t_{k_1} < t_{k_1+1} = \ldots = t_{k_2} < \ldots < t_{k_2+1} < \ldots \\ \ldots & < & t_{k_i+1} = \cdots = t_{k_{i+1}} < \ldots < t_{k_{j+1}} = \cdots = t_N. \end{array}$$

On each segment of constancy, the function $t$ coincides with the solution to the problem of the type (4.3.52):

$$\min_{y \in C} \max_{k_i < k \leq k_{i+1}} p_k |f_k - y_k|. \tag{4.3.54}$$

Indeed, since $t_{k_i} < t_{k_i+1}$, then, by Lemmas 4.3.32 and 4.3.33, from the steps b)–f) of algorithm (4.3.53) it follows that $t_{k_i+1}$ is a solution to the problem (4.3.54) on a segment $\left[x_{k_{i+1}}, x_{k_i}\right] \subseteq \left[x_{k_i+1}, x_{k_{i+1}}\right]$. The segment $\left[x_{k_i+1}, x_{k_{i+1}}\right]$ is a segment of constancy for the function $t$; therefore, $t_{k_{i+1}}$ is a solution to the problem (4.3.54).

Let

$$\max_{1 \leq k \leq N} p_k |f_{k_i} - t_{k_i}| = \sigma(t)$$

be attained on $\left[x_{k_i+1}, x_{k_{i+1}}\right]$ at some points $x_n$ and $x_m$, so that

$$\sigma(t) = p_n(f_n - t_n) = p_m(t_m - f_m). \tag{4.3.55}$$

The existence of such points $x_n$ and $x_m$ follows from Lemma 4.3.32. Note that $f_n > f_m$, since $\sigma(t) > 0$ and $t_n = t_m$. Let us show that $n < m$. Indeed, let us assume that $m < n$ and

$$p_k(f_k - t_k) < p_m(t_m - f_m), \quad k = k_{i+1}, \ldots, n-1.$$

Let

$$\max_{k_i < l < j < n} \sigma(l, j) = \sigma(l_1, l_2).$$

By virtue of the choice of the points $n$ and $m$, we have $\sigma(l_1, l_2) < \sigma(m, n)$. Moreover,

$$\bar{\bar{t}} = (p_{l_1} f_{l_1} + p_{l_2} f_{l_2}) / (p_{l_1} + p_{l_2}) \geq \bar{t} = (p_m f_m + p_n f_n) / (p_m + p_n),$$

since, otherwise, according to steps b) and d) of algorithm (4.3.53), the condition of constancy for the function $t$ on the segment $\left[x_{k_i+1}, x_{k_{i+1}}\right]$ would contradict the condition (4.3.32). Then,

$$\begin{aligned} \sigma(l_1, l_2) &\geq p_m\left(\bar{\bar{t}} - f_m\right) = p_m\left(\bar{\bar{t}} - \bar{t}\right) + p_m(\bar{t} - f_m) = \\ &= p_m\left(\bar{\bar{t}} - \bar{t}\right) + \delta(m, n) \geq \delta(m, n); \end{aligned}$$

hence, $\bar{\bar{t}} = \bar{t}$. This means that if $m < n$, then there exists $l$, $k_l < l < m$, such that

$$p_l(t_l - f_l) = p_m(f_m - t_m) = \delta(t).$$

For the function $f$, thus, the error of its uniform approximation constructed according to the procedure (4.3.53) is reached at points $x_n$ and $x_m$ such that $x_n < x_m$, whereas $f_n > f_m$. However, this error cannot be improved on the set of monotone functions, since to accomplish this, it is necessary to increase the value of the function $t$ at the point $x_n$ and to decrease the value of this function at the point $x_m$, which is impossible within the limits of the set $M$.

Hence, the function $t$ constructed by algorithm (4.3.53) is a solution to the problem 4.3.52. The proof of the theorem is completed.

**Remark**. *In order to construct a uniform approximation for a function $f$ on the set of nonincreasing functions, it is necessary to introduce the following changes into algorithm* (4.3.53)*: at step a) it is necessary to put*

$$L := 1, \quad t_0 := +\infty;$$

*at step e) the operation* min *should be replaced by* max, *whereas at step f) the operation* max *should be replaced by* min.

Algorithm (4.3.53) of uniform approximation by monotone functions is implemented in a FORTRAN subroutine by M. K. Samarin. This subroutine is included into the Numerical Analysis Library of Moscow University Research Computing Center [38].

**3.** Following [197], here we consider an infinite-dimensional analogue of the problem of monotone approximation. Its solution allows us to formulate an algorithm for constructing a minimizing sequence in $\sim n$ steps.

Let $f_0(x) \in L_2[a,b]$. Let us consider the minimization problem

$$\int_a^b [f(x) - f_0(x)]^2 \, dx \to \min_f \tag{4.3.56}$$

on the set of monotonically increasing functions $f_0(x) \in L_2[a,b]$.

Since in the Hilbert space $L_2[a,b]$ the set of monotonically increasing functions is convex and closed, a solution $f(x)$ to the minimization problem (4.3.56) exists and is unique [207].

Let us introduce the concept of a convex envelope of a continuous function. Let a function $F(x)$ be continuous on a segment $[a,b]$ and let $C_0F(x)$ be the upper envelope of the family of convex functions $V(x)$ satisfying the inequality $V(x) \geq F(x)$ at all the points $x$ of the segment $[a,b]$. The function $C_0F(x)$ is called the convex envelope of the function $F$. The choice of this term is based on the fact that the convex envelope of the epigraph of the function $F(x)$ coincides geometrically with the epigraph of the convex envelope $C_0F(x)$ for the function $F(x)$.

**Theorem 4.3.73** *The function*

$$f(x) = \frac{d}{dx} C_0 \int_a^x f_0(\tau) \, d\tau \tag{4.3.57}$$

*is a solution to the minimization problem* (4.3.56).

**Proof.** Let $f(x)$ be a solution to the minimization problem (4.3.56). We consider the functions

$$F(x) = \int_a^x f(\tau) \, d\tau, \quad F_0(x) = \int_a^x f_0(\tau) \, d\tau$$

and prove that $F(x) = C_0F_0(x)$.

Since the function $f(x)$ monotonically increases, the function $F(x)$ is convex. Let a point $x$ be fixed on the segment $[a,b]$. Let us put

$$\lambda_x(\tau)=\begin{cases} -1, & \tau\in[a,x)\,; \\ 0, & \tau\in(x,b]\,. \end{cases}$$

Since for any $\varepsilon\in[0,\infty)$ the function $f(\tau)+\varepsilon\lambda_x(\tau)$ monotonically increases in $\tau$, the point $\varepsilon=0$ is a point of minimum for the function

$$\varphi_x(\varepsilon)=\int\limits_a^b [f(\tau)+\varepsilon\lambda_x(\tau)-f_0(\tau)]^2\,d\tau.$$

Hence,

$$\frac{d\varphi_x(0)}{d\varepsilon}=2\int\limits_a^b [f(\tau)-f_0(\tau)]\,\lambda_x(\tau)\,d\tau=2\int\limits_a^x [f_0(\tau)-f(\tau)]\,d\tau\ge 0.$$

Therefore, $F(x)\le F_0(x)$ for any $x\in(a,b)$.

Since for any $\varepsilon\in(-\infty,\infty)$ the function $f(x)+\varepsilon$ monotonically increases, the point $\varepsilon=0$ is a point of minimum for the function

$$\varphi(\varepsilon)=\int\limits_a^b [f(x)+\varepsilon-f_0(x)]^2\,dx,\quad \varepsilon\in(-\infty,\infty)\,.$$

Hence,

$$\frac{d\varphi(0)}{d\varepsilon}=2\int\limits_a^b [f(x)-f_0(x)]\,dx=0.$$

Therefore, $F(x)\le F_0(x)$. Moreover, it is obvious that

$$F(a)=F_0(a)=0.$$

Let $C$ be the set of points $x\in[a,b]$ such that $F(x)=F_0(x)$. Since the functions $F(x)$ and $F_0(x)$ are continuous, the set $U=(a,b)$ is open on the real axis. Let $(\alpha,\beta)$ be a constituent interval of the open set $U$ [235]. Using simple, but cumbersome reasoning, we can prove that the function $f$ is constant on the interval $(\alpha,\beta)$ (see Subsection 1 of this section):

$$f(x)=\frac{1}{\beta-\alpha}\int\limits_\alpha^\beta f_0(\tau)\,d\tau$$

at all $x \in (\alpha, \beta)$. Therefore, on the segment $[\alpha, \beta]$ the function $F(x)$ is linear.

Thus, the convex function $F(x)$ satisfies the inequality $F(x) \leq F_0(x)$ on the segment $[a, b]$. Let us assume that a convex function $V(x)$ satisfies this inequality and prove that $V(x) \leq F(x)$. Let $x \in [a, b]$. If $x \in C$, then $V(x) \leq F(x)$, since $F(x) = F_0(x)$. If $x \in (\alpha, \beta)$, where $(\alpha, \beta)$ is a constituent interval of the set $U$, then the following relations take place:

$$V(\alpha) \leq F_0(\alpha) = F(\alpha) \quad \text{and} \quad V(\beta) \leq F_0(\beta) = F(\beta).$$

Therefore, $V(x) \leq F(x)$, since the function $F(x)$ is linear on the segment $[\alpha, \beta]$ and the function $V$ is convex. From the inequality $V(x) \leq F(x)$ it follows that $F(x) = C_0 F_0(x)$. The theorem is proved.

**Remark.** *If the function $f_0(x)$ is piecewise constant, then the solution $f(x)$ to the minimization problem* (4.3.56) *is also piecewise constant. If the function $f_0(x)$ is continuous, then the solution $f(x)$ to the minimization problem* (4.3.56) *is also continuous.*

If the function $f_0(x)$ is piecewise constant, then the function $\int_a^x f_0(\tau)\, d\tau$ is piecewise linear. Therefore, in order to solve the minimization problem (4.3.56) with the piecewise constant function $f_0(x)$, it is necessary to construct the convex envelope of a piecewise linear function.

Let $\{Y_1, \ldots, Y_N\}$ be the values of a piecewise linear function $Y$ on a grid $\{x_1, \ldots, x_N\}$. Geometrically it is clear that the convex envelope of this function is a piecewise linear function taking the values $\{Y_{I_1}, \ldots, Y_{I_N}\}$ on a grid $\{x_{I_1}, \ldots, x_{I_N}\}$, where $\{I_1, \ldots, I_k\}$ is a subsequence of the sequence $\{1, 2, \ldots, N\}$.

Let us describe Malyshev's algorithm for constructing the convex envelope of a piecewise linear function, using the above grids and the corresponding values of the function.

First we put $I_1 := 1$ and $k := 1$.

Let at some step of the algorithm a sequence $\{I_1, ..., I_s\}$ be constructed. Now we consider the condition

$$\begin{aligned} k \ &\leq \ 2 \ \text{ or } \ \left(Y_{I_k} - Y_{I_{k-1}}\right)\left(X_{I_{k-1}} - X_{I_{k-2}}\right) \geq \\ &\geq \ \left(X_{I_k} - X_{I_{k-1}}\right)\left(Y_{I_{k-1}} - Y_{I_{k-2}}\right). \end{aligned} \tag{4.3.58}$$

Let us assume that the condition (4.3.58) is fulfilled. Then, the piecewise linear function with its values $\{Y_{I_1}, \ldots, Y_{I_k}\}$ on the grid $\{X_{I_1}, \ldots, X_{I_k}\}$ is convex (this can easily be checked by induction). If $I_k = N$, the algorithm finishes its work. If $I_k < N$, then $I_{k+1} := I_k + 1$, $k := k + 1$ and the algorithm continues its work.

Let us assume that the condition (4.3.58) is not fulfilled. Then, the piecewise linear function with its values $\{Y_{I_{k-2}}, Y_{I_{k-1}}, Y_{I_k}\}$ on the grid

$\{X_{I_{k-2}}, X_{I_{k-1}}, X_{I_k}\}$ is concave. Therefore, the point $X_{I_{k-1}}$ cannot belong to the grid of the convex envelope and its index $I_{k-1}$ should be eliminated from the further consideration. Therefore, $I_{k-1} := I_k$, $k := k-1$ and the algorithm continues its work.

The condition (4.3.58) is checked at each step of the above algorithm; therefore, the number of steps of the algorithm is equal to the number of checks of this condition. The fulfilment of the condition (4.3.58) leads to the increase of the last index $I_k$ by 1. Therefore, the number of fulfillments of the condition (4.3.58) is equal to $N$. The nonfulfillment of the condition (4.3.58) results in removal of a point in the grid from consideration. However, at the end of work of the algorithm the convex envelope has $k$ nodes. Therefore, the number of nonfulfillments of the condition (4.3.58) is equal to $N-k$.

Thus, the convex envelope of a piecewise linear function can be constructed in $2N-k$ steps, where $N$ is the number of nodes for the original piecewise linear function and $k$ is the number of nodes in its convex envelope.

As mentioned above, the problem of finite-dimensional monotone approximation was considered in [238]. From the above-proved theorem it follows that its solution is reduced to the solution of the infinite-dimensional problem of monotone approximation with a piecewise constant function to be approximated. Now we discuss this interrelation in more detail.

Let $\{f_1^0, \ldots, f_n^0\}$ be a given sequence and $p_1 > 0, \ldots, p_n > 0$ be given weighting coefficients. Let us consider the minimization problem

$$\sum_{i=1}^{n} p_i \left(f_i - f_i^0\right)^2 \to \min \tag{4.3.59}$$

on the set of monotonically increasing sequences $\{f_1, \ldots, f_n\}$. We reduce it to the minimization problem (4.3.56). To accomplish this, we take the piecewise constant function $f_0(x) = f^0$ for $x \in (x_{i-1}, x_i)$, where $x_0 = 0$ and $x_i = p_1 + \cdots + p_i$ for $i = 1, \ldots, n$. Let us put $a = 0$ and $b = x_n$.

Suppose a function $f(x)$ is a solution to the minimization problem (4.3.56) with the above piecewise constant function $f_0(x)$. Then, the function $f(x)$ is also piecewise constant and its nodes are among the nodes of the function $f_0(x)$. Putting

$$f_i = f\left(\frac{x_{i-1} + x_i}{2}\right),$$

we obtain

$$\sum_{i=1}^{n} h_i \left(f_i - f_i^0\right)^2 = \int\limits_a^b \left[f(x) - f_0(x)\right]^2 dx.$$

Therefore, the sequence $\{f_1, \dots, f_n\}$ is a solution to the minimization problem (4.3.59). It follows from the above that the minimizing sequence $\{f_1, \dots, f_n\}$ can be constructed in $\sim n$ steps.

**4.** Let us consider a generalization of Malyshev's approach.

A function $f(x)$ is said to be $n$-convex on $[a, b]$ if it is convex with respect to Chebyshev's system $1, x, \dots, x^{n-1}$. The set of 1-convex functions coincides with the set of increasing functions, the set of 2-convex functions coincides with the set of convex functions. For $n \geq 2$ a function is $n$-convex on $[a, b]$ if and only if it has the convex derivative of order $n-2$. The proofs of these assertions are given in [197].

Let $V_n[a, b]$ be the set of functions square integrable on a segment $[a, b]$ and $n$-convex on $[a, b]$. Let a function $f_0(x)$ square integrable on $[a, b]$ be given. Now we consider the minimization problem

$$\int\limits_a^b [f(x) - f_0(x)]^2\, dx \to \min, \quad f(x) \in V_n[a, b]. \tag{4.3.60}$$

In order to solve this problem, we construct an operator $\Omega_n : W_2^{(n)}[a, b] \to V_{2n}[a, b]$, where $W_2^{(n)}$ is the space of functions whose $n$th derivative is square integrable on $[a, b]$. Let $F_0(x) \in W_2^{(n)}[a, b]$ and $F(x) = \Omega_n F_0(x) \in V_{2n}[a, b]$. It is assumed that the function $F(x)$ satisfies the boundary conditions $F^{(i)}(a) = F_0^{(i)}(a)$ and $F^{(i)}(b) = F_0^{(i)}(b)$ for all $i = 0, \dots, n-1$, the inequality $(-1)^n F(x) \geq (-1)^n F_0(x)$ for all $x \in [a, b]$, and the equality $F^{(2n)}(x) = 0$ for all $x \in [a, b]$ such that $F(x) \neq F_0(x)$.

It can be proved that the function

$$f(x) = \frac{d^n}{dx^n} \Omega_n \int\limits_a^x (x - \tau)^{n-1} f_0(\tau)\, d\tau \tag{4.3.61}$$

is a solution to the minimization problem (4.3.60).

In particular, it follows from this assertion that if $f_0(x)$ is a spline of degree not higher than $n-1$, then $f(x)$ is a spline of degree $n-1$.

The proof is based on the fact that in the Hilbert space of functions square integrable on a segment, the set $V_n[a, b]$ forms a convex and closed cone. Therefore, a solution to the minimization problem (4.3.60) under the condition $f(x) \in V_n[a, b]$ exists and is unique [197]. This allows one to prove the correctness of definition of the operator $\Omega_n$ and to obtain the formula (4.3.61).

The operator $\Omega_n$ is especially simple for $n = 1$. In this case, the function $F(x) = \Omega_1 F_0(x)$ is an upper envelope for the family of convex functions situated on the segment $[a, b]$ below the function $F_0(x)$. In particular, if this function is piecewise linear with $N$ nodes, then the function $F(x)$ is piecewise linear with $k \leq N$ nodes and can be calculated in $2N - k$ steps. When $n \geq 2$, the operator $\Omega_n$ is rather complicated. For this case, however, V. A. Malyshev proposed a method of successive approximations to the function $F(x) = \Omega_n F_0(x)$. The main procedure in this method is the construction of an interpolating spline of degree $2n - 1$ with zero boundary conditions.

When $n = 1$, we come to the problem of approximation by increasing functions. Let us consider, for example, the piecewise constant function $f_0(x)$ represented in Figure 4.3.8.A by the dotted line. Then, the function $F_0(x) = \int_a^x f(\tau)\, d\tau$ is piecewise linear (see Figure 4.3.8.B, the dotted line). The function $F(x) = \Omega_1 F_0(x)$ is represented in Figure 4.3.8.B by the solid line. The solution $f(x) = F'(x)$ to the minimization problem (4.3.60) in the class of increasing functions is represented in Figure 4.3.8.A by the solid line.

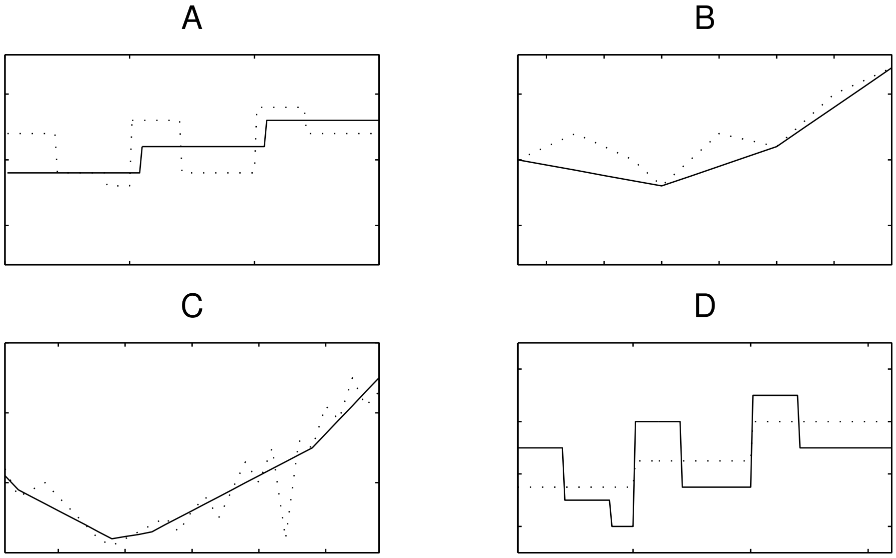


Figure 4.3.8: An example of monotone and convex approximation

When $n = 2$, we come to the problem of approximation by convex functions. An example of its approximate solution is represented in Figure 4.3.8.C. The dotted line represents the function $f_0(x)$, the solid line represents an approximation to the solution $f(x)$. These numerical results were obtained by V. A. Malyshev.

**5.** Now we consider the problem of mean-square minimization on a wedge defined by a differential inequality. Let

$$\mathcal{L}(p) = p^n + a_1 p^{n-1} + \cdots + a_n,$$

where $p = \dfrac{d}{dx}$ is the differentiation operator.

Let us assume that on a segment $[a, b]$ the sought-for function $f_*(x)$ satisfies the differential inequality $\mathcal{L}(p)\, f_*(x) \geq 0$, whereas a given function $f_0(x)$ accessible for processing has the form

$$f_*(x) + \xi(x) = f_0(x),$$

where $\xi(x)$ is a noise. In order to filter out the noise $\xi$, we propose to consider the minimization problem

$$\int_a^b [f(x) - f_0(x)]^2\, dx \to \min_f, \quad f : \mathcal{L}(p)\, f(x) \geq 0,$$

and to adopt its solution as an approximation to the sought-for function $f_*(x)$.

If the natural constraints $f_0(x) \in L_2[a, b]$ and $f \in C^n[a, b]$ are chosen, the above problem has no solution. It has a solution for the less hard constraints $f_0(x) \in L_2[a, b]$ and $f(x) \in L_2[a, b]$. In this case, however, the function $f(x)$ should be considered as a regular generalized function, whereas the differentiation operator $p$ should be considered as the operator of generalized differentiation. In its turn, this leads us to the necessity to consider nonnegative generalized functions, since the generalized function $u(x) = \mathcal{L}(p)\, f(x)$ should be of this type according to the meaning of the problem.

In the framework of the above formulation of the minimization problem, its solution $f(x) \in C^{n-2}[a, b]$ can be approximated by $\mathcal{L}(p)$-splines.

Let $D(a, b)$ be the class of principal functions and $D'(a, b)$ be the class of generalized functions. A generalized function $u(x) \in D'(a, b)$ is said to be nonnegative [197] if it takes a nonnegative value on any nonnegative principal function $\varphi(x) \in D(a, b)$. The fact that a generalized function $u(x)$ is nonnegative is expressed by the inequality $u(x) \geq 0$.

In the space of generalized functions $D'(a, b)$, the nonnegative generalized functions form a convex and closed cone which is the conjugate cone for the cone of nonnegative principal functions. The generalized function $u(x) = \alpha_1 \delta(x - x_1) + \cdots + \alpha_k \delta(x - x_k)$ (where $\alpha_1 \geq 0, \ldots, \alpha_k \geq 0$, $a < x_1 < \ldots < x_k < b$, and $\delta$ is the Dirac generalized function) is an example of a nonnegative generalized function. The set of such generalized functions is dense in the cone of nonnegative generalized functions.

The main property of nonnegativity for generalized functions consists in the fact that any nonnegative generalized function is a generalized function of zeroth order. In the space of generalized functions of zeroth order, the cone of nonnegative generalized functions is a reproducing cone. For any zeroth-order generalized function $u(x)$, there exist nonnegative generalized functions $u_+(x)$ and $u_-(x)$ such that $u(x) = u_+(x) - u_-(x)$.

If a nonnegative generalized function $u(x) \in D'(a,b)$ is of zeroth order, then there exists a functions $\mu(x)$, increasing on $[a,b]$, such that

$$(u(x), \varphi(x)) = \int_a^b \varphi(x)\, d\mu(x)$$

for any principal function $\varphi(x) \in D(a,b)$. The function $\mu(x)$ is defined up to an additive constant and is the indefinite integral of the generalized function $u(x)$.

By definition we have

$$\int_\alpha^\beta u(x)\, dx = \mu(\beta) - \mu(\alpha), \quad a < \alpha, \quad \beta < b.$$

Let $u_1(x), u_2(x), \ldots$ be nonnegative generalized functions such that

$$\lim_{k\to\infty} u_k(x) = u(x).$$

It is assumed that the points $\alpha$ and $\beta$ are the points of continuity for the indefinite integral of the nonnegative generalized function $u(x)$. Then, the following limit takes place:

$$\lim_{k\to\infty} \int_\alpha^\beta u_k(x)\, dx = \int_\alpha^\beta u(x)\, dx.$$

Now we show that the differential equation $\mathcal{L}(p)\, f(x) = u(x)$, where the right-hand side $u(x) \in D'(a,b)$ is a generalized function of zeroth order, can be solved in a standard way.

Since in the space of generalized functions of zeroth order the cone of nonnegative generalized functions is reproducing, it is sufficient to consider the case $u(x) \geq 0$. In fact, only this case is of interest to us.

Let a generalized function $u(x) \in D'(a,b)$ be nonnegative and an increasing function $\mu(x)$ be its indefinite integral. Then,

$$\int_\alpha^x u(\tau)\, d\tau = \mu(x) - \mu(\alpha).$$

By induction, for $k \geq 1$ we put

$$\int\limits_{\alpha}^{x} \frac{(x-\tau)^k}{k!}\, u(\tau)\, d\tau = \int\limits_{\alpha}^{x} \left[ \int\limits_{\alpha}^{y} \frac{(y-\tau)^{k-1}}{(k-1)!}\, u(\tau)\, d\tau \right] dy.$$

Let the power function $z(x) \equiv \sum\limits_{k=0}^{\infty} \gamma_k \frac{x^k}{k!}$ have the infinite radius of convergence. By definition, we put

$$\int\limits_{\alpha}^{x} z(x-\tau)\, u(\tau)\, d\tau = \sum_{k=0}^{\infty} \gamma_k \int\limits_{\alpha}^{x} \frac{(x-\tau)^k}{k!}\, u(\tau)\, d\tau.$$

Now we assume that the function $z(x)$ is a solution to the differential equation $\mathcal{L}(p)\, z(x) = 0$ with the initial conditions $z(0) = \cdots = z^{(n-2)}(0) = 0$ and $z^{(n-1)}(0) = 1$. In this case the function

$$f(x) = \sum_{k=0}^{n-1} c_k z^{(k)}(x) + \int\limits_{\alpha}^{x} z(x-\tau)\, u(\tau)\, d\tau$$

is the general solution to the differential equation $\mathcal{L}(p)\, f(x) = u(x)$.

The function $f(x)$ is classical. Its derivative $f^{(n-2)}(x)$ is continuous on $[a,b]$. Its derivative $f^{(n-1)}(x)$ exists at all points $x$ of $[a,b]$ with the exception of a denumerable set at most. Its derivative $f^{(n)}(x)$ exists at all points $x$ of $[a,b]$ with the exception of a set of measure zero at most. The differential inequality $\mathcal{L}(p)\, f(x) \geq 0$, where $p$ is the operator of classical differentiation, takes place almost everywhere on $[a,b]$.

It follows from the above that a generalized function $f(x) \in D'(a,b)$ satisfying the differential inequality $\mathcal{L}(p)\, f(x) \geq 0$ is a classical function and has the continuous derivative of order $n-2$ on $[a,b]$. We are interested in those functions $f(x)$ for which the integral $\int\limits_a^b f^2(x)\, dx$ is finite.

Let us consider the set of functions

$$V[\mathcal{L}(p)] = \{ f \in L_2[a,b] : \mathcal{L}(p)\, f(x) \geq 0 \}.$$

In the space $L_2[a,b]$, the set $V[\mathcal{L}(p)]$ forms a convex and closed wedge. It is possible to prove that this wedge can be represented in the form of the direct sum: $V[\mathcal{L}(p)] = \mathcal{P}[\mathcal{L}(p)] \oplus K[\mathcal{L}(p)]$, where $\mathcal{P}[\mathcal{L}(p)]$ is the linear $n$-dimensioned space of solutions to the differential equation $\mathcal{L}(p)\, f(x) = 0$

and $K[\mathcal{L}(p)]$ is the infinite-dimensional cone of solutions to the differential inequality

$$\mathcal{L}(p)\, f(x) \geq 0$$

such that these solutions satisfy the relations

$$\int_a^b z^{(i)}(x)\, f(x)\, dx = 0, \quad i = 0, \dots, n-1.$$

Suppose $\varepsilon(x) = z(x)\,\theta(x)$, where $\theta$ is the Heaviside function. The function

$$f(x) = \sum_{i=0}^{n-1} c_i z^{(i)}(x) + \sum_{i=1}^{k} \alpha_i \varepsilon(x - x_i),$$

where $a < x_1 < \ldots < x_k < b$, is called the $\mathcal{L}(p)$-spline with the grid points $x_1, \dots, x_k$. It is said that the $\mathcal{L}(p)$-spline for $f(x)$ has nonnegative coefficients if $\alpha_1 \geq 0, \dots, \alpha_k \geq 0$. It is easy to see that the $\mathcal{L}(p)$-spline for $f(x)$ belongs to $V[\mathcal{L}(p)]$ if and only if its coefficients are nonnegative. The set of $\mathcal{L}(p)$-splines with nonnegative coefficients is dense in the wedge $V[\mathcal{L}(p)]$.

Let $f_1(x),\ f_2(x),\ \ldots$ be functions belonging to the wedge $V[\mathcal{L}(p)]$ and $\lim\limits_{k\to\infty} f_k(x) = f(x)$ in the space $L_2[a,b]$. Then, $\lim_{k\to\infty} f_k^{(i)}(x) = f^{(i)}(x)$ everywhere on $[a,b]$ for $i = 0, 1, \dots, n-2$ and almost everywhere on $[a,b]$ for $i = n-1$. This assertion follows from the fact that the generalized functions $u_k(x) = \mathcal{L}(p)\, f_k(x)$ and $u(x) = \mathcal{L}(p)\, f(x)$ are nonnegative and that the equality $\lim\limits_{k\to\infty} u_k(x) = u(x)$ holds in the space $D'(a,b)$.

Of special interest are the wedges $V[p^n]$, where $n = 1, 2, \dots$. In particular, $V[p]$ is the wedge of increasing functions, $V[p^2]$ is the wedge of convex functions; in the general case, $V[p^n]$ is the wedge of $n$-convex functions [197].

It is possible to prove that the Volterra operator

$$A : f(x) \to f(x) + \int_\alpha^x z^{(n)}(x - \tau)\, f(\tau)\, d\tau$$

maps the wedge of $n$-convex functions $V[p^n]$ onto the wedge $V[\mathcal{L}(p)]$.

Now we turn to the question of solving the above minimization problem.

Let us consider the minimization problem

$$\int_a^b [f(x) - f_0(x)]^2\, dx \to \min_f, \quad f_0(x) \in L_2[a,b], \quad f \in V[\mathcal{L}(p)]. \quad (4.3.62)$$

Since the wedge $V\left[\mathcal{L}(p)\right]$ is convex and closed in the Hilbert space $L_2[a,b]$, the solution $f(x)$ to the minimization problem (4.3.62) exists and is unique [142]. Let us obtain this solution in explicit form.

As above, by $W_2^{(n)}[a,b]$ we denote the space of functions such that their derivatives of order $n-1$ are absolutely continuous and their derivatives of order $n$ are square integrable on $[a,b]$.

Let us consider the nonlinear operator

$$\Omega : W_2^{(n)}[a,b] \to V\left[\mathcal{L}(p)\,\mathcal{L}(-p)\right]$$

acting according to the following rules. Let $F_0(x) \in W_2^{(n)}[a,b]$; then the function $F(x) = \Omega F_0(x)$ belongs to the wedge $V\left[\mathcal{L}(p)\,\mathcal{L}(-p)\right]$ and satisfies

1) the boundary conditions $F^{(i)}(a) = F_0^{(i)}(a)$ and $F^{(i)}(b) = F_0^{(i)}(b)$ for all $i = 0,\ldots,n-1$,

2) the inequality $F(x) \geq F_0(x)$ for all $x \in [a,b]$,

3) and the equality $\mathcal{L}(p)\,\mathcal{L}(-p)\,F(x) = 0$ for all $x \in [a,b]$ such that $F(x) > F_0(x)$.

It is asserted that the function

$$f(x) = \mathcal{L}(-p)\,\Omega\mathcal{L}(-p)^{-1} f_0(x)$$

is a solution to the minimization problem (4.3.62) in the wedge $V\left[\mathcal{L}(p)\right]$. Here

$$\mathcal{L}(-p)^{-1} : f(x) \to \int\limits_x^b z(\tau - x)\,f(\tau)\,d\tau$$

is a linear operator acting from the space $L_2[a,b]$ into the space of functions $F(x) \in W_2^{(n)}[a,b]$ such that $F^{(i)}(b) = 0$ for all $i = 0,1,\ldots,n-1$.

The proof of the above assertion is based on the general theorems on necessary and sufficient conditions for the existence of minimum points for the convex functions on the closed and convex subsets of Hilbert spaces. The completeness of $\mathcal{L}(p)$-splines with nonnegative coefficients in the wedge $V\left[\mathcal{L}(p)\right]$ is essentially used. In particular, this completeness allows one to obtain an approximate solution to the minimization problem (4.3.62) in the class of $\mathcal{L}(p)$-splines with nonnegative coefficients. The main procedure is the construction of the interpolating $\mathcal{L}(p)\,\mathcal{L}(-p)$-spline satisfying given boundary conditions.

The minimization problem (4.3.62) allows one to obtain a solution to the following problem of one-sided approximation (this problem is adjoint with respect to the problem (4.3.62)).

Let $F_0(x) \in W_2^{(n)}[a,b]$. We consider the minimization problem

$$\int_a^b [\mathcal{L}(\pm p) F(x)]^2 \, dx \to \min_F \tag{4.3.63}$$

on the set of the functions $F(x) \in W_2^{(n)}[a,b]$ satisfying

1) the boundary conditions $F^{(i)}(a) = F_0^{(i)}(a)$ and $F^{(i)}(b) = F_0^{(i)}(b)$ for all $i = 0, 1, \ldots, n-1$

2) and the inequality $F(x) \geq F_0(x)$ for all $x \in [a,b]$.

It turns out that the function $F(x) = \Omega F_0(x)$ is a solution to the problem of one-sided approximation (4.3.63); moreover, the function $F(x)$ is a solution both for the cases $+p$ and the case $-p$.

Let $f(x) \in V[\mathcal{L}(p)]$. The set of functions $v(x) \in L_2[a,b]$ such that $f(x) + \varepsilon v(x) \in V[\mathcal{L}(p)]$ for a sufficiently small $\varepsilon > 0$ is called the wedge of feasible directions at the point $f(x)$ and is denoted by $V_{f(x)}[\mathcal{L}(p)]$.

Of special interest is the conjugate wedge $V_{f(x)}[\mathcal{L}(p)]^*$. From the previous subsection it follows that a function $\delta(x)$ belongs to $V_{f(x)}[\mathcal{L}(p)]^*$ if and only if $\delta(x) = \mathcal{L}(p)\Delta(x)$, where the function $\Delta(x)$ belongs to the space $W_2^{(n)}[a,b]$ and satisfies

1) the boundary conditions $\Delta^{(i)}(a) = \Delta^{(i)}(b) = 0$, $i = 0, 1, \ldots, n-1$,

2) the inequality $\Delta(x) \geq 0$ for $x \in [a,b]$,

3) and the equality $\Delta(x) = 0$ for all $x \in \operatorname{supp} \mathcal{L}(p) f(x)$.

Here by $\operatorname{supp} u(x)$ we denote the support of the generalized function $u(x) = \mathcal{L}(p) f(x)$.

In particular, $V_{f(x)}[\mathcal{L}(p)]^* = 0$ if and only if the set $\operatorname{supp} \mathcal{L}(p) f(x) = 0$ is dense in $[a,b]$. For example, if $f(x) \in C^n[a,b]$ and $\mathcal{L}(p) f(x) > 0$ for all $x \in [a,b]$, then $\operatorname{supp} \mathcal{L}(p) f(x) = [a,b]$ and $V_{f(x)}[\mathcal{L}(p)]^* = 0$.

Thus, a function $f(x) \in V[\mathcal{L}(p)]$ may be a solution to the minimization problem (4.3.62) for a function $f_0(x) \in L_2[a,b]$, $f_0(x) \neq f(x)$, if and only if the set $\mathcal{L}(p) f(x)$ is not dense in $[a,b]$.

## 4.4 Convergence of monotone and convex solutions of integral equations

**1.** Let us consider the equation

$$Au = f, \quad f \in F, \tag{4.4.64}$$

where the linear closed operator $A : B \to F$ is defined on a linear set $D_A \subset B$ and acts from the Banach space $B$ of real functions given on a segment

$[a, b]$ into a linear topological space $F$. Let the sought-for solution to equation (4.4.64) be a monotone or convex function from the definition domain of the operator $A$.

The necessity of considering the properties of monotonicity and convexity may be explained by the following reasons. First of all, these properties, as a rule, have the natural physical interpretation and, hence, can often be separated from the *a priori* knowledge about the behavior of the physical process under study. In addition, they correspond to an intuitive understanding of simplicity of the solution's structure. From the mathematical standpoint, finally, the monotonicity and convexity are so strong constraints that in a number of cases they provide a uniform approximation for a solution to the problem (4.4.64). For example, in [241] it is shown that if $B$ and $F$ are functional normalized spaces, $A$ is a linear bounded operator, and the solution to the problem (4.4.64) is a piecewise smooth monotone function $u_0(x)$ bounded by a constant $K$, then, using a sequence $\{f_n\} \in F_n$ convergent to the element $f_0 = Au_0$ in the space $F$, we can construct a sequence $\{u_n\}$ uniformly convergent to the function $u_0(x)$ on any segment $[\alpha, \beta] \subset [a, b]$ which does not contain the endpoints of the segment $[a, b]$ and the discontinuity points of the function $u_0(x)$.

In a number of cases, the necessity of preserving (for physical reasons) the properties of monotonicity or convexity, the simplicity of these conditions, and their positive influence on stability of approximate solutions to the problem under consideration lead to creation of special algorithms based on the use of such *a priori* information on the sought-for solution [193].

Here we study the convergence of a sequence of monotone or convex functions if it is known that the images of the members of this sequence converge to some element in the topology of the space $F$.

Let us make a preliminary remark. The convergence of the sequence $\{f_n\}$ formed by the images $f_n = Au_n$ of elements $u_n$ in the space $F$ is equivalent to the convergence of the sequence $\{u_n\}$ in a special space $T$ whose topology is induced by the operator $A$ and by the topology of the space $F$. In order to simplify our discussion, therefore, in the sequel we consider the operator $A$ as the embedding operator $B$ into $T$. Let us formulate the requirements that are imposed on the space $T$.

Let $T$ be a linear topological space [142] of real functions given on a segment $[a, b]$; it is assumed that the topology of the space $T$ satisfies the following condition: for any $\varepsilon > 0$ and for any segment $[x_1, x_2] \subseteq [a, b]$, $x_1 < x_2$, there exists a zero neighborhood of the space $T$ such that this neighborhood does not contain the functions $u \in T$ for which

$$|u(x)| \geq \varepsilon \quad \text{for all} \quad x \in [x_1, x_2]. \tag{4.4.65}$$

In the sequel, we base on the assumption that a sequence $\{u_n\} \subset F$ of monotone or convex functions converges in the space $T$ to some function $u_0(x)$. Let us consider a necessary criterion for the convergence of a sequence $\{u_n\}$ in the topology of the space $T$.

**Lemma 4.4.35** *In order that a sequence $\{u_n\} \subset T$ be convergent in the space $T$ to zero as $n \to \infty$, it is necessary that for any $\varepsilon > 0$ and any segment $[x_1, x_2] \subseteq [a, b]$, $x_1 < x_2$, there exist only a finite number of members of the sequence $\{u_n\}$ satisfying the condition $|u_n(x)| \geq \varepsilon$ for all $x \in [x_1, x_2]$.*

The validity of this lemma follows from the condition (4.4.65) imposed on the topology of the space $T$.

The next lemmas show that the condition (4.4.65) is not too burdensome.

**Lemma 4.4.36** *Let $T$ be a linear normalized functional space. In order that the topology of $T$ satisfy the condition* (4.4.65)*, it is necessary and sufficient that for any $\varepsilon > 0$ there exists a positive number $\delta$ such that for any function $u \in T$, $\|u\|_T < \delta$, there exists a set $X$ everywhere dense on $[a, b]$ such that*

$$|u(x)| < \varepsilon \quad \textit{for any} \quad x \in X. \tag{4.4.66}$$

The conclusion of this lemma is a reformulation of the condition (4.4.65) to the case when the space $T$ is normalized.

Obviously, all usually-used functional normalized spaces satisfy the condition (4.4.66).

**Lemma 4.4.37** *The weak topology of the space $L_2[a, b]$ satisfies the condition* (4.4.65).

**Proof**. Let the topology of the space $T$ coincide with the weak topology of the space $L_2[a, b]$, i.e., the defining system of zero neighborhoods for the space $T$ is formed by the sets

$$\left\{u : \left|\int_a^b u(x)\, v_k(x)\, dx\right| < \delta, \;\; k = 1, 2, \ldots, n\right\},$$

where $\{v_k\}$ is an arbitrary finite set of elements from $L_2[a, b]$ and $\delta$ is an arbitrary positive number. Let us assume that the topology of the space $T$ does not satisfy the condition (4.4.65), i.e., for some $\varepsilon > 0$ and a segment $[x_1, x_2]$, $x_1 < x_2$, in any zero neighborhood of the space $T$ there exists a function $u_0(x)$ such that $|u_0(x)| \geq \varepsilon$ for any $x \in [x_1, x_2]$.

Without loss of generality, we may assume that $u_0(x) \geq 0$ for all $x \in [x_1, x_2]$, since the space $L_2[a,b]$ and the space $T$ contain the function $u_0(x)$ and the function $|u_0(x)|$. But in the zero neighborhood

$$\left\{ u : \int_{x_1}^{x_2} u(x)\, dx < \varepsilon (x_2 - x_1)/2 \right\}$$

there exist no such nonnegative functions. This contradiction completes the proof of the lemma.

Thus, the condition (4.4.65) is fulfilled under sufficiently weak assumptions on the topology of the space $T$.

**2.** In order to study the uniform convergence of monotone functions, we restrict ourselves to the consideration of nondecreasing functions only, since for nonincreasing functions all assertions can be proved similarly.

Let $M \subset T$ be a set of nondecreasing functions $u \in T$ given on a segment $[a,b]$.

**Theorem 4.4.74** *Let a sequence $\{u_n\} \subset M$ be convergent in the space $T$ to a function $u(x) \in T$, i.e.,*

$$u_n \to u \in T \quad as \quad n \to \infty. \tag{4.4.67}$$

*Then, for any continuity point $x_0 \in (a,b)$ of the function $u(x)$ there exists*

$$\lim_{n\to\infty} u_n(x_0) = u(x_0), \quad x_0 \in (a,b). \tag{4.4.68}$$

**Proof.** Let us assume that the function $u(x)$ is continuous at the point $x_0 \in (a,b)$, but the condition (4.4.68) is not fulfilled, i.e., for some $\varepsilon > 0$ there exists a subsequence $\{u_{n_k}\}$ such that for any $k = 1, 2, \ldots$ the following inequality is valid:

$$u(x_0) - u_{n_k}(x_0) \geq 2\varepsilon, \tag{4.4.69}$$

or

$$u_{n_k}(x_0) - u(x_0) \geq 2\varepsilon. \tag{4.4.70}$$

An infinite number of members of the subsequence $\{u_{n_k}(x_0)\}$ satisfy one of the inequalities (4.4.69) or (4.4.70); therefore, we may assume that the entire subsequence $\{u_{n_k}(x_0)\}$ satisfies, for example, the condition (4.4.69).

Since the point $x_0 \in (a,b)$ is a continuity point of the function $u(x)$, there exists a number $\delta$, $0 < \delta < \min\{x_0 - a, b - x_0\}$, such that for any $x \in [x_0 - \delta, x_0 + \delta]$ the inequality

$$u(x_0) - \varepsilon < u(x) < u(x_0) + \varepsilon \tag{4.4.71}$$

is valid.

Then, from the conditions (4.4.69) and (4.4.71) it follows that for any $x \in [x_0 - \delta, x_0]$ the inequality

$$|u(x) - u_{n_k}(x)| \geq u(x) - u_{n_k}(x) > u(x_0) - \varepsilon - u_{n_k}(x_0) \geq \varepsilon$$

holds for any $k = 1, 2 \ldots$; by Lemma 4.4.35, this result contradicts the condition (4.4.67) of convergence of the sequence $\{u_n\}$. The proof of the theorem is completed.

The following example shows that the continuity of the function $u(x)$ at the endpoints of the segment $[a, b]$ is not a sufficient condition for existence of the limit (4.4.68) for $x_0 = a$ or $x_0 = b$. Indeed, the sequence of monotone functions $u_n(x)$ defined by the relations

$$u_n(x) = \begin{cases} 0, & 0 \leq x < 1 - 1/n, \\ 1, & 1 - 1/n \leq x \leq 1, \end{cases}$$

strongly converges to zero in the space $L_p[0, 1]$, $p > 1$, but does not converge to zero at the endpoint $x = 1$.

The conditions for convergence at the endpoints of the segment $[a, b]$ are formulated in the next theorem.

**Theorem 4.4.75** *Let the condition* (4.4.67) *be fulfilled and a function* $u(x)$ *be continuous at the point* $x = a$ $(x = b)$. *For the existence of the limit*

$$\lim_{n\to\infty} u_n(a) = u(a) \quad \left(\lim_{n\to\infty} u_n(b) = u(b)\right) \tag{4.4.72}$$

*it is necessary and sufficient that for any* $\varepsilon > 0$ *there exists a number* $\delta > 0$ *such that for any* $n$, $n > N(\varepsilon)$, *the inequality*

$$|u_n(a + \delta) - u_n(a)| < \varepsilon \quad (|u_n(b) - u_n(b - \delta)| < \varepsilon) \tag{4.4.73}$$

*is valid.*

**Proof.** Let us consider the left endpoint of the segment $[a, b]$ (the second case is proved similarly).

*Sufficiency.* Let the condition (4.4.73) be fulfilled, but the sequence $\{u_n(a)\}$ do not converge to $u(a)$ as $n \to \infty$, i.e., there exist $\varepsilon > 0$ and a subsequence $\{u_{n_k}(a)\}$ such that either

$$u(a) - u_{n_k}(a) \geq 3\varepsilon, \quad k = 1, 2, \ldots, \tag{4.4.74}$$

or

$$u_{n_k}(a) - u(a) \geq 3\varepsilon, \quad k = 1, 2, \ldots .$$

For definiteness we assume that the inequality (4.4.74) is valid. From the condition (4.4.73) it follows that for a given $\varepsilon$ there exist a number $\delta > 0$ and a number $K$ such that for $k \geq K$ the inequality

$$u_{n_k}(a+\delta) - u_{n_k}(a) < \varepsilon \tag{4.4.75}$$

holds. Finally, since the function $u(x)$ is continuous at the point $x = a$, there exists a number $\delta_1$, $0 < \delta_1 < \delta$, such that for any $x \in [a, a+\delta_1]$ the inequality

$$u(a) - \varepsilon < u(x) < u(a) + \varepsilon \tag{4.4.76}$$

takes place.

Then, from the conditions (4.4.74)–(4.4.76) it follows that for any $x \in [a, a+\delta_1]$ the inequalities

$$|u(x) - u_{n_k}(x)| \geq u(x) - u_{n_k}(x) > u(a) - \varepsilon - u_{n_k}(a) - \varepsilon \geq \varepsilon$$

are valid for any $k \geq K$; by Lemma 4.4.35, this conclusion contradicts the condition (4.4.67) for convergence of the sequence $\{u_n\} \subset M$.

*Necessity.* Let the limit (4.4.72) exist, but the condition (4.4.73) be not fulfilled, i.e., there exists $\varepsilon > 0$ such that for any $x \in [a, b]$ there exists a subsequence $\{u_{n_k}\}$ dependent on a point $x$ such that for each member of this subsequence the inequality

$$u_{n_k}(x) - u_{n_k}(a) \geq 3\varepsilon, \quad k = 1, 2, \dots, \tag{4.4.77}$$

is valid.

Since the function $u(x)$ is continuous at the point $x = a$, there exists $\delta > 0$ such that for any $x \in [a, a+\delta]$ the inequality

$$u(a) - \varepsilon < u(x) < u(a) + \varepsilon \tag{4.4.78}$$

holds.

Suppose $\{u_{n_k}\}$ is a subsequence each member of which satisfies the condition (4.4.77) for $x = a + \delta/2$. From the existence of the limit (4.4.72) it follows that there exists a number $K$ such that

$$u(a) - \varepsilon < u_{n_k}(x) < u(a) + \varepsilon \tag{4.4.79}$$

for $k \geq K$. Then, from the conditions (4.4.77)–(4.4.79) it follows that for any $x \in [a + \delta/2, a + \delta]$ the inequalities

$$|u(x) - u_{n_k}(x)| \geq u_{n_k}(x) - u(x) > u_{n_k}(a) + 3\varepsilon - u(a) - \varepsilon > \varepsilon$$

are valid for any $k \geq K$; by Lemma 4.4.35, this conclusion contradicts the condition (4.4.67) for convergence of the sequence $\{u_n\} \subset M$. Theorem 4.4.75 is proved.

Thus, the convergence of a sequence of monotone functions in the space $T$ to a continuous function $u(x)$ together with other conditions ensuring the existence of the limits (4.4.72) guarantee the pointwise convergence of this sequence to the function $u(x)$. Hence, this convergence is uniform [207]. In other words, the next two theorems are valid.

**Theorem 4.4.76** *If a sequence $\{u_n\} \subset M$ converges in the space $T$ to a continuous function $u(x) \in T$, then $u \in M$ and the sequence $\{u_n\}$ converges to the function $u(x)$ uniformly on any segment $[a+\varepsilon, b-\varepsilon]$, $\varepsilon > 0$.*

**Theorem 4.4.77** *If a sequence $\{u_n\} \subset M$ converges in the space $T$ to a continuous function $u(x) \in T$ and the conditions* (4.4.73) *for convergence at the endpoints of the segment $[a, b]$ are fulfilled, then $u(x) \in M$ and the sequence $\{u_n\}$ converges to the function $u(x)$ uniformly on the entire segment $[a, b]$.*

**3.** Here we formulate some conditions for which a sequence of convex functions converges in the metric of the space $C[a, b]$ of functions continuous on $[a, b]$ and in the metric of the space $C^1[a, b]$ of functions continuously differentiable on $[a, b]$.

A function $u(x)$ given on a segment $[a, b]$ is said to be *convex downwards* (*convex upwards*) if for any $x, y \in [a, b]$ and any number $\alpha$, $0 \leq \alpha \leq 1$, the inequality

$$u(\alpha x + (1-\alpha)y) \leq \alpha u(x) + (1-\alpha)u(y) \tag{4.4.80}$$

holds (respectively, $u(\alpha x + (1-\alpha)y) \geq \alpha u(x) + (1-\alpha)u(y)$).

We restrict ourselves to the consideration of functions convex downwards only, since all properties for functions convex upwards are proved similarly.

Let us recall some facts relating to the properties of convex functions.

1) A function $u(x)$ given on $[a, b]$ and convex in the sense of the definition (4.4.80) is bounded on $[a, b]$ and, hence, is continuous in the interval $(a, b)$.

2) A function $u(x)$ convex on $[a, b]$ has the finite left-hand derivative $u'_-(x_0)$ and the finite right-hand derivative $u'_+(x_0)$ at each point $x_0 \in (a, b)$. At the left (right) endpoint of the segment $[a, b]$, the convex function $u(x)$ has, perhaps, the infinite right-hand derivative $u'_+(x_0)$ (the left-hand derivative $u'_-(x_0)$). For the one-sided derivatives of the convex function $u(x)$ the following inequalities are valid:

$$u'_-(x) \leq u'_-(y) \quad \text{for} \quad a < x < y \leq b,$$

$$u'_+(x) \leq u'_+(y) \quad \text{for} \quad a \leq x < y < b,$$

$$u'_{-}(x) \le u'_{+}(y) \quad \text{for} \quad a < x \le y < b.$$

3) A function $u(x)$ convex on $(a, b)$ can be represented on this interval in the form

$$u(x) = u(c) + \int_{c}^{x} v(t)\,dt, \tag{4.4.81}$$

where $c$ is any point belonging to $(a, b)$, whereas $v(t)$ is a function nondecreasing on $(a, b)$ and bounded on each segment $[p, q] \subseteq [a, b]$.

Let $W \subset T$ be the set of functions $u(x)$ convex downwards on a segment $[a, b]$ and belonging to a linear topological space $T$ whose topology satisfies the condition (4.4.65).

**Theorem 4.4.78** *If a sequence of functions $\{u_n\} \subset W$ converges in the space $T$ and for any $u_n(x)$, $n = 1, 2, \ldots$, the inequalities*

$$u'_{+n}(a) \ge -K, \quad u'_{-n}(b) \le K \tag{4.4.82}$$

*hold for some $K$, where $u'_{+n}(x)$ and $u'_{-n}(x)$ are the right- and left-hand derivatives of the function $u_n(x)$, then the sequence $\{u_n\}$ converges to a continuous convex function $u(x)$ uniformly on $[a, b]$.*

**Proof**. First of all we note that the conditions (4.4.82) ensure the continuity of the functions $u_n(x)$, $n = 1, 2, \ldots$, on the segment $[a, b]$. Therefore, it is sufficient to show that the sequence $\{u_n\}$ is fundamental in the space $C[a, b]$ of continuous functions, since the existence of the limiting continuous function $u(x)$ in this case follows from the completeness of the space $C[a, b]$, whereas the convexity of the function $u(x)$ is obvious.

Let us assume that the sequence $\{u_n\}$ is not fundamental in the space $C[a, b]$, i.e., there exists a number $\varepsilon > 0$ such that for any $n = 1, 2, \ldots$ there exists a number $m$ and a point $x_n \in [a, b]$ such that the inequality

$$|u_n(x_n) - u_{n+m}(x_n)| \ge 2\varepsilon$$

holds. Without loss of generality we may assume that the value of $u_n(x_n) - u_{n+m}(x_n)$ has a constant (for example, positive) sign for all $n = 1, 2, \ldots$, since otherwise we can isolate a sequence for which this assumption takes place. Thus, the following inequalities are valid:

$$u_n(x_n) - u_{n+m}(x_n) \ge 2\varepsilon, \quad n = 1, 2, \ldots, \quad m = m(n). \tag{4.4.83}$$

Let $x_0 \in [a, b]$ be a limiting point of the set $\{x_n\}$ (for simplicity we assume that $x_0$ is the only limiting point of this set, so that $x_n \to x_0$ as $n \to \infty$).

Let $N$ be a number such that for any $n > N$ we have $x_n \in [\alpha, \beta] \subseteq [a, b]$, where $[\alpha, \beta]$ is a segment that contains the point $x_0$ and whose length $\delta$ satisfies the condition

$$\delta = \beta - \alpha < \varepsilon / 2K.$$

By virtue of the condition (4.4.83), then, for any $x \in [\alpha, \beta]$ the inequalities

$$\begin{aligned} |u_n(x) - u_{n+m}(x)| &= u_n(x) - u_{n+m}(x) \geq \\ &\geq u_n(x) + u'_{+n}(x_n)(x - x_n) - u_{n+m}(x_n) - K|x - x_n| \\ &\geq u_n(x_n) - u_{n+m}(x_n) - 2K\delta > \varepsilon \end{aligned}$$

are valid for any $n > N$; by Lemma 4.4.35, this conclusion contradicts the condition for convergence of the sequence $\{u_n\}$ in the space $T$. The proof of the theorem is completed.

**Corollary.** *If a sequence $\{u_n\} \subset W$ converges in the space $T$, then it converges to a function $u(x)$ uniformly on any segment $[a + \varepsilon, b - \varepsilon]$, $\varepsilon > 0$, where $u(x)$ is continuous and convex on the interval $(a, b)$.*

**Proof**. By Theorem 4.4.78, it is sufficient to show that the sequences of the right-hand derivatives $\left\{u'_{+n}(a + \varepsilon)\right\}$ and the left-hand derivatives $\left\{u'_{-n}(b - \varepsilon)\right\}$ taken at the endpoints of the segment $[a + \varepsilon, b - \varepsilon]$ are uniformly bounded. Let us prove, for example, that $u'_{-n}(b - \varepsilon) \leq K$, $n = 1, 2, \ldots$. By assuming the contrary, we may suppose that for any $n = 2, 3, \ldots$ the inequality

$$u'_{-n}(b - \varepsilon) \geq 4(M_{n-1} + 1)/\varepsilon, \quad M_n = \sup_{x, y \in [b - \varepsilon, b - \varepsilon/4]} |u_n(x) - u_n(y)|$$

holds, where $n = 1, 2, \ldots$. If this is not the case, we pass to a sequence for which this inequality is fulfilled. Then, the total variation of the function $u_n(x)$ on each of the three segments $[b - \varepsilon, b - 3\varepsilon/4]$, $[b - 3\varepsilon/4, b - \varepsilon/2]$, and $[b - \varepsilon/2, b - \varepsilon/4]$ should be no less than $M_{n-1} + 1$; since the function $u_n(x)$ strictly increases on the segment $[b - \varepsilon, b]$ by virtue of $u'_{-n}(b - \varepsilon) > 0$, at least on one of these segments (for example, on $[b - \varepsilon/2, b - \varepsilon/4]$) the inequality

$$|u_n(x) - u_{n-1}(x)| \geq 1, \quad x \in [b - \varepsilon/2, b - \varepsilon/4], \tag{4.4.84}$$

should be fulfilled.

From $\{u_n\}$ we choose a subsequence $\{u_{n_k}\}$ such that any two adjacent members of this subsequence satisfy the condition (4.4.84) on the segment $[b - \varepsilon/2, b - \varepsilon/4]$. Like in the proof of Theorem 4.4.78, it is easy to show that the sequence $\{u_{n_k}\}$ does not converge in the space $T$. This contradiction completes the proof of the corollary.

As in the case of monotone functions, thus, we have formulated the conditions under which a sequence $\{u_n\}$ of convex functions convergent in the space

$T$ to a function $u(x)$ converges to the function $u(x)$ uniformly on the segment $[a,b]$. For convex functions it is possible to obtain a stronger convergence, namely, the convergence in the metric of the space $C^1[a,b]$ of continuously differentiable functions.

**Theorem 4.4.79** *If a sequence $\{u_n\}$ of convex functions converges to a function $u(x)$ uniformly on a segment $[a,b]$ and $x \in (a,b)$ is a continuity point of the derivative $u'(x)$ of the limiting function $u(x)$, then the following limits exist:*

$$\lim_{n\to\infty} u'_{-n}(x_0) = u'(x_0) = \lim_{n\to\infty} u'_{+n}(x_0).$$

*Here $u'_{-n}(x_0)$ and $u'_{+n}(x_0)$ are the left-hand and right-hand derivatives of the function $u_n(x)$ at the point $x_0$, respectively.*

**Proof**. According to (4.4.81), any function $u(x)$ convex on $(a,b)$ can be represented on this interval in the form $u(x) = u(c) + \int_c^x v(t)\,dt$, where $c$ is an arbitrary point from $(a,b)$ and, for example, $u'_-(t)$ or $u'_+(t)$ may be chosen as a function $v(t)$.

Let $x_0 \in (p,q) \subset [a,b]$ and $T_1$ be a linear normalized space of functions $v \in T_1$ summable on $[p,q]$; the norm in $T_1$ is defined as

$$\|v\|_{T_1} = \left\| \int_p^x v(t)\,dt \right\|_C = \max_{x\in[p,q]} \left| \int_p^x v(t)\,dt \right|.$$

Obviously, the topology of the space $T_1$ satisfies the condition (4.4.65).

The sequences $\{u'_{-n}(x)\}$ and $\{u'_{+n}(x)\}$ of monotone functions converge in the space $T_1$ to the functions $u'_-(x)$ and $u'_+(x)$, which are continuous at the point $x_0 \in (p,q)$; by Theorem 4.4.74, hence, there exist the limits

$$\lim_{n\to\infty} u'_{+n}(x_0) = u'_+(x_0) = u'(x_0) = u'_-(x_0) = \lim_{n\to\infty} u'_{-n}(x_0).$$

The theorem is proved.

**Corollary 1.** *If a sequence $\{u_n\}$ of convex functions converges in the space $T$ to a continuously differentiable function $u(x)$, then on any segment $[a+\varepsilon, b-\varepsilon]$ there exist the uniform limits*

$$\lim_{n\to\infty} u_n(x) = u(x), \tag{4.4.85}$$

$$\lim_{n\to\infty} u'_{-n}(x) = u'(x) = \lim_{n\to\infty} u'_{+n}(x). \tag{4.4.86}$$

**Corollary 2.** *If a sequence $\{u_n\}$ of convex functions satisfying the condition* (4.4.82) *converges to a continuously differentiable function $u(x)$ in the space $T$ and the conditions $\lim_{n\to\infty} u'_{+n}(a) = u'(a)$, $\lim_{n\to\infty} u'_{-n}(b) = u'(b)$ are fulfilled, then on the entire segment $[a, b]$ there exist the uniform limits* (4.4.85) *and* (4.4.86)*.*

The validity of these corollaries directly follows from the theorems given in [239], [207].

Coming back to the original problem of studying the properties of monotone or convex solutions to the operator equation (4.4.64) and relying upon the above results, it is possible to assert the following.

Let a linear operator $A : B \to F$ and the topology of the space $F$ ensure the fulfilment of the condition (4.4.65), i.e., for any $\varepsilon > 0$ and any segment $[x_1, x_2] \subset [a, b]$ there exists a zero neighborhood of the space $F$ such that this neighborhood does not contain the images $Au$ of those functions $u \in B$ for which $|u(x)| \geq \varepsilon$ for all $x \in [x_1, x_2]$. Then, all the above-proved theorems remain valid if the condition (4.4.67) for convergence of a sequence $\{u_n\}$ to a function $u(x)$ in the topology of the space $T$ is replaced by the condition of convergence of the images $Au_n$ of the sequence $\{u_n\}$ to an element $Au$ in the topology of the space $F$.

**4.** The good approximative qualities of monotone functions are based on the following main facts:

a) under sufficiently general conditions, an operator $A$ mapping a compact set $M$ onto a set $N$ in a one-to-one manner has the continuous inverse operator mapping $N \to M$ (Tikhonov's theorem on the continuity of an inverse operator [265]);

b) a set of bounded monotone functions is compact in the space $L_2[a, b]$ of functions square integrable on the segment $[a, b]$ (the Helly theorem).

Thus, the use of information on monotonicity (or convexity, which is closely related to monotonicity) of the sough-for solution actually results in a narrowing of the domain of feasible solutions such that this narrowing ensures a conditional correctness (in another terminology, the correctness according to A. N. Tikhonov) of the problem under consideration and, hence, makes it possible to use the numerical quasisolution-type methods [115]. This approach is already well studied and allows one to develop very efficient numerical algorithms [193].

Here we also use the information on monotonicity of the sought-for solution, but the requirement of its boundedness we replace by the weaker condition of semiboundedness. Moreover, the above approach allows one to take into account the often-known information on a pointwise error in the right-hand side of an equation, whereas the majority of other regularization methods are based on the information on a rougher integral error estimate.

Thus, let us assume that on a nonempty set $D_A \subset L_2[a,b]$ a closed (not necessarily linear) operator $A : L_2 \to F$ is given; this operator acts into the normalized space $F$ of bounded functions $f(y)$, $y \in [c,d]$, whose norm satisfies the condition

$$\|f\|_C \leq K\|f\|_F, \quad f \in F. \tag{4.4.87}$$

In other words, $F$ is embedded into the space $C$.

Let $M$ be the set of nonnegative nondecreasing functions from $L_2[a,b]$: $M = \{u \in L_2 : u \nearrow, u(a) \geq 0\}$ and $N$ be the image of the set $M$ in the space $F : N = A[M]$. It is assumed that $M \subset D_A$.

Let us consider the operator equation

$$Au = f, \quad f \in N. \tag{4.4.88}$$

The set of its solutions is denoted by $\bar{U}$:

$$\bar{U} = \{\bar{u} : u \in M, \ \ A\bar{u} = f\}.$$

Equation (4.4.88) is always solvable, i.e., $\bar{U} \neq \emptyset$. Further we assume that an approximation $\tilde{f}$ is given instead of the right-hand side $f$ of equation (4.4.88) and

$$\left|\tilde{f}(y) - f(y)\right| \leq \delta(y),$$

where $\delta(y)$ is a known continuous function.

Let us introduce the uncertainty set

$$S_\delta = \left\{u \in M : \left|Au - \tilde{f}\right| \leq \delta(y)\right\}$$

and consider the problem: to find a function $\tilde{u} \in S_\delta$ such that

$$\tilde{u}(b) = \min_{u \in S_\delta} u(b). \tag{4.4.89}$$

Let $U_\delta^1$ be the set of solutions to problem (4.4.89).

**Theorem 4.4.80** (*existence*). *If $A$ is a continuous operator, then the problem* (4.4.89) *has at least a one solution, i.e., $U_\delta^1 \neq \emptyset$.*

**Proof**. Let

$$\mu = \inf_{u \in S_\delta} u(b). \tag{4.4.90}$$

Then, there exists a sequence $\{u_n\} \in S_\delta$ such that $u_n(b) \to \mu$ as $n \to \infty$. However, the set $\{u_n\}$ of nondecreasing bounded functions is compact in $L_2[a,b]$;

hence, there exists a subsequence $\{u_{n_k}\}$ such that $u_{n_k} \to u_0$ in $L_2$ as $k \to \infty$ and $u_0$ is a monotone function satisfying the inequality $u_0\,(b) \geq \mu$.

Without loss of generality, we may assume that $u_0\,(b) = \mu$. Indeed, the case $u_0\,(b) > \mu$ is possible only when the function $u_0\,(x)$ is not continuous for $x = b$, but in this case it is possible to supplement the definition of $u_0\,(x)$ by continuity at this point.

Since $A$ is a continuous operator, we have

$$Au_{n_k} \xrightarrow{F} Au_0 \quad \text{as} \quad k \to \infty. \tag{4.4.91}$$

Now we show that $u_0 \in S_\delta$, i.e.

$$\left| Au_0 - \tilde{f}\,(y) \right| \leq \delta\,(y)\,. \tag{4.4.92}$$

Indeed,

$$\begin{aligned} \left| Au_0 - \tilde{f}\,(y) \right| &\leq |Au_0 - Au_{n_k}| + \left| Au_{n_k} - \tilde{f}\,(y) \right| \leq \\ &\leq |Au_0 - Au_{n_k}| + \delta\,(y)\,; \end{aligned} \tag{4.4.93}$$

by virtue of (4.4.64) and (4.4.68) we have

$$|Au_0 - Au_{n_k}| \leq k_1 \, \|Au_0 - Au_{n_k}\|_F \to 0 \quad \text{as} \quad k \to \infty.$$

Passing in (4.4.93) to the pointwise limit as $k \to \infty$, we obtain the inequality (4.4.92). Thus, we have proved the existence of an element $u_0 \in S_\delta$ such that

$$u_0\,(b) = \mu = \inf_{u \in S_\delta} u\,(b)\,.$$

The proof of the theorem is completed.

Thus, the necessary boundedness of the set of feasible solutions to equation (4.4.88) in the case under consideration is attained by introducing the special minimization problem (4.4.89).

In a more general case, when the operator $A$ is not continuous, the existence of at least a one solution to the problem (4.4.89) cannot be guaranteed. As a set of feasible solutions, we may take the set

$$U_\delta^2 = \{u : u \in S_\delta; \;\; u\,(b) \leq 2\mu\}\,, \tag{4.4.94}$$

where $\mu$ is defined by formula (4.4.90). Obviously, the set $U_\delta^2$ is never empty and $U_\delta^2 \subseteq U_\delta^1$ for $U_\delta^1 \neq \emptyset$.

Now we consider the conditions under which the approximate solutions defined by the problem (4.4.89) or by the inequality (4.4.94) converge to the exact solution to equation (4.4.88). Since we do not assume the uniqueness of a solution to equation (4.4.88), it is natural to describe the proximity between the exact solution and an approximate one in terms of $\beta$-continuity.

Let us recall some definitions.

The semideviation [131] of a set $X \subset B$ from a set $Y \subset B$ in a linear normalized space $B$ is called the quantity

$$\beta(X,Y) = \sup_{x \in X} \rho(x,Y), \quad \rho(x,Y) = \inf_{y \in Y} \|x-y\|_B .$$

**Theorem 4.4.81** (*convergence*). *Let $F = C[a,b]$. Then,*

1) $\beta\left(U_\delta^2, \bar{U}\right) \to 0$ *as* $\|\delta\|_C \to 0$;

2) *if the operator $A$ is continuous, the more exact relation takes place:* $\beta\left(U_\delta^1, \bar{U}\right) \to 0$ *as* $\|\delta\|_C \to 0$.

We prove the theorem only for the first case, since the second case can be proved similarly. It is necessary to show that for any $\varepsilon > 0$ there exists $\delta_0 > 0$ such that $\forall \delta(y) : \|\delta\|_C < \delta_0$ and $\forall \tilde{f} : \left|f - \tilde{f}\right| \le \delta(y)$ the following inequality is valid: $\beta\left(U_\delta^2, \bar{U}\right) < \varepsilon$.

Let this be not the case, i.e., there exists $\varepsilon > 0$ such that $\forall \delta_0 : \|\delta\|_C < \delta_0$ and $\forall \tilde{f}(y) : \left|f(y) - \tilde{f}(y)\right| \le \delta(y)$ the inequality $\beta\left(U_\delta^2, \bar{U}\right) \ge \varepsilon$ holds.

Let $\delta_0 = 1/n,\ \delta = \delta_n(y),\ \tilde{f} = f_n(y)$, and $\beta\left(U_\delta^2, \bar{U}\right) \ge \varepsilon$, i.e., $\exists u_n \in U_\delta^2$ such that

$$\rho\left(u_n, \bar{U}\right) \ge \varepsilon. \tag{4.4.95}$$

But the set $\{u_n\}$ is compact in $L_2$; hence, from $\{u_n\}$ we can choose a subsequence $\{u_{n_k}\}$ converging in the metric of the space $L_2$ to an element $u_0$ such that $|Au_{n_k} - f| \le |Au_{n_k} - f_{n_k}| + |f_{n_k} - f| \le 2\delta_{n_k}$. Then, the similar inequality is also valid for the norm: $\|Au_{n_k} - f\|_C \le 2\|\delta_{n_k}\|_C$. It follows from here that $Au_{n_k} \to f$ in $L_2$ as $k \to \infty$.

Since the operator $A$ is closed, we have $u_0 \in \bar{U}$; this conclusion contradicts the inequality (4.4.95). The theorem is proved.

## 4.5 The descriptive regularization method and the stable solution of linear systems

**1.** Contrary to the traditional method of regularization for which, as a rule, it is required to specify mean-square deviations, here the information on pointwise estimates of errors in input data is used; this peculiarity leads to another type of problems, in particular, to the necessity of solving the standard problems of linear programming. Methods for determining the monotone solutions or the solutions which have some specific properties (the least variation, a given number of extrema, convexity (concavity), etc.) are considered. Note that the solution of linear systems is a main well-studied problem of numerical linear algebra. There exist many methods for solving such systems. The traditional approach consists in the following: a certain scheme of computations [280] is chosen in advance, i.e., a method for solving a linear system is fixed (such a choice already predetermines the quality of resulting approximate solutions) and then, if it is possible, the influence of errors in input data and of errors arising during the process of computations on the final result is estimated.

This approach is well suited for a wide class of well-posed linear systems. It is clear *a priori* (and this is confirmed by the careful analysis of limiting accuracy of solutions to linear systems by various methods) that the majority of known and most widespread methods for solving the linear systems with well-conditioned matrices have almost identical quality and are characterized by the same parameters, being distinguished only in details.

The evolution of computing practice results in the necessity of constructing the solutions, which are stable with respect to small perturbations in matrices and right-hand sides, for ill-posed linear systems, too. The classical methods, as a rule, are rather inefficient and lead to such solutions that are difficult to interpret reasonably. It often happens that a computing process cannot be terminated successfully and stops because of emergency situations, however carefully these situations were taken into account by a programmer.

New opportunities for solving the problems of linear algebra are provided by the regularization method. Its main idea is the stabilization of a problem by the introduction of an additional parameter (of regularization) and by its choice from a compromise condition between a sufficiently accurate approximation of the original problem and the stability of this approximation. In practice, the fulfillment of such a condition is not always a simple problem and requires the application of special algorithms for the choice of the regularization parameter. Usually, all procedures of such a choice are based on matching the accuracy of the solution to the regularized problem with the accuracy of input data (the matrix and the right-hand side). One of the most widespread procedures is the choice of the regularization parameter based on the residual principle: the matching of the residual obtained for the regularized solution with errors in the matrix and in the right-hand side. This approach is substantiated theoretically [172], [189], can easily be expressed algorithmically, and leads to the stabilization of an approximate solution to a linear system with inexact data.

In practice, the application of the regularization method with any procedure of specification of the regularization parameter comes across a number of difficulties. First of all, as a rule, the multiple solution of a regularized problem is required. Second, which is more important, the traditional manner of application of the regularization method is essentially motivated by a way of specification of mean-square estimates for errors in input data. In practice, as our experience shows, only pointwise estimates for the accuracy of input data are known more often. Under these conditions the application of the traditional approach is possible but is not always justified, since such an approach leads to the inadequate use of *a priori* information.

In this section we consider several methods for solving linear systems; these methods are conceptually simple, but are reliable in relation to perturbations of input data. They are based on the use of information on pointwise estimates of errors in input data.

Suppose it is required to solve the linear system

$$Au = f, \quad \bar{U} = \{u : Au = f\}, \tag{4.5.96}$$

where $A$ is an $m \times n$-matrix, $f \in R_m$, and $u \in R_n$ is the sought-for vector. It is assumed that the system (4.5.96) with exact data $\{A, f\}$ is consistent, but, perhaps, is not definite, i.e., the multiplicity of solutions to (4.5.96) is allowed. It is usually required to solve (4.5.96) when approximate data $\{\tilde{A}, \tilde{f}\}$ are known, where $\tilde{f} \in R_m$ and $\tilde{A}$ is a matrix of the same type as $A$. It is possible that the system $\tilde{A}u = \tilde{f}$ is unsolvable in the classical sense; if this system is solvable (assume, uniquely), it may be unstable, i.e., it may not approximate (in one sense or another) the set of solutions to (4.5.96). The above facts may easily be confirmed by trivial examples.

Suppose the data $\{\tilde{A}, \tilde{f}\}$ and $\{A, f\}$ are related by the inequalities $|a_{ij} - \tilde{a}_{ij}| \le \Delta_{ij}$ and $|f_i - \tilde{f}_i| \le \delta_i$, where $\Delta_{ij}$ and $\delta_i$ are given and tend to zero when the accuracy of input data are increasing.

It is easy to see that for any vector $u \in R_n$ satisfying (4.5.96) the following inequalities are valid:

$$\left|\sum_{j=1}^{n} \tilde{a}_{ij} u_j - \tilde{f}_i\right| \le \sum_{j=1}^{n} \Delta_{ij} |u_j| + \delta_i, \quad i = 1, \ldots, m. \tag{4.5.97}$$

Let $\tilde{U}$ be the set of all vectors $u \in R_n$ satisfying (4.5.97); this set is said to be the *uncertainty set of solutions* (the set of formal solutions). It is easy to see that its diameter may tend to infinity. Nevertheless, this set can be used as a basis for a method of determining the meaningful solutions. To accomplish this, it is natural to formulate a rule of selection which is able to reduce the number of feasible solutions and to enhance the level of their reliability.

**2.** Let $D \neq \emptyset$, $D \subseteq R_n$, be a closed set given *a priori*. It is assumed that $\bar{U} \cap D = \bar{U}_D \neq \emptyset$. We call $\bar{U}_D$ the set of $D$-solutions to the linear system (4.5.96). Let us assume that $L$ is a $q \times n$-matrix such that the following complementing condition is fulfilled: if $D$ is not bounded, then there exist constants $a \geq 0$, $b \geq 0$, and $\gamma > 0$ such that

$$a \left\| Au \right\|_1 + b \left\| Lu \right\|_2 \geq \gamma \left\| u \right\| \quad \forall u \in D,$$

where $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|$ are some norms in $R_m$, $R_q$, and $R_n$.

**Definition.** *A vector $u^* \in \bar{U}_D$ is said to be the L-minimal D-solution to the system* (4.5.96) *if*

$$\left\| Lu^* \right\|_2 = \inf_{u \in \bar{U}_D} \left\| Lu \right\|_2 . \tag{4.5.98}$$

The set of all $L$-minimal solutions to (4.5.96) is denoted by $U_D^*$. For $D = R_n$ and $L = E$, according to [265], we come to the conclusion that $u^*$ is the normal solution to (4.5.96).

**Theorem 4.5.82** *The set of L-minimal D-solutions to* (4.5.96) *is nonempty, i.e., $U_D^* \neq \emptyset$.*

**Proof.** Let $\mu^* = \inf_{u \in \bar{U}_D} \|Lu\|_2$ and the vectors $u_n$ satisfying the inequalities

$$\left\| Lu_n \right\|_2 \leq \mu^* + 1/n , \quad n = 1, 2, \ldots , \tag{4.5.99}$$

be chosen. Let us consider the cases:

1) $D$ is bounded and

2) $D$ is not bounded.

In case 1), the sequence $\{u_n\}$ is certainly bounded and, hence, is compact. Without loss of generality, we assume that $\{u_n\}$ converges: $u_n \to u^*$ as $n \to \infty$. Since $D$ is closed, we have $u^* \in D$. It is also obvious that $Au^* = f$, i.e., $u^* \in \bar{U}_D$. From (4.5.99) we obtain $\|Lu^*\|_2 = \mu^*$. Thus, the vector $u^* \in U_D^*$.

In case 2) we have

$$a \left\| Au_n \right\|_1 + b \left\| Lu_n \right\|_2 \leq a \left\| f \right\|_1 + b \left( \mu^* + 1 \right) .$$

By virtue of the complementing condition, from here we conclude that the sequence $\{u_n\}$ is bounded. Further, the reasoning is repeated. The theorem is proved.

**Remark 4.5.1** *Let the set $D$ be convex. By $N_A$ we denote the set of all $u \in R_n$ such that $Au = 0$. If we assume that for all $u, v \in D$ such that $u - v \in N_A$, $u \neq v$, the inequality*

$$\left\| L \left( \frac{u + v}{2} \right) \right\|_2 < \frac{\|Lu\|_2}{2} + \frac{\|Lv\|_2}{2}$$

*holds, then the set* $U_D^* = \{u^*\}$*, i.e., this set is one-element.*

Indeed, if $u_1^*$, $u_2^* \in U_D^*$, then $(u_1^* + u_2^*)/2 \in D$ and $Au_1^* = Au_2^* = f$, i.e., $u_1^* - u_2^* \in N_A$. Assuming that $u_1^* \neq u_2^*$, we obtain

$$\left\| L\left(\frac{u_1^* + u_2^*}{2}\right)\right\|_2 < \frac{\mu^*}{2} + \frac{\mu^*}{2} = \mu^*;$$

this inequality contradicts the definition of $\mu^*$. Hence, $u_1^* = u_2^*$, which was to be proved.

**3.** Now we consider a method for solving the system (4.5.96) with approximate data $\{\tilde{A}, \tilde{f}\}$. Let $\tilde{U}_D = \tilde{U} \cap D$. The set $\tilde{U}_D$ is called the uncertainty set of $D$-solutions (the set of formal $D$-solutions). Let us find $\tilde{u}^* \in \tilde{U}_D$ such that

$$\|L\tilde{u}^*\|_2 = \inf_{u \in \tilde{U}_D} \|Lu\|_2 = \tilde{\mu}^*. \tag{4.5.100}$$

**Theorem 4.5.83** *The set* $\tilde{U}_D^* = \left\{u \in \tilde{U}_D : \|Lu\|_2 = \tilde{\mu}^*\right\}$ *is not empty for sufficiently small values of* $\Delta_{ij}$.

**Proof**. Let us choose a sequence $\{u^n\}$, $u^n \in \tilde{U}_D$, $n = 1, 2, \ldots$, for which

$$\|Lu^n\|_2 \leq \tilde{\mu}^* + 1/n, \quad n = 1, 2, \ldots . \tag{4.5.101}$$

Let $D$ be unbounded. According to the condition (4.5.97), we have

$$\xi_i = \left|\sum_{j=1}^n a_{ij} u_j^n - f_i\right| \leq 2\left(\sum_{j=1}^n \Delta_{ij}\left|u_j^n\right| + \delta_i\right) \leq 2\left(\Delta \|u^n\|_{II} + \delta\right),$$

where

$$\Delta = \max_i \left(\sum_{j=1}^n \Delta_{ij}^2\right)^{1/2}, \quad \delta = \max_i \delta_i, \quad \|u\|_{II} = \left(\sum_{i=1}^n u_i^2\right)^{1/2}.$$

Thus,

$$\|Au^n - f\|_I \equiv \max_i \xi_i \leq 2\left(\Delta \|u^n\|_{II} + \delta\right).$$

Using the equivalence of norms in a finite-dimensional space, we obtain

$$\|Au^n\|_{II} \leq a_1 \|Au^n\|_I \leq a_1 \left\{\|f\|_I + 2\left(\Delta \|u^n\|_{II} + \delta\right)\right\}, \quad a_1 = \text{const}.$$

By virtue of the complementing condition and the above estimates, we come to the inequality

$$\gamma \|u^n\| \leq a\left\{a_1 \|f\|_I + 2a_1\left(\Delta \|u^n\|_{II} + \delta\right)\right\} + b\left(\tilde{\mu}^* + 1\right). \tag{4.5.102}$$

Since $\|u\|_{II} \leq a_2 \|u\| \ \forall u \in R_n$ ($a_2$ = const), for a sufficiently small $\Delta$ it follows from (4.5.102) that $\|u^n\|$ are bounded. Isolating, if it is necessary, a subsequence, we assume that $\{u^n\}$ converges: $u^n \to \tilde{u}^*$. It is easy to see that $\tilde{u}^* \in D$ and $\tilde{u}^* \in \tilde{U}$, i.e., $\tilde{u}^* \in \tilde{U}_D$. From (4.5.101) it follows that $\|\tilde{u}^*\|_2 = \tilde{\mu}^*$. Thus, $\tilde{u}^* \in \tilde{U}_D^*$.

If $D$ is bounded, it is possible to use the compactness of the sequence $\{u^n\}$ and the previous reasoning. The theorem is proved.

**Remark 4.5.2** *Under the hypotheses of Theorem* 4.5.83*, the set* $\tilde{U}_D^*$ *of approximate* $D$*-solutions to the system* (4.5.96) *is bounded. This assertion is easy to establish, using the scheme of the proof for Theorem* 4.5.82*. Note also that the condition of sufficient smallness of* $\Delta_{ij}$ *in Theorem* 4.5.83 *is not necessary if* $a = 0$ *or if* $D$ *is bounded. Let* $\tilde{U}_D^\varepsilon = \{u \in \tilde{U}_D : \|Lu\|_2 \leq \tilde{\mu}^* + \varepsilon\}$*, where* $\varepsilon > 0$ *is given. The set* $\tilde{U}_D^\varepsilon$ *is also bounded. Any vector* $u \in \tilde{U}_D^\varepsilon$ *is said to be the* $\varepsilon$*-solution to the problem* (4.5.100)*.*

**4.** Let $D_1$ and $D_2$ be nonempty sets in $R_n$. As is known, the $\beta$-distance of the set $D_1$ from $D_2$ is called the quantity

$$\beta(D_1, D_2) \equiv \sup_{u \in D_1} \inf_{v \in D_2} \|u - v\|.$$

The main property of $\varepsilon$-solutions to the problem (4.5.100) is established by the next theorem.

**Theorem 4.5.84** *Let* $\varepsilon$, $\Delta_{ij}$, $\delta_i \to 0$ *as* $i \to \infty$*. Then,* $\beta\left(\tilde{U}_D^\varepsilon, U_D^*\right) \to 0$.

Indeed, for any $u \in \tilde{U}_D^\varepsilon$ we have

$$\left|\sum_{j=1}^n a_{ij} u_j - f_i\right| \leq 2\left(\Delta \|u\|_{II} + \delta\right) \leq 2\left(\Delta C + \delta\right).$$

Here $C$ = const bounds the set $\tilde{U}_D^\varepsilon$. Based on the previous inequalities, it is easy to complete the proof if we assume that the theorem is false. The theorem is proved.

**Remark 4.5.3.** *If* $U_D^* = \{u^*\}$*, then the* $\beta$*-convergence means that*

$$\sup_{u \in \tilde{U}_D^\varepsilon} \|u - u^*\| \to 0 \quad as \quad \varepsilon,\ \Delta_{ij},\ \delta_i \to 0.$$

**5.** Now we consider various cases when the complementing condition is fulfilled.

1) Let $b = 0$. Then $L$ may be arbitrary. In the case $L = 0$, the problem (4.5.100) is reduced to searching for any formal $D$-solution, i.e., for a vector $\tilde{u} \in \tilde{U} \cap D$. If under the above conditions we solve $2n$ problems

$$(-1)^l u_j \to \min_{u \in \tilde{U}_D}, \quad l = 0, 1, \tag{4.5.103}$$

then, denoting their solutions by $\tilde{u}_j^{\max}$ and $\tilde{u}_j^{\min}$ for $l = 1$ and $l = 0$, respectively, we obtain the inequalities $\tilde{u}_j^{\min} \le \tilde{u}_j \le \tilde{u}_j^{\max}$, $j = 1, \dots, n$, for any vector $\tilde{u}$ (including for $u \in U_D^*$).

The problem of estimation for any linear form of $u_j$ is considered similarly. If $D$ includes the constraints of the type $u_j \ge 0$, it is easy to see that the problem (4.5.103) is reduced to the standard problem of linear programming. When $A$ is singular, the system $\tilde{A}u = \tilde{f}$ with the constraint $u \in D$ (or without it) may be inconsistent.

2) If $L = E$, the consistency condition is certainly fulfilled with $a = 0$ for any matrix $A$ (including $A = 0$, in this case $\bar{U}_D = D$) and for any $D$.

3) Let $D = \{u \in R_n : 0 \le u_1 \le \ldots \le u_n\}$. This case is often encountered when finding the monotonically increasing solutions of Fredholm integral equations of the first kind (see Subsection 4 of the previous section). Let $\forall u \in D : \|u\|_2 = u_n$. It is obvious that $\|Lu\|_2 = \|u\|_1$, i.e., the consistency condition is fulfilled for $a = 0$, $b = 1$, and $\gamma = 1$.

If the set $D$ does not contain the constraint $u_1 \ge 0$, we may put $\|Lu\|_2 = u_n - u_1$. In this case it should be required that $Ae \ne 0$, where $e = (1, \dots, 1)$. Then, the consistency conditions are again fulfilled. In both the cases, we come to the necessity of solving the standard problem of linear programming. The condition $Ae \ne 0$ is almost always valid in practice and is easy to check, since this condition is equivalent to the inequality $\max_i \left| \sum_{j=1}^n a_{ij} \right| > 0$.

Note that the above algorithms essentially differ from the algorithm following from [77].

In order to obtain more qualitative approximations, it is also possible to include the condition of convexity (concavity) for the sought-for solution into the constraints defining the set $D$ (if this agrees with *a priori* data), i.e., the condition of the type

$$u_{i+1} - 2u_i + u_{i-1} \le 0 \quad (u_{i+1} - 2u_i + u_{i-1} \ge 0)\,.$$

Numerical experiments show [193] that such conditions substantially stabilize the solution even for a sufficiently large $n$ (see also the previous section).

As in case 1), we may also hope that some estimates for approximate solutions can be obtained.

4) Let $D = \{u : 0 \leq u_1 \leq \ldots \leq u_{j_1},\ u_{j_1} \geq u_{j_1+1} \geq \ldots \geq u_n \geq 0\}$, where $1 < j_1 < n$. This case corresponds to the unimodal behavior of the sought-for solution. We may put $\|Lu\|_2 = u_{j_1}$. Again, $\|Lu\|_2 = \|u\|_1$, $u \in D$, i.e., the consistency conditions are fulfilled for $a = 0$, $b = 1$, and $\gamma = 1$. By including the convexity condition into $D$, more qualitative solutions may be obtained. The estimates for solutions are obtained similarly to the previous cases.

If $D$ does not contain the constraints of the type $u_1 \geq 0$ and $u_n \geq 0$, we may put $\|Lu\|_2 = (u_{j_1} - u_1) + (u_{j_1} - u_n)$, $u \in D$. Again, it should be required that the condition $Ae \neq 0$ be fulfilled.

This approach is easily generalized to the case of "multiextremal" solutions if only the extreme points $1 < j_1 < \ldots < j_k < n$ are known and between successive extrema the solution behaves naturally. In this case we may put $\|Lu\|_2 = (u_{j_1} - u_1) + (u_{j_1} - u_{j_2}) + \cdots + (u_n - u_{j_k})$ and to require that $Ae \neq 0$. It is assumed that the solution increases at the beginning and at the end of its definition range. Other cases of natural behavior of the sought-for solutions are considered similarly. The problems to which we come are the problems of linear programming. The estimation of solutions is performed according to (4.5.103) and the constraint $\|Lu\|_2 \leq \tilde{\mu}^*$ may be included into the definition of $\tilde{U}_D$ *a posteriori*.

Note that conceptually these approaches are similar to the descriptive regularization method developed in a number of works [193], [191], [192], [272], but lead to other numerical algorithms.

5) It is easy to see that in the above cases, $\|Lu\|_2$ is a variation of the solution. In the case when *a priori* information is not known, we have

$$\|Lu\|_2 = |u_2 - u_1| + \cdots + |u_n - u_{n-1}|\,.$$

Requiring that $Ae \neq 0$ and putting $D = R_n$, we come to the fulfillment of the complementing condition. The solution to a linear system under the condition of minimal variation is called the solution with minimal variation.

**6.** It is well-known that the normal solution to the system (4.5.96) is called the solution with minimal root-mean-square norm [265].

**Proposition.** *If $u^*$ is the normal solution to* (4.5.96)*, then this solution can be represented in the form $u^* = A^*v^*$, $v^* \in R_m$, where $A^*$ is the transpose of the matrix $A$.*

Indeed, the equation

$$AA^*v = f \tag{4.5.104}$$

is always solvable when the system (4.5.96) is solvable. It is sufficient to use the singular value decomposition of $A$ (see [55]). Let us put $u^* = A^*v$, where $v$ is any solution to (4.5.104).

**Lemma 4.5.38** *The element $u^*$ does not depend on the choice of $v$.*

Indeed, let $v_1$ and $v_2$ be two solutions to (4.5.104). Putting $u_1^* = A^*v_1$ and $u_2^* = A^*v_2$, we obtain

$$\|u_1^* - u_2^*\|_{II}^2 = (u_1^* - u_2^*, A^*v_1 - A^*v_2) = (Au_1^* - Au_2^*, v_1 - v_2) = 0.$$

From here it follows that $u_1^* = u_2^*$. Further, let $z$ be such that $Az = 0$. Then,

$$(u^*, z) = (A^*v, z) = (v, Az) = 0 \quad \forall z,$$

i.e., $u^* \perp z \ \forall z$. But any solution to (4.5.96) can be represented in the form $\hat{u} = u^* + z$. Then, using the Pythagorean theorem, we obtain

$$\|\hat{u}\|_{II}^2 = \|u^* + z\|_{II}^2 = \|u^*\|_{II}^2 + \|z\|_{II}^2 \geq \|u^*\|_{II}^2 .$$

Thus, $u^*$ is the normal solution. The lemma is proved.

Note that the above discussion proves the existence and uniqueness of the normal solution.

Let $v^*$ be the normal solution to (4.5.104). Then, $u^* = A^*v^*$.

Now we use a procedure (similar to that for (4.5.96)) to search for the normal solution to (4.5.104). As an objective functional we take $\|v\|_{II}$. If $\tilde{v}_\varepsilon^*$ is an $\varepsilon$-solution to the corresponding system of inequalities, then $\tilde{u}_\varepsilon^* = \tilde{A}^*\tilde{v}_\varepsilon^*$ is taken as an approximate solution for the normal solution to (4.5.96). Obviously,

$$\begin{aligned} \|\tilde{u}_\varepsilon^* - u^*\|_{II} &\leq \left\|\tilde{A}^*\tilde{v}_\varepsilon^* - A^*\tilde{v}_\varepsilon^*\right\|_{II} + \|A^*\tilde{v}_\varepsilon^* - A^*v^*\|_{II} \leq \\ &\leq \left\|\tilde{A}^* - A^*\right\| \|\tilde{v}_\varepsilon^*\|_{II} + \|A^*\tilde{v}_\varepsilon^* - A^*v^*\|_{II} \leq \\ &\leq \left\|\tilde{A} - A\right\| (\|v^*\|_{II} + \varepsilon) + \|A^*(\tilde{v}_\varepsilon^* - v^*)\|_{II} , \end{aligned} \tag{4.5.105}$$

since $\|\tilde{v}_\varepsilon^*\|_{II} \leq \|v^*\|_{II} + \varepsilon$. Let $\hat{\Delta} = \max_{i,j} \Delta_{ij}$ and $\hat{\delta} = \max_i \delta_i$. Using the result obtained in [187] on the basis of (4.5.105), we come to the validity of the next theorem.

**Theorem 4.5.85** *The following estimate is fulfilled: $\|\tilde{u}_\varepsilon^* - u^*\| = O\left(\hat{\Delta} + \hat{\delta}\right)$. Here $\tilde{u}_\varepsilon^*$ is obtained according to the above procedure.*

**7.** Often it is necessary to search for a solution to the system (4.5.96) in the "sourcewise" form $u = Bv$, where the $n \times s$-matrix $B$ is given. In essence, in the previous discussions we have considered the cases $B = E$ and $B = A^*$. The condition of sourcewise representability imposes certain constraints on the behavior of the sought-for solution. This solution may also be defined by additional conditions. For example, if the solution to (4.5.96) is sought from the set $D = \{u : 0 \leq u_1 \leq \ldots \leq u_n\}$, we may put $u_1 = v_1$, $u_1 = v_1 + v_2, \ldots, u_n = v_1 + \cdots + v_n$, where $v_i \geq 0$, $i = 1, \ldots, n$. In this case, the objective functional takes the form $\|Lu\|_2 = u_n = v_1 + \cdots + v_n$ and the matrix $A$ is replaced by $AB$. Note that here the constraints imposed on $v$ have the simplest form.

The linear system

$$ABv = f \tag{4.5.106}$$

can be solved by all the above methods. If $\tilde{v}$ is a solution to (4.5.106) with approximate data, then an approximate solution to the original system (4.5.96) has the form $\tilde{u} = \tilde{B}\tilde{v}$, where $\tilde{B}$ is an approximation for the matrix $B$. Additional constraints may also be imposed on the sought-for solution to the system (4.5.106).

The replacement $u = Bv$ can be used to reduce the number of required parameters and, hence, to enhance the stability of an approximate solution.

The consistency condition for the system (4.5.96) may be removed. Indeed, the system of equations

$$Au + \xi = f, \quad A'\xi = 0 \tag{4.5.107}$$

for $u$ and $\xi$ is always consistent and the component $\xi$ of the solution is always determined uniquely. Indeed, let $(u_1, \xi_1)$ and $(u_2, \xi_2)$ are two solutions. Then,

$$A(u_2 - u_1) = -(\xi_2 - \xi_1), \;\; A'(\xi_2 - \xi_1) = 0.$$

Multiplying the first equality by $A'$ from the left and using the second one, we obtain

$$A'A(u_2 - u_1) = 0;$$

hence, $A(u_2 - u_1) = 0$, i.e., $\xi_2 = \xi_1$.

Putting

$$\hat{A} = \begin{pmatrix} A & E \\ 0 & A' \end{pmatrix}, \quad \hat{f} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \quad \hat{u} = \begin{pmatrix} u \\ \xi \end{pmatrix},$$

the system (4.5.107) can be rewritten in matrix form as

$$\hat{A}\hat{u} = \hat{f}. \tag{4.5.108}$$

The above-proposed methods are applicable to the system (4.5.108).

A continuous analogue of the above algorithms was studied in [153] for numerical solution of Fredholm integral equations of the first kind with a quadratic objective functional. Some numerical results are given in [192].

**8.** Let $K_n = \{u \in R_n,\ u \geq 0\}$, where the vector inequality $x \geq y$ is equivalent to the set of number inequalities $x_i \geq y_i$.

**Lemma 4.5.39** *Let $D = K_n$ and $\|\cdot\|_2$ be the Euclidean norm. Then there exists a vector $v^* \in R_m$ such that*

$$L'Lu^* = A'v^*,$$

*where $L'$ and $A'$ are the transposes of the matrices $L$ and $A$, respectively.*

**Proof**. The Lagrange function for the problem (4.5.98) is of the form

$$F\,(u, \lambda) \equiv \|Lu\|_2^2 + (\lambda, Au - f)\,.$$

According to the Kuhn–Tucker theorem [126], there exists a vector $\lambda^* \in R_m$ such that

$$\left.\frac{\partial F}{\partial u}\right|_{u=u^*,\ \lambda=\lambda^*} = 2L'Lu^* + A'\lambda^* \geq 0, \quad \left(u^*, 2L'Lu^* + A'\lambda^*\right) = 0.$$

To complete the proof, it is sufficient to put $v^* = -\lambda^*/2$.

A relation between $u^* \in U_D^*$ and $\tilde{u}^* \in \tilde{U}_D^*$ is established by the next lemma.

**Lemma 4.5.40** *Let the hypotheses of Lemma* 4.5.39 *be fulfilled; then, for any $\tilde{u}^* \in \tilde{U}_D^*$ the following inequality is valid:*

$$\|Lu^* - L\tilde{u}^*\|^2 \leq (Lu^*, Lu^* - L\tilde{u}^*)\,.$$

**Proof**. Since $u^* \in \tilde{U}_D$, then from the convexity of $\tilde{U}_D$ it follows that for any $\theta \in (0, 1)$ we have

$$\begin{aligned} \|L\tilde{u}^*\|_2^2 \quad &\leq \quad \|L\,[\theta u^* + (1-\theta)\,\tilde{u}^*]\|_2^2 \leq \|L\tilde{u}^*\|_2^2 + \\ &+ 2\theta\,(L\tilde{u}^*, Lu^* - L\tilde{u}^*) + \theta^2\,\|Lu^* - L\tilde{u}^*\|_2^2\,; \end{aligned}$$

hence,

$$(1-\theta/2)\,\|Lu^* - L\tilde{u}^*\|_2^2 \leq (Lu^*, Lu^* - L\tilde{u}^*)\,. \tag{4.5.109}$$

Passing in (4.5.109) to the limit as $\theta \to 0+$, we come to the conclusion of the lemma.

Now we estimate an accuracy of approximation for the $L$-minimum $D$-solution to the system (4.5.96) by approximate $L$-minimum $D$-solutions.

**Theorem 4.5.86** *Let the hypotheses of Lemma* 4.5.38 *be fulfilled. Then, for sufficiently small values of* $\Delta_{ij}$ *the estimate*

$$\sup_{u\in\tilde{U}_D^*} \|u^* - u\| = O\left(\delta^{1/2} + \Delta^{1/2}\right)$$

*is valid, where* $u^* \in U_D^*$, $\Delta = \max_i \left(\sum_{j=1}^n \Delta_{ij}^2\right)^{1/2}$, $\delta = \max_i \delta_i$.

**Proof.** By Lemmas 4.5.39 and 4.5.40 we obtain

$$\begin{aligned} \|Lu^* - L\tilde{u}^*\|_2^2 &\le \left(L'Lu^*, u^* - \tilde{u}^*\right) = \\ &= \left(L'Lu^* - A'v^*, u^* - \tilde{u}^*\right) + (v^*, Au^* - A\tilde{u}^*) = \\ &= -\left(L'Lu^* - A'v^*, \tilde{u}^*\right) + (v^*, Au^* - A\tilde{u}^*) \le \\ &\le (v^*, Au^* - A\tilde{u}^*) \le \|v^*\|_1 \|Au^* - A\tilde{u}^*\|_1 . \end{aligned}$$

From the complementing condition it follows that

$$\begin{aligned} \gamma \|u^* - \tilde{u}^*\| &\le a \|Au^* - A\tilde{u}^*\|_1 + b \|Lu^* - L\tilde{u}^*\|_2 \le \\ &\le a \|Au^* - A\tilde{u}^*\|_1 + b \|v^*\|_2^{1/2} \|Au^* - A\tilde{u}^*\|_1^{1/2} . \end{aligned}$$

It is proved in [187] that

$$\|Au^* - A\tilde{u}^*\|_\infty = \|f - A\tilde{u}^*\|_\infty \le 2\left(\Delta \|\tilde{u}^*\|_2 + \delta\right),$$

where

$$\|f - A\tilde{u}^*\|_\infty = \max_{1\le i\le m} \left| f_i - \sum_{j=1}^n a_{ij}\tilde{u}_j^* \right|.$$

Since the set $\tilde{U}_D^*$ is bounded for sufficiently small values of $\Delta_{ij}$, we have

$$\sup_{u\in\tilde{U}_D^*} \|Au^* - A\tilde{u}^*\|_\infty = O(\Delta + \delta).$$

To complete the proof, it is sufficient to use the equivalence of vector norms in a finite-dimensional space.

## 4.6 Chebyshev's solutions to linear systems

**1.** As previously mentioned, a wide class of applied problems is reduced to the solution of linear systems of the form

$$u \in D:\ Au = f, \tag{4.6.110}$$

where $A$ is an $m \times n$-matrix, $f \in R_m$ is a given vector, $u \in D$ is the sought-for vector, and $D \subseteq R_n$ is a closed set of *a priori* constraints. It is assumed that (4.6.110) is a general-type linear system, i.e., this system, perhaps, is unsolvable in the classical sense, but if it is solvable, then it admits the multiplicity of solutions [55]. The solutions to the system (4.6.110) are understood in the generalized sense: as the solutions to the variational problem

$$\hat{m} = \inf \left\{ \varphi \left[ u \right], u \in D \right\}, \tag{4.6.111}$$

where $\varphi \left[ u \right] \equiv \left\| Au - f \right\|_\infty$ is the residual functional.

Thus, the Chebyshev pseudosolutions to the linear system (4.6.110) are considered. The set of such pseudosolutions is denoted by $\bar{U}$. It is clear that $\bar{U} \neq \emptyset$. The quantity $\hat{m}$ is called the (Chebyshev) measure of inconsistency for the linear system (4.6.110) [172]. It is easy to see that if $\hat{m} = 0$, then $\bar{U}$ coincides with the set of classical solutions to (4.6.110) satisfying the *a priori* conditions $D$. In the sequel, by the problem of solving the linear system (4.6.110) we mean the problem of finding the quantity $\hat{m}$ and the choice of some element from the set $\bar{U}$.

Let us assume that instead of (4.6.110) we have the following linear system with approximate data $\{\tilde{A}, \tilde{f}\}$:

$$\tilde{A}u = \tilde{f}, \quad u \in D. \tag{4.6.112}$$

Here $\tilde{f} \in R_m$ and the matrix $\tilde{A}$ is of the same type as the matrix $A$. The set $\tilde{U}$ of Chebyshev's pseudosolutions to the linear system (4.6.112) is defined as the set of solutions to the variational problem

$$\tilde{m} = \inf \left\{ \tilde{\varphi} \left[ u \right], \, u \in D \right\}, \tag{4.6.113}$$

where $\tilde{\varphi} \left[ u \right] \equiv \left\| \tilde{A}u - \tilde{f} \right\|_\infty$. Let the approximate input data $\{\tilde{A}, \tilde{f}\}$ be related to $\{A, f\}$ by the inequalities $|a_{ij} - \tilde{a}_{ij}| \leq \Delta_{ij}$ and $\left| f_i - \tilde{f}_i \right| \leq \delta_i$, where $\Delta_{ij}$ and $\delta_i$ are given and tend to zero when the accuracy of input data are enhanced. By simple examples it is easy to see that, in general, $\tilde{m}$ does not converge to $\hat{m}$ and the set $\tilde{U}$ does not approximate $\bar{U}$ in any reasonable sense if $\Delta_{ij}$ and $\delta_i$ converge to zero. Thus, we come to the problem of constructing stable approximations to $\hat{m}$ and to some element of the set $\bar{U}$ when the inexact data $S_\sigma = \{\tilde{A}, \tilde{f}, \sigma\}$, $\sigma \equiv \{\Delta_{ij}, \delta_i\}$, are fixed [172].

**2.** First we formulate a method for stable determination of $\hat{m}$ with the use of $S_\sigma$ and $\sigma$. It is easy to see that the following lemma is valid.

**Lemma 4.6.41** *For any $u \in R_m$ the inequalities*

$$\left|\sum_{j=1}^{n} \tilde{a}_{ij} u_j - \tilde{f}_i\right| \le \left|\sum_{j=1}^{n} a_{ij} u_j - f_i\right| + \sum_{j=1}^{n} \triangle_{ij} |u_j| + \delta_i, \tag{4.6.114}$$

$$\left|\sum_{j=1}^{n} a_{ij} u_j - f_i\right| \le \left|\sum_{j=1}^{n} \tilde{a}_{ij} u_j - \tilde{f}_i\right| + \sum_{j=1}^{n} \triangle_{ij} |u_j| + \delta_i \tag{4.6.115}$$

*take place.*

Let

$$\begin{aligned} \varphi_i [u] \ &\equiv \ \left|\sum_{j=1}^{n} a_{ij} u_j - f_i\right|, \quad \tilde{\varphi}_i [u] \equiv \left|\sum_{j=1}^{n} \tilde{a}_{ij} u_j - \tilde{f}_i\right|, \\ \tilde{\psi}_i [u] \ &\equiv \ \tilde{\varphi}_i [u] + \sum_{j=1}^{n} \triangle_{ij} |u_j| + \delta_i, \quad i = \overline{1, m}, \quad \tilde{\psi} [u] \equiv \max_{1 \le i \le n} \tilde{\psi}_i [u] . \end{aligned}$$

Now we formulate the following variational problem: to find

$$\tilde{\rho} \equiv \inf \left\{ \tilde{\psi} [u] , \, u \in D \right\} . \tag{4.6.116}$$

**Theorem 4.6.87** *The relation $\tilde{\rho} \to \hat{m}+$ is valid if $\triangle_{ij}$, $\delta_i \to 0$.*

**Proof**. Taking (4.6.114) and (4.6.115) into account, for any $\bar{u} \in \bar{U}$ we have

$$\hat{m} = \varphi [\bar{u}] \le \tilde{\rho} \le \tilde{\psi} [\bar{u}] \le \hat{m} + 2 \max_i \left( \sum_{j=1}^{n} \triangle_{ij} |\bar{u}_j| + \delta_i \right).$$

Here we used the obvious inequality $\max_i (b_i + c_i) \le \max_i b_i + \max_i c_i$ for any $b_i$, $c_i \ge 0$, $i = 1, \dots, m$. The theorem is proved.

**Remark 1.** *As a stable approximation to $\hat{m}$, we may take the value of the function $\tilde{\psi} [u]$ for an element $u = u^{\varpi(\sigma)}$ of a minimizing sequence for the problem* (4.6.116) *such that $\tilde{\rho} \le \tilde{\rho}^{\varpi(\sigma)} \le \tilde{\rho} + \varpi(\sigma)$, where $\tilde{\rho}^{\varpi(\sigma)} = \tilde{\psi} \left[u^{\varpi(\sigma)}\right]$ and $\varpi(\sigma) \to 0$ as $\sigma \to 0$.*

Let us consider the question of numerical solution of the variational problem (4.6.116).

**Lemma 4.6.42** *The set of solutions to the problem* (4.6.116) *is not empty.*

The proof follows from the closure of the set $D$ and from the boundedness of the piecewise linear function $\tilde{\psi} [u]$ from below.

**Lemma 4.6.43** *The problem* (4.6.116) *is equivalent to the following problem of mathematical programming: to find*

$$u_{n+1}^* = \min\{u_{n+1} : \tilde{\varphi}_i\,[u] \le u_{n+1},\ i = \overline{1,m},\ u_{n+1} \ge 0,\ u \in D\}. \qquad (4.6.117)$$

*Then,* $u_{n+1}^* = \tilde{\rho}$.

**Proof.** Let $\tilde{u}$ be a solution to the problem (4.6.116). Then, $\tilde{u} \in D$, i.e., the point $\{\tilde{u},\ \tilde{\rho}\}$ is feasible for the problem (4.6.117). Hence, $u_{n+1}^* \le \tilde{\rho}$. Let $\hat{u} \in D$ be an element such that $\tilde{\varphi}_i\,[\hat{u}] \le u_{n+1}^* < \tilde{\rho}$, $i = 1, \ldots, m$. Then, $\tilde{\varphi}\,[\hat{u}] < \tilde{\rho}$; this conclusion contradicts the definition of $\tilde{\rho}$. Thus, $u_{n+1}^* = \tilde{\rho}$. The further proof is obvious.

**Remark 2.** *If the set* $D = R_n^+ = \{u \in R_n\ |u_j \ge 0,\ j = 1, \ldots, n\}$, *then the problem* (4.6.117) *is a problem of linear programming and can efficiently be solved on the basis of standard algorithms.*

**3.** Now we turn to the problem of constructing the stable approximations for Chebyshev's pseudosolutions to the system (4.6.110). For this purpose, we use the approach proposed in [186], [187].

Suppose $L$ is a $q \times n$-matrix such that the complementing condition is fulfilled: if $D$ is not bounded, it is possible to specify constants $a \ge 0$, $b \ge 0$, and $\gamma > 0$ such that

$$a\,\|Au\|_m + b\,\|Lu\|_q \ge \gamma\,\|u\|_n \quad \forall u \in D, \qquad (4.6.118)$$

where $\|\cdot\|_m$, $\|\cdot\|_q$, and $\|\cdot\|_n$ are some norms in the spaces $R_m$, $R_q$, and $R_n$, respectively.

**Definition 7 .** *A vector* $u^* \in \bar{U}$ *is said to be the* $L$*-minimal Chebyshev pseudosolution to the linear system* (4.6.110) *if*

$$\|Lu^*\|_2 = \mu \equiv \inf\left\{\|Lu\|_2\,,\ u \in \bar{U}\right\}. \qquad (4.6.119)$$

The set of all $L$-minimal Chebyshev pseudosolutions to the linear system (4.6.110) is denoted by $\bar{U}^*$. Let $E$ be the unit matrix. When $D = R_n$ and $L = E$, from (4.6.115) we come to the conclusion that $u^*$ is the normal Chebyshev pseudosolution to the system (4.6.110).

**Theorem 4.6.88** *The set of* $L$*-minimal Chebyshev pseudosolutions* $\bar{U}^*$ *is nonempty, closed, and bounded.*

**Proof.** If the set $D$ is bounded, then the conclusion of this theorem follows from the continuity of the function $\Omega\,[u] \equiv \|Lu\|_2$ and from the closure and boundedness of the set $\bar{U}$. Let $D$ be not bounded and $\{u^k\}$ be a minimizing sequence for the problem (4.6.119), i.e.,

$$\mu \le \Omega\left[u^k\right] \le \mu + \varepsilon_k, \tag{4.6.120}$$

where $\varepsilon_k \to 0+$ as $k \to \infty$. Without loss of generality, we may assume that the sequence $\{\varepsilon_k\}$ is decreasing. Then, using the equivalence of norms in finite-dimensional spaces, we obtain

$$\begin{aligned} a\left\|Au^k\right\|_1 + b\left\|Lu^k\right\|_2 &\le a\left(\left\|Au^k - f\right\|_1 + \|f\|_1\right) + b\left(\mu + \varepsilon_1\right) \le \\ &\le a\left(C\hat{m} + \|f\|_1\right) + b\left(\mu + \varepsilon_1\right), \end{aligned}$$

where $C$ is a constant independent of $k$. By virtue of the complementing condition, we can conclude that the sequence $\{u^k\}$ is bounded. Without loss of generality, we may assume that $\{u^k\}$ converges, i.e., $u^k \to u^*$ as $k \to \infty$. Since $\bar{U}$ is closed, then $u^* \in \bar{U}$. From (4.6.120) we obtain $\Omega[u^*] = \mu$. Thus, $u^* \in \bar{U}^*$ and, hence, $\bar{U}^* \neq \emptyset$. The set $\bar{U}^*$ is closed, since the set $D$ is closed and the function $\Omega[u]$ is continuous. The boundedness of the set $\bar{U}^*$ follows from the complementing condition. The theorem is proved.

**Corollary.** *If the set $D$ is convex, then the set $\bar{U}^*$ is also convex.*

Along with the set $\bar{U}^*$ of solutions to the problem (4.6.119), we also consider the set

$$\bar{U}^*_\varepsilon \equiv \left\{u \in \bar{U} : \Omega[u] \le \mu + \varepsilon\right\},$$

where $\varepsilon \ge 0$ is given. It is clear that $\bar{U}^*_0 = \bar{U}^*$. Any element $\bar{u}^*_\varepsilon \in \bar{U}^*_\varepsilon$ is called an $\varepsilon$-solution to the problem (4.6.119). It is easy to see that $\bar{U}^*_\varepsilon$, $\varepsilon \ge 0$, is nonempty, closed, and bounded.

**Definition 8 .** *By an approximate method for solving the linear system* (4.6.110) *we mean the specification of a map $T_\sigma : T_\sigma(S_\sigma) = D_\sigma$, where $D_\sigma \neq \emptyset$ such that $\lim\limits_{\sigma\to 0} \sup\limits_{u \in D_\sigma} \varphi[u] = \hat{m}$.*

**Definition 9 .** *It is said that the sets $D_\sigma$ are $B$-convergent to a set $Z$ from $R_n$ as $\sigma \to 0$ if for any choice of elements $u_\sigma \in D_\sigma$ the set of limiting points of the family $\{u_\sigma\}$ is nonempty and belongs to the set $Z$.*

**Definition 10 .** *The sets $D_\sigma$ are $\beta$-convergent to a set $Z$ from $R_n$ as $\sigma \to 0$ if $\lim\limits_{\sigma\to 0} \beta(D_\sigma, Z) = 0$, where $\beta(D_\sigma, Z) = \sup\limits_{u\in D_\sigma} \inf\limits_{v \in Z} \|u - v\|$.*

**Definition 11 .** *An approximate method $T_\sigma$ for solving the linear system* (4.6.110) *is said to be $B$-regular* ($\beta$*-regular*) *if the sets $D_\sigma$ are $B$-convergent* ($\beta$*-convergent*) *to a set $Z \subseteq U$ as $\sigma \to 0$.*

Let us introduce the following set:

$$\tilde{U} = \left\{ u \in D : \, \tilde{\varphi}_i [u] \le \sum_{j=1}^{n} \Delta_{ij} |u_j| + \delta_i + \tilde{\rho}, \quad i = 1, 2, \ldots, m \right\}.$$

Here $\tilde{\rho}$ is defined in (4.6.116). The set $\tilde{U}$ is called the set of formal approximate solutions to the linear system (4.6.110). Let us consider the variational problem

$$\|L\tilde{u}^*\|_q = \tilde{\mu} = \inf \left\{ \|Lu\|_q \,\middle|\, u \in \tilde{U} \right\}. \tag{4.6.121}$$

By $\tilde{U}^*$ we denote the set of vectors for which (4.6.121) is fulfilled.

**Theorem 4.6.89** *The method based on solving the variational problem* (4.6.121) *is $\beta$-regular for* (4.6.116) *and the sets $\tilde{U}^*$ are $\beta$-convergent to the set $\bar{U}^*$ as $\sigma \to 0$.*

**Proof**. Note that the set $\tilde{U}$ is not empty, since $\bar{u} \subseteq \tilde{U}$. This follows from (4.6.114) and from the inequality $\hat{m} \le \tilde{\rho}$ obtained in Theorem 4.6.87. It is easy to see that the set $\tilde{U}$ is closed. Now we assume that the set $D$ of *a priori* constraints is bounded. Then, from the continuity of the functional $\Omega[u]$ it follows that the set $\tilde{U}^*$ is nonempty, closed, and bounded. Taking into account the inequalities (4.6.114) and (4.6.115), moreover, for any $u \in \tilde{U}^*$ and $i = 1, \ldots, m$ we have

$$\varphi_i [u] \le \left| \sum_{j=1}^{n} a_{ij} u_j - f_i \right| \le 2 \left[ \sum_{j=1}^{n} \Delta_{ij} |u_j| + \delta_i \right] + \tilde{\rho}. \tag{4.6.122}$$

Since the norms are equivalent in finite-dimensional spaces and the set $D$ is bounded, from here we obtain

$$\varphi_i [u] \le 2 (C\Delta + \delta) + \tilde{\rho}, \quad i = 1, 2, \ldots, m,$$

where $C$ is a constant, $\Delta = \max\limits_i \sum\limits_{j=1}^{n} \Delta_{ij}$, and $\delta = \max\limits_i \delta_i$. Thus,

$$\hat{m} \le \varphi [u] \le 2 (C\Delta + \delta) + \tilde{\rho} \quad \forall u \in \tilde{U}^*, \tag{4.6.123}$$

where $\varphi [u] = \max\limits_i \varphi_i [u]$. In addition, taking into account that $\bar{U} \subseteq \tilde{U}$ and, hence, $\tilde{\mu} \le \mu$, we have

$$\Omega [u] \le \mu \quad \forall u \in \tilde{U}^*. \tag{4.6.124}$$

Using (4.6.123) and (4.6.124) as well as Theorem 4.6.87, it is easy to prove that $\tilde{U}^*$ is nonempty, closed and bounded, since the complementing condition (4.6.118) is valid and the set $\tilde{U}$ is closed (the scheme of the proof coincides with the proof of Theorem 4.6.88).

The general case ($D$ is unbounded) is proved similarly with the use of the complementing condition.

**4.** Now we consider the case when the variational problem (4.6.121) is solved approximately. Let $\tilde{U}_\varepsilon^* = \left\{ u \in \tilde{U} : \|Lu\|_q \le \tilde{\mu} + \varepsilon \right\}$, where $\varepsilon \ge 0$.

**Theorem 4.6.90** *The methods based on the choice of the set $\tilde{U}_\varepsilon^*$ are $B$-regular for the linear system* (4.6.110) *if $\varepsilon > 0$ is fixed and is $\beta$-regular as $\sigma, \varepsilon \to 0$. The set $\tilde{U}_\varepsilon^*$ is $B$-convergent to the set $\bar{U}_\varepsilon^*$ as $\sigma \to 0$ ($\varepsilon > 0$ is fixed) and the set $\tilde{U}_\varepsilon^*$ is $\beta$-convergent to the set $\bar{U}^*$ as $\sigma, \varepsilon \to 0$.*

The proof is the same as for Theorem 4.6.89.

# Bibliography


[1] **Ageev, A. L.,** On a regularizing algorithm of searching for the basis of a linear operator kernel //*J. Comp. Math. and Math. Phys.* 23 (5). 1041–1051. 1983.

[2] **Alberg, Jh., Nilson, E. and Walsh, Jh.,** *The theory of splines and its application.* Moscow: Mir, 1972.

[3] **Alifanov, O. M.,** *Inverse problems of heat exchange.* Moscow: Mashinostroenie, 1988.

[4] **Alifanov, O. M. and Rumyantzev, S. B.,** On stability of methods for solving linear ill-posed problems // *Docl. Acad. Nauk USSR.* 248 (6). 1289–1291. 1979.

[5] **Allachverdiev, D. E.,** On the approximation rate of completely continuous operators by finite-dimensional operators // *Sci. Notes of Azerb. Univ.* 27–35. 1957.

[6] **Arestov, V. V.,** On uniform regularization of the problem on determining the values of an operator // *Math. Notes.* 22 (2). 231–244. 1977.

[7] **Arefyeva, M. V.,** A fast regularizing algorithm for solving the convolution-type Fredholm integral equations of the first kind // *Numerical Analysis in FORTRAN. Methods and Algorithms.* Moscow: Moscow University Press, 1978.

[8] **Arefyeva, M. V.,** On optimal regularization of linear operator equations with random errors // *Computing Methods and Programming.* Issue 26. 93–116. Moscow: Moscow University Press, 1977.

[9] **Arefyeva, M. V.,** Some questions of optimal regularization for stochastic operator equations in Hilbert spaces // *Theory and Methods for Solving Ill-Posed Problems.* P. 15. Saratov: Saratov University Press, 1985.

[10] **Arefyeva, M. V.,** Correlation and spectral analysis of stationary random processes. Part 1 // *Numerical Analysis on FORTRAN*. Issue 15. 3–29. Moscow: Moscow University Press, 1976.

[11] **Arefyeva, M. V.,** Correlation and spectral analysis of stationary random processes. Part 2 // *Numerical Analysis on FORTRAN*. Issue 20. 67–72. Moscow: Moscow University Press, 1977.

[12] **Arefyeva, M. V. and Sysoev, A. F.,** Comparative analysis of some filtration algorithms // *Numerical Analysis: Methods, Algorithms, Programs*. 80–93. Moscow: Moscow University Press, 1983.

[13] **Arefyeva, M. V.,** Solving the convolution-type equations by the regularization method with the use of fast Fourier transform and residual criterion // *Computing Methods and Programming*. Issue 37. 51–68. Moscow: Moscow University Press, 1981.

[14] **Arefyeva, M. V. and Sysoev, A. F.,** On image restoration by the regularization method with the use of fast Fourier transform // *Methods and Algorithms in Numerical Analysis*. 3–27. Moscow: Moscow University Press, 1981.

[15] **Arefyeva, M. V. and Sysoev, A. F.,** On some methods of digital filtration // *Methods and Algorithms in Numerical Analysis*. 65–75. Moscow: Moscow University Press, 1982.

[16] **Arefyeva, M. V. and Sysoev, A. F.,** Fast regularizing algorithms for digital image restoration // *Computing Methods and Programming*. Issue 39. Moscow: Moscow University Press, 1983.

[17] **Arsenin, V. Ya.,** *On methods for solving ill-posed problems*. Moscow: MIFI, 1973.

[18] **Arsenin, V. Ya. and Ivanov, V. V.,** On optimal regularization // *Docl. Acad. Nauk USSR*. 182 (1). 9–12. 1968.

[19] **Arushanian, O. B.,** *Automatization of constructing the subroutine libraries*. Moscow: Moscow University Press, 1988.

[20] **Arsenin, V. Ya. and Zyabrev, N. B.** *On construction of approximate and optimal filtration*. Preprint Inst. of Appl. Math. Acad. of Scien. of USSR. Issue 4. Moscow, 1978.

[21] **Akhiezer, N. I.** *The classical problem of moments*. Moscow: Fizmatgiz, 1961.

[22] **Akhiezer, N. I. and Glazman, I. M.** *The theory of linear operators in Hilbert spaces.* Moscow: Nauka, 1966.

[23] **Badeva, V. V.,** Numerical algorithms for designing the Ω-optimal linear systems of automatic control // *Numerical Analysis in FORTRAN. Methods and Algorithms.* 66–90. Moscow: Moscow University Press, 1976.

[24] **Badeva, V. V. and Morozov, V. A.,** Algorithms for fast and accelerated solving of some special systems of linear algebraic equations // *Numerical Analysis in FORTRAN.* Issue 20. 80–88. Moscow: Moscow University Press, 1977.

[25] **Badeva, V. V. and Morozov, V. A.,** Theory and methods for designing the Ω-optimal linear systems of automatic control // *Numerical Analysis in FORTRAN.* Issue 21. 38–52. Moscow: Moscow University Press, 1978.

[26] **Badeva, V. V., Morozov, V. A., and Khadzhijsky M.,** Stable designing of pulse transition functions on the basis of regularization method // *Math. and Math. Educ., Proceedings of the 8th Conf. of Bulgarian Math. Union.* 63–82. Sofia: Bulg. Ac. of Sci. Press, 1979.

[27] **Bakushinsky, A.B.,** *Selected questions of approximate solution of ill-posed problems.* Moscow: Moscow University Press, 1968.

[28] **Bakushinsky, A.B.,** A regularizing algorithm based on the Newton–Kantorovich method for solving variational inequalities // *J. Comp. Math. and Math. Phys.* 16 (6). 1397–1404. 1976.

[29] **Bakushinsky, A.B.,** On the principle of iterative regularization // *J. Comp. Math. and Math. Phys.* 19 (4). 1040–1043, 1979.

[30] **Balakirev, V. S., Volodin, V. M., and Tsirlin, A. M.,** *Optimal control of chemical technology processes.* Moscow: Chemistry, 1978.

[31] **Barbashin, E. A.,** To the theory of generalized dynamic systems // *Sci. Notes Moscow Univ. Math.* 2 (135). 110–134, 1949.

[32] **Bakhvalov, N. S.,** On optimality of linear methods for approximation of operators on convex classes of functions // *J. Comp. Math. and Math. Phys.* 11 (4). 1014–1018, 1971.

[33] **Belotserkovskiy, S. M.,** *A thin bearing surface in subsonic gaseous flow.* Moscow: Nauka, 1965.

[34] **Belotserkovskiy, S. M., Lifanov, I. K., and Soldatov, M. M.,** The method of discrete singularities in the plane problems of elasticity theory // *Appl. Math. and Mech.* 47 (5). 781–789, 1983.

[35] **Belotserkovskiy, S. M. and Lifanov, I. K.,** *Numerical methods in singular integral equations.* Moscow: Nauka, 1985.

[36] **Belotserkovskiy, S. M. and Skripach, B. K.,** *Aerodynamic derivatives of flying vehicles at subsonic speed.* Moscow: Nauka, 1975.

[37] Subroutine Library in FORTRAN. Vol. 1, 2, 3 // *Numerical Analysis in FORTRAN.* Moscow: Moscow University Press, 1979.

[38] Subroutine Library in FORTRAN, Vol. 1–4 // *Numerical Analysis in FORTRAN.* Moscow: Moscow University Press, 1982.

[39] **Borisova, P. P., Zaikin, P. N., and Maslennikov, V. A.,** On numerical solution of linear problems of automatic control // *Computing Methods and Programming.* Issue 12. Moscow: Moscow University Press, 1969.

[40] **Bochner, S.** *Lectures on Fourier integrals.* Moscow: Fizmatgiz, 1961.

[41] **Budak, B. M. and Vasiljev, F. P.,** *Approximate methods for solving the optimal control problems (texts of lectures).* Issue 2. Moscow: Moscow University Press, 1968.

[42] **Vaynikko, G. M.,** *Methods for solving linear ill-posed problems in Hilbert spaces.* Tartu: Tartu State University Press, 1982.

[43] **Vajnikko, G. M. and Veretennikov, A. Ju.,** *Iterative procedures in ill-posed problems.* Moscow: Nauka, 1986.

[44] **Vasilenko, V.A.,** *Spline-functions: theory, algorithms, subroutines.* Novosibirsk: Nauka, 1983.

[45] **Vasilenko, G. I.,** *The theory of signal restoration.* Moscow: Sov. Radio, 1979.

[46] **Vasiljev, F. P.**, On regularization of ill-posed extremal problems // *Docl. Acad. Nauk USSR.* 241 (5). 1001–1004. 1978.

[47] **Vasiljev, F. P.**, On regularization of ill-posed minimization problems based on approximately-given sets // *J. Comp. Math. and Math. Phys.* 20 (1). 38–50. 1980.

[48] **Vasiljev, F. P.**, *Numerical methods for solving extremal problems.* Moscow: Nauka, 1981.

[49] **Vershinin, V. V., Zavjalov, Ju. S., and Pavlov, N. N.,** *Extremal properties of splines and the problem of smoothing.* Novosibirsk: Nauka, 1988.

[50] **Vinokurov, V. A.,** On a necessary condition of regularizability according to Tikhonov // *Docl. Acad. Nauk USSR.* 195 (3). 530–531. 1970.

[51] **Vinokurov, V. A.,** On errors in solutions to linear operator equations // *J. Comp. Math. and Math. Phys.* 10 (4). 830–839. 1970.

[52] **Vladimirov, V. S.,** *Equations of mathematical physics.* Moscow: Nauka, 1986.

[53] **Voevodin, V. V.,** *Numerical methods in linear algebra.* Moscow: Nauka, 1966.

[54] **Voevodin, V. V.,** On the regularization method // *J. Comp. Math. and Math. Phys.* 9 (3). 671–673. 1969.

[55] **Voevodin, V. V.,** *Computational fundamentals of linear algebra.* Moscow: Nauka, 1977.

[56] **Voevodin, V. V. and Arushanian, O. B.,** Structure and organization of Numerical Analysis Library of Moscow State University // *Numerical Analysis in FORTRAN. Computational Methods and Instrumental Systems.* 73–83. Moscow: Moscow University Press, 1979.

[57] **Voevodin, V. V. and Kuznetsov, Ju. A.,** *Matrices and computation.* Moscow: Nauka, 1984.

[58] **Voevodin, V. V. and Tyrtyshnikov, E. E.,** *Computational processes with Toeplitz matrices.* Moscow: Nauka, 1987.

[59] **Voevodina, S. N.,** Solving systems of equations with block-Toeplitz matrices // *Computing Methods and Programming.* Issue 24. 94–100. Moscow: Moscow University Press, 1975.

[60] **Voronin, V. V. and Tsetsokho, V. A.,** Numerical solution of integral equations of the first kind with logarithmic singularity // *Mathematical Problems in Geophysics.* Issue 4. 212–227. Novosibirsk: Comp. Center. of Sib. Branch Ac. of Sci. USSR, 1973.

[61] **Voskobojnikov, Ju. E., Preobrazhensky, N. G., and Sedelnikov, A. I.,** *Mathematical data processing in molecular gas dynamics.* Novosibirsk: Nauka, 1984.

[62] *USSR Conference on Ill-Posed Problems. Theses of Reports.* Frunze: Ilim Press, 1979.

[63] **Gavurin, M. K.,** *Lectures on computational methods.* Moscow: Nauka, 1971.

[64] **Gantmakher, F.P.,** *The theory of matrices.* Moscow: Nauka, 1967.

[65] **Gel'fand, I. M., Rajkov, D. A., and Shilov, G. E.,** *Commutative normalized rings.* Moscow: Fizmatgiz, 1960.

[66] **Gilyazov, S. F.,** On estimates of convergence rate of iterative methods for solving linear operator equations // *Numerical Analysis in FORTRAN.* Issue 14. 97–101. Moscow: Moscow University Press, 1976.

[67] **Gilyazov, S. F.,** On iterative methods for solving inconsistent operator equations // *Moscow Univ. Bul. Comput. Math. Cybern.* 3. 77–81, 1977.

[68] **Gilyazov, S. F.,** *Methods for solving linear ill-posed problems.* Moscow: Moscow University Press, 1987.

[69] **Gold, B. and Reyder, Ch.,** *Digital processing of signals.* Moscow: Sov. Radio, 1973.

[70] **Gol'dman, N. L.,** Approximate solution of Fredholm integral equations of the first kind in the class of piecewise convex functions with the bounded first derivative // *Numerical Analysis in FORTRAN. Methods and Algorithms.* 30–40. Moscow: Moscow University Press, 1978.

[71] **Gol'dman, N. L., Morozov, V. A., and Samarin, M. K.,** The method of descriptive regularization and the quality of approximate solutions // *Eng.-Phys. J.* 23 (6). 1117–1124. 1977.

[72] **Goncharskiy, A.V, Gushchina, L. G., Leonov, A. S., and Yagola, A. G.,** *On some algorithms of searching for approximate solutions of ill-posed problems on the set of monotone functions.* Preprint Comp. Center of MSU. Moscow, 1971.

[73] **Goncharskiy, A.V, Gushchin, G. V., and Yagola, A. G.,** On a numerical experiment of solving two-dimensional Fredholm integral equations of the first kind // *Some Questions of Automatized Processing*

*and Interpretation of Physical Experiments.* Issue 1. 192–201. Moscow: Moscow University Press, 1973.

[74] **Goncharskiy, A.V, Leonov, A. S., and Yagola, A. G.,** Methods for solving the convolution-type Fredholm integral equations of the first kind // *Some Questions of Automatized Processing and Interpretation of Physical Experiments.* Issue 1. 170–191. Moscow: Moscow University Press, 1973.

[75] **Goncharskiy, A.V. and Stepanov, V. V.,** An algorithm for approximate solving ill-posed problems on some compact sets //*Docl. Acad. Nauk USSR.* 245 (6). 1296–1299. 1979.

[76] **Goncharskiy, A.V., Cherepashchuk, A. M., and Yagola, A. G.,** *Numerical methods for solving inverse problems of astrophysics.* Nauka: Moscow, 1978.

[77] **Goncharskiy, A.V. and Yagola, A. G.,** On uniform approximation of monotone solutions to ill-posed problems // *Docl. Acad. Nauk USSR.* 184 (4). 771–773, 1969.

[78] **Goncharskiy, A.V. and Yagola, A. G.,** On regularization of ill-posed problems with approximately-given operators // *J. Comp. Math. and Math. Phys.* 14 (4). 1022–1027. 1974.

[79] **Goncharskiy, A.V., Yagola, A. G., and Leonov, A. S.,** On solving the two-dimensional Fredholm integral equations of the first kind with kernels depending on the difference of arguments // *J. Comp. Math. and Math. Phys.* 11 (5). 1296–1301. 1971.

[80] **Gordonova, V. I. and Morozov, V. A.,** Numerical algorithms for the choice of regularization parameter // *J. Comp. Math. and Math. Phys.* 13 (3). 539–545. 1973.

[81] **Gradshtein, N. S. and Ryzhik, I. M.,** *Tables of integrals, sums, series, and products.* Moscow: Fizmatgiz, 1963.

[82] **Grebennikov, A. I.,** On the choice of nodes for approximation of functions by splines // *J. Comp. Math. and Math. Phys.* 16 (1). 219–223. 1976.

[83] **Grebennikov, A. I.,** *The spline method in numerical analysis.* Moscow: Moscow University Press, 1979.

[84] **Grebennikov, A. I.,** *The spline method and solving ill-posed problems in the theory of approximations.* Moscow: Moscow University Press, 1983.

[85] **Grebennikov, A. I. and Kaze V. O.-L.,** Fast algorithms for approximation of functions by splines and solving some applied problems // *Methods and Algorithms in Numerical Analysis.* 154–162. Moscow: Moscow University Press, 1984.

[86] **Grebennikov, A. I.,** *Algorithms and subroutines for approximation of functions of one and several variables by splines and their applications.* Moscow: Moscow University Press, 1987.

[87] **Grebennikov, A. I. and Morozov, V.A.,** An optimal approximation of operators // *Docl. Acad. Nauk USSR.* 223 (6). 1307–1310, 1975.

[88] **De Boor, C.,** *A practical guide to splines.* Moscow: Radio and Svjaz, 1985.

[89] **Demjanov, V. F. and Malozemov, V. N.,** *Introduction to minimax.* Moscow: Nauka, 1972.

[90] **Denisov, D. V.,** On a random search method in the problems of conditional minimization // *J. Comp. Math. and Math. Phys.* 18 (5). 1315–1317. 1978.

[91] **Denisov, D. V. and Tretjakov, A. A.,** *The questions of optimization for control.* Moscow: Moscow University Press, 1979.

[92] **Dmitriev, V. I. and Zakharov, E. V.,** On numerical solution of some Fredholm integral equations of the first kind // *Computing Methods and Programming.* Issue 10. 49–54. Moscow: Moscow University Press, 1968.

[93] **Dzhumaev, S.,** An analogue of the Lax theorem for nonuniquely solvable problems // *Docl. Tajik Acad. Nauk.* 20 (7). 3–5. 1977.

[94] **Dzhumaev, S.,** A continuous dependence of pseudoinverse operators on small perturbations // *Docl. Tajik Acad. Nauk.* 21 (9). 3–5. 1978.

[95] **Drobyshevich, V. I. and Tambovtsev, B. Z.,** *Application of the methods of quadratic programming to solving ill-posed problems.* Preprint of Comp. Center Sib. Branch Akad. Nauk USSR. Novosibirsk, 1975.

[96] **Emelin, I. V. and Krasnoselsky, M. A.,** To the theory of ill-posed problems // *Docl. Acad. Nauk USSR.* 244 (11). 805–808. 1979.

[97] **Eremin, I. I., Mazurov, V. D., and Astafjev, N. N.,** *Improper problems in linear convex programming.* Moscow: Nauka, 1983.

[98] **Evgrafov, M. A.,** *Asymptotic estimates and integer functions.* Moscow: Fizmatgiz, 1962.

[99] **Zhukovsky, E. L.,** Statistical regularization of systems of linear algebraic equation // *J. Comp. Math. and Math. Phys.* 12 (1). 185–191, 1972.

[100] **Zhukovsky, E. L. and Morozov, V. A.,** On sequential Bayesian regularization for systems of linear algebraic equations // *J. Comp. Math. and Math. Phys.* 12 (2). 464–465. 1972.

[101] **Zabrejko, P. P., et al.** *Integral equations.* Moscow: Nauka, 1968.

[102] **Zavjalov, Ju. S., Kvasov, B. I., and Miroshnichenko, V. L.,** *Methods of spline functions.* Moscow: Nauka, 1980.

[103] **Zavjalov, Ju. S., Leus, V. A., and Skorospelov, V. A.,** *Splines in engineering geometry.* Moscow: Mashinostroenie, 1985.

[104] **Zaikin P. N. and Ufimtsev, M. V.,** *Methodical and information support for automatized computer processing of experimental data.* Moscow: Moscow University Press, 1987.

[105] **Zakharov, E. V. and Pimenov, Ju. V.,** *Numerical analysis of radiowave diffraction.* Moscow: Radio and Svjaz, 1982.

[106] **Zojtendeik, G.,** *Methods of feasible directions.* Moscow: IL-press, 1963.

[107] **Ivanitskij, A. Ju.,** On solving the systems of linear algebraic equations with inexact data // *Numerical Analysis: Methods, Algorithms, Programs.* 136–141. Moscow: Moscow University Press, 1983.

[108] **Ivanov, V. V.,** On optimal-in-accuracy algorithms for approximate solving the first-kind operator equations // *J. Comp. Math. and Math. Phys.* 15 (1). 3–11. 1975.

[109] **Ivanov, V. V., Karagodova, E. A., and Jadrenko, Ja. K.,** Approximate methods for determining the dynamic characteristics of objects under control // *J. Comp. Math. and Math. Phys.* 1 (3). 523–541. 1961.

[110] **Ivanov, V. K.,** On linear ill-posed problems // *Docl. Acad. Nauk USSR.* 145 (2). 270–272. 1962.

[111] **Ivanov, V. K.,** On ill-posed problems // *Matem. Sbornik.* 61 (2). 211–233. 1963.

[112] **Ivanov, V. K.,** On one type of ill-posed linear equations in vector topological spaces // *Sib. Math. J.* 6 (4). 832–839. 1965.

[113] **Ivanov, V. K.,** On approximate solution of operator equations of the first kind // *J. Comp. Math. and Math. Phys.* 6 (6). 1089–1093. 1966.

[114] **Ivanov, V. K.,** An estimate of errors when solving the operator equations of the first kind // *Questions of Accuracy and Efficiency of Computing Algorithms.* Issue 2. 102–116. Kiev: Ac. of Sci. of Ukr. Press, 1963.

[115] **Ivanov, V. K., Vasin, V. V., and Tanana, V. P.,** *The theory of linear ill-posed problems and its applications.* Moscow: Nauka, 1978.

[116] **Ivanov, V. K. and Korolyuk, T. I.,** An estimate of errors when solving linear ill-posed problems // *J. Comp. Math. and Math. Phys.* 9 (1). 30–41. 1969.

[117] **Ivanov, V. V. and Simonenko, I. B.,** On approximate searching for all solutions of a given linear equation in a Banach space // *Docl. Acad. Nauk USSR.* 126 (6). 18–21. 1959.

[118] **Iljin, V. A. and Poznjak, E. G.,** *Foundations of mathematical analysis.* Moscow: Nauka, 1971.

[119] **Iljinskij, A. S., Repin, V. M., and Starostin, L. I.,** On numerical solution of the two-dimensional problem on propagation of E-polarization waves in a system of domains connected through holes // *Computing Methods and Programming.* Issue 20. Moscow: Moscow University Press, 1973.

[120] **Kantorovich, L. V.,** On the method of steepest descent // *Docl. Acad. Nauk USSR.* 56 (3). 1947.

[121] **Kantorovich, L. V.,** An efficient method of solving extremal problems for quadratic functionals // *Docl. Acad. Nauk USSR.* 48 (7). 455–460. 1945.

[122] **Karasik, B. G.,** Regularization of the ill-posed Cauchy problem for differential-operator equations of arbitrary order // *Notes of Azerb. Acad. of Sci. Ser. Phys.-Techn. Math. Sci.* 6. 1976.

[123] **Karasik, B. G.,** *The conditionally correct Cauchy problem for differential-operator equations and its application.* Ph. D. Thesis. IMM of AzSSR Academy Nauk. Baku, 1979.

[124] **Karlin, S. and Stadden, V.,** *Chebyshev systems and their application in analysis and statistics.* Moscow: Nauka, 1976.

[125] **Karmazin, V. N.,** On generalized residual principle for solving the systems of linear inequalities and the problems of linear programming // *Optimization and Control.* 76–84. Moscow: Moscow University Press, 1983.

[126] **Karmanov, V. G.,** *Mathematical programming.* Moscow: Nauka, 1975.

[127] **Kafarov, V. Ja. and Dorokhov, I. N.,** *System analysis of chemical technology processes.* Moscow: Nauka, 1976.

[128] **Kashin, G. M., Pshenichnov, G. I., and Fljorov, Ju. A.** *Methods of automatized design of aeroplanes.* Moscow: Mashinostroenie, 1979.

[129] **Kornejchuk, N. P.,** *Splines in the approximation theory.* Moscow: Nauka, 1984.

[130] **Kochetov, I. I.,** A new method for the choice of regularization parameter // *J. Comp. Math. and Math. Phys.* 16 (2). 499–503. 1976.

[131] **Krein, S. G.,** *Linear equations in Banach spaces.* Moscow: Nauka, 1971.

[132] **Lavrentjev, M. M.,** *On some ill-posed problems of mathematical physics.* Novosibirsk: Siberian Branch of USSR AS, 1962.

[133] **Lavrentjev, M. M.,** *Conditionally-correct problems for differential equations.* Novosibirsk: Novosibirsk State University Press, 1973.

[134] **Lavrentjev, M. M. and Vasiliev, V.G.,** On formulation of some ill-posed problems of mathematical physics // *Sib. Math. J.* 7 (3). 559–576, 1966.

[135] **Lanczos, K.,** *Practical methods of applied analysis.* Nauka: Moscow, 1968.

[136] **Lattes, P. and Lions, Zh. L.,** *Methods of quasiinversion and its applications*. Moscow: Mir, 1970.

[137] **Leonov, A. S.,** On algorithms of approximate solution of nonlinear ill-posed problems with perturbed operators // *Docl. Acad. Nauk USSR*. 245 (2). 300–304. 1979.

[138] **Lions, Zh. L.,** *Optimal control of the systems described by partial differential equations*. Moscow: Mir, 1972.

[139] **Liskovets, O. A.,** Theory and methods for solving ill-posed problems // *Results of Sci. Techn. Ser. Math. Anal.* 116–172. Moscow: VINITI, 1982.

[140] **Lifanov, I. K.,** On ill-posedness and regularization of numerical solutions to singular integral equations // *Docl. Acad. Nauk USSR.* 255 (5). 1046–1050. 1980.

[141] **Likht, M. K.,** On calculation of functionals on solutions of first-kind linear equations // *J. Comp. Math. and Math. Phys.* 7 (3). 667–672. 1967.

[142] **Lyusternik, L. A. and Sobolev, V. I.,** *Elements of functional analysis.* Moscow: Nauka, 1965.

[143] **Marchuk, A. G.,** *The optimal-in-accuracy methods for solving recovery problems*. Preprint of Comp. Center of Siberian Branch. USSR AS. Novosibirsk, 1976.

[144] **Marchuk, A. G. and Osipenko, K. Ju.** The best approximation of functions given with errors at a finite number of points // *Math. Notes.* 17 (3). 359–368. 1975.

[145] **Marchuk, G. I. and Vasilyev, V. G.,** On approximate solution of first-kind operator equations // *Docl. Acad. Nauk USSR*. 195 (4). 773–775. 1970.

[146] **Marchuk, G. I. and Kuznetsov, Yu. I.,** *Iterative methods and quadratic functionals*. Novosibirsk: Nauka, 1972.

[147] **Meleshko, V. I.,** Stable pseudoinversion of closed operators // *J. Comp. Math. and Math. Phys.* 17 (5). 1132–1143. 1977.

[148] *Methods for solving conditionally-correct problems.* Proc. of Inst. Math. Mech. USC USSR AS. Issue 17. Sverdlovsk, 1975.

[149] *Methods of regularization for unstable problems.* Proc. of Inst. Math. Mech. USC USSR AS. Issue 23. Sverdlovsk, 1976.

[150] *Methods for solving ill-posed problems and their applications.* Proc. of USSR School of Young Scientists. Moscow: Moscow University Press, 1974.

[151] **Molchanov, I. N. and Yakovlev, M. F.,** On a class of iterative processes for solving inconsistent systems of linear algebraic equations // *Docl. Acad. Nauk USSR.* 209 (4). 782–784. 1973.

[152] **Morozov, V. A.,** On regularization of ill-posed problems and the choice of regularization parameter // *J. Comp. Math. and Math. Phys.* 6 (1). 170–175. 1966.

[153] **Morozov, V. A.,** On solution of functional equations by the regularization method // *Docl. Acad. Nauk USSR.* 167 (3). 510–512. 1966.

[154] **Morozov, V. A.,** On the choice of regularization parameter when solving the functional equations by the regularization method // *Docl. Acad. Nauk USSR.* 175 (6). 1225–1228. 1967.

[155] **Morozov, V. A.,** On regularizing families of operators // *Computing Methods and Programming.* Issue 8. 63–93. Moscow: Moscow University Press, 1967.

[156] **Morozov, V. A.,** *Methods for solving unstable problems.* Moscow: Moscow University Press, 1967.

[157] **Morozov, V. A.,** On the residual principle for solution of operator equations by the regularization method // *J. Comp. Math. and Math. Phys.* 8 (2). 295–309. 1968.

[158] **Morozov, V. A.,** On pseudosolutions // *J. Comp. Math. and Math. Phys.* 9 (6). 1381–1392. 1969.

[159] **Morozov, V. A.,** On regularization of some classes of extremal problems // *Computing Methods and Programming.* Issue 12. 24–37. Moscow: Moscow University Press, 1969.

[160] **Morozov, V. A.,** On problem of differentiation and some algorithms for approximation of experimental data // *Computing Methods and Programming.* Issue 14. 46–63. Moscow: Moscow University Press, 1970.

[161] **Morozov, V. A.,** On stability of the problem on specification of parameters // *Computing Methods and Programming*. Issue 14. 63–67. Moscow: Moscow University Press, 1970.

[162] **Morozov, V. A.,** On optimal regularization of operator equations // *J. Comp. Math. and Math. Phys.* 10 (4). 818–829. 1970.

[163] **Morozov, V. A.,** On approximate solution of operator equations by the spline method // *Docl. Acad. Nauk USSR*. 200 (1). 35–38. 1971.

[164] **Morozov, V. A.,** An efficient numerical algorithm for constructing pseudosolutions // *J. Comp. Math. and Math. Phys.* 11 (1). 260–262, 1971.

[165] **Morozov, V. A.,** The theory of splines and the problem of stable calculation of unbounded operators // *J. Comp. Math. and Math. Phys.* 11 (3). 545–558, 1971.

[166] **Morozov, V. A.,** On specification of parameters for a linear model according to experimental data // *Proc. USSR Symp. Inverse Problems for Differential Equations. Siberian Branch of USSR AS*. Novosibirsk, 1972.

[167] **Morozov, V. A.,** Linear and nonlinear ill-posed problems // *Itogi Nauki i Tekhniki. Ser. Mathematical Analysis*. Issue 11, Moscow: VINITI, 1973.

[168] **Morozov, V. A.,** On some general conditions of regularizability for ill-posed variational problems // *Proc. First Conf. Young Scientists of Computing Mathematics and Cybernetics Department*. 140–164. Moscow: Moscow University Press, 1973.

[169] **Morozov, V. A.,** On the residual principle for solution of inconsistent equations by Tikhonov's regularization method // *J. Comp. Math. and Math. Phys.* 13 (5). 1009–1111. 1973.

[170] **Morozov, V. A.,** On computing the lower bounds of functionals according to approximate input data // *J. Comp. Math. and Math. Phys.* 13 (4). 1045–1049. 1973.

[171] **Morozov, V. A.,** Convergence of an approximate method for solving the first-kind operator equations // *J. Comp. Math. and Math. Phys.* 13 (1). 3–17. 1973.

[172] **Morozov, V. A.,** *Regular methods for solving ill-posed problems.* Moscow: Moscow University Press, 1974.

[173] **Morozov, V. A.,** On a class of inverse problems and the methods of their solving // *Eng.-Phys. J.* XXIX (1). 165–169. 1975.

[174] **Morozov, V. A.,** On some applications of the spline method to solving operator equations of the first kind // *Docl. Acad. Nauk USSR.* 229 (2). 300–303. 1976.

[175] **Morozov, V. A.,** The $L$-pseudoinversion and its properties // *Docl. Acad. Nauk USSR.* 233 (2). 291–294, 1977.

[176] **Morozov, V. A.,** Seminormalized factor-spaces and the theory of splines // *Variational-Differential Methods.* 44–48. Novosibirsk: Comp. Center of Siberian Branch of USSR AS, 1977.

[177] **Morozov, V. A.,** Estimation of accuracy for solutions to ill-posed problems and the solution of systems of linear algebraic equations // *J. Comp. Math. and Math. Phys.* 17 (6). 1341–1349. 1977.

[178] **Morozov, V. A.,** Ill-posed problems in the theory of automatic control // *Proc. of the First School-Conf. On Control of Complicated Chem.-Techn. Systems.* 333–364. Varna, 1978.

[179] **Morozov, V. A.,** *Regular methods for solving linear and nonlinear ill-posed problems.* Doct. Thesis. Moscow, 1978.

[180] **Morozov, V. A.,** Regular methods for solving nonlinear operator equations // *Izv. Vuzov. Ser. Mat.* 11 (198). 74–86. 1978.

[181] **Morozov, V. A.,** On optimal approximation of linear operators // *Numerical Analysis in FORTRAN. Methods and Algorithms.* 25–29. Moscow: Moscow University Press, 1978.

[182] **Morozov, V. A.,** On solution of operator equations of the first kind by the method of finite-rank approximations // *Docl. Acad. Nauk USSR.* 247 (6). 1317–1320. 1979.

[183] **Morozov, V. A.,** Some peculiarities of numerical solution of integral equations by the method of descriptive regularization // *Numerical Analysis in FORTRAN. Computational Methods and Instrumental Systems.* 41–49. Moscow: Moscow University Press, 1979.

[184] **Morozov, V. A.,** Ill-posed problems in the theory of automatic control // *Numerical Analysis in FORTRAN. Methods and Algorithms.* 3–45. Moscow: Moscow University Press, 1979.

[185] **Morozov, V. A.,** On restoration of images with guaranteed accuracy // *Numerical Analysis in FORTRAN. Methods and Algorithms.* 46–65. Moscow: Moscow University Press, 1979.

[186] **Morozov, V. A.,** On stable numerical methods for solving systems of algebraic equations // *Methods and Algorithms.* 3–10. Moscow: Moscow University Press, 1982.

[187] **Morozov, V. A.,** On stable numerical methods for solving the consistent algebraic equations // *J. Comp. Math. and Math. Phys.* 24 (2). 179–186. 1984.

[188] **Morozov, V. A.,** *Regular methods for solving ill-posed problems.* Moscow: Nauka, 1987.

[189] **Morozov, V. A.,** *The regularization methods for unstable problems.* Moscow: Moscow University Press, 1987.

[190] **Morozov, V. A. and Gilyazov, S. F.,** On numerical solution of some types of inconsistent equations // *Numerical Analysis in FORTRAN. Computing Methods and Instrumental Systems.* 27–37. Moscow: Moscow University Press, 1979.

[191] **Morozov, V. A. and Gol'dman, N. L.**, On algorithms of descriptive regularization for solving Fredholm integral equations of the first kind // *Numerical Analysis in FORTRAN.* 52–72. Issue 15. Moscow: Moscow University Press, 1976.

[192] **Morozov, V. A. and Gol'dman, N. L.**, A numerical algorithm for solving Fredholm integral equations of the first kind by the pointwise residual method // *Methods and Algorithms in Numerical Analysis.* 28–49. Moscow: Moscow University Press, 1981.

[193] **Morozov, V. A., Gol'dman, N. L., and Samarin, M. K.,** The descriptive regularization method and the quality of approximate solutions // *J. Eng. Phys.* 33 (6). 1117–1124. 1977.

[194] **Morozov, V. A. and Gordonova, V. I.,** Numerical algorithms for the choice of parameter in the regularization method // *J. Comp. Math. and Math. Phys.* 13 (3). 539–545. 1973.

[195] **Morozov, V. A. and Grebennikov, A. I.** On an optimal approximation of operators // *Docl. Acad. Nauk USSR.* 233 (6). 1307–1310. 1977.

[196] **Morozov, V. A., Kirsanova, N. N., and Sysoev, A. F.,** A collection of algorithms for fast Fourier transform of discrete series // *Numerical Analysis in FORTRAN.* Issue 15. 30–51. Moscow: Moscow University Press, 1976.

[197] **Morozov, V. A. and Malyshev, V. A.** *Differential inequalities and approximations.* Moscow: Moscow University Press, 1989.

[198] **Morozov, V. A., Medvedev, N. V., and Ivanitskiy, A. Yu.,** *Regularization of problems in algebra and analysis.* Moscow: Moscow University Press, 1987.

[199] **Morozov, V. A. and Mchedlishvili, N.I.,** On approximate solution of discrete Wiener–Hopf equations in $O(N \log N)$ operations // *Numerical Analysis: Methods, Algorithms, Programs.* 30–40. Moscow: Moscow University Press, 1983.

[200] **Morozov, V. A. and Nazimov, A. B.,** To the problem of regularization for systems of linear algebraic equations // *Numerical Analysis: Methods, Algorithms, Programs.* 14–18. Moscow: Moscow University Press, 1985.

[201] **Morozov, V. A. and Pospelov, V. V.**, *Digital processing of signals.* Moscow: Moscow University Press, 1986.

[202] **Morozov, V. A. and Samarin, M. K.,** Finding the semibounded monotone solutions to operator equations // *Methods and Algorithms in Numerical Analysis.* 29–33. Moscow: Moscow University Press, 1982.

[203] **Muravjeva, M. V.,** On optimality and limiting properties of the Bayesian solution to a system of linear algebraic equations // *J. Comp. Math. and Math. Phys.* 13 (4). 819–828. 1973.

[204] **Muravjeva, M. V.,** On searching for the pulse transition function by the regularization method // *J. Comp. Math. and Math. Phys.* 13 (1). 194–199. 1973.

[205] **Muskhelishvili, N. I.**, *Singular integral equations.* Moscow: Nauka, 1968.

[206] **Mchedlishvili, N. I.,** A numerical algorithm for solving Wiener–Hopf integral equations of the first kind by the regularization method // *Numerical Analysis in FORTRAN. Methods and Algorithms.* 23–32. Moscow: Moscow University Press, 1980.

[207] **Natanson, N. I.,** *The theory of functions of real variables.* Moscow: Nauka, 1974.

[208] **Nedyalkov, I. P.,** *On some ill-posed problems in the theory of potential.* Sofia: BAN Press, 1978.

[209] *Ill-posed inverse problems of atomic physics.* Proc. of Inst. of Theor. and Appl. Mech. of Siberian Branch of USSR AS. Novosibirsk, 1976.

[210] **Nikolaev, E. S.,** Nonlinear acceleration of two-level iteration methods // *J. Comp. Math. and Math. Phys.* 16 (6). 1381–1387. 1976.

[211] **Nikolskiy, S. M.,** *Quadrature formulas.* Moscow: Nauka, 1979.

[212] **Nikolskiy, S. M.,** Linear equations in linear normed spaces // *Izv. Acad. Nauk USSR. Ser. Math.* 7 (3). 147–163. 1943.

[213] *Image processing and digital filtration* / Huang, T. M., Ed. Moscow: Mir, 1979.

[214] *Image processing with computers* / Endruce, G. and Inlo, L., Eds. Moscow: Mir, 1973.

[215] *Optimization of numerical methods.* Proc. of Inst. of Cibern. Ukr. AS. Kiev, 1974.

[216] On the third edition of numerical analysis library of SRCC MSU (information) // *Numerical Analysis in FORTRAN. Methods and Algorithms.* 152–153. Moscow: Moscow University Press, 1979.

[217] **Panasuk, V. V., Savruk, M. P., and Datsyshin, A. P.**, *Distribution of stresses near cracks in plates and envelopes.* Kiev: Naukova Dumka, 1976.

[218] **Parton, V. Z. and Permin, P. I.**, *Integral equations of elasticity theory.* Moscow: Nauka, 1977.

[219] **Petkov, T.,** *Identification on the basis of automatization.* Sofia, 1972.

[220] **Petrov, A. P.,** Estimates of linear functionals for solving ill-posed inverse problems // *J. Comp. Math. and Math. Phys.* 7 (3). 648–654. 1967.

[221] **Polia, G. and Cege, G.**, *Problems and theorems of analysis.* Moscow: Nauka, 1977.

[222] **Polyak, B. T.,** Methods of minimization with constraints. *Itogi Nauky i Tekhniki. Mat. Anal.* Issue 12. Moscow: VINITI, 1975.

[223] **Pospelov, V. V.,** On errors of approximation of functions of two variables by the sums of products of functions of one variable // *J. Comp. Math. and Math. Phys.* 18 (5). 1307–1308. 1978.

[224] **Presdorf, Z.,** *Some classes of singular equations.* Moscow: Mir, 1979.

[225] **Pratt, W.,** *Digital processing of images.* Moscow: Mir, 1982.

[226] **Pshenichny, B. N.,** Dual methods in extremal problems // *Cybernetics.* 3. 89–95. 1965; 4. 64–69. 1965.

[227] **Pytyev, Yu. P.,** Morphological concepts in problems of image analysis // *Docl. Acad. Nauk USSR.* 244 (6). 1283–1286. 1975.

[228] **Pytyev, Yu. P.,** Problems of improving the quality of images // *Docl. Acad. Nauk USSR.* 245 (2). 315–319. 1979.

[229] **Pytyev, Yu. P.,** Problems of image restoration // *Docl. Acad. Nauk USSR.* 245 (1). 42–46. 1979.

[230] **Pytyev, Yu. P.,** Suppression of false signals in the problem of resolution enhancement // *Docl. Acad. Nauk USSR.* 225 (3). 540–544. 1980.

[231] **Reider, Ch.,** Discrete Fourier transform for the case when the number of samples is prime // *Proc. of Inst. of Electrotechnical Engineering.* 56 (7). 1968.

[232] **Rosenbrock, H. and Storn, S.,** *Numerical methods for chemical engineers.* Moscow: Mir, 1968.

[233] **Rokafellar, R.,** *Convex analysis.* Moscow: Mir, 1973.

[234] **Rotach, V. Ya. and Hadjisky, M. B.,** On synthesis of a system of automatic control for industrial processes // *Automatics and Telemechanics.* 9. 1966.

[235] **Rudin, U.,** *Functional analysis.* Moscow: Mir, 1979.

[236] **Savelova, T. I. and Tikhomirov, V. V.,** On solution of convolution-type integral equations in the multidimensional case // *J. Comp. Math. and Math. Phys.* 13 (3). 555–563. 1973.

[237] **Samarin, M. K.,** Approximation of grid functions by functions with a given number of interior extrema // *Numerical Analysis in FORTRAN*. Issue 7. Moscow: Moscow University Press, 1974.

[238] **Samarin, M. K.,** Mean-square approximation by piecewise monotone functions // *Numerical Analysis in FORTRAN*. Issue 15. 83–90. Moscow: Moscow University Press, 1976.

[239] **Samarin, M. K.,** *On stability of solving some ill-posed problems on sets of special structure.* Ph. D. Thesis. Moscow, 1979.

[240] **Samarin, M. K.,** The uniform approximation by monotone functions // *Numerical Analysis in FORTRAN. Methods and Algorithms*. Issue 18. 8–14. Moscow: Moscow University Press, 1980.

[241] **Samarin, M. K.,** On convergence of monotone and convex solutions of integral equations // *Computing Methods and Programming*. Issue 35. 41–51. Moscow: Moscow University Press, 1981.

[242] **Sveshnikov, A. G. and Tikhonov, A. N.,** *The theory of functions of complex variables*. Moscow: Nauka, 1979.

[243] **Sidorov, M. A. and Trenogin, V. A.,** On an approach to the regularization problem on the basis of perturbations of linear operators // *Math. Notes*. 20 (5). 747–752. 1976.

[244] **Simonov, V. I. and Shchedrin, B. M.,** The Fourier integral of a minimization function and signs of structural amplitudes // *Crystallography*. 6 (3). 363–374. 1961.

[245] **Smolyak, S. A.,** *On the optimal recovery of functions and functionals.* Ph. D. Thesis. Moscow, 1965.

[246] **Solodovnikov, V. V. and Uskov, A. S.,** *Statistical analysis of control objects*. Moscow, 1960.

[247] **Stechkin, S. B.,** The best approximations of linear operators // *Math. Notes*. 1 (2). 137–148. 1967.

[248] **Strakhov, V. N.,** On numerical solution of ill-posed problems represented by convolution-type integral equations // *Docl. Acad. Nauk USSR*. 178 (2). 1968.

[249] **Strakhov, V. N.,** On solving linear ill-posed problems in Hilbert spaces // *Diff. Eq.* 6 (8). 1490–1495. 1970.

[250] **Strakhov, V. N.,** On algorithms of approximate solution of linear conditionally ill-posed problems // *Docl. Acad. Nauk USSR* . 207 (5). 1057–1059. 1972.

[251] **Sysoev, A. F.,** A comparative analysis of algorithms and subroutines for fast Fourier transform // *Methods and Algorithms in Numerical Analysis.* 76–83. Moscow: Moscow University Press, 1982.

[252] **Tanana, V. P.,** *Methods for solving operator equations.* Moscow: Nauka, 1981.

[253] **Titchmarsh, E. I.,** *Introduction to the theory of Fourier integrals.* Moscow–Leningrad: Gostekhizdat, 1948.

[254] **Tikhonov, A. N.,** On stability of inverse problems // *Docl. Acad. Nauk USSR.* 39 (5). 195–198. 1943.

[255] **Tikhonov, A. N.,** On solving ill-posed problems and the method of regularization // *Docl. Acad. Nauk USSR.* 151 (3). 501–504. 1963.

[256] **Tikhonov, A. N.,** On nonlinear equations of the first kind // *Docl. Acad. Nauk USSR.* 161 (5). 1023–1026. 1975.

[257] **Tikhonov, A. N.,** On ill-posed problems in linear algebra and a stable method for their solution // *Docl. Acad. Nauk USSR.* 163 (3). 591–594. 1965.

[258] **Tikhonov, A. N.,** On stability of problems on optimization of functionals // *J. Comp. Math. and Math. Phys.* 6 (4). 631–634. 1966.

[259] **Tikhonov, A. N.,** On ill-posed problems // *Computing Methods and Programming.* Issue VIII. 3–33. Moscow: Moscow University Press, 1967.

[260] **Tikhonov, A. N.,** On approximate systems of linear algebraic equations // *J. Comp. Math. and Math. Phys.* 20 (6). 1373–1383. 1980.

[261] **Tikhonov, A. N.,** On a reciprocity principle // *Docl. Acad. Nauk USSR.* 253 (2). 306–308. 1980.

[262] **Tikhonov, A. N.,** On approximate functional equations of the first kind // *Proc. MIAN USSR.* 158. 197–202. 1981.

[263] **Tikhonov, A. N.,** On problems with inexact given data // *Docl. Acad. Nauk USSR.* 280 (3). 559–562. 1985.

[264] **Tikhonov, A. N., et al.,** On the question of quasioptimal choice of the parameter for regularized approximation // *Docl. Acad. Nauk USSR.* 248 (3). 531–535. 1980.

[265] **Tikhonov, A. N. and Arsenin, V. Ya.,** *Methods for solving ill-posed problems.* Moscow: Nauka, 1974.

[266] **Tikhonov, A. N., Arsenin, V. Ya., and Timonov, A. A.,** *Mathematical problems in computer tomography.* Moscow: Nauka, 1987.

[267] **Tikhonov, A. N. and Vasilyev, F. P.,** Methods for solving ill-posed extremal problems // *Banach Center Publication.* 3. 297–342. 1978.

[268] **Tikhonov, A. N., Vasilyev, F. P., Potapov, M. M., et al.,** On regularization of minimization problems on the sets given approximately // *Moscow University Bul. Computing Mathematics and Cybernetics.* Ser. 15 (1). 4–19. 1977.

[269] **Tikhonov, A. N. and Gaysaryan, S. S.,** On the optimal choice of grids for the approximate evaluation of quadratures // *J. Comp. Math. and Math. Phys.* 9 (5). 1170–1176. 1969.

[270] **Tikhonov, A. N. and Glasko, V. B.,** On approximate solution of Fredholm integral equations of the first kind // *J. Comp. Math. and Math. Phys.* 4 (3). 564–571. 1964.

[271] **Tikhonov, A. N. and Glasko, V. B.,** Application of the regularization method in ill-posed problems // *J. Comp. Math. and Math. Phys.* 5 (3). 463–473. 1965.

[272] **Tikhonov, A. N., et al.,** *Regularizing algorithms and a priori information.* Moscow: Nauka, 1988.

[273] **Tikhonov, A. N. and Dmitriev, V. I.,** A method for computing the current distribution in a system of linear vibrators and directional diagrams // *Computing Methods and Programming.* Issue 10. 3–8. Moscow: Moscow University Press, 1968.

[274] **Trenogin, V. A. and Sidorov, N. A.,** On Tikhonov's regularization of the problem on the points of bifurcation for nonlinear operators //*Sib. Matem. J.* 17 (2). 402–413. 1976.

[275] **Trikomy, F.,** *Integral Equations.* Moscow: IL, 1960.

[276] **Tseng, Ya. Yu.,** Generalized inverse unbounded operators in unitary spaces // *Docl. Acad. Nauk USSR.* 67 (3). 431–434. 1949.

[277] **Uzhlio, S.,** *Mathematical statistics.* Moscow: Nauka, 1967.

[278] **Uspensky, A. B. and Fedorov, V. V.,** On formulation and solution of the problem on design of experiments for some inverse problems in mathematical physics // *Approximate Methods for Solving the Problems of Optimal Control and some Ill-Posed Problems.* 82–97. Moscow: Moscow University Press, 1972.

[279] **Fadeeva, V. N.,** Shift for systems with ill-conditioned matrices // *J. Comp. Math. and Math. Phys.* 5 (5). 907–911. 1965.

[280] **Fadeev, D. K. and Fadeeva, V. N.,** *Computational methods in linear algebra.* Moscow: Fizmatgiz, 1963.

[281] **Fedorov, V. V.,** *Numerical methods of maximin.* Moscow: Nauka, 1979.

[282] **Fedotov, A. M.,** An estimate of accuracy of approximate solutions to operator equations of the first kind with random errors in input data // *Docl. Acad. Nauk USSR.* 240 (3). 545–548. 1978.

[283] **Fiacko, A. and McCormic, G.,** *Nonlinear programming.* Moscow: Mir, 1972.

[284] **Fikhtengol'ts, G. M.,** *A course of differential and integral calculus.* Moscow: Fizmatgiz, 1960.

[285] **Fridman, V. M.,** New methods for solving linear operator equations // *Docl. Acad. Nauk USSR.* 128 (3). 482–484. 1959.

[286] **Hadzhiyski, M., Badeva, V. V., and Bozhorov, E. E.,** The 11th National Conference with International Participation "Application of SAC in Engineering and Construction". Primorsko, 1975.

[287] **Khalfin, L. A.,** A statistical approach to solving ill-posed problems // *Studies in the Statistical Theory of Estimation.* 67–81. Leningrad: Nauka, 1978.

[288] **Tsypkin, Ja. Z.,** *Fundamentals of the automatic control systems.* Moscow, 1977.

[289] **Shilov, G. E.,** On locally analytic functions //*Uspekhi Mat. Nauk.* 21 (6). 177–182. 1966.

[290] **Shirikov, V. P.,** *The FORTRAN Language.* Moscow: Moscow University Press, 1970.

[291] **Yagola, A. G.,** On the choice of regularization parameter according to the generalized residual principle // *Docl. Acad. Nauk USSR.* 245 (1). 37–39. 1979.

[292] **Yagola, A. G.,** The generalized residual principle in reflexive spaces // *Docl. Acad. Nauk USSR.* 149 (1). 71–73. 1979.

[293] **Yaroslavsky, L. P.,** *Introduction to digital image processing.* Moscow: Sov. Radio, 1979.

[294] **Anselone, P. M. and Loran, J.,** A general method for the construction of interpolating or smoothing spline functions // *Numer. Math.* 12. 66–82. 1968.

[295] **Bertero, M. and Viano, G. A.,** *On probabilistic methods for the solution of improperly posed problems.* Geneva: Instituto di Scienze Fisiche, 1978.

[296] **Boor C. de.,** On calculation with *B*-splines // *J. Approx. Theory.* 6. 50–62. 1972.

[297] **Boor C. de.,** Computational aspects of optimal recovery // *Opt. Estim. in Approx. Theory.* 69–91. New York: Plenum Press, 1977.

[298] **Carasso, A. S., Sanderson, J. G., and Human, J. M.,** Digital removal of random media image degradations by solving the diffusion equation backwards in time // *SIAM. J. Numer. Anal.* 15 (2). 1978.

[299] **Douglas, J.,** Mathematical programming and integral equations // *Sympos. Num. Treatm. Ord. Diff. Eq. Int.-Diff. Eq.* 169–274. Birkhauser, 1960.

[300] **Elden, L.,** *Numerical methods for the regularization of Fredholm integral equations of the first kind.* Rep. LITH-MAT-R-1974-7. Linkoping, 1974.

[301] **Elden, L.,** *Algorithms for least-squares problems with banded inequality constraints.* Rep. LITH-MAT-R-1977-20. Linkoping, 1977.

[302] **Elden, L.,** Regularization of the backward solution of parabolic problems // *Proc. of the Conf. "Inverse and Improperly Posed Problems in Diff. Eq.".* 73–82. Berlin: Akademie-Verlag, 1979.

[303] **Elden, L.,** *A practical examination of four methods for spaceinvariant image restorations.* Report LITH-MAT-R-1981-24. Linkoping, 1981.

[304] **Franclin, J. N.,** Well-posed stochastic extensions of ill-posed linear problems // *J. Math. Anal. and Appl.* 31 (3). 682–716. 1970.

[305] **Fridrich, V.,** Zur stochastischen Regularisierung nichtkorrekter Gleichungen in Hilbertraumen // *Proc. of the Conf. "Inverse and Improperly Posed Problem in Diff. Eq."*. 89–98. Berlin: Akademie-Verlag, 1979.

[306] **Classman, J. A.,** A generalization of the fast Fourier transform // *IEEE. Trans. on Computers.* C-19 (2). 1970.

[307] **Golomb, M. and Weinberger, H. F.,** Optimal approximation and error bounds // *On Numerical Approx.* 117–190. Madison, 1959.

[308] **Hadamard, J.,** Sur les problemes aux derivees pertielles et leur signification physique // *Bull. Univ. Princeton.* 13. 1902.

[309] **Hadamard, J.,** *Le probleme de Cauchy.* Paris, 1932.

[310] **Hadjyiski, M. and Badeva, V. V.,** Synthesis of complex cascade systems for the automatic control of a sequence of chemical and technological apparatus // *The 6th CHISA Congress.* Prague, 1978.

[311] **Hoffman, A. J.,** On approximate solution of systems of linear inequalities // *J. of Research of Standards.* 49. 263–265. 1959.

[312] **Hoffman, B.,** *Regularization for applied inverse and ill-posed problems.* Leipzig: BSB Tenbner, 1986.

[313] *Inverse and improperly posed problems in differential equations.* Proc. of the Conf. in Halle-Salle GDR. Berlin: Akademie-Verlang, 1979.

[314] **Kammerer, W. J. and Nashed, M. Z.,** Iterative methods for the best approximate solutions of linear integral equations of the first and second kinds // *J. Math. Anal. and Appl.* 40 (3). 547–573. 1972.

[315] **Martin, R. S., Reinsch, C., and Wilkinson, J. H.,** Householder tridiagonalization of a symmetric matrix // *Numer. Math.* 11. 181–195. 1968.

[316] **Melkman, A. A. and Micchelli, C. A.,** Optimal estimation of linear operators in Hilbert spaces from inaccurate data // *SIAM J. Numer Anal.* 16 (1). 87–105. 1979.

[317] **Micchelli, C. A. and Rivlin, T. J.,** A survey of optimal recovery // *Optimal Estimation in Approx. Theory.* 1–53. N.Y.: Plenum Press, 1977.

[318] **Miller, K.,** Stabilizer of quasi-reversibility and other-nearly-best-possible-methods for non-well posed problems // *Symposium on Logarithmic Convexity and Non-Well Posed Problems.* 162–176. Edinburg: 1972.

[319] **Morozov, V. A.,** Regular methods for solving nonlinear problems // *Proc. of the Conf. "Inverse and Improperly Posed Problems in Diff. Eq."* 289–296. Berlin: Akademie-Verlag, 1979.

[320] **Morozov, V. A.,** *Methods for solving incorrectly posed problems.* N.Y.; Berlin; Heidelberg; Tokyo: Springer-Verlag, 1984.

[321] **Nashed, M. Z.,** Generalized inverses. Normal solvability and iterations for singular operator equations // *Nonlinear Functional Analysis and Application.* 311–359. N.Y.: Academic Press, 1971.

[322] **Nashed, M. Z. and Wahba, G.,** Generalized inverses in reproducing kernel spaces: an approach to regularization of linear operator equations // *SIAM J. Math. Anal.* 5 (6). 974–987. 1974.

[323] **Nedelkov, I. P.,** Improper problems in computational physics // *Computer Physics Communs.* 157–164. 1972.

[324] **Polge, R. J., Bhagavan, B. K., and Carswell, J. M.,** Fast computational algorithms for bit reversal // *IEEE Trans. On Computers.* C-23 (1). 1974.

[325] **Phillips, D. L.,** A technique for the numerical solution of certain integral equations of the first kind // *JASM.* 9 (1). 84–97. 1962.

[326] **Reinsch, C. H.,** Smoothing by spline functions // *Num. Math.* 10. 177–183. 1967.

[327] **Reinsch, C. H.,** Two extensions of the Sard–Schönberg theory of best approximation // *SIAM J. Num. Anal.* 11 (1). 45–51. 1974.

[328] *Symposium on mathematics and numerical methods for inverse and ill-posed problems.* Abstracts. Report LITN-MAT-R-1977-5. Linkoping, 1977.

[329] **Sard, A.,** Approximations based on nonscalar observations // *J. Approx. Theory.* 4. 315–334. 1973.

[330] **Strand, O. N. and Westwater, E. R.,** Statistical estimation of the numerical solution of Fredholm integral equations of the first kind // *JACM.* 15 (1). 100–114. 1968.

[331] **Wahba, G.,** Practical approximate solutions to linear operator equations when the data are noisy // *SIAM. J. Num. Anal.* 14 (4). 651–667. 1977.

[332] **Wasilkowski, G. W. and Wozniakowski, H.,** *Optimality of spline algorithms.* Report CMU-CS-79-104. Warsaw, 1979.

[333] **Wedin, P. A.,** Perturbation theory for pseudo-inverses // *BIT.* 13. 217–232. 1973.

[334] **Torsti, J. and Aurela, A. A.,** A fast quadratic programming method for solving ill-conditioned systems of equations // *J. of Math. Anal. and Appl.* 38 (1). 193–205. 1972.

**Additional Bibliography to English Edition**

[335] **Morozov, V. A. and Grebennikov, A. I.,** *Methods for solving ill-posed problems: algorithmic aspect.* Moscow: Moscow University Press, 1992.

[336] **Morozov, V. A.,** *Regularization methods for ill-posed problems.* London: CRC Press, 1993.

[337] **Morozov, V. A.,** Estimates of the accuracy of regularization for nonlinear unstable problems // *J. Comp. Math. and Math. Phys.* 35 (9). 1139–1145. 1995.

[338] **Morozov, V. A.,** Regularization for the case of considerable noise // *J. Comp. Math. and Math. Phys.* 36 (9). 1175–1181. 1996.

[339] **Morozov, V. A.,** On the convergence rate for regularized solutions // *Numerical Functional Analysis and Optimization.* 19 (3). 345–352. 1998.

[340] **Grebennikov, A. I.,** The spline approximation method for recovery of functions // *Sov. J. Numer. Anal. Mathem. Mod.* 4 (4). 1–15. 1989.

[341] **Grebennikov, A. I.,** Solving integral equations with singularities in kernels by the spline approximation method // *Sov. J. Numer. Anal. Mathem. Mod.* 5 (3). 199–208. 1990.

[342] **Grebennikov, A. I.,** Fast and stable algorithms for solving some inverse problems // *Algorithmic and Numerical Analysis of Ill-Posed Problems. All-Russia Conference Dedicated to the Memory of V. K. Ivanov.* P. 53. Ekaterinburg, 1995.

[343] **Grebennikov, A. I.,** The spline approximation method and its application // *Collection of Reports. International Conference "Optimization of Computations"*. 81–86. Kiev, 1997.

[344] **Grebennikov, A. I.,** Convergence of the spline approximation method for large errors // *Numerical Methods of Analysis*. 94–105. Moscow: Moscow University Press, 1997.

[345] **Grebennikov, A. I.,** Numerical restoration of coefficients in differential equations // *Algorithmical Analysis of Ill-Posed Problems. All-Russia Conference Dedicated to the Memory of V. K. Ivanov.* P. 75. Ekaterinburg, 1998.

[346] **Morozov, V. A., et al.,** The pointwise residual method for solving some problems in linear algebra and linear programming // *J. Comp. Math. and Math. Phys.* 38 (8). 975–981. 1998.

[347] **Morozov, V. A., et al.,** The pointwise residual method for equations and inequalities with approximate data // *J. Fund. and Appl. Math.* 4 (3). 537–542. 1998.

[348] **Gilyazov, S. F. and Morozov, V. A.,** Regularization of ill-posed problems with normally resolvable operators // *J. Comp. Math. and Math. Phys.* 37 (2). 133–138. 1997.

[349] **Morozov, V. A.,** Some aspects of signal restoration by the regularization method // *Collection of Reports. IV International Conference "Numerical Methods and Applications"*. Sofia, 1998.

[350] **Grebennikov, A. I.,** The spline approximation method and its application for solving some ill-posed problems // *Proceedings of the 11th Baikal International School-Seminar "Optimization Methods and Their Applications"*. 94–97. Irkutsk: Melentjev Institute of Energy Systems SB RAS, 1998.

[351] **Grebennikov, A. I.,** A stable restoration of coefficients of heat-conduction equations by splines // *Collection of Reports. III International Conference "Identification of Dynamical Systems and Inverse Problems"*. 114–119. Moscow, 1998.

[352] **Grebennikov, A. I.,** Spline algorithms for data processing and solving some inverse problems // *The IV International Conference "Numerical Methods and Applications"*. P. 51. Sofia, 1998.

[353] **Grebennikov, A. I.,** Solution of some inverse problems by the spline approximation method // *Proc. of the Conference "Inverse and ill-posed problems"*. P. 25. Moscow: Dialog-MSU Press, 1998.

[354] **Grebennikov, A. I.,** Algorithms of coefficient restoration for ordinary differential equations // *Numerical Analysis: Methods and Algorithms.* 53–61. Moscow: Moscow University Press, 1998.

[355] **Grebennikov, A. I.,** Spline algorithms for multidimensional data processing and solving the coefficient inverse problems for heat-conduction equations // *Numerical Analysis: Methods and Algorithms.* 62–71. Moscow: Moscow University Press, 1998.

[356] **Grebennikov, A. I.,** Spline algorithms for data processing and solving some inverse problems // *Recent Advances in Numerical Methods and Applications II. Proc. of the Fourth International Conference NMA'98.* 375–383. Singapore: World Scientific, 1999.

[357] **Grebennikov, A. I.,** Metodo de aproximacion spline para resolver problemas inestables y su aplicacion. // *Revista Aportaciones Matematicas. Serie Comunicaciones.* 27. 53–70. 2000.

[358] **Fraguela, A., Oliveros, J., and Grebennikov, A. I.,** Planteamiento operacional y analisis del problema inverso electroencefalografico // *Revista Mexicana de Fisica.* 47 (2). 162–174. 2001.

[359] **Fraguela, A., Oliveros, J., and Grebennikov, A. I.,** Uniqueness of the solution to the inverse electroencephalographic problem // *Numerical Analysis and Applications*, Eds. J. Wasniewski, L. Vulkov, and P. Yalamov. Lecture Notes in Computer Science. Vol. 1988. 207–213. Berlin: Springer Verlag, 2001.

[360] **Grebennikov, A. I., Fraguela, A., and Solis, S. S.,** Analysis of one electroencephalographic inverse problem for the human head model consisting of two no-concentric spheres using spherical harmonics. // *Mathematical Modeling of Complex Systems.* Eds. V. A. Sadovnichii, E. G. Doger. 144–146. Moscow: Moscow University Press, 2001.

[361] **Grebennikov, A. I., Fraguela, A., and Leyva, J.F.,** Mathematical and numerical analysis of periodic solutions in a neighborhood of stationary states in a nonlinear model of brain cortex activation // *Mathematical Modeling of Complex Systems.* Eds. V. A. Sadovnichii, E. G. Doger. 135–138. Moscow: Moscow University Press, 2001.

[362] **Grebennikov, A. I.,** The local spline approximation method for solving some coefficient inverse problems for differential equations of the parabolic type // *Inverse Problems in Engineering J.* 9. 455–469. 2001.

[363] **Grebennikov, A. I.,** Solucion de problemas mal-planteados por metodo de aproximacion spline y aplicacion // *Metodos Numericos en Ingenieria y Ciencias Aplicadas.* 1. 167–180. Barcelona : CIMNE, 2002.

[364] **Grebennikov, A. I.,** Regularization of applied inverse problems by the full spline approximation method // *WSEAS Transaction on Systems J.* 1 (2). 124–129. 2002.

[365] **Grebennikov, A. I.,** The spline regularization method for solving inverse problems: theory and algorithms // *Mathematical Modeling J.* 3. 53–60. 2002.

[366] **Grebennikov, A. I.,** Regularization algorithms for electric tomography image reconstruction // *WSEAS Transaction on Systems J.* 2 (2). 487–492. 2003.

[367] **Grebennikov, A. I.,** Local regularization algorithms of solving coefficient inverse problems for some differential equations // *Inverse Problems in Engineering J.* 11 (3). 103–113. 2003.

[368] **Grebennikov, A. I. and Leyva, F.,** Analysis of approximate differential equations in a one-dimensional model of the brain cortex dynamics // *Proc. of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* 2807–2809. Mexico. Cancun, 2003.

[369] **Grebennikov, A. I.,** Spline smoothing algorithms for improving the capacitance of tomography image reconstruction // *Proc. of the 3rd World Congress on Industrial Process Tomography.* 800–805. Canada. Banff, 2003.

[370] **Grebennikov, A. I.,** Fast image reconstruction algorithms for electric tomography // *Proc. of the 25th Silver Anniversary International Conference of the IEEE Engineering in Medicine and Biology Society.* 825–828. Mexico. Cancun, 2003.

[371] **Grebennikov, A. I.,** Mathematical and numerical modeling of scanning and penetrating computer tomography // *WSEAS Transaction on Computers J.* 2 (3). 548–553. 2003.

[372] **Grebennikov, A. I.,** A fast algorithm for solution of the Dirichlet problem for the Laplace equation // *WSEAS Transaction on Computers J.* 4 (2). 1039–1043. 2003.

[373] **Grebennikov, A. I.,** The study of the approximation quality of the GR-method for solving the Dirichlet problem for the Laplace equation // *WSEAS Transaction on Mathematics J.* 4 (2). 312–317. 2003.

[374] **Grebennikov, A. I.,** Fast algorithms for solution of some nonlinear coefficient inverse problems for differential equations and applications // *Memorias de III Congreso Internacional Metodos Numericos en Ingenierea y Ciencias Aplicadas.* 22–24. Mexico. Monterrey, 2004.

[375] **Grebennikov, A. I.,** A novel approach for solution of boundary value problems for differential equations of mathematical physics // *WSEAS Transaction on Systems J.* 4 (3). 1410–1415. 2004.

[376] **Grebennikov, A. I.,** A general ray (GR) method for solution of differential equations for electromagnetic fields // *WSEAS Transaction on Communications J.* 2 (3). 809–811. 2004.

[377] **Grebennikov, A. I.,** Transformation of an arbitrary simply-connected star domain to unit circle and application for solution of boundary value problems for the Laplace equation // *WSEAS Transaction on Communications J.* 2 (3). 794–796. 2004.

[378] **Grebennikov, A. I.,** *Spline approximation method and its applications.* Moscow: MAX Press, 2004.