Numerical Methods for Solving Inverse Problems of Mathematical Physics

A. A. Samarskii
P. N. Vabishchevich

INVERSE AND ILL-POSED PROBLEMS SERIES

de Gruyter

Inverse and Ill-Posed Problems Series

Alexander A. Samarskii
Peter N. Vabishchevich

Numerical Methods for Solving Inverse Problems of Mathematical Physics

Walter de Gruyter · Berlin · New York

Keywords:

Inverse problems, mathematical physics, boundary value problems, ordinary differential equations, elliptic equations, parabolic equations, right-hand side identification, evolutionary inverse problems, ill-posed problems, regularization methods, Tikhonov regularization, conjugate gradient method, discrepancy principle, finite difference methods, finite element methods.

Mathematics Subject Classification 2000:
65-02, 65F22, 65J20, 65L09, 65M32, 65N21

♾ Printed on acid-free paper which falls within the guidelines of the ANSI to ensure permanence and durability.

ISBN 978-3-11-019666-5

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Printed in Germany
Cover design: Martin Zech, Bremen.
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen.

Preface

Applied problems often require solving boundary value problems for partial differential equations. Elaboration of approximate solution methods for such problems rests on the development and examination of numerical methods for boundary value problems formulated for basic (fundamental, model) mathematical physics equations. If one considers second-order equations, then such equations are elliptic, parabolic and hyperbolic equations.

The solution of a boundary value problem is to be found from the equation and from some additional conditions. For time-independent equations, to be specified are boundary conditions, and for time-dependent equations, in addition, initial conditions. Such classical problems are treated in all tutorials on mathematical physics equations and partial differential equations.

The above boundary value problems belong to the class of direct mathematical physics problems. A typical inverse problem is the problem in which it is required to find equation coefficients from some additional information about the solution; in the latter case, the problem is called a coefficient inverse problem. In boundary inverse problems, to be reconstructed are unknown boundary conditions, and so on.

Inverse mathematical physics problems often belong to the class of classically ill-posed problems. First of all, ill-posedness here is a consequence of lacking continuous dependence of solution on input data. In this case, one has to narrow the class of admissible solutions and, to determine a stable solution, apply some special regularizing procedures.

Numerical solution of direct mathematical physics problems is presently a well-studied matter. In solving multi-dimensional boundary value problems, difference methods and the finite element method are widely used. At present, tutorials and monographs on numerical solution methods for inverse problems are few in number. The latter was the primary motivation behind writing the present book.

By no means being a comprehensive guide, this book treats some particular inverse problems for time-dependent and time-independent equations often encountered in mathematical physics. Rather a complete and closed consideration of basic difficulties in approximate solution of inverse problems is given. A minimum mathematical apparatus is used, related with some basic properties of operators in finite-dimensional spaces.

A predominant contribution to the scope of problems dealt with in the theory and solution practice of inverse mathematical physics problems was made by Russian mathematicians, and the pioneer here was Andrei Nikolaevich Tikhonov. His ideas, underlying the modern applied mathematics, are now developed by his numerous disciples. Our work pays tribute to A. N. Tikhonov.

Main definitions and notations

A, B, C, D, S — difference operators;

E — unit (identity) operator;

A^* — adjoint operator;

A^{-1} — operator inverse to A;

$A > 0$ — positive operator ($(Ay, y) > 0$ if $y \neq 0$);

$A \geq 0$ — non-negative operator ($(Ay, y) \geq 0$);

$A \geq \delta E$, $\delta > 0$ — positive definite operator;

$A_0 = (A + A^*)/2$ — self-adjoint part of A;

$A_1 = (A - A^*)/2$ — skew-symmetric part of A;

H — Hilbert space of mesh functions;

$(\cdot, \cdot)$ — scalar product in H;

$\|\cdot\|$ — norm in H;

$(y, v)_A = (Ay, v)$ — scalar product in H_A ($A = A^* > 0$);

$\|\cdot\|_A$ — norm in H_A;

$L_2(\omega)$ — Hilbert space of mesh functions;

$\|\cdot\|$ — norm in L_2;

$\|A\|$ — norm of difference operator A;

M, M_β — positive constants;

Ω — computation domain;

$\partial\Omega$ — boundary of Ω;

ω — set of internal nodes;

$\partial\omega$ — set of boundary nodes;

h, h_β — mesh size in space;

τ — time step;

σ — weighting parameter of difference scheme;

$y_x = \dfrac{y(x+h) - y(x)}{h}$ — right-hand difference derivative at the point x;

$y_{\bar{x}} = \dfrac{y(x) - y(x-h)}{h}$ — left-hand difference derivative at the point x;

$y_{\mathring{x}} = (y_x + y_{\bar{x}})/2$ — central difference derivative at the point x;

$y_{\bar{x}x} = \dfrac{y_x - y_{\bar{x}}}{h}$ — second difference derivative at the point x;

$y = y_n = y(x, t_n)$ — magnitude of mesh function at the point x at the time $t_n = n\tau$, $n = 0, 1, \ldots$;

α — regularization parameter;

δ — typical input-data inaccuracy;

Contents

1 Inverse mathematical physics problems

We associate direct mathematical physics problems with classical boundary value problems often encountered in mathematical physics. In a direct problem, it is required to find a solution that satisfies some given partial differential equation and some initial and boundary conditions. In inverse problems, the master equation and/or initial conditions and/or boundary conditions are not fully specified but, instead, some additional information is available. So separating out inverse mathematical physics problems, we can speak of coefficient inverse problems (in which the equation is not specified completely as some equation coefficients are unknown), boundary inverse problems (in which boundary conditions are unknown), and evolutionary inverse problems (in which initial conditions are unknown). Very often, inverse problems are problems ill-posed in the classical sense. A typical feature here is the violated requirement of solution continuity on input data. An ill-posed problem can be moved up into the class of well-posed problems by narrowing the class of admissible solutions.

1.1 Boundary value problems

The core of applied mathematical models is made up by partial differential equations. Here, the solution can be found from mathematical physics equations and some additional relations. Such additional relations are, first of all, boundary and initial conditions. In courses on mathematical physics equations, equations most important for applications are second-order equations. Among such equations are elliptic, parabolic, and hyperbolic equations.

1.1.1 Stationary mathematical physics problems

As an example, consider several two-dimensional boundary value problems. The solution $u(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, x_2)$ is to be found in some bounded domain Ω with a sufficiently smooth boundary $\partial\Omega$. This solution is defined by the *second-order elliptic equation*

$$-\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big) + q(\boldsymbol{x})u = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{1.1}$$

Normally, the constraints imposed on the equation coefficients look as

$$k(\boldsymbol{x}) \geq \kappa > 0, \qquad q(\boldsymbol{x}) \geq 0, \qquad \boldsymbol{x} \in \Omega.$$

A typical example of an elliptic equation (1.1) is given by the *Poisson equation*

$$-\Delta u \equiv -\sum_{\alpha=1}^{2} \frac{\partial^2 u}{\partial x_\alpha^2} = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \tag{1.2}$$

i. e., in (1.1) we have $k(\boldsymbol{x}) = 1$ and $q(\boldsymbol{x}) = 0$.

For equation (1.1), we consider *first-kind boundary conditions*

$$u(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega. \tag{1.3}$$

On the boundary of Ω or on a part of the boundary, second- or third-kind boundary conditions can also be given. In the case of *third-kind boundary conditions*, we have

$$k(\boldsymbol{x}) \frac{\partial u}{\partial n} + \sigma(\boldsymbol{x})u = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega, \tag{1.4}$$

where n is the external normal to Ω.

Many key features of stationary mathematical physics problems for second-order elliptic equations can be figured out considering simplest boundary value problems for the *second-order ordinary differential equation*. A prototype of (1.1) is the equation

$$-\frac{d}{dx}\Big(k(x)\frac{du}{dx}\Big) + q(x)u = f(x), \qquad 0 < x < l \tag{1.5}$$

with variable coefficients

$$k(x) \geq \kappa > 0, \qquad q(x) \geq 0.$$

For the unknown function $u(x)$ to be uniquely determined, equation (1.5) must be supplemented with two boundary conditions given at the end points of the segment $[0, l]$. Here, either the function $u(x)$ (first-kind boundary condition), the flux $w(x) = -k(x)\dfrac{du}{dx}(x)$ (second-kind boundary condition), or a linear combination of the above conditions (third-kind boundary condition) can be considered:

$$u(0) = \mu_1, \qquad u(l) = \mu_2, \tag{1.6}$$

$$-k(0)\frac{du}{dx}(0) = \mu_1, \qquad k(l)\frac{du}{dx}(l) = \mu_2, \tag{1.7}$$

$$-k(0)\frac{du}{dx}(0) + \sigma_1 u(0) = \mu_1, \qquad k(l)\frac{du}{dx}(l) + \sigma_2 u(l) = \mu_2. \tag{1.8}$$

1.1.2 Nonstationary mathematical physics problems

A fundamental time-dependent equation in mathematical physics is the *one-dimensional second-order parabolic equation*. In a rectangle

$$\overline{Q}_T = \overline{\Omega} \times [0, T], \qquad \overline{\Omega} = \{x \mid 0 \leq x \leq l\}, \qquad 0 \leq t \leq T$$

we consider the equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\frac{\partial u}{\partial x}\Big) + f(x,t), \qquad 0 < x < l, \quad 0 < t \leq T. \tag{1.9}$$

This equation is supplemented (the first boundary value problem) with the boundary conditions

$$u(0,t)=0, \qquad u(l,t)=0, \qquad 0<t\le T \tag{1.10}$$

and with the initial condition

$$u(x,0)=u_0(x), \qquad 0\le x\le l. \tag{1.11}$$

For simplicity, here we have restricted ourselves to homogeneous boundary conditions and to the case in which the coefficient k depends on the spatial variable only and, in addition, $k(x)\ge\kappa>0$.

Instead of first-kind conditions (1.10), other boundary conditions can be considered. For instance, in many applied problems one has to formulate third-type boundary conditions:

$$\begin{aligned} -k(0)\,\frac{du}{dx}(0,t)+\sigma_1(t)u(0,t)&=\mu_1(t),\\ k(l)\,\frac{du}{dx}(l,t)+\sigma_2(t)u(l,t)&=\mu_2(t), \end{aligned} \qquad 0<t\le T. \tag{1.12}$$

Among other nonstationary boundary value problems, the *second-order hyperbolic equation* is to be considered. In the spatially one-dimensional case, we seek the solution of the following equation:

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big)+f(x,t), \qquad 0<x<l, \quad 0<t\le T. \tag{1.13}$$

For the solution to be defined completely, in addition to (1.10) the following two initial conditions are to be considered:

$$u(x,0)=u_0(x), \qquad \frac{\partial u}{\partial t}(0,t)=u_1(x), \qquad 0\le x\le l. \tag{1.14}$$

Particular attention must be given to multi-dimensional nonstationary mathematical physics problems. An example here is the two-dimensional parabolic equation. We seek, in some domain Ω, a function $u(\boldsymbol{x},t)$ which satisfies the equation

$$\begin{aligned} \frac{\partial u}{\partial t}=\sum_{\alpha=1}^{2}\frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big)-q(\boldsymbol{x},t)u+f(\boldsymbol{x},t),\\ \boldsymbol{x}\in\Omega, \qquad 0\le t\le T \end{aligned} \tag{1.15}$$

and the conditions

$$u(\boldsymbol{x},t)=0, \qquad \boldsymbol{x}\in\partial\Omega, \quad 0<t\le T, \tag{1.16}$$

$$u(\boldsymbol{x},0)=u_0(\boldsymbol{x}), \qquad \boldsymbol{x}\in\Omega. \tag{1.17}$$

In a similar way, other nonstationary multidimensional boundary value problems for partial differential equations can be formulated.

1.2 Well-posed problems for partial differential equations

Here, we introduce the notion of well-posed boundary value problem, related with the existence of a unique solution that continuously depends on input data. Results on stability of classical boundary value problems for partial differential equations are presented.

1.2.1 The notion of well-posedness

Boundary and initial conditions are formulated to identify, among the whole set of possible solutions of a partial differential equation, a desired solution. These additional conditions must be not too numerous (solutions must exist), nor they must be few in number (solutions must be not numerous). With this circumstance, the notion of well-posed statement of a problem is related. Let us dwell first on the notion of *well-posedness of a problem* according to J. Hadamard (well-posedness in the classical sense).

A problem is *well-posed* if:

1) the solution of the problem does exist;

2) the solution is unique;

3) the solution continuously depends on input data.

It is the third condition for well-posedness that is of primary significance here. This condition provides for smallness of the solution changes resulting from small input-data variation. The input data are the equation coefficients, the right-hand side and the boundary and initial conditions, taken from an experiment and always known to some limited accuracy. In fact, solution stability with respect to small perturbations of initial and boundary conditions, coefficients, and right-hand side justifies the problem statement itself, as well as its cognitive essence, and makes the whole study valuable.

In consideration of boundary value problems for mathematical physics equations, the existence, uniqueness and stability theorems taken as a whole provide a complete study of well-posedness of a posed problem. Of course, conditions for well-posedness must be rendered concrete considering each particular problem. The latter is related with the fact that the solution of the problem and the input data are considered as elements in a certain fully defined functional space. That is why a given problem can be ill-posed with one choice of spaces and well-posed with another choice of spaces. Hence, a statement that this or that problem is a well- (ill-)posed one is never global: such statements must be supplemented with necessary amendments.

1.2.2 Boundary value problem for the parabolic equation

Some fundamental points in a consideration of well- or ill-posedness of a boundary value mathematical physics problem can be illustrated with the example of a simplest boundary value problem for the one-dimensional parabolic equation (1.9)–(1.11).

Here, we do not touch on points concerning the solution existence; instead, we restrict ourselves to the uniqueness problem and to the property of continuous dependence of solution on input data. We assume that problem (1.9)–(1.11) indeed has a classical solution $u(x, t)$ (for instance, this solution is doubly differentiable with respect to x and continuously differentiable with respect to t).

We write problem (1.9)–(1.11) as a Cauchy problem for a first-order operator differential equation. For functions given in the domain $\Omega = (0, 1)$ and vanishing at the boundary points of this domain (at the boundary $\partial\Omega$), we define a Hilbert space $\mathcal{H} = \mathcal{L}_2(\Omega)$ in which the scalar product is defined as

$$(v, w) = \int_\Omega v(x)w(x)\,dx.$$

For the norm in $\mathcal{H}$ we use

$$\|v\| = (v, v)^{1/2} = \Big(\int_\Omega v^2(x)\,dx\Big)^{1/2}.$$

For functions satisfying the boundary conditions (1.10), we define the operator

$$\mathcal{A}u = -\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l. \tag{1.18}$$

With the notation introduced, equation (1.9) supplemented with conditions (1.10) at the boundary $\partial\Omega$ can be written as an operator differential equation for the function $u(t) \in \mathcal{H}$:

$$\frac{du}{dt} + \mathcal{A}u = f(t), \qquad 0 < t \le T. \tag{1.19}$$

The initial condition (1.11) can be written as

$$u(0) = u_0. \tag{1.20}$$

The following main properties of the operator $\mathcal{A}$ defined by (1.18) are worthy of note. The operator $\mathcal{A}$ is a self-adjoint operator non-negative in $\mathcal{H}$:

$$\mathcal{A}^* = \mathcal{A} \ge 0. \tag{1.21}$$

The property of self-adjointness stems from the expression

$$(\mathcal{A}v, w) = \int_0^l \mathcal{A}v(x)w(x)\,dx = \int_0^l k(x)\,\frac{\partial v}{\partial x}\,\frac{\partial w}{\partial x}\,dx = (v, \mathcal{A}^*w),$$

derived with regard to the fact that the functions $v(x)$ and $w(x)$ vanish at the points $x \in \partial\Omega$. For the functions $u(x) = 0$, $x \in \partial\Omega$, we have

$$(\mathcal{A}v, v) = \int_0^l k(x)\Big(\frac{\partial v}{\partial x}\Big)^2\,dx \ge 0$$

and, hence, $\mathcal{A} \geq 0$.

Let us derive now a simplest a priori estimate for the solution of problem (1.19), (1.20). In consideration of problems for evolutionary equations, the *Gronwall lemma* is of significance. Here, we restrict ourselves to this lemma formulated in its simplest form.

Lemma 1.1 *For a function $g(t)$ satisfying the inequality*

$$\frac{dg}{dt} \leq ag(t) + b(t), \qquad t > 0$$

with $a = \text{const}$ and $b(t) \geq 0$, the following estimate holds:

$$g(t) \leq \exp(at)\Big(g(0) + \int_0^t \exp(-a\theta)b(\theta)\,d\theta\Big).$$

Theorem 1.2 *For the solution of problem* (1.19), (1.20)*, the following a priori estimate holds:*

$$\|u(t)\| \leq \|u_0\| + \int_0^t \|f(\theta)\|\,d\theta. \tag{1.22}$$

Proof. From scalarwise multiplying equation (1.19) by $u(t)$ we obtain the equality

$$\Big(\frac{du}{dt}, u\Big) + (Au, u) = (f, u).$$

Taking the Cauchy–Bunyakowsky (Schwarz) inequality into account, we obtain:

$$\Big(\frac{du}{dt}, u\Big) = \frac{1}{2}\frac{d}{dt}\|u\|^2 = \|u\|\frac{d}{dt}\|u\|,$$

$$(f, u) \leq \|f\|\,\|u\|.$$

From the latter equality and from the non-negativeness of A it follows that

$$\frac{d}{dt}\|u\| \leq \|f\|.$$

The latter inequality yields the estimate (1.22) (in the Gronwall lemma, we have $a = 0$). □

The well-posedness of the problem is related to the existence of a unique solution and with stability of this solution with respect to small input-data perturbations.

Corollary 1.3 *The solution of problem* (1.19), (1.20) *is unique.*

Suppose that we have two solutions $u_1(t)$ and $u_2(t)$. Then, the difference $u(t) = u_1(t) - u_2(t)$ satisfies equation (1.19) with $f(t) = 0$, $0 < t \le T$ and with the homogeneous initial condition ($u_0 = 0$). From the a priori estimate (1.20) it readily follows that $u(t) = 0$ at all times in the interval $0 \le t \le T$.

In the problem of interest, it is necessary to take into account, first of all, the initial conditions. In this case, we speak of *solution stability with respect to initial data*. The input data here are the equation coefficients (*coefficient stability*). In particular, it makes sense to examine the dependence of solution on the right-hand side of the equation or, in other words, *solution stability with respect to the right-hand side*. Let us show, for instance, that the obtained a priori estimate (1.22) guarantees both solution stability with respect to input data and solution stability with respect to the right-hand side.

Apart from problem (1.19), (1.20), consider a problem with perturbed initial condition and with perturbed right-hand side:

$$\frac{d\tilde{u}}{dt} + \mathcal{A}\tilde{u} = \tilde{f}(t), \qquad 0 < t \le T. \tag{1.23}$$

The initial condition (1.11) can be written as

$$\tilde{u}(0) = \tilde{u}_0. \tag{1.24}$$

Corollary 1.4 *Suppose we have*

$$\|u_0 - \tilde{u}_0\| \le \varepsilon, \qquad \|f(t) - \tilde{f}(t)\| \le \varepsilon, \quad 0 \le t \le T,$$

where $\varepsilon > 0$. *Then,*

$$\|u(t) - \tilde{u}(t)\| \le M\varepsilon,$$

where $M = 1 + T$.

The latter inequality guarantees continuous dependence of the solution of problem (1.19), (1.20) on the right-hand side and on the initial conditions. For $\delta u(t) = u(t) - \tilde{u}(t)$, from (1.19), (1.20), (1.23), and (1.24) we obtain the problem

$$\frac{d\delta u}{dt} + \mathcal{A}\delta u = \delta f(t), \qquad 0 < t \le T. \tag{1.25}$$

The initial condition (1.11) can be written as

$$\delta u(0) = \delta u_0, \tag{1.26}$$

where

$$\delta u_0 = u_0 - \tilde{u}_0, \qquad \delta f(t) = f(t) - \tilde{f}(t).$$

For the solution of problem (1.25), (1.26), there holds the a priori estimate (see (1.22))

$$\|\delta u(t)\| \le \|\delta u_0\| + \int_0^t \|\delta f(\theta)\|\, d\theta,$$

and, hence,

$$\|\delta u(t)\| \le (1+T)\varepsilon.$$

Estimates of the coefficient stability of the solution of problem (1.18)–(1.20) (with respect to $k(x)$) are more difficult to obtain.

1.2.3 Boundary value problem for the elliptic equation

In consideration of elliptic boundary value problems, primary attention is normally paid to a priori estimates of solution stability with respect to boundary conditions and right-hand side. As an example of a stationary mathematical physics problem, consider the Dirichlet problem for the Poisson equation (1.2), (1.3).

Similarly to the above case of a parabolic problem, we can try to derive a priori estimates of solutions of boundary value problems for the second-order elliptic equations in Hilbert spaces. Here, worthy of noting is the possibility to obtain a priori estimates in other norms. Our consideration is based on using the well-known *maximum principle*.

Theorem 1.5 (Maximum principle) *Suppose that in problem* (1.2), (1.3) *we have* $f(\boldsymbol{x}) \le 0$ *(*$f(\boldsymbol{x}) \ge 0$*) in a bounded domain* Ω*. Then, the solution* $u(\boldsymbol{x})$ *attains its maximum (minimum) value at the boundary of the domain, i.e.,*

$$\max_{\boldsymbol{x}\in\Omega} u(\boldsymbol{x}) = \max_{\boldsymbol{x}\in\partial\Omega} \mu(\boldsymbol{x}), \qquad \min_{\boldsymbol{x}\in\Omega} u(\boldsymbol{x}) = \min_{\boldsymbol{x}\in\partial\Omega} \mu(\boldsymbol{x}). \tag{1.27}$$

A trivial corollary resulting from the maximum principle is the simple statement about uniqueness of the solution of the Dirichlet problem for the Poisson equation.

Corollary 1.6 *The solution of problem* (1.2), (1.3) *is unique.*

Based on Theorem 1.5, we can also derive a priori estimates showing that the solution of problem (1.2), (1.3) is stable with respect to the right-hand side and with respect to boundary conditions in homogeneous norm. We use the settings

$$\|u(\boldsymbol{x})\|_{C(\Omega)} = \max_{\boldsymbol{x}\in\Omega} |u(\boldsymbol{x})|.$$

Theorem 1.7 *For the solution of problem* (1.2), (1.3) *the following a priori estimate holds:*

$$\|u(\boldsymbol{x})\|_{C(\Omega)} \le \|\mu(\boldsymbol{x})\|_{C(\partial\Omega)} + M\|f(\boldsymbol{x})\|_{C(\Omega)}. \tag{1.28}$$

Here, the constant M *depends on the diameter of* Ω*.*

Proof. We choose a function $v(\boldsymbol{x})$ that satisfies the conditions

$$-\Delta v \ge 1, \qquad \boldsymbol{x} \in \Omega, \tag{1.29}$$

$$v(\boldsymbol{x}) \ge 0, \qquad \boldsymbol{x} \in \partial\Omega. \tag{1.30}$$

Consider the functions

$$\begin{aligned} w_+(\boldsymbol{x}) &= v(\boldsymbol{x})\|f(\boldsymbol{x})\|_{C(\Omega)} + \|\mu(\boldsymbol{x})\|_{C(\partial\Omega)} + u(x), \\ w_-(\boldsymbol{x}) &= v(\boldsymbol{x})\|f(\boldsymbol{x})\|_{C(\Omega)} + \|\mu(\boldsymbol{x})\|_{C(\partial\Omega)} - u(x). \end{aligned} \tag{1.31}$$

Taking into account statement (1.27) and inequality (1.30), we readily obtain that $w_\pm(\boldsymbol{x}) \geq 0$ on the boundary of Ω. Inside the domain Ω, equalities (1.2) and (1.29) yield

$$-\Delta w_\pm = -\|f(\boldsymbol{x})\|_{C(\Omega)}\,\Delta v \pm f(\boldsymbol{x}).$$

With regard to (1.29), we have $-\Delta w_\pm(\boldsymbol{x}) \geq 0$. Based on the maximum principle, we obtain that $w(\boldsymbol{x}) \geq 0$ everywhere in Ω.

Since the functions $w_\pm(\boldsymbol{x})$ are non-negative functions, then it readily follows from (1.31) that

$$\|u(\boldsymbol{x})\|_{C(\Omega)} \leq \|\mu(\boldsymbol{x})\|_{C(\partial\Omega)} + \|v(\boldsymbol{x})\|_{C(\Omega)}\|f(\boldsymbol{x})\|_{C(\Omega)}.$$

In this way, we have derived the a priori estimate (1.28) with $M = \|v(\boldsymbol{x})\|_{C(\Omega)}$.

Render concrete the magnitude of M and its dependence on the calculation domain. We assume that the whole bounded domain Ω lies in a circle of radius R centered at $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_1^{(0)})$.

We put

$$v(\boldsymbol{x}) = c(R^2 - (x_1 - x_1^{(0)})^2 - (x_2 - x_2^{(0)})^2)$$

with some still undetermined positive constant c. Apparently, $v(\boldsymbol{x}) \geq 0$, $\boldsymbol{x} \in \partial\Omega$ and

$$-\Delta v = 4c.$$

Hence, with $c = 1/4$ the conditions (1.29) and (1.30) are fulfilled and, therefore, the constant M in (1.28) is given by

$$M = \max_{\boldsymbol{x}\in\Omega} \frac{1}{4}\,(R^2 - (x_1 - x_1^{(0)})^2 - (x_2 - x_2^{(0)})^2) = \frac{R^2}{4}.$$

This constant depends just on the diameter of Ω. □

The a priori estimate (1.28) guarantees solution stability of problem (1.2), (1.3) with respect to the right-hand side and boundary conditions. Similarly, the case of more general boundary value problems for the second-order elliptic equation (1.1) with third-kind boundary conditions (1.4) can be considered.

1.3 Ill-posed problems

Inverse mathematical physics problems are often assigned to the class of problems incorrect in the classical sense. As an example of well-posed problems, below we consider the inverted-time problem for the second-order parabolic equation, for which continuous dependence of the solution on the input data is lacking. Upon narrowing the class of solutions, stability appears; hence, this problem belongs to the class of conditionally well-posed problems (Tikhonov well-posed problems).

1.3.1 Example of an ill-posed problem

Problems in which some of the three conditions for well-posed statement (existence, uniqueness or stability) are not fulfilled belong to the class of *ill-posed problems*. Here, the determining role is played by continuous dependence of solution on input data. Consider several examples of ill-posed mathematical physics problems.

For elliptic equations, well-posed problems are problems with given boundary conditions (see, for instance, (1.1), (1.4)). One can consider a Cauchy problem for elliptic equations in which conditions are given not on the whole boundary $\partial\Omega$, but only on some part of the boundary $\Gamma \subset \partial\Omega$. The solution $u(\boldsymbol{x})$ is to be determined from equation (1.1) and from the following two conditions given on Γ:

$$u(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \frac{\partial u}{\partial n}(\boldsymbol{x}) = \nu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Gamma. \tag{1.32}$$

Ill-posedness of the Cauchy problem (1.1), (1.32) is a consequence of solution instability with respect to initial conditions. Continuous dependence takes place only if the solution and the initial data are analytic functions. In functional spaces whose norms involve a finite number of derivatives there is no continuous dependence of solution on input data.

For parabolic equations, well-posed problems are problems with given boundary and initial conditions (see (1.9)–(1.11)). By specifying the end-time solution, we obtain an inverted-time problem in which from a given state it is required to reconstruct the pre-history of the process. Let us dwell on the following simple inverted-time problem:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < l, \quad 0 \le t < T, \tag{1.33}$$

$$u(0,t) = 0, \qquad u(l,t) = 0, \qquad 0 \le t < T, \tag{1.34}$$

$$u(x,T) = u_T(x), \qquad 0 \le x \le l. \tag{1.35}$$

To explain the ill-posedness of the problem, consider the solution of problem (1.33)–(1.35) with the condition

$$u_T(x) = \frac{1}{k^p}\sqrt{\frac{2}{l}}\,\sin\Big(\frac{\pi k x}{l}\Big), \tag{1.36}$$

where k and p are positive integer numbers. In the Hilbert space norm $\mathcal{H} = \mathcal{L}_2(\Omega)$, $\Omega = (0,l)$, we have

$$\|u_T(x)\|^2 = \int_\Omega u_T^2(x)dx = \frac{1}{k^{2p}} \to 0$$

as $k \to \infty$; i.e., here, the initial condition is indefinitely small.

The exact solution of problem (1.33)–(1.36) is

$$u(x,t) = u_T(x) \exp\left(\left(\pi \frac{k}{l}\right)^2 (T-t)\right).$$

It follows from the latter representation that in the interval $0 \le t < T$

$$\|u(x,t)\| = \frac{1}{k^p} \exp\left(\left(\pi \frac{k}{l}\right)^2 (T-t)\right) \to \infty$$

as $k \to \infty$. Thus, perturbations in the initial condition, however small, increase indefinitely at $t < T$.

1.3.2 The notion of conditionally well-posed problems

The necessity to solve nonstationary problems similar to that presented above requires more exact determination of problem solution. In problems *conditionally well-posed according to Tikhonov*, we have to do not just with a solution, but with a solution that belongs to some class of solutions. Making the class of admissible solutions narrower allows one in some cases to pass to a well-posed problem.

We say that a problem is *Tikhonov well-posed* if:

1) the solution of the problem is a priori known to exist in some class;

2) in this class, the solution is unique;

3) the solution continuously depends on input data.

The fundamental difference here consists in separating out the class of admissible solutions. The classes of a priori restrictions differ widely. In consideration of ill-posed problems, the problem statement itself undergoes substantial changes: the condition that the solution belongs to a certain set is to be included into the problem statement.

1.3.3 Condition for well-posedness of the inverted-time problem

Previously, we have established continuous dependence of the solution of the evolutionary problem (1.18)–(1.20). Now, consider an ill-posed Cauchy problem (the initial condition (1.20)) for the equation

$$\frac{du}{dt} - \mathcal{A}u = f(t), \qquad 0 < t \le T \tag{1.37}$$

in which the operator $\mathcal{A}$ is defined by (1.18).

Let us derive an estimate of the solution of problem (1.18), (1.20), (1.37) with $f(t) = 0$ from which conditional well-posedness of the problem follows. In our investigation, we lean upon self-adjointness of the operator $\mathcal{A}$ and on the fact that $\mathcal{A}$ is a time-independent (constant) operator. We introduce the setting

$$\Phi(t) = \|u\|^2 = (u,u). \tag{1.38}$$

Direct differentiation of (1.38) with allowance for (1.37) yields:

$$\frac{d\Phi}{dt} \;=\; 2\Big(u,\,\frac{\partial u}{\partial t}\Big) \;=\; 2(u,\,\mathcal{A}u). \tag{1.39}$$

Taking the self-adjointness of $\mathcal{A}$ into account, after repeated differentiation we obtain:

$$\frac{d^2\Phi}{dt^2} \;=\; 4\Big(\mathcal{A}u,\,\frac{\partial u}{\partial t}\Big) \;=\; 4\Big\|\frac{\partial u}{\partial t}\Big\|^2. \tag{1.40}$$

It follows from (1.38)–(1.40) and from the Cauchy–Bunyakowsky inequality that

$$\Phi\,\frac{d^2\Phi}{dt^2} - \Big(\frac{d\Phi}{dt}\Big)^2 \;=\; 4\Big(\|u\|^2\Big\|\frac{\partial u}{\partial t}\Big\|^2 - \Big(u,\,\frac{\partial u}{\partial t}\Big)^2\Big) \;\geq\; 0. \tag{1.41}$$

Inequality (1.41) is equivalent to

$$\frac{d^2}{dt^2}\,\ln\Phi(t) \;\geq\; 0, \tag{1.42}$$

i. e. the function $\ln\Phi(t)$ is a convex function. From (1.42) we obtain

$$\ln\Phi(t) \;\leq\; \frac{t}{T}\,\ln\Phi(T) + \Big(1-\frac{t}{T}\Big)\ln\Phi(0).$$

From that it follows

$$\Phi(t) \;\leq\; (\Phi(T))^{t/T}(\Phi(0))^{1-t/T}.$$

With regard to (1.38), we obtain the desired estimate for the solution of problem (1.20), (1.37):

$$\|u(t)\| \;\leq\; \|u(T)\|^{t/T}\|u(0)\|^{1-t/T}. \tag{1.43}$$

Consider now the solution of problem (1.20, (1.37) in the class of solutions bounded in H, i.e.,

$$\|u(t)\| \leq M, \qquad 0 < t \leq T. \tag{1.44}$$

In the class of a priori constraints (1.44), from (1.43) we obtain the following estimate:

$$\|u(t)\| \leq M^{t/T}\|u(0)\|^{1-t/T}. \tag{1.45}$$

The latter means that for the problem (1.20), (1.37) in the class of bounded solutions there takes place continuous dependence of solution on initial data in the interval $0 < t < T$. On that basis, it makes sense to construct algorithms allowing approximate solution of the ill-posed problem (1.20), (1.37) and such that, this way or another, these algorithms can be used to separate out the class of bounded solutions. Besides, typical of the approximate solution must be an estimate of type (1.45) that admits the growth of the solution norm in time.

1.4 Classification of inverse mathematical physics problems

A boundary value problem for a partial differential equation is characterized by setting a master equation, a calculation domain, and boundary and initial conditions. That is why among inverse problems in heat transfer one can distinguish coefficient inverse problems, geometric inverse problems, boundary value inverse problems, and evolutionary inverse problems.

1.4.1 Direct and inverse problems

In treating data of full-scale experiments, additional indirect measurements are normally used to draw a conclusion about internal inter-relations in the phenomenon or process under study. If the structure of a mathematical model of the process is known, then one can pose a problem on identification of the mathematical model, in which, for instance, coefficients of the differential equation need to be determined. We assign such problems to the class of inverse mathematical physics problems.

Problems encountered in mathematical physics can be classed considering different characteristics. For instance, we can distinguish stationary problems for steady, time-independent processes and phenomena. Nonstationary problems describe dynamic processes, whose solution undergoes time variation. The demarcation line between direct and inverse mathematical physics problems is less obvious.

From the general methodological standpoint, we can call *direct problems* those problems for which causes are given and the quantities to be found are consequences. In view of this, *inverse* problems are problems in which consequences are known and causes are unknown. Yet, in practice such demarcation is not always easy to make.

In traditional courses on mathematical physics, for partial differential equations it is common practice to formulate well-posed boundary value problems, which are classed to direct problems. For second-order elliptic equations, additional conditions on the solution (of the first, second, or third kind) are given on the domain boundary. From the standpoint of cause-effect relations, the boundary conditions are causes, and the solution is a consequence. For parabolic equations, in addition, an initial condition has to be considered, and in the case of second-order hyperbolic equations the initial state is to be specified by setting the solution and its time derivative.

In order not to overload the consideration with subtle terminological points, we can say that it is the classical mathematical physics problems considered above that we assign to the class of direct problems. These problems are characterized by the necessity to find a solution from an equation with given coefficients and given right-hand side, and from additional boundary and initial conditions.

Under inverse mathematical physics problems, we mean problems that cannot be assigned to direct problems. In these problems, it is often required to determine not only the solution, but also some lacking coefficients and/or conditions. It is the necessity to determine not only the solution but also some parts of the mathematical model

that serves an indicator that the problem of interest is an inverse problem.

From this point of view, inverse problems are characterized, first of all, by the lack of some elements, elements in short that otherwise would allow one to assign the problem of interest to the class of direct mathematical physics problems or, in other words, elements making the problem an inverse problem. On the other hand, we must compensate for the lacking information. That is why in inverse problems one has to require additional information allowing him to hope that the solution can be uniquely found.

Using the noted indicators, one can classify inverse mathematical physics problems. It is natural to consider, first of all, those main characteristics that make a problem an inverse problem. For direct mathematical physics problems, the solution is defined by the equation (by the coefficients and by the right-hand side), by boundary conditions, and (in the case of nonstationary problems) by initial conditions. Inverse problems can be classified considering indications showing that some of the above-mentioned conditions are not specified.

1.4.2 Coefficient inverse problems

We distinguish *coefficient inverse problems*, in which equation coefficients and/or the right-hand side are unknown. As a typical example, consider the following parabolic equation:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) + f(x,t), \qquad 0 < x < l, \quad 0 < t \le T. \tag{1.46}$$

In a simplest direct problem, it is required to find a function $u(x,t)$ that satisfies equation (1.46) and the conditions

$$u(0,t) = 0, \qquad u(l,t) = 0, \qquad 0 < t \le T, \tag{1.47}$$

$$u(x,0) = u_0(x), \qquad 0 \le x \le l. \tag{1.48}$$

In applied problems, unknown properties of a medium are often to be determined. In the case under consideration, we can pose an identification problem for $k(x)$. In the simplest case of a homogeneous medium the unknown quantity is the value of $k(x) = \text{const}$, whereas for a piecewise-homogeneous medium, several values of the coefficient are to be determined. In the case of spatially dependent properties of the medium, a coefficient problem on reconstruction of $k = k(u)$ is of interest.

The list of possible statements of coefficient inverse problems is by no means exhausted by the above problems and can be continued. Typical here is the problem for equation (1.46) in which it is required to find two unknown functions $\{u(x,t), k(x)\}$. The main specific feature of this inverse coefficient problem is its nonlinearity.

As a special problem, one can isolate a problem in which it is required to find the unknown right-hand side $f(x,t)$ of parabolic equation (1.46). More specific statements

are related, for instance, with a particular case of

$$f(x,t) = \eta(t)\psi(x). \tag{1.49}$$

Of interest here is the unknown time dependence of the source (right-hand side) with known spatial distribution: in the representation (1.49), the function $\eta(t)$ is unknown, and the function $\psi(x)$ is given.

If coefficients and/or the right-hand side of (1.46) are unknown, then, apart from the conditions (1.47) and (1.48), one has to use some additional conditions. Such conditions must be not few in number in order to make it possible to uniquely found the solution of the inverse problem. If a coefficient is sought in the class of one-dimensional functions, then additional data must also be given in the same class.

Let us, for instance, consider the inverse problem (1.46)–(1.49) in which two functions, $\{u(x, y)$ and $\eta(t)\}$, are to be found. Apart from the solution of the boundary value problem (1.46)–(1.48), it is requited to find how the right-hand side depends on time. In this case, additional information can have the form

$$u(x^*, t) = \varphi(t), \qquad 0 < x^* < l, \quad 0 < t \le T, \tag{1.50}$$

i.e., the solution at each time is known not only on the boundary, but also at some internal point x^* of the calculation domain Ω.

In the case of inverse problems of type (1.46)–(1.50), primary attention must be paid to uniqueness problems for their solutions. This matter is especially important in the case in which nonlinear problems are treated (an example here is the problem on determination of two functions $\{u(x,t), k(x)\}$).

1.4.3 Boundary value inverse problems

In the cases in which direct measurements at the boundary are unfeasible, we deal with *boundary value inverse problems*. Here, missing boundary conditions can be identified, for instance, from measurements performed inside the domain. Consider, as an example, such an inverse problem for the parabolic equation (1.46).

We assume that measurements are unfeasible at the right end point of the segment $[0, l]$ but, instead, the solution is known at some internal point x^*, i.e., instead of conditions (1.47), the following conditions are given:

$$u(0,t) = 0, \qquad u(x^*, t) = \varphi(t), \qquad 0 < t \le T. \tag{1.51}$$

A typical statement of a boundary value inverse problem consists in determining the flux on some part of the boundary inaccessible for measurements (in the case under consideration, at $x = l$). This case corresponds to finding the pair of functions $\{u(x,t), k(x)\,\dfrac{\partial u}{\partial x}(l,t)\}$ from the conditions (1.46), (1.48), (1.51).

1.4.4 Evolutionary inverse problems

The direct problem for nonstationary mathematical physics problems consists in setting some initial conditions (see, for instance, (1.48)). We assign to *evolutionary inverse problems* inverse problems in which initial conditions (lacking in formulation of the problem as a direct problem) need to be identified.

As applied to the direct problem (1.46)–(1.48), a simplest evolutionary inverse problem can be formulated as follows. Initial conditions (1.48) are not specified; instead, the solution at the end time $t = T$ is known:

$$u(x, T) = u_T(x), \qquad 0 < x < l. \tag{1.52}$$

It is required to find the solution of equation (1.46) at preceding times (*retrospective inverse problem*).

One can pose an inverse problem in which it is required to identify the initial state using additional information about the solution at internal points (additional condition of type (1.50)).

1.5 Exercises

Exercise 1.1 On the function set $v(0) = 0$, $v(l) = 0$, we define the operator

$$\mathcal{A}u = -\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l.$$

Establish positive definiteness of $\mathcal{A}$, i.e., derive the estimate $(\mathcal{A}v, v) \geq \delta(v, v)$, where $\delta > 0$.

Exercise 1.2 Prove the *Friedrichs inequality*, which states that

$$\int_\Omega u^2(\boldsymbol{x})\,d\boldsymbol{x} \;\leq\; M_0 \sum_{\alpha=1}^{2} \int_\Omega \Big(\frac{\partial u}{\partial x_\alpha}\Big)^2 d\boldsymbol{x},$$

provided that $u(\boldsymbol{x}) = 0$, $\boldsymbol{x} \in \partial\Omega$.

Exercise 1.3 Prove the Gronwall lemma (Lemma 1.1).

Exercise 1.4 Show that for the solution of problem (1.18)–(1.20) the following a priori estimate holds:

$$\|u(t)\|^2 \;\leq\; \exp(t)\Big(\|u_0\|^2 + \int_0^t \exp(-\theta)\|f(\theta)\|^2\Big)d\theta.$$

Exercise 1.5 Let $\mathcal{A} \geq \delta E$, $\delta = \text{const} > 0$. Then, for the solution of problem (1.19), (1.20) we have

$$\|u(t)\| \;\leq\; \exp(-\delta t)\Big(\|u_0\| + \int_0^t \exp(\delta\theta)\|f(\theta)\|\Big)\,d\theta.$$

Exercise 1.6 Consider the Cauchy problem for the second-order evolution equation

$$\frac{d^2u}{dt^2} + \mathcal{A}u = f(t), \qquad 0 < t \le T,$$
$$u(0) = u_0, \qquad \frac{du}{dt}(0) = u_1,$$

with $\mathcal{A} = \mathcal{A}^* > 0$. Obtain the following estimate of solution stability with respect to initial data and the right-hand side:

$$\|u(t)\|_*^2 \le \exp(t)\Big(\|u^0\|_{\mathcal{A}}^2 + \|v^0\|^2 + \int_0^t \exp(-\theta)\|f(\theta)\|^2\, d\theta\Big),$$

where

$$\|u\|_*^2 = \Big\|\frac{du}{dt}\Big\|^2 + \|u\|_{\mathcal{A}}^2$$

and $\|v\|_{\mathcal{D}}^2 = (\mathcal{D}v, v)$ for the self-adjoint, positive operator $\mathcal{D}$.

Exercise 1.7 Prove the maximum principle (Theorem 1.5).

Exercise 1.8 Prove the maximum principle for the parabolic equation, which states that the solution of the boundary value problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l, \quad 0 < t \le T,$$
$$u(0,t) = \mu_0(t), \qquad u(l,t) = \mu_1(t), \qquad 0 < t \le T,$$
$$u(x,0) = u_0(x), \qquad 0 \le x \le l$$

assumes its maximum and minimum values either at boundary points or at the initial time, i.e.,

$$\min_{0<x<l,\ 0<t\le T}\{\mu_0(t), \mu_1(t), u_0(x)\} \le u(x,t) \le \max_{0<x<l,\ 0<t\le T}\{\mu_0(t), \mu_1(t), u_0(x)\}.$$

Exercise 1.9 Consider the boundary value problem

$$-\sum_{\alpha=1}^{2}\frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big) + q(\boldsymbol{x})u = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega,$$
$$u(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega,$$

where

$$k(\boldsymbol{x}) \ge \kappa > 0, \qquad q(\boldsymbol{x}) \ge 0, \qquad \boldsymbol{x} \in \Omega.$$

Invoking the Friedrichs inequality (Exercise 1.5.2), derive the following estimate for solution stability with respect to the right-hand side:

$$\|u\| \le M_1\|f\|, \qquad \|u\|^2 = \int_\Omega u^2(\boldsymbol{x})\, d\boldsymbol{x}.$$

Exercise 1.10 Taken the problem

$$-\sum_{\alpha=1}^{2}\frac{\partial^2 u}{\partial x_\alpha^2}=0,$$

$$u(0,x_2)=0,\qquad u(l,x_2)=0,$$

$$u(x_1,0)=u_0(x_1),\qquad \frac{\partial u}{\partial x_2}(x_1,0)=u_1(x_1)$$

as an example, show that the Cauchy problem for elliptic equations is ill-posed (example by J. Hadamard).

Exercise 1.11 Examine whether the boundary inverse problem for the parabolic equation

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2},\qquad 0<x<l,\quad 0<t\le T,$$

$$u(0,t)=\mu_0(t),\qquad \frac{\partial u}{\partial x}(0,t)=\mu_1(t),\qquad 0<t\le T,$$

$$u(x,0)=u_0(x),\qquad 0\le x\le l$$

is a well- or ill-posed problem.

Exercise 1.12 Prove that for any solution of the equation

$$\frac{d^2u}{dt^2}-\mathcal{A}u=0,\qquad 0\le t\le T$$

with a self-adjoint operator $\mathcal{A}$ there holds the estimate

$$\|u(t)\|^2\ \le\ \exp(2t(T-t))(\|u(T)\|^2+\chi)^{t/T}(\|u(0)\|^2+\chi)^{1-t/T}-\chi,$$

$$\chi=\frac{1}{2}\Big((\mathcal{A}u(0),u(0))-\Big(\frac{\partial u}{\partial t}(0),\frac{\partial u}{\partial t}(0)\Big)\Big),$$

which simultaneously proves that the Cauchy problem for this equation is a conditionally well-posed problem in the class of bounded solutions.

2 Boundary value problems for ordinary differential equations

We start discussing the matter of numerical solution of mathematical physics problems with the boundary value problem for the second-order ordinary differential equation. Various approaches are used in approximation of the differential problem, primary attention being paid to finite-difference approximations. Based on an estimate of stability of the finite-difference solution with respect to the right-hand side and boundary conditions, we examine the convergence of approximate solution to the exact solution. In solving finite-difference problems arising on discretization of one-dimensional problems, direct linear algebra methods are used. We present a FORTRAN 77 program that solves boundary value problems for the second-order ordinary differential equation and give computation data obtained for several model problems.

2.1 Finite-difference problem

Below, the approaches to the approximation of boundary value mathematical physics problems are illustrated with the example of a boundary value problem for the second-order ordinary differential equation.

2.1.1 Model differential problem

As a basic equation, consider the second-order ordinary differential equation

$$-\frac{d}{dx}\Big(k(x)\,\frac{du}{dx}\Big) + q(x)u = f(x), \qquad 0 < x < l \tag{2.1}$$

with variable coefficients

$$k(x) \geq \kappa > 0, \qquad q(x) \geq 0.$$

Elliptic equations of second order, prototyped by equation (2.1), simulate many physico-mechanical processes.

For the unknown function $u(x)$ to be uniquely found, equation (2.1) must be supplemented with two boundary conditions given at the end points of the segment $[0, l]$. Here, the function $u(x)$ (first-kind boundary condition), the flux $w(x) = -k(x)\,\frac{du}{dx}(x)$ (second-kind boundary condition), or a linear combination of the above conditions (third-kind boundary condition) can be considered:

$$u(0) = \mu_1, \qquad u(l) = \mu_2, \tag{2.2}$$

$$-k(0)\,\frac{du}{dx}(0) = \mu_1, \qquad k(l)\,\frac{du}{dx}(l) = \mu_2, \tag{2.3}$$

$$-k(0)\,\frac{du}{dx}(0) + \sigma_1 u(0) = \mu_1, \qquad k(l)\,\frac{du}{dx}(l) + \sigma_2 u(l) = \mu_2. \tag{2.4}$$

In the case of problems with discontinuous coefficients (contact of two media), additional conditions need to be formulated. Of such additional conditions, a simplest one (ideal-contact condition) for equation (2.1) is given by the requirement that the solution and the flux both must be continuous at the contact point $x = x^*$:

$$[u(x)] = 0, \qquad \left[k(x)\,\frac{du}{dx}\right] = 0, \qquad x = x^*.$$

Here, we use the setting

$$[g(x)] = g(x+0) - g(x-0).$$

Worthy of being considered at length here are problems with a non-self-adjoint operator; in one of such cases, for instance, we have:

$$-\frac{d}{dx}\left(k(x)\,\frac{du}{dx}\right) + v(x)\,\frac{du}{dx} + q(x)u = f(x), \qquad 0 < x < l. \tag{2.5}$$

The convection-diffusion-reaction equation (2.5) is a model one in consideration of processes dealt with in continuum mechanics.

In the description of deformed plates and shells, and also in hydrodynamic problems, mathematical models involve fourth-order elliptic equations. The consideration of such models can be started with the boundary value problem for the fourth-order ordinary differential equation. A simplest such problem is the problem for the equation

$$\frac{d^4u}{dx^4}(x) = f(x), \qquad 0 < x < l. \tag{2.6}$$

Here, two pairs of boundary conditions are considered at the end points of the segment. For instance, equation (2.6) is supplemented with first-kind conditions:

$$u(0) = \mu_1, \qquad u(l) = \mu_2, \tag{2.7}$$

$$\frac{du}{dx}(0) = \nu_1, \qquad \frac{du}{dx}(l) = \nu_2. \tag{2.8}$$

In other statements of boundary value problems for equation (2.6) boundary conditions at the end points can also involve the second and/or third derivative.

2.1.2 Difference scheme

We denote as $\bar{\omega}$ an *uniform grid*, with a step size h over the interval $\bar{\Omega} = [0, l]$:

$$\bar{\omega} = \{x \mid x = x_i = ih, \quad i = 0, 1, \ldots, N, \quad Nh = l\}.$$

Here, ω is the set of inner nodal points, and $\partial\omega$ is the set of boundary nodal points.

For a sufficiently smooth function $u(x)$, the Taylor series expansion in a vicinity of an arbitrary internal node $x = x_i$ yields:

$$u_{i\pm1} = u_i \pm h\,\frac{du}{dx}(x_i) + \frac{h^2}{2}\,\frac{d^2u}{dx^2}(x_i) \pm \frac{h^3}{6}\,\frac{d^3u}{dx^3}(x_i) + \mathcal{O}(h^4).$$

Here we use the setting $u_i = u(x_i)$. Hence, for the *left difference derivative* we have:

$$u_{\bar{x}} \equiv \frac{u_i - u_{i-1}}{h} = \frac{du}{dx}(x_i) - \frac{h}{2}\,\frac{d^2u}{dx^2}(x_i) + \mathcal{O}(h^2). \tag{2.9}$$

The subscript i is omitted here. In this way, the left difference derivative $u_{\bar{x}}$ approximates the first derivative du/dx accurate to $\mathcal{O}(h)$ at each of the internal nodes if $u(x) \in C^{(2)}(\Omega)$.

In a similar manner, for the *right difference derivative* we obtain:

$$u_x \equiv \frac{u_{i+1} - u_i}{h} = \frac{du}{dx}(x_i) + \frac{h}{2}\,\frac{d^2u}{dx^2}(x_i) + \mathcal{O}(h^2). \tag{2.10}$$

With a three-point approximation pattern (involving the nodes x_{i-1}, x_i, and x_{i+1}), one can use the *central difference derivative*

$$u_{\mathring{x}} \equiv \frac{u_{i+1} - u_{i-1}}{2h} = \frac{du}{dx}(x_i) + \frac{h^2}{3}\,\frac{d^3u}{dx^3}(x_i) + \mathcal{O}(h^3) \tag{2.11}$$

that approximates the derivative du/dx accurate to the second order if $u(x) \in C^{(3)}(\Omega)$.

For the second derivative d^2u/dx^2, similar manipulations yield:

$$u_{\bar{x}x} = \frac{u_x - u_{\bar{x}}}{h} = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}.$$

The latter difference operator approximates the second derivative accurate to the second order at the node $x = x_i$ if $u(x) \in C^{(4)}(\Omega)$.

Difference schemes for problem (2.1), (2.2) with sufficiently smooth coefficients can be constructed based on immediate replacement of differential operators with their difference analogues.

Dwell now at greater length on the approximation of the one-dimensional operator

$$\mathcal{A}u = -\frac{d}{dx}\Big(k(x)\,\frac{du}{dx}\Big) + q(x)u. \tag{2.12}$$

Consider the difference expression

$$(au_{\bar{x}})_x = \frac{a_{i+1}}{h}\,u_x - \frac{a_i}{h}\,u_{\bar{x}}.$$

Taking into account the representations (2.9), (2.10) for the local approximation inaccuracy of first derivatives with directed differences, we obtain:

$$(au_{\bar{x}})_x = \frac{a_{i+1} - a_i}{h} \frac{du}{dx}(x_i) + \frac{a_{i+1} + a_i}{2} \frac{d^2u}{dx^2}(x_i) + \frac{a_{i+1} - a_i}{6} h \frac{d^3u}{dx^3}(x_i) + \mathcal{O}(h^2). \quad (2.13)$$

To find the coefficients a_i, compare (2.13) with the differential expression

$$\frac{d}{dx}\Big(k(x) \frac{du}{dx}\Big) = \frac{dk}{dx} \frac{du}{dx} + k(x) \frac{d^2u}{dx^2}.$$

It seems reasonable to choose the coefficients a_i so that we have

$$\frac{a_{i+1} - a_i}{h} = \frac{dk}{dx}(x_i) + \mathcal{O}(h^2), \quad (2.14)$$

$$\frac{a_{i+1} + a_i}{2} = k(x_i) + \mathcal{O}(h^2). \quad (2.15)$$

In this case, the difference operator

$$Ay = -(ay_{\bar{x}})_x + cy, \qquad x \in \omega \quad (2.16)$$

with, for instance, $c(x) = q(x)$, $x \in \omega$, approximates the difference operator (2.12) accurate to $\mathcal{O}(h^2)$.

In particular, conditions (2.15) and (2.16) are satisfied with the following formulas for the coefficients a_i:

$$\begin{aligned} a_i &= k_{i-1/2} = k(x_i - 0.5h), \\ a_i &= \frac{k_{i-1} + k_i}{2}, \\ a_i &= 2\Big(\frac{1}{k_{i-1}} + \frac{1}{k_i}\Big)^{-1}. \end{aligned} \quad (2.17)$$

Alternative (other than (2.16) and (2.17)) possibilities in constructing the difference operator A will be outlined below.

To the differential problem (2.1), (2.2), we put into correspondence the difference problem

$$-(ay_{\bar{x}})_x + cy = \varphi, \qquad x \in \omega, \quad (2.18)$$

$$y_0 = \mu_1, \qquad y_N = \mu_2, \quad (2.19)$$

with $c(x) = q(x)$, $\varphi(x) = f(x)$, $x \in \omega$, for instance.

Boundary value mathematical physics problems can be conveniently considered with homogeneous boundary conditions. The same in full measure applies to finite-difference problems. The transition from inhomogeneous to homogeneous boundary

conditions itself is not always obvious in differential problems. For finite-difference problems, the situation is simpler in a sense: inhomogeneous boundary conditions can be included into the right-hand side of the finite-difference equation at near-boundary nodes. By way of example, consider the finite-difference problem (2.18), (2.19).

We consider the set of mesh functions vanishing at boundary nodes, i.e. such that $y_0 = 0$, $y_N = 0$. Hence, we have to do with a mesh function $y(x)$ that approximates the function $u(x)$ only at internal nodes of the calculation grid. For $x \in \omega$, instead of the difference problem (2.18), (2.19), we use the operator equation

$$Ay = \varphi, \qquad x \in \omega. \tag{2.20}$$

At near-boundary nodes, we use approximations of type

$$-\frac{1}{h}\Big(a_2\,\frac{y_2 - y_1}{h} - a_1\,\frac{y_1 - \mu_1}{h}\Big) + c_1 y_1 = f_1.$$

Hence,

$$(Ay)_1 = \varphi_1,$$

where

$$\varphi_1 = f_1 + \frac{a_1\mu_1}{h^2}.$$

Thus, the difference problem (2.20) with the operator A defined by (2.16) and acting on the set of mesh functions vanishing at $\partial\omega$ is put into correspondence to the differential problem (2.1), (2.2). Here, the right-hand side of (2.20),

$$\varphi(x) = \begin{cases} f_1 + \dfrac{a_1\mu_1}{h^2}, & x = x_1, \\ f(x), & x = x_i, \quad i = 2, 3, \ldots, N-2, \\ f_{N-1} + \dfrac{a_{N-1}\mu_2}{h^2}, & x = x_{N-1}, \end{cases}$$

looks unusual only at near-boundary nodes.

2.1.3 Finite element method schemes

Stationary mathematical physics problems can be discretized using the *finite element method* (FEM). For the model one-dimensional equation (2.1) with the homogeneous boundary conditions

$$u(0) = 0, \qquad u(l) = 0, \tag{2.21}$$

we construct a finite element scheme based on the Galerkin method. Using simplest piecewise linear elements, we represent the approximate solution as

$$y(x) = \sum_{i=1}^{N-1} y_i w_i(x), \tag{2.22}$$

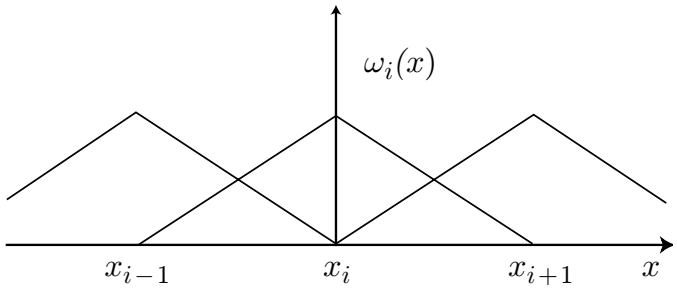

Figure 2.1 Piecewise linear trial functions

where the *trial functions* $w_i(x)$ have the form (see Figure 2.1)

$$w_i(x) = \begin{cases} 0, & x < x_{i-1}, \\ \dfrac{x - x_{i-1}}{h}, & x_{i-1} \le x \le x_i, \\ \dfrac{x_{i+1} - x}{h}, & x_i \le x \le x_{i+1}, \\ 0, & x > x_{i+1}. \end{cases}$$

The expansion coefficient can be found from a system of linear equations obtained by multiplying the initial equation (2.1) by a *verifying function* $w_i(x)$ and integrating the resulting equation over the entire domain. In view of the finiteness of the trial functions, we obtain:

$$\int_{x_{i-1}}^{x_{i+1}} k(x)\,\frac{dy}{dx}\,\frac{dw_i}{dx}\,dx + \int_{x_{i-1}}^{x_{i+1}} q(x)y(x)w_i(x)\,dx = \int_{x_{i-1}}^{x_{i+1}} f(x)w_i(x)\,dx.$$

Substitution of (2.22) yields the three-point difference equation (2.20). For A, we obtain representation (2.16) with a mesh function a_i that depends not only on $k(x)$, but also on $q(x)$:

$$a_i = \frac{1}{h}\int_{x_{i-1}}^{x_i} k(x)\,dx - \frac{1}{h}\int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})(x_i - x)\,dx. \tag{2.23}$$

For smooth coefficients $k(x)$, application of simplest quadrature formulas yields (2.17).

For the right-hand side and for the coefficient c_i in (2.16) we obtain:

$$c_i = \frac{1}{h^2}\Big(\int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})\,dx + \int_{x_{i-1}}^{x_i} q(x)(x_{i+1} - x)\,dx\Big), \tag{2.24}$$

$$\varphi_i = \frac{1}{h^2}\Big(\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1})\,dx + \int_{x_{i-1}}^{x_i} f(x)(x_{i+1} - x)\,dx\Big). \tag{2.25}$$

Even in the simplest case of constant coefficients $k(x)$ and $q(x)$, we arrive at an unusual approximation of the lowest term in A.

Difference schemes with basis functions chosen in the form of piecewise polynomials of higher degree (quadratic, cubic, etc.) can be constructed in a similar manner. In the treatment of convection-diffusion problems (see equation (2.5)), FEM schemes based on the Petrov–Galerkin method have gained acceptance, in which trial and verifying functions differ from each other. Along this line, in particular, finite element analogues of ordinary difference schemes with directed differences can be constructed.

2.1.4 Balance method

Normally, differential equations reflect one or another law of conservation for elementary volumes (integral conservation laws) on contraction of the volumes to zero. In fact, construction of a discrete problem implies a reverse transition from a differential to an integral model. One can reasonably demand that, upon such a transition, the conservation laws remained fulfilled. Difference schemes expressing conservation laws on a grid are called *conservative difference schemes*.

Construction of conservative finite-difference schemes can be reasonably started from conservation laws (balances) for individual meshes of the difference scheme. This construction method for conservative difference schemes received the name *integro-interpolation method* (*balance method*). This approach is also known as *finite-volume method*. The integro-interpolation method was proposed by A. N. Tikhonov and A. A. Samarskii in the early 50ths.

Consider the integro-interpolation method as applied to construction of a difference scheme for the model one-dimensional problem (2.1), (2.2). Let us consider $Q(x) = -k(x)du/dx$. We choose the control volumes as the segments $x_{i-1/2} \le x \le x_{i+1/2}$, where $x_{i-1/2} = (i - 1/2)h$. Integration of (2.1) over the control volume $x_{i-1/2} \le x \le x_{i+1/2}$ yields:

$$Q_{i+1/2} - Q_{i-1/2} + \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)\,dx = \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx. \qquad (2.26)$$

The balance relation (2.26) reflects a conservation law for the segment $x_{i-1/2} \le x \le x_{i+1/2}$. The quantity $Q_{i\pm1/2}$ is the flux through the section $x_{i\pm1/2}$. Unbalance between these fluxes is caused by distributed sources (right-hand side of (2.26)) and by additional sources (integral in the left-hand side of the equation).

To derive a difference equation from the balance relation (2.26), one has to use some completion of mesh functions. We seek the solution itself at integer nodes ($y(x)$, $x = x_i$), and fluxes, at half-integer nodes ($Q(x)$, $x = x_{i+1/2}$). We express the fluxes at half-integer nodes in terms of the values of $u(x)$ at nodal points. To this end, we

integrate the relation $\dfrac{du}{dx} = -\dfrac{1}{k(x)}\,Q(x)$ over the segment $x_{i-1} \le x \le x_i$:

$$u_{i-1} - u_i = \int_{x_{i-1}}^{x_i} \frac{Q(x)}{k(x)}\,dx \approx Q_{i-1/2} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}.$$

We denote

$$a_i = \Big(\frac{1}{h}\int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\Big)^{-1}. \tag{2.27}$$

Then, we obtain

$$Q_{i-1/2} \approx -a_i u_{\bar{x},i}, \qquad Q_{i+1/2} \approx -a_{i+1} u_{x,i}.$$

For the right-hand side, we have:

$$\varphi_i = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx.$$

In view of the chosen completion, we put

$$\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)\,dx \approx c_i u_i,$$

where

$$c_i = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,dx.$$

Finally, in calculating the mesh function a_i by formulas (2.27) (compare (2.17)) we arrive at the difference scheme (2.16).

The above conservative schemes belong to the class of homogeneous difference schemes (here, the coefficients of the difference equation are calculated at all grid nodes by one and the same formulas). In much the same manner to the balance method, difference schemes for more general problems can be constructed, including problems with boundary conditions (2.3), (2.4), problems specified on non-uniform grids, multi-dimensional problems, etc. That is why the integro-interpolation method is distinguished as a generic one in construction of discrete analogues of mathematical physics problems.

2.2 Convergence of difference schemes

The key point in consideration of discrete analogues of boundary value problems is examination of closeness of approximate solution to the exact solution. Based on a priori estimates of solution stability for the difference problem, here we give an estimate for the rate of convergence in mesh Hilbert spaces in numerical solution of the model boundary value problem (2.1), (2.2).

2.2.1 Difference identities

Recall some notions in use in the theory of difference schemes. We assume that a uniform grid

$$\bar{\omega} = \{x \mid x = x_i = ih, \quad i = 0, 1, \ldots, N, \quad Nh = l\},$$

is introduced over the segment $[0, l]$. Here, ω is the set of internal nodes:

$$\omega = \{x \mid x = x_i = ih, \quad i = 1, 2, \ldots, N-1, \quad Nh = l\}.$$

For other parts of $\bar{\omega}$ we use the following settings:

$$\begin{aligned} \omega^+ &= \{x \mid x = x_i = ih, \quad i = 1, 2, \ldots, N, \quad Nh = l\}, \\ \omega^- &= \{x \mid x = x_i = ih, \quad i = 0, 1, \ldots, N-1, \quad Nh = l\}. \end{aligned}$$

On the set of nodes ω and $\omega^{\pm}$ we define the scalar products

$$(y, w) \equiv \sum_{x \in \omega} y(x) w(x) h,$$

$$(y, w)^{\pm} \equiv \sum_{x \in \omega^{\pm}} y(x) w(x) h.$$

Let us also give difference analogues to the differentiation formula for the product of two functions and to the integration-by-parts formula. With the previously introduced definitions of the operators of the right and left difference derivatives, one can immediately check the validity of the following equalities:

$$\begin{aligned} (yw)_{\bar{x},i} &= y_{i-1} w_{\bar{x},i} + w_i y_{\bar{x},i} = y_i w_{\bar{x},i} + w_{i-1} y_{\bar{x},i}, \\ (yw)_{x,i} &= y_{i+1} w_{x,i} + w_i y_{x,i} = y_i w_{x,i} + w_{i+1} y_{x,i}. \end{aligned} \tag{2.28}$$

These equalities give the difference analogue of the differentiation formula

$$\frac{d}{dx}(yw) = y \frac{dw}{dx} + w \frac{dy}{dx}.$$

The analogues for the integration-by-parts formula

$$\int_a^b \frac{dy}{dx} w \, dx = y(b)w(b) - y(a)w(a) - \int_a^b y \frac{dw}{dx} \, dx$$

are the finite-difference identities

$$\begin{aligned} (y_x, w) &= -(y, w_{\bar{x}})^+ + y_N w_N - y_1 w_0, \\ (y_{\bar{x}}, w) &= -(y, w_x)^- + y_{N-1} w_N - y_0 w_0. \end{aligned} \tag{2.29}$$

Replacing y_i with $a_i y_{\bar{x},i}$ in (2.29), we obtain the *first Green difference formula*:

$$((ay_{\bar{x}})_x, w) = -(ay_{\bar{x}}, w_{\bar{x}})^+ + a_N y_{\bar{x},N} w_N - a_1 y_{x,0} w_0. \tag{2.30}$$

The *second Green difference formula* is

$$((ay_{\bar{x}})_x, w) - (y, (aw_{\bar{x}})_x) = a_N(y_{\bar{x},N} w_N - y_N w_{\bar{x},N}) - a_1(y_{x,0} w_0 - y_0 w_{x,0}). \tag{2.31}$$

Formulas (2.30) and (2.31) take a simpler form,

$$((ay_{\bar{x}})_x, w) = -(ay_{\bar{x}}, w_{\bar{x}})^+, \tag{2.32}$$

$$((ay_{\bar{x}})_x, w) = (y, (aw_{\bar{x}})_x), \tag{2.33}$$

for mesh functions $y(x)$ and $w(x)$ vanishing at $x = 0$ and $x = l$ (on $\partial\omega$).

2.2.2 Properties of the operator A

For the model problem (2.1), (2.2), we have constructed the difference scheme (2.20), in which the difference operator A is defined on the set of mesh functions $y(x)$, $x \in \bar{\omega}$ vanishing on $\partial\omega$ by (see (2.16))

$$Ay = -(ay_{\bar{x}})_x + cy, \qquad x \in \omega, \tag{2.34}$$

where, for instance, $a(x) = k(x - 0.5h)$, $c(x) = q(x)$, $x \in \omega^+$. We introduce the norm in the mesh Hilbert space H by the relation $\|y\| = (y, y)^{1/2}$.

Like in the differential case, the difference operator A in H is a self-adjoint operator:

$$A = A^*. \tag{2.35}$$

The equality $(Ay, w) = (y, Aw)$ readily follows from (2.33).

With the usual constraints $k(x) \geq \kappa > 0$, $q(x) \geq 0$, the following lower estimate for A holds:

$$A \geq \kappa\lambda_0 E. \tag{2.36}$$

Here λ_0 is the lowest eigenvalue of the difference operator of the second derivative. In the case of a uniform grid, we have:

$$\lambda_0 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l} \geq \frac{8}{l^2}.$$

Let us obtain such a lower estimate for A based on the following *difference Friedrichs inequality*.

Lemma 2.1 *For all mesh functions vanishing on $\partial\omega$, the following inequality holds:*

$$\|y\|^2 \leq M_0(\|y_{\bar{x}}\|^+)^2, \qquad \|w\|^+ \equiv ((w, w)^+)^{1/2}, \qquad M_0 = l^2/8. \tag{2.37}$$

Proof. For such a mesh function $y_i = y(x_i)$ we have:

$$y_i = \sum_{k=1}^{i} y_{\bar{x},i} h, \tag{2.38}$$

$$y_i = -\sum_{k=i+1}^{N} y_{\bar{x},i} h. \tag{2.39}$$

To estimate the right-hand sides of (2.38) and (2.39), we use the inequality

$$\left| \sum_k a_k b_k \right|^2 \le \sum_k a_k^2 \sum_k b_k^2.$$

We put $a_k = y_{\bar{x},i} h^{1/2}$ and $b_k = h^{1/2}$; then, from (2.38) and (2.39) we obtain:

$$y_i^2 \le x_i \sum_{k=1}^{i} (y_{\bar{x},k})^2 h, \tag{2.40}$$

$$y_i^2 \le (l - x_i) \sum_{k=i+1}^{N} (y_{\bar{x},k})^2 h. \tag{2.41}$$

Let us consider $n = (N-1)/2$ if, for instance, N is an odd number. The case in which N is even will be considered separately. From (2.40) and (2.41) we have:

$$y_i^2 \le x_i \sum_{k=1}^{n} (y_{\bar{x},k})^2 h, \qquad 1 \le i \le n,$$

$$y_i^2 \le (l - x_i) \sum_{k=n+1}^{N} (y_{\bar{x},k})^2 h, \qquad n+1 < i \le N.$$

We multiply each of the above inequalities by h and add the inequalities together:

$$\sum_{i=1}^{N} y_i^2 h \le \sum_{i=1}^{n} x_i h \sum_{k=1}^{n} (y_{\bar{x},k})^2 h + \sum_{i=n+1}^{N} (l - x_i) \sum_{k=n+1}^{N} (y_{\bar{x},k})^2 h. \tag{2.42}$$

Taking into account that

$$\sum_{i=1}^{n} x_i h = \frac{x_n x_{n+1}}{2} < \frac{l^2}{8}, \qquad \sum_{i=n+1}^{N} (l - x_i) h < \frac{l^2}{8},$$

from (2.42) we obtain the estimate (2.37). □

The above proof can be extended, with no substantial changes, to the case of a non-uniform grid. The same note applies to multi-dimensional problems provided that rectangular or general irregular grids are used.

For the difference operator (2.34) with bounded coefficients $k(x)$ and $q(x)$ as for an operator in the finite-dimensional space H, the following upper estimate is useful. Note that it is not valid for the differential operator $\mathcal{A}$ (unbounded operator).

Lemma 2.2 *The estimate*

$$(Ay, y) \le M_1(y, y) \qquad (A \le M_1 E), \tag{2.43}$$

with the constant

$$M_1 = \frac{4}{h^2} \max_{1 \le i \le N-1} \frac{a_i + a_{i+1}}{2} + \max_{1 \le i \le N-1} c_i$$

holds.

Proof. Indeed, we have:

$$(Ay, y) = (a(y_{\bar{x}})^2, 1) + (cy, y) = \sum_{i=1}^{N} \frac{a_i}{h} (y_i - y_{i-1})^2 + \sum_{i=1}^{N} c_i y_i^2 h.$$

We use the inequality $(a + b)^2 \le 2a^2 + 2b^2$, the condition for positiveness of the mesh functions $a(x)$, $c(x)$, $x \in \omega$, and the conditions $y(x) = 0$, $x \notin \omega$; then, we obtain:

$$(Ay, y) \le \sum_{i=1}^{N} \frac{2}{h} (a_i + a_{i+1}) y_i^2 + \sum_{i=1}^{N} c_i y_i^2 h.$$

This yield the desired estimate (2.43). □

2.2.3 Accuracy of difference schemes

To approximately solve the problem (2.1), (2.2), we use the difference scheme

$$Ay = \varphi(x), \qquad x \in \omega, \tag{2.44}$$

$$y_0 = \mu_1, \qquad y_N = \mu_2. \tag{2.45}$$

To examine the accuracy of (2.44), consider the problem for the inaccuracy

$$z(x) = y(x) - u(x), \qquad x \in \bar{\omega}.$$

For this inaccuracy, from (2.44) and (2.45) we obtain the difference problem

$$Az = \psi(x), \qquad x \in \omega, \tag{2.46}$$

which, in view of the exact approximation of boundary conditions (2.2), is considered on the set of mesh functions $z(x)$ vanishing at the boundary. In (2.46) $\psi(x)$ is the approximation inaccuracy:

$$\psi(x) = \varphi(x) - Au, \qquad x \in \omega. \tag{2.47}$$

Considering the case of sufficiently smooth coefficients and a sufficiently smooth right-hand side of (2.1), in using (2.34) for the approximation inaccuracy we obtain:

$$\psi(x) = \mathcal{O}(h^2), \qquad x \in \omega. \tag{2.48}$$

Let us formulate now a simplest statement concerning the accuracy of the difference scheme (2.44), (2.45) used to solve the model one-dimensional problem (2.1), (2.2).

Theorem 2.3 *For the inaccuracy of the difference solution found by formulas* (2.44)*,* (2.45)*, the following a priori estimate holds:*

$$\|z_{\bar{x}}\|^{+} \le \frac{M_0^{1/2}}{\kappa} \|\psi\|. \tag{2.49}$$

Proof. We scalarwise multiply the equation (2.46) for the inaccuracy by $z(x)$, $x \in \omega$ to find that

$$(Az, z) = (\psi, z). \tag{2.50}$$

For the left-hand side of this equality, we have

$$(Az, z) \ge \kappa (\|z_{\bar{x}}\|^{+})^2,$$

whereas the right-hand side can be estimated from below using the Friedrichs inequality (2.37):

$$(\psi, z) \le \|\psi\| \|z\| \le M_0^{1/2} \|\psi\| \|z_{\bar{x}}\|^{+}.$$

Substitution into (2.50) yields the desired estimate (2.49). □

From the expression (2.48) for the local approximation inaccuracy and from the a priori estimate (2.49), convergence of the difference solution to the exact solution with the second-order accuracy follows.

2.3 Solution of the difference problem

On discretization of one-dimensional convection-diffusion problems, we arrive at three-point difference problems. The difference solutions can be found using traditional direct linear algebra methods. Below, the sweep method (Thomas algorithm) is outlined, presenting, as everybody knows, the classical Gauss algorithm for matrices with special banded structure.

2.3.1 The sweep method

Consider, as a model problem, the problem (2.1), (2.2). On a uniform grid $\bar{\omega}$ with a step size h, we put into correspondence to this differential problem for equation (2.1) the difference problem (2.34), (2.44), (2.45). We write the difference problem as follows:

$$-\alpha_i y_{i-1} + \gamma_i y_i - \beta_i y_{i+1} = \varphi_i, \qquad i = 1, 2, \ldots, N-1, \tag{2.51}$$

$$y_0 = \mu_1, \qquad y_N = \mu_2. \tag{2.52}$$

From (2.34) and (2.44) we obtain the following expressions for the coefficients:

$$\alpha_i = \frac{1}{h^2} k_{i-1/2}, \qquad \beta_i = \frac{1}{h^2} k_{i+1/2},$$

$$\gamma_i = \frac{1}{h^2} (k_{i-1/2} + k_{i+1/2}) + c_i, \qquad i = 1, 2, \ldots, N-1.$$

To find the solution of (2.51), (2.52) by the *sweep method (Thomas algorithm)*, we use the representation

$$y_i = \xi_{i+1} y_{i+1} + \vartheta_{i+1}, \qquad i = 0, 1, \ldots, N-1 \tag{2.53}$$

with still undetermined coefficients ξ_i, ϑ_i. Substitution of $y_{i-1} = \xi_i y_i + \vartheta_i$ into (2.51) yields:

$$(\gamma_i - \alpha_i \xi_i) y_i - \beta_i y_{i+1} = \varphi_i + \alpha_i \vartheta_i, \qquad i = 1, 2, \ldots, N-1.$$

Now, we use representation (2.53) and obtain

$$((\gamma_i - \alpha_i \xi_i)\xi_{i+1} - \beta_i) y_{i+1} = \varphi_i + \alpha_i \vartheta_i + (\gamma_i - \alpha_i \xi_i)\vartheta_{i+1},$$
$$i = 1, 2, \ldots, N-1.$$

This equality is fulfilled with arbitrary y_{i+1} if

$$(\gamma_i - \alpha_i \xi_i)\xi_{i+1} - \beta_i = 0,$$

$$\varphi_i + \alpha_i \vartheta_i + (\gamma_i - \alpha_i \xi_i)\vartheta_{i+1} = 0, \qquad i = 1, 2, \ldots, N-1.$$

From here, we obtain the following recurrence formulas for the sweep coefficients ξ_i, ϑ_i:

$$\xi_{i+1} = \frac{\beta_i}{\gamma_i - \alpha_i \xi_i}, \qquad i = 1, 2, \ldots, N-1, \tag{2.54}$$

$$\vartheta_{i+1} = \frac{\varphi_i + \alpha_i \vartheta_i}{\gamma_i - \alpha_i \xi_i}, \qquad i = 1, 2, \ldots, N-1. \tag{2.55}$$

To start the calculations, we write the boundary condition (2.52) at the left end in the form of (2.53), i.e. as $y_0 = \xi_1 y_1 + \vartheta_1$, so that

$$\xi_1 = 0, \qquad \vartheta_1 = \mu_1. \tag{2.56}$$

After the sweep coefficients are calculated by the recurrence formulas (2.54)–(2.56), we can found, using formula (2.53) and the second boundary condition (2.52), the solution itself.

2.3.2 Correctness of the sweep algorithm

Let us formulate conditions sufficient for the use of the above sweep-method formulas. We do not consider here the whole scope of problems that arise in substantiation of the sweep method. Here, we restrict ourselves just to the matter of correctness of the method, which is equivalent in the case of interest to the requirement of nonzero denominator in (2.54), (2.55).

Lemma 2.4 *Let the following conditions be fulfilled for system* (2.51), (2.52)*:*

$$|\alpha_i| > 0, \qquad |\beta_i| > 0, \qquad i = 1, 2, \ldots, N-1, \tag{2.57}$$

$$|\gamma_i| \geq |\alpha_i| + |\beta_i|, \qquad i = 1, 2, \ldots, N-1. \tag{2.58}$$

Then, algorithm (2.53)–(2.56) *is correct, i.e. in the formulas* (2.54), (2.55) *we have* $\gamma_i - \alpha_i \xi_i \neq 0$.

Proof. We are going to show that

$$|\xi_i| \leq 1, \qquad i = 1, 2, \ldots, N-1. \tag{2.59}$$

In view of (2.57), (2.58), under such constraints on the sweep coefficients we have:

$$|\gamma_i - \alpha_i \xi_i| \;\geq\; ||\gamma_i| - |\alpha_i|\,|\xi_i|| \;\geq\; ||\gamma_i| - |\alpha_i|| \;\geq\; |\beta_i| > 0.$$

We will prove (2.59) by induction. For $i = 1$, inequality (2.59) holds. Suppose that inequality (2.59) holds for i; then, by (2.54) we obtain for $i + 1$:

$$|\xi_{i+1}| = \frac{|\beta_i|}{|\gamma_i - \alpha_i \xi_i|} \leq \frac{|\beta_i|}{|\beta_i|} \leq 1.$$

Provided that inequality (2.59) takes place, we also have: $\gamma_i - \alpha_i \xi_i \neq 0$. □

For our model problem (2.1), (2.2), in using the difference scheme (2.51), (2.52) conditions (2.57), (2.58) for sweep correctness will be fulfilled if

$$\alpha_i > 0, \qquad \beta_i > 0, \qquad \gamma_i \geq \alpha_i + \beta_i, \qquad i = 1, 2, \ldots, N-1.$$

These conditions fulfilled, the maximum principle holds for the difference solution of problem (2.51), (2.52), i.e., the difference scheme is monotone. This matter will be discussed in more detail below.

2.3.3 The Gauss method

There exist a variety of versions of the sweep method more accurately taking into account particular specific features of particular problems. For instance, one can adhere the classical direct solution method for systems of linear algebraic equations, the *Gauss method* (LU-decomposition, compact scheme of the Gauss method).

In matrix form, the difference problem (2.51), (2.52) looks as

$$Ay = \varphi, \qquad x \in \omega \tag{2.60}$$

with a properly defined right-hand side (see (2.20)). In view of the established properties of the operator A, the matrix A is a positive-definite matrix. In the latter case, the major (corner) minors of the matrix are all positive and, hence, an LU-decomposition $A = LU$ takes place.

In the notation of (2.60), for the matrix A we have the following representation:

$$A = \begin{bmatrix} \gamma_1 & -\beta_1 & 0 & \dots & 0 \\ -\alpha_2 & \gamma_2 & -\beta_2 & \dots & 0 \\ 0 & -\alpha_3 & \gamma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \gamma_{N-1} \end{bmatrix}.$$

In view of the banded structure of the initial matrix A, for the elements of the LU-decomposition we can conveniently put

$$L = \begin{bmatrix} \xi_2^{-1} & 0 & 0 & \dots & 0 \\ -\alpha_2 & \xi_3^{-1} & 0 & \dots & 0 \\ 0 & -\alpha_3 & \xi_4^{-1} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \xi_N^{-1} \end{bmatrix}, \quad U = \begin{bmatrix} 1 & -\xi_2\beta_1 & 0 & \dots & 0 \\ 0 & 1 & -\xi_3\beta_2 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

The diagonal elements of L are given by the following recurrence relations:

$$\xi_{i+1} = \frac{1}{\gamma_i - \beta_{i-1}\alpha_i\xi_i}, \quad i = 1, 2, \dots, N-1, \qquad \xi_1 = 0. \tag{2.61}$$

The solution of the equation $L\vartheta = \varphi$ is given by

$$\vartheta_{i+1} = \frac{\varphi_i + \alpha_i\vartheta_i}{\gamma_i - \beta_{i-1}\alpha_i\xi_i}, \quad i = 1, 2, \dots, N-1, \qquad \vartheta_1 = 0. \tag{2.62}$$

Now, from the system $Uy = \vartheta$ we obtain the solution of problem (2.60):

$$y_i = \xi_{i+1}\beta_i y_{i+1} + \vartheta_{i+1}, \quad i = 0, 1, \dots, N-1, \qquad y_N = 0. \tag{2.63}$$

Formulas (2.61)–(2.63) for the LU-decomposition stem from the above-described sweep algorithm (2.53)–(2.56) on the substitution

$$\xi_i \longmapsto \beta_{i-1}\xi_i, \qquad i = 1, 2, \ldots, N.$$

And vice versa (which is more correct), sweep formulas (2.53)–(2.56) follow from the formulas of the Gauss-method compact scheme applied to the difference problem (2.51), (2.52).

2.4 Program realization and computational examples

The matter of numerical solution of a boundary value problem for the convection-diffusion equation is considered, the problem being a model one in continuum mechanics. Two types of difference schemes are constructed, with central differences of second approximation order and directed differences of first order used to approximate the convective-transfer operator. A program that solves the model problem and computational results obtained are presented.

2.4.1 Problem statement

In the theoretical consideration of approximate solution methods for boundary value problems for ordinary differential equations, we have restricted ourselves to the case of the simplest boundary value problem (2.1), (2.2). The matter of practical implementation of the numerical methods will be illustrated below for an example of a somewhat more general problem for an equation of type (2.5).

Consider the boundary value Dirichlet problem for the *convection-diffusion equation*:

$$-\frac{d}{dx}\left(k(x)\,\frac{du}{dx}\right) + v(x)\,\frac{du}{dx} = f(x), \qquad 0 < x < l, \tag{2.64}$$

$$u(0) = \mu_1, \qquad u(l) = \mu_2 \tag{2.65}$$

with $k(x) \geq \kappa > 0$.

In many practical problems, the prevailing contribution can be either due to the diffusion term (the term with diffusivity $k(x)$ in (2.64)) or due to convective transfer (the term with velocity $v(x)$). The importance of convective transfer can be evaluated by the *Peclet number*, that arises when one nondimensionalizes the convection-diffusion equation:

$$\mathrm{Pe} = \frac{v_0 l}{k_0}. \tag{2.66}$$

Here v_0 is the characteristic velocity, and k_0 is the characteristic diffusivity. In the equation of motion for a continuum medium, an analogous parameter is the Reynolds number.

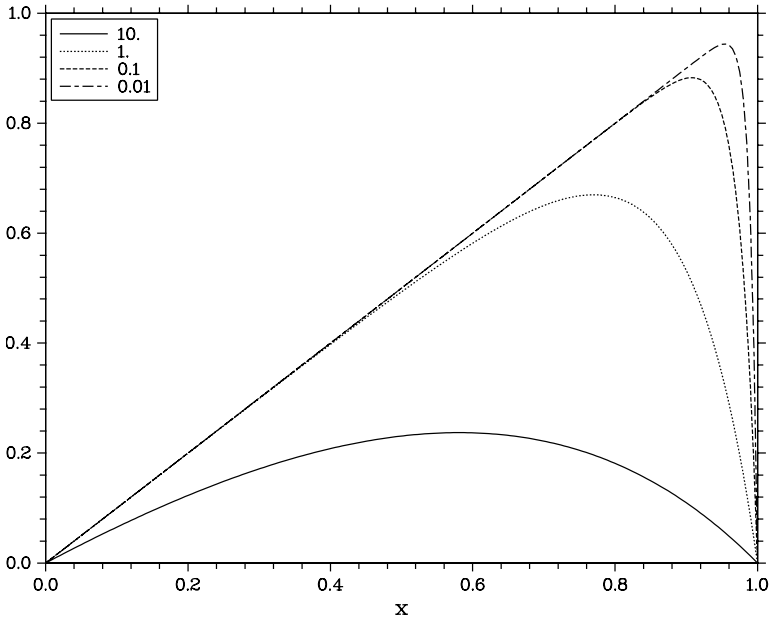

Figure 2.2 Exact solution of the convection-diffusion problem

In the case of $\mathrm{Pe} \ll 1$, we have a process with *dominating diffusion*, whereas in the case of $\mathrm{Pe} \gg 1$ *convection prevails*. In the former case, we arrive at regularly perturbed problems (small parameter Pe^{-1} at low-order derivatives), whereas in the case of strongly dominating convection, we obtain singularly perturbed problems (small parameter Pe^{-1} at hogh-order derivatives). Typical for singularly perturbed problems is the occurrence of regions where the solution displays considerable variations, such regions, for instance, being boundary layers and internal transition layers.

Computational algorithms for approximate solution of convection-diffusion problems can be understood considering the problem with constant coefficients and homogeneous boundary conditions:

$$k(x) = \kappa, \qquad v(x) = 1, \qquad f(x) = 1, \qquad \mu_1 = 0, \qquad \mu_2 = 0. \tag{2.67}$$

The exact solution of problem (2.64), (2.65), (2.66) is

$$u(x) = x - l\,\frac{\exp(x/\kappa) - 1}{\exp(l/\kappa) - 1}.$$

Specific features inherent to the problem with dominating convection (low diffusion coefficients) can be figured out considering Figure 2.2 that shows the solution of the problem with $l = 1$ and $\kappa = 1$, 0.1, and 0.01. As the diffusion coefficient decreases, near the right boundary there forms a boundary layer (region with large solution gradients).

2.4.2 Difference schemes

In consideration of one-dimensional convection-diffusion problems (2.64), (2.65), we choose using difference schemes written at the internal nodes in the form (2.51), (2.52). Consider the difference schemes (2.51), (2.52) in which

$$\alpha_i > 0, \qquad \beta_i > 0, \qquad \gamma_i > 0, \qquad i = 1, 2, \ldots, N-1. \tag{2.68}$$

Let us formulate a criterion for *monotonicity of the difference scheme* or, in other words, formulate conditions under which the difference scheme (2.51), (2.52) satisfies the *difference principle of maximum*.

Theorem 2.5 (Maximum principle) *Let in the difference scheme* (2.51), (2.52), (2.68) $\mu_1 \geq 0$, $\mu_2 \geq 0$ *and* $\varphi_i \geq 0$ *for all* $i = 1, 2, \ldots, N-1$ *(or, alternatively,* $\mu_1 \leq 0$, $\mu_2 \leq 0$ *and* $\varphi_i \leq 0$ *for* $i = 1, 2, \ldots, N-1$*). Then, provided that*

$$\gamma_i \geq \alpha_i + \beta_i, \qquad i = 1, 2, \ldots, N-1, \tag{2.69}$$

we have: $y_i \geq 0$, $i = 1, 2, \ldots, N-1$ ($y_i \leq 0$, $i = 1, 2, \ldots, N-1$).

Proof. Let us follow the line of reasoning assuming the opposite. Let conditions (2.69) be fulfilled, but the difference solution of problem (2.51) with non-negative right-hand side and non-negative boundary conditions be not non-negative at all nodes of the grid. We designate as k the grid node at which the solution assumes the least negative value. If such a value is attained at several nodes, then we choose the node where $y_{k-1} > y_k$. We write the difference equation at this node:

$$-\alpha_k y_{k-1} + \gamma_k y_k - \beta_k y_{k+1} = \varphi_k.$$

The right-hand side is non-negative, and for the left-hand side, in view of (2.68) and (2.69), we have:

$$\begin{aligned} -\alpha_k y_{k-1} + \gamma_k y_k - \beta_k y_{k+1} \\ = \alpha_k (y_k - y_{k-1}) + (\gamma_k - \alpha_k - \beta_k) y_k + \beta_k (y_k - y_{k+1}) > 0. \end{aligned}$$

The obtained contradiction shows that $y_i \geq 0$ at all nodes $i = 1, 2, \ldots, N-1$. □

To approximately solve the problem (2.64), (2.65), we use the simplest difference scheme with central-difference approximation of the convective term. At the internal nodes of the computational grid, we approximate the differential equation (2.64), accurate to the second order, with the difference equation

$$-(a y_{\bar{x}})_x + b y_{\overset{\circ}{x}} = \varphi, \qquad x \in \omega, \tag{2.70}$$

where, for instance, $a(x) = k(x - 0.5h)$ and $b(x) = v(x)$ in problems with smooth coefficients.

We write the difference scheme (2.52), (2.70) in the form (2.51), (2.52) with

$$\alpha_i = \frac{v_i}{2h} + \frac{1}{h^2}\,k_{i-1/2}, \qquad \beta_i = -\frac{v_i}{2h} + \frac{1}{h^2}\,k_{i+1/2},$$
$$\gamma_i = \frac{1}{h^2}\,(k_{i-1/2} + k_{i+1/2}), \qquad i = 1, 2, \ldots, N-1.$$

The sufficient conditions for monotonicity (2.68), (2.69) (conditions for the maximum principle) for the scheme with the convection term approximated with central differences will be fulfilled only in the case of sufficiently small grid steps (low convective-transfer coefficients). Disregarding the sign of velocity, we write the constraints as

$$\mathrm{Pe}_i \equiv \frac{|v_i|h}{\min\{k_{i-1/2}, k_{i+1/2}\}} < 2. \tag{2.71}$$

Here Pe_i is the *mesh Peclet number*. Hence, to obtain a monotone difference scheme, we have to use a sufficiently fine computation grid. With constraints (2.71), we may speak only of conditional monotonicity of the difference scheme (2.52), (2.70).

Absolutely monotone difference schemes for convection-diffusion problems can be constructed using first-order approximations of the convection term with directed differences. We use the settings

$$b(x) = b^+(x) + b^-(x),$$

$$b^+(x) = \frac{1}{2}\,(b(x) + |b(x)|) \geq 0, \qquad b^-(x) = \frac{1}{2}\,(b(x) - |b(x)|) \leq 0$$

for the non-positive and non-negative parts of the mesh function $b(x)$, $x \in \omega$. Instead of (2.70), we use the difference scheme

$$-(ay_{\bar{x}})_x + b^+ y_{\bar{x}} + b^- y_x = \varphi, \qquad x \in \omega. \tag{2.72}$$

For scheme (2.52), (2.72), we have representation (2.5), (2.52) with coefficients

$$\alpha_i = \frac{v_i^+}{h} + \frac{1}{h^2}\,k_{i-1/2}, \qquad \beta_i = -\frac{v_i^-}{h} + \frac{1}{h^2}\,k_{i+1/2},$$

$$\gamma_i = \frac{|v_i|}{h} + \frac{1}{h^2}\,(k_{i-1/2} + k_{i+1/2}), \qquad i = 1, 2, \ldots, N-1.$$

We see immediately that, here, sufficient conditions for monotonicity (conditions (2.68) and (2.69)) are fulfilled.

The scheme with directed differences is monotone with any grid steps. Its deficiencies are related with the fact that, unlike the scheme with central differences, this scheme has first approximation order.

2.4.3 Program

Realization of the difference schemes (2.52), (2.70) and (2.52), (2.72) implies numerical solution of a system of linear algebraic equations with a tridiagonal matrix. As it was noted previously, such problems can be solved using the sweep algorithm. This algorithm is correct under the condition of diagonal prevalence (see Lemma 2.4). For non-monotone difference schemes of type (2.52), (2.70), one has to use more general algorithms. Here we speak of non-monotone sweep, which can be considered as the Gauss method with the choice of major element in solving the tridiagonal system of linear algebraic equations.

We use the subroutine SGTSL from the well-known linear algebra package LINPACK.

```
      SUBROUTINE SGTSL (N, C, D, E, B, INFO)
C***BEGIN PROLOGUE  SGTSL
C***PURPOSE  Solve a tridiagonal linear system.
C***LIBRARY   SLATEC (LINPACK)
C***CATEGORY  D2A2A
C***TYPE      SINGLE PRECISION (SGTSL-S, DGTSL-D, CGTSL-C)
C***KEYWORDS  LINEAR ALGEBRA, LINPACK, MATRIX, SOLVE, TRIDIAGONAL
C***AUTHOR  Dongarra, J., (ANL)
C***DESCRIPTION
C
C     SGTSL given a general tridiagonal matrix and a right hand
C     side will find the solution.
C
C     On Entry
C
C        N       INTEGER
C                is the rank of the tridiagonal matrix.
C
C        C       REAL(N)
C                is the subdiagonal of the tridiagonal matrix.
C                C(2) through C(N) should contain the subdiagonal.
C                On output, C is destroyed.
C
C        D       REAL(N)
C                is the diagonal of the tridiagonal matrix.
C                On output, D is destroyed.
C
C        E       REAL(N)
C                is the superdiagonal of the tridiagonal matrix.
C                E(1) through E(N-1) should contain the superdiagonal.
C                On output, E is destroyed.
C
C        B       REAL(N)
C                is the right hand side vector.
C
C     On Return
C
C        B       is the solution vector.
C
C        INFO    INTEGER
C                = 0 normal value.
C                = K if the Kth element of the diagonal becomes
```

```
C                   exactly zero.  The subroutine returns when
C                   this is detected.
C
C***REFERENCES  J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W.
C                 Stewart, LINPACK Users' Guide, SIAM, 1979.
C***ROUTINES CALLED  (NONE)
C***REVISION HISTORY  (YYMMDD)
C   780814  DATE WRITTEN
C   890831  Modified array declarations.  (WRB)
C   890831  REVISION DATE from Version 3.2
C   891214  Prologue converted to Version 4.0 format.  (BAB)
C   900326  Removed duplicate information from DESCRIPTION section.
C           (WRB)
C   920501  Reformatted the REFERENCES section.  (WRB)
C***END PROLOGUE  SGTSL

      INTEGER N,INFO
      REAL C(*),D(*),E(*),B(*)
C

      INTEGER K,KB,KP1,NM1,NM2
      REAL T
C***FIRST EXECUTABLE STATEMENT  SGTSL
         INFO = 0
         C(1) = D(1)
         NM1 = N - 1
         IF (NM1 .LT. 1) GO TO 40
            D(1) = E(1)
            E(1) = 0.0E0
            E(N) = 0.0E0
C
            DO 30 K = 1, NM1
               KP1 = K + 1
C
C              FIND THE LARGEST OF THE TWO ROWS
C
               IF (ABS(C(KP1)) .LT. ABS(C(K))) GO TO 10
C
C                 INTERCHANGE ROW
C
                  T = C(KP1)
                  C(KP1) = C(K)
                  C(K) = T
                  T = D(KP1)
                  D(KP1) = D(K)
                  D(K) = T
                  T = E(KP1)
                  E(KP1) = E(K)
                  E(K) = T
                  T = B(KP1)
                  B(KP1) = B(K)
                  B(K) = T
   10          CONTINUE
C
C              ZERO ELEMENTS
C
               IF (C(K) .NE. 0.0E0) GO TO 20
                  INFO = K
                  GO TO 100
   20          CONTINUE
```

```
               T = -C(KP1)/C(K)
               C(KP1) = D(KP1) + T*D(K)
               D(KP1) = E(KP1) + T*E(K)
               E(KP1) = 0.0E0
               B(KP1) = B(KP1) + T*B(K)
   30       CONTINUE
   40    CONTINUE
         IF (C(N) .NE. 0.0E0) GO TO 50
            INFO = N
         GO TO 90
   50    CONTINUE
C
C           BACK SOLVE
C
            NM2 = N - 2
            B(N) = B(N)/C(N)
            IF (N .EQ. 1) GO TO 80
               B(NM1) = (B(NM1) - D(NM1)*B(N))/C(NM1)
               IF (NM2 .LT. 1) GO TO 70
               DO 60 KB = 1, NM2
                  K = NM2 - KB + 1
                  B(K) = (B(K) - D(K)*B(K+1) - E(K)*B(K+2))/C(K)
   60          CONTINUE


   70          CONTINUE
   80       CONTINUE
   90    CONTINUE
  100 CONTINUE
C
      RETURN
      END
```

Below we present a program solving the difference convection-diffusion problem in which the convective term in (2.64) is approximated in two ways.

Program PROBLEM1

```
C
C     PROBLEM1  - ONE-DIMENSIONAL STATIONARY
                  CONVECTION-DIFFUSION PROBLEM
C
      REAL KAPPA
      PARAMETER ( KAPPA = 0.01, ISCHEME = 0, N = 20 )
      DIMENSION     Y(N+1),     X(N+1),     PHI(N+1)
     +             ,ALPHA(N+1), BETA(N+1),  GAMMA(N+1)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF SEGMENT;
C     KAPPA     - DIFFUSIVITY
C     ISCHEME   - CENTRAL-DIFFERENCE APPROXIMATIONS (ISCHEME = 0),
C                 SCHEME WITH DIRECTED DIFFERENCES;
C     N + 1     - NUMBER OF NODES;
C     Y(N+1)    - EXACT SOLUTION
C
```

```
      XL = 0.
      XR = 1.

C
      OPEN  ( 01, FILE = 'RESULT.DAT' ) ! FILE TO STORE
                                          THE COMPUTED DATA
C
C     MESH
C
      H = (XR - XL)/N
      DO I = 1, N+1
        X(I) = XL + (I-1)*H
      END DO
C
C     BOUNDARY CONDITION AT THE LEFT END
C
      GAMMA(1) = 1.
      BETA(1)  = 0.
      PHI(1)   = 0.
C
C     BOUNDARY CONDITION AT THE RIGHT END
C
      ALPHA(N+1)  = 0.
      GAMMA(N+1) = 1.
      PHI(N+1)   = 0.
C
C     ELEMENTS OF THE TRIDIAGONAL MATRIX
C
      IF (ISCHEME.EQ.0) THEN
C
C     SCHEME WITH CENTRAL-DIFFERENCE APPROXIMATIONS
C

        DO I = 2,N
          ALPHA(I)  = V(X(I))/(2.*H) + KAPPA/(H*H)
          BETA(I)  = - V(X(I))/(2.*H) + KAPPA/(H*H)
          GAMMA(I) = ALPHA(I) + BETA(I)
          PHI(I)   = F(X(I))
        END DO
      ELSE
C

C     SCHEME WITH DIRECTED DIFFERENCE APPROXIMATIONS
C
        DO I = 2,N
          VP = 0.5*(V(X(I)) + ABS(V(X(I))))
          VM = 0.5*(V(X(I)) - ABS(V(X(I))))
          ALPHA(I)  = VP/H + KAPPA/(H*H)
          BETA(I)  = - VM/H + KAPPA/(H*H)
          GAMMA(I) = ALPHA(I) + BETA(I)
          PHI(I)   = F(X(I))
        END DO
      END IF
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
      DO I = 1,N+1
        ALPHA(I) = - ALPHA(I)
        BETA(I) = - BETA(I)
      END DO
```

```
      CALL SGTSL (N+1, ALPHA, GAMMA, BETA, PHI, INFO)

      IF (INFO.NE.0) STOP  ! POOR MATRIX
C
C     EXACT SOLUTION
C
      AL = XR - XL
      DO I = 1,N+1
        Y(I) = X(I) - AL*EXP((X(I)-AL)/(2.*KAPPA)) *
     +                SINH(X(I)/(2.*KAPPA)) / SINH(AL/(2.*KAPPA))
      END DO
C
C     APPROXIMATE-SOLUTION INACCURACY
C
      EC  = 0.
      EL2 = 0.
      DO I = 1,N+1
        AA = ABS(PHI(I) - Y(I))
        IF (AA.GT.EC) EC = AA
        EL2 = EL2 + AA*AA
      END DO
      EL2 = SQRT(EL2*H)
      WRITE (01,*) KAPPA
      WRITE (01,*) N+1, EC, EL2
      WRITE (01,*) X
      WRITE (01,*) Y
      WRITE (01,*) PHI

      CLOSE (01)
      STOP

      END

      FUNCTION F(X)
C
C     RIGHT-HAND SIDE OF THE EQUATION
C
      F = 1.
      RETURN
      END

      FUNCTION V(X)
C
C     CONVECTIVE-TRANSFER COEFFICIENT
C
      V = 1.
      RETURN
      END
```

2.4.4 Computational experiments

Let us illustrate convergence of the above difference schemes with computational data obtained on a sequence of consecutively refined grids. The calculations were performed for problem (2.64), (2.65) with parameters given as in (2.67). The approximate solution is compared with the exact solution in the mesh norms in $L_2(\omega)$ and $C(\omega)$.

Table 2.1 shows data obtained for the convection-controlled problem ($\kappa = 0.01$) solved by scheme (2.72).

N	10	20	40	80	160
C	0.1599	0.2036	0.1579	0.09219	0.05070
L_2	0.03631	0.03451	0.02339	0.01340	0.007199

Table 2.1 Accuracy of the scheme with directed differences

Theoretical conclusions about the convergence behavior of the schemes with the grid step tending to zero find confirmation here. The scheme with directed differences converges with the first order but, in the case of interest, this asymptotic convergence is achieved with sufficiently fine grids (starting from the grid with $N = 80$). Figure 2.3 shows plots of the approximate solution.

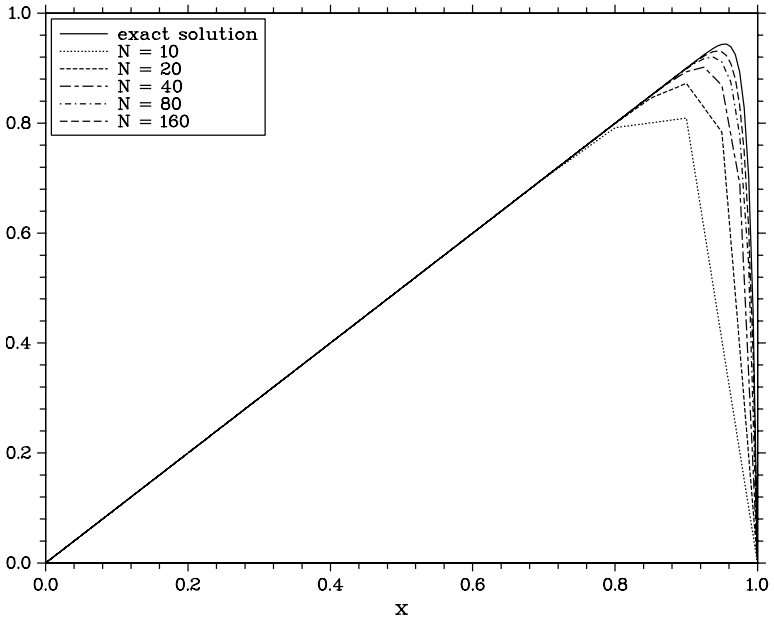

Figure 2.3 Approximate solution (scheme with directed differences)

Analogous data obtained for scheme (2.70) are given in Table 2.2.

N	10	20	40	80	160
C	0.4353	0.1932	0.05574	0.01212	0.003003
L_2	0.1074	0.03056	0.007142	0.001677	0.0004062

Table 2.2 Accuracy of the scheme with central-difference approximations

The inaccuracy decreases approximately fourfold on the twofold refined grid (second-order convergence). It should be noted that, in this example, the condition for monotonicity (2.71) becomes violated on grids with $N = 10, 20, 40$. Violation of this condition results in a substantially distorted solution behavior (see Figure 2.4). We may even say that, here, non-monotone schemes cannot be used.

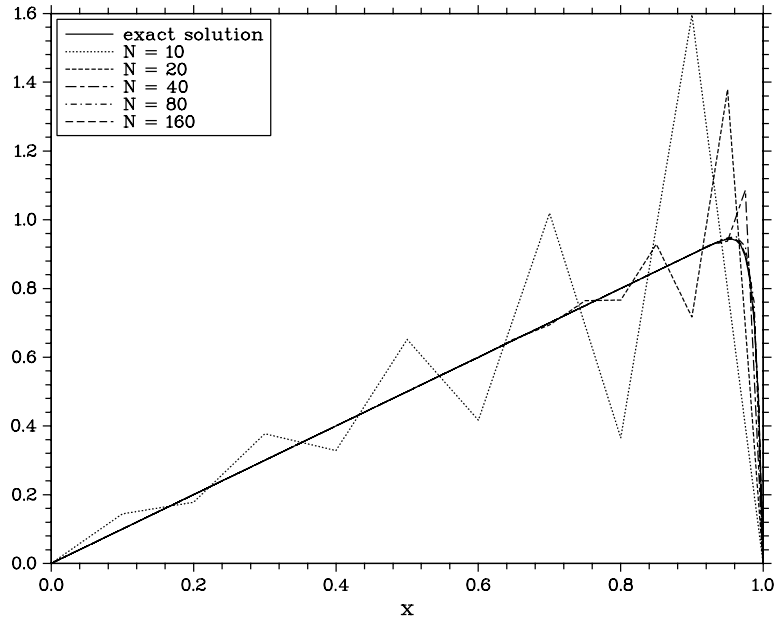

Figure 2.4 Scheme with central-difference approximations

2.5 Exercises

Exercise 2.1 On the solutions of equation (2.1), approximate the third-kind boundary conditions (2.4) accurate to the second order.

Exercise 2.2 On the non-uniform grid

$$\bar{\omega} = \{x \mid x = x_i = x_{i-1} + h_i,\ i = 1, 2, \ldots, N,\ x_0 = 0,\ x_N = l\}$$

construct and examine the difference scheme

$$-\frac{2}{h_{i+1} + h_i}\left(a_{i+1}\,\frac{y_{i+1} - y_i}{h_{i+1}} - a_i\,\frac{y_i - y_{i-1}}{h_i}\right) + c_i y_i = \varphi_i, \qquad i = 1, 2, \ldots, N-1,$$

$$y_0 = \mu_1, \qquad y_N = \mu_2$$

making it possible to solve approximately problem (2.1), (2.2).

Exercise 2.3 Approximate the first-kind boundary conditions (2.4) accurate to the second order so that to solve equation (2.1) on the half-step extended grid

$$x_i = x_0 + ih, \quad i = 0, 1, \dots, N, \quad x_0 = -h/2, \quad x_N = l + h/2.$$

Exercise 2.4 Show that the difference scheme

$$-(ay_{\bar{x}})_x = \varphi(x), \qquad x \in \omega,$$

where

$$a_i = \Big(\frac{1}{h}\int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\Big)^{-1},$$

$$\varphi_i = \frac{1}{h^2}\Big(a_{i+1}\int_{x_i}^{x_{i+1}} \frac{dx}{k(x)}\int_{x_i}^{x} f(s)\,ds + a_i\int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\int_{x}^{x_i} f(s)\,ds\Big)$$

for the approximate solution of the equation

$$-\frac{d}{dx}\Big(k(x)\frac{du}{dx}\Big) = f(x), \qquad 0 < x < l$$

with boundary conditions (2.2) is a perfectly accurate scheme.

Exercise 2.5 Construct an absolutely monotone difference scheme of second approximation order for the boundary value problem (2.64), (2.65) by passing to the equation

$$-\frac{d}{dx}\Big(\tilde{k}(x)\frac{du}{dx}\Big) + \tilde{q}(x)u = \tilde{f}(x), \qquad 0 < x < l.$$

Exercise 2.6 For the boundary value problem (2.1) with homogeneous boundary conditions

$$u(0) = 0, \qquad u(l) = 0,$$

construct a finite element scheme with solution representation in the form (2.22) based on minimization of the functional (the Ritz method)

$$J(v) = \frac{1}{2}\int_0^l \Big(k(x)\Big(\frac{dv}{dx}\Big)^2 + q(x)v^2(x)\Big)\,dx - \int_0^l f(x)v(x)\,dx.$$

Exercise 2.7 Prove the difference Friedrichs inequality (Lemma 2.1) in the case of even N.

Exercise 2.8 Prove the difference Friedrichs inequality for mesh functions given on the non-uniform grid

$$\bar{\omega} = \{x \mid x = x_i = x_{i-1} + h_i, \quad i = 1, 2, \dots, N, \quad x_0 = 0, \quad x_N = l\},$$

where $h_i > 0$, $i = 1, 2, \dots, N$.

Exercise 2.9 Prove that for any mesh function $y(x)$ given on the uniform grid $\bar{\omega}$ and vanishing at $x = 0$ and $x = l$ there holds the following inequality (embedding theorem for mesh functions):

$$\|y(x)\|_\infty \le \frac{\sqrt{l}}{2} \|y_{\bar{x}}\|^+.$$

Exercise 2.10 Find eigenfunctions and eigenvalues of the difference problem

$$y_{\bar{x}x} + \lambda y = 0, \qquad x \in \omega,$$
$$y_0 = 0, \qquad y_N = 0.$$

Exercise 2.11 Suppose that in solving the boundary value problem

$$-\frac{d}{dx}\Big(k(x)\,\frac{du}{dx}\Big) = 0,$$
$$u(0) = 1, \qquad u(1) = 0$$

one uses the difference scheme

$$-k(x) y_{\bar{x}x} - k_{x\circ} y_{x\circ} = 0.$$

Show that this scheme diverges in the class of discontinuous coefficients. Perform numerical experiments.

Exercise 2.12 Let the following conditions be fulfilled:

$$|\alpha_i| > 0, \qquad |\beta_i| > 0, \qquad \delta_i = |\gamma_i| - |\alpha_i| - |\beta_i| > 0,$$
$$i = 1, 2, \ldots, N-1.$$

Prove that under such conditions for the solution of the problem

$$-\alpha_i y_{i-1} + \gamma_i y_i - \beta_i y_{i+1} = \varphi_i, \qquad i = 1, 2, \ldots, N-1,$$
$$y_0 = 0, \qquad y_N = 0$$

there holds the estimate

$$\|y\|_\infty \le \|\varphi/\delta\|_\infty.$$

Exercise 2.13 Let in the difference scheme (2.51), (2.52), (2.68) $\mu_1 \ge 0$, $\mu_2 \ge 0$ and $\varphi_i \ge 0$ for all $i = 1, 2, \ldots, N-1$ (or $\mu_1 \le 0$, $\mu_2 \le 0$ and $\varphi_i \le 0$ for $i = 1, 2, \ldots, N-1$). Then, under the conditions

$$\gamma_i \ge \alpha_{i+1} + \beta_{i-1}, \qquad i = 2, 3, \ldots, N-2,$$
$$\gamma_1 > \alpha_2, \qquad \gamma_{N-1} > \beta_{N-2}$$

(see Theorem 2.5) we have $y_i \ge 0$, $i = 1, 2, \ldots, N-1$ ($y_i \le 0$, $i = 1, 2, \ldots, N-1$).

Exercise 2.14 Construct a sweep method for solving the system of linear equations

$$\begin{aligned} -\alpha_0 y_N + \gamma_0 y_0 - \beta_0 y_1 &= \varphi_0, \\ -\alpha_i y_{i-1} + \gamma_i y_i - \beta_i y_{i+1} &= \varphi_i, \qquad i = 1, 2, \ldots, N-1. \\ -\alpha_N y_{N-1} + \gamma_N y_N - \beta_N y_0 &= \varphi_N \end{aligned}$$

that arises, for instance, when one solves the boundary value problem for equation (2.1) with periodic boundary conditions.

Exercise 2.15 To numerically solve the problem (2.64), (2.65), (2.67), construct a difference scheme on a piecewise-constant grid using uniform grids on the segments $[0, x^*]$ and $[x^*, l]$ with grid densening in the vicinity of $x = l$. Write a program and perform computational experiments to examine the rate of convergence of the difference scheme.

3 Boundary value problems for elliptic equations

Among stationary mathematical physics problems, problems most important for applications are boundary value problems for second-order elliptic equations. For a model two-dimensional problem, we consider the matter of construction of its discrete analogues with the use of regular rectangular grids on the basis of finite-difference approximations. The convergence of approximate solution to the exact solution is examined in mesh Hilbert spaces using the properties of positive definiteness of the elliptic difference operator. The convergence in the homogeneous norm is considered based on the maximum principle for elliptic difference equations. In solving the difference equations that arise upon discretization of boundary value problems for elliptic equations, most frequently used are iteration methods. Here, alternate-triangular iteration methods deserve a detailed consideration. We present a program that solves the Dirichlet problem for the second-order elliptic equation with variable coefficients and give examples of performed calculations.

3.1 The difference elliptic problem

We consider the matter of approximation for model boundary value problems involving second-order elliptic equations. The present approach is primarily based on using the integro-interpolation method (balance method). It is shown that it is possible to construct an appropriate difference problem on the basis of finite element approximation. We briefly discuss the issue of finite-difference approximation of boundary value problems in irregular domains.

3.1.1 Boundary value problems

In the present consideration, primary attention will be paid to two-dimensional boundary value problems in which the calculation domain has a simplest form:

$$\Omega = \big\{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2), \quad 0 < x_\alpha < l_\alpha, \quad \alpha = 1, 2\big\},$$

i.e., is shaped as a rectangle. The main object in the present consideration is the second-order elliptic equation

$$-\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big) + q(\boldsymbol{x})u = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{3.1}$$

The following constraints are imposed on the equation coefficients:

$$k(\boldsymbol{x}) \geq \kappa > 0, \qquad q(\boldsymbol{x}) \geq 0, \qquad \boldsymbol{x} \in \Omega.$$

A typical simplest example of the second-order elliptic equation is given by the Poisson equation:

$$-\Delta u \equiv -\sum_{\alpha=1}^{2} \frac{\partial^2 u}{\partial x_\alpha^2} = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{3.2}$$

In the latter case, in (3.1) we have $k(\boldsymbol{x}) = 1$ and $q(\boldsymbol{x}) = 0$.

Equation (3.1) is to be supplemented with some boundary conditions. In the case of the Dirichlet problem, the boundary conditions are

$$u(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega. \tag{3.3}$$

In more complex cases, second- or third-order boundary conditions can be given on the domain boundary or on a part of the boundary, for instance,

$$k(\boldsymbol{x}) \frac{\partial u}{\partial n} + \sigma(\boldsymbol{x}) u = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega, \tag{3.4}$$

where, recall, n is the external normal to Ω.

3.1.2 Difference problem

We use a grid uniform in both directions. For grids over particular directions x_α, $\alpha = 1, 2$ we use the notation

$$\overline{\omega}_\alpha = \{x_\alpha \mid x_\alpha = i_\alpha h_\alpha, \quad i_\alpha = 0, 1, \ldots, N_\alpha, \quad N_\alpha h_\alpha = l_\alpha\}.$$

Here,

$$\begin{aligned} \omega_\alpha &= \{x_\alpha \mid x_\alpha = i_\alpha h_\alpha, \quad i_\alpha = 1, 2, \ldots, N_\alpha - 1, \quad N_\alpha h_\alpha = l_\alpha\}, \\ \omega_\alpha^+ &= \{x_\alpha \mid x_\alpha = i_\alpha h_\alpha, \quad i_\alpha = 1, 2, \ldots, N_\alpha, \quad N_\alpha h_\alpha = l_\alpha\}. \end{aligned}$$

For a grid in the rectangle Ω, we put

$$\begin{aligned} \overline{\omega} &= \overline{\omega}_1 \times \overline{\omega}_2 = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2), \quad x_\alpha \in \overline{\omega}_\alpha, \quad \alpha = 1, 2\}, \\ \omega &= \omega_1 \times \omega_2. \end{aligned}$$

If the coefficients in (2.1) are smooth, then the difference scheme can be constructed based on immediate passage from differential to difference operators. Similarly to the one-dimensional case, for the boundary value problem (3.1), (3.3), we put into correspondence to the differential equation the difference equation

$$Ay = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \tag{3.5}$$

with

$$A = \sum_{\alpha=1}^{2} A^{(\alpha)}, \qquad \boldsymbol{x} \in \omega, \tag{3.6}$$

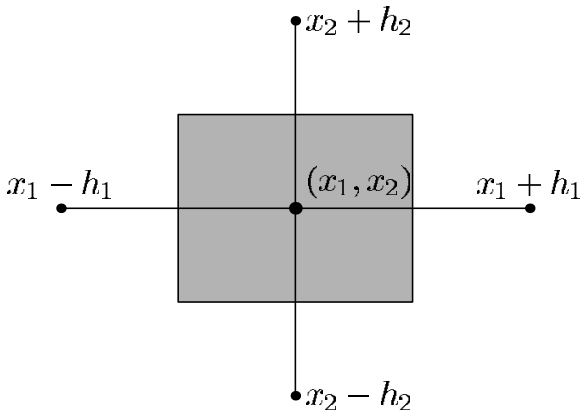

Figure 3.1 Control volume

$$A^{(\alpha)} y = -(a^{(\alpha)} y_{\bar{x}_\alpha})_{x_\alpha} + \theta_\alpha c(\boldsymbol{x}) y, \qquad \alpha = 1, 2, \quad \boldsymbol{x} \in \omega,$$

where $\theta_1 + \theta_2 = 1$.

For the coefficients at higher-order derivatives we put

$$a^{(1)}(\boldsymbol{x}) = k(x_1 - 0.5h_1, x_2), \qquad x_1 \in \omega_1^+, \quad x_2 \in \omega_2,$$
$$a^{(2)}(\boldsymbol{x}) = k(x_1, x_2 - 0.5h_2), \qquad x_1 \in \omega_1, \quad x_2 \in \omega_2^+.$$

For the lowest coefficient and for the right-hand side in (3.5), (3.6) we have:

$$c(\boldsymbol{x}) = q(\boldsymbol{x}), \qquad \varphi(\boldsymbol{x}) = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega.$$

To approximate the boundary conditions (3.3) at the boundary nodes of $\partial\omega$ ($\overline{\omega} = \omega \cup \partial\omega$), we use

$$y(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega. \tag{3.7}$$

The most efficient approach to the construction of difference schemes implies using the integro-interpolation method. First of all, the latter is the case with irregular grids (triangular ones, for instance), but also with simplest rectangular grids. In the case of the uniform rectangular grid under consideration, equation (3.1) is integrated over a control volume around each inner node $\boldsymbol{x}$ of ω (see Figure 3.1):

$$\Omega_{\boldsymbol{x}} = \big\{ \boldsymbol{s} \mid \boldsymbol{s} = (s_1, s_2), \quad x_1 - 0.5h_1 \le s_1 \le x_1 + 0.5h_1, \\ x_2 - 0.5h_2 \le s_2 \le x_2 + 0.5h_2 \big\}.$$

Like in the one-dimensional case, we arrive at the difference scheme (3.5), (3.6) with

$$a^{(1)}(\boldsymbol{x}) = \frac{1}{h_2} \int_{x_2 - 0.5h_2}^{x_2 + 0.5h_2} \Big(\frac{1}{h_1} \int_{x_1 - h_1}^{x_1} \frac{ds_1}{k(\boldsymbol{s})} \Big)^{-1} ds_2,$$

$$a^{(2)}(\boldsymbol{x}) = \frac{1}{h_1} \int_{x_1-0.5h_1}^{x_1+0.5h_1} \Big(\frac{1}{h_2} \int_{x_2-h_2}^{x_2} \frac{ds_2}{k(\boldsymbol{s})} \Big)^{-1} ds_1.$$

The right-hand side and the lower terms are approximated with the expressions

$$\varphi(\boldsymbol{x}) = \frac{1}{h_1 h_2} \int_{x_1-0.5h_1}^{x_1+0.5h_1} \int_{x_2-0.5h_2}^{x_2+0.5h_2} f(\boldsymbol{x})\, d\boldsymbol{x},$$

$$c(\boldsymbol{x}) = \frac{1}{h_1 h_2} \int_{x_1-0.5h_1}^{x_1+0.5h_1} \int_{x_2-0.5h_2}^{x_2+0.5h_2} q(\boldsymbol{x})\, d\boldsymbol{x}.$$

In a similar manner, difference schemes for irregular grids can be constructed.

3.1.3 Problems in irregular domains

Certain difficulties arise in numerical solution of boundary value problems for elliptic equations in complex *irregular calculation domains*. So far we have dealt with problems in a rectangular region of a (*regular calculation domain*). We will not discuss here the latter scope of rather complex problems at much length; instead, we only give a short summary of major lines in this field.

Traditionally, in the approximate solution of stationary mathematical physics problems irregular calculation grids are widely used. Among irregular grids, two main classes of grids can be distinguished.

Structured grids. A most important example of such grids are irregular quadrangular grids, which in many respects inherit the properties of standard rectangular grids or, in other words, present grids topologically equivalent to rectangular grids.

Unstructured grids. Here, the mesh pattern has a variable structure. It is impossible to relate the calculation grid with some regular rectangular grid. In particular, the scheme can be written at each point with different numbers of neighbors.

Approximation on structured grids can be constructed based on the noted closeness of such grids to standard rectangular grids. The latter can be most easily done by introducing new independent variables.

A second possibility bears no relation with formal introduction of new coordinates and can be realized through approximation of the initial problem on such an irregular grid. Of course, in the case of irregular grids the use of simplest approaches for the construction of difference schemes on the basis of undetermined coefficients, although possible, seems not to be a structurally reasonable strategy. In the latter case, one can use the balance method. In both cases, i.e., in the case of general unstructured grids and in the case of structured grids, one can construct difference schemes using finite element approximations.

In a number of cases, for irregular grids it is possible to construct a matched grid formed by the nodes of an ordinary non-uniform rectangular grid and by the boundary

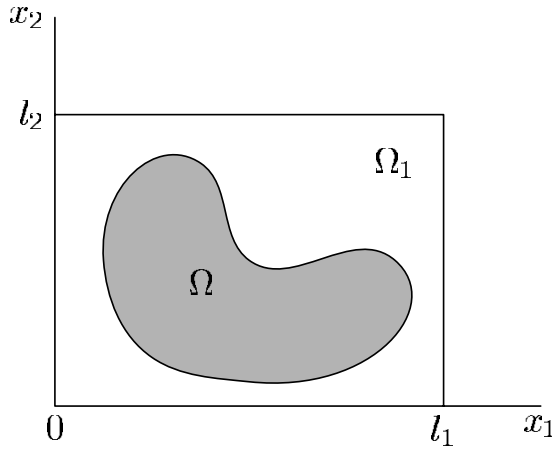

Figure 3.2 Method of fictitious domains

nodes. For the boundary to be formed by nodes, one has to use strongly non-uniform grids. Problems related with construction of difference schemes on a matched grid can be solved in the ordinary way.

A matched difference scheme can be constructed only for rather a narrow class of calculation domains. That is why, normally, other approaches are used to solve the problem related with calculation-domain irregularity. A simplest method here implies using an ordinary rectangular grid in the calculation domain with a boundary condition transferred to the node adjacent to the boundary. In fact, we deal here with the passage from the initial problem (3.1), (3.3) to the problem in another domain whose boundary is adapted to the grid (approximation of boundary).

The most natural and rather universal method here is an approach using a grid formed by the nodes of a regular (uniform) grid (inner nodes) and additional irregular boundary nodes at the domain boundary. The latter nodes are formed by intersections of the lines drawn through the nodes of the regular grid and the boundary of the calculation domain.

More convenient in the approximate solution of boundary value problems in irregular domains is the *method of fictitious domains*. This method is based on extension of the initial calculation domain to some regular domain Ω_0 ($\Omega \subset \Omega_0$), for instance, to a rectangle in two-dimensional problems (Figure 3.2). Afterwards, the problem in Ω_0 can be solved by ordinary difference methods. It is also necessary to so extend the solution of the initial problem into the fictitious domain $\Omega_1 = \Omega_0 \setminus \Omega$ that the difference solution of the problem in the extended domain Ω_0 would give an approximate solution in the initial domain Ω.

The problem in the extended domain involves small (high) coefficients of the differential equation. As a result, it becomes necessary to examine the matter of accuracy and computational realization of various iteration methods for solving resulting problems.

A method alternative to the method of fictitious domains in approximate solution of boundary value problems in irregular calculation domains is the *decomposition method*, in which partition of the calculation domain into simple subdomains is used. This approach has been widely discussed in literature as applied to the development of solution algorithms for boundary value problems using advanced parallel computers.

In each of the subdomains, individual boundary value problems are to be solved, and the solutions are to be matched via boundary conditions. In the subdomains, special grids can be used, matched or not with the grids in other subdomains. That is why difference domain-decomposition schemes can be regarded on the difference level as difference methods on composite grids.

3.2 Approximate-solution inaccuracy

Let us outline possible approaches for the examination of the rate of convergence of approximate solution to the exact solution. The present consideration is based on the examination of the properties of self-adjointness, positive definiteness, and monotonicity of elliptic difference operators.

3.2.1 Elliptic difference operators

Note some basic properties of difference operators that arise in approximate solution of the model problem (3.1), (3.3). Like in the previous consideration for one-dimensional problems, let us reformulate (by modifying the right-hand side at boundary nodes) the difference problem (3.5), (3.7) so that to obtain a problem with homogeneous boundary conditions.

For mesh functions vanishing on the set of boundary nodes $\partial\omega$, we define a Hilbert space $H = L_2(\omega)$ with the scalar product and the norm defined as

$$(y, w) \equiv \sum_{\boldsymbol{x}\in\omega} y(\boldsymbol{x})w(\boldsymbol{x})h_1h_2, \qquad \|y\| \equiv (y, y)^{1/2}.$$

As a most important property, we distinguish the property of self-adjointness of the elliptic difference operator (3.5). The latter property of A stems, with regard to (3.6), from self-adjointness of the one-dimensional operators $A^{(\alpha)}$, $\alpha = 1, 2$. Taking the latter into account, we obtain:

$$\begin{aligned}(Ay, w) &= \sum_{x_2\in\omega_2} h_2 \sum_{x_1\in\omega_1} A^{(1)}y(\boldsymbol{x})w(\boldsymbol{x})h_1 + \sum_{x_1\in\omega_1} h_1 \sum_{x_2\in\omega_2} A^{(2)}y(\boldsymbol{x})w(\boldsymbol{x})h_2 \\ &= \sum_{x_2\in\omega_2} h_2 \sum_{x_1\in\omega_1} w(\boldsymbol{x})A^{(1)}y(\boldsymbol{x})h_1 + \sum_{x_1\in\omega_1} h_1 \sum_{x_2\in\omega_2} w(\boldsymbol{x})A^{(2)}y(\boldsymbol{x})h_2 \\ &= (y, Aw).\end{aligned}$$

For two-dimensional mesh functions vanishing on $\partial\omega$, we define the following difference analogue to the norm in $W_2^1(\omega)$:

$$\|\nabla y\|^2 \equiv \sum_{x_1\in\omega_1^+}\sum_{x_2\in\omega_2}(y_{\bar{x}_1})^2 h_1 h_2 + \sum_{x_1\in\omega_1}\sum_{x_2\in\omega_2^+}(y_{\bar{x}_2})^2 h_1 h_2.$$

Since

$$\begin{aligned}(Ay, y) &= (A^{(1)}y, y) + (A^{(2)}y, y)\\ &= \sum_{x_2\in\omega_2} h_2 \sum_{x_1\in\omega_1^+} a^{(1)}(y_{\bar{x}_1})^2 h_1 + \sum_{x_1\in\omega_1} h_1 \sum_{x_2\in\omega_2^+} a^{(2)}(y_{\bar{x}_2})^2 h_2\end{aligned}$$

and $a_\alpha(\boldsymbol{x}) \geq \kappa$, $\alpha = 1, 2$, then for the two-dimensional difference operator (3.6) the following inequality holds:

$$(Ay, y) \geq \kappa\|\nabla y\|^2. \tag{3.8}$$

To estimate the two-dimensional difference operator of diffusion transport, we use the *Friedrichs inequality for two-dimensional mesh functions*.

Lemma 3.1 *For mesh functions $y(\boldsymbol{x})$ vanishing on $\partial\omega$ the following inequality holds:*

$$\|y\|^2 \leq M_0\|\nabla y\|^2, \qquad M_0^{-1} = \frac{8}{l_1^2} + \frac{8}{l_2^2}. \tag{3.9}$$

Proof. We take into account the Friedrichs inequality for one-dimensional mesh functions (Lemma 2.1); then, we obtain the inequality

$$\sum_{x_1\in\omega_1^+}\sum_{x_2\in\omega_2}(y_{\bar{x}_1})^2 h_1 h_2 + \sum_{x_1\in\omega_1}\sum_{x_2\in\omega_2^+}(y_{\bar{x}_2})^2 h_1 h_2 \geq \left(\frac{8}{l_1^2} + \frac{8}{l_2^2}\right)\sum_{x_1\in\omega_1}\sum_{x_2\in\omega_2} y^2 h_1 h_2.$$

The latter inequality yields (3.9). □

From (3.8) and (3.9), the following lower estimate for A follows:

$$A \geq \kappa\lambda_0 E, \qquad \lambda_0 = M_0^{-1}. \tag{3.10}$$

Let us give now an upper estimate of the elliptic difference operator of interest.

Lemma 3.2 *For the difference operator A there holds the inequality*

$$A \leq M_1 E \tag{3.11}$$

with the constant

$$M_1 = \frac{4}{h_1^2}\max_{\boldsymbol{x}\in\omega}\frac{a^{(1)}(\boldsymbol{x}) + a^{(1)}(x_1 + h_1, x_2)}{2} + \frac{4}{h_2^2}\max_{\boldsymbol{x}\in\omega}\frac{a^{(2)}(\boldsymbol{x}) + a^{(2)}(x_1, x_2 + h_2)}{2} + \max_{\boldsymbol{x}\in\omega} c(\boldsymbol{x}).$$

Proof. This statement can be proved analogously to the one-dimensional case (see Lemma 2.2). □

3.2.2 Convergence of difference solution

Based on the established properties of the elliptic difference operator, one can derive desired a priori estimates. Considering the problem for the approximate-solution inaccuracy, these estimates allow one to draw a conclusion about the rate of convergence in difference schemes for problem (3.1), (3.3). In the present monograph, such an examination was performed previously for one-dimensional stationary problems. Here we give, with some minor changes, an analogous consideration for two-dimensional problems.

With regard to (3.5), the problem for the difference-solution inaccuracy

$$z(\boldsymbol{x}) = y(\boldsymbol{x}) - u(\boldsymbol{x}), \qquad \boldsymbol{x} \in \bar{\omega}$$

has the form

$$Az = \psi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega. \tag{3.12}$$

Here, as usually, $\psi(\boldsymbol{x})$ is the approximation inaccuracy:

$$\psi(\boldsymbol{x}) = \varphi(\boldsymbol{x}) - Au, \qquad \boldsymbol{x} \in \omega.$$

We assume that the boundary value problem (3.1), (3.3) has a sufficiently smooth classical solution. It should be noted in this connection that, apart from the smoothness of equation coefficients, boundary conditions and the right-hand side, for the difference problem in the rectangle Ω certain matching conditions at the corners must be fulfilled. Under such conditions, on a uniform rectangular grid the approximation inaccuracy (3.5), (3.6) is of the second order:

$$\psi(\boldsymbol{x}) = \mathcal{O}(|h|^2), \qquad |h|^2 \equiv h_1^2 + h_2^2, \qquad \boldsymbol{x} \in \omega. \tag{3.13}$$

Theorem 3.3 *For the difference scheme* (3.5)*, the following a priori estimate for the inaccuracy holds:*

$$\|\nabla z\| \le \frac{M_0^{1/2}}{\kappa} \|\psi\|. \tag{3.14}$$

Proof. We scalarwise multiply equation (3.12) for the inaccuracy by $z(\boldsymbol{x})$, $\boldsymbol{x} \in \omega$; then, we obtain

$$(Az, z) = (\psi, z).$$

For the right-hand side, by virtue of (3.9) we have:

$$(\psi, z) \le \|\psi\| \|z\| \le M_0^{1/2} \|\psi\| \, \|\nabla z\|.$$

With (3.8), for the left-hand side we have:

$$(Az, z) \ge \kappa \|\nabla z\|^2;$$

from that, estimate (3.14) follows. □

Under conditions (3.13), Theorem 3.3 guarantees that the approximate solution converges to the exact solution accurate to the second order in the mesh analogue of the Sobolev space $W_2^1(\Omega)$.

3.2.3 Maximum principle

For the solution of the boundary value problem (3.1), (3.3), the maximum principle is valid. Apart from the fact that the maximum principle can be used to examine the convergence of difference schemes in the homogeneous norm, it would be highly desirable if this important property would also be retained for the solution of the difference problem (3.5), (3.7).

Let us formulate the maximum principle for the difference schemes. We write the difference equation (3.5), (3.7) as

$$Sy(\boldsymbol{x}) = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \tag{3.15}$$

where the linear operator S is defined by

$$Sv(\boldsymbol{x}) = A(\boldsymbol{x})v(\boldsymbol{x}) - \sum_{\xi \in \mathcal{W}'(\boldsymbol{x})} B(\boldsymbol{x}, \xi)v(\xi). \tag{3.16}$$

Here $\mathcal{W}(\boldsymbol{x})$ is the mesh pattern, and $\mathcal{W}' \equiv \mathcal{W} \setminus \{\boldsymbol{x}\}$ is the vicinity of the node $\boldsymbol{x} \in \omega$.

Suppose that for the second-order elliptic equations under consideration the mesh pattern $\mathcal{W}$ for inner nodes in the calculation grid involves nodes $(x_1 \pm h_1, x_2)$, $(x_1, x_2 \pm h_2)$ (the mesh pattern is a five-point one at least), and the coefficients satisfy the conditions

$$\begin{aligned} &A(\boldsymbol{x}) > 0, \qquad B(\boldsymbol{x}, \xi) > 0, \qquad \xi \in \mathcal{W}'(\boldsymbol{x}), \\ &D(\boldsymbol{x}) = A(\boldsymbol{x}) - \sum_{\xi \in \mathcal{W}'(\boldsymbol{x})} B(\boldsymbol{x}, \xi) > 0, \qquad \boldsymbol{x} \in \omega. \end{aligned} \tag{3.17}$$

With conditions (3.17) satisfied, the maximum principle holds for the difference equation (3.15), (3.16).

Theorem 3.4 *Suppose that the mesh function $y(\boldsymbol{x})$ satisfies equation* (3.15)*,* (3.16)*, with conditions* (3.17) *fulfilled for the coefficients of this equation, and boundary conditions* (3.7)*. With the conditions*

$$\begin{aligned} &\mu(\boldsymbol{x}) \le 0 \qquad (\mu(\boldsymbol{x}) \ge 0), \\ &\varphi(\boldsymbol{x}) \le 0 \qquad (\varphi(\boldsymbol{x}) \ge 0), \qquad \boldsymbol{x} \in \omega \end{aligned}$$

fulfilled, for the difference solution we have $y(\boldsymbol{x}) \le 0$ ($y(\boldsymbol{x}) \ge 0$), $\boldsymbol{x} \in \omega$.

Based on the maximum principle, comparison theorems for the solutions of elliptic difference equations can be proved. Consider, for instance, the problem

$$\begin{aligned} Sw(\boldsymbol{x}) &= \phi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \\ w(\boldsymbol{x}) &= \nu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega \end{aligned}$$

and let

$$|\varphi(\boldsymbol{x})| \le \phi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega,$$
$$|\mu(\boldsymbol{x})| \le \nu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega.$$

Then, the following estimate is valid for the solution of problem (3.15), (3.7):

$$|y(\boldsymbol{x})| \le w(\boldsymbol{x}), \qquad \boldsymbol{x} \in \overline{\omega}.$$

Invoking such estimates, one can immediately see that for the solution of the homogeneous equation (3.15) ($\varphi(\boldsymbol{x}) = 0$, $\boldsymbol{x} \in \omega$) with the boundary conditions (3.7) the following a priori stability estimate holds:

$$\max_{\boldsymbol{x}\in\omega} |y(\boldsymbol{x})| \le \max_{\boldsymbol{x}\in\partial\omega} |\mu(\boldsymbol{x})|.$$

With similar a priori estimates, convergence of difference schemes in the homogeneous norm can be proved. Let us illustrate this statement with a simple example.

To approximately solve the Dirichlet problem for the Poisson equation (3.2), (3.3), we use the difference equation

$$-y_{\bar{x}_1 x_1} - y_{\bar{x}_2 x_2} = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \tag{3.18}$$

supplemented with the boundary conditions (3.7). For the inaccuracy $z(\boldsymbol{x}) = y(\boldsymbol{x}) - u(\boldsymbol{x})$, $\boldsymbol{x} \in \overline{\omega}$, we obtain the problem

$$-z_{\bar{x}_1 x_1} - z_{\bar{x}_2 x_2} = \psi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega,$$
$$z(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\omega,$$

where $\psi(\boldsymbol{x}) = \mathcal{O}(h_1^2 + h_2^2)$ is the approximation inaccuracy. We choose as a majorizing function the function

$$w(\boldsymbol{x}) = \frac{1}{4}\,(l_1^2 + l_2^2 - x_1^2 - x_2^2)\|\psi(\boldsymbol{x})\|_\infty,$$

where

$$\|v(\boldsymbol{x})\|_\infty = \max_{\boldsymbol{x}\in\omega} |v(\boldsymbol{x})|.$$

Then, we obtain the following estimate for the inaccuracy:

$$\|y(\boldsymbol{x}) - u(\boldsymbol{x})\|_\infty \;\le\; \frac{1}{4}\,(l_1^2 + l_2^2)\|\psi(\boldsymbol{x})\|_\infty.$$

Hence, the difference scheme (3.18), (3.7) converges in $L_\infty(\omega)$ with the second order.

3.3 Iteration solution methods for difference problems

In the numerical solution of stationary mathematical physics problems, iteration methods are used which can be regarded as relaxation methods. Below, we give basic notions in use in the theory of iteration solution methods for operator equations of interest in finite-difference Hilbert spaces. We discuss the choice of iteration parameters and the operator of transition (reconditioner) to a next iteration approximation.

3.3.1 Direct solution methods for difference problems

Upon approximation, the initial differential problem is replaced with a difference problem. The resultant difference (mesh) equations lead to a system of linear algebraic equations for unknown values of the mesh function. These values can be found using direct or iterative linear algebra methods that to the largest possible extent take specific features of particular difference problems into account. A specific feature of difference problems consists in the fact that the resultant matrix of the linear system is a sparse matrix that contains many zero elements and has a banded structure. In the case of multi-dimensional problems the matrix is of a very high order equal to the total number of grid nodes.

First of all, consider available possibilities in the construction of time-efficient direct methods for solving a sufficiently broad range of elliptic difference problems with separable variables. The classical approach to the solution of simplest linear mathematical physics problems is related with the use of the variable separation method. It can be expected that an analogous idea will also receive development in the case of difference equations. Consider the difference problem for the Poisson equation (3.18) with homogeneous boundary conditions

$$y(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\omega. \tag{3.19}$$

As it was noted above, modifying the right-hand side, one can always pass on the difference level from the problem with inhomogeneous boundary conditions to a problem with homogeneous boundary conditions.

To apply the Fourier method to this two-dimensional problem, consider the eigenvalue problem for the difference operator of the second derivative with respect to x_1:

$$-v_{\overline{x}_1 x_1} + \lambda v = 0, \qquad x_1 \in \omega_1,$$
$$v_0 = 0, \qquad v_{N_1} = 0.$$

We denote the eigenvalues and the eigenfunctions as λ_k and $v^{(k)}(x_1)$, $k = 1, 2, \ldots, N_1 - 1$:

$$\lambda_k = \frac{4}{h_1^2} \sin^2 \frac{k\pi h_1}{2l_1},$$

$$v^{(k)}(x_1) = \sqrt{\frac{2}{l_1}} \sin \frac{k\pi x_1}{l_1}, \qquad k = 1, 2, \ldots, N_1 - 1.$$

We seek the approximate solution of (3.18), (3.19) as the expansion

$$y(\boldsymbol{x}) = \sum_{k=1}^{N_1-1} c^{(k)}(x_2) v^{(k)}(x_1), \qquad \boldsymbol{x} \in \omega. \tag{3.20}$$

Let $\varphi^{(k)}(x_2)$ be the right-hand side Fourier coefficients:

$$\varphi^{(k)}(x_2) = \sum_{k=1}^{N_1-1} \varphi(\boldsymbol{x}) v^{(k)}(x_1) h_1. \tag{3.21}$$

For $c^{(k)}(x_2)$, we obtain the following three-point problems:

$$-c^{(k)}_{\bar{x}_2 x_2} - \lambda c^{(k)} = \varphi^{(k)}(x_2), \qquad x_2 \in \omega_2, \tag{3.22}$$

$$c^{(k)}_0 = 0, \qquad c^{(k)}_{N_2} = 0. \tag{3.23}$$

At each $k = 1, 2, \ldots, N_1 - 1$, the difference problem (3.22), (3.23) can be solved by the sweep method.

Thus, the Fourier method involves the determination of eigenfunctions and eigenvalues of a one-dimensional difference problem, the calculation of the right-hand side Fourier coefficients by formula (3.21), the solution of problems (3.22), (3.23) for the expansion coefficients and, finally, finding the solution of the problem by summation formulas (3.20).

Efficient computation algorithms for the variable separation method use the fast Fourier transform (FFT). In this case, one can calculate the right-hand side Fourier coefficients and reconstruct the solution with the computation cost $Q = \mathcal{O}(N_1 N_1 \log N_1)$. In the case of problems with constant coefficients one can use the Fourier transform over both variables (expansion in eigenfunctions of the two-dimensional difference operator).

3.3.2 Iteration methods

Multi-dimensional elliptic difference problems with variable coefficients can be solved using iteration methods. Define the main notions used in the theory of iterative solution methods for systems of linear equations.

In a finite-dimensional Hilbert space H, we seek the function $y \in H$ as the solution of the operator equation

$$Ay = \varphi. \tag{3.24}$$

Here, the operator A is considered as a linear positive operator that acts in H, and φ is a given element in H.

Iteration method is a method in which, starting from some initial approximation $y_0 \in H$, we determine, in succession, approximate solutions $y_1, y_2, \ldots, y_k, \ldots$, of equation (3.24), where k is the iteration number. The values y_{k+1} are being calculated

from the previously found values of $y_k,\ y_{k-1}, \ldots$. If for the calculation y_{k+1} only the values y_k obtained at the previous iteration are used, then the iteration method is called one-sweep (*two-layer*) method. Accordingly, if the values of y_k and y_{k-1} are used, the iteration method is called *three-layer* method.

Any two-layer iteration method can be written as

$$B_k \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = \varphi, \qquad k = 0, 1, \ldots. \tag{3.25}$$

In the theory of difference schemes (3.25), they distinguish the *canonical form of the two-layer iteration method*. Given y_0, one can find all subsequent approximations by formula (3.25). Considering notation (3.25), one can explicitly trace the relation between this method and difference schemes intended for approximate solution of non-stationary problems.

The accuracy of an approximate solution can be adequately characterized by the inaccuracy $z_k = y_k - y$. We consider convergence of the iteration method in the energy space H_D generated by a self-adjoint operator D positively defined in H. In H_D, the scalar product and the norm are

$$(y, w)_D = (Dy, w), \qquad \|y\|_D = ((y, y)_D)^{1/2}.$$

The iteration method converges in H_D if $\|z_k\|_D \to 0$ as $k \to \infty$. As a convergence measure for iterations, they use the relative inaccuracy ε, so that at the nth iteration

$$\|y_n - y\|_D \le \varepsilon \|y_0 - y\|_D. \tag{3.26}$$

Since the exact solution y is unknown, the accuracy of the approximate solution is judged considering the discrepancy

$$r_k = Ay_k - \varphi = Ay_k - Ay,$$

which can be calculated immediately. For instance, the iterative process is continued unless we have

$$\|r_n\| \le \varepsilon \|r_0\|. \tag{3.27}$$

The use of convergence criterion (3.27) implies that in (3.26) we choose $D = A^*A$. We denote as $n(\varepsilon)$ the minimum number of iterations that guarantees the accuracy ε to be achieved (fulfillment of (3.26) or (3.27)).

To construct the iteration method, we have to strive for minimization of the computational work on finding the approximate solution of problem (3.24) with desired accuracy. Let Q_k be the total number of arithmetic operations required for the approximation y_k to be found, and assume that $n \ge n(\varepsilon)$ iterations have been made. Then, the computation costs can be evaluated as

$$Q(\varepsilon) = \sum_{k=1}^{n} Q_k.$$

As applied to the two-layer iteration method (3.25), the quantity $Q(\varepsilon)$ can be minimized through a proper choice of B_k and τ_{k+1}. Normally, the operators B_k are considered based on this or that reasoning, and the iteration method (3.25) can be optimized through a proper choice of iteration parameters.

In the theory of iteration methods, two approaches to the choice of iteration parameters have gained acceptance. The first one is related with invoking some a priori information about the operators of the iteration scheme (B_k and A in (3.25)). In the second approach (variation-type iteration methods), iteration parameters are calculated at each iteration by minimizing some functional, no a priori information about the operators being explicitly used. First, dwell on the general description of iteration methods without specifying the structure of difference operators B_k.

As a basic problem, below we consider the problem (3.24) with a self-adjoint operator A positively defined in a finite-difference Hilbert space H ($A = A^* > 0$). We examine the iteration process

$$B\,\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = \varphi, \qquad k = 0, 1, \ldots, \tag{3.28}$$

i.e., here, in contrast to the general case (3.25), the operator B is constant (not varied in the course of iterations).

3.3.3 Examples of simplest iteration methods

Simple iteration method refers to the case in which the iteration parameter in (3.28) is a constant ($\tau_{k+1} = \tau$), i.e., here, we consider the iterative process

$$B\,\frac{y_{k+1} - y_k}{\tau} + Ay_k = \varphi, \qquad k = 0, 1, \ldots \tag{3.29}$$

under the assumption that

$$A = A^* > 0, \qquad B = B^* > 0. \tag{3.30}$$

The iteration method (3.29) is called *stationary*.

Let a priori information about the operators B and A be given as a two-sided operator inequality

$$\gamma_1 B \le A \le \gamma_2 B, \qquad \gamma_1 > 0, \tag{3.31}$$

i.e., the operators B and A are energetically equivalent with some energy equivalence constants γ_α, $\alpha = 1, 2$. The following basic statement about the rate of convergence in the iteration method under consideration is valid:

Theorem 3.5 *The iteration method* (3.29)–(3.31) *converges in* H_D, $D = A$, B *if* $0 < \tau < 2/\gamma_2$. *The optimum value of the iteration parameter is*

$$\tau = \tau_0 = \frac{2}{\gamma_1 + \gamma_2} \tag{3.32}$$

and, in the latter case, the following estimate holds for the total number of iterations n required for an accuracy ε to be achieved:

$$n \geq n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho_0}. \tag{3.33}$$

Here,

$$\rho_0 = \frac{1-\xi}{1+\xi}, \qquad \xi = \frac{\gamma_1}{\gamma_2}.$$

Note that, generally speaking, $n_0(\varepsilon)$ in (3.33) is not an integer number, and n is the minimum integer number that satisfies the inequality $n \geq n_0(\varepsilon)$. Theorem 3.5 shows how can the iteration process (3.29), (3.30) be optimized through a proper choice of B made according to (3.31), i.e. the operator B must be close to A in energy.

The optimal choice of iteration parameters in (3.28) requires finding the roots of Chebyshev polynomials, and this method therefore is called the *Chebyshev iteration method* (Richardson method). We define the set $\mathcal{M}_n$ as follows:

$$\mathcal{M}_n = \Big\{ -\cos\Big(\frac{2i-1}{2n}\pi\Big), \quad i = 1, 2, \ldots, n \Big\}. \tag{3.34}$$

For the iteration parameters τ_k, we use the formula

$$\tau_k = \frac{\tau_0}{1+\rho_0\mu_k}, \qquad \mu_k \in \mathcal{M}_n, \qquad k = 1, 2, \ldots, n. \tag{3.35}$$

The following key statement about the rate of convergence of the iteration method with the Chebyshev set of iteration parameters can be formulated.

Theorem 3.6 *The Chebyshev iteration method* (3.28), (3.30), (3.31), (3.34), (3.35) *converges in H_D, $D = A, B$, and for the total number n of iterations required for an accuracy ε to be achieved, the following estimate holds:*

$$n \geq n_0(\varepsilon) = \frac{\ln (2\varepsilon^{-1})}{\ln \rho_1^{-1}}, \tag{3.36}$$

where

$$\rho_1 = \frac{1-\xi^{1/2}}{1+\xi^{1/2}}, \qquad \xi = \frac{\gamma_1}{\gamma_2}.$$

In the Chebyshev method, iteration parameters are calculated (see (3.34) and (3.35)) from some given total number of iterations n. Apparently, the degenerate case $n = 1$ gives the simple iteration method considered above. Practical realization of the Chebyshev iteration method is related with the problem of computational stability. The latter stems from the fact that the norm of the transition operator at individual iterations exceeds unity, and the local inaccuracy may therefore increase till the occurrence of an emergency termination of the program. The problem of computational stability can be

solved using special ordering of iteration parameters (achieved by choosing μ_k from the set $\mathcal{M}_n$). The optimal sequences of iteration parameters τ_k can be calculated from the given number of iterations n using various algorithms.

Also, worthy of noting is the widely used three-layered Chebyshev iteration method in which iteration parameters are calculated by recurrence formulas. In this case, the inaccuracy decreases monotonically, and it becomes unnecessary to pre-choose the total number of iterations n, as it was the case with method (3.34), (3.35).

3.3.4 Variation-type iteration methods

Above, we have considered iteration methods for solving the problem (3.24) with a priori information about the operators B and A given in the form of energy equivalence constants γ_1 and γ_2 (see (3.31)). From these constants, optimum values of iteration parameters could be found (see (3.32), (3.35)). Estimation of these constants may turn out to be a difficult problem and, therefore, it makes sense to try construct iteration methods in which iteration parameters could be calculated without such a priori information. Such methods are known as *variation-type iteration methods*. Let us begin with a consideration of the two-layer iteration method (3.28) under the assumption of (3.30).

We denote the discrepancy as $r_k = Ay_k - \varphi$, and the correction, as $w_k = B^{-1}r_k$. Then, the iteration process (3.28) can be written as

$$y_{k+1} = y_k - \tau_{k+1} w_k, \qquad k = 0, 1, \ldots.$$

It seems reasonable to choose the iteration parameter τ_{k+1} from the condition of minimum norm of the inaccuracy of z_{k+1} in H_D. Direct calculations show that the minimum norm is attained with

$$\tau_{k+1} = \frac{(Dw_k, z_k)}{(Dw_k, w_k)}. \tag{3.37}$$

The choice of a most appropriate iteration method can be achieved through the choice of $D = D^* > 0$. The latter choice must obey, in particular, the condition that calculation of iteration parameters is indeed possible. Formula (3.37) involves an uncomputable quantity z_k, and the simplest choice $D = B$ (see Theorem 3.5) cannot be made here. The second noted possibility $D = A$ leads us to the *steepest descend method*, in which case

$$\tau_{k+1} = \frac{(w_k, r_k)}{(Aw_k, w_k)}. \tag{3.38}$$

Among other possible choices of D, the case of $D = AB^{-1}A$ is worth noting *minimum correction method*, in which

$$\tau_{k+1} = \frac{(Aw_k, w_k)}{(B^{-1}Aw_k, Aw_k)}.$$

The two-layer variation-type iteration method converges not slower than the simple iteration method. Let us formulate the latter result with reference to the steepest descend method.

Theorem 3.7 *The iteration method* (3.28), (3.30), (3.31), (3.38) *converges in* H_A, *and for the total number* n *of iterations required for an accuracy* ε *to be achieved, estimate* (3.33) *holds.*

In the computational practice, the most widely used are three-layered variation-type iteration methods. In the rate of convergence, these methods are at least as efficient as the iteration methods with the Chebyshev set of iteration parameters.

In the three-layered (two-sweep) iteration method a next approximation is to be found from the two previous approximations. For the method to be realized, two initial approximations, y_0 and y_1, are necessary. Normally, the approximation y_0 can be given arbitrarily, and the approximation y_1 can be found using the two-layer iteration method. The three-layered method can be written in the following *canonical form for the three-layer iteration method*:

$$\begin{aligned} By_{k+1} &= \alpha_{k+1}(B - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})By_{k-1} + \alpha_{k+1}\tau_{k+1}\varphi, \\ &\qquad k = 1, 2, \ldots, \\ By_1 &= (B - \tau_1 A)y_0 + \tau_1\varphi. \end{aligned} \tag{3.39}$$

Here, α_{k+1} and τ_{k+1} are iteration parameters.

Calculations by formula (3.39) are based on the representation

$$y_{k+1} = \alpha_{k+1}y_k + (1 - \alpha_{k+1})y_{k-1} - \alpha_{k+1}\tau_{k+1}w_k,$$

where, recall, $w_k = B^{-1}r_k$.

In the *conjugate gradient method* intended for calculation of iteration parameters in the three-layer iteration method (3.39), the following formulas are used:

$$\begin{aligned} \tau_{k+1} &= \frac{(w_k, r_k)}{(Aw_k, w_k)}, \qquad k = 0, 1, \ldots, \\ \alpha_{k+1} &= \Big(1 - \frac{\tau_{k+1}}{\tau_k}\,\frac{(w_k, r_k)}{(w_{k-1}, r_{k-1})}\,\frac{1}{\alpha_k}\Big)^{-1}, \quad k = 1, 2, \ldots, \ \alpha_1 = 1. \end{aligned} \tag{3.40}$$

The conjugate gradient method is a method that has gained a most widespread use in computational practice.

Theorem 3.8 *Let conditions* (3.30), (3.31) *be fulfilled. Then the conjugate gradient method* (3.39), (3.40) *converges in* H_A, *and for the total number* n *of iterations required for an accuracy* ε *to be achieved, estimate* (3.36) *holds.*

We gave some reference results concerning iteration solution methods for problem (3.24) with self-adjoint operators A and B (conditions (3.30)). In many applied problems one has to treat more general problems with non-self-adjoint operators. A typical example here is convection-diffusion problems dealt with in continuum mechanics.

3.3.5 Iteration methods with diagonal reconditioner

A second main point in the theory of iteration methods is the choice of *reconditioner* for the operator B in (3.28). We will outline here some major possibilities available along this line. We will not consider general problems (3.24) with non-self-adjoint operators, inviting the reader to address special literature.

A simplest class of iteration methods for solving problem (3.24) can be identified considering the choice of the diagonal operator B. In the latter case, we have

$$B = b(\boldsymbol{x})E, \tag{3.41}$$

and the next iteration approximation can be calculated by explicit formulas. To this class of methods, the *Jacobi iteration method* belongs, which arises when we choose as B the diagonal part of A and $\tau_k = \tau = 1$.

The problem of the optimal choice of B in the class of operators (3.41) has been already solved. The ratio $\xi = \gamma_1/\gamma_2$ in the two-sided operator equality (3.31) for $A = A^* > 0$ will be maximal if we choose, as B, the diagonal part of A. In this respect, the Jacobi method is optimal.

An estimate of efficiency in the used iteration method attaches concrete meaning to the energy equivalence constants γ_1 and γ_2. Let us illustrate the latter with the example of solution of the difference Dirichlet problem for the Poisson equation (3.18), (3.19).

On the basis of Lemmas 3.1 and 3.2, we have:

$$\gamma_1 E \le A \le \gamma_2 E,$$

where

$$\gamma_1 = \frac{8}{l_1^2} + \frac{8}{l_2^2}, \qquad \gamma_2 = \frac{4}{h_1^2} + \frac{4}{h_2^2}. \tag{3.42}$$

For the problem under consideration, we have

$$b(\boldsymbol{x}) = \frac{2}{h_1^2} + \frac{2}{h_2^2} \tag{3.43}$$

and therefore

$$\xi = \frac{\gamma_1}{\gamma_2} = \mathcal{O}(|h|^2), \qquad |h|^2 = h_1^2 + h_2^2.$$

For the number of iterations in the simple iteration method with the optimal value $\tau = \tau_0 = 2/(\gamma_1 + \gamma_2)$ and for the number of iterations in the steepest descend method with the operator B chosen according to (3.41), (3.43) with allowance for (3.33), (3.42), the following estimate holds:

$$n_0(\varepsilon) = \mathcal{O}\Big(\frac{1}{|h|^2} \ln \frac{1}{\varepsilon}\Big). \tag{3.44}$$

Thus, the number of iterations varies in proportion to the total number of nodes (unknowns).

For the Chebyshev iteration method and for the conjugate gradient method, instead of (3.44), we have the estimate

$$n_0(\varepsilon) = \mathcal{O}\Big(\frac{1}{|h|}\,\ln\frac{1}{\varepsilon}\Big). \tag{3.45}$$

Compared to the simple iteration method, the method with the Chebyshev set of iteration parameters converges much faster.

3.3.6 Alternate-triangular iteration methods

Consider the problem (3.24) in the case in which the self-adjoint, positive operator A can be represented as

$$A = A_1 + A_2, \qquad A_1 = A_2^*. \tag{3.46}$$

Let the operator D correspond to the diagonal part of A, and the operator L to the subdiagonal matrix. Then, by virtue of $A = L + D + L^*$, the decomposition (3.46) for the operators A_α, $\alpha = 1,\ 2$ yields

$$A_1 = \frac{1}{2}\,D + L, \qquad A_2 = \frac{1}{2}\,D + L^*. \tag{3.47}$$

For the model problem (3.18), (3.19), we have:

$$Dy = d(\boldsymbol{x})E, \qquad d(\boldsymbol{x}) = \frac{2}{h_1^2} + \frac{2}{h_2^2}. \tag{3.48}$$

For A_α, $\alpha = 1,\ 2$, the expansions (3.46), (3.47) have the following representations on the set of mesh functions vanishing on γ_h:

$$A_1 y = \frac{1}{h_1}\,y_{\bar{x}_1} + \frac{1}{h_2}\,y_{\bar{x}_2}, \qquad A_2 y = -\frac{1}{h_1}\,y_{x_1} - \frac{1}{h_2}\,y_{x_2}. \tag{3.49}$$

In the alternate-triangular method, the operator B is a *factorized* operator chosen as the product of two triangular matrices and one diagonal matrix:

$$B = (D + \omega A_1)\,D^{-1}\,(D + \omega A_2), \tag{3.50}$$

where ω is a numerical parameter.

The rate of convergence in the iteration method (3.29), (3.30), (3.50) is defined by the energy equivalence constants γ_1 and γ_2 in the two-sided inequality (3.31). Find these constants if a priori information is given in the form of inequalities

$$\delta_1 D \le A, \qquad A_1 D^{-1} A_2 \le \frac{\delta_2}{4}\,A, \qquad \delta_1 > 0. \tag{3.51}$$

For the operator B, we have

$$\begin{aligned} B &= (D + \omega A_1)\,D^{-1}\,(D + \omega A_2) \\ &= D + \omega\,(A_1 + A_2) + \omega^2 A_1 D^{-1} A_2. \end{aligned} \tag{3.52}$$

With inequalities (3.51) taken into account, we obtain

$$B \le \Big(\frac{1}{\delta_1} + \omega + \omega^2 \frac{\delta_2}{4}\Big) A.$$

Hence, we have the following expression for γ_1:

$$\gamma_1 = \frac{\delta_1}{1 + \omega\delta + \omega^2 \delta_1 \delta_2 / 4}. \tag{3.53}$$

To evaluate the constant γ_2, we represent the operator B, based on (3.52), in the form

$$\begin{aligned} B &= D - \omega\,(A_1 + A_2) + \omega^2 A_1 D^{-1} A_2 + 2\omega\,(A_1 + A_2) \\ &= (D - \omega A_1)\, D^{-1}\, (D - \omega A_2) + 2\omega A. \end{aligned}$$

Since the operator D is positive, we obtain $(By, y) \ge 2\omega\,(Ay, y)$, i. e. $A \le \gamma_2 B$, where

$$\gamma_2 = \frac{1}{2\omega}. \tag{3.54}$$

Now we can choose the value of ω in (3.50) based on the maximum condition for $\xi = \xi(\omega) = \gamma_1/\gamma_2$. In view of (3.53) and (3.54), we obtain

$$\xi(\omega) = \frac{\gamma_1}{\gamma_2} = \frac{2\omega\delta_1}{1 + \omega\delta + \omega^2 \delta_1 \delta_2 / 4}.$$

The maximum of $\xi(\omega)$ is attained at

$$\omega = \omega_0 = 2\,(\delta_1 \delta_2)^{-1/2}, \tag{3.55}$$

to equal

$$\xi = \xi(\omega_0) = \frac{2\eta^{1/2}}{1 + \eta^{1/2}}, \qquad \eta = \frac{\delta_1}{\delta_2}. \tag{3.56}$$

Based on the obtained estimates, the following statement about the convergence of the alternate-triangular iteration method with the optimal value of ω can be formulated:

Theorem 3.9 *The alternate-triangular iteration method* (3.29), (3.46), (3.50), (3.51), (3.55) *with the Chebyshev set of iteration parameters converges in H_A and H_B, and for the total number of iterations the estimate* (3.36) *with ξ given by* (3.56) *holds.*

Remark 3.10 Taking the smallness of η into account, for the total number of iterations we can obtain a simpler expression:

$$n_0(\varepsilon) = \frac{\ln(2\varepsilon^{-1})}{2\sqrt{2}\,\eta^{1/4}}, \qquad \eta = \frac{\delta_1}{\delta_2}. \tag{3.57}$$

Remark 3.11 The alternate-triangular method can be realized as a conjugate gradient method. In the latter case the total number of iterations obeys the estimate (3.57) as well.

Let us find now the constants δ_1 and δ_2 in equation (3.51) for the model problem (3.18), (3.19). Based on the choice of (3.48) and estimates (3.42) and (3.43), we have $\delta_1 = \mathcal{O}(|h|^2)$. With relations (3.48) and (3.49) taken into account, we obtain

$$(A_1 D^{-1} A_2 y,\, y) = \frac{h_1^2 h_2^2}{2\,(h_1^2 + h_2^2)}\,(A_2 y,\, A_2 y).$$

For the right-hand side we have

$$(A_2 y,\, A_2 y) = \frac{1}{h_1^2}\,(y_{x_1}^2,\, 1) - \frac{2}{h_1 h_2}\,(y_{x_1},\, y_{x_2}) + \frac{1}{h_2^2}\,(y_{x_2}^2,\, 1).$$

On account of the equality

$$-\frac{2}{h_1 h_2}\,(y_{x_1},\, y_{x_2}) \;=\; -2\Big(\frac{1}{h_2}\,y_{x_1},\, \frac{1}{h_1}\,y_{x_2}\Big) \;\le\; \frac{1}{h_2^2}\,(y_{x_1}^2,\, 1) + \frac{1}{h_1^2}\,(y_{x_2}^2,\, 1),$$

we obtain

$$(A_2 y,\, A_2 y) \;\le\; \Big(\frac{1}{h_1^2} + \frac{1}{h_2^2}\Big)(Ay,\, y).$$

In this way, we arrive at the inequality $(A_1 D^{-1} A_2 y,\, y) \le 0.5(Ay,\, y)$. Comparing the latter inequality with the second inequality in (3.51), we obtain: $\delta_2 = 2$.

Hence, for the Chebyshev iteration method the following estimate for the total number of iterations holds:

$$n_0(\varepsilon) = \mathcal{O}\Big(\frac{1}{|h|^{1/2}}\,\ln\frac{1}{\varepsilon}\Big). \tag{3.58}$$

Thus, the total number of iterations varies in proportion to the square root of the number of nodes in one direction (in the considered two-dimensional problem, to the fourth root of the total number of nodes). Estimate (3.58) shows that the considered method is more preferable than the Jacobi method.

The parameter ω in the alternate-triangular method (3.29), (3.50) can be included into the operator D in order the method (3.29) could be used with $B = (D + A_1)\,D^{-1}\,(D + A_2)$. Here, the optimization of the iteration method is achieved through a proper choice of D only. The latter is especially important in the case of problems with varied coefficients. Worthy of note is the alternate-triangular iteration method in the form

$$B = (D + L)\,D^{-1}\,(D + L^*). \tag{3.59}$$

With θD taken as D, where θ is a constant and D is the diagonal part of A, the choice of the reconditioner in the form (3.59) is equivalent to the choice considered previously. Of course, if $D \ne \theta D$, then we have no equivalence between these versions of the alternate-triangular method.

3.4 Program realization and numerical examples

The model Dirichlet problem in a rectangle for the second-order self-adjoint elliptic equation with variable coefficients is considered. To approximately solve this problem, we use the alternate-triangular iteration method of approximate factorization. The program and the calculation results are presented.

3.4.1 Statement of the problem and the difference scheme

In the rectangle

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2), \quad 0 < x_\alpha < l_\alpha, \quad \alpha = 1, 2\}$$

we solve the Dirichlet problem for the second-order elliptic equation:

$$-\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big) + q(\boldsymbol{x})u = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \tag{3.60}$$

$$u(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega. \tag{3.61}$$

We assume that

$$k(\boldsymbol{x}) \geq \kappa > 0, \qquad q(\boldsymbol{x}) \geq 0, \qquad \boldsymbol{x} \in \Omega.$$

In Ω we introduce a grid, uniform in both directions, with step sizes h_α, $\alpha = 1, 2$. We define the set of inner nodes

$$\omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2), \quad x_\alpha = i_\alpha h_\alpha, \quad i_\alpha = 1, 2, \ldots, N_\alpha - 1, \quad N_\alpha h_\alpha = l_\alpha, \quad \alpha = 1, 2\}$$

and let $\partial\omega$ be the set of boundary nodes.

To the differential problem (3.60), (3.61), we put into correspondence the difference problem

$$-(a^{(1)} y_{\bar{x}_1})_{x_1} - (a^{(2)} y_{\bar{x}_2})_{x_2} + c(\boldsymbol{x})y = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \tag{3.62}$$

$$y(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega. \tag{3.63}$$

The coefficients and the right-hand side in (3.62) can be calculated by the simplest formulas

$$a^{(1)}(\boldsymbol{x}) = k(x_1 - 0.5h_1, x_2), \qquad a^{(2)}(\boldsymbol{x}) = k(x_1, x_2 - 0.5h_2),$$

$$c(\boldsymbol{x}) = q(\boldsymbol{x}), \qquad \varphi(\boldsymbol{x}) = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega.$$

If necessary, more complex expressions can be used.

3.4.2 A subroutine for solving difference equations

We start with a brief description of the subroutine `SBAND5` that solves linear systems with symmetric five-diagonal matrices. The subroutine was developed and its description prepared by M. M. Makarov[1].

The subroutine intention. The subroutine `SBAND5` is intended for approximate solution of systems of linear equations with a non-generate symmetric matrix of special form whose nonzero elements are contained only in five diagonals of the matrix, namely, in the principal diagonal, in the two adjacent diagonals (sub- and superdiagonals), and in the two next (remote) diagonals located symmetrically about the principal diagonal. Such matrices are a particular case of banded symmetric matrices (with the lower half-width of the band equal to the upper half-width and to the distance between the principal diagonal and either of the remote diagonals).

Brief information about the solution method. The subroutine `SBAND5` embodies an implicit iteration method of conjugate coefficients. The iteration scheme is constructed based on the condition of minimal inaccuracy of the kth approximation in the energy norm with linear refinement of the current iteration approximation for correction vector from the Krylov subspace of dimension k. To determine the correction vector and subsequent iteration refinement, two iteration parameters are calculated at each iteration step, expressed through scalar products of the iteration vectors. The method is implicit in the sense that, at each iteration step, one has to solve a system of linear equations with a matrix (called reconditioner) close in a sense to the initial matrix of the system and, as a matter of fact, constructed from this matrix. Due to special choice of reconditioner, the subroutine uses an algorithm that contains no multiplication of the initial matrix by a vector, which operation, generally speaking, needs implicit methods to be applied alongside with the inversion of the reconditioner matrix. In the case of a symmetric, positive definite matrix of the system and a symmetric, positive definite reconditioner, the method converges in the energy space.

Algebraic formulation of the problem and the difference interpretation. For an $n \times n$ five-diagonal symmetric matrix A, we introduce the following designations:

`A0` — elements in the principal diagonal taken in the increasing order of row subscripts;

`A1R` — elements in the right diagonal adjacent to the principal diagonal and taken with the opposite sign;

`A2R` — elements in the right remote diagonal taken with the opposite sign.

By convention, the notation

$$Ay = f$$

[1] An earlier version of the subroutine (`SOLVE1`) was described in the book *A. A. Samarskii and P. N. Vabishchevich*, Computational Heat Transfer. Vol. 2. The Finite Difference Methodology. Chichester, Wiley, 1995.

contains no elements on the diagonals A0, A1R, and A2R going beyond the matrix, and no components of the vector y by which these elements multiply. Next, each of the equations can be represented as

$$-\mathrm{A2R}_{j-l}y_{j-l} - \mathrm{A1R}_{j-1}y_{j-1} + \mathrm{A0}_j y_j - \mathrm{A1R}_j y_{j+1} - \mathrm{A2R}_j y_{j+l} = f_j, \quad j = 1, 2, \ldots, n,$$

where l is the band half-width of the matrix.

Systems with similar matrices arise, for instance, in the approximation of elliptic partial differential equations on five-point mesh patterns in orthogonal coordinate systems. If the grid nodes in the enveloping mesh rectangle are ordered from bottom to top in horizontal lines and from left to right in each line, then with the number of nodes in each line equal to l the elements $-\mathrm{A2R}_{j-l}$, $-\mathrm{A1R}_{j-1}$, $\mathrm{A0}_j$, $-\mathrm{A1R}_j$ and $-\mathrm{A2R}_j$ are in fact the coefficients of the difference equation at the jth node for the mesh-function components located respectively at the bottom, left, central, right, and upper nodes of the five-point mesh pattern.

Remark 3.12 The calculation domain is not necessarily a rectangle in some coordinate system; yet, for the subroutine SBAND5 to be used, in writing the difference scheme in the enveloping mesh rectangle the difference equations at all fictitious nodes must be given with coefficients

$$\mathrm{A0}_j = 1, \qquad \mathrm{A1R}_j = \mathrm{A2R}_j = 0,$$

and, as the corresponding components in the initial approximation and in the right-hand side, zeros must be transferred.

Remark 3.13 In interpreting the equations of the system as difference equations it can be expected that at the grid nodes lying on the boundary of the mesh rectangle some of the coefficients must be zero, for instance, the coefficients $\mathrm{A1R}_j$ at the nodes on the right boundary. It is due to this "naturalness" that this remark draws users' attention to the fact that the subroutine SBAND5 allocates no zero elements into the input coefficient array provided that these elements are elements that lie on diagonals not going beyond the matrix. In line with the data structure adopted in SBAND5, the input array can involve diagonal elements that lie outside the matrix. This assumption simplifies the setting of input information as it makes possible to treat all mesh points as inner nodes and all diagonals as having identical lengths. Into the "extra" diagonal components, arbitrary values can be recorded since in SBAND5 these components will all the same be put equal to zero.

Remark 3.14 All matrix elements that do not belong to the principal diagonal are transferred into the subroutine SBAND5 with the opposite sign, and all elements on the principal diagonal must be positive.

Description of subroutine parameters. The heading of the subroutine SBAND5 looks as

```
SUBROUTINE SBAND5 ( N, L, A, Y, F, EPSR, EPSA )
IMPLICIT REAL*8 ( A-H, O-Z )
DIMENSION A(1), Y(N), F(N)
COMMON / SB5 /     IDEFAULT, I1R, I2R, IFREE
COMMON / CONTROL / IREPT, NITER
```

Formal parameters in the subroutine SBAND5:

N — total number of equations in the system, $N > 2$;

L — half-width of the band, $1 < L < N$ (this half-width is defined for a banded matrix as the maximum number of elements in the matrix rows between the element lying in the principal diagonal and the nonzero element most remote from the principal diagonal, including the last and excluding the first element);

A — array that contains at the input the coefficients A0, A1R, A2R and has sufficient length for allocation of the vectors to be calculated and stored at iteration steps (methods used in setting the input information and an estimate of the required array length are outlined below);

Y — array of length not shorter than N that contains at the input in the first N components an initial approximation, and at the output, in this place, the final iteration approximation;

F — array of length not shorter than N that contains at the input in the first N components the right-hand side vector of the system;

EPSR — desired relative accuracy of iteration approximation to the solution; this quantity is used in SBAND5 in the criterion for termination of the iteration process which uses the ratio between the initial and current inaccuracies of some norm;

EPSA — desired absolute accuracy of iteration approximation to the solution; this quantity is used in the criterion for termination of the iterative process, that uses the value of some norm of the current accuracy.

All the above parameters are input parameters. The parameter Y is also an output parameter. The values of N, L, EPSR, EPSA remain unchanged at the output, and the value of A and F at the exit from the subroutine are not defined.

Parameters in the common block /SB5/ of the subroutine SBAND5:

IDEFAULT — input parameter generalizing the flag pointing to the necessity to analyze the remaining values in the common block /SB5/: if this parameter is zero, then specification of all remaining parameters in the common block is unnecessary since the values of these parameters all the same will be ignored in the subroutine SBAND5; otherwise, all parameters in the common block must be specified;

I1R — indicator showing the position of the coefficients A1R in the array A (the description is given below);

I2R — indicator showing the position of the coefficients A2R in the array A;

IFREE — indicator showing the position of the free segment in the array A.

The parameters in the common block /SB5/ specify the method of setting the system coefficients in the input array A. A given value IDEFAULT = 0 indicates that the coefficients in the array appear in the order adopted in the subroutine SBAND5. Otherwise, allocation of coefficients must be specified by parameters in the common block /SB5/.

Parameters in the common block /CONTROL/ of the subroutine SBAND5:

IREPT — indicator of the necessity to perform actions that must be executed following the first call for the subroutine SBAND5 with the given matrix; the user must set the value IREPT = 0 prior to the first call with the given matrix, and the subroutine SBAND5, in its turn, assigns the value IREPT = 1 to this variable; to solve a system with the same matrix but with another right-hand side, the user must call for SBAND5 with newly set values of Y, F but with the same values of A;

NITER — the number of iterations at the output.

Description of data structure and an estimate of required memory capacity. The data structure in the subroutine SBAND5 obeys the concept of through-programming for all operations used to treat matrices and vectors. This concept consists in the possibility to program all operations in one cycle, without specially treating vector and matrix components corresponding to "incomplete" rows of the matrix (in the case of interest, to first L and to last L rows, in which, in turn, the first and last rows would have necessitated special consideration). In the difference-equation-on-grid interpretation, this implies treating all mesh points as inner nodes. This concept makes the program code more concise and easy to read, with each linear or matrix numerical operation realized in one cycle.

In the description of the order of setting the coefficients in the input array A it will be indicated which components in the subroutine SBAND5 are recorded as zeros. The common block /SB5/ provides the user with freedom in setting the coefficients. It is necessary that the coefficients be allocated in the diagonal-by-diagonal manner using one and the same ordering method. It must also be guaranteed that, in writing zeros, the values in other diagonals within the matrix remained unchanged.

Setting IDEFAULT = 0 at the input, the user can omit setting of all other values in the common block /SB5/. In the latter case, the input array A must contain in the first 3*N components the following values:

A(I), I = 1,2,...,N — coefficients A0;

A(I1R+I), I1R = N, I = 1,2,...,N−1 — coefficients A1R within the matrix, the values A(I1R+N) = 0 being assigned;

A(I2R+I), I2R = 2*N, I = 1,2,...,N − L — coefficients A2R within the matrix, the values A(I2R+I) = 0, I = N−L+1, N−L+2,...,N being assigned.

The coefficients A0 must always be allocated in first components of the array A. The coefficients in other diagonals can be recorded as arbitrary components of A. In the latter case it is necessary, however, in the inversion to preset some nonzero

value of IDEFAULT and some values of the indicators I1R and I2R, and also a value of IFREE such that 3*N+2*L components of A, starting from the (IFREE+1)-th component, could be considered independent, available for use in the subroutine SBAND5 as a storage for the iteration vectors.

Remark 3.15 In setting the indicator IFREE never causes a memory allocation error such that into the components of the matrix, following the IFREE-th component, zero values are recorded (for instance, the diagonal A2R is allocated last, and the value IFREE = I2R+N−L is set).

The minimum required length of A can be estimated as follows. With the allocation pattern for coefficients adopted in the subroutine SBAND5, the array must be not shorter than 6*N+L. If another allocation pattern for coefficients is adopted, then 3*N+2*l components are necessary, located behind the components that carry the coefficients.

Example illustrating the use of the subroutine SBAND5. In the program EXSB5, the coefficients of a 10×10 system matrix with band half-width equal to 5, the zero initial approximation, and the right-hand side of the system are set. Then, the subroutine SBAND5 is called for to solve the system, and the obtained approximate solution is printed out (together with a process protocol, in which the relative accuracy achieved at the current iteration step is indicated).

```
      PROGRAM EXSB5
C
C     EXSB5 - EXAMPLE OF SUBROUTINE CALL FOR THE SUBROUTINE SBAND5
C
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     THE SYSTEM MATRIX, THE SOLUTION VECTOR,
C     AND THE RIGHT-HAND SIDE VECTOR ARE
C
C     (  6  -1              -2                  )    (  1 )    ( -8 )
C     ( -1   6  -1              -2              )    (  2 )    ( -6 )
C     (     -1   6  -1              -2          )    (  3 )    ( -4 )
C     (         -1   6  -1              -2      )    (  4 )    ( -2 )

C     (             -1   6  -1              -2  )    (  5 )    (  0 )
C     ( -2              -1   6  -1              )  * (  6 ) = ( 22 )
C     (     -2              -1   6  -1          )    (  7 )    ( 24 )
C     (         -2              -1   6  -1      )    (  8 )    ( 26 )
C     (             -2              -1   6  -1  )    (  9 )    ( 28 )
C     (                 -2              -1   6  )    ( 10 )    ( 41 )
C
C
      PARAMETER ( N = 10, L = 5 )
      DIMENSION A(6*N+L), Y(N), F(N)
      COMMON / SB5 /     IDEFAULT(4)
      COMMON / CONTROL / IREPT, NITER
C
```

```
C     SETTING OF MATRIX DIAGONALS, ZERO INITIAL APPROXIMATION,
C     AND RIGHT-HAND SIDE OF THE SYSTEM:
C
      IDEFAULT(1) = 0
      IREPT       = 0
      DO 1 I = 1, 10
        A(I)     = 6.D0
        A(N+I)   = 1.D0
        A(2*N+I) = 2.D0
        Y(I)     = 0.D0
    1 CONTINUE
      F( 1) = -8.D0
      F( 2) = -6.D0
      F( 3) = -4.D0
      F( 4) = -2.D0
      F( 5) =  0.D0
      F( 6) = 22.D0
      F( 7) = 24.D0
      F( 8) = 26.D0
      F( 9) = 28.D0
      F(10) = 41.D0
C
C     SUBROUTINE CALL FOR SOLVING THE SYSTEM:
C
      EPSR = 1.D-6
      EPSA = 1.D-7
      CALL SBAND5 ( N, L, A, Y, F, EPSR, EPSA )
C
C     PRINTING THE SOLUTION VECTOR:
C
      WRITE ( 06, * ) '  S O L U T I O N:'
      DO 2 I = 1, 10

        WRITE ( 06, * ) Y(I)
    2 CONTINUE
      END
```

Below, the full text of the subroutine `SBAND5` is given.

Subroutine SBAND5

```
      SUBROUTINE  SBAND5 ( N, L, A, Y, F, EPSR, EPSA )
C
      IMPLICIT  REAL*8 ( A-H, O-Z )
      DIMENSION  A(1), Y(N), F(N)
C
      COMMON / SB5 / IDEFAULT,
     *               I1R, I2R,
     *               IFREE
C
      COMMON / CONTROL / IREPT, NITER
C
      IF ( IDEFAULT .EQ. 0 ) THEN
        I1R = N
        I2R = 2*N
        IG  = 3*N
```

```
      ELSE IF ( IDEFAULT .NE. 0 ) THEN
        IG  = IFREE + L
      END IF
C
      IT  = IG  + N
      ITL = IT  + L
      IQ  = ITL + N
C
      IF ( IREPT .EQ. 0 ) THEN
C
      A(I1R+N) = 0.D0
      DO 9 J = 1, L
        A(I2R+J+N-L) = 0.D0
        A(IT+J)      = 0.D0
    9 CONTINUE
C

      DO 10 J = 1, N
        G = A(J) - A(I1R+J-1)*A(ITL+J-1)
     *           - A(I2R+J-L)*A(ITL+J-L)
        G = 1.D0 / G
        A(IG+J) = DSQRT(G)
        A(ITL+J) = G*( A(I1R+J) + A(I2R+J) )
   10 CONTINUE
C
      DO 20 J = 1, N
        A(J)     = A(IG+J)*A(J)*A(IG+J) - 2.D0
        A(I1R+J) = A(IG+J)*A(I1R+J)*A(IG+J+1)
        A(I2R+J) = A(IG+J)*A(I2R+J)*A(IG+J+L)
   20 CONTINUE
C
      IREPT = 1
C
      END IF
C
      DO 30 J = N, 1, -1
        A(ITL+J) = Y(J) / A(IG+J)
        Y(J)     = A(ITL+J) - A(I1R+J)*A(ITL+J+1) - A(I2R+J)*A(ITL+J+L)
        A(ITL+J-L) = 0.D0
   30 CONTINUE
C
      RR0 = 0.D0
      DO 40 J = 1,N
        G = A(ITL+J)
        A(ITL+J) = F(J)*A(IG+J) - A(J)*G - Y(J)
     *           + A(I1R+J-1)*A(ITL+J-1) + A(I2R+J-L)*A(ITL+J-L)
        F(J)     = A(ITL+J) - G
        RR0 = RR0 + F(J)*F(J)
   40 CONTINUE
C
      NIT    = 0
      EPSNIT = 1.D0
      RR     = RR0
C
   50 CONTINUE
C
      IF ( EPSNIT .GE. EPSR .AND. RR .GE. EPSA ) THEN

C
      IF ( NIT .EQ. 0 ) THEN
```

```
          RRI = 1.D0 / RR
          DO 60 J = N, 1, -1
            A(IQ+J)  = F(J)
            A(ITL+J) = A(IQ+J) + A(I1R+J)*A(ITL+J+1) + A(I2R+J)*A(ITL+J+L)

            A(ITL+J-L) = 0.D0
    60    CONTINUE
       ELSE IF ( NIT .GE. 1 ) THEN
          BK  = RR*RRI
          RRI = 1.D0 / RR
          DO 61 J = N, 1, -1
            A(IQ+J)  = F(J) + BK*A(IQ+J)
            A(ITL+J) = A(IQ+J) + A(I1R+J)*A(ITL+J+1) + A(I2R+J)*A(ITL+J+L)
            A(ITL+J-L) = 0.D0

    61    CONTINUE
       END IF
C
       TQ = 0.D0
       DO 70 J = 1, N
         G = A(ITL+J)


         A(ITL+J) = A(IQ+J) + A(J)*G + A(I1R+J-1)*A(ITL+J-1)
      *                              + A(I2R+J-L)*A(ITL+J-L)
         A(IT+J)  = G + A(ITL+J)
         TQ = TQ + A(IT+J)*A(IQ+J)
    70 CONTINUE
C
       AK = RR / TQ
C
       RR = 0.D0
       DO 80 J = 1,N
         Y(J) = Y(J) + AK*A(IQ+J)
         F(J) = F(J) - AK*A(IT+J)
         RR = RR + F(J)*F(J)
    80 CONTINUE
C
       NIT    = NIT + 1
       EPSNIT = DSQRT( RR/RR0 )

C
       GO TO 50
C
       END IF
C
       DO 90 J = N, 1, -1
         A(ITL+J) = Y(J) + A(I1R+J)*A(ITL+J+1) + A(I2R+J)*A(ITL+J+L)
         Y(J)     = A(IG+J)*A(ITL+J)
    90 CONTINUE
       NITER = NIT
C
       RETURN
       END
```

3.4.3 Program

To approximately solve the boundary value problem (3.60), (3.61), the following program was used.

Program PROBLEM2

```
C
C     PROBLEM2 - DIRICHLET BOUNDARY-VALUE PROBLEM FOR THE ELLIPTIC
C                EQUATION WITH VARIABLE COEFFICIENTS IN RECTANGLE
C

      IMPLICIT REAL*8 ( A-H, O-Z )
C
      PARAMETER ( N1 = 101, N2 = 101 )
C
      DIMENSION A(7*N1*N2), Y(N1,N2), F(N1,N2),
     *          BL(N2), BR(N2), BB(N1), BT(N1)
      COMMON / SB5 /     IDEFAULT(4)
      COMMON / CONTROL / IREPT, NITER
C
C     THE ARRAY A MUST BE SUFFICIENTLY LONG TO STORE
C     THE COEFFICIENTS OF THE SYMMETRIC MATRIX OF THE
C     DIFFERENCE PROBLEM, THE SET PRINCIPAL DIAGONAL A0,
C     THE SUPERDIAGONAL A1, THE UPPER REMOTED DIAGONAL A2,
C     THE SOLUTION VECTOR Y, THE RIGHT-HAND SIDE VECTOR F,
C     AND THE VECTORS USED IN THE ITERATION ALGORITHM
C     FOR SOLVING THE DIFFERENCE EQUATION
C     ( SEE SUBROUTINE SBAND5 ).
C
C     DATA INPUT:
C
C     X1L, X2L  - COORDINATES OF THE LEFT BOTTOM CORNER
C                 OF THE RECTANGULAR CALCULATION DOMAIN;
C     X1R, X2R  - COORDINATES OF THE RIGHT UPPER CORNER;
C     X1D, X2D  - COORDINATES OF THE RIGHT UPPER CORNER
C                 OF SUBDOMAIN D2;
C     N1, N2    - NUMBER OF GRID NODES OVER THE CORRESPONDING
C                 DIRECTIONS
C     EPSR      - DESIRED RELATIVE ACCURACY OF THE ITERATIVE
C                 APPROXIMATION TO THE SOLUTION.
C     EPSA      - DESIRED ABSOLUTE ACCURACY OF THE ITERATIVE
C                 APPROXIMATION TO THE SOLUTION.
C
      X1L   = 0.D0
      X1R   = 1.D0
      X2L   = 0.D0
      X2R   = 1.D0

      EPSR  = 1.D-6
      EPSA  = 0.D0
C
      N = N1*N2
      DO I = 1, 7*N
        A(I) = 0.D0
      END DO
C
C     BOUNDARY CONDITIONS
```

```
C
      DO J = 1, N2
        BL(J) = 0.D0
        BR(J) = 0.D0
      END DO
      DO I = 1, N1
        BB(I) = 0.D0
        BT(I) = 0.D0
      END DO
C
C     TO BE FOUND IN THE SUBROUTINE FDS_EL ARE THE CENTRAL, RIGHT,
C     AND UPPER COEFFICIENTS OF THE DIFFERENCE SCHEME
C     ON THE FIVE-POINT MESH PATTERN (A0, A1, AND A2, RESPECTIVELY),
C     AND THE RIGHT-HAND SIDE F.
C

      CALL FDS_EL ( X1L, X1R,  X2L, X2R, N1, N2, BL, BR, BB, BT,
     *                  H1, H2, A(1), A(N+1), A(2*N+1), F )
C
C     THE SUBROUTINE SBAND5 SOLVES THE SOLUTION OF THE DIFFERENCE
C     PROBLEM BY THE ALTERNATE-TRIANGULAR APPROXIMATE
C     FACTORIZATION - CONJUGATE GRADIENT METHOD
C
C
      IDEFAULT(1) = 0
      IREPT       = 0
C

C     INITIAL APPROXIMATION
C
      DO I = 1, N1
        DO J = 1, N2
          Y(I,J) = 0.D0
        END DO

      END DO
      CALL SBAND5 ( N, N1, A, Y, F, EPSR, EPSA )
C
      OPEN  ( 01, FILE = 'RESULT.DAT' )
      WRITE ( 01, * ) NITER
      WRITE ( 01, * ) ((Y(I,J),I=1,N1),J=1,N2)
      CLOSE ( 01 )
C
      STOP
      END
```

Among the main components of the program, the subroutine `FDS_EL`, used to set the difference-problem coefficients, deserves mention.

```
      SUBROUTINE FDS_EL ( X1L, X1R,  X2L, X2R, N1, N2, BL, BR, BB, BT,
     *                  H1, H2, A0, A1, A2, F )
C
C      SETTING OF DIFFERENCE-SCHEME COEFFICIENTS
```

```
C      FOR SOLVING THE DIRICHLET BOUNDARY-VALUE PROBLEM
C      FOR THE SECOND-ORDER ELLIPTIC EQUATION
C      WITH VARIABLE COEFFICIENTS IN RECTANGLE
C
       IMPLICIT REAL*8 ( A-H, O-Z )
       DIMENSION A0(N1,N2), A1(N1,N2), A2(N1,N2), F(N1,N2),
      *          BL(N2), BR(N2), BB(N1), BT(N1)
C
C      PARAMETERS
C

C      INPUT PARAMETERS:
C
C      X1L, X1R  - LEFT AND RIGHT END POINTS OF THE SEGMENT (THE FIRST
C                  VARIABLE);
C      X2L, X2R  - LEFT AND RIGHT END POINTS OF THE SEGMENT (THE SECOND
C                  VARIABLE);
C      N1, N2    - NUMBER OF NODES OVER THE FIRST AND SECOND VARIABLES;
C      BL(N2)    - BOUNDARY CONDITION AT THE LEFT BOUNDARY (X1 = X1L);
C      BR(N2)    - BOUNDARY CONDITION AT THE RIGHT BOUNDARY (X1 = X1RL);
C      BB(N1)    - BOUNDARY CONDITION AT THE LOWER BOUNDARY (X2 = X2L);
C      BT(N1)    - BOUNDARY CONDITION AT THE UPPER BOUNDARY (X2 = X2RL).
C
C      OUTPUT PARAMETERS:
C
C      H1, H2     - STEPS OF THE UNIFORM RESTANGULAR GRID;
C      A0, A1, A2 - DIFFERENCE-SCHEME COEFFICIENTS
C                   (ON THE FIVE-POINT MESH PATTERN WITH REGARD FOR
C                   SYMMETRY);
C      F          - RIGHT-HAND SIDE OF THE DIFFERENCE SCHEME.
C
C      NOTES:
C
C      THE COEFFICIENTS AND THE RIGHT-HAND SIDE OF THE DIFFERENCE
C      EQUATION ARE SET IN THE SUBROUTINES-FUNCTIONS AK, AQ, AF
C
       H1  = (X1R-X1L) / (N1-1)
       H2  = (X2R-X2L) / (N2-1)
       H12 = H1 / H2
       H21 = H2 / H1
       HH  = H1 * H2
C
C      INTERNAL NODES
C


       DO J = 2, N2-1
         X2 = X2L + (J-1)*H2
         DO I = 2, N1-1
           X1 = X1L + (I-1)*H1
           A1(I-1,J) = H21*AK(X1-0.5D0*H1,X2)
           A1(I,J)   = H21*AK(X1+0.5D0*H1,X2)
           A2(I,J-1) = H12*AK(X1,X2-0.5D0*H2)
           A2(I,J)   = H12*AK(X1,X2+0.5D0*H2)
           A0(I,J)   = A1(I-1,J) + A1(I,J) + A2(I,J-1) + A2(I,J)
      *              + HH*AQ(X1,X2)
            F(I,J)   = HH*AF(X1,X2)
         END DO

       END DO
```

```
C
C     LEFT AND RIGHT BOUNDARIES: FIRST-KIND BOUNDARY CONDITION
C
      DO J = 2, N2-1
        A1(1,J)  = 0.D0
        A2(1,J)  = 0.D0

        A0(1,J)  = 1.D0
         F(1,J)  = BL(J)
        A1(N1-1,J) = 0.D0
        A2(N1,J) = 0.D0
        A0(N1,J) = 1.D0
         F(N1,J) = BR(J)
      END DO
C
C     BOTTOM AND UPPER BOUNDARY: FIRST-KIND BOUNDARY CONDITION
C
      DO I = 2, N1-1
        A1(I,1)  = 0.D0
        A2(I,1)  = 0.D0
        A0(I,1)  = 1.D0
         F(I,1)  = BB(I)
        A1(I,N2) = 0.D0
        A2(I,N2-1) = 0.D0
        A0(I,N2) = 1.D0
         F(I,N2) = BT(I)
      END DO
C
C     LEFT BOTTOM CORNER
C
      A1(1,1) = 0.D0
      A2(1,1) = 0.D0
      A0(1,1) = 1.D0
       F(1,1) = 0.5D0*(BL(1) + BB(1))
C
C     LEFT UPPER CORNER
C
      A1(1,N2) = 0.D0

      A0(1,N2) = 1.D0
       F(1,N2) = 0.5D0*(BL(N2) + BT(1))
C
C     RIGHT BOTTOM CORNER
C
      A2(N1,1) = 0.D0
      A0(N1,1) = 1.D0
       F(N1,1) = 0.5D0*(BB(N1) + BR(1))
C
C     RIGHT UPPER CORNER
C
      A0(N1,N2) = 1.D0
       F(N1,N2) = 0.5D0*(BT(N1) + BR(N2))
C
      RETURN
      END

      DOUBLE PRECISION FUNCTION  AK ( X1, X2 )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      COEFFICIENTS AT THE HIGHER DERIVATIVES
```

```
C
      AK = 1.D0
      IF ( X1 .LE. 0.5D0.AND. X2 .LE. 0.5D0 ) AK = 10.D0
C

      RETURN
      END

      DOUBLE PRECISION FUNCTION  AQ ( X1, X2 )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      COEFFICIENTS AT THE LOWER TERMS OF THE EQUATION
C
      AQ = 0.D0
      IF ( X1 .LE. 0.5D0.AND. X2 .LE. 0.5D0 ) AK = 0.D0
C
      RETURN
      END

      DOUBLE PRECISION FUNCTION  AF ( X1, X2 )
      IMPLICIT REAL*8 ( A-H, O-Z )

C
C      RIGHT-HAND SIDE OF THE EQUATION
C
      AF = 1.D0
      IF ( X1 .LE. 0.5D0.AND. X2 .LE. 0.5D0 ) AF = 0.D0
C
      RETURN
      END
```

In the example given below we consider a problem for the elliptic equation with piecewise-constant coefficients.

3.4.4 Computational experiments

Consider data obtained in experiments on numerical solution of a model boundary value problem. The problem simulates processes in a piecewise-homogeneous medium, with equation coefficients assumed constant in separate subdomains. In the calculation domain Ω, a rectangle Ω_1 is singled out (see Figure 3.3). The boundary value problem (3.60), (3.61) is solved in the case in which

$$k(\boldsymbol{x}), q(\boldsymbol{x}), f(\boldsymbol{x}) = \begin{cases} \kappa_1, q_1, f_1, & \boldsymbol{x} \in \Omega_1, \\ \kappa_2, q_2, f_2, & \boldsymbol{x} \in \Omega_2. \end{cases}$$

The above program listing refers to the basic variant in which

$$l_1 = 1.25, \quad l_2 = 1, \quad s_1 = 0.5, \quad s_2 = 0.5,$$
$$\kappa_1 = 1, \quad \kappa_2 = 10, \quad q_1 = 0, \quad q_2 = 0, \quad f_1 = 1, \quad f_2 = 0,$$

and the boundary conditions are assumed homogeneous ($\mu(\boldsymbol{x}) = 0$).

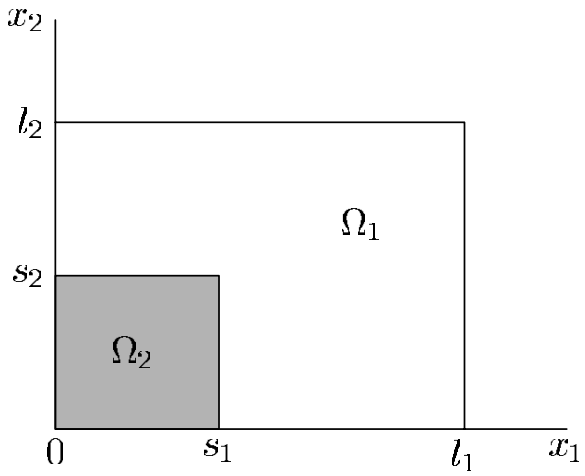

Figure 3.3 Calculation domain

The first point of interest concerns the efficiency of the computational algorithm used. We are speaking, first of all, about the rate of convergence of the iteration method, and about the convergence of approximate solution to the exact solution. Calculated data obtained on a sequence of refined (from $N = N_1 = N_2 = 25$ to $N = 200$) grids are given in Table 3.1.

N	25	50	100	200	400
n	17	22	32	47	69
$y_{\max}$	0.06161	0.06240	0.06266	0.06278	0.06284

Table 3.1 Computations performed on a sequence of consecutively refined grids

Here n is the total number of performed iterations (the program is terminated on reaching the relative accuracy $\varepsilon = 10^{-6}$), and $y_{\max}$ is the maximum magnitude of the approximate solution.

In the case of the problem with variable coefficients, the dependence $n = \mathcal{O}(N^{1/2})$ holds, typical of steepest iteration methods. Considering the problem with discontinuous coefficients, we cannot bargain for a difference-solution convergence rate of $\mathcal{O}(h^2)$, readily achieved in the case of problems with smooth solutions. The actual rate of convergence can be figured out to an extent considering data on the maximum magnitude of the approximate solution.

Figure 3.4 shows the approximate solution of the problem obtained for the basic variant. It is of interest to trace the effect due to the coefficient ratio κ_2/κ_1. We fix $\kappa_1 = 1$ and vary κ_2. Two cases of relatively high and low coefficient κ_2 are illustrated by Figures 3.5 and 3.6.

The use of high (low) coefficients at the higher-order derivatives corresponds to the method of fictitious domains. With high coefficients (see Figure 3.5), on the boundary

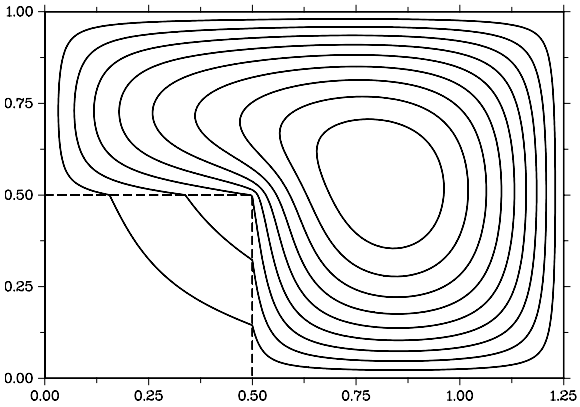

Figure 3.4 Solution obtained on the grid $N = N_1 = N_2 = 100$

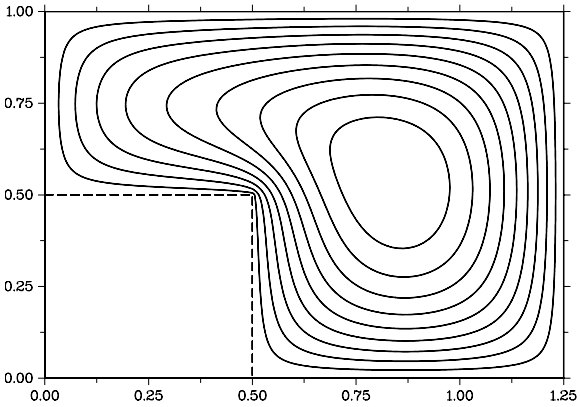

Figure 3.5 Solution obtained with $\kappa_2 = 100$

of the irregular domain Ω_1 we have homogeneous Dirichlet conditions. The second limiting case of low κ_2 gives the homogeneous Neumann approximate boundary conditions (see Figure 3.6).

3.5 Exercises

Exercise 3.1 In an irregular domain Ω, we treat the boundary value problem

$$-\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big) + q(\boldsymbol{x})u = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \tag{3.64}$$

$$u(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega \tag{3.65}$$

with

$$k(\boldsymbol{x}) \geq \kappa > 0, \qquad q(\boldsymbol{x}) \geq 0, \qquad \boldsymbol{x} \in \Omega.$$

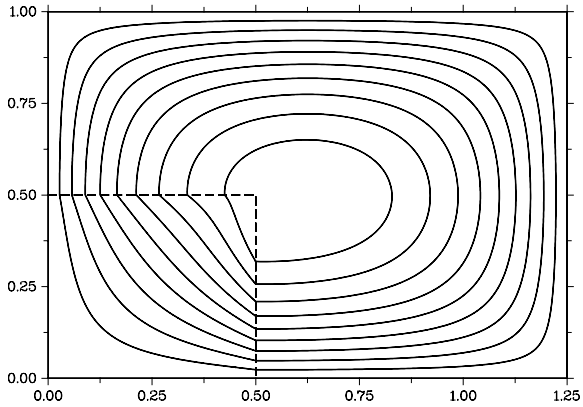

Figure 3.6 Solution obtained with $\kappa_2 = 0.01$

The calculation domain Ω in the method of fictitious domains is dipped into the domain Ω_0. Instead of the approximate solution of problem (3.64), (3.65), we deal with the boundary value problem

$$-\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k_\varepsilon(\boldsymbol{x})\,\frac{\partial u_\varepsilon}{\partial x_\alpha}\Big) + q_\varepsilon(\boldsymbol{x})u_\varepsilon = f_\varepsilon(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega_0, \tag{3.66}$$

$$u_\varepsilon(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega_0. \tag{3.67}$$

Consider a method of fictitious domains with continuation over higher coefficients in which

$$k_\varepsilon(\boldsymbol{x}),\, q_\varepsilon(\boldsymbol{x}),\, f_\varepsilon(\boldsymbol{x}) = \begin{cases} k(\boldsymbol{x}),\, q(\boldsymbol{x}),\, f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ \varepsilon^{-2},\, 0,\, 0, & \boldsymbol{x} \in \Omega_1 = \Omega_0 \backslash \Omega. \end{cases}$$

Show that $\|u_\varepsilon(\boldsymbol{x}) - u(\boldsymbol{x})\|_{W_2^1(\Omega)} \to 0$ as $\varepsilon \to 0$.

Exercise 3.2 Consider a method of fictitious domains with continuation over lower coefficients for approximate solution of problem (3.64), (3.65), when in (3.66), (3.67) we have:

$$k_\varepsilon(\boldsymbol{x}),\, q_\varepsilon(\boldsymbol{x}),\, f_\varepsilon(\boldsymbol{x}) = \begin{cases} k(\boldsymbol{x}),\, q(\boldsymbol{x}),\, f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ 1,\, \varepsilon^{-2},\, 0, & \boldsymbol{x} \in \Omega_1 = \Omega_0 \backslash \Omega. \end{cases}$$

Exercise 3.3 Show that the difference scheme

$$\begin{aligned} -y_{\overline{x}_1 x_1} - y_{\overline{x}_1 x_1} - \frac{h_1^2 + h_2^2}{12}\, y_{\overline{x}_1 x_1 \overline{x}_2 x_2} &= \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \\ y(\boldsymbol{x}) &= \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega \end{aligned}$$

with

$$\varphi(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{h_1^2}{12}\, f_{\overline{x}_1 x_1} + \frac{h_2^2}{12}\, f_{\overline{x}_2 x_2},$$

approximates the boundary value problem (3.2), (3.3) accurate to the fourth order.

Exercise 3.4 In the numerical solution of equation (3.1), approximate the third-kind boundary condition

$$-k(0, x_2)\,\frac{\partial u}{\partial x_1} + \sigma(x_2)u(0, x_2) = \mu(x_2),$$

considered on one side of a rectangle Ω (on the other boundary segments, first-kind boundary conditions are given).

Exercise 3.5 Construct a difference scheme for the boundary value problem (3.1), (3.3) with the following matched conditions at $x_1 = x_1^*$:

$$\begin{aligned} u(x_1^* + 0, x_2) - u(x_1^* - 0, x_2) &= 0, \\ k\,\frac{\partial u}{\partial x_1}(x_1^* + 0, x_2) - k\,\frac{\partial u}{\partial x_1}(x_1^* - 0, x_2) &= \chi(x_2). \end{aligned}$$

Exercise 3.6 Consider the approximation of the second-order elliptic equation with mixed derivatives

$$-\sum_{\alpha,\beta=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k_{\alpha\beta}(\boldsymbol{x})\,\frac{\partial u}{\partial x_\beta}\Big) = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega,$$

in which

$$k_{\alpha\beta}(\boldsymbol{x}) = k_{\beta\alpha}(\boldsymbol{x}), \qquad \alpha, \beta = 1, 2.$$

Exercise 3.7 In cylindrical coordinates, the Poisson equations in a circular cylinder can be written as

$$-\frac{1}{r}\frac{\partial}{\partial r}\Big(r\,\frac{\partial u}{\partial r}\Big) - \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \varphi^2} - \frac{\partial^2 u}{\partial z^2} = f(r, \varphi, z).$$

Construct a difference scheme for this equation with first-kind boundary conditions on the surface of the cylinder.

Exercise 3.8 Consider a difference scheme written as

$$A(\boldsymbol{x})y(\boldsymbol{x}) - \sum_{\xi \in \mathcal{W}'(\boldsymbol{x})} B(\boldsymbol{x}, \xi)y(\xi) = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega$$

with

$$\begin{aligned} &A(\boldsymbol{x}) > 0, \qquad B(\boldsymbol{x}, \xi) > 0, \qquad \xi \in \mathcal{W}'(\boldsymbol{x}), \\ &D(\boldsymbol{x}) = A(\boldsymbol{x}) - \sum_{\xi \in \mathcal{W}'(\boldsymbol{x})} B(\boldsymbol{x}, \xi) > 0, \qquad \boldsymbol{x} \in \omega. \end{aligned}$$

Derive the following estimate for the solution of the problem:

$$\|y(\boldsymbol{x})\|_\infty \le \Big\| \frac{\varphi(\boldsymbol{x})}{D(\boldsymbol{x})} \Big\|_\infty.$$

Exercise 3.9 Suppose that in the iteration method (of alternate directions)

$$B\frac{y_{k+1}-y_k}{\tau}+Ay_k=\varphi, \qquad k=0,1,\dots$$

we have

$$A=A_1+A_2, \qquad A_1A_2=A_2A_1$$

and, in addition,

$$\delta_\alpha E\le A_\alpha\le\Delta_\alpha E, \qquad A_\alpha=A_\alpha^*, \qquad \delta_\alpha>0, \qquad \alpha=1,2.$$

Next, suppose that the operator B is given in the factorized form

$$B=(E+\nu A_1)(E+\nu A_2).$$

Find the optimum value of ν.

Exercise 3.10 To solve the difference problem

$$Ay=\varphi, \qquad A=A^*>0$$

one uses the triangular iteration method

$$(D+\tau A_1)\frac{y^{k+1}-y^k}{\tau}+Ay^k=\varphi,$$

with D being an arbitrary self-adjoint operator and

$$A=A_1+A_2, \qquad A_1=A_2^*.$$

Find the optimum value of τ if a priori information is given in the form

$$\delta D\le A, \qquad A_1D^{-1}A_2\le\frac{\Delta}{4}A.$$

Exercise 3.11 Derive the calculation formula

$$\tau_{k+1}=\frac{(w_k,r_k)}{(Aw_k,w_k)}, \qquad w_k=B^{-1}r_k, \qquad r_k=Ay_k-\varphi$$

for the iteration parameter in the iterative steepest descend method

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=\varphi, \qquad k=0,1,\dots,$$
$$A=A^*>0, \qquad B=B^*>0,$$

from the condition of minimum norm of inaccuracy in H_A at the next iteration.

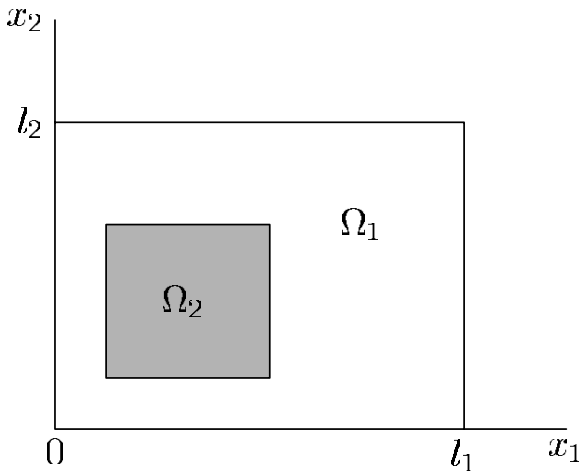

Figure 3.7 On the statement of the modified Dirichlet problem

Exercise 3.12 Using the program PROBLEM2, examine how the rate of convergence in the used iterative process depends on the variable coefficient $k(\boldsymbol{x})$ (or, in the simplest case, how the number of iterations depends on κ_2).

Exercise 3.13 Using the program PROBLEM2, perform numerical experiments on solving the Dirichlet problem in a step domain Ω_1 (Figure 3.3) based on the method of fictitious domains with continuation over lower coefficients.

Exercise 3.14 Modify the program PROBLEM2 so that to solve the third boundary value problem (boundary conditions (3.4)) for equation (3.60).

Exercise 3.15 Consider the modified Dirichlet problem for equation (3.1) in a biconnected domain Ω_1 (Figure 3.7). At the outer boundary, conditions (3.1) are given. At the internal boundary, the solution is constant, but this solution itself must be determined from an additional integral condition:

$$u(\boldsymbol{x}) = \text{const}, \qquad \boldsymbol{x} \in \partial\Omega_2,$$

$$\int_{\partial\Omega_1} k(\boldsymbol{x}) \frac{\partial u}{\partial n}\, d\boldsymbol{x} = 0.$$

To approximately solve the problem, use the method of fictitious domains implemented via the program PROBLEM2.

4 Boundary value problems for parabolic equations

As a typical non-stationary mathematical physiscs problem, we consider here the boundary value problem for the space-uniform second-order equation. On the approximation over space, we arrive at a Cauchy problem for a system of ordinary differential equations. Normally, the approximation over time in such problems can be achieved using two time layers. Less frequently, three-layer difference schemes are used. A theoretical consideration of the convergence of difference schemes for non-stationary problems rests on the theory of stability (correctness) of operator-difference schemes in Hilbert spaces of mesh functions. Conditions for stability of two- and three-layer difference schemes under various conditions are formulated. Numerical experiments on the approximate solution of a model boundary value problem for a one-dimensional parabolic equation are performed.

4.1 Difference schemes

Difference schemes for a model second-order parabolic equation are constructed. The approximation over time is performed using two and three time layers.

4.1.1 Boundary value problems

Consider a simplest boundary value problem for a one-dimensional parabolic equation. The calculation domain is the rectangle

$$\overline{Q}_T = \overline{\Omega} \times [0, T], \quad \overline{\Omega} = \{x \mid 0 \le x \le l\}, \quad 0 \le t \le T.$$

The solution is to be found from the equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) + f(x,t), \qquad 0 < x < l, \quad 0 < t \le T. \tag{4.1}$$

Here, the coefficient k depends just on the spatial variable and, in addition, $k(x) \ge \kappa > 0$.

We consider the first boundary value problem (with the boundary conditions assumed homogeneous) in which equation (4.1) is supplemented with the conditions

$$u(0,t) = 0, \qquad u(l,t) = 0, \qquad 0 < t \le T. \tag{4.2}$$

Also, the following initial conditions are considered:

$$u(x,0) = u_0(x), \qquad 0 \le x \le l. \tag{4.3}$$

In a more general case, one has to use third-kind boundary conditions. In the latter case, instead of (4.2) we have:

$$\begin{aligned} -k(0)\,\frac{du}{dx}(0,t) + \sigma_1(t)u(0,t) &= \mu_1(t), \\ k(l)\,\frac{du}{dx}(l,t) + \sigma_2(t)u(l,t) &= \mu_2(t), \qquad 0 < t \le T. \end{aligned} \tag{4.4}$$

The boundary value problem (4.1)–(4.3) is considered as a Cauchy problem for the first-order differential-operator equation in the Hilbert space $\mathcal{H} = \mathcal{L}_2(\Omega)$ for functions defined in the domain $\Omega = (0, 1)$ and vanishing at the boundary points of the domain (on $\partial\Omega$). For the norm and for the scalar product, we use the settings

$$(v, w) = \int_\Omega v(x)w(x)\,dx, \qquad \|v\|^2 = (v, v) = \int_\Omega v^2(x)\,dx.$$

We define the operator

$$\mathcal{A}u = -\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l, \tag{4.5}$$

for functions satisfying the boundary conditions (4.2).

The boundary value problem (4.1)–(4.3) is written as a problem in which it is required to find the function $u(t) \in \mathcal{H}$ from the differential-operator equation

$$\frac{du}{dt} + \mathcal{A}u = f(t), \qquad 0 < t \le T \tag{4.6}$$

supplemented with the initial condition

$$u(0) = u_0. \tag{4.7}$$

The operator $\mathcal{A}$ is a self-adjoint operator positively defined in $\mathcal{H}$, i. e.,

$$\mathcal{A}^* = \mathcal{A} \ge mE, \qquad m = \kappa\pi^2/l^2 > 0. \tag{4.8}$$

With the aforesaid taken into account (see the proof of Theorem 1.2), we derive the following a priori estimate for the solution of problem (4.6)–(4.8):

$$\|u(t)\| \le \exp(-mt)\Big(\|u_0\| + \int_0^t \exp(m\theta)\|f(\theta)\|\,d\theta\Big). \tag{4.9}$$

A cruder estimate derived by invoking the property of non-negativeness of the operator $\mathcal{A}$ only was obtained previously in Theorem 1.2; this estimate has the form

$$\|u(t)\| \le \|u_0\| + \int_0^t \|f(\theta)\|\,d\theta. \tag{4.10}$$

Estimates (4.9) and (4.10) show that the solution of problem (4.6)–(4.8) is stable with respect to initial data and right-hand side. Such fundamental properties of the differential problem must be inherited when we pass to a discrete problem.

4.1.2 Approximation over space

As usually, we denote as $\bar{\omega}$ a uniform grid with stepsize h over the interval $\bar{\Omega} = [0, l]$:

$$\bar{\omega} = \{x \mid x = x_i = ih,\ i = 0, 1, \ldots, N,\ Nh = l\}$$

and let ω and $\partial\omega$ be the sets of internal and boundary nodes.

At internal nodes, we approximate the differential operator (4.5), accurate to the second order, with the difference operator

$$Ay = -(ay_{\bar{x}})_x, \qquad x \in \omega, \tag{4.11}$$

in which, for instance, $a(x) = k(x - 0.5h)$.

In the mesh Hilbert space H, we introduce a norm defined as $\|y\| = (y, y)^{1/2}$, where

$$(y, w) = \sum_{x \in \omega} y(x)w(x)h.$$

On the set of functions vanishing on $\partial\omega$, for the self-adjoint operator A under the constraints $k(x) \geq \kappa > 0$ and $q(x) \geq 0$ there holds the estimate

$$A = A^* \geq \kappa\lambda_0 E, \tag{4.12}$$

in which

$$\lambda_0 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l} \geq \frac{8}{l^2}$$

is the minimum eigenvalue of the difference operator of second derivative on the uniform grid.

The approximation over space performed, we have the following problem put into correspondence to the problem (4.6), (4.7):

$$\frac{dy}{dt} + Ay = f, \qquad x \in \omega, \quad t > 0, \tag{4.13}$$

$$y(x, 0) = u_0(x), \qquad x \in \omega. \tag{4.14}$$

By virtue of (4.12), for the solution of problem (4.13), (4.14) there holds the estimate

$$\|y(x, t)\| \ \leq\ \exp(-\kappa\lambda_0 t)\Big(\|u^0(x)\| + \int_0^t \exp(\kappa\lambda_0\theta)\|f(x, \theta)\|\, d\theta\Big), \tag{4.15}$$

consistent with estimate (4.9).

4.1.3 Approximation over time

The next step in the approximate solution of the Cauchy problem for the system of ordinary difference equations (4.13), (4.14) is approximation over time. We define a time-uniform grid

$$\bar{\omega}_\tau = \omega_\tau \cup \{T\} = \{t_n = n\tau,\ n = 0, 1, \ldots, N_0,\ \tau N_0 = T\}.$$

We denote as $A, B : H \to H$ linear operators in H dependent, generally speaking, on τ and t_n. In the notation used below, following the manner adopted in the theory of difference schemes, we do not use subscripts:

$$y = y_n, \qquad \hat{y} = y_{n+1}, \qquad \check{y} = y_{n-1},$$
$$y_{\bar{t}} = \frac{y - \check{y}}{\tau}, \qquad y_t = \frac{\hat{y} - y}{\tau}.$$

In a *two-layer difference scheme* used to solve a non-stationary equation, the transition to the next time layer $t = t_{n+1}$ is performed using the solution y_n at the previous time layer.

In the two-layer scheme, equation (4.13) is approximated with the difference equation

$$\frac{y_{n+1} - y_n}{\tau} + A(\sigma y_{n+1} + (1 - \sigma) y_n) = \varphi_n, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{4.16}$$

where σ is a numerical parameter (weight) to be taken from the interval $0 \le \sigma \le 1$. For the right-hand side, one can put, for instance,

$$\varphi_n = \sigma f_{n+1} + (1 - \sigma) f_n.$$

Approximation of (4.14) yields

$$y_0 = u_0(x), \qquad x \in \omega. \tag{4.17}$$

Scheme (4.16) is known as the *weighted scheme*.

For the time-approximation inaccuracy of the first derivative we have:

$$\frac{v_{n+1} - v_n}{\tau} = \frac{dv}{dt}(t^*) + \mathcal{O}(\tau^\nu),$$

where $\nu = 2$ if $t = t_{n+1/2}$; otherwise, $\nu = 1$. As a result, we obtain the difference equation (4.16) that approximates equation (4.13) over time accurate to the second order in the case of $\sigma = 0.5$ and accurate to the first order in the case of $\sigma \ne 0.5$.

Consider also the weighted *three-layer difference scheme* for equation (4.13), in which three time layers (t_{n+1}, t_n and t_{n-1}) are used:

$$\left(\theta \frac{y_{n+1} - y_n}{\tau} + (1 - \theta) \frac{y_n - y_{n-1}}{\tau}\right) + A(\sigma_1 y_{n+1} + (1 - \sigma_1 - \sigma_2) y_n + \sigma_2 y_{n-1}) = \varphi_n, \tag{4.18}$$

$$n = 1, 2, \ldots, N_0 - 1.$$

For the right-hand side, it would appear reasonable to use the approximations

$$\varphi_n \;=\; \sigma_1 f_{n+1} + (1 - \sigma_1 - \sigma_2) f_n + \sigma_2 f_{n-1}.$$

Now, we can enter the three-layer calculation scheme with known y_0 and y_1:

$$y_0 = u_0, \qquad y_1 = u_1. \tag{4.19}$$

For y_0, we use the initial condition (4.14). To find y_1, we invoke a two-layer difference scheme, for instance,

$$\frac{y_1 - y_0}{\tau} + \frac{1}{2} A(y_1 + y_0) \;=\; \frac{1}{2} (f_1 + f_0).$$

The difference scheme (4.18), (4.19) involves a weighting parameter θ used in the approximation of the time derivative and two weighting parameters, σ_1 and σ_2, used in the approximation of other terms. Let us give some typical sets of weighting parameters that have found use in practical computations.

Using a greater number of layers in approximation of time-dependent equations is aimed, first of all, at raising the approximation order. That is why in the consideration of three-layer schemes (4.18), (4.19) it makes sense for us to restrict ourselves to schemes that approximate the initial problem (4.13), (4.14) over time with an order not less than second order.

First of all, consider a one-parametric family of symmetric schemes of the second approximation order

$$\frac{y_{n+1} - y_{n-1}}{2\tau} + A(\sigma y_{n+1} + (1 - 2\sigma) y_n + \sigma y_{n-1}) \;=\; \varphi_n, \tag{4.20}$$

which results from (4.18) with the settings

$$\theta = 1/2, \qquad \sigma = \sigma_1 = \sigma_2.$$

Particular attention should be paid to the scheme

$$\frac{3y_{n+1} - 4y_n + y_{n-1}}{2\tau} + A y_{n+1} = \varphi_n \tag{4.21}$$

that approximates equation (4.13) accurate to $\mathcal{O}(\tau^2)$ with properly given right-hand side (for instance, $\varphi_n = f_{n+1}$). Note that if in (4.18) we choose

$$\theta = 3/2, \qquad \sigma_1 = 1, \qquad \sigma_2 = 0,$$

then we arrive at scheme (4.21).

4.2 Stability of two-layer difference schemes

A highly important place in the robustness study of approximate solution methods for non-stationary problems is occupied by the stability theory. Below, we introduce key notions in use in the theory of stability for operator-difference schemes considered in finite-dimensional Hilbert spaces, formulate stability criteria for two-layer difference schemes with respect to initial data, and give typical estimates of stability with respect to initial data and right-hand side.

4.2.1 Basic notions

Stability conditions are formulated for difference schemes written in the unified general (canonical) form. Any two-layer scheme can be written in the form

$$B(t_n)\,\frac{y_{n+1}-y_n}{\tau} + A(t_n)y_n = \varphi_n, \qquad t_n \in \omega_\tau, \tag{4.22}$$

$$y_0 = u_0, \tag{4.23}$$

where $y_n = y(t_n) \in H$ is the function to be found and the functions $\varphi_n, u^0 \in H$ are given functions. Notation (4.22), (4.23) is called the *canonical form of two-layer schemes*.

For the Cauchy problem at the next time layer to be solvable, we assume that the operator B^{-1} does exist. Then, we can write equation (4.22) as

$$y_{n+1} = Sy_n + \tau\tilde{\varphi}_n, \qquad S = E - \tau B^{-1}A, \qquad \tilde{\varphi}_n = B^{-1}\varphi_n. \tag{4.24}$$

The operator S is called the *transition operator* for the two-layer difference scheme (using this operator, transitions from one time layer to the next time layer can be performed).

A two-layer scheme is called *stable*, if there exist positive constants m_1 and m_2, independent of τ and, also, of u_0 and φ, such that for all $u_0 \in H$, $\varphi \in H$, and $t \in \bar{\omega}_\tau$ for the solution of problem (4.22), (4.23) the following estimate is valid:

$$\|y_{n+1}\| \le m_1\|u_0\| + m_2 \max_{0\le\theta\le t_n} \|\varphi(\theta)\|_*, \qquad t_n \in \omega_\tau. \tag{4.25}$$

Here $\|\cdot\|$ and $\|\cdot\|_*$ are some norms in the space H. Inequality (4.25) implies that the solution of problem (4.22), (4.23) depends continuously on input data.

The difference scheme

$$B(t_n)\,\frac{y_{n+1}-y_n}{\tau} + A(t_n)y_n = 0, \qquad t_n \in \omega_\tau, \tag{4.26}$$

$$y_0 = u_0 \tag{4.27}$$

is called a scheme *stable with respect to initial data*, if for the solution of problem (4.26), (4.27) the following estimate holds:

$$\|y_{n+1}\| \le m_1\|u_0\|, \qquad t_n \in \omega_\tau. \tag{4.28}$$

The two-layer difference scheme

$$B(t_n)\,\frac{y_{n+1}-y_n}{\tau}+A(t_n)y_n=\varphi_n,\qquad t_n\in\omega_\tau, \tag{4.29}$$

$$y_0=0 \tag{4.30}$$

is *stable with respect to the right-hand side*, if for the solution there holds the inequality

$$\|y_{n+1}\|\ \le\ m_2\max_{0\le\theta\le t_n}\|\varphi(\theta)\|_*,\qquad t_n\in\omega_\tau. \tag{4.31}$$

The difference scheme (4.26), (4.27) is called a ρ-*stable* (uniformly stable) scheme with respect to initial data in H_R if there exist constants $\rho>0$ and m_1 independent of τ and n and such that for any n and for all $y_n\in H$ for the solution y_{n+1} of (4.26) there holds the estimate

$$\|y_{n+1}\|_R\le\rho\|y_n\|_R,\qquad t_n\in\omega_\tau \tag{4.32}$$

and, in addition, $\rho^n\le m_1$.

In the theory of difference schemes, the constant ρ is traditionally chosen as one of the following quantities:

$$\rho=1,$$

$$\rho=1+c\tau,\qquad c>0,$$

$$\rho=\exp(c\tau).$$

Here, the constant c does not depend on τ and n.

With regard to (4.24), we rewrite equation (4.26) in the form

$$y_{n+1}=Sy_n. \tag{4.33}$$

The requirement for ρ-stability is equivalent to the fulfillment of the two-sided operator inequality

$$-\rho R\le RS\le\rho R, \tag{4.34}$$

provided that RS is a self-adjoint operator ($RS=S^*R$). For any transition operator in (4.33), the condition for ρ-stability has the form

$$S^*RS\le\rho^2R. \tag{4.35}$$

Let us formulate now the *difference analogue of the Gronwall lemma* (see Lemma 1.1).

Lemma 4.1 *From an estimate of the difference solution at some layer*

$$\|y_{n+1}\|\le\rho\|y_n\|+\tau\|\varphi_n\|_* \tag{4.36}$$

there follows the a priori estimate

$$\|y_{n+1}\|\le\rho^{n+1}\|y_0\|+\sum_{k=0}^{n}\tau\rho^{n-k}\|\varphi_k\|_*. \tag{4.37}$$

In this way, an estimate of the solution at an individual layer yields an a priori estimate of the difference solution at arbitrary time.

4.2.2 Stability with respect to initial data

Let us formulate basic criteria for stability of two-layer operator-difference schemes with respect to initial data. Of primary significance here is the following theorem about exact (sufficient and necessary) stability conditions in H_A.

Theorem 4.2 *Suppose that the operator A in* (4.26) *is a self-adjoint, positive operator being, in addition, a constant (independent of n) operator. The condition*

$$B \geq \frac{\tau}{2} A, \qquad t \in \omega_\tau \tag{4.38}$$

is a necessary and sufficient condition for stability of A in H_A, or for the fulfillment of the estimate

$$\|y_{n+1}\|_A \leq \|u_0\|_A, \qquad t \in \omega_\tau. \tag{4.39}$$

Proof. We scalarwise multiply equation (4.26) by y_t; then, we obtain the identity

$$(By_t, y_t) + (Ay, y_t) = 0. \tag{4.40}$$

Next, we use the representation

$$y = \frac{1}{2}(y + \hat{y}) - \frac{1}{2}\tau y_t,$$

and write the identity (4.40) as

$$\Big(\Big(B - \frac{\tau}{2}A\Big)y_t, y_t\Big) + \frac{1}{2\tau}(A(\hat{y} + y), \hat{y} - y) = 0. \tag{4.41}$$

For the self-adjoint operator A, we have: $(Ay, \hat{y}) = (y, A\hat{y})$ and

$$(A(\hat{y} + y), \hat{y} - y) \;=\; (A\hat{y}, \hat{y}) - (Ay, y).$$

We insert the latter equalities into (4.41) and use the condition (4.38); then, we obtain the inequality

$$\|y_{n+1}\|_A \leq \|y_n\|_A \tag{4.42}$$

that yields the desired estimate (4.39).

To prove the necessity of (4.39), we assume that the scheme under consideration is stable in H_A or, in other words, inequality (4.39) is fulfilled. Prove now that, from here, the operator inequality (4.38) follows. We start from the identity (4.41) at the first layer $n = 0$,

$$2\tau\Big(\Big(B - \frac{\tau}{2}A\Big)w, w\Big) + (Ay_1, y_1) \;=\; (Ay_0, y_0), \qquad w = \frac{y_1 - y_0}{\tau}.$$

In view of (4.39), the latter identity can be fulfilled iff

$$\Big(\Big(B - \frac{\tau}{2}A\Big)w, w\Big) \geq 0.$$

Since $y_0 = u_0 \in H$ is an arbitrary element, then $w = -B^{-1}Au_0 \in H$ is also an arbitrary element. Indeed, since the operator A^{-1} does exist, for any element $w \in H$ we can find an element $u_0 = -A^{-1}Bw \in H$. Thus, the above inequality turns out to be fulfilled for all $w \in H$, i.e., the operator inequality (4.38) holds. □

Condition (4.38) is a necessary and sufficient condition for stability not only in H_A, but also in other norms. Not discussing all possibilities that arise along this line, let us formulate without proof only a statement concerning the stability in H_B.

Theorem 4.3 *Let the operators A and B in* (4.26), (4.27) *be constant operators and, in addition,*

$$B = B^* > 0, \qquad A = A^* > 0. \tag{4.43}$$

Then, condition (4.38) *is a condition necessary and sufficient for the scheme* (4.26), (4.27) *to be stable, with* $\rho = 1$, *with respect to initial data in* H_B.

General non-stationary problems must be treated using conditions for ρ-stability.

Theorem 4.4 *Let A and B be constant operators and, in addition,*

$$A = A^*, \qquad B = B^* > 0.$$

Then, the conditions

$$\frac{1-\rho}{\tau}\,B \leq A \leq \frac{1+\rho}{\tau}\,B \tag{4.44}$$

are conditions necessary and sufficient for ρ*-stability of scheme* (4.26), (4.27) *in* H_B *or, in other words, for the fulfillment of the inequality*

$$\|y_{n+1}\|_B \leq \rho\|y_n\|_B.$$

Proof. We write the scheme (4.26) in the form of (4.33); then, from (4.34) we obtain the following conditions of stability in H_B:

$$-\rho B \leq B - \tau A \leq \rho B.$$

The latter two-sided operator inequality can be formulated as inequalities (4.44) involving the operators of the two-layer difference scheme. □

It should be emphasized here that the conditions of the theorem do not assume positiveness (or even non-negativeness) of A. Under an additional assumption that A is a positive operator, one can show that the conditions (4.44) are necessary and sufficient conditions for ρ-stability of scheme (4.26), (4.27) in H_A.

Like in Theorem 4.2, in the case of $\rho \geq 1$, stability can be established for two-layer difference schemes with a non-self-adjoint operator B.

Theorem 4.5 *Let the operator A be a self-adjoint positive and constant operator. Then, in the case of*

$$B \geq \frac{\tau}{1+\rho} A \tag{4.45}$$

the difference scheme (4.26), (4.27) *is ρ-stable in H_A.*

Proof. We add to and subtract from the basic energy identity (see (4.41))

$$2\tau\Big(\Big(B - \frac{\tau}{2} A\Big)y_t, y_t\Big) + (A\hat{y}, \hat{y}) - (Ay, y) = 0 \tag{4.46}$$

the expression

$$2\tau^2 \frac{1}{1+\rho} (Ay_t, y_t).$$

This yields

$$2\tau\Big(\Big(B - \frac{\tau}{1+\rho} A\Big)y_t, y_t\Big) + (A\hat{y}, \hat{y}) - (Ay, y) - \frac{1-\rho}{1+\rho} \tau^2 (Ay_t, y_t) \ = \ 0.$$

With the condition (4.45) and self-adjointness of A taken into account, after simple manipulations we obtain:

$$(A\hat{y}, \hat{y}) - \rho(Ay, y) + (\rho - 1)(A\hat{y}, y) \ \leq \ 0.$$

Next, we use the inequality

$$|(A\hat{y}, y)| \leq \|\hat{y}\|_A \|y\|_A$$

and introduce the number

$$\eta = \frac{\|\hat{y}\|_A}{\|y\|_A};$$

then, we arrive at the inequality

$$\eta^2 - (\rho - 1)\eta + \rho \leq 0.$$

This inequality holds for all η from the interval $1 \leq \eta \leq \rho$, thus proving the desired estimate

$$\|\hat{y}\|_A \leq \|y\|_A$$

that guarantees stability in H_A. □

Turn now to the derivation of a priori estimates that express stability with respect to right-hand side. These results form a basis the examination of stability of difference schemes for non-stationary problems rests on.

4.2.3 Stability with respect to right-hand side

First of all, show that stability with respect to initial data in H_R, $R = R^* > 0$ implies stability with respect to right-hand side provided that the norm $\|\varphi\|_* = \|B^{-1}\varphi\|_R$ is used.

Theorem 4.6 *Let the difference scheme* (4.22), (4.23) *be a scheme ρ-stable in H_R with respect to initial data, i. e., the estimate* (4.42) *be valid in the case of $\varphi_n = 0$. Then, the difference scheme* (4.22), (4.23) *is also stable with respect to right-hand side and, for the solution, the following a priori estimate holds:*

$$\|y_{n+1}\|_R \le \rho^{n+1}\|u_0\|_R + \sum_{k=0}^{n} \tau\rho^{n-k}\|B^{-1}\varphi_k\|_R. \tag{4.47}$$

Proof. Since the operator B^{-1} does exist, equation (4.22) can be written as

$$y_{n+1} = Sy_n + \tau\tilde{\varphi}_n, \qquad S = E - \tau B^{-1}A, \qquad \tilde{\varphi}_n = B^{-1}\varphi_n. \tag{4.48}$$

From (4.48) we obtain

$$\|y_{n+1}\|_R \;\le\; \|Sy_n\|_R + \tau\|B^{-1}\varphi_n\|_R. \tag{4.49}$$

The requirement for ρ-stability of the scheme with respect to initial data is equivalent to the boundedness of the norm of S:

$$\|Sy_n\|_R \;\le\; \rho\|y_n\|_R, \qquad t \in \omega_\tau.$$

As a result, from (4.49) we obtain

$$\|y_{n+1}\|_R \;\le\; \rho\|y_n\|_R + \tau\|B^{-1}\varphi_n\|_R.$$

We use the difference analogue of the Gronwall lemma and obtain the desired estimate (4.47) that shows the scheme to be stable with respect to initial data and right-hand side. □

In the particular case of $D = A$ or $D = B$ (with $A = A^* > 0$ or $B = B^* > 0$), the a priori estimate (4.47) yields simplest estimates of stability in the energy space H_A or in H_B.

Some new estimates for the two-layer difference scheme (4.22), (4.23) can be obtained using a stability criterion more crude than (4.48).

Theorem 4.7 *Let A be a constant, self-adjoint, positive operator, and let the operator B satisfy the condition*

$$B \ge \frac{1+\varepsilon}{2}\,\tau A \tag{4.50}$$

with some constant $\varepsilon > 0$ independent of τ. Then, for the difference scheme (4.22), (4.23) *there holds the a priori estimate*

$$\|y_{n+1}\|_A^2 \;\le\; \|u_0\|_A^2 + \frac{1+\varepsilon}{2\varepsilon}\sum_{k=0}^{n}\tau\|\varphi\|_{B^{-1}}^2. \tag{4.51}$$

Proof. We scalarwise multiply equation (4.22) by $2\tau y_t$; then, in the same way as for (4.46), we obtain the energy identity

$$2\tau\Big(\Big(B-\frac{\tau}{2}A\Big)y_t,\, y_t\Big) + (A\hat{y},\,\hat{y}) \;=\; (Ay,\,y) + 2\tau(\varphi,\,y_t). \tag{4.52}$$

The right-hand side of (4.52) can be estimated as

$$2\tau(\varphi,\,y_t) \;\le\; 2\tau\|\varphi\|_{B^{-1}}\|y_t\|_B \;\le\; 2\tau\varepsilon_1\|y_t\|_B^2 + \frac{\tau}{2\varepsilon_1}\,\|\varphi\|_{B^{-1}}^2$$

with some still undetermined positive constant ε_1. On substitution of the latter estimate into (4.52), we obtain:

$$2\tau\Big(\Big((1-\varepsilon_1)B-\frac{\tau}{2}A\Big)y_t,\, y_t\Big) + (A\hat{y},\,\hat{y}) \;\le\; (Ay,\,y) + \frac{\tau}{2\varepsilon_1}\,\|\varphi\|_{B^{-1}}^2.$$

With condition (4.50) being fulfilled, one can choose the constant ε_1 so that we have

$$\frac{1}{1-\varepsilon_1} = 1+\varepsilon$$

and, hence,

$$\begin{aligned}(1-\varepsilon_1)B-\frac{\tau}{2}A \;&=\; (1-\varepsilon_1)\Big(B-\frac{1+\varepsilon}{2}\,\tau A\Big)\ge 0,\\ (A\hat{y},\,\hat{y}) \;&\le\; (Ay,\,y) + \frac{1+\varepsilon}{2\varepsilon}\,\tau\|\varphi\|_{B^{-1}}^2.\end{aligned}$$

The latter inequality yields the estimate (4.51). □

Theorem 4.8 *Let A be a constant, self-adjoint and positive operator, and the operator B satisfy the condition*

$$B \ge G + \frac{\tau}{2}A, \qquad G = G^* > 0. \tag{4.53}$$

Then, for the scheme (4.22), (4.23) *the following a priori estimate holds:*

$$\|y_{n+1}\|_A^2 \;\le\; \|u_0\|_A^2 + \frac{1}{2}\sum_{k=0}^{n}\tau\|\varphi_k\|_{G^{-1}}^2. \tag{4.54}$$

Proof. In (4.52), we use the estimate

$$2\tau(\varphi, y_t) \le 2\tau(Gy_t, y_t) + \frac{\tau}{2}(G^{-1}\varphi, \varphi).$$

We insert this estimate into (4.52); then, in view of (4.53), we obtain

$$(A\hat{y}, \hat{y}) \le (Ay, y) + \frac{1}{2}\tau\|\varphi\|^2_{G^{-1}}$$

which, by the difference Gronwall lemma, yields (4.54). □

Convergence of difference schemes can be examined in different classes of solution smoothness for the initial differential problem; we therefore need a broad spectrum of estimates in which, in particular, the right-hand side would be evaluated in different easily calculable norms. Here, we have restricted ourselves only to some typical a priori estimates of the solutions of operator-difference schemes.

4.3 Three-layer operator-difference schemes

Below, three-layer operator-difference schemes are considered based on the passage to an equivalent two-layer operator-difference scheme. Estimates of stability with respect to initial data and right-hand side in various norms are obtained.

4.3.1 Stability with respect to initial data

In the consideration of three-layer difference schemes, we use the following *canonical form of three-layer difference schemes*:

$$B(t_n)\frac{y_{n+1} - y_{n-1}}{2\tau} + R(t_n)(y_{n+1} - 2y_n + y_{n-1}) + A(t_n)y_n = \varphi_n, \qquad n = 1, 2, \dots \tag{4.55}$$

with the values

$$y_0 = u_0, \qquad y_1 = u_1. \tag{4.56}$$

Let us obtain conditions for stability with respect to initial data in the case of constant (independent of n), self-adjoint operators A, B and R; in other words, instead of the general scheme (4.55) here we consider the scheme

$$B\frac{y_{n+1} - y_{n-1}}{2\tau} + R(y_{n+1} - 2y_n + y_{n-1}) + Ay_n = 0. \tag{4.57}$$

Let us derive a simplest a priori estimate for the scheme (4.56), (4.57) that expresses stability with respect to initial data. We put

$$u_n = \frac{1}{2}(y_n + y_{n-1}), \qquad w_n = y_n - y_{n-1} \tag{4.58}$$

and rewrite, using the identity

$$y_n = \frac{1}{4}(y_{n+1} + 2y_n + y_{n-1}) - \frac{1}{4}(y_{n+1} - 2y_n + y_{n-1}),$$

the scheme (4.57) as

$$B\frac{w_{n+1} + w_n}{2\tau} + R(w_{n+1} - w_n) - A(w_{n+1} - w_n) + A\frac{u_{n+1} + u_n}{2} = 0. \quad (4.59)$$

We scalarwise multiply equation (4.59) by

$$2(u_{n+1} - u_n) = w_{n+1} + w_n;$$

this yields the equality

$$\frac{1}{2\tau}(B(w_{n+1} + w_n), w_{n+1} + w_n) + (R(w_{n+1} - w_n), w_{n+1} + w_n) - \frac{1}{4}(A(w_{n+1} - w_n), w_{n+1} + w_n) + (A(u_{n+1} + u_n), u_{n+1} - u_n) = 0. \quad (4.60)$$

For self-adjoint operators R and A and for a non-negative operator B ($B \geq 0$), it follows from (4.60) that

$$\mathcal{E}_{n+1} \leq \mathcal{E}_n, \quad (4.61)$$

where, in view of the notation (4.58), we have

$$\mathcal{E}_{n+1} = \frac{1}{4}(A(y_{n+1} + y_n), y_{n+1} + y_n) + (R(y_{n+1} - y_n), y_{n+1} - y_n) - \frac{1}{4}(A(y_{n+1} - y_n), y_{n+1} - y_n). \quad (4.62)$$

Under certain constraints, the quantity $\mathcal{E}_n$, defined by (4.62), specifies a norm and, hence, inequality (4.61) guarantees stability of the operator-difference scheme with respect to initial data. More accurately, the following statement is valid.

Theorem 4.9 *Let the operators R and A in the operator-difference scheme* (4.57) *be self-adjoint operators. Then, with the conditions*

$$B \geq 0, \qquad A > 0, \qquad R > \frac{1}{4}A \quad (4.63)$$

fulfilled, there holds the a priori estimate

$$\frac{1}{4}\|y_{n+1} + y_n\|_A^2 + \|y_{n+1} - y_n\|_R^2 - \frac{1}{4}\|y_{n+1} - y_n\|_A^2 \leq \frac{1}{4}\|y_n + y_{n-1}\|_A^2 + \|y_n - y_{n-1}\|_R^2 - \frac{1}{4}\|y_n - y_{n-1}\|_A^2 \quad (4.64)$$

that proves the operator-difference scheme (4.57) *to be stable with respect to initial data.*

It is the complex structure of the norm (see (4.62)) that presents a specific feature of the three-layer schemes under consideration. In some important cases, on narrowing the class of difference schemes or on making stability conditions cruder, one can use simpler norms.

4.3.2 Passage to an equivalent two-layer scheme

Multi-layer difference schemes can be conveniently examined considering the passage to an equivalent two-layer scheme. For two-layer schemes, this approach yields most important results, including (coincident) necessary and sufficient conditions for stability.

We set as H^2 the *direct sum of spaces* H: $H^2 = H \oplus H$. For vectors $U = \{u^1, u^2\}$, the addition and multiplication in H^2 is defined coordinate-wise, and the scalar product is

$$(U, V) = (u^1, v^1) + (u^2, v^2).$$

In H^2, we define the operators (operator matrices)

$$\mathbf{G} = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

whose elements $G_{\alpha\beta}$ are operators in H. To a self-adjoint, positively defined operator $\mathbf{G}$, we put into correspondence a Hilbert space $H^2_{\mathbf{G}}$ in which the scalar product and the norm are given by

$$(U, V)_{\mathbf{G}} = (\mathbf{G}U, V), \qquad \|U\|_{\mathbf{G}} = \sqrt{(\mathbf{G}U, U)}.$$

We write the three-layer operator-difference scheme (4.57) as the two-layer vector scheme

$$\mathbf{B}\frac{Y^{n+1} - Y^n}{\tau} + \mathbf{A}Y^n = 0, \qquad n = 1, 2, \dots \tag{4.65}$$

with appropriately defined vectors Y^n, $n = 1, 2, \dots$.

In view of the aforesaid, for each $n = 1, 2, \dots$ we define the vector

$$Y^n = \Big\{\frac{1}{2}(y_n + y_{n-1}), y_n - y_{n-1}\Big\}. \tag{4.66}$$

Under the conditions of Theorem 4.9, in this notation the estimate (4.64) of stability with respect to initial data can be expressed as

$$\|Y^{n+1}\|_{\mathbf{G}} \le \|Y^n\|_{\mathbf{G}}, \tag{4.67}$$

with

$$G_{11} = A, \quad G_{12} = G_{21} = 0, \quad G_{22} = R - \frac{1}{4}A. \tag{4.68}$$

In view of (4.58), the two-layer vector scheme (4.65), (4.66) can be written as

$$B_{11}\frac{u_{n+1}-u_n}{\tau}+B_{12}\frac{w_{n+1}-w_n}{\tau}+A_{11}u_n+A_{12}w_n = 0, \tag{4.69}$$

$$B_{21}\frac{u_{n+1}-u_n}{\tau}+B_{22}\frac{w_{n+1}-w_n}{\tau}+A_{21}u_n+A_{22}w_n = 0. \tag{4.70}$$

Equality (4.69) is put into correspondence to the three-layer operator-difference scheme written in the form (4.57). Taking into account the identities

$$\begin{aligned}\frac{u_{n+1}+u_n}{2} &= u_n+\frac{u_{n+1}-u_n}{2},\\ 2(u_{n+1}-u_n) &= w_{n+1}+w_n,\end{aligned}$$

we rewrite (4.57) in a more convenient form:

$$B\frac{u_{n+1}-u_n}{\tau}+R\frac{w_{n+1}-w_n}{\tau}-\frac{1}{4}A(w_{n+1}-w_n) + \frac{\tau}{2}A\frac{u_{n+1}-u_n}{\tau}+Au_n=0. \tag{4.71}$$

To pass over from (4.69) to (4.71), we put:

$$B_{11}=B+\frac{\tau}{2}A,\quad B_{12}=\tau R-\frac{\tau}{4}A,\quad A_{11}=A,\quad A_{12}=0. \tag{4.72}$$

Equation (4.70) does not affect the three-layer scheme (4.57). Considering the two-layer operator-difference schemes (4.65) with self-adjoint operators $\mathbf{A}$, we define

$$B_{21}=-\tau Q,\quad B_{22}=\frac{\tau}{2}Q,\quad A_{21}=0,\quad A_{22}=Q, \tag{4.73}$$

where Q is some self-adjoint, positive operator.

In the case of (4.72), (4.73), for the operators in (4.65) we have the representation

$$\mathbf{B}=\frac{\tau}{2}\mathbf{A}+\mathbf{Q}, \tag{4.74}$$

where

$$\mathbf{Q}=\begin{pmatrix} Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{pmatrix}$$

with

$$Q_{11}=B,\quad Q_{12}=\tau R-\frac{\tau}{4}A,\quad Q_{21}=-\tau Q,\quad Q_{22}=0. \tag{4.75}$$

Based on the latter notation, under the conditions of Theorem 4.9 we can establish stability of the operator-difference scheme (4.57), i. e., the estimate (4.67), (4.68).

The two-layer vector operator-difference scheme (4.65) with a self-adjoint, positive operator $\mathbf{A}$ is stable with respect to initial data in $H^2_{\mathbf{A}}$ if

$$\mathbf{B} \geq \frac{\tau}{2}\,\mathbf{A}. \tag{4.76}$$

Representation (4.74) taken into account, the condition (4.76) is valid in the case of

$$\mathbf{Q} \geq 0.$$

The latter condition is always fulfilled for an operator $\mathbf{Q}$ defined by (4.75) in the case of $B \geq 0$ and

$$Q = R - \frac{1}{4}\,A. \tag{4.77}$$

In the case of (4.77), stability in $H^2_{\mathbf{A}}$ refers to the case in which inequalities (4.67), (4.68) are fulfilled.

4.3.3 ρ-stability of three-layer schemes

Admitting that norm of the difference solution of a problem can both increase or decrease, we consider here ρ-stable schemes, for which the condition for stability with respect to initial data has the form

$$\|Y^{n+1}\|_{\mathbf{G}} \leq \rho \|Y^n\|_{\mathbf{G}}, \tag{4.78}$$

with $\rho > 0$.

Theorem 4.10 *Let the operators R and A in the difference scheme* (4.57) *be self-adjoint operators. Then, with the conditions*

$$B + \frac{\tau}{2}\,\frac{\rho - 1}{\rho + 1}\,A \geq 0, \qquad A > 0, \qquad R - \frac{1}{4}\,A > 0 \tag{4.79}$$

fulfilled with $\rho > 1$, there holds the a priori estimate (4.78)*,* (4.68) *or, in other words, the operator-difference scheme* (4.57) *is ρ-stable with respect to initial data.*

Proof. The two-layer vector difference scheme (4.65) is ρ-stable with $\rho > 1$ in the case of (see Theorem 4.5)

$$\mathbf{B} \geq \frac{\tau}{\rho + 1}\,\mathbf{A}. \tag{4.80}$$

With (4.74), inequality (4.80) can be rewritten as

$$\mathbf{Q} + \frac{\tau}{2}\,\frac{\rho - 1}{\rho + 1}\,\mathbf{A} \geq 0. \tag{4.81}$$

Under the conditions of Theorem 4.5, the validity of (4.81) can be checked directly. □

Under somewhat more general conditions, ρ-stability estimates (4.78) with arbitrary $\rho > 0$ can be obtained in the case of ρ-dependent norms. In the operator-difference scheme (4.57), we introduce new unknowns $y_n = \rho^n z_n$, which yields

$$B\,\frac{\rho z_{n+1} - \rho^{-1} z_{n-1}}{2\tau} + R(\rho z_{n+1} - 2z_n + \rho^{-1} z_{n-1}) + Az_n = 0. \tag{4.82}$$

We write scheme (4.82) in the canonical form

$$\tilde{B}\,\frac{z_{n+1} - z_{n-1}}{2\tau} + \tilde{R}(z_{n+1} - 2z_n + z_{n-1}) + \tilde{A}z_n = 0. \tag{4.83}$$

Direct manipulations yield

$$\begin{gathered}\tilde{B} = \frac{\rho^2+1}{2}\,B + \tau(\rho^2-1)R, \qquad \tilde{R} = \frac{\rho^2-1}{4\tau}\,B + \frac{\rho^2+1}{2}\,R, \\ \tilde{A} = \frac{\rho^2-1}{2\tau}\,B + (\rho-1)^2 R + \rho A.\end{gathered} \tag{4.84}$$

By Theorem 4.9, in the case of

$$\tilde{B} \geq 0, \qquad \tilde{A} > 0, \qquad \tilde{R} - \frac{1}{4}\,\tilde{A} > 0 \tag{4.85}$$

the scheme (4.83) is stable with respect to initial data and there holds the estimate

$$\|Z^{n+1}\|_{\tilde{\mathbf{G}}} \leq \|Z^n\|_{\tilde{\mathbf{G}}}, \tag{4.86}$$

where (see (4.66))

$$Z^n = \Big\{\frac{1}{2}\,(z_n + z_{n-1}),\, z_n - z_{n-1}\Big\}.$$

Now we define the vector

$$Y^n = \Big\{\frac{1}{2}\Big(\frac{1}{\rho}\,y_n + z_{n-1}\Big),\, \frac{1}{\rho}\,z_n - z_{n-1}\Big\}. \tag{4.87}$$

Then, the estimate (4.86) assumes the form

$$\|Y^{n+1}\|_{\tilde{\mathbf{G}}} \leq \rho\|Y^n\|_{\tilde{\mathbf{G}}}, \tag{4.88}$$

i.e., the initial difference scheme (4.57) is ρ-stable with respect to initial data.

The norm in (4.88) is defined by the operator $\tilde{\mathbf{G}}$, for which

$$\tilde{G}_{11} = \tilde{A}, \qquad \tilde{G}_{12} = \tilde{G}_{21} = 0, \qquad \tilde{G}_{22} = \tilde{R} - \frac{1}{4}\,\tilde{A}. \tag{4.89}$$

The stability conditions can be formulated based on (4.84), (4.85).

Theorem 4.11 *Let the operators B, R and A in the difference scheme* (4.67) *be self-adjoint operators. Then, with the conditions*

$$\begin{aligned} \frac{\rho^2+1}{2}\,B + \tau(\rho^2-1)R \geq 0, \\ \frac{\rho^2-1}{2\tau}\,B + (\rho-1)^2 R + \rho A > 0, \\ \frac{\rho^2-1}{2\tau}\,B + (\rho+1)^2 R - \rho A > 0 \end{aligned} \tag{4.90}$$

fulfilled with $\rho > 0$, there holds the a priori estimate (4.87)–(4.89) *or, in other words, the difference scheme* (4.57) *is ρ-stable with respect to initial data in $H^2_{\widetilde{\mathbf{G}}}$.*

4.3.4 Estimates in simpler norms

Stability of the operator-difference schemes discussed above was established in Hilbert spaces with a complex composite norm (see (4.61), (4.62)). In the consideration of stability of three-layer difference schemes, we have also obtained estimates of stability in simpler (compared to (4.62)) norms. The latter was achieved at the expense of more tight stability conditions. Let us formulate the result.

Theorem 4.12 *Let the operators R and A in the operator-difference scheme* (4.57) *be self-adjoint operators. Then, in the case of*

$$B \geq 0, \qquad A > 0, \qquad R > \frac{1+\varepsilon}{4}\,A \tag{4.91}$$

with $\varepsilon > 0$ there hold the a priori estimates

$$\|y_{n+1}\|_A^2 \;\leq\; 2\,\frac{1+\varepsilon}{\varepsilon}\,(\|y_0\|_A^2 + \|y_1 - y_0\|_R^2), \tag{4.92}$$

$$\|y_{n+1}\|_A^2 + \|y_n - y_{n-1}\|_R^2 \;\leq\; \frac{4+3\varepsilon}{\varepsilon}\,(\|y_0\|_A^2 + \|y_1 - y_0\|_R^2). \tag{4.93}$$

Proof. In the used notation with omitted subscripts, $y_n = y$, and $y_{n+1} = \hat{y}$, $\hat{y} - y = \tau y_t$, and for $\mathcal{E}_{n+1}$ defined by (4.62) we have

$$\begin{aligned} \mathcal{E}_{n+1} &= \frac{1}{4}\,(A(\hat{y}+y), \hat{y}+y) + \tau^2 (Ry_t, y_t) - \frac{\tau^2}{4}\,(Ay_t, y_t) \\ &= (A\hat{y}, y) + \tau^2 (Ry_t, y_t). \end{aligned} \tag{4.94}$$

Substitution of $\hat{y} = y + \tau y_t$ into (4.94) yields

$$\mathcal{E}_{n+1} \;=\; (Ay, y) + \tau(Ay, y_t) + \tau^2 (Ry_t, y_t) \;\leq\; \|y\|_A^2 + \tau \|y\|_A \|y_t\|_A + \tau^2 \|y_t\|_R^2.$$

Taking into account the third inequality in (4.93), we obtain

$$\mathcal{E}_{n+1} \;\le\; \|y\|_A^2 + \frac{2\tau}{(1+\varepsilon)^{1/2}}\,\|y\|_A\|y_t\|_R + \tau^2\|y_t\|_R^2 \;\le\; 2(\|y\|_A^2 + \tau^2\|y_t\|_R^2).$$

Thus, we have established a lower estimate for the composite norm

$$\mathcal{E}_{n+1} \;\le\; 2(\|y_n\|_A^2 + \|y_{n+1} - y_n\|_R^2). \tag{4.95}$$

An upper estimate can be established in a similar manner. In (4.94), we put $y = \hat{y} - \tau y_t$; then, in view of (4.91), we obtain:

$$\begin{aligned}\mathcal{E}_{n+1} &= (A\hat{y}, \hat{y}) - \tau(A\hat{y}, y_t) + \tau^2(Ry_t, y_t)\\ &\ge \|\hat{y}\|_A^2 - \tau\|\hat{y}\|_A\|y_t\|_A + \tau^2\|y_t\|_R^2\\ &\ge \|\hat{y}\|_A^2 - \frac{2\tau}{(1+\varepsilon)^{1/2}}\,\|\hat{y}\|_A\|y_t\|_R + \tau^2\|y_t\|_R^2.\end{aligned}$$

For an arbitrary $\beta > 0$ we have:

$$\mathcal{E}_{n+1} \;\ge\; (1-\beta)\|\hat{y}\|_A^2 + \Big(1 - \frac{1}{\beta(1+\varepsilon)}\Big)\tau^2\|y_t\|_R^2. \tag{4.96}$$

We put $\beta = 1/(1+\varepsilon)$; then, from (4.96) we obtain:

$$\mathcal{E}_{n+1} \;\ge\; \frac{\varepsilon}{1+\varepsilon}\,\|y_{n+1}\|_A^2. \tag{4.97}$$

With estimates (4.95) and (4.97) taken into account, the stability estimate (4.61) yields the desired stability estimate (see (4.92)) for the three-layer difference scheme (4.57) in H_A.

To prove estimate (4.93), we put $\beta = (1+\varepsilon)^{-1/2}$, so that

$$1 - \beta = \frac{(1+\varepsilon)^{1/2} - 1}{(1+\varepsilon)^{1/2}} = \frac{\varepsilon}{1 + \varepsilon + (1+\varepsilon)^{1/2}}.$$

With the inequality $(1+\varepsilon)^{1/2} < 1 + 0.5\varepsilon$ taken into account, starting from (4.96), we arrive at a second lower estimate of the composite norm:

$$\mathcal{E}_{n+1} > \frac{2\varepsilon}{4+3\varepsilon}\,(\|y_{n+1}\|_A^2 + \|y_{n+1} - y_n\|_R^2). \tag{4.98}$$

Inequality (4.61), and estimates (4.95) and (4.98), yield the estimate (4.93). □

Estimates of type (4.92) naturally arise when one considers three-layer schemes for first-order evolutionary equations (second-order parabolic equation), and estimates of type (4.93), in the case of three-layer schemes for second-order equations (second-order hyperbolic equation).

4.3.5 Stability with respect to right-hand side

Let us give some simplest stability estimates for three-layer operator-difference schemes with respect to initial data and right-hand side. Instead of (4.57), now we consider the scheme

$$B\,\frac{y_{n+1}-y_{n-1}}{2\tau}+R(y_{n+1}-2y_n+y_{n-1})+Ay_n=\varphi_n. \tag{4.99}$$

Theorem 4.13 *Let the operators R and A in* (4.99) *be self-adjoint operators. Then, with the conditions*

$$B\geq\varepsilon E,\qquad A>0,\qquad R>\frac{1}{4}\,A \tag{4.100}$$

fulfilled with a constant $\varepsilon>0$, for the difference solution there hold the following a priori estimates

$$\mathcal{E}_{n+1}\leq\mathcal{E}_1+\frac{1}{2\varepsilon}\sum_{k=1}^{n}\tau\|\varphi_k\|^2, \tag{4.101}$$

$$\mathcal{E}_{n+1}\leq\mathcal{E}_1+\frac{1}{2}\sum_{k=1}^{n}\tau\|\varphi_k\|^2_{B^{-1}}. \tag{4.102}$$

Proof. Analogously to the proof of Theorem 4.9 (see (4.62)), we obtain the equality

$$\frac{1}{2\tau}\,(B(w_{n+1}+w_n),w_{n+1}+w_n)+\mathcal{E}_{n+1}\;=\;(\varphi_n,w_{n+1}+w_n)+\mathcal{E}_n.$$

To derive the estimate (4.101) with $\varepsilon>0$, in conditions (4.100) we invoke the inequality

$$(\varphi_n,w_{n+1}+w_n)\;\leq\;\frac{1}{2\tau}\,\varepsilon\|w_{n+1}+w_n\|^2+\frac{\tau}{2\varepsilon}\,\|\varphi_n\|^2.$$

The inequality

$$(\varphi_n,w_{n+1}+w_n)\;\leq\;\frac{1}{2\tau}\,\|w_{n+1}+w_n\|^2_B+\frac{\tau}{2}\,\|\varphi_n\|^2_{B^{-1}},$$

is used in the proof of estimate (4.102). □

Some other stability estimates for three-layer difference schemes (4.99) with respect to right-hand side can be obtained based on the estimates (4.92), (4.93) with somewhat more tight constraints imposed on R.

4.4 Consideration of difference schemes for a model problem

Below, general results obtained in the stability theory for operator-difference schemes are used to examine stability and convergence of difference schemes for a model boundary value problem with one-dimensional parabolic equation.

4.4.1 Stability condition for a two-layer scheme

The boundary value problem (4.1)–(4.3) can be solved using the two-layer difference scheme (4.16), (4.17). We assume this scheme to be written in the canonical form (4.22), (4.23) for two-layer difference schemes with

$$B = E + \sigma\tau A, \qquad A > 0. \tag{4.103}$$

Let us formulate now stability conditions for scheme (4.22), (4.103) under rather general conditions not assuming self-adjointness of A.

Theorem 4.14 *For the weighted difference scheme* (4.16), (4.17) *to be stable with respect to initial data in H, it is necessary and sufficient that the following operator inequality be fulfilled:*

$$A^* + \Big(\sigma - \frac{1}{2}\Big)\tau A^* A \geq 0. \tag{4.104}$$

Proof. In view of $A > 0$, the operator A^{-1} does exist. We multiply (4.16) by A^{-1}; so doing, we pass from (4.22), (4.103) to the difference scheme

$$\tilde{B}\,\frac{y_{n+1} - y_n}{\tau} + \tilde{A} y_n = \tilde{\varphi}_n, \qquad t_n \in \omega_\tau$$

in which

$$\tilde{B} = A^{-1} + \sigma\tau E, \qquad \tilde{A} = E.$$

Necessary and sufficient stability conditions for the latter scheme with respect to initial data in $H = H_{\tilde{A}}$ (Theorem 4.2) are given by the inequality

$$A^{-1} + \Big(\sigma - \frac{1}{2}\Big)\tau E \geq 0.$$

We multiply this inequality from the left by A^*, and from the right, by A (after such operations, the inequality still remains valid); this yields inequality (4.104). □

For weights $\sigma \geq 0.5$, the operator-difference scheme (4.22), (4.103) is absolutely stable (for all $\tau > 0$).

With a weight σ taken from the interval $0 \leq \sigma < 0.5$, one can expect the operator-difference scheme (4.22), (4.103) to be conditionally stable. In the latter case, to derive constraints on the time step size, we can invoke the upper estimate of A (see Lemma 2.2):

$$A \leq M_1 E, \qquad M_1 = \frac{4}{h^2} \max_{1 \leq i \leq N-1} \frac{a_i + a_{i+1}}{2}. \tag{4.105}$$

In the case of a self-adjoint operator A, inequality (4.104) can be written as

$$E + \Big(\sigma - \frac{1}{2}\Big)\tau A \geq 0.$$

Then, in view of (4.105), we have:

$$E + \Big(\sigma - \frac{1}{2}\Big)\tau A \;\geq\; \frac{1}{M_1}A + \Big(\sigma - \frac{1}{2}\Big)\tau A \;\geq\; 0.$$

Hence, in the case of $0 \leq \sigma < 0.5$ the stability condition is fulfilled if

$$\tau \leq \Big(\frac{1}{2} - \sigma\Big)\frac{1}{M_1}. \tag{4.106}$$

Thus (see (4.105)), for schemes with weighting factors taken from the interval $0 \leq \sigma < 0.5$ there exist tight constraints on the time step size. Here, as it follows from (4.106), the maximum possible time step size is $\tau_{\max} = \mathcal{O}(h^2)$, the latter circumstance presenting the most serious obstacle that hampers the use of such schemes in computational practice.

4.4.2 Convergence of difference schemes

The central issue in the theoretical substantiation of the difference schemes in use consists in the necessity to prove that the approximate solution converges to the exact solution as we successively refine the computation grid. Such a consideration can be performed using a priori estimates of stability with respect to initial data and right-hand side.

To investigate into the matter of accuracy of the difference scheme (4.16), (4.17) in yielding the approximate solution of problem (4.1)–(4.3), we write the related problem for the inaccuracy. For the inaccuracy at the time layer $t = t_n$, we put $z_n = y_n - u_n$ and substitute $y_n = z_n + u_n$ into (4.16), (4.17). In this way, we arrive at the problem

$$\frac{z_{n+1} - z_n}{\tau} + A(\sigma z_{n+1} + (1-\sigma)z_n) = \psi_n, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{4.107}$$

$$z_0 = 0, \qquad x \in \omega. \tag{4.108}$$

In (4.107), the right-hand side ψ_n is the approximation inaccuracy for equation (4.1):

$$\psi_n = \frac{u_{n+1} - u_n}{\tau} + A(\sigma u_{n+1} + (1-\sigma)u_n).$$

The approximation over space and time was discussed above. In the case of smooth coefficients and smooth right-hand side in (4.1), and also with smooth initial conditions, we have:

$$\psi_n = \mathcal{O}(h^2 + \tau^{m(\sigma)}), \qquad m(\sigma) = \begin{cases} 2, & \sigma = 0.5, \\ 1, & \sigma \neq 0.5. \end{cases} \tag{4.109}$$

To estimate the norm of the inaccuracy, we can invoke the a priori estimates that express stability of the difference scheme (4.107), (4.108) for the inaccuracy with respect to the right-hand side. We write the scheme (4.107), (4.108) in the canonical form

$$B\,\frac{z_{n+1}-z_n}{\tau}+Az_n=\psi_n,\qquad n=0,1,\ldots,N_0-1, \tag{4.110}$$

where the operator B is defined by (4.103).

In view of (4.103), we obtain an estimate for the difference solution of problem (4.108), (4.110) based on Theorem 4.8. In the case of interest, inequality (4.53) holds with $G=E$ if we choose $\sigma\geq 0.5$. The resulting estimate for the inaccuracy (see (4.54)) is

$$\|z_{n+1}\|_A^2\leq\frac{1}{2}\sum_{k=0}^{n}\tau\|\psi_k\|^2. \tag{4.111}$$

Taking the condition (4.108) into account and based on the estimate (4.111), we conclude that, with $\sigma\geq 0.5$, the weighted difference scheme (4.16), (4.17) converges in H_A at a rate of $\mathcal{O}(h^2+\tau^{m(\sigma)})$.

4.4.3 Stability of weighted three-layer schemes

The problem (4.1)–(4.3) can be approximately solved using the three-layer weighted scheme (4.18), (4.19). Consider conditions for stability of such schemes in the case of $\theta=0.5$, i.e., for the scheme (4.19), (4.20).

We write scheme (4.19), (4.20) in the canonical form (4.55), (4.56) with

$$B=E+\tau(\sigma_1-\sigma_2)A,\qquad R=\frac{\sigma_1+\sigma_2}{2}\,A. \tag{4.112}$$

Consider the case of a varying positive operator A ($A=A^*>0$ in the model problem (4.13), (4.14)).

Theorem 4.15 *Provided that $A>0$ and the conditions*

$$\sigma_1\geq\sigma_2,\qquad \sigma_1+\sigma_2>1/2 \tag{4.113}$$

are fulfilled, the scheme (4.19), (4.20) *is stable with respect to initial data, and for the difference solution (for $\varphi=0$) there holds the estimate*

$$\|Y^{n+1}\|_*\leq\|Y_1\|_*, \tag{4.114}$$

where

$$\|Y^{n+1}\|_*^2=\frac{1}{4}\,\|y_{n+1}+y_n\|^2+\frac{1}{2}\Big(\sigma_1+\sigma_2-\frac{1}{2}\Big)\|y_{n+1}-y_n\|^2.$$

Proof. Here, direct application of results concerning stability of three-layer operator-difference schemes is hampered by non-self-adjointness of A. We therefore start the proof with preliminary transformation of the difference scheme.

We act on (4.19), (4.20) with the operator A^{-1}; then, we obtain:

$$\tilde{B}\,\frac{y_{n+1}-y_{n-1}}{2\tau}+\tilde{R}(y_{n+1}-2y_n+y_{n-1})+\tilde{A}y_n\;=\;\tilde{\varphi}_n,\qquad n=1,2,\ldots, \tag{4.115}$$

where, in view of (4.112),

$$\tilde{B}=A^{-1}+\tau(\sigma_1-\sigma_2)E,\qquad \tilde{R}=\frac{\sigma_1+\sigma_2}{2}\,E,$$
$$\tilde{A}=E,\qquad \tilde{\varphi}=A^{-1}\varphi. \tag{4.116}$$

Let us apply Theorem 4.9 to scheme (4.115), (4.116). Under the assumptions (4.113), the following inequalities, that guarantee stability with respect to initial data, hold:

$$\tilde{R}-\frac{1}{4}\,\tilde{A}=\Big(\frac{\sigma_1+\sigma_2}{2}-\frac{1}{4}\Big)\tau^2E>0,$$
$$\tilde{B}=A^{-1}+(\sigma_1-\sigma_2)\tau E\geq 0.$$

Here, for the solution of the problem with homogeneous right-hand side (with $\varphi=0$) the estimate (4.114) holds. □

Based on the stability estimates with respect to initial data and right-hand side, convergence of three-layer difference schemes of type (4.19), (4.20) can be examined.

4.5 Program realization and computational experiments

Below, we consider a boundary value problem for the quasi-linear parabolic equation with nonlinear right-hand side. In the numerical solution of such problems, primary attention is normally paid to the use of linear difference schemes. Examples of numerical solutions of several model problems are given which illustrate some effects due to nonlinearity.

4.5.1 Problem statement

Consider boundary value problems for the quasi-linear parabolic equation

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\Big(k(x,u)\,\frac{\partial u}{\partial x}\Big)+f(x,t,u),\qquad 0<x<l,\quad 0<t\leq T, \tag{4.117}$$

whose coefficients depend not only on the spatial variable, but also on the solution itself. We restrict ourselves to the case of first-kind homogeneous boundary conditions, so that

$$u(0,t)=0,\qquad u(l,t)=0,\qquad 0<t\leq T, \tag{4.118}$$

$$u(x,0) = u_0(x), \qquad 0 \le x \le l. \tag{4.119}$$

First of all, note conditions under which the problem (4.117)–(4.119) has a unique solution. The study of such problems leans on results concerning the solution uniqueness in the related linear problem.

Suppose that in the domain $0 < x < l$, $0 < t \le T$ a function $u(x,t)$ satisfies the parabolic equation

$$\frac{\partial u}{\partial t} = a(x,t)\,\frac{\partial^2 u}{\partial x^2} + b(x,t)\,\frac{\partial u}{\partial x} - c(x,t)u + f(x,t) \tag{4.120}$$

with continuous coefficients and, in addition, $a(x,t) > 0$. Based on the principle of maximum, we can show that the solution of the linear problem (4.118)–(4.120) is unique. Note that the latter statement is valid irrespective of the sign of $c(x,t)$.

Assume that there exist two solutions of problem (4.117)–(4.119), $u_\alpha(x,t)$, $\alpha = 1, 2$:

$$\frac{\partial u_\alpha}{\partial t} = \frac{\partial}{\partial x}\Big(k(x,u_\alpha)\,\frac{\partial u_\alpha}{\partial x}\Big) + f(x,t,u_\alpha), \qquad 0 < x < l, \quad 0 < t \le T$$

with some given boundary and initial conditions. For the difference of the two solutions, $w(x) = u_2(x) - u_1(x)$, we obtain the following boundary value problem:

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial x}\Big(k(x,u_2)\,\frac{\partial w}{\partial x}\Big) + \frac{\partial}{\partial x}\Big(\frac{\partial k}{\partial u}(x,\bar u)\,\frac{\partial u_1}{\partial x}\,w\Big) + \frac{\partial f}{\partial u}(x,t,\bar u)w, \tag{4.121}$$

$$0 < x < l, \qquad 0 < t \le T,$$

$$w(0,t) = 0, \qquad w(l,t) = 0, \qquad 0 < t \le T, \tag{4.122}$$

$$w(x,0) = 0, \qquad 0 \le x \le l. \tag{4.123}$$

Here, the following settings were used:

$$\frac{\partial q}{\partial u}(x,\bar u) = \int_0^1 \frac{\partial q}{\partial u}(x,u_\theta)\,d\theta, \qquad u_\theta = \theta u_2 + (1-\theta)u_1.$$

The linear boundary value problem (4.121)–(4.123) belongs to the above-mentioned problem class (4.118)–(4.120). Hence, the trivial solution $w(x,t) = 0$ of the problem (4.121)–(4.123) is indeed unique provided that the coefficient $k(x,u)$, the right-hand side $f(x,t,u)$ and the solution of (4.117)–(4.119) are sufficiently smooth functions.

4.5.2 Linearized difference schemes

Above, we have considered difference schemes for the linear parabolic equation. Among these schemes, absolutely stable two- and three-layer implicit schemes have been distinguished. In application of analogous schemes to nonlinear problems it may happen so that we encounter difficulties with computational realization. In the case

of implicit approximations, at the next time layer we arrive at nonlinear difference equations. For the approximate solution to be found, we have to use these or those iteration methods for solving systems of nonlinear equations. To avoid this situation, in computational practice they widely use linearized difference schemes in which the solution at the next layer is found from a system of linear equations. We will illustrate some possibilities that arise along this line with the example of difference schemes for the nonlinear problem (4.117)–(4.119).

By analogy with (4.11), on the set of mesh functions given on $\bar{\omega}$ and vanishing on $\partial\omega$, we define the operator

$$A(v)y = -(a(x, v)y_{\bar{x}})_x, \qquad x \in \omega.$$

Here, the coefficient $a(x, v)$ is to be defined, for instance, as

$$a(x, v) = k(x - 0.5h, 0.5(v(x) + v(x - h))),$$

or

$$a(x, v) = \frac{1}{2}\,(k(x - h, v(x - h)) + k(x, v)).$$

To the initial differential problem (4.117)–(4.119), we put into correspondence the differential-difference problem

$$\frac{dy}{dt} + A(y)y = f(x, t, y), \qquad x \in \omega, \quad t > 0, \tag{4.124}$$

$$y(x, 0) = u_0(x), \qquad x \in \omega. \tag{4.125}$$

Let us discuss difference schemes for problem (4.124), (4.125). To begin with, consider nonlinear difference schemes in which the solution at the next time layer is found as the solution of a nonlinear difference problem. Such schemes can be constructed similarly to difference schemes for the linear parabolic equation. For instance, a purely implicit difference scheme for (4.124), (4.125) is

$$\frac{y_{n+1} - y_n}{\tau} + A(y_{n+1})y_{n+1} = f(x, t_{n+1}, y_{n+1}), \qquad n = 0, 1, \ldots, N_0 - 1, \tag{4.126}$$

$$y_0 = u_0(x), \qquad x \in \omega. \tag{4.127}$$

In the nonlinear scheme (4.126), (4.127), for the difference solution at the next time layer to be found, a nonlinear difference problem must be solved. For the values of y_{n+1} to be determined, these or those iterative processes are used. Some important specific features of the corresponding nonlinear difference problems deserve mention.

The first specific feature is related with the fact that in the iterative embodiment of implicit difference schemes a good initial approximation is always available. This initial approximation can be chosen as the solution at the previous layer. The second,

also beneficial specific feature is the fact that the difference problem for y_{n+1} involves a small parameter, the time step size τ. This parameter essentially affects the rate of convergence of the iterative process (the smaller τ, the higher is the rate of convergence in the process).

We can also identify the class of linearized difference schemes whose specific feature consists in that, here, the solution at the next time layer is to be found from the solution of a linear difference problem. A simplest scheme is characterized by that, in it, the coefficients are taken from the previous time layer. An example here is the difference scheme

$$\frac{y_{n+1} - y_n}{\tau} + A(y_n, y_{n+1}) = f(x, t_n, y_n). \tag{4.128}$$

This difference scheme has, apparently, an approximation inaccuracy of order $\mathcal{O}(\tau + |h|^2)$.

The main drawback of scheme (4.128) is often related with the fact that the right-hand side (source) is taken at the previous time layer. Further development of the difference scheme (4.128) is the scheme with quasi-linearized right-hand side:

$$\frac{y_{n+1} - y_n}{\tau} + A(y_n, y_{n+1}) = f(x, t_{n+1}, y_n) + \frac{\partial f}{\partial y}(x, t_{n+1}, y_n)(y_{n+1} - y_n). \tag{4.129}$$

Scheme (4.129), which remains linear, has wider stability margins with respect to the nonlinear right-hand side.

Linearized schemes can be constructed around the difference schemes "predictor-corrector", which conceptually border on the additive difference schemes (split schemes). Let us restrict ourselves here to a simplest variant of the predictor-corrector scheme:

$$\frac{\tilde{y}_{n+1} - y_n}{\tau} + A(y_n, y_n) = f(x, t_n, y_n). \tag{4.130}$$

Scheme (4.130) is used to calculate the coefficients and the right-hand side; that is why at the correction stage we can use the scheme

$$\frac{y_{n+1} - y_n}{\tau} + A(\tilde{y}_{n+1}, y_{n+1}) = f(x, t_{n+1}, \tilde{y}_{n+1}). \tag{4.131}$$

Alternatively, the linearization scheme (4.129) can be used at the correction stage.

The above schemes show considerable potential in constructing linearized difference schemes. In practical simulations, special methodical studies need to be performed aimed at making a proper choice of difference schemes for a particular class of nonlinear boundary value problems. In the case of nonlinear problems, theoretical considerations give only obscure guiding lines.

Of course, linearized difference schemes for problem (4.124), (4.125) can also be constructed around three-layer difference schemes. Without dwelling on a detailed

description, let us give here only two simplest examples. The problem (4.124), (4.125) can be approximately solved using the three-layer symmetric difference schemes

$$\frac{y_{n+1} - y_{n-1}}{2\tau} + A(y_n, \sigma y_{n+1} + (1 - 2\sigma)y_n + \sigma y_{n-1}) = f(x, t_n, y_n) \qquad (4.132)$$

with appropriate initial and boundary conditions. These linearized schemes are of the second approximation order both over space and time.

4.5.3 Program

Below, we give the text of a program in which two linearized schemes, schemes (4.128) and (4.129) were implemented. To find the solution at the next time layer, the standard sweep algorithm was employed.

Program PROBLEM3

```
C
C     PROBLEM3  - ONE-DIMENSIONAL NON-STATIONARY PROBLEM
C     QUASI-LINEAR 1D PARABOLIC EQUATION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( ISCHEME = 0, N = 1000, M = 1000 )

      DIMENSION  X(N+1), Y(N+1), A(N+1), B(N+1), C(N+1), F(N+1)
     +           ,ALPHA(N+2), BETA(N+2)
C
C     PROBLEM PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF THE SEGMENT;
C     ISCHEME   - PARAMETER FOR THE CHOICE OF THE SCHEME,
C     N + 1     - NUMBER OF GRID NODES OVER SPACE;
C     M + 1     - NUMBER OF GRID NODES OVER TIME;
C
      XL   = 0.D0
      XR   = 10.D0
      TMAX = 2.5D0
C
      OPEN ( 01, FILE = 'RESULT.DAT' ) ! FILE TO STORE THE COMPUTED DATA
C
C     GRID
C
      H    = ( XR - XL ) / N
      TAU  = TMAX / M
      DO I = 1, N+1
        X(I) = XL + (I-1)*H
      END DO

C
C     INITIAL CONDITION
C

      T = 0.D0
      XD = 2.5D0
      DO I = 1, N+1
```

```
        Y(I) = 0.D0
        IF ( X(I).LT.XD ) Y(I) = 1.D0
      END DO
      WRITE ( 01, * ) T
      WRITE ( 01, * ) (Y(I),I=1,N+1)
C
C     NEXT TIME LAYER
C
      DO K = 1, M
        T = K*TAU

C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        IF ( ISCHEME.EQ.0 ) THEN
C
C     LINEARIZED SCHEME WITH EXPLICIT RIGHT-HAND SIDE
C
        DO I = 2, N
          X1 = (X(I) + X(I-1)) / 2

          X2 = (X(I+1) + X(I)) / 2

          U1 = (Y(I) + Y(I-1)) / 2
          U2 = (Y(I+1) + Y(I)) / 2
          A(I) = AK(X1,U1) / (H*H)
          B(I) = AK(X2,U2) / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          F(I) = Y(I) / TAU + AF(X(I),T,Y(I))
        END DO
        END IF
        IF ( ISCHEME.EQ.1 ) THEN
C
C     LINEARIZED SCHEME WITH QUASI-LINEARIZED RIGHT-HAND SIDE
C
        DO I = 2, N
          X1 = (X(I) + X(I-1)) / 2
          X2 = (X(I+1) + X(I)) / 2
          U1 = (Y(I) + Y(I-1)) / 2
          U2 = (Y(I+1) + Y(I)) / 2
          A(I) = AK(X1,U1) / (H*H)
          B(I) = AK(X2,U2) / (H*H)

          C(I) = A(I) + B(I) + 1.D0 / TAU - ADF(X(I),T,Y(I))
          F(I) = Y(I) / TAU + AF(X(I),T,Y(I)) - ADF(X(I),T,Y(I))*Y(I)
        END DO
        END IF
C
C     LEFT BOUNDARY CONDITION
C
        B(1) = 0.D0
        C(1) = 1.D0
        F(1) = 1.D0
C
C     RIGHT BOUNDARY CONDITION
C
        A(N+1) = 0.D0
        C(N+1) = 1.D0
        F(N+1) = 0.D0
```

```
C
C     SOLUTION OF THE PROBLEM AT THE NEXT TIME LAYER
C
        CALL PROG ( N+1, A, C, B, F, ALPHA, BETA, Y )

C
C     SOLUTION RECORDING
C
        IF ( K/200*200.EQ.K ) THEN
          WRITE ( 01, * ) T
          WRITE ( 01, * ) (Y(I),I=1,N+1)
        END IF
      END DO
C
      CLOSE ( 01 )
      STOP
      END
C
      SUBROUTINE PROG ( N, A, C, B, F, AL, BET, Y )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     SWEEP METHOD
C
      DIMENSION A(N), C(N), B(N), F(N), Y(N), AL(N+1), BET(N+1)
C
      AL(1)  = B(1) / C(1)
      BET(1) = F(1) / C(1)
      DO I = 2, N
        SS     = C(I) - AL(I-1)*A(I)
        AL(I)  = B(I) / SS
        BET(I) = ( F(I) + BET(I-1)*A(I) ) / SS
      END DO
      Y(N) = BET(N)
      DO I = N-1, 1, -1
        Y(I) = AL(I)*Y(I+1) + BET(I)
      END DO
      RETURN

      END


      DOUBLE PRECISION FUNCTION  AK ( X, U )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     COEFFICIENT AT THE HIGHER DERIVATIVE
C
      AK = 0.2D0
C
      RETURN
      END

      DOUBLE PRECISION FUNCTION  AF ( X, T, U )
      IMPLICIT REAL*8 ( A-H, O-Z )

C
C     RIGHT-HAND SIDE OF THE EQUATION
C
      AF = 5.D0*U*(1.D0 - U)
C
```

```
      RETURN
      END

      DOUBLE PRECISION FUNCTION  ADF ( X, T, U )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     DIRIVATIVE OF THE RIGHT-HAND SIDE OF THE EQUATION
C
      ADF = 5.D0 - 10.D0*U
C
      RETURN
      END
```

The above program text refers to the case in which equation (4.117) is solved with the coefficients

$$k(x,u) = \kappa, \qquad f(x,t,u) = \chi u(1-u), \tag{4.133}$$

where $\kappa = 0.2$ and $\chi = 5$. The latter nonlinear diffusion-reaction equation is known as the Fisher equation (Kolmogorov–Petrovskii–Piskunov equation).

4.5.4 Computational experiments

Very often, nonlinearity brings about new effects related with an unusual behavior of the solution. The latter is also the case with the quasi-linear parabolic equation (4.117). A typical example here is the Fisher equation (4.117), (4.133).

For the linear parabolic-type equation (4.1), infinite propagation velocity of perturbations is typical. In consideration of equations with nonlinear coefficients and nonlinear right-hand side, one can identify solutions with a finite velocity of perturbations. In the case of the Fisher equation (4.117), (4.133), there exist solutions of the traveling-wave type:

$$u(x,t) = \psi(\xi), \qquad \xi = x - ct,$$

where $c = \text{const}$ is the wave velocity.

A. N. Kolmogorov, I. G. Petrovskii and N. S. Piskunov showed that, for instance, with

$$u(x,0) = \begin{cases} 1, & x < x^*, \\ 0, & x > x^* \end{cases}$$

the solution of the Cauchy problem for equation (4.117), (4.133) is unique, presenting a wave traveling at the velocity $c = 2\sqrt{\kappa\chi}$. Figure 4.1 shows the solution of the

boundary value problem for equation (4.117), (4.133) with $\kappa = 0.2$, $\chi = 5$ and with the following boundary and initial conditions:

$$u(0,t) = 1, \qquad u(10,t) = 1,$$
$$u(x,0) = \begin{cases} 1, & x < x^*, \\ 0, & x > x^*, \end{cases} \qquad 0 < x < 10$$

where $x^* = 2.5$.

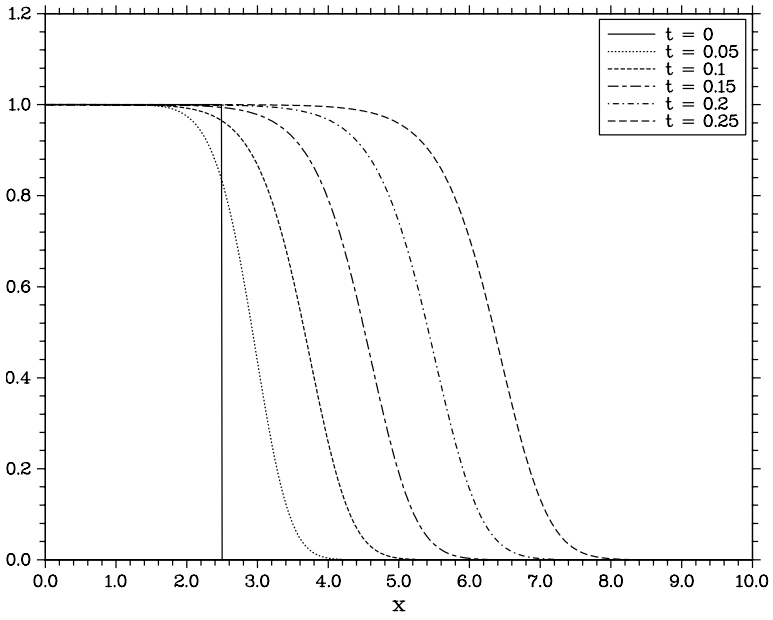

Figure 4.1 The solution at various times

In the calculations, a fine calculation grid with $h = 0.01$, $\tau = 0.0025$ and scheme (4.129) were used. Transformation of the initially rectangular waveform into a solitary wave traveling to the right is observed.

Some specific features of schemes in which linearization of the right-hand side of the equation is used are illustrated by Figures 4.2 and 4.3 that show calculation data obtained with a larger time step size ($\tau = 0.025$) using the linearized schemes (4.128) (Figure 4.2) and (4.129) (Figure 4.3). The scheme with explicit right-hand side yields more accurate results, whereas in the case of the scheme with right-hand side partially transferred to the next time layer the front of the approximate solution propagates somewhat faster. It should be noted that the conclusion about certain advantages of non-linearized schemes concerns only the considered problem; in other cases, a linearized scheme may yield better results.

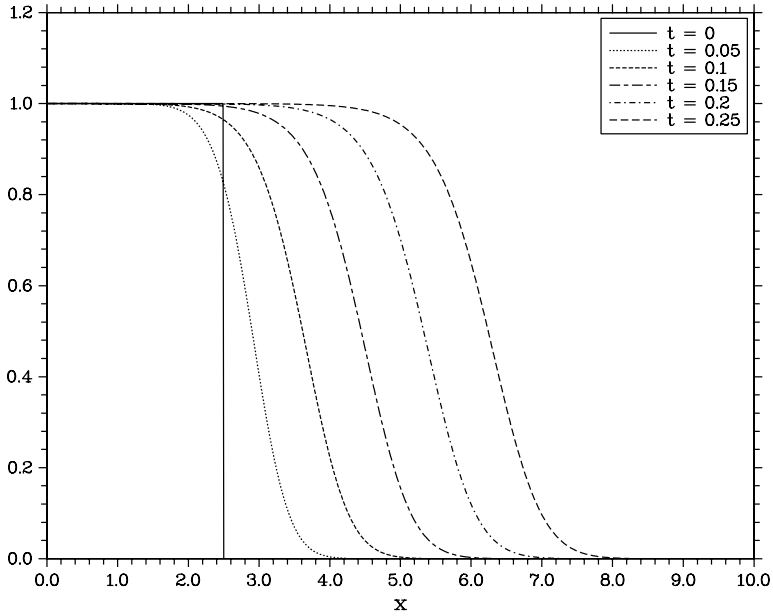

Figure 4.2 Difference scheme with non-linearized right-hand side

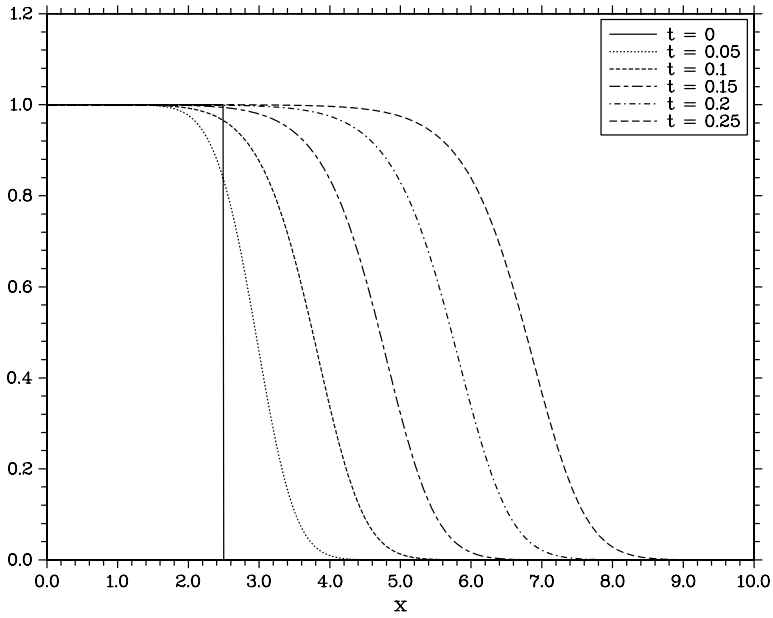

Figure 4.3 Difference scheme with linearized right-hand side

4.6 Exercises

Exercise 4.1 Consider the Cauchy problem for the second-order evolutionary equation:

$$\frac{d^2u}{dt^2} + Au = f(t), \qquad t > 0, \tag{4.134}$$

$$u(0) = u_0, \qquad \frac{du}{dt}(0) = v_0. \tag{4.135}$$

Show that in the case of $A = A^* > 0$ the following a priori estimate holds for the solution of problem (4.134), (4.135):

$$\|u(t)\|_*^2 \le \exp(t)\Big(\|u_0\|_A^2 + \|v_0\|^2 + \int_0^t \exp(-\theta)\|f(\theta)\|^2 d\theta\Big)$$

where

$$\|u\|_*^2 = \left\|\frac{du}{dt}\right\|^2 + \|u\|_A^2.$$

Exercise 4.2 Construct a two-layer weighted scheme for the problem (4.1), (4.3) with the third-kind boundary conditions (4.4).

Exercise 4.3 Suppose that we seek an approximate solution of the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \qquad 0 < x < l, \quad 0 < t \le T,$$

supplemented with conditions (4.2), (4.3). Construct a two-layer scheme with an approximation order $\mathcal{O}(\tau^2 + h^4)$.

Exercise 4.4 In the class of weighted schemes

$$\Big(\theta\,\frac{y_{n+1} - y_n}{\tau} + (1-\theta)\,\frac{y_n - y_{n-1}}{\tau}\Big) + A(\sigma_1 y_{n+1} + (1 - \sigma_1 - \sigma_2)y_n + \sigma_2 y_{n-1}) = \varphi_n,$$

construct a scheme with an approximation order $\mathcal{O}(\tau^3)$.

Exercise 4.5 Prove Theorem 4.3.

Exercise 4.6 Establish stability conditions for the scheme

$$\frac{3y_{n+1} - 4y_n + y_{n-1}}{2\tau} + Ay_{n+1} = 0$$

in the case of $A = A^* > 0$.

Exercise 4.7 Suppose that in an examination of the weighted scheme

$$\frac{y_{n+1} - y_n}{\tau} + A(\sigma y_{n+1} + (1 - \sigma))y_n = \varphi_n \tag{4.136}$$

the pure implicit scheme (with $\sigma = 1$) was found to be stable with respect to initial data and right-hand side. Show that, in this case, all schemes with $\sigma > 1$ are also stable.

Exercise 4.8 Suppose that for the scheme

$$\begin{aligned} B\,\frac{y_{n+1} - y_n}{\tau} + A\,y_n &= \varphi_n, \qquad t_n \in \omega_\tau, \\ y_0 &= u_0 \end{aligned}$$

with a constant operator $A = A^* > 0$ there holds the inequality

$$B \geq \frac{\tau}{2} A.$$

Show that for the solution there holds the following stability estimate with respect to initial data and right-hand side:

$$\|y_{n+1}\|_A \;\leq\; \|u_0\|_A + \|\varphi_0\|_{A^{-1}} + \|\varphi_n\|_{A^{-1}} + \sum_{k=1}^{n} \tau \left\| \frac{\varphi_k - \varphi_{k-1}}{\tau} \right\|.$$

Exercise 4.9 Based on the maximum principle for difference schemes, establish conditions of stability in $C(\omega)$ for the weighted scheme (4.136) with

$$Ay = -(a y_{\bar{x}})_x, \qquad x \in \omega$$

used to solve the boundary value problem (4.1)–(4.3).

Exercise 4.10 Examine stability of the three-layer symmetric scheme

$$\frac{y_{n+1} - y_{n-1}}{2\tau} + A(\sigma y_{n+1} + (1 - 2\sigma)y_n + \sigma y_{n-1}) \;=\; \varphi_n,$$

used to solve the boundary value problem (4.1)–(4.3).

Exercise 4.11 In the solution of problem (4.134), (4.135), one can naturally use the weighted scheme

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} + A(\sigma_1 y_{n+1} + (1 - \sigma_1 - \sigma_2)y_n + \sigma_2 y_{n-1}) \;=\; \varphi_n,$$
$$n = 1, 2, \ldots, N_0 - 1.$$

Obtain stability conditions for this scheme.

Exercise 4.12 Modify the program PROBLEM3 by implementing in it the three-layer difference scheme (4.132). Conduct computational experiments on the use of the scheme for approximate solution of the model problem for equation (4.117), (4.133).

Exercise 4.13 Construct an automodel solution of the traveling-wave type for equation (4.117) with coefficients

$$k(x,u) = u^{\sigma}, \qquad f(x,t,u) = 0.$$

Conduct numerical experiments to model such regimes with the use of the program PROBLEM3.

Exercise 4.14 Numerically examine the development of local perturbations in the Cauchy problem for equation (4.117) with

$$k(x,u) = u^{\sigma}, \qquad f(x,t,u) = u^{\sigma+1}.$$

Exercise 4.15 Write, around the program PROBLEM2, which numerically solves difference elliptic equations, a program that numerically solves the linear Dirichlet boundary value problem for the two-dimensional parabolic equation

$$\frac{\partial u}{\partial t} + \sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x}) \frac{\partial u}{\partial x_\alpha}\Big) = f(\boldsymbol{x},t), \qquad \boldsymbol{x} \in \Omega, \quad t > 0,$$

in the rectangle

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\alpha < l_\alpha,\ \alpha = 1,2\}.$$

5 Solution methods for ill-posed problems

Basic approaches for the approximate solution of ill-posed problems use this or that perturbation of the initial problem; such a perturbation allows one to pass to some "close" yet well-posed problem. In this way, various regularization algorithms can be obtained. With the example of the first-order operator equation, below we consider basic approaches to the solution of unstable problems. In particular, an ill-posed problem can be replaced with a well-posed problem by perturbing the initial equation, by passing to a variational problem, etc. Of primary importance here is an appropriate choice of the regularization parameter, whose value needs to be properly matched with input-data inaccuracies. Illustrative calculations performed for the first-kind integral equation are given.

5.1 Tikhonov regularization method

Following A. N. Tikhonov, consider the general approach of constructing stable computational algorithms for the approximate solution of ill-posed problems. The method is based on the passage from the initial first-kind equation to a minimization problem for a functional with an additional stabilizing term.

5.1.1 Problem statement

Normally, methods for the numerical solution of ill-posed problems are considered as applied to the first-kind linear operator equation

$$Au = f. \tag{5.1}$$

The right-hand side and the solution itself belong to some metric spaces.

To not overload the consideration with technical details, we restrict ourselves to the case of a linear operator A under conditions acting in a Hilbert space H (for simplicity, here we assume that $u \in H$, $f \in H$, i.e., $A : H \to H$). In H, we introduce the scalar product (u, v) and the norm $\|u\|$ for elements $u, v \in H$.

Even more, we assume that the operator A is self-adjoint, positive and compact, with eigenvalues λ_k tending to zero as $k \to \infty$ ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \cdots > 0$), and the related eigenfunction system $\{w_k\}$, $w_k \in H$, $k = 1, 2, \ldots$ is an orthonormal system complete in H. Hence, for each $v \in H$ we have

$$v = \sum_{k=1}^{\infty} v_k w_k, \qquad v_k = (v, w_k).$$

Typical of practical studies is a situation in which input data are given with some inaccuracy. To model this general situation, consider a case in which the right-hand

side of (5.1) is given accurate to some δ. Instead of f, we know f_δ such that

$$\|f_\delta - f\| \le \delta. \tag{5.2}$$

In a more general case, we consider problems (5.1) in which not only the right-hand side but also the operator A are known approximately.

We pose a problem in which it is required to find an approximate solution of equation (5.1) with an approximately given right-hand side f_δ. This approximate solution is denoted by u_α, and the parameter α can be naturally related with the inaccuracy level in the right-hand side, i.e., $\alpha = \alpha(\delta)$.

5.1.2 Variational method

The main idea behind the construction of stable methods for solving ill-posed problems is based on the use of some a priori information about the input-data inaccuracy. Once the right-hand side is given with an inaccuracy, no attempts to solve the equation

$$Au_\alpha = f_\delta \tag{5.3}$$

exactly are necessary. We can try to compensate for the uncertainty in the right-hand side, for instance, by passing to some another yet well-posed problem

$$A_\delta u_\alpha = f_\delta,$$

whose operator A_δ possesses improved properties compared to A.

In variational methods, instead of solving equation (5.3), they minimize the norm of the difference $r = Av - f_\delta$, or the *discrepancy functional*

$$J_0(v) = \|Av - f_\delta\|^2.$$

There are many solutions to such a variational problem that satisfy equation (5.3) accurate to some discrepancy δ. What is only needed is to wisely take into account all available information about the inaccuracy in the right-hand side and, in this way, to single out a most appropriate solution.

In the *Tikhonov regularization method*, they introduce some *smoothing functional*

$$J_\alpha(v) = \|Av - f_\delta\|^2 + \alpha\|v\|^2. \tag{5.4}$$

The approximate solution of the initial problem (5.1), (5.2) is the extremal of the functional:

$$J_\alpha(u_\alpha) = \min_{v \in H} J_\alpha(v). \tag{5.5}$$

In (5.4), $\alpha > 0$ is the *regularization parameter*, whose value must be matched with the right-hand side inaccuracy δ. For a bounded solution to be separated out, the discrepancy functional contains an additional *stabilizing functional* $\|v\|^2$.

The key point in the theoretical consideration of approximate algorithms is related with the proof of convergence of the approximate solution to the exact solution. It is required to find under which conditions the approximate solution u_α found from (5.4), (5.5) converges to the exact solution of problem (5.1). Ideally, it is required not only to establish the fact that convergence takes place but also to find the rate of convergence.

We represent the approximate solution in the operator form

$$u_\alpha = R(\alpha) f_\delta. \tag{5.6}$$

If the approximate solution converges to the exact solution with the right-hand side inaccuracy tending to zero, then they say that the operator $R(\alpha)$ is a *regularizing operator*. With the chosen structure of $R(\alpha)$, it is also required to substantiate the choice of regularization parameter α as a function of δ.

5.1.3 Convergence of the regularization method

In a consideration of conditionally well-posed problems, it is required to identify the class of desired solutions and explicitly give a priori constraints on the solution. In problem (5.1), (5.2), we are interested in bounded solutions and, hence, the a priori constraint on the solution looks as

$$\|u\| \le M, \tag{5.7}$$

where $M = \text{const} > 0$. The main result here can be formulated as follows.

Theorem 5.1 *Let for the right-hand side inaccuracy the estimate* (5.2) *holds. Then, the approximate solution u_α found as the solution of problem* (5.4), (5.5) *with $\alpha(\delta) \to 0$, $\delta^2/\alpha(\delta) \to 0$ converges in H, as $\delta \to 0$, to the bounded exact solution u.*

Proof. Under the indicated assumptions concerning the operator A, the exact solution of (5.1) can be represented as

$$u = \sum_{k=1}^{\infty} \frac{1}{\lambda_k} (f, w_k) w_k. \tag{5.8}$$

Suppose that

$$v = \sum_{k=1}^{\infty} c_k w_k, \qquad c_k = (v, w_k).$$

Then, the functional $J_\alpha(v)$ takes the form

$$J_\alpha(v) = \sum_{k=1}^{\infty} \left((\lambda_k c_k - (f_\delta, w_k))^2 + \alpha c_k^2 \right).$$

Condition (5.5) is equivalent to

$$\frac{\partial J_\alpha}{\partial c_k} = 2\lambda_k(\lambda_k c_k - (f_\delta, w_k)) + 2c_k = 0, \qquad k = 1, 2, \dots .$$

From here, we obtain the following representation for the solution of problem (5.4), (5.5):

$$u_\alpha = \sum_{k=1}^{\infty} \frac{\lambda_k}{\lambda_k^2 + \alpha} (f_\delta, w_k) w_k. \tag{5.9}$$

For the inaccuracy $z = u_\alpha - u$, we use the representation

$$z = z_1 + z_2, \qquad z_1 = u_\alpha - R(\alpha) f, \qquad z_2 = R(\alpha) f - u. \tag{5.10}$$

In (5.10), $R(\alpha) f$ is the solution to the minimization problem for the smoothing functional with the right-hand side given exactly.

In view of (5.8), (5.9), we obtain:

$$\|z_1\|^2 = \sum_{k=1}^{\infty} \frac{\lambda_k^2}{(\lambda_k^2 + \alpha)^2} ((f_\delta, w_k) - (f, w_k))^2.$$

For non-negative x we have:

$$\frac{x}{x^2 + \alpha} = \frac{1}{(\sqrt{x} - \sqrt{\alpha/x})^2 + 2\sqrt{\alpha}} \le \frac{1}{2\sqrt{\alpha}}$$

and, hence, under the assumptions of (5.2) there holds the two-sided inequality

$$\|z_1\|^2 \le \frac{1}{4\alpha} \sum_{k=1}^{\infty} ((f_\delta, w_k) - (f, w_k))^2 \le \frac{\delta^2}{4\alpha}. \tag{5.11}$$

Estimate (5.11) expresses the fact that the solution of problem (5.4), (5.5) is stable with respect to weak right-hand side perturbations.

Let us estimate now z_2 in the representation (5.10) for the approximate-solution inaccuracy. In this case, we deal with the conditions of closeness of the solution of (5.1) to the solution of the minimization problem for the smoothing functional with the same right-hand sides. From (5.8), (5.9) we have:

$$\|z_2\|^2 = \sum_{k=1}^{\infty} \frac{\alpha^2}{\lambda_k^2(\lambda_k^2 + \alpha)^2} (f, w_k)^2.$$

We are going to establish the closeness of $R(\alpha) f$ to u for functions from class (5.7). Let us show that for an arbitrary $\varepsilon > 0$ we can choose $\alpha(\varepsilon)$ such that $\|z_2\|^2 \le \varepsilon$ for all α from the interval $0 < \alpha \le \alpha(\varepsilon)$. For all functions from (5.7), the series

$$\|u\|^2 = \sum_{k=1}^{\infty} \frac{1}{\lambda_k^2} (f, w_k)^2$$

converges and, hence, we can found a number $n(\varepsilon)$ such that

$$\sum_{k=n(\varepsilon)+1}^{\infty} \frac{1}{\lambda_k^2} (f, w_k)^2 \le \frac{\varepsilon}{2}.$$

Under the stated conditions, we obtain the inequality

$$\|z_2\|^2 \le \sum_{k=1}^{n(\varepsilon)} \frac{\alpha^2}{\lambda_k^2(\lambda_k^2+\alpha)^2} (f, w_k)^2 + \sum_{k=n(\varepsilon)+1}^{n(\varepsilon)} \frac{1}{\lambda_k^2}(f, w_k)^2.$$

At the expense of a sufficiently small chosen $\alpha(\varepsilon)$, for the first term we have:

$$\sum_{k=1}^{n(\varepsilon)} \frac{\alpha^2}{\lambda_k^2(\lambda_k^2+\alpha)^2} (f, w_k)^2 \le \frac{\varepsilon}{2}.$$

The latter inequality holds for all α in the interval $0 < \alpha \le \alpha(\varepsilon)$. Hence, we obtain that $s(\alpha) = \|z_2\| \to 0$ as $\alpha \to 0$.

Substitution into (5.10) yields

$$\|z\| \le \|z_1\| + \|z_2\| \le \frac{\delta}{2\sqrt{\alpha}} + s(\alpha). \tag{5.12}$$

By virtue of the above, if $\alpha(\delta) \to 0$ and $\delta^2/\alpha(\delta) \to 0$ as $\delta \to 0$, then $\|z\| \to 0$. □

The proved theorem remains valid under even more general conditions. A fundamental generalization here can be obtained considering problems (5.1), (5.2) whose operator A is not a self-adjoint operator.

5.2 The rate of convergence in the regularization method

In the class of bounded solutions, we have established previously that the approximate solution in the Tikhonov regularization method converges to the exact solution. Under more tight constraints on the exact solution, one can evaluate the rate at which the method converges.

5.2.1 Euler equation for the smoothing functional

Instead of the extremum problem (5.4), (5.5), we can consider a related Euler equation. In the latter case, the approximate solution can be found as the solution of the following second-kind equation:

$$A^*Au_\alpha + \alpha u_\alpha = A^* f_\delta. \tag{5.13}$$

The transfer from the ill-posed problem (5.3) to the well-posed problem (5.13) can be made passing to a problem with a self-adjoint operator A^*A. This can be done by

multiplying equation (5.3) from the left by A^*, followed by a subsequent perturbation of the resulting equation with the operator αE.

The mentioned equivalence between the minimization problem (5.4), (5.5) and equation (5.13) provides us with some freedom in the computational realization of the method. The variational approach is quite general, allowing one to consider various classes of problems from a unified methodological standpoint. The numerical solution can most conveniently be found using the related Euler equation.

In the case of $A = A^* \geq 0$, we can restrict ourselves to perturbing the operator itself:

$$Au_\alpha + \alpha u_\alpha = f_\delta. \tag{5.14}$$

Problem (5.14) refers to the *simplified regularization* algorithm.

In fact, we can say that, in addition to variational solution methods for ill-posed problems, a second class of approximate methods can be identified, which (see, for instance, (5.13), (5.14)) features some perturbation of the operator in the initial or transformed problem.

5.2.2 Classes of a priori constraints imposed on the solution

We have considered the Tikhonov regularization method under the assumption (5.7) about the exact solution of the ill-posed problem (5.1), (5.2). For the solution inaccuracy we have obtained the estimate (5.12), in which $s(\alpha) \to 0$ provided that $\alpha \to 0$. Note that the convergence was established in the same norm in which the a priori constraints on the solution were formulated. We would also like to obtain an estimate for the rate of convergence of the approximate solution to the exact solution with $\alpha(\delta) \to 0$ and $\delta \to 0$.

To clarify the situation, consider the problem of approximate calculation of a derivative. We use the difference relation

$$u_{\overset{\circ}{x}} = \frac{u(x+h) - u(x-h)}{2h}, \tag{5.15}$$

and assume that the function $u(x)$ is differentiable at each x. We would like to know how accurately the difference derivative (5.15) approximates the derivative du/dx at some point x.

To prove the convergence and evaluate the rate with which the central difference derivative (5.15) converges to $du/dx(x)$, let us formulate more tight constraints on $u(x)$. If, for instance, $u(x)$ is a twice differentiable function, then from (5.15) we obtain

$$u_{\overset{\circ}{x}} = \frac{du}{dx} + \mathcal{O}(h).$$

For a three times differentiable function we have

$$u_{\overset{\circ}{x}} = \frac{du}{dx} + \mathcal{O}(h^2).$$

Thus, with reduced smoothness of differentiable functions the approximation inaccuracy decreases.

We encounter a similar situation when examining the convergence rate in the regularization method. Instead of (5.17), we are going to formulate more tight constraints on the exact solution of problem (5.1). It seems reasonable to associate the requirement for enhanced smoothness of the exact solution with the operator A. We assume that the exact solution belongs to the class

$$\|A^{-1}u\| \leq M. \tag{5.16}$$

Then, in the case of a self-adjoint positive operator A the intermediate (between (5.7) and (5.16)) position belongs to the following type of classes of a priori constraints on the solution:

$$\|u\|_{A^{-1}} \leq M. \tag{5.17}$$

Under conditions (5.16) and (5.17), one can try to render more concrete the dependence $s(\alpha)$ in the inaccuracy estimate of type (5.12) in the regularization method.

5.2.3 Estimates of the rate of convergence

We now derive estimates for the solution inaccuracy in the Tikhonov regularization method based on the a priori estimates of the operator equation (5.13) which, in view of $A = A^* > 0$, assumes the form

$$A^2 u_\alpha + \alpha u_\alpha = A f_\delta. \tag{5.18}$$

Theorem 5.2 *For the inaccuracy $z = u_\alpha - u$ of the approximate solution of problem* (5.1)*,* (5.2) *found from* (5.13)*, there hold the a priori estimate*

$$\|z\|^2 \leq \frac{\delta^2}{2\alpha} + \frac{\alpha}{2} M^2, \tag{5.19}$$

for exact solutions from class (5.16) and the estimate

$$\|z\|^2 \leq \frac{\delta^2}{\alpha} + \frac{\sqrt{\alpha}}{2} M^2 \tag{5.20}$$

for solutions from (5.17).

Proof. We subtract the equation

$$A^2 u = A f,$$

from (5.18); this yields the following equation for the inaccuracy $z = u_\alpha - u$:

$$A^2 z + \alpha z = A(f_\delta - f) - \alpha u.$$

We scalarwise multiply this equation by z and obtain

$$\|Az\|^2 + \alpha\|z\|^2 = ((f_\delta - f), Az) - \alpha(u, z).$$

Then, insertion of the inequality

$$((f_\delta - f), Az) \;\le\; \frac{1}{2}\,\|Az\|^2 + \frac{1}{2}\,\|f_\delta - f\|^2$$

gives

$$\frac{1}{2}\,\|Az\|^2 + \alpha\|z\|^2 \;\le\; \frac{1}{2}\,\|f_\delta - f\|^2 - \alpha(u, z). \tag{5.21}$$

Consider first the case of a priori constraints (5.16). The use of the inequality

$$-\alpha(u, z) = -\alpha(A^{-1}u, Az) \;\le\; \frac{1}{2}\,\|Az\|^2 + \frac{\alpha^2}{2}\,\|A^{-1}u\|^2$$

in (5.21) leads us to the estimate

$$\alpha\|z\|^2 \le \frac{1}{2}\,\|f_\delta - f\|^2 + \frac{\alpha^2}{2}\,\|A^{-1}u\|^2.$$

With inequalities (5.2) and (5.16) taken into account, we arrive at the desired estimate (5.19).

In the case of (5.17) the last term in the right-hand side of (5.21) can be estimated as

$$-\alpha(u, z) = -\alpha(A^{-1/2}u, A^{1/2}z) \;\le\; \sqrt{\alpha}\,\|A^{1/2}z\|^2 + \frac{\alpha\sqrt{\alpha}}{4}\,\|A^{1/2}u\|^2.$$

Taking the fact into account that

$$\frac{1}{2}\,\|Az\|^2 + \alpha\|z\|^2 - \sqrt{\alpha}\,\|A^{1/2}z\|^2 \;=\; \left\|\frac{1}{\sqrt{2}}\,Az - \frac{\sqrt{\alpha}}{\sqrt{2}}\,z\right\|^2 + \frac{\alpha}{2}\,\|z\|^2,$$

we obtain

$$\frac{\alpha}{2}\,\|z\|^2 \le \frac{1}{2}\,\|f_\delta - f\|^2 + \frac{\alpha\sqrt{\alpha}}{4}\,\|A^{1/2}u\|^2.$$

In view of (5.2), (5.17), we have the estimate (5.20). □

Tightened constraints on the smoothness of the exact solution result in an improved rate of convergence of the approximations to the exact solution (see estimate (5.19), (5.20)).

5.3 Choice of regularization parameter

Below several choices of the regularization parameter, whose value must be matched with the input-data inaccuracy, are outlined.

5.3.1 The choice in the class of a priori constraints on the solution

In the theory of approximate solution methods for ill-posed problems, considerable attention is paid to the choice of regularization parameter. The most widespread choices here are the choices based on the discrepancy between the solutions, on the generalized discrepancy (which takes into account both the right-hand side inaccuracy and the inaccuracy in the operator A), the quasi-optimal choice, etc. A good choice of regularization parameter largely determines the efficiency of the computational algorithm.

The value of the regularization parameter α must be matched with the input-data inaccuracy: the smaller is the inaccuracy, the smaller is the value of regularization parameter that is to be chosen, i.e., $\alpha = \alpha(\delta)$. To comply with the structure of the inaccuracy (see (5.12), (5.19), (5.20)), the regularization parameter cannot be chosen too small: it is in the fact that, with decreased value of regularization parameter, the inaccuracy also decreases, that the ill-posedness of the problem is manifested. Hence, there exists an optimal value of regularization parameter that minimizes the approximate-solution inaccuracy.

The optimal value of regularization parameter depends not only on the inaccuracy in the right-hand side, but also on the class of a priori constraints on the exact solution. For instance, in the case of bounded solutions (class (5.7)) the above estimate (5.12) for the approximate-solution inaccuracy in the Tikhonov method does not allow one to explicitly give the optimal value of regularization parameter.

On narrowing the class of exact solutions, it becomes possible to render concrete the choice of regularization parameter. In the class of exact solutions (5.16) there holds the a priori estimate (5.19) for the inaccuracy, and for the optimal value of regularization parameter we obtain the expression

$$\alpha_{\text{opt}} = \frac{\delta}{M}, \qquad \|A^{-1}u\| \le M. \tag{5.22}$$

With this value of regularization parameter, a rate

$$\|z\| \le \sqrt{M\delta}$$

of convergence of the approximate solution to the exact solution can be achieved.

A similar consideration for the choice of regularization parameter in the class (5.17) of a priori constraints on the exact solution leads us to

$$\alpha_{\text{opt}} = \left(4\,\frac{\delta^2}{M^2}\right)^{2/3}, \qquad \|u\|_{A^{-1}} \le M \tag{5.23}$$

with

$$\|z\| \le \sqrt{3}\sqrt[3]{\frac{M^2\delta}{4}}.$$

To summarize, we can formulate the following statement.

Theorem 5.3 *If one chooses an optimal value of regularization parameter by the rule* (5.22) *in the class* (5.16) *of exact solutions or by the rule* (5.23) *in the class* (5.17) *of such solutions, then for the approximate-solution inaccuracy we have:*

$$\|z\| = \mathcal{O}(\delta^{\beta}), \qquad \beta = \begin{cases} 1/2, & \|A^{-1}u\| \le M, \\ 1/3, & \|u\|_{A^{-1}} \le M. \end{cases} \tag{5.24}$$

An optimum value of regularization parameter can be chosen provided that the inaccuracy level in the right-hand side (the constant δ in (5.2)) and the class of a priori constraints on the exact solution (the constant M in the estimate (5.16) or (5.17)) are known. In solving practical problems, such a priori information can be partially or even fully lacking. That is why in such cases we have to use another choice of regularization. Consider some possibilities available along this line.

5.3.2 Discrepancy method

In *choosing the regularization parameter from the discrepancy* , the master equation is

$$\|Au_{\alpha} - f_{\delta}\| = \delta. \tag{5.25}$$

This choice of regularization parameter, or, in other words, the convergence of the approximate solution u_{α} with $\alpha = \alpha(\delta)$ to the exact solution of (5.1) with $\delta \to 0$ was considered for many classes of problems. Note some specific features in the computational implementation of this method in choosing the regularization parameter.

The difference between the approximation and the exact solution, or the discrepancy, somehow depends on α. We introduce the notation

$$\varphi(\alpha) = \|Au_{\alpha} - f_{\delta}\|;$$

then, to find the regularization parameter, we must solve, in line with the discrepancy principle (5.25), the equation

$$\varphi(\alpha) = \delta. \tag{5.26}$$

Under rather general conditions, the function $\varphi(\alpha)$ is a non-decreasing function, and equation (5.26) has a solution.

Equation (5.26) can be approximately solved using various computational procedures. For instance, we can use the succession

$$\alpha_k = \alpha_0 q^k, \qquad q > 0. \tag{5.27}$$

Here, the calculations are to be made starting from $k = 0$ and going to a certain $k = K$ at which equality (5.26) becomes fulfilled to an acceptable accuracy. With so defined regularization parameter, we need $K + 1$ calculations of the discrepancy (for the solutions of variational problems of type (5.4), (5.5) or for the solutions of the Euler equations (5.13)).

In finding the approximate solution of (5.26), more rapidly converging iterative methods can also be used. It was found that the function $\psi(\beta) = \varphi(1/\beta)$ is a decreasing convex function. Hence, in solving the equation

$$\psi(\beta) = \delta$$

one can use the Newton iterative method, in which

$$\beta_{k+1} = \beta_k - \frac{\psi(\beta_k) - \delta}{\psi'(\beta_k)}.$$

This method converges for any initial approximation $\beta_0 > 0$. To avoid the calculation of the derivative of $\psi(\beta)$, we can use the iterative secant method, in which

$$\beta_{k+1} = \beta_k - \frac{\beta_k - \beta_{k-1}}{\psi(\beta_k) - \psi(\beta_{k-1})} (\psi(\beta_k) - \delta).$$

The use of such iterative procedures reduces the computational cost in the determination of α.

5.3.3 Other methods for choosing the regularization parameter

With no reliable estimates for the inaccuracy in the input data (of type (5.2)) at hand, the use of the well-accepted discrepancy method meets serious difficulties. As an alternative, many other methods for choosing the regularization parameters have gained widespread use in the computational practice.

The choice of the *quasi-optimal value of regularization parameter* is never directly related with the level of δ. We choose a value of $\alpha > 0$ that minimizes

$$\chi(\alpha) = \left\| \alpha \frac{du_\alpha}{d\alpha} \right\|. \tag{5.28}$$

The quasi-optimal value is most frequently found using the sequence (5.27). Minimization of (5.28) with such values of the regularization parameter corresponds to searching for the minimum of

$$\tilde{\chi}(\alpha_{k+1}) = \|u_{\alpha_{k+1}} - u_{\alpha_k}\|.$$

Hence, it is required to estimate just the norm of the difference between approximate solutions at two neighboring points.

There exist other methods to choose the regularization parameter. It is worth noting that a good choice of regularization parameter allows considerable computational savings and, to this or that extent, can be made iteratively. For each particular value of an iteration parameter, we solve problem (5.4), (5.5) or equation (5.13). Of course, at intermediate values of the regularization parameter there is little point in very accurate solution of such problems. That is why we can combine the determination of the solution of problem (5.13) with optimization of the value of the regularization parameter. In fact, a closely related idea is implemented in iterative solution methods using integrated iteration-parameter and regularization-parameter functions.

5.4 Iterative solution methods for ill-posed problems

In solving ill-posed problems, iteration methods are successfully used. In this case, as a regularization parameter, the number of iterations is used. With the example of problem (5.1), (5.2), below we discuss specific features in the application of iteration methods.

5.4.1 Specific features in the application of iteration methods

Let us discuss specific features of iteration methods as applied to the solution of ill-posed problems. Consider the model problem (5.1), (5.2) with a self-adjoint, positive operator A. The ill-posedness of the problem is related with the fact that the eigenvalues of A, taken in decreasing order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \cdots > 0$), tend to zero as $k \to \infty$.

The general two-layer iteration solution method for equation (5.3) with approximately given right-hand side can be written as

$$B \frac{u_{k+1} - u_k}{\tau_{k+1}} + A u_k = f_\delta, \qquad k = 0, 1, \ldots . \tag{5.29}$$

Here, $B : H \to H$ and the operator B^{-1} does exist.

If the operator in the initial problem (5.1), (5.2) is not a self-adjoint positive operator, then preliminary Gaussian symmetrization must be applied. The use of the iteration method for the symmetrized problem corresponds to the use of the equation

$$B \frac{u_{k+1} - u_k}{\tau_{k+1}} + A^* A u_k = A^* f_\delta, \qquad k = 0, 1, \ldots . \tag{5.30}$$

Depending on a particular context, the iteration method (5.30) can be interpreted as an iteration method for solving the variational problem on minimization of the discrepancy functional

$$J_0(v) = \|Av - f_\delta\|^2.$$

The iteration method (5.29) and its rate of convergence are characterized by the choice of the operator B and by the values of the iteration parameters τ_{k+1}, $k = 0, 1, \ldots$. This matter as applied to the solution of difference problems that arise in difference solution methods for elliptic boundary value problems was discussed in Chapter 3. Yet, the direct use of previous results concerning the rate of convergence of the iteration methods gives little in the consideration of iteration methods for solving the ill-posed problem (5.3).

Dwell first on the case of explicit iteration methods, a constant iteration parameter (simple-iteration method), and $B = E$. Here, we have

$$\frac{u_{k+1} - u_k}{\tau} + A u_k = f_\delta, \qquad k = 0, 1, \ldots . \tag{5.31}$$

The rate of convergence of the iteration method (5.31) (see Theorem 3.5) is defined by the constants γ_1 and γ_2 in the inequality

$$\gamma_1 E \le A \le \gamma_2 E. \tag{5.32}$$

In the case of $\gamma_1 > 0$, the iteration method (5.31) converges in H, H_A for all values of τ in the interval

$$0 < \tau < \frac{2}{\gamma_2}, \tag{5.33}$$

and for the number n of iterations required for an accuracy ε to be achieved the following estimate holds:

$$n \ge n_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho_0},$$

with

$$\rho_0 = \frac{1-\xi}{1+\xi}, \qquad \xi = \frac{\gamma_1}{\gamma_2}.$$

In the case of (5.3), with the adopted assumptions about A for the constants in (5.32) we have $\gamma_2 = \lambda_1$, $\gamma_1 = 0$ and, hence, $\xi = 0$. By virtue of this, we cannot render more concrete the rate of convergence of the iteration method.

A second specific feature in the application of iteration methods for the approximate solution of ill-posed problems is related with the stopping criterion. In the solution of elliptic difference equations the iterations are to be continued unless the initial difference has decreased by the factor of ε^{-1}. The parameter ε has to be given based on this or that reasoning. In the iteration solution of the ill-posed problem (5.3) with regard for the right-hand side inaccuracy (estimate (5.2)) we can choose a condition for terminating the iterations considering the level of the inaccuracy, i.e., the iterations are to be continued unless some $n(\delta)$ is reached.

5.4.2 Iterative solution of ill-posed problems

Let us formulate now conditions under which the iteration method (5.31) gives the approximate solution of problem (5.1), (5.2).

Theorem 5.4 *Let in the iteration method* (5.31)–(5.33) *the number of iterations* $n(\delta) \to \infty$ *and* $n(\delta)\delta \to 0$ *as* $\delta \to 0$. *Then,* $\|u_{n(\delta)} - u\| \to 0$ *provided that* $\delta \to 0$.

Proof. We denote the inaccuracy at the nth iteration as $z_n = u_n - u$. From (5.31), we readily obtain:

$$u_n = (E - \tau A)^n u_0 + \sum_{k=0}^{n-1} (E - \tau A)^k \tau f_\delta, \tag{5.34}$$

where u_0 is some initial approximation.

To obtain an exact solution, we can use a similar representation

$$u = (E - \tau A)^n u + \sum_{k=0}^{n-1} (E - \tau A)^k \tau f.$$

This representation corresponds to the iteration solution of equation (5.1) in which the initial approximation coincides with the exact solution of the problem.

With equality (5.34) taken into account, for the inaccuracy we obtain the expression

$$z_n = z_n^{(1)} + z_n^{(2)}, \tag{5.35}$$

where

$$z_n^{(1)} = (E - \tau A)^n z_0, \qquad z_n^{(2)} = \sum_{k=0}^{n-1} (E - \tau A)^k \tau (f_\delta - f),$$

with $z_0 = u_0 - u$ being the initial inaccuracy. The first term $z_n^{(1)}$ in (5.35) is a standard one in iteration methods, and the term $z_n^{(2)}$ in the right-hand side of (5.35) is related with the inaccuracy in the right-hand side of (5.1).

Under the above (see (5.33)) constraints on τ, we have

$$\|E - \tau A\| \le 1. \tag{5.36}$$

To prove this inequality, we pass from inequality (5.36) to the equivalent inequality

$$(E - \tau A^*)(E - \tau A) \le E.$$

With the properties of self-adjointness and positiveness of A, and with the estimate $A \le \gamma_2 E$ (see (5.32)), taken into account, we obtain

$$(E - \tau A^*)(E - \tau A) - E \;=\; \tau A^{1/2}(\tau A - 2E)A^{1/2} \;\le\; \tau A^{1/2}(\tau\gamma_2 - 2)A^{1/2} \le 0$$

provided that inequality (5.33) is fulfilled.

Taking inequality (5.36) into account, we have

$$\|z_n^{(2)}\| \;\le\; \sum_{k=0}^{n-1} \|E - \tau A\|^k \tau \|f_\delta - f\| \;\le\; n\tau\delta. \tag{5.37}$$

The estimate $z_n^{(1)}$ deserves a more detailed consideration.

To begin with, assume that $z_0 \in H$. Such a situation is met, for instance, with the initial approximation $u_0 = 0$ taken in the solution of problem (5.1), (5.2) in the class (5.7). Let us show that $s(n) = \|z_n^{(1)}\| \to 0$ as $n \to \infty$. We use the representation

$$s^2(n) = \sum_{i=1}^{\infty} (1 - \tau\lambda_i)^{2n} (z_0, w_i)^2.$$

For any small $\varepsilon > 0$, there can be found a sufficiently large N such that

$$\sum_{i=N+1}^{\infty} (z_0, w_i)^2 \le \frac{\varepsilon}{2}.$$

Taking the fact into account that $|1 - \tau\lambda_i| < 1$, we have:

$$s^2(n) \le \sum_{i=1}^{N} (1 - \tau\lambda_i)^{2n} (z_0, w_i)^2 + \sum_{i=N+1}^{\infty} (z_0, w_i)^2.$$

In the case of sufficiently large n, for the first term we have:

$$\sum_{i=1}^{N} (1 - \tau\lambda_i)^{2n} (z_0, w_i)^2 \le \frac{\varepsilon}{2}.$$

Substitution of (5.37) into (5.35) yields the estimate

$$\|z_n\| \le n\tau\delta + s(n), \tag{5.38}$$

where $s(n) \to 0$ as $n \to \infty$. From the obtained estimate (5.38), the desired statement readily follows. □

In the iteration method (5.31)–(5.33), it is the number of iterations, matched with the right-hand side inaccuracy, which serves the regularization parameter. The first term in the right-hand side of (5.38) grows in value with the number of iterations, whereas the second term decreases. For the inaccuracy to be minimized, the number of iteration must be chosen not too large, nor too small.

5.4.3 Estimate of the convergence rate

Like in the proof of Theorem 5.1, we have established the convergence without finding its rate. On narrowing the class of a priori constraints on the solution, we can derive (see Theorem 5.2) the approximate-solution inaccuracy as an explicit function of the inaccuracy in the right-hand side. A similar situation is also observed in using iterative solution methods for ill-posed problems. Let us formulate a typical result along this line.

Consider the iteration method (5.31), (5.32) under constraints on the iteration parameter tighter than (5.33),

$$0 < \tau \le \frac{1}{\gamma_2}. \tag{5.39}$$

Theorem 5.5 *Suppose that the exact solution of problem* (5.3) *belongs to the class*

$$\|A^{-p}u\| \le M, \qquad 0 < p < \infty. \tag{5.40}$$

Then, for the inaccuracy of the iteration method (5.31), (5.32), (5.39) *with* $u_0 = 0$ *there holds the estimate*

$$\|z_n\| \le n\tau\delta + M_1 n^{-p}, \qquad M_1 = M_1(\tau, p, M). \tag{5.41}$$

Proof. Under the conditions of (5.40), it is necessary to find how the quantity $s(n)$ in the estimate (5.38) depends on the number of iterations n.

With $u_0 = 0$, we have $z_0 = u$ and, using the above notation, obtain that

$$z_n^{(1)} = \sum_{i=1}^{\infty} \lambda_i^p (1 - \tau\lambda_i)^n \frac{(u, w_i)}{\lambda_i^p} w_i.$$

Then, in view of (5.40), we obtain:

$$s(n) \le \max_{\lambda_i} \lambda_i^p |1 - \tau\lambda_i|^n \|A^{-p}u\|. \tag{5.42}$$

Under the constraint (5.39) on the iteration parameter we have $0 < \tau\lambda_i \le 1$ and, hence,

$$\max_{\lambda_i} \lambda_i^p |1 - \tau\lambda_i|^n \le \frac{1}{\tau^p} \max_{0<\eta<1} \chi(\eta), \qquad \chi(\eta) = \eta^p (1 - \eta)^n.$$

The function $\chi(\eta)$ attains its maximum at the point

$$\eta = \eta^* = \frac{p}{p + n}$$

and, in addition,

$$\chi(\eta^*) = \Big(\frac{p}{n}\Big)^p \Big(1 - \frac{p}{p+n}\Big)^{p+n} < \frac{p^p}{n^p} \exp(-p).$$

Substitution into (5.42) results in the estimate (5.41), in which the constant

$$M_1 = \frac{p^p}{\tau^p} \exp(-p) M$$

depends on p, τ and M, but does not depend on n. □

Minimization of the right-hand side of (5.41) allows us to formulate the termination criterion for the iterations:

$$n_{\text{opt}} = \Big(\frac{pM_1}{\tau}\Big)^{1/(p+1)} \delta^{-1/(p+1)}, \tag{5.43}$$

i. e., $n(\delta) = \mathcal{O}(\delta^{-1/(p+1)})$. Here, for the approximate-solution inaccuracy we obtain the estimate

$$\|z_{n_{\text{opt}}}\| \le M_2 \delta^{p/(p+1)} \tag{5.44}$$

with the constant

$$M_2 = \tau \Big(\frac{pM_1}{\tau}\Big)^{1/(p+1)} + M_1 \Big(\frac{pM_1}{\tau}\Big)^{-p/(p+1)}.$$

Estimate (5.44) demonstrates the direct dependence of the rate of convergence of the approximate solution to the exact solution on the right-hand side inaccuracy δ and on the smoothness of the exact solution (parameter p).

5.4.4 Generalizations

Above, the possibility to optimally choose the number of iterations in the case of known class of a priori constraints on the exact solution was noted (the constant M in (5.40)). In the practical use of iteration methods for solving ill-posed problems, one often fails to explicitly specify the class of a priori constraints. That is why, instead of (5.43), they often choose the number of iterations from the discrepancy, so that the iterative process is to be continued unless the following inequality becomes fulfilled:

$$\|Au_{n(\delta)} - f_\delta\| \le \delta. \tag{5.45}$$

In the discussion of iteration methods for solving elliptic difference equations, variation-type iteration methods were isolated. In these methods, explicit calculation formulas for iteration parameters are used. For instance, in the steepest descend method the following formula is used:

$$\frac{u_{k+1} - u_k}{\tau_{k+1}} + Au_k = f_\delta, \qquad k = 0, 1, \ldots, \tag{5.46}$$

where

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \qquad r_k = Au_k - f_\delta. \tag{5.47}$$

In this connection, the conjugate gradient method deserves to be mentioned. This method belongs to the class of three-layer iteration methods, when in the solution of problem (5.3) we use the formulas

$$\begin{gathered} u_{k+1} = \alpha_{k+1}(E - \tau_{k+1}A)u_k + (1 - \alpha_{k+1})u_{k-1} + \alpha_{k+1}\tau_{k+1}f_\delta, \\ k = 1, 2, \ldots, \\ u_1 = (E - \tau_1 A)u_0 + \tau_1 f_\delta. \end{gathered} \tag{5.48}$$

For the iteration parameters α_{k+1} and τ_{k+1}, we use the calculation formulas

$$\begin{gathered} \tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \qquad k = 0, 1, \ldots, \\ \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k}\,\frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})}\,\frac{1}{\alpha_k}\right)^{-1}, \qquad k = 1, 2 \ldots, \quad \alpha_1 = 1. \end{gathered} \tag{5.49}$$

We will not dwell here on the regularizing properties of such iteration methods, inviting the interested reader to the special literature. Note only that for the variational iterative solution methods (5.46), (5.47) and (5.48), (5.49) for the ill-posed problem (5.1), (5.2) similar results were obtained as for the simple iteration method (5.31).

The problem of using implicit iteration methods for solving ill-posed problems deserves special mention. Here, we mean, for instance, iteration methods of type (5.29), (5.30), in which $B \ne E$. In the general context, there is a problem concerning the choice of the operator B in solving ill-posed problems.

In the solution of difference problems for well-posed mathematical physics problems, the choice of B should be made considering primarily the demand for improved rate of convergence in the iteration method. In the solution of ill-posed problems the iterative process should be terminated on the achievement of a discrepancy whose value is defined by the input-data inaccuracy. We are interested not only in the rate at which the iterative process converges in this domain of decreasing, but also in the class of smoothness on which this iterative process converges and in the norm in which the desired level of discrepancy can be achieved. The most important feature of the approximate solution of ill-posed problems by iteration methods consists in the fact that an appropriate approximate solution can be isolated from the necessary smoothness class through the choice of B.

5.5 Program implementation and computational experiments

As a routine ill-posed problem, the first-kind integral equation is normally considered. Following the common practice, we consider here the points that arise in the numerical solution of an integral equation used to continue anomalous gravitation fields given on the earth surface towards disturbing masses. The program implementation is based on the use of iteration methods in solving problems with random input-data inaccuracies.

5.5.1 Continuation of a potential

In gravimetrical and magnetic prospecting, and in direct-current geoelectrical prospecting, most important are problems in which it is required to continue a potential fields from the earth surface deep into the earth. The solution of such problems is used to identify, to this or that extent, the position of gravitational and electromagnetic anomalies. Here, we restrict ourselves to the formulation of the gravitational-field continuation problem.

We designate as U the gravitational potential of an anomaly located in depth of the earth. Let x be a horizontal coordinate and the axis Z be directed upward so that at the earth surface we have $z = 0$. We consider a problem in which it is required to determine the gravity potential in the zone $z < 0$ down to anomalies whose depth is H.

We consider a problem in which it is required to continue the gravitational potential from two disturbing masses (D_1 and D_2 in Figure 5.1). This potential $U(x, z)$ satisfies the Laplace equation in the zone outside the anomalies, so that

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial z^2} = 0, \qquad z > -H. \tag{5.50}$$

At the earth surface (according to field observations measured in which is the first

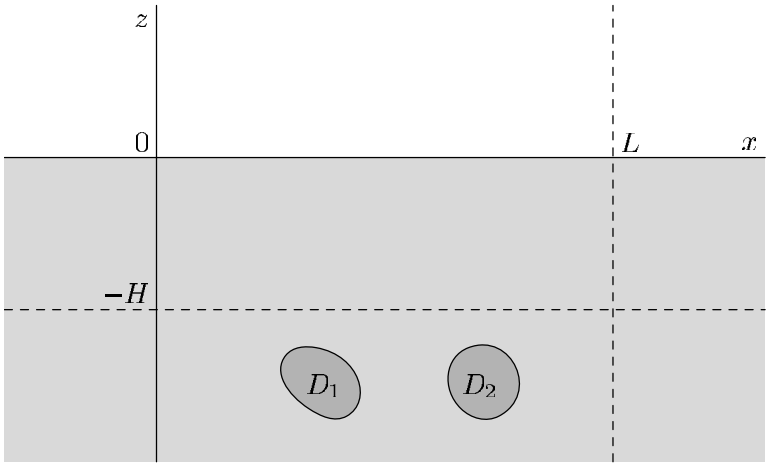

Figure 5.1 Schematic illustrating the continuation problem

vertical derivative of the potential) the following conditions are posed:

$$\frac{\partial U}{\partial z}(x, 0) = \varphi(x). \tag{5.51}$$

One more boundary condition is

$$U(x, \infty) = 0. \tag{5.52}$$

On the behavior of $\varphi(x)$ with $|x| \to \infty$, natural constraints are posed which provide for solution boundedness.

The problem (5.50)–(5.52) considered in the zone $z > 0$) is an ordinary boundary value problem. We pose a continuation problem for the solution of this well-posed problem into the adjacent zone $-H < z < 0$.

The continuation problem (5.50)–(5.52) is not quite convenient to examine because it involves boundary condition (5.52) with angled derivative. This problem can be reformulated as a continuation problem for $u = \partial U/\partial z$:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial z^2} = 0, \qquad z > -H, \tag{5.53}$$

$$u(x, 0) = \varphi(x), \tag{5.54}$$

$$u(x, \infty) = 0. \tag{5.55}$$

In the case of (5.53)–(5.55), we have a continuation problem for the solution of the Dirichlet problem for the Laplace equation.

Let us give some simple reasoning concerning the necessity to solve the continuation problem for gravitational fields toward anomalies. In the performed calculations,

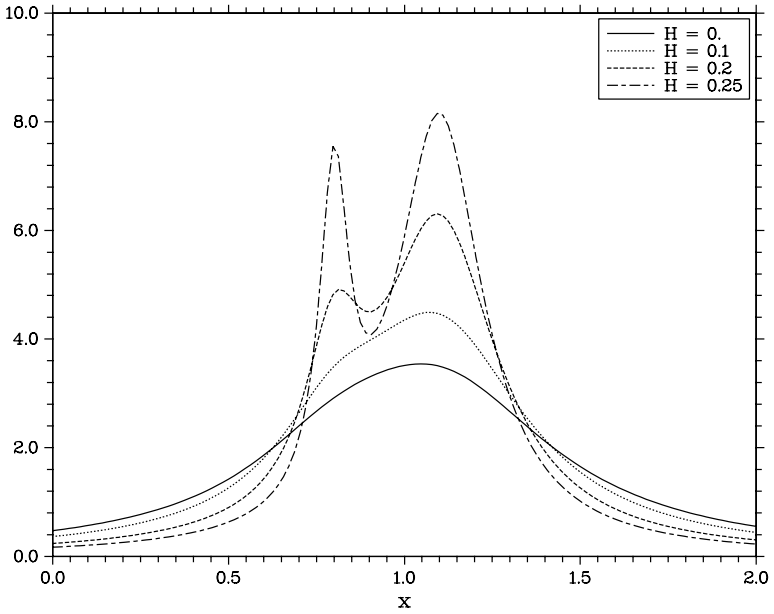

Figure 5.2 Solution of the direct problem at various depths

we consider a model problem with two anomalies, circular in cross-section, with centers at $(x^1, z^1) = (0.8, -0.3)$ and $(x^2, z^2) = (1.1, -0.4)$. The exact solution of problem (5.53)–(5.55) is

$$u(x, z) = c_1 \frac{z - z_1}{(x - x_1)^2 + (z - z_1)^2} + c_2 \frac{z - z_2}{(x - x_2)^2 + (z - z_2)^2}.$$

Suppose that $c_1 = 0.3$ and $c_2 = 1.2$, i. e., the deeper anomaly has fourfold greater power.

The exact solution of the problem at various depths H is shown in Figure 5.2. From the measured data, at $H = 0$ and $H = 0.1$ two sources of gravitational perturbations can be suspected. At larger depths (as we approach the sources), two individual anomalies can be identified (at $H = 0.2$ and especially at $H = 0.25$). It is this circumstance that necessitates the continuation of potential fields towards anomalies.

5.5.2 Integral equation

To find an approximate solution of the continuation problem (5.53)–(5.55), one can use many computational algorithms. We will dwell here on the use of the integral equation method. We approximate the gravitational potential produced by anomalies with a potential generated by an elemental bed with a carrier located at the segment $0 \le x \le L$ and at the depth $z = -H$. Accurate to a multiplier, we have

$$U(x, z) = \int_0^L \ln r \, \mu(s) \, ds, \qquad r = \sqrt{(x - s)^2 + (z + H)^2}, \quad z > -H.$$

For the derivative with respect to the vertical variable we obtain:

$$u(x, z) = \int_0^L \frac{z + H}{(x - s)^2 + (z + H)^2} \mu(s)\, ds, \qquad z > -H. \tag{5.56}$$

For the unknown density, from (5.54) and (5.56) we obtain the first-kind integral equation

$$\int_0^L K(x, s)\mu(s)\, ds = \varphi(x) \tag{5.57}$$

with the symmetric kernel

$$K(x, s) = \frac{H}{(x - s)^2 + H^2}.$$

We write equation (5.57) as the first-kind equation

$$A\mu = \varphi \tag{5.58}$$

in which

$$A\mu = \int_0^L K(x, s)\mu(s)\, ds.$$

The integral operator $A : H \to H$, where $H = L_2(0, L)$, is a self-adjoint operator $(A\mu, \nu) = (\mu, A\nu)$. Besides, this operator is a bounded operator because

$$(A\mu, \mu) = \int_0^L \int_0^L K(x, s)\mu(s)\mu(x)\, ds\, dx$$
$$\leq \Big(\int_0^L \int_0^L K^2(x, s)\, ds\, dx \Big)^{1/2} \Big(\int_0^L \int_0^L \mu^2(s)\mu^2(x)\, ds\, dx \Big)^{1/2} \leq \gamma_2 \|\mu\|^2,$$

where $\gamma_2 < L/H$.

The integral operator under consideration is a self-adjoint, bounded operator. Yet, the iteration method cannot be directly applied to (5.58) because, generally speaking, the operator A here is not a positive (non-negative) operator. Hence, symmetrization of (5.58) is to be applied:

$$A^* A\mu = A^* \varphi, \tag{5.59}$$

which we can finally use in the iteration method.

5.5.3 Computational realization

In the numerical solution of the integral equation (5.59), we use a uniform grid

$$\bar{\omega} = \{x \mid x = (i - 0.5)h,\ i = 1, 2, \ldots, N,\ Nh = L\}.$$

We designate the approximate solution as $y_i = y(x_i)$, $i = 1, 2, \ldots, N$.

We put into correspondence to the integral operator A the difference operator

$$(A_h y)(x_i) = \sum_{j=1}^{N} K(x_i, s_j) y_j h. \tag{5.60}$$

This approximation corresponds to the use of the rectangle quadrature formula. The difference analogue of (5.59) is the equation

$$Ry = f, \qquad R = A_h^* A_h, \qquad f = A_h^* \varphi. \tag{5.61}$$

To approximately solve equation (5.61), we use the iteration method

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ry_k = f, \qquad k = 0, 1, \ldots . \tag{5.62}$$

Two choices of iteration parameters were considered:

1. Simple iteration method, in which $\tau_k = \tau = \text{const}$;
2. Steepest descend method, in which

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Rr_k, r_k)}, \quad r_k = Ry_k - f.$$

To model the input-data inaccuracy, we perturbed the exact solution at the grid nodes by the law

$$\varphi_\delta(x) \;=\; \varphi(x) + 2\zeta(\sigma(x) - 1/2), \qquad x = x_i, \quad i = 1, 2, \ldots, N,$$

where $\sigma(x)$ is a random quantity normally distributed over the interval from 0 to 1. The parameter ζ defines the inaccuracy level in the right-hand side of (5.58), and

$$\|\varphi_\delta(x)) - \varphi(x)\| \le \delta \;=\; \zeta\sqrt{L}.$$

The iterative process is to be terminated considering the discrepancy. Some other specific features of the computational algorithm will be noted below.

5.5.4 Program

To approximately solve the model two-dimensional continuation problem for gravitational fields, we used the program presented below.

Program PROBLEM4

```
C
C     PROBLEM4  - CONTINUATION OF ANAMALOUS GRAVITATIONAL FIELD
C
      PARAMETER ( HH = 0.225, DELTA = 0.1, ISCHEME = 1, N = 100 )
```

```
      DIMENSION  Y(N), X(N), PHI(N), PHID(N), F(N),  A(N,N)
     +          ,V(N), R(N), AR(N),  U(N),    UY(N), UY1(N)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF SEGMENT;
C     HH        - CARRIER DEPTH;
C     H0        - CONTINUATION DEPTH;
C     ISCHEME   - SIMPLE-ITERATION METHOD (ISCHEME = 0),
C                 QUICKEST DESCEND METHOD (ISCHEME = 1);
C     N         - NUMBER OF GRID NODES;
C     Y(N)      - SOLUTION OF THE INTEGRAL EQUATION;
C     U(N)      - EXACT ANOMALOUS FIEKD AT THE DEPTH Z = H0;
C     UY(N)     - CALCULATED FIELD AT THE DEPTH Z = H0;
C     UY1(N)    - CALCULATED FIELD AT THE EARTH SURFACE;
C     PHI(N)    - EXACT ANOMALOUS FIELD AT THE EARTH SURFACE;
C     PHID(N)   - DISTURBED FIELD AT THE EARTH SURFACE;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     A(N,N)    - PROBLEM MATRIX;
C
      XL = 0.
      XR = 2.
C
      OPEN ( 01, FILE = 'RESULT.DAT' ) ! FILE TO STORE THE COMPUTED DATA
C
C     GRID
C
      H = (XR - XL)/N
      DO I = 1, N
        X(I) = XL + (I-0.5)*H
      END DO
C
C     DATA AT THE EARTH SURFACE
C
      C1 = 0.3
      X1 = 0.8
      Z1 = - 0.3
      C2 = 1.2
      X2 = 1.1
      Z2 = - 0.4
      DO I = 1,N
        PHI(I) = - C1/((X(I)-X1)**2 + Z1**2) * Z1
     *           - C2/((X(I)-X2)**2 + Z2**2) * Z2
        PHID(I) = PHI(I) + 2.*DELTA*(RAND(0)-0.5)/SQRT(XR-XL)
      END DO
C
C     EXACT IN-EARTH DATA
C
      H0 = 0.2
      DO I = 1,N
        U(I) = - C1/((X(I)-X1)**2 + (H0+Z1)**2) * (H0+Z1)
     *         - C2/((X(I)-X2)**2 + (H0+Z2)**2) * (H0+Z2)
      END DO
C
C     RIGHT-HAND SIDE
C
      DO I = 1,N
        SUM = 0.
        DO J = 1,N
          SUM = SUM + AK(X(I),X(J),HH) * PHID(J)
```

```
        END DO
        F(I) = SUM * H
      END DO
C
C     MATRIX ELEMENTS
C
      DO I = 1,N
        DO J = 1,N
           SUM = 0.
           DO K = 1,N
             SUM = SUM + AK(X(I),X(K),HH) * AK(X(K),X(J),HH)
           END DO
           A(I,J) = SUM * H**2
        END DO
      END DO
C
C     ITERATIVE PROCESS
C
      IT = 0
      ITMAX = 100
      DO I = 1,N
        Y(I) = 0.
      END DO
  100 IT = IT+1

C
C     CACULATION V = R Y
C
      DO I = 1,N
        SUM = 0.
        DO J = 1,N
          SUM = SUM + A(I,J) * Y(J)

        END DO
        V(I) = SUM
      END DO
C
C     ITERATION PARAMETER
C
      DO I = 1,N
        R(I) = V(I) - F(I)
      END DO
      IF ( ISCHEME.EQ.0 ) THEN
C
C     SIMPLE-ITERATION METHOD
C
      TAU = 0.1
      END IF
      IF ( ISCHEME.EQ.1 ) THEN
C
C     QUICKEST DESCEND METHOD
C
      DO I = 1,N
        R(I) = V(I) - F(I)
      END DO
      DO I = 1,N
        SUM = 0.
        DO J = 1,N
```

```
          SUM = SUM + A(I,J) * R(J)
        END DO
        AR(I) = SUM
      END DO
      SUM1 = 0.
      SUM2 = 0.
      DO I = 2,N-1
        SUM1 = SUM1 + R(I)*R(I)
        SUM2 = SUM2 + AR(I)*R(I)
      END DO
      TAU = SUM1/SUM2
      END IF
C
C     NEXT APPROXIMATION
C
      DO I = 1,N
        Y(I) = Y(I) - TAU*R(I)
      END DO
C
C     ITERATION LOOP EXIT
C
      SUM0 = 0.
      DO I = 1,N
        SUM = 0.
        DO J = 1,N
          SUM = SUM + AK(X(I),X(J),HH) * Y(J)
        END DO
        SUM0 = SUM0 + (SUM * H - PHID(I))**2 * H
      END DO

      SL2 = SQRT(SUM0)
      WRITE ( 01,* ) IT, TAU, SL2
      IF (SL2.GT.DELTA .AND. IT.LT.ITMAX) GO TO 100
      CLOSE ( 01 )
C
C     CALCULATION OF THE FOUND FIELD AT THE DEPTH Z = H0
C
      DO I = 1,N
        SUM = 0.
        DO J = 1,N
          SUM = SUM + AK(X(I),X(J),HH-H0) * Y(J)
        END DO
        UY(I) = SUM * H
      END DO

C
C     CALCULATION OF THE FOUND FIELD AT THE EARTH SURFACE
C
      DO I = 1,N
        SUM = 0.
        DO J = 1,N

          SUM = SUM + AK(X(I),X(J),HH) * Y(J)
        END DO
        UY1(I) = SUM * H
      END DO
      STOP
      END
      FUNCTION AK(X,S,HH)
C
```

```
C     INTEGRAL-EQUATION KERNEL
C
      AK = HH/((X-S)**2 + HH**2)
      RETURN
      END
```

5.5.5 Computational experiments

Consider several examples for the solution of the continuation problem in the case of $H = 0.225$. We use a computation grid whose total number of nodes is $N = 100$. The most interesting dependence here is the accuracy in reconstructing the anomalous gravitational field versus the input-data inaccuracy δ. Figure 5.3 shows data obtained by solving the problem with the input-data inaccuracy $\delta = 0.1$ (in setting the field at the earth surface $z = 0$). Presented are the exact and found fields at the depth $z = -2$.

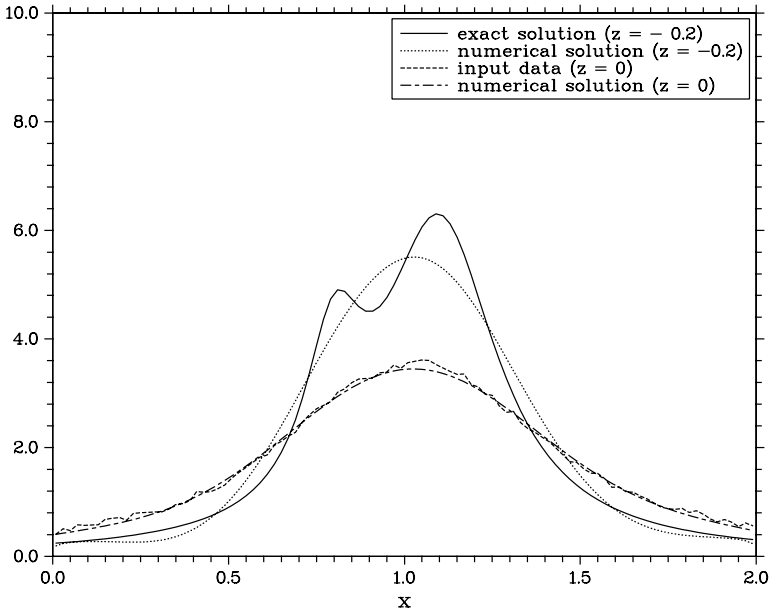

Figure 5.3 Solution of the continuation problem obtained with $\delta = 0.1$

Similar data obtained at a larger inaccuracy level ($\delta = 0.2$) are shown in Figure 5.4. At the latter inaccuracy level, two individual anomalies are hard to identify (here, no two-peak configuration is observed). Even a decrease of δ down to $\delta = 0.01$ little saves the situation (see Figure 5.5).

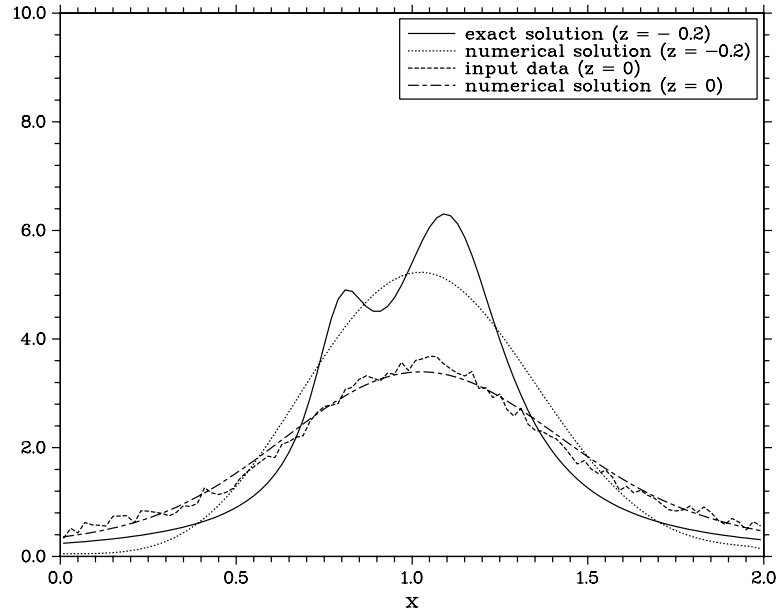

Figure 5.4 Solution of the continuation problem obtained with $\delta = 0.2$

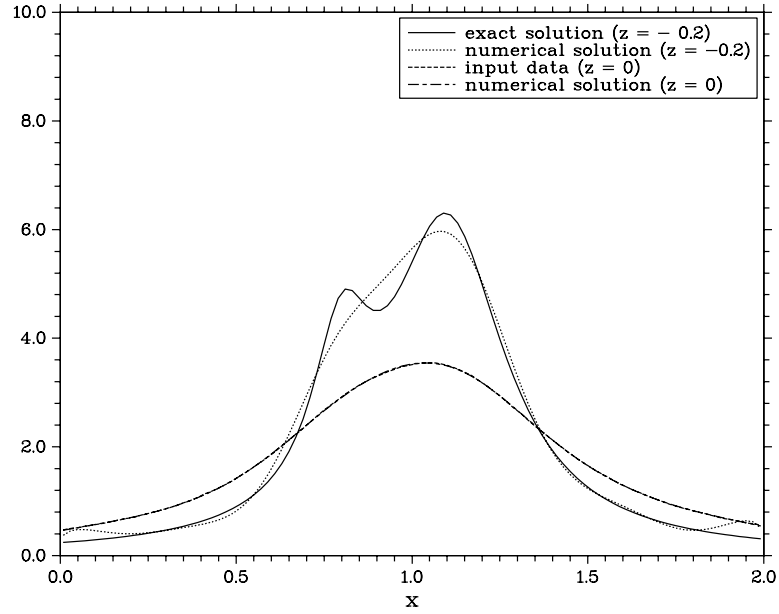

Figure 5.5 Solution of the continuation problem obtained with $\delta = 0.01$

We now provide a few statements concerning the use of these or those iteration methods. In the program PROBLEM4, the possibility to use either the simple-iteration method (constant iteration parameter) or the steepest descend method is provided. A slight modification allows one to implement the iterative conjugate gradient method, to be used in the case of problems with small right-hand side inaccuracies, in which the steepest descend method fails to provide a desired rate of convergence.

The number of iterations necessary for solving the problem with $\delta = 0.1$ by the steepest descend method is $n = 6$. The total number of iterations in the simple-iteration method versus the iteration parameter is illustrated by the data in Table 5.1.

τ	0.05	0.1	0.15	0.2	0.25	0.3
n	43	21	14	11	9	32

Table 5.1 Total number of iterations in the simple-iteration method

In the iterative solution of (5.59), the optimal iteration parameter is $\tau = \tau_0 = 2/\bar{\gamma}_2$, where

$$A^*A \le \bar{\gamma}_2 E,$$

can be estimated invoking the estimate $\bar{\gamma}_2 < L^2/H^2$.

5.6 Exercises

Exercise 5.1 Show that in the Tikhonov regularization method for the solution there holds the a priori estimate

$$\|u_\alpha\| \le \frac{1}{2\sqrt{\alpha}} \|f_\delta\|$$

that expresses stability with respect to the right-hand side.

Exercise 5.2 Formulate conditions for convergence (analogue to Theorem 5.1, 5.2) of the approximate solution in H_D, $D = D^* > 0$ found by minimization of the functional

$$J_\alpha(v) = \|Av - f_\delta\|^2 + \alpha\|v\|_D^2.$$

Exercise 5.3 Construct an example illustrating the necessity of condition $\delta^2/\alpha(\delta) \to 0$ with $\delta \to 0$ (Theorem 5.1) for convergence of the approximate solutions to the exact solution in the Tikhonov method.

Exercise 5.4 Suppose that the exact solution of problem (5.1) is

$$\|A^{-2}u\| \le M,$$

where $M = \text{const} > 0$, and for the right-hand side inaccuracy the estimate (5.2) holds. Then, for the approximate solution u_α found as the solution of problem (5.4), (5.5) there holds the a priori estimate

$$\|z\| \le \frac{\delta}{2\sqrt{\alpha}} + \alpha M.$$

Exercise 5.5 Obtain an estimate for the rate of convergence in the Tikhonov method under the a priori constraints

$$\|A^{-p}u\| \le M, \quad 0 < p < 2$$

posed on the exact solution of problem (5.1), (5.2).

Exercise 5.6 Let u_α be the solution of problem (5.4), (5.5) and

$$\begin{aligned} m(\alpha) &= \|Au_\alpha - f_\delta\|^2 + \alpha\|u_\alpha\|^2, \\ \varphi(\alpha) &= \|Au_\alpha - f_\delta\|^2, \quad \psi(\alpha) = \|u_\alpha\|^2. \end{aligned}$$

Show that in the case of $f_\delta \neq 0$ and $0 < \alpha_1 < \alpha_2$ there hold the inequalities

$$m(\alpha_1) < m(\alpha_2), \qquad \varphi(\alpha_1) < \varphi(\alpha_2), \qquad \psi(\alpha_1) > \psi(\alpha_2).$$

Exercise 5.7 Show that for any $\delta \in (0, \|f_\delta\|)$ there exists a unique solution $\alpha = \alpha(\delta)$ of the equation

$$\varphi(\alpha) = \|Au_\alpha - f_\delta\|^2 = \delta^2,$$

such that $\alpha(\delta) \to 0$ and $\|u_{\alpha(\delta)} - u\| \to 0$ as $\delta \to 0$ (substantiation of the choice of regularization parameter based on the discrepancy).

Exercise 5.8 In solving problem (5.1), (5.2) with $A = A^* > 0$, the algorithm of simplified regularization uses the equation

$$Au_\alpha + \alpha u_\alpha = f_\delta \tag{5.63}$$

for finding the approximate solution. Formulate the related variational problem.

Exercise 5.9 Prove that in the case of

$$\|u\|_{A^{-1}} \le M$$

for the inaccuracy in the approximate solution of problem (5.1), (5.2) found from (5.63) there holds the priori estimate

$$\|z\|^2 \le \frac{\delta^2}{\alpha} + \frac{\alpha}{2} M^2.$$

Exercise 5.10 Suppose that for the approximate solution of problem (5.1), (5.2) one uses the iteration method

$$B\,\frac{u_{k+1} - u_k}{\tau} + Au_k = f_\delta, \qquad k = 0, 1, \ldots$$

with $B = B^* > 0$ and

$$\gamma_1 B \le A \le \gamma_2 B, \qquad \gamma_1 > 0, \qquad 0 < \tau < \frac{2}{\gamma_2}.$$

Show that in the iteration method (5.31)–(5.33) $\|u_{n(\delta)} - u\|_B \to 0$ as $\delta \to 0$ provided that $n(\delta) \to \infty$ and $n(\delta)\delta \to 0$ as $\delta \to 0$.

Exercise 5.11 Suppose that, in the solution of problem (5.1), (5.2), the total number of iterations in the method (5.31)–(5.33) is chosen from the condition

$$\|Au_{n(\delta)} - f_\delta\| \le \delta < \|Au_{n(\delta)-1} - f_\delta\|$$

and, in addition, $\|Au_0 - f_\delta\| > \delta$. Show that, in this case, $\|u_{n(\delta)} - u\| \to 0$ as $\delta \to 0$.

Exercise 5.12 Prove the ill-posedness of the problem in which it is required to solve the first-kind Fredholm integral equation

$$\int_a^b K(x, s)\mu(s)\, ds = \varphi(x), \qquad c \le x \le d$$

with a kernel $K(x, s)$ continuously differentiable with respect to both variables in the case of $\mu(x) \in C[a, b]$, $\varphi(x) \in C[c, d]$.

Exercise 5.13 Consider the continuation problem for a gravitational field disturbed with local anomalies with non-negative anomalous (excessive with respect to surrounding rocks) density. We seek the solution of the continuation problem (5.53), (5.55) in the form (see (5.57))

$$\int_{-\infty}^{\infty} K(x, s)\mu(s)\, ds = \varphi(x),$$

assuming, in addition, that the anomalies are localized at depths greater than H. Show that $\mu(x) \ge 0$, $-\infty < x < \infty$.

Exercise 5.14 In the program `PROBLEM4`, implement the conjugate gradient method. Examine the efficiency of the method as applied to problems with small inaccuracies in the gravitational field at the earth surface ($\delta \le 0.01$).

Exercise 5.15 Extend the program `PROBLEM4` to the case in which the approximate solution of the continuation problem is solved in the form of a double-bed potential. Based on the solution of the model problem, perform an analysis of possibilities offered by this approach.

6 Right-hand side identification

An important class of inverse mathematical physics problem is the determination of unknown right-hand sides of equations. In such problems, additional information about the solution is provided either throughout the calculation domain or over some part of the domain, in particular, on the domain boundary. Another class consists in calculating a differential operator for an approximately given right-hand side and solution which is approximately known throughout the whole domain. Also worth noting are specific features of right-hand side determination algorithms for non-stationary problems. Here the values of the right-hand side are successively determined at fixed times. In the non-stationary case, additional information about the solution is required on parts of the calculation domain. At the end of this chapter, computational algorithms for solving such inverse boundary value problems involving stationary or non-stationary mathematical physics equations are considered. Computational data are provided that are obtained in several numerical experiments on the approximate solution of inverse problems.

6.1 Reconstruction of the right-hand side from known solution in the case of stationary problems

In this section, we consider a problem in which it is required to reconstruct the right-hand side of a second-order ordinary differential equation whose solution is known accurate to some given inaccuracy. A simplest computational algorithm uses standard difference approximations, the mesh size being the regularization parameter. The possibility of using alternative approaches is noted.

6.1.1 Problem statement

As a simplest mathematical model, we use the second-order ordinary differential equation

$$-\frac{d}{dx}\left(k(x)\,\frac{du}{dx}\right) + q(x)u = f(x), \qquad 0 < x < l \tag{6.1}$$

under the following usual constraints imposed on the coefficients:

$$k(x) \geq \kappa > 0, \qquad q(x) \geq 0.$$

In the direct problem, it is required to determine the function $u(x)$ from equation (6.1) with given coefficients $k(x)$, $q(x)$, right-hand side $f(x)$, and additional conditions given on the boundary. In a simplest case, first-kind homogeneous conditions are given:

$$u(0) = 0, \qquad u(l) = 0. \tag{6.2}$$

Formulate now the inverse problem. We assume that the right-hand side $f(x)$ is unknown. To find it, we have to set some additional information about the solution. Since, here, a function of x is being sought, it is therefore desirable that the additional information be also given as a function of x. This allows us to assume that known is the solution $u(x)$, $x \in [0, l]$.

With known exact $u(x)$, $x \in [0, l]$, the right-hand side can be uniquely found, and it follows from (6.1) that

$$f(x) = -\frac{d}{dx}\Big(k(x)\,\frac{du}{dx}\Big) + q(x)u, \qquad 0 < x < l. \tag{6.3}$$

The use of formula (6.3) implies, for instance, that $f(x) \in C(0, l)$ if $u(x) \in C^2[0, l]$ in the case of $k(x) \in C^1[0, l]$ and $q(x) \in C[0, l]$.

The point here is that the input data (in the case of interest, the solution $u(x)$, $x \in [0, l]$) are given approximately and do not belong to the mentioned smoothness class. A typical situation is such that, for instance, instead of the exact solution $u(x)$, $x \in [0, l]$ we know the function $u_\delta(x)$, $x \in [0, l]$ and, in addition, $u_\delta(x) \in C[0, l]$ and

$$\|u_\delta(x) - u(x)\|_{C[0,l]} \le \delta, \tag{6.4}$$

where

$$\|v(x)\|_{C[0,l]} = \max_{x \in [0,l]} |v(x)|.$$

Problem (6.3), (6.4) is a problem in which it is required to calculate the values of a differential operator. This problem belongs to the class of classically ill-posed problems and, to be solved, needs some regularizing algorithm to be applied. Consider some basic possibilities along this line.

6.1.2 Difference algorithms

Most naturally, difference methods can be used in calculating the right-hand side $f(x)$, $0 < x < l$. Over the interval $\bar{\Omega} = [0, l]$, we introduce a uniform grid with a grid size h:

$$\bar{\omega} = \{x \mid x = x_i = ih, \quad i = 0, 1, \ldots, N, \quad Nh = l\}.$$

Here, ω is the set of internal nodes and $\partial\omega$ is the set of boundary nodes.

We denote the approximate solution of problem (6.3), (6.4) at internal nodes as $f_h(x)$. Using the standard notation adopted in the theory of difference schemes, we set

$$f_h(x) = -(ay_{\bar{x}})_x + cy, \qquad x \in \omega, \tag{6.5}$$

where for the mesh function $y(x)$ we have:

$$y(x) = u_\delta(x), \qquad x \in \bar{\omega}.$$

For the difference coefficients, we set:

$$a(x) = k(x - 0.5h), \quad x, x + h \in \omega, \qquad c(x) = q(x), \quad x \in \omega.$$

The difference algorithm (6.5) is a regularizing one provided that a rule is given by which the discretization grid size $h = h(\delta)$ can be chosen as a function of the inaccuracy δ such that the norm of the inaccuracy $z = f_h - f$ vanishes as $\delta \to 0$. We would also like to know the rate of convergence of the approximate solution to the exact solution.

To optimize the discretization grid size and minimize the inaccuracy, we identify two classes of exact solutions. We assume that

$$\|u(x)\|_{C^4[0,l]} \le M, \tag{6.6}$$

or, alternatively,

$$\|u(x)\|_{C^3[0,l]} \le M, \tag{6.7}$$

where

$$\|v(x)\|_{C^n[0,l]} = \max_{x\in[0,l]} \max_{0\le k\le n} \left|\frac{d^k v}{dx^k}(x)\right|.$$

In the case of (6.6) or (6.7), we impose on the coefficients the constraints

$$k(x) \in C^3[0, l], \qquad q(x) \in C^2[0, l]$$

or

$$k(x) \in C^2[0, l], \qquad q(x) \in C^1[0, l].$$

Consider the problem for the inaccuracy $z(x) = f_h(x) - f(x)$, $x \in \omega$. From (6.3), (6.5), we easily obtain:

$$z(x) = z^{(1)}(x) + z^{(2)}(x), \qquad x \in \omega, \tag{6.8}$$

where

$$z^{(1)}(x) = -(au_{\bar{x}})_x + cu - f(x), \qquad x \in \omega, \tag{6.9}$$

$$z^{(2)}(x) = f_h(x) + (au_{\bar{x}})_x - cu, \qquad x \in \omega. \tag{6.10}$$

Let us estimate now the individual terms in (6.8) as dependent on the a priori assumptions about the exact solution of problem (6.3). The first term ($z^{(1)}(x)$) is the inaccuracy in approximating the differential operator with the difference operator. The second term ($z^{(2)}(x)$) reflects the effect induced by the input-data inaccuracy.

For the approximation inaccuracy, we readily obtain:

$$\begin{aligned} z^{(1)}(x) &= -(au_{\bar{x}})_x + cu + \frac{d}{dx}\left(k(x)\,\frac{du}{dx}\right) - q(x)u \\ &= -\frac{a_{i+1} - a_i}{h}\,\frac{du}{dx}(x_i) - \frac{a_{i+1} + a_i}{2}\,\frac{d^2u}{dx^2}(x_i) - \frac{a_{i+1} - a_i}{6}\,h\,\frac{d^3u}{dx^3}(x_i) \\ &\quad + cu + \mathcal{O}(h^\beta) + \frac{dk}{dx}\,\frac{du}{dx} + k(x)\,\frac{d^2u}{dx^2} - q(x)u. \end{aligned}$$

Here $\beta = 1$ under the assumption of (6.7) and $\beta = 2$, provided that the inequality (6.6) holds.

From similar representations, for the inaccuracy we obtain the estimate

$$\|z^{(1)}(x)\| \le M_1 h^\beta, \tag{6.11}$$

with the used settings

$$\|z(x)\| = \|z(x)\|_{C(\omega)} = \max_{x\in\omega} |z(x)|.$$

The constant M_1 in (6.11) depends not only on the solution but also on the coefficient $k(x)$ and on the derivatives of $k(x)$, i. e., $M_1 = M_1(u, k)$.

For the second inaccuracy component, from (6.5), (6.10) we obtain

$$\begin{aligned} \|z^{(2)}(x)\| &= \|-(a(y-u)_{\bar{x}})_x + c(y-u)\| \\ &\le \left(\max\left\{\|a(x)\|, \|a(x+h)\|\right\}\frac{4}{h^2} + \|c(x)\|\right)\max_{x\in\bar{\omega}}\|y(x)-u(x)\|. \end{aligned}$$

With (6.4), we obtain

$$\|z^{(2)}(x)\| \le M_2\delta. \tag{6.12}$$

The constant

$$M_2 = \max\left\{\|a(x)\|, \|a(x+h)\|\right\}\frac{4}{h^2} + \|c(x)\|$$

depends on the discretization grid size, i. e., $M_2 = M_2(h)$, and increases with decreasing h.

With (6.11), (6.12), from (6.8)–(6.10) we obtain the desired estimate for the inaccuracy:

$$\|z(x)\| \le M_1 h^\beta + M_2(h)\delta. \tag{6.13}$$

It follows from here that for the convergence with $\delta \to 0$ it is required that $\delta h^{-2} \to 0$.

In the classes of a priori assumptions (6.6) and (6.7) about the solution, from the estimate (6.13) we obtain the optimal discretization grid size and the rate of convergence

$$h_{\text{opt}} = \mathcal{O}(\delta^{1/4}), \qquad \|z(x)\| = \mathcal{O}(\delta^{1/2})$$

in the case of (6.6) ($\beta = 2$) and

$$h_{\text{opt}} = \mathcal{O}(\delta^{1/3}), \qquad \|z(x)\| = \mathcal{O}(\delta^{1/3})$$

in the case of (6.7) ($\beta = 1$).

Thus, we see that the discretization grid size h in the difference calculation of the right-hand side can be used as regularization parameter. Its value must be matched with the input-data inaccuracy δ. The optimal value of h depends on the constants M_1 and M_2, i. e., $h_{\text{opt}} = h_{\text{opt}}(M_1, M_2)$. Most frequently, no direct calculation of these constants is possible; in such cases, the discretization grid size can be chosen considering the discrepancy criterion.

In numerical solution of unstable problems, discretization (passage to a finite-difference problem) always brings about a regularization effect. This note applies both to difference and projection methods. Here, as the discretization parameter, the dimension of the problem (discretization step) can be used. We can say that the numerical methods possess the self-regularization property. Yet, if we try to raise the dimension of the problem in order to refine the solution, starting from a certain moment we obtain progressively worsening results: here, ill-posedness of the problem is manifested. That is why we have to terminate the calculations in due time and restrict ourselves to an approximate solution obtained with some optimal discretization.

Discretization offers rather restricted opportunities in regularization of problems. In computational practice, they often take another strategy. Methods are used in which the continuous problem possesses regularizing properties. In the course of discretization, the closeness of the discrete problem to the regularized differential problem due to the use, for instance, of a sufficiently fine grid cannot be explicitly traced. In this case, additional regularizing properties resulting from discretization are ignored as exerting only an insignificant influence. More accurate account of discretization inaccuracies is also possible in the cases in which the effect due to the inaccuracies cannot be ignored. In particular, the value of regularization parameter is matched not only with the right-hand side inaccuracy, but also with the inaccuracy of the operator in the first-kind equation (the so-called *general discrepancy principle*).

6.1.3 Tikhonov regularization

Consider problem (6.2)–(6.4) on the operator level. On the set of functions vanishing at the end points of the segment $[0, l]$ (see (6.2)) we define the operator $\mathcal{D}$ as

$$\mathcal{D}u = -\frac{d}{dx}\Big(k(x)\,\frac{du}{dx}\Big) + q(x)u.$$

Then, with exact input data the problem (6.2), (6.3) can be written as

$$f = \mathcal{D}u. \tag{6.14}$$

In $\mathcal{H} = \mathcal{L}_2(0, l)$, we have:

$$\mathcal{D} = \mathcal{D}^* \geq mE, \qquad m = \kappa\,\frac{\pi^2}{l^2}.$$

Hence, the operator $\mathcal{D}^{-1}$ exists and

$$0 < \mathcal{D}^{-1} \leq \frac{1}{m}\,E.$$

The latter allows us to pass from equation (6.14) to the equation

$$\mathcal{A}f = u, \tag{6.15}$$

where $\mathcal{A} = \mathcal{A}^* = \mathcal{D}^{-1}$.

Instead of the exact right-hand side of (6.15), the function u_δ is known to some accuracy given by

$$\|u_\delta - u\| \le \delta. \tag{6.16}$$

The problem (6.15), (6.16) was considered previously. This problem can be stably solved by the Tikhonov regularization method. In this case, the approximate solution f_α, $\alpha = \alpha(\delta)$ is to be found from

$$J_\alpha(f_\alpha) = \min_{v \in \mathcal{H}} J_\alpha(v), \tag{6.17}$$

where

$$J_\alpha(v) = \|\mathcal{A}v - u_\delta\|^2 + \alpha\|v\|^2. \tag{6.18}$$

Using the adopted notation, we can write the functional (6.18) as

$$J_\alpha(v) = \|\mathcal{D}^{-1}v - u_\delta\|^2 + \alpha\|v\|^2.$$

This functional is inconvenient for use because the values of $\mathcal{D}$ can be calculated easier than the inversion of the operator. Instead of (6.18), we can use the more general functional

$$J_\alpha(v) = \|\mathcal{D}^{-1}v - u_\delta\|^2_{\mathcal{G}} + \alpha\|v\|^2_{\mathcal{G}}$$

with $\mathcal{G} = \mathcal{G}^* > 0$. We set $\mathcal{G} = \mathcal{D}^*\mathcal{D}$; this yields

$$J_\alpha(v) = \|v - \mathcal{D}u_\delta\|^2 + \alpha\|\mathcal{D}v\|^2. \tag{6.19}$$

In this way, we arrive at a version of the Tikhonov regularization method for the approximate solution of problem (6.14), (6.16) based on the solution of the variational problem (6.17), (6.19). The situation around this algorithm was clarified previously by considering the intermediate problem (6.15), (6.16).

Method (6.17), (6.19) as applied to the approximate solution of problem (6.14), (6.16) can be substantiated using the same scheme as in the approximate solution of problem (6.15) (6.16) by method (6.17), (6.18). This makes further consideration of this point unnecessary.

The Euler equation for (6.17), (6.19) has the form

$$(E + \alpha\mathcal{D}^*\mathcal{D})f_\alpha = \mathcal{D}u_\delta. \tag{6.20}$$

In the case of (6.2)–(6.4), equation (6.20) yields a boundary value problem for the fourth-order ordinary differential equation. For the solution of (6.20) the following standard a priori estimate is valid:

$$\|f_\alpha\| \le \frac{1}{2\sqrt{\alpha}}\|u_\delta\|. \tag{6.21}$$

This estimate shows that the approximate solution is stable with respect to weak right-hand side perturbations.

6.1.4 Other algorithms

We present some other possibilities for the approximate solution of problem (6.14), (6.16). First of all, considering the case in which one obtains the approximate solution from equation (6.20), we can give another interpretation to the Tikhonov regularization method.

At the first stage, we define an auxiliary function

$$\tilde{f} = \mathcal{D}u_\delta. \tag{6.22}$$

Thus, $\tilde{f}$ is the solution of the problem with noisy right-hand side. At the second stage, the found solution is to be processed by applying a smoothing treatment. This can most naturally be done by minimizing the functional

$$J_\alpha(v) = \|v - \tilde{f}\|^2 + \|v\|^2_{\mathcal{G}}. \tag{6.23}$$

The minimum condition for the functional gives

$$(E + \alpha\mathcal{G}) f_\alpha = \tilde{f}. \tag{6.24}$$

In this interpretation, method (6.17), (6.19) refers to the case with $\mathcal{G} = \mathcal{D}^*\mathcal{D}$ chosen in (6.17), (6.22), (6.23).

With

$$\mathcal{G} \geq \mathcal{D}^*\mathcal{D},$$

for the solution of (6.22), (6.24) there holds the estimate (6.21). Note one such possibility that is of interest from the computational point of view. In the case of $\mathcal{G} = \mathcal{D}^2$ we have $\mathcal{G} = \mathcal{D}^*\mathcal{D} > 0$, and the method requires inversion of the operator $E + \alpha\mathcal{G}$. In the model problem under consideration it is required to solve a boundary value problem for a fourth-order equation. With

$$\mathcal{G} = \mathcal{D}^2 + 2\,\frac{1}{\sqrt{\alpha}}\,\mathcal{D},$$

equation (6.24) takes the form

$$(E + \sqrt{\alpha}\,\mathcal{D})^2 f_\alpha = \tilde{f}. \tag{6.25}$$

Hence, we can restrict ourselves to double inversion of the operator $E + \sqrt{\alpha}\,\mathcal{D}$ (i. e., to solving two boundary value problems for a second-order equation). Note that algorithm (6.22), (6.25) can be considered as double smoothing performed with the operator $\mathcal{G} = \mathcal{D}$.

Iteration methods deserve a special detailed consideration. Here, however, just one brief remark will be made. The approximate solution $f_{n(\delta)}$ of problem (6.15), (6.16) can be found using the simple-iteration method

$$\frac{f_{k+1} - f_k}{\tau} + \mathcal{A} f_k = u_\delta, \qquad k = 0, 1, \ldots,$$

which was considered at length previously. Turning back to problem (6.14), (6.16), rewrite the iterative process as

$$\mathcal{D}\,\frac{f_{k+1}-f_k}{\tau}+f_k=\mathcal{D}u_\delta,\qquad k=0,1,\dots\ . \tag{6.26}$$

Thus, for the solution of the problem at the next iteration step to be obtained, one has to invert the operator $\mathcal{D}$. The regularization effect can be achieved with a properly chosen reconditioner $B=\mathcal{D}$.

6.1.5 Computational and program realization

We construct a computational algorithm for problem (6.14), (6.16) around the Tikhonov regularization method (6.17), (6.19). First, pass to a discrete problem, assuming the discretization step to be sufficiently small.

On the set of mesh functions defined on the grid $\bar{\omega}$ and vanishing on $\partial\omega$, we define the difference operator

$$Dy=-(ay_{\bar{x}})_x+cy,\qquad x\in\omega.$$

On discretization, we pass from problem (6.14) to the problem

$$\varphi=Dy. \tag{6.27}$$

Instead of the mesh function $y(x)$, $x\in\omega$, the function $y_\delta(x)$, $x\in\omega$ is now given. The inaccuracy level is defined by the parameter δ:

$$\|y_\delta-y\|\le\delta, \tag{6.28}$$

where $\|\cdot\|$ is the norm in $H=L_2(\omega)$:

$$\|y\|=(y,y)^{1/2},\qquad (y,v)=\sum_{x\in\omega}y(x)v(x)h.$$

In H, we have:

$$D=D^*\ge\kappa\lambda_0E,\qquad \lambda_0=\frac{4}{h^2}\sin^2\frac{\pi h}{2l}\ge\frac{8}{l^2}.$$

In the Tikhonov method as applied to problem (6.27), (6.28) there arises (see (6.20)) the equation

$$(E+\alpha D^2)\varphi_\alpha=Dy_\delta \tag{6.29}$$

for the approximate solution $\varphi_\alpha(x)$, $x\in\bar{\omega}$. The value of $\alpha=\alpha(\delta)$ can be found using the discrepancy principle:

$$\|y_\delta-D^{-1}\varphi_\alpha\|=\delta. \tag{6.30}$$

Note some specific features in the computational realization of the method.

Equation (6.29) is a difference equation with five-diagonal matrix. This equation can be solved by the *sweep method* (*Thomas algorithm*). Below, we give the sweep-method calculation formulas for the following five-diagonal system:

$$C_0 y_0 - D_0 y_1 + E_0 y_2 = F_0, \tag{6.31}$$

$$-B_1 y_0 + C_1 y_1 - D_1 y_2 + E_1 y_3 = F_3, \tag{6.32}$$

$$A_i y_{i-2} - B_i y_{i-1} + C_i y_i - D_i y_{i+1} + E_i y_{i+2} = F_i,\ i = 2, \ldots, N-2, \tag{6.33}$$

$$A_{N-1} y_{N-3} - B_{N-1} y_{N-2} + C_{N-1} y_{N-1} - D_{N-1} y_N = F_{N-1}, \tag{6.34}$$

$$A_N y_{N-2} - B_N y_{N-1} + C_N y_N = F_N. \tag{6.35}$$

The sweep method for system (6.31)–(6.35) uses the following solution representation:

$$y_i = \alpha_{i+1} y_{i+1} - \beta_{i+1} y_{i+2} + \gamma_{i+1}, \quad i = 0, 1, \ldots, N-2, \tag{6.36}$$

$$y_{N-1} = \alpha_N y_N + \gamma_N. \tag{6.37}$$

Like in the ordinary three-point sweep method, we find first the sweep coefficients. Using (6.36), we express y_{i-1} and y_{i-2} in terms of y_i and y_{i+1}:

$$y_{i-1} = \alpha_i y_i - \beta_i y_{i+1} + \gamma_i, \quad i = 1, 2, \ldots, N-1, \tag{6.38}$$

$$y_{i-2} = (\alpha_i \alpha_{i-1} - \beta_{i-1}) y_i - \beta_i \alpha_{i-1} y_{i+1} + \alpha_{i-1} \gamma_i + \gamma_{i-1}, \quad i = 2, 3, \ldots, N-1. \tag{6.39}$$

Substitution of (6.38) and (6.39) into (6.33) yields

$$\begin{aligned}(C_i - A_i \beta_{i-1} + \alpha_i (A_i \alpha_{i-1} - B_i)) y_i \\ = (D_i + \beta_i (A_i \alpha_{i-1} - B_i)) y_{i+1} - E_i y_{i+2} \\ + F_i - A_i \gamma_{i-1} - \gamma_i (A_i \alpha_{i-1} - B_i), \quad i = 2, 3, \ldots, N-2.\end{aligned}$$

In view of (6.36), we obtain

$$\begin{aligned}\alpha_{i+1} = S_i^{-1} (D_i + \beta_i (A_i \alpha_{i-1} - B_i)), \qquad \beta_{i+1} = S_i^{-1} E_i, \\ \gamma_{i+1} = S_i^{-1} (F_i - A_i \gamma_{i-1} - \gamma_i (A_i \alpha_{i-1} - B_i)),\end{aligned} \tag{6.40}$$

where

$$S_i = (C_i - A_i \beta_{i-1} + \alpha_i (A_i \alpha_{i-1} - B_i)), \quad i = 2, 3, \ldots, N-2. \tag{6.41}$$

The recurrent relations (6.40), (6.41) can be used provided that the coefficients α_i, β_i and γ_i $(i = 1, 2)$ are known. It follows from (6.31) and (6.36) that

$$\alpha_1 = C_0^{-1} D_0, \qquad \beta_1 = C_0^{-1} E_0, \qquad \gamma_1 = C_0^{-1} F_0. \tag{6.42}$$

Substituting (6.38) with $i = 1$ into (6.32), we obtain

$$(C_1 - B_1\alpha_1)y_1 = (D_1 - B_1\beta_1)y_2 - E_1y_3 + F_1 + B_1\gamma_1.$$

Hence, we can set:

$$\begin{gathered}\alpha_2 = S_1^{-1}(D_1 - B_1\beta_1), \qquad S_1 = C_1 - B_1\alpha_1, \\ \beta_2 = S_1^{-1}E_1, \qquad \gamma_2 = S_1^{-1}(F_1 + B_1\gamma_1).\end{gathered} \tag{6.43}$$

To determine the coefficients α_N, γ_N in (6.37), we substitute (6.38), (6.39) with $i = N - 1$ into (6.34). This yields the same formulas (6.40), (6.41) with $i = N - 1$. Thus, all sweep coefficients can be found by the recurrent formulas (6.40), (6.41) together with (6.42), (6.43).

For the solution to be found from (6.36), (6.37), we have to find y_N. To this end, we use equation (6.35). Substitution of (6.36) with $i = N - 2$ yields

$$y_N = \gamma_{N+1}, \tag{6.44}$$

where γ_{N+1} is to be found using formulas (6.40), (6.41) with $i = N$. Equality (6.44) allows the solution of (6.31)–(6.35) to be found from (6.36), (6.37).

The above algorithm is realized in the subroutine `PROG5`:

Subroutine PROG5

```
      SUBROUTINE PROG5 ( N, A, B, C, D, E, F, Y, ITASK )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     SWEEP METHOD
C     FOR FIVE-DIAGONAL MATRIX
C
C     ITASK = 1: FACTORIZATION AND SOLUTION;
C
C     ITASK = 2: SOLUTION ONLY
C
      DIMENSION A(N), B(N), C(N), D(N), E(N), F(N), Y(N)
      IF ( ITASK .EQ. 1 ) THEN
C
      D(1) = D(1) / C(1)
      E(1) = E(1) / C(1)
      C(2) =   C(2) - D(1)*B(2)
      D(2) = ( D(2) - E(1)*B(2) ) / C(2)
      E(2) =          E(2)        / C(2)

C
      DO I = 3, N
        C(I) = C(I) - E(I-2)*A(I) + D(I-1)*( D(I-2)*A(I) - B(I) )
        D(I) = ( D(I) + E(I-1)*( D(I-2)*A(I) - B(I) ) ) / C(I)
        E(I) =          E(I)                            / C(I)
      END DO
C
      ITASK = 2
```

```
      END IF
C
      F(1) = F(1) / C(1)
      F(2) = ( F(2) + F(1)*B(2) ) / C(2)

      DO I = 3, N
        F(I) = ( F(I) - F(I-2)*A(I)
     *               - F(I-1)*( D(I-2)*A(I) - B(I) ) ) / C(I)
      END DO
C
      Y(N) = F(N)
      Y(N-1) = D(N-1)*Y(N) + F(N-1)
      DO I = N-2, 1, -1
        Y(I) = D(I)*Y(I+1) - E(I)*Y(I+2) + F(I)
      END DO
      RETURN
      END
```

The identification problem was solved within the framework of a quasi-real experiment. First, the direct problem (6.1), (6.2) with given coefficients and right-hand side was solved. Afterwards, random perturbations were introduced into the mesh solution and, then, the inverse problem was treated. The difference solution of the direct problem was found using the sweep method for tridiagonal matrices. Like in the case of PROG5, in the subroutine PROG3 the matrix coefficients are not retained at the exit, being replaced instead with the sweep coefficients.

Subroutine PROG3

```
      SUBROUTINE PROG3 ( N, A, C, B, F, Y, ITASK )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     SWEEP METHOD
C     FOR TRIDIAGONAL MATRIX
C
C     ITASK = 1: FACTORIZATION AND SOLUTION;
C
C     ITASK = 2: SOLUTION ONLY
C

      DIMENSION A(N), C(N), B(N), F(N), Y(N)
      IF ( ITASK .EQ. 1 ) THEN
C
      B(1) = B(1) / C(1)
      DO I = 2, N
        C(I) = C(I) - B(I-1)*A(I)
        B(I) = B(I) / C(I)

      END DO
C
      ITASK = 2
      END IF
C
      F(1) = F(1) / C(1)
      DO I = 2, N
```

```
        F(I) = ( F(I) + F(I-1)*A(I) ) / C(I)
      END DO
C
      Y(N) = F(N)
      DO I = N-1, 1, -1
        Y(I) = B(I)*Y(I+1) + F(I)
      END DO
      RETURN
      END
```

The regularization parameter was found from the discrepancy using the sequence

$$\alpha_k = \alpha_0 q^k, \qquad q > 0$$

with set α_0 and q.

Program PROBLEM5

```
C
C     PROBLEM5  - RIGHT-HAND SIDE IDENTIFICATION
C                 ONE-DIMENSIONAL STATIONARY PROBLEM
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.001D0, N = 201 )
      DIMENSION  Y(N), X(N), YD(N), F(N), FA(N), FT(N)
     +          ,A(N), B(N), C(N),  D(N), E(N), FF(N)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF THE SEGMENT;
C     N         - NUMBER OF GRID NODES;
C     DELTA     - INPUT-DATA INACCURACY;
C     ALPHA     - INITIAL REGULARIZATION PARAMETER;
C     Q         - MULTIPLIER IN THE REGULARIZATION PARAMETER;
C     Y(N)      - EXACT DIFFERENCE SOLUTION OF THE BOUNDARY-VALUE
C                 PROBLEM;
C     YD(N)     - DISTURBED DIFFERENCE SOLUTION OF THE BOUNDARY-VALUE
C                 PROBLEM;
C     F(N)      - EXACT RIGHT-HAND SIDE;
C     FA(N)     - CALCULATED RIGHT-HAND SIDE;
C
      XL = 0.
      XR = 1.
C
      OPEN (01, FILE = 'RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H = (XR - XL) / (N - 1)
      DO I = 1, N
        X(I) = XL + (I-1)*H
      END DO
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
```

```
      B(1)  = 0.D0
      C(1)  = 1.D0
      FF(1) = 0.D0
      A(N)  = 0.D0
      C(N)  = 1.D0
      FF(N) = 0.D0

      DO I = 2,N-1
        A(I)  = AK(X(I)-0.5D0*H) / (H*H)
        B(I)  = AK(X(I)+0.5D0*H) / (H*H)
        C(I)  = A(I) + B(I) + AQ(X(I))
        FF(I) = AF(X(I))
      END DO
C
C     SOLUTION OF THE DIFFERENCE PROBLEM

C
      ITASK1 = 1
      CALL PROG3 ( N, A, C, B, FF, Y, ITASK1 )
C
C     NOISE ADDITION TO THE SOLUTION OF THE BOUNDARY-VALUE PROBLEM
C
      YD(1) = 0.D0
      YD(N) = 0.D0
      DO I = 2,N-1
        YD(I) = Y(I) + 2.*DELTA*(RAND(0)-0.5)
      END DO
C

C     RIGHT-HAND SIDE
C
      FT(1) = 0.D0
      FT(N) = 0.D0
      DO I = 2,N-1
        FT(I) = - AK(X(I)-0.5D0*H) * YD(I-1) / (H*H) +
     +  ((AK(X(I)-0.5D0*H) + AK(X(I)+0.5D0*H)) / (H*H) +
     +    AQ(X(I))) * YD(I) -
     +    AK(X(I)+0.5D0*H) * YD(I+1) / (H*H)
      END DO
      WRITE ( 01,* ) (X(I),  I = 1,N)
      WRITE ( 01,* ) (Y(I),  I = 1,N)
      WRITE ( 01,* ) (YD(I), I = 1,N)
      WRITE ( 01,* ) (FT(I), I=1,N)
C
C     ITERATIVE ADJUSTMENT OF THE REGULARIZATION PARAMETER

C
      IT = 0
      ITMAX = 1000
      ALPHA = 0.001D0
      Q = 0.75D0
  100 IT = IT + 1
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE INVERSE PROBLEM
C
      C(1) = 1.D0


      D(1) = 0.D0
      E(1) = 0.D0
```

```
       B(2) = 0.D0
       C(2) = ALPHA*(AK(X(2)+0.5D0*H) / (H**2)) **2 +
      +       ALPHA*( (AK(X(2)-0.5D0*H) +
      +       AK(X(2)+0.5D0))/(H**2) + AQ(X(2)))**2 + 1.D0
       D(2) = ALPHA*AK(X(2)+0.5D0*H) / (H**2) *
      +    ( (AK(X(2)+0.5D0*H) +
      +       AK(X(2)+1.5D0*H))/(H**2) + AQ(X(3)) +
      +      (AK(X(2)-0.5D0*H) +
      +       AK(X(2)+0.5D0*H))/(H**2) + AQ(X(2)))
       E(2) = ALPHA*AK(X(2)+0.5D0*H)* AK(X(2)+1.5D0*H) / (H**4)
       A(N-1) = ALPHA*AK(X(N-1)-0.5D0*H)* AK(X(N-1)-1.5D0*H) / (H**4)
       B(N-1) = ALPHA*AK(X(N-1)-0.5D0*H) / (H**2) *
      +      ( (AK(X(N-1)-0.5D0*H) + AK(X(N-1)-1.5D0*H))/(H**2) +
      +         AQ(X(N-2) ) +
      +        (AK(X(N-1)-0.5D0*H) + AK(X(N-1)+0.5D0*H))/(H**2) +
      +         AQ(X(N-1)))
       C(N-1) = ALPHA*(AK(X(N-1)-0.5D0*H) / (H**2)) **2 +
      +         ALPHA*( (AK(X(N-1)-0.5D0*H) +
      +         AK(X(N-1)+0.5D0*H))/(H**2) + AQ(X(N-1)))**2 + 1.D0
       D(N-1) = 0.D0
       A(N) = 0.D0
       B(N) = 0.D0
       C(N) = 1.D0
       DO I = 2,N-2
         A(I) = ALPHA*AK(X(I)-0.5D0*H)* AK(X(I)-1.5D0*H) / (H**4)
         B(I) = ALPHA*AK(X(I)-0.5D0*H) / (H**2) *
      +      ( (AK(X(I)-0.5D0*H) + AK(X(I)-1.5D0*H))/(H**2) +
      +         AQ(X(I-1)) +
      +        (AK(X(I)-0.5D0*H) + AK(X(I)+0.5D0*H))/(H**2) + AQ(X(I)))
         C(I) = ALPHA*(AK(X(I)-0.5D0*H) / (H**2)) **2 +
      +         ALPHA*(AK(X(I)+0.5D0*H) / (H**2)) **2 +
      +         ALPHA*( (AK(X(I)-0.5D0*H) +
      +         AK(X(I)+0.5D0))/(H**2) + AQ(X(I)))**2 + 1.D0
         D(I) = ALPHA*AK(X(I)+0.5D0*H) / (H**2) *
      +      ( (AK(X(I)+0.5D0*H) + AK(X(I)+1.5D0*H))/(H**2) +
      +         AQ(X(I+1)) +
      +        (AK(X(I)-0.5D0*H) + AK(X(I)+0.5D0*H))/(H**2) + AQ(X(I)))
         E(I) = ALPHA*AK(X(I)+0.5D0*H)* AK(X(I)+1.5D0*H) / (H**4)
       END DO

C
C      SOLUTION OF THE DIFFERENCE PROBLEM
C
       ITASK2 = 1
       DO I = 1,N
         FF(I) = FT(I)
       END DO
       CALL PROG5 ( N, A, B, C, D, E, FF, FA, ITASK2 )
C


C      DISCREPANCY
C

       B(1)  = 0.D0
       C(1)  = 1.D0
       FF(1) = 0.D0
       A(N)  = 0.D0
       C(N)  = 1.D0
       FF(N) = 0.D0
```

```
      DO I = 2,N-1
        A(I)  = AK(X(I)-0.5D0*H) / (H*H)
        B(I)  = AK(X(I)+0.5D0*H) / (H*H)
        C(I)  = A(I) + B(I) + AQ(X(I))
        FF(I) = FA(I)
      END DO

C
      ITASK1 = 1

      CALL PROG3 ( N, A, C, B, FF, D, ITASK1 )
C
      SUM = 0.D0
      DO I = 2,N-1
        SUM = SUM + (D(I)-YD(I))**2*H
      END DO
      SL2 = DSQRT(SUM)
C
      IF ( IT.EQ.1 ) THEN
        IND = 0
        IF ( SL2.LT.DELTA ) THEN
          IND = 1
          Q = 1.D0/Q
        END IF
         ALPHA = ALPHA*Q
        GO TO 100
      ELSE
        ALPHA = ALPHA*Q
        IF ( IND.EQ.0 .AND. SL2.GT.DELTA ) GO TO 100
        IF ( IND.EQ.1 .AND. SL2.LT.DELTA ) GO TO 100
      END IF
C
C     SOLUTION
C
      WRITE ( 01,* ) (FA(I), I=1,N)
      WRITE ( 01,* ) IT, ALPHA, SL2
      CLOSE ( 01 )
C
      STOP
      END

      DOUBLE PRECISION FUNCTION  AK ( X )
      IMPLICIT REAL*8 ( A-H, O-Z )
C

C      COEFFICIENT AT THE HIGHER DERIVATIVES
C
      AK = 0.05D0
C
      RETURN
      END

      DOUBLE PRECISION FUNCTION  AQ ( X )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      COEFFICIENT AT THE LOWEST TERM
C
      AQ = 0.D0
C
      RETURN
```

```
      END



      DOUBLE PRECISION FUNCTION  AF ( X )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      RIGHT-HAND SIDE
C
      AF = X
      IF ( X.GE.0.5D0 ) AF = 0.5D0
C
      RETURN
      END
```

The equation coefficients and the right-hand side are set in the subroutine functions `AK, AQ, AF`.

6.1.6 Examples

As an example, the model problem (6.1), (6.2) with

$$k(x) = 0.05, \quad q(x) = 0, \quad f(x) = \begin{cases} x, & 0 < x < 0.5, \\ 0.5, & 0.5 < x < 1, \end{cases}$$

was considered. The input-data inaccuracies were modeled by perturbing the difference solution of the problem with exact right-hand side at grid nodes:

$$y_\delta(x) = y(x) + 2\delta(\sigma(x) - 1/2), \quad x = x_i, \quad i = 0, 1, \ldots, N-1,$$

where $\sigma(x)$ is a random quantity normally distributed over the interval from 0 to 1.

Note first some direct difference differentiation possibilities in calculating the right-hand side of (6.1). Figure 6.1 shows data calculated on various difference grids. The data were obtained for the perturbed solution of the direct problem on a grid with $N = 512$; here, the noise level was defined by the parameter $\delta = 0.001$. Acceptable results could also be obtained using numerical differentiation on rather crude grids with $h \geq 1/16$.

The solution of the same identification problem obtained by the Tikhonov method is shown in Figure 6.2. Here, a calculation grid with $N = 200$ was used. The effect due to the inaccuracy level can be traced considering the data calculated with the input-data inaccuracy $\delta = 0.01$; these data are shown in Figure 6.3. A substantial loss of accuracy near the right boundary $x = l$ is observed. This loss is related with the fact that the calculated operator $\mathcal{D}$ is defined on the set of functions satisfying the boundary conditions (6.2). If the right-hand side to be found also satisfies these conditions, an improved accuracy in finding the approximate solution can be expected. The latter is illustrated by Figure 6.4 that shows the solution of the identification problem with somewhat modified right-hand side.

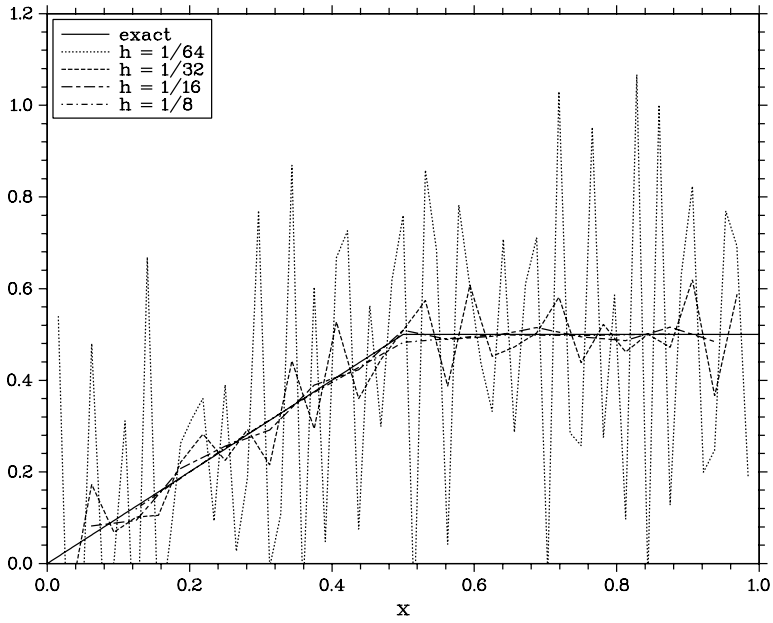

Figure 6.1 Numerical differentiation with $\delta = 0.001$

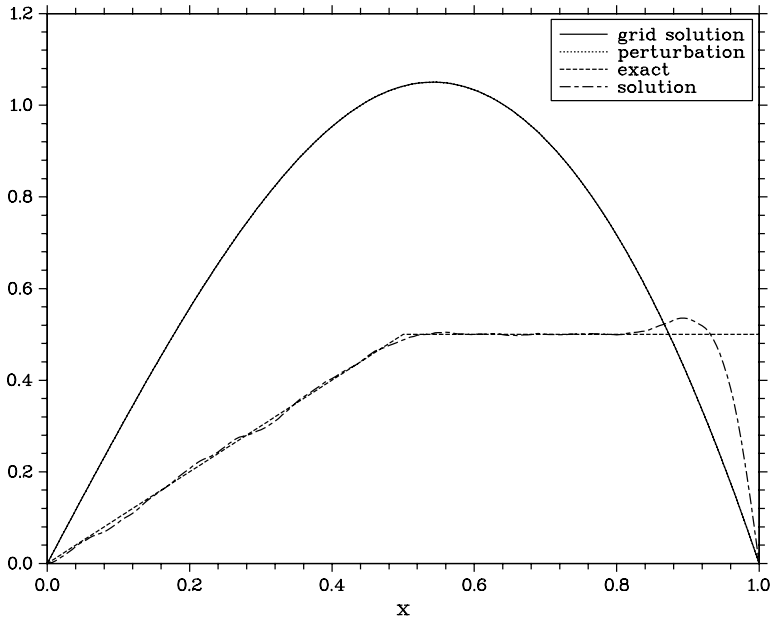

Figure 6.2 Solution of the identification problem obtained with $\delta = 0.001$

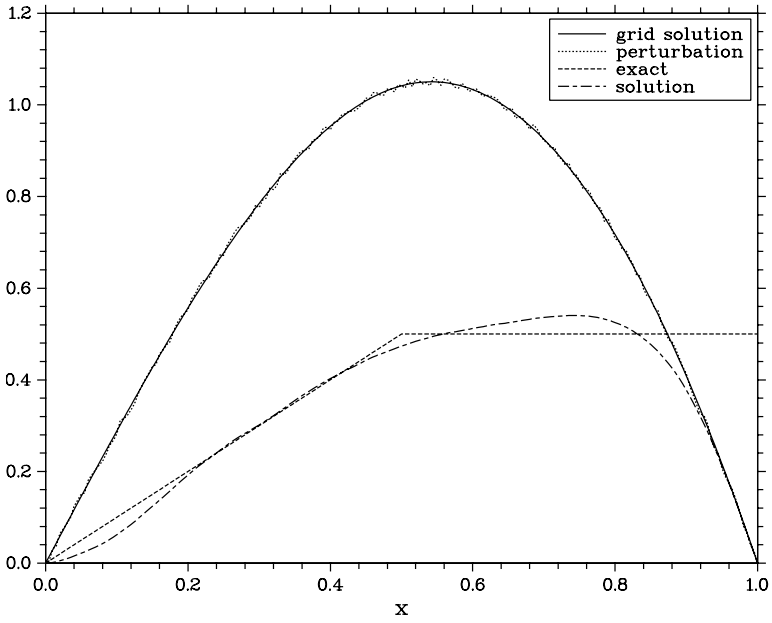

Figure 6.3 Solution of the identification problem obtained with $\delta = 0.01$

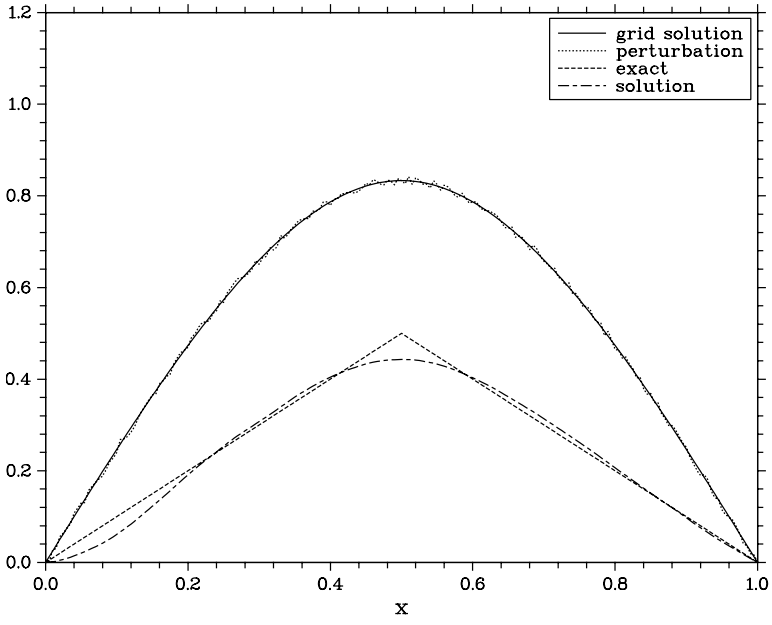

Figure 6.4 Solution of the identification problem obtained with $\delta = 0.01$ and for a modified right-hand side

6.2 Right-hand side identification in the case of a parabolic equation

Below, we consider an inverse problem in which it is required to reconstruct the right-hand side of a one-dimensional parabolic equation from the known solution. Specific features of the problem are discussed, related with the evolutionary character of the problem and with the possibility to sequentially determine the right-hand side with increasing time.

6.2.1 Model problem

As a model problem, consider the problem in which it is required to reconstruct the right-hand side of a one-dimensional parabolic equation. Let us begin the consideration with the statement of the direct problem.

The solution $u(x,t)$ is defined in the rectangle

$$\overline{Q}_T = \overline{\Omega} \times [0,T], \quad \overline{\Omega} = \{x \mid 0 \le x \le l\}, \quad 0 \le t \le T.$$

The function $u(x,t)$ satisfies the equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) + f(x,t), \qquad 0 < x < l, \quad 0 < t \le T \tag{6.45}$$

under the standard constraint $k(x) \ge \kappa > 0$.

For simplicity, homogeneous boundary and initial conditions are assumed:

$$u(0,t) = 0, \qquad u(l,t) = 0, \qquad 0 < t \le T, \tag{6.46}$$

$$u(x,0) = 0, \qquad 0 \le x \le l. \tag{6.47}$$

In the direct problem (6.45)–(6.47), the solution $u(x,t)$ is to be found from the known coefficient $k(x)$ and from the known right-hand side $f(x,t)$. In the inverse problem, the unknown quantity is the right-hand side $f(x,t)$ (source power), with the solution $u(x,t)$ assumed known. The right-hand side can be calculated by the following explicit formula obtained from (6.45):

$$f(x,t) = \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l, \quad 0 < t \le T. \tag{6.48}$$

The input data are given with some inaccuracy; this circumstance makes the direct use of formula (6.48) difficult. Most significant here is the effect resulting from the solution inaccuracy. Let us know, instead of the exact solution $u(x,t)$ of problem (6.45)–(6.47), a perturbed solution $u_\delta(x,t)$, and in some norm the parameter δ defines the inaccuracy level in the solution, i. e.,

$$\|u_\delta(x,t) - u(x,t)\| \le \delta. \tag{6.49}$$

Here, special computational algorithms for stable numerical differentiation must be applied. To approximate the solution, we can use finite-difference regularization algorithms in which the discretization step size serves the regularization parameter. This method was discussed at length previously, when we considered the stationary problem for the second-order ordinary differential equation. In the solution of the inverse problem (6.46)–(6.49), the discretization step sizes over space and time must be matched with the inaccuracy in $u(x, t)$.

Evolutionary problems possess some specific feature of utmost significance. The current solution of the problem depends only on the prehistory, i.e., on the solutions at preceding times, and does not depends on the solutions at future times. This point can (and often must) be taken into account when developing computational algorithms. This requirement is perfectly justified in the case of direct problems of type (6.45)–(6.47) and can also be important when one considers inverse problems similar to problem (6.46)–(6.49).

Discussing the solution of evolutionary problems in the general context, we can speak of computational algorithms of two types. In the computational algorithms of the first type, the current solution is to be calculated from the solution at preceding times. Here, we speak of *local algorithms for solving evolutionary problems*. In the *global algorithms* the current solution is sought using a procedure that involves future times.

6.2.2 Global regularization

To approximately solve the inverse problem (6.46)–(6.49), we use the general regularization scheme proposed by A. N. Tikhonov. First of all, we introduce some new settings.

In the Hilbert space $\mathcal{L}_2(\Omega)$, in the ordinary way we introduce the norm and the scalar product:

$$(v, w) = \int_\Omega v(x)w(x)\,dx, \qquad \|v\|^2 = (v, v) = \int_\Omega v^2(x)\,dx.$$

For functions $v(x, t)$, $w(x, t) \in \mathcal{H}$, where $\mathcal{H} = \mathcal{L}_2(Q_T)$, we set:

$$(v, w)^* = \int_0^T (v, w)\,dt = \int_0^T \int_\Omega v(x)w(x)\,dx\,dt, \qquad \|v\|^* = ((v, v)^*)^{1/2}.$$

We define the operator

$$\mathcal{A}u = -\frac{d}{dx}\left(k(x)\,\frac{du}{dx}\right)$$

on the set of functions satisfying the condition (6.44). In $\mathcal{L}_2(\Omega)$, for the operator $\mathcal{A}$ we have:

$$\mathcal{A} = \mathcal{A}^* \geq mE, \qquad m = \kappa\frac{\pi^2}{l^2}.$$

Then, the inverse problem (6.46)–(6.48) can be written as

$$f = \mathcal{D}u, \tag{6.50}$$

where the operator

$$\mathcal{D}u = \frac{du}{dt} + \mathcal{A}u \tag{6.51}$$

is defined on the set of functions satisfying the initial condition (6.47). For the input data (see (6.49)), we have:

$$\|u_\delta - u\|^* \le \delta. \tag{6.52}$$

In the Tikhonov regularization method, the approximate solution f_α of problem (6.50)–(6.52) is sought as the solution of the following variational problem:

$$(E + \alpha \mathcal{D}^* \mathcal{D}) f_\alpha = \mathcal{D}u_\delta. \tag{6.53}$$

A specific feature of the regularization method as applied to the solution of the inverse evolutionary problem is manifested in the problem operator $\mathcal{D}$ defined by (6.51). In particular, it becomes necessary to explicitly define the operator $\mathcal{D}^*$.

Provided that $v(x, 0) = 0$ and $w(x, T) = 0$, we have:

$$\begin{aligned}(\mathcal{D}v, w)^* &= \int_0^T \Big(\frac{dv}{dt}, w\Big)\, dt + \int_0^T (\mathcal{A}, w)\, dt \\ &= (v, w)\Big|_0^T - \int_0^T \Big(v, \frac{dw}{dt}, w\Big)\, dt + \int_0^T (\mathcal{A}, w)\, dt \;=\; (v, \mathcal{D}w)^*.\end{aligned}$$

Thus, we define the operator $\mathcal{D}^*$ as

$$\mathcal{D}^* w = -\frac{dw}{dt} + \mathcal{A}w \tag{6.54}$$

on the set of functions

$$w(x, T) = 0, \qquad 0 \le x \le l. \tag{6.55}$$

In view of (6.47), (6.51), (6.54) and (6.55), equation (6.53) for the approximate solution f_α is an elliptic equation involving second time derivatives and forth spatial derivatives. For f_α, the boundary conditions at $t = 0, T$ are as follows:

$$f_\alpha(x, 0) = 0, \qquad 0 \le x \le l, \tag{6.56}$$

$$\mathcal{D} f_\alpha(x, T) = 0, \qquad 0 \le x \le l. \tag{6.57}$$

These points should be given particular attention in the numerical realization.

In the case of the identification problem for the right-hand side of the second-order ordinary differential equation, the following two basic possibilities are available. The first possibility implies using the Tikhonov regularization method in interpreting the right-hand side identification problem as a problem in which it is required to solve a

first-kind operator equation. In the second possibility, the right-hand side identification problem is considered as a problem in which it is required to calculate the values of a bounded operator. The latter possibility was realized in the scheme (6.52), (6.53) for problem (6.46)–(6.49). It also makes sense to dwell here on the standard variant of the Tikhonov regularization.

With the given right-hand side $f(x,t)$, the solution of the boundary value problem (6.45)–(6.47) uniquely defines the solution $u(x,t)$. To reflect this correspondence, we introduce the operator $\mathcal{G}$:

$$\mathcal{G}f = u. \tag{6.58}$$

Instead of $u(x,t)$, the function $u_\delta(x,t)$ is given and, in addition, the estimate (6.52) holds.

The approximate solution f_α of problem (6.49), (6.58) can be found as the solution of the following problem:

$$J_\alpha(f_\alpha) = \min_{v \in \mathcal{H}} J_\alpha(v), \tag{6.59}$$

where

$$J_\alpha(v) = (\|\mathcal{G}v - u_\delta\|^*)^2 + \alpha(\|v\|^*)^2. \tag{6.60}$$

The following important circumstance deserves mention. In using algorithm (6.52), (6.53), constructed around the interpretation of the identification problem as an unbounded operator value problem, additional constraints on the sought function (boundary conditions of type (6.56), (6.57)) are to be posed. This is not quite justified a procedure (see the results of numerical experiments on the identification of the right-hand side of the ordinary differential equation in the previous section). No such problems are encountered in using (6.59), (6.60).

6.2.3 Local regularization

In the approximate solution of the problem in which it is required to identify the right-hand side of a non-stationary equation from known solution, one can more conveniently employ an algorithm that makes it possible to determine the right-hand side at a given time using input information only at preceding times. Compared to the global regularization algorithm, here we, speaking generally, loose in the approximate-solution accuracy, but save time. Consider some basic possibilities in the construction of local regularization algorithms for the approximate solution of the inverse problem (6.46)–(6.49). We will dwell here on the local analogue of the regularization of type (6.59), (6.60).

The basic idea uses the fact that the right-hand side is defined by the solution by each fixed moment. In other words, the numerical differentiation procedure is to be regularized only with respect to spatial variables. In fact, the input data are to be smoothed considering only part of all variables. Such algorithms can be realized employing preliminary discretization over time.

We introduce the uniform grid over time

$$\bar{\omega}_\tau = \omega_\tau \cup \{T\} = \{t_n = n\tau,\ n = 0, 1, \dots, N_0,\ \tau N_0 = T\}.$$

The following subscriptless notation generally accepted in the theory of difference schemes will be used:

$$y = y^n, \qquad \hat{y} = y^{n+1}, \qquad \check{y} = y^{n-1},$$

$$y_{\bar{t}} = \frac{y - \check{y}}{\tau}, \qquad y_t = \frac{\hat{y} - y}{\tau}.$$

Let us formulate the inverse right-hand side identification problem for the parabolic equation after the partial discretization. For the main quantities, the same settings as in the continuous case will be used. For simplicity, we restrict the consideration to the case of purely explicit time approximation, in which the right-hand side is determined (compare (6.48)) from the differential-difference relation

$$f^n = \frac{u^n - u^{n-1}}{\tau} + \mathcal{A}u^n, \qquad n = 1, 2, \dots, N_0. \tag{6.61}$$

The input data (direct-problem solution u^n) is set accurate to some inaccuracy. We assume the inaccuracy level to be characterized by a constant δ, so that

$$\|u_\delta^n - u^n\| \le \delta, \qquad n = 1, 2, \dots, N_0. \tag{6.62}$$

We denote the approximate solution of problem (6.61), (6.62) at the moment $t = t_n$ as f_α^n. Consider the matter of reconstruction of the function f_α^n from given u_δ^n, u_δ^{n-1} and f_α^{n-1}. We assume that the approximate right-hand side f_α^n corresponds to the solution of the boundary value problem

$$\frac{w^n - w^{n-1}}{\tau} + \mathcal{A}w^n = f_\alpha^n, \qquad n = 1, 2, \dots, N_0, \tag{6.63}$$

$$w^0 = 0. \tag{6.64}$$

First, we determine f_α^n as the solution of the problem

$$J_\alpha^n(f_\alpha^n) = \min_{v \in \mathcal{H}} J_\alpha^n(v), \tag{6.65}$$

where $\mathcal{H} = \mathcal{L}_2(\Omega)$ and

$$J_\alpha^n(v) = \|u_\delta - w(v)\|^2 + \alpha\|v\|^2. \tag{6.66}$$

Here, in view of (6.63) and (6.64), $w(v)$ is the solution of the difference problem

$$\frac{w - \check{w}}{\tau} + \mathcal{A}w = v. \tag{6.67}$$

This allows us to write equation (6.66) as

$$J_\alpha^n(v) = \left\| u_\delta - \frac{1}{\tau}\,\mathcal{G}_\tau \check{w} - \mathcal{G}_\tau v \right\|^2 + \alpha \|v\|^2,$$

where

$$\mathcal{G}_\tau = \mathcal{G}_\tau^* = \left(\frac{1}{\tau}\,E + \mathcal{A}\right)^{-1}.$$

In this way, the approximate solution at each time $t = t_n$ can be determined from the equation

$$\mathcal{G}_\tau^* \mathcal{G}_\tau f_\alpha + \alpha f_\alpha = \mathcal{G}_\tau^* u_\delta - \frac{1}{\tau}\,\mathcal{G}_\tau^* \mathcal{G}_\tau \check{w}. \tag{6.68}$$

With (6.67), using the introduced notation, we obtain the following problem for w_n:

$$\mathcal{G}_\tau^{-1} w = f_\alpha + \frac{1}{\tau}\,\check{w}. \tag{6.69}$$

6.2.4 Iterative solution of the identification problem

In solving inverse mathematical physics problems, primary attention should be paid to iteration methods that most fully realize the idea of finding the solution of the inverse problem through successive solution of a set of direct problems. Considering approximate solution of problem (6.46)–(6.49) by iteration methods, here we dwell on global regularization in the variant (6.58)–(6.60).

After symmetrization of (6.58), we write the two-layer iteration method as

$$\frac{f_{k+1} - f_k}{\tau_{k+1}} + \mathcal{G}^* \mathcal{G} f_k = \mathcal{G}^* u_\delta, \qquad k = 0, 1, \ldots. \tag{6.70}$$

In using the steepest descend method, the iteration parameters can be calculated by

$$\tau_{k+1} = \left(\frac{\|r_k\|^*}{\|\mathcal{G} r_k\|^*}\right)^2, \quad r_k = \mathcal{G}^* \mathcal{G} f_k - \mathcal{G}^* u_\delta, \quad k = 0, 1, \ldots\,. \tag{6.71}$$

The number of iterations in (6.70), (6.71) is to be matched with the inaccuracy δ (see (6.52)).

This approach can be of use provided that we are able to calculate the operators $\mathcal{G}$ and $\mathcal{G}^*$. Recall that $v = \mathcal{G} f_k$ is the solution of the direct problem

$$\frac{dv}{dt} + \mathcal{A} v = f_k, \qquad 0 < t \le T, \tag{6.72}$$

$$v(0) = 0. \tag{6.73}$$

The values of the conjugate operator $w = \mathcal{G}^* v$ can be found by solving the direct problem

$$-\frac{dw}{dt} + \mathcal{A} w = v, \qquad 0 \le t < T, \tag{6.74}$$

$$w(T) = 0. \tag{6.75}$$

Thus, for a given iteration parameter, the passage to the next iteration is related, in line with (6.70), with the solution of two direct problems, namely, (6.72), (6.73) and (6.74), (6.75).

Consider some specific features in the computational realization of the iteration method related, first of all, with time discretization. We retain for the mesh functions the same settings as for the continuous-argument functions.

We use a uniform grid $\bar{\omega}$ over the interval $\bar{\Omega} = [0, l]$ with a grid size h:

$$\bar{\omega} = \{x \mid x = x_i = ih,\ i = 0, 1, \ldots, N,\ Nh = l\}.$$

Here, ω is the set of internal nodes, and $\partial\omega$ is the set of boundary nodes.

We approximate the differential operator $\mathcal{A}$ at internal nodes accurate to the second order with the difference operator

$$Ay = -(ay_{\bar{x}})_x, \qquad x \in \omega, \tag{6.76}$$

where, for instance, $a(x) = k(x - 0.5h)$.

In the mesh Hilbert space $L_2(\omega)$, we introduce the norm by the relation $\|y\| = (y, y)^{1/2}$, where

$$(y, w) = \sum_{x\in\omega} y(x)w(x)h.$$

Recall that on the set of functions vanishing on $\partial\omega$, for the self-conjugate operator A under the constraints $k(x) \geq \kappa > 0$, $q(x) \geq 0$ there holds the estimate

$$A = A^* \geq \kappa\lambda_0 E, \tag{6.77}$$

where

$$\lambda_0 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l} \geq \frac{8}{l^2}.$$

We put into correspondence to the direct problem (6.72), (6.73) the following symmetric difference problem:

$$\frac{v^n - v^{n-1}}{\tau} + \frac{1}{2} A(v^{n+1} + v^n) = \frac{1}{2} (f_k^{n+1} + f_k^n), \qquad n = 1, 2, \ldots, N_0, \tag{6.78}$$

$$v^0 = 0, \qquad x \in \omega. \tag{6.79}$$

In the latter case, the approximation inaccuracy is of second order both over time and space. In a similar manner, other two-layer difference schemes can be considered. We bring the problem (6.78), (6.79) to the operator notation $v = Gf_k$, which defines the operator G.

For two-dimensional mesh functions, we define a Hilbert space $H = L_2(Q_T)$ in which the scalar product and the norm are defined as

$$(v, w)^* = \sum_{n=1}^{N_0-1} (v^n, w^n)\tau + \frac{\tau}{2} (v^0, w^0) + \frac{\tau}{2} (v^{N_0}, w^{N_0}), \qquad \|v\|^* = \sqrt{(v, v)^*}.$$

A problem conjugate in H to (6.78), (6.79) is the difference problem (see (6.74), (6.75))

$$-\frac{w^n - w^{n-1}}{\tau} + \frac{1}{2} A(w^n + w^{n-1}) = \frac{1}{2}(v^n + v^{n-1}), \quad n = 1, 2, \ldots, N_0, \tag{6.80}$$

$$w^{N_0} = 0, \qquad x \in \omega. \tag{6.81}$$

We can check this by scalarwise multiplying equation (6.78) by w^n. The problem (6.80), (6.81) can be written in a compact form as $w = G^* v$.

In line with (6.70), the iteration method can be written as

$$\frac{f_{k+1} - f_k}{\tau_{k+1}} + G^* G f_k = G^* u_\delta, \qquad k = 0, 1, \ldots . \tag{6.82}$$

At the first stage, we have to calculate the right-hand side $G^* u_\delta$; to this end, a boundary value problem of type (6.80), (6.81) is to be solved. For the difference $r_k = G^* G f_k - G^* u_\delta$ to be calculated, it is required at each step to solve two boundary value problems, ((6.78), (6.79) and (6.80), (6.81)). The iteration parameters can be calculated (see (6.71)) by the formula

$$\tau_{k+1} = \left(\frac{\|r_k\|^*}{\|G r_k\|^*} \right)^2, \qquad k = 0, 1, \ldots . \tag{6.83}$$

Determination of $G r_k$ requires an additional boundary value problem of type (6.78), (6.79) to be solved.

The regularization parameter here is the number $K(\delta)$ of iterations in (6.82). The criterion for the exit from the iterative process is

$$\|G f_{K(\delta)} - u_\delta\|^* \le \delta. \tag{6.84}$$

It should be noted that the algorithm of interest is to be used under certain restrictions imposed on the right-hand side. In accordance with the applied symmetrization at the expense of the operator G, the following constraints on the right-hand side are imposed:

$$f_\alpha^{N_0} = 0, \qquad x \in \omega,$$

$$f_\alpha^n = 0, \qquad n = 0, 1, \ldots, N, \quad x \in \partial\omega.$$

The algorithm (6.78)–(6.84) is embodied in the program PROBLEM6. Given below is the complete text of this program.

Program PROBLEM6

```
C
C     PROBLEM6  - RIGHT-HAND SIDE IDENTIFICATION
C                 ONE-DIMENSIONAL NON-STATIONARY PROBLEM
C
```

```
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.05D0, N = 101, M = 101 )
      DIMENSION  X(N), Y(N,M), YD(N,M), F(N,M), FA(N,M)
     +          ,V(N,M), W(N,M), GU(N,M), RK(N,M), YY(N)
     +          ,A(N), B(N), C(N),  D(N), E(N), FF(N)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF THE SEGMENT;
C     N         - NUMBER OF GRID NODES OVER THE SPATIAL VARIABLE;
C     TMAX      - MAXIMAL TIME;
C     M         - NUMBER OF GRID NODES OVER TIME;
C     DELTA     - INPUT-DATA INACCURACY;
C     Y(N,M)    - EXACT DIFFERENCE SOLUTION OF THE BOUNDARY-VALUE
C                 PROBLEM;
C     YD(N,M)   - DISTURBED DIFFERENCE SOLUTION OF THE BOUNDARY-VALUE
C                 PROBLEM;
C     F(N,M)    - EXACT RIGHT-HAND SIDE;
C     FA(N,M)   - CALCULATED RIGHT-HAND SIDE;
C
      XL = 0.D0
      XR = 1.D0
      TMAX = 1.D0
C
      OPEN (01, FILE = 'RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H = (XR - XL) / (N - 1)
      DO I = 1, N
        X(I) = XL + (I-1)*H
      END DO
      TAU = TMAX / (M-1)
C
C     DIRECT PROBLEM
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I,1)  = 0.D0
        YD(I,1) = 0.D0
      END DO
C
C     NEXT TIME LAYER

C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        DO I = 2, N-1
          X1 = (X(I) + X(I-1)) / 2
          X2 = (X(I+1) + X(I)) / 2

          A(I) = AK(X1) / (2*H*H)
          B(I) = AK(X2) / (2*H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          FF(I) = A(I) * Y(I-1,K-1)
```

```
     +          + (1.D0 / TAU - A(I) - B(I) ) * Y(I,K-1)
     +          + B(I) * Y(I+1,K-1)
     +          + (AF(X(I),T) + AF(X(I),T-TAU)) / 2
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
        B(1)  = 0.D0
        C(1)  = 1.D0
        FF(1) = 0.D0

        A(N)  = 0.D0
        C(N)  = 1.D0
        FF(N) = 0.D0
C
C     SOLUTION AT THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, FF, YY, ITASK )
        DO I = 1, N
          Y(I,K) = YY(I)
        END DO
      END DO
C
C     NOISE ADDITION
C
      DO K = 2, M

        YD(1,K) = 0.D0
        YD(N,K) = 0.D0
        DO I = 2,N-1
          YD(I,K) = Y(I,K)
     +            + 2.D0*DELTA*(RAND(0)-0.5D0) / DSQRT(TMAX*(XR-XL))
        END DO
      END DO
C
C     SYMMETRIZATION (SOLUTION OF THE CONJUGATE PROBLEM)
C
C     INITIAL CONDITION
C
      T = TMAX
      DO I = 1, N
        GU(I,M) = 0.D0
      END DO

C
C     NEXT TIME LAYER
C
      DO K = M-1, 1, -1

        T = T - TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        DO I = 2, N

          X1 = (X(I) + X(I-1)) / 2
          X2 = (X(I+1) + X(I)) / 2
          A(I) = AK(X1) / (2*H*H)
          B(I) = AK(X2) / (2*H*H)
```

```
          C(I) = A(I) + B(I) + 1.D0 / TAU
          FF(I) = A(I) * GU(I-1,K+1)
     +          + (1.D0 / TAU - A(I) - B(I) ) * GU(I,K+1)
     +          + B(I) * GU(I+1,K+1)
     +          + (YD(I,K) + YD(I,K+1)) / 2
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
        B(1)  = 0.D0
        C(1)  = 1.D0
        FF(1) = 0.D0
        A(N)  = 0.D0
        C(N)  = 1.D0
        FF(N) = 0.D0

C
C     SOLUTION OF THE PROBLEM AT THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, FF, YY, ITASK )

        DO I = 1, N
          GU(I,K) = YY(I)
        END DO
      END DO
C
C     INVERSE RIGHT-HAND SIDE IDENTIFICATION PROBLEM
C
      IT = 0
      ITMAX = 1000
C
C     INITIAL APPROXIMATION
C
      DO K = 1, M
        DO I = 1, N
          FA(I,K) = 0.D0
        END DO
      END DO
  100 IT = IT+1
C
C     SOLUTION OF THE DIRECT PROBLEM FOR THE GIVEN RIGHT-HAND SIDE
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        V(I,1)  = 0.D0
      END DO
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        DO I = 2, N
          X1 = (X(I) + X(I-1)) / 2
          X2 = (X(I+1) + X(I)) / 2
```

```
            A(I) = AK(X1) / (2*H*H)
            B(I) = AK(X2) / (2*H*H)
            C(I) = A(I) + B(I) + 1.D0 / TAU
            FF(I) = A(I) * V(I-1,K-1)
     +            + (1.D0 / TAU - A(I) - B(I) ) * V(I,K-1)
     +            + B(I) * V(I+1,K-1)
     +            + (FA(I,K) + FA(I,K-1)) / 2
          END DO

C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
          B(1)  = 0.D0
          C(1)  = 1.D0
          FF(1) = 0.D0
          A(N)  = 0.D0
          C(N)  = 1.D0
          FF(N) = 0.D0
C
C     SOLUTION OF THE PROBLEM AT THE NEXT TIME LAYER
C
          ITASK = 1
          CALL PROG3 ( N, A, C, B, FF, YY, ITASK )
          DO I = 1, N
            V(I,K) = YY(I)
          END DO
        END DO
C
C     SOLUTION OF THE CONJUGATE PROBLEM
C
C     INITIAL CONDITION
C

        T = TMAX
        DO I = 1, N
          W(I,M) = 0.D0
        END DO
C
C     NEXT TIME LAYER
C
        DO K = M-1, 1, -1
          T = T - TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
          DO I = 2, N
            X1 = (X(I) + X(I-1)) / 2

            X2 = (X(I+1) + X(I)) / 2
            A(I) = AK(X1) / (2*H*H)
            B(I) = AK(X2) / (2*H*H)
            C(I) = A(I) + B(I) + 1.D0 / TAU
            FF(I) = A(I) * W(I-1,K+1)
     +            + (1.D0 / TAU - A(I) - B(I) ) * W(I,K+1)
     +            + B(I) * W(I+1,K+1)
     +            + (V(I,K) + V(I,K+1)) / 2
          END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
```

```
        B(1)  = 0.D0
        C(1)  = 1.D0
        FF(1) = 0.D0
        A(N)  = 0.D0
        C(N)  = 1.D0
        FF(N) = 0.D0
C
C     SOLUTION OF THE PROBLEM AT THE NEXT TIME LAYER
C

        ITASK = 1
        CALL PROG3 ( N, A, C, B, FF, YY, ITASK )
        DO I = 1, N
          W(I,K) = YY(I)

        END DO
      END DO
C
C     DISCREPANCY
C
      DO K = 1, M
        DO I = 1, N
          RK(I,K) = W(I,K) - GU(I,K)
        END DO
      END DO
C
C     ITERATION PARAMETER
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        W(I,1)  = 0.D0
      END DO

C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS

C
        DO I = 2, N
          X1 = (X(I) + X(I-1)) / 2
          X2 = (X(I+1) + X(I)) / 2
          A(I) = AK(X1) / (2*H*H)
          B(I) = AK(X2) / (2*H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          FF(I) = A(I) * W(I-1,K-1)
     +          + (1.D0 / TAU - A(I) - B(I) ) * W(I,K-1)
     +          + B(I) * W(I+1,K-1)
     +          + (RK(I,K) + RK(I,K-1)) / 2
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS

C
```

```
        B(1)  = 0.D0
        C(1)  = 1.D0
        FF(1) = 0.D0
        A(N)  = 0.D0
        C(N)  = 1.D0
        FF(N) = 0.D0
C
C     SOLUTION OF THE PROBLEM AT THE NEXT TIME LAYER
C
        ITASK = 1

        CALL PROG3 ( N, A, C, B, FF, YY, ITASK )
        DO I = 1, N
          W(I,K) = YY(I)
        END DO
      END DO
C
C     QUICKEST DESCEND METHOD
C
      SUM1 = 0.D0
      SUM2 = 0.D0

      DO K = 1, M
        DO I = 1, N
          SUM1 = SUM1 + RK(I,K)*RK(I,K)
          SUM2 = SUM2 + W(I,K)*W(I,K)
        END DO
      END DO
      TAUK = SUM1/SUM2
C
C     NEXT APPROXIMATION
C
      DO K = 1, M
        DO I = 1, N
          FA(I,K) = FA(I,K) - TAUK *RK(I,K)
        END DO
      END DO
C
C     EXIT FROM THE ITERATION CYCLE
C
      SUM = 0.D0
      DO K = 1, M
        DO I = 1, N
          SUM = SUM + (V(I,K) - YD(I,K))**2
        END DO

      END DO
      SL2 = DSQRT(SUM*TAU*H)

      WRITE ( 01,* ) IT, TAUK, SL2
      IF (SL2.GT.DELTA .AND. IT.LT.ITMAX) GO TO 100
C
C     EXACT SOLUTION
C
      SUM = 0.D0
      DO K = 1, M
        T = (K-1)*TAU
        DO I = 1, N
          F(I,K) = AF(X(I),T)
          SUM = SUM + (FA(I,K) - F(I,K))**2
```

```
        END DO
      END DO
      STL2 = DSQRT(SUM*TAU*H)
      WRITE ( 01,* ) ((FA(I,K), I = 1,N), K = 1,M)
      WRITE ( 01,* ) ((F(I,K), I = 1,N), K = 1,M)
      WRITE ( 01,* ) STL2
      CLOSE ( 01 )
      STOP

      END


      DOUBLE PRECISION FUNCTION  AK ( X )
      IMPLICIT REAL*8 ( A-H, O-Z )
C

C      COEFFICIENT AT THE HIGHER DERIVATIVES
C
      AK = 1.D0
C
      RETURN
      END

      DOUBLE PRECISION FUNCTION  AF ( X, T )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      RIGHT-HAND SIDE
C
      AF = 10.*T*(1.D0 - T)*X*(1.D0-X)

C
      RETURN
      END
```

The equation coefficients, dependent on the spatial variable, and the right-hand side are to be set in the subroutine-functions `AK and AF`.

6.2.5 Computational experiments

The calculations were carried out on a uniform grid with the number of modes $N = 100$, $N_0 = 100$, the calculation domain being a unit square ($l = 1$, $T = 1$). The inverse problem was solved within the framework of a quasi-real experiment in the case of

$$k(x) = 1, \qquad f(x,t) = 10t(1-t)x(1-x).$$

The solution of the problem at the inaccuracy level $\delta = 0.001$ is shown in Figure 6.5 (the total number of iterations is 3). Shown are the level curves for the exact (dashed lines) and approximate solutions drawn with the step $\Delta = 0.1$. The effect of the input-data inaccuracy on the right-hand side reconstruction accuracy is illustrated by Figure 6.6, 6.7 that show data obtained with greater and lesser input-data inaccuracy. For the problem to be solved with $\delta = 0.0001$, a total of 23 iterations were required.

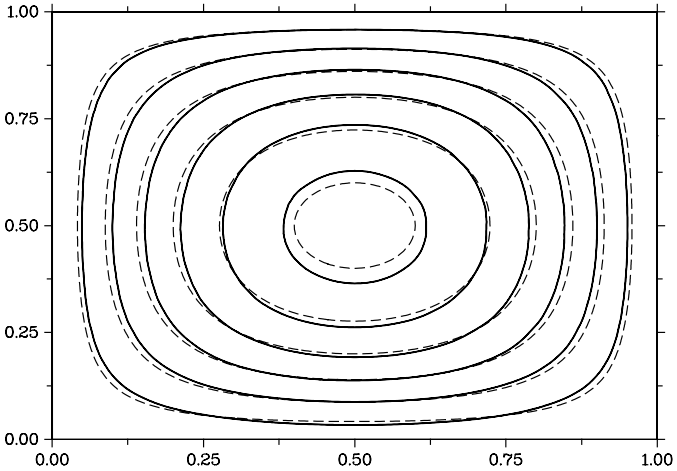

Figure 6.5 Solution of the identification problem obtained with $\delta = 0.001$

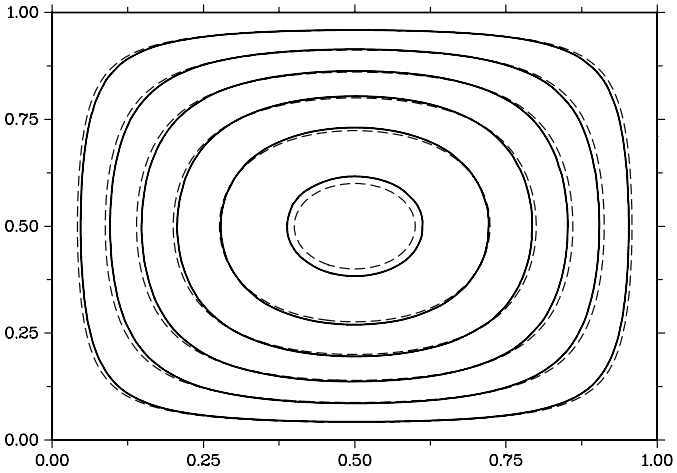

Figure 6.6 Solution of the identification problem obtained with $\delta = 0.0001$

Of course, the identification accuracy essentially depends on the exact solution. In particular, noted above was the necessity to narrow the class of sought right-hand sides due to homogeneous conditions given on some part of the calculation-domain boundary. Figure 6.8 shows data obtained, with $\delta = 0.001$, for the problem with the right-hand side

$$f(x,t) = 2x(1-x).$$

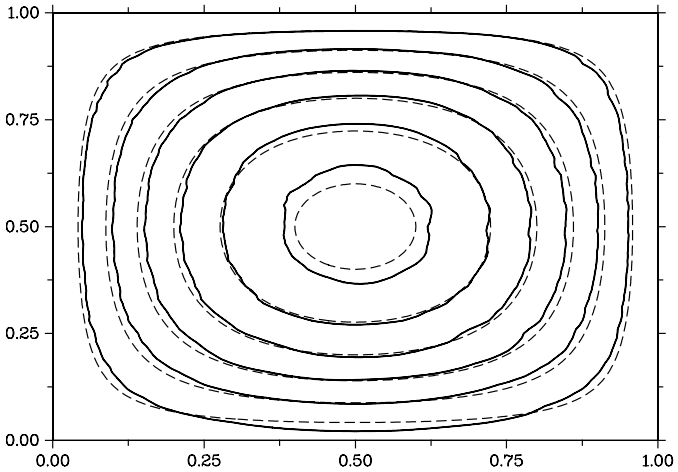

Figure 6.7 Solution of the identification problem obtained with $\delta = 0.01$

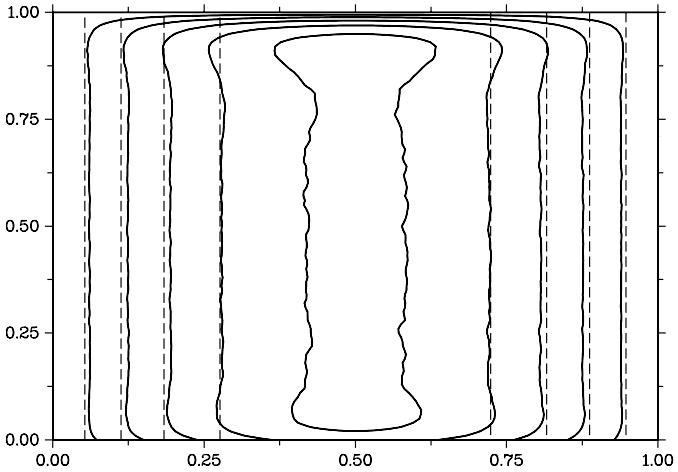

Figure 6.8 Solution of the identification problem obtained with $\delta = 0.001$

6.3 Reconstruction of the time-dependent right-hand side

In the theory of inverse heat-transfer problems, also problems are considered in which it is required to reconstruct unknown heat sources from additional temperature measurements made at some single points. In a similar manner, some important applied problems are formulated that arise in hydrogeology. The solution can be made unique by narrowing the class of admissible right-hand sides of the parabolic equations. In many cases, one can assume the time-dependent right-hand side to be an unknown quantity.

6.3.1 Inverse problem

In this section, we construct a computational algorithm for approximate solution of a simplest one-dimensional (over space) inverse problem in which it is required to reconstruct the time-dependent right-hand side of a parabolic equation from the known spatial distribution. Such a linear inverse problem belongs to the class of classically well-posed mathematical physics problems under some special assumptions concerning the points in space where additional measurements were performed, namely, under the condition that the source acts at the observation points.

We assume the state of the system to be defined by the equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) + f(x,t), \quad 0 < x < l, \quad 0 < t \le T \tag{6.85}$$

with a sufficiently smooth positive coefficient $k(x)$. We restrict ourselves to the problem with the simplest first-kind homogeneous boundary conditions:

$$u(0,t) = 0, \qquad u(l,t) = 0, \qquad 0 \le t \le T. \tag{6.86}$$

Also given is the initial condition:

$$u(x,0) = u_0(x), \qquad 0 < x < l. \tag{6.87}$$

The direct problem is formulated in the form (6.85)–(6.87).

We consider the inverse problem in which, apart from $u(x,t)$, also unknown is the right-hand side $f(x,t)$ of equation (6.85). We assume that the function $f(x,t)$ can be represented as

$$f(x,t) = \eta(t)\psi(x), \tag{6.88}$$

where the function $\psi(x)$ is given, and unknown is the time-dependent source, the function $\eta(t)$ in the representation (6.88). This dependence can be reconstructed from additional observations of $u(x,t)$ made at some internal point $0 < x^* < l$:

$$u(x^*,t) = \varphi(t). \tag{6.89}$$

We arrive at a simple right-hand side identification problem for the parabolic equation (6.85)–(6.89).

6.3.2 Boundary value problem for the loaded equation

We consider the solution of the identification problem under the following constraints:

1) $\psi(x^*) \neq 0$;
2) $\psi(x)$ is a sufficiently smooth function ($\psi \in C^2[0,1]$);
3) $\psi(x) = 0$ on the boundary of the calculation domain.

Particular attention should be given to the first assumption stating that the source to be reconstructed acts at the observation point x^*. It is this assumption that makes the identification problem under consideration a well-posed problem, providing for continuous dependence of the solution on initial data, right-hand side, and data measured at the internal point.

The last assumption is of no fundamental significance; it is used just to simplify the consideration.

We seek the solution of the inverse problem in the form

$$u(x,t) = \theta(t)\psi(x) + w(x,t), \tag{6.90}$$

where

$$\theta(t) = \int_0^t \eta(s)\,ds. \tag{6.91}$$

Substitution of (6.90), (6.91) into (6.85), (6.88) yields the following equation for $w(x,t)$:

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial w}{\partial x}\Big) + \theta(t)\,\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial \psi}{\partial x}\Big). \tag{6.92}$$

With representation (6.90), the condition (6.89) yields the following representation for the unknown function $\theta(t)$:

$$\theta(t) = \frac{1}{\psi(x^*)}\,(\varphi(t) - w(x^*,t)). \tag{6.93}$$

Substitution of (6.93) into (6.92) yields the sought loaded parabolic equation

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial w}{\partial x}\Big) + \frac{1}{\psi(x^*)}\,(\varphi(t) - w(x^*,t))\,\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial \psi}{\partial x}\Big). \tag{6.94}$$

Under the adopted assumptions about the right-hand side, the boundary condition at the domain boundary is

$$w(0,t) = 0, \qquad w(l,t) = 0, \qquad 0 \le t \le T. \tag{6.95}$$

With allowance for (6.91), for the auxiliary function $\theta(t)$ we have:

$$\theta(0) = 0. \tag{6.96}$$

This allows us to use the initial condition

$$w(x,0) = u_0(x), \qquad 0 < x < l. \tag{6.97}$$

In this way, the inverse problem (6.85)–(6.89) can be formulated as a boundary value problem for the loaded equation (6.94)–(6.97) with the representation (6.91), (6.93) for the unknown time-dependent source.

6.3.3 Difference scheme

Note major points in the numerical solution of the identification problem. The computational algorithm is based on the approximate solution of the initial boundary problem for the loaded equation. Today, the numerical solution methods for such non-classical mathematical physics problems still remain poorly developed.

We assume a uniform grid $\bar{\omega}$ with a grid size h to be introduced over the coordinate x. We denote the grid nodes as $x_i = ih$, $i = 0, 1, \ldots, N$, $Nh = 1$, and let $v = v_i = v(x_i)$. For simplicity, we assume that the observation point $x = x^*$ coincides with an internal node $i = k$.

We pass from one time layer $t_n = n\tau$, $n = 0, 1, \ldots, N_0$, $N_0\tau = T$, $\tau > 0$ to the next time layer t_{n+1} using a purely implicit difference scheme for equation (6.94). At the internal grid nodes over space we have:

$$\frac{w^{n+1} - w^n}{\tau} = (aw_{\bar{x}}^{n+1})_x + \frac{1}{\psi_k}(\varphi^{n+1} - w_k^{n+1})(a\psi_{\bar{x}})_x. \tag{6.98}$$

In the case of problems with a sufficiently smooth coefficient $k(x)$ we put, for instance, $a_i = k(0.5(x_i + x_{i-1}))$. We approximate the conditions (6.95) and (6.97); then, we have:

$$w_0^{n+1} = 0, \qquad w_N^{n+1} = 0, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{6.99}$$

$$w_i^0 = u_0(x_i), \qquad i = 1, 2, \ldots, N - 1. \tag{6.100}$$

In accordance with (6.93), from the solution of the difference problem (6.98)–(6.100) we determine

$$\theta^{n+1} = \frac{1}{\psi_k}(\varphi^{n+1} - w_k^{n+1}), \qquad n = 0, 1, \ldots, N_0 - 1 \tag{6.101}$$

and, then, supplement these relations with the condition (see (6.96)) $\theta^0 = 0$. With (6.91), for the sought time dependence of the right-hand side we use the simplest numerical-differentiation procedure:

$$\eta^{n+1} = \frac{\theta^{n+1} - \theta^n}{\tau}, \qquad n = 0, 1, \ldots, N_0 - 1. \tag{6.102}$$

To realize the implicit scheme, it is necessary to dwell here on the matter of solution of the difference problem.

6.3.4 Non-local difference problem and program realization

In spite of non-locality of the difference problem on the next time layer, we encounter no substantial problem in the computational realization of scheme (6.98)–(6.100). We write equation (6.98) at internal nodes in the form

$$\frac{w_i^{n+1}}{\tau} - (aw_{\bar{x}}^{n+1})_{x,i} + \frac{1}{\psi_k}(a\psi_{\bar{x}})_{x,i} w_k^{n+1} = g_i^n \tag{6.103}$$

with given right-hand side g_i^n and given boundary conditions (6.99). We seek the solution of system (6.99), (6.103) in the form

$$w_i^{n+1} = y_i + w_k^{n+1} z_i, \qquad i = 0, 1, \ldots, N. \tag{6.104}$$

Substitution of (6.104) into (6.103) allows us to formulate the following difference problems for the auxiliary functions y_i and z_i:

$$\frac{y_i}{\tau} - (a y_{\bar{x}})_{x,i} = g_i^n, \qquad i = 1, 2, \ldots, N-1, \tag{6.105}$$

$$y_0 = 0, \qquad y_N = 0, \tag{6.106}$$

$$\frac{z_i}{\tau} - (a z_{\bar{x}})_{x,i} + \frac{1}{\psi_k} (a \psi_{\bar{x}})_{x,i} = 0, \qquad i = 1, 2, \ldots, N-1, \tag{6.107}$$

$$z_0 = 0, \qquad z_N = 0. \tag{6.108}$$

Afterwards, with (6.104) we can find w_k^{n+1}:

$$w_k^{n+1} = \frac{y_k}{1 - z_k}. \tag{6.109}$$

Correctness of the algorithm is guaranteed by the fact that the denominator in (6.109) is never zero. For the difference problem (6.107), (6.108) the following a priori estimate can be established based on the maximum principle for difference schemes:

$$\max_{0 \le i \le N} |z_i| \;\le\; \tau \max_{0 < i < N} \left| \frac{1}{\psi_k} (a \psi_{\bar{x}})_{x,i} \right|.$$

As a result, for a sufficiently small $\tau = \mathcal{O}(1)$ we have: $|z_i| < 1$, i. e., the time step size must be sufficiently small.

The difference problems (6.105), (6.106) and (6.107), (6.108) are standard problems, their numerical solution presenting no difficulties. In the one-dimensional case under consideration, the usual three-point sweep algorithm can be employed.

In fact, the computational difficulty of the used computational algorithm is equivalent to double solution of the direct problem. For this reason, the considered method can be classified to optimal methods. Below, we give the text of a program that numerically solves the inverse problem of interest.

Program PROBLEM7

```
C
C     PROBLEM7  - RIGHT-HAND SIDE IDENTIFICATION
C                 ONE-DIMENSIONAL NON-STATIONARY PROBLEM
C                 UNKNOWN TIME DEPENDENCE
C
      IMPLICIT REAL*8 ( A-H, O-Z )
C
      PARAMETER ( DELTA = 0.0005D0, N = 101, M = 101 )
```

```
      DIMENSION VT(M), VY(M), VP(M), VV(M)
     +         ,A(N), B(N), C(N), F(N), Y(N), VB(N), FB(N), W(N), Z(N)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF THE SEGMENT;
C     XD        - OBSERVATION POINT;
C     N         - NUMBER OF GRID NODES OVER THE SPATIAL VARIABLE;
C     TMAX      - MAXIMAL TIME;
C     M         - NUMBER OF GRID NODES OVER TIME;
C     DELTA     - INPUT-DATA INACCURACY;
C     VY(M)     - EXACT SOLUTION AT THE OBSERVATION POINT;
C     VP(M)     - DISTURBED SOLUTION AT THE OBSERVATION POINT;
C     VT(M)     - EXACT TIME DEPENDENCE OF THE RIGHT-HAND SIDE;
C     VB(N))    - DEPENDENCE OF THE RIGHT-HAND SIDE ON THE SPATIAL
C                 VARIABLE;
C     VV(M)     - CALCULATED RIGHT-HAND SIDE VERSUS TIME;
C
      XL   = 0.D0
      XR   = 1.D0
      XD   = 0.3D0
      TMAX = 1.D0
C
      OPEN (01, FILE = 'RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID

C
      H = (XR - XL) / (N - 1)
      TAU = TMAX / (M-1)
      ND   = 1 + (XD + 0.5D0*H) / H
C
C     DIRECT PROBLEM
C
C     SOURCE
C
      DO K = 1, M
        T = (K-0.5D0)* TAU
        VT(K) = (K-0.5D0)* TAU
        IF (T.GE.0.6D0) VT(K) = 0.D0
      END DO

C
C     DEPENDENCE OF THE RIGHT-HAND SIDE ON THE SPATIAL VARIABLE
C
      DO I = 1, N
        VB(I) = DSIN(3.1415926*(I-1)*H)
      END DO
C

C     SOLUTION OF THE PROBLEM
C
C     INITIAL CONDITION
C
      T = 0.D0

      DO I = 1, N
        Y(I)  = 0.D0
      END DO
      VY(1) = Y(ND)
```

```
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
          B(I) = 1.D0 / (H*H)
          C(I) = 2.D0 / (H*H) + 1.D0 / TAU
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C

        B(1) = 0.D0
        C(1) = 1.D0
        F(1) = 0.D0
        A(N) = 0.D0
        C(N) = 1.D0
        F(N) = 0.D0
C
C     RIGHT-HAND SIDE OF THE DIFFERENCE EQUATION
C
        DO I = 2, N-1
          F(I) = VB(I)*VT(K) + Y(I) / TAU
        END DO
C
C     SOLUTION AT THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, F, Y, ITASK )
        VY(K) = Y(ND)
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO K = 1,M
        VP(K) = VY(K) + 2.D0*DELTA*(RAND(0)-0.5D0)
      END DO
C
C     INVERSE PROBLEM
C
C     AUXILIARY FUNCTION
C
      DO I = 2, N-1
        FB(I) = ( 1.D0 / VB(ND) )
     +          * (VB(I+1)-2.D0*VB(I) + VB(I-1)) / (H*H)
      END DO
C
C     SOLUTION OF THE PROBLEM
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I)  = 0.D0
      END DO
```

```
C
C     NEXT TIME LAYER
C

      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
          B(I) = 1.D0 / (H*H)
          C(I) = 2.D0 / (H*H) + 1.D0 / TAU
        END DO

C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
        B(1) = 0.D0
        C(1) = 1.D0
        F(1) = 0.D0
        A(N) = 0.D0
        C(N) = 1.D0
        F(N) = 0.D0
C
C     RIGHT-HAND SIDE OF THE DIFFERENCE EQUATION
C
        DO I = 2, N-1
          F(I) = FB(I)*VY(K) + Y(I) / TAU
        END DO
C
C     SOLUTION OF THE FIRST SUBPROBLEM AT THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, F, W, ITASK )
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
          B(I) = 1.D0 / (H*H)
          C(I) = 2.D0 / (H*H) + 1.D0 / TAU
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
        B(1) = 0.D0
        C(1) = 1.D0
        F(1) = 0.D0
        A(N) = 0.D0
        C(N) = 1.D0
        F(N) = 0.D0
C
C     RIGHT-HAND SIDE OF THE DIFFERENCE EQUATION
C

        DO I = 2, N-1
          F(I) = - FB(I)
        END DO
C
```

```
C     SOLUTION OF THE SECOND SUBPROBLEM AT THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, F, Z, ITASK )
C
        VV(K) = (VP(K)-VP(K-1)
     +        - (W(ND) / (1 - Z(ND)) - Y(ND) )) / (TAU * VB(ND))
        DO I = 1, N

          Y(I) = W(I) + Z(I) * W(ND) / (1 - Z(ND))
        END DO
      END DO
C
C     APPROXIMATE SOLUTION
C
      WRITE ( 01,* ) (VV(K), K = 2,M)
      WRITE ( 01,* ) (VT(K), K = 2,M)
      CLOSE (01)
      STOP
      END
```

6.3.5 Computational experiments

Below, results of calculations are presented which were carried out for the simplest model inverse problem (6.84)–(6.89). Within the concept of quasi-real experiment, we consider the direct problem (6.84)–(6.87) with some given right-hand side. For the equation coefficient and for the initial condition we put:

$$k(x) = 1, \qquad u_0(x) = 0, \qquad 0 \le x \le 1.$$

The right-hand side is defined as

$$\psi(x) = \sin(\pi x), \quad 0 \le x \le 1,$$

$$\eta(t) = \begin{cases} t, & 0 < t < 0.6, \\ 0, & 0.6 < t < T = 1. \end{cases}$$

This problem was solved numerically on a grid with $N = 100$, $N_0 = 100$.

Below, we give data obtained by reconstructing the right-hand side from observations performed at the point $x^* = 0.3$. The calculated data were used to specify the mesh function φ^n.

In the solution of the inverse problem, the mesh function $\varphi(t)$ was perturbed using random inaccuracies. We put

$$\varphi_\delta^n = \varphi^n + 2\delta(\sigma^n - 1/2),$$

where σ^n is a random function normally distributed over the interval $[0, 1]$. The quantity δ defines the inaccuracy level. Figure 6.9 presents, as a function of time, the exact right-hand side and the right-hand side reconstructed by the above algorithm at the inaccuracy level $\delta = 0.001$.

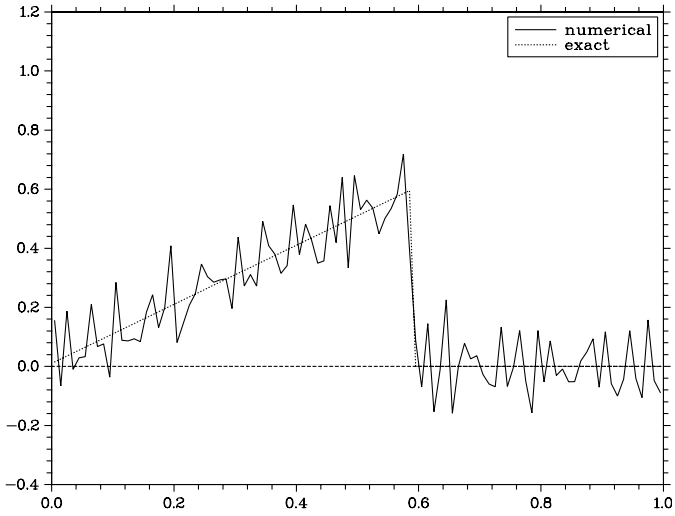

Figure 6.9 Reconstructed right-hand side versus time at $\delta = 0.001$

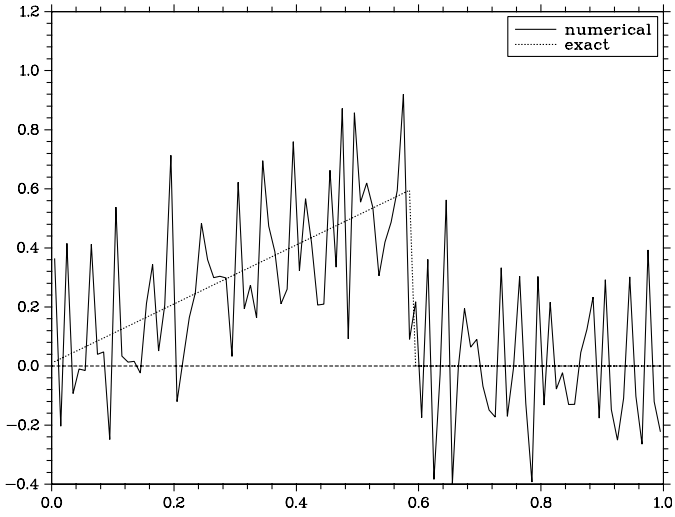

Figure 6.10 Solution of the inverse problem obtained with $\delta = 0.0025$

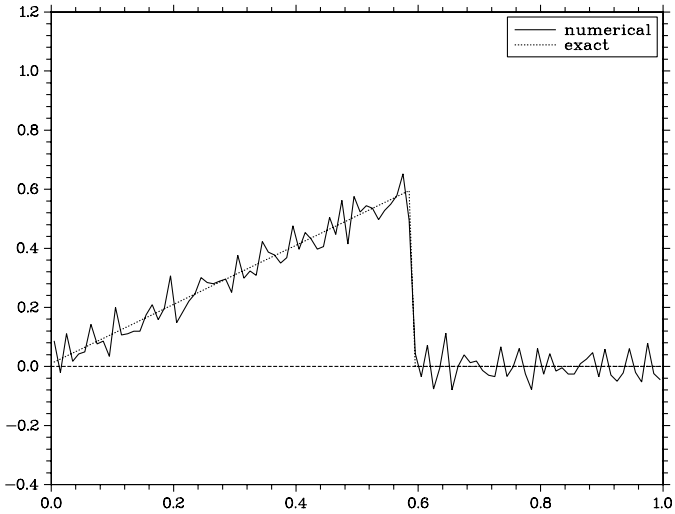

Figure 6.11 Solution of the inverse problem obtained with $\delta = 0.0005$

Correctness of the considered inverse problem is illustrated by calculation data obtained at various inaccuracy levels. Figures 6.10 and 6.11 show the solutions of the problem obtained with $\delta = 0.0025$ and $\delta = 0.0005$, respectively. The lower is the inaccuracy level, the more accurately can the solution be reconstructed.

The considered computational algorithm for solving the inverse problem can be used to solve more general problems. In particular, multi-dimensional problems, problems with many observation points, etc. can be treated much in the same manner. Fundamental difficulties (ill-posedness) arise when we pass to problems with localized sources, with abandoned assumption $\psi(x^*) \neq 0$.

6.4 Identification of a time-independent right-hand side of a parabolic equation

Below, we consider an inverse problem in which it is required to reconstruct the time-independent right-hand side of a parabolic equation. Additional measurements are assumed to be performed at the end time.

6.4.1 Statement of the problem

Consider a process governed by a second-order one-dimensional parabolic equation. We assume that the dynamics of interest is defined by a time-independent, spatially distributed source, so that

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) + f(x), \qquad 0 < x < l, \quad 0 < t \le T. \tag{6.110}$$

We supplement this equation with first-kind homogeneous boundary conditions:

$$u(0,t) = 0, \qquad u(l,t) = 0, \qquad 0 \le t \le T. \tag{6.111}$$

The initial state is defined by the condition

$$u(x,0) = 0, \qquad 0 < x < l. \tag{6.112}$$

With given coefficient $k(x)$ and right-hand side $f(x)$, relations (6.110)–(6.112) define the direct problem.

Consider an inverse problem in which the unknown quantity is the right-hand side $f(x)$ of equation (6.110). We assume that the function $f(x)$ can be reconstructed from the known end-time solution; i.e., this function can be represented as

$$u(x,T) = \varphi(x), \qquad 0 < x < l. \tag{6.113}$$

To begin with, obtain an a priori estimate for the solution $u(x,t)$ of the problem that proves the solution to be stable with respect to weak perturbations of $\varphi(x)$.

6.4.2 Estimate of stability

The simplest approach to the inverse problem (6.110)–(6.113) consists in elimination of the unknown function $\varphi(x)$. To this end, we differentiate the equation (6.110) with respect to time:

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial^2 u}{\partial x\,\partial t}\Big), \qquad 0 < x < l, \quad 0 < t < T. \tag{6.114}$$

For the latter equation, two boundary conditions, (6.112) and (6.113), are given for the variable t.

For the problem (6.111)–(6.114), we use the operator notation that can be used to study more general problems. We introduce a Hilbert space $\mathcal{L}_2(\Omega)$ with the following scalar product and norm:

$$(v,w) = \int_\Omega v(x)w(x)\,dx, \qquad \|v\|^2 = (v,v) = \int_\Omega v^2(x)\,dx.$$

For the functions $v(x,t)$, $w(x,t)$ from $\mathcal{H} = \mathcal{L}_2(Q_T)$, we put

$$(v,w)^* = \int_0^T (v,w)\,dt = \int_0^T \int_\Omega v(x)w(x)\,dx\,dt, \quad \|v\|^* = ((v,v)^*)^{1/2}.$$

On the set of functions satisfying the conditions (6.111), we define the operator

$$\mathcal{A}u = -\frac{d}{dx}\Big(k(x)\,\frac{du}{dx}\Big).$$

Among of the main properties of $\mathcal{A}$ in $\mathcal{L}_2(\Omega)$, the following property deserves mention:

$$\mathcal{A} = \mathcal{A}^* \geq mE, \qquad m > 0.$$

Using the settings introduced, reformulate the problem (6.111)–(6.114) as the following boundary value problem:

$$\frac{d^2u}{dt^2} + \mathcal{A}\frac{du}{dt} = 0, \qquad 0 < t < T, \tag{6.115}$$

$$u(0) = 0, \qquad u(T) = \varphi. \tag{6.116}$$

We pass from (6.115), (6.116) to a problem with homogeneous boundary conditions. To do so, we represent the solution as

$$u(t) = w(t) + \frac{t}{T}\varphi. \tag{6.117}$$

For the function $w(t)$, we obtain the problem

$$\frac{d^2w}{dt^2} + \mathcal{A}\frac{dw}{dt} = -\psi, \qquad 0 < t < T, \tag{6.118}$$

$$w(0) = 0, \qquad w(T) = 0, \tag{6.119}$$

where

$$\psi = \frac{1}{T}\mathcal{A}\varphi.$$

To derive a simplest a priori estimate for the solution of problem (6.118), (6.119), we scalarwise multiply equation (6.118) in $\mathcal{H}$ by w:

$$\Big(\frac{d^2w}{dt^2}, w\Big)^* + \Big(\mathcal{A}\frac{dw}{dt}, w\Big)^* = -(\psi, w)^*. \tag{6.120}$$

Consideration of the constancy of $\mathcal{A}$ and the homogeneous boundary conditions (6.119) gives

$$\Big(\mathcal{A}\frac{dw}{dt}, w\Big)^* = \frac{1}{2}\int_0^T \frac{d}{dt}(\mathcal{A}w, w) = \frac{1}{2}(\mathcal{A}w, w)\Big|_0^T = 0.$$

With (6.119) taken into account, for the first term in (6.120) we obtain

$$\Big(\frac{d^2w}{dt^2}, w\Big)^* = -\Big(\frac{dw}{dt}, \frac{dw}{dt}\Big) = -\Big(\int_0^T \Big(\frac{dw}{dt}\Big)^2 dt, 1\Big).$$

Then, substitution into (6.20) yields:

$$\Big(\int_0^T \Big(\frac{dw}{dt}\Big)^2 dt, 1\Big) = (\psi, w)^*. \tag{6.121}$$

The left-hand side of (6.121) can be estimated by invoking the Friedrichs inequality

$$\int_0^T v^2\,dt \le \mathcal{M}_0 \int_0^T \Big(\frac{dw}{dt}\Big)^2 dt.$$

By virtue of this, we have

$$\Big(\int_0^T \Big(\frac{dw}{dt}\Big)^2 dt, 1\Big) \ge \mathcal{M}_0^{-1}(\|w\|^*)^2.$$

For the right-hand side of (6.121), we use the estimate

$$(\psi, w)^* \le \|\psi\|^* \|w\|^*.$$

This allows us to obtain from (6.121) the desired inequality

$$\|w\|^* \le \mathcal{M}_0 \|\psi\|^*, \tag{6.122}$$

which shows that the solution of problem (6.118), (6.119) is stable with respect to the right-hand side.

With allowance for (6.122) and (6.117), for the solution of problem (6.115), (6.116) we obtain

$$\|u\|^* \le \|\varphi\|^* + \frac{\mathcal{M}_0}{T}\,\|\mathcal{A}\varphi\|^*. \tag{6.123}$$

Thus, in the inverse problem (6.115), (6.116) the solution $u(t)$ continuously depends on the conditions at $t = T$.

With the function $u(t)$ found from the solution of the well-posed problem (6.115), (6.116), the sought right-hand side f is given by

$$f = \frac{du}{dt} + \mathcal{A}u. \tag{6.124}$$

A complete study of well-posedness of the inverse problem for the pair of functions $\{u, f\}$ poses the question how the functions u and f depend on the input data (on the function φ). Above, we have restricted ourselves to obtaining the simplest estimate (6.123) only for u.

6.4.3 Difference problem

The difference problem (6.110)–(6.113) reduces to the non-classical boundary value problem (6.111)–(6.114). It makes sense to consider specific features of the computational algorithm, with special emphasis placed to the matter of time discretization. For the mesh functions we use the same notation as for the continuous-argument functions.

For spatial discretization, we use a uniform grid $\bar{\omega}$ with a grid size h over the interval $\bar{\Omega} = [0, l]$:

$$\bar{\omega} = \{x \mid x = x_i = ih,\ i = 0, 1, \dots, N,\ Nh = l\},$$

where, as usual, ω is the set of internal nodes and $\partial\omega$ is the set of boundary nodes. In the mesh Hilbert space $L_2(\omega)$, we introduce the norm via the relation $\|y\| = (y, y)^{1/2}$, where

$$(y, w) = \sum_{x\in\omega} y(x)w(x)h.$$

At internal nodes, we approximate the differential operator $\mathcal{A}$ with second order, with the difference operator

$$Ay = -(ay_{\bar{x}})_x, \quad x \in \omega, \tag{6.125}$$

where, for instance, $a(x) = k(x - 0.5h)$.

On the set of functions vanishing on $\partial\omega$ (see (6.111)) for the self-adjoint operator A under the constraints $k(x) \geq \kappa > 0$ and $q(x) \geq 0$ there holds the estimate

$$A = A^* \geq \kappa\lambda_0 E \tag{6.126}$$

with a constant

$$\lambda_0 = \frac{4}{h^2}\,\sin^2\frac{\pi h}{2l} \geq \frac{8}{l^2}.$$

Over time, we use the uniform grid

$$\bar{\omega}_\tau = \omega_\tau \cup \{T\} = \{t_n = n\tau,\ n = 0, 1, \ldots, N_0,\ \tau N_0 = T\}$$

and let $y^n = y(t_n)$.

To approximate equation (6.114) with second order over time and space, we can naturally use the difference equation

$$\begin{gathered}\frac{u^{n+1} - 2u^n + u^{n-1}}{\tau^2} + A\,\frac{u^{n+1} - u^{n-1}}{2\tau} = 0,\\ x \in \omega, \qquad n = 1, 2, \ldots, N_0 - 1.\end{gathered} \tag{6.127}$$

This equation is supplemented with the boundary conditions (see (6.112), (6.113))

$$u^0 = 0, \qquad u^{N_0} = \varphi, \qquad x \in \omega. \tag{6.128}$$

The difference problem (6.127), (6.128) can be examined analogously to the differential case. We represent (see (6.117)) the difference solution as

$$u^n = w^n + \frac{t_n}{T}\,\varphi, \qquad n = 0, 1, \ldots, N_0. \tag{6.129}$$

The analogue of (6.118), (6.119) is the problem

$$\begin{gathered}\frac{w^{n+1} - 2w^n + w^{n-1}}{\tau^2} + A\,\frac{w^{n+1} - w^{n-1}}{2\tau} = -\psi,\\ x \in \omega, \qquad n = 1, 2, \ldots, N_0 - 1,\end{gathered} \tag{6.130}$$

$$w^0 = 0, \qquad w^{N_0} = 0, \qquad x \in \omega, \tag{6.131}$$

in which

$$\psi = \frac{1}{T} A\varphi.$$

For two-dimensional mesh functions, we define a Hilbert space $H = L_2(Q_T)$, in which the scalar product and the norm are

$$(v, w)^* = \sum_{n=1}^{N_0-1} (v^n, w^n)\tau, \qquad \|v\|^* = \sqrt{(v, v)^*}.$$

We scalarwise multiply equation (6.130) in $H = L_2(Q_T)$ by w; this yields:

$$\left(w_{\bar{t}t}, w\right)^* + \left(Aw_{\overset{\circ}{t}}, w\right)^* = -(\psi, w)^*. \tag{6.132}$$

Here, the following standard settings adopted in the theory of difference schemes are used:

$$w_{\bar{t}} = \frac{w^n - w^{n-1}}{\tau}, \qquad w_t = \frac{w^{n+1} - w^n}{\tau},$$
$$w_{\bar{t}t} = \frac{w^{n+1} - 2w^n + w^{n-1}}{\tau^2}, \qquad w_{\overset{\circ}{t}} = \frac{w^{n+1} - w^{n-1}}{2\tau}.$$

Further consideration is based on the directly verifiable property of skew symmetry of the operator of central difference derivative, so that

$$\left(Aw_{\overset{\circ}{t}}, w\right)^* = 0.$$

Analogously to (6.121), from (6.132) we obtain

$$\left(\sum_{n=1}^{N_0} (w_{\bar{t}})^2\tau, 1\right) = (\psi, w)^*. \tag{6.133}$$

With the difference Friedrichs inequality taken into account, we obtain

$$\sum_{n=1}^{N_0-1} w^2\tau \le M_0 \sum_{n=1}^{N_0} (w_{\bar{t}})^2\tau,$$

where

$$M_0 = T^2/8.$$

In view of

$$(\psi, w)^* \le \|\psi\|^* \|w\|^*,$$

from (6.133) we obtain the estimate

$$\|w\|^* \le M_0 \|\psi\|^*. \tag{6.134}$$

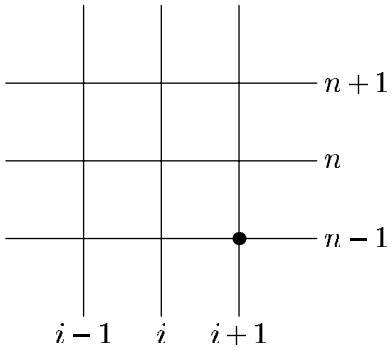

Figure 6.12 Mesh pattern of the difference problem

Then, representation (6.129) and inequality (6.134) yield the desired estimate

$$\|u\|^* \le \|\varphi\|^* + \frac{M_0}{T}\,\|A\varphi\|^*. \tag{6.135}$$

Estimate (6.135) is fully consistent with the a priori estimate (6.123) for the solution of the differential problem.

The right-hand side can be found from the solution of problem (6.127), (6.128) at half-integer nodes using the expression

$$f_{n+1/2} = \frac{u^{n+1} - u^n}{\tau} + A\,\frac{u^{n+1} + u^n}{2}, \tag{6.136}$$

matched with the difference approximation (6.127) of equation (6.114).

6.4.4 Solution of the difference problem

The a priori estimate (6.135) obtained previously in the consideration of the problem for the inaccuracy can be used to establish convergence of the difference solution found from (6.127), (6.128) to the solution of the boundary value problem (6.112)–(6.114) with second order over time and space. Some problems arise in the computational realization and in finding the solution of the difference problem (6.127), (6.128). For problems with the constant coefficient $k(x)$, fast algorithms can be constructed around the variable separation method. For more general problems (in particular, for problems with $k(x) \neq \text{const}$), iteration methods are preferable. Figure 6.12 shows the mesh pattern for the difference problem (6.127), (6.128).

Consider a possible approach to solving the difference problem in question. With (6.129), the problem reduces to finding the mesh function w as the solution of (6.130), (6.131). We write this difference problem as

$$\bar{A}w = \psi, \qquad \bar{A} = A_0 + A_1. \tag{6.137}$$

The operators A_0 and A_1 are defined by

$$A_0 w^n = -\frac{w^{n+1} - 2w^n + w^{n-1}}{\tau^2}, \tag{6.138}$$

$$A_1 w^n = -A\frac{w^{n+1} - w^{n-1}}{2\tau}, \quad n = 1, 2, \ldots, N_0 - 1 \tag{6.139}$$

on the set of mesh functions satisfying the conditions (6.131).

A specific feature of the difference problem (6.137) consists in that the operator $\bar{A}$ in this problem is not a self-adjoint operator. This circumstance substantially hampers the construction of efficient iteration methods. As it was shown above, the operator A_1 defined by (6.139) is a skew-symmetric operator $(A_1 w, w) = 0$). Among the main properties of the operator A_0 (see (6.138), the following property deserves mention:

$$A_0 = A_0^* \geq M_0^{-1} E.$$

Thus, the operators A_0 and A_1 are respectively the self-adjoint and skew-symmetric parts of $\bar{A}$:

$$A_0 = \frac{1}{2}(\bar{A} + \bar{A}^*), \qquad A_1 = \frac{1}{2}(\bar{A} - \bar{A}^*).$$

Among the general approaches to the solution of difference problems with non-self-adjoint operators, methods using preliminary symmetrization are worth noting. In this way, one can pass from an initial problem with a non-self-adjoint operator to the problem with a self-adjoint operator. An example here is the Gauss symmetrization, in which case instead of (6.137) to be solved is the equation

$$\bar{A}^* \bar{A} w = \bar{A}^* \psi. \tag{6.140}$$

A second idea widely used in computational practice, in the approximate solution of problems with non-self-adjoint operator (bad problem) is related with the choice, as the iteration-method reconditioner, the self-conjugate part of the problem operator ($B = A_0$) (good problem at each iteration step). In the solution of problem (6.137), we will use an iteration method developed around the mentioned passage to a problem with self-adjoint operator, with the operator A_0 used as the reconditioner.

At the first stage, the initial problem is to be reconditioned as

$$A_0^{-1}(A_0 + A_1) w = A_0^{-1} \psi.$$

At the second stage, the operator $A_0 + A_1^*$, $A_1^* = -A_1$ (compare (6.140)) is used to perform symmetrization:

$$\tilde{A} w = \tilde{f}, \tag{6.141}$$

$$\tilde{A} = (A_0 + A_1^*) A_0^{-1} (A_0 + A_1), \qquad \tilde{f} = (A_0 + A_1^*) A_0^{-1} \psi.$$

The problem (6.141) can be solved using the iterative conjugate method that takes, with the so-chosen reconditioner operator ($B = A_0$), the form

$$\begin{aligned} A_0 w_{k+1} &= \alpha_{k+1}(A_0 - \tau_{k+1}\tilde{A})w_k + (1-\alpha_{k+1})A_0 w_{k-1} + \alpha_{k+1}\tau_{k+1}\tilde{f}, \\ & \qquad k = 1, 2, \ldots, \\ A_0 w_1 &= (A_0 - \tau_1 \tilde{A})w_0 + \tau_1 \tilde{f}. \end{aligned} \tag{6.142}$$

The iteration parameters α_{k+1} and τ_{k+1} can be calculated by the formulas

$$\begin{aligned} \tau_{k+1} &= \frac{(\tilde{w}_k, r_k)}{(\tilde{A}\tilde{w}_k, \tilde{w}_k)}, \qquad k = 0, 1, \ldots, \\ \alpha_{k+1} &= \left(1 - \frac{\tau_{k+1}}{\tau_k}\,\frac{(\tilde{w}_k, r_k)}{(\tilde{w}_{k-1}, r_{k-1})}\,\frac{1}{\alpha_k}\right)^{-1}, \quad k = 1, 2 \ldots, \quad \alpha_1 = 1, \end{aligned} \tag{6.143}$$

where $r_k = \tilde{A}w_k - \tilde{f}$ is the discrepancy and $\tilde{w} = A_0^{-1} r_k$ is the correction.

The iteration method (6.142), (6.143) is embodied in the program PROBLEM8.

Program PROBLEM8

```
C
C     PROBLEM8  - IDENTIFICATION OF TIME-INDEPENDENT RIGHT-HAND SIDE
C                 ONE-DIMENSIONAL NON-STATIONARY PROBLEM
C                 (UNKNOWN SPATIAL DISTRIBUTION)
C
      IMPLICIT REAL*8 ( A-H, O-Z )
C
      PARAMETER ( DELTA = 0.D0, N = 51, M = 51 )
      DIMENSION X(N), Y(N), FT(N), FY(N), PHI(N), PHID(N)
     +         ,A(M), B(M), C(M), F(M)      ! M .GE. N
     +         ,V(N,M), PSI(N), FP(N,M), FR(N,M), WORK(N,M)
     +         ,W(N,M), WOLD(N,M), RK(N,M), ARK(N,M), FTILDE(N,M)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF THE SEGMENT;
C     N         - NUMBER OF GRID NODES OVER THE SPATIAL VARIABLE;
C     TMAX      - MAXIMAL TIME;
C     M         - NUMBER OF GRID NODES OVER TIME;
C     DELTA     - INPUT-DATA INACCURACY;
C     PHI(N)    - EXACT SOLUTION AT THE END TIME;
C     PHID(N)   - DISTURBED END-TIME SOLUTION;
C     FT(N)     - EXACT SPATIAL DEPENDENCE OF THE RIGHT-HAND SIDE;
C     FY(N))    - CALCULATED SPATIAL DEPENDENCE OF THE RIGHT-HAND SIDE;
C     EPS       - REQUIRED RELATIVE ACCURACY IN THE ITERATIVE APPROACH
C                 TO THE SOLUTION.
C
      XL   = 0.D0
      XR   = 1.D0
      TMAX = 1.D0
C
      OPEN (01, FILE = 'RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
```

```
C     GRID
C
      H   = (XR - XL) / (N - 1)
      TAU = TMAX / (M - 1)
C
C     DIRECT PROBLEM
C
C     EXACT RIGHT-HAND SIDE
C

      DO I = 1, N
        X(I)  = (I-1) * H
        FT(I) = AF(X(I))
      END DO
C
C     SOLUTION OF THE DIRECT PROBLEM
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I)  = 0.D0
      END DO
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        DO I = 2, N-1
          X1 = (X(I) + X(I-1)) / 2
          X2 = (X(I+1) + X(I)) / 2
          A(I) = AK(X1) / (2.D0*H*H)
          B(I) = AK(X2) / (2.D0*H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          F(I) = A(I) * Y(I-1)
     +          + (1.D0 / TAU - A(I) - B(I)) * Y(I)
     +          + B(I) * Y(I+1)
     +          + AF(X(I))
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
        B(1) = 0.D0
        C(1) = 1.D0
        F(1) = 0.D0
        A(N) = 0.D0
        C(N) = 1.D0
        F(N) = 0.D0
C

C     SOLUTION OF THE PROBLEM AT THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, F, Y, ITASK )
      END DO
C
C     SOLUTION AT THE END TIME
```

```
C
      DO I = 1, N
        PHI(I) = Y(I)
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO I = 2, N-1
        PHID(I) = PHI(I) + 2.D0*DELTA*(RAND(0)-0.5D0)
      END DO
      PHID(1) = 0.D0
      PHID(N) = 0.D0
C

C     INVERSE PROBLEM
C
      EPS = 1.D-8

C
C     AUXILIARY FUNCTION AND INITIALIZATION
C
      DO I = 1, N
        DO K = 1, M
          V(I,K)      = (K-1)*PHID(I) / (M-1)
          FP(I,K)     = 0.D0
          FR(I,K)     = 0.D0
          RK(I,K)     = 0.D0
          ARK(I,K)    = 0.D0
          WORK(I,K)   = 0.D0
          FTILDE(I,K) = 0.D0
        END DO
      END DO
C
C     RIGHT-HAND SIDE OF THE SYMMETRIZED EQUATION
C
      DO I = 2, N-1
        X1 = (X(I) + X(I-1)) / 2
        X2 = (X(I+1) + X(I)) / 2
        A(I) = AK(X1) / (H*H)
        B(I) = AK(X2) / (H*H)
        PSI(I) = - A(I)*PHID(I-1) + (A(I) + B(I))*PHID(I)
     +           - B(I)*PHID(I+1)
        PSI(I) = PSI(I) / TMAX
      END DO
      DO I = 2, N-1
        DO K = 2, M-1
          FP(I,K) = PSI(I)
        END DO
      END DO
      CALL A0I ( N, M, TAU, FR, FP, A, B, C, F, Y )
      DO I = 2, N-1
        DO K = 2, M-1
        X1 = (X(I) + X(I-1)) / 2
        X2 = (X(I+1) + X(I)) / 2
        AA = AK(X1) / (H*H)
        BB = AK(X2) / (H*H)
        WORK(I,K) = - AA*(FR(I-1,K+1) - FR(I-1,K-1))
     +              + (AA + BB)*(FR(I,K+1) - FR(I,K-1))
     +              - BB*(FR(I+1,K+1) - FR(I+1,K-1))
        WORK(I,K) = WORK(I,K) / (2.D0*TAU)
```

```
        END DO

      END DO
      DO I = 2, N-1
        DO K = 2, M-1
          FTILDE(I,K) = FP(I,K) + WORK(I,K)
        END DO
      END DO
C
C     ITERATIVE CONJUGATE-GRADIENT METHOD

C
      NIT    = 0
      NITMAX = 5000
      AL     = 1.D0
C
C     INITIAL APPROXIMATION
C
      DO I = 1, N
        DO K = 1, M
          W(I,K)    = 0.D0
          WOLD(I,K) = 0.D0
        END DO
      END DO
C
C     ITERATION CYCLE
C

  100 NIT = NIT + 1

C
C     DISCREPANCY
C
      CALL ATILDE ( N, M, H, TAU, W, RK, WORK, A, B, C, F, Y )
      DO I = 2, N-1
        DO K = 2, M-1
          RK(I,K)  = RK(I,K) - FTILDE(I,K)
        END DO
      END DO
C

C     CORRECTION
C
      CALL A0I ( N, M, TAU, ARK, RK, A, B, C, F, Y )
C
C     ITERATION PARAMETERS
C
      CALL ATILDE ( N, M, H, TAU, ARK, FR, WORK, A, B, C, F, Y )
      RR = 0.D0

      RA = 0.D0
      DO I = 2, N-1


        DO K = 2, M-1
          RR = RR + RK(I,K)*ARK(I,K)
          RA = RA + FR(I,K)*ARK(I,K)
        END DO
      END DO
      IF (NIT.EQ.1) RR0 = RR
```

```
      TAUK = RR / RA
      IF (NIT.GT.1)
     +  AL = 1.D0 / (1.D0 - TAUK * RR / (TAUKOLD * RROLD * ALOLD))
C
C     NEXT APPROXIMATION
C
      DO I = 2, N-1
        DO K = 2, M-1
          AA = AL * W(I,K) + (1.D0 - AL) * WOLD(I,K)
     +       - TAUK * AL * ARK(I,K)
          WOLD(I,K) = W(I,K)
          W(I,K)    = AA
        END DO
      END DO
      RROLD   = RR
      TAUKOLD = TAUK

      ALOLD   = AL
C
C     END OF ITERATIONS
C
      IF (RR .GE. EPS*RR0 .AND. NIT .LT. NITMAX) GO TO 100
C
C     APPROXIMATE SOLUTION OF THE INVERSE PROBLEM
C
      DO I = 1, N
        DO K = 1, M
          W(I,K) = W(I,K) + V(I,K)
        END DO
      END DO
      DO I = 2, N-1
        X1 = (X(I) + X(I-1)) / 2
        X2 = (X(I+1) + X(I)) / 2
        A(I)  = AK(X1) / (2.D0*H*H)
        B(I)  = AK(X2) / (2.D0*H*H)
        FY(I) = (W(I,2) - W(I,1)) / TAU
     +        - A(I)*(W(I-1,2) + W(I-1,1))
     +        + (A(I)+B(I))*(W(I,2) + W(I,1))
     +        - B(I)*(W(I+1,2) + W(I+1,1))
      END DO
C
      WRITE ( 01,* ) NIT
      WRITE ( 01,* ) (FT(I), I = 2,N-1)
      WRITE ( 01,* ) (FY(I), I = 2,N-1)
      WRITE ( 01,* ) (PHI(I), I = 1,N)
      WRITE ( 01,* ) (PHID(I), I = 1,N)
      CLOSE (01)
      STOP
      END
C
      DOUBLE PRECISION FUNCTION  AK ( X )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      COEFFICIENT AT THE HIGHER DERIVATIVES
C
      AK = 1.D-1
C
      RETURN
      END
C
```

```
      DOUBLE PRECISION FUNCTION  AF ( X )

      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      RIGHT-HAND SIDE
C

      AF = 1.D0
      IF (X.GT.0.5D0) AF = 0.D0
C
      RETURN

      END
C
      SUBROUTINE A0I ( N, M, TAU, V, F, A, B, C, FF, YY )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     SOLUTION OF THE DIFFERENCE PROBLEM A0 V = F
C
      DIMENSION V(N,M), F(N,M)
     +         ,A(M), C(M), B(M), FF(M), YY(M)
C
      DO I = 2, N-1
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C

        DO K = 2, M-1
          A(K)  = 1.D0 / (TAU*TAU)
          B(K)  = A(K)
          C(K)  = A(K) + B(K)
          FF(K) = F(I,K)
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
        B(1)  = 0.D0
        C(1)  = 1.D0
        FF(1) = 0.D0
        A(M)  = 0.D0
        C(M)  = 1.D0
        FF(M) = 0.D0
C
C     SOLUTION OF THE PROBLEM
C
        ITASK = 1
        CALL PROG3 ( M, A, C, B, FF, YY, ITASK )
        DO K = 2, M-1
          V(I,K) = YY(K)
        END DO
      END DO
      RETURN
      END

C
      SUBROUTINE ATILDE ( N, M, H, TAU, V, F, WORK, A, B, C, FF, YY )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     CALCULATION OF F = ATILDE V
C
```

```
      DIMENSION V(N,M), F(N,M), WORK(N,M)
     +         ,A(M), C(M), B(M), FF(M), YY(M)
C
      DO I = 2, N-1
        DO K = 2, M-1
          X1 = (I-0.5D0) * H
          X2 = (I+0.5D0) * H

          AA = AK(X1) / (H*H)
          BB = AK(X2) / (H*H)

          F(I,K) = - AA*(V(I-1,K+1) - V(I-1,K-1))
     +             + (AA + BB)*(V(I,K+1) - V(I,K-1))
     +             - BB*(V(I+1,K+1) - V(I+1,K-1))
          F(I,K) = F(I,K) / (2.D0*TAU)
        END DO
      END DO
      CALL A0I ( N, M, TAU, WORK, F, A, B, C, FF, YY )
C
      DO I = 2, N-1

        DO K = 2, M-1
          X1 = (I-0.5D0) * H
          X2 = (I+0.5D0) * H
          AA = AK(X1) / (H*H)
          BB = AK(X2) / (H*H)
          F(I,K) = - AA*(WORK(I-1,K+1) - WORK(I-1,K-1))
     +             + (AA + BB)*(WORK(I,K+1) - WORK(I,K-1))

     +             - BB*(WORK(I+1,K+1) - WORK(I+1,K-1))
          F(I,K) = F(I,K) / (2.D0*TAU)
        END DO
      END DO
C
      AA = 1.D0 / (TAU*TAU)
      DO I = 2, N-1
        DO K = 2, M-1
          F(I,K) = - AA*(V(I,K-1) - 2.D0*V(I,K) + V(I,K+1))
     +             - F(I,K)

        END DO
      END DO
      RETURN
      END
```

The subroutine `A0I` solves the equation $A_0 v = f$, and the subroutine `ATILDE` calculates the value of $f = \tilde{A}v$ from the given v.

6.4.5 Computational experiments

The presented program `PROBLEM8` solves the inverse problem (6.110)–(6.113) in the case of

$$k(x) = 0.1, \quad f(x) = \begin{cases} 1, & 0 < x < 0.5, \\ 0, & 0.5 \le x < 1, \end{cases} \quad l = 1, \quad T = 1,$$

The problem is being solved on a uniform grid with $h = 0.02$ and $\tau = 0.02$.

Correctness of the problem is illustrated by calculation data obtained for the problem with perturbed boundary conditions at $t = T$ (function $\varphi(x)$ in (6.113)). In the framework of the quasi-real experiment, the obtained solution was perturbed at $t = T$ by the law

$$\varphi_\delta(x) = \varphi(x) + 2\delta(\sigma(x) - 1/2), \qquad x \in \omega,$$

where $\sigma(x)$ is a random function normally distributed over the interval $[0, 1]$, and the parameter δ defines the inaccuracy level.

Figure 6.13 shows the exact and reconstructed dependence of the right-hand side on the spatial variable, and also the exact and approximate difference end-time solutions of the direct problem obtained with $\delta = 0.0005$. Similar data obtained with an inaccuracy increased to $\delta = 0.001$ and $\delta = 0.0002$ are shown respectively in Figures 6.14 and 6.15. Considerable sensitiveness of the reconstructed right-hand side to the solution inaccuracy at $t = T$ is worth noting: at a relative inaccuracy of 0.1 % in the given $\varphi(x)$, the right-hand side could be determined accurate to approximately 25 %.

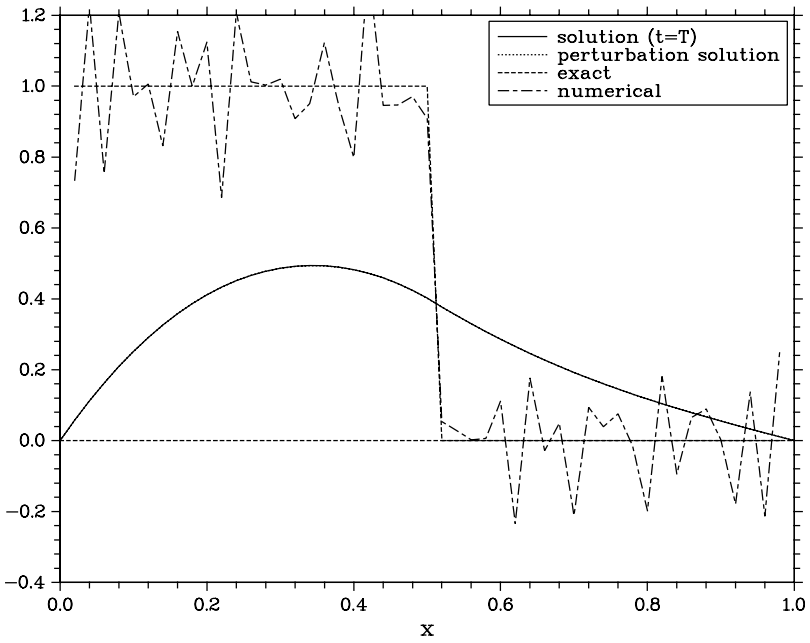

Figure 6.13 Solution of the problem obtained with $\delta = 0.0005$

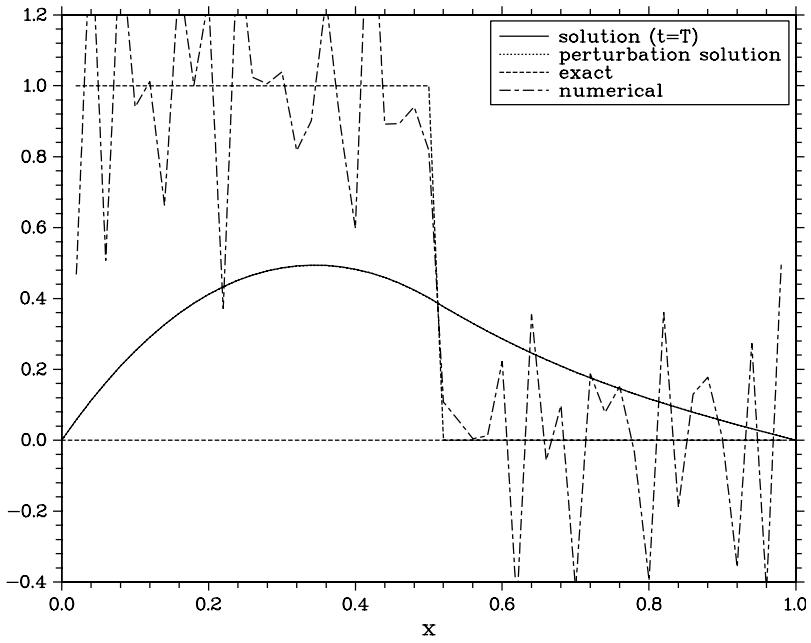

Figure 6.14 Solution of the problem obtained with $\delta = 0.0005$

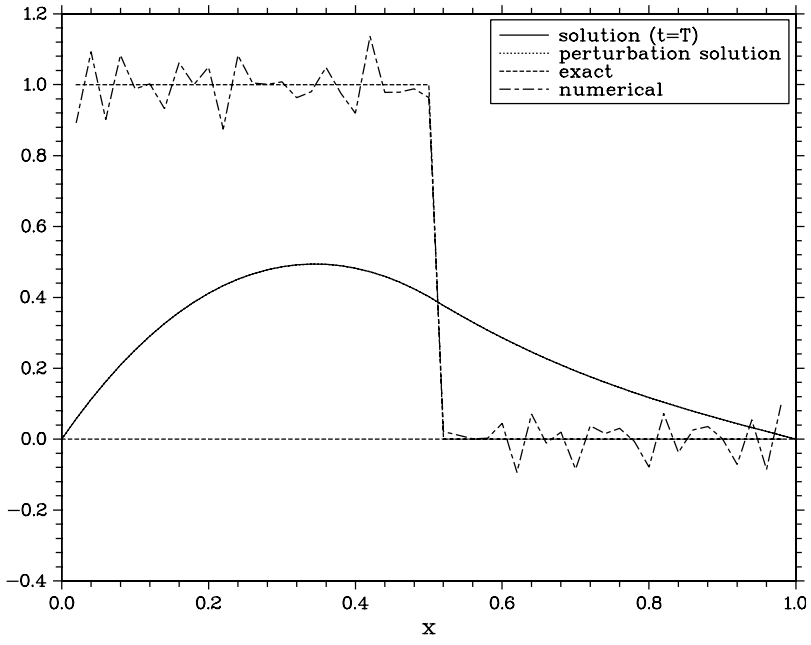

Figure 6.15 Solution of the problem perturbed with $\delta = 0.0002$

6.5 Reconstruction of the right-hand side of an elliptical equation from observation data obtained at the boundary

Below, we consider the classical inverse problem in the potential theory in which it is required to determine the unknown right-hand side of the elliptic equation in the case in which additional data are given on the boundary of the calculation domain.

6.5.1 Statement of the inverse problem

Consider a model inverse problem in which it is required to determine the unknown right-hand side from observation data obtained at the domain boundary. To simplify the consideration, restrict ourselves to the two-dimensional Poisson equation. Consider first the formulation of the direct problem.

In a bounded domain Ω the function $u(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, x_2)$ satisfies the equation

$$-\Delta u \equiv -\sum_{\alpha=1}^{2} \frac{\partial^2 u}{\partial x_\alpha^2} = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{6.144}$$

Consider a Dirichlet problem in which equation (6.144) is supplemented with the following first-kind homogeneous boundary conditions:

$$u(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega. \tag{6.145}$$

The direct problem is formulated in the form (6.144), (6.145), with known right-hand side $f(\boldsymbol{x})$ in (6.144).

Among the inverse problems for elliptic equations, consider the right-hand side identification problem. We assume that additional measurements are feasible only on the domain boundary. In addition to (6.145), the following second-kind boundary conditions are also considered:

$$\frac{\partial u}{\partial n}(\boldsymbol{x}) = \mu(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega, \tag{6.146}$$

where n is the external normal to Ω.

In this general formulation the solution of the inverse problem in which it is required to determine the pair of functions $\{u(\boldsymbol{x}), f(\boldsymbol{x})\}$ from conditions (6.144)–(6.146) is not unique. The latter statement requires no special comments: it suffices to consider the inverse problem in a circle with the right-hand side dependent on the distance from the center of the circle. The non-uniqueness stems from the fact that we are trying to reconstruct a two-dimensional function (the right-hand side $f(\boldsymbol{x})$ from a function with lower dimensionality ($\mu(\boldsymbol{x})$, $\boldsymbol{x} \in \partial\Omega$).

6.5.2 Uniqueness of the inverse-problem solution

Unique determination of the right-hand side is possible in the case in which the unknown right-hand side is independent of one of the variables. In fact, the latter was the case of the right-hand side identification problem for a parabolic equation, with the unknown dependence of the right-hand side on time or spatial variables to be reconstructed.

Not trying to consider the general case, turn to a typical example. We assume that the right-hand side (6.144) can be represented as

$$f(\boldsymbol{x}) = \varphi_1(x_2) + x_1\varphi_2(x_2). \tag{6.147}$$

We pose a problem in which it is required to determine two functions $\varphi_\alpha(x_2)$, $\alpha = 1, 2$, independent of one of the variables (namely, of the variable x_1), from (6.144)–(6.146).

We reformulate the inverse problem (6.144)–(6.147) by eliminating the unknown functions $\varphi_\alpha(x_2)$, $\alpha = 1, 2$. Double differentiation of (6.144) with respect to x_1 with allowance for (6.147) gives:

$$\frac{\partial^2}{\partial x_1^2}\,\Delta u = 0, \qquad \boldsymbol{x} \in \Omega. \tag{6.148}$$

In this way, we arrive at a boundary value problem for the composite equation (6.145), (6.146), (6.148).

Let us show that the solution of problem (6.145), (6.146), (6.148) is unique. For this to be shown, it suffices to prove that the solution of the problem with homogeneous boundary conditions

$$\frac{\partial u}{\partial n}(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega, \tag{6.149}$$

is $u(\boldsymbol{x}) \equiv 0$, $\boldsymbol{x} \in \Omega$.

We multiply equation (6.148) by $u(\boldsymbol{x})$ and perform integration over the whole domain Ω); this yields

$$\int_\Omega \frac{\partial^2}{\partial x_1^2}\,\Delta u\, u\, d\boldsymbol{x} = 0.$$

Taking into account the homogeneous boundary conditions (6.145) and permutability of the operators $\partial/\partial x_1$ and Δ, we obtain

$$\int_\Omega v\,\Delta v\, d\boldsymbol{x} = 0, \qquad v = \frac{\partial u}{\partial x_1}.$$

The pair of homogeneous boundary conditions (6.145), (6.149) guarantees that

$$v(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega.$$

Under these conditions, we have

$$\int_\Omega v\,\Delta v\, d\boldsymbol{x} = \sum_{\alpha=1}^{2} \int_\Omega \Big(\frac{\partial v}{\partial x_\alpha}\Big)^2 d\boldsymbol{x} = 0$$

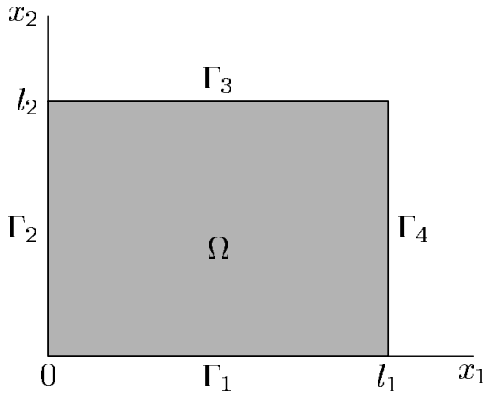

Figure 6.16 Calculation domain

and, hence, $v(\boldsymbol{x}) = 0$ throughout the whole domain Ω. From

$$\frac{\partial u}{\partial x_1} = 0, \qquad \boldsymbol{x} \in \Omega$$

and boundary conditions (6.145) it follows that the only solution of problem (6.145), (6.148), (6.149) is $u(\boldsymbol{x}) \equiv 0$, $\boldsymbol{x} \in \Omega$.

More informative a priori estimates for the solution of the boundary value problem (6.145), (6.146), (6.148) can also be obtained. This matter will be considered on the difference level below.

6.5.3 Difference problem

We assume the calculation domain to be a rectangle:

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\alpha < l_\alpha,\ \alpha = 1, 2\}.$$

For the sides of Ω we use conditions indicated in Figure 6.16, so that

$$\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4.$$

We seek the right-hand side of (6.144) in class (6.147) under the following additional conditions posed on the sides Γ_2 and Γ_4 of the rectangle:

$$\frac{\partial u}{\partial x_1}(0, x_2) = \mu_1(x_2), \qquad \frac{\partial u}{\partial x_1}(l_1, x_2) = \mu_2(x_2). \tag{6.150}$$

With boundary conditions given on Γ_1 and/or Γ_3 (see (6.146)), the problem becomes overspecified.

Along both direction x_α, $\alpha = 1, 2$, we introduce a uniform grid

$$\overline{\omega}_\alpha = \{x_\alpha \mid x_\alpha = i_\alpha h_\alpha,\ i_\alpha = 0, 1, \ldots, N_\alpha,\ N_\alpha h_\alpha = l_\alpha\},$$

so that

$$\begin{aligned} \omega_\alpha &= \{x_\alpha \mid x_\alpha = i_\alpha h_\alpha,\ i_\alpha = 1, 2, \dots, N_\alpha - 1,\ N_\alpha h_\alpha = l_\alpha\}, \\ \partial\omega_\alpha &= \{x_\alpha \mid x_\alpha = 0, l_\alpha\}. \end{aligned}$$

For the grid in the rectangle Ω we use the notations

$$\begin{aligned} \overline{\omega} &= \overline{\omega}_1 \times \overline{\omega}_2 = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ x_\alpha \in \overline{\omega}_\alpha,\ \alpha = 1, 2\}, \\ \omega &= \omega_1 \times \omega_2. \end{aligned}$$

In the standard notation adopted in the theory of difference schemes, at internal nodes we define the difference Laplace operator

$$\Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}, \qquad \boldsymbol{x} \in \omega.$$

We put into correspondence to the direct problem (6.144), (6.145) the difference problem

$$-\Lambda y = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \tag{6.151}$$

$$y(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\omega. \tag{6.152}$$

In the inverse problem, the right-hand side is sought in the class (6.147) from the additional conditions (6.150). To pass to a difference analogue of (6.145), (6.147), (6.150), we define the mesh function $v = -\Lambda y$ not only at internal nodes (see (6.151)), but also on the set of boundary nodes.

We can conveniently introduce fictitious nodes with $i_1 = -1$ and $i_1 = N_1 + 1$ to extend the grid over the variable x_1 by one node from either side. We approximate boundary conditions (6.150) on the extended grid. Accurate to $\mathcal{O}(h_1^2)$, we have:

$$\frac{y(h_1, x_2) - y(-h_1, x_2)}{2h_1} = \mu_1(x_2), \tag{6.153}$$

$$\frac{y(l_1 + h_1, x_2) - y(l_1 - h_1, x_2)}{2h_1} = \mu_2(x_2). \tag{6.154}$$

Taking into account the boundary conditions (6.145) at the left boundary, we obtain

$$v(0, x_2) = -\Lambda y(0.x_2) = -\frac{y(h_1, x_2) - 2y(0, x_2) + y(-h_1, x_2)}{h_1^2}.$$

With (6.153) taken into account, we arrive at the expression

$$v(0, x_2) = -\frac{2}{h_1^2}\, y(h_1, x_2) + \frac{2}{h_1}\, \mu_1(x_2). \tag{6.155}$$

In a similar way, on Γ_4 we obtain

$$v(l_1, x_2) = -\frac{2}{h_1^2}\, y(l_1 - h_1, x_2) - \frac{2}{h_1}\, \mu_1(x_2). \tag{6.156}$$

Double difference differentiation (6.151) yields the equation

$$v_{\bar{x}_1 x_1} = 0, \qquad \boldsymbol{x} \in \omega. \tag{6.157}$$

The boundary conditions for this equation have the form (6.154), (6.155). With known v, the solution is to be determined (see (6.151), (6.152)) from

$$-\Lambda y = v(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \tag{6.158}$$

$$y(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\omega. \tag{6.159}$$

In this manner, we arrive at a system of two difference Poisson equations for the pair $\{y, v\}$. These two equations are interrelated via boundary conditions (6.155), (6.156).

We can conveniently reformulate the boundary value problem with non-homogeneous boundary conditions (6.155)–(6.157) as a problem with homogeneous boundary conditions for a non-homogeneous equation at the internal nodes. With (6.155), at near-boundary nodes we have:

$$\frac{2v(h_1, x_2) - v(2h_1, x_2)}{h_1^2} = \frac{v(0, x_2)}{h_1^2} = -\frac{2}{h_1^4}\, y(h_1, x_2) + \frac{2}{h_1^3}\, \mu_1(x_2).$$

From (6.156), we obtain

$$\frac{2v(l_1 - h_1, x_2) - v(l_1 - 2h_1, x_2)}{h_1^2} = -\frac{2}{h_1^4}\, y(l_1 - h_1, x_2) - \frac{2}{h_1^3}\, \mu_2(x_2).$$

Let us define difference operators A_α, $\alpha = 1, 2$ on the set of mesh functions vanishing at the boundary nodes:

$$A_\alpha y = -y_{\bar{x}_\alpha x_\alpha}, \qquad \boldsymbol{x} \in \omega.$$

Then, the boundary value problem (6.155)–(6.157) can be written as

$$A_1 v = -A_0 y + \phi, \qquad \boldsymbol{x} \in \omega. \tag{6.160}$$

Here, the difference operator A_0 is defined by

$$A_0 y = \begin{cases} \dfrac{2}{h_1^4}\, y(h_1, x_2), & x_1 = h_1, \\ 0, & h_1 < x_1 < l_1 - h_1, \\ \dfrac{2}{h_1^4}\, y(l_1 - h_1, x_2), & x_1 = l_1 - h_1. \end{cases}$$

The right-hand side ϕ in nonzero only at near-boundary nodes:

$$\phi = \begin{cases} -\dfrac{2}{h_1^3}\, \mu_1(x_2), & x_1 = h_1, \\ 0, & h_1 < x_1 < l_1 - h_1, \\ \dfrac{2}{h_1^3}\, \mu_2(x_2), & x_1 = l_1 - h_1. \end{cases}$$

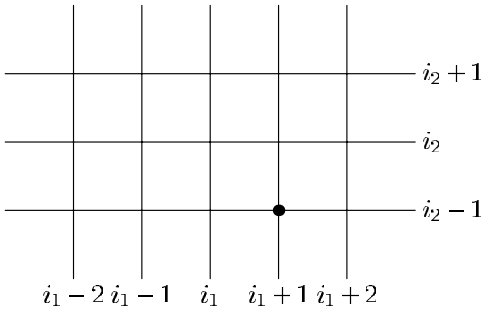

Figure 6.17 The mesh pattern for the difference scheme

In the introduced notation, the boundary value problem (6.158), (6.159) takes the form

$$(A_1 + A_2)y = v, \qquad \boldsymbol{x} \in \omega. \tag{6.161}$$

In this way, we pass from (6.155)–(6.159) to the system (6.160), (6.161). The latter system can be conveniently written as the following single operator equation:

$$Ay = \phi. \tag{6.162}$$

Here,

$$A = (A_1 + A_2)A_1 + A_0. \tag{6.163}$$

The mesh pattern used in this difference scheme is shown in Figure 6.17.

In the ordinary way, in the Hilbert space $H = L_2(\omega)$ we introduce the scalar product and the norm:

$$(y, w) \equiv \sum_{\boldsymbol{x}\in\omega} y(\boldsymbol{x})w(\boldsymbol{x})h_1h_2, \qquad \|y\| \equiv (y, y)^{1/2}.$$

In H,

$$A_\alpha = A_\alpha^* > 0, \qquad A_0 = A_0^* \geq 0$$

and, hence, in (6.162) we have $A = A^* > 0$. By virtue of this, the difference problem (6.162) has a unique solution.

We gave a sufficiently general scheme for constructing a discrete analogue to the non-classical boundary value problem (6.145), (6.148), (6.150) suitable for solving even more complex problems. In the case under consideration, we can substantially simplify the problem by explicitly writing the solution of problem (6.155)–(6.157). It should be noted that such transformation allows for specific features of the inverse problem in the greatest possible extent and most clearly on the difference level.

The general solution of the difference equation (6.157) is a linear function of x_1:

$$v(x_1, x_2) = \Big(1 - \frac{x_1}{l_1}\Big)v(0, x_2) + \frac{x_1}{l_1}\, v(l_1, x_2).$$

With boundary conditions (6.155), (6.156) taken into account, we obtain

$$v(x_1, x_2) = -\Big(1 - \frac{x_1}{l_1}\Big)\frac{2}{h_1^2}\, y(h_1, x_2) - \frac{x_1}{l_1}\,\frac{2}{h_1^2}\, y(l_1 - h_1, x_2) + \psi(x_1, x_2), \quad (6.164)$$

where

$$\psi(x_1, x_2) = \Big(1 - \frac{x_1}{l_1}\Big)\,\frac{2}{h_1}\,\mu_1(x_2) - \frac{x_1}{l_1}\,\frac{2}{h_1}\,\mu_2(x_2).$$

Substitution of (6.164) into (6.151) leads us to the difference equation

$$-\Lambda y + \Big(1 - \frac{x_1}{l_1}\Big)\frac{2}{h_1^2}\, y(h_1, x_2) + \frac{x_1}{l_1}\,\frac{2}{h_1^2}\, y(l_1 - h_1, x_2) = \psi(\boldsymbol{x}), \quad \boldsymbol{x} \in \omega. \quad (6.165)$$

In this way, the solution of the inverse right-hand side identification problem for equation (6.151) in class (6.147) has reduced to the solution of the boundary value problem (6.152), (6.165). Equation (6.165) is a loaded difference equation.

6.5.4 Solution of the difference problem

To find the solution of the difference problem, we use the variable separation method. This approach can be applied to the difference problem (6.152) with the composite difference operator A defined by (6.153). Let us dwell on a second possibility, when to be sought is the solution of the boundary-layer problem for the loaded difference elliptic equation (6.152), (6.165).

We write problem (6.152), (6.165) as the equation

$$(A_1 + A_2)y + q_1(x_1)y(h_1, x_2) + q_2(x_1)y(l_1 - h_1, x_2) = \psi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega, \quad (6.166)$$

in which $q_\alpha \geq 0, \alpha = 1, 2$.

We denote as λ_k, $v_k(x_2)$, $k = 1, 2, \ldots, N_2 - 1$ the eigenvalues and eigenfunctions of A_2:

$$A_2 v = \lambda v.$$

The solution of this difference spectral problem is well known:

$$\lambda_k = \frac{4}{h_2^2}\,\sin^2\frac{k\pi h_2}{2l_2}, \qquad v_k(x_2) = \sqrt{\frac{2}{l_2}}\,\sin\frac{k\pi x_2}{l_2}.$$

The eigenfunctions are orthonormal functions:

$$(v_k, v_m)^{(2)} = \delta_{km}, \qquad \delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m, \end{cases}$$

where

$$(v, w)^{(2)} = \sum_{i_2=1}^{N_2-1} v(x_2)w(x_2)h_2$$

is the scalar product in $L_2(\omega_2)$.

We seek the solution of problem (6.166) as an expansion in the eigenfunctions of A_2:

$$y(x_1, x_2) = \sum_{k=1}^{N_2-1} c_k(x_1) v_k(x_2).$$

Substitution into (6.166) leads us to the necessity to solve the difference problems

$$(A_1 + \lambda_k) c_k(x_1) + q_1(x_1) c_k(h_1) + q_2(x_1) c_k(l_1 - h_1) = \psi_k(x_1), \tag{6.167}$$

where

$$\psi_k(x_1) = (\psi, v_k)^{(2)}, \qquad k = 1, 2, \ldots, N_2 - 1.$$

The matter of solution of these $N_2 - 1$ one-dimensional difference problems should be given particular attention. In the case of (6.167) it is required to find the solution of the difference boundary value problem

$$-w_{\bar{x}_1 x_1} + \lambda w + q_1(x_1) w(h_1) + q_2(x_1) w(l_1 - h_1) = r(x_1), \tag{6.168}$$

$$w(0) = 0, \qquad w(l_1) = 0. \tag{6.169}$$

We represent the solution of problem (6.168), (6.169) as

$$w(x_1) = s(x_1) + s^{(1)}(x_1) w(h_1) + s^{(2)}(x_1) w(l_1 - h_1). \tag{6.170}$$

We insert this expression into (6.168) and isolate the functions s_α, $\alpha = 1, 2$, collecting the terms with $w(h_1)$, $w(l_1 - h_1)$ and equating them to zero. This gives us three three-point difference equations for the auxiliary functions $s(x_1)$, $s_\alpha(x_1)$, $\alpha = 1, 2$. With allowance for (6.169), we assume the boundary conditions to be homogeneous, so that

$$-s_{\bar{x}_1 x_1} + \lambda s = r(x_1), \tag{6.171}$$

$$s(0) = 0, \qquad s(l_1) = 0, \tag{6.172}$$

$$-s^{(1)}_{\bar{x}_1 x_1} + \lambda s^{(1)} = -q_1(x_1), \tag{6.173}$$

$$s^{(1)}(0) = 0, \qquad s^{(1)}(l_1) = 0, \tag{6.174}$$

$$-s^{(2)}_{\bar{x}_1 x_1} + \lambda s^{(2)} = -q_2(x_1), \tag{6.175}$$

$$s^{(2)}(0) = 0, \qquad s^{(2)}(l_1) = 0. \tag{6.176}$$

After solving the three standard problems (6.171)–(6.176), the functions $w(h_1)$ and $w(l_1 - h_1)$ can be found. From representation (6.170), it readily follows that

$$w(h_1) = s(h_1) + s^{(1)}(h_1) w(h_1) + s^{(2)}(h_1) w(l_1 - h_1), \tag{6.177}$$

$$w(l_1 - h_1) = s(l_1 - h_1) + s^{(1)}(l_1 - h_1) w(h_1) + s^{(2)}(l_1 - h_1) w(l_1 - h_1). \tag{6.178}$$

Solvability of this system is controlled by the determinant

$$D = (1 - s^{(1)}(h_1))(1 - s^{(2)}(l_1 - h_1)) - s^{(2)}(h_1)s^{(1)}(l_1 - h_1),$$

whose non-zero value can easily be guaranteed under certain constraints. Taking the inequalities $\lambda > 0$ and $q_\alpha \geq 0$, $\alpha = 1, 2$ into account, we have: $s^{(\alpha)}(x_1) \geq 0$, $0 \leq x_1 \leq l_1$. Hence, we have $D > 0$ at sufficiently small l_1, for instance.

As a matter of fact, the determinant of (6.177), (6.178) is always positive. To show this, we have to recall (see (6.164)) the expressions for the mesh functions $q_\alpha(x_1)$, $\alpha = 1, 2$:

$$q_1(x_1) = \Big(1 - \frac{x_1}{l_1}\Big)\frac{2}{h_1^2}, \qquad q_2(x_1) = \frac{x_1}{l_1}\,\frac{2}{h_1^2}.$$

By virtue of this, for the solutions of the boundary value problems (6.173), (6.174) and (6.175), (6.176) there hold the relations

$$s^{(1)}(x_1) = s^{(2)}(l_1 - x_1), \qquad 0 \leq x_1 \leq l_1.$$

The latter means, in particular, that it is unnecessary for us to solve (in the case of the uniform computational grid used) one of the two difference boundary value problems, (6.173), (6.174) or (6.175), (6.176). Hence,

$$D = 1 - 2s^{(1)}(h_1) + (s^{(1)}(h_1))^2 - (s^{(1)}(l_1 - h_1))^2$$

and, with allowance for $s^{(1)}(h_1) > s^{(1)}(l_1 - h_1)$, we obtain that $D > 1$.

With the mesh functions $s(x_1)$, $s^{(\alpha)}$, $\alpha = 1, 2$ found from (6.171)–(6.176) and with the mesh functions $w(h_1)$ and $w(l_1 - h_1)$ found from (6.177), (6.178), the solution of problem (6.168), (6.169) can be found in the form (6.170). The program realization of this algorithm is discussed below.

6.5.5 Program

The used computational algorithm employs the fast Fourier transform over the variable x_2. In the embodied algorithm, one can conveniently use the non-normalized eigenfunctions

$$\sin\frac{k\pi x_2}{l_2} = \sin\frac{k\pi i_2}{N_2}, \qquad x_2 = i_2 h_2.$$

In the latter case, any mesh function f_{i_2} defined at $i_2 = 1, 2, \ldots, N_2 - 1$ (on ω_2) can be expanded as

$$f_{i_2} = \frac{2}{N_2}\sum_{k=1}^{N_2-1}\xi_k \sin\frac{k\pi i_2}{N_2}, \qquad i_2 = 1, 2, \ldots, N_2 - 1,$$

where

$$\xi_k = \sum_{i_2=1}^{N_2-1} f_{i_2} \sin\frac{k\pi i_2}{N_2}, \qquad k = 1, 2, \ldots, N_2 - 1.$$

Let us use the subroutine SINT from the program library SLATEC. Here, we cite a borrowed description of the subroutine, representing a component of the big program package FFTPACK that embodies fast Fourier transform algorithms. You can use other subroutines for fast Fourier transform available at hand.

Subroutine SINT

```
*DECK SINT
      SUBROUTINE SINT (N, X, WSAVE)
C***BEGIN PROLOGUE  SINT
C***PURPOSE  Compute the sine transform of a real, odd sequence.
C***LIBRARY   SLATEC (FFTPACK)
C***CATEGORY  J1A3
C***TYPE      SINGLE PRECISION (SINT-S)
C***KEYWORDS  FFTPACK, FOURIER TRANSFORM
C***AUTHOR  Swarztrauber, P. N., (NCAR)
C***DESCRIPTION

C
C  Subroutine SINT computes the discrete Fourier sine transform
C  of an odd sequence X(I).  The transform is defined below at
C  output parameter X.
C
C  SINT is the unnormalized inverse of itself since a call of SINT
C  followed by another call of SINT will multiply the input sequence
C  X by 2*(N+1).
C

C  The array WSAVE which is used by subroutine SINT must be
C  initialized by calling subroutine SINTI(N,WSAVE).
C
C  Input Parameters
C
C  N       the length of the sequence to be transformed.  The method
C          is most efficient when N+1 is the product of small primes.
C

C  X       an array which contains the sequence to be transformed
C
C
C  WSAVE   a work array with dimension at least INT(3.5*N+16)
C          in the program that calls SINT.  The WSAVE array must be
C          initialized by calling subroutine SINTI(N,WSAVE), and a
C          different WSAVE array must be used for each different
C          value of N.  This initialization does not have to be
C          repeated so long as N remains unchanged.  Thus subsequent

C          transforms can be obtained faster than the first.
C
C  Output Parameters
C
C  X       For I=1,\dots,N
C
C               X(I)= the sum from K=1 to K=N
C
C                    2*X(K)*SIN(K*I*PI/(N+1))
C
C               A call of SINT followed by another call of
```

```
C               SINT will multiply the sequence X by 2*(N+1).
C               Hence SINT is the unnormalized inverse
C               of itself.
C
C  WSAVE   contains initialization calculations which must not be
C          destroyed between calls of SINT.
C
C***REFERENCES  P. N. Swarztrauber, Vectorizing the FFTs, in Parallel
C                 Computations (G. Rodrigue, ed.), Academic Press,
C                 1982, pp. 51-83.
C***ROUTINES CALLED  RFFTF
C***REVISION HISTORY  (YYMMDD)
C   790601  DATE WRITTEN
C   830401  Modified to use SLATEC library source file format.
C   860115  Modified by Ron Boisvert to adhere to Fortran 77 by
C           (a) changing dummy array size declarations (1) to (*),

C           (b) changing definition of variable SQRT3 by using
C               FORTRAN intrinsic function SQRT instead of a DATA
C               statement.
C   881128  Modified by Dick Valent to meet prologue standards.
C   891009  Removed unreferenced statement label.  (WRB)
C   891009  REVISION DATE from Version 3.2
C   891214  Prologue converted to Version 4.0 format.  (BAB)
C   920501  Reformatted the REFERENCES section.  (WRB)
C***END PROLOGUE  SINT
```

The auxiliary subroutine `SINTI` initializes the subroutine `SINT`.

Subroutine SINTI

```
*DECK SINTI
      SUBROUTINE SINTI (N, WSAVE)
C***BEGIN PROLOGUE  SINTI
C***PURPOSE  Initialize a work array for SINT.
C***LIBRARY   SLATEC (FFTPACK)
C***CATEGORY  J1A3
C***TYPE      SINGLE PRECISION (SINTI-S)
C***KEYWORDS  FFTPACK, FOURIER TRANSFORM
C***AUTHOR  Swarztrauber, P. N., (NCAR)
C***DESCRIPTION
C

C  Subroutine SINTI initializes the array WSAVE which is used in
C  subroutine SINT.  The prime factorization of N together with
C  a tabulation of the trigonometric functions are computed and
C  stored in WSAVE.
C
C  Input Parameter
C
C  N       the length of the sequence to be transformed.  The method
C          is most efficient when N+1 is a product of small primes.
C
C  Output Parameter
C
C  WSAVE   a work array with at least INT(3.5*N+16) locations.
C          Different WSAVE arrays are required for different values
```

```
C          of N.  The contents of WSAVE must not be changed between
C          calls of SINT.
C
C***REFERENCES  P. N. Swarztrauber, Vectorizing the FFTs, in Parallel
C               Computations (G. Rodrigue, ed.), Academic Press,
C               1982, pp. 51-83.
C***ROUTINES CALLED  RFFTI
C***REVISION HISTORY  (YYMMDD)

C   790601  DATE WRITTEN
C   830401  Modified to use SLATEC library source file format.

C   860115  Modified by Ron Boisvert to adhere to Fortran 77 by
C           (a) changing dummy array size declarations (1) to (*),
C           (b) changing references to intrinsic function FLOAT
C               to REAL, and
C           (c) changing definition of variable PI by using
C               FORTRAN intrinsic function ATAN instead of a DATA
C               statement.
C   881128  Modified by Dick Valent to meet prologue standards.
C   890531  Changed all specific intrinsics to generic.  (WRB)
C   890531  REVISION DATE from Version 3.2
C   891214  Prologue converted to Version 4.0 format.  (BAB)
C   920501  Reformatted the REFERENCES section.  (WRB)
C***END PROLOGUE  SINTI
```

The program PROBLEM9 is used to obtain the approximate solution of the inverse problem.

Program PROBLEM9

```
C
C      PROBLEM9  IDENTIFICATION OF THE RIGHT-HAND SIDE
C                OF THE POISSON EQUATION IN A RECTANGLE
C
C
      PARAMETER (  DELTA = 0.0, N1 = 257, N2 = 257 )
C
      DIMENSION U(N1,N2), Y(N1,N2), F(N1,N2), X1(N1), X2(N2)
     +         ,FIT1(N2), FIT2(N2), FI1(N2), FI2(N2)
     +         ,AM1(N2), AM2(N2), AMD1(N2), AMD2(N2)
     +         ,A(N1), B(N1), C(N1), FF(N1), YY(N1), S(N1), S1(N1)
     +         ,WSAVE(4*N2), WW(N2-2)
C
C
C     PARAMETERS:
C
C     X1L, X2L  - COORDINATES OF THE LEFT BOTTOM CORNER
C                 OF THE RECTANGULAR CALCULATION DOMAIN;
C     X1R, X2R  - COORDINATES OF THE RIGHT TOP CORNER;
C     N1, N2    - NUMBER OF DRID NODES OVER THE CORRESPONDING
C                 DIRECTIONS;
C     U(N1,N2)  - DIFFERENCE SOLUTION OF THE DIRECT PROBLEM;
C     FI(N2),
C     FI2(N2)   - EXACT RIGHT-HAND SIDE COMPONENTS;
C     AM1(N2),
```

```
C     AM2(N2)   - BOUNDARY CONDITIONS;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     AMD1(N2),
C     AMD2(N2)  - DISTURBED BOUNDARY CONDITIONS;
C     FIS1(N2),
C     FIS2(N2)  - FOUND RIGHT-HAND SIDE COMPONENTS;
C
      X1L   = 0.
      X1R   = 1.
      X2L   = 0.
      X2R   = 1.
      PI    = 3.1415926
C
      OPEN (01, FILE = 'RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H1 = (X1R - X1L) / (N1 - 1)
      H2 = (X2R - X2L) / (N2 - 1)
      DO I = 1, N1
        X1(I) = X1L + (I-1)*H1
      END DO
      DO J = 1, N2
        X2(J) = X2L + (J-1)*H2
      END DO
C
C     DIRECT PROBLEM
C
C     EXACT RIGHT-HAND SIDE
C
      DO J = 2, N2-1
        FIT1(J) = AF1(X2(J))
        FIT2(J) = AF2(X2(J))
        DO I = 2, N1-1
          F(I,J) = FIT1(J) + X1(I) * FIT2(J)
        END DO
      END DO
C
C     FORWARD FOURIER TRANSFORM
C
      CALL SINTI(N2-2,WSAVE)
      DO I = 2, N1-1
        DO J = 2, N2-1
          WW(J-1) = F(I,J)

        END DO
        CALL SINT(N2-2,WW,WSAVE)
        DO J = 2, N2-1
          F(I,J) = WW(J-1)
        END DO
      END DO
C
C     SOLUTION OF THE ONE-DIMENSIONAL (OVER THE VARIABLE X1) PROBLEM
C
      DO J = 2, N2-1
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        ALAM = 4./H2**2*(SIN(PI*(J-1)/(2.*(N2-1))))**2
        DO I = 2, N1-1
```

```
          A(I)  = 1. / (H1*H1)
          B(I)  = 1. / (H1*H1)
          C(I)  = A(I) + B(I) + ALAM
          FF(I) = F(I,J)
        END DO
C
C     BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
        B(1)  = 0.
        C(1)  = 1.
        FF(1) = 0.
        A(N1)  = 0.
        C(N1)  = 1.
        FF(N1) = 0.
C
C     SOLUTION OF THE ONE-DIMENSIONAL PROBLEM
C
        ITASK = 1
        CALL PROGS3 ( N1, A, C, B, FF, YY, ITASK )

        DO I = 1, N1
          F(I,J) = YY(I)
        END DO
      END DO
C
C     INVERSE FOURIER TRANSFORM
C
      DO I = 2, N1-1
        DO J = 2, N2-1
          WW(J-1) = F(I,J)
        END DO
        CALL SINT(N2-2,WW,WSAVE)

        DO J = 2, N2-1
          U(I,J) = 1./(2.*(N2-1))*WW(J-1)

        END DO
      END DO
C
C     BOUNDARY CONDITIONS IN THE INVERSE PROBLEM
C
      DO J = 1, N2
        AM1(J) = U(2,J) / H1
        AM2(J) = - U(N1-1,J) / H1
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO J = 1, N2
        AMD1(J) = AM1(J) + 2.*DELTA*(RAND(0)-0.5)
        AMD2(J) = AM2(J) + 2.*DELTA*(RAND(0)-0.5)
      END DO
C
C     INVERSE PROBLEM
C
C     RIGHT-HAND SIDE OF THE LOADED EQUATION

C
      DO I = 2, N1-1
        Q1 = 2.*(N1-I) / (X1R - X1L)
```

```
          Q2 = 2.*(I-1) / (X1R - X1L)
          DO J = 2, N2-1
            F(I,J) = Q1 * AMD1(J) - Q2 * AMD2(J)
          END DO
        END DO
C

C       FORWARD FOURIER TRANSFORM
C
        CALL SINTI(N2-2,WSAVE)
        DO I = 2, N1-1
          DO J = 2, N2-1

            WW(J-1) = F(I,J)
          END DO
          CALL SINT(N2-2,WW,WSAVE)
          DO J = 2, N2-1
            F(I,J) = WW(J-1)
          END DO
        END DO

C
C       SOLUTION OF THE ONE-DIMENSIONAL (OVER THE VARIABLE X1) PROBLEM
C
        DO J = 2, N2-1
C
C       AUXILIARY FUNCTIONS
C
C       DIFFERENCE-SCHEME COEFFICIENTS
C
          ALAM = 4./H2**2*(SIN(PI*(J-1)/(2.*(N2-1))))**2
          DO I = 2, N1-1
            A(I)  = 1. / (H1*H1)
            B(I)  = 1. / (H1*H1)
            C(I)  = A(I) + B(I) + ALAM
            FF(I) = F(I,J)
          END DO
C
C       BOUNDARY CONDITIONS AT THE LEFT AND RIGHT END POINTS
C
          B(1)  = 0.
          C(1)  = 1.
          FF(1) = 0.
          A(N1)  = 0.
          C(N1)  = 1.
          FF(N1) = 0.
C
C       SOLUTION OF THE ONE-DIMENSIONAL PROBLEM
C
          ITASK = 1
          CALL PROGS3 ( N1, A, C, B, FF, S, ITASK )
          DO I = 2, N1-1
            FF(I) = - 2.*(N1-I) / ((N1-1.)*H1*H1)
          END DO
          ITASK = 2
          CALL PROGS3 ( N1, A, C, B, FF, S1, ITASK )

C
C       SOLUTION OF THE SYSTEM OF TWO EQUATIONS
C
```

```
        DD = (1. - S1(2))**2 - (S1(N1-1))**2
        W1 = ((1. - S1(2))*S(2) + S1(N1-1)*S(N1-1))/DD
        W2 = ((1. - S1(2))*S(N1-1) + S1(N1-1)*S(2))/DD
        DO I = 2, N1-1
          F(I,J) = S(I) + W1*S1(I) + W2*S1(N1+1-I)
        END DO
      END DO
C
C     INVERSE FOURIER TRANSFORM
C
      DO I = 2, N1-1
        DO J = 2, N2-1
          WW(J-1) = F(I,J)
        END DO
        CALL SINT(N2-2,WW,WSAVE)
        DO J = 2, N2-1
          Y(I,J) = 1./(2.*(N2-1))*WW(J-1)
        END DO
      END DO
C
C     RIGHT-HAND SIDE

C
      DO I = 2, N1-1
        DO J = 2, N2-1

          F(I,J) = - (Y(I+1,J) - 2.*Y(I,J) + Y(I-1,J)) / (H1*H1)
     +             - (Y(I,J+1) - 2.*Y(I,J) + Y(I,J-1)) / (H2*H2)
        END DO
      END DO
C
      DO J = 2, N2-1
        FI1(J)  = 2.*AMD1(J)/H1 - 2.*Y(2,J)/(H1*H1)
        FI2(J)  = - 2.*(AMD2(J)+AMD1(J))/((X1R-X1L)*H1)
     +            - 2.*(Y(N1-1,J)-Y(2,J))/((X1R-X1L)*H1*H1)
      END DO
      WRITE ( 01, * ) (FI1(J),J=2,N2-1)
      WRITE ( 01, * ) (FIT1(J),J=2,N2-1)
      WRITE ( 01, * ) (FI2(J),J=2,N2-1)
      WRITE ( 01, * ) (FIT2(J),J=2,N2-1)
      CLOSE ( 01 )
C
      STOP
      END

      FUNCTION  AF1 ( X2 )
C
C      FIRST COMPONENT OF THE RIGHT-HAND SIDE OF THE EQUATION
C
      AF1 = 1.
      IF (X2.GT.0.5) AF1 = 0.
C
      RETURN
      END

      FUNCTION  AF2 ( X2 )
C
C     SECOND COMPONENT OF THE RIGHT-HAND SIDE OF THE EQUATION
C
      AF2 = X2
```

```
C
      RETURN
      END

      SUBROUTINE PROGS3 ( N, A, C, B, F, Y, ITASK )
C
C     SWEEP METHOD
C     FOR TRIDIAGONAL MATRIX
C
C     ITASK = 1: FACTORIZATION AND SOLUTION;
C
C     ITASK = 2: ONLY SOLUTION
C
      DIMENSION A(N), C(N), B(N), F(N), Y(N)
      IF ( ITASK .EQ. 1 ) THEN
C
      B(1) = B(1) / C(1)
      DO I = 2, N
        C(I) = C(I) - B(I-1)*A(I)
        B(I) = B(I) / C(I)
      END DO
C
      ITASK = 2
      END IF
C
      F(1) = F(1) / C(1)
      DO I = 2, N
        F(I) = ( F(I) + F(I-1)*A(I) ) / C(I)
      END DO
C
      Y(N) = F(N)
      DO I = N-1, 1, -1
        Y(I) = B(I)*Y(I+1) + F(I)
      END DO
      RETURN
      END
```

The function of the additional programs requires no comments.

6.5.6 Computational experiments

The presented program PROBLEM9 solves the inverse problem (6.144), (6.145), (6.150), in which to be found is the right-hand side (6.147) with

$$\varphi_1(x_2) = \begin{cases} 1, & 0 < x_2 < 0.5, \\ 0, & 0.5 < x_2 < 1, \end{cases} \qquad \varphi_1(x_2) = x_2.$$

The problem is solved in the unit square ($l_1 = l_2 = 1$). To obtain the input data for the inverse problem, the direct problem (6.144), (6.145) is preliminarily solved at a given right-hand side.

First of all, consider the solution data obtained for the inverse problem with unperturbed input data. Of interest here are numerical data obtained on a sequence of progressively refined grids (see Figures 6.18–6.20). The approximate solution is seen

to converge to the exact solution. A sufficiently high accuracy can be obtained using refined grids.

More sensitive to input-data inaccuracies are calculation data obtained at different levels of boundary-condition inaccuracies (6.150). These inaccuracies were modeled in the ordinary way, for instance, as

$$\tilde{\mu}_1(x_2) = \mu_1(x_2) + 2\delta(\sigma(x_2) - 1/2), \qquad x_2 \in \omega_2,$$

where $\sigma(x_2)$ is a random function normally distributed over the interval [0, 1], and the parameter δ defines the inaccuracy level. Figure 6.21 shows data obtained by solving the inverse problem with $\delta = 0.0003$. This inaccuracy level corresponds to a relative inaccuracy of 0.1%. The calculation grid $N_1 = N_2 = 129$ was used.

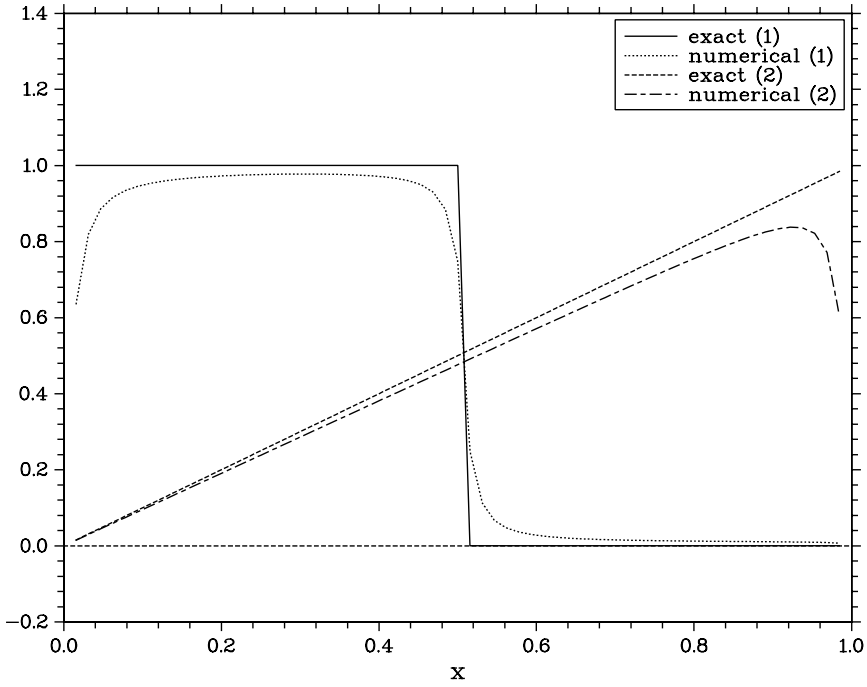

Figure 6.18 Solution of the problem obtained on the grid $N_1 = N_2 = 65$

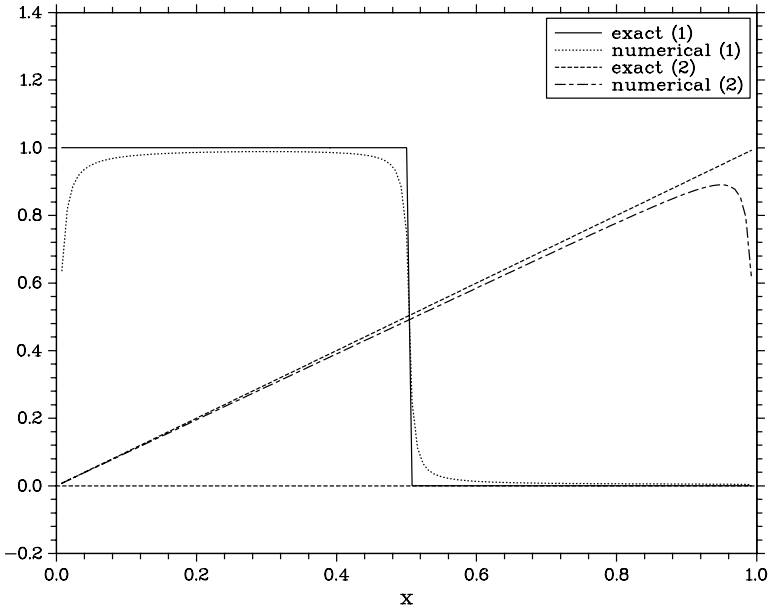

Figure 6.19 Solution of the problem obtained on the grid $N_1 = N_2 = 129$

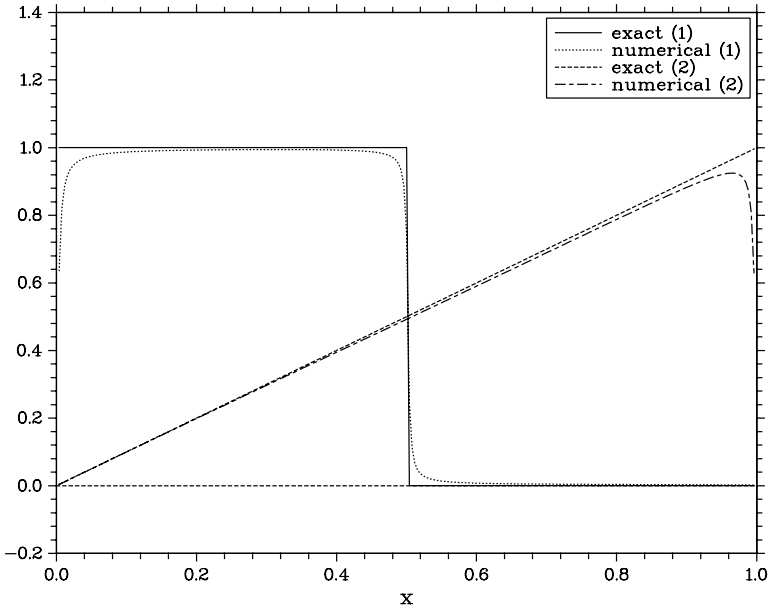

Figure 6.20 Solution of the problem obtained on the grid $N_1 = N_2 = 257$

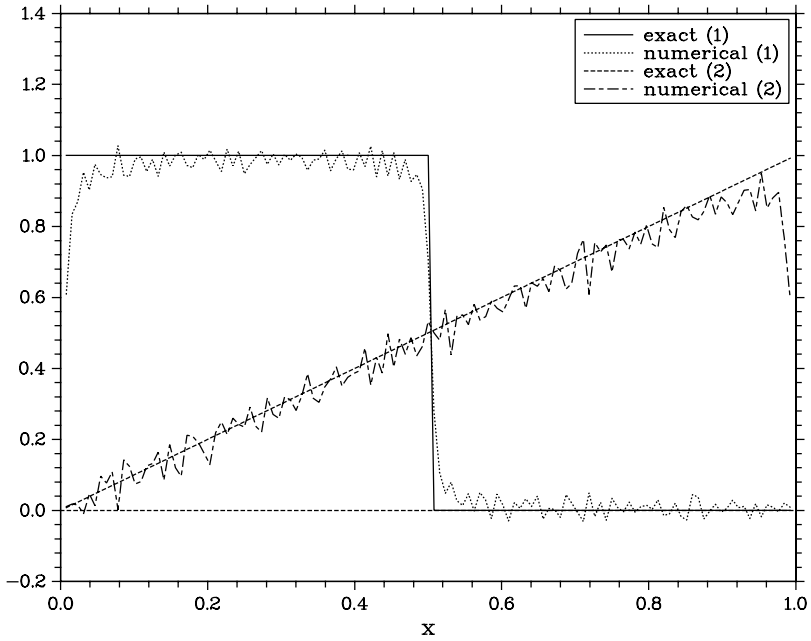

Figure 6.21 Solution of the problem obtained at the inaccuracy level $\delta = 0.0003$

6.6 Exercises

Exercise 6.1 Formulate the condition for convergence of the approximate solution determined from (6.16), (6.17), (6.19) to the solution of problem (6.14).

Exercise 6.2 Substantiate the iteration method (6.26) as applied to the approximate solution of problem (6.14), (6.16).

Exercise 6.3 Modify the program `PROBLEM5` so that to make it suitable for solving the inverse problem (6.14), (6.16) by the iteration method (6.26). Give a comparative analysis of the methods.

Exercise 6.4 Consider the basic specific features of the Tikhonov regularization method as applied to solving the inverse problem on the identification of the right-hand side of parabolic equation with lower terms

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) + b(x)\,\frac{\partial u}{\partial x} + f(x,t), \qquad 0 < x < l, \qquad 0 < t \le T,$$

from a given function $u(x,t)$.

Exercise 6.5 Construct an iterative local-regularization algorithm making it possible to identify the right-hand side of the parabolic equation (6.45) from an approximately given solution of the boundary value problem (6.45)–(6.47).

Exercise 6.6 Using the program PROBLEM6, numerically examine the rate of convergence of the approximate solution of the inverse problem to the exact solution as dependent on the input-data inaccuracy.

Exercise 6.7 Consider the first boundary value problem for the loaded parabolic equation:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) + c(x)u(x^*, t) + f(x, t), \qquad 0 < x < l, \qquad 0 < x^* < l,$$
$$u(0, t) = 0, \qquad u(l, t) = 0, \qquad 0 \le t \le T,$$
$$u(x, 0) = u_0(x), \qquad 0 < x < l.$$

In this problem, formulate the applicability conditions for the maximum principle and, on this basis, establish solution uniqueness.

Exercise 6.8 Examine the convergence of the solution of the difference problem (6.98)–(6.100) to the solution of the differential problem (6.94), (6.95).

Exercise 6.9 Based on calculations performed on a sequence of progressively refined grids and using the program PROBLEM7, examine the accuracy in reconstructing the time-dependent right-hand side of parabolic equation (6.84), (6.88) under conditions (6.86), (6.87), (6.89).

Exercise 6.10 Examine the rate of convergence of the difference scheme (6.127), (6.128) as applied to problem (6.111)–(6.114).

Exercise 6.11 Suppose that we solve the equation

$$Ay = f$$

in which

$$A = A_0 + A_1, \quad A_0 = A_0^* > 0, \quad A_1 = -A_1^*.$$

To modify the initial problem, we use symmetrization:

$$\tilde{A}y = \tilde{f},$$

where

$$\tilde{A} = A^* A_0^{-1} A, \qquad \tilde{f} = A^* A_0^{-1} f.$$

We use the iteration method

$$A_0 \frac{y_{k+1} - y_k}{\tau_{k+1}} + \tilde{A} y_k = \tilde{f}, \qquad k = 0, 1, \ldots.$$

Examine the rate of convergence in this method under the conditions

$$\|A_1 y\|^2 \le M(y, A_0 y), \qquad M = \text{const} > 0.$$

Exercise 6.12 Using the program PROBLEM8, numerically examine the convergence of the right-hand side as reconstructed in different norms.

Exercise 6.13 Consider the inverse problem (6.144)–(6.146) under the condition that

$$f(\boldsymbol{x}) = \varphi(x_1^2 + x_2^2).$$

In which case the solution of this problem is unique?

Exercise 6.14 Obtain an a priori estimate for the solution of the difference problem (6.162), (6.163).

Exercise 6.15 Propose procedures for preliminary treatment of noisy input data in the solution of the right-hand side identification problem (6.144), (6.145), (6.147), (6.150) making it possible to obtain possibly more smooth solution. Modify the program PROBLEM9 to practically examine the possibilities offered by these procedures.

7 Evolutionary inverse problems

In the present chapter, inverse problems for non-stationary mathematical physics equations are considered whose specific feature consists in that the initial state of the system is not specified. A most typical example here is given by the inverted-time problem for the second-order parabolic equation. Such problems belong to the class of problems ill-posed in the classical sense; they can be approximately solved by various versions of main regularizing algorithms. Among the latter algorithms, variational methods can be identified in which non-locally perturbed initial conditions are used. In the second class of methods, perturbed equations are used for which a well-posed problem can be posed (generalized inverse method). Regularizing algorithms enabling the solution of unstable evolutionary problems can be constructed using the regularization principle, a general guiding principle that makes it possible to obtain operator-difference schemes of desired quality. Of great potential are iterative algorithms for solving evolutionary inverse problems, capable of adequately taking into account specific features of particular problems and based on successive solution of several direct problems. The consideration is performed on the differential and difference levels. Numerical results for model problems are cited to illustrate theoretical results.

7.1 Non-local perturbation of initial conditions

In this section, we consider the inverse inverted-time problem for the parabolic equation. To approximately solve the problem, we use a regularizing algorithm with non-locally perturbed initial conditions. A close relation between this algorithm and variational solution methods for such ill-posed problems is established.

7.1.1 Problem statement

As a model problem, consider the inverted-time problem for the one-dimensional parabolic equation. In the rectangle

$$\overline{Q}_T = \overline{\Omega} \times [0, T], \qquad \overline{\Omega} = \{x \mid 0 \le x \le l\}, \qquad 0 \le t \le T$$

the function $u(x, t)$ satisfies the equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big) = f(x, t), \qquad 0 < x < l, \quad 0 < t \le T, \tag{7.1}$$

with $k(x) \ge \kappa > 0$. We restrict the present consideration to the boundary value problem with the first-kind homogeneous boundary conditions

$$u(0, t) = 0, \qquad u(l, t) = 0, \qquad 0 < t \le T \tag{7.2}$$

and with the initial condition

$$u(x, 0) = u_0(x), \qquad 0 \le x \le l. \tag{7.3}$$

We can conveniently consider problem (7.1)–(7.3) as a Cauchy problem for the second-order differential-operator equation. For functions defined in the domain $\Omega = (0, 1)$ and vanishing at the boundary $\partial\Omega$, we define a Hilbert space $\mathcal{H} = \mathcal{L}_2(\Omega)$, in which the scalar product is defined as

$$(v, w) = \int_\Omega v(x)w(x)\,dx.$$

For the norm in $\mathcal{H}$, we use the usual setting

$$\|v\| = (v, v)^{1/2} = \Big(\int_\Omega v^2(x)\,dx\Big)^{1/2}.$$

On the set of functions satisfying the boundary conditions (7.2), we define the operator

$$\mathcal{A}u = -\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l. \tag{7.4}$$

The operator $\mathcal{A}$ is a positive definite self-adjoint operator in $\mathcal{H}$:

$$\mathcal{A}^* = \mathcal{A} \ge mE, \qquad m > 0. \tag{7.5}$$

Equation (7.1), supplemented with conditions (7.2) at the boundary, is written as a differential-operator equation for the function $u(t) \in \mathcal{H}$:

$$\frac{du}{dt} - \mathcal{A}u = f(t), \qquad 0 < t \le T. \tag{7.6}$$

The initial condition (7.3) gives

$$u(0) = u_0. \tag{7.7}$$

Problem (7.6), (7.7) is an ill-posed problem because continuous dependence on input data, namely, on the initial conditions, is lacking here.

7.1.2 General methods for solving ill-posed evolutionary problems

The evolutionary inverse problem (7.6), (7.7) can be solved using, in this or another version, all the above-discussed approaches to solving ill-posed problems which were previously considered as applied to the first-kind operator equation

$$Au = f.$$

The right-hand side of the latter equation is given with some inaccuracy and, in addition,

$$\|f_\delta - f\| \le \delta.$$

In the Tikhonov method, the approximate solution u_α is to be found as follows:

$$J_\alpha(u_\alpha) = \min_{v \in H} J_\alpha(v),$$

$$J_\alpha(v) = \|Av - f_\delta\|^2 + \alpha\|v\|^2.$$

As applied to the evolutionary problem (7.6), (7.7), the Tikhonov method corresponds to the solution of an optimal control problem for equation (7.6). Here, the point of interest concerns the start control and the final observation.

The optimal control problem for (7.6), (7.7) can be formulated as follows. Let $v \in H$ be an unconstrained control. Define the function $u_\alpha = u_\alpha(t; v)$ as the solution of the following well-posed problem:

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha = f(t), \qquad 0 < t < T, \tag{7.8}$$

$$u_\alpha(T; v) = v. \tag{7.9}$$

We adopt the following simplest form of the quadratic quality functional (smoothing functional):

$$J_\alpha(v) = \|u_\alpha(0; v) - u_0\|^2 + \alpha\|v\|^2. \tag{7.10}$$

The optimal control w can be found as a minimum of $J_\alpha(v)$:

$$J_\alpha(w) = \min_{v \in H} J_\alpha(v), \tag{7.11}$$

and the related solution $u_\alpha(t, w)$ of problem (7.8), (7.9) can be considered as an approximate solution of the ill-posed problem (7.6), (7.7).

The variational problem (7.8)–(7.11) can be approximately solved by various numerical methods. In gradient methods, the minimizing sequence $\{v_k\}$, $k = 1, 2, \ldots, K$ can be constructed by the rule

$$v_{k+1} = v_k + \gamma_k p_k,$$

where p_k is the descend direction and γ_k is the descend parameter. In the simplest case, the descend direction can be directly related with the gradient of (7.10): $p_k = -\operatorname{grad} J_\alpha(v_k)$. An unusual feature here consists in the necessity to calculate a functional gradient. That is why in the solution of applied inverse problems by variational methods the latter point deserves special attention.

In the Tikhonov method, instead of the extremum problem, we solve a related Euler equation. In the latter case, the approximate solution is to be found by solving the second-kind equation

$$A^*Au_\alpha + \alpha u_\alpha = A^* f_\delta.$$

Thus, the transition to a well-posed problem is performed by passing to a problem with the self-adjoint operator A^*A, which can be done by multiplying the initial equation

from the left by A^*, and by subsequent perturbation of the resulting equation with an operator αE.

Note the class of approximate solution methods for ill-posed evolutionary problems of type (7.6), (7.7) based on the passage to a perturbed well-posed problem; these methods are known as *generalized inverse methods*. Consider briefly major variants of the generalized inverse method for problem (7.6), (7.7) in which the perturbed initial equation (7.6) is used.

In the classical variant of the generalized inverse method, the approximate solution $u_\alpha(t)$ is to be found from the equation

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha + \alpha\mathcal{A}^*\mathcal{A}u_\alpha = 0. \tag{7.12}$$

In the class of bounded solutions, one can establish the regularizing properties, namely, the convergence of the approximate solution to the exact solution in the class of bounded solutions.

Also, a variant of the generalized inverse method for stable solution of problem (7.6), (7.7) with $\mathcal{A} = \mathcal{A}^*$ can be applied; in this method, the following pseudo-parabolically perturbed equation is treated:

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha + \alpha\mathcal{A}\frac{du_\alpha}{dt} = 0. \tag{7.13}$$

In the approximate solution of inverse mathematical physics problems (problem (7.1)–(7.3)), the first variant of the generalized inverse method (see (7.12)) is based on raising the order of the differential operator over space (instead of $\mathcal{A}$, the operator $\mathcal{A}^*\mathcal{A}$ is used). In the second variant of the generalized inverse method (see (7.13)), the problem suffers no such dramatic changes.

Some other possibilities in the regularization of the problem at the expense of additional terms deserve mention. In the context of the general theory of approximate solution of ill-posed problems, these variants border on the variant of simplified regularization in which in the problem with $A = A^* \geq 0$ to be solved is the equation

$$Au_\alpha + \alpha u_\alpha = f_\delta,$$

i. e., here, one can restrict himself to perturbing the operator of the initial problem only.

7.1.3 Perturbed initial conditions

A most significant class of approximate methods for solving ill-posed evolutionary problems is based on using perturbed initial conditions instead of passing to some new equation like in the ordinary variant of the generalized inverse method. This approach can appear to be more reasonable in problems with initial conditions specified with an inaccuracy.

In methods with perturbed initial conditions, the extremum formulation of problems has found a most widespread use. The optimum control problem for systems governed by the evolutionary equations of interest can be solved using these or those regularization methods. To this class of approximate solution methods for unstable evolutionary problems, methods with non-locally perturbed initial conditions can be assigned. In the latter case, the regularizing effect is achieved due to the established relation between the initial solution and the end-time solution.

Special attention should be paid to the equivalence between the extremum formulation of ill-posed evolutionary problems and non-local problems. The latter allows us to perform a uniform consideration of non-local perturbation methods and extremum solution methods for evolutionary problems. Moreover, the equivalence between these methods makes it possible to construct computational algorithms based on this or another formulation of the problem. For instance, in some cases, instead of solving a related functional-minimization problem, one can use simple computational algorithms for solving non-local difference problems. Nonetheless, opposite examples are also known in which construction of computational algorithms around an extremum formulation is more preferable.

To approximately solve the ill-posed problem (7.6), (7.7) (with $f(t) = 0$), apply the method with non-locally perturbed initial condition. We find the approximate solution $u_\alpha(t)$ as the solution of the equation

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha = 0, \qquad 0 < t \le T \tag{7.14}$$

with the initial condition (7.7) replaced with the simplest non-local condition

$$u_\alpha(0) + \alpha u_\alpha(T) = u_0. \tag{7.15}$$

Here, the regularization parameter α is positive ($\alpha > 0$).

Let us derive estimates for the solution of the non-local problem with regard to the above-formulated constraint on the operator $\mathcal{A}$ ($\mathcal{A} = \mathcal{A}^* > 0$). Of primary concern here is stability of the approximate solution $u_\alpha(t)$ with respect to initial data.

Our consideration is based on using the expansion of the solution in eigenfunctions of $\mathcal{A}$. Not restricting ourselves to the case of (7.4), (7.5), we denote as $\mathcal{A}$ a linear constant (t-independent) operator with a domain of definition $\mathcal{D}(\mathcal{A})$, dense in $\mathcal{H}$. We assume that the operator $\mathcal{A}$ is positive definite self-adjoint in $\mathcal{H}$; generally speaking, this operator is an unbounded operator. For simplicity, we assume that the spectrum of $\mathcal{A}$ is discrete, consisting of eigenvalues $0 < \lambda_1 \le \lambda_2 \le \cdots$, and the system of eigenfunctions $\{w_k\}$, $w_k \in \mathcal{D}(\mathcal{A})$, $k = 1, 2, \ldots$ is an orthonormal complete system in $\mathcal{H}$. That is why for each $v \in \mathcal{H}$ we have:

$$v = \sum_{k=1}^{\infty} v_k w_k, \qquad v_k = (v, w_k).$$

Theorem 7.1 *For the solution of problem* (7.14), (7.15) *the following estimates are valid:*

$$\|u_\alpha(t)\| \le \frac{1}{\alpha}\,\|u_0\|, \tag{7.16}$$

$$\|u_\alpha(t)\| \ge \frac{1}{1+\alpha}\|u_0\|. \tag{7.17}$$

Proof. To prove the above theorem, we write the solution of problem (7.14), (7.15) as

$$u_\alpha = R(t,\alpha)u_0. \tag{7.18}$$

In problem (7.14), (7.15), the operator $R(t,\alpha)$ can be written as

$$R(t,\alpha) = \exp(\mathcal{A}t)(E + \alpha\exp(\mathcal{A}T))^{-1}. \tag{7.19}$$

In view of (7.18) and (7.19), the solution of the non-local problem admits the following representation:

$$u_\alpha(t) = \sum_{k=1}^{\infty}(u_0, w_k)\exp(\lambda_k t)(1+\alpha\exp(\lambda_k T))^{-1}w_k. \tag{7.20}$$

With (7.20), we have:

$$\|u_\alpha(t)\|^2 = \sum_{k=1}^{\infty}(u_0, w_k)^2\exp(2\lambda_k)(1+\alpha\exp(\lambda_k T))^{-2}.$$

From here, with allowance for $\lambda_k > 0,\ k = 1, 2, \ldots$, the desired inequalities (7.16), (7.17) readily follow. □

Remark 7.2 Estimate (7.16) guarantees stability of the solution with respect to initial data (upper estimate), and inequality (7.17) gives a lower estimate of the solution of the non-local problem.

Remark 7.3 Note that the lower estimate (7.17) can be obtained directly from (7.14), (7.15) under an assumption that the time-independent operator $\mathcal{A}$ is not necessarily a self-adjoint yet non-negative operator ($\mathcal{A} \ge 0$). To derive this estimate, we scalarwise multiply equation (7.14) by $u_\alpha(t)$ to obtain the following estimate typical of the parabolic equation with inverted time:

$$\|u_\alpha(t)\| \le \|u_\alpha(T)\|. \tag{7.21}$$

With (7.21), the non-local condition (7.15) yields

$$\|u_0\| = \|u_\alpha(0) + \alpha u_\alpha(T)\| \le \|u_\alpha(0)\| + \alpha\|u_\alpha(T)\| \le (1+\alpha)\|u_\alpha(T)\|.$$

We have proved that the problem with conditions (7.14), (7.15), non-local over time, is a well-posed problem. Now, we can discuss a more fundamental matter, namely, under which conditions solution of problem (7.14), (7.15) gives an approximate solution of the ill-posed problem (7.6), (7.7) (with $f(t) = 0$).

7.1.4 Convergence of approximate solution to the exact solution

Let us formulate a statement concerning the convergence of the approximate solutions to the exact solution of problem (7.6), (7.7). Suppose that the initial condition (7.7) is given with an inaccuracy

$$\|u_0^\delta - u_0\| \le \delta. \tag{7.22}$$

Instead of the non-local condition (7.15), we consider the condition

$$u_\alpha(0) + \alpha u_\alpha(T) = u_0^\delta. \tag{7.23}$$

Theorem 7.4 *Suppose that for the initial-condition inaccuracy the estimate* (7.22) *is valid. Then, the approximate solution* $u_\alpha(t)$ *defined as the solution of problem* (7.14), (7.23) *converges, with* $\delta \to 0$, $\alpha(\delta) \to 0$, $\delta/\alpha \to 0$, *to the exact (bounded in* $\mathcal{H}$*) solution* $u(t)$ *of problem* (7.6), (7.7) *with* $f(t) = 0$.

Proof. In the operator form, the approximate solution $u_\alpha(t)$ obtained with the inaccurately given initial condition u_0^δ can be written as

$$u_\alpha(t) = R(t, \alpha) u_0^\delta. \tag{7.24}$$

In the latter notation, the operator $R(t, 0)$ gives the exact solution of the problem and, therefore, we have $u(t) = R(t, 0)u_0$. In the adopted notation, for the inaccuracy v we obtain:

$$\begin{aligned} \|u_\alpha(t) - u(t)\| &= \|R(t, \alpha)(u_0^\delta - u_0) - (R(t, \alpha) - R(t, 0))u_0\| \\ &\le \|R(t, \alpha)\| \, \|u_0^\delta - u_0\| + \|(R(t, \alpha) - R(t, 0))u_0\|. \end{aligned} \tag{7.25}$$

The first term in (7.25) refers to stability with respect to initial data, or to the boundedness of $R(t, \alpha)$. The second term in the right-hand side of (7.25) requires that the solution $R(t, \alpha)u_0$ of the perturbed problem with exact input data be close to the exact solution $u(t)$. It is with this aim that a certain class of solutions (well-posedness class) was isolated with the used regularizing operator $R(t, \alpha)$.

In view of the derived estimate (7.16) for stability with respect to initial data and estimate (7.22), we have:

$$\|R(t, \alpha)\| \, \|u_0^\delta - u_0\| \le \frac{1}{\alpha} \delta. \tag{7.26}$$

Consider the second term in the right-hand side of (7.25). With (7.19), for the approximate solution found from (7.14), (7.23) we obtain the representation

$$\|u_\alpha(t)\|^2 = \sum_{k=1}^{\infty} (u_0^\delta, w_k)^2 \exp(2\lambda_k t)(1 + \alpha \exp(\lambda_k T))^{-2}. \tag{7.27}$$

By virtue of (7.27), we have:

$$\chi(t) \equiv \|(R(t,\alpha) - R(t,0))u_0\|^2 = \sum_{k=1}^{\infty} \exp(2\lambda_k t)(1 - (1 + \alpha \exp(\lambda_k T))^{-1})^2 (u_0, w_k)^2. \quad (7.28)$$

We consider stability in the class of solutions bounded in $\mathcal{H}$:

$$\|u(t)\| \le M, \qquad 0 \le t \le T. \quad (7.29)$$

In view of (7.29), for any $\varepsilon > 0$ there exists a number $r(\varepsilon)$ such that

$$\sum_{k=r(\varepsilon)+1}^{\infty} \exp(2\lambda_k t)(u_0, w_k)^2 \le \frac{\varepsilon}{8}.$$

Hence, from (7.28) we obtain:

$$\begin{aligned}\chi(t) &\le \sum_{k=1}^{r(\varepsilon)} \exp(2\lambda_k t)(1 - (1 + \alpha \exp(\lambda_k T))^{-1})^2 (u_0, w_k)^2 \\ &\qquad + \sum_{k=r(\varepsilon)+1}^{\infty} \exp(2\lambda_k t)(u_0, w_k)^2 \\ &\le M^2 \sum_{k=1}^{r(\varepsilon)} (1 - (1 + \alpha \exp(\lambda_k T))^{-1})^2 + \frac{\varepsilon^2}{8}.\end{aligned}$$

For any $r(\varepsilon)$, there exists a number α_0 such that

$$M^2 \sum_{k=1}^{r(\varepsilon)} (1 - (1 + \alpha \exp(\lambda_k T))^{-1})^2 \le \frac{\varepsilon^2}{8}$$

if $\alpha \le \alpha_0$.

Hence, the representation (7.29) yields:

$$\|(R(t,\alpha) - R(t,0))u_0\| \le \frac{\varepsilon}{2}. \quad (7.30)$$

With (7.26) and (7.30), from (7.25) we obtain the estimate

$$\|u_\alpha(t) - u(t)\| \le \frac{1}{\alpha}\delta + \frac{\varepsilon}{2}. \quad (7.31)$$

For an arbitrary $\varepsilon > 0$, there exists a number $\alpha = \alpha(\delta) \le \alpha_0$ and a sufficiently small number $\delta(\varepsilon)$ such that $\delta/\alpha \le \varepsilon/2$. Hence, the estimate (7.31) assumes the form

$$\|u_\alpha(t) - u(t)\| \le \varepsilon.$$

Thus, the approximate solution converges to the exact solution. □

To establish the fact that the approximate solution converges to the exact solution, it suffices for us to assume that the exact solution is bounded in $\mathcal{H}$ (see (7.29)). One can expect that under stronger assumptions about the smoothness of the exact solution we will be able to obtain an estimate for the rate of convergence. We discussed such a situation in detail when we considered methods for the stable solution of first-kind equations. Similar possibilities in the approximate solution of the ill-posed Cauchy problem for the first-order evolutionary equation (7.6), (7.7) also deserve mention.

A natural narrowing of the class of exact solutions of problem (7.6), (7.7) is

$$\|\mathcal{A}u(t)\| \leq M,$$

or

$$\|\mathcal{A}^2 u(t)\| \leq M$$

for all $t \in [0, T]$. Yet, this does not allow us to directly derive an estimate for the rate of convergence of the approximate solution obtained by (7.14), (7.23) to the exact solution of the problem.

Under conditions of the above theorem, we assume that the exact solution is bounded over the double time interval $[0, 2T]$ and consider the convergence of the approximate solution found as the solution of the non-local problem (7.14), (7.23) only at times $t \in [0, T]$. Suppose that the a priori conditions are

$$\|u(t)\| \leq M, \qquad 0 \leq t \leq 2T. \tag{7.32}$$

With the latter conditions (see (7.28)), we have:

$$\chi(t) = \alpha^2 \sum_{k=1}^{\infty} \exp(2\lambda_k(t+T))(1+\alpha\exp(\lambda_k T))^{-2}(u_0, w_k)^2 \;\leq\; M^2\alpha^2.$$

For the inaccuracy, we obtain the explicit estimate

$$\|u_\alpha(t) - u(t)\| \leq \frac{\delta}{\alpha} + \alpha M. \tag{7.33}$$

It makes sense to give here an estimate for the rate of convergence under constraints on the exact solution less tight than (7.32). Let for the solution of problem (7.6), (7.7) we have:

$$\|u(t)\| \leq M, \qquad 0 \leq t \leq (1+\theta)T, \quad 0 < \theta \leq 1. \tag{7.34}$$

Under the conditions of (7.34), we obtain:

$$\begin{aligned}
\chi(t) &= \alpha^2 \sum_{k=1}^{\infty} \exp(2\lambda_k(t+\theta T)) \\
&\qquad \times \exp(2\lambda_k((1-\theta)T))(1+\alpha\exp(\lambda_k T))^{-2}(u_0, w_k)^2 \\
&\leq M^2\alpha^2 \max_{1\leq k<\infty} \exp(2\lambda_k((1-\theta)T))(1+\alpha\exp(\lambda_k T))^{-2}.
\end{aligned}$$

We denote $\eta = \exp(\lambda_k T)$ and consider the maximum of the function

$$\varphi(\eta) = \frac{\eta^{(1-\theta)}}{1+\alpha\eta}, \qquad \eta > 0, \quad 0 < \theta \le 1.$$

This maximum, attained at

$$\eta_{\text{opt}} = \frac{\gamma}{\alpha}, \qquad \gamma = \frac{1-\theta}{\theta},$$

is equal to

$$\varphi(\eta_{\text{opt}}) = \alpha^{\theta-1}\,\frac{\gamma^{1-\theta}}{1+\gamma}.$$

In view of this, we have

$$\chi(t) \;\le\; M^2\alpha^{2\theta}\,\frac{\gamma^{2(1-\theta)}}{(1+\gamma)^2}$$

and, therefore, under the a priori assumptions (7.34) about the exact solution we have the following estimate for the inaccuracy:

$$\|u_\alpha(t) - u(t)\| \;\le\; \frac{\delta}{\alpha} + \alpha^\theta\,\frac{\gamma^{1-\theta}}{1+\gamma}\,M. \tag{7.35}$$

The estimate (7.35) shows that the accuracy of the approximate solution $u_\alpha(t)$ depends on the smoothness of the exact solution of problem (7.6), (7.7).

Let us briefly mention here the possibility of using non-local conditions more general than (7.23). Suppose that, for instance, instead of (7.23) we use the conditions

$$u_\alpha(0) + \alpha\mathcal{S}u_\alpha(T) = u_0^\delta. \tag{7.36}$$

The operator $\mathcal{S}$ is assumed to be a self-adjoint, positively defined operator.

Not dwelling on the formulation of general results, note a special case of non-local condition (7.36) with $\mathcal{S} = \mathcal{A}$. Under such conditions, it is an easy matter to derive the following stability estimate in $\mathcal{H}_\mathcal{D}$ for the solution of the non-local problem (7.14), (7.36):

$$\|u_\alpha(t)\|_\mathcal{D} \le \frac{1}{\alpha}\,\|u_0\|. \tag{7.37}$$

Here, $\mathcal{D} = \mathcal{A}^2$. Thus, stability takes place in the case of less smooth initial conditions. In $\mathcal{H}_\mathcal{D}$, one can also prove convergence of the approximate solution to the exact solution (an analogue of Theorem 7.1) under the a priori assumption that the exact solution is bounded in the same space.

7.1.5 Equivalence between the non-local problem and the optimal control problem

Let us dwell now on the equivalence between the non-local problem and the optimal control problem. We denote the control as $v \in \mathcal{H}$ (there are no constraints imposed on the control). We define $u_\alpha = u_\alpha(t; v)$ as the solution of the well-posed problem

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha = 0, \qquad 0 \le t < T, \tag{7.38}$$

$$u_\alpha(T; v) = v. \tag{7.39}$$

We specify the quadratic quality functional (cost function) in the form

$$J_\alpha(v) = \|u_\alpha(0; v) - u_0^\delta\|^2 + \alpha(\mathcal{S}v, v), \tag{7.40}$$

where $\mathcal{S}$ is a self-adjoint, positively defined operator. The choice of (7.40) refers to the start observation. The optimal control w is determined as the minimum of the functional $J_\alpha(v)$:

$$J_\alpha(w) = \min_{v \in \mathcal{H}} J_\alpha(v), \tag{7.41}$$

and the related solution $u_\alpha(t; w)$ of problem (7.38), (7.39) is considered as an approximate solution of the ill-posed problem (7.6), (7.7) ($f(t) = 0$).

Let us show that, under certain conditions, the non-local problem (7.14), (7.36) presents an Euler equation for the functional (7.40) on the set of constraints given by (7.38), (7.39). To this end, we conveniently introduce the conjugate state. We introduce the setting

$$(u, p)^* = \int_0^T (u, p)\, dt.$$

For (7.38), in view of the self-adjointness of $\mathcal{A}$, we have:

$$\begin{aligned} \left(\left(\frac{dy}{dt} - \mathcal{A}y\right), p\right)^* &= \left(\frac{dy}{dt}, p\right)^* - (y, \mathcal{A}p)^* \\ &= (y(T), p(T)) - (y(0), p(0)) - \left(y, \frac{dy}{dt}\right)^* - (y, \mathcal{A}p)^*. \end{aligned} \tag{7.42}$$

With (7.42), we assume that the function $p(t)$ is determined from the equation

$$\frac{dp}{dt} + \mathcal{A}p = 0, \qquad 0 \le t < T. \tag{7.43}$$

The initial condition for (7.43) will be chosen below.

For the optimal control w of interest, the following Euler equation holds:

$$(u_\alpha(0; w) - u_0^\delta, u_\alpha(0; v) - u_\alpha(0; w)) + \alpha(\mathcal{S}w, v - w) = 0, \quad \forall\, v \in \mathcal{H}. \tag{7.44}$$

We take into consideration the equation

$$(u_\alpha(T; v),\, p(T)) = (u_\alpha(0; v),\, p(0)),$$

which follows from (7.38), (7.42) and (7.43), and assume that

$$p(0) = u_\alpha(0) - u_0^\delta, \tag{7.45}$$

where $u_\alpha(t) = u_\alpha(t; w)$. Then, from (7.44) we obtain:

$$(p(T) + \alpha \mathcal{S} w,\, v - w) = 0, \qquad \forall\, v \in H.$$

This yields

$$p(T) + \alpha \mathcal{S} u_\alpha(T) = 0. \tag{7.46}$$

Equality (7.46) is well known in control theory for parabolic equations.

Thus, the problem of finding the optimal control w and the related solution $u_\alpha(t)$ is reduced to the solution of system (7.38), (7.43) with correlatons (7.45), (7.46).

For the time-independent operator $\mathcal{S}$, permutable with $\mathcal{A}$, there holds the equation

$$p(T - \theta) + \alpha \mathcal{S} u_\alpha(T + \theta) = 0 \tag{7.47}$$

with a constant θ. Here, the functions $p(t)$ and $u_\alpha(t)$ satisfy respectively equations (7.43) and (7.14), being correlated at $t = T$ with relations (7.46).

Indeed, assuming that $g(t) = -\alpha \mathcal{S} u_\alpha(t)$, we rewrite equation (7.46) in the form

$$p(T) = g(T). \tag{7.48}$$

In the case of $\mathcal{S} \neq \mathcal{S}(t)$ and $\mathcal{S}\mathcal{A} = \mathcal{A}\mathcal{S}$, the function $g(t)$ satisfies the equation

$$\frac{dg}{dt} - \mathcal{A} g = 0. \tag{7.49}$$

Equation (7.49) coincides with equation (7.43) for the function $p(t)$ with sign change applied to the variable t. With equation (7.48) taken into account, we obtain that

$$p(T - \theta) = g(T + \theta)$$

for any θ. It is from here that the desired correlation (7.47) between the solution of the conjugate equation (7.43) and the solution of the optimal control problem follows.

We assume that in equality (7.47) we have $\theta = T$; then, we arrive at the equality

$$p(0) + \alpha \mathcal{S} u(2T) = 0;$$

with the initial condition (7.45), this equality assumes the form

$$u_\alpha(0) + \alpha \mathcal{S} u_\alpha(2T) = u_0^\delta. \tag{7.50}$$

In this way, we eliminate the auxiliary function $p(t)$, which defines the conjugate state, and arrive at a non-local condition for the approximate solution. The latter allows the following statement to be formulated:

Theorem 7.5 *Let self-adjoint, time-independent and positively defined operators $\mathcal{S}$ and $\mathcal{A}$ be mutually permutable operators. Then, the solution of the variational problem* (7.38)–(7.41) *satisfies the equation*

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha = 0, \qquad 0 < t \le 2T \tag{7.51}$$

and the non-local conditions (7.50).

Equations (7.50) and (7.51) are the Euler equations for the variational problem (7.38)–(7.41). The established relation between the non-local problem and the optimal control problem is a very useful relation of utmost significance. This statement can be supported by a line of reasoning similar to that used for the Euler equations in classical variational problems.

In the above consideration of ill-posed evolutionary problems of type (7.6), (7.7) we showed the equivalence of the two different approaches to finding the approximate solution. The methods based, first, on non-local perturbation of initial conditions and, second, on using the extremum formulation of the problem give rise to the same regularizing algorithms. A fundamental difference between the two approaches consists only in the manner in which the approximate solution is obtained, i. e., in the computational realization.

7.1.6 Non-local difference problems

Let us show that, under certain conditions, statements analogous to Theorems 7.1 and 7.5 are valid for the difference analogues of the non-local problem (7.14), (7.15). Here, we restrict ourselves to the consideration of time approximations. We routinely approximate the operator $\mathcal{A}$ with a difference operator, self-adjoint and positive in the corresponding mesh space. We introduce a uniform grid over the variable t,

$$\bar{\omega}_\tau = \omega_\tau \cup \{T\} = \{t_n = n\tau,\ n = 0, 1, \dots, N_0,\ \tau N_0 = T\},$$

with a step size $\tau > 0$. As usually, we denote the approximate solution of the non-local problem (7.14), (7.15) at $t = t_n$ as y_n. To find this solution, we use the simplest explicit scheme, often more preferable than implicit schemes for unstable problems. From (7.14) and (7.15), we obtain:

$$\frac{y_{n+1} - y_n}{\tau} - \mathcal{A}y_n = 0, \qquad n = 0, 1, \dots, N_0 - 1, \tag{7.52}$$

$$y_0 + \alpha y_{N_0} = u_0. \tag{7.53}$$

For the solution of the difference equation (7.52), we have the following representation:

$$y_n = \sum_{k=1}^{\infty} (1 + \tau\lambda_k)^n (y_0, w_k) w_k, \qquad n = 0, 1, \dots, N_0. \tag{7.54}$$

Starting from (7.54), we can easily establish the following direct analogue of Theorem 7.1.

Theorem 7.6 *For the solution of the non-local operator-difference problem* (7.52), (7.53)*, there hold the estimates*

$$\|y_n\| \le \frac{1}{\alpha}\,\|u_0\|, \qquad n = 0, 1, \dots, N_0, \tag{7.55}$$

and

$$\|y_{N_0}\| \ge \frac{1}{1+\alpha}\,\|u_0\|. \tag{7.56}$$

Proof. From (7.53) and (7.54), the following representation for the solution of the non-local problem (7.52), (7.53) can be obtained:

$$y_n = \sum_{k=1}^{\infty}(1+\tau\lambda_k)^n(1+\alpha(1+\tau\lambda_k)^{N_0})^{-1}(u_0, w_k)w_k. \tag{7.57}$$

Since $\lambda_k > 0$, $k = 1, 2, \dots$, then for the norm in $\mathcal{H}$ the representation (7.57) yields

$$\|y_n\|^2 = \sum_{k=1}^{\infty}(1+\tau\lambda_k)^{2n}(1+\alpha(1+\tau\lambda_k)^{N_0})^{-2}(u_0, w_k)^2 \le \frac{1}{\alpha^2}\,\|u_0\|^2.$$

Thus, the estimate (7.55) is proved.

By analogy with the continuous case (see Theorem 7.1), the lower estimate (7.56) can be derived under more general assumptions on the operator $\mathcal{A}$. Suppose that $\mathcal{A} \ge 0$; then, we scalarwise multiply the difference equation (7.52) in $\mathcal{H}$ by y_n and obtain

$$(y_{n+1}, y_n) = \|y_n\|^2 + \tau(\mathcal{A}y_n, y_n) \ge \|y_n\|^2. \tag{7.58}$$

For the left-hand side of (7.58), we have

$$(y_{n+1}, y_n) \le \|y_{n+1}\|\,\|y_n\|$$

and, hence,

$$\|y_n\| \le \|y_{n+1}\| \le \cdots \le \|y_{N_0}\|. \tag{7.59}$$

Inequality (7.59) is a difference analogue of (7.21). With the non-local condition (7.53) taken into account and in view of (7.59), we obtain

$$\|u_0\| \;=\; \|y_0 + \alpha y_{N_0}\| \;\le\; \|y_0\| + \alpha\|y_{N_0}\| \;\le\; (1+\alpha)\,\|y_{N_0}\|.$$

The latter inequality yields the estimate (7.56). □

It should be emphasized here that, with the use of the explicit scheme (7.52), the above estimates for the difference solution were obtained assuming no constraints on the time step size. Simultaneously, conditional stability takes place if the weighted scheme is used.

Consider now the difference problem that corresponds to the optimal control problem (7.38), (7.39). We denote the approximate solutions of the two problems at the time t_n under the control v as $y_n(v)$ and $y_n = y_n(w)$, respectively.

To problem (7.38), (7.39), we put into correspondence the difference problem

$$\frac{y_{n+1} - y_n}{\tau} - \mathcal{A} y_n = 0, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{7.60}$$

$$y_{N_0}(v) = v. \tag{7.61}$$

In view of (7.40), the quality functional is

$$J_\alpha(v) = \|y_0(v) - u_0\|^2 + \alpha(\mathcal{S} v, v), \tag{7.62}$$

with some positive self-adjoint operator $\mathcal{S}$. The optimal control w is defined by

$$J_\alpha(w) = \min_{v \in \mathcal{H}} J_\alpha(v). \tag{7.63}$$

For the conjugate problem to be formulated, we obtain a difference analogue of formula (7.42). To this end, we consider the following extended grid:

$$\bar{\omega}_\tau^+ = \{t_n = n\tau,\ n = -1, 0, \ldots, N_0,\ \tau N_0 = T\}.$$

We use the following settings adopted in the theory of difference schemes. On the extended grid, we have:

$$\{y, v\} = \sum_{n=0}^{N_0 - 1} y_n v_n \tau, \quad \{y, v] = \sum_{n=0}^{N_0} y_n v_n \tau, \quad [y, v\} = \sum_{n=-1}^{N_0 - 1} y_n v_n \tau.$$

The difference summation-by-parts formulas yield the equation

$$\{y, v_{\bar{t}}] + [y_t, v\} = y_{N_0} v_{N_0} - y_{-1} v_{-1}, \tag{7.64}$$

where, recall,

$$y_{\bar{t}} = \frac{y_n - y_{n-1}}{\tau}, \qquad y_t = \frac{y_{n+1} - y_n}{\tau}.$$

Using (7.64) and the obvious identity

$$\{y, v] - [y, v\} = y_{N_0} v_{N_0} \tau - y_{-1} v_{-1} \tau,$$

we obtain

$$\{y, p_{\bar{t}} + \mathcal{A} p] + [y + t - \mathcal{A} y, p\} = y_{N_0}(p_{N_0} + \tau \mathcal{A} p_{N_0}) - p_{-1}(y_{-1} + \tau \mathcal{A} y_{-1}). \tag{7.65}$$

Equation (7.65) is a difference analogue of (7.42). By analogy with the continuous problem and with allowance for (7.65), we determine the conjugate state from the difference equation

$$\frac{p_n - p_{n-1}}{\tau} + \mathcal{A}p_n = 0, \qquad n = 0, 1, \ldots, N_0. \tag{7.66}$$

Then, by virtue of (7.65), we have the equation

$$y_{N_0}(p_{N_0} + \tau\mathcal{A}p_{N_0}) = p_{-1}(y_{-1} + \tau\mathcal{A}y_{-1}). \tag{7.67}$$

It readily follows from (7.66) that

$$p_{N_0} + \tau\mathcal{A}p_{N_0} = p_{N_0-1}. \tag{7.68}$$

We assume that the difference equation (7.60) is also valid for $n = -1$. Then we obtain

$$y_{-1} + \tau\mathcal{A}y_{-1} = y_0. \tag{7.69}$$

With (7.68) and (7.69), equation (7.67) can be written as

$$y_{N_0}p_{N_0-1} = y_0p_{-1}. \tag{7.70}$$

For the functional (7.62), we obtain (see (7.44))

$$(y_0(w) - u_0, y_0(v) - y_0(w)) + \alpha(\mathcal{S}w, v - w) \tag{7.71}$$

for all $v \in \mathcal{H}$. Now, we choose

$$p_{-1} = y_0(w) - u_0. \tag{7.72}$$

Then, in view of (7.70), it follows from (7.71) that

$$p_{N_0-1} + \alpha\mathcal{S}y_{N_0} = 0. \tag{7.73}$$

Let a time-independent operator $\mathcal{S}$ be permutable with the operator $\mathcal{A}$. Then, for any integer number m the equation

$$p_{N_0-1-m} + \alpha\mathcal{S}y_{N_0+m} = 0 \tag{7.74}$$

is valid provided that the functions p_n and y_n satisfy respectively equations (7.66) and (7.60) and are correlated with relations (7.73). This statement can be proved analogously to the continuous case (see (7.47)).

To pass to the non-local problem, in (7.74) we put $m = N_0$; then, with (7.72), we arrive at the condition

$$y_0 + \alpha\mathcal{S}y_{2N_0} = u_0. \tag{7.75}$$

The latter allows the following statement to be formulated:

Theorem 7.7 *Let self-adjoint, time-independent and positively defined operators S and A be mutually permutable. Then, the solution of the difference optimal control problem* (7.60)–(7.63) *satisfies the equation*

$$\frac{y_{n+1} - y_n}{\tau} - Ay_n = 0, \qquad n = 0, 1, \ldots, 2N_0 - 1, \tag{7.76}$$

supplemented with the non-local condition (7.75).

Thus, for the difference optimal control problem we have established their relation with problems with time-non-local conditions over the double time interval. Above, the relation between non-local problems and optimal control problems was established under the assumption that the initial operator A is a self-adjoint, time-independent operator positively defined in H. These restrictions are not related with the essence of the problems, and for more general optimal control problems, in a similar manner, an Euler equation in the form of a non-local problem can be constructed.

7.1.7 Program realization

For methods with non-locally perturbed initial conditions, there exists a problem with computational realization of such regularizing algorithms. It should be noted that, presently, simple and convenient computational schemes for numerical solution of general non-local boundary value problems for mathematical physics equations are lacking. In order to avoid computational difficulties that obscure the essence of the problem, below we consider a simple inverse problem for the one-dimensional parabolic equation with constant coefficients. The solution of the difference problem will be constructed using the variable separation method in the fast Fourier transform technique.

Within the framework of a quasi-real experiment, we first solve the direct, initially boundary value problem:

$$\begin{aligned}
&\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \qquad 0 < x < l, \quad 0 < t \le T, \\
&u(0, t) = 0, \qquad u(l, t) = 0, \qquad 0 < t \le T, \\
&u(x, 0) = u_0(x), \qquad 0 \le x \le l.
\end{aligned}$$

The solution of the problem at the end time ($u(x, T)$) is perturbed to subsequently use this solution as input data for the inverse problem of reconstructing the initial solution at $t = 0$.

The realization is based in the fast Fourier transform (subroutines `SINT` and `SINTI`). A more detailed description was given above (program `PROBLEM9`).

The value of the regularization parameter is chosen considering the discrepancy. The computational realization is based on using the sequence

$$\alpha_k = \alpha_0 q^k, \qquad q > 0$$

with given α_0 and q.

Program PROBLEM10

```
C
C     PROBLEM10 - PROBLEM WITH INVERTED TIME
C                 ONE-DIMENSIONAL PROBLEM
C                 NON-LOCALLY DISTURBED INITIAL CONDITION
C
      PARAMETER ( DELTA = 0.005, N = 65, M = 65 )
      DIMENSION U0(N), UT(N), UTD(N), UT1(N), U(N), U1(N)
     +         ,X(N), ALAM(N-2), W(N-2), W1(N-2), WSAVE(4*N)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF THE SEGMENT;
C     N         - NUMBER OF GRID NODES OVER SPACE;
C     TMAX      - MAXIMAL TIME;
C     M         - NUMBER OF GRID NODES OVER TIME;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     Q         - MULTIPLIER IN THE REGULARIZATION PARAMETER;
C     U0(N)     - INITIAL CONDITION TO BE RECONSTRUCTED;
C     UT(N)     - END-TIME SOLUTION OF THE DIRECT PROBLEM;
C     UTD(N)    - DISTURBED SOLUTION OF THE DIRECT PROBLEM;
C     U(N)      - APPROXIMATE SOLUTION OF THE INVERSE PROBLEM;
C
      XL   = 0.
      XR   = 1.
      TMAX = 0.1
      PI    = 3.1415926
C
      OPEN  (01, FILE = 'RESULT.DAT') !FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H = (XR - XL) / (N - 1)
      TAU = TMAX / (M-1)
      DO I = 1, N
        X(I) = XL + (I-1)*H
      END DO
C
C     DIRECT PROBLEM
C
C     INITIAL CONDITION
C
      DO I = 1, N
        U0(I) = AU(X(I))
      END DO
C
C     EIGENVALUES OF THE DIFFERENCE OPERATOR
C
      DO I = 1, N-2
        ALAM(I) = 4./H**2*(SIN(PI*I/(2.*(N-1))))**2
      END DO
C
C     FORWARD FOURIER TRANSFORM
C
      CALL SINTI(N-2,WSAVE)
      DO I = 2, N-1
```

```
        W(I-1) = U0(I)
      END DO
      CALL SINT(N-2,W,WSAVE)
C
C     FOURIER INVERSION
C
      DO I = 1, N-2
        QQ = 1. / (1. + TAU*ALAM(I))
        W1(I) = QQ**(M-1) * W(I)
      END DO
      CALL SINT(N-2,W1,WSAVE)
      DO I = 2, N-1
        UT(I) = 1./(2.*(N-1))*W1(I-1)
      END DO
      UT(1) = 0.
      UT(N) = 0.
      DO I = 1, N-2
        QQ = 1. / (1. + TAU*ALAM(I))
        W1(I) = QQ**((M-1)/2) * W(I)
      END DO
      CALL SINT(N-2,W1,WSAVE)
      DO I = 2, N-1
        UT1(I) = 1./(2.*(N-1))*W1(I-1)
      END DO
      UT1(1) = 0.
      UT1(N) = 0.
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO I = 2, N-1
        UTD(I) = UT(I) + 2.*DELTA*(RAND(0)-0.5)
      END DO
      UTD(1) = 0.
      UTD(N) = 0.
C
C     INVERSE PROBLEM
C
C     INPUT-DATA FOURIER TRANSFORM
C
      DO I = 2, N-1
        W(I-1) = UTD(I)
      END DO
      CALL SINT(N-2,W,WSAVE)
C
C     ITERATIVE PROCESS TO ADJUST THE REGULARIZATION PARAMETER
C
      IT = 0
      ITMAX = 1000
      ALPHA = 0.001
      Q = 0.75
  100 IT = IT + 1
C
C     FOURIER INVERSION
C
      DO I = 1, N-2
        QQ = 1. + TAU*ALAM(I)
        W1(I) = W(I) / (1. + ALPHA*QQ**(M-1))
      END DO

      CALL SINT(N-2,W1,WSAVE)
```

```
      DO I = 2, N-1
        U(I) = 1./(2.*(N-1))*W1(I-1)
      END DO
C
C     DISCREPANCY
C
      SUM = 0.D0

      DO I = 2,N-1
        SUM = SUM + (U(I)-UTD(I))**2*H
      END DO
      SL2 = SQRT(SUM)
C
      IF ( IT.EQ.1 ) THEN
        IND = 0


        IF ( SL2.LT.DELTA ) THEN
          IND = 1
          Q = 1.D0/Q
        END IF
         ALPHA = ALPHA*Q
        GO TO 100
      ELSE
        ALPHA = ALPHA*Q
        IF ( IND.EQ.0 .AND. SL2.GT.DELTA ) GO TO 100
        IF ( IND.EQ.1 .AND. SL2.LT.DELTA ) GO TO 100
      END IF
C
C     SOLUTION
C
      DO I = 1, N-2
        QQ = 1. + TAU*ALAM(I)
        W1(I) = QQ**((M-1)/2) * W(I) / (1. + ALPHA*QQ**(M-1))
     +        * QQ**((M-1)/2)
      END DO
      CALL SINT(N-2,W1,WSAVE)
      DO I = 2, N-1
        U(I) = 1./(2.*(N-1))*W1(I-1)
      END DO
      U(1) = 0.

      U(N) = 0.
      DO I = 1, N-2
        QQ = 1. + TAU*ALAM(I)
        W1(I) = QQ**((M-1)/2) * W(I) / (1. + ALPHA*QQ**(M-1))
      END DO
      CALL SINT(N-2,W1,WSAVE)
      DO I = 2, N-1
        U1(I) = 1./(2.*(N-1))*W1(I-1)
      END DO
      U1(1) = 0.
      U1(N) = 0.
C
      WRITE ( 01, * ) (U0(I),I=1,N)
      WRITE ( 01, * ) (UT(I),I=1,N)
      WRITE ( 01, * ) (UT1(I),I=1,N)
      WRITE ( 01, * ) (UTD(I),I=1,N)
      WRITE ( 01, * ) (U(I),I=1,N)
      WRITE ( 01, * ) (U1(I),I=1,N)
```

```
      WRITE ( 01, * ) (X(I),I=1,N)
      WRITE ( 01, * ) IT, ALPHA
      CLOSE (01)
      STOP
      END


      FUNCTION  AU ( X )
C
C      INITIAL CONDITION

C
      AU = X / 0.3
      IF (X.GT.0.3) AU = (1. - X) / 0.7
C
      RETURN
      END
```

The initial condition is set in the subroutine AU. With regard to the previous consideration of the convergence, the approximate solution is calculated at the end time ($t = T$) and at the time $t = T/2$.

7.1.8 Computational experiments

We solved a model problem whose input data were taken equal to the difference solution of the direct problem ($l = 1$, $t = 0.1$) with the initial condition

$$u_0(x) = \begin{cases} x/0.3, & 0 < x < 0.3, \\ (1-x)/0.7, & 0.3 < x < 1. \end{cases}$$

The computations were performed on a uniform grid with $h = 1/64$, $\tau = 1/640$. The results obtained by solving the direct problem are shown in Figure 7.1.

The effect due to inaccuracies was modeled by perturbing the input data (solution of the direct problem at the time $t = T$):

$$\tilde{u}(x, T) = u(x, T) + 2\delta\Big(\sigma(x) - \frac{1}{2}\Big), \qquad x \in \omega.$$

Here, $\sigma(x)$ is a random function normally distributed over the interval $[0, 1]$. Figure 7.2 shows the exact and approximate solutions of the inverse problem at two characteristic times, $t = T$ and $t = T/2$, for $\delta = 0.005$. Similar data obtained with a greater ($\delta = 0.015$) and lower ($\delta = 0.0025$) inaccuracy are shown respectively in Figures 7.3 and 7.4.

The data presented show that the solution reconstruction accuracy essentially depends on the time. The end-time solution is reconstructed inaccurately, and the convergence related with decreased input-data inaccuracy is indistinctly observed. A more favorable situation is observed at $t = T/2$. It is by this time, as it was shown previously in the theoretical consideration of the problem, that the rate of convergence of the approximations to the exact solution attains its highest.

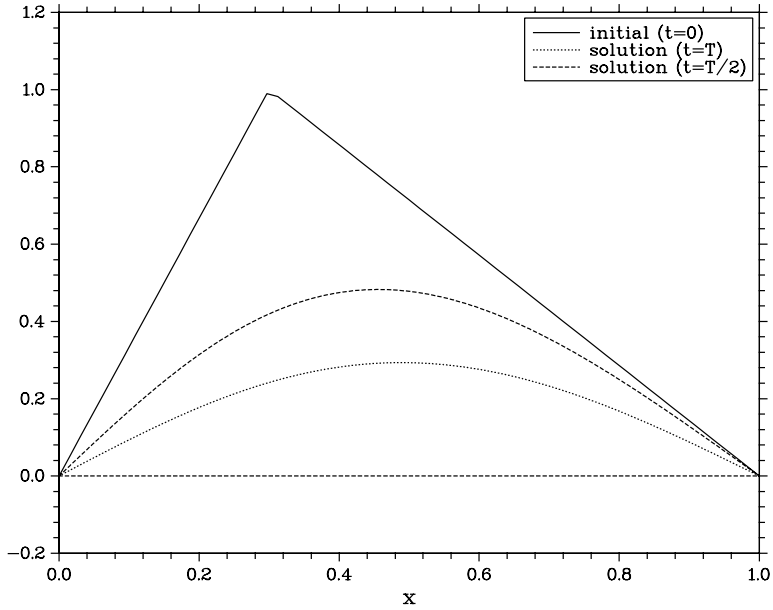

Figure 7.1 Mesh solution of the direct problem

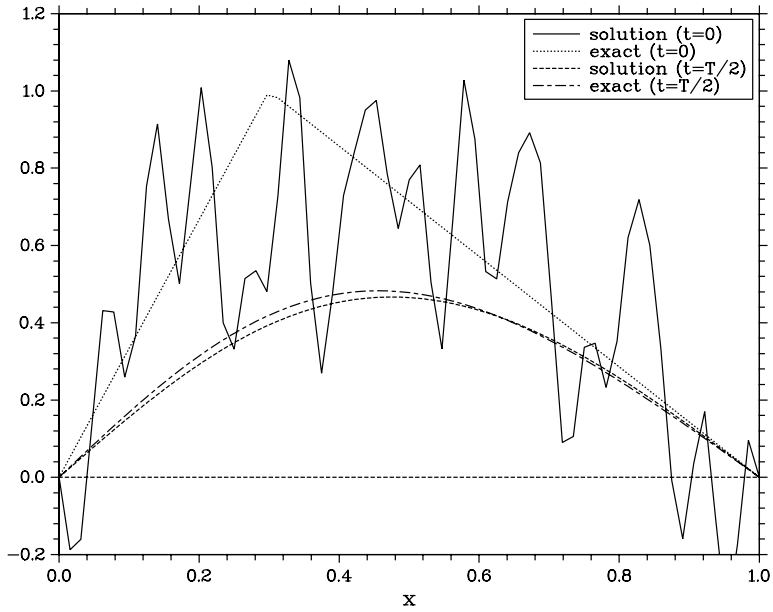

Figure 7.2 Solution of the inverse problem obtained with $\delta = 0.005$

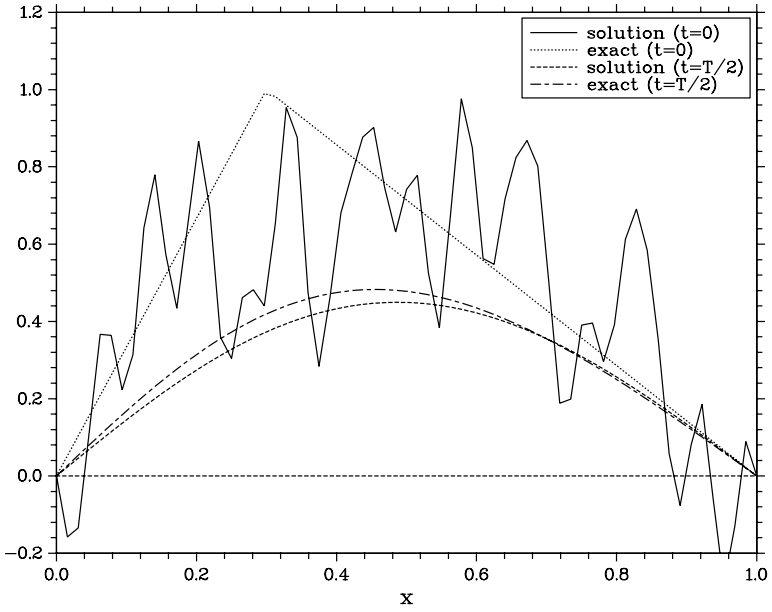

Figure 7.3 Solution of the inverse problem obtained with $\delta = 0.01$

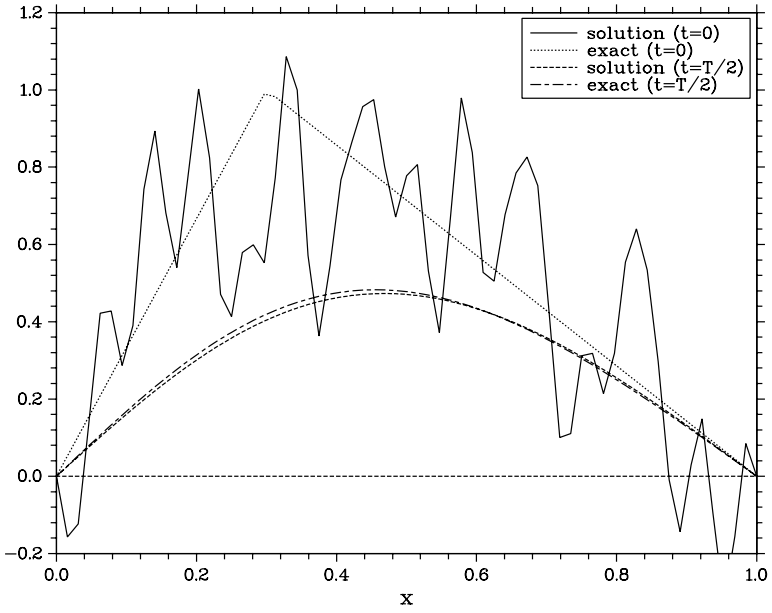

Figure 7.4 Solution of the inverse problem obtained with $\delta = 0.0025$

7.2 Regularized difference schemes

An important class of solution methods for ill-posed evolutionary problems is related with some perturbation applied to the initial equation. In the generalized inverse method, such a perturbation can be applied immediately to the initial differential problem.

In a more natural approach, no auxiliary differential problem is considered; instead, this approach uses a perturbation applied to the difference problem. In this way, regularized difference schemes can be constructed.

7.2.1 Regularization principle for difference schemes

The regularization principle for difference schemes is presently recognized as a guiding principle in improvement of difference schemes. For general two- and three-layer schemes, recipes aimed at improving the quality of the difference schemes (their stability, accuracy and efficiency) can be formulated. Using this principle, one can examine stability and convergence of a broad class of difference schemes for boundary value mathematical physics problems and develop solution algorithms for difference problems.

Traditionally, the regularization principle is widely used to develop stable difference schemes for well-posed problems involving partial differential equations. The same uniform methodological base is used to construct difference schemes for conditionally well-posed non-stationary mathematical physics problems. Weakly perturbing problem operators, one can exert control over the growth of the solution norm on passage from one to the next time layer.

Absolutely stable difference schemes can be constructed, with the help of the regularization principle, in the following manner:

1. For a given initial problem we construct some simplest difference scheme (generating difference scheme) that does not possesses the required properties; i. e., a scheme conditionally stable or even absolutely unstable.
2. We write the difference scheme in the standard (canonical) form for which stability conditions are known.
3. The properties of the difference scheme (its stability) can be improved by perturbing the operators in the difference scheme.

Thus, the regularization principle for a difference scheme is based on using already known results concerning conditional stability. Such criteria are given by the general stability theory for difference schemes. From the latter standpoint, we can consider the regularization principle as a means enabling an efficient use of general results yielded by the stability theory for difference schemes. The latter can be achieved by writing difference schemes in rather a general canonical form and by formulation of easy-to-check stability criteria.

To illustrate the considerable potential the regularization principle offers in the realization of difference schemes, we will use this general approach to construct absolutely stable schemes for direct mathematical physics problems. As a model problem, below we consider the first boundary value problem for the parabolic equation. As a generating function, consider a conditionally stable explicit scheme. For an absolutely stable scheme to be obtained, we perturb the operators in the explicit difference scheme. We separately consider the variants with additively (standardly) and multiplicative (nonstandardly) perturbed operators of the generating difference scheme.

Consider a two-dimensional problem in the rectangle

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\alpha < l_\alpha,\ \alpha = 1, 2\}.$$

In Ω, we seek the solution of the parabolic equation

$$\frac{\partial u}{\partial t} - \sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\left(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\right) = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T, \tag{7.77}$$

supplemented with the simplest homogeneous boundary conditions of the first kind:

$$u(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \partial\Omega, \quad 0 < t < T. \tag{7.78}$$

Also, an initial condition is given:

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{7.79}$$

We assume that the coefficient in (7.77) is sufficiently smooth and $k(x) \geq \kappa$, $\kappa > 0$, $\boldsymbol{x} \in \Omega$.

We put into correspondence to the differential problem (7.77)–(7.79) a differential-difference problem by performing discretization over space. We assume for simplicity that a grid with step sizes h_α, $\alpha = 1, 2$, uniform along each of the directions, is introduced in the domain Ω. As usually, ω is the set of internal nodes.

On the set of mesh functions $y(\boldsymbol{x})$ such that $y(\boldsymbol{x}) = 0$, $\boldsymbol{x} \neq \omega$, we define, by the relation

$$\Lambda y = -\sum_{\alpha=1}^{2} (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}, \tag{7.80}$$

a difference operator Λ, putting, for instance,

$$a_1(\boldsymbol{x}) = k(x_1 - 0.5h_1, x_2), \qquad a_2(\boldsymbol{x}) = k(x_1, x_2 - 0.5h_2).$$

In the Hilbert space $H = L_2(\omega)$, we introduce the scalar product and the norm by the relations

$$(y, w) = \sum_{\boldsymbol{x} \in \omega} y(\boldsymbol{x}) w(\boldsymbol{x}) h_1 h_2, \qquad \|y\| = \sqrt{(y, y)}.$$

In H, we have $\Lambda = \Lambda^* \geq mE$, $m > 0$. We pass from (7.77)–(7.79) to the differential-operator equation

$$\frac{dy}{dt} + \Lambda y = 0, \qquad 0 < t < T \tag{7.81}$$

at a given

$$y(0) = u_0, \qquad \boldsymbol{x} \in \omega. \tag{7.82}$$

Now, we are going to construct absolutely stable two-layer difference schemes for problem (7.81), (7.82) around the regularization principle.

In compliance with the regularization principle, we first choose some difference scheme to start from. As such generating a scheme, we can consider the simplest explicit scheme

$$\frac{y_{n+1} - y_n}{\tau} + \Lambda y_n = 0, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{7.83}$$

$$y_0 = u_0, \qquad \boldsymbol{x} \in \omega, \tag{7.84}$$

where $N_0\tau = T$.

We write the difference scheme (7.83), (7.84) in the canonical form for two-layer operator-difference schemes

$$B\,\frac{y_{n+1} - y_n}{\tau} + Ay_n = 0, \qquad t_n \in \omega_\tau \tag{7.85}$$

with the operators

$$B = E, \qquad A = \Lambda.$$

By Theorem 4.2, the condition

$$B \geq \frac{\tau}{2} A, \qquad t \in \omega_\tau \tag{7.86}$$

is a condition necessary and sufficient for the scheme (7.84), (7.85) to be stable in H_A or, in other words, for the estimate

$$\|y_{n+1}\|_A \leq \|u_0\|_A, \qquad t \in \omega_\tau$$

to hold.

With the inequality $\Lambda \leq \|\Lambda\| E$ taken into account, we would like to obtain, from the necessary and sufficient stability conditions (7.86), the following constraints on the time step size in the explicit scheme (7.83), (7.84):

$$\tau \leq \frac{2}{\|\Lambda\|}.$$

In the case under consideration, $\|\Lambda\| = \mathcal{O}(|h|^2)$, where $|h|^2 = h_1^2 + h_2^2$, and for the maximum admissible step size we have $\tau_0 = \mathcal{O}(|h|^2)$.

According to (7.86), improved stability can be gained in two ways. In the first case, stability can be improved due to increased energy (By, y) of the operator B (left-hand side of inequality (7.86)) or, alternatively, due to decreased energy of the operator A (right-hand side of inequality (7.86)). Consider first the opportunities related with addition of operator terms to the operators B and A. In this case, we will speak of additive regularization.

We can most naturally start from an additive perturbation applied to the operator B, i.e., from the transition $B \mapsto B + \alpha R$, where R is the regularizing operator and α is the regularization parameter. Taking the fact into account that in the generating scheme of interest we have $B = E$, we put:

$$B = E + \alpha R. \tag{7.87}$$

To retain the first approximation order in the scheme (7.85), (7.87), it suffices for us to choose $\alpha = \mathcal{O}(\tau)$.

Consider two typical choices of the regularizing operator:

$$R = \Lambda, \tag{7.88}$$

$$R = \Lambda^2. \tag{7.89}$$

We can directly establish that the regularized difference scheme (7.85), (7.87) is stable in H_A provided that $\alpha \geq \tau/2$ in the case of (7.88) and $\alpha \geq \tau^2/16$ in the case of (7.89).

The regularized scheme (7.85), (7.87), (7.88) corresponds to the case in which we use the standard weighted scheme

$$\frac{y_{n+1} - y_n}{\tau} + \Lambda(\sigma y_{n+1} + (1 - \sigma) y_n) = 0 \qquad n = 0, 1, \ldots, N_0 - 1,$$

with $\alpha = \sigma\tau$.

In the standard approach for the construction of stable schemes, additive regularization is used. The alternative approach uses multiplicative perturbation of difference operators in the generating scheme. Within the framework of the latter approach, consider simplest examples part of which can be considered as a new interpretation of the above regularized schemes.

In multiplicative regularization of B, apply, for instance, the change $B \mapsto B(1 + \alpha R)$ or $B \mapsto (1 + \alpha R)B$. With such a perturbation, we still remain in the class of schemes with self-adjoint operators provided that $R = R^*$. In this case, we have the previously examined regularized scheme (7.85), (7.87).

An example of more complex regularization is given by the transformation

$$B \mapsto (E + \alpha R^*) B (E + \alpha R).$$

In the case of $R = A$, the condition for stability is $\alpha \geq \tau/8$. In another interesting example, the alternate triangular method is used, in which $A = R^* + R$ and $\alpha \geq \tau/2$.

In a similar manner, multiplicative regularization due to perturbed operator A can be applied. With allowance for the inequality (7.86), we can use the transformation $A \mapsto A(1+\alpha R)^{-1}$ or $A \mapsto (1+\alpha R)^{-1}A$. For simplest two-layer schemes, such regularization can be considered as a new version of regularization applied to B. To remain in the class of schemes with self-adjoint operators, it suffices for us to choose $R = R(A)$. A higher potential is offered by the regularization

$$A \mapsto (E+\alpha R^*)^{-1} A (E+\alpha R)^{-1}.$$

In the latter case, the regularizing operator R can be chosen not directly related with the operator A.

7.2.2 Inverted-time problem

In the development of computational algorithms for the approximate solution of ill-posed problems involving evolutionary equations, broad possibilities are offered by the regularization principle. In the rectangle Ω, we seek the solution of the parabolic equation

$$\frac{\partial u}{\partial t} + \sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big) = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T, \tag{7.90}$$

that differs from (7.77) only by the sign at the spatial derivatives, which corresponds to the substitution of t with $-t$, yielding the equation with inverted time. The boundary and initial conditions here remain the same as previously (see (7.78) and (7.79)).

We put into correspondence to the differential inverse problem (7.78), (7.78), (7.90) a Cauchy problem for the differential-operator equation

$$\frac{dy}{dt} - \Lambda y = 0, \qquad 0 < t < T. \tag{7.91}$$

Let us construct now absolutely stable difference schemes for (7.82), (7.91) using the regularization principle for difference schemes.

To construct the difference scheme for the ill-posed problem under consideration, we use the regularization principle. We start with the simplest explicit difference scheme

$$\frac{y_{n+1} - y_n}{\tau} - \Lambda y_n = 0, \qquad \boldsymbol{x} \in \omega, \quad n = 0, 1, \ldots, N_0 - 1 \tag{7.92}$$

supplemented with the initial condition (7.84).

In compliance with the regularization principle, we write the scheme (7.92) in the canonical form with

$$A = -\Lambda, \qquad B = E, \tag{7.93}$$

i. e., $A = A^* < 0$.

We use (see Lemma 3.2) the upper estimate for the difference operator Λ:

$$\Lambda \le ME \tag{7.94}$$

with the constant

$$M = \frac{4}{h_1^2} \max_{\boldsymbol{x}\in\omega} \frac{a^{(1)}(\boldsymbol{x}) + a^{(1)}(x_1 + h_1, x_2)}{2} + \frac{4}{h_2^2} \max_{\boldsymbol{x}\in\omega} \frac{a^{(2)}(\boldsymbol{x}) + a^{(2)}(x_1, x_2 + h_2)}{2}.$$

Theorem 7.8 *The explicit scheme* (7.84), (7.92) *is ϱ-stable in H with*

$$\varrho = 1 + M\tau. \tag{7.95}$$

Proof. The latter statement stems from the general conditions for ϱ-stability for two-layer operator-difference schemes. The difference scheme (7.84), (7.85) with self-adjoint, time-independent operators $B > 0$, A is ϱ-stable in H_B in the case of (see Theorem 4.4)

$$\frac{1-\rho}{\tau} B \le A \le \frac{1+\rho}{\tau} B. \tag{7.96}$$

For the scheme (7.84), (7.92), we have $B > 0$, $A < 0$, and for $\varrho > 1$ the right-hand side of the two-sided operator inequality (7.96) is fulfilled for all $\tau > 0$. The left-hand side of (7.96) assumes the form $(\varrho - 1)/\tau E \ge \Lambda$; in view of (7.94), this inequality holds with ϱ chosen according to (7.95). □

Remark 7.9 In the approximate solution of ill-posed problems, the value of the regularization parameter must be matched with the input-data inaccuracy. Here, we restrict ourselves to the construction of stable computational algorithms for ill-posed evolutionary problems and to the examination of the effect of regularization parameter on the stability of the difference scheme only. For a given value of α, the minimum value of ϱ is to be specified according to (7.95).

Starting from the explicit scheme (7.84), (7.92), we write the regularized scheme for problem (7.82), (7.91) in the canonical form (7.85) with

$$A = -\Lambda, \qquad B = E + \alpha R. \tag{7.97}$$

Theorem 7.10 *The regularized scheme* (7.85), (7.97) *is ϱ-stable in H_B with*

$$\varrho = 1 + \frac{\tau}{\alpha} \tag{7.98}$$

if the regularizing operator is chosen according to (7.88) and with

$$\varrho = 1 + \frac{\tau}{2\sqrt{\alpha}} \tag{7.99}$$

if the operator (7.89) is used.

Proof. To prove the theorem, it suffices to check that the left-hand side of the two-sided inequality (7.96) is fulfilled. For (7.97), this inequality assumes the form

$$\frac{\varrho - 1}{\tau}\,(E + \alpha R) \geq \Lambda. \tag{7.100}$$

In the case of $R = \Lambda$ and with ϱ chosen by (7.98), inequality (7.100) is fulfilled.

In the case of $R = \Lambda^2$, inequality (7.100) can be transformed into

$$E + \alpha\Lambda^2 - \frac{\tau}{\varrho - 1} = \Big(\sqrt{\alpha}\Lambda - \frac{\tau}{2\sqrt{\alpha}(\varrho - 1)}\,E\Big)^2 + \Big(1 - \frac{\tau^2}{4\alpha(\varrho - 1)^2}\Big)E \geq 0.$$

The latter inequality will be fulfilled with ϱ chosen in the form of (7.99). □

In a similar manner, other regularized difference schemes can be constructed. In particular, by now second-order evolutionary problems, problems with non-self-adjoint operators, additive schemes for multi-dimensional inverse problems, etc. have been considered. With constructed regularized difference schemes, these or those (standard or non-standard) variants of the generalized inverse method can be related.

7.2.3 Generalized inverse method

First of all, consider methods for the approximate solution of ill-posed problems involving evolutionary equations based on some perturbation of the initial equation that makes the perturbed problem a well-posed problem. Here, the perturbation parameter serves the regularization parameter. Such methods are known as generalized inverse methods. Let us give a priori estimates for the solution of the perturbed problem in various versions of the generalized inverse method for problems with self-adjoint, positive operators. Convergence of the approximate solution u_α to the exact solution u takes place in some well-posedness classes. It makes sense to consider here the matter of approximate solution of unstable evolutionary problems using difference methods based on various versions of the generalized inverse method. Primary attention will be paid to the matter of stability of the difference schemes.

First of all, let us dwell on the examination of the main version of the generalized inverse method. In the Hilbert space $\mathcal{H}$, consider an ill-posed Cauchy problem for the first-order evolutionary equation

$$\frac{du}{dt} - \mathcal{A}u = 0, \tag{7.101}$$

$$u(0) = u_0. \tag{7.102}$$

In the model inverse problem (7.78), (7.79), (7.90), we can put $\mathcal{H} = \mathcal{L}_2(\Omega)$ and

$$\mathcal{A}u = -\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\alpha}\Big)$$

on the set of functions satisfying (7.78). We assume that the operator $\mathcal{A}$ is a positive definite self-adjoint operator in $\mathcal{H}$. Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots$ be the eigenvalues of $\mathcal{A}$. The corresponding system of eigenfunctions $\{w_k\}$, $w_k \in D(\mathcal{A})$, $k = 1, 2, \ldots$ is an orthonormal complete system in $\mathcal{H}$.

For the stable approximate solution of the ill-posed problem (7.101), (7.102), apply the generalized inverse method. Taking into account the self-adjointness of $\mathcal{A}$, we determine $u_\alpha(t)$ as the solution of the equation

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha + \alpha\mathcal{A}^2 u_\alpha = 0 \tag{7.103}$$

with the initial condition

$$u_\alpha(0) = u_0. \tag{7.104}$$

Let us formulate a typical statement about the approximate-solution stability.

Theorem 7.11 *For the solution of problem* (7.103)*,* (7.104)*, the following estimate holds:*

$$\|u_\alpha(t)\| \ \leq \ \exp\Big(\frac{1}{4\alpha}\, t\Big)\|u_\alpha(0)\|. \tag{7.105}$$

Proof. To derive the estimate (7.105), we scalarwise multiply equation (7.103) in $\mathcal{H}$ by $u_\alpha(t)$. This yields the equality

$$\frac{1}{2}\,\frac{d}{dt}\,\|u_\alpha\|^2 + \alpha\|\mathcal{A}u_\alpha\|^2 \ = \ (\mathcal{A}u_\alpha, u_\alpha). \tag{7.106}$$

The right-hand side of (7.106) can be estimated as

$$(\mathcal{A}u_\alpha, u_\alpha) \ \leq \ \varepsilon\|\mathcal{A}u_\alpha\|^2 + \frac{1}{4\varepsilon}\,\|u_\alpha\|^2. \tag{7.107}$$

Substitution of (7.107) with $\varepsilon = \alpha$ into (7.106) yields:

$$\frac{d}{dt}\,\|u_\alpha\|^2 \leq \frac{1}{2\alpha}\,\|u_\alpha\|^2.$$

From here, and also from the Gronwall inequality, the desired estimate (7.105) follows. □

Let us briefly discuss the matter of convergence of the approximate solution $u_\alpha(t)$ to the exact solution $u(t)$. From (7.101) and (7.103), for the inaccuracy $v(t) = u_\alpha(t) - u(t)$ we obtain the equation

$$\frac{dv}{dt} - \mathcal{A}v + \alpha\mathcal{A}^2 v = -\alpha\mathcal{A}^2 u. \tag{7.108}$$

The ill-posedness of problem (7.101), (7.102) results from instability of the solution with respect to initial data. Instead of the exact initial condition (7.104), we use the condition

$$u_\alpha(0) = u_0^\delta. \tag{7.109}$$

For the initial-condition inaccuracy, we can use the estimate of type

$$\|u_0^\delta - u_0\| \le \delta. \tag{7.110}$$

In view of this, equation (7.108) is supplemented with the initial condition

$$v(0) = u_0^\delta - u_0, \tag{7.111}$$

in which $\|v(0)\| \le \delta$.

Convergence of the approximate solution to the solution of the ill-posed problem can be established in certain well-posedness classes. Most frequently, the a priori constraints on the solution of the unstable problem are related with the assumption about solution boundedness. Concerning the problem (7.101), (7.102), we can consider the following classes of solutions:

$$\|\mathcal{A}u(t)\| \le M, \tag{7.112}$$

$$\|\mathcal{A}^2 u(t)\| \le M \tag{7.113}$$

for all $t \in [0, T]$.

By examining stability with respect to initial data and right-hand side, we can derive an estimate for the inaccuracy in the classes of a priori constraints (7.112), (7.113). Under the constraints (7.112), we obtain:

$$\|v(t)\| \le \delta \exp\left(\frac{1}{2\alpha}\, t\right) + \left(\exp\left(\frac{t}{\alpha}\right) - 1\right)^{1/2} M. \tag{7.114}$$

In the class (7.113), the corresponding estimate is

$$\|v(t)\| \le \delta \exp\left(\frac{1}{2\alpha}\, t\right) + \sqrt{2}\left(\exp\left(\frac{t}{\alpha}\right) - 1\right)^{1/2} \alpha M. \tag{7.115}$$

The derived estimates (7.114) and (7.115) are not optimal estimates. In particular, these estimates do not prove convergence of the approximate solution $u_\alpha(t)$ found from equation (7.103) and initial condition (7.109) to the exact solution $u(t)$ with $\delta \to 0$. Here, a more subtle consideration is required. Let us show, for instance, the convergence in $\mathcal{H}$ in the class of solutions bounded in $\mathcal{H}$.

We will adhere to a scheme closely following the proof of convergence of approximate solution to the exact solution in the case of approximate solution obtained with non-locally perturbed initial condition. We write the approximate solution $u_\alpha(t)$ for the inaccurately given initial condition (7.109) in the operator form:

$$u_\alpha(t) = R(t, \alpha) u_0^\delta. \tag{7.116}$$

In this notation, $R(t, 0)$ is the exact solution of the problem and, hence, $u(t) = R(t, 0)u_0$. With the introduced settings, for the inaccuracy v we obtain

$$\begin{aligned} \|u_\alpha(t) - u(t)\| &= \|R(t, \alpha)(u_0^\delta - u_0) - (R(t, \alpha) - R(t, 0))u_0\| \\ &\le \|R(t, \alpha)\|\, \|u_0^\delta - u_0\| + \|(R(t, \alpha) - R(t, 0))u_0\|. \end{aligned} \tag{7.117}$$

The first term in (7.117) refers to stability with respect to initial data (boundedness of $R(t,\alpha)$). The second term in the right-hand side of (7.117) requires that the solution $R(t,\alpha)u_0$ of the perturbed problem obtained with exact input data be close to the exact solution $u(t)$. It is with this aim that a certain class of solutions (well-posedness class) has been isolated.

In view of the derived initial-data stability estimate (7.105) and estimate (7.110), we have:

$$\|R(t,\alpha)\|\,\|u_0^\delta - u_0\| \le \exp\Big(\frac{1}{4\alpha}\,t\Big)\delta. \tag{7.118}$$

Consider the second term in the right-hand side of (7.117).

The operator $R(t,\alpha)$ has the form

$$R(t,\alpha) = \exp\left((\mathcal{A} - \alpha\mathcal{A}^2)t\right). \tag{7.119}$$

For the approximate solution determined from (7.113), (7.109), with (7.119) we obtain the representation

$$u_\alpha(t) = \sum_{k=1}^{\infty}(u_0^\delta, w_k)\exp\left((\lambda_k - \alpha\lambda_k^2)t\right)w_k. \tag{7.120}$$

In view of (7.120), we have

$$\begin{aligned}\chi(t) &\equiv \|(R(t,\alpha) - R(t,0))u_0\|^2 \\ &= \sum_{k=1}^{\infty}\exp(2\lambda_k t)(1-\exp(-\alpha\lambda_k^2 t))^2(u_0, w_k)^2.\end{aligned} \tag{7.121}$$

We consider stability in the class of solutions bounded in $\mathcal{H}$:

$$\|u(t)\| \le M, \qquad 0 \le t \le T. \tag{7.122}$$

In view of (7.122), for any $\varepsilon > 0$ there exists a number $r(\varepsilon)$ such that

$$\sum_{k=r(\varepsilon)+1}^{\infty}\exp(2\lambda_k t)(u_0, w_k)^2 \;\le\; \frac{\varepsilon}{8}.$$

Hence, from (7.121) we obtain:

$$\begin{aligned}\chi(t) &\le \sum_{k=1}^{r(\varepsilon)}\exp(2\lambda_k t)(1-\exp(-\alpha\lambda_k^2 t))^2(u_0, w_k)^2 + \sum_{k=r(\varepsilon)+1}^{\infty}\exp(2\lambda_k t)(u_0, w_k)^2 \\ &\le M^2\sum_{k=1}^{r(\varepsilon)}(1-\exp(-\alpha\lambda_k^2 t))^2 + \frac{\varepsilon^2}{8}.\end{aligned}$$

For any $r(\varepsilon)$, there exists a number α_0 such that in the case of $\alpha \le \alpha_0$ we have

$$M^2 \sum_{k=1}^{r(\varepsilon)} (1 - \exp(-\alpha \lambda_k^2 t))^2 \le \frac{\varepsilon^2}{8}.$$

Hence, (7.122) yields:

$$\|(R(t,\alpha) - R(t,0))u_0\| \le \varepsilon/2. \tag{7.123}$$

With relations (7.118) and (7.123) taken into account, from (7.117) we obtain the estimate

$$\|u_\alpha(t) - u(t)\| \le \exp\left(\frac{1}{4\alpha}\, t\right)\delta + \frac{\varepsilon}{2}. \tag{7.124}$$

For any $\varepsilon > 0$ there exist a number $\alpha = \alpha(\delta) \le \alpha_0$ and a sufficiently small $\delta(\varepsilon)$ for which $\delta \exp(t/(4\alpha)) \le \varepsilon/2$. Hence, (7.124) assumes the form

$$\|u_\alpha(t) - u(t)\| \le \varepsilon.$$

Thus, the following statement can be formulated:

Theorem 7.12 *Suppose that for the initial-condition inaccuracy the estimate* (7.110) *is valid. Then, the approximate solution* $u_\alpha(t)$ *found as the solution of problem* (7.103), (7.109) *with* $\delta \to 0$, $\alpha(\delta) \to 0$, $\delta \exp(t/(4\alpha)) \to 0$ *converges to the bounded exact solution* $u(t)$ *of problem* (7.101), (7.102) *in* $\mathcal{H}$.

Remark 7.13 The proved statement admits various generalizations. For instance, in the narrower class of a priori constraints (7.113), one can derive a direct estimate of $\|(R(t,\alpha) - R(t,0))u_0\|$ in terms of α. It follows from (7.121) and (7.113) that

$$\|(R(t,\alpha) - R(t,0))u_0\|^2 \le \sum_{k=1}^{\infty} \lambda_k^4 \exp(2\lambda_k t)\alpha^2 t^2 (u_0, w_k)^2 \le \alpha^2 t^2 M^2$$

and, hence, the following inequality holds:

$$\|u_\alpha(t) - u(t)\| \le \exp\left(\frac{1}{4\alpha}\, t\right) + \alpha t M.$$

In the second variant of the generalized inverse method, transition to a pseudo-parabolic equation is used. In the case of a self-adjoint, positive operator $\mathcal{A}$, instead of the unstable problem (7.101), (7.102), we solve the equation

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha + \alpha\, \mathcal{A}\frac{du_\alpha}{dt} = 0, \tag{7.125}$$

supplemented with the initial condition (7.104).

Theorem 7.14 *For the solution of problem* (7.104), (7.125), *the following estimate holds:*

$$\|u_\alpha(t)\| \le \exp\left(\frac{1}{\alpha}\, t\right)\|u_\alpha(0)\|. \tag{7.126}$$

Proof. To prove the theorem, we rewrite equation (7.125) as

$$\frac{du_\alpha}{dt} - (E + \alpha\mathcal{A})^{-1}\mathcal{A}u_\alpha = 0. \tag{7.127}$$

We scalarwise multiply (7.127) in $\mathcal{H}$ by $u_\alpha(t)$; then, we obtain:

$$\frac{1}{2}\,\frac{d}{dt}\,\|u_\alpha\|^2 = ((E + \alpha\mathcal{A})^{-1}\mathcal{A}u_\alpha, u_\alpha). \tag{7.128}$$

It is easy to check that for the right-hand side of (7.128) there holds the estimate

$$((E + \alpha\mathcal{A})^{-1}\mathcal{A}u_\alpha, u_\alpha) \le \frac{1}{\alpha}\,(u_\alpha, u_\alpha).$$

On the operator level, the latter inequality corresponds to the inequality

$$(E + \alpha\mathcal{A})^{-1}\mathcal{A} \le \frac{1}{\alpha}\,E. \tag{7.129}$$

To prove inequality (7.129), we multiply (7.129) from the right and from the left by $(E + \alpha\mathcal{A})$. In this way, we pass to the equivalent inequality

$$(E + \alpha\mathcal{A})\mathcal{A} \le \frac{1}{\alpha}\,(E + 2\alpha\mathcal{A} + \alpha^2\mathcal{A}^2),$$

which obviously holds in the case of $\alpha > 0$.

With inequality (7.129) taken into account, from (7.128) we obtain the inequality

$$\frac{1}{2}\,\frac{d}{dt}\,\|u_\alpha\| \le \frac{1}{\alpha}\,\|u_\alpha\|^2,$$

which yields the desired estimate (7.126). □

Convergence of the approximate solution to the exact solution can be established in the case of $\delta \to 0$, $\alpha(\delta) \to 0$, $\delta \exp(t/(\alpha)) \to 0$ under the previously discussed constraints.

Let us briefly discuss the matter of construction and examination of numerical solution methods for ill-posed evolutionary problems with perturbed equations. As a model problem, consider the problem (7.78), (7.79), (7.90), a retrospective heat-transfer problem in the rectangle Ω. On discretization over space, we arrive at a Cauchy problem for the differential-operator equation (7.82), (7.91).

The difference analogue of the basic generalized inverse method (see (7.103)) has the form

$$\frac{y_{n+1}-y_n}{\tau} - \Lambda(\sigma_1 y_{n+1} + (1-\sigma_1)y_n) + \alpha\Lambda^2(\sigma_2 y_{n+1} + (1-\sigma_2)y_n) = 0, \quad n = 0, 1, \ldots, N_0 - 1. \tag{7.130}$$

Let us examine stability of this weighted scheme using the results of the general stability theory for operator-difference schemes. Here, stability criteria are formulated as operator inequalities for difference schemes written in the canonical form.

The scheme (7.130) can be written in the canonical form (7.85) with

$$\begin{aligned} A &= -\Lambda + \alpha\Lambda^2, \\ B &= E - \sigma_1\tau\Lambda + \sigma_2\tau\alpha\Lambda^2. \end{aligned} \tag{7.131}$$

Theorem 7.15 *The scheme* (7.85), (7.131) *is* ρ*-stable in* $H = L_2(\omega)$ *for any* $\tau > 0$ *with*

$$\rho = 1 + \frac{1}{4\alpha}\,\tau, \tag{7.132}$$

provided that $\sigma_1 \le 0$, $\sigma_2 \ge 1/2$.

Proof. First of all, note that the choice of ρ is perfectly consistent with the estimate (7.105) for the solution of the continuous problem. The proof of (7.132) is based on checking the necessary and sufficient conditions for ρ-stability.

First of all, we check that under the formulated constraints on the weights of (7.130) we have: $B = B^* > 0$. With (7.131), the right-hand side of (7.96) can be transformed as follows:

$$\begin{aligned} 0 \;&\le\; \frac{1+\rho}{\tau}\,B - A \\ &=\; \Lambda - \alpha\Lambda^2 + \frac{1+\rho}{\tau}\,E - (1+\rho)\sigma_1\Lambda + (1+\rho)\alpha\sigma_2\Lambda^2 \\ &=\; \frac{1+\rho}{\tau}\,E + (1-(1+\rho)\sigma_1)\Lambda + \alpha((1+\rho)\sigma_2 - 1)\Lambda^2. \end{aligned}$$

Under the indicated constraints on the weights, this inequality holds for all $\rho > 1$.

An estimate for ρ can be obtained from the left inequality of (7.96). With $\rho = 1 + c\tau$, this inequality assumes the form

$$cB \ge -A. \tag{7.133}$$

Substitution of (7.131) into (7.133) yields:

$$c(E - \sigma_1\tau\Lambda + \sigma_2\tau\alpha\Lambda^2) \;\ge\; \Lambda - \alpha\Lambda^2 \;=\; -\Big(\frac{1}{2\sqrt{\alpha}}\,E - \sqrt{\alpha}\,\Lambda\Big)^2 + \frac{1}{4\alpha}\,E.$$

In the case of $\sigma_1 \le 0$, $\sigma_2 \ge 1/2$, for the inequality to be fulfilled, we can put $c = 1/(4\alpha)$. The latter yields for ρ the expression (7.132). ∎

In a similar manner, difference schemes for other variants of the generalized inverse method can be constructed. Let us examine, for instance, two-layer difference schemes for a pseudo-parabolic perturbation. To approximately solve equation (7.91) with the initial condition (7.82), we use the difference scheme

$$\frac{y_{n+1} - y_n}{\tau} - \Lambda(\sigma y_{n+1} + (1 - \sigma) y_n) + \alpha \Lambda \frac{y_{n+1} - y_n}{\tau} = 0. \tag{7.134}$$

The scheme (7.134) can be written in the canonical form (7.85) with

$$A = -\Lambda, \qquad B = E + (\alpha - \sigma\tau)\Lambda. \tag{7.135}$$

Theorem 7.16 *The difference scheme* (7.85), (7.135) *is* ρ*-stable in* H *for all* $\tau > 0$ *with*

$$\rho = 1 + \frac{1}{\alpha}\tau, \tag{7.136}$$

provided that $\sigma \leq 0$.

Proof. In the case under consideration, the condition $B = B^* > 0$ and the right inequality of (7.96) obviously hold. With (7.135), the left inequality of (7.96) assumes the form

$$c(E + (\alpha - \sigma\tau)\Lambda) \geq \Lambda$$

and, hence, we can put $c = 1/\alpha$; then, for ρ we have (7.136). □

Remark 7.17 The estimate of the difference solution of scheme (7.134) with ρ given by (7.136) is perfectly consistent with the estimate for the solution of the differential problem (see estimate (7.126)).

Remark 7.18 Accurate to the adopted notation, the scheme (7.134) coincides with the ordinary weighted scheme immediately written for (7.82), (7.91):

$$\frac{y_{n+1} - y_n}{\tau} - \Lambda(\sigma' y_{n+1} + (1 - \sigma') y_n) = 0. \tag{7.137}$$

In suffices to put in (7.137) $\sigma' = \sigma - \alpha/\tau$. The only essential thing here is that the weight σ' in (7.137) is negative.

The presented schemes, as well as all other known schemes, can be obtained by regularization of unstable difference schemes. Moreover, based on an analysis of regularized difference schemes, other versions of the generalized inverse method can be constructed.

The difference scheme (7.134) of the pseudo-parabolic generalized inverse method is nothing else than a regularized scheme of type (7.85), (7.88), (7.97). The difference scheme of the basic variant (7.130) of the generalized inverse method refers to the use of regularized difference schemes with perturbed operators A and B (see (7.131)).

The regularized difference scheme (7.85), (7.89), (7.97) can be used to construct a new version of the generalized inverse method in which the approximate solution is sought as the solution of a Cauchy problem for the equation

$$\frac{du_\alpha}{dt} - \mathcal{A}u_\alpha + \alpha\mathcal{A}^2\frac{du_\alpha}{dt} = 0. \tag{7.138}$$

In construction of regularized difference schemes, the perturbed differential problem itself becomes unnecessary. Note also that for difference problems the construction of a stable scheme around the regularization principle is facilitated substantially because general results are available concerning the necessary and sufficient conditions for ρ-stability of the difference schemes.

7.2.4 Regularized additive schemes

Certain difficulties may arise in the computational realization of difference schemes constructed within the framework of the generalized inverse method. The latter is related, first of all, with the fact that the perturbed problem is a singularly perturbed problem (with small perturbation parameters at higher derivatives). It should be noted that a perturbed equation is a higher-order equation. To discuss this matter, here we imply that the standard variant of the generalized inverse method (7.103), (7.104) is used.

In the approximate solution of problem (7.77)–(7.79), it seems reasonable to use the difference scheme

$$\frac{y_{n+1} - y_n}{\tau} - \Lambda y_n + \alpha\Lambda^2(\sigma y_{n+1} + (1-\sigma)y_n) = 0. \tag{7.139}$$

The latter implies that in the two-parametric scheme (7.130) we have $\sigma_1 = 0$ and $\sigma_2 = \sigma$. The scheme (7.139) is ρ-stable (see Theorem 7.15) if $\sigma \geq 1/2$.

Above, we showed that the difference schemes constructed within the framework of the generalized inverse method can be obtained in two ways. The first approach implies approximation of the perturbed differential problem. The second approach is based on using the regularization principle for operator-difference schemes. In connection with the present consideration of scheme (7.139), another interpretation of the generalized inverse method scheme can prove useful, in which this scheme is considered as a scheme with smoothed mesh solution.

We consider scheme (7.139) as a scheme of the predictor-corrector type. Consider the simplest case of $\sigma = 1$. In this case, at the predictor stage we seek the mesh solution $\tilde{y}_{n+1}$ using the explicit scheme

$$\frac{\tilde{y}_{n+1} - y_n}{\tau} - \Lambda y_n = 0. \tag{7.140}$$

In accordance with (7.139), in the case of $\sigma = 1$ the corrector stage is

$$\frac{y_{n+1} - \tilde{y}_{n+1}}{\tau} + \alpha\Lambda^2 y_{n+1} = 0. \tag{7.141}$$

The explicit scheme (7.140) is unstable; a stable solution can be obtained using (7.141). Equation (7.141) can be considered as an Euler equation in the smoothing problem for the mesh function $\tilde{y}_{n+1}$:

$$J_\alpha(y_{n+1}) = \min_{v \in H} J_\alpha(v), \tag{7.142}$$

$$J_\alpha(v) = \|v - \tilde{y}_{n+1}\|^2 + \tau\alpha\|\Lambda v\|^2. \tag{7.143}$$

In the latter interpretation, we can speak of the generalized inverse method in the form (7.140), (7.141) (or (7.140), (7.142), (7.143)) as of a local regularization algorithm. Such a relation can be traced especially distinctly on the difference level.

Realization of the difference scheme (7.139) requires solution of the following forth-order difference elliptic equation:

$$\alpha\sigma\Lambda^2 v + \frac{1}{\tau} v = f.$$

At present, this task presents rather a difficult computational problem. It would therefore be desirable to have regularization difference schemes simpler from the standpoint of computational realization. In this connection, additive spatial-variable split schemes deserve mention in which transition to the next time layer necessitates solution of one-dimensional difference problems.

We construct additive operator-difference schemes starting from scheme (7.139). We begin with the interpretation of scheme (7.139) as a regularization scheme. As a generating scheme, we consider, as usual, an explicit scheme in which for the generalized inverse method we have:

$$\frac{y_{n+1} - y_n}{\tau} + (\alpha\Lambda^2 - \Lambda)y_n = 0. \tag{7.144}$$

To an additive regularization scheme, the following form of (7.139) corresponds:

$$(E + \alpha\sigma\tau\Lambda^2)\,\frac{y_{n+1} - y_n}{\tau} + (\alpha\Lambda^2 - \Lambda)y_n = 0.$$

In the latter case, the transition from (7.144) is interpreted as

$$B = E \mapsto B + \alpha\sigma\tau\Lambda^2.$$

We can change scheme (7.139) to the following somewhat different yet equivalent form:

$$\frac{y_{n+1} - y_n}{\tau} + (E + \alpha\sigma\tau\Lambda^2)^{-1}(\alpha\Lambda^2 - \Lambda)y_n = 0. \tag{7.145}$$

Scheme (7.145) corresponds to the use of the following version of multiplicative regularization:

$$A = \alpha\Lambda^2 - \Lambda \mapsto (E + \alpha\sigma\tau\Lambda^2)^{-1}A. \tag{7.146}$$

It is the regularization (7.145), (7.146) that will be used as the basis for the desired regularized additive operator-difference schemes.

In view of (7.80), we have the following additive representation for Λ:

$$\Lambda = \sum_{\beta=1}^{2} \Lambda_\beta, \qquad \Lambda_\beta y = -(a_\beta y_{\bar{x}_\beta})_{x_\beta}, \qquad \beta = 1, 2. \tag{7.147}$$

Either of these operator terms is positive,

$$\Lambda_\beta = \Lambda_\beta^* > 0, \qquad \beta = 1, 2,$$

and both are generally non-permutable:

$$\Lambda_1 \Lambda_2 \neq \Lambda_2 \Lambda_1.$$

In the case of three-dimensional problems, the operator Λ can be factorized to three one-dimensional pairwise non-permutable positive operators.

We construct the additive scheme using constructions analogous to (7.145), and used for each operator term in (7.147). The latter yields the regularized scheme

$$\frac{y_{n+1} - y_n}{\tau} + \sum_{\beta=1}^{2} (E + \alpha\sigma\tau\Lambda_\beta^2)^{-1}(\alpha\Lambda_\beta^2 - \Lambda_\beta) y_n = 0. \tag{7.148}$$

To obtain the stability condition, we write scheme (7.148) in the form

$$\frac{y_{n+1} - y_n}{\tau} + \sum_{\beta=1}^{2} A_\beta y_n = 0, \tag{7.149}$$

where

$$A_\beta = (E + \alpha\sigma\tau\Lambda_\beta^2)^{-1}(\alpha\Lambda_\beta^2 - \Lambda_\beta), \qquad \beta = 1, 2. \tag{7.150}$$

In the case of interest, for the operators we have $A_\beta = A_\beta^*$, $\beta = 1, 2$; hence,

$$A = \sum_{\beta=1}^{2} A_\beta = A^*. \tag{7.151}$$

The scheme (7.149), (7.151) is ρ-stable provided that the two-sided inequality (7.96) is fulfilled; in the case of interest, this inequality assumes the form

$$\frac{1-\rho}{\tau} E \leq A \leq \frac{1+\rho}{\tau} E. \tag{7.152}$$

Inequality (7.152) will be fulfilled, for instance, in the case of

$$\frac{1-\rho}{2\tau} E \leq A_\beta \leq \frac{1+\rho}{2\tau} E, \qquad \beta = 1, 2.$$

With allowance for (7.150), we rewrite this inequality as follows:

$$\frac{1-\rho}{2\tau}\,(E+\alpha\sigma\tau\Lambda_\beta^2) \;\le\; \alpha\Lambda_\beta^2-\Lambda_\beta \;\le\; \frac{1+\rho}{2\tau}\,(E+\alpha\sigma\tau\Lambda_\beta^2). \tag{7.153}$$

Taking into account the inequality

$$\alpha\Lambda_\beta^2-\Lambda_\beta \ge -\frac{1}{4\alpha}\,E,$$

we obtain that for the left inequality in (7.153) to be fulfilled it suffices for us to put

$$\rho = 1+\frac{\tau}{2\alpha}. \tag{7.154}$$

The right inequality (7.153) is fulfilled in all cases if $\sigma \ge 1$. This allows the following statement to be formulated:

Theorem 7.19 *The regularized additive scheme* (7.148) *is ρ-stable in H, with ρ defined by* (7.154) *with $\tau > 0$, provided that $\sigma \ge 1$.*

The realization of the additive scheme (7.148) requires the inversion of the one-dimensional difference operators $E+\alpha\sigma\tau\Lambda_\beta^2$, $\beta = 1, 2$. The following computational scheme can be proposed which demonstrates an intimate interrelation between the regularized schemes under consideration and additive-averaged difference schemes.

We determine the auxiliary mesh functions $y_{n+1}^{(\beta)}$, $\beta = 1, 2$ by solving the equations

$$\frac{y_{n+1}^{(\beta)}-y_n}{2\tau}+(E+\alpha\sigma\tau\Lambda_\beta^2)^{-1}(\alpha\Lambda_\beta^2-\Lambda_\beta)y_n = 0, \qquad \beta = 1, 2. \tag{7.155}$$

Afterwards, the next-layer solution is given by

$$y_{n+1} = \frac{1}{2}\sum_{\beta=1}^{2} y_{n+1}^{(\beta)}. \tag{7.156}$$

We change equation (7.155) to the form analogous to (7.139):

$$\frac{y_{n+1}^{(\beta)}-y_n}{2\tau}-\Lambda_\beta y_n+\alpha\Lambda_\beta^2(\sigma y_{n+1}^{(\beta)}+(1-\sigma)y_n) \;=\; 0.$$

Thus, the determination of $y_{n+1}^{(\beta)}$, $\beta = 1, 2$ corresponds to the case in which the generalized inverse method is applied to each operator term in (7.147).

Above (see (7.140), (7.141)), a close relation between the generalized inverse method and local regularization algorithms (smoothing at each time layer, see (7.140), (7.142), (7.143)) was noted. In the latter interpretation, the regularized scheme (7.148) corresponds to the case in which, as a local regularization algorithm, we use smoothing along individual directions (see (7.155)) followed by averaging according to (7.156).

We can relate with the regularized additive scheme (7.148) or, more precisely, with the additive-averaged scheme (7.155), (7.156), the generalized inverse method used to solve problem (7.101), (7.102) with

$$\mathcal{A} = \sum_{\beta=1}^{p} \mathcal{A}_\beta, \qquad \mathcal{A}_\beta = \mathcal{A}_\beta^* \geq 0, \qquad \beta = 1, 2, \ldots, p. \tag{7.157}$$

In the case of two-component splitting, we have $p = 2$. The approximate solution of problem (7.101), (7.102), (7.157) is to be found as the solution of a Cauchy problem for the equation

$$\frac{du_\alpha}{dt} - \sum_{\beta=1}^{p} \mathcal{A}_\beta u_\alpha + \alpha \sum_{\beta=1}^{p} \mathcal{A}_\beta^2 u_\alpha = 0. \tag{7.158}$$

For the solution of problem (7.104), (7.158), the following estimate of stability with respect to initial data holds:

$$\|u_\alpha(t)\| \ \leq \ \exp\left(\frac{p}{4\alpha}\, t\right) \|u_0\|.$$

The latter estimate is perfectly consistent with the above ρ-stability estimate (see (7.154)) for the regularized difference scheme (7.148).

7.2.5 Program

To approximately solve the inverted-time problem, we use the regularized additive scheme (7.148). The computational realization here is based on the interpretation of this scheme as the additive-averaged scheme (7.155), (7.156). The five-diagonal difference problems are solved by the five-point sweep algorithm using the subroutine `PROG5`. A detailed description of the algorithm is given in Section 6.1.

The input data for the inverted-time problem are taken from the solution of the direct problem (7.77)–(7.79). To find the approximate solution, the following purely implicit additive-averaged scheme is used:

$$\begin{aligned} \frac{y_{n+1}^{(\beta)} - y_n}{2\tau} + \Lambda_\beta y_{n+1} &= 0, \qquad \beta = 1, 2, \\ y_{n+1} &= \frac{1}{2} \sum_{\beta=1}^{2} y_{n+1}^{(\beta)}. \end{aligned}$$

The end-time difference solution of the direct problem is perturbed to use this solution as input data in the inverse problem.

The value of the regularization parameter was chosen considering the discrepancy, using the sequence

$$\alpha_k = \alpha_0 q^k, \qquad q > 0$$

with given $\alpha_0 = 0.$ and $q = 0.75$. To calculate the discrepancy, we additionally solved the direct problem in which initial data were taken equal to the solution of the inverse problem.

For simplicity, in the program presented below we have restricted ourselves to the simplest case of time-independent coefficients, with $k(x) = 1$ in (7.77), and the weighting parameter in the regularized additive scheme (7.155), (7.156) is $\sigma = 1$.

Program PROBLEM11

```
C
C     PROBLEM11  INVERTED-TIME PROBLEM
C                 TWO-DIMENSIONAL PROBLEM
C                 ADDITIVE REGULARIZED SCHEME
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.01, N1 = 101, N2 = 101, M = 101 )
      DIMENSION U0(N1,N2), UT(N1,N2), UTD(N1,N2), U(N1,N2)
     +         ,X1(N1), X2(N2), Y(N1,N2), Y1(N1,N2), Y2(N1,N2), YY(N1)
     +         ,A(N1), B(N1), C(N1), D(N1), E(N1), F(N1)  ! N1 >= N2
C
C     PARAMETERS:
C
C     X1L, X2L  - COORDINATES OF THE LEFT CORNER;
C     X1R, X2R  - COORDINATES OF THE RIGHT CORNER;
C     N1, N2    - NUMBER OF NODES IN THE SPATIAL GRID;
C     H1, H2    - STEP OVER SPACE;
C     TAU       - TIME STEP;
C     DELTA     - INPUT-DATA INACCURACY;
C     Q         - MULTIPLIER IN THE REGULARIZATION PARAMETER;
C     U0(N1,N2) - RECONSTRUCTED INITIAL CONDITION;
C     UT(N1,N2) - END-TIME SOLUTION OF THE DIRECT PROBLEM;
C     UTD(N1,N2)- DISTURBED SOLUTION OF THE DIRECT PROBLEM;
C     U(N1,N2)  - APPROXIMATE SOLUTION OF THE INVERSE PROBLEM;
C
      X1L    = 0.D0
      X1R    = 1.D0
      X2L    = 0.D0
      X2R    = 1.D0
      TMAX   = 0.025D0
      PI     = 3.1415926D0
C
      OPEN (01, FILE = 'RESULT.DAT') !FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H1  = (X1R-X1L) / (N1-1)
      H2  = (X2R-X2L) / (N2-1)

      TAU = TMAX / (M-1)
      DO I = 1, N1
        X1(I) = X1L + (I-1)*H1
      END DO
      DO J = 1, N2
        X2(J) = X2L + (J-1)*H2
      END DO
C
C     DIRECT PROBLEM
```

```
C     ADDITIVE-AVERAGE DIFFERENCE SCHEME
C
C     INITIAL CONDITION
C
      DO I = 1, N1

        DO J = 1, N2
          U0(I,J) = AU(X1(I),X2(J))
          Y(I,J)  = U0(I,J)
        END DO
      END DO
      DO K = 2, M
C
C     SWEEP OVER X1
C
        DO J = 2, N2-1
C
C     COEFFICIENTS OF THE DIFFERENCE SCHEME IN THE DIRECT PROBLEM
C
          B(1)  = 0.D0
          C(1)  = 1.D0
          F(1)  = 0.D0
          A(N1) = 0.D0
          C(N1) = 1.D0
          F(N1) = 0.D0
          DO I = 2, N1-1
            A(I) = 1.D0 / (H1*H1)
            B(I) = 1.D0 / (H1*H1)
            C(I) = A(I) + B(I) + 0.5D0 / TAU
            F(I) = 0.5D0 * Y(I,J) / TAU
          END DO
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          ITASK1 = 1

          CALL PROG3 ( N1, A, C, B, F, YY, ITASK1 )
          DO I = 1, N1
            Y1(I,J) = YY(I)
          END DO

        END DO
C
C     SWEEP OVER X2
C
        DO I = 2, N1-1
C
C     COEFFICIENTS OF THE DIFFERENCE SCHEME IN THE DIRECT PROBLEM
C
          B(1)  = 0.D0
          C(1)  = 1.D0
          F(1)  = 0.D0
          A(N2) = 0.D0

          C(N2) = 1.D0
          F(N2) = 0.D0
          DO J = 2, N2-1
            A(J) = 1.D0 / (H2*H2)
            B(J) = 1.D0 / (H2*H2)
            C(J) = A(J) + B(J) + 0.5D0 / TAU
```

```
            F(J) = 0.5D0 * Y(I,J) / TAU

          END DO
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          ITASK1 = 1
          CALL PROG3 ( N2, A, C, B, F, YY, ITASK1 )
          DO J = 1, N2
            Y2(I,J) = YY(J)
          END DO
        END DO
C
C     ADDITIVE AVERAGING
C
        DO I = 1, N1
          DO J = 1, N2
            Y(I,J) = 0.5D0 * (Y1(I,J) + Y2(I,J))
          END DO
        END DO
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO I = 1, N1
        DO J = 1, N2
          UT(I,J)  = Y(I,J)
          UTD(I,J) = Y(I,J)
        END DO
      END DO

      DO I = 2, N1-1
        DO J = 2, N2-1
          UTD(I,J) = UT(I,J) + 2.*DELTA*(RAND(0)-0.5)
        END DO
      END DO
C
C     INVERSE PROBLEM
C
C     ITERATIVE PROCESS TO ADJUST
C     THE VALUE OF THE REGULARIZATION PARAMETER
C
      IT = 0
      ITMAX = 100
      ALPHA = 0.001D0

      Q = 0.75D0
  100 IT = IT + 1
C

C     ADDITIVE-AVERAGED REGULARIZATION SCHEMES
C
C     INITIAL CONDITION
C
      DO I = 1, N1
        DO J = 1, N2
          Y(I,J)  = UTD(I,J)
        END DO
      END DO
      DO K = 2, M
```

```
C
C     SWEEP OVER X1
C
        DO J = 2, N2-1
C
C     COEFFICIENTS OF THE DIFFERENCE SCHEME IN THE INVERSE PROBLEM
C
          DO I = 2,N1-1
            A(I) = ALPHA / (H1**4)
            B(I) = 4.D0 * ALPHA / (H1**4)
            C(I) = 6.D0 * ALPHA / (H1**4) + 0.5D0 / TAU
            D(I) = 4.D0 * ALPHA / (H1**4)
            E(I) = ALPHA / (H1**4)
            F(I) = 0.5D0 * Y(I,J) / TAU
     +           - (Y(I+1,J) - 2.D0*Y(I,J) + Y(I-1,J)) / (H1**2)
          END DO
          C(1)    = 1.D0
          D(1)    = 0.D0

          E(1)    = 0.D0
          F(1)    = 0.D0
          B(2)    = 0.D0
          C(2)    = 5.D0 * ALPHA / (H1**4) + 0.5D0 / TAU
          C(N1-1) = 5.D0 * ALPHA / (H1**4) + 0.5D0 / TAU
          D(N1-1) = 0.D0
          A(N1)   = 0.D0
          B(N1)   = 0.D0
          C(N1)   = 1.D0
          F(N1)   = 0.D0
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          ITASK2 = 1
          CALL PROG5 ( N1, A, B, C, D, E, F, YY, ITASK2 )
          DO I = 1, N1
            Y1(I,J) = YY(I)
          END DO
        END DO
C
C     SWEEP OVER X2
C
        DO I = 2, N1-1
C
C     COEFFICIENTS IN THE DIFFERENCE SCHEME FOR THE INVERSE PROBLEM
C
          DO J = 2, N2-1
            A(J) = ALPHA / (H2**4)
            B(J) = 4.D0 * ALPHA / (H2**4)
            C(J) = 6.D0 * ALPHA / (H2**4) + 0.5D0 / TAU
            D(J) = 4.D0 * ALPHA / (H2**4)
            E(J) = ALPHA / (H2**4)
            F(J) = 0.5D0 * Y(I,J) / TAU
     +           - (Y(I,J+1) - 2.D0*Y(I,J) + Y(I,J-1)) / (H2**2)
          END DO
          C(1)    = 1.D0
          D(1)    = 0.D0
          E(1)    = 0.D0
          F(1)    = 0.D0
          B(2)    = 0.D0
          C(2)    = 5.D0 * ALPHA / (H2**4) + 0.5D0 / TAU
```

```
          C(N2-1) = 5.D0 * ALPHA / (H2**4) + 0.5D0 / TAU
          D(N2-1) = 0.D0
          A(N2)   = 0.D0
          B(N2)   = 0.D0
          C(N2)   = 1.D0
          F(N2)   = 0.D0
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          ITASK2 = 1
          CALL PROG5 ( N2, A, B, C, D, E, F, YY, ITASK2 )
          DO J = 1, N2

            Y2(I,J) = YY(J)
          END DO
        END DO
C
C     ADDITIVE AVERAGING
C
        DO I = 1, N1
          DO J = 1, N2
            Y(I,J) = 0.5D0 * (Y1(I,J) + Y2(I,J))
          END DO
        END DO
      END DO
      DO I = 1, N1
        DO J = 1, N2
          U(I,J) = Y(I,J)
        END DO

      END DO
C
C     DIRECT PROBLEM
C
      DO K = 2, M
C
C     SWEEP OVER X1
C
        DO J = 2, N2-1
C
C     COEFFICIENTS IN THE DIFFERENCE SCHEME FOR THE DIRECT PROBLEM
C
          B(1)  = 0.D0
          C(1)  = 1.D0
          F(1)  = 0.D0
          A(N1) = 0.D0
          C(N1) = 1.D0
          F(N1) = 0.D0
          DO I = 2, N1-1
            A(I) = 1.D0 / (H1*H1)
            B(I) = 1.D0 / (H1*H1)
            C(I) = A(I) + B(I) + 0.5D0 / TAU
            F(I) = 0.5D0 * Y(I,J) / TAU
          END DO
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          ITASK1 = 1
          CALL PROG3 ( N1, A, C, B, F, YY, ITASK1 )
```

```
          DO I = 1, N1

            Y1(I,J) = YY(I)
          END DO
        END DO
C
C     SWEEP OVER X2
C
        DO I = 2, N1-1
C
C     COEFFICIENTS IN THE DIFFERENCE SCHEME FOR THE DIRECT PROBLEM
C
          B(1)  = 0.D0
          C(1)  = 1.D0
          F(1)  = 0.D0
          A(N2) = 0.D0
          C(N2) = 1.D0
          F(N2) = 0.D0
          DO J = 2, N2-1
            A(J) = 1.D0 / (H2*H2)
            B(J) = 1.D0 / (H2*H2)
            C(J) = A(J) + B(J) + 0.5D0 / TAU

            F(J) = 0.5D0 * Y(I,J) / TAU
          END DO
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          ITASK1 = 1
          CALL PROG3 ( N2, A, C, B, F, YY, ITASK1 )
          DO J = 1, N2
            Y2(I,J) = YY(J)
          END DO
        END DO
C
C     ADDITIVE AVERAGING
C
        DO I = 1, N1
          DO J = 1, N2
            Y(I,J) = 0.5D0 * (Y1(I,J) + Y2(I,J))
          END DO

        END DO
      END DO
C
C     CRITERION FOR THE EXIT FROM THE ITERATIVE PROCESS
C
      SUM = 0.D0
      DO I = 2, N1-1
        DO J = 2, N2-1
          SUM = SUM + (Y(I,J) - UTD(I,J))**2*H1*H2
        END DO
      END DO
      SL2 = DSQRT(SUM)

C
      IF ( IT.EQ.1 ) THEN
        IND = 0
        IF ( SL2.LT.DELTA ) THEN
          IND = 1
```

```
            Q = 1.D0/Q
          END IF
           ALPHA = ALPHA*Q
          GO TO 100
        ELSE
          ALPHA = ALPHA*Q
          IF ( IND.EQ.0 .AND. SL2.GT.DELTA ) GO TO 100
          IF ( IND.EQ.1 .AND. SL2.LT.DELTA ) GO TO 100
        END IF
C
C       SOLUTION
C
        WRITE ( 01, * ) ((UTD(I,J),I=1,N1), J=1,N2)
        WRITE ( 01, * ) ((U(I,J),I=1,N1), J=1,N2)

        CLOSE (01)
        STOP
        END


        DOUBLE PRECISION FUNCTION  AU ( X1, X2 )
        IMPLICIT REAL*8 ( A-H, O-Z )
C
C        INITIAL CONDITION
C
        AU = 0.D0
        IF ((X1-0.6D0)**2 + (X2-0.6D0)**2.LE.0.04D0) AU = 1.D0
C
        RETURN
        END
```

7.2.6 Computational experiments

The data presented below were obtained on a uniform grid with $h_1 = 0.01$, $h_2 = 0.01$ for the problem in the unit square. As input data, the solution of the direct problem at the time $T = 0.025$ was taken, the time step size being $\tau = 0.00025$. In the direct problem, the initial condition, taken as the exact end-time solution of the inverse problem, is given in the form

$$u_0(\boldsymbol{x}, 0) = \begin{cases} 1, & (x_1 - 0.6)^2 + (x_2 - 0.6)^2 \le 0.04, \\ 0, & (x_1 - 0.6)^2 + (x_2 - 0.6)^2 > 0.04. \end{cases}$$

The end-time solution of the direct problem (the exact initial condition for the direct problem) is shown in Figure 7.5. Here, contour lines obtained with $\Delta u = 0.05$ are shown. Figure 7.6 shows the solution of the inverse problem obtained with $\delta = 0.01$ (here, $\triangle u = 0.1$). Since, in the present example, a discontinuous function is reconstructed, one cannot expect that a very good accuracy can be achieved. The effect due to the inaccuracy level can be traced considering Figures 7.7 and 7.8.

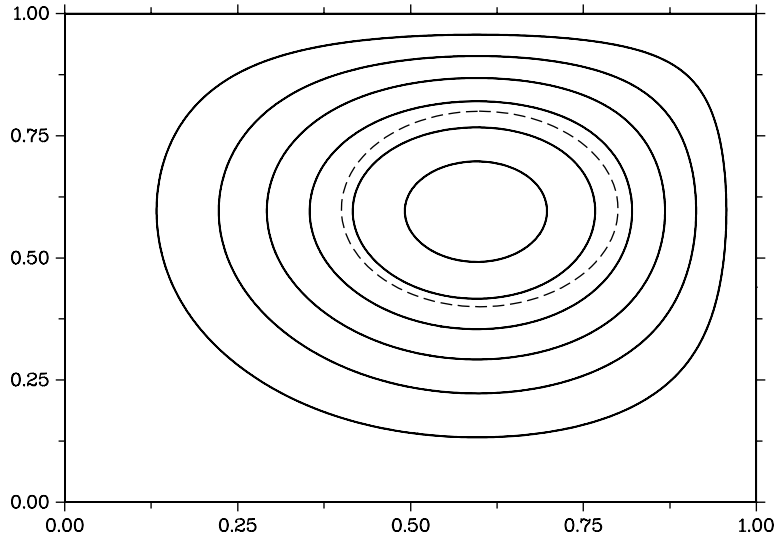

Figure 7.5 Solution of the direct problem at $t = T$

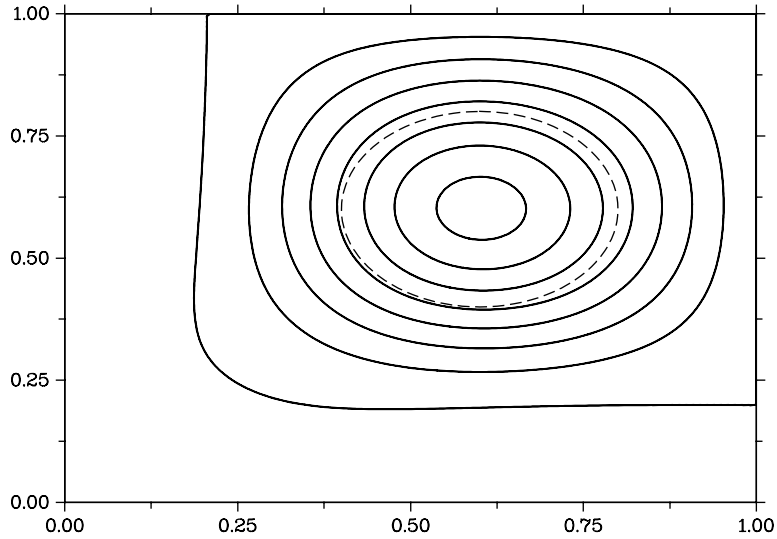

Figure 7.6 Solution of the inverse problem obtained with $\delta = 0.01$

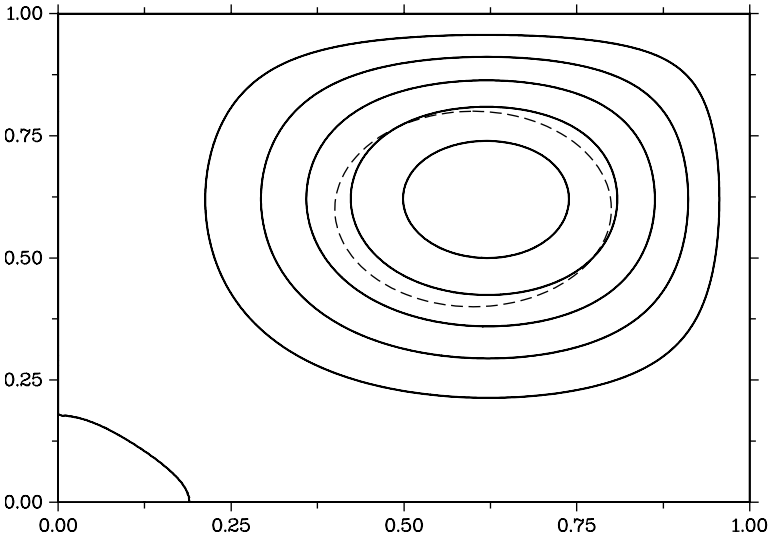

Figure 7.7 Solution of the inverse problem obtained with $\delta = 0.02$

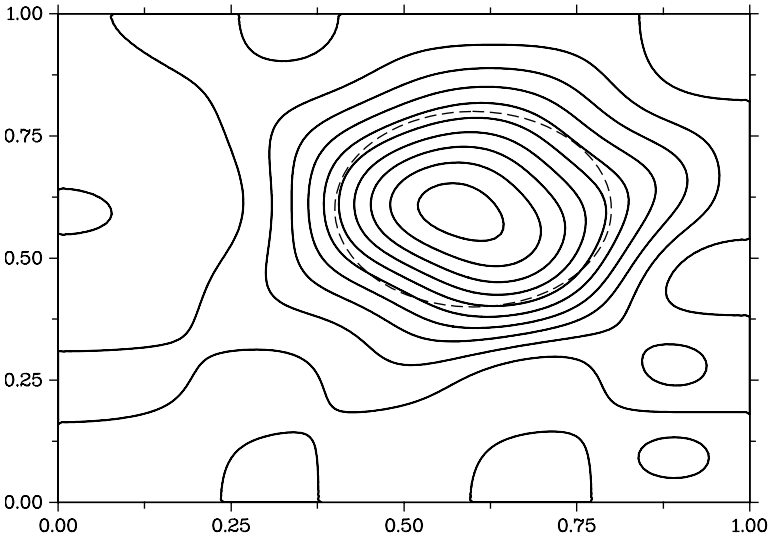

Figure 7.8 Solution of the inverse problem obtained with $\delta = 0.005$

7.3 Iterative solution of retrospective problems

Key features of algorithms intended for the approximate solution of inverted-time problems by iteration methods with refinement of initial conditions are outlined. A model problem for the two-dimensional non-stationary parabolic equation is considered.

7.3.1 Statement of the problem

In solving inverse problems for mathematical physics equations, gradient iteration methods are applied to the variational formulation of the problem. Below, we consider a simplest iteration method for the approximate solution of the retrospective inverse problem for the second-order parabolic equation. For this inverse problem, the initial condition is refined iteratively, which requires solving, at each iteration step, an ordinary boundary value problem for the parabolic equation.

Based on the general theory of iteration solution methods for operator equations, sufficient conditions for convergence of the iterative process can be established, and the iteration parameters are chosen. In such problems, the operator of transition to the next approximation makes it possible to identify the approximate solution in a desired class of smoothness.

As a model problem, consider a two-dimensional problem in the rectangle

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\beta < l_\beta,\ \beta = 1, 2\}.$$

In the domain Ω, we seek the solution of the parabolic equation

$$\frac{\partial u}{\partial t} - \sum_{\beta=1}^{2} \frac{\partial}{\partial x_\beta}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\beta}\Big) = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T, \tag{7.159}$$

supplemented with the simplest first-kind homogeneous boundary conditions:

$$u(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \partial\Omega, \quad 0 < t < T. \tag{7.160}$$

In the inverse problem, instead of setting the zero-time solution (the solution at $t = 0$), the end-time solution is specified:

$$u(\boldsymbol{x}, T) = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{7.161}$$

Such an inverse problem is a well-posed one, for instance, in the classes of bounded solutions.

7.3.2 Difference problem

Performing discretization over space, we put into correspondence to the differential problem (7.159)–(7.161) some differential-difference problem. For simplicity, we assume that a grid uniform along either direction with steps h_β, $\beta = 1, 2$ is introduced in the domain Ω, and let ω be the set of internal nodes.

On the set of mesh functions $y(\boldsymbol{x})$ such that $y(\boldsymbol{x}) = 0$, $\boldsymbol{x} \notin \omega$, we define the difference operator Λ:

$$\Lambda y = -\sum_{\beta=1}^{2} (a_\beta y_{\bar{x}_\beta})_{x_\beta} \tag{7.162}$$

where, for instance,

$$a_1(\boldsymbol{x}) = k(x_1 - 0.5h_1, x_2), \qquad a_2(\boldsymbol{x}) = k(x_1, x_2 - 0.5h_2).$$

In the mesh Hilbert space H, we introduce the scalar product and the norm:

$$(y, w) = \sum_{\boldsymbol{x}\in\omega} y(\boldsymbol{x}) w(\boldsymbol{x}) h_1 h_2, \qquad \|y\| = (y, y)^{1/2}.$$

In H, we have $\Lambda = \Lambda^* > 0$. We pass from (7.159)–(7.161) to the differential-operator equation

$$\frac{dy}{dt} + \Lambda y = 0, \qquad \boldsymbol{x} \in \omega, \quad 0 < t < T \tag{7.163}$$

with some given

$$y(T) = \varphi, \qquad \boldsymbol{x} \in \omega. \tag{7.164}$$

Previously, possible ways in constructing regularizing algorithms for the approximate solution of problem (7.163), (7.164) based on using perturbed equations or perturbed initial conditions were discussed. In this chapter, the inverse problem (7.163), (7.164) will be solved by iteration methods.

7.3.3 Iterative refinement of the initial condition

Here, we are going to employ methods in which, at each iteration step, the emerging well-posed problems are solved using standard two-layer difference schemes.

Suppose that, instead of the inverse problem (7.163), (7.164), we treat the direct problem for equation (7.163), in which, instead of (7.164), we use the initial condition

$$y(0) = v, \qquad \boldsymbol{x} \in \omega. \tag{7.165}$$

We denote by y_n the difference solution at the time $t_n = n\tau$, where $\tau > 0$ is the time step size, so that $N_0\tau = T$. In the ordinary two-layer weighted scheme the passage to

the next time layer in problem (7.163), (7.165) is to be made in accordance with

$$\frac{y_{n+1}-y_n}{\tau}+\Lambda(\sigma y_{n+1}+(1-\sigma)y_n)=0, \qquad n=0,1,\dots,N_0-1, \tag{7.166}$$

$$y_0=v, \qquad \boldsymbol{x}\in\omega. \tag{7.167}$$

As it is well known, the weighted scheme (7.166), (7.167) is absolutely stable if $\sigma \geq 1/2$, and there holds the following estimate of stability:

$$\|y_{n+1}\| \leq \|y_n\| \leq \cdots \leq \|y_0\| = \|v\|, \qquad n=0,1,\dots,N_0-1. \tag{7.168}$$

Thereby, the solution norm decreases with time.

To approximately solve the inverse problem (7.163), (7.164), we use a simplest iterative process based on successive refinement of the initial condition and on solving at each iteration step the direct problem. Let us formulate the problem in the operator form.

From (7.166) and (7.167), for the given y_0 at the end time we obtain

$$y_{N_0}=S^N v, \tag{7.169}$$

where S is the operator of transition from one time layer to the next time layer:

$$S=(E+\sigma\tau\Lambda)^{-1}(E+(\sigma-1)\tau\Lambda). \tag{7.170}$$

In view of (7.163), (7.164), and (7.169), the approximate solution of the inverse problem can be put into correspondence to the solution of the following difference operator equation:

$$Av=\varphi, \quad \boldsymbol{x}\in\omega, \qquad A=S^N. \tag{7.171}$$

Since the operator Λ is a self-adjoint operator, the transition operator S and the operator A in (7.171) are also self-adjoint operators. The difference equation (7.171) can be solved uniquely if, for instance, the operator A is a positive operator. In turn, the latter condition is satisfied if the transition operator S is a positive operator. Taking into account the representation (7.170), we obtain that $S>0$ in the case of

$$\sigma \geq 1. \tag{7.172}$$

Condition (7.172) imposed on the weight of scheme (7.166), (7.167) is a more stringent condition than the ordinary stability condition. In the case of interest, under the constraints (7.172) for the operator A defined by (7.171) we have:

$$0<A=A^*<E. \tag{7.173}$$

Equation (7.171), (7.173) can be solved using the explicit two-layer iteration method; this method can be written as

$$\frac{v_{k+1} - v_k}{s_{k+1}} + Av_k = \varphi. \tag{7.174}$$

Here, s_{k+1} are iteration parameters. We denote the difference solution obtained with the initial condition v_k as $y^{(k)}$.

The iteration method under consideration implies the following organization of calculations in the approximate solution of the retrospective inverse problem (7.159), (7.160).

First, with the given v_k, we solve the direct problem, using, for determining $y_{N_0}^{(k)}$, the difference scheme

$$\frac{y_{n+1}^{(k)} - y_n^{(k)}}{\tau} + \Lambda(\sigma y_{n+1}^{(k)} + (1-\sigma)y_n^{(k)}) = 0, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{7.175}$$

$$y_0^{(k)} = v_k, \qquad x \in \omega. \tag{7.176}$$

Then, with the found end-time solution of the direct problem, we use formula (7.174) to refine the initial condition:

$$v_{k+1} = v_k - s_{k+1}(y_{N_0}^{(k)} - \varphi). \tag{7.177}$$

As it follows from the general theory of iterative solution methods, the rate of convergence in method (7.174), used to solve equation (7.171), is defined by the energy equivalence constants $\gamma_\beta,\ \beta = 1, 2$:

$$\gamma_1 E \le A \le \gamma_2 E, \qquad \gamma_1 > 0. \tag{7.178}$$

Regarding (7.173), we can put $\gamma_2 = 1$. The positive constant γ_1, close to zero, depends on the grid.

In the notation used, in the stationary iteration method ($s_k = s_0 = \text{const}$) the conditions for convergence in (7.174) have the form $s_0 \le 2$. For the optimal constant value of the iteration parameter we have: $s_0 \approx 1$. For the convergence to be improved, it makes sense to use variation-type iteration methods. In the iteration method of minimal discrepancies, for the iteration parameters we have:

$$s_{k+1} = \frac{(Ar_k, r_k)}{(Ar_k, Ar_k)}, \qquad r_k = Av_k - \varphi.$$

Here, at each iteration step we minimize the discrepancy norm which implies the following estimate:

$$\|r_{k+1}\| < \|r_k\| < \cdots < \|r_0\|.$$

In the more general implicit iteration method, instead of (7.174) we have:

$$B\,\frac{v_{k+1}-v_k}{s_{k+1}}+Av_k=\varphi, \tag{7.179}$$

where $B=B^*>0$. In the method of minimal corrections, iteration parameters can be calculated by the formulas

$$s_{k+1}=\frac{(Aw_k,w_k)}{(B^{-1}Aw_k,Aw_k)},\qquad w_k=B^{-1}r_k.$$

Here, to be minimized at each iteration step is the correction w_{k+1} which implies the estimate

$$\|w_{k+1}\|<\|w_k\|<\cdots<\|w_0\|.$$

In a similar manner, more rapidly converging three-layer variational iteration methods can be considered.

Note the following specific features in the choice of B in solving ill-posed problems. In ordinary iteration methods the operator B is to be chosen so that just to raise the rate of convergence in the method. In solving ill-posed problem, the iterative process is to be terminated on attainment of a discrepancy value defined by the input-data inaccuracy. Of importance for us is not only the rate with which the iterative process converges on the descending portion of the characteristic curve but also the class of smoothness in which this iterative process converges and the norm with which the required discrepancy level can be achieved. A key specific feature of the approximate solution of ill-posed problems by iteration methods consists in that the approximate solution can be identified in the desired class of smoothness using a proper choice of B.

7.3.4 Program

The iteration method under consideration is based on refinement of the initial condition for the solution of a well-examined direct problem. For the implicit difference schemes (7.175), (7.176) to be realized, we have to solve at each time step two-dimensional difference elliptic problems. To this end, we use iteration methods (embedded iterative process).

Program PROBLEM12

```
C     PROBLEM12 - PROBLEM WITH INVERTED TIME
C                 TWO-DIMENSIONAL PROBLEM
C                 ITERATIVE REFINEMENT OF THE INITIAL CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.01D0, N1 = 51, N2 = 51, M = 101 )
      DIMENSION A(17*N1*N2), X1(N1), X2(N2)
      COMMON / SB5 /     IDEFAULT(4)
```

```
      COMMON / CONTROL / IREPT, NITER
C
C     PARAMETERS:
C
C     X1L, X2L  - COORDINATES OF THE LEFT CORNER;
C     X1R, X2R  - COORDINATES OF THE RIGHT CIRNER;
C     N1, N2    - NUMBER OF NODES IN THE SPATIAL GRID;
C     H1, H2    - MESH SIZES OVER SPACE;
C     TAU       - TIME STEP;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     U0(N1,N2) - INITIAL CONDITION TO BE RECONSTRUCTED;
C     UT(N1,N2) - END-TIME SOLUTION IN THE DIRECT PROBLEM;
C     UTD(N1,N2)- DISTURBED DIRECT-PROBLEM SOLUTION;
C     U(N1,N2)  - APPROXIMATE SOLUTION IN THE INVERSE PROBLEM;
C     EPSR      - RELATIVE INACCURACY OF THE DIFFERENCE SOLUTION;
C     EPSA      - ABSOLUTE INACCURACY OF THE DIFFERENCE SOLUTION;
C
C     EQUIVALENCE ( A(1),         A0            ),
C    *            ( A(N+1),       A1            ),
C    *            ( A(2*N+1),     A2            ),
C    *            ( A(9*N+1),     F             ),
C    *            ( A(10*N+1),    U0            ),
C    *            ( A(11*N+1),    UT            ),
C    *            ( A(12*N+1),    UTD           ),
C    *            ( A(13*N+1),    U             ),
C    *            ( A(14*N+1),    V             ),
C    *            ( A(15*N+1),    R             ),
C    *            ( A(16*N+1),    BW            )
C
      X1L    = 0.D0
      X1R    = 1.D0
      X2L    = 0.D0
      X2R    = 1.D0
      TMAX   = 0.025D0
      PI     = 3.1415926D0
      EPSR   = 1.D-5
      EPSA   = 1.D-8
C
      OPEN (01, FILE ='RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H1  = (X1R-X1L) / (N1-1)
      H2  = (X2R-X2L) / (N2-1)
      TAU = TMAX / (M-1)
      DO I = 1, N1
        X1(I) = X1L + (I-1)*H1
      END DO
      DO J = 1, N2
        X2(J) = X2L + (J-1)*H2
      END DO
C
      N  = N1*N2
      DO I = 1, 17*N
        A(I) = 0.0
      END DO
C
C     DIRECT PROBLEM
C     PURELY IMPLICIT DIFFERENCE SCHEME
C
```

```
C     INITIAL CONDITION
C
      T  = 0.D0
      CALL INIT (A(10*N+1), X1, X2, N1, N2)
      DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        CALL FDST (A(1), A(N+1), A(2*N+1), A(9*N+1), A(10*N+1),
     +             H1, H2, N1, N2, TAU)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO I = 1, N
        A(11*N+I) = A(10*N+I)
        A(12*N+I) = A(11*N+I) + 2.*DELTA*(RAND(0)-0.5)
      END DO
C
C     INVERSE PROBLEM
C     ITERATION METHOD
C
      IT = 0
C
C     STARTING APPROXIMATION
C
      DO I = 1, N
        A(14*N+I) = 0.D0
      END DO
C
  100 IT = IT + 1
C

C     DIRECT PROBLEM
C
      T  = 0.D0
C
C     INITIAL CONDITION
C
      DO I = 1, N
        A(10*N+I) = A(14*N+I)
      END DO
      DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        CALL FDST (A(1), A(N+1), A(2*N+1), A(9*N+1), A(10*N+1),
     +             H1, H2, N1, N2, TAU)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
```

```
      END DO
C
C     DISCREPANCY
C
      DO I = 1, N
        A(15*N+I) = A(10*N+I) - A(12*N+I)
      END DO
C
C     CORRECTION
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
      CALL FDSB (A(1), A(N+1), A(2*N+1), A(9*N+1), A(15*N+1),
     +           H1, H2, N1, N2)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
      IDEFAULT(1) = 0
      IREPT       = 0
      CALL SBAND5 (N, N1, A(1), A(13*N+1), A(9*N+1), EPSR, EPSA)
C
C     ITERATION PARAMETERS
C
      T  = 0.D0
C
C     INITIAL CONDITIONS
C
      DO I = 1, N
        A(10*N+I) = A(13*N+I)
      END DO
      DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        CALL FDST (A(1), A(N+1), A(2*N+1), A(9*N+1), A(10*N+1),
     +             H1, H2, N1, N2, TAU)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
      END DO
C
C     METHOD OF MINIMAL DISCREPANCIES
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
      CALL FDSB (A(1), A(N+1), A(2*N+1), A(9*N+1), A(10*N+1),
     +           H1, H2, N1, N2)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
      IDEFAULT(1) = 0
      IREPT       = 0
      CALL SBAND5 (N, N1, A(1), A(16*N+1), A(9*N+1), EPSR, EPSA)
C
      SUM  = 0.D0
      SUM1 = 0.D0
      DO I = 1, N
```

```
        SUM  = SUM + A(10*N+I)*A(13*N+I)
        SUM1 = SUM1 + A(16*N+I)*A(10*N+I)
      END DO
      SS = SUM / SUM1
C
C     NEXT APPROXIMATION
C
      DO I = 1, N
        A(14*N+I) = A(14*N+I) - SS*A(13*N+I)
      END DO
C
C     EXIT FROM THE ITERATIVE PROCESS BY THE DISCREPANCY CRITERION
C
      SUM = 0.D0
      DO I = 1, N
        SUM = SUM + A(15*N+I)**2*H1*H2
      END DO
      SL2 = DSQRT(SUM)
      IF ( SL2.GT.DELTA ) GO TO 100
C
C     SOLUTION
C
      DO I = 1, N
        A(13*N+I) = A(14*N+I)
      END DO
      WRITE ( 01, * ) (A(11*N+I), I=1,N)
      WRITE ( 01, * ) (A(13*N+I), I=1,N)
      CLOSE (01)
      STOP
      END
C
      SUBROUTINE INIT (U, X1, X2, N1, N2)
C
C     INITIAL CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  U(N1,N2), X1(N1), X2(N2)
      DO I = 1, N1
        DO J = 1, N2
          U(I,J) = 0.D0
          IF ((X1(I)-0.6D0)**2 + (X2(J)-0.6D0)**2.LE.0.04D0)
     +            U(I,J) = 1.D0
        END DO
      END DO
C
      RETURN
      END
C
      SUBROUTINE FDST (A0, A1, A2, F, U, H1, H2, N1, N2, TAU)
C
C     GENERATION OF DIFFERENCE-SCHEME COEFFICIENTS
C     FOR THE PARABOLIC EQUATION WITH CONSTANT COEFFICIENTS
C     IN THE CASE OF PURELY IMPLICIT SCHEME
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  A0(N1,N2), A1(N1,N2), A2(N1,N2), F(N1,N2), U(N1,N2)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          A1(I-1,J) = 1.D0/(H1*H1)
```

```
            A1(I,J)   = 1.D0/(H1*H1)
            A2(I,J-1) = 1.D0/(H2*H2)
            A2(I,J)   = 1.D0/(H2*H2)
            A0(I,J)   = A1(I,J) + A1(I-1,J) + A2(I,J) + A2(I,J-1)
     +                + 1.D0/TAU
            F(I,J)    = U(I,J)/TAU
          END DO
        END DO
C
C     FIRST-KIND HOMOGENEOUS BOUNDARY CONDITION
C
        DO J = 2, N2-1
          A0(1,J) = 1.D0
          A1(1,J) = 0.D0
          A2(1,J) = 0.D0
          F(1,J)  = 0.D0
        END DO
C
        DO J = 2, N2-1
          A0(N1,J)   = 1.D0
          A1(N1-1,J) = 0.D0
          A1(N1,J)   = 0.D0
          A2(N1,J)   = 0.D0
          F(N1,J)    = 0.D0
        END DO
C
        DO I = 2, N1-1
          A0(I,1) = 1.D0
          A1(I,1) = 0.D0
          A2(I,1) = 0.D0
          F(I,1)  = 0.D0
        END DO
C
        DO I = 2, N1-1
          A0(I,N2)   = 1.D0
          A1(I,N2)   = 0.D0
          A2(I,N2)   = 0.D0
          A2(I,N2-1) = 0.D0
          F(I,N2)    = 0.D0
        END DO
C
         A0(1,1) = 1.D0
         A1(1,1) = 0.D0
         A2(1,1) = 0.D0
         F(1,1)  = 0.D0
C
         A0(N1,1) = 1.D0
         A2(N1,1) = 0.D0
         F(N1,1)  = 0.D0
C
         A0(1,N2) = 1.D0
         A1(1,N2) = 0.D0
         F(1,N2)  = 0.D0
C
         A0(N1,N2) = 1.D0
         F(N1,N2)  = 0.D0
C
        RETURN
        END
C
```

```
      SUBROUTINE FDSB (A0, A1, A2, F, U, H1, H2, N1, N2)
C
C     GENERATION OF DIFFERENCE-SCHEME COEFFICIENTS
C     FOR THE DIFFERENCE ELLIPTIC EQUATION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  A0(N1,N2), A1(N1,N2), A2(N1,N2), F(N1,N2), U(N1,N2)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          A1(I-1,J) = 1.D0/(H1*H1)
          A1(I,J)   = 1.D0/(H1*H1)
          A2(I,J-1) = 1.D0/(H2*H2)
          A2(I,J)   = 1.D0/(H2*H2)
          A0(I,J)   = A1(I,J) + A1(I-1,J) + A2(I,J) + A2(I,J-1)
     +              + 1.D0
          F(I,J)    = U(I,J)
        END DO
      END DO
C
C     FIRST-KIND HOMOGENEOUS BOUNDARY CONDITIONS
C
      DO J = 2, N2-1
        A0(1,J) = 1.D0
        A1(1,J) = 0.D0
        A2(1,J) = 0.D0
        F(1,J)  = 0.D0
      END DO
C
      DO J = 2, N2-1
        A0(N1,J)   = 1.D0
        A1(N1-1,J) = 0.D0
        A1(N1,J)   = 0.D0
        A2(N1,J)   = 0.D0
        F(N1,J)    = 0.D0
      END DO
C
      DO I = 2, N1-1
        A0(I,1) = 1.D0
        A1(I,1) = 0.D0
        A2(I,1) = 0.D0
        F(I,1)  = 0.D0
      END DO
C
      DO I = 2, N1-1
        A0(I,N2)   = 1.D0
        A1(I,N2)   = 0.D0
        A2(I,N2)   = 0.D0
        A2(I,N2-1) = 0.D0
        F(I,N2)    = 0.D0
      END DO
C
       A0(1,1) = 1.D0
       A1(1,1) = 0.D0
       A2(1,1) = 0.D0
       F(1,1)  = 0.D0
C
       A0(N1,1) = 1.D0
       A2(N1,1) = 0.D0
       F(N1,1)  = 0.D0
```

```
C
      A0(1,N2) = 1.D0
      A1(1,N2) = 0.D0
      F(1,N2)  = 0.D0
C
      A0(N1,N2) = 1.D0
      F(N1,N2)  = 0.D0
C
      RETURN
      END
```

In the subroutine FDST, coefficients of the difference elliptic problem are generated for the problem to be solved at the next time step. In the subroutine FDSB, coefficients of the difference elliptic operator B in the iterative process (7.179) are generated, so that

$$By = -\sum_{\beta=1}^{2} y_{\bar{x}_\beta x_\beta} + y.$$

Difference elliptic problems are solved in the subroutine SBAND5.

7.3.5 Computational experiments

Like in the case of regularized difference schemes intended for approximate solution of inverted-time problems, a uniform grid with $h_1 = 0.02$ and $h_2 = 0.02$ for the model problem with $k(\boldsymbol{x}) = 1$ in unit square was used. In the framework of a quasi-real experiment, the direct problem with $T = 0.025$ is solved using a time grid with $\tau = 0.00025$. A purely implicit difference scheme ($\sigma = 1$) was employed. Again, in the direct problem the initial condition (the exact end-time solution of the inverse problem) is given by

$$u_0(\boldsymbol{x}, 0) = \begin{cases} 1, & (x_1 - 0.6)^2 + (x_2 - 0.6)^2 \le 0.04, \\ 0, & (x_1 - 0.6)^2 + (x_2 - 0.6)^2 > 0.04. \end{cases}$$

The end-time solution of the direct problem is shown in Figure 7.5.

To illustrate the capability of the adopted scheme in reconstructing a piecewise-discontinuous initial condition, here we present computational data obtained with unperturbed input data (the difference solution of the direct problem at $t = T$). The iterative process was terminated when the difference $r_k = Av_k - \varphi$ attained the estimate $\|r_k\| \le \varepsilon$. Approximate solutions obtained with $\varepsilon = 0.001$ and $\varepsilon = 0.0001$ are shown in Figures 7.9 and 7.10, respectively (in the figures, contour lines with the step size $\triangle u = 0.05$ are plotted). A substantial (tenfold) change in the solution accuracy for the inverse problem results in slight refinement of the approximate solution. In the case of interest, we cannot expect that a more accurate reconstruction of the piecewise-discontinuous initial condition can be achieved.

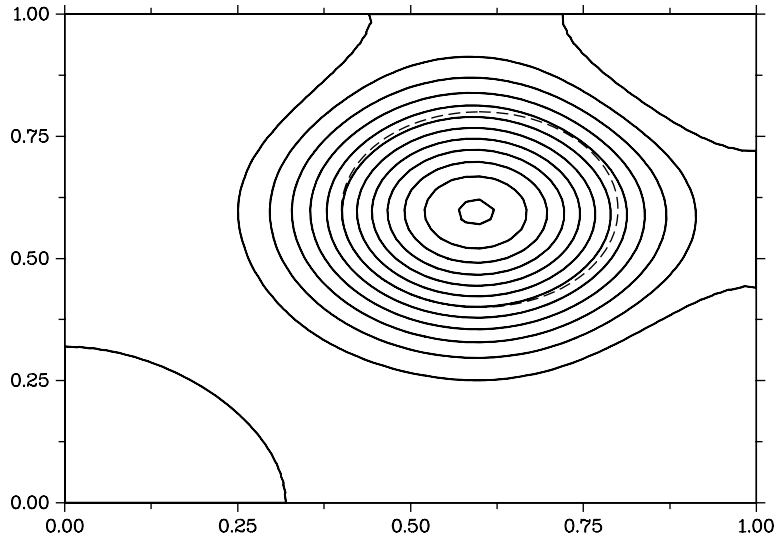

Figure 7.9 Inverse-problem solution obtained with $\varepsilon = 0.001$

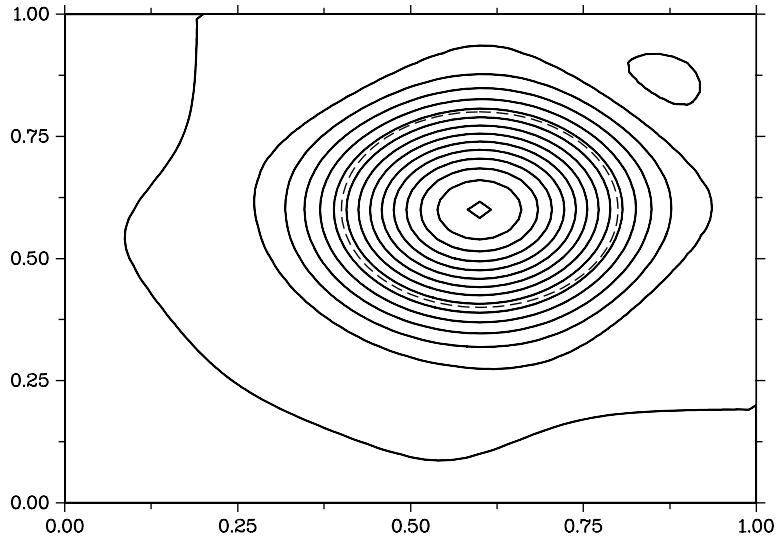

Figure 7.10 Inverse-problem solution obtained with $\varepsilon = 0.0001$

In perturbing input data, we have to try identify the smooth initial solution of the inverse problem using a proper choice of the operator $B \neq E$ in the iterative process (7.179). Figure 7.11 shows the solution of the inverse problem obtained at the inaccuracy level $\delta = 0.01$. Solutions of the problem obtained at a higher and lower inaccuracy levels are shown in Figures 7.12 and 7.13.

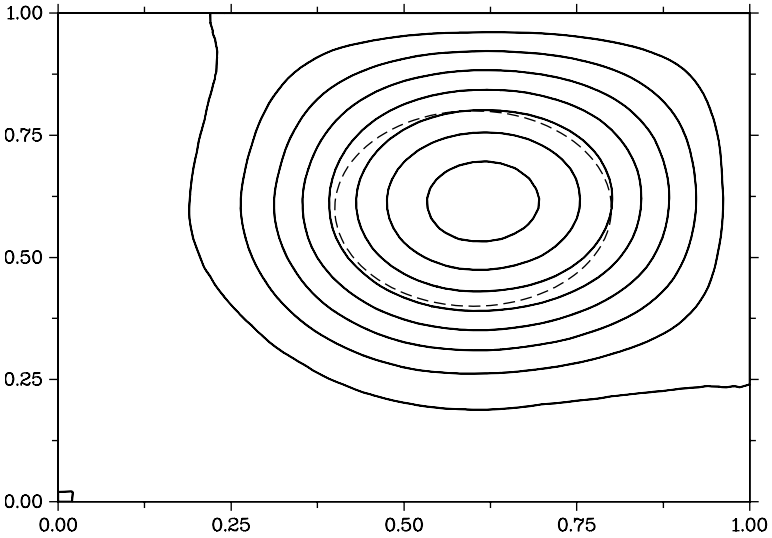

Figure 7.11 Inverse-problem solution obtained with $\delta = 0.01$

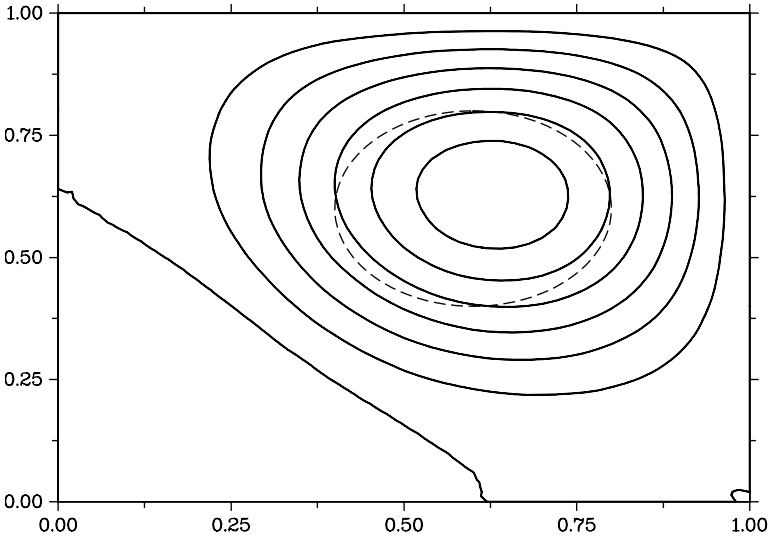

Figure 7.12 Inverse-problem solution obtained with $\delta = 0.02$

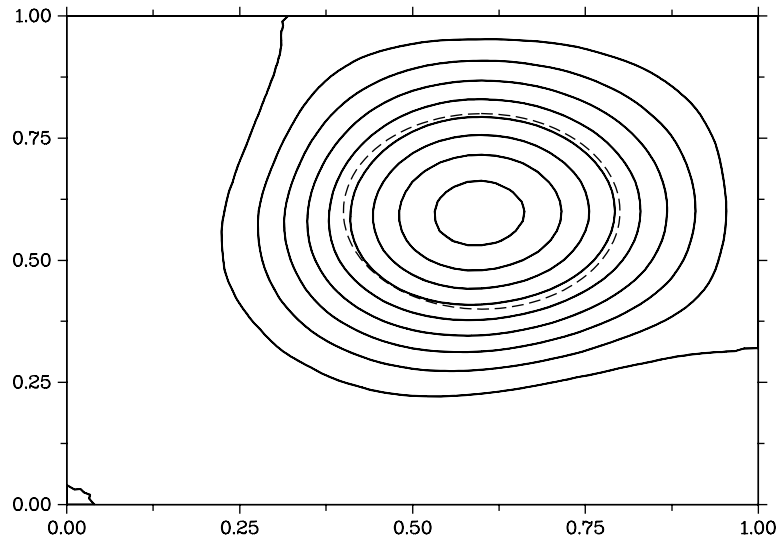

Figure 7.13 Inverse-problem solution obtained with $\delta = 0.005$

7.4 Second-order evolution equation

A classical example of ill-posed problems is the Cauchy problem for the elliptic equation (Hadamard examples). In this section, major possibilities available in the construction of stable solution algorithm for such evolutionary inverse problems are considered. Methods using the perturbation of initial conditions and the perturbation of the initial second-order evolution equation are briefly discussed. A program in which, to approximately solve the Cauchy problem for the Laplace equation, regularized difference schemes are used, is presented.

7.4.1 Model problem

Consider the simplest inverse problem for the two-dimensional Laplace equation. Let us start with formulating the direct problem. Suppose that in the rectangle

$$\overline{Q}_T = \overline{\Omega} \times [0, T], \qquad \overline{\Omega} = \{x \mid 0 \le x \le l\}, \quad 0 \le t \le T$$

the function $u(x, t)$ satisfies the equation

$$\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2 u}{\partial x^2} = 0, \qquad 0 < x < l, \quad 0 < t < T. \tag{7.180}$$

We supplement this equation with some boundary conditions. On the lateral sides of the rectangle, we put:

$$u(0, t) = 0, \qquad u(l, t) = 0, \qquad 0 < t < T. \tag{7.181}$$

On the top and bottom sides, the following boundary conditions are considered:

$$\frac{\partial u}{\partial t}(x,0) = 0, \qquad 0 \le x \le l, \tag{7.182}$$

$$u(x,T) = u_T(x), \qquad 0 \le x \le l. \tag{7.183}$$

Among the inverse evolutionary problems for equation (7.180), we can identify the Cauchy problem and the continuation problem for the solution of the boundary value problem. The Cauchy problem is formulated as follows. Suppose that the boundary condition at $t = T$ is not specified, but we know the solution at $t = 0$, i.e., instead of (7.183), the following condition is considered:

$$u(x,0) = u_0(x), \qquad 0 \le x \le l. \tag{7.184}$$

In practical cases, of interest can be the problem in which it is required to protract the solution of the direct problem (7.180)–(7.183) beyond the calculation-domain boundary. For instance, we are interested in the solution $u(x,t)$ in the domain $\overline{Q}_{T+\triangle T}$ with $\triangle T > 0$, i.e., here the solution is to be continued into a region adjacent to the top portion of the boundary. After the solution of the problem inside the calculation domain $\overline{Q}_T$ is found, the continuation problem reduces to a Cauchy problem similar to problem (7.180)–(7.182), (7.184). Also of interest is the possibility of passing from the Cauchy problem to a continuation problem. With this remark, in what follows we will concentrate our attention on the Cauchy problem.

We will consider problem (7.180)–(7.182), (7.184) from a somewhat more general standpoint as the Cauchy problem for the second-order differential-operator equation. For functions given on the interval $\Omega = (0,1)$ we define, in the traditional manner, the Hilbert space $\mathcal{H} = \mathcal{L}_2(\Omega)$. On the set of functions vanishing at $\partial\Omega$ we define the operator

$$\mathcal{A}u = -\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < l. \tag{7.185}$$

Among the key properties of this operator, note that in $\mathcal{H}$ we have:

$$\mathcal{A}^* = \mathcal{A} \ge mE, \qquad m > 0. \tag{7.186}$$

Equation (7.180), supplemented with condition (7.181) at the boundary, can be written as the following differential-operator equation for $u(t) \in \mathcal{H}$:

$$\frac{d^2 u}{dt^2} - \mathcal{A}u = 0, \qquad 0 < t \le T. \tag{7.187}$$

Initial conditions (7.182) and (7.184) yield:

$$u(0) = u_0, \tag{7.188}$$

$$\frac{du}{dt}(0) = 0. \tag{7.189}$$

In (7.187), the operator $\mathcal{A}$ is a self-adjoint, positively defined operator. Problem (7.187)–(7.189) is an ill-posed problem because, here, continuous dependence on input data (initial conditions) is lacking. Conditional well-posedness takes place in the class of solutions bounded in $\mathcal{H}$.

7.4.2 Equivalent first-order equation

In the construction of regularizing algorithms for the approximate solution of problem (7.187)–(7.189), passage to a Cauchy problem for the first-order evolution equation may prove useful. In the latter case, we can follow the above-considered approaches to the solution of ill-posed problems for evolution equations using perturbed initial conditions and/or perturbed equation.

The simplest transformation related with the traditional introduction of the vector of unknown quantities $U = \{u_1, u_2\}$, $u_1 = u$, $u_2 = du/dt$ results in a system of first-order equations with a non-self-adjoint operator. This approach will be discussed in more detail below.

In the case of problem (7.187)–(7.189), the self-adjointness and positive definiteness of $\mathcal{A}$ can be taken into account.

We perform the following change of variables:

$$v(t) = \frac{1}{2}\Big(u - \mathcal{A}^{-1/2}\,\frac{du}{dt}\Big), \quad w(t) = \frac{1}{2}\Big(u + \mathcal{A}^{-1/2}\,\frac{du}{dt}\Big). \tag{7.190}$$

Then, from (7.187) it readily follows that the new unknown quantities $v(t)$ and $w(t)$ satisfy the following first-order equations:

$$\frac{dv}{dt} + \mathcal{A}^{1/2}v = 0, \qquad \frac{dw}{dt} - \mathcal{A}^{1/2}w = 0. \tag{7.191}$$

With regard to (7.188), (7.189) and for the introduced notation (7.190), for equations (7.191) we pose the initial conditions

$$v(0) = \frac{1}{2}\,u_0, \qquad w(0) = \frac{1}{2}\,u_0. \tag{7.192}$$

Thus, starting from the ill-posed problem (7.187)–(7.189), we arrive at a well-posed problem in which it is required to determine $v(t)$, and also at an ill-posed problem for $w(t)$. Hence, regularizing algorithms for problem (7.187)–(7.189) can be constructed based on regularization of the split system (7.191). Below, in the consideration of the generalized inverse method for problem (7.187)–(7.189), we will outline some possibilities available along this direction. In practical realizations, it seems reasonable to use perturbations related with the calculation of the square root of $\mathcal{A}$.

We define the vector $U = \{u_1, u_2\}$ and the space $\mathcal{H}^2$ as the direct sum of spaces $\mathcal{H}$: $\mathcal{H}^2 = \mathcal{H} \oplus \mathcal{H}$. The addition in $\mathcal{H}^2$ is to be performed coordinatewise, and the scalar product there is defined as follows:

$$(U, V) = (u_1, v_1) + (u_2, v_2).$$

Suppose that $u_1 = v$, $u_2 = w$; then, system (7.191) assumes the form

$$\frac{dU}{dt} - \mathcal{L}U = 0, \tag{7.193}$$

where

$$\mathcal{L} = \begin{bmatrix} -\mathcal{A}^{1/2} & 0 \\ 0 & \mathcal{A}^{1/2} \end{bmatrix}. \tag{7.194}$$

Equation (7.194) is supplemented with the boundary condition (see (7.192))

$$U(0) = U_0, \qquad U_0 = \left\{\frac{1}{2}u_0, \frac{1}{2}u_0\right\}. \tag{7.195}$$

Problem (7.193), (7.195) belongs to the above-considered class of ill-posed problems for the first-order equation. The specific feature of the problem is due to fact that the operator $\mathcal{L}$ (see (7.194)) may change its sign.

Consider some possibilities in the passage to an equivalent system of first-order equations under more general conditions, for instance, in the case in which the operator $\mathcal{A}$ is a non-self-adjoint operator. Suppose that the components of $U = \{u_1, u_2\}$ from $\mathcal{H}^2$ are defined as

$$u_1 = u, \qquad u_2 = \frac{du}{dt}. \tag{7.196}$$

Then, the initial problem (7.187)–(7.189) can be written as (compare with (7.193)–(7.195)) equation (7.193), supplemented now with the initial condition

$$U(0) = U_0, \qquad U_0 = \{u_0, 0\}. \tag{7.197}$$

For the operator $\mathcal{L}$ we obtain the representation

$$\mathcal{L} = \begin{bmatrix} 0 & E \\ \mathcal{A} & 0 \end{bmatrix}. \tag{7.198}$$

Thus, here again from the Cauchy problem for the evolution second-order equation we have passed to the Cauchy problem for the evolution first-order equation.

7.4.3 Perturbed initial conditions

Let us briefly outline the available possibilities in using methods based on non-local perturbation of initial conditions as applied to the approximate solution of the ill-posed Cauchy problem for the second-order evolution equation. Generally speaking, in the case of general problems, two initial conditions are to be perturbed. Suppose that we pose first an ill-posed problem in which it is required to determine the function $u = u(t) \in \mathcal{H}$ from equation (7.187) supplemented with initial conditions (7.188) and (7.189).

The approximate solution $u_\alpha(t)$ is defined as the solution of the following non-local problem:

$$\frac{d^2u_\alpha}{dt^2} - \mathcal{A}u_\alpha = 0, \qquad 0 < t \le T, \tag{7.199}$$

$$u_\alpha(0) + \alpha u_\alpha(T) = u_0, \tag{7.200}$$

$$\frac{du_\alpha}{dt}(0) = 0. \tag{7.201}$$

A specific feature in the non-local problem (7.199)–(7.201) consists in the fact that here only one initial condition (see (7.200)) is to be perturbed. The stability estimate is given by the following statement.

Theorem 7.20 *For the solution of the non-local problem* (7.199)–(7.201) *there holds the estimate*

$$\|u_\alpha(t)\| \le \frac{1}{\alpha}\,\|u_0\|. \tag{7.202}$$

Proof. The solution of problem (7.199)–(7.201) can be written in the traditional operator form

$$u_\alpha(t) = R(t,\alpha)u_0, \tag{7.203}$$

where

$$R(t,\alpha) = \operatorname{ch}(\mathcal{A}^{1/2}t)(E + \alpha \operatorname{ch}(\mathcal{A}^{1/2}T))^{-1}. \tag{7.204}$$

We assume that the spectrum of $\mathcal{A}$ is discrete, consisting of eigenvalues $0 < \lambda_1 \le \lambda_2 \le \cdots$, and the related system of eigenfunctions $\{w_k\}$, $w_k \in D(\mathcal{A})$, $k = 1, 2, \ldots$ is an orthonormal complete system in $\mathcal{H}$. In such conditions, for the solution of problem (7.199)–(7.201) we have the representation

$$u_\alpha(t) = \sum_{k=1}^{\infty}(u_0, w_k) \operatorname{ch}(\lambda_k^{1/2}t)(1 + \alpha \operatorname{ch}(\lambda_k^{1/2}T))^{-1}w_k.$$

From here, the stability estimate (7.203) follows. □

Using the perturbed initial condition, one can establish regularizing properties of the algorithm perfectly analogous to the case of the first-order equation (see Theorem 7.4). Suppose that, instead of the exact initial condition u_0, an approximate solution u_0^δ is given. As usually, we assume that there exists a stability estimate of form

$$\|u_0^\delta - u_0\| \le \delta. \tag{7.205}$$

The approximate solution for inaccurate initial conditions is to be found from equation (7.199), condition (7.201), and from the non-local condition

$$u_\alpha(0) + \alpha u_\alpha(T) = u_0^\delta. \tag{7.206}$$

Theorem 7.21 *With estimate* (7.205) *fulfilled, the solution of problem* (7.199), (7.201), (7.206) *converges to the solution of problem* (7.187)–(7.189), *bounded in* $\mathcal{H}$, *provided that* $\delta \to 0$, $\alpha(\delta) \to 0$, $\delta/\alpha \to 0$.

Consider the relation between the non-local problem (7.199)–(7.201) and the optimal control problem. In the case of interest, both the solution of the optimal control problem and the solution of the non-local problem are obtained based on the variable separation method.

Let $v \in \mathcal{H}$ be the sought control and $u_\alpha(t; v)$ be defined as the solution of the problem

$$\frac{d^2 u_\alpha}{dt^2} - \mathcal{A} u_\alpha = 0, \qquad 0 < t \le T, \tag{7.207}$$

$$u_\alpha(T; v) = v, \tag{7.208}$$

$$\frac{du_\alpha}{dt}(0) = 0. \tag{7.209}$$

We consider the quality functional in the following simplest form:

$$J_\alpha(v) = \|u_\alpha(0; v) - u_0\|^2 + \alpha \|v\|^2. \tag{7.210}$$

For the optimal control w, we have:

$$J_\alpha(w) = \min_{v \in \mathcal{H}} J_\alpha(v). \tag{7.211}$$

The solution of the optimal control problem (7.207)–(7.211) can be written in the form (7.203) provided that

$$R(t, \alpha) = \operatorname{ch}(\mathcal{A}^{1/2} t)(E + \alpha \operatorname{ch}^2(\mathcal{A}^{1/2} T))^{-1}. \tag{7.212}$$

Through comparison of (7.204) with (7.212), the following statement can be established.

Theorem 7.22 *The solution of the optimal control problem* (7.207)–(7.211) *coincides with the solution of the non-local problem for equation* (7.199) *on the double interval* $(0, 2T)$ *with initial condition* (7.201) *and non-local*

$$u_\alpha(0) + \frac{\alpha}{2 + \alpha} u_\alpha(2T) = \frac{2}{2 + \alpha} u_0. \tag{7.213}$$

Proof. To derive (7.213) we use for the solution the representation (7.212) and formula $2 \operatorname{ch}(x) = \operatorname{ch}(2x) + 1$. □

Some other non-local conditions can be obtained by writing the Cauchy problem (7.187)–(7.189) as a Cauchy problem for the first-order evolution equation. An example here is problem (7.193)–(7.195).

We determine the approximate solution of problem (7.193)–(7.195) as the solution of the non-local problem

$$\frac{dU_\alpha}{dt} - \mathcal{L}U_\alpha = 0, \tag{7.214}$$

$$U_\alpha(0) + \alpha U_\alpha(T) = U_0. \tag{7.215}$$

To formulate the related non-local problem for the second-order equation, in problem (7.214), (7.215) we perform the reverse transition. By analogy with (7.190), we represent the solution of (7.214) as

$$\begin{aligned} u_1(t) &= \frac{1}{2}\Big(u_\alpha - \mathcal{A}^{-1/2}\,\frac{du_\alpha}{dt}\Big), \\ u_2(t) &= \frac{1}{2}\Big(u_\alpha + \mathcal{A}^{-1/2}\,\frac{du_\alpha}{dt}\Big). \end{aligned} \tag{7.216}$$

From (7.216) it follows that

$$\begin{aligned} u_\alpha(t) &= u_1(t) + u_2(t), \\ \frac{du_\alpha}{dt}(t) &= \mathcal{A}^{-1/2}(u_2(t) - u_1(t)). \end{aligned} \tag{7.217}$$

Then, with (7.195) the non-local condition (7.215) yields:

$$\begin{aligned} u_1(0) + \alpha u_1(T) &= \frac{1}{2}\,u_0, \\ u_2(0) + \alpha u_2(T) &= \frac{1}{2}\,u_0. \end{aligned} \tag{7.218}$$

With (7.216), (7.217) taken into account, from (7.218) we obtain:

$$\begin{aligned} u_\alpha(t) + \alpha u_\alpha(T) &= u_0, \\ \frac{du_\alpha}{dt}(0) + \frac{du_\alpha}{dt}(T) &= 0. \end{aligned} \tag{7.219}$$

We arrive at two non-local conditions for the solution $u_\alpha(t)$ of equation (7.199). Here (compare with (7.200) and (7.201)), two, instead of one, initial conditions are to be perturbed.

7.4.4 Perturbed equation

Consider the possibilities available in the construction of regularized solution algorithms suitable for inverse problems for second-order equations based on the perturbed initial equation. Here, we can use the above-considered versions of the generalized inverse method for first-order evolution equations.

To approximately solve problem (7.187)–(7.189), we use the generalized inverse method in its version analogous to the basic version of this method for first-order equations. The approximate solution $u_\alpha(t)$ is to be found from the equation

$$\frac{d^2u_\alpha}{dt^2} - \mathcal{A}u_\alpha + \alpha\mathcal{A}^2u_\alpha = 0, \qquad 0 < t \le T. \tag{7.220}$$

The initial conditions are given with some inaccuracy. Suppose that

$$u_\alpha(0) = u_0^\delta, \tag{7.221}$$

$$\frac{du_\alpha}{dt}(0) = 0. \tag{7.222}$$

We assume that for the initial-condition inaccuracy the estimate (7.205) holds.

The solution of (7.220)–(7.222) can be written in the operator form

$$u_\alpha(t) = R(t,\alpha)u_0^\delta, \tag{7.223}$$

where

$$R(t,\alpha) = \operatorname{ch}\left((\mathcal{A} - \alpha\mathcal{A}^2)^{1/2}t\right).$$

Using the variable separation method, we can prove convergence of the approximate solution to the exact solution.

Theorem 7.23 *Let for the initial-condition inaccuracy the estimate* (7.205) *hold. Then, in the case of* $\delta \to 0$, $\alpha(\delta) \to 0$, $\delta \exp\left(T/(4\alpha)^{1/2}\right) \to 0$ *the approximate solution* $u_\alpha(t)$*, determined as the solution of problem* (7.220)–(7.222)*, converges to the exact solution* $u(t)$ *of problem* (7.187)–(7.189)*, bounded in* $\mathcal{H}$*, and the following stability estimate with respect to initial data holds:*

$$\|u_\alpha(t)\| \le \operatorname{ch}\left(\frac{t}{2\sqrt{\alpha}}\right)\|u_\alpha(0)\|. \tag{7.224}$$

Proof. To prove the statement and, in particular, to show the validity of estimate (7.224), we can follow the procedure previously used to prove analogous results for the first-order equation (see Theorem 7.12, for instance). □

The version of the generalized inverse method for equation (7.187) analogous to pseudo-parabolic perturbation of the first-order equation consists in the determination of the approximate solution from the equation

$$\frac{d^2u_\alpha}{dt^2} - \mathcal{A}u_\alpha + \alpha\mathcal{A}\frac{d^2u_\alpha}{dt^2} = 0, \qquad 0 < t \le T, \tag{7.225}$$

supplemented with initial conditions (7.221), (7.222).

The approximate solution converges to the exact solution provided that the value of the regularization parameter is matched with the input-data inaccuracy, and for the solution of problem (7.221), (7.222), (7.225) there holds the following stability estimate with respect to initial data:

$$\|u_\alpha(t)\| \le \operatorname{ch}\Big(\frac{t}{\sqrt{\alpha}}\Big)\|u_\alpha(0)\|. \tag{7.226}$$

Variants of the generalized inverse method for the approximate solution of the ill-posed Cauchy problem (7.187)–(7.189) can be constructed based on the passage to an equivalent Cauchy problem for the system of first-order equations. Instead of (7.187)–(7.189), consider the problem (7.193)–(7.195), in which

$$\begin{aligned} u_1(t) = v(t) &= \frac{1}{2}\Big(u - \mathcal{A}^{-1/2}\frac{du}{dt}\Big), \\ u_2(t) = w(t) &= \frac{1}{2}\Big(u + \mathcal{A}^{-1/2}\frac{du}{dt}\Big). \end{aligned} \tag{7.227}$$

Taking the fact into account that the operator $\mathcal{L}$ is a self-adjoint operator that can change its sign, we find the approximate solution $U_\alpha(t)$ of problem (7.193), (7.195) from the equation

$$\frac{dU_\alpha}{dt} - \mathcal{L}U_\alpha + \alpha\mathcal{L}^2U_\alpha = 0 \tag{7.228}$$

and the initial conditions

$$U_\alpha(0) = U_0. \tag{7.229}$$

Based on the previously obtained results (see Theorem 7.11), derive the following estimate of stability of the approximate solution $U_\alpha(t)$ with respect to initial data:

$$\|U_\alpha(t)\| \le \exp\Big(\frac{t}{4\alpha}\Big)\|U_\alpha(0)\|. \tag{7.230}$$

First, we derive an equation from which the approximate solution $u_\alpha(t)$ that corresponds to system (7.228) can be found. In view of (7.194), we have

$$\mathcal{L}^2 = \begin{bmatrix} \mathcal{A} & 0 \\ 0 & \mathcal{A} \end{bmatrix}.$$

With regard to (7.228), for $v_\alpha(t)$, $w_\alpha(t)$ (with $U_\alpha = \{v_\alpha, w_\alpha\}$) we obtain the following system of equations:

$$\begin{aligned} \frac{dv_\alpha}{dt} + \mathcal{A}^{1/2}v_\alpha + \alpha\mathcal{A}v_\alpha = 0, \\ \frac{dw_\alpha}{dt} - \mathcal{A}^{1/2}w_\alpha + \alpha\mathcal{A}w_\alpha = 0. \end{aligned} \tag{7.231}$$

By analogy with the exact solution $u(t)$ (see (7.227)), we define the approximate solution $u_\alpha(t)$ by the following equation:

$$u_\alpha(t) = v_\alpha(t) + w_\alpha(t). \tag{7.232}$$

From (7.231) and (7.232), it follows immediately that the function $u_\alpha(t)$ satisfies the equation

$$\frac{d^2u_\alpha}{dt^2} - \mathcal{A}u_\alpha + 2\alpha\mathcal{A}\frac{du_\alpha}{dt} + \alpha^2\mathcal{A}^2u_\alpha = 0, \quad 0 < t \le T. \tag{7.233}$$

It follows from (7.232) that the solution of equation (7.233), supplemented with appropriate initial conditions, satisfies

$$\begin{aligned} \|u_\alpha\|^2 &\le (\|v_\alpha\| + \|w_\alpha\|)^2 \le 2(\|v_\alpha\|^2 + \|w_\alpha\|^2) = 2\|U_\alpha\|^2 \\ &\le 2\exp\left(\frac{t}{2\alpha}\right)\|U_\alpha(0)\|^2 \le \exp\left(\frac{t}{2\alpha}\right)\|u_\alpha(0)\|^2. \end{aligned}$$

Of course, the same estimate can be obtained starting from the estimates for $v_\alpha(t)$ and $w_\alpha(t)$ derived from equation (7.231).

Equation (7.233) was obtained by perturbing two terms. The regularization performed by perturbing one of the terms $(\alpha^2\mathcal{A}^2u_\alpha)$ was considered previously (see equation (7.220)). Of interest here is to consider, in its pure form, the regularization performed at the expense of the second term. For this reason, we will seek the approximate solution from the equation (see (7.233))

$$\frac{d^2u_\alpha}{dt^2} - \mathcal{A}u_\alpha + \alpha\mathcal{A}\frac{du_\alpha}{dt} = 0, \qquad 0 < t \le T, \tag{7.234}$$

supplemented with the initial conditions (7.221), (7.222). Such regularization, applied to well-posed problems for evolution equations, is called "parabolic" regularization.

Similarly to Theorem 7.23, we can prove an analogous statement about regularizing properties of the generalized inverse method in its "parabolic"-regularization version. Here, we give only the stability estimate with respect to initial data:

$$\|u_\alpha(t)\| \le \exp\left(\frac{t}{\alpha}\right)\|u_\alpha(0)\|.$$

We have restricted ourselves to the application of the generalized inverse method in its standard version (7.228), (7.229) applied to problem (7.193)–(7.195). Some additional possibilities are given by pseudo-parabolic perturbation of equation (7.193).

7.4.5 Regularized difference schemes

Let us turn now to constructing regularized difference schemes for the approximate solution of the ill-posed Cauchy problem for the second-order evolution equation.

On the interval $\bar{\Omega} = [0, l]$, we introduce a uniform grid with a grid size h:

$$\bar{\omega} = \{x \mid x = x_i = ih,\ i = 0, 1, \dots, N,\ Nh = l\}.$$

Here, ω is the set of internal nodes and $\partial\omega$ is the set of boundary nodes.

At the internal nodes, we approximate the differential operator (7.185), accurate to second order, with the difference operator

$$\Lambda y = -y_{\bar{x}x}, \qquad x \in \omega. \tag{7.235}$$

In the mesh Hilbert space H, we introduce the norm by the relation $\|y\| = (y, y)^{1/2}$, in which

$$(y, w) = \sum_{x \in \omega} y(x) w(x) h.$$

On the set of functions vanishing on $\partial\omega$, for the self-adjoint operator Λ there holds the estimate

$$\Lambda = \Lambda^* \geq \lambda_0 E, \tag{7.236}$$

in which

$$\lambda_0 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l} \ \geq \ \frac{8}{l^2}.$$

On the approximation over the space, to problem (7.187)–(7.189) we put into correspondence the problem

$$\frac{d^2 y}{dt^2} - \Lambda y = 0, \qquad x \in \omega, \quad t > 0, \tag{7.237}$$

$$y(x, 0) = u_0(x), \qquad x \in \omega, \tag{7.238}$$

$$\frac{dy}{dt}(x, 0) = 0, \qquad x \in \omega. \tag{7.239}$$

Regularized difference schemes for problem (7.237)–(7.239) can be constructed based on the regularization principle for difference schemes. As a generic (initial) difference scheme, we adopt the explicit symmetric difference scheme

$$\begin{gathered} \frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} - \Lambda y_n = 0, \\ \boldsymbol{x} \in \omega, \qquad n = 1, 2, \ldots, N_0 - 1 \end{gathered} \tag{7.240}$$

with some initial conditions.

With regard to (7.238), we have:

$$y_0 = u_0(x), \qquad x \in \omega.$$

The solution at the time $t = \tau$ can be approximated, with conditions (7.238), (7.239) on the solutions of (7.237). Taking into account the relation

$$y_1 = y_0 + \tau \frac{dy}{dt}(0) + \frac{\tau^2}{2} \frac{d^2 y}{dt^2}(0) + \mathcal{O}(\tau^3),$$

to approximate (7.239), we use the difference relation

$$\frac{y_1 - y_0}{\tau} + \frac{\tau}{2} \Lambda y_0 = 0.$$

Scheme (7.240) can be written in the canonical form for three-layer difference schemes

$$B\,\frac{y_{n+1}-y_{n-1}}{2\tau}+R(y_{n+1}-2y_n+y_{n-1})+Ay_n=0 \tag{7.241}$$

with operators

$$B=0,\qquad R=\frac{1}{\tau^2}\,E,\qquad A=-\Lambda, \tag{7.242}$$

i. e., with $A=A^*<0$.

Theorem 7.24 *The explicit scheme* (7.240) *is* ρ*-stable with*

$$\rho=\exp\left(M^{1/2}\tau\right). \tag{7.243}$$

Proof. For scheme (7.240), we check that the following conditions for ρ-stability (see Theorem 4.11) are fulfilled:

$$\frac{\rho^2+1}{2}\,B+\tau(\rho^2-1)R\geq 0, \tag{7.244}$$

$$\frac{\rho^2-1}{2\tau}\,B+(\rho-1)^2R+\rho A>0, \tag{7.245}$$

$$\frac{\rho^2-1}{2\tau}\,B+(\rho+1)^2R-\rho A>0. \tag{7.246}$$

Apparently, in the case of $B\geq 0$, $A\leq 0$, $R\geq 0$ and $\rho>1$ (see (7.242)) inequalities (7.244) and (7.246) are fulfilled for all $\tau>0$.

With (7.242), inequality (7.245) yields:

$$(\rho-1)^2E-\tau^2\rho\Lambda\ \geq\ ((\rho-1)^2M^{-1}-\tau^2\rho)\Lambda>0.$$

Derivation of the estimates is based on the following useful result.

Lemma 7.25 *Inequality*

$$(\rho-1)^2\chi-\tau^2\rho>0 \tag{7.247}$$

is fulfilled for positive χ, τ *and for* $\rho>1$ *in the case of*

$$\rho\geq\exp\left(\chi^{-1/2}\tau\right).$$

Proof. Inequality (7.247) is fulfilled in the case of $\rho>\rho_2$, where

$$\rho_2=1+\frac{1}{2}\,\tau^2\chi^{-1}+\tau\chi^{-1/2}\Big(1+\frac{1}{4}\,\tau^2\chi^{-1}\Big)^{1/2}.$$

Taking into account the inequality

$$\Big(1+\frac{1}{4}\,\tau^2\chi^{-1}\Big)^{1/2}<1+\frac{1}{8}\,\tau^2\chi^{-1},$$

we obtain

$$\begin{aligned}\rho_2 &< 1+\tau\chi^{-1/2}+\frac{1}{2}\tau^2\chi^{-1}+\frac{1}{8}\tau^3\chi^{-3/2}\\ &< 1+\tau\chi^{-1/2}+\frac{1}{2}\tau^2\chi^{-1}+\frac{1}{6}\tau^3\chi^{-3/2} < \exp(\chi^{-1/2}\tau).\end{aligned}$$

Thus, the lemma is proved. □

In the case of interest, we have $\chi = M^{-1}$ and, hence, for ρ we obtain the desired estimate (4.32) for the explicit scheme (7.240). □

With regard to the boundedness of Λ (in Cauchy problems for elliptic equations), we conclude that the grid step over space limits the solution growth, i.e., serves the regularization parameter.

Let us construct now, at the expense of introducing penalty terms into the difference operators of difference schemes, unconditionally stable difference schemes for the approximate solution of problem (7.237)–(7.239). Starting from the explicit scheme (7.240), we write the regularized scheme in the canonical form (7.241) with

$$B = 0, \qquad R = \frac{1}{\tau^2}(E+\alpha G), \qquad A = -\Lambda. \tag{7.248}$$

Theorem 7.26 *The regularized scheme* (7.241), (7.248) *is ρ-stable with the regulizer $G = \Lambda$ with*

$$\rho = \exp\Big(\frac{\tau}{\sqrt{\alpha}}\Big), \tag{7.249}$$

and with the regulizer $G = \Lambda^2$, with

$$\rho = \exp\Big(\frac{\tau}{\sqrt{2\sqrt{\alpha}}}\Big). \tag{7.250}$$

Proof. Starting from (7.245) for (7.248), we arrive at the inequality

$$(\rho-1)^2(E+\alpha G)-\tau^2\rho\Lambda \geq 0. \tag{7.251}$$

In the case of $G = \Lambda$, analogously to the proof of Theorem 7.24 ($\chi = \alpha + M^{-1}$) we obtain the following expression for ρ:

$$\rho = \exp\big((\alpha + \|\Lambda\|^{-1})^{-1/2}\tau\big).$$

Making ρ cruder yields the estimate (7.249).

In the case of $G = \Lambda^2$, from inequality (7.245) we obtain:

$$E+\alpha\Lambda^2-\frac{\tau^2\rho}{(\rho-1)^2}\Lambda = \Big(\sqrt{\alpha}\Lambda-\frac{\tau^2\rho}{2\sqrt{\alpha}(\rho-1)^2}E\Big)^2+\Big(1-\frac{\tau^4\rho^2}{4\alpha(\rho-1)^4}\Big)E.$$

This inequality is fulfilled for some given ρ if

$$\alpha \geq \frac{\tau^4 \rho^2}{4(\rho-1)^4}. \tag{7.252}$$

Let us evaluate now the quantity ρ for given α from inequality (7.252) rewritten in the form

$$(\rho-1)^2 2\sqrt{\alpha} - \tau^2 \rho.$$

By the above lemma with $\chi = 2\sqrt{\alpha}$, the latter inequality is fulfilled for the values of ρ defined by (7.250). □

The regularized scheme (7.241), (7.248) with the regularizer $G = \Lambda$ can be written as the weighted scheme

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} - \Lambda(\sigma y_{n+1} + (1-2\sigma)y_n + \sigma y_{n-1}) = 0 \tag{7.253}$$

with $\sigma = -\alpha/\tau^2$. Thereby, the regularizing parameter here is the negative weight in (7.253). The latter scheme can also be related to a version of the generalized inverse method (7.225) for the approximate solution of the ill-posed problem (7.187)–(7.189).

The construction of the regularized difference scheme is based on a perturbation of the operators in the generic difference scheme (7.241), (7.242) chosen so that to fulfill the operator inequality (7.245). In Theorem 7.26, the latter can be achieved at the expense of some additive perturbation (increase) of R. There are many other possibilities. In particular, note the possibility of additive perturbation of B:

$$B = \alpha G, \qquad R = \frac{1}{\tau^2} E, \qquad A = -\Lambda. \tag{7.254}$$

Theorem 7.27 *The regularized scheme* (7.241), (7.254) *with* $G = \Lambda$ *is* ρ*-stable with*

$$\rho = \exp\left(\frac{\tau}{\alpha}\right). \tag{7.255}$$

Proof. Inequality (7.245) can be rearranged as

$$\frac{\rho^2-1}{2\tau} B + (\rho-1)^2 R + \rho A \; > \; \frac{\rho^2-1}{2\tau} \alpha\Lambda - \rho\Lambda \; \geq \; 0.$$

This inequality is fulfilled with $\rho \geq \rho_2$, where

$$\rho_2 = \frac{\tau}{\alpha} + \left(1 + \frac{\tau^2}{\alpha^2}\right)^{1/2} \; < \; 1 + \frac{\tau}{\alpha} + \frac{1}{2}\frac{\tau^2}{\alpha^2} \; < \; \exp\left(\frac{\tau}{\alpha}\right).$$

From here, expression (7.255) for ρ in the difference scheme (7.241), (7.254) follows. □

The regularized scheme (7.241), (7.254) can be directly related to the use of the generalized inverse method in its version (7.234). In a similar way, regularized schemes can be constructed which can be related to the basic variant (7.225) of the generalized inverse method.

7.4.6 Program

Here, we do not has as our object to check the efficiency of all mentioned methods for the approximate solution of the model Cauchy problem for elliptic equation (7.180)–(7.182), (7.184). As judged from the standpoint of computational realization, the simplest approach here is related with using regularized schemes of type (7.241), (7.248) (or (7.241), (7.254)) with the regulizer $G = \Lambda$.

In the case of (7.241), (7.248), the approximate solution is to be found from the difference equation

$$(E + \alpha\Lambda)\,\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} - \Lambda y_n = 0,$$

and in the case of (7.241), (7.254), from

$$\alpha\Lambda\,\frac{y_{n+1} - y_{n-1}}{2\tau} + \frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} - \Lambda y_n = 0.$$

For such schemes, the computational realization is not much more difficult than for direct problems.

Here, to be controlled (bounded) is the growth of the solution norm, this very often being not sufficient for obtaining a satisfactory approximate solution. The latter circumstance is related with the fact that, here, we do not use any preliminary treatment of the approximate solution burdened with input-data inaccuracy. That is why we have to use regularized difference schemes with stronger regulizers.

The program `PROBLEM13` realizes the regularized difference scheme (7.241), (7.248) with $G = \Lambda^2$:

$$(E + \alpha\Lambda^2)\,\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} - \Lambda y_n = 0. \tag{7.256}$$

For the model problem (7.180)–(7.182), (7.184), the realization of (7.256) is based on using the five-point sweep algorithm.

Note some possibilities available in choosing the regularization parameter. In the most natural approach, the regularization parameter is to be chosen considering the discrepancy; here, we compare, at $t = 0$, the solutions of the direct problem of type (7.180)–(7.183), in which the boundary condition (7.183) is formulated from the solution of the inverse problem. Here, two circumstances are to be mentioned, which make this approach very natural as used with regularized difference schemes of type (7.256). First, the used algorithm becomes a global regularization algorithm (we have to solve the problem for all times t at ones). Second, here the computational realization of the direct problem, i.e., the boundary value problem for the elliptic equation, is much more difficult than for the inverse problem, a problem for the second-order evolution equation.

With the aforesaid, for choosing the value of the regularization parameter we have to apply such algorithms that retain the local regularization property (making a consecutive determination of the approximate solution at the next time layer possible). A simplest such algorithm is realized in the program PROBLEM13.

Previously (see (7.140)–(7.143)), the relation between the regularized schemes and the difference-smoothing algorithms was elucidated. We rewrite scheme (7.256) as

$$\tilde{w}_n - \Lambda y_n = 0,$$

$$(E + \alpha\Lambda^2) w_n = \tilde{w}_{n+1},$$

where

$$w_n = \frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2}.$$

In this way, the second difference derivative is first to be calculated by an explicit formula and, then, to be smoothed:

$$\begin{aligned} J_\alpha(w_n) &= \min_{v \in H} J_\alpha(v), \\ & J_\alpha(v) = \|v - \tilde{w}_n\|^2 + \alpha \|\Lambda v\|^2. \end{aligned}$$

In the case under consideration, the regularization is related with smoothing of mesh functions.

First of all, the smoothing procedure has to be applied to the inaccurate input data. Here, these data are the mesh function y_0. In accordance with the discrepancy principle, the regularization (smoothing) parameter can be found from the condition

$$\|y_0 - u_0^\delta\| = \delta,$$

with

$$\begin{aligned} J_\alpha(y_0) &= \min_{v \in H} J_\alpha(v), \\ & J_\alpha(v) = \|v - u_0^\delta\|^2 + \alpha \|\Lambda v\|^2. \end{aligned}$$

Program PROBLEM13

```
C
C     PROBLEM13 - CAUCHY PROBLEM FOR THE LAPLACE EQUATION
C                 TWO-DIMENSIONAL PROBLEM
C                 REGULARIZED SCHEME
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.005D0, N = 65, M = 65 )
      DIMENSION U0(N), U0D(N), UT(N), U(N), U1(N), Y(N), Y1(N)
     +         ,X(N), A(N), B(N), C(N), D(N), E(N), F(N)
C
C     PARAMETERS:
C
```

```
C     XL, XR    - LEFT AND RIGHT ENDS OF THE SEGMENT;
C     N         - NUMBER OF NODES IN THE SPATIAL GRID;
C     TMAX      - MAXIMAL TIME;
C     M         - NUMBER OF NODES OVER TIME;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     Q         - FACTOR IN THE FORMULA FOR THE REGULARIZATION PARAMETER;
C     U0(N)     - INITIAL CONDITION;
C     U0D(N)    - DISTURBED INITIAL CONDITION;
C     UT(N)     - EXACT END-TIME SOLUTION;
C     U(N)      - APPROXIMATE SOLUTION OF THE INVERSE PROBLEM;
C
      XL   = 0.D0
      XR   = 1.D0
      TMAX = 0.25D0
C
      OPEN (01, FILE = 'RESULT.DAT')! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H = (XR - XL) / (N - 1)
      TAU = TMAX / (M-1)
      DO I = 1, N
        X(I) = XL + (I-1)*H
      END DO
C
C     EXACT SOLUTION OF THE PROBLEM
C
      DO I = 1, N
        U0(I) = AU(X(I), 0.D0)
        UT(I) = AU(X(I), TMAX)
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO I = 2, N-1
        U0D(I) = U0(I) + 2.*DELTA*(RAND(0)-0.5)
      END DO
      U0D(1) = U0(1)
      U0D(N) = U0(N)
C
C     INVERSE PROBLEM
C
C     SMOOTHING OF INITIAL CONDITIONS
C
      IT = 0
      ITMAX = 100
      ALPHA = 0.00001D0
      Q = 0.75D0
  100 IT = IT + 1
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE INVERSE PROBLEM
C
      DO I = 2, N-1
        A(I) = ALPHA / (H**4)
        B(I) = 4.D0 * ALPHA / (H**4)
        C(I) = 6.D0 * ALPHA / (H**4) + 1.D0
        D(I) = 4.D0 * ALPHA / (H**4)
        E(I) = ALPHA / (H**4)
      END DO
      C(1)   = 1.D0
```

```
      D(1)   = 0.D0
      E(1)   = 0.D0
      F(1)   = 0.D0
      B(2)   = 0.D0
      C(2)   = 5.D0 * ALPHA / (H**4) + 1.D0
      C(N-1) = 5.D0 * ALPHA / (H**4) + 1.D0
      D(N-1) = 0.D0
      A(N)   = 0.D0
      B(N)   = 0.D0
      C(N)   = 1.D0
      F(N)   = 0.D0
      DO I = 2, N-1
        F(I) = U0D(I)
      END DO
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
      ITASK = 1
      CALL PROG5 ( N, A, B, C, D, E, F, U1, ITASK )
C
C     CRITERION FOR THE EXIT FROM THE ITERATIVE PROCESS
C
      WRITE (01,*) IT, ALPHA
      SUM = 0.D0
      DO I = 2, N-1
        SUM = SUM + (U1(I) - U0D(I))**2*H
      END DO
      SL2 = DSQRT(SUM)
C
      IF (IT.GT.ITMAX) STOP
      IF ( IT.EQ.1 ) THEN
        IND = 0
        IF ( SL2.LT.DELTA ) THEN
          IND = 1
          Q = 1.D0/Q
        END IF
        ALPHA = ALPHA*Q
        GO TO 100
      ELSE
        ALPHA = ALPHA*Q
        IF ( IND.EQ.0 .AND. SL2.GT.DELTA ) GO TO 100
        IF ( IND.EQ.1 .AND. SL2.LT.DELTA ) GO TO 100
      END IF
C
C     INVERSE-PROBLEM SOLUTION WITH CHOSEN REGULARIZATION PARAMETER
C     REGULARIZED SCHEME
C
C     INITIAL CONDITION
C
      DO I = 1, N
        Y1(I)  = U1(I)
      END DO
      DO I = 2, N-1
        Y(I) = Y1(I)
     +       + 0.5D0*TAU**2/H**2 * (Y1(I+1)-2.D0*Y1(I) + Y1(I-1))
      END DO
      Y(1) = 0.D0
      Y(N) = 0.D0
      DO K = 3, M
C
```

```
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE INVERSE PROBLEM
C
        DO I = 2, N-1
          A(I) = ALPHA / (H**4)
          B(I) = 4.D0 * ALPHA / (H**4)
          C(I) = 6.D0 * ALPHA / (H**4) + 1.D0
          D(I) = 4.D0 * ALPHA / (H**4)
          E(I) = ALPHA / (H**4)
        END DO
        C(1)   = 1.D0
        D(1)   = 0.D0
        E(1)   = 0.D0
        F(1)   = 0.D0
        B(2)   = 0.D0
        C(2)   = 5.D0 * ALPHA / (H**4) + 1.D0
        C(N-1) = 5.D0 * ALPHA / (H**4) + 1.D0
        D(N-1) = 0.D0
        A(N)   = 0.D0
        B(N)   = 0.D0
        C(N)   = 1.D0
        F(N)   = 0.D0
        DO I = 3, N-2
          F(I) = A(I)*(2.D0*Y(I-2) - Y1(I-2))
     +         - B(I)*(2.D0*Y(I-1) - Y1(I-1))
     +         + C(I)*(2.D0*Y(I) - Y1(I))
     +         - D(I)*(2.D0*Y(I+1) - Y1(I+1))
     +         + E(I)*(2.D0*Y(I+2) - Y1(I+2))
     +         - TAU**2/(H*H) * (Y(I+1)-2.D0*Y(I) + Y(I-1))
        END DO
        F(2) = - B(2)*(2.D0*Y(1) - Y1(1))
     +         + C(2)*(2.D0*Y(2) - Y1(2))

     +         - D(2)*(2.D0*Y(3) - Y1(3))
     +         + E(2)*(2.D0*Y(4) - Y1(4))
     +         - TAU**2/(H*H) * (Y(3)-2.D0*Y(2) + Y(1))
        F(N-1) = A(N-1)*(2.D0*Y(N-3) - Y1(N-3))
     +         - B(N-1)*(2.D0*Y(N-2) - Y1(N-2))
     +         + C(N-1)*(2.D0*Y(N-1) - Y1(N-1))
     +         - D(N-1)*(2.D0*Y(N) - Y1(N))
     +         - TAU**2/(H*H) * (Y(N)-2.D0*Y(N-1) + Y(N-2))
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        DO I = 1, N
          Y1(I) = Y(I)
        END DO
        ITASK = 1
        CALL PROG5 ( N, A, B, C, D, E, F, Y, ITASK )
      END DO
      DO I = 1, N
        U(I) = Y(I)
      END DO
C
C     SOLUTION
C
      WRITE ( 01, * ) (U0(I),I=1,N)
      WRITE ( 01, * ) (UT(I),I=1,N)
      WRITE ( 01, * ) (U0D(I),I=1,N)
      WRITE ( 01, * ) (X(I),I=1,N)
      WRITE ( 01, * ) (U(I),I=1,N)
```

```
      CLOSE (01)
      STOP
      END

      DOUBLE PRECISION FUNCTION  AU ( X, T )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      EXACT SOLUTION
C
      PI = 3.1415926D0
      C1 = 0.1D0
      C2 = 0.5D0 * C1
      AU = C1 * (DEXP(PI*T) + DEXP(-PI*T)) * DSIN(PI*X)
     +   + C2 * (DEXP(2*PI*T) + DEXP(-2*PI*T)) * DSIN(2*PI*X)
C
      RETURN
      END
```

7.4.7 Computational experiments

The data presented below were obtained on a uniform grid with $h = 1/64$ and $l = 1$. The inverse problem was solved till the time $T = 0.25$, the time step size of the grid being $\tau = T/64$. The exact solution of the inverse problem is

$$u(x,t) = \operatorname{ch}(\pi t)\sin(\pi x) + \frac{1}{2}\operatorname{ch}(2\pi t)\sin(2\pi x).$$

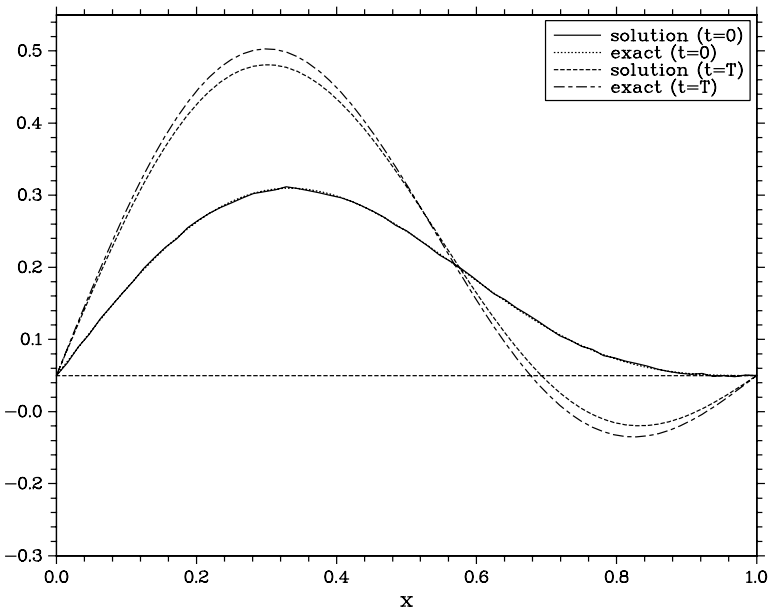

Figure 7.14 Solution of the problem obtained with $\delta = 0.002$

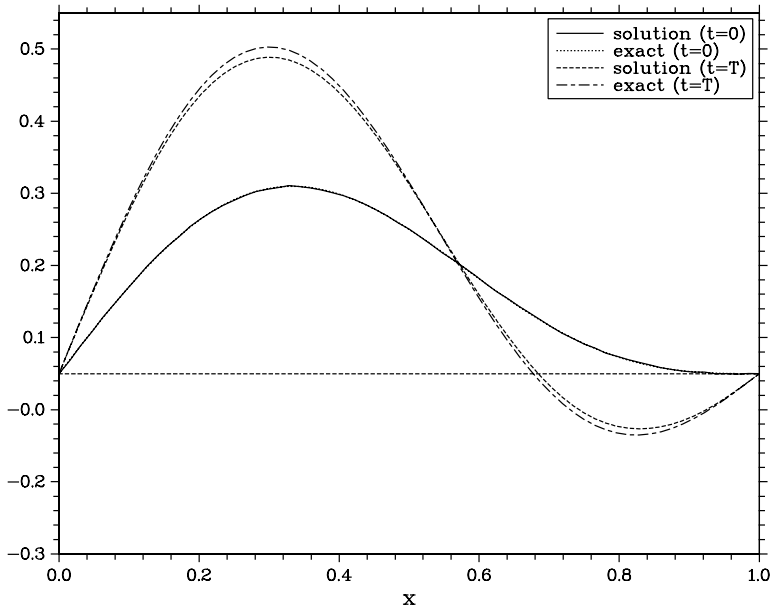

Figure 7.15 Solution of the problem obtained with $\delta = 0.001$

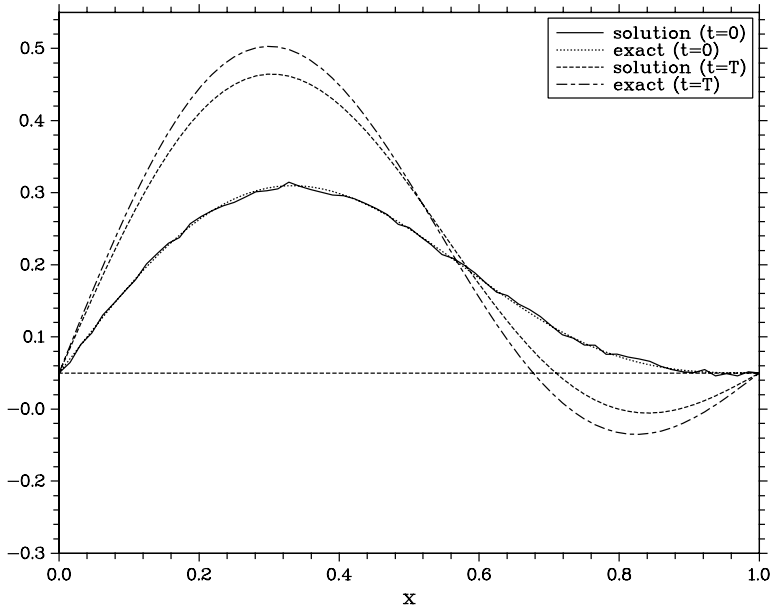

Figure 7.16 Inverse-problem solution obtained with $\delta = 0.005$

Figure 7.14 shows the solution of the inverse problem obtained for the inaccuracy level defined by the quantity $\delta = 0.002$. Here, the exact and perturbed initial condi-

tions are shown, as well as the exact and approximate end-time solution of the inverse problem. Solutions of the problem obtained for two other inaccuracy levels are shown in Figures 7.15 and 7.16. The calculation data prove it possible to reconstruct smooth solutions at inaccuracy levels amounting to one percent.

7.5 Continuation of non-stationary fields from point observation data

In this section, we consider the inverse problem for a model non-stationary parabolic equation with unknown initial condition and with information about the solution available at some points of the two-dimensional calculation domain. We describe a computational algorithm developed around the variational formulation of the problem and using the Tikhonov regularization algorithm.

7.5.1 Statement of the problem

We consider a problem in which it is required to determine a non-stationary field $u(\boldsymbol{x}, t)$ that satisfies a second-order parabolic equation in a bounded two-dimensional domain Ω and certain boundary conditions provided that information about the solution at some points in the domain is available. The initial state $u(\boldsymbol{x}, 0)$ is assumed unknown. This problem is an ill-posed one; in particular, we cannot rely on obtaining a unique solution. Such inverse problems often arise in hydrogeology, for instance, in the cases in which all available information about the solution is given by variation of physical quantities in observation holes.

In a two-dimensional bounded domain Ω ($\boldsymbol{x} = (x_1, x_2)$), we seek the solution of the parabolic equation

$$\frac{\partial u}{\partial t} - \sum_{\beta=1}^{2} \frac{\partial}{\partial x_\beta}\Big(k(\boldsymbol{x}) \frac{\partial u}{\partial x_\beta}\Big) = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T. \tag{7.257}$$

This equation is supplemented with first-kind homogeneous boundary conditions:

$$u(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \partial\Omega, \quad 0 < t < T. \tag{7.258}$$

For a well-posed problem to be formulated, we have to set the initial state, the function $u(\boldsymbol{x}, 0)$.

In the inverse problem of interest, the initial condition is unknown. The additional information is gained in observations performed over the solution at some individual points in the calculation domain; this information is provided by functions $u(\boldsymbol{x}, t)$ given at points $\boldsymbol{z}_m \in \Omega$, $m = 1, 2, \ldots, M$ (see Figure 7.17). With regard to the measurement inaccuracies, we put:

$$u(\boldsymbol{z}_m, t) \approx \varphi_m(t), \qquad 0 < t < T, \quad m = 1, 2, \ldots, M. \tag{7.259}$$

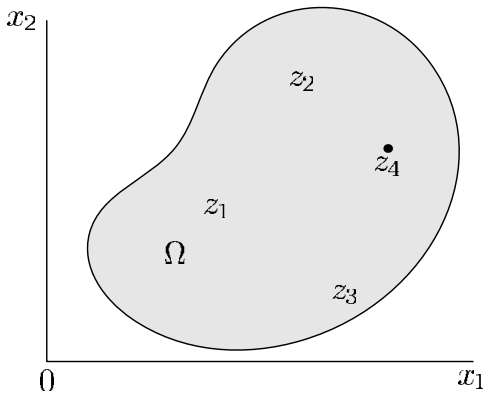

Figure 7.17 Schematic illustrating the statement of the problem

It is required to find, from equation (7.257), boundary conditions (7.258) and additional measurements (7.259), the function $u(\boldsymbol{x}, t)$.

With the given data measured at individual points (see (7.259)), we can pose a problem aimed at treating the data. At each time, we can try to interpolate (extrapolate) these data towards all points in the calculation domain. From this standpoint, the above-posed inverse problem (7.257)–(7.259) can be considered as a problem in which it is required to treat the data maximally taking into account a priori information about the solution; here, the interpolation problem is to be solved in the class of functions that satisfy equation (7.257) and boundary conditions (7.258).

7.5.2 Variational problem

In the consideration of inverse problem (7.257)–(7.259), we will restrict ourselves to the cases in which it is required to solve such problems approximately. We will solve these problems on the basis of the Tikhonov regularization method. To this end, we will use the formulation of the inverse problem as an optimal control problem.

As the control $v(\boldsymbol{x})$, it seems reasonable to choose the initial condition. We denote the corresponding solution as $u(\boldsymbol{x}, t; v)$. The state $u(\boldsymbol{x}, t; v)$ of the system is pre-defined by equation (7.257), by boundary conditions (7.258), and by the initial condition

$$u(\boldsymbol{x}, 0; v) = v(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega. \tag{7.260}$$

We assume that the control $v(\boldsymbol{x})$ belongs to the Hilbert space $\mathcal{H} = \{v(\boldsymbol{x}) \mid v(\boldsymbol{x}) \in \mathcal{L}_2(\Omega),\ v(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \partial\Omega\}$, in which the scalar product and the norm are defined as

$$(v, w) = \int_\Omega v(\boldsymbol{x}) w(\boldsymbol{x})\, d\boldsymbol{x}, \qquad \|v\| = \sqrt{(v, v)}.$$

In line with (7.259), we choose the smoothing functional in the form

$$J_\alpha(v) = \sum_{m=1}^{M} \int_0^T (u(z_m, t; v) - \varphi_m(t))^2 \, dt + \alpha \|v\|^2, \tag{7.261}$$

where $\alpha > 0$ is the regularization parameter, whose value must be chosen considering the inaccuracy in the measurements (7.259).

The optimal control $w(x)$ is to be chosen from the minimum of (7.261), i.e.,

$$J_\alpha(w) = \min_{v \in \mathcal{H}} J_\alpha(v). \tag{7.262}$$

The solution of the inverse problem is $u(x, t) = u(x, t; w)$.

For the variational problem (7.257), (7.258), (7.260)–(7.262), we use a somewhat more general differential-operator statement, considering the problem in $\mathcal{H}$, so that $u(x, t; v) = u(t; v) \in \mathcal{H}$. We write equation (7.257) with boundary conditions (7.258) in $\mathcal{H}$ in the form of an evolutionary first-order equation:

$$\frac{du}{dt} + \mathcal{A}u = 0, \qquad 0 < t < T. \tag{7.263}$$

The operator $\mathcal{A}$, defined as

$$\mathcal{A}u \equiv -\sum_{\beta=1}^{2} \frac{\partial}{\partial x_\beta}\Big(k(x) \frac{\partial u}{\partial x_\beta}\Big),$$

is self-adjoint and positively defined in $\mathcal{H}$:

$$\mathcal{A} = \mathcal{A}^* \geq \kappa \lambda_0 E, \tag{7.264}$$

where $k(x) \geq \kappa > 0$, and $\lambda_0 > 0$ is the minimal eigenvalue of the Laplace operator.

We introduce the function $\chi \in \mathcal{H}$:

$$\chi = \sum_{m=1}^{M} \delta(x - z_m).$$

Here, $\delta(x)$ is the δ-function, and let $\varphi(t) \in \mathcal{H}$ be a function such that

$$\chi \varphi(t) = \sum_{m=1}^{M} \delta(x - z_m) \varphi_m(t).$$

With the notation introduced, the functional (7.261) can be rewritten as

$$J_\alpha(v) = \int_0^T (\chi, (u(t; v) - \varphi(t))^2) \, dt + \alpha \|v\|^2. \tag{7.265}$$

Equation (7.263) is supplemented with the initial condition (see (7.260))

$$u(0) = v. \tag{7.266}$$

So, we arrive at the minimization problem (7.262), (7.265) for the solutions of problem (7.263), (7.266).

7.5.3 Difference problem

For simplicity, we restrict ourselves to the case in which Ω is a rectangle:

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\beta < l_\beta,\ \beta = 1, 2\}.$$

In the domain Ω, we introduce a grid, uniform over either direction, with grid sizes h_α, $\alpha = 1, 2$. Suppose that, as usually, ω is the set of internal nodes.

On the set of mesh functions $y(\boldsymbol{x})$ such that $y(\boldsymbol{x}) = 0$, $\boldsymbol{x} \neq \omega$, by the relation

$$\Lambda y = -\sum_{\beta=1}^{2} (a_\beta y_{\bar{x}_\beta})_{x_\beta}, \tag{7.267}$$

we define the difference operator Λ, where, for instance,

$$a_1(\boldsymbol{x}) = k(x_1 - 0.5h_1, x_2), \qquad a_2(\boldsymbol{x}) = k(x_1, x_2 - 0.5h_2).$$

In the mesh Hilbert space $H = L_2(\omega)$, we introduce the scalar product and the norm with the relations

$$(y, w) = \sum_{\boldsymbol{x}\in\omega} y(\boldsymbol{x}) w(\boldsymbol{x}) h_1 h_2, \qquad \|y\| = \sqrt{(y, y)}.$$

In H we have $\Lambda = \Lambda^* \geq \gamma E$, $m > 0$, where

$$\gamma = \kappa\Big(\frac{8}{l_1^2} + \frac{8}{l_2^2}\Big).$$

From (7.263) and (7.266), we pass to the differential-operator equation

$$\frac{dy}{dt} + \Lambda y = 0, \qquad 0 < t < T \tag{7.268}$$

with the given initial condition

$$y(0) = v, \qquad \boldsymbol{x} \in \omega. \tag{7.269}$$

We denote by y_n the difference solution at the time $t_n = n\tau$, $n = 0, 1, \ldots, N_0$, $N_0\tau = T$, where $\tau > 0$ is the time step size. For problem (7.268), (7.269), we will use the two-layer scheme with the weights

$$\begin{gathered}\frac{y_{n+1} - y_n}{\tau} + \Lambda(\sigma y_{n+1} + (1 - \sigma) y_n) = 0, \\ n = 0, 1, \ldots, N_0 - 1,\end{gathered} \tag{7.270}$$

supplemented with the initial condition

$$y_0(\boldsymbol{x}) = v(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega. \tag{7.271}$$

We write scheme (7.270) in the canonical form

$$B\,\frac{y_{n+1}-y_n}{\tau}+Ay_n=0,\qquad n=0,1,\dots,N_0-1. \tag{7.272}$$

For the difference operators B and A we have:

$$B=E+\sigma\tau A,\qquad A=\Lambda. \tag{7.273}$$

The scheme with weights (7.272) and (7.273) in the case of $A=A^*>0$ is stable (see Theorem 4.14) in H provided that

$$E+\tau\Big(\sigma-\frac{1}{2}\Big)A\geq 0.$$

This (necessary and sufficient) condition is fulfilled for all $\sigma\geq 0.5$, i.e., here we have an unconditionally stable difference scheme. In the case of $\sigma<0.5$ scheme (7.272), (7.273) is conditionally stable. For the difference solution of problem (7.270), (7.271) in the case of $\sigma\geq 0.5$ there holds the following a priori estimate:

$$\|y_n\|\leq\|v\|,\qquad n=1,2,\dots,N_0;$$

this estimate shows the scheme to be stable with respect to initial data.

We assume that the observation points $\boldsymbol{z}_m$, $m=1,2,\dots,M$ coincide with some internal nodes of the calculation grid. Like in the continuous case, we define the mesh function $\chi_h\in H$ as

$$\chi_h(\boldsymbol{x})=\frac{1}{h_1h_2}\sum_{m=1}^{M}\int_\Omega\delta(\boldsymbol{x}-\boldsymbol{z}_m)\,d\boldsymbol{x},$$

i.e., $\chi_h\in H$ is the sum of the corresponding difference δ-functions. In a similar way, we can introduce $\varphi_h(\boldsymbol{x},t_n)$, $\boldsymbol{x}\in\omega$, $n=1,2,\dots,N_0$ so that

$$\chi_h(\boldsymbol{x})\varphi_h(\boldsymbol{x},t_n)=\frac{1}{h_1h_2}\sum_{m=1}^{M}\int_\Omega\delta(\boldsymbol{x}-\boldsymbol{z}_m)\varphi_m(t_n)\,d\boldsymbol{x}.$$

To functional (7.265), we put into correspondence the following difference functional:

$$J_\alpha(v)=\sum_{n=1}^{N_0}(\chi_h,(y(\boldsymbol{x},t_n;v)-\varphi_h(\boldsymbol{x},t_n))^2)\tau+\alpha\|v\|^2. \tag{7.274}$$

The minimization problem

$$J_\alpha(w)=\min_{v\in H}J_\alpha(v) \tag{7.275}$$

is to be solved under constraints (7.270) and (7.271).

7.5.4 Numerical solution of the difference problem

To approximately solve the discrete variational problem of interest, we will use gradient iteration methods. First, we are goint to derive the Euler equation (optimality condition) for the variational problem to be approximately solved by the iteration methods.

To formulate the optimality conditions for problem (7.271), (7.272), (7.274), (7.275), consider the problem for the increments. From (7.271) and (7.272) we have:

$$B\,\frac{\delta y_{n+1}-\delta y_n}{\tau}+A\delta y_n=0,\qquad n=0,1,\ldots,N_0-1, \tag{7.276}$$

$$\delta y_0=\delta v. \tag{7.277}$$

To formulate the problem for the conjugate state ψ_n, we multiply equation (7.276) (scalarwise in H) by $\tau\psi_{n+1}$ and calculate the sum over n from 0 to N_0-1. This yields

$$\sum_{n=0}^{N_0-1}((B\delta y_{n+1},\psi_{n+1})+((\tau A-B)\delta y_n,\psi_{n+1}))=0. \tag{7.278}$$

With the constancy and self-adjointness of the operators A and B taken into account, for the first term we have:

$$\sum_{n=0}^{N_0-1}(B\delta y_{n+1},\psi_{n+1})=\sum_{n=1}^{N_0}(\delta y_n,B\psi_n).$$

We consider the mesh function ψ_n for $n=0,1,\ldots,N_0$ under the following additional conditions:

$$\psi_{N_0+1}=0. \tag{7.279}$$

The second term in (7.278) can be rearranged as

$$\begin{aligned}\sum_{n=0}^{N_0-1}((\tau A-B)\delta y_n,\psi_{n+1})&=\sum_{n=0}^{N_0}((\tau A-B)\delta y_n,\psi_{n+1})\\&=\sum_{n=1}^{N_0}(\delta y_n,(\tau A-B)\psi_{n+1})+(\delta y_0,(\tau A-B)\psi_1).\end{aligned} \tag{7.280}$$

Substitution of (7.279) and (7.280) into (7.278) yields the equation

$$\sum_{n=0}^{N_0}(\delta y_n,(B\psi_n+(\tau A-B)\psi_{n+1}))=(\delta v,B\psi_0). \tag{7.281}$$

With the form of (7.274) taken into account, we determine the conjugate state from the difference equation

$$B\,\frac{\psi_n - \psi_{n+1}}{\tau} + A\psi_{n+1} = \chi_h(y_n - \varphi_h(\boldsymbol{x}, t_n)), \quad n = N_0, N_0 - 1, \ldots, 0, \tag{7.282}$$

supplemented with conditions (7.279). Scheme (7.279), (7.282) is stable under the same conditions as scheme (7.271), (7.272) for the ground state.

From (7.281), for the gradient of (7.274) the following representation can be derived:

$$J'_\alpha(v) = 2\Big(\frac{1}{\tau}\,B\psi_0 + \alpha v\Big). \tag{7.283}$$

With (7.283), the necessary and sufficient condition for the minimum of (7.274) can be written as

$$B\psi_0 + \tau\alpha v = 0. \tag{7.284}$$

For equation (7.284) to be solved, we construct the iteration method.

The iteration algorithm for initial-state correction consists in the following organization of the computational procedure.

- At a given w^k (k is the iteration number), we solve the identification problem for the ground state:

$$\begin{aligned} B\,\frac{y^k_{n+1} - y^k_n}{\tau} + Ay^k_n &= 0, \qquad n = 0, 1, \ldots, N_0 - 1, \\ y^k_0(\boldsymbol{x}) &= w^k(\boldsymbol{x}), \qquad \boldsymbol{x} \in \omega. \end{aligned}$$

- Then, we calculate the conjugate state:

$$\begin{aligned} B\,\frac{\psi^k_n - \psi^k_{n+1}}{\tau} + A\psi^k_{n+1} &= 2\chi_h(y^k_n - \varphi_h(\boldsymbol{x}, t_n)), \quad n = N_0, N_0 - 1, \ldots, 1, \\ B\,\frac{\psi^k_0 - \psi^k_1}{\tau} + A\psi^k_1 &= 0, \\ \psi^k_{N_0+1} &= 0, \qquad \boldsymbol{x} \in \omega. \end{aligned}$$

- Next, we refine the initial condition:

$$\frac{w^{k+1} - w^k}{s^{k+1}} + B\psi^k_0 + \alpha w^k = 0, \qquad \boldsymbol{x} \in \omega.$$

Thus, the algorithm is based on the solution of two non-stationary difference problems and on refinement of the initial condition at each iteration step.

A second, more attractive possibility is related with the construction of the iteration method of minimized discrepancy functional (with $\alpha = 0$ in (7.274)). The latter situation arises when the iteration method is used to solve the equation (see (7.284))

$$B\psi_0 = 0. \tag{7.285}$$

In the case of (7.285), we use the same computational procedure as in the solution of (7.284). Yet, there are no problems with the special definition of the regularization parameter, this circumstance being the major advantage of iteration methods over the Tikhonov regularization method.

7.5.5 Program

In the program listing presented below, the iteration method for the minimization of the discrepancy functional is realized (equation (7.285) is solved). In order to make the problem not too complicated, we have restricted ourselves to the iterative method of simple iteration, in which

$$\frac{w^{k+1} - w^k}{s} + B\psi_0^k = 0, \qquad x \in \omega$$

and the value of s is set explicitly.

Program PROBLEM14

```
C
C     PROBLEM14 - CONTINUATION OF NON-STATIONARY FIELDS
C                 FROM POINT OBSERVATIONS
C                 TWO-DIMENSIONAL PROBLEM
C                 ITERATIVE REFINEMENT OF THE INITIAL CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.02D0, N1 = 51, N2 = 51, M = 101, L = 10 )
      DIMENSION A(13*N1*N2), X1(N1), X2(N2)
     +         ,XP(L), YP(L), IM(L), JM(L)
     +         ,FF(L), FI(L,M), FID(L,M), FIK(L,M)
      COMMON / SB5 /     IDEFAULT(4)
      COMMON / CONTROL / IREPT, NITER
C
C     PARAMETERS:
C
C     X1L, X2L  - COORDINATES OF THE LEFT CORNER;
C     X1R, X2R  - COORDINATES OF THE RIGHT CIRNER;
C     N1, N2    - NUMBER OF NODES IN THE SPATIAL GRID;
C     H1, H2    - MESH SIZES OVER SPACE;
C     TAU       - TIME STEP;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     U0(N1,N2) - INITIAL CONDITION TO BE RECONSTRUCTED;
C     XP(L),    - COORDINATES OF THE OBSERVATION POINTS;
C       YP(L)
C     FI(L,M)   - SOLUTION AT THE OBSERVATION POINTS;
C     FID(L,M)  - DISTURBED SOLUTION AT THE OBSERVATION POINTS;
```

```
C
C     EPSR      - RELATIVE INACCURACY OF THE DIFFERENCE SOLUTION;
C     EPSA      - ABSOLUTE INACCURACY OF THE DIFFERENCE SOLUTION;
C
C     EQUIVALENCE ( A(1),         A0            ),
C    *            ( A(N+1),       A1            ),
C    *            ( A(2*N+1),     A2            ),
C    *            ( A(9*N+1),     F             ),
C    *            ( A(10*N+1),    U0            ),
C    *            ( A(11*N+1),    V             ),
C    *            ( A(12*N+1),    B             ),
C
      X1L    = 0.D0
      X1R    = 1.D0
      X2L    = 0.D0
      X2R    = 1.D0
      TMAX   = 0.025D0
      PI     = 3.1415926D0
      EPSR   = 1.D-5
      EPSA   = 1.D-8
      SS     = - 2.D0
C
      OPEN (01, FILE = 'RESULT.DAT')! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H1  = (X1R-X1L) / (N1-1)
      H2  = (X2R-X2L) / (N2-1)
      TAU = TMAX / (M-1)
      DO I = 1, N1
        X1(I) = X1L + (I-1)*H1
      END DO
      DO J = 1, N2
        X2(J) = X2L + (J-1)*H2
      END DO
C
      N  = N1*N2
      DO I = 1, 13*N
        A(I) = 0.0
      END DO
C
C     DIRECT PROBLEM
C     PURELY IMPLICIT DIFFERENCE SCHEME
C
C     INITIAL CONDITION
C
      T  = 0.D0
      CALL MEGP (XP, YP, IM, JM, L, H1, H2)
      CALL INIT (A(10*N+1), X1, X2, N1, N2)
      CALL PU (IM, JM, A(10*N+1), N1, N2, FF, L)
      DO IP = 1,L
        FI(IP,1) = FF(IP)
      END DO
      DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        CALL FDS (A(1), A(N+1), A(2*N+1), A(9*N+1), A(10*N+1),
     +            H1, H2, N1, N2, TAU)
C
```

```
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
C
C     OBSERVATIONAL DATA
C
        CALL PU (IM, JM, A(10*N+1), N1, N2, FF, L)
        DO IP = 1,L
          FI(IP,K) = FF(IP)
        END DO
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO K = 1,M
        DO IP = 1,L
          FID(IP,K) = FI(IP,K) + 2.*DELTA*(RAND(0)-0.5)
        END DO
      END DO
C
C     INVERSE PROBLEM
C     ITERATION METHOD
C
      IT = 0
C
C     STARTING APPROXIMATION
C
      DO I = 1, N
        A(11*N+I) = 0.D0
      END DO
C
  100 IT = IT + 1
C
C     GROUND STATE
C
      T  = 0.D0
C
C     INITIAL CONDITION
C
      DO I = 1, N
        A(10*N+I) = A(11*N+I)
      END DO
      CALL PU (IM, JM, A(10*N+1), N1, N2, FF, L)
      DO IP = 1,L
        FIK(IP,1) = FF(IP)
      END DO
      DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        CALL FDS (A(1), A(N+1), A(2*N+1), A(9*N+1), A(10*N+1),
     +            H1, H2, N1, N2, TAU)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
```

```
C
C     SOLUTION AT THE OBSERVATION POINTS
C
        CALL PU (IM, JM, A(10*N+1), N1, N2, FF, L)
        DO IP = 1,L
          FIK(IP,K) = FF(IP)
        END DO
      END DO
C
C     CONJUGATE STATE
C
      T  = TMAX + TAU
C
C     INITIAL CONDITION
C
      DO I = 1, N
        A(10*N+I) = 0.D0
      END DO
      DO K = 2, M+1
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
        CALL FDS (A(1), A(N+1), A(2*N+1), A(9*N+1), A(10*N+1),
     +             H1, H2, N1, N2, TAU)
        CALL RHS (A(9*N+1), N1, N2, H1, H2, FID, FIK, IM, JM, L, M, K)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
      END DO
C
C     SIMPLE ITERATION METHOD
C     NEXT APPROXIMATION
C
      CALL BS (A(10*N+1), A(1), A(N+1), A(2*N+1), A(12*N+1), N1, N2 )
      DO I = 1, N
        A(11*N+I) = A(11*N+I) - SS*A(12*N+I)
      END DO
C
C     EXIT FROM THE ITERATIVE PROCESS BY THE DISCREPANCY CRITERION
C
      SUM = 0.D0
      DO K = 1, M
        DO IP = 1, L
          SUM = SUM + (FID(IP,K) - FIK(IP,K))**2*TAU
        END DO
      END DO
      SUM = SUM/(L*TMAX)
      SL2 = DSQRT(SUM)
      WRITE (*,*) IT, SL2
      IF ( SL2.GT.DELTA ) GO TO 100
C
C     SOLUTION
C
      WRITE ( 01, * ) (A(11*N+I), I=1,N)
      WRITE ( 01, * ) ((FI(IP,K), IP=1,L), K=1,M)
      CLOSE (01)
      STOP
```

```
      END
C
      SUBROUTINE INIT (U, X1, X2, N1, N2)
C
C     INITIAL CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  U(N1,N2), X1(N1), X2(N2)
      DO I = 1, N1
        DO J = 1, N2
          U(I,J) = 0.D0
          IF ((X1(I)-0.6D0)**2 + (X2(J)-0.6D0)**2.LE.0.04D0)
     +            U(I,J) = 1.D0
        END DO
      END DO
C
      RETURN
      END
C
      SUBROUTINE FDS (A0, A1, A2, F, U, H1, H2, N1, N2, TAU)
C
C     GENERATION OF DIFFERENCE-SCHEME COEFFICIENTS
C     FOR PARABOLIC EQUATION WITH CONSTANT COEFFICIENTS
C     IN THE CASE OF PURELY IMPLICIT SCHEME
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  A0(N1,N2), A1(N1,N2), A2(N1,N2), F(N1,N2), U(N1,N2)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          A1(I-1,J) = 1.D0/(H1*H1)
          A1(I,J)   = 1.D0/(H1*H1)
          A2(I,J-1) = 1.D0/(H2*H2)
          A2(I,J)   = 1.D0/(H2*H2)
          A0(I,J)   = A1(I,J) + A1(I-1,J) + A2(I,J) + A2(I,J-1)
     +              + 1.D0/TAU
          F(I,J)    = U(I,J)/TAU
        END DO
      END DO
C
C     FIRST-KIND HOMOGENEOUS BOUNDARY CONDITION
C
      DO J = 2, N2-1
        A0(1,J) = 1.D0
        A1(1,J) = 0.D0
        A2(1,J) = 0.D0
        F(1,J)  = 0.D0
      END DO
C
      DO J = 2, N2-1
        A0(N1,J)   = 1.D0
        A1(N1-1,J) = 0.D0
        A1(N1,J)   = 0.D0
        A2(N1,J)   = 0.D0
        F(N1,J)    = 0.D0
      END DO
C
      DO I = 2, N1-1
        A0(I,1) = 1.D0
        A1(I,1) = 0.D0
```

```
        A2(I,1) = 0.D0
        F(I,1)  = 0.D0
      END DO
C
      DO I = 2, N1-1
        A0(I,N2)   = 1.D0
        A1(I,N2)   = 0.D0
        A2(I,N2)   = 0.D0
        A2(I,N2-1) = 0.D0
        F(I,N2)    = 0.D0
      END DO
C
       A0(1,1) = 1.D0
       A1(1,1) = 0.D0
       A2(1,1) = 0.D0
       F(1,1)  = 0.D0
C
       A0(N1,1) = 1.D0
       A2(N1,1) = 0.D0
       F(N1,1)  = 0.D0
C

       A0(1,N2) = 1.D0
       A1(1,N2) = 0.D0
       F(1,N2)  = 0.D0
C
       A0(N1,N2) = 1.D0
       F(N1,N2)  = 0.D0
C
      RETURN
      END
C
      SUBROUTINE RHS (F, N1, N2, H1, H2, FI, FIK, IM, JM, L, M, K)
C
C     RIGHT-HAND SIDE IN THE EQUATION FOR THE CONJUGATE STATE
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  F(N1,N2), FI(L,M), FIK(L,M), IM(L), JM(L)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          DO IP = 1, L
            IF (I.EQ.IM(IP).AND.J.EQ.JM(IP)) THEN
             F(I,J) = F(I,J) + (FIK(IP,M+2-K)-FI(IP,M+2-K))/(H1*H2)
            END IF
          END DO
        END DO
      END DO
C
      RETURN
      END
C
      SUBROUTINE MEGP (XP, YP, IM, JM, L, H1, H2)
C
C     OBSERVATION POINTS
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION IM(L), JM(L), XP(L), YP(L)
C
      XP(1) = 0.17D0
```

```
      YP(1) = 0.53D0
      XP(2) = 0.36D0
      YP(2) = 0.87D0
      XP(3) = 0.42D0
      YP(3) = 0.39D0
      XP(4) = 0.58D0
      YP(4) = 0.48D0
      XP(5) = 0.83D0
      YP(5) = 0.25D0
      XP(6) = 0.11D0
      YP(6) = 0.15D0
      XP(7) = 0.76D0
      YP(7) = 0.71D0
      XP(8) = 0.28D0
      YP(8) = 0.33D0
      XP(9) = 0.35D0
      YP(9) = 0.65D0
      XP(10) = 0.49D0
      YP(10) = 0.24D0
      DO IP = 1,L
        IM(IP) = XP(IP)/H1 + 1
        IF ((XP(IP) - IM(IP)*H1).GT. 0.5D0*H1) IM(IP) = IM(IP)+1
        JM(IP) = YP(IP)/H2 + 1
        IF ((YP(IP) - JM(IP)*H2).GT. 0.5D0*H2) JM(IP) = JM(IP)+1
      END DO
C
      RETURN
      END
C
      SUBROUTINE PU (IM, JM, U, N1, N2, F, L)
C
C     SOLUTION AT THE MEASUREMENT POINTS
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION IM(L), JM(L), F(L), U(N1,N2)
C
      DO IP = 1,L
        I = IM(IP)
        J = JM(IP)
        F(IP) = U(I,J)
      END DO
C
      RETURN
      END
C
      SUBROUTINE BS ( U, A0, A1, A2, B, N1, N2 )
C
C     FUNCTIONAL GRADIENT
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION U(N1,N2), A0(N1,N2), A1(N1,N2), A2(N1,N2), B(N1,N2)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          B(I,J) = A0(I,J)*U(I,J)
     *           - A1(I-1,J)*U(I-1,J)
     *           - A1(I,J)*U(I+1,J)
     *           - A2(I,J-1)*U(I,J-1)
     *           - A2(I,J)*U(I,J+1)
        END DO
```

```
      END DO
C
      RETURN
      END
```

We have restricted ourselves to the case in which in equation (7.257) we have $k(\boldsymbol{x}) = 1$ and, in addition, purely implicit schemes are used (the difference-scheme coefficients for the ground state are generated in the subroutine `FDS`, and for the conjugate state, in the subroutines `FDS` and `RHS`.

7.5.6 Computational experiments

The basic grid was the uniform grid with $h_1 = 0.02$ and $h_2 = 0.02$ for the problem in unit square. In the realization of the quasi-real experiment, as input data, the solution of the direct problem at some points of the calculation domain was taken. The problem is solved till the time $T = 0.025$, the time step size of the grid being $\tau = 0.00025$. In the direct problem, the initial condition is given in the form

$$u_0(\boldsymbol{x}, 0) = \begin{cases} 1, & (x_1 - 0.6)^2 + (x_2 - 0.6)^2 \le 0.04, \\ 0, & (x_1 - 0.6)^2 + (x_2 - 0.6)^2 > 0.04. \end{cases}$$

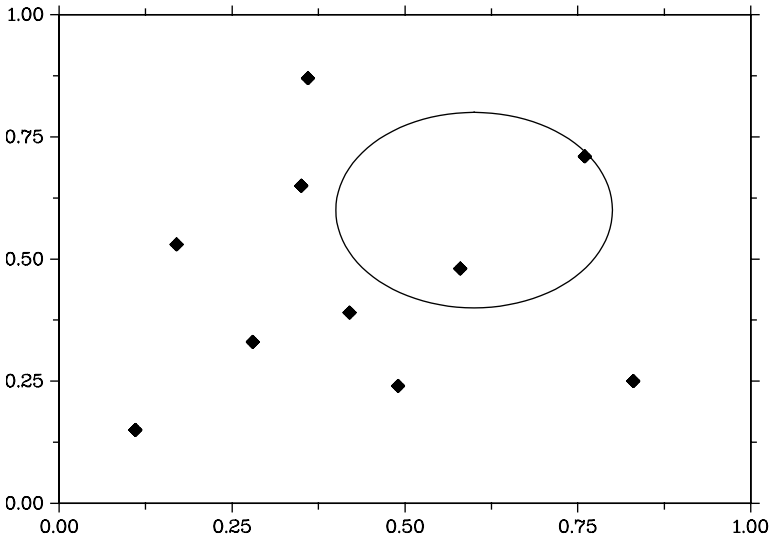

Figure 7.18 Observation points and initial conditions in the direct problem

Figure 7.18 shows the location of the observation points and indicates the portion of the calculation domain in which the initial condition was localized. As input data, the solution at the observation points is used (see Figure 7.19).

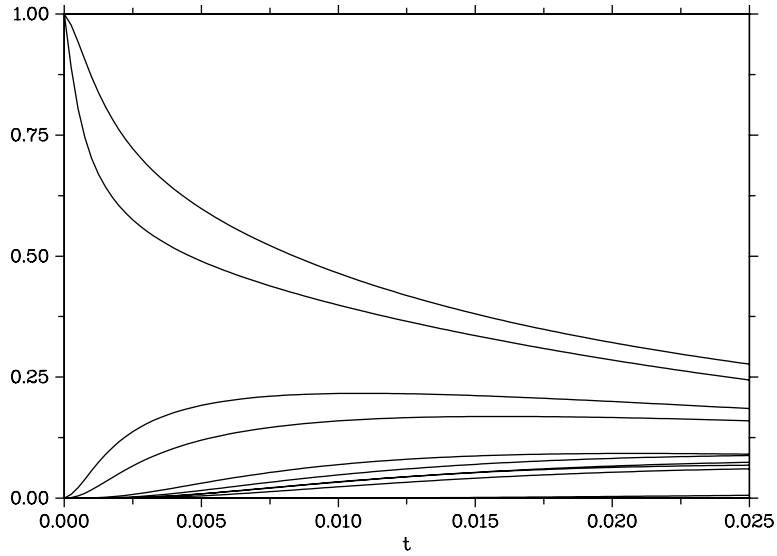

Figure 7.19 Solution at the observation points

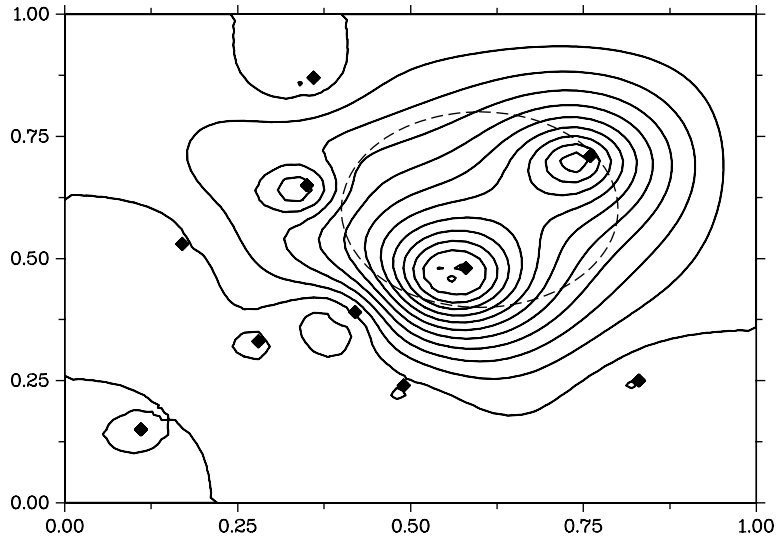

Figure 7.20 Inverse-problem solution obtained with $\delta = 0.02$

Consider calculated data that illustrate possibilities in reconstructing the initial condition from point observation data. Figure 7.20 shows the solution of the inverse problem obtained with the inaccuracy level $\delta = 0.02$ (plotted are contour lines following with the grid size $\triangle u = 0.1$). The effect due to the inaccuracy level is illustrated in Figures 7.21 and 7.22.

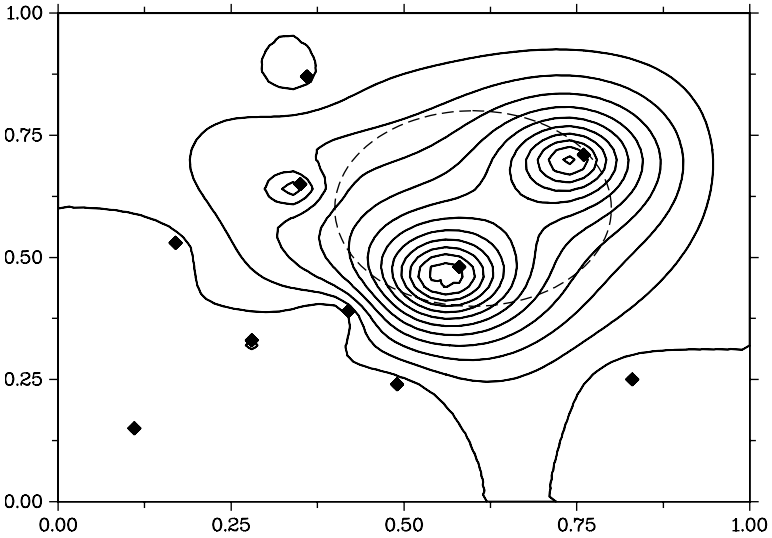

Figure 7.21 Inverse-problem solution obtained with $\delta = 0.04$

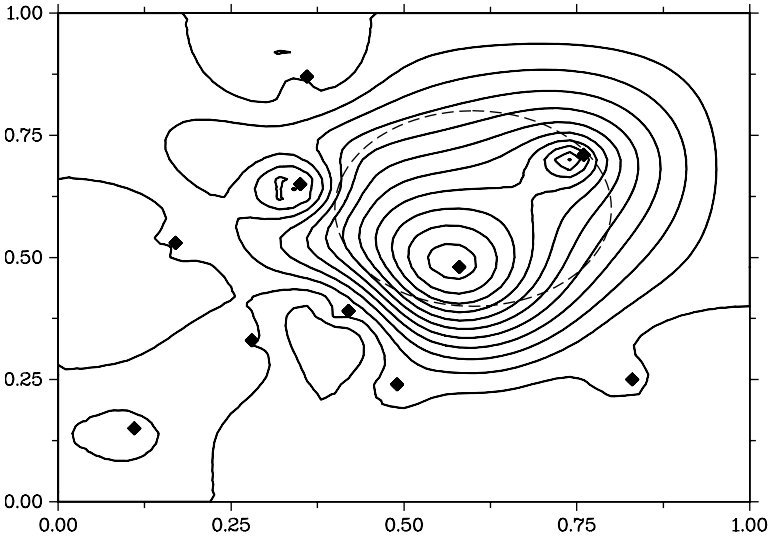

Figure 7.22 Inverse-problem solution obtained with $\delta = 0.01$

7.6 Exercises

Exercise 7.1 Examine convergence of the approximate solution found by formulas (7.14), (7.36) to the exact solution of problem (7.6), (7.7) $\mathcal{H}_{\mathcal{D}}$, $\mathcal{D} = \mathcal{S}^2$.

Exercise 7.2 Formulate the non-local problem that arises in the optimal control problem (7.38)–(7.41) with a non-self-adjoint operator $\mathcal{A}$.

Exercise 7.3 Using the program `PROBLEM10`, examine the time dependence of the approximate-solution inaccuracy. Compare the experimental data with the theoretical results (see estimate (7.35)).

Exercise 7.4 Examine the generalized inverse method (7.109), (7.138) to approximately solve the ill-posed problem (7.101), (7.102).

Exercise 7.5 Consider the additive scheme of component-by-component splitting (total approximation scheme)

$$\frac{y_{n+1/2} - y_n}{\tau} - \Lambda_1 y_n + \alpha \Lambda_1^2 (\sigma y_{n+1/2} + (1-\sigma) y_n) = 0,$$

$$\frac{y_{n+1} - y_{n+1/2}}{\tau} - \Lambda_2 y_{n+1/2} + \alpha \Lambda_2^2 (\sigma y_{n+1} + (1-\sigma) y_{n+1/2}) = 0$$

for the Cauchy problem posed for equation (7.158).

Exercise 7.6 Based on the data calculated by the program `PROBLEM11`, examine the effect of exact-solution smoothness on the initial-condition reconstruction inaccuracy.

Exercise 7.7 Consider the possibility of using iteration methods constructed around the minimization of the discrepancy functional in solving the inverted-time problem (7.159)–(7.161).

Exercise 7.8 Construct an iteration method for refining the initial condition in scheme (7.166), (7.167) as applied to the approximate solution of the retrospective inverse problem with a non-self-adjoint, positively defined operator Λ.

Exercise 7.9 Using the program `PROBLEM12`, experimentally examine how the choice of B in (7.179) affects the possibility of identifying a solution with desired smoothness.

Exercise 7.10 Construct a difference scheme for the non-local boundary value problem

$$\begin{aligned} -\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u &= f(x), \qquad 0 < x < l, \\ u(0) + \alpha u(l) &= \mu_1, \\ -k(0)\frac{du}{dx}(0) &= \mu_2. \end{aligned}$$

Modify the sweep algorithm so that to make it appropriate for the solution of the related non-local difference problem.

Exercise 7.11 Construct the difference scheme for the generalized inverse method (7.220)–(7.222) as applied to the approximate solution of the ill-posed problem (7.187)–(7.189) and perform a stability study for this scheme.

Exercise 7.12 Perform computational experiments (program `PROBLEM13`) to investigate into the solution inaccuracy in the Cauchy problem for the Laplace equation as dependent on the smoothness of exact initial conditions.

Exercise 7.13 Formulate optimality conditions for the minimization problem for (7.274), (7.275) under constraints (7.270), (7.271).

Exercise 7.14 Derive optimality conditions for the difference problem (7.271)–(7.274), (7.275) in the case of a non-self-adjoint operator $\Lambda > 0$.

Exercise 7.15 Modify the program `PROBLEM14` so that the initial condition be refined by the method of minimal discrepancies instead of the simple iteration method.

8 Other problems

Above, two classes of inverse problems for mathematical physics equations have been considered in which it was required to identify the right-hand side of an equation or the initial condition for the equation. Among other problems important for applications, boundary value inverse problems deserve mention in which to be reconstructed are the boundary conditions. For approximate solution of the latter problems, methods using some perturbation of the equation or methods using non-locally perturbed boundary conditions can be applied. In the case in which the generalized inverse method is used to solve some boundary value inverse problem, namely, in treating the spatial coordinate as the evolutionary coordinate, special emphasis is to be placed on the *hyperbolic* regularization method, used to pass from the hyperbolic to a parabolic equation. In the present chapter, possibilities offered by the generalized inverse method as applied to problems with perturbed boundary conditions are discussed with the example of the boundary value inverse problem for the one-dimensional parabolic equation of second order. For a more general two-dimensional problem, an algorithm with iteratively refined boundary condition is used. The problems most difficult for examination are coefficient inverse problems for mathematical physics equations. Here, we have restricted ourselves to the matter of numerical solution of two coefficient problems. In the first problem it is required to determine the higher coefficient as a function of the solution for a one-dimensional parabolic equation. We describe a computational algorithm that solves the coefficient inverse problem for the two-dimensional elliptic equation in the case in which the unknown coefficient does not depend on one of the two coordinates.

8.1 Continuation over the spatial variable in the boundary value inverse problem

In this section, we consider the boundary value inverse problem for the one-dimensional parabolic equation of second order (heat conduction equation). In this problem, it is required to reconstruct the boundary condition from measurements performed inside the calculation domain. This problem belongs to the class of conditionally well-posed problems and, to be solved stably, it requires the use of regularization methods. Here, the generalized inverse method is to be applied under conditions in which the problem is considered as an evolutionary one with respect to the spatial variable. The use of the generalized inverse method leads, in particular, to the well-known *hyperbolic* regularization of boundary value inverse problems.

8.1.1 Statement of the problem

Among inverse mathematical physics problems, of primary significance for practical applications is the boundary value inverse problem. This problem is often encountered in diagnostics, in the cases in which it is required to reconstruct, from additional measurements made inside the calculation domain, the thermal boundary condition at the domain boundary, where direct measurements are unfeasible.

This problem belongs to the class of conditionally well-posed problems and, for its approximate solution, development of special regularization methods is under way. A general approach for the solution of unstable problems for partial equations is the generalized inverse method. This method uses some perturbation of the initial equation, the problem for the perturbed equation being a well-posed one. Here, the perturbation parameter serves as regularization parameter.

In the consideration of the boundary value inverse problem for the one-dimensional parabolic equation of the second order, the generalized inverse method can be developed considering the initial problem as a problem for the evolutionary equation of the first order. A second possibility is related with the consideration of the boundary value inverse problem as a problem with initial data for the evolutionary equation of the second order. Here, as the evolutionary variable, the spatial variable is used. That is why here we speak of the continuation over the spatial variable in a boundary value inverse problem.

Consider a heat conduction boundary value inverse problem in which it is required to continue the solution over the spatial variable, which serves as time variable. The problem is to be constructed as follows. The solution $v(x, t)$ is to be found from the equation

$$\frac{\partial v}{\partial t} - \frac{\partial^2 v}{\partial x^2} = 0, \qquad 0 < x < l, \quad 0 < t < T, \tag{8.1}$$

supplemented with initial conditions, written in terms of the variables x and t, of the form

$$v(0, t) = \varphi(t), \qquad 0 < t < T, \tag{8.2}$$

$$\frac{\partial v}{\partial x}(0, t) = 0, \qquad 0 < t < T, \tag{8.3}$$

$$v(x, 0) = 0, \qquad 0 < x < l. \tag{8.4}$$

Let us apply in the boundary value inverse problem (8.1)–(8.4) the following change of variables: the variable x is changed for t, and t, for x ($l \to T$, $T \to l$). Next, we denote the solution to be found as $u(x, t)$ ($= v(t, x)$). For $u(x, t)$, we obtain the

problem

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial u}{\partial x} = 0, \qquad 0 < x < l, \quad 0 < t < T, \tag{8.5}$$

$$u(0,t) = 0, \qquad 0 < t < T, \tag{8.6}$$

$$u(x,0) = u_0(x), \qquad 0 < x < l, \tag{8.7}$$

$$\frac{\partial u}{\partial t}(x,0) = 0, \qquad 0 < x < l. \tag{8.8}$$

In the new settings, the function $\varphi(t)$ refers to $u_0(x)$.

The problem (8.5)–(8.8) is written as the operator equation

$$\frac{d^2 u}{dt^2} - \mathcal{A}u = 0 \tag{8.9}$$

with the initial conditions (8.7), (8.8). The operator $\mathcal{A}$ is defined by

$$\mathcal{A}u = \frac{\partial u}{\partial x} \tag{8.10}$$

with the domain of definition

$$\mathcal{D}(\mathcal{A}) = \{u \mid u = u(x,t),\ x \in [0,l],\ u(0,t) = 0\}.$$

The introduced operator $\mathcal{A}$ is not a self-adjoint, sign-definite operator in $\mathcal{H} = \mathcal{L}_2(0,l)$.

8.1.2 Generalized inverse method

To approximately solve the inverse problem (8.7)–(8.9), we use the generalized inverse method. We use this method in its version in which the approximate solution $u_\alpha(x,t)$ is to be determined from the perturbed equation

$$\frac{d^2 u_\alpha}{dt^2} - \mathcal{A}u_\alpha + \alpha \mathcal{A}^* \mathcal{A} u_\alpha = 0, \tag{8.11}$$

supplemented with the initial conditions

$$u_\alpha(x,0) = u_0(x), \qquad 0 < x < l, \tag{8.12}$$

$$\frac{\partial u_\alpha}{\partial t}(x,0) = 0, \qquad 0 < x < l. \tag{8.13}$$

By virtue of (8.10), the operator conjugate in $\mathcal{H} = \mathcal{L}_2(0,l)$ to $\mathcal{A}$ is given by

$$\mathcal{A}^* u = -\frac{\partial u}{\partial x}, \tag{8.14}$$

and, in addition,

$$\mathcal{D}(\mathcal{A}^*) = \{u \mid u = u(x,t),\ x \in [0,l],\ u(l,t) = 0\}.$$

With regard to (8.10), (8.14), in the variant (8.11)–(8.13) of the generalized inverse method it is required to solve the problem

$$\frac{\partial^2 u_\alpha}{\partial t^2} - \frac{\partial u_\alpha}{\partial x} + \alpha \frac{\partial^2 u_\alpha}{\partial x^2} = 0, \qquad 0 < x < l, \quad 0 < t < T, \tag{8.15}$$

$$u_\alpha(0, t) = 0, \qquad 0 < t < T, \tag{8.16}$$

$$\frac{\partial u_\alpha}{\partial x}(l, t) = 0, \qquad 0 < t < T, \tag{8.17}$$

$$u_\alpha(x, 0) = u_0(x), \qquad 0 < x < l, \tag{8.18}$$

$$\frac{\partial u_\alpha}{\partial t}(x, 0) = 0, \qquad 0 < x < l. \tag{8.19}$$

Thus, the generalized inverse method leads us to a hyperbolic perturbation (see (8.15)) of the initial parabolic equation (8.5), this perturbation being a well-known one in the computational practice.

Theorem 8.1 *For the solution of the boundary value problem* (8.15)–(8.19)*, there holds the following a priori estimate:*

$$\left\| \frac{\partial u_\alpha}{\partial t}(x, t) \right\|^2 + \alpha \left\| \frac{\partial u_\alpha}{\partial x}(x, t) \right\|^2 \le \alpha \exp\left(\frac{1}{\sqrt{\alpha}}\, t \right) \left\| \frac{\partial u_0}{\partial x}(x) \right\|^2. \tag{8.20}$$

Proof. Let us prove the above statement under more general conditions (namely, for problem (8.11)–(8.13)). We multiply equation (8.11) scalarwise by du/dt; this yields:

$$\frac{1}{2} \frac{d}{dt} \left(\left\| \frac{du_\alpha}{dt} \right\|^2 + \alpha \|\mathcal{A} u_\alpha\|^2 \right) = \left(\mathcal{A} u_\alpha, \frac{du_\alpha}{dt} \right). \tag{8.21}$$

For the right-hand side of (8.21), we have:

$$\left(\mathcal{A} u_\alpha, \frac{du_\alpha}{dt} \right) \le \frac{1}{2} \frac{1}{\sqrt{\alpha}} \left(\left\| \frac{du_\alpha}{dt} \right\|^2 + \alpha \|\mathcal{A} u_\alpha\|^2 \right). \tag{8.22}$$

Substitution of (8.22) into (8.21) yields the estimate

$$\left\| \frac{du_\alpha}{dt} \right\|^2 + \alpha \|\mathcal{A} u_\alpha\|^2 \le \alpha \exp\left(\frac{1}{\sqrt{\alpha}}\, t \right) \|\mathcal{A} u_0\|^2. \tag{8.23}$$

Inequality (8.23) yields the following simpler estimate:

$$\|\mathcal{A} u_\alpha\| \le \alpha \exp\left(\frac{1}{\alpha^{1/4}}\, t \right) \|\mathcal{A} u_0\|.$$

The desired estimate (8.20) for the perturbed problem (8.15)–(8.19) is a simple corollary to the estimate (8.23). □

Some other variants of the generalized inverse method can also be used. For instance, we can determine the approximate solution from the equation

$$\frac{d^2u_\alpha}{dt^2} - \mathcal{A}u_\alpha + \alpha\mathcal{A}\mathcal{A}^* \frac{du_\alpha}{dt} = 0, \tag{8.24}$$

supplemented with conditions (8.12), (8.13).

Hence, with (8.10) and (8.14), the approximate solution $u_\alpha(x, t)$ is to be found from the equation

$$\frac{\partial^2 u_\alpha}{\partial t^2} - \frac{\partial u_\alpha}{\partial x} + \alpha \frac{\partial^3 u_\alpha}{\partial x^2\, \partial t} = 0, \qquad 0 < x < l, \quad 0 < t < T \tag{8.25}$$

and conditions (8.16)–(8.19). Similarly to Theorem 8.1, we formulate the following statement:

Theorem 8.2 *For the solution of the boundary value problem* (8.16)–(8.19), (8.25), *there holds the a priori estimate*

$$\|u_\alpha(x,t)\|^2 + \left\| \frac{\partial u_\alpha}{\partial t}(x,t) \right\|^2 \;\le\; \exp\Big(\frac{1+2\alpha}{2\alpha}\, t\Big) \|u_0(x)\|^2. \tag{8.26}$$

Proof. Here again, the consideration will be performed for the general perturbed evolutionary equation. We can conveniently reformulate equation (8.24) as a system of first-order equations.

We define the vector $U = \{u_1, u_2\}$ and the space $\mathcal{H}^2$ as the direct sum of spaces $\mathcal{H}$: $\mathcal{H}^2 = \mathcal{H} \oplus \mathcal{H}$. The addition in $\mathcal{H}^2$ is performed coordinatewise, and the scalar product is defined as

$$(U, V) = (u_1, v_1) + (u_2, v_2).$$

We define $u_1 = u_\alpha$, $u_2 = du_\alpha/dt$ and write equation (8.24) as a system of first-order equations ($U_\alpha = \{u_1, u_2\}$):

$$\frac{dU_\alpha}{dt} + \mathcal{P}U_\alpha = 0, \tag{8.27}$$

where

$$\mathcal{P} = \begin{bmatrix} 0 & -E \\ -\mathcal{A} & \alpha\mathcal{A}\mathcal{A}^* \end{bmatrix}. \tag{8.28}$$

Equation (8.27) is supplemented (see (8.12), (8.13)) with the initial conditions

$$U_\alpha(0) = U_0 = \{u_0, 0\}. \tag{8.29}$$

In the notation introduced, we have:

$$\|U_\alpha\|^2 \;=\; \|u_1\|^2 + \|u_2\|^2 \;=\; \|u_\alpha\|^2 + \left\| \frac{du_\alpha}{dt} \right\|^2.$$

For problem (8.27)–(8.29), there holds the a priori estimate

$$\|U_\alpha(t)\|^2 \le \exp\Big(\frac{1+2\alpha}{2\alpha}\,t\Big)\|U_\alpha(0)\|^2. \tag{8.30}$$

With (8.28), we have:

$$\mathcal{P}U = \{-u_2, -\mathcal{A}u_1 + \alpha\mathcal{A}\mathcal{A}^* u_2\}.$$

We multiply equation (8.27) scalarwise by U_α and obtain:

$$\frac{1}{2}\frac{d}{dt}\left(\|u_1\|^2 + \|u_2\|^2\right) + \alpha\|\mathcal{A}^* u_2\|^2 = (\mathcal{A}u_1, u_2) + (u_1, u_2). \tag{8.31}$$

For the right-hand side, we use the estimate

$$\begin{aligned} (\mathcal{A}u_1, u_2) &\le \alpha\|\mathcal{A}^* u_2\|^2 + \frac{1}{4\alpha}\,\|u_1\|^2, \\ (u_1, u_2) &\le \varepsilon\|u_2\|^2 + \frac{1}{4\varepsilon}\,\|u_1\|^2. \end{aligned} \tag{8.32}$$

Choosing $\varepsilon = 1/2$ and substituting (8.32) into (8.31), we arrive at the estimate (8.30).

From (8.30), the estimate (8.26) for the solution of problem (8.16)–(8.19), (8.25) follows. □

8.1.3 Difference schemes for the generalized inverse method

Consider now the proposed variants of the generalized inverse method on the mesh level. For simplicity, we restrict ourselves to the case of a one-dimensional boundary value problem. To pass to a difference problem, we introduce the uniform grid

$$\bar{\omega} = \{x \mid x = x_i = ih,\ i = 0, 1, \dots, N,\ Nh = l\},$$

where ω is the set of internal nodes, and $\partial\omega$ is the set of boundary nodes. The operators A and A^* are introduced in the traditional manner. Let in the subscriptless notation we have

$$Ay = \begin{cases} 0, & i = 0, \\ y_{\bar{x}}, & i = 1, 2, \dots, N \end{cases}$$

with the domain of definition

$$\mathcal{D}(A) = \{y \mid y(x),\ x \in \bar{\omega},\ y_0 = 0\}.$$

We define the scalar product and the norm in the mesh Hilbert space $H = L_2(\omega)$ as

$$(z, y) = \sum_{x\in\bar{\omega}} z(x)y(x)h, \qquad \|y\| = \sqrt{(y, y)}.$$

Then, for the adjoint operator A^* we have

$$A^*y = \begin{cases} -y_x, & i = 0, 1, \dots, N-1, \\ 0, & i = N, \end{cases}$$

and

$$\mathcal{D}(A^*) = \{y \mid y(x),\ x \in \bar{\omega},\ y_N = 0\}.$$

For the operator A^*A we have:

$$A^*Ay = \begin{cases} -y_x/h, & i = 0, \\ -y_{\bar{x}x}, & i = 1, 2, \dots, N-1, \\ 0, & i = N. \end{cases}$$

Let us dwell first on the variant (8.11) of the generalized inverse method. We determine the difference solution of problem (8.11)–(8.13) from the equation

$$\frac{d^2y}{dt^2} - Ay + \alpha A^*Ay = 0 \tag{8.33}$$

for $y(x, t) \in H$, supplemented with the initial conditions

$$y(x, 0) = u_0(x), \qquad x \in \bar{\omega}, \tag{8.34}$$

$$\frac{dy}{dt}(x, 0) = 0, \qquad x \in \bar{\omega}. \tag{8.35}$$

To approximately solve problem (8.33)–(8.35), we use a time-uniform grid with the time step size τ. Consider the difference scheme

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} - \frac{1}{2}A(y_n + y_{n-1}) + \frac{\alpha}{4}A^*A(y_{n+1} + 2y_n + y_{n-1}) = 0, \qquad n = 1, 2, \dots. \tag{8.36}$$

To perform a stability study for the scheme (8.36), we introduce the settings

$$v_n = \frac{1}{2}(y_n + y_{n-1}), \qquad w_n = \frac{1}{\tau}(y_n - y_{n-1}). \tag{8.37}$$

Then, the difference scheme (8.36) can be written as

$$\frac{w_{n+1} - w_n}{\tau} - Av_n + \frac{\alpha}{2}A^*A(v_{n+1} + v_n) = 0. \tag{8.38}$$

We multiply the difference equation (8.38) scalarwise by

$$2(v_{n+1} - v_n) = y_{n+1} - y_{n-1} = \tau(w_{n+1} + w_n);$$

then we obtain

$$\|w_{n+1}\|^2 - \|w_n\|^2 + \alpha\|Av_{n+1}\|^2 - \alpha\|Av_n\|^2 = \tau(Av_n, w_{n+1} + w_n). \tag{8.39}$$

The right-hand side in (8.39) can be estimated as follows:

$$\begin{aligned}(Av_n,\, w_{n+1}+w_n) &\le \beta\|Av_n\|^2+\frac{1}{4\beta}\,\|w_{n+1}+w_n\|^2\\ &\le \beta\|Av_n\|^2+\frac{1+\varepsilon}{4\beta}\,\|w_{n+1}\|^2+\frac{1+\varepsilon}{4\beta\varepsilon}\,\|w_n\|^2.\end{aligned}$$

We choose $\varepsilon = 1/2$; then, we substitute the latter inequality into (8.39). This yields

$$\Big(1-\frac{3}{8}\,\frac{\tau}{\beta}\Big)\|w_{n+1}\|^2+\alpha\|Av_{n+1}\|^2 \le \Big(1+\frac{3}{4}\,\frac{\tau}{\beta}\Big)\|w_n\|^2+\alpha\Big(1+\frac{\beta}{\alpha}\Big)\|Av_n\|^2. \quad (8.40)$$

Using in the case of $\tau \le 4\beta/3$ the estimate

$$\Big(1-\frac{3}{8}\,\frac{\tau}{\beta}\Big)^{-1} \le \exp\Big(\frac{3}{4\beta}\,\tau\Big)$$

and choosing $\beta = (3\alpha/2)^{1/2}$, from (8.40) we obtain the a priori estimate

$$\Big(1-\frac{3}{8}\,\frac{\tau}{\beta}\Big)\|w_{n+1}\|^2+\alpha\|Av_{n+1}\|^2 \le \varrho^2\Big(\Big(1-\frac{3}{8}\,\frac{\tau}{\beta}\Big)\|w_n\|^2+\alpha\|Av_n\|^2\Big) \quad (8.41)$$

with

$$\varrho = \exp\Big(\sqrt{\frac{3}{2\alpha}}\,\tau\Big). \quad (8.42)$$

Thus, the following statement is proved:

Theorem 8.3 *For the difference scheme* (8.36)*, in the case of* $\tau \le 2(2\alpha/3)^{1/2}$ *there holds the estimate* (8.37)*,* (8.40)*.*

In the variant (8.24) of the generalized inverse method, it is required to solve the equation

$$\frac{d^2y}{dt^2}-Ay+\alpha AA^*\,\frac{dy}{dt}=0$$

with the initial conditions (8.34), (8.35). In the numerical realization, we use the scheme

$$\frac{y_{n+1}-2y_n+y_{n-1}}{\tau^2}-Ay_n+\alpha AA^*\,\frac{y_{n+1}-y_{n-1}}{2\tau} = 0, \quad n=1,2,\ldots. \quad (8.43)$$

Theorem 8.4 *The difference scheme* (8.43) *is* ϱ*-stable with*

$$\varrho = \exp\Big(\frac{1+4\alpha}{4\alpha}\,\tau\Big)$$

and, for this scheme, there holds the a priori estimate

$$\|y_n\|^2+\Big\|\frac{y_{n+1}-y_n}{\tau}\Big\|^2 \le \varrho^2\Big(\|y_{n-1}\|^2+\Big\|\frac{y_n-y_{n-1}}{\tau}\Big\|^2\Big). \quad (8.44)$$

Proof. To obtain the desired a priori estimate, we scalarwise multiply the difference equation (8.43) by

$$y_t^{\circ} = \frac{y_{n+1} - y_{n-1}}{2\tau}$$

and, using the standard subscriptless notation adopted in the theory of difference schemes, arrive at the equality

$$(y_{\bar{t}t}, y_t^{\circ}) - (Ay, y_t^{\circ}) + \alpha \|A^* y_t^{\circ}\|^2 = 0. \tag{8.45}$$

For the first two terms in the left side of (8.45) we have

$$\begin{aligned} (y_{\bar{t}t}, y_t^{\circ}) &= \frac{1}{2} (y_{\bar{t}}, y_{\bar{t}})_t, \\ (Ay, y_t^{\circ}) &\le \alpha \|A^* y_t^{\circ}\|^2 + \frac{1}{4\alpha} \|y\|^2. \end{aligned} \tag{8.46}$$

We insert (8.46) into (8.45); then, we arrive at the inequality

$$(y_{\bar{t}}, y_{\bar{t}})_t \le \frac{1}{2\alpha} \|y\|^2. \tag{8.47}$$

We add to the both sides of (8.47) the term

$$\begin{aligned} (\|y\|^2)_{\bar{t}} &= \frac{1}{\tau} (\|y_n\|^2 - \|y_{n-1}\|^2) \\ &= \Big(y_n + y_{n-1}, \frac{y_n - y_{n-1}}{\tau}\Big) = 2(y_{n-1}, y_{\bar{t}}) + \tau \|y_{\bar{t}}\|^2. \end{aligned}$$

Using the fact that

$$\|y\|^2 = \|y_{n-1} + \tau y_{\bar{t}}\|^2 \le (1+\varepsilon)\|y_{n-1}\|^2 + \Big(1 + \frac{1}{4\varepsilon}\Big)\tau^2 \|y_{\bar{t}}\|^2$$

and taking the inequality

$$(y_{n-1}, y_{\bar{t}}) \le \beta \|y_{n-1}\|^2 + \frac{1}{4\beta} \|y_{\bar{t}}\|^2$$

into account, from (8.47) we obtain the inequality

$$\begin{aligned} &(\|y_{n-1}\|^2 + \|y_{\bar{t}}\|^2)_t \\ &\qquad \le \Big(\frac{1+\varepsilon}{2\alpha} + 2\beta\Big)\|y_{n-1}\|^2 + \Big(\frac{1}{2\beta} + \tau + \frac{\tau^2}{2\alpha}\,\frac{1+4\varepsilon}{4\varepsilon}\Big)\|y_{\bar{t}}\|^2, \end{aligned} \tag{8.48}$$

that holds for all positive ε and β.

We choose $\varepsilon = \tau/2$ and $\beta = 1/2$; then, we obtain

$$\frac{1+\varepsilon}{2\alpha} + 2\beta = 1 + \frac{1}{2\alpha} + \frac{\tau}{4\alpha} < \frac{1}{\tau}(\varrho^2 - 1),$$
$$\frac{1}{2\beta} + \tau + \frac{\tau^2}{2\alpha}\,\frac{1+4\varepsilon}{4\varepsilon} = 1 + \tau + \frac{\tau}{4\alpha} + \frac{\tau^2}{2\alpha} < \frac{1}{\tau}(\varrho^2 - 1).$$

Here, the value of ϱ is defined in the conditions of the theorem. With regard to these inequalities, inequality (8.48) yields the desired estimate (8.44). □

8.1.4 Program

We numerically solve the boundary value inverse problem (8.1)–(8.4). In the framework of the quasi-real computational experiment, to formulate the boundary conditions (8.2), we consider the direct problem in which it is required to determine the function $v(x,t)$ from equation (8.1), initial condition (8.4), boundary condition (8.3) at the left boundary, and the condition

$$v(l,t) = \psi(t), \qquad 0 < t < T$$

at the right boundary. To find the approximate solution, we use the purely implicit difference scheme.

The approximate solution of the inverse problem can be found using the variant (8.24) of the generalized inverse method. In terms of the initial variables, this corresponds to the solution of the problem (see (8.16)–(8.19), (8.25))

$$\frac{\partial^2 v_\alpha}{\partial x^2} - \frac{\partial v_\alpha}{\partial t} + \alpha\,\frac{\partial^3 v_\alpha}{\partial t^2\,\partial x} = 0, \qquad 0 < x < l, \quad 0 < t < T,$$
$$v_\alpha(0,t) = \varphi(t)), \qquad 0 < t < T,$$
$$\frac{\partial v_\alpha}{\partial x}(0,t) = 0, \qquad 0 < t < T,$$
$$v_\alpha(x,0) = 0, \qquad 0 < x < l,$$
$$\frac{\partial v_\alpha}{\partial t}(x,T) = 0, \qquad 0 < x < l.$$

This boundary value problem is solved by the difference scheme (8.43).

The value of α is determined using the discrepancy criterion. We assume that the input-data inaccuracy (the inaccuracy in setting $\varphi(t)$)) plays a decisive role and, hence, the approximation inaccuracy in choosing the regularization parameter can be ignored.

Program PROBLEM15

```
C
C     PROBLEM15 - IDENTIFICATION OF THE BOUNDARY CONDITION
C                 NON-STATIONARY ONE-DIMENSIONAL PROBLEM
```

```
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.01D0, N = 101, M = 101 )
      DIMENSION  X(N), Y(N,M), F(N,M), YY(M)
     +          ,FI(M), FID(M), FIY(M), Q(M), QA(M)
     +          ,A(M), B(M), C(M), FF(M) ! M >= N
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT ENDS OF THE SEGMENT;
C     N         - NUMBER OF GRID NODES OVER SPACE;
C     TMAX      - MAXIMAL TIME;
C     M         - NUMBER OF GRID NODES OVER TIME;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     FI(M)     - EXACT DIFFERENCE BOUNDARY CONDITION;
C     FID(M)    - DISTURBED DIFFERENCE BOUNDARY CONDITION;
C     Q(M)      - EXACT SOLUTION OF THE INVERSE PROBLEM
C                   (BOUNDARY CONDITION);
C     QA(M)     - APPROXIMATE SOLUTION OF THE INVERSE PROBLEM;
C
      XL = 0.D0
      XR = 1.D0
      TMAX = 1.D0
C
      OPEN  (01, FILE='RESULT.DAT')! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H = (XR - XL) / (N - 1)
      DO I = 1, N
        X(I) = XL + (I-1)*H
      END DO
      TAU = TMAX / (M-1)
C
C     DIRECT PROBLEM
C
C     BOUNDARY REGIME
C
      DO K = 1, M
        T    = (K-1)*TAU
        Q(K) = AF(T, TMAX)
      END DO
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I,1)  = 0.D0
      END DO
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C     PURELY IMPLICIT SCHEME
C
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
```

```
          B(I) = 1.D0 / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          FF(I) = Y(I,K-1) / TAU
        END DO
C
C     BOUNDARY CONDITION AT THE LEFT AND RIGHT END POINTS
C
        B(1)  = 2.D0 / (H*H)
        C(1)  = B(1) + 1.D0 / TAU
        FF(1) = Y(1,K-1) / TAU
        A(N)  = 0.D0
        C(N)  = 1.D0
        FF(N) = Q(K)
C
C     SOLUTION OF THE PROBLEM ON THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, FF, YY, ITASK )
        DO I = 1, N
          Y(I,K) = YY(I)
        END DO
      END DO
C
C      SOLUTION ON THE LEFT BOUNDARY
C
      DO K = 1, M
        FI(K)  = Y(1,K)
        FID(K) = FI(K)
      END DO
C
C     NOISE ADDITION TO THE SOLUTION OF THE BOUNDARY-VALUE PROBLEM
C
      DO K = 2, M
        FID(K) = FI(K) + 2.D0*DELTA*(RAND(0)-0.5D0)
      END DO
C
C     INVERSE PROBLEM
C
C     GENERALIZED INVERSE METHOD
C     CONTINUATION OVER THE SPATIAL VARIABLE
C
      IT = 0
      ITMAX = 100
      ALPHA = 0.001D0
      QQ = 0.75D0
  100 IT = IT + 1
C
C      INITIAL CONDITIONS
C      (LEFT BOUNDARY)
C
      DO K = 2, M
        Y(1,K) = FID(K)
        Y(2,K) = FID(K) + 0.5D0*H**2*(FID(K)-FID(K-1))/TAU
      END DO
C
C     NEXT LAYER
C
      DO I = 3, N
C
C     DIFFERENCE-SCHEME COEFFICIENTS
```

```
C
        DO K = 2, M-1
          A(K) = ALPHA / (H*TAU**2)
          B(K) = ALPHA / (H*TAU**2)
          C(K) = A(K) + B(K) + 1.D0 / (H*H)
          FF(K) = (Y(I-1,K)-Y(I-1,K-1))/TAU
     +          + (2.D0*Y(I-1,K) - Y(I-2,K))/(H*H)
     +          - A(K)*(Y(I-2,K+1)-2.D0*Y(I-2,K)+Y(I-2,K-1))
        END DO
C
C     BOUNDARY CONDITION ON THE BOTTOM AND ON THE TOP
C
        B(1)  = 0.D0
        C(1)  = 1.D0
        FF(1) = 0.D0
        A(M)  = ALPHA / (H*TAU**2)
        C(M)  = A(M) + 1.D0 / (H*H)
        FF(M) = (Y(I-1,M)-Y(I-1,M-1))/TAU
     +          + (2.D0*Y(I-1,M) - Y(I-2,M))/(H*H)
     +          - A(M)*(-Y(I-2,M)+Y(I-2,M-1))
C
C     SOLUTION OF THE PROBLEM ON THE NEXT LAYER
C
        ITASK = 1
        CALL PROG3 ( M, A, C, B, FF, YY, ITASK )
        DO K = 1, M
          Y(I,K) = YY(K)
        END DO
      END DO
C
C     SOLUTION
C
      DO K = 1, M
        QA(K) = Y(N,K)
      END DO
C
C     SOLUTION OF THE DIRECT PROBLEM WITH THE FOUND BOUNDARY CONDITION
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I,1)  = 0.D0
      END DO
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
          B(I) = 1.D0 / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          FF(I) = Y(I,K-1) / TAU
        END DO
C
C     BOUNDARY CONDITION AT THE LEFT AND RIGHT END POINTS
C
```

```
          B(1)  = 2.D0 / (H*H)
          C(1)  = B(1) + 1.D0 / TAU
          FF(1) = Y(1,K-1) / TAU
          A(N)  = 0.D0
          C(N)  = 1.D0
          FF(N) = QA(K)
C
C     SOLUTION OF THE PROBLEM ON THE NEXT TIME LAYER
C
          ITASK = 1
          CALL PROG3 ( N, A, C, B, FF, YY, ITASK )
          DO I = 1, N
            Y(I,K) = YY(I)
          END DO
        END DO
C
C     CRITERION FOR THE EXIT FROM THE ITERATIVE PROCESS
C
        SUM = 0.D0
        DO K = 1, M
          FIY(K) = Y(1,K)
          SUM = SUM + (FIY(K) - FID(K))**2*TAU
        END DO
        SL2 = DSQRT(SUM)
C
        IF (IT.GT.ITMAX) STOP
        IF ( IT.EQ.1 ) THEN
          IND = 0
          IF ( SL2.LT.DELTA ) THEN
            IND = 1
            QQ = 1.D0/QQ
          END IF
          ALPHA = ALPHA*QQ
          GO TO 100
        ELSE
          ALPHA = ALPHA*QQ
          IF ( IND.EQ.0 .AND. SL2.GT.DELTA ) GO TO 100
          IF ( IND.EQ.1 .AND. SL2.LT.DELTA ) GO TO 100
        END IF
C
C     RECORDING OF CALCULATED DATA
C
        WRITE ( 01,* ) (Q(K), K = 1,M)
        WRITE ( 01,* ) (FID(K), K = 1,M)
        WRITE ( 01,* ) (QA(K), K = 1,M)
        WRITE ( 01,* ) (FIY(K), K = 1,M)
        CLOSE ( 01 )
        STOP
        END
C
        DOUBLE PRECISION FUNCTION  AF ( T, TMAX )
        IMPLICIT REAL*8 ( A-H, O-Z )
C
C       BOUNDARY CONDITION ON THE RIGHT BOUNDARY
C
        AF = 2.D0*T/TMAX
        IF (T.GT.(0.5D0*TMAX)) AF = 2.D0*(TMAX-T)/TMAX
C
        RETURN
```

```
      END
```

8.1.5 Examples

As the basic one, a uniform grid with $h = 0.01$ and $\tau = 0.01$ for the problem with $l = 1$ and $T = 1$ was used. In the realization of the quasi-real experiment for the direct problem, the boundary condition at the right end point was set as follows:

$$\psi(t) = \begin{cases} 2t/T, & 0 < t < T/2, \\ 2(T-t)/T, & T/2 < t < T. \end{cases}$$

Figure 8.1 shows the contour lines for the direct-problem solution obtained with the chosen boundary conditions.

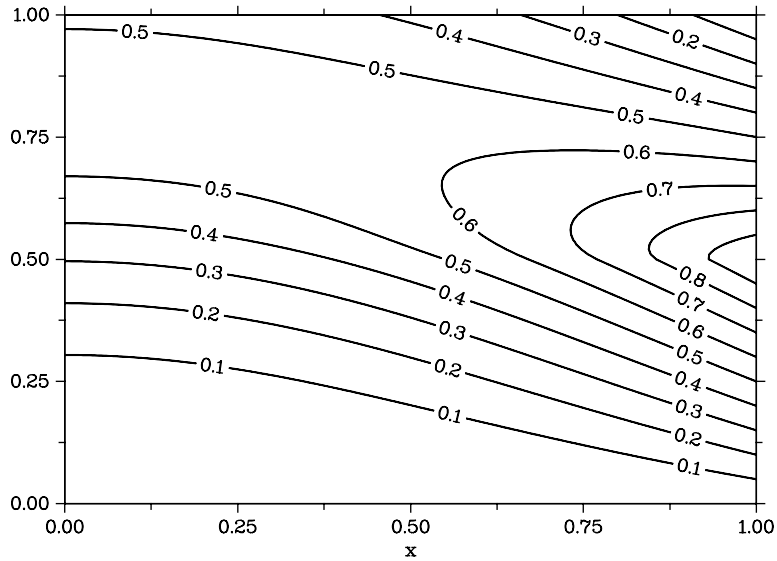

Figure 8.1 Direct-problem solution

First of all, we would like to know how the inaccuracy in setting the boundary conditions affects the solution accuracy in the inverse problem. Figure 8.2 shows the solution of the inverse problem obtained with the inaccuracy level defined by $\delta = 0.01$. Here, plotted are the exact and perturbed solutions at the left boundary (at $x = 0$), serving the input data in solving the inverse problem. From these conditions, the solution in the interval $0 < x \leq l$ is to be reconstructed. Plotted in the figure are the exact solution and the found solution at $x = 1$. The direct-problem solution at $x = 1$ for the found boundary condition is shown in Figure 8.3 (compare with Figure 8.1). The effect due to the inaccuracy level is illustrated by Figures 8.4 and 8.5.

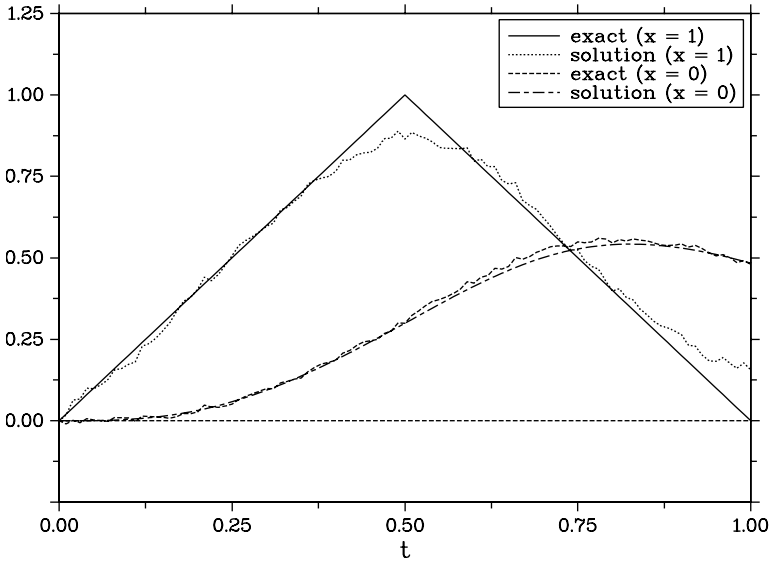

Figure 8.2 Inverse-problem solution obtained with $\delta = 0.01$

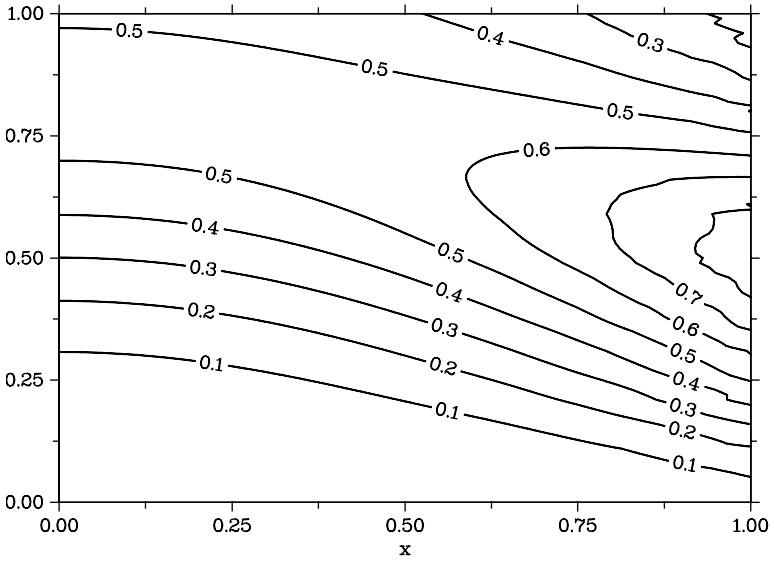

Figure 8.3 Direct-problem solution obtained with the found boundary conditions

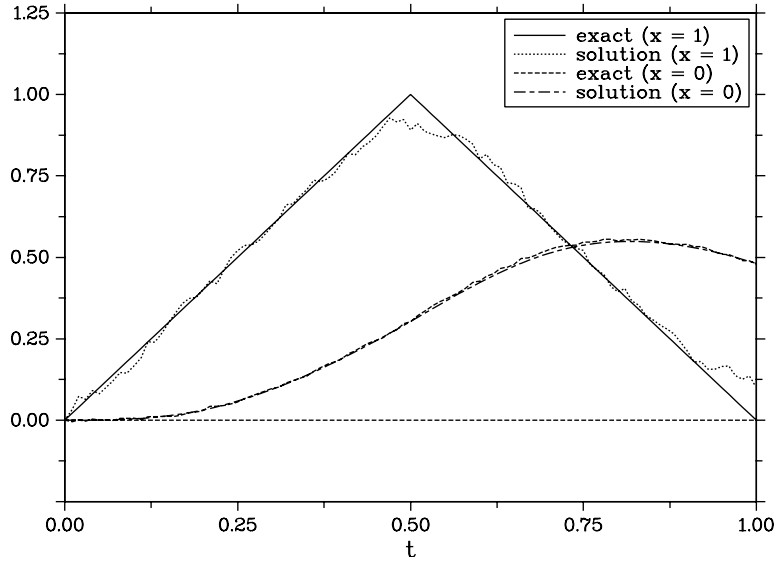

Figure 8.4 Inverse-problem solution obtained with $\delta = 0.005$

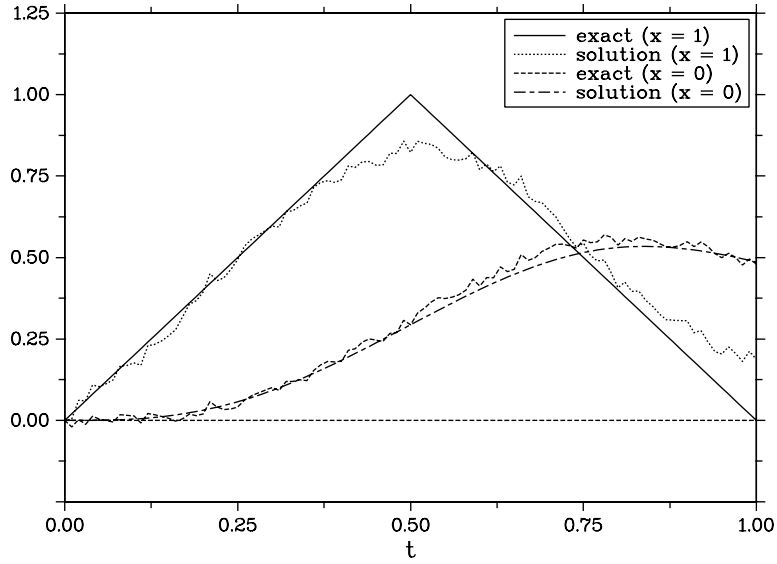

Figure 8.5 Inverse-problem solution obtained with $\delta = 0.02$

8.2 Non-local distribution of boundary conditions

For the approximate solution of the boundary value inverse problem for the one-dimensional parabolic equation, a computational algorithm is often used based on non-local perturbation of the boundary condition. This approach can be related to the local Tikhonov regularization.

8.2.1 Model problem

We assume that the process of interest obeys the one-dimensional parabolic equation of the second order. The related direct problem can be formulated as follows:

The solution $u(x,t)$ is to be determined in the rectangle

$$\overline{Q}_T = \overline{\Omega} \times [0,T], \qquad \overline{\Omega} = \{x \mid 0 \le x \le l\}, \qquad 0 \le t \le T.$$

The function $u(x,t)$ satisfies the equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l, \quad 0 < t \le T, \tag{8.49}$$

with the usual constraints $k(x) \ge \kappa > 0$. The adopted boundary and initial conditions are as follows:

$$k(x)\,\frac{\partial u}{\partial x}(0,t) = 0, \qquad 0 < t \le T, \tag{8.50}$$

$$u(l,t) = \psi(t), \qquad 0 < t \le T, \tag{8.51}$$

$$u(x,0) = 0, \qquad 0 \le x \le l. \tag{8.52}$$

We consider the boundary value inverse problem in which the boundary condition at the right boundary is not given (the function $\psi(t)$ in (8.51) is unknown). Instead, given is the additional condition at the left boundary:

$$u(0,t) = \varphi(t), \qquad 0 < t \le T. \tag{8.53}$$

Additionally, we assume that, as it is often the case in practice, the latter boundary condition is given with some inaccuracy.

8.2.2 Non-local boundary value problem

Among the various possible approaches to the approximate solution of inverse problems for evolutionary equations, we choose to treat methods with perturbed initial equation and methods with perturbed initial (boundary) conditions. The variant of the generalized inverse method with the passage to a well-posed problem for a perturbed equation was realized above by considering the spatial variable as the evolutionary variable. It is of interest here to use the variant of this method with perturbed boundary (initial, with the interpretation of the variable x as the evolutionary variable) condition.

We denote the approximate solution of the boundary value inverse problem (8.49), (8.50), (8.52), (8.53) as $u_\alpha(x,t)$ and determine it from the equation

$$\frac{\partial u_\alpha}{\partial t} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u_\alpha}{\partial x}\Big), \qquad 0 < x < l, \quad 0 < t \le T. \tag{8.54}$$

We leave the boundary condition (8.50) and the initial condition (8.52) unchanged:

$$k(x)\frac{\partial u_\alpha}{\partial x}(0,t) = 0, \qquad 0 < t \le T, \tag{8.55}$$

$$u_\alpha(x,0) = 0, \qquad 0 \le x \le l. \tag{8.56}$$

We replace the boundary condition (8.53), which makes the inverse problem (8.49), (8.50), (8.52), (8.53) an ill-posed problem, with the following non-local condition:

$$u_\alpha(0,t) + \alpha u_\alpha(l,t) = \varphi(t), \qquad 0 < t \le T. \tag{8.57}$$

In (8.54)–(8.57), the passage to a non-local boundary value problem can be made immediately. A second possibility in formulating such a non-classical problem is based on the consideration of the Tikhonov regularization method for problem (8.54)–(8.57) interpreted as a boundary control problem (the boundary condition at the right end point is (8.51)) with boundary observation (at the left end point, the condition (8.53) is adopted). Next, we can try to formulate a related Euler equation, which, as we saw, leads to non-classical boundary value problems. What is necessary is to only take the fact into account that in the case of interest both for the ground and conjugate states we have evolutionary problems with non-selfadjoint operators.

8.2.3 Local regularization

As it was repeated over and over again, in the application of regularization methods to evolutionary problems we have two possibilities. In global regularization methods the solution is to be determined at all times simultaneously, whereas in local regularization methods the solution depends only on the pre-history, and can be determined sequentially at separate times. Local regularization methods take into account the specific feature of inverse problems for evolutionary problems in maximal possible measure.

Over time, we introduce the uniform grid

$$\bar{\omega}_\tau = \{t_n = n\tau,\ n = 0, 1, \dots, N_0,\ \tau N_0 = T\},$$

and let $u^n(x) = u(x,t_n)$. In the approximate solution of the inverse problem (8.54)–(8.57), we perform the transition to the next time layer using the purely implicit scheme

for the direct problem:

$$\frac{u^{n+1}-u^n}{\tau}-\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u^{n+1}}{\partial x}\Big)=0,\qquad 0<x<l,\quad n=0,1,\ldots,N_0-1, \tag{8.58}$$

$$k(x)\,\frac{\partial u^{n+1}}{\partial x}(0)=0,\qquad n=0,1,\ldots,N_0-1, \tag{8.59}$$

$$u^{n+1}(l)=v^{n+1},\qquad n=0,1,\ldots,N_0-1, \tag{8.60}$$

$$u^0(x)=0,\qquad 0\le x\le l. \tag{8.61}$$

Using, on each time layer, the Tikhonov regularization for determining the boundary condition at the right boundary (see (8.60)) implies minimization of the smoothing functional

$$J_\alpha(v^{n+1})=(u^{n+1}(0)-\varphi^{n+1})^2+\alpha(v^{n+1})^2. \tag{8.62}$$

Let us show that the minimization problem for the functional (8.62) under constraints (8.58)–(8.61) is in fact equivalent to solving the difference problem with non-local boundary conditions of type (8.57) on each time layer.

We represent the solution of problem (8.58)–(8.60) in the form

$$u^{n+1}(x)=z^{n+1}(x)+w^{n+1}(x). \tag{8.63}$$

In the latter representation, $z^{n+1}(x)$ is the solution of the difference problem

$$\frac{z^{n+1}-u^n}{\tau}-\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial z^{n+1}}{\partial x}\Big)=0,\qquad 0<x<l,\quad n=0,1,\ldots,N_0-1, \tag{8.64}$$

$$k(x)\,\frac{\partial z^{n+1}}{\partial x}(0)=0,\qquad n=0,1,\ldots,N_0-1, \tag{8.65}$$

$$z^{n+1}(l)=0,\qquad n=0,1,\ldots,N_0-1, \tag{8.66}$$

$$z^0(x)=0,\qquad 0\le x\le l. \tag{8.67}$$

Thereby, $z^{n+1}(x)$ is the solution of the direct problem with homogeneous condition of the first kind at the right boundary.

From (8.63) and (8.64)–(8.67), for $w^{n+1}(x)$ we obtain:

$$\frac{w^{n+1}}{\tau}-\frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial w^{n+1}}{\partial x}\Big)=0,\qquad 0<x<l,\quad n=0,1,\ldots,N_0-1, \tag{8.68}$$

$$k(x)\,\frac{\partial w^{n+1}}{\partial x}(0,t)=0,\qquad n=0,1,\ldots,N_0-1, \tag{8.69}$$

$$w^{n+1}(l)=v^{n+1},\qquad n=0,1,\ldots,N_0-1, \tag{8.70}$$

$$w^0(x)=0,\qquad 0\le x\le l. \tag{8.71}$$

Taking into account the linearity of the coefficient $k(x)$ and its independence of time, for the solution of the difference problem (8.68)–(8.71) we obtain the representation

$$w^{n+1}(x)=q(x)v^{n+1}, \tag{8.72}$$

in which $q(x)$ is the solution of the difference problem

$$\frac{q}{\tau} - \frac{d}{dx}\Big(k(x)\,\frac{dq}{dx}\Big) = 0, \qquad 0 < x < l, \tag{8.73}$$

$$k(x)\,\frac{dq}{dx}(0) = 0, \qquad q(l) = 1. \tag{8.74}$$

Substitution of (8.63) and (8.72) into (8.62) yields:

$$J_\alpha(v^{n+1}) = (z^{n+1}(0) + q(0)v^{n+1} - \varphi^{n+1})^2 + \alpha(v^{n+1})^2. \tag{8.75}$$

The minimum of (8.75) is attained at

$$(z^{n+1}(0) + q(0)v^{n+1} - \varphi^{n+1}))q(0) + \alpha v^{n+1} = 0. \tag{8.76}$$

Thereby, for the boundary condition at the right boundary we have:

$$v^{n+1} = q(0)\,\frac{\varphi^{n+1} - z^{n+1}(0)}{\alpha + q^2(0)}. \tag{8.77}$$

In the local regularization algorithm as applied to the solution of the inverse problem (8.54)–(8.57), the transition to the next time layer implies the solution of the direct problems (8.64)–(8.67) and (8.73), (8.74), the calculation of the constant v^{n+1} by formula (8.76), and the use of representation (8.63), (8.72) for the solution of the inverse problem.

Using the maximum principle for problem (8.64)–(8.67), we have $q(0) > 0$; hence, the condition (8.76) can be brought (see (8.70)) to the form

$$u^{n+1}(0) + \frac{\alpha}{q(0)}\,u^{n+1}(l) = \varphi^{n+1}. \tag{8.78}$$

Thus, the minimization problem for the functional (8.62) under constraints (8.58)–(8.61) is equivalent to the solution of the problem with the non-local boundary conditions (8.58), (8.59), (8.61), and (8.77). With all these taken into account, we can say that the local regularization algorithm is a discrete variant of the method with non-locally perturbed boundary conditions (8.54)–(8.57) for the approximate solution of the boundary value inverse problem (8.49), (8.50), (8.52), (8.53).

8.2.4 Difference non-local problem

To the differential problem with the non-local boundary condition (8.54)–(8.57), we put in correspondence a difference problem. Along the spatial variable, we introduce a uniform grid $\bar{\omega}$ with a grid size h over the interval $\bar{\Omega} = [0, l]$:

$$\bar{\omega} = \{x \mid x = x_i = ih,\ i = 0, 1, \dots, N,\ Nh = l\}.$$

In this grid, ω is the set of internal nodes, and $\partial\omega$ is the set of boundary nodes.

At the internal nodes, in using the purely implicit scheme we approximate equation (8.54) with the difference equation

$$\frac{y^{n+1} - y^n}{\tau} - (a y_{\bar{x}}^{n+1})_x = 0, \qquad x \in \omega, \quad n = 0, 1, \ldots, N_0 - 1, \tag{8.79}$$

with, for instance, $a(x) = k(x - 0.5h)$. The initial condition (8.56) yields:

$$y^0(x) = 0, \qquad x \in \omega. \tag{8.80}$$

The second-kind boundary condition (8.55) is approximated on the solutions of (8.54):

$$\frac{y_0^{n+1} - y_0^n}{\tau} - \frac{2}{h} a_1 \frac{y_1^{n+1} - y_0^{n+1}}{h} = 0, \qquad n = 0, 1, \ldots, N_0 - 1. \tag{8.81}$$

To the non-local boundary condition (8.57), we put in correspondence the non-local difference condition

$$y_0^{n+1} + \alpha y_N^{n+1} = \varphi^{n+1}, \qquad n = 0, 1, \ldots, N_0 - 1. \tag{8.82}$$

Realization of the difference scheme (8.79)–(8.82) implies solution of the three-point difference problem at each time step with the non-local boundary conditions (8.82). To this end, we can use some modification of the standard sweep algorithm. A second possibility was considered previously, in the discussion of non-local regularization; this possibility is related with the use of some special representation of the solution (see (8.63) and (8.72)).

We seek the solution of the difference problem (8.79)–(8.82) in the form

$$y^{n+1}(x) = z^{n+1}(x) + q(x) v^{n+1}, \qquad x \in \bar{\omega}. \tag{8.83}$$

Here, similarly to (8.64)–(8.67), the mesh function $z^n(x)$ is defined as the solution of the following direct problem:

$$\frac{z^{n+1} - y^n}{\tau} - (a z_{\bar{x}}^{n+1})_x = 0, \qquad x \in \omega, \quad n = 0, 1, \ldots, N_0 - 1, \tag{8.84}$$

$$z^0(x) = 0, \qquad x \in \omega, \tag{8.85}$$

$$\frac{z_0^{n+1} - y_0^n}{\tau} - \frac{2}{h} a_1 \frac{z_1^{n+1} - z_0^{n+1}}{h} = 0, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{8.86}$$

$$z_N^{n+1} = 0, \qquad n = 0, 1, \ldots, N_0 - 1. \tag{8.87}$$

The mesh function $q(x)$ (see (8.73), (8.74)) is defined as the solution of the boundary value problem

$$\frac{q}{\tau} - (a q_{\bar{x}})_x = 0, \qquad x \in \omega, \tag{8.88}$$

$$\frac{q_0}{\tau} - \frac{2}{h}\frac{q_1 - q_0}{h} = 0, \qquad q_N = 1. \tag{8.89}$$

Substitution of (8.83) into (8.82) yields:

$$v^{n+1} = \frac{\varphi^{n+1} - z_0^{n+1}}{\alpha + q_0}, \qquad n = 0, 1, \ldots, N_0 - 1. \tag{8.90}$$

For the solution of the difference problem (8.79)–(8.82) to be found, we have to solve two standard problems, problems (8.84)–(8.87) and (8.88), (8.89) and, then, find the function $v^n,\ n = 1, 2, \ldots, N$ by formula (8.90); subsequently, the sought solution is to be represented in the form (8.83).

8.2.5 Program

The above solution algorithm for the boundary value inverse problem (8.49), (8.50), (8.52), (8.53) based on a non-local perturbation of the boundary condition is realized in the program PROBLEM16.

Program PROBLEM16

```
C
C     PROBLEM16 - IDENTIFICATION OF THE BOUNDARY CONDITION
C                 ONE-DIMENSIONAL NON-STATIONARY PROBLEM
C                 NON-LOCAL DISTURBANCE OF THE BOUNDARY CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.01D0, N = 101, M = 21 )
      DIMENSION  X(N), Y(N), Z(N), Q(N)
     +          ,FI(M), FID(M), FIY(M), U(M), UA(M)
     +          ,A(N), B(N), C(N), F(N)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT ENDS OF THE GEGMENT;
C     N         - NUMBER OF GRID NODES OVER SPACE;
C     TMAX      - MAXIMAL TIME;
C     M         - NUMBER OF GRID NODES OVER TIME;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     FI(M)     - EXACT DIFFERENCE BOUNDARY CONDITION;
C     FID(M)    - DISTURBED DIFFERENCE BOUNDARY CONDITION;
C     U(M)      - EXACT SOLUTION OF THE INVERSE PROBLEM
C                 (BOUNDARY CONDITION);
C     UA(M)     - APPROXIMATE SOLUTION OF THE INVERSE PROBLEM;
C
      XL = 0.D0
      XR = 1.D0
      TMAX = 1.D0
C
      OPEN  ( 01, FILE='RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H = (XR - XL) / (N - 1)
```

```
      DO I = 1, N
        X(I) = XL + (I-1)*H
      END DO
      TAU = TMAX / (M-1)
C
C     DIRECT PROBLEM
C
C     BOUNDARY REGIME
C
      DO K = 1, M
        T    = (K-1)*TAU

        U(K) = AF(T, TMAX)
      END DO
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I)  = 0.D0
      END DO
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C     PURELY IMPLICIT SCHEME
C
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
          B(I) = 1.D0 / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          F(I) = Y(I) / TAU
        END DO
C
C     BOUNDARY CONDITION AT THE LEFT AND RIGHT ENDS
C
        B(1) = 2.D0 / (H*H)
        C(1) = B(1) + 1.D0 / TAU
        F(1) = Y(1) / TAU
        A(N) = 0.D0
        C(N) = 1.D0
        F(N) = U(K)
C
C     SOLUTION OF THE PROBLEM ON THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, F, Y, ITASK )
C
C      SOLUTION AT THE LEFT BOUNDARY
C
        FI(K)  = Y(1)
        FID(K) = FI(K)
      END DO
C
C     NOISE ADDITION TO THE SOLUTION OF THE BOUNDARY-VALUE PROBLEM
C
      DO K = 2, M
```

```
        FID(K) = FI(K) + 2.D0*DELTA*(RAND(0)-0.5D0)
      END DO
C
C     INVERSE PROBLEM
C
C     NON-LOCAL DISTURBANCE OF THE BOUNDARY CONDITION
C
C     AUXILIARY MESH FUNCTION
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
      DO I = 2, N-1
        A(I) = 1.D0 / (H*H)
        B(I) = 1.D0 / (H*H)
        C(I) = A(I) + B(I) + 1.D0 / TAU
        F(I) = 0.D0
      END DO
C
C     BOUNDARY CONDITION AT THE LEFT AND RIGHT END POINTS
C
      B(1) = 2.D0 / (H*H)
      C(1) = B(1) + 1.D0 / TAU
      F(1) = 0.D0
      A(N) = 0.D0
      C(N) = 1.D0
      F(N) = 1.D0
C
C     SOLUTION OF THE PROBLEM
C
      ITASK = 1
      CALL PROG3 ( N, A, C, B, F, Q, ITASK )
C
C     ITERATIVE PROCESS FOR THE REGULARIZATION PARAMETER
C
      IT = 0
      ITMAX = 100
      ALPHA = 0.001D0
      QQ = 0.75D0
  100 IT = IT + 1
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I)  = 0.D0
      END DO
      UA(1) = Y(N)
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C     PURELY IMPLICIT SCHEME
C
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
          B(I) = 1.D0 / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
```

```
          F(I) = Y(I) / TAU
        END DO
C
C     BOUNDARY CONDITION AT THE LEFT AND RIGHT END POINTS
C
        B(1) = 2.D0 / (H*H)
        C(1) = B(1) + 1.D0 / TAU
        F(1) = Y(1) / TAU
        A(N) = 0.D0
        C(N) = 1.D0
        F(N) = 0.D0
C
C     SOLUTION OF THE AUXILIARY PROBLEM ON THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, F, Z, ITASK )
C
C     SOLUTION AT THE RIGHT BOUNDARY
C
        UA(K) = (FID(K) - Z(1)) / (ALPHA + Q(1))
C
C     SOLUTION AT ALL NODES
C
        DO I = 1, N
          Y(I) = Z(I) + Q(I)*UA(K)
        END DO
      END DO
C
C     SOLUTION OF THE DIRECT PROBLEM WITH THE FOUND BOUNDARY CONDITION
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N
        Y(I)  = 0.D0
      END DO
      FIY(1) = Y(1)
C
C     NEXT TIME LAYER
C
      DO K = 2, M
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
        DO I = 2, N-1
          A(I) = 1.D0 / (H*H)
          B(I) = 1.D0 / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          F(I) = Y(I) / TAU
        END DO
C
C     BOUNDARY CONDITION AT THE LEFT AND RIGHT END POINTS
C
        B(1) = 2.D0 / (H*H)
        C(1) = B(1) + 1.D0 / TAU
        F(1) = Y(1) / TAU
        A(N) = 0.D0
        C(N) = 1.D0
        F(N) = UA(K)
C
```

```
C     SOLUTION OF THE PROBLEM ON THE NEXT TIME LAYER
C
        ITASK = 1
        CALL PROG3 ( N, A, C, B, F, Y, ITASK )
        FIY(K) = Y(1)
      END DO
C
C     CRITERION FOR THE EXIT FROM THE ITERATIVE PROCESS
C
      SUM = 0.D0
      DO K = 1, M
        SUM = SUM + (FIY(K) - FID(K))**2*TAU
      END DO
      SL2 = DSQRT(SUM)
C
      IF (IT.GT.ITMAX) STOP
      IF ( IT.EQ.1 ) THEN
        IND = 0

        IF ( SL2.LT.DELTA ) THEN
          IND = 1
          QQ = 1.D0/QQ
        END IF
        ALPHA = ALPHA*QQ
        GO TO 100
      ELSE
        ALPHA = ALPHA*QQ
        IF ( IND.EQ.0 .AND. SL2.GT.DELTA ) GO TO 100
        IF ( IND.EQ.1 .AND. SL2.LT.DELTA ) GO TO 100
      END IF
C
C     RECORDING OF CALCULATED DATA
C
      WRITE ( 01,* ) (U(K), K = 1,M)
      WRITE ( 01,* ) (FID(K), K = 1,M)
      WRITE ( 01,* ) (UA(K), K = 1,M)
      WRITE ( 01,* ) (FIY(K), K = 1,M)
      CLOSE ( 01 )
      STOP
      END
C
      DOUBLE PRECISION FUNCTION  AF ( T, TMAX )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      BOUNDARY CONDITION AT THE RIGHT BOUNDARY
C
      AF = 2.D0*T/TMAX
      IF (T.GT.(0.5D0*TMAX)) AF = 2.D0*(TMAX-T)/TMAX
C
      RETURN
      END
```

The program implements the algorithm with non-locally perturbed boundary conditions for problem (8.49), (8.50), (8.52), (8.53) with the coefficient $k(x) = \text{const} = 1$ and with the non-local perturbation parameter chosen from the discrepancy.

8.2.6 Computational experiments

The problem was solved using a uniform grid with $h = 0.01$ and $\tau = 0.05$ in the calculation domain with $l = 1$ and $T = 1$. The input data in the inverse problem were taken from the solution of the direct problem with the right-end boundary condition

$$\psi(t) = \begin{cases} 2t/T, & 0 < t < T/2, \\ 2(T-t)/T, & T/2 < t < T. \end{cases}$$

The same model problem was considered above, when we discussed the algorithm with continuation over the spatial variable.

The data calculated with the various input-data inaccuracy levels are shown in Figures 8.6–8.8. A comparison with the data obtained using the algorithm with continuation over the spatial variable (see Figures 8.2, 8.4, 8.5) shows that the algorithm with non-locally perturbed boundary condition is inferior in terms of data accuracy because it poorly takes into account the specific features of the boundary value inverse problems of interest. With this approach, we hardly can count on time filtration of high-frequency inaccuracies because here, in fact, we have regularization with respect to the spatial variable.

In many respects, the effect due to the regularization is provided at the expense of cruder calculation grids along time (self-regularization effect). An illustration here are the data calculated with a finer grid along time (see Figure 8.9). The input-data inaccuracies can be most distinctly identified by reducing the time step size. The latter can be explicitly traced considering the formula for the solution at the right boundary, since $q_0 \to 0$ as $\tau \to 0$.

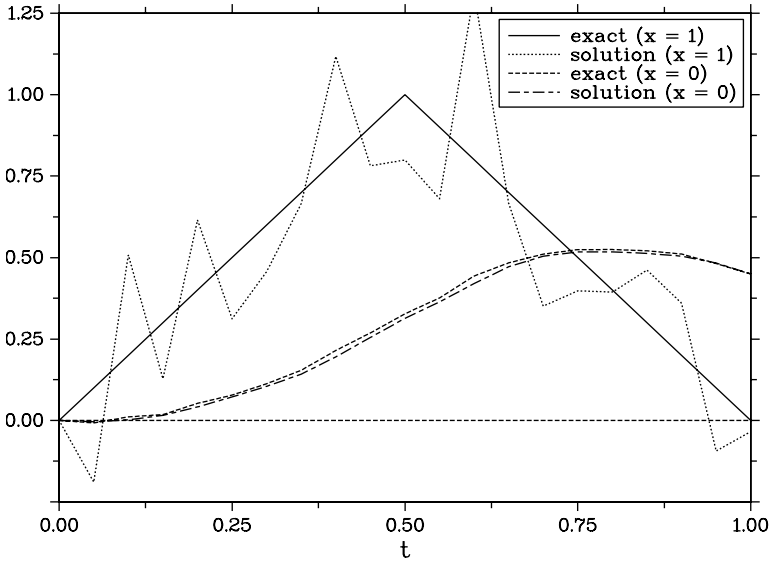

Figure 8.6 Inverse-problem solution obtained with $\delta = 0.01$

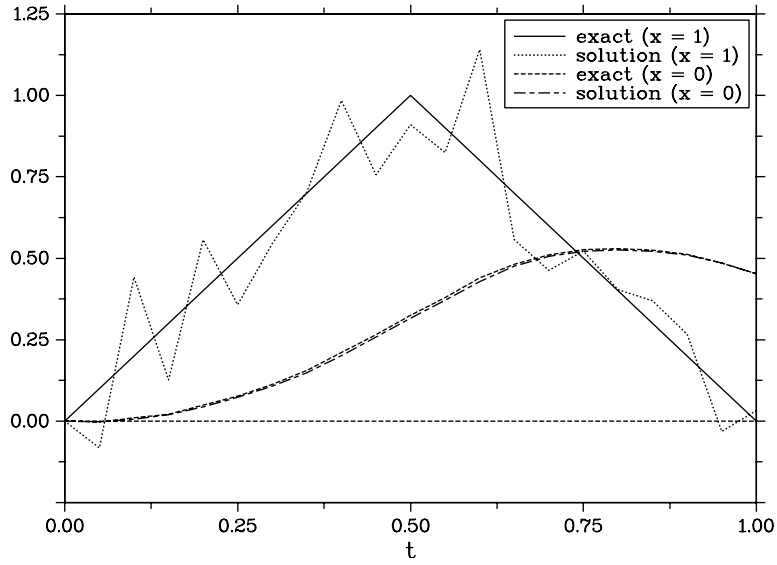

Figure 8.7 Inverse-problem solution obtained with $\delta = 0.005$

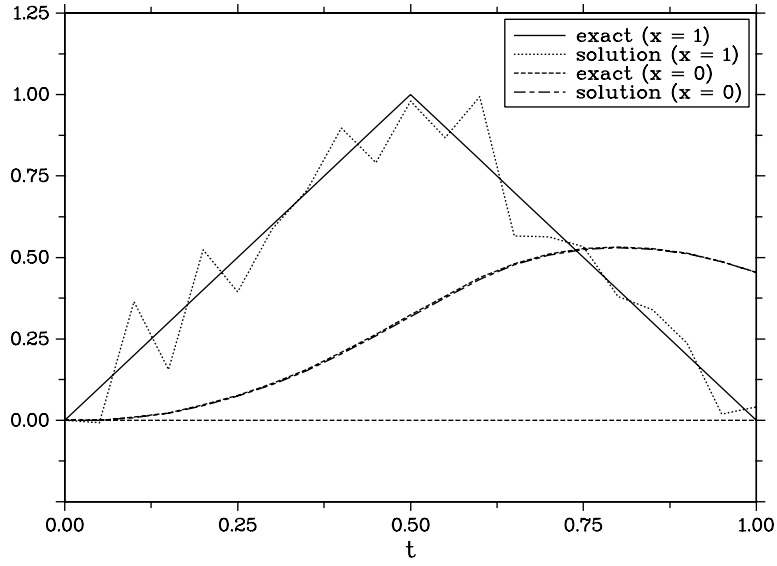

Figure 8.8 Inverse-problem solution obtained with $\delta = 0.0025$

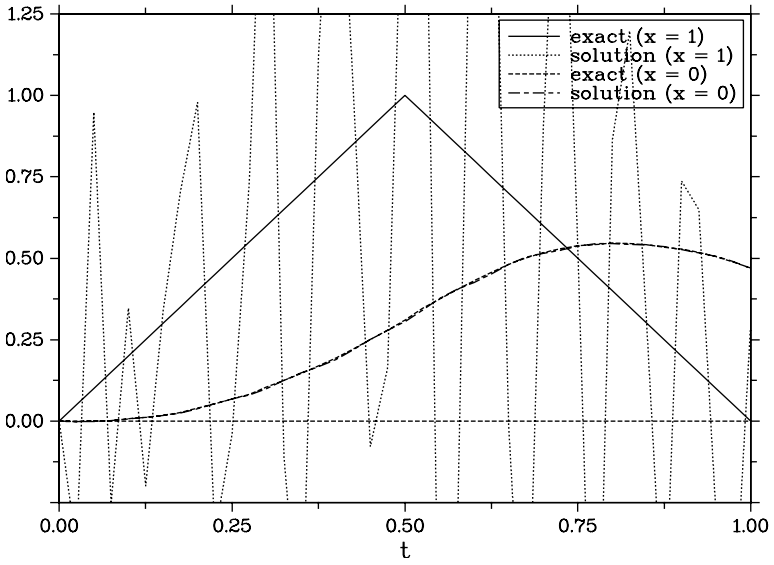

Figure 8.9 Solutions obtained with $\delta = 0.0025$ and $\tau = 0.025$

8.3 Identification of the boundary condition in the two-dimensional problem

In this section, we consider the boundary-layer inverse problem for the two-dimensional parabolic equation of second order. From data on some portion of the boundary, it is required to reconstruct the boundary data on the other portion of the boundary. To approximately solve the problem, we use the iteration method. Primary attention is given to accurate formulation of the symmetrized operator equation of the first order at the differential and mesh levels.

8.3.1 Statement of the problem

To most powerful methods intended for the approximate solution of inverse problems for mathematical physics equations, iteration methods belong. These methods rather adequately take into account the general specific features of the problems. Very often, the correct use of such methods is hampered by the necessity to perform certain analytical work primarily related with the fact that symmetrization of the corresponding operator equation of the first order is necessary. Similar problems are encountered in the formulation of necessary minimum conditions in the Tikhonov regularization method as applied to related optimum control problems for systems governed by mathematical physics equations. Here, the matter of obtaining the symmetrized operator equation is considered both on the differential level and on the mesh level.

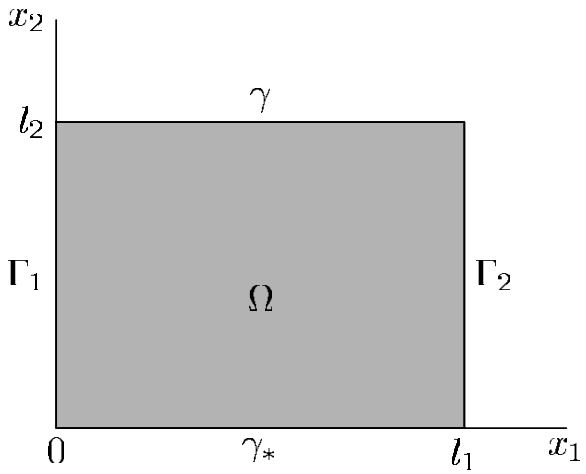

Figure 8.10 Calculation domain

As a model one, consider the two-dimensional problem in the rectangle

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\beta < l_\beta,\ \beta = 1, 2\}.$$

For the sides of Ω, we use the following notations (see Figure 8.10):

$$\partial\Omega = \gamma_* \cup \Gamma_1 \cup \gamma \cup \Gamma_2, \qquad \Gamma = \Gamma_1 \cup \Gamma_2.$$

In Ω, we seek the solution of the parabolic equation

$$\frac{\partial u}{\partial t} - \sum_{\beta=1}^{2} \frac{\partial}{\partial x_\beta}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\beta}\Big) = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T. \tag{8.91}$$

We assume that $k(x) \geq \kappa$, $\kappa > 0$, $\boldsymbol{x} \in \Omega$.

The starting point in the present consideration is the direct initial-boundary value problem for equation (8.91), in which the boundary and initial conditions look as

$$k(\boldsymbol{x})\,\frac{\partial u}{\partial n}(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \partial\gamma_*, \tag{8.92}$$

$$k(\boldsymbol{x})\,\frac{\partial u}{\partial n}(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \Gamma, \tag{8.93}$$

$$k(\boldsymbol{x})\,\frac{\partial u}{\partial n}(\boldsymbol{x}, t) = \mu(x_1, t), \qquad \boldsymbol{x} \in \gamma, \tag{8.94}$$

$$u(\boldsymbol{x}, 0) = 0, \qquad \boldsymbol{x} \in \Omega. \tag{8.95}$$

Consider an inverse problem in which it is required to identify the boundary condition on some portion of the boundary (on γ). In the latter case, instead of (8.94) the following condition is given:

$$u(\boldsymbol{x}, t) = \varphi(x_1, t), \qquad \boldsymbol{x} \in \gamma_*. \tag{8.96}$$

To approximately solve the boundary value inverse problem (8.91)–(8.93), (8.95), (8.96), we will use the iteration method related with refinement of a boundary condition of type (8.94).

8.3.2 Iteration method

We write the boundary value inverse problem as an operator equation of the first kind. To the boundary condition (8.94) we put in correspondence equation (8.96), i. e.,

$$\mathcal{A}\mu = \varphi. \tag{8.97}$$

The linear operator $\mathcal{A}$ is defined for functions given on some portion of the boundary γ, and its values are functions given on the other portion of the boundary (on γ_*). To calculate the values of $\mathcal{A}$, we solve the boundary value problem (8.91)–(8.95).

To approximately solve the ill-posed problem (8.97), we employ an iteration method based on the passage to a problem with a symmetric operator, where, instead of (8.97), we solve the equation

$$\mathcal{A}^*\mathcal{A}\mu = \mathcal{A}^*\varphi. \tag{8.98}$$

In using the explicit iteration method for (8.98), we have:

$$\frac{\mu_{k+1} - \mu_k}{s_{k+1}} - \mathcal{A}^*\mathcal{A}\mu_k = \mathcal{A}^*\varphi, \qquad k = 0, 1, \ldots. \tag{8.99}$$

For non-stationary functions given on γ and γ_*, we define the Hilbert spaces $\mathcal{L}(\gamma, [0, T])$ and $\mathcal{L}(\gamma_*, [0, T])$ with the scalar products and norms

$$(u, v) = \int_0^T \int_\gamma u(\boldsymbol{x})v(\boldsymbol{x})\, d\boldsymbol{x}\, dt, \qquad \|u\| = \sqrt{(u, u)},$$

$$(u, v)_* = \int_0^T \int_{\gamma_*} u(\boldsymbol{x})v(\boldsymbol{x})\, d\boldsymbol{x}\, dt, \qquad \|u\|_* = \sqrt{(u, u)_*},$$

respectively. In the steepest descend method, the iteration parameters can be calculated by the rule

$$s_{k+1} = \frac{\|r_k\|^2}{\|\mathcal{A}r_k\|_*^2}, \qquad r_k = \mathcal{A}^*\mathcal{A}\mu_k - \mathcal{A}^*\varphi. \tag{8.100}$$

Let us dwell on a point of primary concern in the practical use of (8.99) related with the necessity to calculate the values of the operator conjugate in $\mathcal{A}$. For functions μ and ν given respectively on γ and γ_* we have:

$$(\mathcal{A}\mu, \nu)_* = (\mu, \mathcal{A}^*\nu). \tag{8.101}$$

We obtain the values of $\mathcal{A}$ as

$$\mathcal{A}\mu = u(\boldsymbol{x}, t), \qquad \boldsymbol{x} \in \gamma_*.$$

Here, $u(\boldsymbol{x}, t)$ is the solution of the boundary value problem (8.91)–(8.95) (the function $u(\boldsymbol{x}, t)$ defines the ground state of the system). The values of the conjugate operator can be found from the function $\psi(\boldsymbol{x}, t)$, to be found from the solution of an auxiliary boundary value problem (conjugate state). For this boundary value problem to be formulated, we multiply equation (8.91) through by $\psi(\boldsymbol{x}, t)$, and then perform integration over Ω and time:

$$\int_0^T \int_\Omega \Big(\frac{\partial u}{\partial t} + \mathcal{L}u \Big) \psi \, d\boldsymbol{x} \, dt = 0, \tag{8.102}$$

where

$$\mathcal{L}u = -\sum_{\beta=1}^{2} \frac{\partial}{\partial x_\beta} \Big(k(\boldsymbol{x}) \frac{\partial u}{\partial x_\beta} \Big).$$

For the first term in (8.102), we have

$$I_1 = \int_0^T \int_\Omega \frac{\partial u}{\partial t} \psi \, d\boldsymbol{x} \, dt = -\int_0^T \int_\Omega u \frac{\partial \psi}{\partial t} \, d\boldsymbol{x} \, dt, \tag{8.103}$$

provided that, in view of (8.95),

$$\psi(\boldsymbol{x}, T) = 0, \qquad \boldsymbol{x} \in \Omega. \tag{8.104}$$

Similar manipulations with the second term in (8.102), performed with due regard for the boundary conditions (8.92)–(8.94), yield:

$$\begin{aligned} I_2 &= \int_0^T \int_\Omega \mathcal{L}u\psi \, d\boldsymbol{x} \, dt \\ &= -\int_0^T \int_\gamma \mu\psi \, d\boldsymbol{x} \, dt + \int_0^T \int_{\partial\Omega} uk \frac{\partial \psi}{\partial n} \, d\boldsymbol{x} \, dt + \int_0^T \int_\Omega u\mathcal{L}\psi \, d\boldsymbol{x} \, dt. \end{aligned} \tag{8.105}$$

We choose the boundary conditions for the conjugate state in the form

$$k(\boldsymbol{x}) \frac{\partial \psi}{\partial n}(\boldsymbol{x}, t) = \nu(x_1, t), \qquad \boldsymbol{x} \in \partial\gamma_*, \tag{8.106}$$

$$k(\boldsymbol{x}) \frac{\partial \psi}{\partial n}(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \Gamma, \tag{8.107}$$

$$k(\boldsymbol{x}) \frac{\partial u}{\partial \psi}(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \gamma. \tag{8.108}$$

With such boundary conditions, substitution of (8.103), (8.105) yields the equality

$$\int_0^T \int_\Omega u \Big(-\frac{\partial \psi}{\partial t} + \mathcal{L}\psi \Big) \, d\boldsymbol{x} \, dt - \int_0^T \int_\gamma \mu\psi \, d\boldsymbol{x} \, dt + \int_0^T \int_{\gamma_*} u\nu \, d\boldsymbol{x} \, dt = 0. \tag{8.109}$$

We assume that the function $\psi(\boldsymbol{x}, t)$ satisfies the equation

$$-\frac{\partial \psi}{\partial t} + \mathcal{L}\psi = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T. \tag{8.110}$$

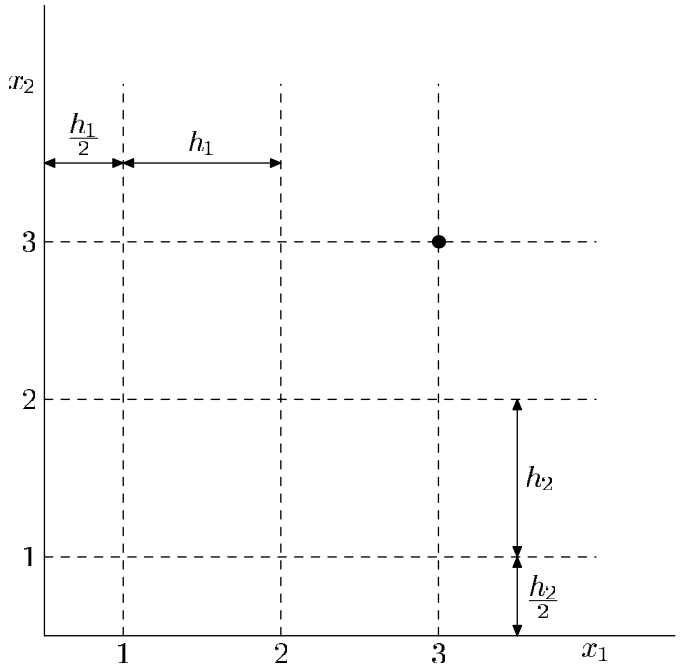

Figure 8.11 Calculation grid over space

Thereby, the conjugate state is to be found as the solution of the direct problem (8.104), (8.106)–(8.108), (8.110).

Under the present conditions, from (8.109) we obtain:

$$\int_0^T \int_\gamma \mu\psi \, d\boldsymbol{x} \, dt = \int_0^T \int_{\gamma_*} u\nu \, d\boldsymbol{x} \, dt.$$

The latter equality can be brought to the form (8.101), and for the values of $\mathcal{A}^*$ we obtain the representation

$$\mathcal{A}^*\nu = \psi(\boldsymbol{x}, t), \qquad \boldsymbol{x} \in \gamma,$$

where $\psi(\boldsymbol{x}, t)$ is the solution of the boundary value problem (8.104), (8.106)–(8.108), (8.110).

8.3.3 Difference problem

Let us start the present consideration with the formulation of a difference problem for the direct problem (8.91)–(8.95). Taking the fact into account that on the calculation-domain boundary second-kind boundary conditions are given, we use a grid whose nodes are displaced by half the step size from the domain boundary (see Figure 8.11).

Along each of the directions x_β, $\beta = 1, 2$, the grid is uniform, and let

$$\omega_\beta = \{x_\beta \mid x_\beta = (i_\beta - 1/2)h_\beta,\ i_\beta = 1, 2, \ldots, N_\beta,\ N_\beta h_\beta = l_\beta\}.$$

For the grid in the rectangle Ω we use the settings $\omega = \omega_1 \times \omega_2$.

With regard to the boundary conditions (8.92)–(8.94), to the differential operator $\mathcal{L}$ we put in correspondence a two-dimensional difference operator defined as the sum of one-dimensional operators:

$$\Lambda = \Lambda_1 + \Lambda_2. \tag{8.111}$$

With the grid displaced by half the grid size, it seems reasonable to define the difference operator Λ_1 as follows:

$$\Lambda_1 y = \begin{cases} -a_1(x_1 + h_1, x_2)y_{x_1}/h_1, & x_1 = h_1/2, \\ -(a_1 y_{\bar{x}_1})_{x_1}, & h_1/2 < x_1 < l_1 - h_1/2, \\ a_1 y_{\bar{x}_1}/h_1, & x_1 = l_1 - h_1/2, \end{cases}$$

where, for instance, $a_1(x_1, x_2) = k(x_1 - 0.5h_1, x_2)$. In a similar way, we define the difference operator Λ_2:

$$\Lambda_2 y = \begin{cases} -a_2(x_1, x_2 + h_2)y_{x_2}/h_2, & x_2 = h_2/2, \\ -(a_2 y_{\bar{x}_2})_{x_2}, & h_2/2 < x_2 < l_2 - h_2/2, \\ a_2 y_{\bar{x}_2}/h_2, & x_2 = l_2 - h_2/2 \end{cases}$$

with $a_2(x_1, x_2) = k(x_1, x_2 - 0.5h_2)$.

In the Hilbert space $L_2(\omega)$ of mesh functions set on ω, we define the scalar product as

$$(y, v)_\omega = \sum_{\boldsymbol{x} \in \omega} y(\boldsymbol{x})v(\boldsymbol{x})h_1 h_2.$$

It is an easy matter to check that the difference operator Λ defined by (8.111) is a self-adjoint operator non-negative in $L_2(\omega)$, i.e. $\Lambda = \Lambda^* \geq 0$.

We denote as y_n the difference solution at the time $t_n = n\tau$, where $\tau > 0$ is the time step size ($N_0\tau = T$). An approximate solution of the direct problem (8.91)–(8.95) can be found as the solution of the difference problem

$$\frac{y_{n+1} - y_n}{\tau} + \Lambda(\sigma y_{n+1} + (1 - \sigma)y_n) = f_n, \qquad n = 0, 1, \ldots, N_0 - 1, \tag{8.112}$$

$$y_0 = 0, \qquad \boldsymbol{x} \in \omega. \tag{8.113}$$

The weighted difference scheme (8.112), (8.113) is absolutely stable in the case of $\sigma \geq 1/2$. For simplicity, we restrict ourselves to the case of the purely implicit scheme ($\sigma = 1$), in which for the solution of the difference problem (8.112), (8.113) there holds the following stability estimate:

$$\|y_{n+1}\| \leq \|y_n\| + \tau\|f_n\|, \qquad n = 0, 1, \ldots, N_0 - 1.$$

The inhomogeneity of the right-hand side of (8.112) results from the boundary conditions (8.94), so that in the case of $\sigma = 1$ we can put:

$$f_n(\boldsymbol{x}) = \begin{cases} 0, & x_1 \in \omega_1,\ x_2 = h_2/2, \\ 0, & x_1 \in \omega_1,\ h_2/2 < x_2 < l_2 - h_2/2, \\ \mu_{n+1}(x_1)/h_2, & x_1 \in \omega_1,\ x_2 = l_2 - h_2/2. \end{cases} \tag{8.114}$$

Thus, the boundary conditions are included on the operator level into the difference equation. With the form of the inhomogeneous right-hand side taken into account, the above estimate proves that the solution of the difference problem is stable with respect to the initial data and boundary conditions.

Let us formulate now the boundary value inverse problem on the mesh level. In the case under consideration (see (8.91)–(8.93), (8.95), (8.96)) the function $\mu_n(x_1)$, $n = 1, 2, \ldots, N_0$ in the right-hand side of (8.114) is not given. Instead of this function, the solution at the nodes adjacent to the boundary γ_* is assumed to be known:

$$y_n(x_1, 0.5h_2) = \varphi_n(x_1), \qquad x_1 \in \omega_1, \quad n = 1, 2, \ldots, N_0. \tag{8.115}$$

Hence, on the mesh level we can speak about a problem in which it is required to identify the right-hand side of special form (8.114) from boundary observations (8.115).

8.3.4 Iterative refinement of the boundary condition

Similarly to the continuous case, here we use the formulation of the boundary value inverse problem in the form of a first-kind operator equation. From the given boundary condition (right-hand side set in the form (8.114) with known mesh function $\mu_n(x_1)$, $n = 1, 2, \ldots, N_0$), we put in correspondence the function $\varphi_n(x_1)$, $n = 1, 2, \ldots, N_0$ (see (8.115)):

$$A\mu = \varphi. \tag{8.116}$$

To calculate $\varphi_n(x_1)$, $n = 1, 2, \ldots, N_0$, we have to solve the difference boundary value problem (8.112)–(8.114). The operator equation (8.116) is the difference analogue to (8.97).

The iteration method is to be applied to the symmetrized equation (see (8.98))

$$A^*A\mu = A^*\varphi. \tag{8.117}$$

The explicit iteration method for the approximate solution of (8.117) can be written in the form

$$\frac{\mu^{k+1} - \mu^k}{s_{k+1}} + A^*A\mu^k = A^*\varphi, \qquad k = 0, 1, \ldots\ . \tag{8.118}$$

The matter of calculating the values of A^* deserves special consideration. To this end, first of all we have to give mesh analogues of the Hilbert spaces $\mathcal{L}(\gamma, [0, T])$ and

$\mathcal{L}(\gamma_*, [0, T])$, to be subsequently referred to as H and H_*, respectively. In H and H_*, the scalar product and the norm are defined as

$$(u, v) = \sum_{n=1}^{N_0} \sum_{x_1 \in \omega_1} u_n(x_1, l_2) v_n(x_1, l_2) h_1 \tau, \qquad \|u\| = \sqrt{(u, u)},$$

$$(u, v)_* = \sum_{n=1}^{N_0} \sum_{x_1 \in \omega_1} u_n(x_1, 0) v_n(x_1, 0) h_1 \tau, \qquad \|u\|_* = \sqrt{(u, u)_*}.$$

Then, in the steepest descend method, for the iteration parameters in (8.118) we have:

$$s_{k+1} = \frac{\|r^k\|^2}{\|Ar^k\|_*^2}, \qquad r^k = A^* A\mu^k - A^*\varphi. \tag{8.119}$$

The operator adjoint to A can be found from the equality

$$(A\mu, v)_* = (\mu, A^* v). \tag{8.120}$$

Here, the mesh functions $\mu_n(x_1)$, $n = 1, 2, \ldots, N_0$ and $v_n(x_1)$, $n = 1, 2, \ldots, N_0$ are given in the calculation domain ω at the nodes adjacent to the boundaries γ and γ_*, respectively.

To formulate the difference problem for the conjugate state, we multiply the difference equation for the ground state

$$\frac{y_{n+1} - y_n}{\tau} + \Lambda y_{n+1} = f_n, \tag{8.121}$$

by the mesh function $\psi_{n+1} h_1 h_2 \tau$ and sum it up over all nodes $\boldsymbol{x} \in \omega$ and all $n = 0, 1, \ldots, N_0 - 1$:

$$\sum_{n=1}^{N_0} \sum_{\boldsymbol{x} \in \omega} \Big(\frac{y_n - y_{n-1}}{\tau} + \Lambda y_n \Big) \psi_n h_1 h_2 \tau = \sum_{n=1}^{N_0} \sum_{\boldsymbol{x} \in \omega} f_{n-1} \psi_n h_1 h_2 \tau. \tag{8.122}$$

With regard to (8.114) and with regard to the introduced notation, for the right-hand side of (8.122) we immediately obtain:

$$\sum_{n=1}^{N_0} \sum_{\boldsymbol{x} \in \omega} f_{n-1} \psi_n h_1 h_2 \tau = (\mu, \psi). \tag{8.123}$$

As a result, we can relate this right-hand side with the right-hand side in (8.120).

Using the initial condition (8.113), we can rearrange the first term in the left-hand side in (8.122) as

$$I_1 = \sum_{n=1}^{N_0} \sum_{\boldsymbol{x} \in \omega} \frac{y_n - y_{n-1}}{\tau} \psi_n h_1 h_2 \tau = - \sum_{n=1}^{N_0} \sum_{\boldsymbol{x} \in \omega} \frac{\psi_{n+1} - \psi_n}{\tau} y_n h_1 h_2 \tau, \tag{8.124}$$

provided that the following additional condition holds:

$$\psi_{N_0+1} = 0, \qquad \boldsymbol{x} \in \omega. \tag{8.125}$$

With regard to the self-adjointness of Λ and $L_2(\omega)$, for the second term in the right-hand side of (8.122) we have:

$$I_2 = \sum_{n=1}^{N_0} \sum_{\boldsymbol{x}\in\omega} \Lambda y_n \psi_n h_1 h_2 \tau = \sum_{n=1}^{N_0} \sum_{\boldsymbol{x}\in\omega} y_n \Lambda \psi_n h_1 h_2 \tau. \tag{8.126}$$

Substitution of (8.123), (8.124), and (8.126) into (8.122) yields:

$$\sum_{n=1}^{N_0} \sum_{\boldsymbol{x}\in\omega} y_n \Big(- \frac{\psi_{n+1} - \psi_n}{\tau} + \Lambda \psi_n \Big) h_1 h_2 \tau = (\mu, \psi). \tag{8.127}$$

We choose the mesh function $\psi_n(\boldsymbol{x})$ so that the left side of (8.127) is related with the left side in (8.120).

We assume that the function $\psi_n(\boldsymbol{x})$ satisfies the difference equation

$$-\frac{\psi_{n+1} - \psi_n}{\tau} + \Lambda \psi_n = g_n, \qquad n = 1, 2, \ldots, N_0 \tag{8.128}$$

supplemented with the initial condition (8.125). Hence, with the given right-hand side $g_n(\boldsymbol{x})$ the conjugate state is to be found using the purely implicit scheme (8.125), (8.128) on the grid displaced by the time step size.

From (8.127) and (8.128), we obtain

$$\sum_{n=1}^{N_0} \sum_{\boldsymbol{x}\in\omega} y_n g_n h_1 h_2 \tau = (\mu, \psi). \tag{8.129}$$

In view of (8.115) and (8.116), the left-hand side of (8.129) coincides with the left-hand side in (8.120) provided that the right-hand side in (8.128) is given as

$$g_n(\boldsymbol{x}) = \begin{cases} \nu_n(x_1)/h_2, & x_1 \in \omega_1, \ x_2 = h_2/2, \\ 0, & x_1 \in \omega_1, \ h_2/2 < x_2 < l_2 - h_2/2, \\ 0, & x_1 \in \omega_1, \ x_2 = l_2 - h_2/2. \end{cases} \tag{8.130}$$

This choice of the right-hand side corresponds to the solution of the difference boundary value problem with boundary conditions of the second kind (see (8.106)).

From the equality between the right-hand sides of (8.120) and (8.129), for the values of A^* we obtain:

$$A^* \nu_n = \psi_n(x_1, l_2), \qquad x_1 \in \omega_1, \quad n = 1, 2, \ldots, N_0. \tag{8.131}$$

Here, the mesh function $\psi_n(\boldsymbol{x})$ is to be found as the solution of the difference problem (8.125), (8.128), (8.130).

8.3.5 Program realization

The program presented below embodies the above iteration solution method for the boundary value inverse problem (8.91)–(8.93), (8.95), (8.96) in the case of $k(x) = 1$. The iteration method is terminated based on discrepancy.

Program PROBLEM17

```
C
C     PROBLEM17 - BOUNDARY-VALUE INVERSE PROBLEM
C                 TWO-DIMENSIONAL PROBLEM
C                 ITERATIVE REFINEMENT OF THE BOUNDARY CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.005D0, N1 = 50, N2 = 50, M = 101 )
      DIMENSION A(9*N1*N2), X1(N1), X2(N2), T(M)
     +          ,Y(N1,N2), F(N1,N2), G(N1)
     +          ,FI(N1,M), FID(N1,M), FS(N1,M), FIK(N1,M)
     +          ,U(N1,M),UA(N1,M), UK(N1,M), R(N1,M), AR(N1,M)
      COMMON / SB5 /     IDEFAULT(4)
      COMMON / CONTROL / IREPT, NITER
C
C     PARAMETERS:
C
C     X1L, X2L  - COORDINATES OF THE LEFT CORNER;
C     X1R, X2R  - COORDINATES OF THE RIGHT CORNER;
C     N1, N2    - NUMBER OF NODES IN THE SPATIAL GRID;
C     H1, H2    - MESH SIZE OVER SPACE;
C     TAU       - TIME STEP;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     FI(N1,M)  - EXACT DIFFERENCE BOUNDARY CONDITION;
C     FID(N1,M) - DISTURBED DIFFERENCE BOUNDARY CONDITION;
C     U(N1,M)   - EXACT SOLUTION OF THE INVERSE PROBLEM
C                 (BOUNDARY CONDITION);
C     UA(N1,M)  - APPROXIMATE SOLUTION OF THE INVERSE PROBLEM;
C     EPSR      - RELATIVE SOLUTION INACCURACY IN THE DIFFERENCE
C                 PROBLEM;
C     EPSA      - ABSOLUTE SOLUTION INACCURACY IN THE DIFFERENCE
C                 PROBLEM;
C
C     EQUIVALENCE ( A(1),         A0            ),
C    *            ( A(N+1),       A1            ),
C    *            ( A(2*N+1),     A2            ),
C
      X1L    = 0.D0
      X1R    = 1.D0
      X2L    = 0.D0
      X2R    = 0.5D0
      TMAX   = 1.D0
      PI     = 3.1415926D0
      EPSR   = 1.D-5
      EPSA   = 1.D-8
C
      OPEN (01, FILE='RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H1  = (X1R-X1L) / N1
```

```
      H2  = (X2R-X2L) / N2
      TAU = TMAX / (M-1)
      DO I = 1, N1
        X1(I) = X1L + (I-0.5D0)*H1
      END DO
      DO J = 1, N2
        X2(J) = X2L + (J-0.5D0)*H2

      END DO
      DO K = 1, M
        T(K) = (K-1)*TAU
      END DO
C
      N  = N1*N2
      DO I = 1, 9*N
        A(I) = 0.0
      END DO
C
C     DIRECT PROBLEM
C     PURELY IMPLICIT DIFFERENCE SCHEME
C
C     INITIAL CONDITION
C
      DO I = 1, N1
        DO J = 1, N2
          Y(I,J) = 0.D0
        END DO
      END DO
C
C     BOUNDARY CONDITION
C
      CALL FLUX (U, X1, T, N1, M)
      DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        DO I = 1, N1
          G(I) = U(I,K)
        END DO
        CALL FDS (A(1), A(N+1), A(2*N+1), F, Y,
     +            H1, H2, N1, N2, TAU, G)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), Y, F, EPSR, EPSA)
C
C     INPUT DATA FOR THE INVERSE PROBLEM
C
        DO I = 1, N1
          FI(I,K) = Y(I,1)
        END DO
      END DO
C
C     DISTURBING OF MEASURED QUANTITIES
C
      DO I = 1, N1
        DO K = 2, M
          FID(I,K) = FI(I,K) + 2.*DELTA*(RAND(0)-0.5)
```

```
        END DO
      END DO
C
C     INVERSE PROBLEM
C
C     RIGHT SIDE OF THE SYMMETRIZED EQUATION
C
C     DIRECT PROBLEM
C     PURELY IMPLICIT DIFFERENCE SCHEME
C
C     INITIAL CONDITION (AT t = T)
C
      DO I = 1, N1
        DO J = 1, N2
          Y(I,J) = 0.D0
        END DO
      END DO
      DO K = M,2,-1
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        DO I = 1, N1
          G(I) = FID(I,K)
        END DO
        CALL FDSS (A(1), A(N+1), A(2*N+1), F, Y,
     +             H1, H2, N1, N2, TAU, G)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), Y, F, EPSR, EPSA)
C
C     RIGHT-HAND SIDE
C
        DO I = 1, N1
          FS(I,K) = Y(I,N2)
        END DO
      END DO
C
C     ITERATION METHOD
C
      IT = 0
C
C     INITIAL APPROXIMATION
C
      DO I = 1, N1
        DO K = 2, M
          UK(I,K) = 0.D0
        END DO
      END DO
C
  100 IT = IT + 1
C
C     GROUND STATE
C
C     INITIAL CONDITION
C
      DO I = 1, N1
        DO J = 1, N2
```

```
            Y(I,J) = 0.D0
          END DO
        END DO
        DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
          DO I = 1, N1
            G(I) = UK(I,K)
          END DO
          CALL FDS (A(1), A(N+1), A(2*N+1), F, Y,
     +              H1, H2, N1, N2, TAU, G)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          IDEFAULT(1) = 0
          IREPT       = 0
          CALL SBAND5 (N, N1, A(1), Y, F, EPSR, EPSA)
C
C     INPUT DATA FOR THE CONJUGATE PROBLEM
C
          DO I = 1, N1
            FIK(I,K) = Y(I,1)
          END DO
        END DO
C
C     CONJUGATE STATE
C     PURELY IMPLICIT DIFFERENCE SCHEME
C
C     INITIAL CONDITION (AT t = T)
C
        DO I = 1, N1
          DO J = 1, N2
            Y(I,J) = 0.D0
          END DO
        END DO
        DO K = M,2,-1
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
          DO I = 1, N1
            G(I) = FIK(I,K)
          END DO
          CALL FDSS (A(1), A(N+1), A(2*N+1), F, Y,
     +              H1, H2, N1, N2, TAU, G)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
          IDEFAULT(1) = 0
          IREPT       = 0
          CALL SBAND5 (N, N1, A(1), Y, F, EPSR, EPSA)
C
C     DISCREPANCY
C
          DO I = 1, N1
            R(I,K) = Y(I,N2) - FS(I,K)
          END DO
        END DO
C
C     QUICKEST DESCEND METHOD
```

```
C
C     AUXILIARY PROBLEM
C
C     INITIAL CONDITION
C
      DO I = 1, N1
        DO J = 1, N2
          Y(I,J) = 0.D0
        END DO
      END DO
      DO K = 2, M
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
        DO I = 1, N1
          G(I) = R(I,K)
        END DO
        CALL FDS (A(1), A(N+1), A(2*N+1), F, Y,
     +             H1, H2, N1, N2, TAU, G)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
        IDEFAULT(1) = 0
        IREPT       = 0
        CALL SBAND5 (N, N1, A(1), Y, F, EPSR, EPSA)
        DO I = 1, N1
          AR(I,K) = Y(I,1)
        END DO
      END DO
C
C     ITERATION PARAMETER
C
      SUM1 = 0.D0
      SUM2 = 0.D0
      DO I = 1, N1
        DO K = 2, M
          SUM1 = SUM1 + R(I,K)*R(I,K)
          SUM2 = SUM2 + AR(I,K)*AR(I,K)
        END DO
      END DO
      SS = SUM1/SUM2
C
C     NEXT APPROXIMATION
C
      DO I = 1, N1
        DO K = 2, M
          UK(I,K) = UK(I,K) - SS*R(I,K)
        END DO
      END DO
C
C     EXIT FROM THE ITERATIVE PROCESS BY THE DISCREPANCY CRITERION
C
      SUM = 0.D0
      DO I = 1, N1
        DO K = 2, M
          SUM = SUM + (FID(I,K) - FIK(I,K))**2*H1*TAU
        END DO
      END DO
      SUM = SUM/((X1R-X1L)*TMAX)
      SL2 = DSQRT(SUM)
```

```
      IF ( SL2.GT.DELTA ) GO TO 100
C
C     SOLUTION
C
      WRITE ( 01, * ) ((UK(I,K), I=1,N1), K=2,M)
      WRITE ( 01, * ) ((FID(I,K), I=1,N1), K=2,M)
      CLOSE (01)
      STOP
      END
C
      SUBROUTINE FLUX (U, X1, T, N1, M)
C
C     BOUNDARY CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  U(N1,M), X1(N1), T(M)
      DO I = 1, N1
        DO K = 1, M
          U(I,K) = X1(I)*T(K)/T(M)
C          IF (T(K).GT.0.5D0*T(M)) U(I,K) = 2.D0*X1(I)*(T(M)-T(K))
        END DO
      END DO
C
      RETURN
      END
C
      SUBROUTINE FDS (A0, A1, A2, F, U, H1, H2, N1, N2, TAU, G)
C
C     GENERATION OF DIFFERENCE-SCHEME COEFFICIENTS
C     FOR THE PARABOLIC EQUATION WITH CONSTANT COEFFICIENTS
C     IN USING THE PURELY IMPLICIT SCHEME
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  A0(N1,N2), A1(N1,N2), A2(N1,N2)
     +          ,F(N1,N2), U(N1,N2), G(N1)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          A1(I-1,J) = 1.D0/(H1*H1)
          A1(I,J)   = 1.D0/(H1*H1)
          A2(I,J-1) = 1.D0/(H2*H2)
          A2(I,J)   = 1.D0/(H2*H2)
          A0(I,J)   = A1(I,J) + A1(I-1,J) + A2(I,J) + A2(I,J-1)
     +              + 1.D0/TAU
          F(I,J)    = U(I,J)/TAU
        END DO
      END DO
C
C     BOUNDARY CONDITION OF THE SECOND KIND
C
      DO J = 2, N2-1
        A2(1,J)   = 1.D0/(H2*H2)
        A2(1,J-1) = 1.D0/(H2*H2)
        A0(1,J)   = A1(1,J) + A2(1,J) + A2(1,J-1)+ 1.D0/TAU
        F(1,J)    = U(1,J)/TAU
      END DO
C
      DO J = 2, N2-1
        A2(N1,J)   = 1.D0/(H2*H2)
        A2(N1,J-1) = 1.D0/(H2*H2)
```

```
        A0(N1,J)   = A1(N1-1,J) + A2(N1,J) + A2(N1,J-1)+ 1.D0/TAU
        F(N1,J)    = U(N1,J)/TAU
      END DO
C
      DO I = 2, N1-1
        A1(I,1)   = 1.D0/(H1*H1)
        A1(I-1,1) = 1.D0/(H1*H1)
        A0(I,1)   = A1(I,1) + A1(I-1,1) + A2(I,1) + 1.D0/TAU
        F(I,1)    = U(I,1)/TAU
      END DO
C
      DO I = 2, N1-1
        A1(I,N2)   = 1.D0/(H1*H1)
        A1(I-1,N2) = 1.D0/(H1*H1)
        A0(I,N2)   = A1(I,N2) + A1(I-1,N2) + A2(I,N2-1) + 1.D0/TAU
        F(I,N2)    = U(I,N2)/TAU + G(I)/H2
      END DO
C
       A0(1,1) = A1(1,1) + A2(1,1) + 1.D0/TAU
       F(1,1)  = U(1,1)/TAU
C
       A0(N1,1) = A1(N1-1,1) + A2(N1,1) + 1.D0/TAU
       F(N1,1)  = U(N1,1)/TAU
C
       A0(1,N2) = A1(1,N2) + A2(1,N2-1) + 1.D0/TAU
       F(1,N2)  = U(1,N2)/TAU + G(1)/H2
C
       A0(N1,N2) = A1(N1-1,N2) + A2(N1,N2-1) + 1.D0/TAU
       F(N1,N2)  = U(N1,N2)/TAU + G(N1)/H2
C
      RETURN
      END
C
      SUBROUTINE FDSS (A0, A1, A2, F, U, H1, H2, N1, N2, TAU, G)
C
C     GENERATION OF DIFFERENCE-SCHEME COEFFICIENTS
C     FOR THE CONJUGATE STATE
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  A0(N1,N2), A1(N1,N2), A2(N1,N2)
     +          ,F(N1,N2), U(N1,N2), G(N1)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          A1(I-1,J) = 1.D0/(H1*H1)
          A1(I,J)   = 1.D0/(H1*H1)
          A2(I,J-1) = 1.D0/(H2*H2)
          A2(I,J)   = 1.D0/(H2*H2)
          A0(I,J)   = A1(I,J) + A1(I-1,J) + A2(I,J) + A2(I,J-1)
     +              + 1.D0/TAU
          F(I,J)    = U(I,J)/TAU
        END DO
      END DO
C
C     BOUNDARY CONDITION OF THE SECOND KIND
C
      DO J = 2, N2-1
        A2(1,J)   = 1.D0/(H2*H2)
        A2(1,J-1) = 1.D0/(H2*H2)
        A0(1,J)   = A1(1,J) + A2(1,J) + A2(1,J-1)+ 1.D0/TAU
```

```
        F(1,J)    = U(1,J)/TAU
      END DO
C
      DO J = 2, N2-1
        A2(N1,J)   = 1.D0/(H2*H2)
        A2(N1,J-1) = 1.D0/(H2*H2)
        A0(N1,J)   = A1(N1-1,J) + A2(N1,J) + A2(N1,J-1)+ 1.D0/TAU
        F(N1,J)    = U(N1,J)/TAU
      END DO
C
      DO I = 2, N1-1
        A1(I,1)   = 1.D0/(H1*H1)
        A1(I-1,1) = 1.D0/(H1*H1)
        A0(I,1)   = A1(I,1) + A1(I-1,1) + A2(I,1) + 1.D0/TAU
        F(I,1)    = U(I,1)/TAU + G(I)/H2
      END DO
C
      DO I = 2, N1-1
        A1(I,N2)   = 1.D0/(H1*H1)
        A1(I-1,N2) = 1.D0/(H1*H1)

        A0(I,N2)   = A1(I,N2) + A1(I-1,N2) + A2(I,N2-1) + 1.D0/TAU
        F(I,N2)    = U(I,N2)/TAU
      END DO
C
       A0(1,1) = A1(1,1) + A2(1,1) + 1.D0/TAU
       F(1,1)  = U(1,1)/TAU + G(1)/H2
C
       A0(N1,1) = A1(N1-1,1) + A2(N1,1) + 1.D0/TAU
       F(N1,1)  = U(N1,1)/TAU + G(N1)/H2
C
       A0(1,N2) = A1(1,N2) + A2(1,N2-1) + 1.D0/TAU
       F(1,N2)  = U(1,N2)/TAU
C
       A0(N1,N2) = A1(N1-1,N2) + A2(N1,N2-1) + 1.D0/TAU
       F(N1,N2)  = U(N1,N2)/TAU
C
      RETURN
      END
```

In the subroutine `FLUX`, boundary conditions of the second kind for the test direct problem (exact inverse-problem solution) are introduced. In the subroutines `FDS` and `FDSS`, difference-scheme coefficients for the ground and conjugate states are generated, respectively.

8.3.6 Computational experiments

The input data for the inverse problem are obtained as the solution of the direct problem (8.91)–(8.95) with the boundary condition (8.94) given as

$$\mu(x_1, t) = x_1 \frac{t}{T}.$$

The direct problem is being solved in the rectangle Ω with $l_1 = 1,\ l_2 = 0.5$ at $T = 1$ on the grid ω with $N_1 = N_2 = 50$ and $\tau = 0.01$. The solution of the direct problem at

the times $t = 0.25, 0.5, 0.75, 1$ are shown in Figures 8.12–8.15.

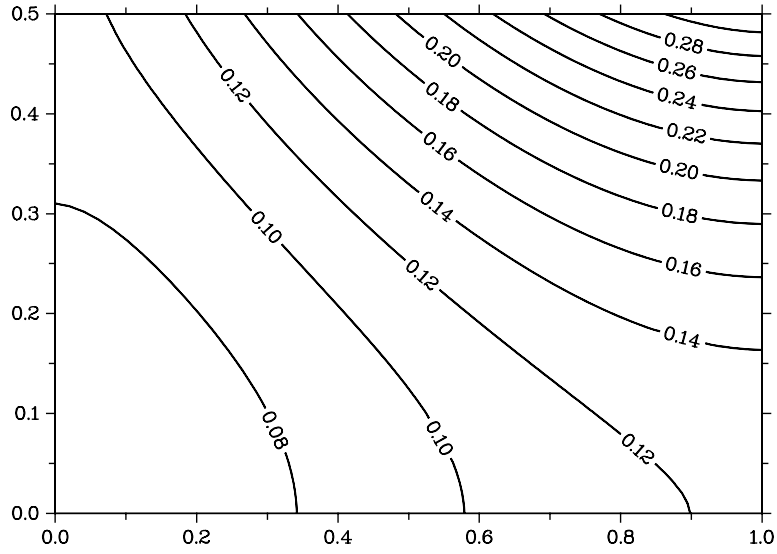

Figure 8.12 Direct-problem solution ($t = 0.25$)

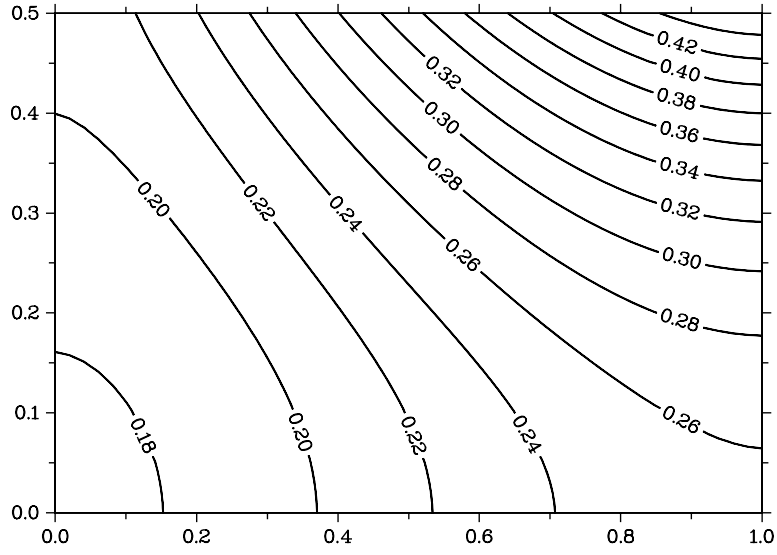

Figure 8.13 Direct-problem solution ($t = 0.5$)

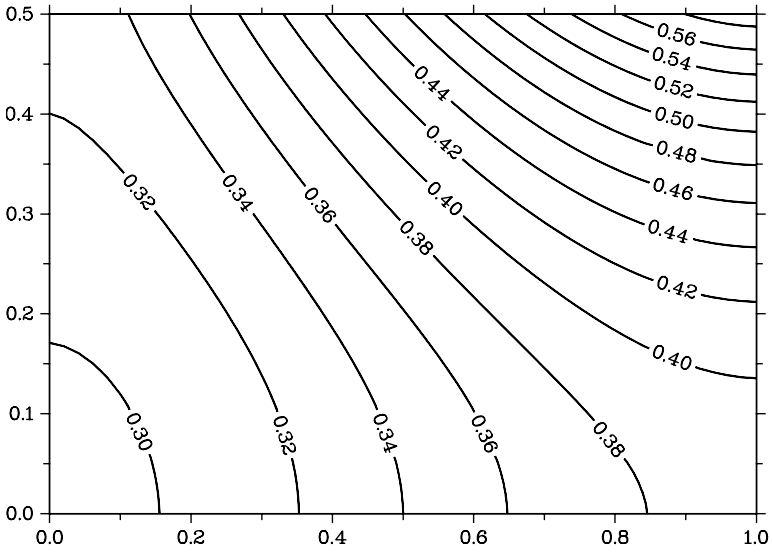

Figure 8.14 Direct-problem solution ($t = 0.75$)

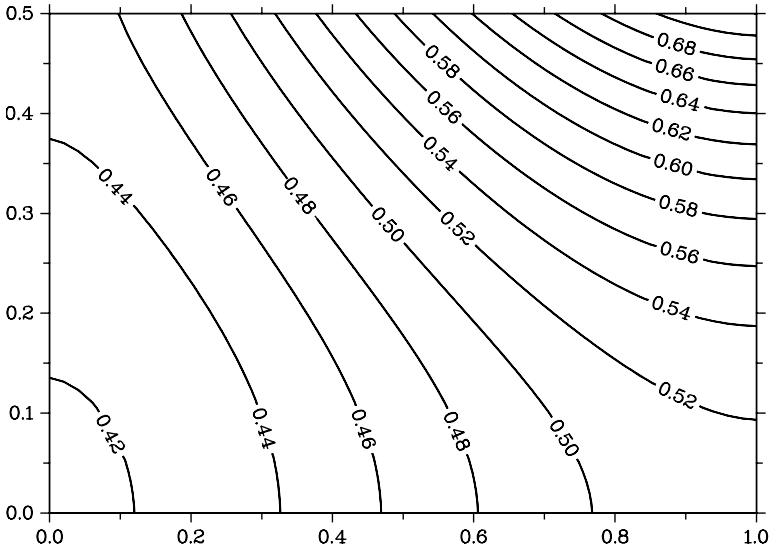

Figure 8.15 Direct-problem solution ($t = 1$)

The found solution of the direct problem at the grid nodes closest to γ_* (the functions $\varphi_n(x_1)$ in (8.115)) at the same times are shown in Figure 8.16. These data were randomly perturbed to be subsequently considered as input data for the inverse problem:

$$\bar{\varphi}_n(x_1) = \varphi_n(x_1) + 2\delta(\sigma(x_1, t_n) - 1/2), \qquad x_1 \in \omega, \quad n = 1, 2, \ldots, N_0.$$

Here, $\sigma(x_1, t_n)$ is a random function distributed normally over the interval [0, 1]. The approximate solution of the inverse problem obtained with the inaccuracy level defined by the parameter $\delta = 0.005$ is shown in Figure 8.17. The same figure shows the exact and found boundary conditions at the times $t = 0.25, 0.5, 0.75$, and 1. The end-time reconstruction accuracy (at $t = 1$) is seen to be very poor. In fact, here there is no other thing to expect since the changes in the boundary condition at times around $t = T$ do not affect the solution at the observation points (on the other portion of the boundary): the perturbations do not arrive at the observation points. The effect due to the accuracy in setting the input data can be figured out considering Figures 8.18–8.19.

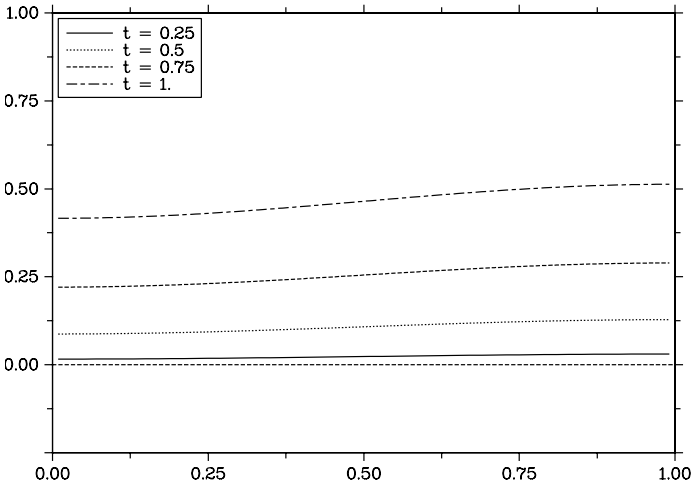

Figure 8.16 Unperturbed initial data for the inverse problem

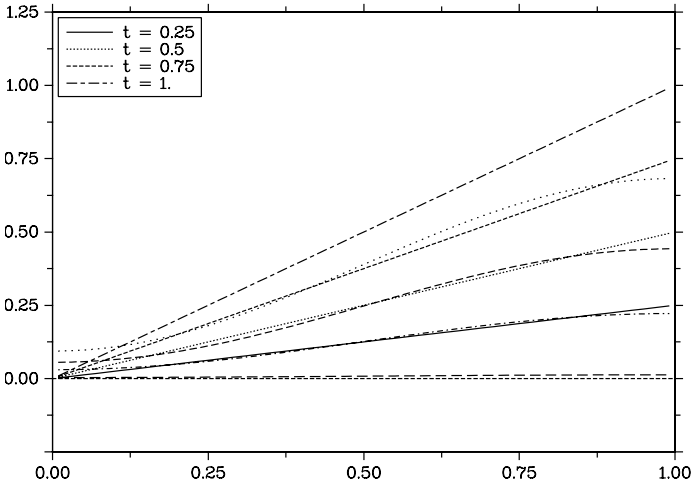

Figure 8.17 Inverse-problem solution obtained with $\delta = 0.005$

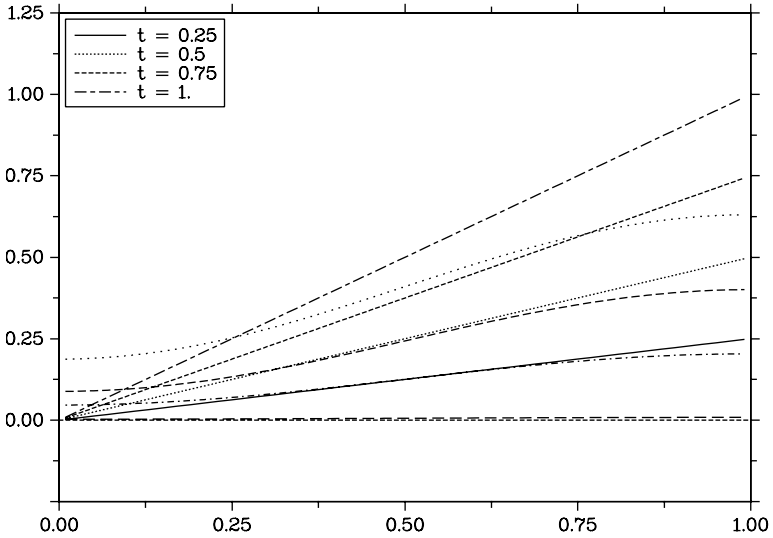

Figure 8.18 Inverse-problem solution obtained with $\delta = 0.01$

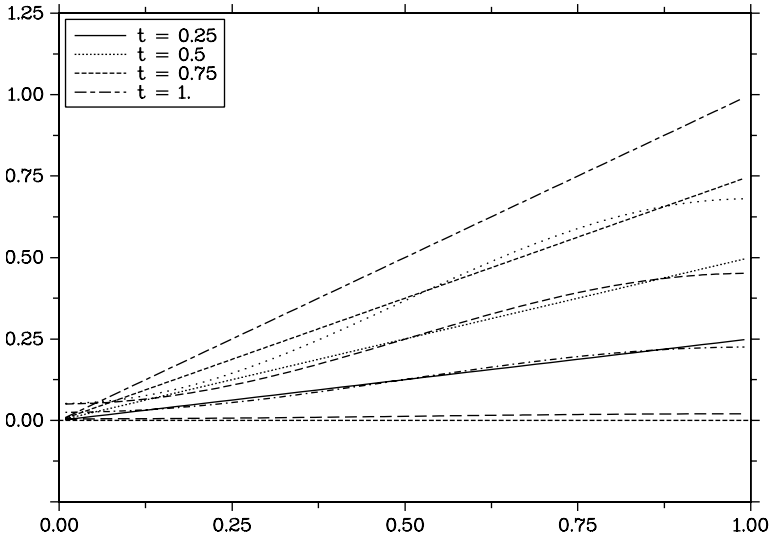

Figure 8.19 Inverse-problem solution obtained with $\delta = 0.0025$

8.4 Coefficient inverse problem for the nonlinear parabolic equation

In this section, we consider the inverse problem in which it is required to determine a coefficient that depends on the solution of the one-dimensional parabolic equation from the solution observed at some internal point/points. Algorithms of functional and

parametric optimization are discussed. Primary attention is paid to the construction and realization of computational algorithms.

8.4.1 Statement of the problem

In many applied problems, there arises a problem in which it is required to identify coefficients in a partial differential equation. Coefficient inverse problems for second-order parabolic equations are typical of heat- and mass-transfer problems and problems encountered in hydrogeology. The inverse problems in which it is required to identify coefficients in linear equations are nonlinear problems. The latter circumstance substantially hampers the construction of computational algorithms for the approximate solution of coefficient problems and makes complete and rigorous substantiation of their convergence hardly possible. That is why the emphasis here is placed on maximum possible approbation of numerical methods aimed at obtaining most informative solution examples for inverse problems.

Very often, of primary interest are problems in which it is required to find nonlinear coefficients that depend on the solution. Let us formulate a simplest of such a problem. Suppose that in the rectangle

$$\overline{Q}_T = \overline{\Omega} \times [0, T], \qquad \overline{\Omega} = \{x \mid 0 \le x \le l\}, \quad 0 \le t \le T$$

the function $u(x, t)$ satisfies the equation

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\Big(k(u)\,\frac{\partial u}{\partial x}\Big) = 0, \qquad 0 < x < l, \quad 0 < t \le T. \tag{8.132}$$

We assume that $k(u) \ge \kappa > 0$. Consider the boundary value problem with the first-kind boundary conditions

$$u(0, t) = 0, \qquad u(l, t) = g(t), \qquad 0 < t \le T \tag{8.133}$$

and the homogeneous initial conditions

$$u(x, 0) = 0, \qquad 0 \le x \le l. \tag{8.134}$$

The direct problem is formulated in the form (8.132)–(8.134).

In the coefficient inverse problem, the unknown function $k(u)$ is to be found, for instance, from additional observations performed at some internal points $z_m \in \Omega$, $m = 1, 2, \ldots, M$. Considering the fact that these data are given inaccurate, we put

$$u(z_m, t) \approx \varphi_m(t), \qquad 0 < t \le T, \quad m = 1, 2, \ldots, M. \tag{8.135}$$

It is required to find the functions $u(x, t)$ and $k(u)$ from the conditions (8.132)–(8.135).

In the consideration of coefficient inverse problems similar to problem (8.132)–(8.135), much attention is paid to problems of solution unicity in the case of exact

measurements. In the case at hand, under the assumptions that the coefficient $k(u)$ and the solution itself are smooth functions, it suffices to additionally demand that the function $g(t)$ be a monotonic function. For definiteness, we assume that

$$\frac{dg}{dt}(t) > u_{\min} = 0, \quad 0 < t < T, \qquad g(0) = 0, \quad g(T) = u_{\max}. \tag{8.136}$$

In the inverse problem (8.132)–(8.135), we can pose a problem in which it is required to find the functional relation $k(u)$ in the case of $u_{\min} \le u \le u_{\max}$.

8.4.2 Functional optimization

In the approximate solution of the inverse problem (8.132)–(8.135), we can use the variational formulation of this problem. We use the settings

$$K = L_2(u_{\min}, u_{\max}), \quad (k, r)_K = \int_{u_{\min}}^{u_{\max}} k(u) r(u) du, \quad \|k\|_K = \sqrt{(k, k)_K}.$$

In gradient iteration methods, we have to minimize the discrepancy functional that, with regard to (8.135), looks as

$$J(k) = \sum_{m=1}^{M} \int_0^T (u(z_m, t; k) - \varphi_m(t))^2 \, dt \tag{8.137}$$

under the conditions (8.132)–(8.134). In the Tikhonov regularization method, to be regularized is the smoothing functional

$$J_\alpha(k) = \sum_{m=1}^{M} \int_0^T (u(z_m, t; k) - \varphi_m(t))^2 \, dt + \alpha \|k\|_K^2.$$

In using gradient iteration methods for the minimization of the functional $J(k)$, we have to calculate the gradient of the functional. Some difficulties arise owing to the fact that the functional here is not quadratic. The gradient of $J'(k)$ referring to the increment δk is given by

$$\delta J(k) = (J'(k), \delta k)_K + s,$$

where $\delta J(k) = J(k + \delta k) - J(k)$ is the functional increment and

$$\frac{|s|}{\|\delta k\|_K} \to 0 \quad \text{for} \quad \delta k \to 0.$$

As we saw previously, the gradient of the discrepancy functional can be expressed through the solution of some conjugate initial-boundary value problem. The most general approach to the formulation of this problem is related with the consideration of a problem on conditional minimization of the discrepancy functional on the solutions

of the boundary value problem for the ground state as a problem on unconditional minimization by introducing Lagrange multipliers. As applied to the minimization problem (8.137) under constraints (8.132)–(8.134), the Lagrange functional has the form

$$G(k) = J(k) + \int_0^T \int_0^l \psi\Big(\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\Big(k(u)\,\frac{\partial u}{\partial x}\Big)\Big)\,dx\,dt, \tag{8.138}$$

where $\psi(x, t)$ is the Lagrange multipliers.

Let δG be the increment of the functional G corresponding to the increment δk and

$$\delta G = \delta J + \delta Q. \tag{8.139}$$

We denote as δu the increment of u; then, we have for δJ:

$$\delta J = 2\sum_{m=1}^{M} \int_0^T \int_0^l \delta u(u - \varphi_m)\delta(x - z_m)\,dx\,dt, \tag{8.140}$$

where $\delta(x)$ is the δ-function. Neglecting the second-order terms in (8.139), for the second term in this formula we obtain

$$\delta Q = \int_0^T \int_0^l \psi\Big(\frac{\partial \delta u}{\partial t} - \frac{\partial}{\partial x}\Big(k(u)\,\frac{\partial \delta u}{\partial x}\Big) - \frac{\partial}{\partial x}\Big(\delta k\,\frac{\partial u}{\partial x}\Big)\Big)\,dx\,dt. \tag{8.141}$$

For the solution increments, from (8.133) and (8.134) we obtain:

$$\delta u(0, t) = 0, \qquad \delta u(l, t) = 0, \qquad 0 < t \le T,$$
$$\delta u(x, 0) = 0, \qquad 0 \le x \le l.$$

Hence, we have

$$\int_0^T \int_0^l \psi\,\frac{\partial \delta u}{\partial t}\,dx\,dt = -\int_0^T \int_0^l \delta u\,\frac{\partial \psi}{\partial t}\,dx\,dt,$$

provided that

$$\psi(0, T) = 0, \qquad 0 \le x \le l. \tag{8.142}$$

In a similar manner, on setting the boundary conditions for $\psi(x, t)$ in the form

$$\psi(0, t) = 0, \qquad \psi(l, t) = 0, \qquad 0 < t \le T \tag{8.143}$$

we arrive at the equality

$$\int_0^T \int_0^l \psi\,\frac{\partial}{\partial x}\Big(k(u)\,\frac{\partial \delta u}{\partial x}\Big)\,dx\,dt = \int_0^T \int_0^l \delta u\,\frac{\partial}{\partial x}\Big(k(u)\,\frac{\partial \psi}{\partial x}\Big)\,dx\,dt.$$

Besides, in the case of boundary conditions (8.143) we have

$$\int_0^T \int_0^l \psi\,\frac{\partial}{\partial x}\Big(\delta k\,\frac{\partial u}{\partial x}\Big)\,dx\,dt = -\int_0^T \int_0^l \delta k\,\frac{\partial u}{\partial x}\,\frac{\partial \psi}{\partial x}\,dx\,dt.$$

Now, we can rewrite expression (8.141) in the form

$$\delta Q = \int_0^T \int_0^l \delta u \Big(- \frac{\partial \psi}{\partial t} - \frac{\partial}{\partial x}\Big(k(u)\,\frac{\partial \psi}{\partial x}\Big)\Big)\,dx\,dt + \int_0^T \int_0^l \delta k\,\frac{\partial u}{\partial x}\,\frac{\partial \psi}{\partial x}\,dx\,dt. \quad (8.144)$$

With (8.140) and (8.144), we determine the function $\psi(x,t)$ as the solution of the equation

$$-\frac{\partial \psi}{\partial t} - \frac{\partial}{\partial x}\Big(k(u)\,\frac{\partial \psi}{\partial x}\Big) + 2\sum_{m=1}^{M}(u-\varphi_m)\delta(x-z_m) = 0, \qquad 0 < x < l, \qquad 0 \le t < T. \quad (8.145)$$

Thus, for the conjugate state to be found, we have to solve the well-posed initial-boundary problem (8.142), (8.143), (8.145).

From (8.139) and (8.144), we obtain

$$\delta G = \int_0^T \int_0^l \delta k\,\frac{\partial u}{\partial x}\,\frac{\partial \psi}{\partial x}\,dx\,dt.$$

With regard to (8.138), this increment of the functional can be expressed in terms of $J'(k)$:

$$\delta G = (J'(k), \delta k)_K.$$

In the calculation domain Q_T we introduce new independent variables $u(x,t)$ and $v(x,t)$, with the transformed Jacobian $D(x,t)/D(u,v) \neq 0$. With regard to our assumptions (8.136) concerning the boundary condition (function $g(t)$), such a transform is indeed possible. We therefore have:

$$\int_0^T \int_0^l \delta k\,\frac{\partial u}{\partial x}\,\frac{\partial \psi}{\partial x}\,dx\,dt = \int_{u_{\min}}^{u_{\max}} \delta k(u) \int_{v_1(u)}^{v_2(u)} \frac{\partial u}{\partial x}\,\frac{\partial \psi}{\partial x}\,\frac{D(x,t)}{D(u,v)}\,dv\,du$$

and, hence,

$$J'(k) = \int_{v_1(u)}^{v_2(u)} \frac{\partial u}{\partial x}\,\frac{\partial \psi}{\partial x}\,\frac{D(x,t)}{D(u,v)}\,dv, \qquad u_{\min} \le u \le u_{\max}. \quad (8.146)$$

The obtained representation (8.146) is not convenient for use in the development of practical iteration solution methods for the coefficient inverse problem (8.132)–(8.135) because, in this case, we have a rather complex computational procedure for the gradient of the discrepancy functional.

8.4.3 Parametric optimization

In the approximate solution of coefficient inverse problems, special attention should be paid to parametric identification methods. In the gradient methods discussed above the approximate solution is sought as a function of a continuous (or discrete, in the case of difference approximation) argument. That is why here we use the term "functional optimization". Yet, another approach is possible, which can be considered as the projection method for solving inverse problems. In this method, the approximate solution is represented in parametric form, and what is required is to find the parameters in this representation.

In a function space K, we choose a finite-difference subspace K_p with some basis $\eta_\beta(u)$, $\beta = 1, 2, \ldots, p$. In the approximate solution of the coefficient inverse problem (8.132)–(8.135), the coefficient to be found is represented in the form

$$k_p(u) = \sum_{\beta=1}^{p} a_\beta \eta_\beta(u). \tag{8.147}$$

In solving the inverse problem, the unknown coefficients a_β, $\beta = 1, 2, \ldots, p$ are to be found.

The parametric identification algorithm can be realized in two variants. We assume that the accuracy in setting the input information is defined by some quantity δ. Next, we assume that in the model problem (8.132)–(8.135)

$$u(z_m, t) = \varphi_m^\delta(t), \qquad 0 < t \le T, \quad m = 1, 2, \ldots, M, \tag{8.148}$$

$$\sum_{m=1}^{M} \int_0^T (\varphi_m^\delta(t) - \varphi_m(t))^2 \, dt \le MT\delta^2. \tag{8.149}$$

In solving the coefficient problem (8.132)–(8.134), (8.148), (8.149), the first variant of the parametric identification algorithm is related with using a sufficiently high dimensionality p of K_p, in which case the inaccuracy in the approximation of $k(u)$ with the function $k_p(u)$ results in much lower solution inaccuracies at observation points compared to initial inaccuracies (see (8.149)). In other words, the inaccuracies in (8.147) can be ignored in solving the inverse problem. Similar situation takes place in difference discretization of the inverse problem: here, with sufficiently fine calculation grids used, we neglect the inaccuracies generated by the discretization.

Like with functional identification, in the case under consideration the approximate solution (8.147) can be regularized by minimizing the Tikhonov smoothing functional for the vector $\boldsymbol{a} = \{a_1, a_2, \ldots, a_p\}$:

$$J_\alpha(\boldsymbol{a}) = \sum_{m=1}^{M} \int_0^T (u(z_m, t; \boldsymbol{a}) - \varphi_m(t))^2 \, dt + \alpha \sum_{\beta=1}^{p} a_\beta^2 \tag{8.150}$$

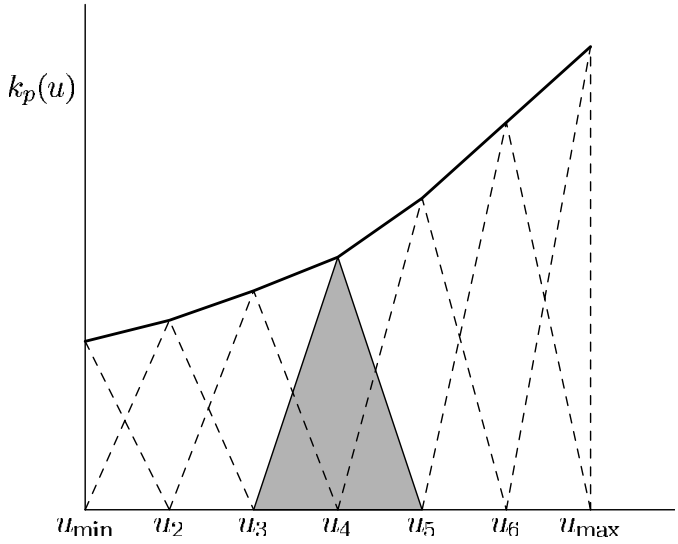

Figure 8.20 Piecewise-linear approximation

provided that the value of the regularization parameter α is properly matched with the input-data inaccuracy (i.e., with the value of δ in (8.149)). An alternative here is an iteration method for determining the vector $\boldsymbol{a}$.

In the second variant of the parametric identification algorithm, specific features of parametric identification are taken into account more fully. Here, the dimensionality of K_p, or the number of elements in the expansion (8.147), is used as the regularization parameter. Here, we can speak of self-regularizing properties exhibited by the discretization algorithm (8.147).

As a typical example of (8.147), consider piecewise linear approximation. We assume that a grid

$$u_\beta = u_{\min} + (\beta - 1)\,\frac{u_{\max} - u_{\min}}{p - 1}, \qquad \beta = 1, 2, \ldots, p.$$

is introduced, uniform over the variable u. In this case (see Figure 8.20), the piecewise linear finite functions $\eta_\beta(u)$, $\beta = 1, 2, \ldots, p$ are given in the form

$$\eta_\beta(u) = \begin{cases} 0, & u < u_{\beta-1}, \\ \dfrac{u - u_{\beta-1}}{u_\beta - u_{\beta-1}}, & u_{\beta-1} < u < u_\beta, \\ \dfrac{u_{\beta+1} - u}{u_{\beta+1} - u_\beta}, & u_{\beta-1} < u < u_\beta, \\ 0, & u > u_{\beta+1}, \end{cases} \qquad \beta = 2, 3, \ldots, p-1,$$

and the coefficients are $a_\beta = k_p(u_\beta)$, $\beta = 1, 2, \ldots, p$. In this representation of the

approximate solution, the mesh size over u or the total number p of nodes can be used as a regularization parameter.

In the use of the regularization method (8.150), the computational algorithms for parametric identification (8.147) are related with the minimization problem for the function p of the variable $J_\alpha(\boldsymbol{a})$. Let us formulate a necessary condition for a minimum that can be used to construct numerical algorithms for the parameters a_β, $\beta = 1, 2, \ldots, p$. Immediately from (8.150), we have:

$$\frac{\partial J_\alpha}{\partial a_\beta} = 2 \sum_{m=1}^{M} \int_0^T (u(z_m, t; \boldsymbol{a}) - \varphi_m(t)) \frac{\partial u}{\partial a_\beta}\, dt + 2\alpha a_\beta. \tag{8.151}$$

Relations (8.132)–(8.134) and representation (8.147) for $\partial u / \partial a_\beta$ yield the following boundary value problem:

$$\frac{\partial v}{\partial t} - \frac{\partial}{\partial x}\Big(k(u)\, \frac{\partial v}{\partial x}\Big) = \frac{\partial}{\partial x}\Big(\eta_\beta(u)\, \frac{\partial u}{\partial x}\Big), \qquad 0 < x < l, \qquad 0 < t \le T, \tag{8.152}$$

$$v(0, t) = 0, \qquad v(l, t) = 0, \qquad 0 < t \le T, \tag{8.153}$$

$$v(x, 0) = 0, \qquad 0 \le x \le l. \tag{8.154}$$

To obtain a more convenient representation for the first term in the right-hand side of (8.151), we formulate a boundary value problem for the conjugate state. Let the function $\psi(x, t)$ with some known $u(x, t)$ be the solution of the boundary value problem (8.142), (8.143), (8.145). We multiply equation (8.152) through by $\psi(x, t)$; then, integrating the resulting equation over x and t, by direct calculations we obtain the desired system of equations:

$$\int_0^T \int_0^l \eta_\beta(u)\, \frac{\partial u}{\partial x}\, \frac{\partial \psi}{\partial x}\, dx\, dt + 2\alpha a_\beta = 0, \qquad \beta = 1, 2, \ldots, p. \tag{8.155}$$

Around (8.155), we can construct the computational algorithms. In particular, we can use iteration methods for determining the parameters a_β, $\beta = 1, 2, \ldots, p$. Here, the complication owing to the nonlinear dependence of the ground state $u(x, t)$ on the sought coefficient $k(u) = k_p(u)$ must be taken into account.

Under monotonic conditions for the boundary conditions of type (8.136), successive identification algorithms can be constructed. A similar local regularization procedure was discussed above, in the consideration of evolutionary inverse problems in which it was required to reconstruct the initial condition and the boundary conditions. In the case of (8.132)–(8.136), at each time in the interval $t \le t_* < T$ we can find the relation $k(u)$ for $u \le g(t_*)$. Such specific features of the coefficient problem under consideration can most easily be taken into account in the case of parametric optimization of (8.147) in the class of piecewise constant functions. In the latter case (see also Figure 8.21), with the use of the uniform grid

$$u_\beta = u_{\min} + \beta\, \frac{u_{\max} - u_{\min}}{p}, \qquad \beta = 0, 1, \ldots, p$$

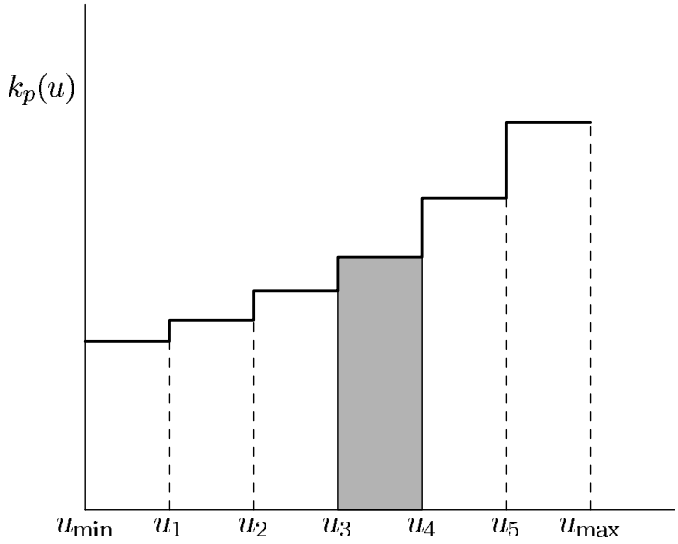

Figure 8.21 Piecewise-constant approximation

the trial functions are given in the form

$$\eta_\beta(u) = \begin{cases} 0, & u < u_{\beta-1}, \\ 1, & u_{\beta-1} \le u \le u_\beta, \\ 0, & u > u_\beta, \end{cases} \qquad \beta = 1, 2, \ldots, p. \tag{8.156}$$

In the case of the latter parameterization, with $u_{\beta-1} \le u \le u_\beta$ it is required to find just one numerical parameter a_β since the parameters a_ν, $\nu = 1, 2, \ldots, \beta - 1$ were found previously. Such a procedure is possible not only with piecewise constant completion of the unknown coefficient, but also in using other approximations of (8.147), for instance, with the piecewise linear approximation (see Figure 8.20).

8.4.4 Difference problem

Consider problems that arise when the parametric optimization algorithm in the variant of local regularization is used in the approximate solution of the coefficient inverse problem (8.132)–(8.135). Under the assumptions of (8.136), we solve the identification problem for the coefficient $k(u)$ in the class (8.147), (8.156).

We begin with constructing the difference analogue to the direct problem (8.132)–(8.134). Over time, we introduce the simplest uniform grid

$$\bar{\omega}_\tau = \omega_\tau \cup \{T\} = \{t_n = n\tau,\ n = 0, 1, \ldots, N_0,\ \tau N_0 = T\}$$

with some step size $\tau > 0$. To find the approximate solution by the time $t = t_n$, we use the notations $y_n(x) = y(x, t_n)$.

We complement the time-uniform grid with a grid non-uniform over u. With (8.133), (8.136), we define

$$g_n = g(t_n), \qquad n = 0, 1, \ldots, N_0,$$
$$u_{\min} = g_0 < g_1 < \cdots < g_{n-1} < g_n < \cdots < g_{N_0} = u_{\max}.$$

Using these data, one can construct a cruder grid over u:

$$\begin{gathered} u_\beta = u_{\beta-1} + N_0^{(\beta)}\tau, \\ u_0 = u_{\min}, \qquad u_p = u_{\max}, \qquad \beta = 1, 2, \ldots, p. \end{gathered} \tag{8.157}$$

In this way, between the nodes $u_{\beta-1}$ and u_β we make $N_0^{(\beta)}$ time steps.

We denote as $\bar{\omega}$ a uniform grid with a step size h over the interval $\bar{\Omega} = [0, l]$:

$$\bar{\omega} = \{x \mid x = x_i = ih,\ i = 0, 1, \ldots, N,\ Nh = l\}$$

Here, ω is the set of internal nodes, and $\partial\omega$ is the set of boundary nodes. We define the difference operator

$$A(v)y = -(d(v)y_{\bar{x}})_x, \qquad x \in \omega.$$

Then, we set the coefficient $d(v)$ in the form

$$d(v) = k(0.5(v(x) + v(x-h))), \qquad d(v) = 0.5(k(v(x-h)) + k(v(x))).$$

At the internal grid nodes over space, we put in correspondence to equation (8.132) the purely implicit difference scheme

$$\begin{gathered} \frac{y_{n+1} - y_n}{\tau} + A(y_{n+1})y_{n+1} = 0, \\ x \in \omega, \qquad n = 0, 1, \ldots, N_0 - 1. \end{gathered} \tag{8.158}$$

The approximation of the boundary conditions (8.133) yields

$$y_{n+1}(0) = 0, \qquad y_{n+1}(l) = g_{n+1}, \qquad n = 0, 1, \ldots, N_0 - 1. \tag{8.159}$$

To the initial condition (8.134), the following condition corresponds:

$$y_0 = 0, \qquad x \in \omega. \tag{8.160}$$

The matter of construction and computational realization of difference schemes for the approximate solution of direct initially boundary problems of type (8.132)–(8.134) was considered in more detail in Chapter 4. Of our primary concern here is the following question: how, based on the difference scheme (8.158)–(8.160), computational algorithms for numerical solution of the coefficient inverse problem (8.132)–(8.134) can be constructed.

In using the piecewise constant approximation (8.147), (8.156), (8.157) of the unknown coefficient $k(u)$, we can use the step-by-step identification method. We assume that some approximation for the nonlinear coefficient is available for $0 \le u < u_{\beta-1}$. The difference solution at the corresponding time, and at preceding times, is also found. We seek the solution of the inverse problem for $u_{\beta-1} \le u < u_\beta$.

We denote as $L^{(\beta)}$ the number of the time layer at which the solution u_β is achieved. With (8.157), we obtain:

$$L^{(\beta)} = \sum_{\gamma=1}^{\beta} N_0^{(\gamma)}, \qquad L^{(0)} = 0, \quad \beta = 1, 2, \dots, p.$$

In solving the inverse problem, the solution y_l, $0 \le l \le L^{(\beta-1)}$ is known; according to (8.147), (8.156), (8.157)), also known is the sought coefficient k_u in the same interval $(0 < u < u_{\beta-1})$.

Next, we solve the difference problem

$$\begin{aligned} &\frac{y_{n+1} - y_n}{\tau} + A(y_{n+1})y_{n+1} = 0, \qquad x \in \omega, \\ &n = L^{(\beta-1)} + 1, L^{(\beta-1)} + 2, \dots, L^{(\beta)} \end{aligned} \tag{8.161}$$

$$\begin{aligned} &y_{n+1}(0) = 0, \qquad y_{n+1}(l) = g_{n+1}, \\ &n = L^{(\beta-1)} + 1, L^{(\beta-1)} + 2, \dots, L^{(\beta)}. \end{aligned} \tag{8.162}$$

Here, only the coefficient a_β is unknown. To determine this coefficient, we invoke additional available information (see (8.148), (8.149)).

We assume that the observation points z_m, $m = 1, 2, \dots, M$ are some internal nodes of the calculation grid over space. As the criterion for closeness of the approximate solution at these points to measured values, in solution of problem (8.161), (8.162), in line with (8.148), it seems reasonable to use the criterion

$$J^{(\beta)} = \sum_{m=1}^{M} \sum_{n=L^{(\beta-1)}+1}^{L^{(\beta)}} (y_n(z_m) - \varphi_m^\delta(t_n))^2 \tau. \tag{8.163}$$

In view of $J^{(\beta)} = J^{(\beta)}(a_\beta)$, the parameter a_β can be found from the minimum of the function $J^{(\beta)}(a_\beta)$. In computational realization, to find the minimum of (8.163), we can use standard minimization methods for a function of one variable (the golden-section method, the method of parabolas, etc.).

In the local regularization under consideration, matching with the input-data inaccuracy can be achieved through a proper choice of the interval $[u_{\beta-1}, u_\beta]$ (choice of $N_0^{(\beta)} = L^{(\beta)} - L^{(\beta-1)}$ in (8.161), (8.162). Using the discrepancy principle and inequality (8.149), we choose a maximum $N_0^{(\beta)}$ for which

$$J^{(\beta)} \le M N_0^{(\beta)} \delta^2. \tag{8.164}$$

In some critical cases, relation (8.164) can be violated at all $N_0^{(\beta)}$; then, we have to choose $N_0^{(\beta)} = 1$.

8.4.5 Program

The computational algorithm for the successive determination of the nonlinear coefficient $k(u)$ is realized in the program PROBLEM18. Consider some specific features of this program.

The time interval is divided into equal subintervals; the total number of these subintervals is $p = 2^{\nu}$, $\nu = 0, 1, \ldots, \nu_{\max}$ (in the case of $N_0 = 2^{\nu_{\max}}$). The refinement is terminated on achievement of a certain level of discrepancy or on the condition that subsequent densening of the grid does not decrease the discrepancy. To approximately solve the nonlinear equation for the constant, we use the golden-section method.

Program PROBLEM18

```
C
C     PROBLEM18 - COEFFICIENT INVERSE PROBLEM
C     QUASI-LINEAR 1D PARABOLIC EQUATION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.02D0, N = 100, M = 128 )
      DIMENSION     X(N+1), Y(N+1), Y1(N+1), YT(N+1)
     +             ,U(N+1,M+1), AKS(M+1), PHI(M+1), PHID(M+1)
     +             ,BR(M+1), UL(M+1)
     +             ,A(N+1), B(N+1), C(N+1), F(N+1)
     +             ,ALPHA(N+2), BETA(N+2)
C
C     PARAMETERS:
C
C     XL, XR    - LEFT AND RIGHT END POINTS OF THE GEGMENT;
C     N + 1     - NUMBER OF NODAL POINTS OVER SPACE;
C     M + 1     - NUMBER OF NODAL POINTS OVER TIME;
C     XD        - OBSERVATION POINT;
C     PHI(M+1)  - EXACT SOLUTION AT THE OBSERVATION POINT;
C     PHID(M+1) - DISTURBED SOLUTION AT THE OBSERVATION POINT;
C
      XL   = 0.D0
      XR   = 1.D0
      TMAX = 1.D0
      XD   = 0.6D0
      EPSA = 1.D-4
      EPSH = 1.D-4
C
      OPEN (01, FILE='RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H    = ( XR - XL ) / N
      TAU  = TMAX / M
      DO I = 1, N+1
        X(I) = XL + (I-1)*H
      END DO
C
```

```
C     NODE CLOSEST TO THE OBSERVATION POINT XD
C
      ND = (XD-XL) / H + 1
      IF ( (XD-XL - (ND-1)*H).GT.0.5*H ) ND = ND + 1
C
C     DIRECT PROBLEM
C
C     BOUNDARY CONDITION
C
      DO K = 1, M
        T = K*TAU
        BR(K) = AG(T)
      END DO
C
C     INITIAL CONDITION
C
      T = 0.D0
      DO I = 1, N+1
        Y(I)   = 0.D0
        U(I,1) = Y(I)
      END DO
C
C     NEXT TIME LAYER
C
      DO K = 1, M
        T = K*TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C     PURELY IMPLICIT LINEARIZED SCHEME
C
        DO I = 2, N
          U1 = (Y(I) + Y(I-1)) / 2
          U2 = (Y(I+1) + Y(I)) / 2
          A(I) = AK(U1) / (H*H)
          B(I) = AK(U2) / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          F(I) = Y(I) / TAU
        END DO
C
C     BOUNDARY CONDITION AT THE LEFT END POINT
C
        B(1) = 0.D0
        C(1) = 1.D0
        F(1) = 0.D0
C
C     BOUNDARY CONDITION AT THE RIGHT END POINT
C
        A(N+1) = 0.D0
        C(N+1) = 1.D0
        F(N+1) = AG(T)
C
C     SOLUTION OF THE PROBLEM ON THE NEXT TIME LAYER
C
        CALL PROG ( N+1, A, C, B, F, ALPHA, BETA, Y )
        DO I = 1, N + 1
          U(I,K+1) = Y(I)
        END DO
      END DO
C
C     SOLUTION AT THE OBSERVATION POINT
```

```
C
      DO K = 1, M+1
        PHI(K)  = U(ND,K)
        PHID(K) = PHI(K)
      END DO
C
C     DISTURBING OF MEASURED VALUES
C
      DO K = 2, M+1
          PHID(K) = PHI(K) + 2.*DELTA*(RAND(0)-0.5)
      END DO
C
C     INVERSE PROBLEM
C
C     CHOICE OF THE STEP IN THE PIECEWISE-CONSTANT APPROXIMATION
C
      L = 1
  100 CONTINUE
      ML = M / L
      DO LK = 1, L+1
        T = (LK-1)*ML*TAU
        UL(LK) = AG(T)
      END DO
C
C     INITIAL CONDITION
C
      T  = 0.D0
      SD = 0.D0
      DO I = 1, N+1
        Y1(I) = 0.D0
      END DO
      DO LK = 1, L
C
C     DETERMINATION OF THE UNKNOWN COEFFICIENT OVER THE SUB-INTERVAL
C     ITERATION GOLDEN-SECTION METHOD
C
        AS = 0.1D0
        BS = 10.D0
        R1 = (DSQRT(5.0D0)-1.D0)/2
        R2 = R1**2
        HS = BS - AS
C
C    SOLUTION OF THE PROBLEM OVER THE SUB-INTERVAL
C    AT A GIVEN COEFFICIENT
C
        CALL STEPB ( N, ND, M, ML, LK, H, TAU, A, C, B, F, AL, BET
     +              ,Y, Y1, AS, AKS, UL, PHID, YA )
        CALL STEPB ( N, ND, M, ML, LK, H, TAU, A, C, B, F, AL, BET
     +              ,Y, Y1, BS, AKS, UL, PHID, YB )
        CS = AS + R2*HS
        DS = AS + R1*HS
        CALL STEPB ( N, ND, M, ML, LK, H, TAU, A, C, B, F, AL, BET
     +              ,Y, Y1, CS, AKS, UL, PHID, YC )
        CALL STEPB ( N, ND, M, ML, LK, H, TAU, A, C, B, F, AL, BET
     +              ,Y, Y1, DS, AKS, UL, PHID, YD )
        KS = 0
  200   KS = KS + 1
        IF (YC.LT.YD) THEN
          BS = DS
          YB = YD
```

```
            DS = CS
            YD = YC
            HS = BS - AS
            CS = AS + R2*HS
            CALL STEPB ( N, ND, M, ML, LK, H, TAU, A, C, B, F, AL, BET
     +                  ,Y, Y1, CS, AKS, UL, PHID, YC )
          ELSE
            AS = CS
            YA = YC
            CS = DS
            YC = YD
            HS = BS - AS
            DS = AS + R1*HS
            CALL STEPB ( N, ND, M, ML, LK, H, TAU, A, C, B, F, AL, BET
     +                  ,Y, Y1, DS, AKS, UL, PHID, YD )
          END IF
          IF (DABS(YB-YA).GT.EPSA.OR.HS.GT.EPSH) GO TO 200
C
C     FOUND SOLUTION
C
          PS = AS
          YP = YA
          IF (YB.LT.YA) THEN
            PS = BS
            YP = YB
          ENDIF
          DO I = 1, N+1
            Y1(I) = Y(I)
          END DO
C
C     TOTAL INACCURACY
C
          SD = SD + YP
        END DO
        WRITE(01,*) L, KS, SD
C
C     EXIT FROM THE PROGRAM BY THE DISCREPANCY CRITERION
C
        L = L*2
        IF (L.EQ.2) THEN
           SD0 = SD
           IF (SD.GT.TMAX*DELTA**2 .AND. L .LE. M) GO TO 100
        ELSE
           IF (SD.LT.SD0) THEN
             IF (SD.GT.TMAX*DELTA**2 .AND. L .LE. M) GO TO 100
           END IF
        END IF
C
C     RECORDING OF THE SOLUTION IN A FILE
C
        WRITE ( 01, * ) ((U(I,K),I=1,N+1), K=1,M+1)
        WRITE ( 01, * ) (PHI(K), K=1,M+1)
        WRITE ( 01, * ) (PHID(K), K=1,M+1)

        WRITE ( 01, * ) (AKS(K), K=1,M+1)
        CLOSE ( 01 )
        STOP
        END
C
        SUBROUTINE PROG ( N, A, C, B, F, AL, BET, Y )
```

```
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     SWEEP METHOD
C
      DIMENSION A(N), C(N), B(N), F(N), Y(N), AL(N+1), BET(N+1)
C
      AL(1)  = B(1) / C(1)
      BET(1) = F(1) / C(1)
      DO I = 2, N
        SS     = C(I) - AL(I-1)*A(I)
        AL(I)  = B(I) / SS
        BET(I) = ( F(I) + BET(I-1)*A(I) ) / SS
      END DO
      Y(N) = BET(N)
      DO I = N-1, 1, -1
        Y(I) = AL(I)*Y(I+1) + BET(I)
      END DO
      RETURN
      END
C
      DOUBLE PRECISION FUNCTION  AK ( U )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      COEFFICIENT AT HIGHER DERIVATIVES
C
      AK = 0.5D0 + U
C
      RETURN
      END
C
      DOUBLE PRECISION FUNCTION  AG ( T )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C      BOUNDARY CONDITION AT THE RIGHT END POINT
C
      AG = T
C
      RETURN
      END
C
      SUBROUTINE STEPB ( N, ND, M, ML, LK, H, TAU, A, C, B, F, AL, BET
     +                  ,Y, Y1, AK, AKS, UL, PHI, FD )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     CALCULATION OF THE DISCREPANCY ON THE SUBINTERVAL
C     AT A GIVEN CONSTANT COEFFICIENT
C
      DIMENSION A(N+1), C(N+1), B(N+1), F(N+1), Y(N+1), Y1(N+1)
     +         ,ALPHA(N+2), BETA(N+2), PHI(M+1), AKS(M+1), UL(M+1)
C
      AKS(LK+1) = AK
      FD = 0.D0
      DO I = 1, N+1
        Y(I) = Y1(I)
      END DO
      T = (LK-1)*ML*TAU
      DO K = 1, ML
        T = T + TAU
C
C     DIFFERENCE-SCHEME COEFFICIENTS
```

```
C     PURELY IMPLICIT SCHEME
C
        DO I = 2, N
          U1 = (Y(I) + Y(I-1)) / 2
          U2 = (Y(I+1) + Y(I)) / 2
          A(I) = AFK(U1, LK+1, UL, AKS, M) / (H*H)
          B(I) = AFK(U2, LK+1, UL, AKS, M) / (H*H)
          C(I) = A(I) + B(I) + 1.D0 / TAU
          F(I) = Y(I) / TAU
        END DO
C
C     BOUNDARY CONDITION AT THE LEFT END POINT
C
        B(1) = 0.D0
        C(1) = 1.D0
        F(1) = 0.D0
C
C     BOUNDARY CONDITION AT THE RIGHT END POINT
C
        A(N+1) = 0.D0
        C(N+1) = 1.D0
        F(N+1) = AG(T)
C
C     SOLUTION OF THE PROBLEM ON THE NEXT TIME LAYER
C
        CALL PROG ( N+1, A, C, B, F, ALPHA, BETA, Y )
C
C     DISCREPANCY AT THE OBSERVATION POINT
C
        FD = FD + (Y(ND) - PHI((LK-1)*ML+K+1))**2*TAU
      END DO
      RETURN
      END
C
      DOUBLE PRECISION FUNCTION  AFK ( U, L, UL, AL, M )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
      DIMENSION UL(M+1),  AL(M+1)
C
C     PIECEWISE-CONSTANT COEFFICIENT
C
      AFK = AL(2)
      DO IL = 2, L
        IF (U.LE.UL(IL) .AND. U.GT.UL(IL-1)) AFK = AL(IL)
      END DO
C
      RETURN
      END
```

In the subroutine-function AK, the nonlinear coefficient for the direct problem is specified; the data obtained by solving this program are used as input data for the inverse problem. The subroutine-function AG calculates the boundary conditions at the right boundary. The piecewise constant completion is effected in AFK.

8.4.6 Computational experiments

For input-data setting, we solve the direct problem for $l = 1$, $T = 1$ on the uniform grid with $h = 0.01$ and $\tau = 2^{-7}$. Here, the simplest linearized difference scheme is used in which the nonlinear coefficient was taken from the previous time layer, i. e.,

$$\frac{y_{n+1} - y_n}{\tau} + A(y_n)y_{n+1} = 0, \qquad x \in \omega, \quad n = 0, 1, \ldots, N_0 - 1.$$

Below, calculation data are given, obtained for the case in which the exact solution was obtained for

$$k(u) = 0.5 + u,$$

so that the coefficient to be found exhibits three-fold variation in the calculation domain. The direct-problem solution is shown in Figure 8.22.

In solving the inverse problem, we have restricted ourselves to one observation point and to the case in which

$$z_1 = 0.3, \quad M = 1, \qquad \varphi_1(t_n) = y_n(z_1), \quad n = 1, 2, \ldots, N_0,$$

where $y_n(x)$ is the mesh solution of the direct problem. These data were perturbed with a normally distributed function:

$$\varphi_1^\delta(t_n) = \varphi_1(t_n) + 2\delta(\sigma(t_n) - 1/2), \qquad n = 1, 2, \ldots, N_0.$$

The solution of the inverse problem obtained at the inaccuracy level defined by $\delta = 0.02$ is shown in Figure 8.23. Considering the fact that the boundary condition was set in the form $g(t) = t$, in one and the same graph we have plotted the exact and approximate input data for the inverse problem (the solution at the point $x = z_1 = 0.6$ and the sought coefficient $k(u)$). At the chosen level of inaccuracy, for the coefficient to be reconstructed it suffices to have two subintervals ($p = 2$ in (8.147), (8.156)). The effect due to inaccuracies can be figured out considering Figures 8.24 and 8.25.

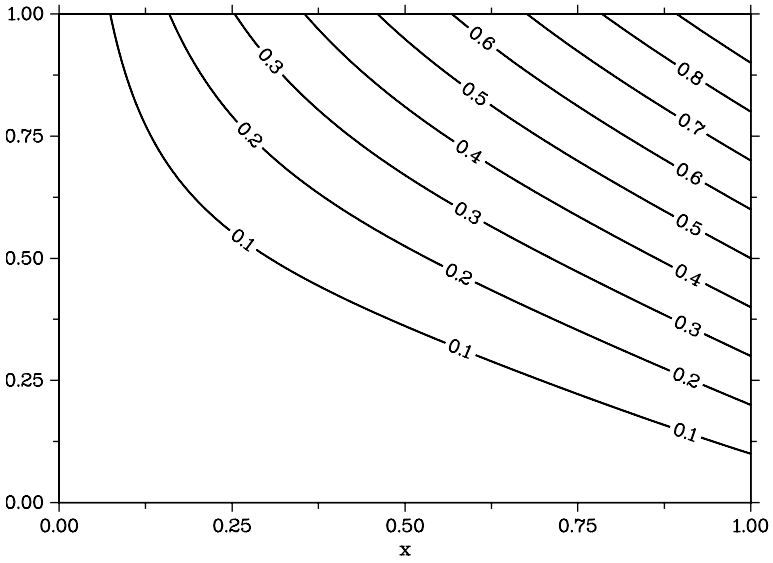

Figure 8.22 Direct-problem solution

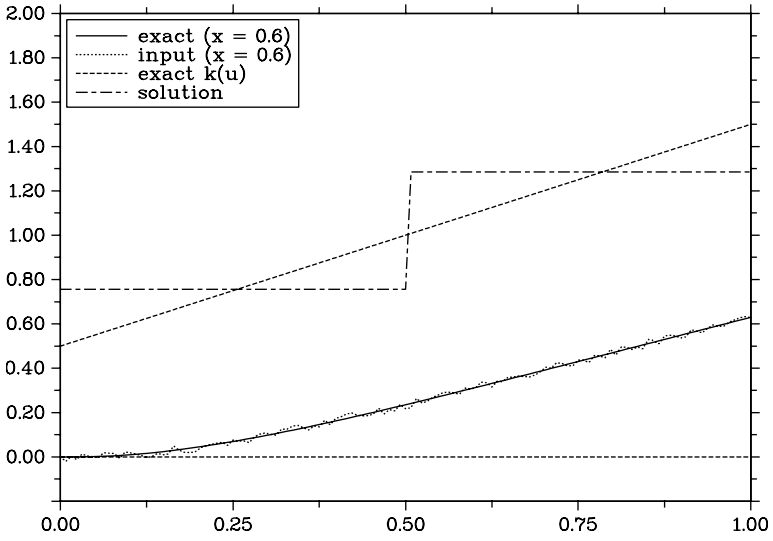

Figure 8.23 Inverse-problem solution obtained with $\delta = 0.02$

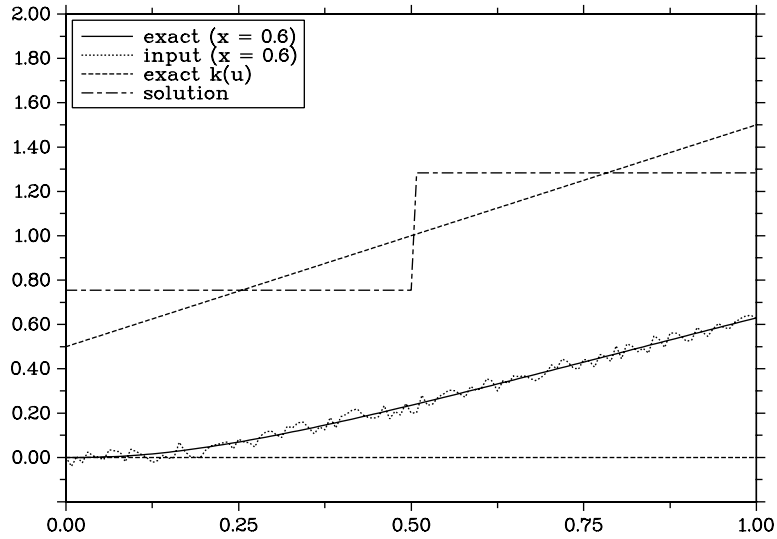

Figure 8.24 Inverse-problem solution obtained with $\delta = 0.04$

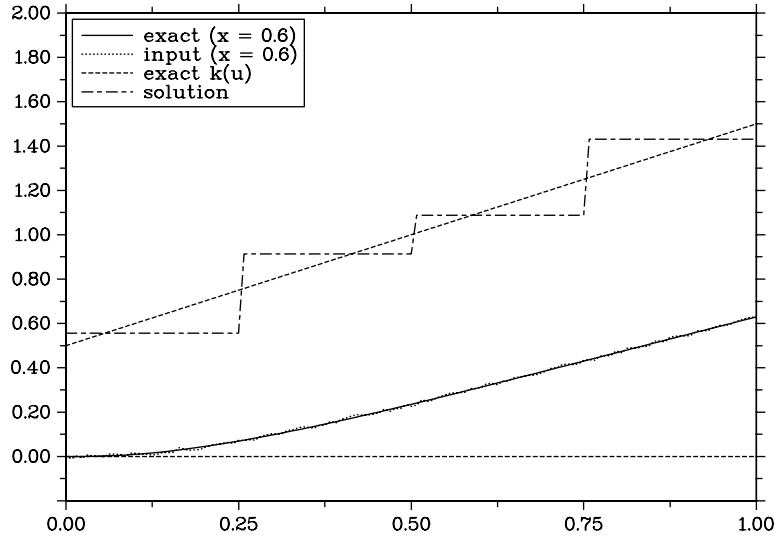

Figure 8.25 Inverse-problem solution obtained with $\delta = 0.01$

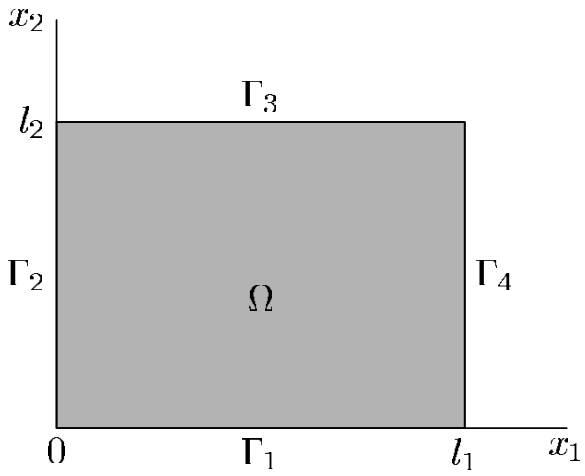

Figure 8.26 Calculation domain

8.5 Coefficient inverse problem for elliptic equation

In this section, we consider a problem in which it is required to find the lowest coefficient in the second-order elliptic equation from data given at the calculation-domain boundary. We assume that the unknown coefficient is independent of one of the variables. With this simplest example, we illustrate the possibility to examine the solution unicity for coefficient inverse problems. The computational algorithm is constructed for a model problem in a rectangle.

8.5.1 Statement of the problem

The problem is considered in its simplest form, with the calculation domain being a rectangle:

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\alpha < l_\alpha,\ \alpha = 1, 2\}.$$

For individual segments of the boundary of Ω, we use the following settings (see Figure 8.26):

$$\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4.$$

As usually, we start from the formulation of the direct problem. We assume that the function $u(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, x_2)$ satisfies the equation

$$-\Delta u + c(x_2)u = 0, \qquad \boldsymbol{x} \in \Omega, \tag{8.165}$$

where

$$\Delta u \equiv \sum_{\alpha=1}^{2} \frac{\partial^2 u}{\partial x_\alpha^2}.$$

Equation (8.165) is supplemented with boundary conditions of the first kind:

$$u(\boldsymbol{x}) = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega. \tag{8.166}$$

In the direct problem (8.165), (8.166), the lowest coefficient is assumed to depend just on x_2.

In the inverse problem of interest, the coefficient $c(x_2)$ is unknown. This coefficient is to be determined from some additional data. We consider the case in which the additional data are obtained from measurements performed on the domain boundary in the form

$$\frac{\partial u}{\partial n}(\boldsymbol{x}) = \psi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega, \tag{8.167}$$

where n is the external normal to Ω.

Previously, we considered a linear inverse problem in which it was required to determine the unknown right-hand side independent of one of the variables. The posed inverse problem (8.165)–(8.167), in which it is required to find a function pair $\{u(\boldsymbol{x}), c(x_2)\}$, is a more complex problem. The difficulties in the consideration of this problem largely result from its nonlinearity.

8.5.2 Solution uniqueness for the inverse problem

We can speak of a certain overdetermination of the coefficient inverse problem under consideration. In problem (8.165)–(8.167), identification of the unknown coefficient $c(x_2)$ implies determination of a function of one variable over the interval $[0, l_1]$ from two functions of the variable x_2 (the function $\psi(\boldsymbol{x})$ on the sides Γ_2, Γ_4) and from two functions of the variable x_1 (the function $\psi(\boldsymbol{x})$ on the sides Γ_1, Γ_3). Hence, in general it would be reasonable to pose the question about sufficient additional data for determining the coefficient $c(x_2)$.

To examine solution uniqueness for the inverse problem (8.165)–(8.167), one can follow the traditional approach that is normally used in the cases of direct nonlinear boundary value mathematical physics problems and nonlinear inverse problems. Suppose that there exist two solutions of the inverse problem (8.165)–(8.167); we denote these solutions as $\{u_\beta(\boldsymbol{x}), c_\beta(x_2), \beta = 1, 2\}$, i. e.,

$$-\Delta u_\beta + c_\beta(x_2)u_\beta = 0, \qquad \boldsymbol{x} \in \Omega, \tag{8.168}$$

$$u_\beta(\boldsymbol{x}) = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega, \tag{8.169}$$

$$\frac{\partial u_\beta}{\partial n}(\boldsymbol{x}) = \psi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega, \quad \beta = 1, 2. \tag{8.170}$$

For the differences

$$v(\boldsymbol{x}) = u_1(\boldsymbol{x}) - u_2(\boldsymbol{x}), \qquad \theta(x_2) = c_1(x_2) - c_2(x_2),$$

from (8.168)–(8.170) we obtain:

$$-\Delta v + c_1(x_2)v + \theta(x_2)u_2(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \Omega, \tag{8.171}$$

$$v(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega, \tag{8.172}$$

$$\frac{\partial v}{\partial n}(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega. \tag{8.173}$$

The solution uniqueness for the inverse problem (8.165)–(8.167) will be proved if we show that equalities (8.171)–(8.173) are only valid with $v(\boldsymbol{x}) = 0$, $\theta(x_2) = 0$, $\boldsymbol{x} \in \Omega$.

Problem (8.171)–(8.173) is considered with given $c_1(x_2)$ and $u_2(\boldsymbol{x})$. Here, we deal with an inverse (but linear) problem in which it is required to determine the pair $v(\boldsymbol{x})$, $\theta(x_2)$, the identification problem for the right-hand side. Let us formulate some sufficient conditions that guarantee that $v(\boldsymbol{x}) = 0$, $\theta(x_2) = 0$, $\boldsymbol{x} \in \Omega$. Apart from the usual assumptions that the solution and the coefficient are smooth functions, we assume that the solution $u(\boldsymbol{x})$ in (8.165)–(8.167) is a constant-sign function, for instance,

$$u(\boldsymbol{x}) > 0, \qquad \boldsymbol{x} \in \overline{\Omega}.$$

This condition is guaranteed (maximum principle) by the assumption that

$$c(x_2) \geq 0, \quad \boldsymbol{x} \in \Omega, \qquad \psi(\boldsymbol{x}) > 0, \quad \boldsymbol{x} \in \partial\Omega.$$

Under such constraints, from (8.171) we obtain the following composite equation:

$$\frac{\partial}{\partial x_1}\Big(\frac{1}{u_2}\,(-\Delta v + c_1(x_2)v)\Big) = 0, \qquad \boldsymbol{x} \in \Omega. \tag{8.174}$$

It is required to prove that the solution of the boundary value problem (8.172)–(8.174) is $v(\boldsymbol{x}) = 0$. Then, from (8.171) it immediately follows ($u_2 > 0$) that $\theta(x_2) = 0$ as well.

We multiply equation (8.174) by some function $\eta(\boldsymbol{x})$ such that

$$\eta(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega, \tag{8.175}$$

and integrate the resulting equation over the domain Ω. In view of (8.175), we obtain

$$\int_\Omega w(-\Delta v + c_1(x_2)v)\,d\boldsymbol{x} = 0, \tag{8.176}$$

where the following notation is used:

$$w(\boldsymbol{x}) = \frac{1}{u_2(\boldsymbol{x})}\,\frac{\partial \eta(\boldsymbol{x})}{\partial x_1}.$$

With regard to the homogeneous boundary conditions (8.172) and (8.173), from (8.176) we obtain:

$$\int_\Omega (-\Delta w + c_1(x_2)w)v\,d\boldsymbol{x} = 0.$$

This equality yields $v(\boldsymbol{x}) = 0$, $\boldsymbol{x} \in \Omega$ provided that we can choose $\eta(\boldsymbol{x})$ so that

$$-\Delta w + c_1(x_2)w = v(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega \tag{8.177}$$

holds. For the calculation domain under consideration (see Figure 8.26), consider the following boundary value problem for equation (8.177) with mixed boundary conditions

$$w(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \Gamma_1 \cup \Gamma_3, \tag{8.178}$$

$$\frac{\partial w}{\partial n}(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \Gamma_2 \cup \Gamma_4. \tag{8.179}$$

There is no doubt that the solution of the standard boundary value problem (8.177)–(8.179) does exist. In more general problems, we can use the result about solution existence for equation (8.177), in the case in which the boundary conditions are given in the form

$$\frac{\partial w}{\partial x_1}(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\Omega;$$

here we have a boundary value problem for the second-order elliptic equation with given angled derivatives on the boundary.

It only remains to identify the form of $\eta(\boldsymbol{x})$. To this end, using the function $w(\boldsymbol{x})$ found from (8.177)–(8.179), we solve the boundary value problem

$$\frac{\partial}{\partial x_1}\Big(\frac{1}{u_2(\boldsymbol{x})}\,\frac{\partial \eta(\boldsymbol{x})}{\partial x_1}\Big) = \frac{\partial w}{\partial x_1}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega$$

with boundary conditions (8.175).

8.5.3 Difference inverse problem

Let us formulate the difference analogue of the coefficient inverse problem (8.165)–(8.167). In Ω, we introduce a grid with step sizes h_α, $\alpha = 1, 2$ uniform along both directions. First, we define the set of internal grid nodes

$$\omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ x_\beta = i_\beta h_\beta,\ i_\beta = 1, 2, \ldots, N_\beta - 1,\ N_\beta h_\beta = l_\beta,\ \beta = 1, 2\}$$

and let $\partial\omega$ be the set of boundary nodes. We designate as $\partial\omega^*$ the set of near-boundary nodes, i.e.,

$$\partial\omega^* = \{\boldsymbol{x} \mid \boldsymbol{x} \in \omega,\ x_\beta = h_\beta,\ x_\beta = l_\beta - h_\beta,\ \beta = 1, 2\}.$$

First, we approximate the direct problem (8.165), (8.166). For internal nodes, we define the standard two-dimensional difference Laplace operator Λ on a five-point mesh pattern:

$$\Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}, \qquad \boldsymbol{x} \in \omega.$$

We put in correspondence to the boundary value problem (8.165), (8.166) the difference Dirichlet problem

$$-\Lambda y + c(x_2)y = 0, \qquad \boldsymbol{x} \in \omega, \tag{8.180}$$

$$y(\boldsymbol{x}) = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega. \tag{8.181}$$

In the inverse problem, the mesh function $c(x_2),\ x_2 \in \omega_2$ is unknown, but additional conditions are given which correspond to (8.167). On the passage to the discrete problem, additional conditions can be put in correspondence with setting the mesh solution at the nodes of $\partial\omega^*$. The latter is the case in using the simplest approximation with directed differences (with first-order approximation accuracy). A similar situation is observed in the case in which second-order approximations of the boundary condition (8.167) are used. As an illustration, consider approximations of boundary conditions on Γ_2.

For the near-boundary node we have

$$\frac{u(h_1, x_2) - u(0, x_2)}{h_1} = \frac{\partial u}{\partial x_1}(0, x_2) + \frac{h_1}{2}\frac{\partial^2 u}{\partial x_1^2}(0, x_2) + \mathcal{O}(h_1^2). \tag{8.182}$$

With regard to the boundary conditions (8.166), (8.167), we arrive at the expression

$$y(h_1, x_2) = \varphi(h_1, x_2) - h_1\psi(h_1, x_2).$$

The use of this expression corresponds to the approximation of boundary conditions (8.167) of first order. The higher (second) approximation order is achieved for the solution of the boundary value problem (8.165), (8.166). In this case, it follows from (8.165) that

$$\frac{\partial^2 u}{\partial x_1^2}(0, x_2) = \Delta u(0, x_2) - \frac{\partial^2 u}{\partial x_2^2}(0, x_2) = c(x_2)\varphi(0, x_2) - \frac{\partial^2 \varphi}{\partial x_2^2}(0, x_2).$$

With regard to (8.182), for the near-boundary nodes we can put

$$y(h_1, x_2) = \varphi(h_1, x_2) - h_1\psi(h_1, x_2) + c(x_2)\varphi(0, x_2) - \varphi_{\bar{x}_2 x_2}(0, x_2).$$

Yet, such approximations are not too much suitable for us in solving the inverse problem since, here, the value at the near-boundary node depends on the unknown (sought) coefficient $c(x_2)$. That is why we will use the simplest first-kind approximations; in this case, the additional conditions can be formulated in the form

$$y(\boldsymbol{x}) = \phi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega^*. \tag{8.183}$$

In the solution of applied problems the input data are normally given with some inaccuracy. We assume that, compared to these inaccuracies, the approximation inaccuracies can be disregarded with sufficiently fine calculation grids used. In the example under consideration (8.165)–(8.167), we restrict ourselves to the case in which the

largest inaccuracies are the measurement inaccuracies of the normal derivative at the boundary. Hence, in approximating the solution of the inverse problem with input-data inaccuracies, we can leave the boundary condition (8.181) unchanged, and instead of (8.183), put

$$y(\boldsymbol{x}) \approx \phi_\delta(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\omega^*, \tag{8.184}$$

where the parameter δ defines the inaccuracy level.

8.5.4 Iterative solution of the inverse problem

In the approximate solution of the inverse problem (8.180), (8.181), (8.184), we will use gradient iteration methods. Here, the sought mesh function $c(x_2)$ is to be refined at each iteration step using the criterion for minimization of the discrepancy functional (the discrepancy here is the accuracy with which the condition (8.184) is fulfilled).

We define the discrepancy as

$$J(c) = \sum_{\boldsymbol{x}\in\partial\omega^*} (y(\boldsymbol{x}) - \phi_\delta(\boldsymbol{x}))^2 h(\boldsymbol{x}), \tag{8.185}$$

where

$$h(\boldsymbol{x}) = \begin{cases} h_1, & x_2 = h_2, l_2 - h_2,\ x_1 \neq h_1, l_1 - h_1, \\ h_2, & x_1 = h_1, l_1 - h_1,\ x_2 \neq h_2, l_2 - h_2, \\ (h_1 + h_2)/2, & x_1 = h_1, l_1 - h_1,\ x_2 = h_2, l_2 - h_2. \end{cases}$$

We assume that the sought function $c(x_2)$ belongs to the space of mesh functions $L_2(\omega_2)$, in which for the scalar product and norm we use the settings

$$(c, d) = \sum_{x_2\in\omega} c(x_2) d(x_2) h_2, \qquad \|c\| = \sqrt{(c, c)}.$$

To derive an expression for the gradient of $J(c)$, we assume that, to the increment δc, some increment δJ of the functional (8.185) and some increment δy of the solution of problem (8.180), (8.181) corresponds.

Accurate to terms of second-order smallness, from (8.180) and (8.181) we obtain:

$$-\Lambda\, \delta y + c(x_2)\, \delta y + \delta c\, y = 0, \qquad \boldsymbol{x} \in \omega, \tag{8.186}$$

$$\delta y(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \partial\omega. \tag{8.187}$$

For the discrepancy functional gradient, we immediately obtain:

$$\delta J(c) = 2 \sum_{\boldsymbol{x}\in\partial\omega^*} (y(\boldsymbol{x}) - \phi_\delta(\boldsymbol{x}))\, \delta y\, h(\boldsymbol{x}). \tag{8.188}$$

The gradient of $J'(c)$ corresponds to the functional increment represented as

$$\delta J(c) = (J'(c), c).$$

To rearrange the right-hand side of (8.188), we multiply the equation (8.186) by some mesh function $\xi(\boldsymbol{x})\,h_1h_2,\ \boldsymbol{x}\in\omega$ and sum up the obtained equation over all nodes in ω:

$$\sum_{\boldsymbol{x}\in\omega}\xi(\boldsymbol{x})\,(-\Lambda\,\delta y + c(x_2)\,\delta y + \delta c\,y)h_1h_2 = 0. \tag{8.189}$$

We assume that

$$\xi(\boldsymbol{x}) = 0, \qquad \boldsymbol{x}\in\partial\omega. \tag{8.190}$$

Under such constraints, equation (8.189) yields:

$$\sum_{\boldsymbol{x}\in\omega}\delta y\,(-\Lambda\xi + c(x_2)\xi)h_1h_2 + \sum_{\boldsymbol{x}\in\omega}\delta c\,y\xi h_1h_2 = 0. \tag{8.191}$$

To derive the desired representation for $J'(c)$, we relate the first term in (8.191) with the right-hand side of (8.188).

Let the function $\xi(\boldsymbol{x})$ be defined as the solution of the equation

$$-\Lambda\xi + c(x_2)\xi = -F(\boldsymbol{x}), \qquad \boldsymbol{x}\in\omega. \tag{8.192}$$

Here, the right-hand side $F(\boldsymbol{x})$ is

$$F(\boldsymbol{x}) = \begin{cases} 2(y(\boldsymbol{x}) - \phi_\delta(\boldsymbol{x}))/h_2, & x_2 = h_2, l_2 - h_2,\ x_1 \neq h_1, l_1 - h_1, \\ 2(y(\boldsymbol{x}) - \phi_\delta(\boldsymbol{x}))/h_1, & x_1 = h_1, l_1 - h_1,\ x_2 \neq h_2, l_2 - h_2, \\ \dfrac{h_1+h_2}{h_1h_2}\,(y(\boldsymbol{x}) - \phi_\delta(\boldsymbol{x})), & x_1 = h_1, l_1 - h_1,\ x_2 = h_2, l_2 - h_2, \\ 0, & \boldsymbol{x}\in\omega,\ \boldsymbol{x}\notin\partial\omega^*. \end{cases}$$

For this right-hand side, from (8.188), (8.191), and (8.192) we obtain:

$$\delta J(c) = \sum_{\boldsymbol{x}\in\omega}\delta c\,y\xi h_1h_2 = \sum_{x_2\in\omega_2}\delta c\Big(\sum_{x_1\in\omega_1} y\,\xi h_1\Big)h_2.$$

With this equality, for the functional gradient we have the representation

$$J'(c) = \sum_{x_1\in\omega_1} y\,\xi h_1, \qquad x_2\in\omega. \tag{8.193}$$

To calculate the gradient, we have to solve the boundary value problem (8.180), (8.181) for the ground state (the mesh function $y(\boldsymbol{x})$) and the boundary value problem (8.190), (8.192) for the conjugate state (the mesh function $\xi(\boldsymbol{x})$).

In the case of the two-layer gradient iteration method the quantity $c^k(x_2)$ (k is the iteration number) is to be refined using the scheme

$$\frac{c^{k+1} - c^k}{s_{k+1}} + J'(c^k) = 0, \qquad x_2\in\omega_2, \quad k = 0, 1, \ldots\ . \tag{8.194}$$

Here, it should be taken into account that the equation $J'(c) = 0$, to be solved, is a nonlinear equation. The iterative process (8.194) can be terminated by the discrepancy criterion.

8.5.5 Program

The gradient method described above was realized in its simplest version with constant iteration parameter s_{k+1} (simple iteration method). In the perturbation procedure for input data, a situation with perturbed boundary conditions of the second kind (8.167) is modeled.

Program PROBLEM19

```
C
C     PROBLEM19 - IDENTIFICATION OF THE LOWEST COEFFICIENT
C                 IN THE ELLIPTIC EQUATION OF SECOND ORDER
C                 TWO-DIMENSIONAL PROBLEM
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      PARAMETER ( DELTA = 0.02D0, N1 = 51, N2 = 51 )
      DIMENSION A(12*N1*N2), X1(N1), X2(N2), CK(N2), GR(N2)
     +         ,YG1(N1), YG2(N2), YG3(N1), YG4(N2)
     +         ,YD1(N1), YD2(N2), YD3(N1), YD4(N2)
     +         ,YY1(N1), YY2(N2), YY3(N1), YY4(N2)
      COMMON / SB5 /     IDEFAULT(4)
      COMMON / CONTROL / IREPT, NITER
C
C     PARAMETERS:
C
C     X1L, X2L  - COORDINATES OF THE LEFT CORNER;
C     X1R, X2R  - COORDINATES OF THE RIGHT CORNER;
C     N1, N2    - NUMBER OF NODES IN THE SPATIAL GRID;
C     H1, H2    - SPATIAL STEP OF THE GRID;
C     TAU       - TIME STEP;
C     DELTA     - INPUT-DATA INACCURACY LEVEL;
C     CK(N2)    - COEFFICIENT TO BE RECONSTRUCTED;
C     YG1(N1),
C     YG2(N2),
C     YG3(N1),
C     YG4(N2)   - MESH FUNCTION AT THE NEAR-BOUNDARY NODES;
C     YD1(N1),
C     YD2(N2),
C     YD3(N1),
C     YD4(N2)   - DISTURBED MESH FUNCTION AT THE NEAR-BOUNDARY NODES;
C     EPSR      - RELATIVE ACCURACY FOR THE DIFFERENCE-PROBLEM SOLUTION;
C     EPSA      - ABSOLUTE ACCURACY FOR THE DIFFERENCE-PROBLEM SOLUTION;
C
C     EQUIVALENCE ( A(1),         A0            ),
C    *            ( A(N+1),       A1            ),
C    *            ( A(2*N+1),     A2            ),
C    *            ( A(9*N+1),     F             ),
C    *            ( A(10*N+1),    U             ),
C    *            ( A(11*N+1),    US            )
C
      X1L    = 0.D0
      X1R    = 1.D0
      X2L    = 0.D0
      X2R    = 1.D0
      EPSR   = 1.D-6
      EPSA   = 1.D-9
      SS     = - 40.D0
C
```

```
      OPEN (01, FILE='RESULT.DAT') ! FILE TO STORE THE CALCULATED DATA
C
C     GRID
C
      H1  = (X1R-X1L) / (N1-1)
      H2  = (X2R-X2L) / (N2-1)
      DO I = 1, N1
        X1(I) = X1L + (I-1)*H1
      END DO
      DO J = 1, N2
        X2(J) = X2L + (J-1)*H2
      END DO
C
      N  = N1*N2
      DO I = 1, 12*N
        A(I) = 0.D0
      END DO
C
C     DIRECT PROBLEM
C
C     BOUNDARY CONDITIONS
C
      CALL BNG (A(10*N+1), X1, X2, N1, N2)
C
C     LOWEST COEFFICIENT
C
      DO J = 1, N2
        CK(J) = AC(X2(J))
      END DO
C
C     DIFFERENCE-SCHEME COEFFICIENT IN THE DIRECT PROBLEM
C
      CALL FDS (A(1), A(N+1), A(2*N+1), CK, A(9*N+1), A(10*N+1),
     +          H1, H2, N1, N2)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
      IDEFAULT(1) = 0
      IREPT       = 0
      CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
C
C     OBSERVATIONAL DATA
C
      CALL BNGDER (A(10*N+1), H1, H2, N1, N2, YG1, YG2, YG3, YG4)
C
C     DISTURBING OF MEASURED VALUES
C
      DO I = 2, N1-1
        H = H2
        IF (I.EQ.2 .OR. I.EQ.N1-1) H = (H1+H2) / 2.D0
        YD1(I) = YG1(I) + 2.*H*DELTA*(RAND(0)-0.5)
        YD3(I) = YG3(I) + 2.*H*DELTA*(RAND(0)-0.5)
      END DO
      DO J = 2, N2-1
        H = H1
        IF (J.EQ.2 .OR. J.EQ.N2-1) H = (H1+H2) / 2.D0
        YD2(J) = YG2(J) + 2.*H*DELTA*(RAND(0)-0.5)
        YD4(J) = YG4(J) + 2.*H*DELTA*(RAND(0)-0.5)
      END DO
C
```

```
C     INVERSE PROBLEM
C
C     ITERATION METHOD
C
      IT = 0
C
C     INITIAL APPROXIMATION
C
      DO J = 1, N2
        CK(J) = 0.D0
      END DO
C
  100 IT = IT + 1
C
C     GROUND STATE
C
C     DIFFERENCE-SCHEME COEFFICIENTS IN THE DIRECT PROBLEM
C
      CALL FDS (A(1), A(N+1), A(2*N+1), CK, A(9*N+1), A(10*N+1),
     +          H1, H2, N1, N2)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
      IDEFAULT(1) = 0
      IREPT       = 0
      CALL SBAND5 (N, N1, A(1), A(10*N+1), A(9*N+1), EPSR, EPSA)
C
C     SOLUTION AT THE OBSERVATION POINTS
C
      CALL BNGDER (A(10*N+1), H1, H2, N1, N2, YY1, YY2, YY3, YY4)
C
C     CONJUGATE STATE
C
C     DIFFERENCE-SCHEME COEFFICIENTS
C
      CALL FDS (A(1), A(N+1), A(2*N+1), CK, A(9*N+1), A(10*N+1),
     +          H1, H2, N1, N2)
C
C     RIGHT-HAND SIDE
C
      CALL RHS (A(9*N+1), N1, N2, H1, H2, YD1, YD2, YD3, YD4,
     +          YY1, YY2, YY3, YY4)
C
C     SOLUTION OF THE DIFFERENCE PROBLEM
C
      IDEFAULT(1) = 0
      IREPT       = 0
      CALL SBAND5 (N, N1, A(1), A(11*N+1), A(9*N+1), EPSR, EPSA)
C
C     SIMPLE-ITERATION METHOD
C
C     FUNCTIONAL GRADIENT
C
      CALL GRAD (A(10*N+1), A(11*N+1), H1, H2, N1, N2, GR)
C
C     NEXT APPROXIMATION
C
      DO J = 2, N2-1
        CK(J) = CK(J) + SS*GR(J)
      END DO
```

```
C
C     EXIT FROM THE ITERATION PROCEDURE BY THE DISCREPANCY CRITERION
C
      SUM  = 0.D0
      SUMD = 0.D0
      DO I = 3, N1-2
        SUM = SUM + (YD1(I) - YY1(I))**2 * H1
        SUM = SUM + (YD3(I) - YY3(I))**2 * H1
      END DO
      SUMD = SUMD + 2.D0*(N1-4)*H1*H2*DELTA**2
      DO J = 3, N2-2
        SUM = SUM + (YD2(J) - YY2(J))**2 * H2
        SUM = SUM + (YD4(J) - YY4(J))**2 * H2
      END DO
      SUMD = SUMD + 2.D0*(N2-4)*H1*H2*DELTA**2
      SUM = SUM + (YD1(2) - YY1(2))**2 * (H1+H2) / 2.D0
      SUM = SUM + (YD4(2) - YY4(2))**2 * (H1+H2) / 2.D0
      SUM = SUM + (YD3(2) - YY3(2))**2 * (H1+H2) / 2.D0
      SUM = SUM + (YD3(N1) - YY3(N1))**2 * (H1+H2) / 2.D0
      SUMD = SUMD + 2.D0*(H1+H2)*DELTA**2
      IF ( SUM.GT.SUMD ) GO TO 100
C
C     SOLUTION
C
      WRITE ( 01, * ) (A(10*N+I), I=1,N)
      WRITE ( 01, * ) (CK(J), J=1,N2)
      CLOSE (01)
      STOP
      END
C
      DOUBLE PRECISION FUNCTION  AC ( X2 )
      IMPLICIT REAL*8 ( A-H, O-Z )
C
C     LOWEST COEFFICIENT IN THE ELLIPTIC EQUATION
C
      AC = 10.D0*X2
C
      RETURN
      END
C
      SUBROUTINE FDS (A0, A1, A2, CK, F, U,  H1, H2, N1, N2)
C
C     GENERATION OF THE COEFFICIENT ARRAY FOR THE DIFFERENCE SCHEME
C     FOR THE ELLIPTIC EQUATION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  A0(N1,N2), A1(N1,N2), A2(N1,N2)
     +          ,CK(N2), F(N1,N2), U(N1,N2)
C
      DO J = 2, N2-1
        DO I = 2, N1-1
          A1(I-1,J) = 1.D0/(H1*H1)
          A1(I,J)   = 1.D0/(H1*H1)
          A2(I,J-1) = 1.D0/(H2*H2)
          A2(I,J)   = 1.D0/(H2*H2)
          A0(I,J)   = A1(I,J) + A1(I-1,J) + A2(I,J) + A2(I,J-1) + CK(J)
           F(I,J)   = 0.D0
        END DO
      END DO
C
```

```
C     FIRST-KIND HOMOGENEOUS BOUNDARY CONDITIONS
C
      DO J = 2, N2-1
        A0(1,J) = 1.D0
        A1(1,J) = 0.D0
        A2(1,J) = 0.D0
         F(1,J) = U(1,J)
         F(2,J) = F(2,J) + U(1,J) / (H1*H1)
      END DO
C
      DO J = 2, N2-1
        A0(N1,J)   = 1.D0
        A1(N1-1,J) = 0.D0
        A1(N1,J)   = 0.D0
        A2(N1,J)   = 0.D0
         F(N1,J)   = U(N1,J)
         F(N1-1,J) = F(N1-1,J) + U(N1,J) / (H1*H1)
      END DO
C
      DO I = 2, N1-1
        A0(I,1) = 1.D0
        A1(I,1) = 0.D0
        A2(I,1) = 0.D0
         F(I,1) = U(I,1)
         F(I,2) = F(I,2) + U(I,1) / (H2*H2)
      END DO
C
      DO I = 2, N1-1
        A0(I,N2)   = 1.D0
        A1(I,N2)   = 0.D0
        A2(I,N2)   = 0.D0
        A2(I,N2-1) = 0.D0
         F(I,N2)   = U(I,N2)
         F(I,N2-1) = F(I,N2-1) + U(I,N2) / (H2*H2)
      END DO
C
       A0(1,1) = 1.D0
       A1(1,1) = 0.D0
       A2(1,1) = 0.D0
        F(1,1) = U(1,1)
C
       A0(N1,1) = 1.D0
       A2(N1,1) = 0.D0
        F(N1,1) = U(N1,1)
C
       A0(1,N2) = 1.D0
       A1(1,N2) = 0.D0
        F(1,N2) = U(1,N2)
C
       A0(N1,N2) = 1.D0

        F(N1,N2) = U(N1,N2)
C
      RETURN
      END
C
      SUBROUTINE RHS (F, N1, N2, H1, H2, YD1, YD2, YD3, YD4,
     +                YY1, YY2, YY3, YY4)
C
C     RIGHT-HAND SIDE IN THE EQUATION FOR THE CONGUGATE STATE
```

```
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  F(N1,N2), YD1(N1), YD2(N2), YD3(N1), YD4(N2)
     +          ,YY1(N1), YY2(N2), YY3(N1), YY4(N2)
C
      DO I = 1, N1
        DO J = 1,N2
          F(I,J) = 0.D0
        END DO
      END DO
      DO I = 3, N1-2
        F(I,2)    = 2.D0*(YD1(I) - YY1(I)) / H2
        F(I,N1-1) = 2.D0*(YD3(I) - YY3(I)) / H2
      END DO
      DO J = 3, N2-2
        F(2,J)    = 2.D0*(YD2(J) - YY2(J)) / H1
        F(N2-1,J) = 2.D0*(YD4(J) - YY4(J)) / H1
      END DO
      F(2,2)       = (H1+H2)* (YD1(2) - YY1(2)) / (H1*H2)
      F(2,N2-1)    = (H1+H2)* (YD4(2) - YY4(2)) / (H1*H2)
      F(N1-1,2)    = (H1+H2)* (YD3(2) - YY3(2)) / (H1*H2)
      F(N1-1,N2-1) = (H1+H2)* (YD3(N1) - YY3(N1)) / (H1*H2)
C
      RETURN
      END
C
      SUBROUTINE BNG (U, X1, X2, N1, N2)
C
C     FIRST-KIND BOUNDARY CONDITION
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  U(N1,N2), X1(N1), X2(N2)
      DO I = 1, N1
        DO J = 1, N2
          U(I,J) = 1.D0 + X1(I)
        END DO
      END DO
C
      RETURN
      END
C
      SUBROUTINE BNGDER (U, H1, H2, N1, N2, YD1, YD2, YD3, YD4)
C
C     MESH FUNCTION AT THE NEAR-BOUNDARY NODES
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  U(N1,N2), YD1(N1), YD2(N2), YD3(N1), YD4(N2)
      DO I = 2, N1-1
        YD1(I) = U(I,2)
        YD3(I) = U(I,N2-1)
      END DO
      DO J = 2, N2-1
        YD2(J) = U(2,J)
        YD4(J) = U(N1-1,J)
      END DO
C
      RETURN
      END
C
      SUBROUTINE GRAD (U, XI, H1, H2, N1, N2, GR)
```

```
C
C     MESH FUNCTION AT THE NEAR-BOUNDARY NODES
C
      IMPLICIT REAL*8 ( A-H, O-Z )
      DIMENSION  U(N1,N2), XI(N1,N2), GR(N2)
      DO J = 2, N2-1
        SUM = 0.D0
        DO I = 2, N1-1
          SUM = SUM + U(I,J)*XI(I,J)*H1
        END DO
        GR(J) = SUM
      END DO
C
      RETURN
      END
```

In solving the direct problem, the coefficient $c(x_2)$ is set in the subroutine-function AC. In the subroutine BNG, the boundary condition (8.166) is set. The coefficient array for the difference equation is generated in the subroutine FDS. The difference problem for the conjugate state (8.190), (8.192) differs from the difference problem for the ground state (8.180), (8.181) in the right-hand side and in the boundary conditions only (see the subroutine RHS).

8.5.6 Computational experiments

In solving the problem on identification of the lower coefficient, the input data were taken from the solution of the direct problem (8.165), (8.166) in the unit square Ω ($l_1 = l_2 = 1$) with

$$c(x_2) = 10x_2, \qquad \varphi(\boldsymbol{x}) = 1 + x_1.$$

The problem was solved on the uniform grid with $h_1 = h_2 = 0.02$. Difference-solution contour lines spaced 0.05 apart are shown in Figure 8.27.

From the approximate solution of the direct problem, the values at the near-boundary nodes of the calculation grid are extracted; in this way, a condition of type (8.183) is being set. These input data for the inverse problem are perturbed with some random function. Conditions are modeled in which an exact first-kind boundary condition (see (8.166)) is set, whereas the additional, second-kind boundary conditions (see (8.167)) are set with some inaccuracy. In these conditions, for instance, at the near-boundary nodes closest to the boundary Γ_2 we have:

$$y(h_1, x_2) = \phi_\delta((h_1, x_2)) = \phi((h_1, x_2)) + 2h_1\delta(\sigma(h_1, x_2) - 1/2),$$

where

$$\phi((h_1, x_2)) = \varphi(h_1, x_2) - h_1\psi(h_1, x_2),$$

and $\sigma(h_1, x_2)$ is a random function.

The data obtained in the reconstruction of the coefficient with $\delta = 0.02$ are shown in Figure 8.28. The solution of the direct problem obtained with this value of the

coefficient is shown in Figure 8.29. The effect due to the inaccuracy level can be figured out considering Figures 8.30 and 8.31 (twice decreased and twice increased inaccuracy, respectively).

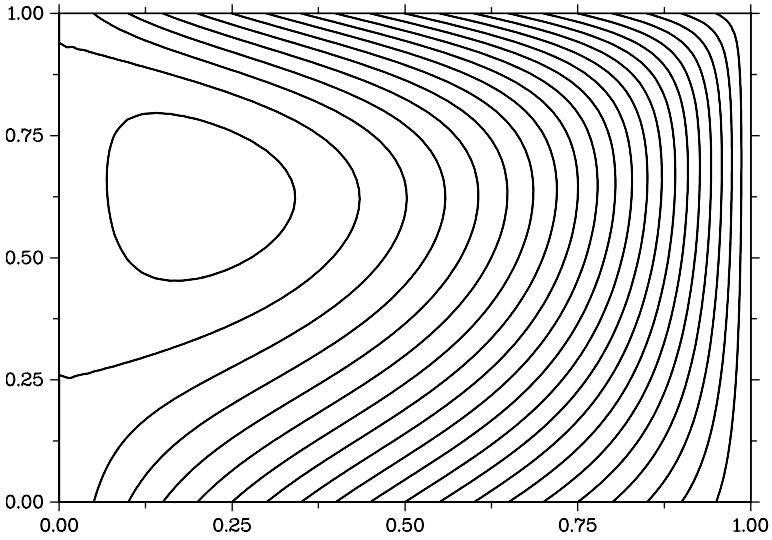

Figure 8.27 Solution of the direct problem

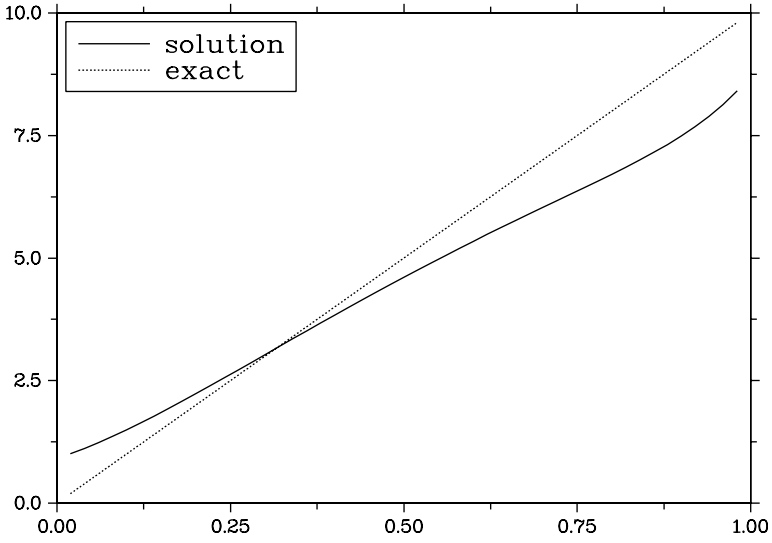

Figure 8.28 Reconstruction of the coefficient with $\delta = 0.02$

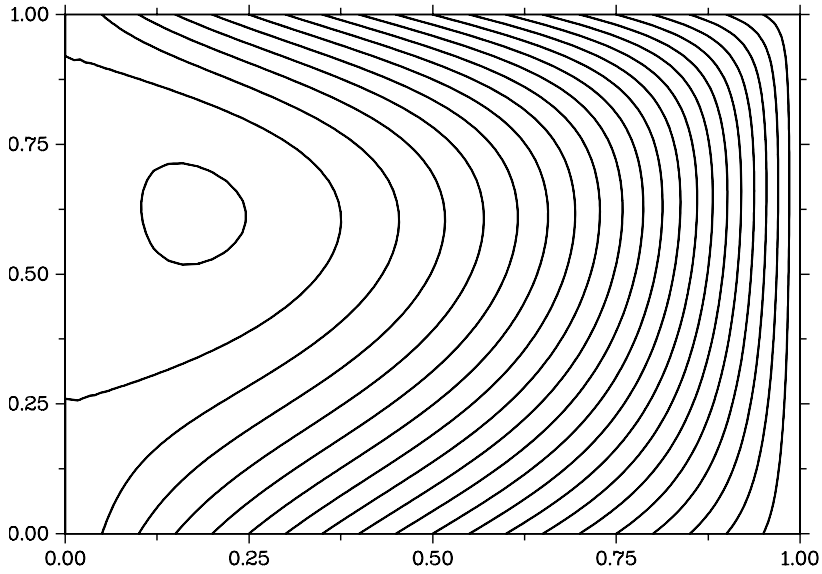

Figure 8.29 Contour lines for the inverse-problem solution obtained with $\delta = 0.02$

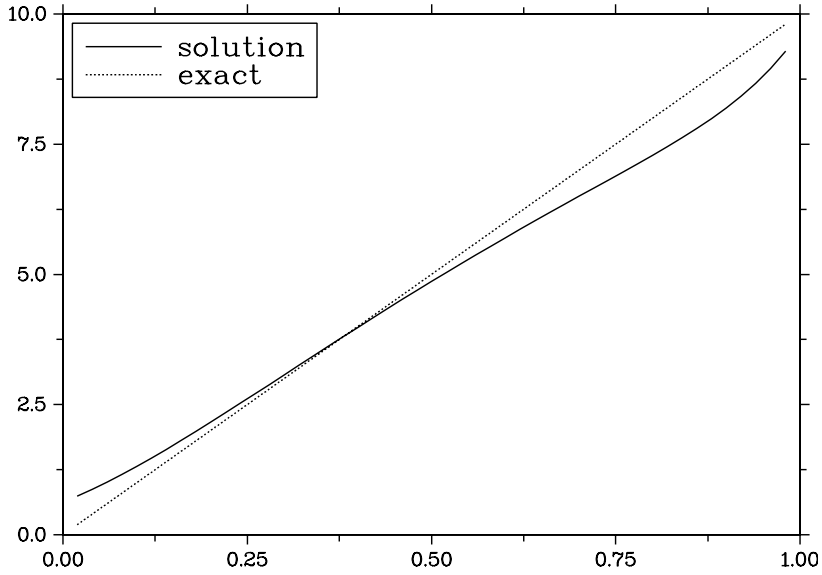

Figure 8.30 Inverse-problem solution obtained with $\delta = 0.01$

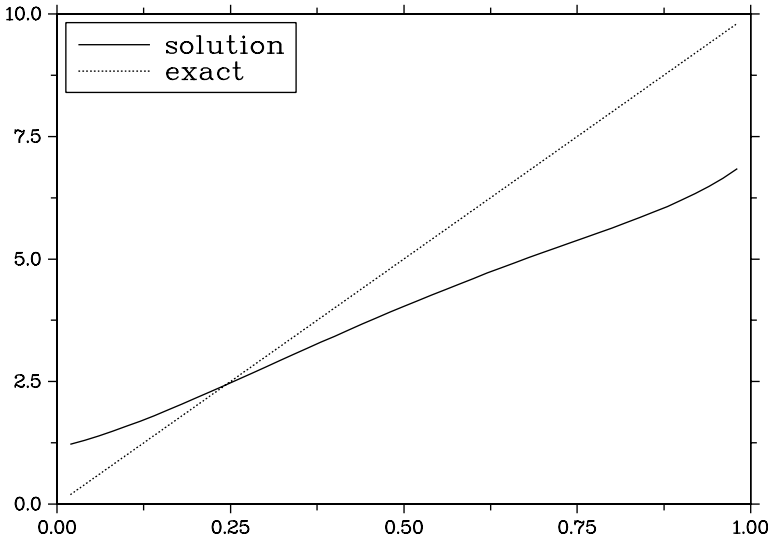

Figure 8.31 Inverse-problem solution obtained with $\delta = 0.04$

8.6 Exercises

Exercise 8.1 Consider the matter of using the quasi-inversion method in continuation over the spatial variable in solving the boundary value inverse problem in the rectangle

$$\Omega = \{\boldsymbol{x} \mid \boldsymbol{x} = (x_1, x_2),\ 0 < x_\alpha < l_\alpha,\ \alpha = 1, 2\}$$

for the two-dimensional parabolic equation

$$\frac{\partial u}{\partial t} - \sum_{\beta=1}^{2} \frac{\partial^2 u}{\partial x_\beta^2} = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T,$$

$$u(0, x_2, t) = \varphi(x_2, t), \qquad \frac{\partial u}{\partial x_1}(0, x_2, t) = 0,$$

$$u(x_1, 0, t) = 0, \qquad u(x_1, l_2, t) = 0, \qquad 0 < t < T,$$

$$u(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \Omega.$$

Exercise 8.2 Examine the convergence of the difference scheme (8.43) in the approximate solution of the Cauchy problem (8.12), (8.13), (8.25).

Exercise 8.3 Based on the program `PROBLEM15`, write a program that realizes the quasi-inversion method in the variant in which the equation

$$\frac{\partial^2 v_\alpha}{\partial x^2} - \frac{\partial v_\alpha}{\partial t} + \alpha \frac{\partial^2 v_\alpha}{\partial t^2} = 0, \qquad 0 < x < l, \quad 0 < t < T$$

is to be solved. Perform a comparative analysis of the variants of the quasi-inversion method as applied to the approximate solution of the model problems.

Exercise 8.4 In the Tikhonov regularization method as applied to approximate solution of the boundary value inverse problem for one-dimensional parabolic equation, to be minimized is the functional

$$J_\alpha(v) = \int_0^T (u(0,t) - \varphi_\delta(t))^2\,dt + \alpha \int_0^T v^2(t)\,dt$$

with the constraints

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l, \quad 0 < t \le T, \\ k(x)\,\frac{\partial u}{\partial x}(0,t) &= 0, \qquad 0 < t \le T, \\ u(l,t) &= v(t), \qquad 0 < t \le T, \\ u(x,0) &= 0, \qquad 0 \le x \le l. \end{aligned}$$

Obtain the optimality conditions for this optimal control problem (Euler equation).

Exercise 8.5 In the uniform norm, examine stability of the difference scheme (8.79)–(8.82) as applied for the solution of the problem with a non-local boundary condition.

Exercise 8.6 In the program `PROBLEM16`, provide a scheme for mesh smoothing of input data for obtaining a smooth solution of the inverse problem. Perform numerical experiments showing how the smoothing parameter affects the approximate-solution accuracy in model inverse problems.

Exercise 8.7 Construct a gradient iteration method for the approximate solution of the boundary value inverse problem

$$\frac{\partial u}{\partial t} + b(x)\,\frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\Big(k(x)\,\frac{\partial u}{\partial x}\Big), \qquad 0 < x < l, \quad 0 < t \le T, \tag{8.195}$$

$$k(x)\,\frac{\partial u}{\partial x}(0,t) = 0, \qquad 0 < t \le T, \tag{8.196}$$

$$u(0,t) = \varphi(t), \qquad 0 < t \le T, \tag{8.197}$$

$$u(x,0) = 0, \qquad 0 \le x \le l, \tag{8.198}$$

by refining the boundary condition (function $v(t)$) on the right boundary:

$$u(l,t) = v(t), \qquad 0 < t \le T.$$

Exercise 8.8 Construct an additive difference scheme (splitting scheme) over spatial variables for the realization of the method with non-locally perturbed boundary conditions in solving the boundary value inverse problem in the rectangle Ω:

$$\frac{\partial u}{\partial t} - \sum_{\beta=1}^{2} \frac{\partial}{\partial x_\beta}\Big(k(\boldsymbol{x})\,\frac{\partial u}{\partial x_\beta}\Big) = 0, \qquad \boldsymbol{x} \in \Omega, \quad 0 < t < T,$$

$$\frac{\partial u}{\partial x_1}(0, x_2, t) = 0, \qquad u(0, x_2, t) + \alpha u(l_1, x_2, t) = \varphi(x_2, t),$$
$$u(x_1, 0, t) = 0, \qquad u(x_1, l_2, t) = 0, \qquad 0 < t < T,$$
$$u(\boldsymbol{x}, t) = 0, \qquad \boldsymbol{x} \in \Omega.$$

Exercise 8.9 The program PROBLEM17 implements an algorithm for iterative refinement of a second-kind boundary condition. Modify this program to realize an algorithm for iterative refinement of the third-kind boundary condition

$$\frac{\partial u}{\partial n}(\boldsymbol{x}, t) + \sigma u(\boldsymbol{x}, t) = \mu(x_1, t), \qquad \boldsymbol{x} \in \gamma.$$

Perform computational experiments to investigate the effect due to the numerical parameter $\sigma \geq 0$.

Exercise 8.10 Suppose that in the problem

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\Big(k(x, t)\,\frac{\partial u}{\partial x}\Big) = 0, \qquad 0 < x < l, \quad 0 < t \leq T,$$
$$u(0, t) = 0, \qquad u(l, t) = g(t), \qquad 0 < t \leq T,$$
$$u(x, 0) = 0, \qquad 0 \leq x \leq l$$

the coefficient $k(x, t)$ is a piecewise constant function, but the position of the interface between the medium is unknown. In the representation

$$k(x, t) = \begin{cases} k_1, & 0 < x < \gamma(t), \\ k_2, & \gamma(t) < x < l \end{cases}$$

the constants k_β, $\beta = 1, 2$ are given, and the function $\gamma(t)$ is to be determined from additional observations of the solution performed at the internal points $z_m \in \Omega$, $m = 1, 2, \ldots, M$:

$$u(z_m, t) \approx \varphi_m(t), \qquad 0 < t \leq T, \quad m = 1, 2, \ldots, M.$$

Examine the possibility of constructing gradient iterative methods for the approximate solution of this coefficient inverse problem.

Exercise 8.11 Examine the possibility of successive identification (local regularization) of the nonlinear coefficient $k(u)$ in the approximate solution of the inverse problem (8.132)–(8.136) using the parametric identification (8.147) in the class of piecewise linear functions.

Exercise 8.12 Modify the program PROBLEM18 using the gradient iterative method for determining the piecewise constant coefficient $k(u)$ instead of the golden-section method applied to minimize the discrepancy functional on an individual subinterval. Compare the computational efficiencies of these approaches as exemplified by the solutions of model problems.

Exercise 8.13 Consider the matter of solution unicity for the coefficient inverse problem on finding the function pair $\{u(\boldsymbol{x}),\ k(x_2)\}$ from the conditions

$$\sum_{\beta=1}^{2} \frac{\partial}{\partial x_\beta}\Big(k(x_2)\,\frac{\partial u}{\partial x_\beta}\Big) = 0, \qquad \boldsymbol{x} \in \Omega,$$

$$u(\boldsymbol{x}) = \varphi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega,$$

$$\frac{\partial u}{\partial n}(\boldsymbol{x}) = \psi(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial\Omega.$$

Exercise 8.14 Consider on the difference level the coefficient inverse problem formulated in the previous exercise. Obtain the gradient of the discrepancy functional under the assumption that the first-kind boundary conditions are given exactly and that the second-kind boundary conditions are given approximately.

Exercise 8.15 Using the program `PROBLEM19`, perform numerical experiments on the determination of the lowest coefficient $c(x_2)$ in equation (8.165) with additional information available not on the whole boundary $\partial\Omega$, but only on some individual segments of the boundary, namely, on the sides Γ_β, $\beta = 1, 2, 3, 4$.

Bibliography

Alifanov, O. M. (1995). *Inverse Heat Transfer Problems*. Springer-Verlag Telos.

Alifanov, O. M., Artiukhin, E. A., and Rumiantsev, S. V. (1995). *Extreme Methods for Solving Ill-Posed Problems With Applications to Inverse Heat Transfer Problems*. Begell House Publishers.

Bakushinskii, A. B., and Goncharskii, A. V. (1989). *Iterative Solution Methods for Ill-Posed Problems*. Nauka, Moscow.

Bakushinskii, A. B., and Goncharskii, A. V. (1989). *Ill-Posed Problems: Numerical Methods*. Moscow State Univ. Publ., Moscow.

Denisov, A. M. (1999). *Elements of the Theory of Inverse Problems (Inverse and Ill-Posed Problems)*. Brill Academic Publishers.

Ivanov, V. K., Vasin, V. V., and Tanana, V. P. (2002). *Theory of Linear Ill-Posed Problems and Its Applications*. Brill Academic Publishers.

Isakov, V. (1997). *Inverse Problems for Partial Differential Equations*. Springer-Verlag.

Lavrent'ev, M. M., Reznitskaya, K. G., and Yakhno, V. G. (1982). *One-Dimensional Ill-Posed Problems in Mathematical Physics*. Nauka, Novosibirsk.

Lavrent'ev, M. M., Romanov, V. G., and Shishatskii, S. P. (1980). *Ill-Posed Problems in Mathematical Physics*. Nauka, Moscow.

Lattes, R., and Lions, J.-L. (1969). *Method of Quasireversibility*. Elsevier.

Lions, J.-L. (1972). *Optimal Control of Systems Governed by Partial Differential Equations*. Mir, Moscow.

Marchuk, G. I. (1989). *Computational-Mathematics Methods*. Nauka, Moscow.

Morozov, V. A. (1987). *Regularization Methods for Non-Stationary Problems*. Moscow State Univ. Publ., Moscow.

Osipov, Yu. S., Vasil'ev, F. P., and Potapov, M. M. (1999). *Basics of the Dynamic Regularization Method*. Moscow State Univ. Publ., Moscow.

Prilepko, A. I., Orlovsky, D. G., and Vasin, I. A. (2000). *Methods for Solving Inverse Problems in Mathematical Physics*. Marcel Dekker.

Romanov, V. G. (2002). *Investigation Methods for Inverse Problems*. Brill Academic Publishers.

Romanov, V. G., and Kabanikhin, S. I. (1991). *Inverse Problems in Geoelectricity*. Nauka, Moscow.

Samarskii, A. A. (2001). *Theory of Difference Schemes*. Marcel Dekker.

Samarskii, A. A., and Andreev, V. B. (1976). *Difference Methods for Elliptic Equations*. Nauka, Moscow.

Samarskii, A. A., and Gulin, A. V. (1973). *Stability of Difference Schemes*. Nauka, Moscow.

Samarskii, A. A., and Nikolaev, E. (2001). *Numerical Methods for Grid Equations, Volume II*. Birkhäuser.

Samarskii, A. A., and Vabishchevich, P. N. (1995). *Computational Heat Transfer. Vol. 2. The Finite Difference Methodology*. Wiley, Chichester.

Samarskii, A. A., and Vabishchevich, P. N. (1997). Difference solution methods for inverse mathematical-physics problems. In: *Fundamentals of Mathematical Modeling*. Nauka, Moscow, 5–97.

Tikhonov, A. N., and Arsenin, V. Ya. (1986). *Solution Methods for Ill-Posed Problems*. Nauka, Moscow.

Tikhonov, A. N., Goncharskii, A. V., Stepanov, V. V., and Yagola, A. G. (1983). *Regularizing Algorithms and A Priori Information*. Nauka, Moscow.

Tikhonov, A. N., Goncharskii, A. V., Stepanov, V. V., and Yagola, A. G. (1995). *Numerical Methods for the Solution of Ill-Posed Problems*. Springer-Verlag.

Tikhonov, A. N., Leonov, A. S., and Yagola, A. G. (1997). *Nonlinear Ill-posed Problems*. CRC Press.

Tikhonov, A. N., and Samarskii, A. A. (1990). *Equations of Mathematical Physics*. Dover Publications.

Vabishchevich, P. N., and Samarskii, A. A. (1999). *Additive Schemes for Mathematical-Physics Problems*. Nauka, Moscow.

Vainikko, G. M. and A.Yu. Veretennikov. (1986). *Iterative Procedures in Ill-Posed Problems*. Nauka, Moscow.

Vasil'ev, F. P. (1981). *Solution Methods for Extremal Problems*. Nauka, Moscow.

Index