$$\lim_{h\to 0}\frac{\omega(u,h)}{\|h\|}=0$$

$$A_0 u=-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_j}\left(A_{i,j}\frac{\partial u}{\partial x_j}\right)=$$

Methods for Solving Mathematical Physics Problems

$$u(A)=\int_L \rho(P)\frac{\partial}{\partial n}\ln\left(\frac{1}{r}\right)dl=\int_L\frac{\rho(P)\cos\theta}{r}dl$$

V. I. Agoshkov, P. B. Dubovski
and V. P. Shutyaev

$$\sum_{k=0}^{n}F_k u^k\to F(u$$

CISP
Cambridge International Science Publishing

METHODS FOR SOLVING MATHEMATICAL PHYSICS PROBLEMS

METHODS FOR SOLVING MATHEMATICAL PHYSICS PROBLEMS

V.I. Agoshkov, P.B. Dubovski, V.P. Shutyaev

CAMBRIDGE INTERNATIONAL SCIENCE PUBLISHING

Published by

Cambridge International Science Publishing
7 Meadow Walk, Great Abington, Cambridge CB1 6AZ, UK
http://www.cisp-publishing.com

First published October 2006

© V.I. Agoshkov, P.B. Dubovskii, V.P. Shutyaev
© Cambridge International Science Publishing

Conditions of sale
All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 10: 1-904602-05-3
ISBN 13: 978-1-904602-05-7

Cover design Terry Callanan
Printed and bound in the UK by Lightning Source (UK) Ltd

Preface

The aim of the book is to present to a wide range of readers (students, postgraduates, scientists, engineers, etc.) basic information on one of the directions of mathematics, *methods for solving mathematical physics problems*.

The authors have tried to select for the book methods that have become classical and generally accepted. However, some of the current versions of these methods may be missing from the book because they require special knowledge.

The book is of the handbook-teaching type. On the one hand, the book describes the main definitions, the concepts of the examined methods and approaches used in them, and also the results and claims obtained in every specific case. On the other hand, proofs of the majority of these results are not presented and they are given only in the simplest (methodological) cases.

Another special feature of the book is the inclusion of many examples of application of the methods for solving specific mathematical physics problems of applied nature used in various areas of science and social activity, such as power engineering, environmental protection, hydrodynamics, elasticity theory, etc. This should provide additional information on possible applications of these methods.

To provide complete information, the book includes a chapter dealing with the main problems of mathematical physics, together with the results obtained in functional analysis and boundary-value theory for equations with partial derivatives.

Chapters 1, 5 and 6 were written by V.I. Agoshkov, chapters 2 and 4 by P.B. Dubovski, and chapters 3 and 7 by V.P. Shutyaev. Each chapter contains a bibliographic commentary for the literature used in writing the chapter and recommended for more detailed study of the individual sections.

The authors are deeply grateful to the editor of the book G.I. Marchuk, who has supervised for many years studies at the Institute of Numerical Mathematics of the Russian Academy of Sciences in the area of computational mathematics and mathematical modelling methods, for his attention to this work, comments and wishes.

The authors are also grateful to many colleagues at the Institute for discussion and support.

Contents

PREFACE

1. MAIN PROBLEMS OF MATHEMATICAL PHYSICS 1

Main concepts and notations 1
1. Introduction 2
2. Concepts and assumptions from the theory of functions and functional analysis 3
2.1. Point sets. Class of functions $C^p(\Omega)$, $C^p(\overline{\Omega})$ 3
2.1.1. Point Sets 3
2.1.2. Classes $C^p(\Omega)$, $C^p(\overline{\Omega})$ 4
2.2. Examples from the theory of linear spaces 5
2.2.1. Normalised space 5
2.2.2. The space of continuous functions $C(\overline{\Omega})$ 6
2.2.3. Spaces $C^\lambda(\Omega)$ 6
2.2.4. Space $L_p(\Omega)$ 7
2.3. $L_2(\Omega)$ Space. Orthonormal systems 9
2.3.1. Hilbert spaces 9
2.3.2. Space $L_2(\Omega)$ 11
2.3.3. Orthonormal systems 11
2.4. Linear operators and functionals 13
2.4.1. Linear operators and functionals 13
2.4.2. Inverse operators 15
2.4.3. Adjoint, symmetric and self-adjoint operators 15
2.4.4. Positive operators and energetic space 16
2.4.5. Linear equations 17
2.4.6. Eigenvalue problems 17
2.5. Generalized derivatives. Sobolev spaces 19
2.5.1. Generalized derivatives 19
2.5.2. Sobolev spaces 20
2.5.3. The Green formula 21
3. Main equations and problems of mathematical physics 22
3.1. Main equations of mathematical physics 22
3.1.1. Laplace and Poisson equations 23
3.1.2. Equations of oscillations 24
3.1.3. Helmholtz equation 26
3.1.4. Diffusion and heat conduction equations 26
3.1.5. Maxwell and telegraph equations 27
3.1.6. Transfer equation 28
3.1.7. Gas- and hydrodynamic equations 29
3.1.8. Classification of linear differential equations 29

3.2. Formulation of the main problems of mathematical physics 32
3.2.1. Classification of boundary-value problems 32
3.2.2. The Cauchy problem 33
3.2.3. The boundary-value problem for the elliptical equation 34
3.2.4. Mixed problems 35
3.2.5. Validity of formulation of problems. Cauchy–Kovalevskii theorem 35
3.3. Generalized formulations and solutions of mathematical physics problems ... 37
3.3.1. Generalized formulations and solutions of elliptical problems 38
3.3.2. Generalized formulations and solution of hyperbolic problems 41
3.3.3. The generalized formulation and solutions of parabolic problems 43
3.4. Variational formulations of problems 45
3.4.1. Variational formulation of problems in the case of positive definite ..... operators 45
3.4.2. Variational formulation of the problem in the case of positive operators . 46
3.4.3. Variational formulation of the basic elliptical problems 47
3.5. Integral equations 49
3.5.1. Integral Fredholm equation of the 1st and 2nd kind 49
3.5.2. Volterra integral equations 50
3.5.3. Integral equations with a polar kernel 51
3.5.4. Fredholm theorem 51
3.5.5. Integral equation with the Hermitian kernel 52
Bibliographic commentary 54

2. METHODS OF POTENTIAL THEORY 56
Main concepts and designations 56
1. Introduction 57
2. Fundamentals of potential theory 58
2.1. Additional information from mathematical analysis 58
2.1.1 Main orthogonal coordinates 58
2.1.2. Main differential operations of the vector field 58
2.1.3. Formulae from the field theory 59
2.1.4. Main properties of harmonic functions 60
2.2 Potential of volume masses or charges 61
2.2.1. Newton (Coulomb) potential 61
2.2.2. The properties of the Newton potential 61
2.2.3. Potential of a homogeneous sphere 62
2.2.4. Properties of the potential of volume-distributed masses 62
2.3. Logarithmic potential 63
2.3.1. Definition of the logarithmic potential 63
2.3.2. The properties of the logarithmic potential 63
2.3.3. The logarithmic potential of a circle with constant density 64
2.4. The simple layer potential 64
2.4.1. Definition of the simple layer potential in space 64
2.4.2. The properties of the simple layer potential 65
2.4.3. The potential of the homogeneous sphere 66
2.4.4. The simple layer potential on a plane 66

2.5. Double layer potential 67
2.5.1. Dipole potential 67
2.5.2. The double layer potential in space and its properties 67
2.5.3. The logarithmic double layer potential and its properties 69
3. Using the potential theory in classic problems of mathematical physics 70
3.1. Solution of the Laplace and Poisson equations 70
3.1.1. Formulation of the boundary-value problems of the Laplace equation 70
3.1.2 Solution of the Dirichlet problem in space 71
3.1.3. Solution of the Dirichlet problem on a plane 72
3.1.4. Solution of the Neumann problem 73
3.1.5. Solution of the third boundary-value problem for the Laplace equation 74
3.1.6. Solution of the boundary-value problem for the Poisson equation 75
3.2. The Green function of the Laplace operator 76
3.2.1. The Poisson equation 76
3.2.2. The Green function 76
3.2.3. Solution of the Dirichlet problem for simple domains 77
3.3 Solution of the Laplace equation for complex domains 78
3.3.1. Schwarz method 78
3.3.2. The sweep method 80
4. Other applications of the potential method 81
4.1. Application of the potential methods to the Helmholtz equation 81
4.1.1. Main facts 81
4.1.2. Boundary-value problems for the Helmholtz equations 82
4.1.3. Green function 84
4.1.4. Equation $\Delta v - \lambda v = 0$ 85
4.2. Non-stationary potentials 86
4.2.1 Potentials for the one-dimensional heat equation 86
4.2.2. Heat sources in multidimensional case 88
4.2.3. The boundary-value problem for the wave equation 90
Bibliographic commentary 92

3. EIGENFUNCTION METHODS 94
Main concepts and notations 94
1. Introduction 94
2. Eigenvalue problems 95
2.1. Formulation and theory 95
2.2. Eigenvalue problems for differential operators 98
2.3. Properties of eigenvalues and eigenfunctions 99
2.4. Fourier series 100
2.5. Eigenfunctions of some one-dimensional problems 102
3. Special functions 103
3.1. Spherical functions 103
3.2. Legendre polynomials 105
3.3. Cylindrical functions 106
3.4. Chebyshef, Laguerre and Hermite polynomials 107
3.5. Mathieu functions and hypergeometrical functions 109

4. Eigenfunction method 110
4.1. General scheme of the eigenfunction method 110
4.2. The eigenfunction method for differential equations of mathematical physics 111
4.3. Solution of problems with nonhomogeneous boundary conditions 114
5. Eigenfunction method for problems of the theory of electromagnetic phenomena 115
5.1. The problem of a bounded telegraph line 115
5.2. Electrostatic field inside an infinite prism 117
5.3. Problem of the electrostatic field inside a cylinder 117
5.4. The field inside a ball at a given potential on its surface 118
5.5 The field of a charge induced on a ball 120
6. Eigenfunction method for heat conductivity problems 121
6.1. Heat conductivity in a bounded bar 121
6.2. Stationary distribution of temperature in an infinite prism 122
6.3. Temperature distribution of a homogeneous cylinder 123
7. Eigenfunction method for problems in the theory of oscillations 124
7.1. Free oscillations of a homogeneous string 124
7.2. Oscillations of the string with a moving end 125
7.3. Problem of acoustics of free oscillations of gas 126
7.4. Oscillations of a membrane with a fixed end 127
7.5. Problem of oscillation of a circular membrane 128
Bibliographic commentary 129

4. METHODS OF INTEGRAL TRANSFORMS 130
Main concepts and definitions 130
1. Introduction 131
2. Main integral transformations 132
2.1. Fourier transform 132
2.1.1. The main properties of Fourier transforms 133
2.1.2. Multiple Fourier transform 134
2.2. Laplace transform 134
2.2.1. Laplace integral 134
2.2.2. The inversion formula for the Laplace transform 135
2.2.3. Main formulae and limiting theorems 135
2.3. Mellin transform 135
2.4. Hankel transform 136
2.5. Meyer transform 138
2.6. Kontorovich–Lebedev transform 138
2.7. Meller–Fock transform 139
2.8 Hilbert transform 140
2.9. Laguerre and Legendre transforms 140
2.10 Bochner and convolution transforms, wavelets and chain transforms 141
3. Using integral transforms in problems of oscillation theory 143
3.1. Electrical oscillations 143
3.2. Transverse vibrations of a string 143

3.3. Transverse vibrations of an infinite circular membrane 146
4. Using integral transforms in heat conductivity problems 147
4.1. Solving heat conductivity problems using the Laplace transform 147
4.2. Solution of a heat conductivity problem using Fourier transforms 148
4.3. Temperature regime of a spherical ball 149
5. Using integral transformations in the theory of neutron diffusion 149
5.1. The solution of the equation of deceleration of neutrons for a moderator of infinite dimensions 150
5.2. The problem of diffusion of thermal neutrons 150
6. Application of integral transformations to hydrodynamic problems 151
6.1. A two-dimensional vortex-free flow of an ideal liquid 151
6.2. The flow of the ideal liquid through a slit 152
6.3. Discharge of the ideal liquid through a circular orifice 153
7. Using integral transforms in elasticity theory 155
7.1. Axisymmetric stresses in a cylinder 155
7.2. Bussinesq problem for the half space 157
7.3. Determination of stresses in a wedge 158
8. Using integral transforms in coagulation kinetics 159
8.1. Exact solution of the coagulation equation 159
8.2. Violation of the mass conservation law 161
Bibliographic commentary 162

5. METHODS OF DISCRETISATION OF MATHEMATICAL PHYSICS PROBLEMS 163
Main definitions and notations 163
1. Introduction 164
2. Finite-difference methods 166
2.1. The net method 166
2.1.1. Main concepts and definitions of the method 166
2.1.2. General definitions of the net method. The convergence theorem 170
2.1.3. The net method for partial differential equations 173
2.2. The method of arbitrary lines 182
2.2.1. The method of arbitrary lines for parabolic-type equations 182
2.2.2. The method of arbitrary lines for hyperbolic equations 184
2.2.3. The method of arbitrary lines for elliptical equations 185
2.3. The net method for integral equations (the quadrature method) 187
3. Variational methods 188
3.1. Main concepts of variational formulations of problems and variational methods 188
3.1.1. Variational formulations of problems 188
3.1.2. Concepts of the direct methods in calculus of variations 189
3.2. The Ritz method 190
3.2.1. The classic Ritz method 190
3.2.2. The Ritz method in energy spaces 192
3.2.3. Natural and main boundary-value conditions 194
3.3. The method of least squares 195

3.4. Kantorovich, Courant and Trefftz methods 196
3.4.1. The Kantorovich method 196
3.4.2. Courant method 196
3.4.3. Trefftz method 197
3.5. Variational methods in the eigenvalue problem 199
4. Projection methods 201
4.1. The Bubnov–Galerkin method 201
4.1.1. The Bubnov-Galerkin method (a general case) 201
4.1.2 The Bubnov–Galerkin method ($A = A_0 + B$) 202
4.2. The moments method 204
4.3. Projection methods in the Hilbert and Banach spaces 205
4.3.1. The projection method in the Hilbert space 205
4.3.2. The Galerkin–Petrov method 206
4.3.3. The projection method in the Banach space 206
4.3.4. The collocation method 208
4.4. Main concepts of the projection-grid methods 208
5. Methods of integral identities 210
5.1. The main concepts of the method 210
5.2. The method of Marchuk's integral identity 211
5.3. Generalized formulation of the method of integral identities 213
5.3.1. Algorithm of constructing integral identities 213
5.3.2. The difference method of approximating the integral identities 214
5.3.3. The projection method of approximating the integral identities 215
5.4. Applications of the methods of integral identities in mathematical physics problems 217
5.4.1. The method of integral identities for the diffusion equation 217
5.4.2. The solution of degenerating equations 219
5.4.3. The method of integral identities for eigenvalue problems 221
Bibliographic Commentary 223

6. SPLITTING METHODS 224
1. Introduction 224
2. Information from the theory of evolution equations and difference schemes . 225
2.1. Evolution equations 225
2.1.1. The Cauchy problem 225
2.1.2. The nonhomogeneous evolution equation 228
2.1.3. Evolution equations with bounded operators 229
2.2. Operator equations in finite-dimensional spaces 231
2.2.1. The evolution system 231
2.2.2. Stationarisation method 232
2.3. Concepts and information from the theory of difference schemes 233
2.3.1. Approximation 233
2.3.2. Stability 239
2.3.3. Convergence 240
2.3.4. The sweep method 241
3. Splitting methods 242

3.1. The method of component splitting (the fractional step methods) 243
3.1.1. The splitting method based on implicit schemes of the first order of ..... accuracy 243
3.1.2. The method of component splitting based on the Cranck–Nicholson ..... schemes 243
3.2. Methods of two-cyclic multi-component splitting 245
3.2.1. The method of two-cyclic multi-component splitting 245
3.2.2. Method of two-cyclic component splitting for quasi-linear problems 246
3.3. The splitting method with factorisation of operators 247
3.3.1. The implicit splitting scheme with approximate factorisation of the operator 247
3.3.2. The stabilisation method (the explicit–implicit schemes with approximate factorisation of the operator) 248
3.4. The predictor–corrector method 250
3.4.1. The predictor–corrector method. The case $A = A_1 + A_2$. 250
3.4.2. The predictor–corrector method. Case $A = \sum_{\alpha=1}^{n} A_\alpha$. 251
3.5. The alternating-direction method and the method of the stabilising correction 252
3.5.1. The alternating-direction method 252
3.5.2. The method of stabilising correction 253
3.6. Weak approximation method 254
3.6.1. The main system of problems 254
3.6.2. Two-cyclic method of weak approximation 254
3.7. The splitting methods – iteration methods of solving stationary problems . 255
3.7.1. The general concepts of the theory of iteration methods 255
3.7.2. Iteration algorithms 256
4. Splitting methods for applied problems of mathematical physics 257
4.1. Splitting methods of heat conduction equations 258
4.1.1. The fractional step method 258
4.2.1. Locally one-dimensional schemes 259
4.1.3. Alternating-direction schemes 260
4.2. Splitting methods for hydrodynamics problems 262
4.2.1. Splitting methods for Navier–Stokes equations 262
4.2.2. The fractional steps method for the shallow water equations 263
4.3. Splitting methods for the model of dynamics of sea and ocean flows 268
4.3.1. The non-stationary model of dynamics of sea and ocean flows 268
4.3.2. The splitting method 270
Bibliographic Commentary 272

7. METHODS FOR SOLVING NON-LINEAR EQUATIONS 273
Main concepts and Definitions 273
1. Introduction 274
2. Elements of nonlinear analysis 276
2.1. Continuity and differentiability of nonlinear mappings 276
2.1.1. Main definitions 276

2.1.2. Derivative and gradient of the functional 277
2.1.3. Differentiability according to Fréchet 278
2.1.4. Derivatives of high orders and Taylor series 278
2.2. Adjoint nonlinear operators 279
2.2.1. Adjoint nonlinear operators and their properties 279
2.2.2. Symmetry and skew symmetry 280
2.3. Convex functionals and monotonic operators 280
2.4. Variational method of examining nonlinear equations. 282
2.4.1. Extreme and critical points of functionals 282
2.4.2. The theorems of existence of critical points 282
2.4.3. Main concept of the variational method 283
2.4.4. The solvability of the equations with monotonic operators 283
2.5 Minimising sequences 284
2.5.1. Minimizing sequences and their properties 284
2.5.2. Correct formulation of the minimisation problem 285
3. The method of the steepest descent 285
3.1. Non-linear equation and its variational formulation 285
3.2. Main concept of the steepest descent methods 286
3.3. Convergence of the method 287
4. The Ritz method 288
4.1. Approximations and Ritz systems 289
4.2. Solvability of the Ritz systems 290
4.3. Convergence of the Ritz method 291
5. The Newton–Kantorovich method 291
5.1. Description of the Newton iteration process 291
5.2. The convergence of the Newton iteration process 292
5.3. The modified Newton method 292
6. The Galerkin–Petrov method for non-linear equations 293
6.1. Approximations and Galerkin systems 293
6.2. Relation to projection methods 294
6.3. Solvability of the Galerkin systems 295
6.4. The convergence of the Galerkin–Petrov method 295
7. Perturbation method 296
7.1. Formulation of the perturbation algorithm 296
7.2. Justification of the perturbation algorithms 299
7.3. Relation to the method of successive approximations 301
8. Applications to some problem of mathematical physics 302
8.1. The perturbation method for a quasi-linear problem of non-stationary heat conduction 302
8.2. The Galerkin method for problems of dynamics of atmospheric processes .. 306
8.3. The Newton method in problems of variational data assimilation 308
Bibliographic Commentary 311
Index 317

Chapter 1

MAIN PROBLEMS OF MATHEMATICAL PHYSICS

Keywords: point sets, linear spaces, Banach space, Hilbert space, orthonormal systems, linear operators, eigenvalues, eigenfunctions, generalised derivatives, Sobolev spaces, main problems of mathematical physics, Laplace equation, Poisson equation, oscillation equation, Helmholtz equation, diffusion equation, heat conductivity equation, Maxwell equations, telegraph equations, transfer equation, equations of gas and hydrodynamics, boundary conditions, initial conditions, classification of equations, formulation of problems, generalised solution, variational formulation of problems, integral equations, Fredholm theorem, Hilbert–Schmidt theorem.

MAIN CONCEPTS AND NOTATIONS

Domain	– open connected set.
Compact	– closed bounded set.
R^n	– n-dimensional Euclidean space.
$\partial\Omega$	– the boundary of the bounded set Ω.
$\|\|f\|\|_X$	– the norm of element f from the normalised space X.
$C(T)$	– Banach space of functions continuous on T.
$C^p(T)$	– Banach space of functions, continuous on T together with derivatives of the p-th order.
$C^\lambda(T)$, $0<\lambda<1$	– space of continuous Hölder function.
$L_2(\Omega)$	– the Hilbert space of functions, quadratically integrated according to Lebesgue.
$L_p(\Omega)$, $1\le p<\infty$	– Banach space with norm. $$\|f\|_p \equiv \|f\|_{L_p(\Omega)} = \left(\int_\Omega |f|^p\,dx\right)^{1/p}$$
$L_\infty(\Omega)$	– Banach space with the norm $\|\|f\|\|_{L_\infty}(\Omega)=\text{supvrai}_{\chi\in\Omega}\|f(x)\|$.
$W_p^1(\Omega)$, $1\le p<\infty$,	– Sobolev space, consisting of functions $f(x)$ with the generalized derivatives up to the order of l.

$C^{\infty}(\Omega)$ – the set of functions infinitely differentiated in Ω.

$C_0^{\infty}(\Omega)$ – the set of functions infinitely differentiated and finite in Ω.

supp f – support of $f(x)$.

$D(A)$ – the domain of definition of operator A.

$R(A)$ – the domain of the values of operator A, the range.

$\overline{f}(x)$ – the function complexly adjoint with $f(x)$.

$L(X,Y)$ – the space of linear continuous operators, acting from space X to space Y.

Eigenvalue – the numerical parameter λ which together with the eigenfunction φ is the solution of the equation $A\varphi=\lambda\varphi$.

$\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$ – the Laplace operator.

$\Box_a = \frac{\partial^2}{\partial t^2} - a^2\Delta$ – D'Alembert operator.

$D(f) = \|\nabla f\|^2_{L_2(\Omega)} = \sum_{i=1}^{n} \int_{\Omega} \left(\frac{\partial f}{\partial x_i}\right)^2 dx$ – Dirichlet integral.

$\int_{\Omega} K(x,y)u(y)dy = f(x)$ – the Fredholm equation of the first kind.

$u(x) = \lambda \int_{\Omega} K(x,y)u(y)dy + f(x)$ – the Fredholm equation of the second kind.

1. INTRODUCTION

Mathematical physics examines mathematical models of physical phenomena. Mathematical physics and its methods were initially developed in the 18th century in examination of the oscillations of strings and bars, in the problems of acoustics, hydrodynamics, analytical mechanics (J. D'Alembert, L. Euler, J. Lagrange, D. Bernoulli, P. Laplace). The concepts of mathematical physics were again developed more extensively in the 19th century in connection with the problems of heat conductivity, diffusion, elasticity, optics, electrodynamics, nonlinear wave processes, theories of stability of motion (J. Fourier, S. Poisson, K. Gauss, A. Cauchy, M.V. Ostrogradskii, P. Dirichlet, B. Riemann, S.V. Kovalevskaya, G. Stokes, H. Poincaré, A.M. Lyapunov, V.S. Steklov, D. Hilbert). A new stage of mathematical physics started in the 20th century when it included the problems of the theory of relativity, quantum physics, new problems of gas dynamics, kinetic equations, theory of nuclear reactors, plasma physics (A. Einstein, N.N. Bogolyubov, P. Dirac, V.S. Vladimirov, V.P. Maslov).

Many problems of classic mathematical physics are reduced to boundary-value problems for differential (integro-differential) equations – equations of mathematical physics which together with the appropriate boundary (or initial and boundary) conditions form mathematical models of the investigated physical processes.

The main classes of these problems are elliptical, hyperbolic, parabolic

problems and the Cauchy problem. Classic and generalised formulations are distinguished in the group of formulation of these problems. An important concept of the generalised formulation of the problems and generalised solutions is based on the concept of the generalised derivative using the Sobolev space.

One of the problems, examined in mathematical physics, is the problem of eigenvalues. Eigenfunctions of specific operators and of expansion of solutions of problems into Fourier series can often be used in theoretical analysis of problems, and also to solve them (the eigenfunction method).

The main mathematical means of examining the problems of mathematical physics is the theory of differential equations with partial derivatives, integral equations, theory of functions and functional spaces, functional analysis, approximate methods and computing mathematics.

Here, we present information from a number of sections of mathematics used in examination of the problems of mathematical physics and methods of solving them [13, 25, 69, 70, 75, 84, 91, 95].

2. CONCEPTS AND ASSUMPTIONS FROM THE THEORY OF FUNCTIONS AND FUNCTIONAL ANALYSIS

2.1. Point sets. Class of functions $C^p(\Omega)$, $C^p(\overline{\Omega})$

2.1.1. Point Sets

Let R^n ($\mathsf{R}^1 = \mathsf{R}$) be the n-dimensional real Euclidean space, $x = (x_1,\ldots,x_n)$ – is the point in R^n, where x_i, $i = 1,2, \ldots,n$, are the coordinates of point x. *The scalar product and the norm (length)* in R^n are denoted by respectively

$(x,y) = \sum_{i=1}^{n} x_i\, y_i$, $|x| = (x,x)^{1/2} = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$. Consequently, the number $|x–y|$ is the Euclidean distance between the points x and y.

A set of points x from R^n, satisfying the inequality $|x–x_0| < R$, is *an open sphere* with radius R with the centre at point x_0. This sphere is denoted by $U(x_0;R)$, $U_R = U(0;R)$.

The set is referred to as bounded in R^n, if there is a sphere containing this set.

Point x_0 is referred to as *the internal point of the set*, if there is a sphere, $U(x_0;\varepsilon)$, present in this set. The set is referred to *as open* if all its points are *internal*. A set is referred to as *connected* if any two points in this set may be connected by a piecewise smooth curve, located in this set. The connected open set is referred to the domain. Point x_0 is referred to as the *limiting point* of set A, if there is a sequence x_k, k =1,2,..., such that $x_k \in A$, $x_k \neq x_0$, $x_k \to x_0$, $k \to \infty$. If all limiting points are added to the set A, the resultant set is referred to as *closure* of set A and denoted by $\overline{A}$. If the set coincides with its closure, it is referred to as *closed*. The closed bounded set is referred to as *a compact*. The *neighbourhood* of the set A is any open set containing A; ε-*neighbourhood* A_ε of the set A is the in-

tegration of spheres $U(x;\varepsilon)$, when x outstrips A: $A_\varepsilon = U_{x\in A}U(x;\varepsilon)$.

Function $\chi_A(x)$, equal to 1 at $x\in A$ and 0 at $x\notin A$, and is referred to as the *characteristic function* of set A.

Let Ω be a domain. The closure points $\bar{\Omega}$, not belonging to Ω, form the closed set $\partial\Omega$, referred to as the *boundary of the domain* Ω, since $\partial\Omega = \bar{\Omega}\setminus\Omega$.

We shall say that the surface $\partial\Omega$ belongs to the *class* C^p, $p\geq 1$, if in some neighbourhood of every point $x_0 \in \partial\Omega$ it is represented by the equation $\omega_{x0}(x) = 0$ and grad $\omega_{x0}(x) \neq 0$, and the function $\omega_{x0}(x)$ is continuous together with all derivatives to order p inclusive in the given neighbourhood. The surface $\partial\Omega$ is referred to as *piecewise smooth*, if it consists of a finite number of surfaces of class C^1. If the function $x_0 \in \partial\Omega$ in the vicinity of any point $\omega_{x0}(x)$ satisfies the Lipschitz condition $|\omega_{x0}(x)-\omega_{x0}(y)|\leq C|x-y|$, C = const, then $\partial\Omega$ is the Lipschitz boundary of domain Ω.

If $\partial\Omega$ is a piecewise smooth boundary of class C^1 (or even Lipschitz boundary) then almost at all points $x \in \partial\Omega$ there is the unit vector of the external normal $n(x)$ to $\partial\Omega$.

It is assumed that point x_0 is situated on the piecewise smooth surface $\partial\Omega$. The *neighbourhood of the point* x_0 on the surface $\partial\Omega$ is the connected part of the set $\partial\Omega\cap U(x_0;R)$ which contains point x_0.

The bounded domain Ω' is referred to as a subdomain strictly situated in the domain Ω if $\bar{\Omega}'\subset\Omega$; in this case, it is written that $\Omega'\subset\Omega$.

2.1.2. Classes $C^p(\Omega)$, $C^p(\bar{\Omega})$

Let $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_n)$ be the integer vector with non-negative components α_j (multi-index). $D^\alpha f(x)$ denotes the derivative of the function $f(x)$ of the order $|\alpha|=\alpha_1+\alpha_2+\ldots+\alpha_n$:

$$D^\alpha f(x) = D_1^{\alpha_1}\ldots D_n^{\alpha_n} f(x) = \frac{D^{|\alpha|} f(x_1,x_2,\ldots,x_n)}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\ldots\partial x_n^{\alpha_n}},\ D^0 f(x) = f(x);$$

$$D = (D_1, D_2,\ldots,D_n),\ D_j = \frac{\partial}{\partial x_j},\ j = 1,2,\ldots,n.$$

For lower derivatives it is common to use the notations $f_{x_i}, f_{x_i x_j}$. The following shortened notations are also used:

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2}\ldots x_n^{\alpha_n},\quad \alpha! = \alpha_1!\,\alpha_2!\ldots\alpha_n!.$$

The set of (complex-valued) functions f, which are continuous together with derivatives $D^\alpha f(x)$, $|\alpha|\leq p (0\leq p<\infty)$ in domain Ω, form the *class of functions* $C^p(\Omega)$. Function f of class $C^p(\Omega)$ in which all derivatives $D^\alpha f(x)$, $|\alpha|\leq p$, permit continuous continuation to closure $\bar{\Omega}$, form the class of functions $C^p(\bar{\Omega})$; in this case, the value $D^\alpha f(x)$, $x \in\in \partial\Omega$, $|\alpha| \leq p$, indicates lim $D^\alpha f(x')$ at $x'\to x$, $x'\in \Omega$. The class of function belonging to $C^p(\Omega)$ at all p, is denoted by $C^\infty(\Omega)$; similarly, the class of functions $C^\infty(\bar{\Omega})$ is also determined. Class $C(\Omega) \equiv C^0(\Omega)$ consists of all continuous functions in Ω, and class $C(\bar{\Omega}) \equiv C^0(\bar{\Omega})$ may be regarded as identical with the set of all continuous functions $\bar{\Omega}$.

Let the function $f(x)$ be given on some set containing domain Ω. in this

case, the affiliation of f to the class $C^p(\bar{\Omega})$ shows that the *restriction* of f on Ω belongs to $C^p(\bar{\Omega})$.

The introduced classes of the functions are *linear sets*, i.e. from the affiliation of the functions f and g to some of these classes we obtain affiliation to the same class also of their linear combination $\lambda f+\mu g$, where λ and μ are arbitrary complex numbers.

The function f is piecewise continuous in R^n, if there is a finite or countable number of domains Ω_k, k =1,2, ..., without general points with piecewise smooth boundaries which are such that every sphere is covered by a finite number of closed domains $\{\bar{\Omega}_k\}$ and $f \in C(\bar{\Omega}_k)$, k = 1,2,...

The piecewise continuous function is referred to as *finite* if it does not revert to zero outside some sphere.

Let it be that $\varphi \in C(\mathrm{R}^n)$. The *support* supp φ of the continuous function φ is the closure of the set of the points for which $\varphi(x) \neq 0$.

$C_0^\infty(\mathrm{R}^n)$ denotes the set of infinitely differentiated functions with finite supports, and $C_0^\infty(\Omega)$ denotes functions whose supports belong to $\Omega\subset\mathrm{R}^n$.

2.2. Examples from the theory of linear spaces Spaces $C(\Omega)$, $C^\lambda(\Omega)$, $L_p(\Omega)$

2.2.1. Normalised space

Lets us assume that X is a linear set. It is said that the norm $||\cdot||_X$. is *introduced* on X if every element $f\in X$ is related to a non-negative number $||f||_X$ (*norm* f) so that the following three axioms are fulfilled:

a) $||f||_X \geq 0$; $||f||_X = 0$ if and only if $f = 0$;

b) $||\lambda f||_X, = |\lambda|||f||_X$,where λ – any complex number;

c) $||f+g||_X \leq ||f||_X+||g||_X$ (triangular inequality)

Any linear set having a norm is referred to as the *linear normalised space*.

Let X be a linear normalized space. Sequence $x_n \in X$ is *fundamental* (*converging in itself*) if for any $\varepsilon > 0$ there is such $N = N(\varepsilon)$ that for any $n > N$ and for all natural p the inequality $||f_{n+p}- f_n||< \varepsilon$ is satisfied. Space X is referred to as *total* if any fundamental sequence converges in this space. The total linear normalized space is referred to as *Banach space*.

Let it be that X is a linear normalized space. Set $A \subset X$ is referred to as *compact* if every sequence of its elements contains a sub-sequence converging to the element from X.

Two norms $|| f ||_1$ and $|| f ||_2$ in the linear space X are referred to as *equivalent* if there are such numbers $\alpha > 0$, $\beta > 0$ that for any $f \in X$ the inequality $\alpha|| f ||_1\leq|| f ||_2\leq\beta|| f ||_1$ is satified.

The linear normalized spaces X and Y are termed *isomorphous* if the image $J : X \to Y$ is defined on all X. This image is linear and carries out isomorphism X and Y as linear spaces and is such that there are constants $\alpha > 0$, $\beta > 0$, such that for any $f \in X$ the inequality $\alpha|| f ||_x\leq|| J(f) ||_Y \leq \beta|| f ||_X$ is fullfillied. If $||J(f)||_Y = || f ||_X$, the spaces X and Y are referred to as *isometric*.

The linear normalised space X is referred to as *inserted* into the linear normalised space Y if on all X the image $J : X \to Y$ is determined, this image is linear and mutually unambiguous on the domain of values, and there is such a constant $\beta > 0$ that for any $f \in X$ the inequality $\|J(f)\|_Y \le \beta \| f \|_X$ is satsified.

The Banach space $\hat{X}$ is the *supplement* of the linear normalized space X, if X is the linear manifold, dense everywhere in space $\hat{X}$.

Theorem 1. *Each linear normalized space X has a supplement, and this supplement is unique with the accuracy to the isometric image, converting X in itself.*

2.2.2. The space of continuous functions $C(\bar{\Omega})$

Let Ω be the domain from R^n. The set of functions continuous on $\bar{\Omega} = \partial\Omega \cup \Omega$ for which the norm

$$\|f\|_{C(\bar{\Omega})} = \sup_{x \in \bar{\Omega}} |f(x)|,$$

is finite, is referred to as the normalized space $C(\bar{\Omega})$. It is well known that the space $C(\bar{\Omega})$ is Banach space. Evidently, the convergence $f_k \to f$, $k \to \infty$, in $C(\bar{\Omega})$ is equivalent to the uniform convergence of the sequence of functions f_k, $k = 1,2,\ldots$, to the function $f(x)$ on the set $\bar{\Omega}$. The following theorem is also valid.

Theorem 2 (Weierstrass theorem). *If Ω is the bounded domain and $f \in C^p(\bar{\Omega})$, then for any $\varepsilon > 0$ there is a polynomial P such that*

$$\left\| D^\alpha f - D^\alpha P \right\|_C < \varepsilon \text{ for } |\alpha| \le P.$$

A series, consisting of the functions $f_k \in C(\bar{\Omega})$, is referred to as regularly converging on $\bar{\Omega}$, if a series of the absolute values $|f_k(x)|$ converges at $C(\bar{\Omega})$, i.e. converges uniformly on $\bar{\Omega}$.

The set $M \subset C(\bar{\Omega})$ is equicontinuous on $\bar{\Omega}$ if for any $\varepsilon > 0$ there is a number δ_ε which is such that the inequality $|f(x_1) - f(x_2)| < \varepsilon$ holds at all $f \in M$ as long as $|x_1 - x_2| < \delta_\varepsilon$, $x_1, x_2 \in \bar{\Omega}$.

The conditions of compactness of the set on $C(\bar{\Omega})$ are determined by the following theorem.

Theorem 3 (Arzelà–Ascoli theorem). *For the compactness of the set* $M \subset C(\bar{\Omega})$ *it is necessary and sufficient that it should be:*

a) *uniformly bounded, i.e.* $\| f \| \le K$ *for any function* $f \in M$;

b) *equicontinuous on* $\bar{\Omega}$.

2.2.3. Spaces $C^\lambda(\Omega)$

Let Ω be a bounded connected domain. We determine spaces $C^\lambda(\Omega)$, where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$, $0 < \lambda_i \le 1$, $i = 1,\ldots,n$. Let $e_i = (0,\ldots 0,1,0,\ldots,0)$, where unity stands on the i-th position. $[x_1, x_2]$ denotes a segment connecting points $x_1, x_2 \in E_n$. It is assumed that:

$$\Delta_i(h) f(x) = \begin{cases} f(x + he_i) - f(x) & \text{at } [x, x + e_i h] \subset \Omega, \\ 0 & \text{at } [x, x + e_i h] \not\subset \Omega. \end{cases}$$

We define the norm:

$$\|f\|_{C^{\lambda}(\Omega)} = \|f\|_{C(\Omega)} + \sum_{i=1}^{n} \sup_{x\in\Omega, |h|\le\delta} \frac{|\Delta_i(h) f(x)|}{|h|^{\lambda_i}}.$$

The set of the functions $f \in C(\bar{\Omega})$, for which the norm $\| f \|_{C^{\lambda}(\Omega)}$ is finite, forms the *Hölder space* $C^{\lambda}(\Omega)$. The Arzelà–Ascoli theorem shows that the set of the functions, bounded in $C^{\lambda}(\Omega)$, is compact in $C(\Omega_{\delta})$, where Ω_{δ} is the set of the points $x \in \Omega$ for which $\rho(x,\partial\Omega) \equiv \inf_{y\in\partial\Omega}|x-y| > \delta = \text{const}>0$.

If $\lambda_1=\ldots=\lambda_n=\lambda=1$, function $f(x)$ is *Lipschitz continuous on* Ω, ($f(x)$ is *Lipschitz function on* Ω).

2.2.4. Space $L_p(\Omega)$

The set $M\subset[a,b]$ has the *measure zero* if for any $\varepsilon>0$ there is such finite or countable system of segments $[\alpha_n,\beta_n]$ that $M\subset \cup_n[\alpha_n,\beta_n]$, $\sum_n(\beta_n-\alpha_n)<\varepsilon$. If for the sequence $f_n(t)$ ($n\in \mathrm{N}$) everywhere on $[a,b]$, with the exception of, possibly, the set of the measure zero, there is a limit equal to $f(t)$, it is then said that $f_n(t)$ converges to $f(t)$ almost everywhere on $[a,b]$, and we can write $\lim_{n\to\infty} f_n(t) \overset{\text{a.e.}}{=} f(t)$.

Let $\tilde{L}_1[a,b]$ be the space of functions continuous on $[a,b]$ with the norm

$$\|f\| = \int_a^b |f(t)|\,dt;$$

the convergence in respect of this norm is referred to as the *convergence in the mean*. The space $\tilde{L}_1[a,b]$ is not complete; its completion is referred to as the *Lebesgue space* and denoted by $L_1[a,b]$. Function $f(t)$ is referred to as *integrable in respect to Lebesgue* on the segment $[a,b]$ if there is such a fundamental in the mean sequence of continuous functions $f_n(t)$ ($n \in N$), so that

$$\lim_{n\to\infty} f_n(t) \overset{\text{a.e}}{=} f(t).$$

Therefore, the *Lebesgue integral* on $[a,b]$ of function $f(t)$ is the number

$$\int_a^b f(t)dt = \lim_{n\to\infty}\int_a^b f_n(t)dt.$$

The elements of the space $L_1[a,b]$ are the functions $f(t)$ for which

$$\int_a^b |f(t)|\,dt < \infty.$$

We now examine the set $A\subset \mathrm{R}^n$. It is said that A has the *measure zero* if for any $\varepsilon > 0$ it can be covered by spheres with the total volume smaller than ε.

Let it be that $\Omega\subset \mathrm{R}^n$ is a domain. It is said that some property is satisfied *almost everywhere* in Ω if the set of points of the domain Ω which does not have this property, has the measure of zero.

Function $f(x)$ is referred to as *measurable* if it coincides almost everywhere

with the limit of almost everywhere converging sequence of piecewise continuous functions.

The set $A \subset \mathrm{R}^n$ is referred to as *measurable* if its characteristic function $\chi_A(x)$ is measurable.

Let Ω be the measureable set from R^n. Therefore, by analogy with the previously examined case of functions of a single independent variable we can introduce the concept of *function $f(x)$ integrable according to Lebesgue*, on Ω, determine the Lebesgue integral of $f(x)$ *and the space $L_1(\Omega)$ of integrable functions* – Banach space of the functions $f(x)$, for which the finite norm is:

$$\|f\|_{L_1(\Omega)} = \int_\Omega |f(x)|\,dx,$$

where $\int_\Omega$ is the Lebesgue integral.

Function $f(x)$ is referred to as *locally integrable according to Lebesgue* in the domain Ω, $f \in L_{\mathrm{loc}}(\Omega)$, if $f \in L_1(\Omega')$ for all measureable $\Omega' \subset \Omega$.

Let $1 \le p \le \infty$. The set of functions $f(x)$, measureable according to Lebesgue, defined on Ω for which the finite norm is

$$\|f\|_p \equiv \|f\|_{L_p(\Omega)} = \left(\int_\Omega |f(x)|^p\,dx \right)^{1/p},$$

forms the space $L_p(\Omega)$. We shall list some of the properties of the spaces L_p.

Theorem 4. Let Ω be a bounded domain in R^n. Therefore:

1) $L_p(\Omega)$ is the completed normalized space;

2) The set of the finite nonzero functions $C_0^\infty(\Omega)$ is dense in $L_p(\Omega)$;

3) The set of the finite function $C_0^\infty(\mathrm{R}^n)$ is dense in $L_p(\mathrm{R}^n)$;

4) Any linear continuous functional $l(\varphi)$ in $L_p(\Omega)$, $1<p<\infty$, is presented in the form

$$l(\varphi) = \int_\Omega f(x)\varphi(x)\,dx,$$

where $f \in L_{p'}(\Omega)$, $1/p+1/p'=1$;

5) *Function $f(x) \in L_p(\Omega)$, $1 \le p < \infty$, is continuous alltogether, i.e. for any $\varepsilon>0$ we find $\delta(\varepsilon)>0$ such that*

$$\left(\int_\Omega |f(x+y) - f(x)|^p\,dx \right)^{1/p} < \varepsilon,$$

when $|y| \le \delta(\varepsilon)$ (here $f(x)=0$ for $x \notin \Omega$).

Theorem 5 (Riesz theorem). For compactness of the set $M \subset L_p(\Omega)$, where $1 \le p < \infty$, Ω is the bounded domain in R^n, it is necessary and sufficient to satisfy the following conditions:

a) $\| f \|_{Lp(\Omega)} \le K$, $f \in M$;

b) the set M is equicontinuous altogether, i.e. for every $\varepsilon>0$ we find $\delta(\varepsilon)>0$ such that

$$\left(\int_{\Omega}|f(x+y)-f(x)|^p dx\right)^{1/p} \le \varepsilon$$

for all $f\in M$ if only $|y| \le \delta(\varepsilon)$ (here $f(x) = 0$ for $x \notin \Omega$).

For functions from spaces L_p the following inequalities are valid:

1) *Hölder inequality*. Let $f_1\in L_p(\mathrm{R}^n)$, $f_2\in L_{p'}(\mathrm{R}^n)$, $1/p+1/p' = 1$. Consequently $f_1\cdot f_2$ is integrable on R^n and

$$\int_{R^n}|f_1\cdot f_2|dx \le \|f_1\|_p \|f_2\|_{p'}, \quad (\|\cdot\|_{Lp} = \|\cdot\|_p).$$

2) *Generalized Minkovskii inequality*. Let it be that $f(x,y)$ is the function measurable according Lebesgue, defined on $\mathrm{R}^n \times \mathrm{R}^m$, then

$$\left\|\int_{\mathrm{R}^m} f(x,y)dy\right\|_p \le \int_{\mathrm{R}^m}\|f(x,y)\|_p\, dy, \quad 1\le p<\infty.$$

3) *Young inequality*. Let p, r, q be real numbers, $1\le p\le q<\infty$, $1-1/p+1/q=1/r$, functions $f\in L_p$, $K\in L_r$. We examine the convolution

$$f*K = \int_{\mathrm{R}^n} f(y)K(x-y)dy.$$

$$\|f*K\|_q \le \|K\|_r \|f\|_p .$$

4) *Hardy inequality*. Let $1<p<\infty$. Then

$$\int_0^\infty x^{-r}\left|\int_0^x f(t)dt\right|^p dx \le c\int_0^\infty x^{p-r}|f(x)|^p dx \text{ for } r>1,$$

$$\int_0^\infty x^{-r}\left|\int_x^\infty f(t)dt\right|^p dx \le c\int_0^\infty x^{p-r}|f(x)|^p dx \text{ for } r<1.$$

2.3. $L_2(\Omega)$ Space. Orthonormal systems

2.3.1. Hilbert spaces

Let X be a linear set (real or complex). Each pair of the elements f, g from X will be related to a complex number $(f,g)_X$, satisfying the following axioms:

a) $(f,f)_X \ge 0$; $(f,f)_X = 0$ at $f = 0$ and only in this case;

b) $(f,g)_X = \overline{(g,f)}_X$ (the line indicates complex conjugation);

c) $(\lambda f,g)_X = \lambda(f,g)_X$ for any number λ;

d) $(f + g,h)_X = (f,h)_X + (g,h)_X$.

If the axioms a)–d) are satisified, the number $(f,g)_X$ is the *scalar product* of the elements f,g from X.

If $(f,g)_X$ is a scalar product then a norm can be imposed on X setting that

$\| f \|_X = (f, f)_X^{1/2}$. The axioms of the norm a), b) are evidently fulfilled and the third axiom follows from the Cauchy–Bunyakovskii inequality

$$|(f,g)_X| \le \|f\|_X \|g\|_X ,$$

which is valid for any scalar product $(f,g)_X$ and the norm $\|f\|x = (f, f)_X^{1/2}$, generated by the scalar product $(f,g)_X$.

If the linear space X with the norm $\|f\|_X = (f, f)_X^{1/2}$, is complete in relation to this norm, X is referred to as a Hilbert space.

Let it be that X is a space with a scalar product $(f,g)_X$. If $(f,g)_X = 0$, then the elements f,g are orthogonal and we can write $f \perp g$. It is evident that the zero of the space X is orthogonal to any element from X.

We examine in X elements $f_1,\ldots, f_m$, all of which differ from zero. If $(f_k, f_l)_X = 0$ for any $k, l = 1,\ldots m$ $(k \ne l)$, then the system of elements $f_1,\ldots,f_m$ is the *orthogonal system*. This system is referred to as *orthonormalized (orthonormal)* if

$$(f_k, f_l)_X = \delta_{kl} = \begin{cases} 1 & \text{for} \quad k = l, \\ 0 & \text{for} \quad k \ne l. \end{cases}$$

It should be mentioned that if $f_1,\ldots, f_m$ *is the orthogonal system*, then $f_1,\ldots, f_m$ *are linearly independent*, i.e. from the relationship $\lambda_1 f_1 + \ldots + \lambda_m f_m = 0$, where $\lambda_1,\ldots,\lambda_m$ are some numbers, we obtain $\lambda_k = 0$, $k = 1,\ldots,m$. If the 'infinite' system f_k is given, $k = 1,2,\ldots, m \to \infty$, it is referred to as *linearly independent* if at any finite m system $f_1,\ldots, f_m$ is linearly independent.

Theorem 6. *Let $h_1,h_2,\ldots \in X$ be a linearly independent system of elements. Consequently, in X there is some orthogonal system of elements $f_1, f_2\ldots$, such that*

$$f_k = a_{k1}h_1 + a_{k2}h_2 + \ldots + a_{kk}h_k; \quad a_{ki} \in \mathbb{C}, \quad a_{kk} \ne 0, \quad k = 1,2,\ldots,$$
$$h_j = b_{j1}f_1 + b_{j2}f_2 + \ldots + b_{jj}f_j; \quad b_{ji} \in \mathbb{C}, \quad b_{jj} \ne 0, \quad j = 1,2,\ldots,$$

where $\mathbb{C}$ *is the set of complex numbers.*

The construction of the orthogonal system in respect of the given linearly independent system is referred to as *orthogonalization*.

The orthogonal system $\varphi_1,\varphi_2,\ldots \in X$ is referred to as *complete* if every element from X can be presented in the form of the so-called *Fourier series* $f = \sum_\kappa c_k \varphi_k$, where $c_k = (f,\varphi_k)/\|\varphi_k\|^2$ are the Fourier coefficient (i.e. the series $\sum_\kappa c_k \varphi_k$ converges in respect of norm X, and its sum is equal to f). The complete orthogonal system is referred to as the *orthogonal basis* of space X.

Theorem 7. *Let M be a closed convex set in the Hilbert space X and element $f \notin M$. Consequently, there is a unique element $g \in M$ such that* $\rho(f, M) = \| f - g \| \equiv \inf_{\tilde{g} \in M} \| f - \tilde{g} \|_x$.

Element g is the *projection* of the element f on M.

Several examples of the Hilbert space will now be discussed.

Example 1. *The Euclidean space* $\mathbb{R}^n$. Elements of $\mathbb{R}^n$ are real vectors $x = (x_1,...,x_n)$, and the scalar product is given by the equation $(x, y) = \sum_{k=1}^{n} x_k y_k$.

Example 2. *Space* l_2. In the linear space of real sequences $x = (x_k)_{k=1}^{\infty}$, $y = (y_k)_{k=1}^{\infty}$, such that $\sum_{k=1}^{\infty} x_k^2 < \infty$, $\sum_{k=1}^{\infty} y_k^2 < \infty$, the scalar product is defined by equation $(x, y) = \sum_{k=1}^{\infty} x_k y_k$.

Example 3. *Space* L_2 $[a,b]$. In a linear space of complex valued functions, defined (almost everywhere) on $[a,b]$, the scalar product is given as $(x, y) = \int_a^b x(t)\overline{y}(t)dt$, where $\overline{y}$ (t) is the function complexly conjugated with $y(t)$.

2.3.2. Space $L_2(\Omega)$

The set of all functions $f(x)$ for which function $|f(x)|^2$ is integrable according to Lebesgue on domain Ω, is denoted by $L_2(\Omega)$. The scalar product and norm in $L_2(\Omega)$ are determined respectively by the equations:

$$(f, g) = \int_{\Omega} f(x)\overline{g}(x)dx, \ \|f\| = \left(\int_{\Omega} |f(x)|^2 dx \right)^{1/2} = (f, f)^{1/2},$$

and, subsequently, $L_2(\Omega)$ converts to a linear normalized space.

The sequence of functions f_k, $k = 1,2,...$, $L_2(\Omega)$ is referred to as *converging* to function $f \in L_2(\Omega)$ in space $L_2(\Omega)$ (or *in the mean in* $L_2(\Omega)$) if $||f_k - f|| \to 0$, $k \to \infty$; in this case we can write $f_k \to f$, $k \to \infty$ in $L_2(\Omega)$.

The following theorem expresses the property of completeness of the space $L_2(\Omega)$.

Theorem 8. (Reiss–Fischer theorem). If the sequence of functions f_k, $k = 1,2,...$, from $L_2(\Omega)$ converges in itself in $L_2(\Omega)$, i.e.$\| f_k - f_p \| \to 0$, $k \to \infty$, $p \to \infty$, but there is the function $f \in L_2(\Omega)$ such that $\| f_k - f \| \to 0$, $k \to \infty$; in this case function f is unique with the accuracy to the values of the measure zero.

Space $L_2(\Omega)$ is a Hilbert space.

The set of the function $M \subset L_2(\Omega)$ is *dense* in $L_2(\Omega)$ if for any $f \in L_2(\Omega)$ there is a sequence of functions from M, converging to f in $L_2(\Omega)$. For example, the set $C(\overline{\Omega})$ is dense in $L_2(\Omega)$; from this it follows that the set of polynomials is dense in $L_2(\Omega)$ if Ω is the bounded domain (because of the Weierstrass theorem).

2.3.3. Orthonormal systems

According to general definition for the Hilbert spaces, the system of functions $\{\varphi_k(x)\}$ in $L_2(\Omega)$ is *orthonormal* if $(\varphi_k, \varphi_i) = \int_{\Omega} \varphi_k(x)\overline{\varphi}_i(x)dx = \delta_{ki}$. Any system $\{\varphi_k(x)\}$ orthonormal in $L_2(\Omega)$ consists of linearly independent functions. If ψ_1, ψ_2... is a system of functions linearly independent in $L_2(\Omega)$ then it converts to the orthonormal system φ_1, φ_2,....by the following process of

Hilbert–Schmidt orthogonalization:

$$\varphi_1 = \frac{\psi_1}{\|\psi_1\|},\ \varphi_2 = \frac{\psi_2 - (\psi_2;\varphi_1)\varphi_1}{\|\psi_2 - (\psi_2,\varphi_1)\varphi_1\|}, \ldots$$

$$\ldots,\ \varphi_k = \frac{\psi_k - (\psi_k,\varphi_{k-1})\varphi_{k-1} - \ldots - (\psi_k,\varphi_1)\varphi_1}{\|\psi_k - (\psi_k,\varphi_{k-1})\varphi_{k-1} - \ldots - (\psi_k,\varphi_1)\varphi_1\|}\ \ldots$$

Let the system of functions φ_k, $k = 1,2,\ldots$, be orthonormal in $L_2(\Omega)$, $f \in L_2(\Omega)$. The numbers (f, φ_k) are the *Fourier coefficients*, and the formal series is $\sum_{k=1}^{\infty}(f,\varphi_k)\varphi_k(x)$ is *the Fourier series* of function f in respect of the orthonormal system $\{\varphi_k(x)\}$.

If the system of functions φ_k, $k = 1,2,\ldots$, is *orthonormal* in $L_2(\Omega)$, then for every $f \in L_2(\Omega)$ and any (complex numbers) $a_1, a_2, \ldots a_N$, $N = 1,2,\ldots$, the following equality is valid

$$\left\| f - \sum_{k=1}^{N} a_k\varphi_k \right\|^2 = \left\| f - \sum_{k=1}^{N}(f,\varphi_k)\varphi_k \right\|^2 + \sum_{k=1}^{N}\left|(f,\varphi_k) - a_k\right|^2.$$

Assuming in this equality that $a_k = 0$, $k = 1,2,\ldots,N$, we obtain the following equality:

$$\left\| f - \sum_{k=1}^{N}(f,\varphi_k)\varphi_k \right\|^2 = \|f\|^2 - \sum_{k=1}^{N}\left|(f,\varphi_k)\right|^2,$$

leading to *the Bessel inequality*:

$$\sum_{k=1}^{N}\left|(f,\varphi_k)\right|^2 \le \|f\|^2.$$

In addition to this, it should be noted that: *in order to ensure that the Fourier series converges to function f in $L_2(\Omega)$, it is necessary and sufficient to fulfil the Parseval–Steklov equality (completeness equation):*

$$\sum_{k=1}^{\infty}\left|(f,\varphi_k)\right|^2 = \|f\|^2.$$

Let the system φ_k, $k \ge 1$, be orthonormal in $L_2(\Omega)$. If for any $f \in L_2(\Omega)$ its Fourier series in respect of system $\{\varphi_k\}$ converges to f in $L_2(\Omega)$, then the system is referred to as *complete* (*closed*) in $L_2(\Omega)$) (*orthonormal basis in* $L_2(\Omega)$). This definition and the claims formulated previously in this section result in:

Theorem 9. *To ensure that the orthogonal system $\{\varphi_k\}$ is complete in L_2 (Ω) it is necessary and sufficient that the Parseval–Steklov equality (completeness equation) $L_2(\Omega)$ is satisfied for any function f from $L_2(\Omega)$.*

The following theorem is also valid.

Theorem 10. *To ensure that the orthogonal system $\{\varphi_k\}$ is complete in L_2 (Ω), it is necessary and sufficient that each function f from the set M,*

dense in $L_2(\Omega)$ *can be approximated with sufficient accuracy* (*as close to as required*) *by linear combinations of functions of this system.*

Corollary. *If* Ω *is a bounded domain, then in* $L_2(\Omega)$*, there is a countable complete orthogonal system of polynomials.*

We shall formulate the following claim, specifying one of the possibilities of constructing orthonormal systems in the case $G\subset R^n$ at a high value of n.

Lemma 1. *Let it be that the domains* $\Omega\subset R^n$ and $D\subset R^n$ *are bounded, the system of functions* $\psi_j(y)$, $j = 1,2,\ldots$, *is orthonormal and complete in* $L_2(D)$ *and for every* $j = 1,2,\ldots$ *this system of functions* $\varphi_{kj}(x)$, $k = 1,2,\ldots$, *is orthonormal and complete in* $L_2(\Omega)$. *Consequently, this systems of functions,* $\chi_{kj} = \varphi_{kj}(x)\psi_j(y)$, $k,j = 1,2,\ldots$, *is orthonormal and complete in* $L_2(\Omega\times D)$.

Comment. All we have said about the space $L_2(\Omega)$ is also applicable to the space $L_2(\Omega;\rho)$ or $L_2(\partial\Omega)$ with scalar products:

$$(f,g)_\rho = \int_\Omega \rho(x) f(x)\overline{g}(x)dx, \quad (f,g) = \int_{\partial\Omega} f(x)\overline{g}(x)d\Gamma,$$

where $\rho\in C(\overline{\Omega})$, $\rho(x) > 0$, $x\in\overline{\Omega}$ and $\partial\Omega$ is a piecewise smooth boundary of the domain Ω.

2.4. Linear operators and functionals

2.4.1. Linear operators and functionals

Let X,Y be linear normalized spaces, $D(A)$ is some linear set from X, and $R(A)$ is a linear set from Y. Let the elements from $D(A)$ be transformed to elements $R(A)$ in accordance with some rule (law). Consequently, it is said that the operator A is defined with the domain of definition $D(A)$ and the range of values $R(A)$, acting from X in Y, i.e. A: $X\to Y$. If $Af = f$ at all $f\in D(A)$, then A is the *identical* (*unique*) and it is denoted by I.

Let X,Y be linear normalized spaces, $A : X\mapsto Y$ be the mapping or the *operator*, determined in the neighbourhood of point $f_0 \in X$. It is referred to as *continuous* at point f_0 if $A(f)\to A(f_0)$ at $f\to f_0$.

Let A be the operator with the definition domain $D(A)\subset X$ and with the *range of values* $R(A)\subset Y$. It is referred to as *bounded* if it transfers any bounded set from $D(A)$ to the set bounded in space Y.

Let X,Y be linear normalized spaces, both are real or both are complex. The operator A: $X\to Y$ with the domain of definition $D(A)\subset X$ is referred to as *linear* if $D(A)$ is a linear manifold in X and for any f_1, $f_2\in D(A)$ and any $\lambda_1,\lambda_2\in R(\lambda_1,\lambda_2\in C)$ the equality $A(\lambda_1+\lambda_2 f_2) =\lambda_1 A f_1 + \lambda_2 A f_2$ is satisfied.

The set $N(A)=\{f\in D(A):A(x)=0\}$ is referred to as the *zero manifold* or the *kernel of the operator A*.

Theorem 11. *The linear operator* A: $X\mapsto Y$ *, defined on all X and continuous at point* $0\in X$*, is continuous at any point* $f_0\in X$.

The linear operator $A:X\mapsto Y$ with $D(A)=X$ is *continuous* if it is continuous at point $0\in X$.The linear operator A: $X\to Y$ with $D(A)=X$ is referred as *bounded*

if there is $c\in$ R, $c > 0$, such that for any $f\in \bar{S}_1(0)\equiv\{f:||f||_x\leq 1\}$ the inequality $||Af|| \leq c$ holds.

Theorem 12. *The linear operator* $A:X\mapsto Y$ with $D(A)=X$ *is bounded if and only if the inequality* $||Af||\leq c||f||$ *is satisfied for every* $f\in X$.

Theorem 13. *The linear operator* $A:X\mapsto Y$ with $D(A)=X$ *is continuous if and only if it is bounded.*

The *norm* of the bounded linear *operator* $A:X\mapsto Y$ with $D(A)=X$ is the number $\|A\| = \sup\limits_{f\in X, \|f\|\leq 1} \|Af\|$.

The set of the operators from X in Y with a finite norm forms a linear normalized space of bounded linear operators $\mathcal{L}(X,Y)$.

The linear operator from X in Y is referred to as *completely continuous* if it transfers every bounded set from X to a compact set from Y.

Let it be that A is a linear operator, determined on the set $D(A)\subset X$ and acting in Y. Operator A is referred to as *closed* if for any sequence $\{f_n\}$ of elements $D(A)$ such that $f_n\to f_0\in X$, $Af_n\to g_0\in Y$, we have $f_0\in D(A)$ and $Af_0=g_0$. Operator A is referred to as *weakly closed* if for any sequence of the elements $\{f_n\}$ such that f_n weakly converges to $f_0\in X$, and Af_n weakly converges to $g_0\in Y$, it follows that $f_0\in D(A)$ and $Af_0=g_0$.

A partial case of linear operators are linear functionals. If the linear operator l transforms the set of elements $M \subset X$ to the set of complex numbers lf, $f\in M$, i.e. $l : X \to C$, then l is referred to as a *linear functional* on the set M, the value of the function l on element f – complex number lf – will be denoted by $(l,f) \equiv l(f)\equiv\langle f,l\rangle$. The continuity of the linear functional l denotes the following: if $f_k\to 0$, $k\to\infty$, in M, then the sequence of complex numbers (l, f_k), $k\to\infty$, tends to zero.

Let the norm $||l||=\sup_{||x||=1}|(l,x)|$ be introduced in the linear space of all linear functionals of X. Consequently, the set of bounded functionals on X, i.e. functionals for which the norm is finite and forms Banach space, is referred to as adjoint to X and is denoted by X^*.

It is assumed that the sequence l_1, l_2,... of linear functionals on M *weakly converges* to the (linear) functional l on M, if it converges to l on every element f from M, i.e. $(l_k, f)\to(l, f)$, $k\to\infty$.

The sequence $\{f_n\}$ of elements from X is referred as *weakly converging* to $f_0\in X$ if $\lim_{n\to\infty}(1,f_n) = (1,f_0)$ for any $l\in X^*$.

Some examples of linear operators and functions will now be discussed.

Example 1. The linear operator of the form

$$Kf = \int_\Omega K(x,y)\, f(y)dy, \quad x\in \Omega,$$

is referred to as a (linear) *integral operator*, and the function $K(x,y)$ is its *kernel*. If the kernel $K\in L_2(\Omega\times\Omega)$, i.e.

$$\int_{\Omega\times\Omega} |K(x,y)|^2\, dxdy = C^2 < \infty,$$

operator K is bounded (and, consequently, continuous) from $L_2(\Omega)$ to $L_2(\Omega)$.

Example 2. The linear operator of the form

$$Af = \sum_{|\alpha|\le m} a_\alpha D^\alpha f(x), \quad \sum_{|\alpha|=m} |a_\alpha(x)| \equiv 0, \ m > 0,$$

is referred to as (linear) *differential operator of order m*, and the functions $a_\alpha(x)$ are its coefficients. If the coefficients $a_\alpha(x)$ are continuous functions on the domain $\Omega\subset R^n$, then the operator A transforms $C^m(\bar{\Omega})=D(A)$ to $C(\bar{\Omega})=R(A)$. However, operator A is not continuous from $C(\bar{\Omega})$ to $C(\bar{\Omega})$. It should also be mentioned that operator A is not defined on the entire space $C(\bar{\Omega})$ and only in its part – on the set of the functions $C^m(\bar{\Omega})$.

Example 3. Linear operator

$$Af = \sum_{|\alpha|\le m} \left[\int_\Omega k_\alpha(x,y) f(y)dy + a_\alpha(x) D^\alpha f(x) \right]$$

is referred to as a (linear) integro-differential operator.

Example 4. An example of a linear continuous functional l on $L_2(\Omega)$ is the scalar product $(l,f) = (f,g)$, where g is a fixed function from $L_2(\Omega)$. The linearity of this functional follows from the linearity of the scalar product in respect to the first argument, and because of the Cauchy–Bunyakovskii inequality it is bounded: $|(l, f)| = |(f,g)| \le ||g||\cdot||f||$, and, consequently, continuous.

2.4.2. Inverse operators

Let X,Y be linear normalized spaces, $A:X \mapsto Y$ is a linear operator, mapping $D(A)$ on $R(A)$ one-to-one. Consequently, there is an inverse operator $A^{-1}:Y\to X$, mapping $R(A)$ on $D(A)$ one-to-one and is also linear.

Linear operator $A : X\to Y$ is referred to as *continuously invertible*, if $R(A)= Y$, A^{-1} exists and is bounded, i.e. $A^{-1}\in \mathcal{L}(X,Y)$.

Theorem 14. *Operator A^{-1} exists and is bounded on $R(A)$ iff the inequality $||Ax|| \ge m||x||$ is fulfilled for some constant $m > 0$ and any $x \in D(A)$.*

Theorem 15. *Let X,Y be Banach spaces, $A\in \mathcal{L}(X,Y)$, $R(A)=Y$ and A is invertible. Consequently, A is continuously invertible.*

2.4.3. Adjoint, symmetric and self-adjoint operators

Let X,Y be linear normalized spaces, $A:X\to Y$ is a linear operator with the domain of definition $D(A)$, dense in X, possibly unbounded. We introduce the set $D^*\subset Y^*$ of such $f\in Y^*$ for which $\varphi\in X^*$ there is the equality $\langle Ax, f\rangle = \langle x,\varphi\rangle$. The operator $A^* f = \varphi$ with the domain of definition $D(A^*) = D^*\subset Y^*$ and with the values in X^* is referred to as *adjoint* to the

operator A. Thus, $\langle Ax, f\rangle = \langle x,f\rangle$ for any x from $D(A)$ and for any f from $D(A^*)$.

Linear operator A is referred to as *symmetric* if $A\subset A^*$ (i.e. $D(A)\subset D(A^*)$ and $A=A^*$ in $D(A)$) and closure of $D(A)$ coincides with X, i.e. $\overline{D(A)}=X$. The linear operator A with $\overline{D(A)}=X$ is referred to as *self-adjoint* if $A=A^*$.

Theorem 16. A^* is *a closed linear operator*.

Theorem 17. *Equality* $D(A^*)=Y^*$ *holds iff* A *is bounded on* $D(A)$. *In this case,* $A^*\in \mathcal{L}(X^*,Y^*)$, $||A^*||=||A||$.

2.4.4. Positive operators and energetic space

Symmetric operator A acting in some Hilbert space is referred to as *positive* if the inequality $(Au,u)\geq 0$ holds for any element from the domain of definition of the operator and the equality holds only when $u = 0$, i.e. only when u is the zero element of the space.

If A is a positive operator, the scalar product (Au,u) is the *energy* of the element u in relation to A.

Symmetric operator A is *positive definite* if there is a positive constant γ such that for any element u from the domain of definition of operator A the inequality $(Au,u)\geq\gamma^2||u||^2$ holds.The unique Hilbert space referred to as the *energetic space*, can be linked with any positive (in particular, positive definite) operator. Let A be a positive operator acting in some Hilbert space H, and let $M = D(A)$ be the domain of definition of this operator. On M we introduce a new scalar product (which will be denoted by square brackets): if u and v are elements of M, we set $[u,v] = (Au,v)$. The quantity $[u,v]$ is referred to as the *energetic product* of elements u and v. It is easily verified that the energetic product satisfies the axioms of the scalar product.

In a general case, M is incomplete, and we complete it in respect of the norm$[u] = [u,u]^{1/2}$. The new Hilbert space, constructed in this manner, is referred to as the *energetic space* and denoted by H_A.

The norm in the energetic space is referred to as the *energetic norm* and denoted by symbol $[u]$. For the elements of the domain of definition of M of operator A, the energetic norm is defined by the formula $|u| = \sqrt{(Au,u)}$. The convergence in the energetic spaces is referred to as *convergence in respect to energy*.

An important role is played by the problem of the nature of elements that are used for completing and constructing the energetic space. If operator A is positive definite, we have a theorem because of which all elements of space H_A also belong to the initial Hilbert space H; if u is the element of space H_A we then have the inequality $||u||\leq(1/\gamma)|u|$, where the symbol $||\cdot||$ denotes the norm in initial space H.

Some boundary conditions are often imposed on elements from M. The elements from H_A, used for completing M to H_A, may not satisfy some of the boundary conditions which are referred to as *natural* in the present case. The boundary conditions which are satisfied by both of the elements from M, and all elements from H_A are referred to as *main*.

If A is a positive definite operator, the convergence of some sequence in respect of energy also leads to its convergence in the norm of the initial space: if $u_n \in H_A$, $u \in H_A$ and $|u_n - u| \to 0$, also $||u_n - u|| \to 0$.

Symmetric positive operators and corresponding energetic spaces play an important role in examination of variational formulations of the mathematical physics problems.

2.4.5. Linear equations

Let A be a linear operator with the domain of definition $D(A) \subset X$ and range $R(A) \subset Y$. The equation

$$Au = F \tag{1}$$

is a linear (inhomogeneous) equation. In equation (1), the given element F is the *free term* (or the *right-hand side*), and the unknown element u from $D(A)$ is the *solution* of this equation. If the free term F in equation (1) is assumed to be equal to zero, the resultant equation

$$Au = 0 \tag{2}$$

is a linear *homogenous* equation corresponding to equation (1). Since operator A is linear, the set of the solutions of the homogeneous equation (2) forms a linear set; in particular, $u = 0$ is always a solution of this equation.

Any solution u of the linear inhomogeneous equation (1) (if such a solution exists) is presented in the form of the sum of the partial solution u_0 *of this equation and general solution* $\tilde{u}$ *of the appropriate linear homogeneous equation (2)*:

$$u = u_0 + \tilde{u}. \tag{3}$$

From this, we directly conclude: *to ensure that the solution of equation* (1) *is unique in* $D(A)$, *it is necessary and sufficient that the appropriate homogeneous equation* (2) *has only zero solution in* $D(A)$.

Let it be that the homogeneous equation (2) has only zero solution in $D(A)$. Consequently for any $F \in R(A)$ equation (1) has the unique solution $u \in D(A)$ and this also defines operator A^{-1}, i.e. the operator inverse to A, since

$$u = A^{-1}F. \tag{4}$$

The relations (1),(4) finally give:

$$AA^{-1}F = F, \qquad F \in R(A), \qquad A^{-1}Au = u, \qquad u \in D(A),$$

i.e. $AA^{-1} = I$ and $A^{-1}A = I$.

2.4.6. Eigenvalue problems

We examine a linear homogeneous equation

$$Au = \lambda u, \tag{5}$$

where λ is a numerical parameter. This equation has zero solution for all λ. It may be that at some λ it has nonzero solutions from $D(A)$. Complex values λ at which equation (5) has nonzero solutions from $D(A)$ are referred to as *eigenvalues* of the operator A, and the appropriate solutions are *eigenelements* (functions) corresponding to this eigenvalue.The total number r ($1 \le r \le \infty$) of linearly independent eigenelements, corresponding to the given eigenvalue λ, is referred to as the *multiplicity* of this eigenvalue; if the multiple

$r = 1$, λ is referred to as the *simple eigenvalue*.

The set of eigenvalues (numbers) of the operator A is referred to as its *point spectrum*.

If the eigenelements u_1, u_2, ...u_n from equation (5) correspond to the eigenvalue λ, then any linear combination of this elements, differing from zero element $c_1u_1 + c_2u_2 + \ldots c_nu_n$, where $c_1,\ldots,c_n$ are arbitrary constants, is also an eigenelement of this equation corresponding to this eigenvalue λ. Thus, the set of the eigenelements of the given equation supplemented by the zero element and corresponding to the given eigenvalue, is linear.

In very wide conditions (if the operator $A–\lambda I$ is close), this set is a subspace referred to as the *eigen subspace* of equation (5), corresponding to eigenvalue λ; the dimension of this subspace is equal to the multiplicity of the eigenvalue.

If the multiplicity r of the eigenvalue λ of the operator A is finite u_1, u_2,..., and u_r are the appropriate linearly independent eigenelements, then any linear combination of these eigenelements

$$u_0=c_1u_1+c_2u_2+\ldots+c_ru_r \tag{6}$$

is also an eigenelement corresponding to this eigenvalue, and equation (6) gives the general solution of equation (5). From this and also from equation (3) we obtain: *if the solution of the equation*

$$Au=\lambda u + f \tag{7}$$

does exists, then its general solution is represented by the formula

$$u = u^* + \sum_{k=1}^{r} c_k u_k, \tag{8}$$

where u^* is the partial solution of (7) and c_k, $k = 1,2,\ldots., r$, are arbitrary constants.

It is assumed that the set of the eigenvalues of the symmetric operator A is not greater than countable, and every eigenvalue has a finite multiplicity. We numerate all its eigenvalues: λ_1, λ_2..., repeating λ_k as many times as the value of its multiplicity. The corresponding eigenfunction is denoted by u_1, u_2, ..., in such a manner that every eigenvalue corresponds to only one eigenfunction u_k:

$$Au_k= \lambda_k u_k, \quad k = 1,2,\ldots$$

The eigenfunctions, corresponding to the same eigenvalue, can be selected as orthonormal, using the process of Hilbert–Schmidt orthogonolization. This again gives the eigenfunctions corresponding to the same eigenvalue.

We shall list the main properties of the eigenvalues and eigenelements of symmetric operators.

1. The eigenvalues of a symmetric operator are real.

2. The eigenelements of the symmetric operator, corresponding to different eigenvalues, are orthogonal.

3. If the eigenvalue corresponds to several linearly independent eigenelements then, using the orthogonolization process for these elements, we make them orthogonal. Taking this into account, it may be assumed that the set of all eigenelements of the symmetric operator forms an orthogonal system.

4. The symmetric operator may either have the finite or countable set of eigenvalues which can therefore be written in the form of a finite or countable succession $\lambda_1,\lambda_2,\ldots,\lambda_n,\ldots$ Of course, cases are also possible in which the symmetric operator does not have any eigenvalues.

5. The eigenelements of the positive definite operator are orthogonal in the energetic space.

6. The eigenvalues of the positive definite operator are positive.

Comment. Many of these properties of the eigenvalues and the eigenelements remain valid in examination of the general problem for eigenvalues

$$Au=\lambda Bu, \tag{9}$$

where A and B are symmetric positive definite operators, and $D(A)\subset D(B)$.

2.5. Generalized derivatives. Sobolev spaces

2.5.1. Generalized derivatives

According to S.L. Sobolev, we defined a generalized derivative for a locally summable function.

Function Ω locally summable on ω_α is referred to as the *generalized derivative of the function* $f \in L_{loc}(\Omega)$ of the order $\alpha = (\alpha_1,\ldots,\alpha_n)$ (α_k are non-negative integrals) $k = 1,\ldots,n$, if for any function $\varphi\in C_0^\infty(\Omega)$ we have the equality:

$$\int_\Omega fD^\alpha\varphi dx = (-1)^{|\alpha|}\int_\Omega \omega_\alpha\varphi dx. \tag{10}$$

We also denote $\omega_\alpha = D^\alpha f$.

We examine some of the properties of generalized derivatives. Equality (10) makes the locally summable function f correspond to the unique generalized derivative of order α. This results from the Dubois–Raymond lemma.

Lemma 2 (Dubois–Raymond lemma). *For the locally summable function f to be equal to zero almost everywhere in domain Ω, it is necessary and sufficient that for any function* $\varphi\in C_0^\infty(\Omega)$ *the following is fulfilled:*

$$\int_\Omega f\varphi dx = 0.$$

Theorem 18 (weak closure of the operator of generalized differentiation). *Let it be that f_n is a sequence of locally summable functions on Ω. If there* are ω_0, $\omega_\alpha\in L_{\text{loc}}$ *such that for any finite functions* $\varphi\in C_0^\infty(\Omega)$ *the equalities*

$$\lim_{n\to\infty}\int_\Omega f_u\varphi dx = \int_\Omega \omega_0\varphi dx, \tag{11}$$

$$\lim_{n\to\infty}\int_\Omega f_n D^\alpha\varphi dx = (-1)^{|\alpha|}\int_\Omega \omega_\alpha\varphi dx, \tag{12}$$

are satisfied, then the locally summable function ω_α is a generalized derivative of the order α of function ω_0.

Corollary. Let it be that the sequence $f_n\in L_p(\Omega)$, $1 < p < \infty$, weakly converges to $f_0\in L_p(\Omega)$, and the sequence of generalized derivatives $D^\alpha f_n\in L_p(\Omega)$ weakly converges to $\omega_\alpha\in L_p(\Omega)$. Consequently, f_0 has the generalized derivative of the order α and:

$$D^{\alpha} f_0 = \omega_{\alpha}. \tag{13}$$

2.5.2. Sobolev spaces

Theorem 18 shows that the generalized derivatives according to Sobolev can be regarded as limiting elements of converging in $L_p(\Omega)$ sequences of derivatives of smooth functions. This property of the generalized derivatives is used widely in different boundary-value problems of mathematical physics. The examined problems are usually used to examine some operator, initially given on smooth functions, which must be expanded to a closed operator in some normalized space. Large groups of the differential operators, examined in the space of type L_p, will be closed if they are extended to functions having generalized derivatives. This method, proposed in studies by S.L. Sobolev and K.O. Friedrichs, has made it possible to solve a large number of difficult problems in the theory of differential equations and has become a classic method. A very important role is played here by the classes of the function $W_p^l(\Omega)$ introduced by S.L. Sobolev.

We determine Sobolev classes $W_p^l(\Omega)$ where $l = (l_1,...,l_n)$, $l_i > 0$ are integers, $1 < p < \infty$. Let it be that function $f \in L_p(\Omega)$ has non-mixed generalized derivatives $D^i f \in L_p(\Omega)$, $i = 1,...,n$. For these functions, we determine the norm:

$$\|f\|_{W_p^l(\Omega)} = \|f\|_{L_p(\Omega)} + \sum_{i=1}^{n} \left\|D^h f\right\|_{L_p(\Omega)}. \tag{14}$$

The set of the functions $f \in L_p(\Omega)$, having generalized derivatives $D^i f$, $i = 1,...,n$, for which the norm (14) is finite, is referred to as the Sobolev space $W_p^l(\Omega)$. The spaces $W_p^l(\Omega)$ were introduced and examined for the first time by Sobolev at $l_j = l_i$, $i, j = 1,...,n$.

We formulate some of the properties of spaces $W_p^l(\Omega)$, assuming that Ω is the bounded domain from $\mathbb{R}^n$.

Theorem 19. *Space* $W_p^l(\Omega)$, $l = (l_1,...,l_n)$, *is a complete normalized space.*

The set $M \subset L_p(\Omega)$, $1<p<\infty$, is *weakly compact* if each sequence $f_n \subset M$ contains a sub-sequence weakly converging to some function $f_0 \in L_p(\Omega)$.

Lemma 3. *Any bounded set* $M \subset L_p(\Omega)$, $1<p<\infty$, *is weakly compact.*

We say that the sequence $f_m \in W_p^1(\Omega)$, $l = (l_1,...,l_n)$, $1<p<\infty$, *weakly converges in* $W_p^l(\Omega)$ to $f_0 \in W_p^l(\Omega)$, if:

1. f_m *weakly* converges in $L_p(\Omega)$ to $f_0 \in L_p(\Omega)$;
2. Generalized derivatives $D^{l_i} f_m$ *weakly* converge in $L_p(\Omega)$ to $D^{l_i} f_0 \in L_p(\Omega)$ for $i = 1,...,n$.

The set $M \subset W_p^l(\Omega)$, $l = (l_1,...,l_n)$, $1<p<\infty$, is referred to as *weakly compact* if every sequence $f_m \subset M$ contains a sub-sequence *weakly* converging in $W_p^l(\Omega)$ to some function $f_0 \in W_p^l(\Omega)$.

Theorem 20. *Any bounded set* $M \subset W_p^l(\Omega)$ $l = (l_1,\ldots,l_n)$, $1 < p < \infty$, is *weakly compact*.

$W_p^k(\Omega)$ denotes the subspace from $W_p^k(\Omega)$ consisting of the limits of sequences of functions from $C_0^\infty(\Omega)$ in respect of the norm $\|\cdot\|_{W_p^k(\Omega)}$. At $p = 2$, the spaces $W_p^k(\Omega)$, $\overset{\circ}{W}{}_p^k(\Omega)$ are Hilbert spaces with the scalar product

$$(u,v)_{W_2^k(\Omega)} = \sum_{|\alpha|\le k} \int_\Omega D^\alpha u \overline{D^\alpha v} dx$$

and the norm

$$\|u\|_{W_2^k(\Omega)} = \left(\sum_{|\alpha|\le k} \left\| D^\alpha u \right\|_{L_2(\Omega)}^2 \right)^{1/2}$$

(which is equivalent to the norm of $W_2^k(\Omega)$, introduced previously).

Theorem 21. *Let* $\Omega \subset \mathrm{R}^n$ *be a bounded domain with the Lipschitz boundary* $\partial\Omega$.

1. If $1 < p$, then constant C exists and is such that

$$\|u\|_{L_p(\partial\Omega)} \le C \|u\|_{W_p^l(\Omega)} \quad \forall u \in C^1(\overline{\Omega}),$$

2. If $n < kp$, $\lambda < k - n/p$, then C exists and is such that

$$\|u\|_{C^\lambda(\Omega)} \le C \|u\|_{W_p^k(\Omega)} \quad \forall u \in W_p^k(\Omega).$$

3. If $u \in W_p^k(\Omega)$ and $D^\alpha u = 0$ on $\partial\Omega$ for $|\alpha| \le k-1$, then also $u \in \overset{\circ}{W}{}_p^k(\Omega)$.

4. There is a linear bounded operator L_{ext}, mapping $W_p^k(\Omega)$ such that $\overset{\circ}{W}{}_p^k(\mathrm{R}^n)$, into $L_{\text{ext}} u(x) = u(x)$ at $x \in \Omega$.

5. There is a constant $C = C(\Omega,k)$ such that

$$\|u\|_{W_2^k(\Omega)} \le C \|u\|_{W_2^m(\Omega)}^{k/m} \|u\|_{L_2(\Omega)}^{1-k/m} \quad \forall u \in W_2^m(\Omega), \quad 0 \le k \le m.$$

2.5.3. The Green formula

Let $\Omega \in \mathrm{R}^n$ be the bounded domain with the Lipschitz boundary $\partial\Omega$, $n(x)$ is the unit vector of the external normal to $\partial\Omega$. We examine function $u(x)$ from the class $C^1(\overline{\Omega})$ (or even from $W_1^1(\overline{\Omega})$). Consequently, the *formula of integration by parts* holds

$$\int_\Omega D_j u dx = \int_{\partial\Omega} u n_j d\Gamma \tag{15}$$

where n_j is the j-th coordinate of the vector $n(x)$.

We examine the linear differential operator A of the order m:

$$A = \sum_{|\alpha|\le m} a_\alpha D^\alpha. \tag{16}$$

Let $m = 2$ be the operator A is written in the form:

$$A = \sum a^{jk} D_j D_k + \sum a^j D_j + a, \quad a^{jk} = a^{kj}. \tag{17}$$

Consequently, the *formally adjoint* (*transposed*) operator is:

$$A^{\mathrm{T}} = \sum D_j (a^{jk} D_k) + D_j (a^{ij}) + a, \tag{18}$$

where $a^{jk} \in C^2$, $a^j \in C^1$ and $a^{ij} = \sum D_k a^{jk} - a^j$. Summation is carried out everywhere in respect of j, $k = 1,\ldots,n$. Since

$$vAu - uA^{\mathrm{T}}v = \sum (D_j (va^{jk} D_k u - ua^{jk} D_k v) + D_j ((-D_k a^{jk} + a^j) uv)),$$

then, using the formula of integration by parts, we obtain the following well-known claim.

Theorem 22. *If* Ω *is a bounded domain with the Lipschitz boundary and* coefficient α_a *in the differential operator of the second order A is from the class* $C^{|\alpha|}(\overline{\Omega})$, then for arbitrary functions u, $v \in W_2^1(\Omega)$ the Green equation is valid

$$\int_{\Omega} (vAu - uA^{\mathrm{T}}v)\, dx = \int_{\partial\Omega} \left(a_0 \left(v \frac{\partial u}{\partial N} - u \frac{\partial v}{\partial N} \right) + a_* uv \right) d\Gamma, \tag{19}$$

where $a_0 \left(\sum \left(\sum a^{jk} n_j \right)^2 \right)^{1/2}$, $a_* = -\sum a^{ij} n_j$ *u* $N_k = \sum a^{jk} n_j / a_0$ *are components of the co-normal to* $\partial\Omega$, *corresponding to operator A.*

The Green equation is used widely in analysis and development of numerical methods for solving greatly differing problems of mathematical physics.

3. MAIN EQUATIONS AND PROBLEMS OF MATHEMATICAL PHYSICS

3.1. Main equations of mathematical physics

We examine characteristic physical processes described by different mathematical models, and differential equations in partial derivatives together with typical boundary conditions, included in these models.

Differential equations are equations in which unknown quantities are functions of one or several variables, and the equations include not only the functions themselves but also their derivatives. If functions of many (at least two) variables are unknown, these equations are referred to as *equations in partial derivatives*.

An equation in partial derivatives of the unknown function u of variables $x_1,\ldots,x_n$ is the equation of the *N-th order* if it contains at least one derivative of order N and does not contain derivatives of a higher order, i.e. the equation of the type

$$A\left(x_{1,}...,x_u,u,\frac{\partial u}{\partial x_1},...,\frac{\partial u}{\partial x_n},\frac{\partial^2 u}{\partial x_1^2},\frac{\partial^2 u}{\partial x_1 \partial x_2},...,\frac{\partial^N u}{\partial x_n^N}\right)=0. \tag{20}$$

Equation (20) is linear if A as a function of variables $u,\frac{\partial u}{\partial x_1},...,\frac{\partial^N u}{\partial x_n^N}$ is linear.

An important group of equations in partial derivatives is described by a linear equation of the second order having the form:

$$Au \equiv \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j}+\sum_{i=1}^{u} a_i(x)\frac{\partial u}{\partial x_i}+a(x)u=F(x); \tag{21}$$

where $x = (x_1,...,x_n)$. Functions $a_{ij}(x)$, $i, j = 1,...,n, a_i(x)$, $i = 1,...,n$, $a(x)$ are referred to as *coefficients* of equation (21), and function $F(x)$ is a *free term*.

3.1.1. Laplace and Poisson equations

The Laplace equation has the form

$$\Delta u=0, \tag{22}$$

where $u=u(x),\ x\in \mathrm{R}^n,\ \Delta=\frac{\partial^2}{\partial x_1^2}+...+\frac{\partial^2}{\partial x_n^2}$, is the Laplace operator in R^n. The corresponding heterogeneous equation

$$-\Delta u=F, \tag{23}$$

(F is a known function) is the *Poisson equation*. The Laplace and Poisson equations are found in greatly different problems. For example, the stationary, (i.e. constant with time) temperature distribution in a homogeneous medium and the steady form of a stretched membrane satisfy the Laplace equation, and the identical distribution of temperature in the presence of heat sources (with the density constant with time) and the shape of the membrane in the presence of stationary external forces satisfy the Poisson equation. The potential of the electrostatic field satisfies the Poisson equation with the function F, proportional to the density of charges (at the same time, in the domain when there are no charges it satisfies the Laplace equation).

Laplace and Poisson equations describe the stationary state of objects. For them, it is not necessary to specify initial conditions, and the typical boundary conditions in the case of the bounded domain $\Omega \subset \mathrm{R}^n$ is *Dirichlet boundary condition* (*the condition of the first kind*)

$$u\big|_{\partial\Omega}=\varphi, \tag{24}$$

the *Neumann condition* (*the condition of the second kind*)

$$\left.\frac{\partial u}{\partial n}\right|_{\partial\Omega}=\varphi \tag{25}$$

and the *third boundary-value condition* (*the condition of the third kind*)

$$\left.\left(\frac{\partial u}{\partial n}+\gamma u\right)\right|_{\partial\Omega}=\varphi, \tag{26}$$

where λ,φ are the functions given on $\partial\Omega$.

3.1.2. Equations of oscillations

Many problems in mechanics (oscillations of strings, bars, membranes in three-dimensional volumes) and physics (electromagnetical oscillations) are described by the equation of oscillations of the type

$$\rho\frac{\partial^2 u}{\partial t^2}=\operatorname{div}(p\operatorname{grad}u)-qu+F(x,t), \tag{27}$$

with the unknown function $u(x,t)$ which depends on n (n = 1,2,3) spatial coordinates, $x=(x_1,x_2,\ldots x_n)$, and time t; coefficients ρ, p and q are determined by the properties of the medium in which the oscillation process takes place; the free term $F(x,t)$ expresses the intensity of an external perturbation. In equation (27) in accordance with the definition of the operators div and grad

$$\operatorname{div}(p\operatorname{grad}u)=\sum_{i=1}^{u}\frac{\partial}{\partial x_j}\left(p\frac{\partial u}{\partial x_i}\right).$$

In the case of *small transverse oscillations of a string*, represented by a stretched filament not resisting to bending (i.e. $|T(x,t)|=T_0$=const), equation (27) has the form

$$\rho\frac{\partial^2 u}{\partial t^2}=T_0\frac{\partial^2 u}{\partial x^2}+F, \tag{28}$$

where (x,u) are the coordinates of the plane in which the string carries out transverse oscillations around its equilibrium position, coinciding with the axis x.

At $F\neq 0$ the oscillations of the string are referred to as *induced*, and if F = 0 as *free*.

If density ρ is constant, $\rho(x)=\rho$, the equation of oscillations of the string has the form

$$\frac{\partial^2 u}{\partial t^2}=a^2\frac{\partial^2 u}{\partial x^2}+f, \tag{29}$$

where $f=F/\rho$, and $a^2=T_0/\rho$ is a constant. Equation (29) is also referred to as the *one-dimensional wave equation*.

The equation of type (27) also describes *small longitudinal oscillations of an elastic bar*:

$$\rho S\frac{\partial^2 u}{\partial t^2}=\frac{\partial}{\partial x}\left(ES\frac{\partial u}{\partial x}\right)+F(x,t), \tag{30}$$

where $S(x)$ is the cross section area of the bar, and $E(x)$ is Young's modulus at point x.

Physical considerations show that for the unambiguous description of the process of oscillations of a string or a bar it is necessary to specify also displacement u and velocity u_t at the initial moment of time (*initial conditions*) and the conditions at the end (*boundary conditions*).

Examples of boundary conditions

a) If the end x_0 of a string or a bar moves in accordance with the law $\mu(t)$, then

$$u\big|_{x=x_0}=\mu(t).$$

b) If the right end x_1 of the string is subjected to the effect of the given force $\nu(t)$, then

$$\left.\frac{\partial u}{\partial x}\right|_{x=x_1}=\frac{\nu(t)}{T_0}.$$

c) If the right end x_1 of the string is fixed elastically and α is the coefficient of fixing, then

$$E\frac{\partial u}{\partial x}+\alpha u\big|_{x=x_1}=0$$

in accordance with Hooke's law.

A partial case of equation (27) is also *an equation of small transverse oscillations of a membrane*:

$$\rho\frac{\partial^2 u}{\partial t^2}=T_0\left(\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}\right)+F. \tag{31}$$

If density ρ is constant, the equation of oscillations of the membrane

$$\frac{\partial^2 u}{\partial t^2}=a^2\left(\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}\right)+f,\ a^2=\frac{T_0}{\rho},\quad f=\frac{F}{\rho}, \tag{32}$$

is also referred to as the *two-dimensional wave equation*.

The three-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2}=a^2\left(\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}+\frac{\partial^2 u}{\partial x_3^2}\right)+f \tag{33}$$

describes the processes of the propagation of sound in a homogeneous medium and of electromagnetic waves in a homogeneous non-conducting medium. This equation is satisfied by the density of the gas, its pressure and the potential of velocities, and also by the components of the strength of electrical and magnetic fields and appropriate potentials.

The wave equations (29),(32),(33) can be expressed by a single equation:

$$\Box_\alpha u = f \tag{34}$$

where $\Box_\alpha$ *is the wave operator (d'Alembert operator),*

$$\Box_\alpha = \frac{\partial^2}{\partial t^2} - a^2\Delta \quad (\Box = \Box_1).$$

3.1.3. Helmholtz equation

Let the external perturbation $f(x, t)$ in the wave equation (34) be periodic with frequency ω with amplitude $a^2 f(x)$:

$$f(x,t) = a^2 f(x) e^{i\omega r}.$$

If we find periodic perturbations $u(x,t)$ with the same frequency and the unknown amplitude $u(x)$: $u(x,t) = u(x)e^{i\omega t}$, then the following stationary equation is obtained for the function $u(x)$:

$$\Delta u + k^2 u = -f(x), \quad k^2 = \frac{\omega^2}{a^2}. \tag{35}$$

which is referred to as *the Helmholtz equation.*

Problems of scattering (diffraction) lead to boundary-value problems for the Helmholtz equation. For example, let us assume that an incident (from infinity) flat wave $e^{i\omega t}$, $|a| = 1$, $k > 0$, is given. This wave changes as a result of the presence of an obstacle at the boundary ∂Ω of the bounded domain Ω. The obstacle may be specified, using the condition $u\big|_{\partial\Omega} = 0$ or $\left.\frac{\partial u}{\partial u}\right|_{\partial\Omega} = 0.$ This obstacle generates a scattered wave $v(x)$. Away from scattering centres, the wave is close to a diverging spherical wave

$$v(x) = f\left(\frac{x}{|x|}\right)\frac{e^{ik|x|}}{|x|} + o\left(|x|^{-1}\right), \tag{36}$$

Therefore, at $|x| \to \infty$ the wave $v(x)$ should satisfy the conditions of the type:

$$v(x) = O\left(|x|^{-1}\right), \quad \frac{\partial v(x)}{\partial |x|} - ikv(x) = o\left(|x|^{-1}\right), \tag{37}$$

referred to as the *Sommerfeld radiation conditions*. The total disturbance $u(x)$ outside the domain Ω consists of flat and scattered waves;

$$u(x) = e^{ik(a,x)} + v(x). \tag{38}$$

It should also be mentioned that function $f(s), \; s = \frac{x}{|x|}$, from (36) is referred to as the *scattering amplitude*.

3.1.4. Diffusion and heat conduction equations

The processes of propagation of heat and of diffusion of particles in a medium are described by the following general diffusion equation:

$$\rho\frac{\partial u}{\partial t} = \operatorname{div}(p\operatorname{grad} u) - qu + F(x,t), \tag{39}$$

where ρ is the coefficient of porosity of the medium and p and q characterize its properties. In examining heat propagation, $u(x,t)$ is the temperature of the medium at the point $x = (x_1, x_2, x_3)$ at time t. Assuming that the medium is isotropic, $\rho(x)$, $c(x)$ and $k(x)$ denote respectively its density, specific heat capacity and heat conduction coefficient, and $F(x,t)$ denotes the intensity of heat sources. The process of heat propagation is described by a function satisfying the equation of type:

$$c\rho\frac{\partial u}{\partial t} = \operatorname{div}(k\operatorname{grad} u) + F(x,t). \tag{40}$$

If the medium is homogeneous, i.e. c, ρ and k are constant, equation (40) has the form:

$$\frac{\partial u}{\partial t} = a^2\Delta u + f, \tag{41}$$

where

$$a^2 = \frac{k}{c\rho}, \quad f = \frac{F}{c\rho}.$$

Equation (41) is the *heat conductivity equation*. The number n of spatial variables $x_1, x_2, \ldots x_n$ in this equation may have any value.

As in the case of the oscillation equation, for complete description of the sources of heat propagation it is necessary to specify the initial distribution of temperature u in medium (in initial condition) and the conditions at the boundary of this medium (boundary condition).

Examples of boundary conditions

a) if the given temperature distribution u_0 is maintained at boundary $\partial\Omega$, then

$$u\big|_{\partial\Omega} = u_0. \tag{42}$$

b) if the given heat flow u_1 is maintained on $\partial\Omega$, then

$$k\frac{\partial u}{\partial n}\bigg|_{\partial\Omega} = u_1. \tag{43}$$

c) if heat exchange on $\partial\Omega$ takes place in accordance with Newton's law, then

$$k\frac{\partial u}{\partial n} + h(u - u_0)\bigg|_{\partial\Omega} = 0, \tag{44}$$

where h is the heat exchange coefficient and u_0 is the temperature of the medium surrounding Ω.

3.1.5. Maxwell and telegraph equations

Maxwell equations are a system of equations for vectors $E = (E_1, E_2, E_3)$, giving the strength of electrical and magnetic fields in some medium. In the Gaussian system of units CGS, the system has the following form:

$$\operatorname{div} D = 4\pi\rho, \quad \operatorname{rot} E = -\frac{1}{c}\frac{\partial B}{\partial t},$$
$$\operatorname{div} B = 0, \qquad \operatorname{rot} H = -\frac{4\pi}{c} j + \frac{1}{c}\frac{\partial D}{\partial t}, \tag{45}$$

where ρ is the density of electrical charges, c is the velocity of light in vacuum and in the case of fields in vacuum $D = E$, $B = H$, $j = 0$, and for any isotropic media $E = \varepsilon E$, $B = \mu H$, $j = \sigma E + j_s$, where ε in the dielectric permittivity of the medium, μ is the magnetic permittivity of the medium, σ is specific electrical conductivity (ε, μ, σ may be functions of t, x), j_s is the density of secondary currents, i.e. currents sustained by any forces, with exception of the forces of the electrical field (for example, by a magnetic field or diffusion).

The Maxwell system (45) is the basis of the theory of electromagnetic waves and is used for all radio electrical engineering calculations, for example, in waveguide theory. The boundary and initial conditions for the system are usually specified on the basis of physical considerations.

From the Maxwell equations, in particular, it is possible to derive *telegraph equations*, important for electrical engineering, which describe the variation of the intensity of current and voltage in a conductor

$$\frac{\partial i}{\partial x} + C\frac{\partial v}{\partial t} + Gv = 0,$$
$$\frac{\partial v}{\partial x} + L\frac{\partial i}{\partial t} + Ri = 0, \tag{46}$$

where x is the coordinate along the conductor, v is the voltage at the given point of the conductor (counted from an arbitrary initial level), i is current intensity, R is specific resistance (per unit length), L is specific self-induction, C is specific capacity, G is specific leakage.

3.1.6. Transfer equation

Instead of the diffusion equation, the process of the propagation of particles is also described by more accurate equations, the so-called *transfer equations* (*kinetic equations*). One of the representatives of this class of equations is the *single-speed transfer equation* of the type

$$\frac{1}{v}\frac{\partial \varphi}{\partial t} + (s, \operatorname{grad})\varphi + \sigma\varphi = \frac{\sigma_s}{4\pi}\int_{S_1} \varphi(x, s', t)\,ds' + F, \tag{47}$$

where $\varphi = vN(x,s,t)$ is the flux of particles flying with (the same) velocity v in the direction $s = (s_1, s_2, s_3)$, $|s| = 1$; $N(x,s,t)$ is the density of particles; $F(x,s,t)$ is the density of sources, coefficients $\sigma(x,t)$, $\sigma_s(x,t)$ characterize the properties of the medium; S_1 is the sphere with unit radius in $\mathbb{R}^3$.

For a complete description of the process of transfer of particles it is necessary to specify the initial distribution of the flux of particles φ in the domain $\Omega \subset \mathbb{R}^3$ (initial condition) and the regime at the boundary of this domain (boundary condition). For example, if the domain Ω in which transfer

takes place is convex, the boundary condition of the type

$$\varphi(x,s,t)=0, \quad x\in\partial\Omega, \quad (s,n)<0, \tag{48}$$

where $n = n(x)$ is the unit vector of the external normal to the boundary of the domain Ω, expresses the absence of an incident flux of the particles on the domain from the outside.

It should be mentioned that the transfer equation describes the processes of transfer of neutrons in a nuclear reactor, the transfer of radiant energy, the passage of γ-quanta through matter, movement of gases, and others.

3.1.7. Gas- and hydrodynamic equations

We examine the movement of an ideal liquid (gas), i.e. the liquid with no viscosity. Let $V(x,t)=(v_1,v_2,v_3)$ be the vector of velocity of the motion of the liquid, $\rho(x,t)$ is its density, $p(x,t)$ is pressure, $f(x,t)$ is the intensity of sources, and $F(x,t)= (F_1,F_2,F_3)$ is the intensity of mass forces. Consequently, these quantities satisfy the following non-linear system of equations, referred to as the hydrodynamics (gas dynamics) equations:

$$\frac{\partial\rho}{\partial t}+\operatorname{div}(\rho V)=f, \tag{49}$$

$$\frac{\partial V}{\partial t}+(V,\operatorname{grad})V+\frac{1}{\rho}\operatorname{grad}p=F. \tag{50}$$

Equations (49) and (50) are the *equations of continuity* and the *Euler equation of motion*. To close this system of equations, it is also necessary to specify the link between pressure and density:

$$\Phi(p,\rho) = 0 \tag{51}$$

which is the so called *equation of state*. For example, for an incompressible liquid the equation of state has the form ρ = const, and for the adiabatic motion of gas

$$p\rho^{-\kappa}=\text{const}, \quad \kappa=\frac{c_p}{c_v},$$

where c_p and c_v are the specific heat capacities of the gas at constant pressure and constant volume, respectively.

In particular, if the liquid is incompressed (ρ = const) and its motion is potential ($V = -\operatorname{grad} u$), the equation of continuity (49) shows that the potential u satisfies Poisson's equation.

3.1.8. Classification of linear differential equations

When examining linear differential equations, with partial derivatives in mathematical physics, we define three main types of equations: elliptical, parabolic and hyperbolic. The simplest equations of these types are respectively the Laplace equation, the heat conductivity equation and the wave equation.

We examined the general linear equation of the second order in $\mathbb{R}^n$

$$\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \ldots = 0 \tag{52}$$

where the coefficients $a_{ij}(x) \equiv a_{ij}(x)$ are real, and the dotted line denotes the younger numbers (the numbers containing only u and $\frac{du}{dx_j}$ but not the second derivatives in respect to u). We introduce the quadratic form associated with equation (52)

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i \xi_j. \tag{53}$$

Direct calculations show that when replacing independent variables $y=f(x)$ the quadratic form does not change if the vector $\xi=(\xi_1,\ldots,\xi_n)$ is transformed using the matrix T^{t-1}, transposed and reciprocal to the Jacobi matrix $T=f'(x)$, examined at point x. In particular, the invariants of linear transformations of the quadratic form (rank, number of positive coefficients and the number of negative coefficients at squares in its canonic form) do not change when the independent variables of the equation are replaced. The canonic form of quadratic equation (53) is determined by the eigenvalues of the symmetric matrix $||a_{ij}(x)||^n_{i,j=1}$. In particular, the elliptical form of equation (52) at point x is equivalent to the point that all these eigenvalues are of the same sign, the hyperbolic form to $n-1$ eigenvalues of the same sign, and one has the opposite sign; finally, the parabolicity at point x indicates that there is one zero value and all other values have the same sign.

Fixing point x, as a result of the linear substitution of the independent variables of the equation (52) we can achieve that the quadratic form (53) acquires the canonic form. This means that the equation itself acquires the following canonical form at point x:

$$\sum_{j=1}^{r} \pm\frac{\partial^2 u}{\partial x_j^2} + \ldots = 0, \tag{54}$$

where r is the rank of the quadratic form (53). In particular, if the initial equation is elliptical, all signs in (54) will be the same since, changing the sign if necessary, we obtain an equation with the main part at point x which is the same as in the Laplace equation. For the hyperbolic equation, the main part in the canonic form at point x coincides with the main part of the wave equation in $\mathbb{R}^n$, and for the parabolic equation the main part becomes a Laplacian in respect of $n-1$ variables in $\mathbb{R}^n$. (It should be mentioned that the reduction of the equation to the form (54) by the described transformation in the entire domain (not at a single point) is, generally speaking, not possible).

If it is also decided to multiply equation (52) by a real number different from zero (or by a real-valued function which does not convert to zero), the positive and negative coefficients of the canonic form of equation (53) can swap places. This gives the meaning to the following definitions.

Definition 1. a) Equation (52) is elliptical at point x if the canonic form of the quadratic equation (53) contains n positive or n negative coefficients, i.e. the form is positively or negatively defined.

b) Equation (52) is referred to as *hyperbolic at point x* if the quadratic form (53) has the rank n and its canonic form contains (after a possible sign change) n–1 positive and one negative coefficient.

c) Equation (52) is *parabolic at point x* if the quadratic form (53) has rank n–1 and after a possible change of the sign becomes non-negatively defined, i.e. its canonic form contains n–1 positive or n–1 negative coefficients.

If any of the conditions a),b),c) is valid at all $x \in \Omega$, where Ω is the domain R^n, then we talk about the ellipticity, hyperbolicity or parabolicity in domain Ω.

In mathematical physics, it is also necessary to solve *mixed equations*, i.e., equations having different type at different points of the examined domain. For example, the *Tricomi equation*

$$yu_{xx}+u_{yy}= 0 \tag{55}$$

examined in R^2 is elliptical at $y > 0$, hyperbolic at $y < 0$ and parabolic at the straight line $y = 0$. This equation forms when describing the motion of a solid in a gas with the velocity close to supersonic velocity: the hyperbolic domain $y < 0$ corresponds to motion with subsonic velocity, and the elliptical domain $y > 0$ to motion with supersonic velocity.

We examine a general non-linear differential operator

$$A=\sum_{|\alpha|\le m} a_\alpha(x)D^\alpha \tag{56}$$

in domain $\Omega\subset\mathrm{R}^n$ and the appropriate equation

$$Au=F. \tag{57}$$

We introduce the *main symbol* of operator A defined by the equation

$$a_m(x,\xi)=\sum_{\alpha=m} a_\alpha(x)\xi^\alpha. \tag{58}$$

Definition 2. Operator (56) and equation (57) are *elliptical at point x*, if $a_m(x,\xi) \neq 0$ at all $\xi\in\mathrm{R}^n\backslash\{0\}$. If this is fulfilled at all $x \in \Omega$, operator A and equation (57) are *elliptical in domain* Ω or simply *elliptical*.

The hyperbolic form of the equation or system is usually determined in the presence of a specific variable (usually it is time) or only for a specific direction (in the presence of a specific variable this direction is represented by the direction of axis t).

Definition 3. Operator A of type (56) and equation (57) are *hyperbolic in the direction of the vector* ν (at point x) if $a_m(x,\nu) \neq 0$, (i.e. direction ν is not characteristic) and for any vector $\xi\in\mathrm{R}^n$, not proportional to ν, all roots λ of the equation

$$a_m(x,\xi+\lambda\nu)=0 \tag{59}$$

are real. The operator (58) and equation (57) are strictly hyperbolic in the

direction of vector ν (at point x) if all roots of equation (59) (there are m such roots, because of the characteristicity condition) are real and different.

The same procedure is used to determine hyperbolicity for the matrix operator A of type (56) (size $N \times N$) and the appropriate system (57): the condition of noncharacteristicity has the form $a_m(x,\nu) \neq 0$ and instead of equation (59) the following equation should be examined in this case

$$\det a_m(x,\xi + \lambda v) = 0$$

The systems of the first order with the specific variable t having the form

$$\frac{\partial u}{\partial t} + \sum_{j=1}^{n} A_j \frac{\partial u}{\partial x_j} + Bu = F, \tag{60}$$

are often encountered. Here u is the N-component vector-function, A_j, B are the matrices $N \times N$ (depending on t,x), f is the known vector-function from t, x, the hyperbolicity condition (strict hyperbolicity) of such a system (in relation to the direction of axis t) indicates that for any real $\xi_1,\ldots,\xi_n$ all eigenvalues of the matrix $\sum_{j=1}^{n} \xi_j A_j$ are real (real and different, respectively). In particular, if all matrices A_j are symmetric, the system is hyperbolic (these systems are referred to as *symmetric hyperbolic systems*).

3.2. Formulation of the main problems of mathematical physics

We formulate the main boundary-value (initial boundary-value) problems of mathematical physics.

3.2.1. Classification of boundary-value problems

As mentioned previously, the linear differential equation of the second order

$$\rho \frac{\partial^2 u}{\partial t^2} = \operatorname{div}(p \operatorname{grad} u) - qu + F(x,t) \tag{61}$$

describes the vibration processes, the equation

$$\rho \frac{\partial u}{\partial t} = \operatorname{div}(p \operatorname{grad} u) - qu + F(x,t) \tag{62}$$

describes the diffusion processes and, finally, the equation

$$-\operatorname{div}(p \operatorname{grad} u) + qu = F(x) \tag{63}$$

describes the appropriate stationary process.

Let it be that $\Omega \subset \mathrm{R}^n$ is a domain in which a physical process takes place, and $\partial\Omega$ is its boundary which is assumed to be a piecewise smooth surface. The domain of variation of arguments x – domain Ω – in the case of equation (63) is the *domain of definition of the equation*. The time variable t belongs to $(0, T)$.

It is assumed that the coefficients ρ, p and q of equations (61)–(63) do not depend on t. In addition, in accordance with their physical meaning it is assumed that $\rho(x)>0$, $p(x)>0$, $q(x)>0$, $\rho(x)>0$ $x \in \bar{\Omega}$. Also, in accordance

with the mathematical meaning of equation (61)–(63) it should be assumed that $\rho \in \bar{\Omega}$, $p \in C^1(\bar{\Omega})$, and $q \in C(\bar{\Omega})$.

Under these assumptions, according to the given classification, the equation of vibration (61) *is a hyperbolic equation, the diffusion equation* (62) *is a parabolic equation, and the stationary equation* (66) *is elliptical.*

As already mentioned, in order to describe completely some physical processes, it is necessary, in addition to the equation describing this process, to define the initial state of the process (*initial conditions*) and the regime at the boundary of the domain in which the process takes place (*boundary conditions*). Mathematically, this is associated with the non-uniqueness of the solution of differential equations. Therefore, in order to define the solution describing the real physical process, it is necessary to specify additional conditions. These additional conditions are also *boundary-value conditions*: initial and boundary conditions. The corresponding problem is the *boundary-value problem*. Thus, the boundary-value problem of mathematical physics is a differential (integro-differential) equation (or a system of equations with the given boundary-value conditions.

Thus, there are three main types of boundary-value problems for differential equations.

a) *The Cauchy problem* for hyperbolic and parabolic equations: the initial conditions are specified, the domain Ω coincides with the entire space R^n, there are no boundary conditions.

b) The *boundary-value* elliptical problem: the boundary conditions at the boundary $\partial\Omega$ are specified, there are no initial conditions.

c) The *mixed problem* for hyperbolic and parabolic equations: initial and boundary conditions are given, $\Omega \neq \mathrm{R}^n$.

We describe in greater detail the formulation of each of these boundary-value problems for the examined equation (61)–(63).

3.2.2. The Cauchy problem

For the equations of vibrations (61) (hyperbolic type) the Cauchy problem is defined as follows: to find the function $u(x,t)$ of the class $C^2(t > 0)\cap C^1(t \geq 0)$ satisfying the equation (61) in the half space $t > 0$ and the initial conditions at $t = +0$:

$$u\big|_{t=0} = u_0(x), \quad \left.\frac{\partial u}{\partial t}\right|_{t=0} = u_1(x). \tag{64}$$

For this purpose it should be that $F \in C(t > 0)$, $u_0 \in C^1(\mathrm{R}^n)$, $u_1 \in C(\mathrm{R}^n)$.

For the diffusion equation (62) (parabolic equation), the Cauchy problem is defined as follows: to find function $u(x,t)$ of the class $C^2(t>0)\cap C(t\geq 0)$ satisfying equation (62) in the half space $t > 0$ and the initial condition at $t = +0$:

$$u\big|_{t=0} = u_0(x). \tag{65}$$

In this case it must be that $F \in C(t > 0)$, $u_0 \in C(\mathrm{R}^n)$.

The given formulation of the Cauchy problem permits the following generalization. Let the quasi-linear differential equation of the second order

of the hyperbolic type be given:

$$\frac{\partial^2 u}{\partial t^2} = \sum_{ij=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} a_{i0} \frac{\partial^2 u}{\partial x_i \partial t} = \Phi\left(x, t, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}, \frac{\partial u}{\partial t} \right) \tag{66}$$

together with the piecewise smooth surface $\Sigma=[t = \sigma(x)]$ and functions u_0, u_1 on Σ (*Cauchy data*). The Cauchy problem for equation (66) consists of the determination, in some part of the domain $t > \sigma(x)$ adjacent to surface Σ, of the solution $u(x,t)$ satisfying the boundary-value conditions on Σ

$$u\big|_{\Sigma} = u_0, \quad \left.\frac{\partial u}{\partial n}\right|_{\Sigma} = u_1, \tag{67}$$

where n is the normal to Σ directed to the side of increasing values of t.

3.2.3. The boundary-value problem for the elliptical equation

The boundary-value problem for equation (63) (elliptical type) consists of the determination of the function $u(x,t)$ of the class $C^2(\Omega)\cap C^1(\Omega)$ satisfying, in domain Ω the equation (63) and the boundary condition on $\partial\Omega$ of the type

$$\alpha u + \beta \left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = \nu, \tag{68}$$

where α, β and ν are the given piecewise-continuous functions on $\partial\Omega$, and $\alpha(x)>0$, $\beta(x)\geq 0$, $\alpha(x)+\beta(x)>0$, $x\in\partial\Omega$. The following types of boundary condition (68) are defined: the boundary condition of the first kind ($\alpha=0$, $\beta=0$).

$$u\big|_{\partial\Omega} = u_0; \tag{69}$$

the boundary condition of the second kind ($\alpha=0$, $\beta=1$)

$$\left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = u_1; \tag{70}$$

the boundary condition of the third kind ($\alpha\geq 0$, $\beta=0$)

$$\alpha u + \left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = u_2. \tag{71}$$

The appropriate boundary-value problems are referred to as the *boundary-value problems of the first, second and third kind.*

For Laplace and Poisson equations, the boundary-value problem of the first kind

$$\Delta u = -f, \quad u\big|_{\partial\Omega} = u_0 \tag{72}$$

is referred to as *Dirichlet problem*; the boundary-value problem of the second kind

$$\Delta u = -f, \quad \left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = u_1 \tag{73}$$

is the *Neumann problem*.

The same procedure is used for the defining the boundary-value problems

for equation (63) and outside the bounded domain Ω (*the external boundary-value problems*). The only difference is that in addition to the boundary condition (68) on $\partial\Omega$, the conditions at infinity are also specified. These conditions can be, for example: the conditions of Sommerfeld radiation – for the Helmholtz equation; the conditions of the type

$$u(x)=O(1) \text{ or } u(x)=o(1),\ |x|\to\infty \tag{74}$$

for the Poisson equation.

3.2.4. Mixed problems

For the equation of vibration (62) (hyperbolic type) the mixed problem is defined as follows: to find the function $u(x,t)$ of the class $C^2(Q_T)\cap C^1(\overline{Q}_T)$ where $Q_T \equiv \Omega \times (0,T)$ satisfying the equation (61) in the cylinder Q_T, the initial condition (64) at $t = 0$, $x = \overline{\Omega}$ (on the lower base of the cylinder Q_T) and the boundary condition (68) (on the side surfaces of the cylinder Q_T). The conditions of smoothness

$$F \in C(Q_T),\ \ u_0 \in C^1(\overline{\Omega}),\ \ u_1 \in C(\overline{\Omega}),$$

here v is the piecewise function continuous on $\partial\Omega \times [0,T]$, and the compatibility conditions

$$\alpha u_0 + \beta \left.\frac{\partial u_0}{\partial n}\right|_{\partial\Omega} = v\big|_{t=0},\quad \alpha u_1 + \beta \left.\frac{\partial u_1}{\partial n}\right|_{\partial\Omega} = \left.\frac{\partial v}{\partial t}\right|_{t=0}. \tag{75}$$

must be fulfilled. (The second of the equalities (75) has any meaning if solution $u(x,t)$ is sufficiently smooth up to the lower base Q_T).

Similarly, the mixed problem for the diffusion equation (62) (parabolic type) is defined as follows: find the function $u(x,t)$ of the class $C^2(Q_T)\cap C(\overline{Q}_T)$ satisfying the equation (62) in Q_T, the initial condition (65) and boundary condition (68).

In mathematical physics it is often necessary to solve other boundary-value problems, differing from those formulated previously (f*or example, Goursat problem for a linear hyperbolic equation, Zaremba problem for the Laplace equation, and others*).

3.2.5. Validity of formulation of problems. The Cauchy–Kovalevskii theorem

Since the problems of mathematical physics are in fact mathematical models of real physical processes, the following actual requirements are often imposed on their formulation:

a) the solution should *exist* in some class of functions X_1;

b) the solution must be *unique* in some class of functions X_2;

c) the solution should *depend continuously* on the given data of the problem (initial and boundary data, the free member, the coefficients of the equation, etc).

The continuous dependence of the solution u on the data of the problem

F denotes the following: let it be that the sequence of data F_k (k = 1,2,...) tends in some sense to F and u_k (k = 1,2,...), u are the respective solutions of the problem; consequently, it should be that $u_k \to u$, $k \to \infty$, in the sense of convergence selected in the appropriate manner. For example, let it be that the problem is reduced to the equation $Au = F$, where A is a linear operator transforming X to Y, where X and Y are the linear normalized spaces. In this case, the continuous dependence of the solution u on the free member F is ensured, if the operator A^{-1} does exist and is bounded from Y in X. The requirement for the continuous dependence of the solution is determined by the fact that the physical data are usually determined approximately from the experiments and it is therefore necessary to be sure that the solution of the problem within the framework of the selected mathematical model does not depend greatly on the error of measurements.

The problem, satisfying these requirements, is referred to as a *correctly* (*well*) *posed problem* (according to Hadamard) and the set of function $X_1 \cap X_2$ is the class of validity. The problem which does not satisfy at least one of the conditions a)– b) is referred to as an *ill-posed problem*.

The ill-posed problems often result from *inverse problems of mathematical physics*: use information on the solution of the direct problem to restore some unknown physical quantities (sources, boundary-value conditions, the coefficients of an equation, etc) determining this problem.

We specify a relatively large class of the Cauchy problems for which there is a solution and it is a unique solution. However, initially we introduce two definitions.

1. The system of N differential equations with N unknown function u_1, $u_2,\ldots,u_n$

$$\frac{\partial^{ki} u_i}{\partial t^{k_i}} = \Phi_i(x,t,u_1,u_2,\ldots,u_n,\ldots,D_t^{\alpha_0} D_x^{\alpha} u_j,\ldots);\ i = 1,2,\ldots,N, \tag{76}$$

is referred to as *normal in relation to variable t* if the right hand sides Φ_i do not contain any derivatives of the order higher than k_i and derivatives in the respect of t of the order higher than k_i–1, i.e. therefore

$$\alpha_0+\alpha_1+\ldots+\alpha_n \le k_i,\ \ \alpha_0 \le k_i - 1.$$

For example, the wave equation, the Laplace equation and the heat conductivity equation are normal in relation to every variable x; the wave equation, in addition to this, is normal in relation to t.

2. Function $f(x)$, $x = (x_1,x_2,\ldots,x_n)$ is *analytical at point* x_0 if at some vicinity of this point it is represented in the form of a uniformly converging exponential series

$$f(x) = \sum_{|\alpha| \ge 0} c_\alpha (x - x_0)^\alpha = \sum_{|\alpha| \ge 0} \frac{D^\alpha f(x_0)}{\alpha 1}(x - x_0)^\alpha$$

(point x_0 may also be complex). If function $f(x)$ is analytical at every point of domain Ω, then it is said that it is *analytical in* Ω.

For the system of equations (76) normal in relation to t we define the following Cauchy problem: to find solution $u_1,u_2,\ldots,u_N$ of this system, satisfying initial conditions at $t = t_0$

$$\left.\frac{\partial^k u_i}{\partial t^k}\right|_{t=t_0} = \varphi_{ik}(x),\ k = 0,1,...,k_i - 1;\ i = 1,2,...,N, \tag{77}$$

where $\varphi_{ik}(x)$ is the given function in some domain $\Omega \subset R^n$.

Theorem 23 (Cauchy–Kovalevskii theorem). *If all functions* $\varphi_{ik}(x)$ *are analytical in some neighbourhood of point* x_0 *and all function* $\Phi_i(x,t,...,(u_j)\alpha_0, \alpha_1,...,\alpha_n,...)$ *are analytical in some neighbourhood of the point* $(x_0,t,...,D^\alpha\varphi_{j\alpha}(x_0),...)$, then *the Cauchy problem* (76), (77) *has an analytical solution in some neighbourhood of point* (x_0, t_0) *and it is the only solution in the class of analytical functions.*

It should be mentioned that the Cauchy–Kovalevskii theorem, regardless of its general nature, does not completely solve the problem of the validity of formulation of the Cauchy problem of the normal system of differential equations. This theorem guarantees only existence and uniqueness of the solution in a relatively small neighbourhood or, in other words, *in the small*; usually, these facts must be determined in the domain determined in advance (and by no means small ones) or, *on the whole*. In addition to this, the initial data and the free term of the equation are usually non- analytical functions. Finally, there may not be any continuous dependence of the solution on the initial data (this is indicated by the well-known Hadamard example).

3.3. Generalized formulations and solutions of mathematical physics problems

The formulations of the boundary-value problems, described in previous paragraphs, are characterized by the fact that the solutions of these problems are assumed to be relatively smooth and they should satisfy the equation at every point of the domain of definition of this equation. These solutions will be referred to as *classic*, and the formulation of the appropriate boundary problem as *classic formulation*. Thus, the classic formulations of the problems already assume the sufficient smoothness of the data included in the problem. However, in the most interesting problems these data may have relatively strong special features. Therefore, for these problems classic formulations are already insufficient. In order to formulate such problems, it is necessary to avoid (completely or partially) using the requirement of the smoothness of the solution in the domain or upto the boundary, introduce the so-called *generalized solutions* and *generalized formulations* of mathematical physics problems.

One of the directions in the theory of generalized solutions and formulation of boundary-value problems is based on the application of Sobolev functional spaces. In this case, the classification theorem and theorems of existence of traces (boundary value), determined for these spaces, give meaning to the boundary conditions for the mathematical physics equations, treating these conditions as additional equations in the corresponding spaces ('space of traces'). In a number of problems it is also possible to exclude the obvious presence of the boundary conditions in the generalized formulation of the

problem, 'including' them together with the main equation in some integral identity (the so-called 'natural boundary conditions').

We formulate the main approaches to introducing generalized formulations of the problems and generalized solutions on examples of several main problems of mathematical physics using the Sobolev space.

3.3.1. Generalized formulations and solutions of elliptical problems

Dirichlet problems

We examine the simplest elliptical boundary-value problems – the Dirichlet problem for the Laplace or Poisson equation, and allocate it a generalized formulation. Initially, we examine the problem for the Poisson equation with the zero boundary conditions:

$$-\Delta u(x) = f(x), \quad x \in \Omega,$$
$$u\big|_{\partial\Omega} = 0. \tag{78}$$

Instead of the boundary conditions $u|_{\partial\Omega} = 0$ we write $u \in \overset{\circ}{W}{}_2^1(\Omega)$ (this inclusion in the case of bounded domains with a smooth (piecewise smooth) boundary is equivalent to $u \in W_2^1(\Omega)$ and $u|_{\partial\Omega} = 0$). Multiplying both parts of the equation $-\Delta u = f$ by $\overline{v}(x)$, where $v \in C_0^\infty(\Omega)$, and integrating by parts, we obtain

$$[u,v] = (f,v), \tag{79}$$

where (.,.) denotes the scalar product in $L_2(\Omega)$, and

$$[u,v] = \int_\Omega \sum_{j=1}^n \frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx,$$

since [.,.] is the form continuous in space $W_2^1(\Omega)$, i.e. $|[u,v]| \le C\|u\|_{W_2^1(\Omega)}\|v\|_{W_2^1(\Omega)}$, where the constant $C > 0$ does not depend on u, v. The quantity

$$\mathcal{D}(u) = [u,u] = \int_\Omega |\nabla u(x)|^2\,dx = \int_\Omega \sum_{j=1}^n \left|\frac{\partial u(x)}{\partial x_j}\right|^2 dx$$

is the *Dirichlet integral*.

Equality (79) has the meaning for any function u, $v \in \overset{\circ}{W}{}_2^1(\Omega)$ and for $f \in L_2(\Omega)$. It will be examined instead of problem (78). In this case, we can select the function v which are such that $v \in C_0^\infty(\Omega)$. In the case of a classic solution of u (i.e. solution $u \in C^2(\overline{\Omega})$ of problem (78)), equality (79) is obtained by the previously described procedure at $v \in C_0^\infty$ and then using the limiting transition at $v \in \overset{\circ}{W}{}_2^1(\Omega)$.

Thus, we obtain the following *generalized formulation of the problem* (78): *at the given function* $f \in L_2(\Omega)$ *it is necessary to find a function* $u \in \overset{\circ}{W}{}_2^1(\Omega)$ *so that for any function* $v \in C_0^\infty(\Omega)$ *equation (79) is satisfied.*

As already shown, instead of $v \in C_0^\infty(\Omega)$ we can write $v \in \overset{\circ}{W}{}_2^1(\Omega)$ which results in an equivalent formulation. In addition to this, transferring the derivatives from v to u by integration by parts, we obtain that (79) is equivalent to equation $-\Delta u = f$, regarded as generalized functions because the formulated generalized formulation of the problem is equivalent to the following: the function $f \in L_2(\Omega)$ is given; *it is required to find a function* $u \in \overset{\circ}{W}{}_2^1(\Omega)$ *such that* $-\Delta u = f$ *in the sense of generalized functions.*

Any solution u of problem (79) will be referred to as a *generalized* or *weak solution* (in contrast to the classic solution, which can be discussed at $f \in C(\bar{\Omega})$). On the other hand, any classic solution $u \in C^2(\Omega)$ is generalized.

It should be mentioned that $[.,.]$ can be regarded as the scalar product in the space $\overset{\circ}{W}{}_2^1(\Omega)$, . This is equivalent to the situation in which the expression $\|u\|_1 = \sqrt{\mathcal{D}(u)} = [u,u]^{1/2}$ is a norm equivalent to the norm $\|\cdot\|_{W_2^1(\Omega)}$ on $C_0^\infty(\Omega)$. Because of the obvious relationship $\|u\|_{W_2^1}^2 = \|u\|^2 + \mathcal{D}(u)$, the equivalence of the norms $\|\cdot\|_{W_2^1}$ and $\|\cdot\|_1$ results from the so-called *Steklov inequality*

$$\|u\|^2 \le C\mathcal{D}(u), \quad u \in C_0^\infty(\Omega), \tag{80}$$

where $C > 0$ does not depend on u, and $\|\cdot\|$ is the norm in $L_2(\Omega)$.

Using the equivalence of the norms $\|\cdot\|_{W_2^1(\Omega)}$ and $\|\cdot\|_1$ for functions from $\overset{\circ}{W}{}_2^1(\Omega)$ and using the Riesz theorem on the representation of the linear bounded functional it is easy to make the following claim.

Theorem 24. *If* Ω *is any bounded domain in* R^n *and* $f \in L_2(\Omega)$, *then the generalized solution of problem* (78) *does exist and is unique.*

We now examine briefly the Dirichlet problem for the Laplace equation:

$$\begin{aligned} &\Delta u(x) = 0, \quad x \in \Omega, \\ &u\big|_{\partial\Omega} = \varphi \end{aligned} \tag{81}$$

In transition to the generalized formulation it is initially necessary to solve the problem of interpretation of the boundary condition. If the boundary $\partial\Omega$ is smooth and if $\varphi \in W_2^{3/2}(\partial\Omega)$, then, according to the theorem of existence of a trace, there is a function $v \in W_2^2(\Omega)$ such that $v|_{\partial\Omega} = \varphi$. However, if $u \in W_2^1(\Omega)$ is the solution of the problem (81), then for the $w = u - v$ we obtain the problem of type (78) with $f = -\Delta v \in L_2(\Omega)$ since we can transfer to the

generalized formulation (79) and in the case of the bounded domain (Ω) use theorem 24 which now indicates the existence of uniqueness of solution of the problem (81). If the boundary $\partial\Omega$ is not smooth, we can immediately fix the function $v \in W_2^1(\Omega)$ giving the boundary condition and define the problem as follows: *the function* $v \in W_2^2(\Omega)$ *is given; it is required to find a function u such that* $u - v \in \overset{\circ}{W}{}_2^1(\Omega)$ *and also* $\Delta u\ (x) = 0$ at $x \in \Omega$.

Theorem 25. *If* Ω *is some bounded domain in* R^n *and* $v \in W_2^1(\Omega)$, *then the generalized solution u of the problem* (81) *does exist and is unique. This solution has a strictly minimal Dirichlet integral* $\mathcal{D}(u)$ *amongst all functions* $u \in W_2^1(\Omega)$ *for which* $u - v \in \overset{\circ}{W}{}_2^1(\Omega)$. *Conversely, if u is a stationary point of the Dirichlet integral in the class of all function* $u \in W_2^1(\Omega)$ for which $u - v \in \overset{\circ}{W}{}_2^1(\Omega)$, *then u is a generalized solution of problem* (81) (*and, at the same time, the Dirichlet integral has a strict minimum on function u*).

The Neumann problem. The homogenous Neumann problem for the Poisson equation has the form:

$$-\Delta u(x) = f(x), \quad x \in \Omega,$$
$$\left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = 0. \tag{82}$$

For transition to its generalized formulation, it is assumed that Ω is the bounded domain with a smooth boundary, and it is also assumed that initially $u \in C^\infty(\overline{\Omega})$. Multiplying both parts of the equation $-\Delta u = f$ by the function $\overline{v}$, where $v \in C^\infty(\overline{\Omega})$ and subsequently integrating in respect of Ω, we use the Green equation

$$\int_\Omega \Delta u(x)\overline{v}(x)dx = -\int_\Omega \nabla u(x)\nabla\overline{v}(x)dx + \int_{\partial\Omega} \overline{v}(x)\frac{\partial u(x)}{\partial n}dS_x,$$

where dS_x is the element of the surface area of the boundary. Consequently, because of (82):

$$[u,v] = (f,v). \tag{83}$$

As regards continuity, here instead of $v \in C^\infty(\overline{\Omega})$ we can use $v \in W_2^1(\Omega)$ also in the case in which we only know that $u \in W_2^1(\Omega)$ and $f \in L_2(\Omega)$. This gives a *generalized formulation of the Neumann problem: use the function* $f \in L_2(\Omega)$ *to find some function of* $u \in W_2^1(\Omega)$ *such that* (83) *is satisfied for any function* $v \in W_2^1(\Omega)$.

The solution of this problem is unique with the accuracy to the arbitrary constant: if u_1 is another solution of the Neumann problem (with the same function f), and $w = u_1 - u_1$ then $[w,v] = 0$ for any function $v \in W_2^1(\Omega)$. Assuming that $v = w$, we obtain that $[w,w] = 0$. This means that all generalized

derivatives $\frac{\partial w}{\partial x_j}$, $j = 1,2,\ldots,n$, convert to zero and w = const.

The generalized solution of the Neumann problem exists only for the function $f \in L_2(\Omega)$ for which the condition:

$$(f,1) = \int_\Omega f(x)dx = 0$$

is satisfied, i.e. for the functions with the zero mean value. The necessity of this condition follows directly from (83) at $v \equiv 1$.

To confirm the existence of a generalized solution of the Neumann problem, we can use the Poincaré inequality

$$\|u\|^2_{L_2(\Omega)} \le C\left(\mathcal{D}(u) + \left(\int_\Omega u dx\right)^2\right), \; u \in C^\infty(\Omega), \tag{84}$$

where C = const does not depend on u and, consequently, on functions from $W_2^1(\Omega)$, also orthogonal to unity, the norms

$$\|u\|_{W_2^1(\Omega)}, \; \|u\|_1 = \left(\mathcal{D}(u) + \left(\int_\Omega u dx\right)^2\right)^{1/2}$$

are equivalent. Using the well-known Riesz theorem, *we obtain the existence of the unique generalized solution* $u \in W_2^1(\Omega)$ *of the Neumann problem under the condition* $f \in L_2(\Omega)$, $\int_\Omega f(x)dx = 0.$

The boundary-value problem for the general elliptical equation of the second order can be re-formulated and examined by approaches demonstrated previously on the example of the Dirichlet and Neumann problems for the Laplace operator.

3.3.2. Generalized formulations and solution of hyperbolic problems

Let Ω be some bounded domain of the n-dimensional space R^n ($x = (x_1,x_2,\ldots,x_n)$ is the point of this space). In the $(n+1)$-dimensional space $\mathsf{R}^{n+1} = \mathsf{R}^n \times \{-\infty<t<+\infty\}$ we examine a bounded cylinder $Q_T = \{x \in \Omega, 0<t<\mathrm{T}\}$ with the height $T > 0$. Γ_T denotes the side surface $\{x \in \partial\Omega, 0 < t < \mathrm{T}\}$ of the cylinder Q_T and Ω_τ is the section $\{x \in \Omega, t = \tau\}$ of the cylinder by the plane $t = \tau$; in particular, the upper base of the cylinder Q_T is $\Omega_T = \{x \in \Omega, t = T\}$, and its lower base $\Omega_0 = \{x \in \Omega, t = 0\}$.

In the cylinder Q_T at some $T > 0$ we examine the hyperbolic equation

$$Lu \equiv u_t - \operatorname{div}(k(x)\nabla u) + a(x)u = f(x,t), \tag{85}$$

where $k(x) \in C^1\left(\overline{Q}_T\right)$, $a(x) \in C\left(\overline{Q}_T\right)$, $k(x) \ge k_0 = \text{const} > 0$.

The function $u(x,t)$ belonging to the space $C^2(Q_T) \cap C^1(Q_T \cup \Gamma_T \cup \overline{\Omega}_0)$, which satisfies in Q_T equation (85), the initial condition on $\overline{\Omega}_0$

$$u\big|_{t=0} = \varphi, \tag{86}$$

$$u_t\big|_{t=0} = \psi, \tag{87}$$

and one of the boundary conditions on Γ_T:

$$\mu\big|_{\Gamma_T} = \chi$$

or

$$\left(\frac{\partial u}{\partial n} + \sigma u\right)\Bigg|_{\Gamma_T} = \chi,$$

where σ is some function continuous on Γ_T, is referred to as the (*classic*) *solution of the first or correspondingly third mixed problem for equation* (85). If $\sigma \equiv 0$ on Γ_T, then the third mixed problem is referred to as *the second mixed problem*.

Since the case of inhomogeneous boundary conditions is easily reduced to the case of homogeneous boundary conditions, we confine ourselves to the case of homogeneous boundary conditions

$$u\big|_{\Gamma_T} = 0 \tag{88}$$

$$\left(\frac{\partial u}{\partial n} + \sigma u\right)\Bigg|_{\Gamma_T} = 0. \tag{89}$$

It is assumed that coefficient $a(x)$ in equation (85) is non-negative in Q_T, and function σ in the boundary condition (89) depends only on x and is non-negative on Γ_T.

Let it be that function $u(x,t)$ is a solution of one of the problems (85)–(88) or (85), (86), (87), (89) and the right-hand side $f(x,t)$ of equation (85) belongs to $L_2(Q_T)$. Multiplying (85) by $v(x,t) \in W_2^1(Q_T)$, for which the condition (88) and $v|_{QT}=0$ is satisfied, we integrate over Q_T using the integration by parts and the Green equation. Consequently, we obtain the identity

$$\int_{Q_T} (k\nabla u \nabla v + auv - u_t v_t)dxdt = \int_{\Omega_0} \psi v dx + \int_{Q_T} fvdxdt \tag{90}$$

for all $v \in W_2^1(Q_T)$, for which the conditions (88) and the condition

$$v\big|_{\Omega_T} = 0, \tag{91}$$

is satisfied, or

$$\int_{Q_T} (k\nabla u \nabla v + auv - u_t v_t)dxdt + \int_{Q_T} k\sigma uvdSdt = \int_{\Omega_0} \psi v dx + \int_{Q_T} fvdxdt \tag{92}$$

for all $v \in W_2^1(Q_T)$ for which the condition (91) is satisfied.

Using the resultant identities we produced the concept of generalized solution of the examined mixed problems. It is assumed that $f(x,t) \in L_2(Q_T)$, and $\Psi(x) \in L_2(\Omega)$.

The function u, belonging to the space $W_2^1(Q_T)$, is referred to as the generalized solution in Q_T of the first mixed problem (85)–(88) if it satisfies the initial condition (86), the boundary condition (88) and identity (90).

The function u, belonging in the space with $W_2^1(Q_T)$, is referred to as the generalized solution in Q_T of the third (second at σ=0) mixed problem (85)–(87), (89), if it satisfies the condition (86) and identity (92).

It should be mentioned that, like classic solution, generalized solutions have the following properties. If *u is the generalized solution of the problem* (85)–(88) *or of the problem* (85)–(87), (89) *in the cylinder* Q_T, *it is then a generalized solution of the corresponding problem in the cylinder* $Q_{T'}$ at $T' < T$.

Theorem 26. *Let* $f \in L_2(Q_T)$, $\Psi \in L_2(\Omega)$, $\varphi \in \overset{\circ}{W}{}_2^1(Q)$ in the case of the *first mixed problem* (85)–(88), *and* $\varphi \in W_2^1(\Omega)$ *in the case of the third (second) mixed problem* (85)–(87), (89). *Consequently, the generalized solution of the corresponding problem does exist and is unique. The following inequality holds in this case*

$$\|u\|_{W_2^1(Q_T)} \le C\left(\|\varphi\|_{W_2^1(\Omega)} + \|\psi\|_{L_2(\Omega)} + \|f\|_{L_2(Q_T)}\right), \tag{93}$$

In this inequality, the constant C does not depend on φ, ψ, f.

3.3.3. The generalized formulation and solutions of parabolic problems

Let Ω be the bounded domain of the n-dimensional space R^n, and $x = (x_1, x_2, \ldots, x_n)$ is a point of this space. As in the case of mixed problems for hyperbolic equations, we examine in the $(n+1)$-dimensional space $\mathsf{R}^{n+1} = \{-\infty < t < +\infty\}$ a bounded cylinder $Q_T = \{x \in \Omega, 0 < t < T\}$ with the height $T > 0$, and let it be that Γ_T is the side surface of this cylinder $\Gamma_T = \{x \in \partial\Omega, 0 < t < T\}$, and Ω_τ, $\tau \in [0,T]$ is the set $\{x \in \Omega,\ t = \tau\}$, in particular, the upper base of the cylinder Q_T is $\Omega_T = \{x \in \Omega,\ t = T\}$ and its lower base $\Omega_0 = \{x \in \Omega,\ t = 0\}$. $C^{2,1}(Q_T)$ denotes the set of the functions continuous in Q_T having continuous in Q_T derivatives u_t, $u_{x_i}, u_{x_i x_j}$, $C^{1,0}(Q_T \bigcup \Gamma_T)$ is the set of the continuous functions $(Q_T \bigcup \Gamma_T)$ with continuous derivatives $u_{xi}(i, j = 1.2, \ldots, n)$.

In the cylinder Q_T we examine a parabolic equation at $T > 0$:

$$Lu = u_t - \mathrm{div}(k(x)\nabla u) + a(x)u = f(x,t), \tag{94}$$

where $k(x) \in C^1\left(\overline{Q}_T\right)$, $a(x) \in C^1\left(\overline{Q}_T\right)$, $k(x) \ge k_0 = \text{const} > 0$.

The function $u(x,t)$ belonging to the space $C^{2,1}(Q_T) \cap C(Q_T \cup \Gamma_T \cup \overline{\Omega}_0)$, which satisfies in Q_T equation (94) on $\overline{\Omega}_0$ – the initial condition

$$u\big|_{t=0} = \varphi, \tag{95}$$

and on Γ_T – the boundary condition

$$u\big|_{\Gamma_T} = \chi,$$

is the *classic solution of the first mixed problem for equation* (94).

Function $u(x,t)$ belonging to the space

$$C^{2,1}(Q_T)\cap C(Q_T\cup\Gamma_T\cup\bar{\Omega}_0)\cap C^{1,0}(Q_T\cup\Gamma_T),$$

satisfying in Q_T equation (94), on $\bar{\Omega}_0$ the initial condition (95) and on Γ_T the boundary condition

$$\left(\frac{\partial u}{\partial n}+\sigma(x)u\right)\Bigg|_{\Gamma_T} = \chi,$$

where $\sigma(x)$ is some function continuous on Γ_T, is the *classic solution of the third mixed problem for equation* (94). If $\sigma \equiv 0$, the third mixed problem is the second mixed problem.

Since the case of heterogeneous boundary conditions is reduced to homogenous boundary conditions, we examine only the homogeneous boundary conditions:

$$u_t\big|_{\Gamma_T} = 0 \tag{96}$$

$$\left(\frac{\partial u}{\partial n}+\sigma(x)u\right)\Bigg|_{\Gamma_T} = 0. \tag{97}$$

It is assumed that coefficient $a(x)$ in equation (94) is non-negative in Q_T, and function $\sigma(x)$ in the boundary condition (97) is non-negative on Γ_T.

Let function u be a classic solution of the third (second) mixed problem (94), (95), (97) or it is the classic solution of the first mixed problem (94)–(96) belonging to $C^{1,0}(Q_T\cup\Gamma_T)$, and function $f(x,t) \in L_2(Q_T)$. Equation (95) is multiplied by the arbitrary function $v(x,t)\in C^1\left(\bar{Q}_T\right)$ satisfying the condition

$$v\big|_{\Omega_T} = 0, \tag{98}$$

and the resultant equality is integrated over the cylinder Q_T. Using integration by parts and the Green equation we obtain the following claims.

The classic solution $u(x,t)$, belonging to $C^{1,0}(Q_T\cup\Gamma_T)$, of the first mixed problem satisfied the integral identity

$$\int_{Q_T}(-uv_i + k\nabla u\nabla v + auv)dxdt = \int_{\Omega_0}\varphi v dx + \int_{Q_T} fvdxdt \tag{99}$$

at all $v\in C^1\left(\bar{Q}_T\right)$ satisfying the condition (98) $v\big|_{\Gamma_T} = 0$ and, consequently for all $v\in W_2^1(Q_T)$, satisfying condition (98) and $v\big|_{\Gamma_T} = 0$.

Classic solution $u(x,t)$ of the third (second at $\sigma = 0$) mixed problem satisfies the integral identity

$$\int_{Q_T}(-uv_t + k\nabla u\nabla v + auv)dxdt + \int_{\Gamma_T} k\sigma uvdSdt = \int_{\Omega_0}\varphi vdx + \int_{\Omega_T} fvdxdt \tag{100}$$

at all $v \in C^1\left(\overline{Q}_T\right)$ satisfying condition (98), and consequently, for all $v \in W_2^1(Q_T)$ satisfying condition (98).

The resultant identities may be used for introducing the concept of the generalized solutions of the investigated mixed problem.

It is assumed that $f(x,t) \in L_2(Q_T)$ and $\varphi(x) \in L_2(\Omega)$. Function $u(x,t)$, belonging to the space $W_2^{1,0}(Q_T)$, determined from

$$(u,v)_{W_2^{1,0}(Q_T)} = \int_{Q_T}(uv + \nabla u\nabla v)dx, \ \|u\|_{W_2^{1,0}(Q_T)} = (u,u)^{1/2}_{W_2^{1,0}(Q_T)},$$

is referred to as the *generalized solution of the first mixed problem* (94-96) if it satisfies the boundary condition (96) and the identity (99) for all $v(x,t) \in W_2^1(Q_T)$, satisfying the conditions (96) and (98).

The function $u(x,t)$, belonging in the space $W_2^{1,0}(Q_T)$ is the *generalized solution of the third* (*second at* $\sigma = 0$) *mixed problem* (94, 95, 97) if it satisfies identity (100) at all $v(x,t) \in W_2^1(Q_T)$ satisfying condition (98).

It should also be mentioned that the generalized solution of the mixed problem for the parabolic equation, like the classic solution, has the following property: if $u(x,t)$ is a generalized solution of the mixed problem (94)–(96) or problem (94), (95), (97) in the cylinder Q_T, then it is a generalized solution of the appropriate problem and in the cylinder $Q_{T'}$ at any T', $0 < T' < T$.

Theorem 27. *Let* $f \in L_2(Q_T)$, $\varphi \in L_2(Q_T)$, *then each mixed problem* (94)–(96) or (94), (95), (97) *has a generalized solution* $u \in W_2^{1,0}(Q_T)$. *The inequality*

$$\|u\|_{W_2^{1,0}(Q_T)} \le C\left(\|\varphi\|_{L_2(\Omega)} + \|f\|_{L_2(Q_T)}\right) \tag{101}$$

holds in this case and in this inequality positive constant C does not depend on φ, f.

3.4. Variational formulations of problems

Many problems of mathematical physics may be reformulated as variational problems, representing one of the approaches to introducing generalized formulations of the initial boundary-value problems. We examine this approach to consideration of the generalized formulations of problems also referred to as the *energy method*.

3.4.1. Variational formulation of problems in the case of positive definite operators

Let a problem of mathematical physics be reduced to an equation in the real Hilbert space H and is written in the form

$$Au = f, \tag{102}$$

where u is the required element of some functional space, A is the *operator of the boundary-value problem* whose definition domain is dense in H, f is a given element.

If the operator A is symmetric and positive, the solution of equation (102) may be reduced to solution of some variational problem, such as indicated by the following theorem.

Theorem 28. *Let A be a symmetric and positive operator. If the equation $Au = f$ has a solution, then this solution gives the lowest value for the following functional:*

$$J(u) = (Au,u) - 2(u, f). \tag{103}$$

Conversely, if there is an element realizing the minimum of the functional (103), *then this element satisfies equation $Au = f$.*

The method of solution of the boundary-value problems, consisting of replacing equation (102) by a problem of the minimum of the functional (103), is referred to as the *energy method* in the literature. The functional (103) is referred to as the *functional of the energy method.*

Theorem 28 refers neither to the conditions of existence of the solution of the variational problem nor to how to construct this solution. These indications may be given if the operator of the boundary-value problem is positive definite. In this case, the energy space H_A is examined in which $(Au,u)=|u|^2$. Further, according to the Cauchy–Bunyakovskii inequality and the inequality $||u||\le(|u|/\gamma)$, we have $|(u, f)|\le||f||\ ||u||\le||f||/(|u|/\gamma)$. This means that the linear function (u, f) is bounded in H_A; according to the Riesz theorem, there is an element $u_0\in H_A$ such that $(u, f) =[u,u_0]$ if $u \in H_A$. Now, the function (103) leads to the form

$$J(u) = |u|^2 - 2(u, f) = |u - u_0|^2 - |u_0|^2. \tag{104}$$

Two simple and important consequences follow from equation (104):

1) *This equation makes it possible to determine the functional $J(u)$ not only on the elements of the domain of definition of the operator A but also on all elements of the energy space H_A;*

2) *In space H_A functional $J(u)$ reaches the minimum at $u = u_0$.*

If $u_0 \in D(A)$, then according to theorem 28 u_0 is a solution of equation $Au = f$; however, generally speaking, the energy space H_A is wider than D(A) and it may be that element u_0 constructed on the basis of the Riesz theorem and realising the minimum of the function of $F(u)$ in the energy space does not fit in D(A). In this case, u_0 may be regarded as a generalized solution of equation $Au = f$.

3.4.2. Variational formulation of the problem in the case of positive operators

We examine equation $Au = f$, assuming that in the selected Hilbert space the symmetric operator A is positive but is not positive definite. According to theorem 28, our equation is as previously equivalent to the problem of

the minimum of the functional (103) but in this case this variational problem is, generally speaking, insolvable even in the generalized sense. We shall show the essential and sufficient condition of solvability of this problem.

As in the case of a positive definite operator, we can construct the energy space H_A; at this time, it contains not only the elements of the initial space but also some new elements. We have $(Au, u) = |u|^2$ since

$$J(u) = |u|^2 - 2(u, f).$$

I*n order to ensure that the problem of the minimum of the functional* $F(u)$ *has a solution in* H_A*, it is necessary and sufficient that the linear functional* (u, f) *is bounded in this space*. In this case, according to the Reisz theorem, there is an element $u_0 \in H_A$ such that $(u, f) = [u,u_0]$ if $u \in H_A$; element u_0 realizes the minimum of the functional $J(u)$ in space H_A.

The boundary-value problems for infinite domains often lead to positive operators.

The fact that $u_0 \in H_A$ may be interpreted physically that this element has a finite energy; if the condition of restriction of the function (u,f) in H_A is fulfilled, then the corresponding element u_0 is referred to as *a solution with finite energy for the equation* $Au = f$.

3.4.3. Variational formulation of the basic elliptical problems

In $L_2(\Omega)$ we examine the self-adjoint elliptical equation of the second order:

$$Au = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(A_{ij} \frac{\partial u}{\partial x_j} \right) + C(x)u = f(x), \quad A_{ij} = A_{ji}. \tag{105}$$

Coefficients A_{ij} and C in a general case are the functions of the coordinates $x_1, x_2, \ldots x_n$ of the variable point x; in partial cases, these coefficients may also be constant. It is assumed that the required function should be determined in some finite domain Ω.

For elliptical equations it is in most cases necessary to formulate the following problems differing in the type of conditions which will be regarded as homogeneous.

The Dirichlet problem, or *the first boundary-value problem:*

$$u\big|_{\partial\Omega} = 0. \tag{106}$$

The Neumann problem, or *the second boundary-value problem:*

$$\left[\sum_{i,j=1}^{n} A_{ij} \frac{\partial u}{\partial x_j} \cos(n, x_i) \right]\Bigg|_{\partial\Omega} = 0. \tag{107}$$

The third boundary-value problem

$$\left[\sum_{i,j=1}^{n} A_{ij}\frac{\partial u}{\partial x_j}\cos(n,x_i)+\sigma u\right]\Bigg|_{\partial\Omega}=0. \tag{108}$$

Here n is an external normal to surface $\partial\Omega$, σ is a function which is non-negative and differs from the actual zero and is determined on surface $\partial\Omega$.

If coefficient $C(x) \geq 0$ then in the boundary-value conditions (106) and (108) operator A is positive definite. The Dirichlet problem is reduced to the problem of the minimum of the functional:

$$J(u)=\int_{\Omega}\left(\sum_{i,j=1}^{n} A_{ij}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}+Cu^2-2fu\right)dx \tag{109}$$

(dx is the element of the volume) on the set of the functions satisfying the conditions (106). The third boundary-value problem is reduced to the problem of the minimum of a slightly different functional:

$$J(u)=\int_{\Omega}\left(\sum_{i,j=1}^{n} A_{ij}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}+Cu^2-2fu\right)dx+\int_{\partial\Omega}\sigma u^2 dS \tag{110}$$

in the class of functions in which this functional has a finite value, i.e. in class $W_2^1(\Omega)$. It is not necessary to ensure that these functions are governed by the boundary-value condition (108) because this condition is natural.

If coefficient $C(x)$ is not only non-negative but also differs from zero, the operator A is positive definite, and on the set of functions, satisfying condition (107) the Neumann problem is equivalent to the variational problem of the minimum of the integral (109) on the functions of the class $W_2^1(\Omega)$. The boundary-value condition (107) is natural.

Special attention will be given the Neumann problem for the case in which $C \equiv 0$. Equation (107) has the form

$$A_0 u=-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_j}\left(A_{ij}\frac{\partial u}{\partial x_j}\right)=f(x). \tag{111}$$

The Neumann problem for this equation in a general case is insolvable; the equality

$$(f,1)=\int_{\Omega} f(x)dx=0 \tag{112}$$

is the necessary and sufficient condition for solving this equation. On the other hand, if the Neumann problem is solvable, it has an infinite set of solutions which differ by the constant term. This term can be selected in such a manner that $(u,1) = 0$. In equation (111) it is now possible to examine the given function $f(x)$ and the required function $u(x)$ as the element of the subspace orthogonal to unity. In this subspace operator A_0 is positive definite on the set of functions satisfying condition (107). The Neumann problem

is equivalent to the problem of the minimum of the integral

$$\int_{\Omega} \sum_{i,j=1}^{n} A_{ij} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} dx,$$

$$(u,1) = \int_{\Omega} u(x) dx = 0; \tag{113}$$

on the set of the functions from $W_2^1(\Omega)$ satisfying the condition (113) this variational problem is solvable and has a unique solution.

In some cases, the boundary-value conditions of mixed type are examined: the boundary $\partial\Omega$ is divided into two parts $\partial\Omega'$ and $\partial\Omega''$ and the desired solution is governed by the conditions

$$u\big|_{\partial\Omega'} = 0, \quad \left[\sum_{i,j=1}^{n} A_{ij} \frac{\partial u}{\partial x_i} \cos(n, x_j) + \sigma u \right] \Bigg|_{\partial\Omega^n} = 0. \tag{114}$$

Operator A in equation (105) is in this case positive definite and the 'mixed' boundary-value problem is equivalent to the problem of the minimum of the functional

$$J(u) = \int_{\Omega} \left(\sum_{i,j=1}^{n} A_{ij} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} + Cu^2 - 2fu \right) dx + \int_{\partial\Omega''} \sigma u^2 dS$$

on the set of the functions satisfying the first of the conditions (114); the second of these conditions is natural.

Comment: the energy method may be also be often used if the boundary-value conditions (107), (108) of the initial problem are heterogeneous.

In conclusion, it should be mentioned that in addition to the previously examined variational formulations of problems, mathematical physics uses widely a number of other variational principles and variational methods of examination of problems (*the method of least squares, Treftz method,* and others).

3.5. Integral equations

The integral equations are those containing the unknown function under the integral sign.

3.5.1. Integral Fredholm equation of the 1st and 2nd kind

Many problems of mathematical physics are reduced to linear integral equations of the type

$$\int_{\Omega} \mathcal{K}(x,y)\varphi(y)\, dy = f(x), \tag{115}$$

$$\varphi(x) = \lambda \int_{\Omega} \mathcal{K}(x,y)\varphi(y)\, dy + f(x) \tag{116}$$

in relation to the unknown function $\varphi(x)$ in the domain $\Omega \subset \mathbb{R}^n$. Equations (115) and (116) are referred to as the *Fredholm integral equations* of the 1st and 2nd kind, respectively. The available functions $\mathcal{K}(x,y)$ and $f(x)$ are referred to respectively as the *kernel* and *free member* of the integral equations; λ is a complex parameter.

Integral equation (116) at $f = 0$

$$\varphi(x) = \lambda \int_\Omega \mathcal{K}(x,y)\varphi(y)\,dy \tag{117}$$

is referred to as a *homogeneous* Fredholm integral equation of the second kind, corresponding to equation (116). The integral Fredholm equations of the second kind

$$\psi(x) = \bar{\lambda} \int_\Omega \mathcal{K}^*(x,y)\psi(y)\,dy + g(x), \tag{118}$$

$$\psi(x) = \bar{\lambda} \int_\Omega \mathcal{K}^*(x,y)\psi(y)dy, \tag{119}$$

where $\mathcal{K}^*(x,y) = \overline{\mathcal{K}}(y,x)$ are referred to as *associate* to equations (116) and (117), respectively. The kernel $\mathcal{K}^*(x,y)$ is referred to as the Hermitian-adjoint (associate) kernel for kernel $\mathcal{K}^*(x,y)$.

The integral equations (116), (117), (118) and (119) can be presented in the operator form:

$$\varphi = \lambda K\varphi + f, \quad \varphi = \lambda K\varphi, \quad \psi = \bar{\lambda} K^*\psi + g, \quad \psi = \bar{\lambda} K^*\psi,$$

where the integral operators K and K^* are determined by the kernel $\mathcal{K}(x,y)$ and $\mathcal{K}^*(x,y)$ respectively

$$(Kf)(x) = \int_\Omega \mathcal{K}(x,y) f(y)dy, \quad (K^* f)(x) = \int_\Omega \mathcal{K}^*(x,y) f(y)dy.$$

Consequently, the value of λ at which the homogeneous integral equation (117) has non-zero solutions from $L_2(\Omega)$ is referred to as the *characteristic number* of the kernel $\mathcal{K}(x,y)$ and the appropriate solution are the *eigenfunctions* of this kernel corresponding to this characteristic number. Thus, the characteristic numbers of the kernel $\mathcal{K}(x,y)$ and the eigenvalue of the operator K are mutually convertible and their eigenfunctions coincide.

3.5.2. Volterra integral equations

Let $n = 1$, domain Ω is the interval $(0,a)$ and kernel $\mathcal{K}(x,y)$ converts to zero in a triangle $0 < x < y < a$. This kernel is referred to as the *Volterra kernel*. The integral equations (115) and (116) with the *Volterra kernel* have the form

$$\int_0^x \mathcal{K}(x,y)\varphi(y)dy = f(x), \quad \varphi(x) = \lambda \int_0^x \mathcal{K}(x,y)\varphi(y)dy + f(x)$$

and are referred to as the *Volterra integral equations* of the 1st and 2nd kind, respectively.

The integral Volterra equations of the 1st kind can be reduced by differentiation to equations of the 2nd kind

$$\mathcal{K}(x,x)\varphi(x)+\int_0^x \frac{\partial \mathcal{K}(x,y)}{\partial x}\varphi(y)\,dy = f'(x),$$

where $\mathcal{K}(x,y)$ and $\mathcal{K}_x(x,y)$ are continuous for $0\le x\le y\le a$, $\mathcal{K}(x,x)\ne 0$, $x\in[0,a]$, $f\in C^1([0,a])$ and $f(0) = 0$.

3.5.3. Integral equations with a polar kernel

The kernel

$$\mathcal{K}(x,y)=\frac{\mathcal{H}(x,y)}{|x,y|^{\alpha}},\quad \alpha<n,$$

where $\mathcal{H}(x,y)\in C(\bar{\Omega}\times\bar{\Omega})$ is referred to as *the polar kernel*; if $\alpha < n/2$, then $\mathcal{K}(x,y)$ is a *weakly polar kernel*.

It is well known that:

1) *to ensure that kernel $\mathcal{K}(x,y)$ is polar, it is necessary and sufficient that this kernel is continuous at $x \ne y$, $x\in\bar{\Omega}$, $y\in\bar{\Omega}$ and satisfies the estimate*

$$|\mathcal{K}(x,y)|\le\frac{A}{|x-y|^{\alpha}},\quad \alpha<n,\quad x\in\Omega,\quad y\in\Omega;$$

2) *If $\mathcal{K}_i(x,y)$ are polar kernels*

$$|\mathcal{K}(x,y)|\le\frac{A_i}{|x,y|^{\alpha_i}},\quad \alpha_i<n,\quad i=1,2,$$

and domain Ω is bounded, then the kernel

$$\mathcal{K}_3(x,y)=\int_{\Omega}\mathcal{K}_2(x,y')\mathcal{K}_1(y',y)dy'$$

is also polar, and

$$|\mathcal{K}_3(x,y)|\le\frac{A_3}{|x-y|^{\alpha_1+\alpha_2-n}},\ \text{if}\quad \alpha_1+\alpha_2>n_1$$

$$|\mathcal{K}_3(x,y)|\le A_4\left|\ln|x-y|\right|+A_5,\ \text{if}\quad \alpha_1+\alpha_2=n;$$

$\mathcal{K}_3(x,y)$ *is continuous on* $\bar{\Omega}\times\bar{\Omega}$, if $\alpha_1+\alpha_2<n$.

3.5.4. Fredholm theorem

The basis of the theory of Fredholm integral equations

$$\varphi=\lambda K\varphi+f \tag{120}$$

with the continuous kernel $\mathcal{K}(x,y)$ and the associated equation

$$\psi=\lambda K^{*}\psi+g \tag{121}$$

is represented by the *Fredholm theorems* referred to as the *Fredholm alternative*.

Fredholm Alternative. *If the integral equation* (120) *with a continuous kernel is solvable in $C(\bar{\Omega})$ for any free term $f\in C(\bar{\Omega})$, then the associate*

equation (121) *is solvable in* $C(\bar{\Omega})$ *for any free member of* $g \in C(\bar{\Omega})$, *and these solutions are unique* (*the first Fredholm theorem*).

If the integral equation (120) *is solvable in* $C(\bar{\Omega})$ *but not at every free term f, then*:

1) *homogeneous equations* (120) *and* (121) *have a* (*finite*) *number of linearly independent solutions* (*the second Fredholm theorem*);

2) *To solve the equation* (120) *it is necessary and sufficient that the free term f is orthogonal to all solutions of the associate homogeneous equation* (121) (*third Fredholm theorem*);

3) *in every circle* $|\lambda| \le R$ *there may be only a finite number of characteristic numbers of the kernel* $\mathcal{K}(x,y)$ (*fourth Fredholm theorem*).

We re-formulate the Fredholm alternative in terms of characteristic numbers and eigenfunctions.

If $\lambda \ne \lambda_k$, $k = 1,2,\ldots$, *then the integral equations* (120) *and* (121) *are uniquely solvable at any free term.*

If $\lambda = \lambda_k$, *then homogeneous equations*

$$K\varphi=\lambda_k\varphi \quad u \quad K^*\psi=\bar{\lambda}_k\psi$$

have one a (*finite*) *number* $r_k \ge 1$ *of linearly independent solutions – eigenfunctions* $\varphi_k, \varphi_{k+1},\ldots,\varphi_{k+r_k-1}$ of the *kernel* $\mathcal{K}(x,y)$ and eigenfunctions $\Psi_k,\Psi_{k+1},\ldots,\Psi_{k+r_k-1}$ *of kernel* $\mathcal{K}^*(x,y)$, corresponding to the characteristic numbers λ_k and $\bar{\lambda}_k$ (r_k is the multiplicity of λ_k and $\bar{\lambda}_k$).

If $\lambda = \lambda_k$ *then to solve the equation* (120) *it is necessary and sufficient that*

$$(f,\Psi_{k+1})=0, \quad i=0,1,\ldots,r_k-1.$$

Comment: The Fredholm theorems are also transferred to integral equations with a polar kernel, and all eigenfunctions of the polar kernel $\mathcal{K}(x,y)$, belonging to $L_2(\Omega)$ belong to $C(\bar{\Omega})$.

3.5.5. Integral equation with the Hermitian kernel

Kernel $\mathcal{K}(x,y)$ is referred to as Hermitian, if it coincides with its Hermitian adjoint kernel $\mathcal{K}(x,y) = \mathcal{K}^*(x,y)$. The corresponding integral equation

$$\varphi(x)=\lambda\int_{\Omega}K(x,y)\varphi(y)dy+f(x) \tag{122}$$

at real λ coincide with its associate equation because $K = K^*$. This equation can be examined conventionally in the space $L_2(\Omega)$.

Let K be the integral operator with the Hermitian continuous kernel $\mathcal{K}(x,y)$. This operator transfers $L_2(\Omega)$ (Ω is the bounded domain) to $L_2(\Omega)$ and is Hermitian:

$$(K f, g)=(f, Kg), \; f,g\in L_2(\Omega). \tag{123}$$

Conversely, if the integral operator K with the continuous kernel $\mathcal{K}(x,y)$ *is Hermitian, then this kernel is also Hermitian.*

Theorem 29. *Any Hermitian continuous kernel* $\mathcal{K}(x,y) \neq 0$ *has at least one characteristic number, and the smallest characteristic number in respect of modulus* λ_1 *satisfies the variational principle*

$$\frac{1}{|\lambda_1|} = \sup_{f \in L_2(\Omega)} \frac{\|Kf\|}{\|f\|}. \tag{124}$$

Theorem 30. *The set of the characteristic numbers* $\{\lambda_k\}$ *is not empty, located on the real axis, does not have finite limiting points; every characteristic number has a finite multiplicity, the system of eigenfunctions* $\{\varphi_k\}$ *may be selected as orthogonal*

$$(\varphi_k, \varphi_i) = \delta_{ki}. \tag{125}$$

If $\lambda \neq \lambda_k$, $k = 1,2,\ldots$, *then equation (122) is uniquely solvable at any free term* $f \in C(\overline{\Omega})$. *If* $\lambda = \lambda_k$, *then to solve equation (122) it is necessary and sufficient that*

$$(f, \varphi_{k+i}) = 0, \quad i = 0,1,\ldots,r_k - 1, \tag{126}$$

where $\varphi_k, \varphi_{k+1}, \ldots, \varphi_{k+rk-1}$ *are eigenfunctions corresponding to the characteristic number* λ_k, *and* r_k *is the multiplicity of* λ_k.

Comment: All the previously formulated claims for the integral equations with the Hermitian continuous kernel remain valid also for the integral equations with the Hermitian polar kernel.

Let it be that λ_1, λ_2, ... are the characteristics numbers of the Hermitian continuous kernel $\mathcal{K}(x,y) \neq 0$ distributed in the order of increase in the modulus, $|\lambda_1| \leq |\lambda_2| \leq$ and $\varphi_1, \varphi_2, \ldots$ are the corresponding orthonormal eigenfunctions $(\varphi_k, \varphi_i) = \delta_{ki}$.

It is said that function $f(x)$ is represented by means of the kernel $\mathcal{K}(x,y)$ if there is a function $h \in L_2(\Omega)$ such

$$f(x) = \int_\Omega K(x,y)h(y)dy, \quad x \in \Omega. \tag{127}$$

Theorem 31 (*Hilbert–Schmidt Theorem*). *If the function* $f(x)$ *is represented by means of the Hermitian continuous kernel* $\mathcal{K}(x,y)$, $f = Kh$, *then its Fourier series in respect of eigenfunctions of the kernel* $\mathcal{K}(x,y)$ *converges regularly (and, therefore, uniformly) on* $\overline{\Omega}$ *to this function*

$$f(x) = \sum_{k=1}^{\infty} \frac{(h, \varphi_k)}{\lambda_k} \varphi_k(x). \tag{128}$$

We examine a heterogeneous integral equation

$$\varphi = \lambda K\varphi + f \tag{129}$$

with a Hermitian continuous kernel $\mathcal{K}(x,y)$.

The following claim is made on the basis of the Hilbert–Schmidt theorem: *if* $\lambda \neq \lambda_k$, $k = 1,2,\ldots,$ and $f \in C(\bar{\Omega})$, *then* (*the unique*) *solution* φ *of the integral equation* (129) *is presented in the form of a series uniformly converging on* $\bar{\Omega}$ (*the Schmidt equation*):

$$\varphi(x)=\lambda\sum_{k=1}^{\infty}\frac{(f,\varphi_k)}{\lambda_k-\lambda}\varphi_k(x)+f(x). \tag{130}$$

Equation (130) remains also valid at $\lambda = \lambda_j$ if in accordance with the third Fredholm theorem

$$(f, \varphi_{i+j}) = 0, \quad i = 0,1,\ldots, r_j-1.$$

In this case, the solution of equation (129) is not unique, and its generalized solution is given by the formula

$$\varphi(x)=\lambda_j\sum_{\substack{k=1\\ \lambda_k\neq\lambda_j}}^{\infty}\frac{(f,\varphi_k)}{\lambda_k-\lambda_j}\varphi_k(x)+f(x)+\sum_{i=0}^{r_j-1}c_i\varphi_{j+i}(x) \tag{131}$$

where c_i are arbitrary constants.

Many problems of mathematical physics are reduced to integrally equations with a real Hermitian kernel. These kernels are referred to as *symmetric*; they satisfy the relationship $\mathcal{K}(x,y) = \mathcal{K}(y,x)$. For these integral equations, the claims formulated previously for the equations with the Hermitian kernel are valid. However, there is also a number of specific results. For example, in particular, the *eigenfunctions of the symmetric kernel* $\mathcal{K}(x,y)$ *may be selected as real.*

Comment. The Hilbert–Schmidt theorem is also extended to the integral equations with the Hermitian weakly polar kernel

$$K(x,y)=\frac{H(x,y)}{|x,y|^{\alpha}},\quad \alpha<\frac{n}{2},\quad H^*(x,y)=H(x,y).$$

BIBLIOGRAPHIC COMMENTARY

The main sections of mathematical physics have been described in [13] where the concept of the generalized solution is used widely. The review of the results of the classic theory of linear differential equations with partial derivatives and also a brief description of the functions were published in [25]. The main problems of mathematical physics and methods for solving

them (including the Fourier method) are discussed in [91]. In [110] the authors describe the fundamentals of the theory of problems for eigenvalues, the special functions and the method of eigenfunctions for problems of mathematical physics; the theory of the Fourier series is substantiated.

Special attention to variational formulation of the problems and the energetic method is given in [70,71] where the elements of variational calculations are also presented. Generalized formulations of the problems of mathematical physics have been published in [69] which deals with functional spaces, the Sobolev embedding theorems, the fundamentals of the boundary-value problems for equations in partial derivatives and problems of eigenumbers. A classic work in the theory of the Sobolev spaces and applications to the problems of mathematical physics is the study [84]. Main sections of the current theory of functions, functional spaces and embedding theorems were published [75]. The authors of [95] describe the theory of embedding of spaces of differentiable functions and applications to differential equations. The results are presented for the theory of traces for non-isotropic classes of the functions and the solvability of the mixed boundary-value problems for the equations not resolved with respect to the older derivative.

Chapter 2

METHODS OF POTENTIAL THEORY

Keywords: potential, volume potential, Newton potential, simple layer potential, double layer potential, logarithmic potential, Fredholm equation, Schwarz method, cylindrical coordinates, spherical coordinates, Dirichlet problem, Neumann problem, Green function, sweep method, Helmholtz equations, retardation potential, heat equation, telegraph equation,

MAIN CONCEPTS AND DESIGNATIONS

Vector field – vector-function, given at every point of the given domain.

Scalar field – function given at every point of the examined domain.

The fundamental solution of differential operator L – generalized function, satisfying the equation $Lu = \delta(x)$, where $\delta(x)$ is the Dirac delta function.

The potential – scalar function represented in the integral of the product of some function (potential density) and the fundamental solution or its derivative.

Laplace equation – equation $\Delta u = 0$, where Δ is the Laplace operator.

Poisson equation – equation $-\Delta u = f$.

Helmholtz equation – equation $\Delta u + \chi^2 u = 0$.

Newton (volume) potential – integral $u(A) = \iiint_V \frac{\rho(P)}{r}\, dV$, where r is the distance between the fixed point A and the variable point P, ρ is the density of the potential $V \subset \mathrm{R}^3$.

The simple layer potential – integral $u(A) = \iint_S \frac{\rho(P)}{r}\, dS$, $S = \partial V$, $V \subset \mathrm{R}^3$.

The double layer potential – integral $u(A) = \iint_S \frac{\rho(P)\cos\theta_{PA}}{r^2}\, dS =$

$= \iint_S \rho(P) \frac{\partial}{\partial n}\left(\frac{1}{r}\right) dS$, where θ is the angle between the normal $\mathbf{n}$ to the surface $S = \partial V$ at point $P \in S$ and the direction PA.

The logarithmic potential – integral $u(A) = \iint_S \rho(P) \ln\left(\frac{1}{r}\right) dS,\ S \in \mathrm{R}^2$.

The logarithmic simple layer potential – integral

$$u(A) = \int_L \rho(P) \ln\left(\frac{1}{r}\right) dl,\ L = \partial S,\ S \in \mathrm{R}^2.$$

The logarithmic double layer potential – integral

$$u(A) = \int_L \rho(P) \frac{\partial}{\partial n} \ln\left(\frac{1}{r}\right) dl = \int_L \frac{\rho(P)\cos\theta}{r} dl,$$

where θ is the angle between the normal **n** to the contour $L = \partial S$ at point $P \in L$ and the direction PA.

The integal Green formula – representation of the twice differentiable function u in the form of the sum of three potentials: volume with density Δu, the simple layer potential with surface density $\partial u/\partial n$ and the double layer potential with density u.

The harmonic function – the twice differentiable function, satisfying the Laplace equation.

1. INTRODUCTION

The concept of the Newton potential was introduced for the first time at the end of the 18th century by R. Laplace and G. Lagrange and, at a later stage, by L. Euler for hydrodynamic problems. Examination of the concept of the potential as a function whose gradient is equal to the vector field, was carried out by Gauss. The properties of the simple layer were examined for the first time by Coulomb and Poisson, and Green provided a significant contribution to the development of potential theory. At present, potential theory is the actively developed method of examination and solution of the problems in different areas of mathematical physics.

Let us consider the vector field $\mathbf{F} = \sum_{i=1}^{3} F_i \mathbf{e}_i$, where $F_i = F_i(x,y,z)$ are the coordinates of vector **F** applied at point (x,y,z), $\mathbf{e}_i$ are the directing unit vectors of the orthogonal system of the coordinates; let $u(x,y,z)$ be the scalar function (scalar field). The *potential of the vector field* **F** is the scalar field $u(x,y,z)$ whose potential is equal to **F**: $\nabla u = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z}\right) = \mathbf{F}$. Therefore, if the potential function (potential) is known, the acting forces can be calculated. In many problems of electromagnetism, hydrodynamics and acoustics, heat conductivity and diffusion, it is necessary to solve boundary value problems for elliptical equations whose simplest and most important representatives are the Laplace function $\Delta u = 0$ and Poisson equation – $\Delta u = f$.

A key role in the methods of potential theory is played by the fundamental solutions of the Laplace equation, equal to $1/(4\pi r)$ in the three-dimensional and $(1/(2\pi))\ln(1/r)$ in two-dimensional cases.

These solutions are used to construct potentials presented in the form

of the integral of some function (potential density) and the fundamental solution (or its derivative). Depending on the integration domain and the application of the fundamental solution or its normal derivative, there are volume potentials, the simple layer and double layer potentials. To find the potential (solution of the appropriate elliptical equation) in the form of the integral of density, we obtain an integral equation of unknown density, and since the solution can be sought in the form of different potentials, it is necessary to select a potential at which the resulting integral equation is the simplest. For example, to obtain the Fredholm equation of the second kind, the Dirichlet problem is solved using the double layer potential, and the Neumann problem – the simple layer potential. Below, we examine potentials for Laplace, Helmholtz, wave and heat equations – the main equations of mathematical physics arising in different applied problems [5,13,20,47,49,83,85–87, 91].

2. FUNDAMENTALS OF POTENTIAL THEORY

2.1. Additional information from mathematical analysis

2.1.1 Main orthogonal coordinates

Let us assume that we have a system of three (one-to-one)) functions from three variable each

$$\begin{aligned} x_1 &= \varphi_1(u_1,u_2,u_3), \\ x_2 &= \varphi_2(u_1,u_2,u_3), \\ x_3 &= \varphi_3(u_1,u_2,u_3). \end{aligned} \tag{1}$$

If it is assumed that every system of the values u_1, u_2, u_3 is linked with a specific point M in the space with the Cartesian coordinates x_1,x_2,x_3, the numbers u_1,u_2, u_3, may be regarded as curvilinear coordinates of point M. The system of coordinates determined by them is referred to as *curvilinear*. The system of coordinates is referred to as *orthogonal* if at every point coordinate lines, passing through this point, intersect at right angles.

We examine two main examples of curvilinear orthogonal coordinates.

1°. *Cylindrical coordinates*:

$$x = r\cos\varphi, \quad y = r\sin\varphi, \quad z = z \quad (\varphi\in[0,2\pi], \ r > 0)$$

(here instead of x_1,x_2,x_3, we write x,y,z and instead of $(u_1,u_2,u_3 - r, \varphi, z)$. In a two-dimensional case with the cylindrical coordinates not depending on z, the latter are referred to as polar.

2°. *Spherical coordinates*:

$$x = r\sin\theta\cos\varphi, \quad y = r\sin\theta\sin\varphi, \quad z = r\cos\theta$$
$$(\theta\in[0,\pi], \ \varphi\in[0,2\pi], \ r > 0).$$

2.1.2. Main differential operations of the vector field

Let $\varphi = \varphi(u_1,u_2,u_3)$ be the scalar field, $\mathbf{F} = \mathbf{F}(u_1,u_2,u_3)$ be the vector field, $\mathbf{F} = \sum_{i=1}^{3} F_i \mathrm{e}_i$. In the Cartesian orthogonal coordinates the following operations are detemined:

Gradient

$$\text{grad}\varphi=\nabla\varphi=\sum_{i=1}^{3}\partial_i\varphi \mathrm{e}_i;$$

Divergence

$$\text{div}\,\mathbf{F}=(\nabla,\mathbf{F})_3=\sum_{i=1}^{3}\partial_i F_i;$$

Rotor (vortex)

$$\text{rot}\,\mathbf{F}=[\nabla,F]=\begin{vmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3\\ \partial_1 & \partial_2 & \partial_3\\ F_1 & F_2 & F_3\end{vmatrix};$$

Laplace operator (Laplacian)

$$\Delta\varphi=\text{div grad}\,\varphi=\sum_{i=1}^{3}\partial_i^2\varphi;$$

and to simplify considerations, we introduce the notation:

$$\partial_i=\frac{\partial}{\partial u_i},\ \partial_i^2=\frac{\partial^2}{\partial u_i^2},\ \nabla=(\partial_1,\partial_2,\partial_3).$$

The Laplace operator in the *cylindrical coordinates* has the form

$$\Delta v=\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 v}{\partial\varphi^2}+\frac{\partial^2 v}{\partial z^2}. \tag{2}$$

In the *spherical coordinates*

$$\Delta v=\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial v}{\partial r}\right)+\frac{1}{r^2\sin\theta}\left(\sin\theta\frac{\partial v}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 v}{\partial\varphi^2}. \tag{3}$$

2.1.3. Formulae from the field theory

Let u and v be two arbitrary functions having continuous partial derivatives to the second order, inclusively. Instead of writing $u = u(x,y,z)$, we write $u = u(A)$, where point A has the coordinates (x,y,z). The distance between the point $A(x,y,z)$ and point P (ξ,η,ζ) is equal to:

$$r_{AP}=\sqrt{(x-\xi)^2+(y-\eta)^2+(z-\zeta)^2}.$$

The symbols of the differential operators of A and P are fitted with indices A or P depending on whether differentiation is carried out with respect to x, y, z or ξ,η,ζ; for example:

$$\Delta_A u=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2},\qquad \text{grad}_p u=\frac{\partial u}{\partial\xi}\mathrm{i}+\frac{\partial u}{\partial\eta}\mathrm{j}+\frac{\partial u}{\partial\zeta}\mathrm{k}.$$

The symbol $\left(\frac{\partial u}{\partial n}\right)_P$ denotes the derivative in the direction of the normal to the surface $\mathbf{n}$ passing through P:

$$\left(\frac{\partial u}{\partial n}\right)_p=\frac{\partial u}{\partial\xi}\cos\alpha+\frac{\partial u}{\partial\eta}\cos\beta+\frac{\partial u}{\partial\zeta}\cos\gamma,$$

where $\cos\alpha$, $\cos\beta$, $\cos\gamma$, are the direction cosines of the normal $\mathbf{n}$.

We write the *Ostrogradskii–Gauss formula*:

$$\iiint_V \left(\frac{\partial P}{\partial x}+\frac{\partial Q}{\partial y}+\frac{\partial R}{\partial z}\right)dV = \iint_S \left(P\cos\alpha + Q\cos\beta + R\cos\gamma\right)dS,$$

where the cosines are the direction cosines of the outer normal **n**. Assuming here that $P = u\partial_1 v$, $Q = u\partial_2 v$, $R = u\partial_3 v$, we obtain the *first Green formula*

$$\iiint_V (\mathrm{grad}u, \mathrm{grad}v)\, dV + \iiint_V u\Delta v dV = \iint_S u\frac{\partial v}{\partial n}dS. \tag{4}$$

We change the positions of u and p in equation (4) and the resulting formula is subtracted from (4). Consequently, we obtain the *second Green formula*

$$\iiint_V \{u\Delta v - v\Delta u\}dV = \iint_S \left\{u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right\}dS. \tag{5}$$

If $A\in V$ then $v = 1/r_{AP}$ cannot be substituted directly into (5). Surrounnding the point A with a sphere with a small radius, using the second Green formula (5) for the functions u and v outside the sphere, and tending the radius of the auxiliary sphere to zero, we obtain *the integral Green formula*

$$\Omega \cdot u(A) = \iint_S \left\{\frac{1}{r_{AP}}\frac{\partial u}{\partial n} - u(P)\frac{\partial(1/r_{AP})}{\partial n}\right\}dSp - \iiint_V \frac{\Delta u(P)}{r_{AP}}dVp. \tag{6}$$

Depending on the position of point A, coefficient Ω has the values $\Omega = 4\pi$, $A\in V$, $\Omega = 2\pi$, $A\in\partial V$, $\Omega = 0$, $A\notin V$. Similarly, we use D in the two-dimensional case to denote some domain of the plane (x,y) bounded by the smooth closed curve L (or by several curves). Consequently, for the arbitrary functions u and v, having continuous partial derivatives to the second order inclusively, we can write the following expressions

$$\iint_D \left\{\frac{\partial u}{\partial \xi}\frac{\partial v}{\partial \xi} + \frac{\partial u}{\partial \eta}\frac{\partial v}{\partial \eta}\right\}dS + \iint_D u\Delta v dS = \oint_L u\frac{\partial v}{\partial n}dl, \tag{7}$$

$$\iint_D \{u\Delta v - v\Delta u\}\, dS = \oint_L \left\{u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right\}dl, \tag{8}$$

$$u(A) = \frac{1}{2\pi}\oint_L \left\{1\mathrm{n}\frac{1}{r}\frac{\partial u}{\partial n} - u(P)\frac{\partial\left(1\mathrm{n}\frac{1}{r}\right)}{\partial n}\right\}dl - \frac{1}{2\pi}\iint_D \Delta u 1\mathrm{n}\frac{1}{r}dS, \tag{9}$$

where $\dfrac{\partial}{\partial n}$ is the operator of differentiation in the direction of the outer normal to L, $\Delta = \dfrac{\partial^2}{\partial\xi^2} + \dfrac{\partial^2}{\partial\eta^2}$, $r = r_{AP}$ is the distance between point A and variable point P.

2.1.4. Main properties of harmonic functions

The functions, harmonic in domain V, are the functions satisfying the Laplace

equation in this domain

$$\Delta u = 0.$$

The following properties of the harmonic function U are valid

1°.

$$\iint_S \frac{\partial U}{\partial n} dS = 0,$$

i.e. *the integral of the normal derivative of the harmonic function on the surface of the domain is equal to zero.*

2°. *The value of the harmonic function at any point inside the domain is expressed through the values of this function and its normal derivative on the surface of the domain by the formula:*

$$U(A) = \frac{1}{4\pi} \iint_S \left[\frac{1}{r} \frac{\partial U}{\partial n} - U \frac{\partial (1/r)}{\partial n} \right] dS.$$

3°. *The value of the harmonic function in the centre A of the sphere S_R with radius R is equal to the mean arithmetic value of the function on the surface of sphere:*

$$U(A) = \frac{1}{4\pi R^2} \iint_{S_R} U \, dS.$$

4°. From 3° we obtain the *maximum principle: a function, harmonic inside the domain and continuous up to its boundary, reaches its highest and smallest values at the boundary of the domain.*

2.2 Potential of volume masses or charges

2.2.1. Newton (Coulomb) potential

Let V be a bounded domain of the space bounded by a piecewise smooth closed surface S. Let function $\rho(P)$ be given in V. This function is assumed to be continuous and bounded in V. Then

$$u(A) = \iiint_V \frac{\rho(p)}{r} dV \tag{10}$$

is *the potential of infinite masses, or Newton potential of masses distributed over the volume V with density* ρ. Function $u(A)$ can also be interpreted as a *Coulomb potential of volume-distributed charges.*

2.2.2. The properties of the Newton potential

At all points A outside V, function $u(A)$ from (10) is continuous and differentiable any number of times with respect to x, y, z. In particular

$$grad\, u(A) = \iiint_V \rho(p) \operatorname{grad}\left(\frac{1}{r}\right) dV = -\iiint_V \rho(p) \left(\frac{\mathbf{r}}{r^3}\right) dV, \tag{11}$$

where $\mathbf{r}$ is the radius-vector, $\mathbf{r} = \mathbf{r}_{AP} = (x-\xi)\mathbf{i} + (y-\eta)\mathbf{j} + (z-\zeta)\mathbf{k}$, $A = A(x,y,z)$, $P = P(\xi,\eta,\zeta)$. Since $\Delta(1/r_{AP}) = 0$, $A \notin V$, $P \in V$, then

$$\Delta u(A) = \iiint_V \rho(P)\Delta\left(\frac{1}{r_{AP}}\right)dV = 0, \quad A \notin V.$$

Thus, *potential* $u(A)$ *of masses or charges, distributed in volume* V, *satisfies the Laplace equation at all point outside* V.

At a very large distance from the origin or, which is the same, from domain V, the following approximate equality is valid

$$u(A) \approx \frac{1}{r}\iiint_V \rho(P)dV = \frac{M}{r}, \tag{12}$$

where $M = \iiint \rho dV$ is the total mass of the body. In other words, at infinity the potential of the volume-distributed masses (or charges) behaves as a potential of a material point (or point charge) located at the origin of the coordinates, and the mass (or charge) concentrated there is equal to the entire mass (or charge) distributed in volume V. In particular, $u(\mathrm{A}) \to 0$ at $r \to \infty$. Following estimates are obtained for partial derivatives of potential of the volume-distributed masses:

$$\left|\frac{\partial u}{\partial x}\right| < \frac{C}{r^2}, \quad \left|\frac{\partial u}{\partial y}\right| < \frac{C}{r^2}, \quad \left|\frac{\partial u}{\partial z}\right| < \frac{C}{r^2}, \tag{13}$$

where C is some constant.

2.2.3. Potential of a homogeneous sphere

Let V be a sphere with radius R with the centre at the origin of the coordinates with constant density ρ = const. Transferring to spherical coordinates, r, φ, θ, where $\xi = r\sin\theta\cos\varphi$, $\eta = r\sin\theta\sin\varphi$, $\zeta = r\cos\theta$, we obtain the potential of the homogeneous sphere at the point r

$$u(r) = \begin{cases} \dfrac{M}{r} & \text{if} \quad r > R, \\ \dfrac{M}{2R}\left\{3 - \left(\dfrac{r}{R}\right)^2\right\}, & \text{if} \quad r < R. \end{cases}$$

It may easily be seen that $u(r)$ and its first derivative $u'(r)$ are continuous for all $r \geq 0$, but the second derivative $u''(r)$ shows a discontinuity at point $r = R$.

At all external points the potential of the homogeneous sphere is equal to the potential of the material point of the same mass placed in its centre, and satisfies the Laplace equation. At all internal points of the sphere, the potential satifies the Poisson equation $\Delta u = -4\pi\rho$.

2.2.4. Properties of the potential of volume-distributed masses

The potential of finite bodies of arbitrary form and variable bounded density have the following two characteristic properties.

1°. u and grad u are continuous in the entire space

2°. $u(A) \to 0$ at $r = \sqrt{x^2 + y^2 + z^2} \to \infty$ and $r^2\left|\mathrm{grad}\, u\right| < C$.

Conversely, if some function $u(A)$ has these two properties, then there exists the volume V such that

$$u(A) = \iiint_V \frac{\rho(P)}{r} dV,$$

i.e. it is the Newton potential of the masses. Here $\rho = -(1/4\pi))\Delta u$.

2.3. Logarithmic potential

2.3.1. Definition of the logarithmic potential

Let V be some finite domain of the plane Oxy, bounded by the piecewise smooth closed curve L. Let the continuous function $\rho(P)$ be the continuous density of the domain D. Consequently,

$$u(A) = \iint_D \rho(P) \ln \frac{1}{r} dS \tag{14}$$

is *the logarithmic potential of the domain with density* ρ. Potential $u(A)$ has a property that its gradient is approximately equal to the force of Newton (or Coulomb) attraction at point A caused by a cylinder with density $(1/2)\rho$ (constant along every straight line parallel to axis z). Function $\ln(1/r)$ is a fundamental solution of the two-dimensional Laplace equation.

2.3.2. The properties of the logarithmic potential

At all points A of the plane not belonging to D, $u(A)$ is a continuous function of A, differentiable any times with respect to x and y under the sign of the integral. In particular,

$$\mathrm{grad}_A u = \iint_D \rho(P) \mathrm{grad}_A \ln \frac{1}{r_{AP}} dS = -\iint_D \rho(P) \frac{\mathbf{r}}{r^2} dS, \tag{15}$$

where $\mathbf{r} = (x-\xi)\mathbf{i} + (y-\eta)\mathbf{j}$. Since

$$\Delta_A u = \iint_D \rho(P) \Delta_A \left(\ln \frac{1}{r_{AP}} \right) dS,$$

the logarithmic potential of the domain satisfies the Laplace equation at all points outside this domain. It should be mentioned that the Newton potential (section 2.2.2) has the same property.

The logarithmic potential of the domain is represented in the form of the sum of the logarithmic potential of the point located at the origin of the coordinates with the mass equal to the mass of the entire domain, and of some function which behaves at infinity as the potential of volume-distributed masses, i.e.,

$$u(A) = M \ln(1/r) + u^*(A) \quad \text{where} \quad u^*(A) \to 0 \quad \text{at} \quad r \to \infty,$$

and the inequality $r^2 |\mathrm{grad}_A u^*| < C$ holds, where C is some constant. In particular, if point A moves to infinity, the absolute value of the logarithmic potential of the domain increases as $\ln r$.

The logarithmic potential of the domain and its first order partial derivatives are continuous on the entire plane and equation (15) also holds for points A belonging to D.

If the first derivatives of the function ρ are continuous, then the logarithmic potential inside domain D satisfies the two-dimensional Poisson equation

$$\Delta u = -2\pi\rho$$

and outside the domain the Laplace equation.

The logarithmic potentials of the domain satisfy the following three properties.

1°. $u(A)$ and grad u are continuous on the entire plane.

2°. Δu does exist and is equal to zero at some finite domain D bounded by the piecewise smooth curve L, and inside D the Laplacian Δu is continuous and has continuous derivatives of the first order.

We denote $\rho(A) = -\Delta u/(4\pi)$

3°. If

$$M = \iint_D \rho(P)dS$$

and $u^*(A) = u(A) - M\ln(1/r)$, where $r = \sqrt{x^2 + y^2}$ then $u^*(A) \to 0$ at $r \to \infty$, and $r^2|\mathrm{grad}_A u^*| < C$, where C is some constant.

Conversely, if some function $u(A)$ has the property 1°–3°, then it is the logarithmic potential of some domain D with density $\rho(P)$, i.e.

$$u(A) = \iint_D \rho(P)\ \ln\frac{1}{r}dS.$$

2.3.3. The logarithmic potential of a circle with constant density

We examine a circle K_R with a radius R: $\xi^2 + \eta^2 \le R^2$ and set ρ = const. Consequently, outside K_R

$$u(A) = \pi R^2\rho\,\ln\frac{1}{r} = M\,\ln\frac{1}{r},$$

where $M = \pi R^2\rho$ is the total mass of the circle K_R.

Inside K_R

$$u(A) = M\left\{\frac{1}{2} - \ln R - \frac{1}{2}\left(\frac{r}{R}\right)^2\right\}.$$

Thus, the logarithmic potential outside the circle is equal to the logarithmic potential of the point placed in the centre of the circle with the mass equal to the mass of the entire circle. This property coincides with the corresponding result for the volume potential outside the sphere (section 2.2.3).

2.4. The simple layer potential

2.4.1. Definition of the simple layer potential in space

Let V be the bounded domain of the three-dimensional space, $\rho(P)$ be a continuous function of the point in this domain, and r is the distance from point A to variable point $P \in V$. The potential of the volume mass is determined, as is well-known (section 2.9.1.), by the equation

$$u(A) = \iiint_V \frac{\rho(P)}{r}dV.$$

The simple layer potential, distributed on the surface S with density $\rho(P)$, is determined by the equation

$$u(A) = \iint_S \frac{\rho(P)}{r} dS, \tag{16}$$

where S is the bounded smooth surface on which the continuous bounded function $\rho(P)$ is given. Usually, the 'tortuosity' (fractal nature) of surface S is subject to additional restrictions; these surfaces are referred to as *Lyapunov surfaces*.

Potential (14) is the *Newton potential of the mass* (or the *Coulomb potential of the charges*) distributed on S with surface density ρ. Potential $u(A)$ is referred to as the *simple layer potential*, and surface S as the *carrier surface of the layer*.

If the carrier surface is not closed, it is assumed that it is bounded by a piecewise smooth curve.

2.4.2. The properties of the simple layer potential

The simple layer potential corresponds to the Laplace equation at all points of the space not located on the carrier surface of the layer.

At infinity, the simple layer potential behaves as the potential of the material point located at the origin of the coordinates, and the mass concentrated there is equal to the entire mass distributed on S. For partial derivatives of the first order of the simple layer potential, we can write the following inequalities, similar to those for the potential of the volume-distributed masses

$$\left|\frac{\partial u}{\partial x}\right| < \frac{C}{r^2}, \quad \left|\frac{\partial u}{\partial y}\right| < \frac{C}{r^2}, \quad \left|\frac{\partial u}{\partial z}\right| < \frac{C}{r^2}.$$

The simple layer potential is continuous in the entire space. The normal derivative of the simple layer potential shows a discontinuity on intersecting the layer, and the magnitude of the jump on intersecting the layer at point A in the direction of differentiation is

$$\left(\frac{\partial u}{\partial n}\right)_{A_+} - \left(\frac{\partial u}{\partial n}\right)_{A_-} = -4\pi\rho(A). \tag{17}$$

The value of the normal derivative at point A is:

$$\left(\frac{\partial u}{\partial n}\right)_{A} = \frac{1}{2}\left\{\left(\frac{\partial u}{\partial n}\right)_{A_-} + \left(\frac{\partial u}{\partial n}\right)_{A_+}\right\}.$$

It should be mentioned that at the point of the layer in which the density is equal to zero, the normal derivative of the simple layer potential is continuous. Equality (17) maybe made more accurate as follows:

$$\left(\frac{\partial u}{\partial n}\right)_{A_+} = \iint_S \rho(P)\frac{\cos(\mathbf{r},\mathbf{n})}{r^2} dS + 2\pi\rho(A), \tag{18}$$

$$\left(\frac{\partial u}{\partial n}\right)_{A_-} = \iint_S \rho(P)\frac{\cos(\mathbf{r},\mathbf{n})}{r^2} dS - 2\pi\rho(A), \tag{19}$$

where **n** is the unit vector of the external normal to the surface S at point A, $\mathbf{r} = \overrightarrow{AP}$. It should be mentioned that the normal **n** is fixed in this case.

2.4.3. The potential of the homogeneous sphere

The potential of the homogeneous sphere (simple layer) with constant density ρ is continuous and equal to

$$u(r) = \begin{cases} 4\pi R^2 \rho \dfrac{1}{r} = \dfrac{M}{r}, & \text{if} \quad r > R, \\ 4\pi R \rho = \dfrac{M}{R}, & \text{if} \quad r > R, \end{cases}$$

where $M = 4\pi R^2 \rho$ is the mass distributed on the sphere.

This result shows directly that in intersecting the simple layer the normal derivative of the potential shows a discontinuity. If in this case the layer is intersected in the direction of differentiaition, i.e. in the direction of increasing r, then the *jump of the normal derivative* is equal to:

$$\lim_{r \to R+0} u'(r) - \lim_{r \to R-0} u'(r) = -4\pi\rho.$$

2.4.4. The simple layer potential on a plane

Let in plane (x,y) we have a smooth curve L (closed or not closed) and on it a continuous function $\rho(P)$ be given. Since the logarithmic potential of the domain is written in the form (14), the expression

$$u(A) = \int_L \rho(P) \ln \frac{1}{r_{AP}} dl \tag{20}$$

is the *logarithmic simple layer potential*. Curve L is the *carrier line of the layer*, and $\rho(P)$ is its *density*. As in the case of the logarithmic potential of the domain, in this case we can treat grad u as the limit of the attraction force at point A located in the plane $z = 0$, from a simple layer distributed with density $(1/2)\rho$ not dependent on z, on the side surface of the cylinder with height $2h$, restored on L in the direction normal to the plane Oxy and symmetric in relation to it. The limit is examined at $h \to \infty$. The simple layer potential of the cylinder tends to infinity in this case, but its gradient tends to a finite limit.

At points A, not located on L, potential (20) is a continuous function of A differentiable any number of times in x and y under the integral sign. In particular,

$$\operatorname{grad} u = -\int_L \rho(P) \frac{\mathbf{r}}{r^2} dl, \quad \Delta u = \int_L \rho(P) \Delta\left(\ln \frac{1}{r}\right) dl = 0.$$

If the conditions of continuity of curvature are imposed on the carrier line of the layer L, then the logarithmic potential is continuous in intersecting the layer, i.e. it is continuous on the entire plane.

At infinity, the logarithmic simple layer potential behaves in the same manner as the logarithmic potential of the domain, i.e. satisfies the condition

$$u(A) = M \ln \frac{1}{r} + u^*(A),$$

where $M = \int_L \rho(P)dl$ and $u^*(A) \to 0$ at $r = \sqrt{x^2 + y^2} \to \infty$, also $|\text{grad}_A u^*| < C/r^2$. As an example, we examine the logarithmic simple layer potential of the interval $[-a,a]$ with constant density ρ = const. From (20) we obtain

$$u(x, y) = \rho \int_{\xi=-a}^{a} \ln \frac{1}{\sqrt{(\xi - x)^2 + y^2}} d\xi = -\frac{1}{2}\rho \int_{\xi=-a}^{a} \ln\{(\xi - x)^2 + y^2\} d\xi.$$

2.5. Double layer potential

2.5.1. Dipole potential

Let us have two electrical charges $+e$ and $-e$ of the same magnitude but opposite signs. The charges are situated at distance h from each other (dipole). The straight line, passing through them, is regarded as the direction from the negative to positive charge. This is the axis of the dipole **n**.

Let P be a point situated in the middle between the charges, A be an arbitrary point, and θ be the angle between the axis **n** and direction PA. Writing the potential at point A and transferring to the limit at $h \to 0$ in such a manner that the charges are directed to point P along the straight line connecting them, and the product eh tends to some finite limit ν, referred to as *the dipole coefficient*, we obtain the equation:

$$u(A) = \nu \frac{\cos\theta}{r^2} = \nu \frac{\partial(1/r)}{\partial n}, \tag{21}$$

which is the *dipole potential of the moment* ν situated at point P and having the oriented straight line **n** as its axis.

2.5.2. The double layer potential in space and its properties

We generalize the method of determination of the dipole potential for determining the *double layer potential*. We consider a smooth Lyapunov surface S which is the carrier surface of the potential, and on this surface we specify the continuous function $\nu(P)$ – the density of the moments of the double layer. About S we also assume that it is the *oriented surface*, i.e. the external and internal sides are shown. This means that if at some point P on S we select the positive direction of the normal **n** and point P is transferred on S along an arbitrary closed curve, and direction n changes continuously in this case, then on return to the initial point the direction of the normal coincides with the initial direction. On the oriented surface it is necessary to determine at every point the positive direction of the normal in such a manner that the unit vector **n** of this direction is continuous on it.

On the basis of equation (21) it is concluded that the spatial potential of the dipole satisfies the Laplace equation at all points $A \neq P$.

The following consideration will be made. On the normal at every point P of the surface we place intervals of the length $(1/2)h$ on both sides of P, where h is a sufficiently small quantity. At the ends of these intervals

we place charges $\pm(1/h)\nu(P)$ which are such that the direction from the negative charge to the positive one coincides with the direction of the positive normal. Thus, we obtain two simple layers with surface densities $\pm(1/h)\nu(P)$ situated at a close distance on both sides of S. In the limit at $h \to 0$ we obtain *a double layer on the carrier surface S with the density of the moments* ν with the potential

$$u(A) = \iint_S \nu(P)\frac{\cos\theta}{r^2}dS = \iint_S \nu(P)\frac{\partial(1/r)}{\partial n}dS. \tag{22}$$

Here θ is the angle between the unit vector of the normal **n** at point P and direction PA (A is an arbitrary point).

This potential is a continuous function of A at all points not located on the carrier surface S, and at such points $u(A)$ can be differentiated any number of times in x,y,z under the integral sign. In particular,

$$\mathrm{grad}_A u = \iint_S \nu(P)\mathrm{grad}_A\left(\frac{\cos\theta_{PA}}{r_{AP}^2}\right)dS_P,$$

$$\Delta_A u = \iint_S \nu(P)\Delta_A\left(\frac{\partial(1/r_{AP})}{\partial n}\right)_P dS_P = 0.$$

Thus, the *double layer potential satisfies the Laplace equation at all points not located on the carrier surface*.

From (22) it is quite easy to determine the behaviour of the double layer potential at infinity: $r^2|u(A)|<C$, $r^3|\mathrm{grad}_A u| < C$, where C is a constant. Thus, the *double layer potential decreases at infinity at the rate not slower than* C/r^2, *and its first partial derivatives not slower than* C/r^3. In particular, for the double layer potential $\lim_{r\to\infty} ru(A) = 0$.

We examine an important case when the *density of the moment of the layer is constant*. Let ν = const. Therefore

$$u(A) = \nu\iint_S \frac{\cos\theta}{r^2}dS = \pm\nu\sigma,$$

where σ is the solid angle under which the surface S is visible from point A. The sign of the right side part coincides with the sign of $\cos\theta$, i.e. depends on the selection of the positive direction of the normal. If point A is inside the closed surface, *the double layer potential with constant density of moments* ν *has a constant value for* $4\pi\nu$ *inside this surface (if the internal normal is regarded as positive). At all points, located outside this surface, the potential is equal to zero and equal to* $2\pi\nu$ *at all points on the layer* (i.e. when point A is on the surface).

In a general case, *the double layer potential on intersecting the layer shows a discontinuity, and the value of the jump on intersection in the direction on the positive normal to the point A of the layer is equal to*

$$u_+(A) - u_-(A) = 4\pi\nu(A),$$

the direct value of the potential at point A is equal to

$$u_0(A) = \frac{1}{2}\{u_+(A) + u_-(A)\},$$

and the normal derivative of the potential double layer remains continuous on intersecting the layer. More accurately

$$u_+(A) = \iint_S \rho(P)\frac{\cos(\mathbf{r},\mathbf{n})}{r^2}dS + 2\pi\rho(A), \tag{23}$$

$$u_-(A) = \iint_S \rho(p)\frac{\cos(\mathbf{r},\,\mathbf{n})}{r^2}dS - 2\pi\rho(A). \tag{24}$$

Here $\mathbf{r} = \overrightarrow{AP}$, $\mathbf{n}$ is the external normal at variable point P, $u_+(A)$ is the internal limit of the values of the potential at point A, $u_-(A)$ is the outer limit.

2.5.3. The logarithmic double layer potential and its properties

Examining the logarithmic potential of two points P_1 and P_2, with the charges $\pm\rho$ of the same magnitude but different signs and spaced from each other at distance h, writing the potential at point A and transferring to the limit $h\to\infty$ on the condition $h\rho = \nu$, we obtain

$$u(A) = \nu\frac{\cos\theta}{r} = \nu\frac{\partial\ln(1/r)}{\partial n}. \tag{25}$$

This equation is a *logarithmic potential of the dipole with moment* ν. On the basis of equation (25) it may be concluded that the logarithmic potential of the dipole satisfies the Laplace equation at all point $A\neq P$.

One the basis of the logarithmic dipole potential (25), we introduce the expression

$$u(A) = \int_L \nu(P)\frac{\cos\theta}{r}dl = \int_L \nu(P)\frac{\partial}{\partial n}\left(\ln\frac{1}{r}\right)dl, \tag{26}$$

where L is the smooth curve with a continuous curvature, and $\nu(P)$ is a function continuous on L. Equation (26) is the *logarithmic double layer potential, distributed on the carrier line L with the density of the moments* $\nu(P)$. Its physical interpretation, identical to the interpretation of the previously examined logarithmic potentials, may be easily defined.

When ν = const, integral (26) has a simple geometrical meaning, namely

$$\nu\int_L \frac{\cos\theta}{r}dL_P = \pm\nu\varphi,$$

where φ is the angle under which the chord of the curve L is visible from point A. This sign depends on the direction of selection of the normal to L. In particular, if L is a closed curve, then for point A located outside L

$$\nu\oint_L \frac{\cos\theta}{r}dl = 0,$$

and for point A located inside L

$$\nu\oint_L \frac{\cos\theta}{r}dl = \pm 2\pi\nu,$$

and the positive sign occurs if the normal is directed into L and the negative – for the opposite direction of the normal because in the former case θ is the acute angle, and in the second case it is the obtuse angle. At sufficiently large $r = \sqrt{x^2 + y^2}$

$$|u(A)| \le \nu_m \int_L \frac{dl}{r} < \frac{C}{r},$$

where ν_m is the highest value $|\nu(P)|$ on L and $C > 0$ is some constant. The following inequality holds for the gradient of the logarithmic double layer potential (26)

$$|\mathrm{grad}_A\, u| < 2\nu_m \int_L \frac{dl}{r^2} < \frac{C}{r^2}.$$

The logarithmic double layer potential (26) is continuous at all points A not located on L, and at these points it is differentiable unlimited number of times in x and y below the integral sign, and on intersecting the layer in the direction of the positive normal it has a discontinuity with the jump $2\pi\nu(A)$, where $\nu(A)$ is the density at the point of intersection of the layer. If $u_0(A)$ denotes the value of the logarithmic double layer potential at the point of the double layer A, and $u_+(A)$ and $u_-(A)$ denote the limits $u(B)$ when B tends to A from and, respectively, negative normal, then

$$u_+(A) = u_0(A) + \pi\nu(A), \qquad u_-(A) = u_0(A) - \pi\nu(A).$$

The behaviour of the normal derivatives of the logarithmic potentials of the simple and double layer is also identical to the behaviour of the appropriate derivatives of the Newton potential. In particular, the normal derivative of the *logarithmic simple layer potential* undergoes a discontinuity with the jump $-2\pi\rho(A)$ on intersecting the layer at point A in the direction of the positive normal:

$$\left(\frac{\partial u}{\partial n}\right)_{A+} - \left(\frac{\partial u}{\partial n}\right)_{A-} = -2\pi\rho(A)$$

(compare with (17)), and the normal derivative of the *logarithmic double layer potential* remains continuous on intersecting the layer.

3. USING THE POTENTIAL THEORY IN CLASSIC PROBLEMS OF MATHEMATICAL PHYSICS

3.1. Solution of the Laplace and Poisson equations

3.1.1. Formulation of the boundary-value problems of the Laplace equation

The Laplace equation

$$\Delta u = 0 \tag{27}$$

arises in many sections of the theory of electricity, heat conductivity, diffusion,

astrophysics, oceanology and atmospheric physics, and in analysis of other processes.

Mathematically, the Laplace equation has am infinite number of solutions because in order to separate one physically justified solution from them, it is necessary to impose additional boundary-value conditions. Usually, the boundary-value conditions on S are obtained from the physical formulation of the problem. If the boundary-value conditions of the first kind are known, i.e. the values of the unknown function u at the boundary

$$u|_S = g, \tag{28}$$

are given, it is then assumed that the Dirichlet problem is formulated for function u. If the solution is sought inside the bounded domain V, then we deal with the internal boundary-value problem, outside the external boundary-value problem. There is the uniqueness of the solution of the internal Dirichlet problem (27), (28). The uniqueness of the solution of the external Dirichlet problem is found only with the additional condition on the solution: in the three dimensional case it is the condition of tendency to zero at infinity, and in the two-dimensional case it is a weaker condition of tendency of the function to a finite limit at infinity.

It may easily be seen that to define the external Dirichlet problem in the case of the three-dimensional space it is insufficient to require that u has a finite limit at infinity. Really, let some amount of electricity be in equillibrium on the conducting surface S. This electrostatic simple layer potential has a constant value C on surface S and it may be easily shown that $u(A)$ gives the harmonic function outside S and will tend to zero at infinity. Constant C is also a harmonic function outside S and on S has the same boundary values, but it no longer tends to zero at infinity and, consequently, a non-uniqueness of the solution occurs. For a planar case this consideration is no longer applicable because the electrostatic simple layer potential on the line L converts to infinity at an infinitely distant point.

3.1.2 Solution of the Dirichlet problem in space

We examine an internal Dirichlet problem for domain V bounded by surface S. We search for its solution in the form of the double layer potential:

$$u(A) = \iint_S \rho(P) \frac{\cos(\mathbf{r}, \mathbf{n})}{r^2} dS, \quad r = |AP|, \tag{29}$$

where $\mathbf{r} = \overrightarrow{AP}$, $\mathbf{n}$ is the direction of the external normal at point P on the surface. It is required to find density $\rho(P)$. According to (23), the internal Dirichlet problem with the boundary-value $u|_S = f(P)$ is equivalent to the following integral equation for density $\rho(P)$:

$$f(A) \iint_S \rho(P) \frac{\cos(\mathbf{r}, \mathbf{n})}{r^2} dS + 2\pi\rho(A), \quad \mathbf{r} = \overrightarrow{AP}.$$

Introducing the kernel

$$K(A;P) = -\frac{1}{2\pi}\frac{\cos(\mathbf{r},\mathbf{n})}{r^2},$$

we can write the last equation in the form

$$\rho(A) = \frac{1}{2\pi} f(A) + \iint_S \rho(P) K(A;P)\, dS. \tag{30}$$

It should be mentioned that the kernel $K(A;P)$ is non-symmetric, because the normal is taken at point P and $\mathbf{r}$ denotes the direction $\vec{AP}$. The integral kernel of the adjoint equation is therefore determined by the formula

$$K^*(A;P) = K(P;A) = \frac{\cos(\mathbf{r},\mathbf{n})}{2\pi r^2}.$$

Consequently, determination of $\rho(A)$ is reduced to solving the integral Fredholm equation of the second kind (30).

Similarly, the solution of the external Dirichlet problem is reduced to solving the integral equation

$$\rho(A) = -\frac{1}{2\pi} f(A) - \iint_S \rho(P) K(A;P) dS. \tag{31}$$

As an example, we examine the first boundary-value problem for the Laplace equation in the half space $z \geq 0$ with the boundary condition $u|_S = f(P)$ where $S=\{(x,y,0)\}$. We find its solution in the form of the double layer potential

$$u(x,y,z) = \int_{-\infty}^{+\infty}\int \rho(\xi,\eta)\frac{\cos(\mathbf{r},\mathbf{n})}{r^2} d\xi d\eta. \quad r^2 = (x-\xi)^2 + (y-\eta) + z^2.$$

In this case, $\cos(\mathbf{r}, \mathbf{n})/r^2 = z/r^2$. Hence the kernel of the integral equation (3) is equal to zero since the values are taken at the boundary $z = 0$. This means that the density of the double layer potential $\rho(P) = f(P)/(2\pi)$ and the required solution is equal to (compare with (50).

$$u(x,y,z) = \frac{z}{2\pi}\int_{-\infty}^{\infty}\int \frac{f(\xi,\eta)}{[(\xi-x)^2 + (\eta-y)^2 + z^2]^{3/2}} d\xi d\eta. \tag{32}$$

It should be mentioned that if we look for the solution of the Dirichlet problem in the form of the simple layer potential, then to determine its density we use the integral Fredholm equation of the first kind which is an ill-posed problem and is considerably more complicated than (30), (31). In the case of simple domains V the method of the Green function (section 3.2.3) is a more suitable method of solving the boundary-value problems.

3.1.3. Solution of the Dirichlet problem on a plane

We examine the Dirichlet problem $\Delta u = 0$, $u|_L = f$. As in the case of considerations in the previous points, its solution is sought in the form of the double layer potential

$$u(A) = \int_L \rho(P)\frac{\cos(\mathbf{r},\mathbf{n})}{r} dl.$$

For the density of the potential ρ we obtain the integral equation:

$$\pi\rho(A) = f(\mathrm{A}) + \int_L \rho(P)K(A;P)dl. \tag{33}$$

The Fredholm equation (33) corresponds to the internal Dirichlet problem, and on changing the sign of the right-hand side it corresponds to the external Dirichlet problem. Here **n** is the variable normal, restored at point P, $\mathbf{r} = \overrightarrow{AP}$, $K(A;P) = -(\cos(\mathbf{r},\mathbf{n}))/r$. Equation (33) maybe written in the form

$$\pi\rho(l_0) = f(l_0) + \int_0^{|L|} \rho(l)K(l_0;l)\,dl, \tag{34}$$

where l and l_0 are the lengths of the arcs LP and LA of the contour L, counted from some fixed point L in a specific direction, and $|L|$ is the length of the contour L.

As an example, we solve the Dirichlet problem for the circle K_R with the radius R centred at the origin. If points A and P are located on the circle, then $(\cos(\mathbf{r}_{AP},\mathbf{n}))/r_{AP}=1/(2R)$. Integral equation (34) takes the form

$$\rho(l_0) + \frac{1}{\pi}\int_{K_R} \frac{1}{2R}\rho(l)dl = \frac{1}{\pi} f(l_0).$$

and its solution is the function

$$\rho(l) = \frac{1}{\pi} f(l) - \frac{1}{4\pi^2 R}\int_{K_R} f(l)dt.$$

The corresponding solution – the double layer potential – is equal to

$$u(\rho,\theta) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} f(t)\frac{R^2-\rho^2}{R^2-2\rho R\cos(t-\theta)+\rho^2}dt. \tag{35}$$

This solution, referred to as *the Poisson integral*, is also found in section 3.2.3 using the method of the Green function.

3.1.4. Solution of the Neumann problem

The boundary-value Neumann problem is given by the equation

$$\Delta u(A)=0,\ A\in V \tag{36}$$

and the boundary-value condition

$$\left.\frac{\partial u}{\partial n}\right|_S = f, \tag{37}$$

where **n** is the external normal to S.

The necessary condition of solvability of the internal Neumann problem is the equality

$$\iint_S f(P)dS = 0. \tag{38}$$

It should be mentioned that if some function $u(A)$ gives a solution inside the Neumann problem, the function $u(A) + C$, where C is an arbitrary constant, also gives a solution for the same boundary-value condition $f(P)$. The theorem of the uniqueness of the solution of the internal Neumann problem claims that if $u_1(A)$ and $u_2(A)$ are two solutions of the Neumann problem for

the same boundary-value condition $f(P)$, then the difference $u_2(A)–u_1(A)$ must be constant in domain V.

For the external three-dimensional Neumann problem the condition (38) is omitted but, as for the external Dirichlet problem, the condition of a tendency to zero at infinity is imposed on the solution. The solution of the Neumann problem (36) is sought as the simple layer potential.

$$u(A) = \iint_S \frac{\rho(P)}{r} dS.$$

Using (18),(19) we arrive at

$$2\pi\rho(A) = f(A) - \iint_S \rho(P) K^*(A;P)\, dS,$$

where

$$K^*(A;P) = \frac{\cos(\mathbf{r},\mathbf{n})}{r^2},$$

$\mathbf{n}$ is the fixed external normal at point A, $\mathbf{r} = \vec{AP}$. The right hand side of the integral equation for the external Neumann problem has the opposite sign.

If S is the Lyapunov surface, and the angle Θ between the normals at any two points of the surface with the distance r between them does not exceed the value $\Theta \le Cr$, then the kernels of the integral equations satisfy the estimate $|K(A;P)| \le C/r$ and the Fredholm theorems hold for these integral equations.

The solution of the Neumann problem in a plan is sought in the form of the simple layer potential

$$u(A) = \int_L \rho(P) \ln \frac{1}{r} dl,$$

and the following integral equation is obtained for the required density ρ for the internal problem

$$\pi\rho(A) = f(A) - \int_L \rho(R) K^*(A;P) dl,$$

and for the external problem the sign of its right-hand side changes. Here

$$K^*(A;P) = \frac{\cos(\mathbf{r},\mathbf{n})}{r},$$

where $\mathbf{r} = \vec{AP}$, $\mathbf{n}$ is the fixed external normal to L at point A.

Similar to equation (34), we can write the appropriate Fredholm equation of the second kind for determining the density of the simple layer potential.

It should be mentioned that the integral equations are uniquely solvable if the appropriate conditions of the existence and uniqueness of solution of the initial boundary-value problem are fulfilled.

3.1.5. Solution of the third boundary-value problem for the Laplace equation

This problem arises, for example, when examining the thermal equilibrium of a radiating body. For a steady heat flow, temperature $u(A)$ inside the body

should satisfy the Laplace equation, and the condition

$$\frac{\partial u}{\partial n} + h(u - u_0) = 0,$$

must be satisfied at boundary S. Here h is the coefficient of external heat conductivity and u_0 is the temperature of the external medium. In more general cases the following bounday-value condition forms

$$\frac{\partial u(P)}{\partial n} + q(P)u(P) = f(P), \tag{39}$$

where $q(P)$ and $f(P)$ are the functions defined on S and $q(P) > 0$. We search for the solution of this boundary-value problem in the form of the simple layer potential. The boundary-value condition (39) leads to the following integral equation for density:

$$\rho(A) = \frac{1}{2\pi} f(A) - \iint_S \rho(P)\left(\frac{q(A)}{2\pi r} + \frac{\cos(\mathbf{r},\mathbf{n})}{2\pi r^2}\right) dS. \tag{40}$$

In particular, if $q(P)$ is a positive constant h, and the surface S is the sphere with a unit radius, then from (40) we obtain

$$\rho(A) = \frac{1-2h}{4\pi} \iint_S \frac{\rho(P)}{r} dS + \frac{1}{2\pi} f(A).$$

Eigenvalues for this equation are $h = 0, -1, -2, \ldots$, and spherical functions are their eigenfunctions.

3.1.6. Solution of the boundary-value problem for the Poisson equation

We examine the Poisson equation

$$-\Delta u = f. \tag{41}$$

To reduce the solution of the boundary-value problem for the Poisson equation to the problem for the Laplace equation, it is sufficient to find some continuous partial solution v. We set $u = v + w$ and obtain a boundary-value problem for harmonic function w which is solved by one of the previously described methods, depending on the type of boundary-value conditions. It should be mentioned that the boundary-value conditions for function w depend on the values (and/or their derivatives) which the auxiliary function v. The particular auxiliary solution of the equation (41) is the volume potential

$$v(A) = -\frac{1}{4\pi} \iiint_V f(P) \frac{1}{r} dV,$$

and for a planar problem it is the logarithmic potential

$$v(A) = -\frac{1}{2\pi} \iint_S f(P) \ln \frac{1}{r} dS.$$

3.2. The Green function of the Laplace operator

3.2.1. The Poisson equation

If the continuous function $\rho(P)$ is continuously differentiable in V, the potential

$$u(A)=\iiint_V \frac{\rho(P)}{r}dV$$

satisfies the Poisson equation

$$\Delta u=-4\pi\rho$$

On the basis of this formula (which is used for examining the auxiliary solution in section 3.1.6), we introduce the Green function.

3.2.2. The Green function

If we substitute into the second Green fomula (5) the harmonic function v that is continuous in the entire volume V together with the first derivatives, and then sum up the resulting identity with the main integral Green formula (6), we obtain

$$u(A)=\iint_S\left(G\frac{\partial u}{\partial n}-u\frac{\partial G}{\partial n}\right)dS-\iiint_V \Delta u\cdot G dV, \tag{42}$$

where $G(A,P)=1/(4\pi r_{AP})+v$ is the function of two points A and P, point A is fixed. Equation (42) contains the values u and $\partial u/\partial n$ at boundary S. At the same time, when solving the first boundary-value problem, we specify only $u|_S$, and when solving the second boundary value problem – its normal derivative at boundary S. The function v is selected so that $G|_S=0$ for the first boundary-value problem and $\partial G/\partial n|_S=0$ for the second one.

We examine the Laplace equation for one of the following homogeneous boundary-value conditions

$$u|_S=0, \tag{43}$$

$$\frac{\partial u}{\partial n}+q(P)u\big|_S=0,\quad q(P)>0. \tag{44}$$

Definition. *The Green function of the Laplace operator*, corresponding to the boundary conditions (43) or (44), is the function $G(A,P)$ which satisfies as the function of P the following conditions for the arbitrarily fixed point $A\in V$:

1) inside V, with the exception of A, this function is harmonic;
2) the function satisfies the boundary-value condition (43) or (44);
3) it maybe presented in the form

$$G(A,P)=G(x,y,z;\xi,\eta,\zeta)=\frac{1}{4\pi r}+v(A;P), \tag{45}$$

where $r=|AP|$ and $v(A,P)$ is the harmonic function everywhere inside V.

The Green function does exist if S is the *Lyapunov surface*. It is symmetric and converts to infinity when A and P coincide.

Construction of the Green function is reduced to finding the harmonic function v satisfying the Laplace equation and specific boundary-value conditions. Thus, to determine the solution u of the boundary-value problem

it is necessary to find a solution v of the same problem but without derivatives and with special boundary conditions. This is a far simpler procedure.

In the case of the boundary-value conditions (43), the harmonic inside V function $v(A,P)$ should have the following boundary values on S:

$$v(A,P) = -\frac{1}{4\pi r}, \quad P \in S, \quad r = |AP|, \tag{46}$$

In the case of (44), the boundary-value conditions for $v(A,P)$ have the form

$$\left(\frac{\partial v(A,P)}{\partial n}\right) + q(P)v(A,P) = -\frac{1}{4\pi}\left[\frac{\partial(1/r)}{\partial n} + \frac{q(P)}{r}\right], \quad P \in S. \tag{47}$$

In the planar case, the definition of the Green function is completely identical, but the following formula should be introduced instead of (45)

$$G(A;P) = \frac{1}{2\pi}\ln\frac{1}{r} + v(A;P). \tag{48}$$

Let $u(A)$ be the solution of the internal Dirichlet problem for domain V, bounded by surface S, with the boundary values $f(P)$. Then, from equation (42) we obtain

$$u(A) = -\iint_S f(P)\frac{\partial G(A,P)}{\partial n}dS. \tag{49}$$

This equation gives the solution of the Dirichlet problem for any continuous function $f(P)$ from the boundary-value condition.

On the basis of (42), the solution of the Poisson equation $\Delta u = -\varphi$ in the bounded domain V with homogeneous boundary-value conditions of the first kind has the form

$$u(A) = \iiint_V G(A,P)\varphi(P)dV.$$

3.2.3. Solution of the Dirichlet problem for simple domains

The solution of the Dirichlet problem $\Delta u = 0$, $u|_s = f$ can be found by presenting the solution in the form of the double layer potential (section 3.1.2.) and the solution of the resulting integral equation (30). However, in the case of simple domains, the application of the method of the Green function may facilitate finding the solution.

For a half space, the Green function has the value

$$G(A,P) = \frac{1}{4\pi r} - \frac{1}{4\pi r_1},$$

where r is the distance from the variable point P to A, r_1 is the distance from P to A', symmetric in relation to the boundary of the half space at $z = 0$. Substituting this expression into (42) together with $\Delta u = 0$, $u|_S = f$, we obtain the solution of the Dirichlet problem for the half space (compare with (32)):

$$u(x,y,z)=\frac{z}{2\pi}\int\limits_{-\infty}^{+\infty}\int\int\frac{f(\xi,\eta)}{[(\xi-x)^2+(\eta-y)^2+z^2]^{3/2}}d\xi d\eta. \tag{50}$$

A sphere with radius R centered at the origin has the following Green function:

$$G(A,P)=\frac{1}{r}-\frac{R}{\rho}\frac{1}{r_1},$$

where distance r_1 is determined to point A', inverse to A in relation to the given sphere, $\rho = |OA|$, $r = |AP|$, $|OA|\cdot|OA'| = R^2$. The solution of the internal Dirichlet problem is as follows:

$$u(A)=\frac{1}{4\pi R}\iint\limits_S f(P)\frac{R^2-\rho^2}{r^3}dS.$$

If we introduce angle γ, formed by the vectors $\vec{OA}$ and $\vec{OA'}$, the spherical coordinates (ρ, θ, φ) of point A, then the solution of the internal Dirichlet problem for the sphere has the form

$$U(\rho,\theta,\varphi)=\frac{R}{4\pi}\int\limits_0^{2\pi}\int\limits_0^{\pi}f(\theta',\varphi')\frac{R^2-\rho^2}{(R^2-2\rho R\cos\gamma+\rho^2)^{3/2}}\sin\theta' d\theta' d\varphi. \tag{51}$$

The solution of the external Dirichlet problem for the sphere coincides with (51) if the sign of the formula is changed. Identical considerations lead to the solution of the Dirichlet problem for a circle, i.e. Poisson integral

$$u(\rho,\theta)=\frac{1}{2\pi}\int\limits_{-\pi}^{+\pi}f(t)\frac{R^2-\rho^2}{R^2-2\rho R\cos(t-\theta)+\rho^2}dt. \tag{52}$$

The same formula with accuracy to the sign gives the solution of the external problem.

3.3 Solution of the Laplace equation for complex domains

3.3.1. Schwarz method

It is assumed that the Dirichlet problem for the domains V_1 and V_2 at continuous boundary values has been solved, and the domains have the common part V_0. The Schwarz method makes it possible to solve the Dirichlet problem for the union $V=V_1\cup V_2$. For definiteness, we examine a planar case. The contours of the domains V_1 and V_2 are split by their intersection points into parts α_1 and β_1 for V_1 and α_2 and β_2 for V_2. Let continuous function $\omega(P)$ be given at the contour $l = \alpha_2\cup\alpha_2$ of the domain V. We continuously extend the function ω from α_1 to β_1. Thus, function ω_1 is set. Solving the Dirichlet problem for V_1 we construct in V_1 the harmonic function $u_1(A)$, equal to ω on α_1 and ω_1 on β_1. The values of this function on β_2 together with the values ω and α_2 are treated as the boundary values of the new harmonic function v_1 in V_2. In V_1 we now construct the harmonic function u_2 with the boundary values ω on α_1 and v_1 on β_1, and so on.

The Schwarz method is used not only for the Laplace equation but also for other elliptic equations. This method is also suitable for a three-dimensional case.

We mention another possibility of using the Schwarz method. We examine the solution of the external Dirichlet problem in the three-dimensional space. Let the space contains n closed surfaces S_k (k = 1,2,...,n), and the bodies restricted by them do not have any common points. V denotes the part of the space located outside all surfaces S_k, and V_k is the part of the space outside S_k. It is assumed that the Dirichlet problem is solved for all V_k at any continuous values on S_k and we show how we can solve in this case the Dirichlet problem for V. All domains V_k and domain V contain an infinitely remote point and, as usually, when solving the Dirichlet problem, it is assumed that the harmonic function is equal to zero at infinity.

Thus, it is necessary to find the function harmonic inside V and having the given continuous values on the surfaces S_k

$$u|s_k = f_k(P) \qquad (k=1,2,\ldots,n). \tag{53}$$

In the first step, we find for every k the function $u_{0,k}(A)$ (k = 1,2,...,n), harmonic inside V_k and having the values $f_k(P)$ on S_k. We also determine harmonic inside V_k functions $u_{1,k}(A)$ (k = 1,2,...,n) with the boundary values

$$u_{1,k}(P) = -\sum_{i\neq k} u_{0,i}(P) \quad \text{on} \quad S_k \quad (k=1,2,...,n), \tag{54}$$

and summation is carried out from i = 1 to i = n, with the exception of i = k.

Generally, for any integer positive m we determine functions $u_{m,k}(A)$ (k = 1,2,...,n), harmonic inside V_k, with the boundary values

$$u_{m,k}(P) = -\sum_{i\neq k} u_{m-1,i}(P) \quad \text{on} \quad S_k \quad (k=1,2,...,n). \tag{55}$$

Functions

$$\sum_{m=0}^{P} u_{m,k}(A) \quad (k=1,2,...,n)$$

are harmonic inside V_k with the boundary values

$$\sum_{m=0}^{p} u_{m,k}(P) = f_k(P) - \sum_{m=0}^{p-1}\sum_{i\neq k} u_{m,i}(P) \quad \text{on} \quad S_k \quad (k=1,2,...,n).$$

Subtracting the following sum from both parts

$$\sum_{m=0}^{p-1} u_{m,k}(P),$$

the previous equality may be written in the new form

$$\sum_{m=0}^{p-1}\sum_{i=0}^{n} u_{m,i}(P) = f_k(P) - u_{p,k}(P) \quad \text{on} \quad S_k \quad (k=1,2,...,n). \tag{56}$$

If P tends to infinity, the limiting function

$$u(A) = \sum_{m=0}^{\infty}\sum_{i=1}^{n} u_{m,i}(A) \tag{57}$$

gives the solution of the examined external boundary-value problem. It should

be mentioned that this form is not applicable to the external problem in the planar case.

3.3.2. The sweep method

When solving the Dirichlet problem by the potential method, comparatively strict restrictions must be imposed on the boundary of the domain. We describe another method of solving the Dirichlet problem suitable at very general assumptions about the boundary of the domain and boundary values at this boundary. This method is often referred to as the 'sweep method'. It was proposed by Poincaré and subsequently made more accurate by Perron.

We examine a planar case (for space the considerations are identical). We introduce some definitions. If $\Delta f \geq 0$, function f is *sub-harmonic*. If $\Delta f \leq 0$, it is a *super-harmonic function*. These functions have the following properties: the sub-harmonic function has the highest value at the contour. In addition to this, the function cannot have inside a maximum in the vicinity of which it is not constant. In exactly the same manner, the super-harmonic function has the lowest value of the contour.

Let $f_k(M)$ ($k = 1,\ldots,m$) be the functions continuous in the planar domain D and sub-harmonic inside D. We construct function $\varphi(A)$, which at every point D is equal to the highest of the values $f_k(A)$ ($k = 1,\ldots,m$):

$$\varphi(A) = \max[f_1(A),\ldots,f_m(A)]. \tag{58}$$

Consequently, $\varphi(A)$ is continuous in D and sub-harmonic inside D. Similarly, if $f_k(A)$ are super-harmonic and

$$\psi(\mathrm{A}) = \min[f_1(A),\ldots,f_m(A)], \tag{59}$$

then $\Psi(A)$ is also super-harmonic.

Let $f(A)$ be a sub-harmonic function inside D and continuous in D, K be a circle in D, and $u_K(A)$ is the function harmonic inside K whose values at the circumference of the circle K coincide with the values of $f(A)$. Therefore

$$f(A) \leq u_K(A) \text{ (in } K) \tag{60}$$

Similarly, if $f(A)$ is a super-harmonic function, then

$$f(A) \geq u_K(A) \text{ (in } K) \tag{61}$$

If the values of $f(A)$ in the circle K are substituted by the values $u_K(A)$ and the new function is denoted by $f_K(A)$, then this function, continuous in D, is super-harmonic inside D.

The same construction for the sub-harmonic function gives the sub-harmonic function $f_K(A)$.

We now describe the *Poincarè–Perron method*. Let a bounded domain D be on a plane and L be its boundary for which no assumptions have been made so far. We assume that on L we have the function $\omega(P) = \omega(x,y)$, for which it is only assumed at the moment that it is bounded, i.e. there are two such numbers a and b that

$$a \leq \omega(P) \leq b. \tag{62}$$

The *lower function* will refer to any function $\varphi(A)$ which is a continuous function in a closed domain, sub-harmonic inside the domain, and the condition $\omega(N) \leq \omega(P)$ is satisfied on the contour of this domain. Similarly, the upper function $\Psi(A)$ should be super-harmonic inside and on the contour it should

satisfy the condition $\Psi(N) \geq \omega(P)$.

If $f_1(A), f_2(A),..., f_m(A)$ are lower functions, then function $\varphi(A)$, determined by formula (58), is also a lower function. The same comment applies to the upper functions in formula (59). The set of the values of all possible upper functions $\Psi(A)$ at any fixed point A, located inside D, has the exact lower boundary $u(A)$, and $a \leq u\ (A) \leq b$ and function $u(A)$ is harmonic. If ω is continuous, then this upper boundary of the lower functions coincides with $u(A)$, which is a solution (generalized) of the Dirichlet internal boundary-value problem.

The generalized solution of the Dirichlet problem $u(A)$ at continuity of the function $\omega(P)$ at the boundary L maybe obtained by another method. The function $\omega(P)$ is extended to the entire plane, retaining its continuity. It is also assumed that $D_n (n = 1,2,...)$ is a sequence of domains located together with the boundaries L_n inside D and tending to D, so that any point A situated inside D is located inside all domains D_n, starting at some number n. Domains D_n may be, for example, produced from a finite number of circles. It is assumed that for domain D_n we are capable of solving the Dirichlet problem with continuous values on L_n.

Let us assume that $u_n(A)$ is the solution of the Dirichlet problem for D_n, and the boundary values on L_n are given as a continuation of function $\omega(P)$, which we have already discussed. In the case of unbounded increase of n, functions $u_n(A)$ tend to the previously constructed generalized solution $u(A)$ of the Dirichlet problem and this trend is uniform throughout any closed domain, located inside D. Thus, the limit of $u_n(A)$ does not depend neither on the method of extending $\omega(P)$, nor on the selection of domains D.

4. OTHER APPLICATIONS OF THE POTENTIAL METHOD

4.1. Application of the potential methods to the Helmholtz equation

4.1.1. Main facts

To find the solution of the homogeneous equation $u_{tt} = a^2\Delta u$ in the form of a steady sinusoidal regime with a given frequency, we obtain the Helmholtz equation

$$\Delta v + k^2 v = 0. \tag{63}$$

At infinity, the solutions of this equation should satisfy the irradiation principle

$$v = O(r^{-1}), \quad \frac{\partial v}{\partial r} + ikv = o(r^{-1}),$$

and for the two-dimensional case

$$v = O(r^{-1/2}), \quad \frac{\partial v}{\partial r} + ikv = o(r^{-1/2}).$$

in this case, the uniqueness theorem is valid.

A fundamental solution, satisfying the irradiation principle in the three-dimensional case, is the solution

$$v(A)=\frac{e^{-ikr}}{r}, \tag{64}$$

where r is the distance between some fixed point O and point A.

In a planar case, the fundamental solution, satisfying the emission principle is the solution $H_0^{(2)}(kr)$, where $H_0^{(2)}$ is the second Hankel function.

The irradiation condition gives the unique solution also for the heterogeneous equation

$$\Delta v + k^2 v = -F(A) \qquad (k>0) \tag{65}$$

determined in the entire space. $F(A)$ is a continuously differentiable function of the point of the three-dimensional Euclidean space, determined in the entire space and equal to zero outside some bounded domain V. Consequently, the given solution is determined by the equality

$$V(A)=\frac{1}{4\pi}\iiint\limits_V \frac{e^{-ikr}}{r}F(P)d\tau \quad (r=|PQ|). \tag{66}$$

As in the case of the boundary-value problems for the Poisson equation (section 3.1.6.), to solve the boundary-value problems for the inhomogeneous equation (65) it is necessary to present its solution in the form of the sum of the solution of the boundary-value problem for the homogeneous equation (63) (Helmholtz equation) and function (66).

4.1.2. Boundary-value problems for the Helmholtz equations

The solution (64) of equation (63) has, at $r = 0$, singularity $1/r$; consequently, for equation (63) we can construct a potential theory completely identical to the theory of the Newton potential for the Laplace equation. Denoting by r the distance between the variable point P on the surface S and point A, in the three-dimensional case we obtain the following analogues of the potentials of the simple and double layer:

$$v(A)=\iint\limits_S \mu(P)\frac{e^{-ikr}}{r}dS, \quad w(A)=-\iint\limits_S \mu(P)\frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{r}\right)dS, \tag{67}$$

where $\mathbf{n}$ is the direction of the external normal to S at variable point P. Separating the polar term $1/r$ from the kernel, we obtain the potentials in which the limit A tends to the surface is completed using the formulas identical to (18), (19), (23), (24):

$$\left(\frac{\partial v(A)}{\partial n}\right)_+ = 2\pi\mu(A)+\iint\limits_S \mu(P)\frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{r}\right)dS,$$

$$\left(\frac{\partial v(A)}{\partial n}\right)_- = -2\pi\mu(A)+\iint\limits_S \mu(P)\frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{r}\right)dS, \tag{68}$$

$$r=|AP|,$$

$$w_+(A) = 2\pi\mu(A) - \iint_S \mu(P)\frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{r}\right) dS,$$

$$w_-(A) = -2\pi\mu(A) - \iint_S \mu(P)\frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{r}\right) dS, \tag{69}$$

and in (68) the kernel is the value of the derivative in the direction of the normal **n** at point A, and in (69) in the direction of the normal **n** at point P.

In the planar case, the potentials have the form

$$v(A) = \int_l \mu(P)\frac{\pi}{2i}H_0^{(2)}(kr)dS, \quad w(A) = \int_l \mu(P)\frac{\partial}{\partial n}\left[\frac{\pi}{2i}H_0^{(2)}(kr)\right] dS, \tag{70}$$

and can be described by the equations completely identical to the equations (68) and (69), and the multiplier 2π on the right hand side should be replaced by π. These potentials satisfy the equation (63), and because of special selection of the kernels, each element of the written integrals and the integrals themselves satisfy the irradiation principle.

We introduce a kernel

$$K(A,P;k) = \frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{2\pi r}\right) = -\frac{e^{-ikr}(ikr+1)}{2\pi r^2}\cos\varphi_0,$$

where φ is the angle between the normal at point P and the direction AP. The transposed kernel has the form

$$K(P,A;k) = \frac{\partial}{\partial n}\left(\frac{e^{-ikr}}{2\pi r}\right) = -\frac{e^{ikr}(ikr+1)}{2\pi r^2}\cos\psi,$$

where ψ is the angle between the normal at point A and the direction AP. As in the case of the Laplace equation, we can formulate the Dirichlet and Neumann problems.

The internal Dirichlet problem consists of finding inside S the solution of equation (63) satisfying the boundary condition on S

$$u|_S = f(A).$$

The external problem is formulated in the same manner and the irradiation principle should be fulfilled at infinity. In the case of Neumann's problem, we have the boundary condition

$$\left.\frac{\partial u}{\partial n}\right|_S = f(A).$$

The uniqueness theorem shows that the external problems can have only one solution. For internal problems, the uniqueness does not occur.

Number k^2 is the *eigenvalue of the internal Dirichlet problem*, if inside S there is a solution of equation (63) satisfying the homogeneous boundary condition $u|_S = 0$ on S. The eigenvalues of the internal Neumann problem are determined in the same manner. To find the solution of the external Dirichlet problem in the form of the double layer potential and the internal

Neumann problem in the form of the simple layer potential, we arive at the adjoint integral equations

$$\mu(A)+\iint_S \mu(P)K(A,P;k)\,dS=-\frac{1}{2\pi}f(A), \tag{71}$$

$$\mu(A)+\iint_S \mu(P)K(P,A;k)\,dS=\frac{1}{2\pi}f(A). \tag{72}$$

If k^2 is not the eigenvalue of the internal Neumann problem, the homogeneous equation (71) has only a zero solution and, consequently, the inhomogeneous equation is solvable at any $f(A)$, i.e at any $f(A)$ the external Dirichlet problem has a solution in the form of the double layer potential. If k^2 is not the eigenvalue of the internal Dirichlet problem, then in the same manner the external Neumann has a solution in the form of the simple layer potential.

4.1.3. Green function

For equation (63) we can construct the Green function in exactly the same manner as carried out for the Laplace equation. In the three-dimensional case, the fundamental solution of this equation may be written in the form ($\cos kr/r$). The Green function, corresponding to the condition

$$u|_S=0 \tag{73}$$

should be found in the form

$$G_1(A,P;k^2)=\frac{\cos kr}{4\pi r}+g_1(A,P;k^2)\quad (r=|AP|), \tag{74}$$

where $g_1(A,P;k^2)$ satisfies equation (63) inside V and on S it satisfies the boundary-value condition

$$g_1(A,P;k^2)\Big|_s=-\frac{\cos kr}{4\pi r}\Bigg|_S. \tag{75}$$

If k^2 is not the eigenvalue of equation (63) at the boundary-value condition (73), then such a Green function may be constructed.

In the planar case, the solution of equation (63), which depends only on the distance $r=|AP|$, has the form $Z_0(kr)$, where $Z_0(z)$ is any solution of the Bessel equation of the zero order

$$Z_0''(z)+\frac{1}{z}Z_0'(z)+Z_0(z)=0. \tag{76}$$

To solve this equation, we use the Neumann function

$$N_0(z)=\frac{2}{\pi}J_0(z)\left(\ln\frac{z}{2}+C\right)-\frac{2}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{(k!)^2}\left(\frac{z}{2}\right)^{2k}\left(\frac{1}{k}+\frac{1}{k-1}+\ldots+1\right). \tag{77}$$

The fundamental solution with the singularity $(1/(2\pi)\ln(1/r)$ is the function

$$-\frac{1}{4}N_0(kr). \tag{78}$$

Therefore, the Green function should be found in the form

$$G_1(P,Q;k^2) = -\frac{1}{4}N_0(kr) + g_1(A,P;k^2). \tag{79}$$

Since the first term in the right hand side satisfies the equation and has the required singularity, the problem is reduced to determining the term $g_1(A,P;k^2)$ in such a manner that it has no singularity, satisfies equation (63) and on the contour L would satisfy the following heterogenous boundary-value condition

$$g_1\big|_L = \frac{1}{4}N_0(kr),$$

which gives the zero value of the Green function at boundary L.

4.1.4. Equation $\Delta v - \lambda v = 0$

We examine the equation

$$\Delta v - \lambda v = 0. \tag{80}$$

where λ is the given positive number, and formulate the internal Dirichlet problem with the boundary-value condition

$$v|_S = f(A) \tag{81}$$

The solutions of equation (80) cannot have inside V neither positive nor negative minima and this leads to the uniqueness of the solution of the given Dirichlet problem. If function $f(A)$ satisfies the inequality $-a \le f(A) \le b$, where a and b are some positive numbers, then the solution of the Dirichlet problem should also satisfy this inequality in V.

We examine the heterogeneous equation

$$\Delta v - \lambda v = -\varphi(A) \text{ inside } V \tag{82}$$

with a homogeneous boundary-value condition

$$v|_S = 0. \tag{83}$$

Let $\varphi(A)$ be continuous in the closed domain V and has continuous derivatives inside V. Consequently, the problem (82), (83) is equivalent to the integral equation

$$v(A) = \lambda \iiint_V G(A,P)v(P)d\tau + \iiint_V G(A,P)\varphi(P)d\tau, \tag{84}$$

where $G(A,P)$ is the Green function of the Laplace equation with the boundary-value condition (83). Since $(-\lambda)$ is the negative number, and all eigenvalues of the integral operator with the kernel $G(A,P)$ are positive, equation (84) has a single solution which is the solution of the problem (82), (83).

We now transfer to the solution of the Dirichlet problem (80) and (81). Let $w(A)$ be the solution of the Dirichlet problem for the Laplace equation with the boundary-value condition (81). The function

$$u(A) = v(A) - w(A) \tag{85}$$

should satisfy the condition

$$\Delta u - \lambda u = \lambda w$$

and the homogeneous boundary-value condition $u|_S = 0$. The solution of this problem does exist. Knowing $u(P)$ we can now find the solution of the Dirichlet problem $v(A)$ in accordance with formula (85).

The fundamental solution of equation (80) is the function

$$v_0(A) = \exp\left(-\sqrt{\lambda}r\right)/r, \tag{86}$$

where r is the distance from point A to some fixed point O. Using this solution, we can propose the theory of the potential in exactly the same manner at done previously.

4.2. Non-stationary potentials

4.2.1 Potentials for the one-dimensional heat equation

We examine the one-dimensional heat equation

$$u_t = a^2 u_{xx} \tag{87}$$

and set that for the interval $0 \le x \le l$ we have the boundary-value problem with the boundary-value conditions

$$u|_{x=0} = \omega_1, \qquad u|_{x=l} = \omega_2(t) \tag{88}$$

and, without the loss of the generality, with the homogeneous initial condition

$$u|_{t=0} = 0 \qquad (0 \le x \le l). \tag{89}$$

The fundamental solution, corresponding to the source placed at moment $t = \tau$ at point $x = \xi$, is the thermal potential

$$u = \frac{1}{2a\sqrt{\pi(t-\tau)}} \exp\left\{-\frac{(\xi - x)^2}{4a^2(t-\tau)}\right\}. \tag{90}$$

As in the construction of the dipole potential, we differentiate the fundamental solution in respect of ξ (along the 'normal' derivative) leading to a second singular solution – the heat potential of the 'dipole'. Multiplying this by the density of the potential $\varphi(\tau)$ and integrating, we obtain the solution (87), corresponding to the dipole at point $x = \xi$, acting from the moment $\tau = 0$ with intensity $\varphi(\tau)$, in the form (see [83])

$$u(x,t) = \int_0^t \frac{\varphi(\tau)}{2a\sqrt{\pi}(t-\tau)^{3/2}}(x-\xi)\exp\left\{-\frac{(\xi - x)^2}{4a^2(t-\tau)}\right\} d\tau. \tag{91}$$

If x tends to ξ from the left or right (compare with (23), (24)), the function (91) satisfies the following boundary-value relationships:

$$u(\xi+0,t) = \varphi(t), \qquad u(\xi-0,t) = -\varphi(t). \tag{92}$$

Moreover, the solution of (91) evidently satisfies the homogeneous initial condition

$$u|_{t=0} = 0. \tag{93}$$

The solution of the initial boundary-value problem (87), (88) should be sought in the form of the sum of two potentials: one, placed at point $x = 0$ and the other one at point $x = l$; the required density of the first potential is denoted by $\varphi(\tau)$ and that of the second one by $\psi(\tau)$:

$$u(x,t)=\int_0^t \frac{x\varphi(\tau)}{2a\sqrt{\pi}(t-\tau)^{3/2}}\exp\left\{-\frac{x^2}{4a^2(t-\tau)}\right\}d\tau+$$

$$+\int_0^t \frac{(x-l)\psi(\tau)}{2a\sqrt{\pi}(t-\tau)^{3/2}}\exp\left\{-\frac{(l-x)^2}{4a^2(t-\tau)}\right\}d\tau. \tag{94}$$

The boundary conditions (88), because of (92), are written in the form

$$\varphi(t)\text{-}l\int_0^t \frac{\psi(\tau)}{2a\sqrt{\pi}(t-\tau)^{3/2}}\exp\left\{-\frac{l^2}{4a^2(t-\tau)}\right\}d\tau=\omega_1(t),$$

$$\text{-}\psi(t)+l\int_0^t \frac{\varphi(\tau)}{2a\sqrt{\pi}(t-\tau)^{3/2}}\exp\left\{-\frac{l^2}{4a^2(t-\tau)}\right\}d\tau=\omega_2(t). \tag{95}$$

These equations represent a system of integral Volterra equations for $\varphi(\tau)$ and $\psi(\tau)$, and the kernels of these equations depend only on the difference $t-\tau$. Thus, in the given case, the potential method leads to the solution of a system of two integral equations.

If the function u is not given at one of the ends and its derivative $\partial u/\partial x$ is given instead, then the potential, generated by this boundary-value condition, should be determined on the basis of a retardation potential (90), and not its derivative. In these considerations it is quite simple to see the similarity with the analysis of the Dirichlet problem whose solution is found in the form of the double layer potential, and the Neumann problem with the solution in the form of the simple layer potential.

If, for example, the boundary-value conditions have the form

$$u\big|_{x=0}=\omega_1(t),\quad \frac{\partial u}{\partial x}\bigg|_{x=l}=\omega_2(t), \tag{96}$$

the solution for the problem with the homogeneous initial conditions should be found in the form of the sum of two potentials of the type

$$u(x,t)=\int_0^t \frac{x\varphi(\tau)}{2a\sqrt{\pi}(t-\tau)^{3/2}}\exp\left\{-\frac{x^2}{4a^2(t-\tau)}\right\}d\tau+$$

$$+\int_0^t \frac{a\psi(\tau)}{\sqrt{\pi}\sqrt{t-\tau}}\exp\left\{-\frac{(l-x)^2}{4a^2(t-\tau)}\right\}d\tau. \tag{97}$$

The first of the conditions (96) gives

$$\varphi(t)+\int_0^t \frac{a\psi(\tau)}{\sqrt{\pi}\sqrt{t-\tau}}\exp\left\{-\frac{l^2}{2a^2(t-\tau)}\right\}d\tau=\omega_1(t).$$

Differentiating equation (97) is respect of x and tending x to l, we obtain, in view of (92) and the second of the conditions (96)

$$\psi(t)+\int_0^t \frac{\varphi(\tau)}{2a\sqrt{\pi}(t-\tau)^{3/2}}\exp\left\{-\frac{l^2}{4a^2(t-\tau)}\right\}d\tau-$$

$$-l^2\int_0^t \frac{\varphi(\tau)}{4a^3(t-\tau)^{5/2}}\exp\left\{-\frac{l^2}{4a^2(t-\tau)}\right\}d\tau=\omega_2(t),$$

so that for $\varphi(\tau)$ and $\psi(\tau)$ we obtain a system of differential equations with the kernels dependent on $t-\tau$. It should be mentioned that this type of kernels (in the form of a convolution) indicates the feasibility of using the Laplace transform for solving the system of integral equations.

4.2.2. Heat sources in multidimensional case

The concept of the potential may also be applied to multidimensional heat conductivity problems. We restrict ourselves to discussing the results identical with those obtained previously. We examine a planar case, i.e. the equation

$$u_t=a^2(u_{xx}+u_{yy}). \tag{98}$$

Let domain D be given on the Oxy plane with the boundary L. The fundamental solution, corresponding to the source at point (ξ,η) acting from the time moment τ, has the form

$$u=\frac{1}{4\pi a^2(t-\tau)}\exp\left\{-\frac{r^2}{4a^2(t-\tau)}\right\},\quad r^2=(\xi-x)^2+(\eta-y)^2.$$

The analogue of the simple layer potential is given by the following equation:

$$u(x,y,t)=\frac{1}{2\pi}\int_0^t d\tau\int_L \frac{a(l,\tau)}{t-\tau}\exp\left\{-\frac{r^2}{4a^2(t-\tau)}\right\}dl, \tag{99}$$

where l is the length of the arc of the contour L counted from a fixed point, and $a(l,\tau)$ is the function of the variable point l of the contour and parameter τ. r denotes the distance from the point (x,y) to the variable point l of the contour L. The heat double layer potential is represented by the equation

$$v(x,y,t)=\frac{1}{2\pi}\int_0^t d\tau\int_L \frac{b(l,\tau)}{t-\tau}\frac{\partial}{\partial n}\exp\left\{-\frac{r^2}{4a^2(t-\tau)}\right\}dl, \tag{100}$$

where **n** is the direction of the external normal at the variable point of integration, or

$$v(x,y,t)=\int_0^t d\tau\int_L \frac{b(l,\tau)}{4\pi a^2(t-\tau)^2}\exp\left\{-\frac{r^2}{4a^2(t-\tau)}\right\}r\cos(\mathbf{r},\mathbf{n})\,dl, \tag{101}$$

where direction **r** is counted from point P, travelling on the contour L during integration, to point (x,y). If we introduce angle $d\varphi$ under which the element with length dl is visible from the point (x,y), the previous equation can be rewritten in the form

$$v(x,y,t)=-\int_L d\varphi\int_0^t \frac{b(l,\tau)}{4\pi a^2(t-\tau)^2}\exp\left\{-\frac{r^2}{4a^2(t-\tau)}\right\}r^2d\tau. \tag{102}$$

The boundary-values of the double layer potential at point $A(x_0,y_0)$ of the contour are determined by the equation

$$v(x_0, y_0, t) = (-1)^k b(A,t) + \\ + \int_0^t d\tau \int_L \frac{b(l,\tau)}{4\pi a^2 (t-\tau)^2} \exp\left\{-\frac{r_0^2}{4a^2(t-\tau)}\right\} r_0 \cos(\mathbf{r}_0, \mathbf{n})\, d\sigma, \tag{103}$$

where parameter k determines the change of the sign: $k = 1$ for the internal and $k = 2$ for the external problem; r_0 is the distance from the variable point of integration to point $A(x_0,y_0)$. The simple layer potential (99) is continuous at transition through the contour L, and its derivative along the normal at point A of the contour has at this point boundary values, determined from the formula

$$\frac{\partial u(x_0, y_0, f)}{\partial n} = -a(A,t) - \\ - \int_0^t d\tau \int_L \frac{a(l,\tau)}{4\pi a^2 (t-\tau)^2} \exp\left\{-\frac{r_0^2}{4a^2(t-\tau)}\right\} r_0 \cos(\mathbf{r}_0, \mathbf{n})\, dl. \tag{104}$$

Using these equations, we may transform the solution of the boundary-value problem to integral equations. Let us, for example, require to find a function $v(x,y,t)$, satisfying inside D the equation (98) which has the given boundary values at the contour L:

$$v|_L = \omega(s,t), \tag{105}$$

where s is the coordinate of the point of the contour, determined by the arc length s counted from some point. The initial data are assumed to be equal to zero. Searching the solution in the form of the double layer potential (100), we obtain, in view of the first of the equalities (103), the integral equation for the function $b(l,t)$:

$$-b(s,t) + \int_0^t d\tau \int_L \frac{b(l,\tau)}{4\pi a^2 (l-\tau)^2} \exp\left\{-\frac{r^2}{4a^2(t-\tau)}\right\} r \cos(\mathbf{r}, \mathbf{n})\, dl = \omega(s,t), \tag{106}$$

where r is the distance between the point s and l of the contour L, and direction $\mathbf{r}$ is counted from l to s. In the given equation, integration in l is carried out over the fixed interval $(0,|L|)$, where $|L|$ is the length of the contour L, and in integration in τ the upper limit is variable. In other words, the given integral equation has the form of Fredholm equations with respect to variable l and the nature of Volterra equations with respect to variable τ. Regardless of this mixed nature of equation (106), the conventional method of successive approximations for the Volterra equations is also converging

in the case of equation (106). The method is also suitable for the domain bounded by several contours. It is also easily generalized for the three-dimensional case and applicable for external problems. The reduction of the initial condition $u(x,y,0) = f(x,y)$ is usually achieved by transferring the heterogeneity from the initial condition to the boundary-value conditions using the representation of the required solution in the form of the sum of solution of the problem with the homogeneous initial conditions and the solution of the equation for the entire plane or the entire space. In the two-dimensional case, the solution for the entire plane has the form

$$u(x,y,t) = \frac{1}{4\pi a^2 t} \int_{-\infty}^{+\infty} \int \exp\left\{-\frac{r^2}{4a^2 t}\right\} f(\xi,\eta)\, d\xi\, d\eta.$$

4.2.3. The boundary-value problem for the wave equation

When solving the boundary-value problems for elliptical and parabolic equations, the construction is based on the fundamental solution of the appropriate differential equation. We shall use concept for the hyperbolic equation.

We examine a one-dimensional equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + c^2 u \tag{107}$$

in the interval $0 \le x \le l$ with homogeneous initial conditions

$$u|_{t=0} = u_t|_{t=0} = 0 \tag{108}$$

and boundary-value conditions

$$u|_{x=0} = \omega_1(t); \qquad u|_{x=1} = \omega_2(t). \tag{109}$$

It should be mentioned that the initial condition may always be reduced to homogeneous. The Bessel function of the imaginary variable $I_0\left(c\sqrt{t^2 - x^2}\right)$ is the fundamental solution of equation (107).

Placing the continuously acting sources, corresponding to this solution, at the ends of the interval $[0,l]$, we obtain, as may easily be seen, the ‘simple layer’ potentials – solutions of equation (107):

$$\int_0^{t-x} \varphi(\tau) I_0\left(c\sqrt{(t-\tau)^2 - x^2}\right) d\tau,$$

$$\int_0^{t-(l-x)} \psi(\tau) I_0\left(c\sqrt{(t-\tau)^2 - (l-x)^2}\right) d\tau,$$

where $\varphi(\tau)$ and $\psi(\tau)$ are some differentiable functions. Since the boundary-value conditions are the boundary-value conditions of the first kind, then as in the solution of the Dirichlet problem and the heat equation it is necessary to search for a solution in the form of the dipole potential (double layer). To determine this potential, we differentiate the fundamental 'simple layer' potentials in x and search for the solution of problem (107)–(109) in the form of the sum

$$u(x,t)=\frac{\partial}{\partial x}\int_0^{t-x}\varphi(\tau)I_0\left(c\sqrt{(t-\tau)^2-x^2}\right)d\tau+$$
$$+\frac{\partial}{\partial x}\int_0^{t-(l-x)}\psi(\tau)I_0\left(c\sqrt{(t-\tau)^2-(l-x)^2}\right)d\tau, \tag{110}$$

and it is assumed that $\varphi(\tau) = \psi(\tau) = 0$ at $\tau < 0$. Equation (107) and the initial conditions (108) are satisfied at any selection of $\varphi(\tau)$ and $\psi(\tau)$. The boundary-value conditions (109) lead to the following system of equation for $\varphi(\tau)$ and $\psi(\tau)$:

$$-\varphi(t)+\psi(t-l)+\int_0^{t-l}\psi(\tau)\frac{clI_0'\left(c\sqrt{(t-\tau)^2-l^2}\right)}{\sqrt{(t-\tau)^2-l^2}}d\tau=\omega_1(t),$$
$$-\varphi(t-l)+\psi(t)-\int_0^{t-l}\varphi(\tau)\frac{c\,l\,I_0'\left(c\sqrt{(t-\tau)^2-l^2}\right)}{\sqrt{(t-\tau)^2-t^2}}d\tau=\omega_2(t). \tag{111}$$

The functions $\omega_1(\tau)$ and $\omega_2(t)$ are assumed to be continuously differentiable. We denote

$$\psi(t)-\varphi(t)=\varphi_1(t),\qquad \psi(t)+\varphi(t)=\psi_1(t)$$

Adding and subtracting equation (111) term by term, we obtain separate equations for $\varphi_1(t)$ and $\psi_1(t)$:

$$\varphi_1(t)+\varphi_1(t-l)+cl\int_0^{t-l}\varphi_1(\tau)\frac{I_0'\left(c\sqrt{(t-\tau)^2-l^2}\right)}{\sqrt{(t-\tau)^2-l^2}}d\tau=\omega_1(t)+\omega_2(t),$$
$$-\psi_1(t)+\psi_1(t-l)+cl\int_0^{t-l}\psi_1(\tau)\frac{I_0'\left(c\sqrt{(t-\tau)^2-l^2}\right)}{\sqrt{(t-\tau)^2-l^2}}d\tau=\omega_1(t)-\omega_2(t), \tag{112}$$

where $\varphi_1(\tau) = \psi_1(\tau) = 0$ at $\tau < 0$.

In contrast to the system of integral equations for the heat equation, in this case we obtain a system of integral equations with a functional dependence. It can be solved using successive steps in intervals $[0,l]$, $[l,2l]$ and so on.

It should be mentioned that at $c = 0$ the fundamental solutions of the wave equation

$$u_{tt} = a^2 \Delta u$$

are the following functions (depending on the dimension of the space)

$$W_1(x,t) = \frac{1}{2a}\theta(at - r), \quad W_2(x,t) = \frac{\theta(at - r)}{2\pi a\sqrt{a^2t^2 - r^2}},$$

$$W_2(x,t) = \frac{\theta(t)}{2\pi a}\delta(a^2t^2 - r^2).$$

Here $r \equiv |x|$, $\theta(z)$ is the step function equal to zero at $z < 0$ and to unity in the opposite case. In the three-dimensional space, the wave potential is expressed in the form of the generalized δ-function. The wave potential is determined as a convolution of the fundamental solution and the density of the potential ρ, which is also assumed to be zero at $t < 0$. Therefore, we obtain the integral over interval $[0,t]$. Depending on the dimension, the wave potentials have the form

$$V_1(x,t) = \frac{1}{2a}\int_0^t \int_{x-a(t-\tau)}^{x-a(t-\tau)} \rho(\xi,\tau)d\xi d\tau,$$

$$V_2(x,t) = \frac{1}{2\pi a}\int_0^t \int_{K(a(t-\tau))} \frac{\rho(\xi,\tau)d\xi d\tau}{\sqrt{a^2(t-\tau)^2 - |x-\xi|^2}},$$

$$V_3(x,t) = \frac{1}{4\pi a^2}\int_{K(at)} \frac{\rho\left(\xi, t - \frac{|x-\xi|}{a}\right)}{|x-\xi|}d\xi.$$

Here $K(z)$ is a circle with a centre at the point x with radius z. The three-dimensional wave potential V_3 is retarded because its value at point x at time $t > 0$ is determined by the values of the source $\rho(\xi,\tau)$, taken at early moments of time $\tau = t - |x-\xi|/a$, and the *retardation time* $|x-\xi|/a$ is the time required for a pertubation to move from point ξ to point x.

Further examination of the wave potentials results in the determination of the surface wave potentials of the simple and double layers. The examination of the boundary-value problem described previously for the wave equation uses these constructions.

BIBLIOGRAPHIC COMMENTARY

The history of the development of the mathematical theory of the potential, starting with the studies by Laplace and Lagrange, and also the history of the theory of boundary-value problems for the Laplace equations and equations containing a Laplacian is described in [86]; this study also

gives the theorems and describes the approaches of the potential theory in the primary formulation and shows their gradual improvement and development.

The fundamentals of the potential theory with detailed proofs are given in text books [83,85,91]. The methods of retardation potentials for parabolic and hyperbolic equations are described in [83,91]. A review and handbook material for the fundamentals of the potential theory is given in [49].

A fundamental study of the theory of the potential proving the properties of the potential and describing the methods of solving Neumann and Dirichlet problems, is in the monograph [20]. The potential theory with specific reference to the elliptical problems with detailed proofs is described in [87]. The single approach to introducing and using the concept of the potential irrespectively of the type of equation is given in [13].

The current theory of the potential is described in [47], and its abstract generalization is given in [5]. The book [97] is an introduction to the modern potential theory and is at the contact of applied and abstract potential theory; explanation is given there for the space with dimension $n \geq 3$. The boundary-value problems of the theory of electromagnetic oscillations are discussed in [41] in which antennae (wave) potentials and other non-stationary potentials of oscillation theory are also constructed. Fundamentals of the theory of hydrodynamic potentials are described in [43].

Chapter 3

EIGENFUNCTION METHODS

Keywords: Eigenvalues, eigenfunctions, Fourier methods, special functions, the eigenfunction method, orthonormalized systems, Fourier series, spherical functions, cylindrical functions, orthogonal polynomials, Sturm–Liouville problem, problems of the theory of electromagnetic phenomena, heat conductivity problems, problems of the theory of oscillations.

MAIN CONCEPTS AND NOTATIONS

Eigenvalues – values of parameter λ, at which the homogeneous equation of the type $Au = \lambda u$, where A is an operator, has non-zero solutions.

Eigenfunctions – Non-zero solutions u of homogeneous equations of the type $Au = \lambda u$.

Eigenvalue problems – Problems of finding eigenvalues λ and eigenfunctions u as solutions of the equation of the type $Au = \lambda u$, where A is an operator.

Eigenfunction method – The method of finding solutions of problems by expansion in respect of eigenfunctions.

Orthonormalized systems – A finite or infinite system of functions in which all functions are normalized, and any two functions are orthogonal.

Fourier series – A series consisting of functions of the orthonormalized system with the special coefficients – weights.

Spherical functions – eigenfunctions of the Laplace operator on a sphere.

Cylindrical functions – The solutions of the Bessel equation $x^2y^u+xy'+(x^2-n^2)y = 0$.

Orthogonal polynomials – Polynomials forming an orthogonal system.

1. INTRODUCTION

Many problems of mathematical physics lead to the so-called eigenvalue problems represented by linear homogeneous equations with a parameter. The non-zero solutions of these equations – eigenfunctions – play an important role in finding the solutions of initial problems. In a number of cases, it is

necessary to use special functions which are often eigenfunctions of specific eigenvalue problems.

One of the most frequently used methods in mathematical physics is the eigenfunction method based on finding solutions in the form of expansions in respect of the eigenfunctions of the operators, closely linked with the examined problem. These expansions are usually formed by orthonormalized functions with special weights – Fourier series. The eigenfunction method makes it possible to solve a wide range of mathematical physics problems, including the problems of the theory of electromagnetic phenomena, heat conductivity problems, the problems of the theory of oscillations, including acoustics.

The individual applications of the eigenfunction method relate to L. Euler, and the general formulation of the methods was proposed for the first time by M.V. Ostrogradskii. The method was strictly substantiated by V.A. Steklov.

The eigenfunction method is closely linked with the Fourier method – the method of separating variables designed for finding partial solutions of differential equations. These methods often lead to special functions which are solutions of eigenvalue problems. The method of separation of variables was proposed by J.D'Alembert (1749) and already used in the eighteenth century by L. Euler, D. Bernoulli and J. Lagrange for solving the problem of oscillations of a string. The method was developed quite extensively at the start of the nineteenth century by J. Fourier and used for the problems of heat dissipation. The method was formulated fully by M.V. Ostrogradskii (1828).

In this chapter, we describe the fundamentals of the method of expansion in respect of the eigenfunctions and examine applications of the method for solving specific problems of mathematical physics, including the problems of the theory of electromagnetic phenomena, heat conductivity problems, problems of the theory of oscillations, including acoustics [1,4,13,19,26,37,49,110].

2. EIGENVALUE PROBLEMS

Many problems of mathematical physics lead to the so-called eigenvalue problem which usually represents homogeneous equations with a parameter. The values of parameter at which the equation has non-zero solutions are referred to as eigenvalues, and the corresponding solutions as eigenfunctions.

The simplest eigenvalue problems were solved by L. Euler, and these problems were studied extensively in the nineteenth century when formulating the classic theory of mathematical physics equations.

2.1. Formulation and theory

We examine the simplest example leading to an eigenvalue problem. It is assumed that we have a homogeneous string with length l, whose ends are fixed and the string is not subjected to any external sources. The origin of the coordinates is located in one of the ends of the string and the x axis

is directed along the string. Function $u(x;t)$, describing free small oscillations of such a string, satisfies the homogeneous differential equation:

$$\frac{\partial^2 u}{\partial t^2} - a^2 \frac{\partial^2 u}{\partial x^2} = 0 \qquad (1)$$

and homogeneous boundary-value conditions

$$u(0;t) = 0; \qquad u(l;t) = 0.$$

Every specific movement of the string is determined, with the exception of the equation and the boundary-value conditions, also by some initial conditions.

We examine the simplest possible movement of the examined string – the so-called standing waves. A *standing wave* is a motion during which the forms of the string at different moments of time are similar to each other. The standing wave is given by a function having the form

$$u(x;t) = X(x)T(t),$$

where function $T(t)$, which depends only on time t, is the *law of oscillation* and describes the nature of motion of the individual points of the string, and function $X(x)$, which depends only on coordinate X, describes the form of the string at different moments of time, with the same accuracy to the multiplier $T(t)$.

For a string with fixed ends, it is above all evident that function $X(x)$ should satisfy the conditions $X(0) = 0$; $X(l) = 0$. In addition to this, $X(x)$ and $T(t)$ should satisfy some equations resulting from equation (1). In order to obtain these equations, we substitute in the equation u through X and T; consequently, we obtain the equality

$$X(x)T''(t)=a^2T(t)X''(x)$$

Dividing both parts of the equality by $a^2X(x)T(t)$ gives:

$$\frac{T''(t)}{a^2T(t)} = \frac{X''(x)}{X(x)}.$$

Since the left hand part of the equation depends only on t, and the right-hand side is independent of t, this means that the left-hand side and, consequently, the right-hand side are equal to the same constant. This constant is denoted by $-\lambda$:

$$\frac{T''(t)}{a^2T(t)} = \frac{X''(x)}{X(x)} = -\lambda,$$

and consequently, $T''+\lambda a^2T = 0$, $X'' + \lambda X = 0$. This is the simplest eigenvalue problem. It may easily be seen that constant λ may have the values only $\lambda_n = \pi^2n^2/l^2$ ($n = 1,2,3,\ldots$), and *in the string with fixed ends there may be only standing waves of the following form*:

$$X_n(x) = c\sin\frac{\pi n x}{l}, \quad c = \text{const}.$$

We now determine functions $T_n(t)$, corresponding to the form of the wave $X_n(x)$. For this purpose, the value λ_n is substituted into the equation for T:

$$T_n'' + \frac{a^2\pi^2}{l^2} n^2 T_n = 0.$$

The general integral of this equation has the form::

$$T_n(t) = B_n \sin\frac{\pi an}{l} t + C_n \cos\frac{\pi an}{l} t = A_n \sin\left(\frac{\pi an}{l} t + \varphi_n\right),$$

where B_n and C_n or A_n and φ_n are arbitrary constants.

Using X_n and T_n, we can write the final expression for all possible standing waves:

$$u_n(x,t) = A_n \sin\left(\frac{\pi an}{l} t + \varphi_n\right) \sin\frac{\pi nx}{l},$$

where n = 1,2,3,... Thus, the *n-th standing wave describes motion of the string during which each point of the string carries out harmonic oscillations with the frequency which is the same for all points* $\pi an/l$. The amplitudes of these oscillations change from point to point and are equal to $A_n|\sin(\pi nx/l)|$ (A_n changes arbitrarily).

Since the free oscillations of the examined string are unambiguously determined by the initial form $u|_{t=0}$ and by the initial velocity of its points $\left.\frac{\partial u}{\partial t}\right|_{t=0}$, it is evident that the standing wave forms iff the initial deviation and initial velocity have the form:

$$u\big|_{t=0} = D\sin\frac{\pi nx}{l}, \quad \left.\frac{\partial u}{\partial t}\right|_{t=0} = E\sin\frac{\pi nx}{l}, \quad D,\ E = \text{const}.$$

this standing wave is:

$$u(x,t) = \left(\frac{l}{\pi an} E\sin\frac{\pi an}{l} t + D\cos\frac{\pi an}{l} t\right)\sin\frac{\pi nx}{l}.$$

Quantities λ_n are also *eigenvalues*, and functions $X_n(x)$ are *eigenfunctions*.

We introduce the general definition of the eigenvalues and eigenfunctions. Let L be a linear operator with the definition domain $D(L)$. We examine a linear homogeneous equation:

$$Lu = \lambda u \tag{2}$$

where λ is a complex parameter. This equation has a zero solution at all λ. It may be that at some λ it has non-zero solutions from $D(L)$. The complex values of λ at which equation (2) has non-zero solutions from $D(L)$ are the *eigenvalues* of the operator L, and the corresponding solutions are the *eigenelements* (functions) corresponding to this eigenvalue. The total number $r(1 \le r \le \infty)$ of linearly independent eigenelements, corresponding to the given eigenvalue λ, is referred to as the *multiplicity* of this eigenvalue; if multiplicity r = 1, λ is the *simple eigenvalue*.

If multiplicity r of eigenvalue λ of the operator L is finite and $u_1, u_2,\ldots,u_r$ are the corresponding linearly independent eigenvalues, then any linear combination of these eigenvalues $u_0 = c_1u_1 + c_2u_2 + \ldots + c_ru_r$ is also an eigenelement corresponding to this eigenvalue, and this equation gives the general solution

of equation (2). From this it follows that: *if the solution of equation $Lu = \lambda u + f$ does exist, then its general solution is represented by the formula:*

$$u = u^* + \sum_{k=1}^{r} c_k u_k,$$

where u^* is a partial solution and c_k, $k = 1,2,\ldots,r$, are arbitrary constants.

The eigenvalues and eigenfunctions often have a distinctive physical meaning: in the example examined previously, the eigenvalues λ_n determine the frequency of harmonic oscillations of the string, and the eigenfunctions X_n the amplitude of oscillations.

2.2. Eigenvalue problems for differential operators

We examine a more general case of a mixed problem for homogeneous differential equations and homogeneous boundary-value conditions.

Let us have a bounded domain Ω in the space of one, two or three dimensions. P denotes an arbitrary point in domain Ω and $u(P)$ the function of the coordinates of this point. We examine a linear differential operator of the function u of the type:

$$L[u] \equiv \operatorname{div} p \operatorname{grad} u - qu,$$

where $p(P)$ and $q(P)$ are functions continuous in Ω and at its boundary $\partial\Omega$. In addition to this, we assume that $p(P) > 0$ inside and at the boundary of the domain Ω.

If the problem is one-dimensional, domain Ω is reduced to the interval (a,b) of the x axis. In this case, the operators grad and div refer to $\frac{d}{dx}$, and, consequently

$$L[u] = \frac{d}{dx}\left[p(x)\frac{du}{dx}\right] - q(x)u \equiv p(x)\frac{d^2u}{dx^2} + p'(x)\frac{du}{dx} - q(x)u.$$

At the boundary $\partial\Omega$ of the domain Ω we examine homogeneous boundary-value conditions $\Lambda[u] = 0$, $\Lambda[u] \equiv p\frac{\partial u}{\partial n} - \gamma u$, or $\Lambda[u] \equiv u$. In the first case, we are concerned with the *boundary-value conditions of the third kind* (or, if $\gamma = 0$, of the second kind), and in the second case the *boundary-value conditions of the first kind*. Here γ is continuous and non-negative function given on $\partial\Omega$, and n is the direction of the internal normal to $\partial\Omega$.

In a one-dimensional case, the boundary of the domain consists of two ends of the interval a and b and under the derivative $\frac{\partial}{\partial n}$ we should understand $\frac{d}{dx}$ at point a and $-\frac{d}{dx}$ at point b. The definition of function λ is therefore reduced to specifying two non-negative numbers γ_a and λ_b, and the definition of the operator $\Lambda[u]$ to the operator:

$$\Delta_a[u] = p(a)\frac{du(a)}{dx} - \gamma_a u(a), \quad \Delta_b[u] = -p(b)\frac{du(b)}{dx} - \gamma_b u(b).$$

In some cases we examine the boundary-value conditions of a different type – the so-called *periodicity conditions*. For example, in a one-dimensional case these conditions are given by the equalities $u(a) = u(b)$, $p(a)u'(a) = p(b)u'(b)$. These boundary-value conditions are also homogeneous but the main difference between them is that each of the equalities includes both points a and b.

We examine the following problem:

$$L[u]+\lambda\rho u=0 \quad \text{in} \quad \Omega, \tag{3}$$

$$\Lambda[u] = 0 \quad \text{on} \quad \partial\Omega, \tag{4}$$

where $\rho = \rho(P)$ is a non-negative function continuous in the domain Ω referred to as the *weight function* of the given problem or simply the *weight*. As shown in the case of the string, the solutions u, satisfying the boundary-value conditions, are possible at all values of λ.

The values of parameter λ, at which there are solutions of equations (3) which are not exactly equal to zero and which satisfy the boundary-value conditions (4), are the eigenvalues and the corresponding solutions u are the eigenfunctions of operator L. If the number of eigenfunctions is so large that any function given on the domain Ω (satisfying some natural smoothness conditions) can be expanded into a series in respect of these eigenfunctions, then we find the solution of heterogeneous problems in the form of a series in respect of the appropriate eigenfunctions.

In the following section we present a number of well-known properties of the eigenvalue problem (3), (4).

2.3. Properties of eigenvalues and eigenfunctions

We examine the problem (3), (4) and describe a number of the properties of eigenvalue and eigenfunctions (assuming that they exist).

1. *If there is an eigenvalue λ_0, then the eigenfunction corresponding to this eigenvalue is continuous together with its derivatives to the second order in the domain Ω.*

This property follows from the general theorems of the theory of differential equations if it is taken into account that u_0 is the solution of equation (3).

2. *If there is an eigenvalue λ and eigenfunction u_0 corresponding to this eigenvalue, then at any constant C function Cu_0 is also an eigenfunction corresponding to the same eigenvalue λ.*

3. *If there is an eigenvalue λ and there are two (or several) eigenfunctions u_1 and u_2 corresponding to this eigenvalue, then their sum is again an eigenfunction corresponding to this eigenvalue.*

Since the eigenfunctions are determined with the accuracy to a constant multiplier, it is convenient to describe the selection of this multiplier. It is often convenient to select this multiplier in such a manner as to satisfy the relationship

$$\int_{\Omega} \rho u_0^2 d\mu = 1. \tag{5}$$

(Here and in the following the integral over domain Ω of an arbitrary dimension is denoted by the single sign $\int$, and the element of the length, area or volume of this domain is denoted by $d\mu$).

The *function satisfying condition* (5) *is referred to as normalized with weight* ρ. It is assumed (if not stated otherwise) that all *examined eigenfunctions are normalized with weight* ρ.

4. *Two eigenfunctions* u_1 *and* u_2, *corresponding to different eigenvalues* λ_1 *and* λ_2 *are orthogonal in relation to each other in domain* Ω *at the weight* ρ, *i.e.* $\int_{\Omega} \rho(P) u_1(P) u_2(P) d\mu = 0.$

5. *If the normalized eigenfunction* u_0 *corresponds to eigenvalue* λ_0, *then* $K[u_0, u_0] = \lambda_0$, where

$$K[u,v] = -\int_{\Omega} L[u] v d\mu.$$

Since according to the assumption always $p > 0$ and $\gamma \geq 0$, then

$$K[u,u] \geq \int_{\Omega} q u^2 d\mu,$$

and if function q/ρ is lower bounded by the number m, then for the normalized eigenfunctions we have:

$$\lambda = K[u,u] \geq m.$$

The last inequality gives a simple lower estimate of the eigenvalue. *In particular, if function q is non-negative, then* (*since* ρ > 0) *all eigenvalues are non-negative. If function q is lower bounded by a positive number, then all eigenvalues are positive.*

2.4. Fourier series

The theoretical basis of the eigenfunction method is the theory of the Fourier series.

Let it be that Ω is a bounded domain and ρ is the weight function given in this domain and is continuous and non-negative in Ω and at its boundary and is strictly positive inside Ω.

The finite or infinite system of functions $u_1, u_2, u_3, \ldots, u_n, \ldots$, given in the domain Ω, is referred to as the *orthonormalized system* with the weight ρ if:

a. All functions u_n are normalized with the weight ρ: $\int_{\Omega} \rho u_n^2 d\mu = 1;$

b. any two functions u_i and u_k ($i \neq k$) are orthogonal in relation to each other with weight ρ.

If the weight is equal to unity, then it is usually assumed that the system is *orthonormalized* (without indicating the weight).

Let us have an orthonormalized system u_1, u_2 ... with the weight, and the function f in Ω, represented in the form of a linear combination of the functions

of this system: $f = \sum_{i=1}^{\infty} a_i u_i$.

The expression $a_k = \int_\Omega f u_k \rho d\mu$ is the *Fourier coefficient of the function f in respect of the system* $\{u_k\}$. The series $\sum_{i=1}^{\infty} a_i u_i$, in which coefficients a_i are the Fourier coefficients of function *f*, is the *Fourier series of this function in respect of the orthonormalized system* $\{u_i\}$.

In different problems, the following issue plays an important role. We have function *f* in the domain Ω, which, generally speaking, cannot be represented by the linear combination of the finite number of functions $\{u_i\}$. At a fixed *n* it is necessary to select number c_i such that the function $\sum_{i=1}^{n} c_i u_i$ gives the best approximation of function *f*, and to find the error of this approximate representation. The so-called quadratic error is used quite often.

The *quadratic error with* ρ *in replacing function f by another function* φ gives the expression:

$$\delta = \sqrt{\int_\Omega [f - \varphi]^2 \rho d\mu}.$$

Thus, we obtain the problem: to select coefficient c_i in the sum $\sum_{i=1}^{n} c_i u_i$ such that the expression

$$\delta_n^2 = \int_\Omega \left[f - \sum_{i=1}^{n} c_i u_i \right]^2 \rho d\mu$$

is the smallest, and calculate δ_n^2 for these values of c_i.

It is well-known that *the best approximation of function f in the sense of quadratic error with weight* ρ *by the linear combination of the functions* u_i *of the orthonormalized system is obtained if the coefficients of the linear combination are equal to the Fourier coefficients of function f in respect of the system* u_i. The error of this approximation is determined by equation:

$$\delta_n^2 = \int_\Omega f^2 \rho d\mu - \sum_{i=1}^{n} a_i^2,$$

This leads to a very important inequality:

$$\sum_{i=1}^{n} a_i^2 \le \int_\Omega f^2 \rho d\mu.$$

Passing now to the limit at $n \to \infty$, we obtain the inequality

$$\sum_{i=1}^{\infty} a_i^2 \le \int_\Omega f^2 \rho d\mu,$$

which is usually referred to as the *Bessel inequality*.

The condition $\delta_n \to 0$ as $n \to \infty$ is equivalent to the requirement according

to which the Bessel inequality should be an equality. Consequently, we obtain the *Parseval equality*

$$\sum_{i=1}^{\infty} a_i^2 = \int_{\Omega} f^2 \rho d\mu.$$

If the mean quadratic error of representing the function by the n-th partial sum of the series $\sum_{k=1}^{\infty} \varphi_k$ tends to zero as $n \to \infty$, it is said that this series *converges in the mean* to the function f or that the function f is expanded into series $\sum_{k=1}^{\infty} \varphi_k$ in the sense of convergence in the mean. Thus, the Parseval equality is the necessary and sufficient condition of expandibility of function f in its Fourier series in the sense of convergence in the mean.

Following V.A. Steklov, the orthonormalized system $\{u_k\}$ is referred to as *closed in domain* Ω, if any function f for which $\int_{\Omega} f^2 \rho d\mu$ converges can be represented accurately (in the sense of the quadratic error) by the linear combination of the functions of this system or, in other words, if each function f may be expanded into a Fourier series converging in the mean in respect of the function of the given system. This shows that the *requirement of closure of the orthnormalized system is equivalent to the requirement that the Parseval equality is fulfilled for any function f*. Therefore, this equality, following again V.A. Steklov, is referred to as the *closure condition*.

2.5. Eigenfunctions of some one-dimensional problems

We present several important examples of the eigenvalues and eigenfunctions of the one-dimensional problems of the type $L[u]+\lambda\rho u = 0$.

Example 1. The interval $(0,l)$. Operator $L[u]=u''$. The boundary-value conditions $u(0)=u(l)=0$. Weight $\rho=1$. Equation $u''+\lambda u=0$. The eigenvalues have the form $\lambda_n=\pi^2 n^2/l^2$ $(n=1,2,\ldots)$. The corresponding eigenfunctions are defined by equation $u_n=\sin(\pi n x/l)$.

Example 2. Interval $(0,l)$. Operator $L[u]=u''$. The boundary-value conditions $u'(0)=u'(l)=0$. Weight $\rho=1$. Eigenvalues λ_n are defined by $\lambda_n = b(n-1)^2\pi^2/l^2$ $(n=1,2,\ldots,)$ and the eigenfunctions have the form $u_1=\text{const}$, $u_n=\cos((n-1)\pi x/l)$, $n = 2,3,\ldots$.

Example 3. Interval $(0,l)$. Operator $L[u]=u''$. The boundary-value conditions $u(0) = u'(l) = 0$. Weight $\rho = 1$. The eigenvalues are defined by $\lambda_n = \pi^2(n-1/2)^2/l^2$, $n=1,2,\ldots$, and the eigenfunctions have the form $u_n = \sin(\pi(n-1/2)x/l)$, $n = 1,2,\ldots$

Example 4. Interval $(0,l)$. Operator $L[u]=u''$. The boundary-value conditions $u(0)=0$, $u'(l)+\beta u(l)=0$ (where $b>0$). Weight $\rho=1$. The eigenvalues λ_n are determined as the solution of the equation tg/ $\sqrt{\lambda} = -\sqrt{\lambda}\,\beta$, and in this case

$$\frac{\pi^2}{l^2}\left(n-\frac{1}{2}\right)^2 < \lambda_n < \frac{\pi^2}{l^2} n^2, \quad n = 1,2,\ldots$$

The eigenfunctions are $u_n = \sin\sqrt{\lambda}_n x$, $n = 1,2,\ldots$

The same procedure is used for examining the case with $u'(0)-\alpha u(0) = 0$, $u(l) = 0$.

Example 5. Interval $(0,l)$. Operator $L[u]=u''$. The boundary-value conditions $u'(0)-\alpha u(0) = u'(l)+\beta u(l) = 0$ (where $\alpha \geq 0$, $\beta \geq 0$, $\alpha + \beta>0$). Weight $\rho = 1$. The eigenvalues λ_n are determined as the solutions of the equation

$$\operatorname{ctg} l\sqrt{\lambda} = \frac{\lambda-\alpha\beta}{\sqrt{\lambda}(\alpha+\beta)},$$

where

$$\frac{\pi^2}{l^2}(n-1)^2 < \lambda_n < \frac{\pi^2}{l^2}n^2, \quad n=1,2,\ldots$$

the eigenfunctions have the following form $u_n = \sin(\sqrt{\lambda}_n x+\varphi_n)$, where $\varphi = \operatorname{arctg}\left(\sqrt{\lambda_n}/\alpha\right)$, $n = 1,2\ldots$

Example 6. Interval $(0,l)$. Operator $L[u]=u''$. The boundary-value conditions are periodic: $u(0) = u(l)$; $u'(0) = u'(l)$. Weight $\rho = 1$. The eigenvalues are $\lambda_0 = 0$, $\lambda_n=\{\pi^2(n+1)^2/l^2$, n is odd; $\pi^2 n^2/l^2$, n is even$\}$ $n=1,2,\ldots$ and the appropriate eigenfunctions have the form u_0=const, $u_n=\{\sin(\pi(n+1)x/l)$, n is odd; $\cos(\pi nx/l$, n is even$\}$, $n = 1,2,\ldots$

More complicated examples lead to eigenfunctions of a special type – the so-called special functions which will be examined in the following section.

3. SPECIAL FUNCTIONS

When solving equations of mathematical physics and boundary-value problems for them it is sometimes not possible to manage without a reserve of standard elementary functions. Every equation generates a class of solutions which are not always elementary functions. Non-elementary functions, encountered when solving the simplest and most important equations, include functions which appear many times and, consequently, there have been sufficiently studied and given specific names. These functions are referred to as *special functions*. Usually, they are eigenfunctions or specific problems of mathematical physics. In most cases, these are functions of a single variable formed in separating variables, for example, the eigenfunctions of the Sturm–Liouville operator L of the type

$$Ly = -(k(x)y')'+q(x)y$$

in some finite or infinite interval. There are many cases in which function $k(x)$ converts to zero at least at one of the ends of this interval. In this section we examine some of the most important special functions [25].

3.1. Spherical functions

A spherical function of the order $k = 1,2,\ldots$ is the restriction to the unit sphere $S^{n-1}\subset\mathbb{R}^n$ of a homogeneous polynomial of the degree k harmonic in $\mathbb{R}^n$. The type of spherical function can be determined by the methods of separation of variables.

We examine a Dirichlet problem for a Laplace equation in a unit sphere R^3:

$$\Delta u = 0 \text{ at } r<1, \; u(r,0,\varphi) = f(\theta,\varphi) \text{ at } r=1.$$

Here (r,θ,φ) are the spherical coordinates in R^3.

We examine the solution of the Laplace equation having the form $u(r,\theta,\varphi)= =R(r)Y(\theta,\varphi)$. To determine R, we use the *Euler equation*

$$r^2R'' + 2rR' - \lambda R = 0,$$

and to determine $Y(\theta,\varphi)$ the following equation

$$\frac{1}{\sin\theta}\frac{\partial}{\partial 0}\left(\sin\theta\frac{\partial\gamma}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2 Y}{\partial\varphi^2}+\lambda Y=0,$$

and the function Y should be bounded at $0 \le \varphi \le 2\pi$, $0 \le \theta \le \pi$ and periodic in respect of φ.

The solution of this problem for function $Y(\theta,\varphi)$ is also found by the method of separation of variables, assuming that $Y(\theta,\varphi) = \Theta(\theta)\Phi(\varphi)$. This leads to the equation:

$$\Phi^{u}+\mu\Phi=0,$$

$$\frac{1}{\sin 0}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d0}\right)+\left(\lambda-\frac{\mu}{\sin^2 0}\right)\Theta=0.$$

From the condition of periodicity of the function $\Phi(\varphi)$ it follows that $\mu = m^2$ and $\Phi(\varphi) = c_1\cos m\varphi + c_2\sin m\varphi$, where $m = 0,1,\ldots$ Function Θ should be bounded at $\theta = 0$ and $\theta = \pi$. Let $\cos\varphi = t$, $\Theta(\theta) = X(t)$. We obtain the equation

$$\frac{d}{dt}\left[(1-t^2)\frac{dX}{dt}\right]+\left(\lambda-\frac{m^2}{1-t^2}\right)X=0, \quad -1\le t\le 1.$$

The solutions of this equation, bounded at $|t| \le 1$, exist only at $\lambda = k(k+1)$, where k is integer, $m = 0, +1,\ldots,+k$, and are referred to as *adjoint Legendre functions* $P_k^{(m)}(t)$. These functions (at every fixed $m = 0,1,\ldots$) form a complete orthogonal system in the segment $\{t:|t|\le 1\}$.

Every solution of the Laplace equation in the sphere of the type $R(r)Y(\theta,\varphi)$ coincides with one of the *internal spherical functions* $u = r^kY_k(\theta,\varphi)$, where $Y_k(\theta,\varphi)$ are *spherical functions*: $Y_k(\theta,\varphi)=\sum_{m=0}^{k}(A_{km}\cos m\varphi + B_{km}\sin m\varphi)P_k^{(m)}(\cos\theta)$. Thus, the spherical functions represent restrictions to the unit sphere of spherical functions.

The solution of the Dirichlet equation has the form $u(r,\theta,\varphi)=\sum_{k=0}^{\infty}r^kY_k(\theta,\varphi)$, where coefficients A_{km} and B_{km} are selected such that $\sum_{k=0}^{\infty}Y_k(\theta,\varphi)=f(\theta,\varphi)$. Functions $Y_k(\theta,\varphi)$ form an orthogonal system on the sphere. Spherical functions in R^n have identical properties.

Writing the Laplace operator in spherical coordinates

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\delta,$$

where δ is the Laplace–Beltrami operator on the sphere:

$$\delta v = \sum_{j-1}^{a-1} \frac{1}{q_j \sin^{n\ \ j-1}\theta_j} \frac{\partial}{\partial \theta_j}\left(\sin^{n-j-1}\theta_j \frac{\partial v}{\partial \theta_j}\right),$$

$$q_1=1; \qquad q_j=(\sin\theta_1 \sin\theta_2 \ldots \sin\theta_{j-1})^2, \qquad j\geq 2,$$

we obtain that the spherical functions $Y_{k,n}(w)$ satisfy the equation $\delta Y_{k,n}(w) = k(k+n-2)Y_{k,n}(w) = 0$. Thus, $Y_{k,n}(w)$ are *eigenfunctions of operator* δ *with the eigenvalue* $\lambda_k = k\,(k+n-2)$. The multiplicity of this eigenvalue is equal to $m_{k,n}$. Being eigenfunctions of the symmetric operator δ, the spherical functions of different orders are orthogonal in $L_2(S^{n-1})$. The system of spherical functions is complete in $L_p(S^{n-1})$ at any p, $1\leq p\leq\infty$.

3.2. Legendre polynomials

The Legendre polynomials are closely linked with the Laplace operator and can be determined by the following procedure.

Let x,y be points in R^3 and θ be the angle between their position vectors. Therefore, $|x-y|^2=|x|^2+|y|^2-2|x||y|\cos\theta$. We set $|x| = r$, $y = r_0$, $\cos\theta = t$. The fundamental solution of Laplace operator in R^3 has the form $-\dfrac{1}{4\pi|x-y|}$. The function $\Psi(\rho,t) = (1+\rho^2-2\rho t)^{-1/2}$, where $0<\rho<1$, $-1\leq t\leq 1$, is referred to as the *generating function of Legendre polynomials*. If this function is expanded into a series in respect of degrees of ρ:

$$\Psi(\rho,t) = \sum_{k=0}^{\infty} P_k(t)\rho^k,$$

then the coefficients $P_k(t)$ are the Legendre polynomials.

The following recurrent equation holds: $(k+1)\,P_{k+1}(t) - t(2k+1)P_k(t) + kP_{k-1}(t) = 0$, $P'_{k-1}(t) - 2tP'_k(t) + P'_{k-1}(t) = P'_k(t)$, which show that $(1-t^2)\,P''_k(t) - 2t\,P'_k(t) + k(k-1)P_k(t) = 0$. Thus, the polynomials $P_k(t)$ are the eigenfunctions in the Sturm–Liouville problem for the operator L:

$$Ly = \frac{d}{dt}\left((1-t^2)\frac{dy}{dt}\right), \quad -1\leq t\leq 1,$$

with the eigenvalues $-k\,(k+1)$. Here, the role of the boundary-value conditions is always played by the condition of finiteness of $y(1)$ and $y(-1)$, i.e. the condition that the solution is bounded at at $t\to 1-0$ and at $t\to -1+0$.

The degree of the polynomial $P_k(t)$ is equal to k at $k = 0,1,\ldots$ Therefore, the polynomials $P_k(t)$ form a complete system on $[-1,1]$. In this case, the equality $\int_{-1}^{1} P_k(t)P_l(t)\,dt = 0$ at $k \neq l$ holds. This shows that the equation

$Ly = \lambda y$ does not have any nontrivial solution bounded on $[-1,1]$ at $\lambda \neq k(k+1)$.

Direct calculation shows that the *Rodrig formula*

$$P_k(t) = \frac{1}{2^k k!}\frac{d^k}{dt^k}[(t^2-1)^k].$$

is valid.

Functions $P_k^m(t) = (1-t^2)^{m/2}\dfrac{d^m}{dt^m}P_k(t)$, $m = 0,1,\ldots, k$, satisfies the Legendre equation $(1-t^2)y''-2ty'+\lambda y=0$ and are bounded at $|t| \leq 1$ if $\lambda = k(k+1)$. Functions $\left\{P_k^{(m)}(t)\right\}$ at $k = m, m+1,\ldots$ form a complete orthogonal system of functions in the segment $[-1,1]$.

3.3. Cylindrical functions

Many problems of mathematical physics lead to an ordinary differential equation

$$x^2y''+xy'+(x^2-n^2)y=0 \tag{6}$$

referred to as the *Bessel equation*. Solutions of this equation are referred to as *cylindrical functions* of the n-th order.

We examine the Bessel equation (6) where $n = \nu$ is an arbitrary real number. Its solution can be found in the form $y = x^\sigma \sum_{j=0}^{\infty} a_j x^j$. Substituting this series into the equation, we obtain:

$$a_{2k} = (-1)^k \frac{a_0}{4^k k1(v+1)(v+2)\ldots(v+k)}, \quad a_{2k-1} = 0, \quad k = 1,2,\ldots$$

If $\nu \neq -m$, where m is the natural number, coefficients a_{2k} are determined for all k. We assume $a_0 = 1/(2^\nu\Gamma(\nu+1))$. Consequently

$$a_{2k} = (-1)^k \frac{1}{2^{2k-v}\Gamma(k+1)\Gamma(k+v+1)}.$$

At $\nu \geq 0$ the series

$$J_v(x) = \sum_{k=0}^{\infty}(-1)^k \frac{1}{\Gamma(k+1)\Gamma(k+v+1)}\left(\frac{x}{2}\right)^{2k-v} \tag{7}$$

converges in $\mathbb{R}^1$ (and even on the entire complex plane). Its sum $J_\nu(x)$ is the *Bessel function of the first kind* and of the ν-th order.

The following functions are encountered most frequently in applications:

$$J_0(x) = 1-\left(\frac{x}{2}\right)^2 + \frac{1}{(21)^2}\left(\frac{x}{2}\right)^4 - \frac{1}{(31)^2}\left(\frac{x}{2}\right)^6 + \ldots,$$

$$J_1(x) = \frac{x}{2} - \frac{1}{(21)}\left(\frac{x}{2}\right)^3 + \frac{1}{(2131)}\left(\frac{x}{2}\right)^5 - \ldots$$

Function $J_\nu(x)$ at $\nu > 0$ has at point $x = 0$ zero of order ν, and function $J_{-\nu}(x)$ a pole or the order ν. At $\nu = 0$, function $J_0(x)$ has a finite value at $x = 0$. Every solution of the Bessel equation at $n = 0$, linearly independent of $J_0(x)$, has a logarithmic singularity at point $x = 0$.

There are other important classes of cylindrical functions. For example, the *Neumann function* or the *cylindrical function of the second kind*, is the solution $N_\nu(x)$ of the Bessel equation for which at $x\to\infty$,

$$N_\nu(x)=\sqrt{\frac{2}{\pi x}}\sin\left(x-\frac{\pi\nu}{2}-\frac{\pi}{4}\right)-O\left(\frac{1}{x\sqrt{x}}\right).$$

The Hankel functions of the first and second type $H_\nu^{(1)}(x)$ and $H_\nu^{(2)}(x)$ are cylindrical functions for which at $x\to\infty$

$$H_\nu^{(1)}(x)=\sqrt{\frac{2}{\pi x}}\exp\left\{i\left(x-\frac{\pi}{2}\nu-\frac{\pi}{4}\right)\right\}+O\left(\frac{1}{x\sqrt{x}}\right),$$

$$H_\nu^{(2)}(x)=\sqrt{\frac{2}{\pi x}}\exp\left\{-i\left(x-\frac{\pi}{2}\nu-\frac{\pi}{4}\right)\right\}+O\left(\frac{1}{x\sqrt{x}}\right).$$

The *Bessel functions of the imaginary argument* play a significant role in mathematical physics. The function $I_\nu(x) = i^{-\nu}J_\nu(ix)$ may be determined as the sum of the series:

$$I_\nu(x)=\sum_{k=0}^{\infty}\frac{1}{\Gamma(k+1)\Gamma(k+\nu+1)}\left(\frac{x}{2}\right)^{2k+\nu}$$

or as the solution of the equation

$$y''+\frac{1}{x}y'-\left(1+\frac{v^2}{x^2}\right)y=0,$$

bounded at $x = 0$ (at $\nu = 0$ the condition $I_\nu(0) = 1$ is imposed).

3.4. Chebyshef, Laguerre and Hermite polynomials

The solution of the equation

$$(1-z^2)\frac{d^2w}{dz^2}-z\frac{dw}{dz}+n^2z=0,\quad -1<z<1,$$

having the form

$$T_n(z)=\cos(n\arccos z),$$
$$U_n(z)=\sin(n\arccos z),$$

are referred to as *Chebyshef polynomials* of the first and second kind.

The Chebyshef polynomials of the first kind can be determined using the generating function:

$$\frac{1-t^2}{1-2tz+t^2}=T_0(z)+2\sum_{n=1}^{\infty}T_u(z)t^n.$$

They satisfy the recurrent relationships

$$T_{n-1}(z)-2zT_n(z)+T_{n-1}(z)=0$$

and the orthogonality relationships

$$\int_{-1}^{1}\frac{T_n(z)T_m(z)}{\sqrt{1-z^2}}\,dz=\begin{cases}0, & \text{if } m\neq n,\\ \dfrac{\pi}{2}, & \text{if } m=n\neq 0,\\ \pi & \text{if } m=n=0\end{cases}$$

In the group of all polynomials of the n-th degree with the leading coefficient 1 the polynomial $2^{1-n}T_n(z)$ is characterised by the fact that its deviation from zero in the segment $[-1,1]$ is the smallest.

We present several first Chebyshef polynomials:

$$T_0(z)=1,\quad T_1(z)=z,\quad T_2(z)=2z^2-1,$$
$$T_3(z)=4z^3-3z,\quad T_4(z)=8z^4-8z^2+1.$$

Polynomials $T_n(z)$ form a complete system in the segment $[-1,1]$.

The polynomial solutions of the differential equation

$$z\frac{d^2w}{dz^2}+(\alpha+1-z)\frac{dw}{dz}+nw=0,$$

where $n = 0,1,2,....,\ \alpha\in C$, are referred to as *Laguerre polynomials*. In particular, the following function corresponds to this equation:

$$L_m^{(\alpha)}(z)=e^z z^{-\alpha}\frac{d^n}{dz^n}(e^{-z}z^{n|\alpha}).$$

For example, at $\alpha = 0$ we obtain the solution

$$L_n(z)=1-C_n^1\frac{z}{1!}+C_n^2\frac{z^2}{2!}-...+(-1)^n\frac{z^u}{n!}.$$

The Laguerre polynomials can be determined using the generating function:

$$e^{-a}(1+t)^{\alpha}=\sum_{n=0}^{\infty}L_n^{(\alpha-n)}(z)\frac{t^n}{n!}.$$

The differential equation

$$\frac{d^2w}{dz^3}-2z\frac{dw}{dz}+2nw=0$$

at $n = 0,1,2,...$determines *Hermite polynomials*

$$H_n(z)=(-1)^n e^{-2}\frac{d^n}{dz^n}(e^{-z^2}).$$

The Hermite polynomials are linked with the Laguerre polynomials by the relationship:

$$H_{2m}(z)=(-1)^m 2^{2m}L_m^{(-1/2)}(z^2),\quad H_{2m+1}(z)=(-1)^m 2^{2m+1}L_m^{(1/2)}(z^2).$$

The generating function for the Hermite polynomials is the function:

$$e^{2zt-t^2}=\sum_{n=0}^{\infty}H_n(z)\frac{t^n}{n!}.$$

Functions

$$D_n(z)=2^{-n/2}e^{-z^2/4}H_n\left(\frac{z}{\sqrt{2}}\right)$$

are the functions of the *parabolic cylinder*. They satisfy the equation

$$\frac{d^2w}{dz^2}+\left(n+\frac{1}{2}-\frac{z^2}{4}\right)w=0.$$

These functions, like the Hermite polynomials, form a complete orthogonal system of functions in R^1.

3.5. Mathieu functions and hypergeometrical functions

Mathieu functions, or the *functions of an elliptical cylinder* are the solutions of the *Mathieu equation*:

$$\frac{1}{4}\frac{d^2w}{dz^2}+(\alpha-4q\cos(2z))w=0.$$

Periodic Mathieu functions are solutions of the equation having the period 2π. In this case, α can be regarded as the eigenvalue of the operator $\frac{1}{4}\frac{d^2}{dz^2}-4q\cos(2z)$ with periodic conditions. For every real q there is an infinite sequence of eigenvalues to which the eigenfunctions $\varphi_n(z,q)$ correspond. These functions are integer in respect of z and form a complete orthogonal system on the segment $[0,2\pi]$.

The confluent hypergeometrical functions are solutions of a degenerate hypergeometrical equation

$$z\frac{d^2w}{dz^2}+(c-z)\frac{dw}{dz}-aw=0.$$

At $c = 2a$ these functions are the Bessel functions, at $c = 1/2$ they are the functions of the parabolic cylinder, and at $a = -n$ they are Laguerre polynomials.

If $c \neq 0, -1, -2, \ldots$, then this equation is satified by the *Kummer function*

$$\Phi(a,c,z)=1+\frac{a}{c}\frac{z}{1!}+\frac{a(a+1)z^2}{c(c+1)2!}+\frac{a(a+1)(a+2)}{c(c+1)(c+2)}\frac{z^3}{3!}+\ldots$$

The hypergeometrical functions are solutions of the hypergeometrical equation

$$z(1-z)\frac{d^2w}{dz^2}+(c-(a+b+1)z)\frac{dw}{dz}-abw=0,$$

where a,b,c are complex parameters. This equation was examined by Euler, Gauss, Rieman, Klein, and many others. It is satisfied by the *Gauss hypergeometrical series*

$$F(a,b,c,z)=1+\sum_{n=1}^{\infty}\frac{(a)_n(b)_n}{(c)_n(1)_n}z^n;$$

converging at $|z|<1$. Here

$$(a)_0=1,\quad (a)_n=\frac{\Gamma(a+n)}{\Gamma(a)}=a(a+1)\ldots(a+n-1),\ \ n=1,2,\ldots$$

The Kummer function is obtained from F by passage to the limit

$$\Phi(a,c,z)=\lim_{b\to\infty}F(a,b,c,z/b).$$

It should be mentioned that $F(n+1,-n,1,\ 1/2-z/2)$ at natural n coincides with the Legendre polynomial. The adjoint Legendre function is obtained from F if $2c = a+b+1$.

4. EIGENFUNCTION METHOD

The eigenfunction method is based on finding solutions of problems of mathematical physics in the form of series in respect of the eigenfunctions of the operators, included in the original problem.

4.1. General scheme of the eigenfunction method

It is assumed that we examine a mathematical physics problem written in the form of a linear (heterogeneous) equation

$$Lu = F, \tag{8}$$

where L is the linear operator with the definition domain $D(L)$, F is the given function, $u \in D(L)$ is the unknown solution.

We examine a problem of eigenvalues $Lu = \lambda u$ and assume that we know the orthonormalized functions $\{u_n\}$ of the operator L, corresponding to the eigenvalues λ_n, $n = 1,2,\ldots$, and λ_n does not convert to zero. Let it be that the right-hand part F of equation (8) is presented in the form of a series (finite or infinite) in respect of these eigenfunctions: $F = \sum F_n u_n$, where F_n are the known coefficients.

The eigenfunction method is based on the following. We require a solution of the problem (8) in the form of a series in respect of the eigenfunctions u_n; $u = \sum_n c_n u_n$. Substituting these series into equation (8) we obtain $\sum_n c_n L u_n = \sum_n F_n u_n$. Since $Lu_n = \lambda_n u_n$, then using the orthogonality $\{u_n\}$ we obtain the relationship $c_n \lambda_n = F_n$. This means that $c_n = F_n/\lambda_n$, and the solution u has the form

$$u = \sum_n \frac{F_n}{\lambda_n} u_n.$$

If in the infinite series we restrict ourselves to N first terms, then the function of the type

$$u^{(N)} = \sum_{n=1}^{N} \frac{F_n}{\lambda_n} u_n$$

is the *approximation of the N-th order to u*.

In the case in which at some $n = m$ the eigenvalue λ_m is zero with multiplicity r and $F_m = 0$, the general solution of equation (8) is represented by the formula

$$u = \sum_{n \neq m} \frac{F_n}{\lambda_n} u_n + \sum_{k=1}^{r} c_k u_{m_k},$$

where c_k are arbitrary constants and u_{m_k}, $k = 1,2,\ldots,r$, are the linear independent eigenfunctions corresponding to λ_m.

Thus, the eigenfunction method makes it possible to present the solution of problem (8) in the form of a series in respect of the eigenfunctions of operator L, involved in the original equation.

The eigenfunction method is used for a wide range of mixed problems of mathematical physics.

4.2. The eigenfunction method for differential equations of mathematical physics

We examine:

1) *The differential equation*

$$\left[\alpha\frac{\partial^2 u}{\partial t^2}+\beta\frac{\partial u}{\partial t}\right]-\frac{L[u]}{\rho(P)}=f(P;t), \tag{9}$$

where L[u] is the elliptical operator of the type described in section 2 given in domain Ω of the space of one, two or three dimensions, P is the point in domain Ω, $\rho(P)$ is the function given in Ω. I_α denotes the interval on the axis t, namely: $(0,\infty)$ if $\alpha \geq 0$ and $(0,T)$ if $\alpha < 0$. Coefficients α and β are given functions of t which do not change their sign on I_α. The right hand part of equation $f(P;t)$ is the function determined for points P from Ω and the value of t from I_α;

2) *The boundary-value condition* at the boundary $\partial\Omega$ of the domain Ω:

$$\Delta[u]=\chi(P;t), \tag{10}$$

where the function $\chi(P;t)$ is given for the point P on $\partial\Omega$ and values t from l_α. Operator $\Lambda\,[u]$ of the type described in section 2, is given on $\partial\Omega$, and its form and coefficients do not depend on t;

3) at $\alpha<0$ *the boundary-value conditions*

$$\Delta_1[u]\big|_{t=0}=\varphi(P),\quad \Delta_2[u]\big|_{t=T}=\psi(P); \tag{11}$$

at $\alpha = 0$ the initial condition

$$u\big|_{t=0}=\varphi(P), \tag{12}$$

and at $\alpha > 0$ the initial conditions

$$u\big|_{t=0}=\varphi(P),\quad \left.\frac{\partial u}{\partial t}\right|_{t=0}=\psi(P), \tag{13}$$

where the functions $\varphi(P)$ and $\psi(P)$ are given and are continuous in Ω, $\Lambda_1[u]$ and $\Lambda_2[u]$ are the one-dimensional operators in respect of variable t with constant coefficients of the type described in section 2 and given in Ω.

The systems of equation (9)–(13) give the following problems:

a) At $\alpha > 0$ equation (9) is *hyperbolic*, the problem is *mixed* with the boundary-value condition (10) and the initial conditions (13);

b) At $\alpha = 0$, $\beta > 0$, equation (9) is *parabolic*, and for this equation we also have a mixed problem, but with one initial condition (12);

c) At $\alpha = \beta = 0$ we have a boundary-value problem for domain Ω with the boundary-value condition (10) on its boundary;

d) At $\alpha < 0$ equation (9) is *elliptical*. This case may be presented when t denotes one of the spatial coordinates, and domain Ω is less than three-dimensional. Domain Ω and interval I_α determine (if the domain Ω is two-dimensional) a cylinder with the generating lines parallel to the axis t, and the bases in the planes $t = 0$ and $t = T$ or (if Ω is one-dimensional) a rectangle. For this elliptical equation we have a boundary-value problem in domain Ω with the boundary-value conditions (10) on the side surface of the cylinder (or the side surfaces of the rectangle) and (11) on its basis.

It is assumed that the problem of the eigenfunctions for domain Ω, with the operator $L[u]$, the weight ρ and the boundary-value condition $\Lambda[v] = 0$ has been solved; λ_k ($k = 1,2,...$) denotes the eigenvalues and v_k (normalized) eigenfunctions.

It is also assumed that $u(P;t)$ is a solution of one of the previously mentioned problems. We expand this solution into a Fourier series in respect of functions $\{v_k\}$. The coefficients of this function are evidently functions of t, and they will be denoted by $w_k(t)$, since

$$u(P;t) = \sum_{k=2}^{\infty} w_k(t) v_k(P), \quad w_k(t) = \int_\Omega u(P;t) v_k(P) \rho(P) d\mu.$$

To determine coefficients $w_k(t)$, equation (9) is multiplied by the function $v_k(P)\rho(P)$ and the resultant equality is integrated over domain Ω. Consequently, we obtain

$$\alpha w_k^u + \beta w_k^{'} + \lambda_k w_k = a_k(t) + X_k(t), \tag{14}$$

where

$$a_k(t) = \int_\Omega f v_k \rho d\mu,$$

$$X_k(t) = \begin{cases} -\int_{\partial\Omega} v_k \chi d\sigma & \text{for the boundary condition of the third kind,} \\ \int_{\partial\Omega} \frac{\partial_{\gamma k}}{\partial n} \chi d\sigma & \text{for the boundary condition of the first kind.} \end{cases}$$

Thus, we have obtained an *ordinary differential equation for the k-th Fourier coefficient* w_k of the unknown function u. This differential equation is a consequence of the original differential equation in partial derivatives (9) and the boundary-value condition (10). We examine the fulfilment of additional conditions of the problem in each of the four cases combined in our scheme.

Cases a) and b). If $\alpha > 0$ or $\alpha = 0$ and $\beta > 0$, i.e. for a mixed problem, the initial conditions (12) and (13) give the initial conditions for w_k. In fact

$$w_k(0) = \int_\Omega u(P;0) v_k \rho \, d\mu = \int_\Omega \varphi v_k \rho \, d\mu = b_k,$$

$$\left.\frac{dw_k}{dt}\right|_{l\,0} = \int_{\Omega}\left.\frac{\partial u(p,0)}{\partial t}\right|_{l\,0} v_k \rho d\mu = \int_{\Omega}\psi v_k \rho d\mu = c_k,$$

where b_k denotes the Fourier coefficient of the function φ, and c_k of the function ψ (if α > 0). At α > 0 equation (14) and two initial conditions determine the unique solution. At α = 0, β > 0, equation (14) of the first order and one (first) condition are sufficient for unique determination of w_k.

Case c) At α = β = 0, equation (14) is not differential and, consequently, it is therefore obvious that there are no additional conditions. We examine in greater detail equation (14) in this case. If none of λ_k is equal to zero, then the equation is evidently solvable at any k and $w_k = \frac{1}{\lambda}(a_k + X_k)$. Coefficients w_k are constant and the Fourier series gives the required solution. However, if one of the values of λ_k, for example λ_j, is equal to zero, the equation (14) may have a solution at $k = j$ only on the condition that $a_j + X_j = 0$. If this condition is fulfilled, then w_k is arbitrary and, consequently, the Fourier series gives an infinite number of solutions differing from each other by eigenfunctions corresponding to zero eigenvalues.

Thus, the boundary-value problem is not always correct. In particular, the problem is correct if zero is not the eigenvalue of the operator $L[u]$ for the examined boundary-value conditions and weight. Otherwise, i.e. if $\lambda_j = 0$, the boundary-value problem has a solution and, in this case an infinite set, only for the right-hand sides of equation (9) and the boundary-value condition (10) for which $a_j + X_j = 0$. The following conclusions can be drawn regarding the operator $L[u] \equiv \Delta u$. Operator Δu has a zero eigenvalue only at the boundary-value condition $\frac{\partial u}{\partial n} = 0$. In this last case, the corresponding eigenfunction is constant. Therefore, the boundary-value problem for the Laplace operator is correct for any boundary-value conditions, with the exception of $\frac{\partial u}{\partial n} = 0$ at the boundary. Consequently, with the exception of this case, its unique solution at any right-hand side of equation $-\Delta u = f$ at the boundary-value condition (10) may be written in the form of a series $\sum_{k=1}^{\infty}\frac{1}{\lambda_k} v_k (a_k + X_k)$, where v_k is the eigenfunction of the problem, and a_k are the Fourier coefficients of function f, and X_k were determined previously. In the case of the boundary-value condition $\frac{\partial u}{\partial u} = \chi$ the problem is ill-posed. It is solvable only if the function f and χ satisfy the relationship

$$\int_{\Omega} f\rho d\mu + \int_{\partial\Omega}\chi d\sigma = 0.$$

In this case, the solution is determined by the equation

$$u = \sum_{k=1}^{\infty}\frac{1}{\lambda_k} z_k (a_k + X_k) + c,$$

where v_k are the eigenfunctions of the problem, different from the constant, and c is an arbitrary constant.

Case d). At $\alpha > 0$, equation (14) is examined in the interval $(0,T)$ and at the ends of this interval we obtain from (11) the boundary-value conditions for w_k. Infact,

$$\Delta_1[w]\big|_{t=0} = \left\{ \Delta_1 \left[\int_\Omega u(P;d) u_k \rho d\mu \right] \right\}_{t=0} = \int_\Omega \Delta_1[u]\big|_{t=0} v_k \rho d\mu = b_k,$$

$$\Delta_2[w]\big|_{t=T} = \int_\Omega \Delta_2[u]\big|_{t=T} v_k \rho d\mu = \int_\Omega \psi v_k \rho d\mu = c_k.$$

Thus, for w_k we have a boundary-value problem in the interval $(0,T)$. To examine this problem, it is convenient to represent the first two terms $\alpha \frac{d^2 w}{dt^2} + \beta \frac{dw}{dt}$ in the form $\tilde{L}[u] = \frac{d}{dt}\left(\tilde{p} \frac{du}{dt} \right)$. For this purpose, it is evidently sufficient to multiply (14) by $-\tilde{\rho} = \frac{1}{\alpha} \exp\left(\int \frac{\beta}{\alpha} dt \right)$ and set $\tilde{\rho} = \frac{1}{\alpha} \exp\left(\int \frac{\beta}{\alpha} dt \right)$. Consequently, equation has the form $\tilde{L}[u] - \lambda_k \tilde{\rho} u = -\tilde{\rho}(a_k + X_k)$. On the basis of what was said about the problem c), it may be noted that the problem of finding w_k is uniquely solvable for all k at any a_k, X_k, b_k and c_k if zero is not the eigenvalue for the operator $\tilde{L}[w] + \lambda_k \tilde{\rho} w$, or, which is the same, no $-\lambda_k$ is the eigenvalue for the operator $\tilde{L}[w]$ at the weight $\tilde{\rho}$ and the boundary conditions $\Lambda_1[w]|_{t=0} = \Lambda_2[w]|_{t=T} = 0$. If this condition is fulfilled, the examined problem is correct.

It should also be mentioned that all eigenvalues of the operator $\tilde{L}$ are non-negative, and the lowest eigenvalue may be equal to zero only in the case $\Lambda_1[w] = \Lambda_2[w] \equiv \frac{\partial w}{\partial t}$.

4.3. Solution of problems with nonhomogeneous boundary conditions

In the case of a nonhomogeneus problem it is not possible to use the operators $L[u]$ term by term in respect of the Fourier series. This can be carried out if u satisfies the homogeneous boundary-value conditions. Therefore, the problems with nonhomogeneus boundary-value conditions can sometimes be solved efficiently by different methods. We shall describe some of them.

The boundary-value problems. A further two methods can be used in this case.

1°. Select the coordinate system in such a manner that domain Ω converts to a rectangle or a parallelepiped. The problem can be split into several problems in such a manner that in each of these problems the heterogeneity of the boundary-value condition remains only in one pair of the opposite

sides or faces. Subsequently, in each of these problems we use the method of solving the problem c) in the present section (if this is possible).

2°. Find some function u_0 twice differentiable in Ω and satisfying the boundary-value condition, and introduce the new unknown function $\bar{u}$ setting $u = u_0 + \bar{u}$. For $\bar{u}$ we obtain the identical boundary-value problem, but already with the homogeneous boundary-value condition (and with the other right-hand side of the equation).

Mixed problems. Here we have the following possibilities.

1°. As in the method 2° of solving purely boundary-value problems, described previously, we find some function u_0 satisfying the boundary-value condition. We make the substitution $u = u_0 + \bar{u}$. For function $\bar{u}$ we obtain a homogeneous boundary-value condition and the changed right-hand sides of the equation and the initial conditions.

2°. To solve in advance the boundary-value problem which is obtained if the terms with the derivative in respect of the time and the initial conditions are rejected in the equation. The solution of this problem is denoted by u_4 and we use the method described previously. This method may be used if the solution u_0, which depends on t as a parameter when χ and f depend on t, is twice differentiable in respect of t.

5. EIGENFUNCTION METHOD FOR PROBLEMS OF THE THEORY OF ELECTROMAGNETIC PHENOMENA

5.1. The problem of a bounded telegraph line

We examine a bounded telegraph line with length l with the distributed constants C, L, R, G. The origin of the coordinates is placed in the left hand of the line. The right hand of the line is earthed, and the left end at the moment $t = 0$ is connected to the given EMF V. At $t = 0$ the line contains no current or voltage. Voltage u in this line satisfies the equation

$$\alpha^2 \frac{\partial^2 u}{\partial t^2} - 2\beta \frac{\partial u}{\partial t} + \gamma^2 u - \frac{\partial^2 u}{\partial x^2} = 0,$$

where $\alpha^2 = CL$, $\beta = \dfrac{CR + LG}{2}$, $\gamma^2 = RG$; the initial conditions

$$u(x;0) = \frac{\partial u(x,0)}{\partial t} = 0$$

and the boundary-value conditions $u(l;t) = 0$, $u(0;t) = V$. The equation of the corresponding problem of the eigenvalues has the form $v'' - \gamma^2 v + \lambda v = 0$, and the boundary-value conditions have the form $v(0) = v(l) = 0$. This problem of the eigenvalues is reduced to the problem solved in section 2 if $\lambda - \gamma^2$ is denoted by μ. Thus, $\mu_n = (n^2\pi^2/l^2)$ and, consequently, $\lambda_n = \gamma^2 + (n^2\pi^2)/l^2$ and $v_n = \sin\ (n\pi x/l)$.

We find the solution in the form of a Fourier series $u(x;t) = \sum_{n=1}^{\infty} w_n\ (t)\sin(n\pi x/l)$. Multiplying the equation by $(2/l)\sin(n\pi x/l)$ and

integrating with respect to x in the range from 0 to l, we obtain the equation:

$$\alpha^2 w_n^u + 2\beta w_n^{'} + \left(\gamma^2 + \frac{\pi^2 n^2}{l^2}\right) w_n = \frac{2n\pi}{l^2} V$$

with the initial conditions $w_n(0) = w_n^{'}(0) = 0$. Integrating this equation, we determine w_n and the solution u of our problem.

For definiteness, we examine the case V = const. The characteristic equation has the form $\alpha^2 r^2 + 2\beta r + (\gamma^2 + (\pi^2 n^2)/l^2) = 0$. Its roots are $r_{1,2} = -\lambda \pm i w_n$, where $\lambda = \beta/\alpha^2 = (1/2)((R/L) + (G/C))$ and

$$\omega_n = \frac{1}{2}\sqrt{\frac{4\pi^2 n^2}{l^2 CL} - \left(\frac{R}{L} - \frac{G}{C}\right)^2}.$$

If $(R/L)-(G/C) = 0$, i.e. the line 'without distortion', then $\omega_n = (\pi n)/(l\sqrt{LC})$. The general solution of the equation is:

$$w_n(t) = (C_n^{'} \cos(\omega_n t) + C_n^{''} \sin(\omega_n t)) e^{-\lambda r} + \frac{2n\pi}{n^2\pi^2 + l^2\gamma^2} v.$$

The initial conditions give:

$$C_n^{'} = -\frac{2n\pi}{n^2\pi^2 + l^2\gamma^2} V, \quad C_n^u = -\frac{2n\pi\lambda}{\omega_n (n^2\pi^2 + l^2\gamma^2)} V,$$

and consequently

$$w_u(t) = \frac{2n\pi}{n^2\pi^2 + l^2\gamma^2}\left\{1 - e^{-\lambda t}(\cos(\omega_n t) + \frac{\lambda}{\omega_n}\sin(\omega_n t))\right\} V.$$

Thus, the solution of our problem is given by the series

$$u(x,t) = V\sum_{n=1}^{\infty} \frac{2n\pi}{n^2\pi^2 + l^2\gamma^2}\left\{1 - e^{-\lambda t}(\cos(\omega_n t) + \frac{\lambda}{\omega_n}\sin(\omega_n t))\right\} \sin\frac{n\pi x}{l}.$$

It should be mentioned that the first values of n (at $n < \frac{\sqrt{LCl}}{2\pi}\left|\frac{R}{L} - \frac{G}{C}\right|$) frequencies ω_n may prove to be imaginary. Consequently, $\sin\omega_n t$ is also imaginary. Denoting ω_n for such n by $i\xi$, we obtain

$$\cos\omega_n t = \mathrm{ch}\xi_n l, \quad \frac{\sin\omega_n l}{\omega_n} = \frac{\mathrm{sh}\xi_{nt}}{\xi_n}$$

and

$$w_n(t) = \frac{2n\pi V}{n^2\pi^2 + l^2\lambda^2}\left\{1 - e^{-\lambda t}\left(\mathrm{ch}\xi_u t + \lambda\frac{\mathrm{sh}\xi_n t}{\xi_n}\right)\right\},$$

and the corresponding terms of the Fourier series has the form

$$\frac{2n\pi V}{n^2\pi^2 + l^2\lambda^2}\left\{1 - e^{-\lambda r}\left(\mathrm{ch}\xi_n t + \lambda\frac{\mathrm{sh}\xi_n t}{\xi_n}\right)\right\}\sin\frac{n\pi x}{l}.$$

5.2. Electrostatic field inside an infinite prism

We examine *a right angle infinite prism where each face is a conducting electrode and the faces are insulated from each other*. The potentials of the faces are denoted by u_1, u_2, u_3, u_4. The potential of the field u does not depend on z. This potential satisfies the Laplace equation $\Delta u = 0$ and the boundary-value conditions

$$u\big|_{x=0} = u_1, \quad u\big|_{y=0} = u_2, \quad u\big|_{x=a} = u_3, \quad u\big|_{x=b} = u_4.$$

The corresponding problem of the eigenvalues has the form

$$\Delta v + \lambda v = 0, \quad v\big|_{x=0} = v\big|_{y=0} = v\big|_{x=a} = v\big|_{y=b} = 0.$$

The eigenvalues of each problem are determined from the equations

$$\lambda_{m,n} = \frac{\pi^2 m^2}{a^2} + \frac{\pi^2 n^2}{b^2},$$

and the eigenfunctions have the form

$$v_{m,n} = \sin\frac{\pi m x}{a}\sin\frac{\pi n y}{b}.$$

We obtain the solution in the form of a Fourier series

$$u = \sum_{m,n=1}^{\infty} w_{m,n}\sin\frac{\pi m x}{a}\sin\frac{\pi n y}{b}.$$

Multiplying the equation $\Delta u = 0$ by $(4/ab)\sin(\pi m x/a)\sin(\pi n y/b)$, integrating over the rectangle Ω, we obtain the expression for $w_{m,n}$:

$$w_{m,n} = \frac{4}{nm\left[\dfrac{\pi^2 m^2}{a^2} + \dfrac{\pi^2 n^2}{b^2}\right]}\left\{\frac{m^2}{a^2}[1+(-1)^{n+1}][u_1 + (-1)^{m+1}u_3] + \right.$$

$$\left. + \frac{n^2}{b^2}[1+(-1)^{m+1}][u_2 + (-1)^{n+1}u_4]\right\}.$$

5.3. Problem of the electrostatic field inside a cylinder

We examine *an infinite circular cylinder with radius R. The potential of the wall of the cylinder is given and is constant along every generating line of the cylinder.*

The axis z is directed along the cylinder axis. Evidently, the potential u does not depend on z, and we again have a plane problem. The circle, produced in the cross-section of the cylinder by the plane (x,y), is denoted as Ω, and its contour $\partial\Omega$. Potential u satisfies the equation $\Delta u = 0$ and the boundary-value condition $u|_{\partial\Omega} = f$, where f is a given function on the circumference $\partial\Omega$.

We introduce the polar coordinates r and φ. Domain Ω is a coordinate quadrangle in variables r,φ. The equation has the form

$$r\frac{\partial}{\partial}r\frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial\varphi^2} = 0.$$

On two sides of the quadrangle we have the boundary-value conditions for

$$u\big|_{\varphi=0} = u\big|_{\varphi=2\pi}, \quad \left.\frac{\partial u}{\partial \varphi}\right|_{\varphi=0} = \left.\frac{\partial u}{\partial \varphi}\right|_{\varphi=2\pi}.$$

In addition to this, in the vicinity of the origin of the coordinates, i.e. in the vicinity of $r = 0$, function u should remain bounded.

We examine a system of eigenfunctions of the operator Δv in the circle Ω with the boundary-value conditions $v|_{\partial\Omega} = 0$. The eigenfunctions have the form

$$v_{0,n} = \frac{1}{2} J_0\left(\kappa_n^0 \frac{r}{R}\right) \quad (n = 1, 2, \ldots),$$

$$v_{2m-1,n} = J_m\left(\kappa_n^m \frac{r}{R}\right)\sin m\varphi, \quad v_{2m,n} = J_m\left(\kappa_n^m \frac{r}{R}\right)\cos m\varphi$$

$$(n = 1, 2, 3, \ldots, \quad m = 1, 2, 3, \ldots).$$

where j_m is Bessel function, κ_n^m is the n-th root of the equation $J_m(x) = 0$, and the corresponding eigenvalue is equal to $\lambda_{2m-1,n} = \lambda_{2m,n} = (\kappa_n^m)^2/R^2$. We obtain the solution of our problem in the form of a Fourier series in respect of the functions of this system: $u = \sum_{k=0,\, n=1}^{\infty} w_{k,n} v_{k,n}$. To determine the coefficients $w_{k,n}$ the equation is multiplied by $v_{k,n}$ and integrated over Ω. We obtain

$$w_{0,n} = -\frac{2a_0}{\kappa_n^0 J_0'(\chi_n^0)}, \quad w_{2m-1,n} = -\frac{2a_{2m-1}}{\kappa_n^m J_m(\kappa_n^m)}, \quad w_{2m,n} = -\frac{2a_{2m}}{\kappa_n^m J_m(\kappa_n^m)},$$

where a_0, a_{2m-1}, a_{2m} are the Fourier coefficients of function f in the system 1/2, sin $m\varphi$, cos $m\varphi$. Therefore, the solution of the given problem is the function u written in the form of a series

$$\sum_{n=1}^{\infty}\left[\frac{-a_0}{\kappa_n^0 J_0'(\kappa_n^0)} J_0\left(\kappa_n^0 \frac{r}{R}\right) - 2\sum_{m=1}^{\infty}(a_{2m-1}\sin mx + a_{2m}\cos mx)\frac{J_m\left(\kappa_n^m \frac{r}{R}\right)}{\kappa_n^m J_m'(\kappa_n^m)}\right].$$

5.4. The field inside a ball at a given potential on its surface

We examine the following problem: *to determine a function satisfying inside a ball with radius R the Laplace equation and having the given value f on the sphere* (i.e determining the electrostatic field inside the ball, if the potential on its surface is given).

Placing the origin of the coordinates in the centre of the sphere and introducing spherical coordinates, the Laplace equation is transformed to the form

$$\frac{\partial}{\partial r}\left(r^2 \frac{\partial u}{\partial r}\right) - \frac{1}{\sin\theta}\left[\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u}{\partial \theta}\right) + \frac{1}{\sin\theta}\frac{\partial^2 u}{\partial \varphi^2}\right] = 0.$$

At r = const function u is given on the sphere with radius r and, conse-

quently, it can be expanded in respect of the spherical functions into a series whose coefficients depend on r:

$$u=\sum_{n=0}^{\infty}\left\{\frac{1}{2}w_{n.0}P_n(\cos\theta)+\sum_{k=1}^{\infty}w_{n,2k-1}\sin k\varphi+w_{n.2k}\cos k\varphi]P_{n,k}(\cos\theta)\right\}.$$

Initially, we determine coefficients $w_{n,0}$. Multiplying the initial equation by $((2n+1)/(2\pi))P_n(\cos\theta)\sin\theta$ and integrating with respect to θ from zero to π and with respect to φ from 0 to 2π. Consequently, we obtain that

$$w_{n,0}=a_{n,0}\left(\frac{r}{R}\right)^n=\frac{2n+1}{2\pi}\left(\frac{r}{R}\right)^n\int_0^{2\pi}\int_0^{\pi}f\,P_n(\cos\theta)\sin\theta\,d\theta\,d\varphi.$$

Knowing these coefficients we can determine the value u at the points of the positive half axis z (i.e. at $\theta = 0$). Actually, since $P_n(\cos 0) = 1$ and at $k > 0$, $P_{n,k}(\cos 0) = 0$, then $u=\sum_{n=0}^{\infty}\frac{a_{n,0}}{2}\left(\frac{r}{R}\right)^n$.

It is now required to calculate u at an arbitrary point (r,θ,φ). We draw a beam from the centre through this point and its axis z_1 will be accepted as the new coordinate system. Consequently,

$$u(r,\theta,\varphi)=\sum_{n=0}^{\infty}\frac{b_{n,0}}{2}\left(\frac{r}{R}\right)^n,$$

$$b_{n,0}=\frac{2n+1}{2\pi}\int_0^{2\pi}\int_0^{\pi}f(\theta_1,\varphi_1)P_n(\cos\gamma)\sin\theta_1 d\theta_1 d\varphi_1,$$

where γ is the angle between the beams directed to the point $(r,\theta,\varphi,)$ and the variable point (r_1,θ_1,φ_1). This angle is the latitude in our new coordinate system. For $b_{n,0}$ we obtain

$$b_{n,0}=a_{n,0}P_n(\cos\theta)+2\sum_{k=1}^{n}[a_{n.2k}\cos k\varphi+a_{n,2k-1}\sin k\theta]P_{n,k}(\cos\theta),$$

where $a_{n,i}$ are the Fourier coefficients of expansion of function f in respect of the basic spherical functions. Consequently, the equation for $u(r,\theta,\varphi)$, in the form

$$u=\sum_{n=0}^{\infty}\left\{\frac{1}{2}a_{n,0}P_n(\cos\theta)+\sum_{k=1}^{n}[a_{n,2k-1}\sin k\varphi+a_{n,2k}\cos k\varphi]P_{n,k}(\cos\theta)\right\}\left(\frac{r}{R}\right)^n,$$

gives the solution of the problem. It should be mentioned that the terms of the sum are the main internal spherical functions.

This result may be formulated as follows: *the function u harmonic inside the ball with radius R and continuous up to its surface may be expanded into a series in respect of the main spherical internal functions. The coefficients of this expansion are the Fourier coefficients of the value of the expanded function of the surface of the ball, divided by* R^n.

5.5 The field of a charge induced on a ball

We examine an *electrostatic field formed by charge* ε, *placed at distance a from the centre of the conducting ball with the radius R* ($a > R$), also the density of the charges induced on the surface of this sphere.

The origin of the coordinates is placed in the centre of the call and the axis z is directed through point P at which the charge is located.

The potential u of the field may be represented in the form of a sum of two potentials $u = u_1+u_2$, where u_1 is the potential of the charge ε, placed at point P, and u_2 is the potential of the charges induced on the surface of the ball.

Potential u_1 inside the sphere with radius a can be represented by the series

$$u_1 = \varepsilon \sum_{n=0}^{\infty} \frac{r^n}{a^{n+1}} P_n(\cos\theta),$$

and outside by the series

$$u_1 = \varepsilon \sum_{n=0}^{\infty} \frac{a^n}{r^{n-1}} P_n(\cos\theta),$$

where P_n are Legendre polynomials.

As regards u_2, then (because of the axial symmetry of the field) it can be represented inside the sphere with radius R by the series

$$u_2 = \sum_{n=0}^{\infty} \frac{1}{2} a_{n,0} \left(\frac{r}{R}\right)^n P_n(\cos\theta),$$

and outside it by the series

$$u_2 = \sum_{n=0}^{\infty} \frac{1}{2} a_{n,0} \left(\frac{R}{r}\right)^{n+1} P_n(\cos\theta),$$

where coefficients $a_{n,0}$ are to be determined. However, inside the ball with radius R the total potential $u = u_1 + u_2$ is constant (u = const) (the ball is conducting). Therefore, we obtain $(1/2)a_{n,0} = -\varepsilon(R^n/a^{n+1})(n > 0)$, $(1/2)a_{0,0} = C-(\varepsilon/a)$, where C is still not determined. Substituting the coefficient into the expression for u_2, we obtain u_2 outside the ball with radius R. Therefore, in the layer between the ball with the radii R and $a(R < r < a)$ we have

$$u = \left[\left(C - \frac{\varepsilon}{a}\right)\frac{R}{r} + \frac{\varepsilon}{a}\right] + \varepsilon \sum_{n=1}^{\infty} [r^{2n+1} - R^{2n-1}] \frac{P_n(\cos\theta)}{(ar)^{n+1}},$$

and outside the sphere with radius a

$$u = \left(CR - \frac{\varepsilon R}{a} + \varepsilon\right)\frac{1}{r} + \varepsilon \sum_{n=1}^{\infty} [a^{2n+1} - R^{2n+1}] \frac{P_n(\cos\theta)}{(ar)^{n-1}}.$$

This gives $\lim_{r\to\infty}(ur) = CR - \frac{\varepsilon R}{a} + \varepsilon.$ It is well-known that this limit is equal to the total charge ε present in the field. Therefore, C = ε/a, and, consequently, the required density ρ of the charge on the sphere has the form

$$\rho=\frac{\varepsilon}{4\pi aR}-\frac{\varepsilon}{4\pi R}-\frac{a^2-R^2}{(R^2-2aR\cos\theta+a^2)^{3/2}}.$$

6. EIGENFUNCTION METHOD FOR HEAT CONDUCTIVITY PROBLEMS

6.1. Heat conductivity in a bounded bar

We examine a homogeneous cylindrical bar with length l, where one end of the bar x = 0 is maintained at temperature $u_0 = c(t)$, *and the other one x = l is cooled down in accordance with the Newton law*. The temperature of the external medium is set to equal zero. Temperature $u(x;t)$ in the section x of the bar at the moment of time t satisfies the equation

$$\frac{\partial u}{\partial t}-a^2\frac{\partial^2 u}{\partial x^2}+b^2u=0,$$

where $a^2 = k/(c\rho)$, $b^2 = (h\rho)/(c\rho q)$. It is assumed that the initial temperature in the bar is $\varphi(x)$, where $\varphi(x)$ is the given function of x. Consequently, we have the initial condition $u|_{t=0}= \varphi(x)$ and the boundary-value conditions

$$u\big|_{x=0}=\chi(t),\quad k\frac{\partial u}{\partial x}\bigg|_{x-l}=-hu\big|_{x=l}.$$

The corresponding problem of the eigenvalue has the form $v''-(b^2/a^2)v+\lambda v=0$ with the boundary-value conditions $v\big|_{x=0}=0,\ \left[\frac{\partial v}{\partial x}+\gamma v\right]\bigg|_{x=l}=0,$ where $\gamma= h/k$. According to section 2, the eigenfunctions of the problem are: $v_n = \sin \mu_n x$, where μ_n is determined from the equation $\mathrm{tg}/\mu = -\mu/\gamma$. We find the solution of our problem in the form of the series $u=\sum_{n=1}^{\infty}w_n(t)\sin\mu_n x.$ The equation is multiplied by $\sin \mu_n x$, and integrated from 0 to l. Consequently, to determine w_n we obtain the differential equation

$$w_n'+(b^2+a^2\mu_n^2)w_n=\frac{\mu_n a^2}{N_n^2}\chi(t)$$

and the initial condition

$$w_n(0)=\frac{1}{N_n^2}\int_0^l\varphi(x)\sin\mu_n x dx=b_n.$$

For definition, it is assumed that $\chi(t) = U_0 = \text{const}$, at $\varphi(x) = 0$. Consequently $b_n = 0$ and

$$w_n=U_0\frac{a^3\mu_n}{N_n^2(a^2\mu_a^2+b^2)}[1-\exp(-(b^2+a^2\mu_n^2)t)].$$

The solution of our problem is obtained in the form

$$U = U_0 \sum_{n=1}^{\infty} \frac{a^2 \mu_n}{N_n^2 (a^2 \mu_n^2 + b^2)} [1 - \exp(-(b^2 + a^2 \mu_n^2)t)] \sin \mu_n x.$$

6.2. Stationary distribution of temperature in an infinite prism

L*et it be that the heat sources with density Q are distributed uniformly along the line (l) passing inside the prism parallel to its edges.* The temperature of the outer space is assumed to be equal to zero, and the heat conductivity coefficient k is constant.

The axis z is directed along one of the edges of the prism, and the axis x, y along its faces. a and b denote the dimensions of the cross-section Ω of the prism and x_0 and y_0 are the coordinates of the line (l). The heat sources are assumed to be distributed in the infinite prism with cross-section $\partial\Omega$ with the small diameter δ, containing the straight line (l). The density of distribution of the heat sources in the examined prism is denoted by $q(x,y)$. In this case, $q(x,y)=0$ outside $\partial\Omega$ and $\iint_{\partial\Omega} q(x,y)dxdy = Q.$ Evidently, the temperature u of the prism is distributed in the same manner in any cross-section, i.e. u does not depend on z. Thus, this is a two-dimensional problem for the equation $k\Delta n = -q$ with the boundary condition of the type $k\dfrac{\partial u}{\partial n} - hu = 0,$ which can be written in the form: $\Delta u = -q/k$ and $\dfrac{\partial u}{\partial n} - \gamma u = 0,$ where $\lambda = h/k$.

The appropriate problem of the eigenvalues with the equation $\Delta v + \lambda v = 0$ was solved in section 2. In accordance with the results, the eigenvalues are the numbers $\lambda_{m,n} = \mu_m + \nu_n$, where μ_m and ν_n are the roots of the equation

$$\frac{\mu_m - \gamma^2}{2\gamma\sqrt{\mu_m}} = \operatorname{ctg} a\sqrt{\mu_m}, \quad \frac{\nu_n - \gamma^2}{2\gamma\sqrt{G\nu_n}} = \operatorname{ctg} b\sqrt{\nu_n}.$$

The eigenfunctions are $v_{m,n} = \sin(x\sqrt{\mu_m} + \varphi_m)\sin(y\sqrt{\nu_n} + \psi_n)$, $\varphi_m = \operatorname{arctg}(\gamma/\mu_m)$, $\psi_n = \operatorname{arctg}(\gamma/\nu_n)$. The Fourier coefficients $c_{m,n}$ of the function q are calculated from the system $\nu_{m,n}$:

$$c_{m,n} = \frac{4Q}{\left[a + \dfrac{2\gamma}{\gamma^2 + \mu_m}\right]\left[b + \dfrac{2\gamma}{\gamma^2 + \nu_n}\right]} \sin(x_0\sqrt{\mu_m} + \varphi_m)\sin(y_0\sqrt{\nu_n} + \psi_n),$$

We now find the solution of our problem in the form of a series in respect of the function $\nu_{m,n}$:

$$u(x,y) = \sum_{m,n=1}^{\infty} w_{m,n} \sin(x\sqrt{\mu_m} + \varphi_m)\sin(y\sqrt{\nu_n} + \psi_n).$$

To determine the coefficient $w_{m,n}$ the initial equations are multiplied by $v_{m,n}$, and integrated over rectangle Ω. Consequently, the expression for $w_{m,n}$ has the form

$$w_{m,n} = \frac{4Q\sin(x_0\sqrt{\mu_m}+\varphi_m)\sin(y_0\sqrt{\nu_0}+\psi_n)}{k(\mu_m+\nu_n)\left[a+\dfrac{2\gamma}{\gamma^2-\mu_m}\right]\left[b+\dfrac{2\gamma}{\gamma^2+\nu_n}\right]}.$$

6.3. Temperature distribution of a homogeneous cylinder

We examine a homogeneous cylinder with radius R and height H. The *coefficient of external heat conductivity at one of the bases of the cylinder is* h_1, *and on the side surface it is* h_2, *where* h_1, h_2 *are constant*. The second base is held at temperature U_0. The temperature of the external medium is equal to zero. There are no heat sources inside the cylinder. Temperature u inside the cylinder is determined by the equation $\Delta u = 0$ and the boundary-value conditions:

$$u\big|_{z=0} = U_0, \quad \frac{\partial u}{\partial z}+\beta u\big|_{z=H} = 0, \quad \frac{\partial u}{\partial r}+\gamma u\big|_{r=R} = 0,$$

where $\beta = h_1/k$, $\gamma = h_2/k$. The eigenfunctions and eigenvalues of the Laplace operator in the cylinder at the boundary-value conditions $v\big|_{z=0} = 0$, $\dfrac{\partial y}{\partial z}+ +\beta v\big|_{z=H} = 0$, $\dfrac{\partial v}{\partial r}+\gamma v\big|_{r=R} = 0$ have the following form

$$v_{0,m,n} = \frac{1}{2}J_0\left(\kappa_m^0\frac{r}{R}\right)\sin\sqrt{\nu_n z}, \quad v_{2k-1,m,n} = J_k\left(\kappa_m^k\frac{r}{R}\right)\sin k\varphi\sin\sqrt{\nu_n z},$$

$$v_{2k,m,n} = J_k\left(\kappa_m^k\frac{r}{R}\right)\cos k\varphi\sin\sqrt{\nu_n z},$$

where κ_m^k and ν_n are respectively m-th and n-th roots of the equations $\kappa J_k'(\kappa) = -\gamma R J_k(\kappa)$ and $\operatorname{tg} H\sqrt{\nu} = -\sqrt{\nu}/\beta$. The appropriate eigenvalues are:

$$\lambda_{2k-1,m,n} = \lambda_{2k,m,n} = \frac{(\kappa_m^n)^2}{R^2}+\nu_n.$$

We determine U in the form

$$U = \sum_{m,n=1}^{\infty}\left\{\frac{w_{0,m,n}}{2}J_0\left(\kappa_m^0\frac{r}{R}\right)+\right.$$

$$\left.-\sum_{k=1}^{\infty}J_k\left(\kappa_m^k\frac{r}{R}\right)(w_{2k-1,m,n}\sin k\varphi+w_{2k,m,n}\cos k\varphi)\right\}\sin\sqrt{\nu_n z}.$$

To determine $w_{i,m,n}$ we multiply the initial equation by $v_{i,m,n}$ and integrate over the volume of the cylinder. We obtain

$$w_{i,m,n} = 0, \quad i > 0,$$

$$w_{0,m,n} = \frac{8U_0\gamma R\sqrt{\nu_n}}{(\kappa_m^0)^2 J_0(\kappa_m^0)\left[1+\frac{R^2\lambda^2}{(\kappa_m^0)^2}\right]\left[H-\frac{\beta}{\beta^2+\nu_n}\right]\left[\frac{(\kappa_m^0)^2}{R^2}+\nu_n\right]}.$$

Consequently,

$$u = \frac{1}{2}\sum_{m,n,1}^{\infty} w_{0,m,n}\ J_0\left(\kappa_m^0\frac{r}{R}\right)\sin\sqrt{\nu_n z}.$$

7. EIGENFUNCTION METHOD FOR PROBLEMS IN THE THEORY OF OSCILLATIONS

7.1. Free oscillations of a homogeneous string

We now return to the *problem of a string* and examine in greater detail the solution of this problem not only for free but also *forced oscillations*.

Function $u(x;t)$, describing the oscillations of the string with length l with fixed ends, satisfies the equation

$$\frac{\partial^2 u}{\partial t^2} - a^2\frac{\partial^2 u}{\partial x^2} = f(x;t),$$

where $f(x;t)$, is the density of the external force, the homogeneous boundary-value conditions of the first kind and the initial conditions $u\big|_{t=0} = \varphi(x),\ \left.\frac{\partial u}{\partial t}\right|_{t=0} = \psi(x)$. The corresponding eigenvalue problem has the form $\frac{\partial^2 v}{\partial x^2} + \lambda v = 0.$ For this equation and the Dirichlet boundary-value conditions the eigenvalue problem was solved in section 2. The eigenvalues and eigenfunctions have the form $\lambda_n = \pi^2 n^2/l^2$, $v_n = \sin(\pi n x/l)$. We determine the solution of our problem in the form of a Fourier series:

$$u = \sum_{n=1}^{\infty} w_n(t)\sin\frac{n\pi x}{l}.$$

The initial equation is multiplied by $(2/l)\sin(n\pi x/l)$ and integrated from 0 to l. Consequently,

$$\frac{d^2 w_n}{dt^2} + \frac{n^2\pi^2 a^2}{l^2} w_n = a_n(l),$$

where a_n are the Fourier coefficients of the right-hand side f. The initial conditions for w_n have the form

$$w_n(0) = \frac{2}{l}\int_0^t \varphi(x)\sin\frac{n\pi x}{l}dx = b_n,\quad \frac{\partial w_n(0)}{\partial t} = \frac{2}{l}\int_0^l \psi(x)\sin\frac{n\pi x}{l}dx = c_n,$$

Solving the equation for w_n, we obtain the solution u of our problem in the

form of the Fourier series. Every individual term of the Fourier series is a standing wave on the string. We examine several specific examples.

Example 1. $f(x;t) = \psi(x) = 0$, i.e. *the problem of free oscillations of the string caused only by initial deformation. In this case,* $w_n = b_n \cos(n\pi a/l)t$ and $u(x;t) = \sum_{m=1}^{\infty} b_n \cos(n\pi a/l)t \sin(n\pi x/l)$.

Example 2. $\varphi(x) = \psi(x) = 0$, $f(x;t) = F(x)\sin \omega t$, i.e., *the oscillations of the string, induced by the sinusoidal force of the density of the amplitude* $F(x)$. In this case

$$u(x,t) = \sum_{n=1}^{\infty} w_n(t) \sin \frac{n\pi x}{l},$$

where

$$w_n(t) = \frac{\tilde{a}_n l^2}{a^2 n^2 \pi^2 - \omega^2 t^2}\left[\sin \omega t - \frac{\omega l}{n\pi a}\sin \frac{n\pi a}{l}\right], \quad \omega \neq \frac{n\pi a}{l},$$

$$\tilde{a}_n = \frac{2}{1}\int_0^l F(x)\sin \frac{n\pi x}{l} dx.$$

However, if $\omega = (k\pi a/l)$ at some $n = k$, then $w_k = (\tilde{a}_k / 2\omega^2)[\sin \omega t - \omega t \cos \omega t]$. In this case it is said that the external force is resonance with the k-th frequency of the string.

Example 3. $\varphi(x) = 0$, $f(x;t) = 0$,

$$\psi(x) = \begin{cases} 0 & x < x_0 - \delta, \\ \psi(x) > 0, & x_0 - \delta < x < x_0 + \delta, \\ 0, & x > x_0 + \delta, \end{cases}$$

and $\int_{x_0-\delta}^{x_0+\delta} \psi(x)dx = A$. The problem describes the *free oscillations of the string formed under the effect of the initial pulse A concentrated in the vicinity of the point* x_0 (δ is assumed to be very small). In this case, the solution of the problem is presented in the form of a series:

$$u(x,t) = \frac{2A}{\pi a}\sum_{n=1}^{\infty} \frac{1}{n}\sin \frac{n\pi x_0}{l}\sin \frac{n\pi a t}{l}\sin \frac{n\pi x}{l}.$$

7.2. Oscillations of the string with a moving end

We examine a problem with heterogeneous boundary-value conditions. We examine the *oscillations of a string being at rest at* $t = 0$, *with the right end of the string at* $t > 0$ *forced to move in accordance with the law* $u(l,t) = \chi(t)$, *whereas the left hand of the string is fixed*. There are no forces distributed along the length of the string.

The eigenvalues and eigenfunctions here are the same as in the previous problem, and the solution is found as previously in the form of a Fourier series. The equations for w_n have the form:

$$w_n'' + \frac{n^2\pi^2a^2}{l^2}w_n = -\frac{2n\pi a^2}{l^2}(-1)^n\chi(t),$$

and the initial conditions will be $w_n(0) = w_n'(0) = 0.$

Let for definiteness $\chi(t) = A \sin \omega t$. Consequently, if $\omega \neq n\pi a/l$, we have

$$w_n(t) = -\frac{2n\pi a^2 A}{a^2n^2\pi^2 - \omega^2 l^2}(-1)^n\left[\sin\omega t - \frac{\omega l}{n\pi a}\sin\frac{n\pi at}{l}\right],$$

$$u(x,t) = \sum_{n=1}^{\infty} w_n(t)\sin\frac{n\pi x}{l}.$$

However, if one of the eigenfrequencies, for example, k-th, resonates with the forced oscillation of the end of the string (i.e. if $\omega = k\pi a/l$), the corresponding terms of the series will be:

$$w_k(t)\sin\frac{k\pi x}{l} = (-1)^{k-1}\frac{a}{l\omega}A[\sin\omega t - \omega t\cos\omega t]\sin\frac{k\pi x}{l}.$$

7.3. Problem of acoustics of free oscillations of gas

We now examine a problem of *free oscillations of gas in a closed thin pipe*. The mathematical model of this problem differs from the previous problem by only the boundary-value conditions which now have the form

$$\left.\frac{\partial u}{\partial x}\right|_{x-0} = \left.\frac{\partial u}{\partial x}\right|_{x=l} = 0.$$

In this problem, the function u indicates densening of the gas. The problem of the eigenvalues for these boundary-value conditions was solved in section 2. The eigenvalues and eigenfunctions have the form $\lambda_n = (\pi^2n^2/l^2)$, $n = 0,1....$, $u_0 = 1/2$, $u_n = \cos(n\pi l)$, $n = 1,2,...$ Therefore, for $n = 0,1,2,....$we have

$$a_n = 0,\quad b_n = \frac{2}{l}\int_0^l \varphi(x)\cos\frac{n\pi x}{l}dx,\quad c_n = \frac{2}{l}\int_0^l \psi(x)\cos\frac{n\pi x}{l}dx.$$

The solution of u will be determined in the form:

$$u(x;t) = \frac{1}{2}w_0(t) + \sum_{n=1}^{\infty} w_n(t)\cos\frac{n\pi x}{l}.$$

As in paragraph 7.1, for coefficients w_n we obtain the equation and the initial conditions.

We examine a simple specific example of oscillations, caused only by the densening ($\psi(x) = 0$). In this case, $w_a = b_n\cos(n\pi at/l)$ and therefore

$$u(x;t) = \frac{1}{2}b_0 + \sum_{n=1}^{\infty} b_n\cos\frac{n\pi at}{l}\cos\frac{n\pi x}{l}.$$

It should be mentioned that the oscillation of the gas can be regarded as the result of superposition of direct and reversed half waves reflected from the ends of the pipe, but in this case reflection takes place without any change of the sign.

7.4. Oscillations of a membrane with a fixed end

We examine a *right angled membrane with sides p and q with a fixed end*. Function u, describing the oscillations of this membrane, satisfies the equation

$$\frac{\partial^2 u}{\partial t^2} - a^2 \Delta u = F(x, y),$$

where F is the density of the external force, the boundary-value conditions

$$u\big|_{x=0} = u\big|_{x=p} = u\big|_{y=0} = u\big|_{y=q} = 0$$

and the initial conditions

$$u\big|_{t=0} = f_1(x, y), \quad \left.\frac{\partial u}{\partial t}\right|_{t=0} = f_2(x, y),$$

where f_1 (initial deformation) and f_2 (initial velocity of membrane particles) are given functions.

The system of eigenfunctions, corresponding to the rectangle and the boundary-value conditions, has the form

$$v_{m,n} = \sin\frac{m\pi x}{P}\sin\frac{n\pi y}{q}.$$

We determine the Fourier coefficients of the functions F, f_1, f_2, in respect of the functions of this system and denote them by $a_{m,n}$, $b_{m,n}$, $c_{m,n}$, respectively. We find the solution of the problem in the form of a Fourier series for the same functions:

$$u = \sum_{m,n=1}^{\infty} w_{m,n} \sin\frac{m\pi x}{p}\sin\frac{n\pi y}{q}.$$

Multiplying the equation by $(4/pq)\sin(n\pi y/p)\sin(n\pi y/q)$ and integrating over the rectangle Ω, after normal transformation we obtain the equation for $w_{m,n}(t)$:

$$\frac{d^2 w_{m,n}}{dt^2} + a^2\pi^2\left[\frac{m^2}{P^2} + \frac{n^2}{q^2}\right] w_{m,n} = a_{m,n}$$

and the initial conditions

$$w_{m,n}\big|_{t=0} = b_{m,n}, \quad \left.\frac{dw_{m,n}}{dt}\right|_{t=0} = c_{m,n}.$$

Determining $w_{m,n}$ and substituting it into the Fourier series, we obtain the required solution.

As an example, *we examine the free oscillation of a membrane under the effect of one only initial deformation*: $F = 0$, $f_2 = 0$.

In this case $a_{m,n} = c_{m,n} = 0$ the equation for $w_{m,n}$ is reduced to the following form:

$$w_{m,n}^{n} + a^2\pi^2\left[\frac{m^2}{p^2} + \frac{n^2}{q^2}\right] w_{m,n} = 0.$$

Its general solution has the form

$$w_{m,n} = A_{m,n}\cos a\pi\sqrt{\frac{m^2}{p^2}+\frac{n^2}{q^2}1} + B_{m,n}\sin a\pi\sqrt{\frac{m^2}{p^2}+\frac{n^2}{q^2}}t.$$

The initial conditions give $B_{m,n}= 0$, $A_{m,n}= b_{m,n}$. Therefore:

$$u = \sum_{m,n=1}^{\infty} b_{m,n}\cos a\pi\sqrt{\frac{m^2}{p^2}+\frac{n^2}{q^2}}t\sin\frac{m\pi x}{p}\sin\frac{n\pi y}{q}.$$

Thus, the frequencies corresponding to every standing wave, are the numbers:

$$\omega_{m,n} = a\pi\sqrt{\frac{m^2}{p^2}+\frac{n^2}{q^2}}.$$

If $\omega_{m,n}$ are the same for two different pairs m and n, we have two standing waves with the same frequency. Their sum also gives a standing wave with the same frequency.

7.5. Problem of oscillation of a circular membrane

We examine *the oscillation of a circular membrane with radius R with a fixed end*. This problem is solved in the same manner as the previous problem, with the only difference being that for the circle we have a system of eigenfunctions:

$$v_{0,n} = \frac{1}{2}J_0\left(\kappa_n^0\frac{r}{R}\right),$$

$$v_{2m-1,n} = J_m\left(\kappa_n^m\frac{r}{R}\right)\sin m\varphi,\quad v_{2m,n} = J_m\left(\kappa_n^m\frac{r}{R}\right)\cos m\varphi,$$

corresponding to the eigenvalues $\lambda_{2m-1,n} = \lambda_{2m,n} = (\kappa_n^m)^2/R^2$, where χ_n^m is the root of the equation $J_m(x) = 0$. The value u will be determined as

$$u = \sum_{n=0}^{\infty}\left\{\frac{w_{0,n}(t)}{2}J_0\left(\kappa_n^0\frac{r}{R}\right)+\right.$$

$$\left.+\sum_{m-1}^{\infty}[w_{2m-1,n}(t)\sin m\varphi + w_{2m,n}(t)\cos m\varphi]J_m\left(\kappa_n^m\frac{r}{R}\right)\right\}.$$

For $w_{k,n}$ we obtain the equation:

$$w''_{k,n}(t) + a^2\frac{(\kappa_n^m)^2}{R^2}w_{k,n}(t) = a_{k,n}\quad (k = 2m-1;\quad k = 2m)$$

and the initial condition $w_{k,n}(0) = b_{k,n}$, $w'_{k,n}(0) = c_{k,n}$.

If the external force is not present, then $a_{k,n}$=0 and therefore

$$w_{k,n}(t) = A_k\cos a\frac{\kappa_n^m}{R}t + B_k\sin a\frac{\kappa_n^m}{R}t\quad (k = 2m-1,\quad k = 2m).$$

Thus, the numbers $a\kappa_n^m/R$ are the frequencies of eigenoscillations of the membrane.

It should now be mentioned that if the initial deformation, the initial velocity and the external force are characterized by central symmetry (i.e. do not depend on φ), then at $k > 0$ $a_{k,n}$, $b_{k,n}$, $c_{k,n}$ are equal to zero. Therefore, all $w_{k,n}$ are equal to $k > 0$, and the Fourier series is reduced to:

$$u = \sum_{n\,0}^{\infty} \frac{w_{0,n}(t)}{2} l_0\left(\kappa_n^0 \frac{r}{R}\right).$$

BIBLIOGRAPHIC COMMENTARY

The fundamentals of the theory of eigenvalue problems and special functions and also the eigenfunction method for the problem of mathematical physics were described in a classic work [110], where the theory of the Fourier series and their application to solution of the boundary problems were also justified. The method of solving mathematical physics problems, including the eigenfunction method, and justification of these methods were described in [13,49,70,88,91].

The fundamentals of the theory of special functions are discussed in [76] and also in individual chapters of the books [1,85]. A brief description of the special functions, encountered in mathematical physics, is also given in [25].

The systematic presentation of the main methods of solving mathematical physics problems, relating to the calculation of electrical, magnetic and wave fields, is in [19], where special attention is also given to the eigenfunction method for solving problems with non-negative boundary-value conditions. The formulations and methods of solving problems for eigenvalues with applications in technical mechanics were published in [37]. The authors of [26] examined a number of equations of linear mechanics and theoretical physics and present exact solutions of linear and non-linear equations by different methods, including the eigenfunction method. The method of expansion in respect of the eigenfunctions for the problems of the theory of oscillations, heat conductivity problems and elliptical problems, is described in [4,69].

Chapter 4

METHODS OF INTEGRAL TRANSFORMS

Keywords: Integral transform, Fourier transform, Laplace transform, Mellin transform, Hankel transform, Meyer transform, Kontorovich–Lebedev transform, Mehler–Fock transform, Hilbert transform, Laguerre transform, Legendre transform, convolution, Bochner transform, chain transforms, wavelet transform, Z-transform, generating function, problems of the theory of oscillations, heat conductivity problems, problems of the theory of deceleration of neutrons, hydrodynamic problems, problems of elasticity theory, Bussinesq problem, coagulation equation, physical kinetics.

MAIN CONCEPTS AND DEFINITIONS

Integral transform – functional transform of the type $F(x)=\int_\Gamma K(x,t)f(t)dt$, where $f(t)$ is the original function, $F(x)$ is the map (transform), Γ is the domain of the complex Euclidean space.

Fourier transform – integral transform at $K(x,t) = e^{-ixt}$, $\Gamma = \mathsf{R}^n$.

Fourier sine transform – integral transform in which $K(x,t) = \sin(xt)$, $\Gamma=\mathsf{R}^1_+$.

Fourier cosine transform – integral transform at $K(x,t) = \cos(xt)$, $\Gamma=\mathsf{R}^1_+$.

Laplace transform – integral transform at $K(x,t) = e^{-xt}$, $\Gamma=\mathsf{R}^1_+$.

The Laplace discrete transform of the sequence $f(n)$, $0\le n<\infty$, – function of the complex variable $F(x)=\sum_n f(n)e^{-nx}$.

Generating function of the sequence $\{f(n)\}_{n=0}^{\infty}$ – function $F(z)=\sum_{n=0}^{\infty} z^n f(n)$ if we set $z = e^{-x}$, we obtain a discrete Laplace transform.

Z-transform of the sequence $\{f(n)\}_{n=0}^{\infty}$ – function $F(z)=\sum_{n=0}^{\infty} z^{-n} f(n)$ if we set $z = e^x$, we obtain a discrete Laplace transform.

Mellin transform – integral transform at $K(x,t) = e^{x-1}$, $\Gamma=\mathbb{R}_+^1$.

Hankel transform – integral transform at $K(x,t) = J_\nu(xt)t$, J_ν – the Bessel function, $\Gamma=\mathbb{R}_+^1$.

Meyer transform – integral transform at $K(x,t)=K_\nu(xt)(xt)^{1/2}$, K_ν – the MacDonald function, $\Gamma=\mathbb{R}_+^1$.

The Kontorovich–Lebedev transform – integral transform at $K(x,t)=\sqrt{2x\sinh(\pi x)}K_{ix}(t)/\sqrt{t}, K_\nu$ – the MacDonald function, $\Gamma=\mathbb{R}_+^1$.

Mehler–Fock transform – integral transform at $K(x,t)=\sqrt{t\tanh(\pi t)}P_{-1/2-ix}(t)$, $P_\nu(x)$ is the spherical Legendre function of the first kind, $\Gamma=\mathbb{R}_+^\infty$.

Gilbert transform – integral transform at $K(x,t)=(t-x)^{-1}$, or $K(x,t) = \text{cotan}\,(x-2)/2$, $\Gamma = \mathbb{R}^1$.

Laguerre transform – integral transform at $K(n,t) = e^{-1}L_n(t)$, L_n is the Laguerre polynomial, $\Gamma=\mathbb{R}_+^1$.

Legendre transform – integral transform at $K(n,t) = P_n(t)$, P_n is the Legendre polynomial, $\Gamma = [-1,1]$.

Bochner transform – integral transform at $K(r,t) = 2\pi r^{1-n/2}$ $J_{n/2-1}(2\pi rt)t^{n/2}$, $\Gamma=\mathbb{R}_+^1$.

Convolution transform – integral transform at $K(x,t) = G(x-t)$, $\Gamma=\mathbb{R}^1$.

Wavelet transform – integral transform at $K(a,b;t) = \psi((t-b)/a)$, $\Gamma=\mathbb{R}^n$, where ψ is the 'wavelet', i.e. the function with the zero mean decaying quite rapidly at infinity.

1. INTRODUCTION

The *integral transforms* are the functional transformations of the type

$$F(x)=\int_\Gamma K(x,t)f(t)dt,$$

where Γ is the finite or infinite contour in the complex plane, $K(x,t)$ is the transform kernel. In most cases, we examine integral transforms for which $K(x,t) = K(xt)$, and Γ is the real axis or its part (a,b). If $-\infty<a,b<\infty$, the transform is the finite integral transform. The formulae which make it possible to restore the function $f(t)$ in respect of the available $F(x)$, are the *formulae of the conversion* of the integral transforms.

If x, $t\in\mathbb{R}^n$, and Γ is the domain of the n-dimensional Euclidean space, then we examine multiple (*multidimensional*) *integral transforms*.

The method of integral transforms is an effective method of solving and examining the differential and integral equations in mathematical physics, based on integrating the equations with some weight functions, depending

on two arguments, which often leads to simplification of the initial problem. The main condition for using the integral transforms is the availability of a theorem of conversion which makes it possible to find the initial function, knowing its map. Depending on the weight function and the integration range, we examine Fourier, Laplace, Mellin, Hankel, Meyer, Hilbert and other transforms.These transforms may be used to solve many problems of the theory of vibrations, conductivity, diffusion and deceleration of the neutrons, hydrodynamics, theory of elasticity, physical kinetics.

The integral transforms are used in most cases in solving the differential and integral equations, and their selection depends on the type of examined equations [21,57,89,93,96]. The main condition in selecting the integral transform is the possibility of converting the differential or integral equation to a simpler differential equation (or, even better, to an algebraic relationship) for function $F(x)$, and it is understood that the inversion formula is available. If the contour Γ is finite (for example, a segment), the transform $F(x)$ is the finite transform $f(t)$. Evidently, the number of integral transforms may be greatly increased as a result of new kernels.

In this section, we examine mainly the transforms in which the contour Γ is the real axis or half-axis, and it is also assumed that all the resultant integrals are finite.

2. MAIN INTEGRAL TRANSFORMATIONS

2.1. Fourier transform

The Fourier transform of the function $f(t)$ is the expression

$$F(x) \equiv \mathcal{F}[f] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ixt} f(\tau) d\tau.$$

Function $F(x)$ is the *Fourier map* of the function f. The inverse Fourier transform has the form

$$f(t) \equiv F^{-1}[F(x)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} F(x) dx.$$

Combining these expressions, we obtain the *exponential Fourier formula*

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \int_{-\infty}^{\infty} e^{-ix\tau} f(\tau) d\tau\, dx, \tag{1}$$

which is equivalent to the *integral Fourier formula*

$$f(t) = \frac{1}{\pi} \int_{0}^{\infty} dx \int_{-\infty}^{\infty} f(\tau) \cos(x(\tau - t)) d\tau. \tag{2}$$

Expanding the cosine, we obtain the identity

$$f(t) = \int_{0}^{\infty} [a(x)\cos(tx) + b(x)\sin(tx)] dx, \tag{3}$$

where

$$a(x)=\frac{1}{\pi}\int_{-\infty}^{\infty} f(t)\cos xt\ dt,\quad b(x)=\frac{1}{\pi}\int_{-\infty}^{\infty} f(t)\sin xt\ dt.$$

If $f(t)$ is an even function, formula (3) has the form

$$f(t)=\frac{2}{\pi}\int_{0}^{\infty}\cos tx\ du\int_{0}^{\infty} f(\tau)\cos x\tau\ d\tau. \tag{4}$$

Similarly, if $f(t)$ is an odd function, then

$$f(t)=\frac{2}{\pi}\int_{0}^{\infty}\sin tx\ dx\int_{0}^{\infty} f(\tau)\sin x\tau\ d\tau. \tag{5}$$

The conditions for function f at which the formulas (1), (2) and also the direct and inverse Fourier transforms are valid, are determined in the following theorem, in which $L(-\infty,+\infty)$ denotes the space of the Lebesque integral functions on $(-\infty,+\infty)$.

Theorem 1. *Let f belong to* $L(-\infty,+\infty)$ *and is a function of bounded variation at any finite range. Then formulas* (1), (2) *are valid if at the points of discontinuity of the function* $f(t)$ *the left parts are replaced by* $(1/2)\{f(t+0)+f(t-0)\}$.

2.1.1. The main properties of Fourier transforms

If function $f(t)$ is integrable in the interval $(-\infty,+\infty)$, function $F(x)$ exists for all t. The functions $F(x)$ and $f(t)$, being both Fourier transforms of each other, are referred to as a pair of *Fourier transforms*. Assuming that

$$F_c(x)=\sqrt{\frac{2}{\pi}}\int_{0}^{\infty} f(t)\cos xt\ dt, \tag{6}$$

from formula (4) we obtain

$$f(t)=\sqrt{\frac{2}{\pi}}\int_{0}^{\infty} F_c(x)\cos tx\ dx. \tag{7}$$

The functions, linked in this manner, are referred to as the pair of *Fourier cosine transforms*. Similarly, from equation (5) we can obtain the pair of *Fourier sine transforms*:

$$F_s(x)=\sqrt{\frac{2}{\pi}}\int_{0}^{\infty} f(t)\sin xt\ dt, \tag{8}$$

$$f(t)=\sqrt{\frac{2}{\pi}}\int_{0}^{\infty} F_s(x)\sin tx\ dx. \tag{9}$$

If $f(t)$ is an even function, then

$$F(x)=F_c(x);$$

if $f(t)$ is an odd function, then

$$F(x)=iF_s(x).$$

Let functions $F(x)$ and $G(x)$ be the Fourier transforms of the functions $f(t)$ and $g(t)$, determined by equations (6), (7).

The functions

$$F(u)G(u) \quad \text{and} \quad h(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(\tau)f(t-\tau)d\tau$$

are the pair of Fourier transforms. The function $h(t)$ is the *convolution of the functions* $f(t)$ and $g(t)$ and is denoted by $h = f*g = g*f$.

Theorem 2 (convolution). *Let* $f, g\in L(-\infty,\infty)$. *Then* $h(t) = f*g(t)$ *belongs to* $L(-\infty,\infty)$ *and its Fourier transform is the function* $\sqrt{2\pi}F(x)G(x)$. *Conversely, the product* $\sqrt{2\pi}F(x)G(x)$ *belongs to* $L(-\infty,\infty)$, *and its Fourier transform is determined by the formula* $f*g(t)$.

The Parseval formulae. Let $f(t)\in L(-\infty,\infty)$, $g(t)$ be integrable in every finite range, and let

$$G(x)=\frac{1}{\sqrt{2}}\lim_{l\to\infty}\int_{-l}^{l} g(\tilde{\ })e^{-ix\tilde{\ }}d\tilde{\ }$$

for all x, and $G(x)$ is finite everywhere and belongs to $L(-\infty,\infty)$. Then, we obtain the *Parseval identity*

$$\int_{-\infty}^{\infty} F(x)G(x)dx=\int_{-\infty}^{\infty} f(t)g(t)dt. \tag{10}$$

In particular, at $f = g$ we have the *Parseval formula*

$$\int_{-\infty}^{\infty} |F(x)|^2\,dx=\int_{-\infty}^{\infty} |f(t)|^2 dt. \tag{11}$$

2.1.2. Multiple Fourier transform

According to definition, we have

$$F(x)\equiv\mathcal{F}[f(t)]=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-ixt}f(t)dt,$$

where $x = (x_1,x_2)$, $t = (t_1,t_2)$, $xt = x_1t_1+x_2t_2$, $dt = dt_1dt_2$. Function $F(x)$ is the *Fourier transform of the function of two variables* $f(t)$. For the functions $f(t)$ and $F(x)$, belonging in $L(\mathrm{R}^2)$, we have the following inversion formula:

$$f(t)\equiv\mathcal{F}^{-1}[F(x)]=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{ixt}F(x)dx.$$

If x, $t\in\mathrm{R}^n$, then

$$F(x)=\frac{1}{(2\pi)^{n/2}}\int_{\mathrm{R}^n} e^{-ixt}f(t)dt, \quad f(t)=\frac{1}{(2\pi)^{n/2}}\int_{\mathrm{R}^n} e^{ixt}F(x)dx.$$

2.2. Laplace transform

2.2.1. Laplace integral

$f(t)$ denotes the function of the real variable t, $0 \le t < +\infty$, integrable according

to Lebesgue in any bounded interval $(0,A)$. Let p be a complex number. The function

$$F(p) \equiv \mathcal{L}[f(t)] = \int_0^\infty e^{-pt} f(t)dt \tag{12}$$

is the Laplace transform of function $f(t)$.

2.2.2. The inversion formula for the Laplace transform

Using the definition of the Laplace transform and assuming $p = \gamma + iy$, from (1), (12) we obtain:

$$\int_{\gamma-i\omega}^{\gamma+i\omega} e^{pt} F(p)dp = ie^{\gamma t} \int_{-\omega}^{+\omega} e^{ity} dy \int_0^\infty e^{-iy\tau}[e^{-\gamma\tau} f(\tau)]d\tau. \tag{13}$$

However, at $\omega \to \infty$ the double integral in the right hand side of equation (13) according to (1) is equal to $2\pi e^{-\gamma t} f(t)$ for $t > 0$ and to zero for $t < 0$. Therefore, equation (13) gives

$$f(t) = \frac{1}{2\pi i} \lim_{\omega\to\infty} \int_{\gamma-i\omega}^{\gamma+i\omega} e^{pt} F(p)dp \tag{14}$$

for $t > 0$ and zero for $t < 0$. Equation (14) is the *inversion formula* for the Laplace transform. The condition ensuring the existence of the Laplace transform (12) are superposed on $f(t)$ and γ should be larger than the real parts of all singular points of the Laplace transform $F(p)$.

2.2.3. Main formulae and limiting theorems

At Re $p > \gamma$ the following relationships are valid:

$$\mathcal{L}[f(\alpha t)] = \frac{1}{\alpha} F\left(\frac{p}{\alpha}\right), \qquad \mathcal{L}[f_* g(t)] = F(p)G(p),$$

$$\mathcal{L}\left[\frac{f(t)}{t}\right] = \int_p^\infty F(q)dq, \qquad \mathcal{L}[f(0)g(t) + (f'_* g)(t)] = pF(p)G(p).$$

$$\mathcal{L}[f'(t)] = pF(p) - f(0), \qquad \mathcal{L}\left[\int_0^t f(s)ds\right] = \frac{F(p)}{p}.$$

If $F(p)$ is the Laplace transform and $\mathcal{L}(f'(t))$ does exist, then

$$\lim_{p\to\infty} pF(p) = f(0+0);$$

if, in addition, there is a limit $f(t)$ at $t\to\infty$, then

$$\lim_{p\to 0} pF(p) = \lim_{t\to\infty} f(t).$$

2.3. Mellin transform

The Mellin transform

$$F(s) \equiv \mathcal{M}[f(t)] = \int_0^\infty f(t)t^{s-1}dt, \quad s = \sigma + i\tau, \tag{15}$$

is closely linked with the Fourier and Laplace transforms.

The Mellin transform may be used efficiently when solving a specific group of planar harmonic problems in the sectorial region, the problems of elasticity, and also when examining special functions, summation of series and calculation of integrals. The theorems, belonging in the Mellin transform, may be obtained from the appropriate theorems for the Fourier and Laplace transforms by substituting the variable.

Theorem 4. *Let us assume that* $t^{\sigma-1}f(t)\in L(0,+\infty)$. *Then, we obtain the inversion formula:*

$$\frac{f(t+0)+f(t-0)}{2}=\frac{1}{2\pi i}\lim_{\lambda\to\infty}\int_{\sigma-i\lambda}^{\sigma+i\lambda}F(s)t^{-s}ds. \tag{16}$$

Theorem 5. *Let us assume that* $F=\mathcal{M}[f]$, $G=\mathcal{M}[g]$. *Let either* $t^{k-1}f(t)\in L(0,+\infty)$ *and* $G(1-k-ix)\in L(-\infty,+\infty)$, or $F(k+ix)\in L(-\infty,+\infty)$, $t^k g(t)\in L(0,+\infty)$. *Then*

$$\frac{1}{2\pi i}\int_{k-i\infty}^{k+i\infty}F(s)G(1-s)ds=\int_0^\infty f(t)g(t)dt. \tag{17}$$

The following relationship is also valid

$$\frac{1}{2\pi i}\int_{k-i\infty}^{k+i\infty}F(s)G(s)ds=\int_0^\infty g(t)f\left(\frac{1}{t}\right)\frac{dt}{t}. \tag{18}$$

Theorem 6 (on convolution). *Let* $t^k f(t)$ *and* $t^k g(t)$ *belong to the space* $L(0,+\infty)$ *and*

$$h(t)=\int_0^\infty f(\tau)g\left(\frac{t}{\tau}\right)\frac{d\tau}{\tau}.$$

Then, the function $t^k h(t)$ *belongs* $L(0,+\infty)$ *and its Mellin transform is* $F(s)G(s)$.

2.4. Hankel transform

The integral transformations of the type

$$F(\lambda)=\int_0^\infty f(t)K(\lambda t)\,dt,$$

where $K(z)$ is a Bessel function, are known as *Bessel transforms*. This type includes Hankel, Meyer, Kontorovich–Lebedev, and other transforms.

The formulas of type (1), (2) giving the expansion of the arbitrary function $f(x)$ in the Fourier integral, are of considerable interest in many problems of mathematics and physics. The expansions of this type include the expansion over cylindrical functions, known as the *Fourier–Bessel integral:*

$$f(t)=\int_0^\infty J_\nu(xt)x\,dx\int_0^\infty f(\tau)J_\nu(x\tau)\tau\,d\tau\quad(0<t<\infty), \tag{19}$$

where J_ν is the Bessel function, $\nu>-1/2$.

Theorem 7. *Let $f(t)$ be the function of bounded variation in any finite segment and*

$$\int_0^\infty |f(t)| t^{1/2} dt < \infty.$$

Then, at $\nu > -1/2$

$$\frac{1}{2}[f(t+0)+f(t-0)] = \int_0^\infty J_\nu(xt)x\,dx \int_0^\infty f(\tau)J_\nu(x\tau)\tau\,d\tau. \tag{20}$$

At the continuity points we have the formula (19).

The Hankel transformation is the integral

$$F_\nu(x) = \mathcal{H}_\nu[f(t)] = \int_0^\infty f(t)tJ_\nu(xt)dt \quad (0 < x < +\infty). \tag{21}$$

The integral expansion (19) gives the inversion formula

$$f(t) = H_\nu^{-1}[F_\nu(x)] = \int_0^\infty F_\nu(x)J_\nu(xt)x\,dx \quad (0 < t < +\infty). \tag{22}$$

It should be mentioned that if $f(t)$ is such that $f(t) = O(t^\alpha)$, $t\to 0$, $\alpha+\nu+2>0$ and $f(t) = O(t^\beta)$, where $t\to\infty$, $\beta+(3/2)<0$, then integral (21) converges.

Another expansion of the identical type can be added to the expansion (19)

$$f(t) = \int_0^\infty H_\nu(tx)(tx)^{1/2}dx \int_0^\infty Y_\nu(x\tau)(x\tau)^{1/2} f(\tau)d\tau, \tag{23}$$

where Y_ν is the *Bessel function of the second kind*, H_ν *is the Struve function*. Formula (23) is a basis for introducing the appropriate integral transformation.

A generalisation of the integral expansion (19) is the formula

$$f(t) = \int_0^\infty \frac{\varphi_x(t)x\,dx}{J_\nu^2(ax)+Y_\nu^2(ax)} \int_a^\infty f(\tau)\varphi_x(\tau)\tau\,d\tau \quad (a < t < +\infty), \tag{24}$$

where $\varphi_x(t) = J_\nu(ax)Y_\nu(xt)-Y_\nu(ax)J_\nu(xt)$, $\nu > -1/2$, is the linear combination of the Bessel function of the first and second kind of the ν-th order. Expansion (24) holds if $f(t)$ is a piecewise continuous function of bounded variation in any finite interval (a, R) and the integral

$$\int_a^\infty |f(t)| t^{1/2} dt < \infty.$$

Equation (24) leads to the appropriate integral transform referred to as the *Weber transform*:

$$F(x,a) = \int_a^\infty c_\nu(tx,ax)t\,f(t)dt, \quad a \le t < \infty,$$

where $c_\nu(\alpha,\beta) \equiv J_\nu(\alpha)Y_\nu(\beta)-Y_\nu(\alpha)J_\nu(\beta)$. The inversion formula has the following form:

$$f(t)=\int_0^\infty \frac{c_\nu(tx,ax)}{J_\nu^2(ax)+Y_\nu^2(ax)}\,xF(x,a)dx.$$

At $a\to 0$ the Weber transform changes to the Hankel transform which at $\nu = \pm(1/2)$ changes to the sine and cosine Fourier transforms. There is also the Parseval equality: if $\nu \geq -(1/2)$, $F(x)$ and $G(x)$ are the Hankel transforms of the functions f,g and f, $G\in L_1(0,\infty)$, then

$$\int_0^\infty f(t)g(t)dt=\int_0^\infty F(x)G(x)dx.$$

The Hankel and Weber transforms may be efficiently used in solving the boundary-value problems for Laplace and Helmholtz equations and some problems of elasticity theory.

2.5. Meyer transform

When solving differential equations of the Bessel type, the integral *Meyer transform* is of considerable importance. It is determined by the integral

$$F(s)\equiv \mathcal{M}[f(t)]=\sqrt{\frac{2}{\pi}}\int_0^\infty K_\nu(st)(st)^{1/2}f(t)dt, \tag{25}$$

where $K_\nu(st)$ *is the MacDonald function*. The inversion formula has the form

$$f(t)=\frac{1}{i\sqrt{2\pi}}\lim_{\lambda\to\infty}\int_{\beta-i\lambda}^{\beta+i\lambda} I_\nu(ts)(ts)^{1/2}F(s)ds. \tag{26}$$

Here I_ν is the Bessel function of the imaginary argument.

Theorem 8. *Let $f(t)$ be the function of real variable, $0 \leq t \leq +\infty$, integrable and having bounded variation in any finite interval. Let also*

$$\int_0^\infty e^{-\beta t}\left|f(t)\right|dt<\infty,\quad \beta>a\geq 0.$$

Then

$$\frac{f(t+0)+f(t-0)}{2}=\frac{1}{i\pi}\lim_{\lambda\to\infty}\int_{\beta-i\lambda}^{\beta-i\lambda} I_\nu(ts)(xs)^{1/2}ds\int_0^\infty K_\nu(st)(st)^{1/2}f(t)dt. \tag{27}$$

2.6. Kontorovich–Lebedev transform

When solving some problems of mathematical physics, special attention is given to a number of integral transforms, containing integration in respect of the index of the Bessel functions. This form of integral transforms was examined for the first time by N.I. Kontorovich and N.N. Lebedev in 1938. Of primary significance in the theory of the Kontorovich–Lebedev transform is the expansion of an integral of the Fourier type

$$t\,f(t)=\frac{2}{\pi^2}\int_0^\infty K_{i\tau}(t)\tau\sinh(\pi\tau)\int_0^\infty K_{i\tau}(t')f(t')dt', \tag{28}$$

where $K_\nu(t)$ is the *MacDonald function*, $t > 0$, $f(t)$ is an arbitrary function with continuous derivative, satisfying the conditions:

$$t^2 f(t),\quad tf(t)\in L(0,+\infty).$$

We introduce the Kontorovich–Lebedev transform

$$F(\tau)=\frac{1}{\pi}\int_0^\infty f(t)\frac{\sqrt{2\tau\,\mathrm{sh}(\pi\tau)}K_{i\tau}(t)}{\sqrt{t}}dt,\quad 0\le\tau<+\infty. \tag{29}$$

Expression (28) yields directly the following inversion formula:

$$f(t)=\frac{1}{\pi}\int_0^\infty\frac{\sqrt{2\tau\,\mathrm{sh}(\pi\tau)}K_{i\tau}(t)}{\sqrt{t}}F(\tau)d\tau,\quad t>0. \tag{30}$$

To calculate several types of integrals, it is necessary to use the equations indentical to Parseval formulae in the theory of series and Fourier integrals.

Theorem 9. *Let $f(t)$ be an arbitrary real function such that $f(t)t^{-3/4}\in L(0,+\infty)$, $f(t)\in L_2(0,+\infty)$. Then*

$$\int_0^\infty[F(\tau)]^2d\tau=\int_0^\infty[g(t)]^2dt. \tag{31}$$

Theorem 10. *Let $f_1(t)$ and $f_2(t)$ be arbitrary real functions satisfying the conditions of the previous theorem. Then*

$$\int_0^\infty F_1(\tau)F_2(\tau)d\tau=\int_0^\infty f_1(x)f_2(x)dx.$$

2.7. Meller–Fock transform

The Meller–Fock integral transform is determined by the expression

$$F(x)=\int_0^\infty\sqrt{t\,\mathrm{th}(\pi t)}P_{-1/2+it}(x)f(t)dt,\quad 1\le x<\infty, \tag{32}$$

where $P_\nu(x)$ is the *spherical Legendre function of the first kind*. If $f(t)\in L(0,\infty)$, $|f'(t)|$ is locally integrable on $[0,\infty)$ and $f(0)=0$, then we have the inversion formula

$$f(t)=\sqrt{t\,\mathrm{th}(\pi t)}\int_1^\infty P_{-1/2+it}(x)F(x)dx,\quad t\ge 0. \tag{33}$$

Under additional conditions, we obtain the Parseval identity:

$$\int_0^\infty f_1(t)f_2(t)dt=\int_1^\infty F_1(x)F_2(x)dx.$$

2.8 Hilbert transform

We examine the Fourier integral formula (3). Replacing formally a by b and b by $-a$, we obtain the Hilbert transform

$$g_1(x) = \frac{1}{\pi}\int_0^\infty \frac{f(x+t)-f(x-t)}{t}\,dt. \tag{34}$$

If $f \in L(-\infty,+\infty)$, then function g_1 exists for almost all values of x. If $f \in L_p(-\infty,+\infty)$, $1 < p < \infty$, then $g_1 \in L_p(-\infty,+\infty)$ and the inverse Hilbert transform is valid almost everywhere

$$f(t) = -\frac{1}{\pi}\int_0^\infty \frac{g_1(x+t)-g_1(x-t)}{t}\,dt. \tag{35}$$

The formulas (34), (35) are equivalent to the formulae

$$g_1(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{f(t)}{t-x}\,dt, \quad f(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g_1(t)}{t-x}\,dt, \tag{36}$$

in which the integrals are understood as the principal value.

The Hilbert transform is also the following integral treated as the principal value

$$g_2(x) = \frac{1}{2\pi}\int_0^{2\pi} f(t)\operatorname{ctg}\frac{t-x}{2}\,dt. \tag{37}$$

In the theory of Fourier series, function g_2 is referred to as adjoint to f. The integral operators, generated by the Hilbert transforms, are bounded in the spaces L_p.

If f satisfies the Lipschits condition or $f \in L_p(0, 2\pi)$, and in addition

$$\int_0^{2\pi} g_2(x)\,dx = 0,$$

then the following inversion formula is valid

$$f(t) = -\frac{1}{2\pi}\int_0^{2\pi} g_2(x)\operatorname{ctg}\frac{x-t}{2}\,dx, \tag{38}$$

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

It should be mentioned that there is a simple relationship between the integral kernels of the Hilbert transforms:

$$\frac{d\tau}{\tau-\xi} = \frac{1}{2}\left(\operatorname{ctg}\frac{t-x}{2} + i\right)dt, \tag{39}$$

where $\xi = e^{ix}$, $\tau = e^{it}$.

2.9. Laguerre and Legendre transforms

The integral transform

$$F(n) \equiv \mathcal{T}_1[f(t)] = \int_0^\infty e^{-t} L_n(t) f(t)\,dt \quad (n = 0,1,2,\ldots), \tag{40}$$

where $L_n(t)$ are the Laguerre polynomials of the n-th order, is referred to as the Laguerre transform. The latter is used for solving the differential Laguerre equation

$$Lf + nf = 0, \quad Lf(t) = tf''(t) + (1-t)f'(t). \tag{41}$$

The application of the Laguerre transform reduces the differential operation Lf to the algerbraic one in accordance with the formula

$$\mathcal{T}_1[Lf(t)] = -nF(n) \quad (n = 0,1,2,...).$$

The integral transform of the type

$$F(n) \equiv \mathcal{T}_2[f(t)] = \int_{-1}^{1} P_n(t) f(t)dt, \quad n = 0,1,2,...,$$

where $P_n(t)$ is the Legendre polynomial of the order n, is referred to as the Legendre transform. The inversion formula has the form

$$f(t) = \sum_{n=0}^{\infty} \left(n + \frac{1}{2}\right) P_n(t) F(n), \quad -1 < t < 1,$$

if the series converges. The Legendre transform reduces the differential operation $d(1-t^2)d$ to an algebaic one in accordance with formula

$$\mathcal{T}_2 \left\{ \frac{d}{dt}(1-t^2)\frac{df(t)}{dt} \right\} = -n(n+1)F(n), \quad n = 0,1,2,...$$

2.10 Bochner and convolution transforms, wavelets and chain transforms

The Bochner transform has the form

$$\mathcal{B}[f](r) = 2\pi r^{1-n/2} \int_0^{\infty} J_{n/2-1}(2\pi r\rho)\rho^{n/2} f(\rho)d\rho,$$

where $J_\nu(x)$ is the cylindrical function of the first kind of the order ν, ρ is the distance in $\mathbb{R}^n$. The inversion formula $f = \mathcal{B}^2 f$ is valid. The Parseval identity in this case has the form

$$\int_0^{\infty} |\mathcal{B}[f](r)|^2 r^{k-1} dr = \int_0^{\infty} |f(\rho)|^2 \rho^{k-1} d\rho.$$

The convolution transform with the kernel $G(x-t)$ has the form

$$F(x) = \int_{-\infty}^{\infty} G(x-t) f(t)dt.$$

It is assumed that for the kernel G we have a sequence of differentiation and shift operators $\{P_n\}_{n=1}^{\infty}$, which transforms $G(x-t)$ to the delta-like sequence $G_n(x-t) = P_n G(x-t)$, i.e. to a sequence of functions converging in some sense to the delta function $\delta(x-t)$. Consequently, using assumptions on the kernel G, the inversion formula for the convolution transform has the form $f = \lim_{n\to\infty}(P_n F)$, where the limit is understood in some generalised sense.

It should be mentioned that some examined integral transforms maybe regarded as a partial case of the convolution transform. For example, at $G(t) = e^t \exp(-e^t)$ and substitution of the variables $y = e^x$, $\tau = e^{-t}$ we obtain $G(x-t) = y\tau e^{y\tau}$ and

$$y^{-1}F(\ln y) = \int_0^\infty f(-\ln\tau)e^{-y\tau}d\tau, \quad 0 < y < \infty.$$

The last equation maybe regarded as identical with the Laplace transform. In the last fifteen years, special attention has been given to the application of the theory of wavelet transforms which maybe used as the kernel of a *wavelet integral transform*. A *wavelet* in the most general form is a function ψ determined on the numerical axis and having the zero mean and decaying quite rapidly at infinity. More accurately it is assumed that

$$\psi \in L_1(\mathrm{R}^n), \quad \int_{\mathrm{R}^n} \psi(x)dx = 0$$

and the Fourier transform $\Psi(\omega)$, $\omega \in \mathrm{R}^n$, satisfies the condition

$$\int_0^\infty |\Psi(t\omega)|^2 \frac{dt}{t} = 1 \quad \text{for any} \quad \omega \neq 0. \tag{42}$$

For example, if ψ is sufficiently regular, localised, has the zero mean and is radial, then there is a constant $c > 0$ such that $c\psi(x)$ satisfies (42).

We set

$$\psi_a(x) = a^{-n/2}\psi(x/a) \text{ and } \psi_{a,b}(x) = a^{-n/2}\psi((x-b)/a).$$

A classic example of the wavelet system is the Haar system of basic functions on a straight line where the initial wavelet is represented by the function $\psi(t)$, equal to unity at $t \in (0,1/2)$, $\psi(t) = -1$, $t \in (1/2,1)$ and equal to zero in other cases. We determine the wavelet integral transform:

$$F(a,b) \equiv \mathcal{W}_\psi[f(t)] = \int_{\mathrm{R}^n} f(t)\psi_{a,b}(t)dt, \quad b \in \mathrm{R}^n, \quad a > 0.$$

The following inversion formula is also valid in this case:

$$f(x) = \int_0^\infty \left[\int_{\mathrm{R}^n} F(a,b)\psi_{a,b}(x)db \right] \frac{da}{a^{n+1}}.$$

The integral wavelet transform also gives the local information on the function and its Fourier transform. To analyse high-frequency components of the function the localization is stronger (to increase accuracy), and for low frequency components the localization is weaker (to obtain complete information). This explains the popularity of wavelets in applications associated with analysis of the properties of acoustic and seismic signals, in processing and synthesis of various signals, for example, speech signals, in image analysis, etc. In addition to the kernel of the integral transforms, the wavelets are used as a generating function for constructing a basis using dilations, i.e. compressions with the conservation of the norm in $L_2(\mathrm{R})$: $\Psi_j(t) \equiv \Psi_{j0}(t) = 2^{j/2}\Psi(2^j t)$, $j \in Z$, and shifts $\Psi_{jk}(t) \equiv \Psi_j(t-k2^{-j}) = 2^{j/2}\Psi(2^j t-k)$, $k \in Z$.

The examined integral transforms are a partial case of *chain transforms* (at n=2) for which

$$f_{i+1}(x) = \int_0^\infty f_i(t) K_i(xt)\,dt, \quad i = 1, 2, \ldots, n,$$

and $f_{n+1}(x) = f_1(x)$. This sequence of integral transforms is the chain of integral transformation.

3. USING INTEGRAL TRANSFORMS IN PROBLEMS OF OSCILLATION THEORY

3.1. Electrical oscillations

We examine electrical oscillations in a circuit containing resistance R, inductance L, capacitance C and a source of EMF $e(t)$. To determine the charge q on the plates of a condenser, we use the following differential equation:

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = e(\tau). \tag{43}$$

It is assumed that at the initial moment of time the charge on the condenser plates is equal to q_0 and the current, flowing in the circuit, is i_0. We use the Laplace transform and introduce maps $Q(p)$ and $S(p)$ of the functions q, s. It may easily be seen that

$$\int_0^\infty e^{-pt}\frac{dq}{dt}\,dt = -q_0 + pQ(p), \quad \int_0^\infty e^{-pt}\frac{d^2q}{dt^2}\,dt = -i_0 - pq_0 + p^2Q(p).$$

Therefore, after multiplying both parts of equation (43) by e^{-pt} and integrating in respect of t from 0 to ∞, solving the resultant algebraic equation and using the inversion formula for the Laplace transform, we obtain the final solution of the examined problem

$$q(t) = \exp(-Rt/2L)\left[q_0\cos\omega t + \frac{1}{\omega}\left(i_0 + \frac{Rq_0}{2L}\right)\sin\omega t\right] + \\ + \frac{1}{\omega L}\exp(-Rt/2L)\int_0^t e(\tau)\exp(R\tau/2L)\sin\omega(t-\tau)\,d\tau. \tag{44}$$

In particular, if the resistance of the circuit is equal to zero, i.e. if $R = 0$, then this equality has the form

$$q(t) = q_0\cos\omega t + \frac{i_0}{\omega}\sin\omega t + \frac{1}{\omega L}\int_0^t e(\tau)\sin\omega(t-\tau)\,d\tau. \tag{45}$$

3.2. Transverse vibrations of a string

Small free transverse vibrations of a string are described by the equation

$$\frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}, \tag{46}$$

where $u(x,t)$ is the deviation of the string at a point x at the moment t from the equilibrium position.

We examine the movement of an infinite string $-\infty < x < \infty$ in the absence of external forces. We define the initial conditions $u(x,0) = \varphi(x)$, $\partial u(x,0)/\partial r = \psi(x)$. In order to solve equation (46), we use the Fourier transform. We multiply (46) by $e^{-i\xi x}$, and integrate in respect of x. Consequently, assuming that u and $\partial u/\partial x$ tend to zero at $|x| \to\infty$, and introducing the notations

$$U(\xi,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x,t) e^{-i\xi x} dx, \tag{47}$$

equation (46) maybe written in the new form

$$\frac{d^2U}{dt^2} + a^2\xi^2 U = 0. \tag{48}$$

Thus, using the Fourier transform (47), the problem of solving the differential equation in the partial derivatives can be reduced to the problem of solving the ordinary differential equation (48). Using the initial conditions, we obtain the solution of this equation:

$$U(\xi,t) = \frac{1}{2}\Phi(\xi)(e^{iat\xi} + e^{-iat\xi}) + \frac{\psi(\xi)}{2ia\xi}(e^{iat\xi} - e^{iat\xi}),$$

where Φ, Ψ are Fourier images of the initial functions φ and ψ. The relationship between the function u and U is expressed by the Fourier inversion formula

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} U(\xi,t) e^{i\xi x} d\xi$$

Therefore, replacing function $U(\xi,t)$ by its value, we obtain

$$u(x,t) = \frac{1}{2}\left\{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Phi(\xi)[e^{i\xi[x-at]} + e^{i\xi(x+at)}]d\xi\right\} +$$

$$+ \frac{1}{2a}\left\{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\psi(\xi)}{i\xi}[e^{i\xi(x-at)} - e^{i\xi(x-at)}]d\xi\right\}. \tag{49}$$

Since $\varphi(x)$ and $\Phi(\xi)$ are linked together by the Fourier transform, then

$$\varphi(x \pm at) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} \Phi(\xi) e^{i\xi(x\pm at)} d\xi. \tag{50}$$

From the same considerations, we obtain

$$\int_{x-at}^{x+at} \psi(y)dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\Psi(\xi)}{i\xi}\left[e^{i\xi(x+at)} - e^{i\xi(x-at)}\right]d\xi. \tag{51}$$

Substituting (50), (51) into expression (49), we finally obtain the solution

$$u(x,t)=\frac{1}{2}[\varphi(x+at)+\varphi(x-at)]+\frac{1}{2a}\int_{x-at}^{x+at}\psi(y)dy. \tag{52}$$

We now examine a case in which a string is fixed at the origin and stretched along the positive half axis $x \geq 0$. Free transverse vibrations are determined using the equation

$$\frac{1}{a^2}\frac{\partial^2 u}{\partial t^2}=\frac{\partial^2 u}{\partial x^2}, \quad x \geq 0, \tag{53}$$

and it is assumed that $u(x,0) = \varphi(x)$ and $\partial u/\partial t|_{t=0} = \psi(x)$. Since the deviations at $x = 0$ are zero, it is convenient to use the Fourier sine transform ($\sin 0 = 0$). Both parts of equation (53) are multiplied by $\sin(\xi x)$ and integrated in respect of x from 0 to ∞.

Assuming that when x tends to infinity u and $\partial u/\partial x$ tends to zero, we can write

$$\left(\frac{2}{\pi}\right)^{1/2}\int_0^\infty \frac{\partial^2 u}{\partial x^2}\sin(\xi x)dx=-\xi^2 U_s(\xi),$$

where

$$U_s(\xi,t)=\left(\frac{2}{\pi}\right)^{1/2}\int_0^\infty u(x,t)\sin(\xi x)dx.$$

Thus, equation (53) is equivalent to the equation

$$\frac{1}{a^2}\frac{d^2 U_s}{dt^2}+\xi^2 U_s=0,$$

whose solution is

$$U_s(\xi,t)=\Phi_s(\xi)\cos(a\xi t)+\frac{\Psi_s(\xi)}{a\xi}\sin(a\xi t),$$

where Φ_s, Ψ_s are the sine transforms of the initial functions.

Using the inversion formula for the Fourier sine transform, we obtain,

$$u(x,t)=\frac{1}{\sqrt{2\pi}}\int_0^\infty \Phi_s(\xi)[\sin\xi(x+at)+\sin\xi(x-at)]d\xi+$$

$$+\frac{1}{a\sqrt{2\pi}}\int_0^\infty \Psi_s(\xi)[\cos\xi(x-at)-\cos\xi(x+at)]\frac{d\xi}{\xi}.$$

If $x > at$, from equation (8), (9) we obtain, as in the case of (50), (51)

$$\varphi(x\pm at)=\left(\frac{2}{\pi}\right)^{1/2}\int_0^\infty \Phi_s(\xi)\sin\xi(x\pm at)\,d\xi, \tag{54}$$

$$\int_{x-at}^{x+at}\psi(y)dy=\left(\frac{2}{\pi}\right)^{1/2}\int_0^\infty \frac{\Psi_s(\xi)}{\xi}\left[\cos\xi(x-at)-\cos\xi(x+at)\right]d\xi. \tag{55}$$

Consequently, at $x \geq at$ the solution $u(x,t)$ is expressed by equation (52).

However, if $x < at$, then in (54), (55) we make substitution $\sin(x-at) = -\sin(at-x)$ and taking into account the evenness of the cosine, we obtain the solution:

$$u(x,t) = \frac{1}{2}[\varphi(x+at) - \varphi(at-x)] + \frac{1}{2a}\left\{\int_0^{at-x} \psi(y)dy + \int_0^{x+at} \psi(y)dy\right\}.$$

It should be mentioned that if the string is not fixed at zero, which corresponds to the boundary-value condition of the second kind $\partial u/\partial x|_{x=0} = 0$, it is necessary to use the Fourier cosine transform (instead of sine transform) because the cosines satisfy this boundary-value condition.

3.3. Transverse vibrations of an infinite circular membrane

Free two- and three-dimensional vibrations are described by the equation

$$u_{tt}(x,y,t) = a^2 \Delta u, \tag{56}$$

where Δ is the Laplace operator which has the form $\Delta u = u_{xx} + u_{yy}$ in the Cartesian coordinates in the two-dimensional case. In this case, when the vibrations take place symmetrically in relation to the axis passing through the origin of the coordinates normal to the plane, coinciding with the membrane in the equilibrium position, it is convenient to transfer to the polar coordinates r,φ. Because of symmetry, there is no dependence on angle φ and equation (56) has the form

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} = \frac{1}{a^2}\frac{\partial^2 u}{\partial t^2}. \tag{57}$$

This equation describes the free oscillations in the case of their symmetric distribution.

In order to solve equation (57), we introduce the Hankel image $U(\xi,t)$ of displacement $u(r,t)$. It may easily be seen that

$$\int_0^\infty r\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r}\right) J_0(\xi r)dr = -\xi^2 U(\xi,t), \tag{58}$$

assuming that $r \cdot \partial u/\partial r$ tends to zero at $r = 0$ and $r = \infty$. Multiplying both parts of equation (57) by $rJ_0(\xi r)$ and integrating in respect of r from 0 to ∞, we obtain an ordinary differential equation

$$\frac{d^2 U}{dt^2} + a^2\xi^2 U = 0,$$

which has the solution

$$U(\xi,t) = A(\xi)\cos(a\xi t) + B(\xi)\sin(a\xi t). \tag{59}$$

Let at the initial moment $t = 0$ be $u = \varphi(r), \partial u/\partial t = \psi(r)$. Consequently, $A = \Phi(\xi)$, $B = \Psi(\xi)/(a\xi)$. Therefore, substituting these values into the solution (59) and using the Hankel inversion formula (22), we obtain

$$u(r,t)=\int_0^\infty \xi\Phi(\xi)\cos(a\xi t)J_0(\xi r)d\xi+\frac{1}{a}\int_0^\infty \Psi(\xi)\sin(a\xi t)J_0(\xi r)d\xi. \tag{60}$$

In order to express $u(r,t)$ explicitly through the functions $\varphi(r)$ and $\psi(r)$, we may use the Parseval theorem for the Hankel images. In this case, it is necessary to calculate the integrals of the type

$$\int_0^\infty \xi J_0(\xi\eta)J_0(\xi r)\cos(a\xi t)d\xi,$$

which are rather complicated. However, a general solution can be found by a different method. For example, using the multiple Fourier transform, as in the previous considerations, we obtain the following formula for solving the two-dimensional equation (56)

$$u(x,y,t)=-\frac{1}{2\pi a}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\partial}{\partial t}\frac{\varphi(\alpha,\beta)\,d\alpha\,d\beta}{[a^2t^2-(x-\alpha)^2-(y-\beta)^2]^{1/2}}-$$

$$-\frac{1}{2\pi a}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\psi(\alpha,\beta)\,d\alpha\,d\beta}{[a^2t^2-(x-\alpha)^2-(y-\beta)^2]^{1/2}}, \tag{61}$$

where $u(x,y,0)=\varphi(x,y)$, $u_t(x,y,0)=\psi(x,y)$.

4. USING INTEGRAL TRANSFORMS IN HEAT CONDUCTIVITY PROBLEMS

4.1. Solving heat conductivity problems using the Laplace transform

We examine a classic heat conductivity problem for a semi-infinite solid $x>0$ on the condition that the boundary $x=0$ is maintained at constant temperature T, and the initial temperature is equal to 0.

Let $u(x,t)$ be the temperature at the point x at the time t, and k be the thermal diffusivity coefficient. The problem is reduced to solving the differential equation in partial derivatives

$$\frac{\partial u(x,t)}{\partial t}=k\frac{\partial^2 u(x,t)}{\partial x^2},\quad x>0,\quad t>0, \tag{62}$$

at the boundary-value condition $u(0,t)=T$ and the zero initial condition. Multiplying the differential equation and the boundary-value condition by the kernel e^{-pt} of the Laplace transform and integrating in respect of t in the range from 0 to ∞, we obtain the following equation for the Laplace image $U(P,t)=\int_0^\infty e^{-pt}u(x,t)dt$:

$$k\frac{d^2U}{dx^2}=pU,\quad x>0, \tag{63}$$

and the appropriate boundary-value condition

$$U(p,t)=\int_0^\infty e^{-pt}T\,dt=\frac{T}{P}. \tag{64}$$

Thus, the problem is reduced to solving an *ordinary differential equation*. The bounded at infinity solution of this equation under the condition (64) has the form

$$U(p,t)=\frac{T}{p}e^{-x\sqrt{p/k}}. \tag{65}$$

Transferring from U to the function u using the inversion formula or using Laplace transform tables gives

$$u=T\operatorname{erfc}\left(\frac{x}{2\sqrt{kt}}\right). \tag{66}$$

This is the required solution of the problem. Here erfc x denotes the complementary error function determined by the integral

$$\operatorname{erfc} x=1-\operatorname{erf} x=\frac{2}{\sqrt{\pi}}\int_x^\infty e^{-u^2}du. \tag{67}$$

4.2. Solution of a heat conductivity problem using Fourier transforms

We examine a heat conductivity problems for a semi-infinite solid $x > 0$ on the condition that the boundary $x = 0$ is held at constant temperature T and the initial temperature is equal to zero. This problem was solved in section 4.1 using the Laplace transform. As shown by physical considerations, $u\to 0$ and $\partial u/\partial x\to 0$ at $x\to\infty$. The given boundary condition is the boundary-value condition of the first kind. Therefore, the problem can be solved by the Fourier sine transform $U(\xi,t)=\int_0^\infty \sin(\xi x)u(x,t)dx$. Multiplying the differential equation (62) by the kernel $\sin(\xi x)$ and integrating in respect of x in the range from 0 to ∞, we obtain an auxiliary equation

$$\frac{dU}{dt}=k(\xi T-\xi^2U),\quad t>0, \tag{68}$$

with the zero initial condition. Thus, the sine transform again reduces the solution of our problem to solving an ordinary differential equation. The solution of this equation, bounded at $t > 0$ and satisfying the initial condition, has the form $U=T\left(1-e^{-\xi^2kt}\right)/\xi$. The inversion formula gives

$$u(x,t)=\frac{2T}{\pi}\int_0^\infty(1-e^{-\xi^2kt})\sin(x\xi)\frac{d\xi}{\xi}=T\left[1-\frac{2}{\pi}\int_0^\infty e^{-\xi^2kt}\sin(x\xi)\frac{d\xi}{\xi}\right], \tag{69}$$

which coincides with the previously obtained representation (66). It should be mentioned that the transition from U to function u in the case of the sine transform is far easier than when using the Laplace transform.

4.3. Temperature regime of a spherical ball

It is assumed that the homogeneous ball with the unit radius is symmetrically heated from the bottom for a period of time so that we can examine the stationary temperature regime. Consequently, to determine the temperature u inside the ball we obtain an internal Dirichlet problem, and in the spheric coordinate system

$$\{(r,\theta,\varphi):\ \ 0 \le r < 1,\quad 0 \le \theta \le \pi,\quad 0 \le \varphi < 2\pi\}$$

the dependence on the 'longitude' φ does not exist. Hence $u = u(r,\theta)$. Assuming that $\mu = \cos\theta$, the Laplace equation can be written in the following form

$$r\partial_r^2(ru) + \partial_\mu(1-\mu^2)\partial_\mu u = 0,\quad u = u(r, \arccos\mu),\quad 0 < r < 1,\quad -1 < \mu < 1.$$

The boundary-value condition has the form $u(1, \arccos\mu) = f(\mu)$. We use the Legendre transform in respect of μ for the Laplace equation.Transposing operations $\mathcal{T}_2$ and $r\partial_r^2$ and assuming

$$U(r,n) = \mathcal{T}_2 u(r, \arccos\mu),\quad n = 0,1,2,...,$$

we obtain the equation

$$r\partial_r^2(rU) - n(n+1)U = 0,$$

whose bounded solution has the form $U(r,n) = A(n)r^n$. The functions $A(n)$ are determined from the boundary-value conditions transformed according to Legendre, $A(n) = \mathcal{T}_2 f(t)$. This leads to the equality

$$U(r,n) = r^n \int_{-1}^{1} f(\mu)P_n(\mu)d\mu \equiv (f, \psi_n)r^n,\quad n = 0,1,2,...,$$

where $\psi_n(\mu) = \sqrt{n+1/2}P_n(\mu)$, P_n are the Legendre polynomials. Using the inverse Legendre transform, we obtain the solution

$$u(r,\theta) = u(r, \arccos\mu) = \sum_{n=0}^{\infty}(f, \psi_n)r^n\psi_n(\cos\theta).$$

5. USING INTEGRAL TRANSFORMATIONS IN THE THEORY OF NEUTRON DIFFUSION

The stationary equation of the transfer of neutrons with some simplifying assumptions may lead to the equation:

$$\frac{\partial u(x,\tau)}{\partial\tau} = \frac{\partial^2 u(x,\tau)}{\partial x^2} + S(x,\tau),\quad x \in \mathrm{R}^1, \tau > 0, \tag{70}$$

where the function S describes the sources of neutrons, the required quantity $u(x,t)$ is concentration of neutrons per unit time, reaching the age τ; consequently, u is the *density of deceleration*, τ denotes the symbolic age of the neutrons.

5.1. The solution of the equation of deceleration of neutrons for a moderator of infinite dimensions

We examine the solution of the partial differential equation (70) for the case in which the medium is unbounded and the function of the source in the terms of generalized functions has the form $S\delta(x)\delta(t)$, S = const. At $\tau = 0$ the density of deceleration $u(x,\tau)$ is matched with equation (70) and is equal to $S\delta(x)$. The boundary-value condition, imposed on the density of deceleration, shows that the density should tend to zero with $|x|$ tending to infinity.

The solution of (70) can be found introducing the Fourier map $U(\xi,\tau)$ of the density of deceleration $u(x,\tau)$. Taking into account the behaviour of the density of deceleration at infinity and integrating by parts, we obtain that the equation (70) is equivalent to the ordinary differential equation:

$$\frac{dU}{d\tau}+\xi^2 U=\frac{1}{\sqrt{2\pi}}\delta(\tau).$$

The solution is equal to $U(\xi,\tau)=(2\pi)^{-1/2}e^{-\xi^2\tau}$. Using the inversion theorem, we obtain the solution

$$u(x,\tau)=\frac{1}{2\sqrt{\pi\tau}}e^{-x^2/4\tau}. \tag{71}$$

5.2. The problem of diffusion of thermal neutrons

When the neutrons reach a specific velocity, they cease to lose energy and their motion can be described using the classic theory of diffusion. If $\rho(\mathbf{r},t)$ denotes the density of neutrons, we obtain the diffusion equation

$$\Delta\rho-\frac{\rho}{\Lambda^2}=\frac{\tau}{\Lambda^2}\frac{\partial\rho}{\partial t}-\frac{\tau}{\Lambda^2}q(\mathbf{r}), \tag{72}$$

where Λ is the *diffusion length*. Therefore, in a one-dimensional stationary case (ρ is independent of time) we have the equation

$$\frac{d^2\rho}{dz^2}-\frac{\rho}{\Lambda^2}=-\frac{\tau}{\Lambda^2}q(z). \tag{73}$$

If the medium is infinite, we can solve this equation introducing Fourier images for ρ and q; consequently, the equation has the form

$$R(\xi)=\frac{\tau}{\Lambda^2}\frac{Q(\xi)}{\xi^2+1/\Lambda^2}.$$

Using the inversion theorem we have

$$\rho(z)=\frac{\tau/\Lambda^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{Q(\xi)e^{i\xi z}}{(\xi^2+1/\Lambda^2)}d\xi. \tag{74}$$

The value of this integral on the basis of the theorem of convolution for the Fourier transform can be expressed through the value of the function $q(z)$. Consequently, we obtain

$$\rho(z)=\frac{\tau}{2\Lambda}\int_{-\infty}^{\infty}q(u)e^{-|z-u|/\Lambda}du. \tag{75}$$

6. APPLICATION OF INTEGRAL TRANSFORMATIONS TO HYDRODYNAMIC PROBLEMS

6.1. A two-dimensional vortex-free flow of an ideal liquid

We examine a vortex-free two-dimensional flow of an ideal liquid, filling the half plane $y \geq 0$, and the velocity components are denoted by u and v. It is well known that:

$$u=-\frac{\partial\varphi}{\partial x},\quad v=-\frac{\partial\varphi}{\partial y}, \tag{76}$$

where the scalar velocity potential satisfies the two-dimensional Laplace equation

$$\Delta\varphi\equiv\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}=0. \tag{77}$$

It is assumed that the liquid flows into the examined half-plane through the segment $|x| \leq a$ of the boundary $y = 0$. Initially, it is assumed that the liquid flows in with the given velocity directed normal to the segment through which it flows in. Thus, the boundary condition along the line $y = 0$ has the form

$$\frac{\partial\varphi}{\partial y}=\begin{cases}-f(x), & 0<|x|<a,\\ 0, & |x|>a,\end{cases} \tag{78}$$

where $f(x)$ is the given function of the point. In addition, we assume that at a large distance from the line $y = 0$ the liquid is in the still condition, i.e. $(u, v) \to 0$ at $\to \infty$. It may be easily shown that the equations (77) and (78) are reduced to an ordinary differential equation

$$\frac{d^2\Phi}{dy^2}-\xi^2\Phi=0$$

with the boundary condition $d\Phi(\xi, 0)/dy = -F(\xi)$. Here Φ, F are the Fourier images of the functions φ, f. Evidently, the solution of the examined problem has the form $\Phi=(F(\xi)e^{-|\xi|y})/|\xi|$, from which, on the basis of the conversion theorem, we obtain

$$\varphi(x,y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{F(\xi)}{|\xi|}e^{i\xi x-|\xi|y}d\xi. \tag{79}$$

In particular, if

$$f(x)=\begin{cases}U, & 0<|x|<a,\\ 0, & |x|>a,\end{cases}$$

then

$$F(\xi)=\frac{U}{\sqrt{2\pi}}\frac{\sin\xi a}{\xi},$$

and consequently

$$\varphi(x,y)=\frac{U}{2\pi}\int_{-\infty}^{\infty}\frac{\sin(\xi a)}{\xi}\frac{e^{i\xi x}}{|\xi|}e^{-|\xi|y}d\xi.$$

The component of the velocity of the liquid in the direction of the axis y gives the expression

$$v(x,y)=\frac{\partial\varphi}{\partial y}=\frac{U}{2\pi}\int_{-\infty}^{\infty}\frac{\sin(\xi a)}{\xi}e^{-|\xi|y+i\xi x}d\xi.$$

Taking into account the value of the integral $\int_0^{\infty}\frac{\sin\xi a}{\xi}e^{-\xi y}d\xi=\frac{\pi}{2}-\operatorname{arctg}\frac{y}{a}$, we obtain

$$v(x,y)=\frac{U}{2\pi}(\theta_1-\theta_2),\quad \theta_1=\operatorname{arctg}\frac{y}{x-a},\quad \theta_2=\operatorname{arctg}\frac{y}{x+a}.$$

Using the same method for the component of the velocity of the liquid in the direction of the x axis we obtain the expression

$$u(x,y)=-\frac{\partial\varphi}{\partial x}=\frac{U}{2\pi}\ln\frac{r_2}{r_1},$$

where $r_2^2=(x+a)^2+y^2$ and $r_1^2=(x-a)^2+y^2$.

If we introduce the complex potential $w(z)=\varphi+i\psi$, $z=x+iy$, then

$$\frac{dw}{dz}=\frac{\partial\varphi}{\partial x}-i\frac{\partial\varphi}{\partial y}=-u+iv, \tag{80}$$

and consequently, taking into account the values of the velocity components, we obtain

$$\frac{dw}{dz}=\frac{U}{2\pi}\ln\frac{z-a}{z+a}.$$

Integrating this expression in respect of z, we obtain the expression for the complex potential:

$$w(z)=\frac{U}{2\pi}[2a+(z-a)\ln(z-a)-(z+a)\ln(z+a)].$$

6.2. The flow of the ideal liquid through a slit

We examine a steady two-dimensional flow of the ideal liquid through a slit in a rigid flat boundary. The origin of the co-ordinates is placed in the centre of the slit, and the z axis is normal to the plane of the barrier. Consequently, the problem is reduced to solving the differential equation (77) for the following boundary-value conditions:

at $y = 0$ and $|x| \le a$

$$\varphi = \text{const}$$

at $y = 0$ $|x| > a$

$$v = -\frac{\partial\varphi}{\partial y} = 0.$$

If at $y = 0$

$$v = \begin{cases} (a^2 - x^2)^{-1/2}, & 0 < |x| < a, \\ 0, & |x| < a, \end{cases}$$

then on the basis of the previous results it is concluded that the velocity potential is given by the equation

$$\varphi(x, y) = \frac{1}{2}\int_{-a}^{a} \frac{J_0(a\xi)}{|\xi|} e^{i\xi x - |\xi| y}\, d\xi,$$

and, consequently, from equations (76) we obtain the following expressions for the components of the velocity vector

$$v(x, y) = -\frac{\partial\varphi}{\partial y} = \int_0^\infty e^{-\xi y} J_0(a\xi)\cos(\xi x)\, d\xi, \tag{81}$$

$$u(x, y) = -\frac{\partial\varphi}{\partial x} = \int_0^\infty e^{-\xi y} J_0(a\xi)\sin(\xi x)\, d\xi. \tag{82}$$

Thus, at $y = 0$

$$\frac{\partial\varphi}{\partial x} = -\int_0^\infty \sin(\xi x) J_0(a\xi)\, d\xi = 0, \quad \text{if} \quad |x| < a,$$

and this means that φ is a constant in the segment $y = 0$, $|x| < a$. Substituting the expressions for the components for the velocity into equation (80), we obtain that the complex potential of the flow in the examined case is the solution of the equation

$$\frac{dw}{dz} = i\int_0^\infty e^{i\xi z} J_0(a\xi) d\xi = (z^2 - a^2)^{-1/2}.$$

Integrating this equation gives $z = a \operatorname{ch} w$.

6.3. Discharge of the ideal liquid through a circular orifice

Let the origin of the co-ordinate be situated in the centre of a circular orifice and the axis z be normal to the plane of a thin rigid screen. We use cylindrical coordinates r, z. The solution of the problem of the steady flow of the liquid is reduced to determining the potential of velocities $\varphi(r, z)$, satisfying the Laplace equation in the examined coordinates, i.e.

$$\frac{\partial^2\varphi}{\partial r^2} + \frac{1}{r}\frac{\partial\varphi}{\partial r} + \frac{\partial^2\varphi}{\partial z^2} = 0 \tag{83}$$

under the boundary-value conditions on the plane $z = 0$

$$\varphi = g(r), \quad r < a, \quad \frac{\partial \varphi}{\partial z} = 0, \quad r > a. \tag{84}$$

Function $g(r)$ is the given function.

Both parts of the equation (83) are multiplied by $rJ_0(\xi r)$ and integrated in respect to r from 0 to ∞. We find that this equation is equivalent to the ordinary differential equation of the second order:

$$\frac{d^2\Phi}{dz^2} - \xi^2\Phi = 0, \tag{85}$$

where Φ (ξ, z) is the Hankel map of the velocity potential. If the liquid flows into the half space $z \geq 0$ through the examined orifice, the velocity potential should tend to 0 at $z \to \infty$ and, consequently, the solution of equation (85) should be taken in the form:

$$\Phi = A(\xi)e^{-\xi z}, \tag{86}$$

where $A(\xi)$ should be determined using conditions (84). Differentiating the last equality in respect of z we obtain

$$\int_0^\infty r\frac{\partial \varphi}{\partial z} J_0(\xi r)\, dr = -\xi A(\xi)e^{-\xi z}.$$

Using the inversion theorem for the Hankel transform, we obtain the relationship

$$\varphi(r,z) = \int_0^\infty \xi A(\xi)e^{-\xi z} J_0(\xi r)\, d\xi, \quad \frac{\partial \varphi(r,z)}{\partial z} = -\int_0^\infty \xi^2 A(\xi)e^{-\xi z} J_0(\xi r)\, d\xi.$$

Substituting these relationships into the conditions (84) and assuming that

$$\rho = \frac{r}{a}, \quad A_1(u) = uA\left(\frac{u}{a}\right), \quad g_1(\rho) = a^2 g(r),$$

we obtain the following dual integral equations for determining the function $F(u)$:

$$\int_0^\infty A_1(u)J_0(\rho u)du = g_1(\rho), \quad 0 < \rho < 1; \quad \int_0^\infty uA_1(u)J_0(\rho u)du = 0, \quad \rho > 1.$$

Function $A(\xi)$ is expressed through $A_1(u)$ using the formula $A(\xi) = A_1(a\xi)/(a\xi)$. The solution of the system of equations for $A_1(u)$ has the form

$$A_1(u) = \frac{2}{\pi}\cos u \int_0^1 \frac{yg_1(y)dy}{(1-y^2)^{1/2}} + \frac{2}{\pi}\int_0^1 \frac{ydy}{(1-y^2)^{1/2}} \int_0^1 g_1(yx)xu\sin(xu)dx.$$

In a partial case, when function $g_1(\rho)$ is reduced to constant C, we obtain

$$A_1(u) = \frac{2C}{\pi}\frac{\sin u}{u}.$$

In this case, function $g(r)$ is equal to constant γ, where $\gamma = C/a^2$, and

$$A(\xi) = \frac{2\gamma \sin(\xi a)}{(\pi\xi^2)}.$$

Consequently,

$$\varphi(r,z)=\frac{2\gamma}{\pi}\int_0^\infty \frac{\sin\xi a}{\xi}e^{-\xi z}J_0(\xi r)d\xi. \tag{87}$$

The same procedure can be used for examining a case in which the value $\partial\varphi/\partial z$ is given for the entire plane $z = 0$. If $\partial\varphi/\partial\zeta = -f(r)$ (it is assumed that $\partial\varphi/\partial z$ converts to zero for the values of r exceeding a), and if the Hankel image of the zero order of this function is denoted by

$$F(\xi)=\int_0^a rf(r)J_0(\xi r)dr,$$

then it is easy to show that $\Phi(\xi) = -F(\xi)/\xi$. Consequently

$$\varphi(r,z)=-\int_0^\infty F(\xi)e^{-\xi z}J_0(\xi r)d\xi.$$

If the orifice is sufficiently small, we can set in the terms of the generalised functions $f(r) = S\delta(r)/(2\pi r)$, and the Hankel image of this function is equal to $F(\xi) = S/(2\pi)$. Finally, we obtain

$$\varphi(r,z)=-\frac{S}{2\pi}\int_0^\infty e^{-\xi z}J_0(\xi r)\,d\xi=-\frac{S}{2\pi(r^2+z^2)^{1/2}}. \tag{88}$$

7. USING INTEGRAL TRANSFORMS IN ELASTICITY THEORY

7.1. Axisymmetric stresses in a cylinder

We examine the stresses formed in the an unbounded circular cylinder with the radius equal to unity, when the normal stress, equal to unity at $z > 0$ and to zero at $z < 0$, is applied to the side surface of the cylinder. The method of solving this problem can be used with small modifications for solving a more complicated problem in which the applied stresses represent exponential-type functions. We use the cylindrical coordinate system. The axis of the cylinder is combined with the axis z. The origin of the coordinates is placed in the central cross section. We introduce the notations for the stresses:

1) σ_r – normal stresses in the direction of radii, or radial stresses;
2) σ_0 – normal stresses in the normal direction, or tangential stresses;
3) σ_z – are the normal stresses parallel to the axis z;
4) τ_{rz} – are the tangential stresses directed parallel to the axis z in tangential planes to the cylinder.

According to symmetric conditions, there cannot be any other tangential stresses. It is known that these stresses may be expressed by means of the stress function u from the equations

$$\sigma_r=\frac{\partial}{\partial z}\left(\sigma\nabla^2-\frac{\partial^2}{\partial r^2}\right)u,\quad \tau_{rz}=\frac{\partial}{\partial r}\left((1-\sigma)\nabla^2-\frac{\partial^2}{\partial z^2}\right)u, \tag{89}$$

$$\sigma_0 = \frac{\partial}{\partial z}\left(\sigma\nabla^2 - \frac{1}{r}\frac{\partial}{\partial r}\right)u, \quad \sigma_z = \frac{\partial}{\partial z}\left((2-\sigma)\nabla^2 - \frac{\partial^2}{\partial z^2}\right)u. \tag{90}$$

Here σ = const denotes the Poisson coefficient for the material of the cylinder, $u = u(r, z)$ is the function of stresses satisfying the biharmonic equation

$$\nabla^4 u = 0, \quad -\infty < z < \infty, \quad r < 1. \tag{91}$$

The solution of the problem is reduced to integrating this equation for the boundary conditions:

$$\sigma_r(1,z) = -1, \ 0 < z < \infty, \quad \sigma_r(1,z) = 0, \quad -\infty < z < 0, \tag{92}$$

$$\tau_{rz}(1,z) = 0, \quad -\infty < z < +\infty, \tag{93}$$

where σ_r, τ_{rz} depend on the function of stresses u in accordance with (89).

In this case, the required function u does not disappear when $z \to +\infty$. Therefore, we use the complex Fourier transform

$$U(r,\xi) = \int_{-\infty}^{\infty} e^{-i\xi z} u(r,z)dz,$$

where the imaginary part of ξ is negative in order to ensure that the given integral converges. Multiplying the equation and the boundary-value conditions on the kernel $e^{-i\xi z}$, integrating in respect of z from $-\infty$ to $+\infty$ and denoting $L = \frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} - \xi^2$, we obtain an auxiliary ordinary differential equation of the fourth order for $U(r,\xi)$

$$L^2\bar{u} = 0, \quad r < 1, \tag{94}$$

and the appropriate boundary conditions

$$\left[\sigma L - \frac{d^2}{dr^2}\right]U(1,\xi) = \xi^{-2}, \quad \frac{d}{dr}[(1-\sigma)L + \xi^2]U(1,\xi) = 0, \tag{95}$$

The solution of equation (94), bounded at $r = 0$, has the following form

$$U(r,\xi) = AI_0(\xi r) + B\xi r I_1(\xi r), \tag{96}$$

where A, B are constants. Using the recurrent equation for the Bessel function, we obtain

$$LU(r,\xi) = 2B\xi^2 I_0(\xi r). \tag{97}$$

Substituting (96) and (97) into (95), we determine the constants A and B. Substituting them into (96), we obtain the transformed function of stresses $U(r, \xi)$. To determine the stresses (89), (90) we can use two methods: either use the inversion formula to find $u(r, z)$ and substitute it into (89), (90); or multiply the equalities (89), (90) by $e^{-i\xi z}$, integrate them in respect of z and, consequently find equations for the Fourier images of the stresses, and only after this we can use the inverse Fourier transform. Usually, the second method is preferred; for example, for the map Σ_θ in this case we obtain

$$\sum\nolimits_\theta(r,\xi) = -i\xi\left[\sigma L - \frac{1}{r}\frac{d}{dr}\right]U(r,\xi). \tag{98}$$

The images of the remaining stresses are produced in the same manner.

7.2. Bussinesq problem for the half space

We examine a problem of elasticity theory for the following partial case of the distributed load. To the circular region of the unit radius, situated on the surface $z = 0$ of the elastic half space $z > 0$, we apply a distributed load with the intensity equal to unity. The remaining part of the surface $z = 0$ is free from load. It is required to find the stress on the axis of the circular region at the point situated at distance z from the surface.

As previously, normal stress σ_z and tangential stress τ_{rz} may be expressed through the function of stresses u using equations (89) and (90). The problem is reduced to integrating the biharmonic equation (91) at the boundary conditions

$$\sigma_z(r,0) = -1, \quad 0 < r < 1, \quad \sigma_z(r,0) = 0, \quad r > 1, \tag{99}$$

$$\tau_{rz}(r,0) = 0, \quad 0 < r < \infty. \tag{100}$$

Multiplying equation (91) and the boundary conditions (99) by $rJ_0(pr)$ and integrating in respect of r in the range from 0 to ∞, we obtain the ordinary differential equation for the Hankel image $U(\xi,z) = \int_0^\infty rJ_0(\xi r)u(r,z)\,dr$

$$\left(\frac{d^2}{dz^2} - \xi^2\right)^2 U(\xi,z) = 0, \tag{101}$$

and, taking into account (90), (99), the appropriate boundary-volume condition, $z = 0$

$$(1-\sigma)\frac{d^3U}{dz^3} - (2-\sigma)\xi^2\frac{dU}{dz} = -\int_0^1 rJ_0(\xi r)\,dr = -\frac{J_1(\xi)}{\xi}. \tag{102}$$

The equalities (89_2), (100) give the second boundary-value condition at $z = 0$

$$\sigma\frac{d^2U}{dz^2} + (1-\sigma)\xi^2 U = 0. \tag{103}$$

Thus, the problem is reduced to integrating the ordinary differential equation of the fourth order (101) at the boundary conditions (102), (103). The general solution of the differential equation (101), finite at high positive z, has the form $U(\xi, z) = (A + Bz)e^{-\xi z}$. Setting $z = 0$ and substituting the resultant expressions into (102), (103), we obtain $A = -2\sigma\xi^{-4}J_1(\xi)$, $B = -\xi^{-3}J_1(\xi)$. Consequently

$$\Sigma_z(\xi,z) = -(z+\xi^{-1})J_1(\xi)e^{-\xi z}. \tag{104}$$

Using now the inversion formula for the Hankel transform, we get

$$\sigma_z = -\int_0^\infty (1+\xi z)\,e^{-\xi z}J_1(\xi)J_0(\xi r)\,d\xi. \tag{105}$$

In particular, at $r = 0$ we have $\sigma_z = -1 + z^3(z^2 + 1)^{-3/2}$. Stresses σ_r and σ_0 are determined by the same procedure. The image of the stress τ_{rz} can be produced using the Hankel transform with the kernel $rJ_1(pr)$; in this case, the inversion fomula has the kernel $\xi J_1(r\xi)$.

7.3. Determination of stresses in a wedge

It is assumed that each of the flat surfaces of an infinite wedge with the angle 2α is subjected to the effect of a distributed load with a stress equal to the unity along the strip with length a. It is required to find the tangential stresses in the wedge. In this case, the normal and tangential stresses can be expressed by means of the stress function $u(r,\theta)$ as follows:

$$\sigma_\theta = \frac{\partial^2 u}{\partial r^2}, \tag{106}$$

$$\sigma_{r\theta} = -\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial u}{\partial \theta}\right), \tag{107}$$

where function u satisfies the biharmonic equation

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\right)^2 u = 0, \quad 0 < r < \infty, \quad -\alpha < \theta < \alpha. \tag{108}$$

The problem is reduced to solving the equation at the boundary-value conditions

$$\sigma_\theta(r,\pm\alpha) = -1, \quad 0<r<a, \quad \sigma_\theta(r,\pm\alpha) = 0, \quad r>a, \tag{109}$$

$$\sigma_{r\theta}(r,\pm\alpha)=0, \quad 0<r<\infty. \tag{110}$$

It is assumed that the function u is such that

$$r^{p+n}\frac{\partial^n u}{\partial r^n}, \quad r^p\frac{\partial^n u}{\partial \theta^n} \quad (n = 0,1,2) \quad \text{and} \quad r^{p+1}\frac{\partial^3 u}{\partial r \partial \theta^2}$$

tend to zero when $r \to \infty$, and we denote by $U(p, \theta)$ the Mellin transform of the function $u(r, \theta)$:

$$U(p,\theta) = \int_0^\infty r^{p-1} u(r,\theta)\, dr. \tag{111}$$

Multiplying the biharmonic equation by r^{p+3} and integrating in respect of r from 0 to ∞, we obtain the ordinary differential equation in relation to the transform function:

$$\frac{d^4 U}{d\theta^4} + [(p+2)^2 + p^2]\frac{d^2 U}{d\theta^2} + p^2(p+2)^2 U = 0. \tag{112}$$

The appropriate boundary conditions for U are obtained if we substitute equations (106), (107) into (109) and (110), multiply the results by r^{p+1} and integrate in respect of r from 0 to ∞.

The general solution of the given ordinary differential equation (112) has the form:

$$U(p,\theta)=A\sin p\theta + B\cos p\theta + C\sin(p+2)\theta + D\cos(p+2)\theta, \qquad (113)$$

where A, B, C, D depend on p and α. Since the solution should be symmetric in relation to the plane $\theta = 0$, then $A = C = 0$. Constants B, D are determined from the boundary conditions.

We now turn to tangential stresses. According to (107) we have

$$r^2(\sigma_{r\theta}) = \frac{\partial u}{\partial\theta} - r\frac{\partial^2 u}{\partial r\partial\theta}.$$

The Mellin transform of this function is $(p+1)\dfrac{dU}{d\theta}$. After calculations we obtain

$$\frac{\pi r}{a}\sigma_{r\theta} = \int_0^\infty R(p)\cos\left[p\ln\frac{a}{r}\right]dp, \qquad (114)$$

where

$$R(p) = \frac{\sin(\alpha-\theta)\,\mathrm{sh}\,(\alpha+\theta)\,p - \sin(\alpha+\theta)\,\mathrm{sh}(\alpha-\theta)p}{p\sin 2\alpha + \mathrm{sh}\,2\alpha p}. \qquad (115)$$

In particular, at $\alpha = \pi/2$, when the wedge becomes a semi-infinite hard body, the resultant integral can be calculated using the expression:

$$\int_0^\infty \frac{\mathrm{sh}\,qx}{\mathrm{sh}\,(\pi x/2)}\cos(mx)dx = \frac{\sin 2q}{\cos 2q + \mathrm{ch}2m}.$$

Consequently, we obtain

$$\sigma_{r\theta}(r,\theta) = \frac{2ar\cos\theta}{\pi}\left(\frac{a^2\cos 2\theta}{r^4 + 2a^2r^2\cos 2\theta + a^4}\right). \qquad (116)$$

For the remaining values of α, the stress can be determined by approximate calculation of the integral in (114).

8. USING INTEGRAL TRANSFORMS IN COAGULATION KINETICS

8.1. Exact solution of the coagulation equation

We examine a disperse system consisting of a mixture of two phases, with one of the phases distributed in the other one in the form of fine particles (small crystals, droplets, bubbles, etc.). One of the main mechanisms of evolution of this system is the process of coagulation (coalescence) of the particles described by the following *kinetic coagulation equation*:

$$\frac{\partial c(x,t)}{\partial t} = \frac{1}{2}\int_0^x K(x-y,y)c(-y,t)c(y,t)\,dy - c(x,t)\int_0^\infty K(x,y)c(y,t)dy, \qquad (117)$$

with the initial condition

$$c(x,0)=c_0(x)\geq 0, \quad x\geq 0. \tag{118}$$

Here $c(x,t)$ is the function of distribution of the particles with the masses $x \in [0, \infty)$ at the moment of time $t \geq 0$, $K(x, y)$ is the so-called *coagulation kernel* characterising the intensity of coalescence of the particles with masses x and y. It is known from the physics of the process.

To solve the equation, we use the Laplace transform. Let $C(p,t)$ be the Laplace map of the function $c(x,t)$ and let K = const. Multiplying (117) by $\exp(-px)$ and integrating, we obtain the ordinary differential equation for the image

$$\frac{\partial C(p,t)}{\partial t} = \frac{K}{2}C(p,t)^2 - KC(p,t)C(0,t) \tag{119}$$

with the initial condition $C_0(p) = C(p,0)$. Substituting $p = 0$ in (119), we initially obtain $C(0,t)$ and, after that, a solution in the Laplace images

$$C(p,t) = \left(1+\frac{1}{2}tKC_0(0)\right)^{-2}\left(\frac{1}{C_0(p)}-\frac{1}{C_0(0)}+\frac{1}{1+(1/2)tKC_0(0)}\right)^{-1}.$$

Substituting specific initial distributions, we can find the solution of the coagulation equation (117), (118). In particular, at

$$c_0(x) = \exp(-ax)$$

we have

$$c(x,t) = \frac{1}{(1+Kt/2a)^2}\exp\left(-\frac{ax}{1+Kt/(2a)}\right). \tag{120}$$

If it is assumed that the dimensions of particles are multiples of some minimum value with mass x_0 (by analogy with polymers such a particle is referred to as a *monomer*), the integrals in equation (117) are transformed into sums:

$$\frac{dc_i(t)}{dt} = \frac{K}{2}\sum_{j=1}^{i-1} c_{i-j}(t)c_j(t) - Kc_i(t)\sum_{j=1}^{\infty} c_j(t), \quad x\geq 0, \quad t>0, \tag{121}$$

where $c_i(t)$, $i \geq 1$ is the function of distribution of the particles with the mass ix_0. For this discrete equation, the Laplace transform is transformed to constructing the *generating function* $C(z,t)$ which is a discrete analogue of the Laplace transform:

$$C(z,t) = \sum_{j=1}^{\infty} z^{i-1}c_i(t).$$

Multiplying (121) by z^{i-1} and after summation we obtain, for the solution of the resultant differential equation, an expression for the generating function in the form of a series over z^{i-1}. The coefficients at the powers z give the required solution. In particular, if the initial distribution consists only of monomers, i.e. $c_1(0) = A$, $c_i(0) = 0$, $i \geq 2$, then we have the expression identical with (120)

$$c_i(t) = \frac{1}{(1+Kt/(2A))^2} A^i \left(\frac{Kt}{2(1+Kt/(2A))} \right)^{i-1}, \quad i \geq 1.$$

8.2. Violation of the mass conservation law

The key property of the coagulation equation (117) is the validity of the law of conservation of the total mass of the system which is mathematically equal to the first moment of the solution:

$$M_1(t) = \int_0^\infty xc(x,t)dx,$$

Multiplying (117) by x and integrating we obtain, assuming the boundedness of the resultant double integrals, constancy of the total mass: $dM_1/dt = 0$. However, these double integrals are not always bounded. In a number of physical–chemical processes (for example in polycondensation), the coagulation kernel is multplicative: $K(x, y) = Axy$, and it may easily be seen that in this case we have some critical moment of time t_{cr}, at which the second moment of the solution M_2 converts to infinity. To examine the solution of the coagulation equation with the multiplicative kernel, we use the Laplace transform at real p. We denote

$$\frac{\partial C(p,t)}{\partial p} = F(p,t)$$

and take into account that $F(0, t) = -M_1(t)$. Writing the quasi-linear partial differential equation in partial derivatives, formed for the function $F(p, t)$, and solving it by the method of characteristics, we obtain the functional equation

$$F(p,t) = F_0\left(p + A\int_0^t M_1(s)\,ds + AtF(p,t) \right). \tag{122}$$

We take into account that $M_1(t) = -F(0,t)$, introduce a notation for the argument of the function F_0 at $p = 0$

$$\rho(t) = A\int_0^t M_1(s)ds - AtM_1(t).$$

It should be noted that from (122) we have $M_1 = -F_0(\rho)$. Differentiating in respect of t we obtain

$$\frac{d\rho(t)}{dt} = -At\frac{dM_1(t)}{dt}. \tag{123}$$

On the other hand, at $p = 0$ from equation (122) after differentiation in respect of t we obtain the relationship:

$$\frac{dM_1(t)}{dt} = -F_0'(\rho)\dot{\rho}(t). \tag{124}$$

Comparing the last two equalities, we get

$$\dot{\rho}\left(1 - At\,F_0'(\rho)\right) = 0. \tag{125}$$

The special features of the coagulation system depend strongly on whether the equation for

$$1 - At F_0'(\rho) = 0 \tag{126}$$

has a root on the condition $\rho(t) \to 0$, $t \to 0$. If the second moment of the initial distribution is finite, we carry out expansion into a Taylor series $F_0'(p) = (1/2)M_2(0) + o(p)$, $p \to 0$. Substituting this expression to formula (126), we come to a conclusion that equation (126) has no roots for the time period $0 \le t \le t_{cr}$, consequently, in order to fulfil equality (125), it is necessary to set $\rho = 0$. This means that $\rho(t) = 0$, $0 \le t \le t_{cr}$, and, therefore, $M_1 = \text{const}$. A completely different situation arises at $t \ge t_{cr}$ when a unique root forms in the equation (126) and is situated on the real axis and is displaced monotonically with time along this axis to the right (it should be mentioned that the Laplace transform at real p is a monotonically decreasing function). However, if $\rho(t)$ monotonically increases with time, the mass of the system M_1 in accordance with the previously noted dependence $M_1 = -F_0(\rho)$ monotonically decreases at $t > t_{cr}$. Generalising these results, we conclude that the kinetic coagulation equation at the multiplicative kernel for all initial distributions with the finite second moment has a unique continuous non-negative solution whose total mass monotonically decreases, starting from the critical moment of time t_{cr}.

BIBLIOGRAPHIC COMMENTARY

The Fourier transform is used in solving the boundary-value problems of mathematical physics for unbounded regions: plane, half-plane, square, band, half-band, infinite cylinder, half-cylinder, etc. It is also used when solving the integral equations with difference kernels, and the main role is played by convolution theorems. An important role in the theory of singular integral equations is played by the results of Fourier transforms or analytical functions [13, 21, 89, 96].

The Laplace transform is used for solving non-stationary problems by the operational method. In addition to this, the convolution theorem for the Laplace transform makes it possible to solve integral Volterra equations with different kernels [13, 89, 93].

The Mellin transform is used for solving planar harmonic problems in the sectorial domain, in the elasticity theory, in solving singular integral equations at the half-axis with the kernel which depends on the ratio of the arguments, and in solving paired integral equations [93, 96].

The Hilbert transform is used in boundary-value problems of theory of analytical functions. The Hankel transform is used most frequently when solving equations with the Laplace operator for cylindrical regions in the polar or cylindrical coordinates (axisymmetric problems), and the Legendre transform is used when solving equations in spherical coordinates [96].

Chapter 5

METHODS OF DISCRETISATION OF MATHEMATICAL PHYSICS PROBLEMS

Keywords: finite-difference methods, net methods, method of arbitrary lines, quadrature method, variational methods, Ritz method, the method of least squares, Kantorovich method, Courant method, Trefftz method, projection methods, Bubnov–Galerkin method, the method of moments, projection methods in Hilbert spaces, projection methods in Banach spaces, projection-net methods, method of integral identities, the method of Marchuk integral identity, generalised formulation of the method of integral identities, the finite element method.

MAIN DEFINITIONS AND NOTATIONS

Net – a set of points in the domain of definition of the differential equation and boundary conditions on which the appropriate approximation of the problem is constructed.

The difference scheme – a system of equations approximating a differential equation and boundary conditions on the exact solution of the problem.

Sweep method (factorisation) – the method of Gauss exclusion of the solution of the system of linear algebraic equations with a tri-diagonal matrix.

$Au = f$ – symbolic (operator) form of writing the mathematical physics problems with operator A.

$D(A)$ – the domain of definition of the operator A.

$R(A)$ – the domain of the values of the operator A.

Variational problem – problem of finding functions resulting in the extremum of some functional $J(u)$.

Variational (*direct*) *method* – method of finding functions approximately realising the extremum of the functional and not reducing the variational problem to differential equations.

Minimising sequence – the sequence of functions $u_1, u_2, \ldots, u_n, \ldots$ such that $\lim_{n\to\infty}(u_n) = d \equiv \inf_u J(u)$.

Projection methods – a class of methods of finding approximate solutions $u_N = \sum_{i=1}^{N} a_i\varphi_i$, presented in the form $P_N(Au_N - f) = 0$, where P_N is the projector on some system of basis functions $\psi_1, \psi_2, \ldots, \psi_N$.

The method of integral identities – a set of methods of approximate solution of the problem $Au = f$, consisting of constructing a system of integral identities on the basis of the exact formulation of the problem $Au = f$ and subsequent approximation of these identities.

The quadrature method – a method of nets used for approximate solution of integral equations.

The projection-net method (*finite element method*) – modification of the corresponding projection method in which the base functions have carriers of the order of the step of the net.

1. INTRODUCTION

Different sections of computational mathematics, its concepts and approaches are concerned with the construction and examination of numerical methods of solving mathematical physics problems.We shall note one of the features of computational mathematics: in most cases they may only give approximate results.

Another feature of these methods is that in any calculations we can carry out operations only with a finite number of numbers and after calculations obtain only a finite number of results. Therefore, every problem which should be solved numerically, can be reduced in advance to such a form that it is possible to obtain all the results after a finite number of arithmetic actions. In this case, the initial problem is approximately replaced by the solution of a new problem in which the finite number of parameters is not known. Knowledge of these parameters makes it possible to compute approximately the required solution. This process of replacing the solution of the initial problem by a new one with a finite number of unknown parameters is referred to as *discretisation* of the given problem of mathematical physics.

Discretisation of mathematical physics problems may be carried out by many methods which are often also referred to as the appropriate methods of approximate solutions of the initial problem.

We shall discuss several requirements made on the method of discretisation (approximate methods of computational mathematics) from the viewpoint of calculations. One of them is the approximation requirement (for example, the extent to which the initial equation can be accurately approximated by a finite system of equations, whose solutions are then accepted as the approximate

solution of the initial equation). Examination of the approximation problem in the methods of discretisation is closely linked with a special section of mathematics referred to as the theory of approximation of functions which is of considerable importance for computational mathematics.

Another main requirement on the discretisation method is the requirement to find the relevant quantities with the selected degree of accuracy. Of special importance for computations are therefore approximate methods and processes which make it possible to find the results with any required high degree of accuracy. These methods are referred to as converging methods. For example, let it be that u is the exact solution of the examined problem and we also assume that the selected method is used to construct a sequence of approximations $u_1,\ldots,u_N$ to the solution u. Now one of the first problems of the selected method is the determination of the convergence of the requirements to the exact solution $u_N \to u$ and $N \to \infty$, and if this convergence cannot be achieved in all cases, then it is the explanation of the conditions at which it does occur.

When the convergence is established, we obtain a more difficult problem of the evaluation of the *rate of convergence*, i.e. evaluation of the rate at which u_N tends to solution of u at $N \to \infty$. The rate of convergence of the method is one of the factors determining the general calculation losses in solving the problem with the required accuracy. In order to evaluate the rate of convergence of u_N to u, it is often attempted to evaluate the absolute value of the error $u-u_N$, i.e. construct the quantity $\varepsilon(N)$ such that we have $|u - u_N| \le \varepsilon(N)$, and it is referred to as the *estimate of the error*. In this case to ensure that the estimate reflects the actual degree of closeness of u_N to u, it is essential that $\varepsilon(N)$ differs only slightly from $|u - u_N|$. In addition to this, the estimate $\varepsilon(N)$ should be effective, i.e. such that it can be determined, otherwise it cannot be used.

The computations impose another requirement on the theory of approximate methods – *the requirement of stability of the computing process*. The problem arising in this case may be described as follows. Every approximate method leads to some calculation scheme. It often appears that to obtain all required results it is necessary to carry out long-term calculations using this scheme. The calculations are not carried out completely accurately, and only for a specific number of significant digits and, therefore, a small error forms in every step. All these errors affect the results. The accepted computing scheme may prove to be so unsuitable that the small errors, permitted at the very start of calculations, will have a stronger and stronger effect on the results in the course of subsequent calculations and may cause large deviations from the exact values. This indicate the instability of the selected computing scheme in relation to the small errors in intermediate steps. However, if the calculations using the selected scheme can be carried out for any suitably large number of steps and the required results are obtained, then it may be concluded that the selected scheme of calculations is stable.

In the following sections we present several classes of the methods for discretisation of the problems of mathematical physics and present a number of results for the theory of these methods [3, 29, 35, 40, 60, 64, 65, 71, 72, 79, 81].

2. FINITE-DIFFERENCE METHODS

One of the most widely used methods of numerical solution of different problems of mathematical physics are the *finite-difference methods*. In various variants of these methods, a net is introduced in the domain of definition of the unknown functions and its solution is found on the net.

For the values of the unknown net function (i.e. the function given in the nodes of the net) we construct a system of scalar equations whose solution may also be used as a table of the approximate values of solution of the initial problem. One of the methods of constructing this system of scalar equations is based on the approximate substitution of the derivatives, included in the solved differential equation and in the boundary-value conditions, by difference relationships. This also explains the name of this class of the method of computing mathematics.

2.1. The net method

2.1.1. Main concepts and definitions of the method

We shall explain the main concepts of the net method and examine it initially with special reference to the simplest linear boundary problem for a ordinary differential equation. Let it be that at $a \le x \le b$ we examine a boundary value of the type:

$$Au \equiv -u''+q(x)u=f(x), \qquad x\in (a,b), \qquad u(a)=u(b)=0, \tag{1}$$

where $q(x) \ge q_0 = \text{const} > 0$. It is assumed that the boundary problem (1) has a unique solution, this solution is continuous on $[a,b]$ and has continuous derivatives in this segment to the fourth order, inclusive.

The net method for solving the boundary problem (1), as for many other problems, consists of the following.

1. The domain of definition of the differential equation (1), i.e. the segment $[a,b]$, is replaced by some discrete (net) domain. This means that in the segment $[a,b]$ we select some system of points. The set of these points is referred to as the *net*. If the position of each point is determined in accordance with the rule $x_k = a + kh$, $k = 0.1, \ldots, N$, $h = (b - a)/N$, the net is referred to as *uniform*. Points x_k are nodes of the net.

2. The boundary value problem (1) on the set of the nodes, belonging to the net, is replaced by some net problem. The term *the net problem* refers to some relationships between the approximate values of solution of the boundary problem (1) in the nodes of the net. In the examined case, this is a system of linear algebraic equations.

3. The resultant net problem is solved using some numerical method and, at the same time, we determine the approximate values of solution of the boundary problem in the nodes of the net. This is also the final aim of the net method.

The following problems may be regarded as the problems in the net method.

a) How to replace the domain of definition of differential equations, and in the case of differential equations in partial derivatives also the boundary of the

domain by some net domain;

b) How to replace the differential equation and boundary condition by some net relationships?

c) If the resultant net problem is uniquely solvable, will it be stable, converging?

We shall explain the meaning of the concepts and give answers to the posed problems with reference to (1).

Construction of the difference scheme

We select the uniform net: $x_k = a + kh, \quad k = 0,1\ldots,N, \quad h = \dfrac{(b-a)}{N}.$

The differential equation from (1) is examined only in the internal nodes of the net, i.e. we examine the equation at the points x_k, $k = 1, \ldots, N-1$:

$$Au\big|_{x=x_k} \equiv -u''(x_k) + q(x_k)u(x_k) = f(x_k), \qquad k = 1,2, \ldots, N-1.$$

We express the derivatives, included in this equality, through the values $u(x_k)$ in the nodes of the net, using appropriate finite-difference representations for them:

$$u'(x_k) = \frac{u(x_k) - u(x_{k-1})}{h} + r_k^{(1)}(h), \quad r_k^{(1)}(h) = \frac{h}{2}u''\left(x_k^{(1)}\right), \quad x_{k-1} < x_k^{(1)} < x_k;$$

$$u'(x_k) = \frac{u(x_{k+1}) - u(x_k)}{h} + r_k^{(2)}(h), \quad r_k^{(2)}(h) = -\frac{h}{2}u''\left(x_k^{(2)}\right), \quad x_k < x_k^{(2)} < x_{k+1};$$

$$u'(x_k) = \frac{u(x_{k+1}) - u(x_{k-1})}{2h} + r_k^{(3)}(h), \quad r_k^{(3)}(h) = -\frac{h^2}{6}u'''\left(x_k^{(3)}\right), \; x_{k-1} < x_k^{(3)} < x_{k+1};$$

$$u''(x_k) = \frac{u(x_{k+1}) - 2u(x_k) + u(x_{k-1})}{h^2} - r_k^{(4)}(h);$$

$$r_k^{(4)}(h) = -\frac{h^2}{12}u^{IY}(x_k^{(4)}), \quad x_{k-1} < x_k^{(4)} < x_{k+1}.$$

It should be noted that if we were examining the boundary conditions of a more complicated type, containing derivatives $u'(x_0)$, $u'(x_N)$, if necessary, we could have also used the following finite-difference representations:

$$u'(x_0) = \frac{-u(x_2) - 4u(x_1) - 3u(x_0)}{2h} + O(h^2),$$

$$u'(x_N) = \frac{3u(x_N) - 4u(x_{N-1}) + u(x_{N-2})}{2h} + O(h^2).$$

Taking into account the finite-difference representations (in respect of (1) only for $u''(x_k)$, we obtain

$$Au\big|_{x=x_k} \equiv A_h u(x_k) - R_k(h) = f(x_k),$$

where

$$A_h u(x_k) = \frac{-u(x_{k-1}) + 2u(x_k) - u(x_{k-1})}{h^2} + q(x_k)u(x_k), \qquad R_k(h) = r_k^{(4)}(h).$$

If the condition $|R_k(h)| \le Mh^2$, $k = 1,2, \ldots, N-1$, is satisfied for $R_k(h)$ where M = const does not depend on h, then we may conclude that the difference operator A_h approximates on the solution the differential operator A with the error of the second order in respect of h. Let h be relatively small; consequently, $R_k(h)$ can be ignored and we obtain

$$A_h u_k = f(x_k), \qquad k = 1,2,\ldots, N-1, \tag{2}$$

where if some conditions are satisfied it may be assumed that $u \approx u(x_i)$, $i = 0, 1, 2,\ldots,N$. Generally speaking, in all cases $u_i \ne u(x_i)$, and only at $R_k(h) \equiv 0$ it may be expected that $u_i = u(x_i)$, $i = 0, 1, 2,\ldots,N$. Equality (2) is referred to as the difference scheme approximating the equation $Au = f(x)$.

It should also be mentioned that (2) is a system of linear algebraic equations, the number of these equations is $N - 1$ and the matrix of this system is tri-diagonal. Unknown quantities are $u_0, u_1,\ldots,u_N$. The number of these unknown quantities in the system is $N + 1$.

We return to the boundary conditions from (1). Using the boundary conditions, we obtain the simplest (in the view of the problem) additional equations

$$u_0 = 0, \qquad u_N = 0. \tag{3}$$

The equations (2), (3) form the system of $N + 1$ linear algebraic equations with unknown $u_0, u_1, \ldots, u_N$. Sometimes (as in the case of the examined problem) some of these unknown quantities can be determined from the 'boundary equations' of the type (3) and, consequently, we have already obtained a problem for $N - 1$ unknown $u_1, u_2, \ldots, u_{N-1}$. In other cases, from the boundary conditions we can express u_0, u_N through $u_1, u_2, \ldots, u_{N-1}$. Substituting these expressions into equation (3) instead of u_0, u_N, we again obtain a system of equations for $u_1, u_2, \ldots, u_{N-1}$. However, it should be mentioned that in the latter case, operator A_h and the expression in the right-hand part in (2) may change their form.

If we now solve the system (2), (3) with respect to some algorithm, it may be subsequently accepted that $u(x_k) \approx u_k$, $k = 0, 1, \ldots, N$.

Solvability of the systems of difference equations

In the net method, the approximate solution of the examined problem is calculated as the solution of a system of difference equations.

We explain the conditions of solvability of these equations on an example of the system (2), (3). Here, from the 'boundary equations' we have $u_0 = 0$, $u_N = 0$. Therefore, we examine (2) for $u_1, \ldots, u_{N-1}$. The matrix of this system is with *diagonal dominance*. It should be mentioned that some matrix

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots\dots\dots\dots & & \\ a_{n1} & \dots & a_{nm} \end{pmatrix}$$

is referred to as the diagonal dominance matrix of $\delta > 0$, if $|a_{ii}| \geq \sum_{j, j \neq i}^{n} |a_{ij},| + \delta,\ i = 1, 3, \dots, n.$ For this matrix, there is an inverse matrix A^{-1}, and its norm, linked with the norm $||x||_\infty = \max_j |x_j|$, satisfies the estimate $||A^{-1}||_\infty \leq 1/\delta$.

Thus, if $q(x) \geq q_0 = \text{const} > 0$, the matrix of the system (2) is a diagonal dominance matrix; at any $\{f(x_k)\}$, the system (2) has the unique solution $\{y_k\}$, and

$$\max_j |y_k| \leq \frac{1}{q_0} \max_k |f(x_k)|.$$

This inequality also confirms that the scheme (2) is stable in relation to possible errors in the values of $\{f(x_k)\}$.

The solution of the system (2) can be found by the well known *method of Gauss exclusion* which when used for the systems of equations with tri-diagonal matrices is also referred to as the *sweep method* or *factorisation method*.

The estimate of the error and convergence of the net method

We examine these questions on the example of problem (1). Let it be that

$$\varepsilon_k = u(x_k) - u_k, \qquad \varepsilon(h) = \max_{0 \leq k \leq N} |\varepsilon_k|.$$

The net method is referred to as uniformly converging, if at $h \to 0$ the condition $\varepsilon(h) \to 0$ is satisfied. For the problem (1) we have a system of the type

$$A_h \varepsilon_k = R_k(h), \qquad k = 1, 2, \dots, N-1, \qquad \varepsilon_0 = \varepsilon_N = 0.$$

Assuming that the exact solution of (1) has bounded derivatives of the fourth order, we obtain $|R_k(h)| \leq Mh^2$. Since the matrix of the system for the errors $\{\varepsilon_k\}$ is with diagonal dominance, we conclude that the estimate

$$\max_k |\varepsilon_k| \leq \frac{1}{q_0} M\ h^2 \to 0,$$

is valid. This estimate also guarantees the convergence of the net method in the examined case of the problem with the rate of $O(h^2)$.

Examination of the main problems of justification of the net method in more complicated problems is carried out using different special approaches and claims, developed and produced in the theory of the given methods (*the maximum principle, comparison theorem, etc.*).

2.1.2. General definitions of the net method. The convergence theorem

We present the main definitions and concepts of the theory of the net method in more general formulations.

We examine some boundary-value problem which is written in the form of a symbolic equality:

$$Au=f \quad \text{in} \quad D, \tag{4}$$

where the initial data and the exact solution $u(x)$ are determined in the domains D and $\overline{D}=D\cup\partial D$, respectively. (In the problem examined in paragraph 2.1.1 for the ordinary differential equation we have $D=\{a<x<b\}\subset\mathsf{R}$, $\overline{D}=\{a\le x\le b\}$.) It is assumed that the solution of the problem (4) does exist and is sufficiently smooth in $\overline{D}$. To calculate this solution using the finite difference method or the net method, it is necessary to select, in the domain $\overline{D}$, a finite number of points whose set will be referred to as a net and will be denoted by D_h, and then to find, instead of the solution $u(x)$ of the problem (4), the table $[u]_h$ of the values of the solution at the points of the net D_h. It is assumed that the net D_h depends on the parameter h which may have any suitably small positive values. When the step of the net h tends to zero, the net should become finer and finer. The unknown net function $[u]_h$ in this case at the point x_n of the net has the values denoted by u_h.

Let it be that for the approximate calculation of the solution of the differential boundary-value problem (4), i.e. for the approximate calculation of the net function $[u]_h$ we use the equality (4) to compile a system of equations, i.e. the difference scheme which will be symbolically written, as equation (4), in the form of the equality

$$A_h u^{(h)} = f^{(h)}. \tag{5}$$

The solution $u^{(h)}=(u_0^{(h)},u_1^{(h)},\ldots,u_N^{(h)})$ of the system (5) is determined on the same net D_h as the required net function $[u]_h$. Its values $u_0^{(h)},u_1^{(h)},\ldots,u_N^{(h)}$ at the points x_0, $x_1,\ldots,x_N$ are consecutively calculated from (5) at n = 0, 1, …, N. To shorten the equations, the symbol h at $u_N^{(h)}$ in equation (5) is omitted. The method of construction and examination of the converging difference schemes (5) is the subject of the entire direction of computational mathematics referred to as the *theory of difference schemes*. We shall give the exact meaning to the concept of 'the converging scheme' and the requirement of convergence $u^{(h)}\to[u]_h$. For this purpose, we examine the linear normalised space of the functions, determined on the net D_h. The norm $\left\|u^{(h)}\right\|_{U_h}$ of the net function $u^{(h)}\in U_h$ is a non-negative number, regarded as the measure of deviations of the function $u^{(h)}$ from the actual zero.

The norm can be determined by different methods. For example, as the norm of the function we can use the exact upper bound of the modulus of its values at the points of the net, setting

$$\left\|u^{(h)}\right\|_{U_h} = \max_n |u_n|.$$

If $D \subset \mathrm{R}^n$, the norm determined by the equalities

$$\left\|u^{(h)}\right\|_{U_h} = \left(h^2 \sum_{k=0}^{N} u_k^2\right)^{1/2},$$

is also used quite frequently; here $N + 1$ is the number of nodes of the net and unknown u_0, u_1, …, u_N, determined by the difference scheme. This norm is similar to the norm $\|u\| = \left(\int_D |u|^2 \, dx\right)^{1/2}$, for the functions $u(x)$ integrated on D with a square.

After introducing the normalised space U_h, the concept of the deviation of one function from another has some meaning. If $a^{(h)}$, $b^{(h)}$ are arbitrary net functions from U_h, then the measure of their deviation from each other is the norm of the difference, i.e. the number $\left\|b^{(h)} - a^{(h)}\right\|_{U_h}$.

We now present the exact definition of the converging difference scheme.

It may be seen that the system (5) depends on h and should be written for all values of h for which the net D_h and the net function $[u]_h$ is examined. Thus, the difference boundary-value problem (5) is not only one system but it is a family of systems which depend on parameter h. It is assumed that for each examined sufficiently small h there is the solution $u^{(h)}$ of the problem (5) belonging to the space U_h.

It is assumed that solution $u^{(h)}$ of the difference boundary-value problem (5) during refining of the net *converges* to the solution of the differential boundary-value problem (4), if $\|[u]_h - u^{(h)}\|_{U_h} \to 0$, $h \to 0$. If in addition to this, the inequality $\|[u]_h - u^{(h)}\|_{U_h} \le ch^k$ is satisfied, where $c > 0$, $k > 0$ are some constants independent of h, then it is said that we have the convergence of the order h^k, or that the difference scheme has the k-th order of accuracy.

Having the property of convergence is the fundamental requirement presented in the difference scheme (5) for the numerical solution of the differential boundary-value problem (4). If it does occur, then the difference scheme (5) can be used to calculate the solution $[u]_h$ with any given accuracy, selecting the value of h sufficiently small for this purpose.

We shall now give the exact meaning to the concept of 'approximation of the problem (4) on the solution $u(x)$' by the difference scheme (5).

It is assumed that the right-hand side of the equations, written in the form of the symbolic equality (5), are the components of vector $f^{(h)}$ from some linear normalised space F_h. Consequently, A_h can be treated as an operator placing in correspondence some element $f^{(h)}$ from $F^{(h)}$ to each net function $u^{(h)}$ from

U_h. In this case, the expression $A_h[u]_h$ has some meaning. This expression forms as a result of the application of operator A to the net function $[u]_h$ from U_h and is an element of the space F_h. The discrepancy $\delta f^{(h)} = A_h[u]_h - f^{(h)}$ belongs to space F_h as the difference of two elements of the space. The discrepancy is represented by $\|\delta f^{(h)}\|_{F_h}$.

Definition 1. We say that the difference scheme $A_h u^{(h)} = f^{(h)}$ approximates the problem $Au = f$ on the solution u if $\| f^{(h)}\|_{F_h} \to 0$ at $h \to 0$. If in addition to this, there is the inequality $\| f^{(h)}\|_{F_h} \le ch^k$, where $c > 0$, $k > 0$ are some constants, then it is said that we have the approximation of the order h^k, or the order k in relation to h.

We return to the general definition of the stability of the scheme (5) which approximates the problem (4) on the solution $u(x)$ with some order h^k.

Definition 2. The difference scheme (5) will be referred to as stable if there are numbers $h_0 > 0$, $\delta > 0$ such that at any $h < h_0$ and any $\varepsilon^{(h)} \in F_h$, $\|\varepsilon^{(h)}\|_{F_h} < \delta$ the difference problem

$$A_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}$$

obtained from the problem (5) by adding perturbation $\varepsilon^{(h)}$ to the right-hand side has one and only one solution $z^{(h)}$, and this solution deviates from the solution $u^{(h)}$ of the non-perturbed problem (5) on the net function $z^{(h)} - u^{(h)}$, satisfying the estimate

$$\|z^{(h)} - u^{(h)}\|_{U_h} \le c_1 \|\varepsilon^{(h)}\|_{F_h},$$

where c_1 is some constant independent of h.

In particular, the last inequality indicates that the small perturbation $\varepsilon^{(h)}$ of the right-hand side of the difference scheme (5) causes uniformly a small perturbation $z^{(h)} - u^{(h)}$ of the solution in relation to h.

Let us assume that the operator A_h, mapping U_h in F_h, is linear. Consequently, the previous definition of stability is equivalent to the following definition.

Definition 3. The difference scheme (5) with the linear operator A_h will be referred to as stable if at any $f^{(h)} \in F_h$ the equation $A_h u^{(h)} = f^{(h)}$ has a unique solution $u^{(h)} \in U_h$, and

$$\|u^{(h)}\|_{U_h} \le c_1 \|f^{(h)}\|_{F_h}, \tag{6}$$

where c_1 is some constant independent of h.

It may be shown that the definitions 2 and 3 are equivalent.

We shall now prove one of the fundamental results of the theory of difference schemes, in particular, we shall show that convergence follows from approximation and stability.

The convergence theorem

Let it be that the difference scheme $A_h u^{(h)} = f^{(h)}$ *approximates the problem* $Au = f$ *on the solution with the order* h^k and is stable. *Consequently, the solution* $u^{(h)}$ *of the difference problem* $A_h u^{(h)} = f^{(h)}$ *converges to* $[u]_h$, *and we have the estimate*

$$\|[u]_h - u^{(h)}\|_{U_h} \le c c_1 h^k,$$

where c, c_1 *are the numbers included in the estimates in the definitions of approximation and stability.*

Proof

We set $\varepsilon^{(h)} = \delta f^{(h)}$, $[u]_h = z^{(h)}$. Consequently, from definition 2 we have:

$$\|[u]_h - u^{(h)}\|_{U_h} \le c_1 \|\delta f^{(h)}\|_{F_h},$$

Taking into account the estimates from definition 1, we immediately obtain the proven inequality. □

In conclusion, it should be stressed that the scheme of proof of the convergence of the solution of problem (5) to the solution of the problem (4) by verification of approximation and stability is of a general nature. In this case, $Au = f$ represents any functional equation which can be used as a basis for constructing the 'difference problem' (5).

2.1.3. The net method for partial differential equations

Partial differential equations have extensive application in mathematical physics, hydrodynamics, acoustics and other areas of science. In the majority of cases, these equations are not solved in the explicit form. Therefore, the methods of approximate solution of these equations, in particular, the net method, are used widely.

The construction of different schemes of the net method in the case of equations and partial derivatives depends on the type of equations and the type of boundary (initial) conditions linked with the equations. Partial examples of these equations are the following:

The Poisson equation (elliptical equation)

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x, y);$$

The heat conduction equation (parabolic equation)

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t);$$

The wave equation (hyperbolic equation)

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, t).$$

We shall discuss several approaches to construct in the different schemes with special reference to these equations.

The difference schemes for parabolic equations

We examine a Cauchy problem for a heat conduction equation

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \varphi(x,t), \qquad -\infty < x < +\infty, \quad t > 0, \\ &u(x,0) = \psi(x), \qquad -\infty < x < +\infty. \end{aligned} \tag{7}$$

It is assumed that the problem (7) has a unique solution $u(x, t)$ continuous together with its derivatives $\frac{\partial^i u}{\partial t^i}, i = 1, 2,$ and $\frac{\partial^k u}{\partial x^k}, k = 1, 2, 3, 4.$ The problem (7) will be written in the form

$$Au = f.$$

For this purpose it is sufficient to set that

$$Au \equiv \begin{cases} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}; & -\infty < x < +\infty, \ \ t > 0, \\ u(x,0), & -\infty < x < +\infty, \ \ t = 0; \end{cases}$$

$$f \equiv \begin{cases} \varphi(x,t), & -\infty < x < +\infty, \ \ t > 0, \\ \psi(x), & -\infty < x < +\infty, \ \ t = 0. \end{cases}$$

It is also assumed that t changes in the range $0 \le t \le T \le +\infty$.

In the examined case $D = \{-\infty < x < +\infty, 0 \le t \le T < \infty\}$, Γ is the union of the straight lines $t = 0$ and $t = T$.

We select a right angled net and replace $\overline{D} = D \cup \Gamma$ by the net domain D_h. The domain D_h will be related to the set of points (nodes) (x_m, t_n), whose co-ordinates are determined on the basis of the rule

$$x_m = m\,h, \ \ m = 0, \pm 1, \pm 2, \ldots, \qquad h > 0,$$

$$t_n = n\tau, \qquad n = 0, 1, \ldots, N, \qquad \tau > 0.$$

The problem $Au = f$ will be replaced by the difference scheme of the type $Ah^{(h)} = f^{(h)}$. $u(x_m, t_n)$ denotes the exact solution of the problem $Au = f$ in the node (x_m, t_n), u_m^n corresponds to the approximated net value. We have

$$Au\,|_{(x_m,t_n)} \equiv \begin{cases} \left.\frac{\partial u}{\partial t}\right|_{(x_m,t_n)} - \left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_m,t_n)}, & m = 0, \pm 1, \ldots. \quad n = 1, \ldots, N, \\ u(x,0)\big|_{(x_m,t_n)}; \end{cases}$$

$$f\,\Big|_{(x_m,t_n)} \equiv \begin{cases} \varphi(x,t)\big|_{(x_m,t_n)}, & m = 0, 1, \pm 2, \ldots. \quad n = 1, 2. \ldots, N, \\ \psi(x)\big|_{(x_m,t_n)}. \end{cases}$$

To replace derivatives $\left.\frac{\partial u}{\partial t}\right|_{(x_m,t_n)}$ and $\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_m,t_n)}$ by the difference relationships we use the formulae of numerical differentiation. We have

$$\left.\frac{\partial u}{\partial t}\right|_{(x_m,t_n)} = \frac{u(x_m,t_{n+1})-u(x_m,t_n)}{\tau} - \frac{\tau}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{(x_m,t_n^{(1)})}, \quad t_n < t_n^{(1)} < t_{n+1},$$

$$\left.\frac{\partial u}{\partial t}\right|_{(x_m,t_n)} = \frac{u(x_m,t_n)-u(x_m,t_{n-1})}{\tau} - \frac{\tau}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{(x_m,t_n^{(2)})}, \quad t_{n-1} < t_n^{(2)} < t_n,$$

$$\left.\frac{\partial u}{\partial t}\right|_{(x_m,t_n)} = \frac{u(x_m,t_{n-1})-u(x_m,t_{n-1})}{2\tau} - \frac{\tau^2}{6}\left.\frac{\partial^3 u}{\partial t^3}\right|_{(x_m,t_n^{(3)})}, \quad t_{n-1} < t_n^{(3)} < t_{n-1},$$

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_m,t_n)} = \frac{u(x_{m+1},t_n)-2u(x_m,t_n)+u(x_{m-1},t_n)}{h^2} - \frac{h^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{(x_m,t_n)}.$$

We examine the following two-layer difference approximation:

$$Au\big|_{(x_m,t_n)} \equiv$$

$$\equiv \begin{cases} \dfrac{u(x_m,t_{n-1})-u(x_m,t_n)}{\tau} - \dfrac{u(x_{m+1},t_n)-2u(x_m,t_n)+u(x_{m-1},t_n)}{h^2} + r_{mn}^{(h)}, \\ u(xm,0), \quad \text{n=0}, \end{cases}$$

where

$$r_{mn}^{(h)} = -\frac{\tau}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{(x_m,t_n^{(1)})} - \frac{h^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{(x_m,t_n)}.$$

We introduce the following notation:

$$f^{(h)} \equiv \begin{cases} \varphi(x_m,t_n), \\ \psi(x_m) \end{cases}$$

and write the difference scheme for the problem $Au =$

$$A_h^{(1)} u^{(h)} = f^{(h)}, \tag{8}$$

where the difference operator $A_h^{(1)}$ is determined from the rule

$$A_h^{(1)} u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau} - \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \\ u_m^0, \quad m = 0, \pm 1, \pm 2, \ldots. \quad n = 0, 1, \ldots, N-1. \end{cases}$$

Similarly, we obtain the difference scheme of the type

$$A_h^{(2)} u^{(h)} = f^{(h)}, \tag{9}$$

where

$$A_h^{(2)} u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2}, \\ u_m^0. \end{cases} \qquad f^{(h)} \equiv \begin{cases} \varphi(x_m, t_n), \\ \psi(x_m). \end{cases}$$

$$m = 0, \pm 1, \pm 2, \ldots, \qquad n = 0, 1, \ldots, N-1,$$

It should be mentioned that

$$A_h^{(1)} u_h(x, y) = f^{(h)} + \delta^{(1)} f^{(h)}, \qquad A_h^{(2)} u_h(x, y) = f^{(h)} + \delta^{(2)} f^{(h)},$$

where

$$\delta^{(1)} f^{(h)} \equiv \begin{cases} r_{mn}^{(h)}. \\ 0, \end{cases} \qquad \delta^{(2)} f^{(h)} \equiv \begin{cases} \dfrac{\tau}{2} \dfrac{\partial^2 u}{\partial t^2}\Bigg|_{(x_m, t_n^{(2)})} - \dfrac{h^2}{12} \dfrac{\partial^4 u}{\partial x^4}\Bigg|_{(x_m, t_{n+1})}. \\ 0. \end{cases}$$

We clarify the order of approximation of difference schemes (8) and (9). F_h is represented by the linear set of all pairs of restricted functions

$$g^{(h)} = \begin{cases} \alpha_m^n, \\ \beta_m. \end{cases}$$

The norm in F_h is determined on the basis of the rule

$$\left\| g^{(h)} \right\| = \max_{m,n} \left| \alpha_m^n \right| + \max_m \left| \beta_m \right|.$$

If $\max_{m,n} |\alpha_m^n|$ or $\max_m |\beta_m|$ is not reached, the norm refers to the quantity $\|g^{(h)}\| = \sup_{m,n} |\alpha_m^n| + \sup_m |\beta_m|$.

Let be $\tau = rh^s$, r,s be some positive numbers. It is assumed that for $\dfrac{\partial^2 u}{\partial t^2}$ and $\dfrac{\partial^4 u}{\partial x^4}$ we have the following estimates

$$\max_{(x,t) \in D} \left| \frac{\partial^2 u}{\partial t^2} \right| \le M_2, \qquad \max_{(x,t)} \left| \frac{\partial^4 u}{\partial x^4} \right| \le M_4.$$

It is now easy to obtain

$$\left\| \delta^{(1)} f^{(h)} \right\|_{F_h} = \max_{m,n} \left| r_{mn}^{(1)}(h) \right| \le \left(\frac{r}{2} M_2 + \frac{h^{2-s}}{12} M_4 \right) h^s,$$

$$\left\| \delta^{(2)} f^{(h)} \right\| = \max_{m,n} \left| r_{mn}^{(2)}(h) \right| \le \left(\frac{r}{2} M_2 + \frac{h^{2-s}}{12} M_4 \right) h^s.$$

In the case of the scheme (8) we may set $s = 2$, and for scheme (9) $s = 1$. These equations show that the difference schemes (8), (9) approximate the problem $Au = f$ on $u(x, y)$ with the error of the order of s ($1 \le s \le 2$) in relation to h.

The difference scheme (8) makes it possible to calculate the values in the first layer u_m^1, $m = \pm 1, \pm 2, \ldots$, from the values of the solution in the zero layer, i.e. using the values u_m^0, $m = 0, \pm 1, \pm 2$. For this purpose it is sufficient to set $n = 0$ in equation (8) and carry out recursive calculations. Subsequently, using the values of u_m^1 we can calculate in the same manner at $n = 1$ the values of u_m^2 and so on. Because of these computing properties, the difference scheme (8) is referred to as explicit.

The difference scheme (9) does not have these properties. In fact, if we set $n = 0$ in (9), then in the left-hand side of the resultant equation we have a linear combination of the values u_{m-1}^1, u_m^1, u_{m+1}^1, u_m^1, and the right-hand side contains values of the known function $\varphi(x_m, 0)$ and $\psi(x_m)$. To calculate the values in the first layer $\ldots, u_{-2}^1, u_{-1}^1, u_0^1, u_1^1, u_2^1, \ldots$ it is necessary to solve in this case an infinite system of linear equations. For this reason, the difference scheme (9) is referred to as implicit.

We examine the problem of stability of the schemes (9), (10) determining the norm in the space U_h using the rule

$$\left\| u^{(h)} \right\|_{U_h} = \max_{m,n} \left| u_m^n \right|.$$

We examine the difference scheme (8). We explain at which values of r, $\tau = rh^2$ can this scheme be stable.

To confirm stability, it must be shown that the difference scheme is uniquely solvable and at any

$$g^{(h)} = \begin{cases} \alpha_m^n, \\ \beta_m, \end{cases} \qquad g^{(h)} \in F_h,$$

we have the estimate

$$||z^{(h)}||_{U_h} \le M ||g^{(h)}||_{F_h},$$

where M is a constant which does not depend on h and $g^{(h)}$ and $A_h^{(1)} z^{(h)} = g^{(h)}$.

The difference scheme (8) is explicit and its unambiguous solvability is evident.

We rewrite the formula $A_h^{(1)} z^{(h)} = g^{(h)}$ in the form

$$z_m^{n+1} = r\left(z_{m+1}^n + z_{m-1}^n\right) + (1-2r) z_m^n + \tau \alpha_m^n, \quad z_m^0 = \beta_m,$$

$$m = 0, \pm 1, \pm 2, \ldots, \quad n = 0, 1, 2, \ldots, N-1.$$

From these equations when fulfilling restrictions of the type $r \equiv (/h^2) \le 1/2$ we obtain the inequality

$$\max_m \left| z_m^{n+1} \right| \le \max_m \left| z_m^n \right| + \tau \max_{m,n} \left| \alpha_m^n \right|$$

which is the *maximum principle*. Assuming that $n = 0, 1, \ldots, N-1$ and summing up the resultant relationships, we obtain

$$\max_m \left|z_m^N\right| \le \max_m \left|\beta_m\right| + N\tau \max_{m,n} \left|\alpha_m^n\right| \le \max_m \left|\beta_m\right| + T \max_{m,n} \left|\alpha_m^n\right| \le$$

$$\le \max(1,T)\left(\max_{m,n} \left|\alpha_m^n\right| + \max_m \left|\beta_m\right|\right) = M \cdot \left\|g^{(h)}\right\|_{F_h},$$

with the notations $M = \max(1, T)$; $M = 1$, if $T < 1$, and $M = T$, if $T \ge 1$. Consequently,

$$\left\|z^{(h)}\right\|_{U_h} \le M \left\|g^{(h)}\right\|_{F_h}.$$

Thus, the scheme (8) is stable if the restriction $r \le 0.5$ is satisfied. It should be mentioned that this restriction imposes strict requirements on the selection of τ: $\tau \le 0.5h^2$, and causes that if it is necessary to maintain stability, then in calculations using the scheme (8) the time step should be selected very small.

We return to the difference (9) and write it in the new form

$$\begin{gathered} -r\left(u_{m+1}^{n+1} + u_{m-1}^{n+1}\right) + (1+2r)u_m^{n+1} = u_m^n + \tau\varphi(x_m, t_n), \\ u_m^0 = \psi(x_m), \quad m = 0, \pm1, \pm2, \ldots, \quad n = 0, 1, \ldots, N-1. \end{gathered} \tag{10}$$

This scheme represents an infinite system of linear equations in relation to the unknown quantity $\ldots u_{-1}^h, u_0^1, u_1^1, \ldots$. The solution of these systems is a complicated and time-consuming task and, therefore, the difference schemes of the type (9) are not suitable for Cauchy problems at infinite segments and are used only seldom. However, if the segment of the axis x on which the Cauchy problem is examined, is finite: $a \le x \le b$, $b - a \le K$, and on the straight lines $x = a$ and $x = b$ some restrictions are additionally specified for the solution $u(x, t)$, the difference schemes of the type (9) appear to be highly efficient. In particular, these schemes are *absolutely stable*, i.e. stable for any values of $r = \tau/h^2$.

If, for example, on the segments of the straight lines $x = a$ and $x = b$ the conditions $u(a, t) = \gamma_0(t)$, $u(b, t) = \gamma_1(t)$, are specified, then the type of system (10) at $n = 0$ greatly changes:

$$\begin{gathered} -r\left(u_{m+1}^1 + u_{m-1}^1\right) + (1+2r)u_m^1 = \psi(x_m) + \tau\varphi(x_m, 0), \\ u_0^1 = \gamma_0(t_m). \quad u_M^1 = \gamma_1(t_1), \quad m = 1, 2, \ldots, M-1, \quad h = (b-a)/M. \end{gathered} \tag{11}$$

Equation (11) represents a system of $M+1$ linear algebraic equations in relation to $u_0^1, u_1^1, \ldots, u_M^1$. This system can be solved by the sweep method, and subsequently all u_m^n are determined for other values of n.

These considerations show that the realisation of the implicit difference schemes require long computing times when calculating the solution in a single time layer, but the number of these time layers may be small owing to the fact that in this case there are no restrictions on the ratio τ/h^2. If we use explicit difference schemes, the solution in the next layer is calculated using the recursion rule and is linked with the minimum computing time. However,

because of the restriction $\tau/h^2 \le 1/2$ the number of time layers in the case of explicit schemes may be considerable larger in comparison with the number of time layers for implicit schemes.

We shall note the following confirmation of the convergence of the difference scheme (8). This scheme approximates the problem (7) on the solution $u(x, t)$ with the error of the order $O(\tau + h^2)$ and stable at $r \le 1/2$. Therefore, the *scheme* (8) *is converging*; in this case, *the error for the approximate solution is the value of the order* $O(\tau + h^2)$.

The difference scheme for the Dirichlet problem for the Poisson equation

Let it be that in the region $D = \{0 < x < a,\ 0 < y < b\}$ we examine a Poisson equation with the Dirichlet condition at the boundary Γ of the domain D:

$$Au \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x, y), \qquad u\big|_{\Gamma} = 0. \tag{12}$$

It is assumed that (12) has a unique solution $u(x, y)$ in $\overline{D} = D \cup \Gamma$, and this solution has derivatives $\dfrac{\partial^4 u}{\partial x^4}$ and $\dfrac{\partial^4 u}{\partial y^4}$ continuous in D.

We use a right-angled net, setting $x_m = mh$, $m = 0, 1, \ldots, M$, $h = a/M$; $y_n = nl$, $n = 0,1, \ldots, N$, $l = b/N$. To the set of the internal nodes D_h^0 we relate all nodes located in D, and to the set of boundary nodes Γ_h we relate the nodes located on Γ.

Let us assume that we have the node $(m, n) \in D_h^0$. Substitution of the differential equation from (12) by a difference one will be carried out only in the internal nodes. We have

$$-\left(\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_m, y_n)} + \frac{\partial^2 u}{\partial y^2}\bigg|_{(x_m, y_n)}\right) = f(x_m, y_n),$$

$$-\frac{u(x_{m+1}, y_n) - 2u(x_m, y_n) + u(x_{m-1}, y_n)}{h^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}_{(x_m^{(1)}, y_n)} -$$

$$-\frac{u(x_m, y_{n+1}) - 2u(x_m, y_n) + u(x_m, y_{n-1})}{l^2} + \frac{l^2}{12}\frac{\partial^4 u}{\partial y^4}\bigg|_{(x_m, y_n^{(1)})} = f(x_m, y_n),$$

$$(m, n) \in D_h^0, \quad x_{m-1} < x_m^{(1)} < x_{m+1}, \quad y_{n-1} < y_n^{(1)} < y_{n+1}.$$

Let the functions $\dfrac{\partial^4 u(x, y)}{\partial x^4}$ and $\dfrac{\partial^4 u(x, y)}{\partial y^4}$ be bounded in respect of the absolute value in the domain $\overline{D}$; consequently, at sufficiently small h and l we can ignore the members present as multipliers h^2 and l^2, and obtain the required difference equation

$$A_h u^{(h)} = f^{(h)}, \tag{13}$$

where

$$A_h u^{(h)} \equiv \frac{-u_{m+1,n} + 2u_{m,n} - u_{m-1,n}}{h^2} + \frac{-u_{m,n+1} + 2u_{m,n} - u_{m,n-1}}{l^2},$$

$$f^{(h)} \equiv f(x_m, y_n), \qquad (m,n) \in D_h^0, \qquad u^{(h)}|_{\Gamma_h} \equiv 0.$$

Here u_{mn} denotes the approximate net value of the solution of the problem (12): $u_{mn} \approx u\,(x_m, y_n)$.

Because of the definitions and concepts introduced in the paragraphs 2.1.1 and 2.1.2 we obtain

$$A_h u^h(x,y) = f^{(h)} + \delta f^{(h)},$$

where

$$\delta f^{(h)} \equiv -\frac{h^2}{12} \cdot \frac{\partial^4 u}{\partial x^4}\bigg|_{(x_m^{(1)}, y_n)} - \frac{l^2}{12} \cdot \frac{\partial^4 u}{\partial y^4}\bigg|_{(x_m, y_n^{(1)})}, \qquad (m,n) \in D_h^0.$$

Consequently, under the assumptions made previously in respect of $\dfrac{\partial^4 u}{\partial x^4}$ and $\dfrac{\partial^4 u}{\partial y^4}$ we obtain the estimate

$$\left\| \delta f^{(h)} \right\| F_h \le C(h^2 - l^2) \qquad \left(\left\| u^{(h)} \right\|_{F_h} \equiv \left(\sum_{m=1}^{M-1} \sum_{l=1}^{N-1} hl \,|\, u_{mn} \,|^2 \right)^{1/2} \right),$$

where C is a constant independent of h. This estimate shows that the scheme (13) approximates the problem (12) on the solution $u(x, y)$ with the error $O(h^2 + l^2)$.

It should also be mentioned that operator A_h is symmetric and positive definite in the real Hilbert space F_h with the norm $\|\cdot\|_{F_h}$ and the scalar product $\left(u^{(h)}, v^{(h)}\right)_{F_h} = \sum_{m=1}^{M-1} \sum_{l=1}^{N-1} hl u_{mn} v_{ml}$. The eigenvalues of these operators have the form

$$\lambda_h^{(m,n)} = \frac{4}{h^2} \sin^2 \frac{m\pi h}{2a} + \frac{4}{l^2} \sin^2 \frac{n\pi l}{2b}, \qquad 1 \le m \le M-1, \quad 1 \le n \le N-1.$$

The exact eigenvalues for the problem $Au = \lambda u$, $u|_\Gamma = 0$ are the numbers

$$\lambda^{(m,n)} = \left(\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2 \right) \pi^2, \qquad 1 \le m \le \infty, \qquad 1 \le n \le \infty.$$

It may easily be seen that

$$\lambda_h^{(m,n)} = \lambda^{(m,n)} = -\frac{\pi^2}{12} \left[\left(\frac{m}{a}\right)^4 h^2 + \left(\frac{n}{b}\right)^4 l^2 \right] + O(h^4 + l^4).$$

If we use the method of expansion in respect of the eigenfunctions of the operator A_h, then to solve the problem (13) we can easily obtain the estimate

$$\left\| u^{(h)} \right\|_{F_h} \le C \cdot \left\| f^{(h)} \right\|_{F_h},$$

where $C = 1/\lambda_h^{(1,1)}$, which guarantees the stability of the scheme (13).

On the basis of assumptions regarding the stability and approximation we obtain for (13) that the given scheme is *converging and the estimate* $||u^{(h)} - [u]_h||_{F_h} \le C \cdot (h^2 + l^2)$ is valid, where $[u]_h$ is the table of the values of the exact solution on D_h^0, and constant C is independent of h and l.

The difference scheme for the hyperbolic-type equation

We examine a Cauchy problem for one of the simplest hyperbolic equations – the equation of oscillations of a string (one-dimensional wave equation)

$$Au \equiv \quad u \equiv \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x,t), \qquad -\infty < x < -\infty, \quad t > 0,$$

$$u\,|_{t=0} = \varphi(x), \qquad \left.\frac{\partial u}{\partial t}\right|_{t=0} = g(x), \tag{14}$$

where $f(x, t)$, $\varphi(x)$, $g(x)$ are the given functions. The solution of the problem (14) is determined by the formula (D'Alembert solution)

$$u(x_0,t_0) = \frac{\varphi(x_0 - t_0) + \varphi(x_0 + t_0)}{2} + \frac{1}{2}\int\limits_{\varphi(x_0 - t_0)}^{(x_0 + t_0)} g(x)dx + \frac{1}{2}\int\limits_{D(x_0,t_0)}\int f dxdt,$$

where $D(x_0, t_0)$ is the triangle of the plane Oxt, restricted by the axis $t = 0$ and straight lines ('characteristics') $x + t = x_0 + t_0$, $x - t = x_0 - t_0$, passing through (x_0, t_0). This equation shows that $u(x_0, t_0)$ is determined by the values of φ and g at the points of the base of the triangle $D(x_0, t_0)$ and the values of f inside and on the contour of the triangle.

We examine a right-angled net $x_k = kh$, $t_j = j\tau$ $(k = 0, \pm 1, \ldots; j = 0, 1, 2, \ldots)$. The set of the nodes (k, j) $(-\infty < k < \infty)$ is referred to as the j-th row of the net. The simplest replacement of the problem (14) by the net problem is as follows:

$${}_h u_h \equiv \frac{u_{k,j+1} - 2u_{k,j} + u_{k,j-1}}{\tau^2} - \frac{u_{k,j+1} - 2u_{k,j} + u_{k-1,j}}{h^2} = (f)_{k,j},$$

$$u_{k,0} = \varphi(kh), \qquad \frac{u_{k,1} - u_{k,0}}{\tau} = g(kh). \tag{15}$$

Here $k = 0, \pm 1, \pm 2, \ldots; j = 1,2, \ldots$. From equations (15) u_h is determined in the zeroth and first rows of the net: $u_{k,0} = \varphi(kh)$, $u_{k,1} = \varphi(k, h) + \tau g(kh)$. The equations (15) are such that $u_{k,j+1}$ is uniquely determined by the values of the net function u_h at the nodes of j-th and $(j - 1)$-th rows $(j = 1, 2, \ldots)$. This leads to the unique solvability of the system (15).

We introduce the notations: $\tau / h = r$, $D_r(x_0, t_0)$ is the triangle bounded by the axis $t = 0$ and straight lines $rx + t = rx_0 + t_0$, $rx - t = rx_0 - t_0$, passing through (x_0, t_0). It may easily be seen that the value of the net solution u_h of the system (15) at the node (k_0, j_0) is uniquely determined by the values of φ, g in the nodes located on the base of the triangle $D_r(k_0 h, j_0 \tau)$ and the values of f in the nodes inside and at the boundary $D_r(k_0 h, j_0 \tau)$.

Let us assume that we have a sequence of nets such that $h \to 0$, $\tau = rh$, $r = \text{const} > 1$ and some point (x_0, t_0) $(t_0 > 0)$ is a node of every net of the sequence. If $u_h(x_0, t_0)$ has the limit $u_0(x_0, t_0)$ at $h \to 0$, the latter is determined by the values of f, φ, g in $D_r(x_0, t_0)$ and, generally speaking, differs from $u(x_0, t_0)$ – the value in (x_0, t_0) of the solution of the problem (14) determined by the values of f, φ, g and the triangle $D(x_0, t_0)$, containing $D_r(x_0, t_0)$ as part in the case $r > 1$. Therefore, at $r > 1$ sequence u_h, generally speaking, is not reduced to the exact solution of the problem (14).

Let the functions $f(x, t)$, $\varphi(x)$ and $g(x)$ be sufficiently smooth in the half plane $t \geq 0$ and at the axis $-\infty < x < \infty$, respectively. Consequently, the following claim is valid: if $r = \text{const} \leq 1$, then at $h \to 0$ the sequence u_h of the solution of the problem (15) is converging uniformly to the exact solution $u(x, t)$ of the problem (14) in any finite domain of the half plane.

The approaches to constructing difference schemes, presented previously in the application to the simplest equations in partial derivatives, can be extended to more complex equations (to equations with variable coefficients, multidimensional and/or non-linear equations, and to systems of equations).

2.2. The method of arbitrary lines

The main concept of the method of arbitrary lines is that the derivatives in respect of specific independent variables are replaced by the approximate expressions through finite differences, whereas the derivatives in respect of other variables remain without changes. We examine several variants of this method in application to linear differential equations of the second order.

2.2.1. The method of arbitrary lines for parabolic-type equations

Let the following boundary-value problem be set in the semi-strip $0 \leq x \leq 1$, $t \geq 0$:

$$\begin{aligned} u_t &= Au + f(x,t) \qquad u(x,0) = \varphi(x), \\ u(0,t) &= \psi_0(t), \qquad u(1,t) = \psi_1(t). \end{aligned} \tag{16}$$

Here

$$Au = a(x,t)u_{xx} + b(x,t)u_x + c(x,t)u.$$

We assume the existence and uniqueness of a sufficiently smooth solution of the examined problem. We shall find it approximately by the method of arbitrary lines in rectangle $\Pi = \{0 \leq x \leq 1,\ 0 \leq t \leq T < \infty\}$. Depending on whether the derivatives are approximated in respect of x or t, we obtain a longitudinal or transverse variant of the method. Initially, we examine its transverse variant.

Π denotes the lattice domain $0 < x < 1$, $t = t_n = n\tau$, $n = 1, 2, \ldots, N$ $(N\tau = T)$, and its boundary Γ_τ is a set consisting of the segments $0 \leq x \leq 1$ of the straight line $t = 0$ and the points $(0, t_N)$, $(1, t_n)$, $n = 1, 2, \ldots, N$.

We examine problem (16) of $\Pi_\tau + \Gamma_\tau$. The derivative $u_t(x, t)$ at $t = t_n$, $n = 1, 2, \ldots, N$, is replaced approximately by the left-sided difference relationship $u_t(x, t_n) \approx (u(x, t_n) - u(x, t_{n-1}))/\tau$. Consequently, for the lattice function, given by the system of functions $u_n = u_n(x)$, $n = 0, 1, \ldots, N$, approximately representing

the function $u(x, t)$ and the segments $0 \le x \le 1$ of the straight line $t = t_n$, $n = 0, 1, \ldots, N$, in $\Pi_\tau + \Gamma_\tau$ we can set the following differential-difference boundary problem:

$$\frac{u_n(x) - u_{n-1}(x)}{\tau} = A_n u_n + f_n(x),$$
$$u_0(x) = \varphi(x), \quad u_n(0) = \psi_0(t_n), \quad u_n(1) = \psi_1(t_n), \tag{17}$$

where

$$A_n u_n = a_n(x) u_n''(x) - b_n(x) u_n'(x) + c_n(x) u_n(x),$$
$$a_n(x) = a(x, t_n), \quad b_n(x) = b(x, t_n), \quad c_n(x) = c(x, t_n), \quad f_n(x) = f(x, t_n).$$

The problem (17) breaks down into N boundary problems for ordinary differential equations of the second order, because the function $u_0(x)$ is available from the initial condition, and the unknown functions $u_n(x) \approx u(x, t_n)$, $n = 1, 2, \ldots, N$ can be found gradually, solving at every value of $n = 1, 2, \ldots, N$ one ordinary differential equation of the second order with the appropriate boundary conditions.

The constructed computing scheme of the method of arbitrary lines explains quite efficiently the geometrical meaning of the name of the method. Therefore, in the multi-dimensional case, the examined method is often referred to as the *plane method* or the *hyperplane method*.

We examine one of the longitudinal schemes of the method of arbitrary lines for the problem (16). Here in constructing the computing schemes of the method of arbitrary lines we approximate the differentiation operation not in respect of time t but in respect of the spatial variable x. For this case, as in the case of transverse schemes, the original domain is replaced in advance and approximately by the lattice domain, drawing the straight lines $x = x_n = nh$, $n = 0, 1, \ldots, N$ ($Nh = 1$). On each of the internal straight lines $x = x_n$, $n = 1, 2, \ldots, N - 1$, the derivatives u_{xx} and u_x are approximated by means of the values of the function u on several adjacent straight lines. If we use the simplest symmetric expression for these purposes then, as in the case of the transverse schemes of the method, we can write

$$u_n' = a_n(t) \frac{u_{n+1}(t) - 2u_n(t) + u_{n-1}(t)}{h^2} + b_n(t) \frac{u_{n+1}(t) - u_{n-1}(t)}{2h} +$$
$$+ c_n(t) u_n(t) + f_n(t), \quad t > 0.$$
$$u_0 t = \psi_0(t), \quad u_N(t) = \psi_1(t), \quad t \ge 0, \tag{18}$$
$$u_n(0) = \psi(x_n), \quad n = 1, 2, \ldots, N - 1,$$

where $u_n(t) \approx u(x_n, t)$, $a_n(t) = a(x_n, t)$, $b_n(t) = b(x_n, t)$, $c_n(t) = c(x_n, t)$, $f_n(t) = f(x_n, t)$. Thus, the determination of the approximate values of $u_n(t)$ on the straight lines $x = x_n$, $n = 1, 2, \ldots, N - 1$, is reduced to solving the Cauchy problem for the system of $N - 1$ linear ordinary differential equations of the first order.

The error of approximation of the problem (16) by the problem (18) is predetermined by the inaccuracy of replacement of the derivatives of u_{xx}, u_x by the corresponding simplest symmetric expressions through the values of the function u and it is evidently the value of the order h^2 if

the solution $u(x, t)$ is sufficiently smooth.

It should be mentioned that in the case of longitudinal schemes in approximation of the differential equation on the straight lines, close to boundary ones, we usually experience difficulties associated with disruption of the homogeneity of the approximation process. In some partial cases, of course, these difficulties can be greatly weakened as a result of the application of boundary data.

2.2.2. The method of arbitrary lines for hyperbolic equations

Let the following boundary problem be formulated in semi-strip $0 \le x \le 1$, $t \ge 0$:

$$\begin{gathered} u_u + g(x,t)u_t = Au + f(x,t), \\ u(x,0) = \varphi(x), \qquad u_t(x,0) = \Phi(x), \\ u(0,t) = \psi_0(t), \qquad u(1,t) = \psi_1(t). \end{gathered} \tag{19}$$

Here, Au has the same form as in (16). It is assumed that $a(x, t) \ge a > 0$. The existence and uniqueness of a sufficiently smooth solution of the problem (19) is assumed. We shall find this solution by the method of arbitrary lines in the rectangle $\Pi = \{0 \le x \le 1,\ 0 \le t \le T < \infty\}$.

As in the case of parabolic equations, here we can also talk about the longitudinal and transverse variants of the method. Initially, we present an example of a transverse scheme.

The rectangle Π is approximated by the lattice region $\Pi_\tau + \Gamma_\tau$; in this case, in contrast to the parabolic case (see paragraph 2.2.1), to the boundary Γ_τ of the examined region we also relate the interval $0 < x < 1$ of the straight lines $t = t_1 = \tau$. On each straight line $t = t_{n+1}$ for $n = 1, 2, \ldots, N-1$ the derivatives u_t and u_{tt} are replaced by the following left-sided difference relationships of the first and second order respectively:

$$u(x,t_{n-1}) \approx \frac{u(x,t_{n+1}) - u(x,t_n)}{\tau},$$

$$u_{tt}(x,t_{n+1}) \approx \frac{u(x,t_{n+1}) - 2u(x,t_n) + u(x,t_{n-1})}{\tau^2}.$$

In addition to this, on the straight line $t = 0$ to approximate the derivative u_t we use the right-sided substitution $u_t(x, 0) \approx (u(x,\tau) - u(x, 0))/\tau$. For the lattice function, given by the system of function $u_n(x) \approx u(x, t_n)$, $n = 1, 2, \ldots, N-1$, this makes it possible to formulate in the domain $\Pi_\tau + \Gamma_\tau$ the following differential-difference boundary problem:

$$\begin{gathered} \frac{u_{n+1}(x) - 2u_n(x) + u_{n-1}(x)}{\tau^2} + g_{n+1}(x)\frac{u_{n+1}(x) - u_n(x)}{\tau} = A_{n+1}u_{n+1} + f_{n+1}(x), \\ u_0(x) = \varphi(x), \quad \frac{u_1(x) - u_0(x)}{\tau} = \Phi(x), \\ u_{n+1}(0) = \psi_0(t_{n-1}), \qquad u_{n+1}(1) = \psi_1(t_{n+1}). \end{gathered} \tag{20}$$

Here, we use the notations completely identical with those used in the parabolic case.

The process of solving the approximating problem (20) can be evidently divided into the following stages. Initially, from the given function $u_0(x) = \varphi(x)$ we directly obtain $u_1(x) = \varphi(x) + \Phi(x)$. Consequently, we find the functions $u_2(x), u_3(x), \ldots, u_N(x)$; in this case, the determination of each function is reduced to solving one ordinary differential equation of the second order with the boundary condition of the first kind.

The constructed computing scheme is characterised by only the first order of the error of approximation. If in approximation of the differential equation we replace the derivative u_t using a more accurate equality

$$u_t(x,t_{n-1}) \approx \frac{3u(x,t_{n+1}) - 4u(x,t_n) + u(x,t_{n-1})}{2\tau},$$

whose error is $O(\tau^2)$, then at small values of $\tau > 0$, as usually, we can reduce the part of the approximation error of the original equation which depends on the inaccuracy of substitution of the first derivative in respect of time.

The process of construction of the transverse schemes of the method of arbitrary lines for the hyperbolic equations is very similar to the case of parabolic equations. Similar analogues are even more evident in the case of the longitudinal schemes of the method. For example, if in the examined domain the straight lines, parallel to the axis x, are substituted by the straight lines $x = x_n = nh$, $n = 0, 1, \ldots, N$ ($Nh = 1$), then for the problem (19) we can construct the following longitudinal scheme of the method of arbitrary lines:

$$\begin{gathered} u_n''(t) + g_n(t)u_n'(t) = a_n(t)\frac{u_{n+1}(t) - 2u_n(t) + u_{n-1}(t)}{h^2} + \\ +b_n(t)\frac{u_{n+1}(t) - u_{n-1}(t)}{2h} - c_n(t)u_n(t) + f_n(t), \quad t > 0, \\ u_0(t) = \psi_0(t), \quad u_N(t) = \psi_1(t), \quad t \ge 0, \\ u_n(0) = \varphi(x_n), \quad u_N'(0) = \Phi(x_n), \quad n = 1, 2, \ldots, N-1. \end{gathered} \tag{21}$$

When writing this scheme we use the notations completely identical with those presented previously.

The approximating problem (21) may be interpreted as the Cauchy problem for the system of $N - 1$ linear ordinary differential equations of the second order in relation to unknown functions $u_n(t)$, $n = 1, 2, \ldots, N-1$, approximately representing the required function $u(x, t)$ on the straight line $x = x_n$, $n = 1, 2, \ldots, N-1$.

2.2.3. The method of arbitrary lines for elliptical equations

We examine this method with special reference to the elliptical Dirichlet problem of the partial type.

Let it be that in a rectangle $\Pi = \{0 \le x \le X,\ 0 \le y \le Y\}$ we have the unique solution of the Dirichlet problem

$$u_{yy} + Au = f(x, y).$$
$$u(x,0) = \varphi(x), \quad u(x,Y) = \Phi(x), \tag{22}$$
$$u(0, y) = \psi(y), \quad u(X, y) = \Psi(y),$$

where

$$Au=a(x,y)u_{xx}+b(x,y)u_x+c(x,y)u, \quad a(x,y)\geq a>0.$$

We assume that the solution is sufficiently smooth and we shall find it approximately by the method of arbitrary lines.

The rectangle Π is approximated by a lattice domain, forming the straight lines $y = y_n = nh$, $n = 0, 1, \ldots, N$ ($Nh = Y$). The derivative u_{yy} at $y = y_n$, $n = 1, 2, \ldots, N–1$, is expressed approximately through the value of the function u taking into account the well known equation

$$u_{yy}(x, y_n) = \frac{u(x, y_{n+1}) - 2u(x, y_n) + u(x, y_{n-1})}{h^2} - \frac{h^2}{12}\frac{\partial^4 u(x, y_n + \theta h)}{\partial y^4},$$

where $-1 < \theta < 1$. Consequently, using the notations completely identical with those employed previously, we can write

$$\frac{u_{n+1}(x) - 2u_n(x) + u_{n-1}(x)}{h^2} + A_n u_n = f_n(x), \quad 0<x<X,$$
$$u_0(x) = \varphi(x), \quad u_N(x) = \Phi(x), \quad 0 \le x \le X, \tag{23}$$
$$u_n(0) = \Psi(y_n), \quad u_n(X) = \Psi(y_n), \quad n = 1, 2, \ldots, N-1.$$

The differential-difference problem (23) approximate the initial Dirichlet problem with the error on the order of h^2. It is reduced to solving a system of ordinary differential equations of the second order with the boundary conditions of the first kind.

The problem of constructing the computing schemes of the method of arbitrary lines was discussed previously on the example of rectangular regions oriented in respect of the axis. In the case of domains of a more complicated form, the application of the method may be linked with certain difficulties. In fact, the projections g_n, $n = 0, 1, \ldots, N$, on the co-ordinate axis of the internal sections of the domain by the selected straight lines coincide at any number of the straight lines only for the rectangle oriented in the appropriate direction. In a general case, the n-th equation of the examined computing scheme of the method, linking the approximate values of the required solution on the adjacent straight lines with the numbers from $n - p$ to $n + q$, without additional assumptions, was determined only on the intersection

$$G_n = \bigcap_{-p \le i \le q} g_{n+i}.$$

However, if it is assumed that these equations are also fulfilled outside G_n, then even for very simple domains the approximating problem at some values of N may prove to be insolvable. In order to overcome these difficulties, it is sometimes preferred to extend continuously inside the domain of the function

giving the boundary conditions, and the equations of the approximated scheme are examined only on the appropriate intersections G_n, and the computing algorithm must be constructed using a special procedure. It is often possible to avoid these difficulties, dividing the examined domain into bands by not straight but by curved lines selected taking into account the shape of the domain.

It should also be mentioned that the method of arbitrary lines is used for the numerical solution of differential equations of high order, non-linear problems and systems of equations.

2.3. The net method for integral equations (the quadrature method)

The net method in application to integral equations is also referred to as the quadrature method (the mechanical quadrature method). We examine the main concepts of this method for calculating a solution of the integral equation

$$Au \equiv u(x) - \int K(x,y)u(y)dy = f(x), \quad 0 \le x \le 1,$$

with a smooth kernel $K(x,y)$, assuming that $|K(x,y)| < \rho < 1$.

Specifying N, we set $h = 1/N$ and determine the table $[u]_h$ of the values of the solution on the net

$$x_n = nh \quad (n = 0,1,\ldots,N).$$

To obtain a difference scheme in the equality

$$u(x_n) - \int K(x_n,y)u(y)dy = f(x_n), \quad n = 0,1,\ldots,N,$$

we approximately substitute the integral with a sum using, for example, the quadrature formula of trapezoids

$$\int \varphi(y)dy \approx h\left(\frac{\varphi_0}{2} + \varphi_1 + \ldots + \varphi_{N-1} + \frac{\varphi_N}{2}\right), \quad h = \frac{1}{N},$$

whose error is the quantity $O(h^2)$, if $\varphi(y)$ is the twice continuously differentiable function. After this substitution of the integral we obtain

$$u_n - h\left(\frac{K(x_n,0)}{2}u_0 + K(x_n,h)u_1 + \ldots + K(x_n,y_{N-1})u_{N-1} + \right.$$
$$\left. + \frac{K(x_n,1)}{2}u_N\right) = f_n, \quad n = 0,1,\ldots,N,$$

or

$$A_h u^{(h)} = f^{(h)}, \tag{24}$$

where

$$A_h u^{(h)} = \begin{cases} g_0, \\ g_1, \\ \ldots \\ g_N, \end{cases} \qquad f^{(h)} = \begin{cases} f(0), \\ f(h), \\ \ldots \\ f(Nh), \end{cases}$$

$$g_n = u_n - h\left[\frac{K(x_n,0)}{2}u_0 + K(x_n,h)u_1 + \ldots + \frac{K(x_n,1)}{2}u_N\right].$$

The constructed difference scheme $A_h u^{(h)} = f^{(h)}$ approximates the problem $Au = f$ on the solution u with the second order in relation to the step h, since the quadrature formula of the trapezoids has the second order of accuracy.

Let $u^{(h)} = (u_0, u_1, \ldots, u_N)$ be some solution of the system (24), and let u_s be one of the components of the solution which, according to their modules, are not less than any of the remaining $|u_s| \geq |u_n|$, $n = 0, 1, \ldots, N$. The equation with the number $n = s$ of system (24) gives the inequality

$$|f(x_s)| = \left| u_s - h\left(\frac{K(x_s,0)}{2} u_0 + K(x_s,h)u_1 + \ldots + \frac{K(x_s,1)}{2} u_N \right) \right| \geq$$

$$\geq |u_s| - h\left(\frac{\rho}{2} + \rho + \ldots + \rho + \frac{\rho}{2} \right) |u_s| = (1 - nh\rho)|u_s| = (1-\rho)|u_s|.$$

Therefore, the scheme (24) is stable:

$$\left\| u^{(h)} \right\|_{U_h} = \max_n |u_n| = |u_s| \leq \frac{1}{1-\rho} |f(x_s)| \leq \frac{1}{1-\rho} \left\| f^{(h)} \right\|_{F_h}.$$

In particular, at $f(x_n) \equiv 0$ it follows that the system (24) does not have non-trivial solutions, and consequently is uniquely solvable at any right-hand side. The solution $u^{(h)}$ of the problem $A_h u^{(k)}$ because of the theorem of convergence satisfies the inequality

$$\left\| [u]_h - u^{(h)} \right\|_{U_h} = \max_n |u(mh) - u_n| \leq ch^2,$$

where c is some constant.

According to the above considerations, the net method is used for solving multi-dimensional integral equations.

3. VARIATIONAL METHODS

The principle of many of the variational methods is based on the formulation of the examined problem of mathematical physics in the variational form as a problem of finding a function realising a minimum (or, in a general case, an extremum) of some functional, followed by determination of approximation to these functions.

3.1. Main concepts of variational formulations of problems and variational methods

3.1.1. Variational formulations of problems

Many problems of mathematical physics may be formulated as variational problems, i.e. as problems of finding functions giving an extremum to some functionals. Writing the required conditions of extrema of these functions, we obtain equations referred to as Euler equations which can be symbolically written in the form $Au = f$, where u is a function realising the extremum of the specific functional $J(u)$ examined on the set $D(A)$.

On the other hand, let us assume that in some real Hilbert space H with the scalar product $(\cdot, \cdot)$ and the norm $\| \cdot \| = (\cdot, \cdot)^{1/2}$ we examine a problem

written as the equation

$$Au = f, \qquad f \in H \tag{25}$$

where f is a given element, A is a linear, symmetric and positive definite operator with the definition domain $D(A)$ dense in H. Then, if (25) gives the solution $u \in D(A)$, then, assuming $J(u) \equiv (Au,u) - 2(u, f)$, it may easily be verified that u is also the solution of the variational problem of the type

$$J(u) = \inf_{v \in D(J)} J(v), \tag{26}$$

where $D(J) \equiv D(A)$. Thus, the problems (25), (26) are equivalent.

The equivalence of the problems (25), (26) offers new possibilities of examining (25), (26) and also constructing approximate solutions of this problem by variational (direct) methods.

3.1.2. Concepts of the direct methods in calculus of variations

One of the methods of confirming the existence and actual determination of the solution of a specific variational problem is to reduce this problem to the problem of the existence of a solution of some differential equation or a system of differential equations (Euler equations). However, this route does not always lead to the required result. The difficulties formed here resulted in the effort to find, in calculus of variations, other direct methods, i.e. the method not using the reduction of the variational problem to differential equations. In addition to this, integration of the resultant differential equations of the variational problems can be carried out in the final form only in a small number of cases. Therefore, it is necessary to obtain approximate solutions of these problems. This can be carried out examining the variational formulation (bypassing the problems of type (25)) or reducing them to the problems of type (26) and using direct methods which are also referred to as variational methods.

The development of direct methods of calculus of variations has proved to be useful not only directly for variational problems but also for other areas of mathematics, in particular, they have been used widely in the theory of differential equations. The main concept using variational methods in differential equations may be described as follows.

If the given differential equation can be treated as the Euler equation for some functional and if it has been established, by some method, that this functional has an extremum in the class of some class of functions, then it has been established that the initial equation has a solution satisfying the boundary-volume conditions, corresponding to the examined variational problem. The direct methods of variational calculations make it possible not only to confirm the existence of the appropriate solution but also find this solution with any degree of accuracy.

There are many different procedures grouped under the common name *direct methods*, but the majority of these methods form the general concept which may be described as follows.

For definiteness, we examine a problem of determination of the minimum of some functional $J(u)$ defined on some class $D(J)$ of permissible functions. To ensure that a problem has any meaning, it is assumed that the class $D(J)$ contains functions for which $J(u) < +\infty$, and that $\inf J(u) = d > -\infty$. In this case, according to the definition of the greatest lower bound, there is a sequence of functions $u_1, u_2, \ldots, u_n, \ldots$ (it is referred to as the *minimising sequence*) such that $\lim_{n\to\infty} J(u_n) = d$. If for the sequence $\{u_n\}$ we have the limiting function $u^{(0)}$ and if the limiting transition $J(u^{(0)}) = \lim_{n\to\infty} J(u_n)$ appears to be regular, then $J(u^{(0)}) = d$, i.e. the limiting function $u^{(0)}$ is also the solution of the examined problem.

Thus, the solution of the variational problem (or problem (25)) after reducing it to the appropriate variational problem by the direct method consists of:

1) constructing the minimising sequence $\{u_n\}$;
2) confirmation of the existence for this sequence of the limiting function $u^{(0)}$;
3) proof of the regularity of the passing to the limit

$$J(u^{(0)}) = \lim_{n\to\infty} J(u_n).$$

The members of the minimising sequence in this case may be regarded as approximate solutions of the corresponding variational problem. The method of constructing the minimising sequences, generally speaking, characterises each of the direct methods used in calculus of variations, and the appropriate algorithm of construction of the approximate solutions of the examined variational problems.

3.2. The Ritz method

3.2.1. The classic Ritz method

Let us assume that in the real Hilbert space H with a scalar product $(\cdot, \cdot)$ and the norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$ we examine a problem written in the form of the operator equation (25), where A is the linear, symmetric, positive definite operator, i.e., A is linear, also $(Au, v) = (v, Au)$, $(Au, u) \geq \gamma^2\|u\|^2$, $\gamma = \text{const} > 0$, at any u, v from the domain of definition $D(A)$ of the operator A, assumed to be dense in H. In addition to (25) we examine the problem

$$J(u) = \inf_{v\in D(A)} J(v), \qquad J(v) \equiv (Au, u) - 2(u, f). \tag{27}$$

The following claims must be taken into account.

Theorem 1. To ensure that some element $u_0 \in D(A)$ gives the minimum value to the functional $J(u)$, it is necessary and sufficient that this element satisfies the condition $Au_0 = f$. This element is unique.

This theorem shows that the problems (25), (27) are equivalent.

We specify the functions $\varphi_1^{(N)}, \varphi_2^{(N)}, \ldots, \varphi_N^{(N)}$, $N = 1, 2, \ldots$, where each of these functions belongs to $D(A)$. H_N denotes the span of the functions $\varphi_i^{(N)}$, $i = 1, \ldots, N$. It is assumed that the following conditions are satisfied:

1) at any N the functions $\varphi_1^{(N)}, \ldots, \varphi_N^{(N)}$ are linearly independent;

2) the sequence of the sub-spaces $\{H_N\}$ is characterised by the limiting density in H, i.e. for any function $u \in H$ there are elements such as $\bar{u}_N \in H_N, N = 1, 2, \ldots,$ that $\left\| u - \bar{u}_N \right\|_A = \inf_{v \in H_N} \left\| u - v \right\|_A \le \varepsilon(u, N) \to 0$, $N \to \infty$ where $||u||_A = (Au, u)^{1/2}$, and $\varepsilon(u, N)$ is the estimate of the error of approximation of the function u by a means of $\{\varphi_i^{(N)}\}$.

The set of the functions $\varphi_1^{(N)}, \ldots, \varphi_N^{(N)}$, satisfying the above conditions, is referred to as the *basis* in H_N, and the functions included there $\varphi_i^{(N)}$ as *base* or *co-ordinate* functions.

The Ritz method of construction of the minimising sequence u_N, $N = 1, 2, \ldots$, i.e. a sequence of approximate solutions of the problem (25) (correspondingly problem (27)) may be formulated as follows. Let H_N be the span of the system $\varphi_1^{(N)}, \ldots, \varphi_N^{(N)}$. We formulate a problem of finding the minimum of the functional $J(u)$ on H_N, i.e. determination of the function $u_N \in H_N$, for which

$$J(u_N) = \min_v J(v), \qquad v \in H_N.$$

Since $v = \sum_{i=1}^{N} b_i \varphi_i$, then $J(u_N) = \min_{b_i} J(v)$, where

$$J(v) = J(b_1, \ldots, b_N) = \sum_{i,j=1}^{N} b_i b_j (A\varphi_i, \varphi_j) - 2\sum_{i=1}^{N} b_j (f, \varphi_j).$$

In order to find the minimum of the functional $J(v)$, we calculate its derivatives in respect of b_i and equate these derivatives to zero. We obtain the system of equations

$$\frac{\partial J(v)}{\partial b_i} = 0, \qquad i = 1, \ldots, N,$$

which is equivalent to the system of equations of the type

$$(Au_N, \varphi_i) = (f, \varphi_i), \quad i = 1, \ldots, N, \tag{28}$$

or, which is the same, to the system

$$\hat{A}\, a = f,$$

where $\hat{A}$ is the matrix with the elements $A_{ij} = (A\varphi_j, \varphi_i)$, $a = (a_1, \ldots, a_N)^T$, $f = (f_1, \ldots, f_N)^T$, $f_i = (f, \varphi_i)$.

Since A is a positive definite operator and $\{\varphi_i\}$ is the linearly independent system, then at $v_N = \sum_{i=1}^{N} b_i \varphi_i$, $b = (b_1, \ldots, b_N)^T \ne 0 = (0, \ldots, 0)^T$ we have

$$(\hat{A}b, b)_2 = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij} b_i b_j = (Av_N, v_N) \ge \gamma^2 \left\| v_N \right\|^2 > 0,$$

where $\hat{A}$ is a positive definite matrix and, therefore, a non-degenerate matrix. Consequently, the system (28) has a unique solution a determining at the same time, the unique function u_N.

For u_N we can obtain the apriori estimate. Multiplying equations (28) by a_i and summing them in respect in respect of i, we obtain $(Au_N, u_N) = (f, u_N)$; however, $(Au_N, u_N) \ge \gamma^2 ||u_N||^2$ and therefore

$$\|u_N\|^2 \le \frac{|(f,u_N)|}{\gamma^2} \le \frac{\|f\|\cdot\|u_N\|}{\gamma^2}.$$

Thus, the estimate $||u_N|| \le ||f||/\gamma^2$ is valid.

Let $u_0 \in D(A)$ be the exact solution of equation (25) (problem (27)). Consequently, at the arbitrary function $v \in D(A)$ the following equality is valid:

$$(A(u_0-v),\ u_0-v)=J(v) - J(u_0).$$

Then (since u_N minimises $J(v)$ on H_N) at the arbitrary function $v_N = \sum_{i=1}^{N} c_i\,\varphi_i$ from H_N

$$(A(u_0-u_N),\ u_0-u_N) = J(u_N) - J(u_0) \le J(v_N) - J(u_0) = (A(u_0-v_N),u_0-v_N).$$

Consequently,

$$\|u_0 - u_N\|A \le \inf_{v_N} \|u_0 - v_N\|A \le \varepsilon(u_{0,N}) \to 0, \qquad N \to \infty. \tag{29}$$

Thus, the sequence u_1, u_2, …, u_N, … is a minimising sequence, and functions u_N, $N = 1, 2, \ldots$, are the approximate solutions of problem (25) (or correspondingly of the problem (27)) and the estimate (29) is valid.

3.2.2. The Ritz method in energy spaces

Theorem 1 establishes the equivalence of the problems (25) and (27) but it does not contain any assertions indicating the existence of the solution $u_0 \in D(A)$ of these problems. In paragraph 3.2.1. we examined the classic formulation of the problem when the solution of the equation $Au = f$ is a function belonging to the domain of definition $D(A)$ of the operator A and satisfies this equation. It appears that in this formulation this solution can not exist. However, it does exist in a slightly wider (than $D(A)$) space. It is therefore necessary to change the formulation of the variational problem of minimisation $J(u)$ in order to guarantee that its solution exists.

Let it be that in examining (25) A is a symmetric positive definite operator with the domain of definition $D(A)$ dense in H. To $D(A)$ we introduce the scalar product and the norm $[\varphi,\psi] = (A\varphi,\psi)$, $[\varphi] = [\varphi,\varphi]^{1/2}$. Completing $D(A)$ using the given norm, we obtain the complete Hilbert space H_A, which is referred to as the *energy space* generated by the operator A. Each function from $D(A)$ belongs to space H_A but as a result of completion H_A may also contain elements not included in $D(A)$ (therefore, the representation of the scalar product $[\varphi,\psi]$ at arbitrary φ, $\psi \in H_A$ in the form $(A\varphi,\psi)$ is no longer required).

Let $u \in D(A)$; $F(u)$ is presented in the form

$$J(u)=[u,u]-2(f,u). \tag{30}$$

This form makes it possible to examine $F(u)$ not only in the domain of definition of operator A but also on all elements of the energy space H_A. Therefore, we expand the functional (30) (leaving behind it the usual

notation $J(u)$) on the entire space H_A and we find its minimum on this space. Since the operator is assumed to be positive definite, i.e. $(Au, u) = [u, u] \geq \gamma^2||u||^2$, $u \in D(A)$, $\gamma > 0$, then on complementing $D(A)$ in H_A the ratio of definiteness $[u, u] \geq \gamma^2||u||^2$ remains valid for any element $u \in H_A$.

Functional (f, u) is bounded in H_A:

$$|(f,u)| \leq \|f\| \cdot \|u\| \leq \|f\| \frac{[u]}{\gamma} = c[u].$$

Consequently, according to the Riesz theorem regarding the representation of the linear bounded functional in the Hilbert space there is an element $u_0 \in H_A$, uniquely defined by the element f and is such that for any $u \in H_A$ the equality $(f, u) = [u, u]$ holds. However, in this case $J(u)$ can be presented in the form

$$J(u)=[u,u]-2(f,u)=[u,u]-2[u,u_0]=[u-u_0]^2-[u_0]^2.$$

Consequently, in space H_A the functional $F(u)$ reaches the minimum at $u = u_0$. As already mentioned, u_0 is unique and belongs to H_A. It may appear that $u_0 \in D(A)$; consequently, u_0 is also the classic solution of the examined problem, i.e. satisfies (25). However, if $u_0 \in H_A$, but $u_0 \notin D(A)$, it will be referred to as the *generalised solution* of equation (25).

Thus, the initial problem has been reduced to the problem of minimisation of functional $J(u)$ in the energy space H_A. We now examine the Ritz method for the approximate solution of the last variational problem which in this case will be referred to as the *Ritz method in energy spaces*.

Let us assume that the linearly independent functions $\{\varphi_i\} \subset H_A$ are given; H_N denotes their span. It is assumed that the sequence of the sub-spaces $\{H_N\}$, $N = 1, 2, \ldots$, has limiting density in H_A, i.e. for any function $u \in H_A$ there are such elements $\bar{u}_N \in H_N$, $N=1,2,\ldots$, that

$$[u-\tilde{u}_N] = \inf_w [u-w] \leq \varepsilon(u,N) \to 0,$$

$N \to \infty$, $w \in H_N$, where $\varepsilon(u, N)$ are the estimates of the approximation error. The *Ritz method* can now be formulated as follows: it is necessary to find the element $u_N \in H_N$, minimising $F(u)$ on the sub-space H_A. The realisation of the algorithm may be described as follows:

1) specific N and $\{\varphi_i\}$, $\varphi_i \in H_A$ are defined;

2) the approximate solution is found in the form $u_N = \sum_{i=1}^{N} a_i \varphi_i$;

3) coefficients a_i are determined from the conditions of minimisation of the functional $J(u_N)$, which lead to the system of equations $\partial F(u_N)/\partial a_i = 0$, $i = 1, \ldots, N$. This system can also be written in the form

$$\hat{A}a = f \quad \text{or} \quad [u_N, \varphi_i] = (f, \varphi_i), \quad i = 1,\ldots,N,$$

where $a = (a_1, \ldots, a_N)^T$ and $f = (f_1, \ldots, f_N)^T$ are the N-dimensional vectors, and $f_i = (f, \varphi_i)$, and $\hat{A}$ is the Gram matrix of the system $\{\varphi_i\}$ in the scalar product of the space H_A with the elements $A_{ij} = [\varphi_j, \varphi_i]$, $1 \leq i, j \leq N$. However $A_{ij} = [\varphi_i, \varphi_j] = A_{ji}$, hence $\hat{A}$ is symmetric, and because of the inequality

$$(\hat{A}b,b)_2 = \sum_{i,j=1}^{N} A_{ij} b_i b_j = \left[\sum_{i=1}^{N} b_i \varphi_i\right]^2 \geq \gamma^2 \left|\sum_{i=1}^{N} b_i \varphi_i\right|^2 > 0$$

at $b = (b_1, \ldots, b_N)^T \neq 0$ the matrix $\hat{A}$ is also positive definite. Therefore, the system $\hat{A}a = f$ has the unique solution a, unambiguously defining the element u_N, for which the inequality $[u_N] \leq ||f||/\gamma$ is valid.

The following claim is justified.

Theorem 2. *If the sequence of the sub-space $\{H_N\}$ has limiting density in H_A, then the approximate solution u_N, obtained by the Ritz method, converges at $N \to \infty$ to the generalised solution u_0 of the problem in the metrics of the space H_A, and the following estimate holds:*

$$[u_0 - u_N] \leq \varepsilon(u_0, N) \to 0, \quad N \to \infty.$$

3.2.3. Natural and main boundary-value conditions

The affiliation of element u to the domain of definition $D(A)$ of the operator A often indicates that u satisfies the boundary-value conditions written in the form $T_k u = 0$, $k = 1, \ldots, K$ (here T_k is the operator defining the k-th boundary-value condition). As a result of completion of $D(A)$ in respect of the norm $[\cdot]$ elements may appear in the resultant energy space H_A which do not satisfy some of the conditions $T_k u = 0$. If H_A contains elements not satisfying some condition $T_k u = 0$, then this boundary-value condition is *natural* for the operator A. The boundary-value condition, satisfied by both the elements from $D(A)$ and elements from H_A, is referred to as the *main* boundary-value condition.

The practical importance of being able to distinguish between these conditions is that the basis functions $\{\varphi_i\}$ are not inevitably governed by the natural boundary-value conditions because it is sufficient to take them from the energy space (and not inevitably from $D(A)$). This circumstance greatly facilitates the selection of φ_i when solving many problems important for practice, especially in the case of the multi-dimensional domain with the complicated form of the boundary. It should be mentioned that in the case of the main boundary-value condition, the problem of constructing functions φ_i satisfying these conditions, remains.

We shall describe an approach which makes it possible to determine, for a specific problem, whether the given boundary-value condition is natural. We examine a problem of minimisation of the functional $J(u)$ and assume that there is a function u_0, realising the minimum $J(u)$ in the class of functions, generally speaking, not satisfying this condition. Using the means of variational calculations, we can find the necessary conditions of realisation of the minimum of the functional $J(u)$ by function u_0. If it appears that they also include the examined boundary-value condition, then this condition is natural.

Finally, we present a simple feature (without its theoretical substantiation) which makes it possible to distinguish the natural boundary-value condition from the main conditions and is suitable for solving a number

of boundary-value problems. Let it be that in (25) A is the differential operator of the order $2m$ satisfying some boundary-value condition of the type $T_k u = 0$. Consequently, the boundary-value condition is natural if the expression $T_k u$ contains derivatives of u of the order m and higher (in this case, $T_k u$ may include derivatives of the order smaller than m and also the function u itself with some weights). If $T_k u$ does not contain derivatives of u of the order m and higher, the condition $T_k u = 0$ is the main condition.

3.3. The method of least squares

Let A be a linear operator defined on some linear set $D(A)$ dense in the given Hilbert space H, and let us assume that it is necessary to solve the equation

$$Au = f, \tag{31}$$

where f is the given element of H. This can be carried out using the method of least squares based on the following: we select a sequence of linearly independent co-ordinate elements ($\varphi_k \in D(A)\ \forall k$); the approximate solution of equation (31) is constructed in the form $u_N = \sum_{k=1}^{N} a_k \varphi_k$, where a_k are constants determined from the requirement according to which the value of the residual functional $J(u_N) \equiv ||Au_N - f||^2$ is minimum.

The sequence $\{\varphi_k\}$ is assumed to be *A-dense*, i.e. the following condition is fulfilled: regardless of $u \in D(A)$ and the number $\varepsilon > 0$, we can find N and constants $c_1, \ldots, c_N$ such that $||Au - Au_N|| < \varepsilon$, where $u_N = \sum_{k=1}^{N} c_k \varphi_k$.

The minimisation condition for $J(u_N)$ leads to a system of linear equations for unknown $a_1, a_2, \ldots, a_N$. To determine the type of this system, it is sufficient to differentiate $J(u_N)$ in respect of $\bar{a}_m$. Consequently, we obtain a system of equations for $a_1, a_2, \ldots, a_N$:

$$\sum_{k=1}^{n} a_k (A\varphi_k, A\varphi_m) = (f, A\varphi_m), \quad m = 1, 2, \ldots, N. \tag{32}$$

It should be mentioned that the system (32) is symmetric. The determinant of the matrix of the system is the Gram determinant of the elements $A\varphi_1$, $A\varphi_2, \ldots, A\varphi_N$.

Lemma. *If the homogeneous equation $Au = 0$ has only a trivial solution, the approximate solutions obtained by the method of least squares can be constructed for any N and are determined by the unique procedure.*

Sufficient convergence conditions of the approximate solutions, obtained by the method of least squares, are given by the following theorem.

Theorem 3. *The method of least squares gives a sequence of approximate solutions converging to an accurate solution if the following conditions are fulfilled:*

1) *the sequence of the co-ordinate elements is A-dense*
2) *equation* (32) *is solvable;*

3) *there is a constant K such that* $\|u\| \le K\|Au\|$ *for any* $u \in D(A)$.

If the conditions of theorem 3 are fulfilled, then u_N is the approximate solution of equation (31). Therefore

$$\|u_N - u\| \le K\|Au_N - Au\| = K\|Au_N - f\|.$$

Consequently, if u_N is determined by the method of least squares, then $Au_N \to f$ at $N \to \infty$, and the last equation can be used to estimate the error of the approximate solution.

3.4. Kantorovich, Courant and Trefftz methods

3.4.1. The Kantorovich method

This method of solving variational problems greatly differs from the Ritz method. We describe its concept with special reference to the problem

$$Au = f(x,y) \text{ in } \Omega, \quad u|_\Gamma = 0, \tag{33}$$

where A is an elliptical operator of the second order, $\Omega = \{(x, y) : \Phi_1(x) < y < \Phi_2(x), x \in (a, b)\}$, Φ_1, Φ_2 are the functions smooth in respect of $x \in [a, b]$, Γ is the boundary of the domain Ω. It is assumed that the operator A acts in the Hilbert space $H \equiv L_2(\Omega)$, and it is symmetric and positive definite. Consequently, the problem (33) is reduced to the problem of the minimum of the functional

$$J(u) = (Au,u)-(u,f)-(f,u).$$

The approximate solution is found in the form

$$u_N(x,y) = \sum_{k=1}^{N} f_k(x)\varphi_k(x,y),$$

where $\varphi_k(x, y)$ are the known functions equal to zero on Γ with the exception possibly of the straight lines $x = a$ and $x = b$. The functions of one variable $f_k(x)$ are determined from the requirement that the functional $J(u_N)$ has the minimum value. Using the conventional method of calculus of variations, we obtain for $f_k(x)$ a system of differential equations; to these, we add the boundary-value conditions at $x = a$ and $x = b$ resulting from the boundary-value conditions of problem $f_k(a) = f_k(b) = 0$, $k = 1, 2, \ldots, N$.

Thus, the principle of the Kantorovich method is to reduce (approximately) integration of the equation in partial derivatives to the integration of the system of ordinary differential equations.

3.4.2. Courant method

R. Courant proposed a method of constructing a minimising sequence which for the known conditions converges not only in the mean but also uniformly together with the sequences of the derivatives to some order.

Let us assume that we have the differential equation

$$Au=f \tag{34}$$

and it is required to find its solution defined in some finite domain Ω of

the m-dimensional space $(x_1, x_2, \ldots, x_m)$ and satisfying some homogeneous boundary-value conditions of the boundary Γ of this volume. We examine the Hilbert space $L_2(\Omega)$. It is assumed that operator A is positive definite on the linear set of sufficiently smooth functions satisfying the boundary-value conditions of our problem. Consequently, equation (34) has a solution satisfying (possibly, in the generalised sense) the given boundary-value conditions; this solution realises the minimum of the functional

$$J(u) = (Au,u) - (u,f) - (f,u).$$

It is now assumed that f has continuous derivatives in respect of $(x_1, x_2, \ldots, x_m)$ to the order $k - 1$ inclusive and at least the quadratically summable derivatives of the order k. We compile a functional

$$\Phi(u) = J(u) + \sum_{j=0}^{k} \sum_{a_1+a_2+\ldots+a_m=j} \left\| \frac{\partial^j (Au - f)}{\partial x_1^{a_1} \partial x_2^{a_2} \ldots \partial x_m^{a_m}} \right\|^2 .$$

Evidently, $\Phi(u) \geq J(u)$. Also, the function u, realising the minimum $J(u)$ realises the minimum $\Phi(u)$ since this function makes both the first and second term smallest in $\Phi(u)$, with the second term converted to zero by the function. It follows from this that the solution of our boundary-value problem can be found by finding the solution of the problem of the minimum of $\Phi(u)$. For this functional we construct the minimising sequence $\{u_N\}$, for example, using the Ritz method. Consequently, it is evident

$$\left\| \frac{\partial^j (Au_N - f)}{\partial x_1^{a_1} \partial x_2^{a_2} \ldots \partial x_m^{a_m}} \right\|_{N \to \infty} \to 0, \quad j = 1, 2, \ldots, k.$$

These relationships make it possible to propose additional conclusions on the convergence of the minimum sequence.

The introduction of additional terms in $\Phi(u)$ complicates calculations by the Ritz method; however, this complication may be justified if according to the meaning of the task it is required to obtain a uniformly converging sequence.

3.4.3. Trefftz method

The Ritz method gives the value of the minimum of the functional with an excess. It is also desirable to have a method of constructing an approximate solution giving this quantity with a shortage. E. Trefftz proposed a method which in some cases can be used to construct a sequence of functions giving the approximation to the required minimum of the functional from below. The concept of the Trefftz method may be described as follows. Whereas in the Ritz method the approximate solution is found in the class of the functions exactly satisfying the boundary-value condition but not the differential equation, in the Trefftz method the approximate solution exactly satisfies the differential equation but, generally speaking, it does not satisfy the given boundary-value conditions. We explain the concept of this method with special reference to the Dirichlet problem for the Laplace

equation (although it can be used for a considerably wider range of the problems).

Let it be required to find a function harmonic in the domain Ω and satisfying the boundary-value condition

$$u|_{\Gamma} = f(x), \tag{35}$$

where $f(x)$ is a function which, to simplify considerations, is assumed to be continuous on the boundary Γ. The required function can be determined as a function minimising the integral

$$J(u) = J_1(u,u) = \int_{\Omega} \{\mathrm{grad} u\}^2 d\Omega$$

in comparison with any other function satisfying the condition (35). The Trefftz method may be described as follows.

It is assumed that the given sequence of the linearly independent functions φ_k harmonic in Ω is complete in the following sense: for any function φ harmonic in Ω, quadratically-summable in Ω together with its first derivatives, on the basis of the given number $\varepsilon > 0$ we can find the natural number N and constants $a_1, a_2, \ldots, a_N$ such that

$$J\left(\varphi - \sum_{k=1}^{n} a_k \varphi_k\right) = \int_{\Omega} \left\{\mathrm{grad}(\varphi - \sum_{k=1}^{N} a_k \varphi_k)\right\}^2 d\Omega < \varepsilon.$$

We find the approximate solution of our problem in the form

$$u_N = \sum_{k=1}^{N} a_k \varphi_k,$$

where N is an arbitrary number; coefficients a_k are found from the condition $J(u - u_N) = \min$, where u is the required solution of the problem. Equating the derivatives $\partial J(u - u_N)/\partial a_k$ to zero, we obtain a system of equations

$$J_1(u_N - u, \varphi_k) = 0, \qquad k = 1,2,\ldots,N.$$

After integrating by parts, we obtain

$$\sum_{j=1}^{N} a_j \int_{\Gamma} \varphi_j \frac{\partial \varphi_k}{\partial \nu} d\Gamma = \int_{\Gamma} f \frac{\partial \varphi_k}{\partial \nu} d\Gamma, \quad k = 1,2,\ldots,N.$$

The given system has the unique solution $\{a_j\}$ which determines the unambiguous approximate solution u_N. It may be shown that $(u - u_N) \to 0$ *uniformly in any closed domain, located completely inside* Ω, *and the derivatives of any order of* u_N *uniformly converge to the appropriate derivative of* u.

The well known shortcoming of the Trefftz method is the difficult actual construction of the complete system of harmonic functions. If the domain Ω is plane, simply connected with a sufficiently smooth boundary, the system of harmonic polynomials will be complete; if Ω is multiply connected, the complete system is formed by some harmonic rational functions. It is far more difficult to show such a system in multi-dimensional domains.

3.5. Variational methods in the eigenvalue problem

The variational methods are used successfully for the approximate solution of the eigenvalue problem, namely the problem

$$Au = \lambda u. \tag{36}$$

We examine several approaches used here, using the Ritz method to solve (36).

Let A be a linear, self-adjoint operator lower bounded and acting in the Hilbert space H. We set

$$d = \inf \frac{(Au,u)}{(u,u)}. \tag{37}$$

The number d is the *smallest eigenvalue of the operator A* if there is element u_0 such that

$$d = \frac{(Au_0,u_0)}{(u_0,u_0)}.$$

If it is assumed that such an element does exist, the determination of the smallest eigenvalue of operator A is reduced to determining the lower bound of the values (37) or, which is the same, to the problem of determination of the lower bound of the quantity (Au, u) at the additional condition $(u,u) = 1$.

We shall show that this problem can be solved by the Ritz method. We take a sequence of linearly independent elements $\{\varphi_n\}$ included in the domain of definition of operator A and assume that this sequence has the following properties:

1) it is complete in H;

2) for any element u from the domain of definition of the operator A, we can find a natural number N and also constants $\alpha_1, \alpha_2, \ldots, \alpha_N$ so that $(A(u - u_N)) < \varepsilon$, where $u_N = \sum_{k=1}^{N} \alpha_k \varphi_k$, ε is an arbitrary positive number.

We set $u_N = \sum_{k=1}^{N} a_k \varphi_k$, and select constant coefficient a_k such that u_N satisfies the relationship $(u_N, u_N) = 1$ and quantity (Au_N, u_N) is minimum. Therefore, to find the minimum of the function of N variables (generally speaking, complex)

$$(Au_N, u_N) = \sum_{k,m=1}^{N} (A\varphi_k, \varphi_m) a_k \bar{a}_m,$$

linked by the equation $(u_N, u_N) = 1$ we use the method of the Lagrange multipliers. We compile the function $\Phi = (Au_N, u_N) - \lambda(u_N, u_N)$, where λ is the so-far undetermined numerical multiplier, and equate to zero its partial derivative in respect of α_m and in respect of β_m, where α_m and β_m denote the real and imaginary parts of the coefficient a_m respectively. Therefore, we obtain a system of equations $\partial\Phi / \partial\bar{a}_m = 0, m = 1, 2, \ldots, N$, or, in the open form, the system

$$\sum_{k=1}^{N} a_k [(A\varphi_k, \varphi_m) - \lambda(\varphi_k, \varphi_m)] = 0, \qquad m = 1, 2, \ldots, N. \tag{38}$$

The system (38) is linear homogeneous in relation to unknown a_k which cannot simultaneously convert to zero, otherwise the equation $(u_N, u_N) = 1$ would be violated. It follows from here that the determinant of the system (38) should convert to zero; this gives the equation for

$$\begin{vmatrix} (A\varphi_1,\varphi_1)-\lambda(\varphi_1,\varphi_1) & \dots & (A\varphi_N,\varphi_1)-\lambda(\varphi_N,\varphi_1) \\ \dots\dots & \dots & \dots\dots \\ (A\varphi_1,\varphi_N)-\lambda(\varphi_1,\varphi_N) & \dots & (A\varphi_N,\varphi_N)-\lambda(\varphi_N,\varphi_N) \end{vmatrix} = 0. \tag{39}$$

Equation (39) has exactly N routes. Let it be that λ_0 is one of these routes. Substituting it into system (38) we make its determinant equal to zero, and this system has non-trivial solutions. Let $a_k^{(0)}$, $k = 1, 2, \dots, n$, be such a solution. Therefore, $\mu a_k^{(0)}$, where μ is the arbitrary numerical multiplier, also satisfies system (38). Substituting $\mu a_k^{(0)}$ into equation $(u_N, u_N) = 1$, we obtain the value μ. Replacing the notation $\mu a_k^{(0)}$ by $a_k^{(0)}$, we can use $a_k^{(0)}$ to denote the solution of (38) that satisfies equation $(u_N, u_N) = 1$. Substituting $\lambda=\lambda_0$ and $a_k = a_k^{(0)}$ into (38) we obtain an identity which can written in the following form

$$\sum_{k=1}^{N} a_k^{(0)}(A\varphi_k,\varphi_m) = \lambda_0 \sum_{k=1}^{N} a_k^{(0)}(\varphi_k,\varphi_m), \qquad m = 1, 2, \dots, N.$$

Multiplying it by $\bar{a}_m^{(0)}$ and summing up in respect of all m, and also taking into account the equality $(u_N, u_N) = 1$ we obtain

$$\lambda_0 = (Au_N^{(0)}, u_N^{(0)}), \tag{40}$$

where $u_N^{(0)} = \sum_{k=1}^{N} a_k^{(0)}\varphi_k$.

Formula (40) shows that:

1) equation (39) has only real roots, if the operator A is self-adjoint;
2) one of the elements $u_N^{(0)}$ realises the minimum (Au_N, u_N);
3) this minimum is equal to the smallest of the roots of the equation (39).

With increase of N, this minimum which will be denoted by $\lambda_N^{(0)}$ does not increase; at the same time, it is not lower than d. It follows from here that at $N \to \infty$ quantity $\lambda_N^{(0)}$ tends to the limit which is greater than or equal to d. In addition to this, it may be proven that this limit is equal to d; $\lambda_N^{(0)} \to d$, $N \to \infty$, which also justifies the application of the Ritz method to the problem of eigenvalues.

To obtain the approximate value of the second eigenvalue, we find the minimum of the scalar product (Au_N, u_N) at additional conditions $(u_N, u_N) = 1$ and $(u_N^{(0)}, u_N) = 0$, where $u_N^{(0)} = \sum_{k=1}^{N} a_k^{(0)}\varphi_k$ is the approximate

value of the first normalised eigenfunction of operator A. This problem can again be solved by the Lagrange method.

Similarly, we can obtain approximate values of the next eigenvalues; they all are routes of the equation (39).

4. PROJECTION METHODS

The large group of methods of approximate solution of the equations of the type $Au = f$ uses the following approach: the solution is found in the form $u_N = \sum_{i=1}^{N} a_i \varphi_i$, where the coefficients a_i are determined from the condition of equality to zero of the projection of the residual $r_N = Au_N - f$ on the span of basis functions ψ_i, $i = 1,2,\ldots,N$, generally speaking, differing from φ_i. One of the partial cases of this equation is the requirement of orthogonality of r_N in respect of all ψ_i: $(r_N,\psi_i) = (Au_N - f,\psi_i) = 0$, $i = 1,2,\ldots,N$. Since methods of this type are not linked directly with minimization of some function, they are included in the group of projection methods. Some of the methods used most frequently in practical calculations are representatives of this group, such as the Bubnov–Galerkin methods, the Galerkin–Petrov methods, the methods of moments, the collocation method.

4.1. The Bubnov–Galerkin method

The main shortcoming of the Ritz method is that it can be used only for equations with symmetric positive definite operators. This shortcoming does not exist in the Bubnov–Galerkin method (sometimes, it is referred to simply as the Galerkin method).

4.1.1. The Bubnov-Galerkin method (a general case)

Let us assume that we examine the equation

$$Au = f, \; f \in H, \tag{41}$$

in the Hilbert space H; here A is the linear operator with the definition domain $D(A)$ which may not be symmetric, bounded and positive definite. The Bubnov–Galerkin method of construction of the approximate solution of equation (41) consists of the following:

1) We select the basis functions $\{\varphi_i\}$, $i=1,\ldots,N$, $\varphi_i \in D(A)$;

2) The approximate solution is found in the form $u_N = \sum_{i=1}^{N} a_i \varphi_i$;

3) Coefficients a_i are determined from the condition of orthogonality of residual $Au_N - f$ in relation to $\varphi_1,\ldots,\varphi_N$: $(Au_N - f,\varphi_i) = 0$, $i = 1,\ldots,N$, or

$$\sum_{k=1}^{N} (\varphi_i, A\varphi_k) a_k = (f, \varphi_i), \quad i = 1,\ldots,N. \tag{42}$$

It should be mentioned that the form of the equation (42) coincides with the appropriate equations of the Ritz algorithm (if $\varphi_i \in D(A)$). If A is a symmetric positive definite operator, the Bubnov–Galerkin and Ritz method coincide.

H_N denotes the span of the system $\{\varphi_i\}$, $i = 1,\ldots N$, $\varphi_i \in D(A)$ and AH_N the span of the function $\{A\varphi_i\}$. It should be mentioned that if the homogeneous equation $Au = 0$ has only a zero solution, then the functions $A\varphi_1,\ldots,A\varphi_N$ are linearly independent. P_N denotes the operator of orthogonal projection on H_N. It may easily be shown that the *requirements of equality to zero of the orthogonal projection* $\Phi_1 \equiv P_N \Phi$ *of some element* $\Phi \in H$ *is equivalent to the system* $(\Phi,\varphi_i) = 0$, $i = 1,2,\ldots,N$. Thus, system (42) is equivalent to the equation

$$P_N \; Au_N = P_N f,$$

used widely in examining the convergence of the general case of the Bubnov-Galerkin algorithm.

Let $P_N^{(1)}, P_N^{(2)}$ be the operators of the orthogonal projection on H_N, AH_N, respectively. We introduce the notation:

$$\tau_N = \min_{v_N} \frac{\left\| p_N^{(1)} v_N \right\|}{\left\| v_N \right\|},$$

where $\nu_N \in AH_N, \nu_N \neq 0$.

Theorem 4. Let it be that:

1. $\tau_N \geq \tau > 0$ *where constant* τ *is independent of* N;
2. *The system* $\{\varphi_i\}$ *is A-dense.*

Consequently, the sequence $\{Au_N\}$ *converges to* Au *at any* $f \in H$; *the following estimate is valid in this case:*

$$\left\| Au - Au_N \right\| \leq \left(1 + \frac{1}{\tau_N}\right) \left\| f - P_N^{(2)} f \right\| \leq \left(1 + \frac{1}{\tau}\right) \left\| f - P_N^{(2)} f \right\|.$$

However, if there is a bounded operator A^{-1}, then u_N *converges to* u *as* $N \to \infty$ *and the error estimate is valid:*

$$\left\| u - u_N \right\| \leq \left\| A^{-1} \right\| \left(1 + \frac{1}{\tau}\right) \left\| f - P_N^{(2)} f \right\|.$$

Thus, if in the general Bubnov–Galerkin algorithm we can estimate the quantity $\tau_N > \tau > 0$ from below, we can prove the convergence of the $r_N = Au_N - f$ to zero, and also $u_N \to u$, $N \to \infty$ in respect of the norm $\|\cdot\|$ and the norm of the type $\|u\|_A = \|Au\|$ under the condition of invertibility of operator A.

4.1.2 The Bubnov–Galerkin method ($A = A_0 + B$)

Let the operator in (41) be represented in the form $A = A_0 + B$, where A_0 ('the main part' of operator A) is a self-adjoint positive definite operator with the domain of definition $D(A)$ dense in H. We introduce the energy space H_A of the operator A_0 with the scalar product $[u,v]$ and the norm $[u] = [u,u]^{1/2}$. We multiply (41) by the arbitrary function $v \in H_A$. Consequently, we obtain the equality

$$[u,v] + (Bu,v) = (f,v) \qquad \forall v \in H_A, \tag{43}$$

which permits the introduction of the generalized formulation of the problem (41).

Definition. Function $u \in H_A$ is referred to as the *generalized solution* of equation (41) if u satisfies (43).

It is assumed that this generalized solution does exist. If it appears that $u \in D(A)$ then because of the relationship $[u,v] = (A_0 u, v)$ we obtain $(A_0 u + Bu - f, v) = 0$. Since according to the assumption $D(A_0)$ is dense in H, then H_A will also be dense in H. Therefore, from the last relationship it maybe concluded that u also satisfies (41).

We formulate the Bubnov–Galerkin method for solving the problem examined here:

1) We select basis functions $\{\varphi_i\}$ in H_A, i.e., in this case, it is sufficient that φ_i belongs to H_A and not to $D(A)$;

2) The approximate solution u_N is found in the form $u_N = \sum_{i=1}^{N} a_i \varphi_i$;

3) Coefficients a_i are determined from the systems of equations of the type

$$[u_N, \varphi_i] + (Bu_N, \varphi_i) = (f, \varphi_i), \qquad i = 1, \ldots, N, \tag{44}$$

or in the matrix form $\hat{A}a = f$, where, $\hat{A}a = (A_{ij})$, $A_{ij} = [\varphi_j, \varphi_i] + (B\varphi_j, \varphi_i)$, $a = (A_1, \ldots, A_N)^T$, $f = (f_1, \ldots, f_N)^T$, $f = (f_i, \varphi_i)$ After determining a from the system $\hat{A}a = f$, we construct the approximate solution using the equation

$$u_N = \sum_{i=1}^{N} a_i \varphi_i.$$

Theorem 5. *Let* (41) *have a unique generalized solution and the operator* $T = A^{-1}B$ *is completely continuous in* H_A. *It is also assumed that the sequence of the subspaces* H_N, *being the spans of* $\{\varphi_i\}$, *is characterised by the limiting density in* H_A. *Consequently, at a sufficiently high N, the Bubnov–Galerkin method gives the unique approximate solution* u_N. *The sequence* u_N *converges in accordance with the norm* H_A *to the generalized solution* u, *and the estimates of the type*

$$\min_{c_i}\left[u - \sum_{i=1}^{N} c_i \varphi_i\right] \le [u_N - u] \le (1 + \varepsilon_N) \min_{c_i}\left[u - \sum_{i=1}^{N} c_i \varphi_i\right],$$

are also valid, where $\varepsilon_N \to 0$ *as* $N \to \infty$.

The following claim is also valid.

Theorem 6. Let it be that:

1) *Equation* (41) *has a unique generalized solution* $u \in H_A$;

2) *The form* $a(u,v) = [u,v] + (Bu,v)$ is H_A*-defined and* H_N*-bounded,* i.e. *the following relationships are satisfied for it:*

$$a(u,u) \ge \gamma_0^2 [u]^2, \quad a(u,u) \le \gamma_1^2 [u][v], \quad \gamma_0, \ \gamma_1 = \text{const};$$

3) *The sequence of the subspaces* $\{H_N\}$, *where* H_N *is the span of the functions* $\{\varphi_i\}$, $i = 1, \ldots N$, *with limiting density in* H_A, i.e.

$$\min_{c_i}\left[u-\sum_{i=1}^{N}c_i\varphi_i\right]\le\varepsilon(u,N)\to 0,\quad N\to\infty,$$

where $\varepsilon(u,N)$ *is the estimate of the error of approximation.*

Consequently, at any finite N the system (43) *is uniquely solvable, the approximate solution* u_N *converges to u as* $N\to\infty$ *in the metrics* $[\cdot]$, *and the error estimate* $[u-u_N]\le c\varepsilon(u,N)$ *is also valid, where constant c is independent of N.*

It should be mentioned that in the examined case of the Bubnov–Galerkin method the basis functions (as in the Ritz method) maybe selected as not satisfying the boundary-value conditions, if they are natural.

4.2. The moments method

Let us examine the following equation in the Hilbert space H (which is assumed to be complex)

$$Au + Bu = f,\ f\in H, \tag{45}$$

where the operator A is K-positive definite (or positive definite in a general sense), i.e.

$$(Au,\ Ku)\ge\gamma^2\|u\|^2,\qquad (Au,\ Ku)\ge\beta^2\|Ku\|^2,$$

where β,γ are constants, $\beta,\gamma>0$, $u\in D(A)$.

The moments method of the approximate solution of (45) consists of the following:

1) We select the basis system $\{\varphi_i\}\subset D(A)$;

2) The approximate solution u_N is found in the form $u_N=\sum_{i=1}^{N}a_i\varphi_i$;

3) Coefficients a_i are determined from the system of equations

$$(Au_N + Bu_N - f,K\varphi_i) = 0,\qquad j = 1,\ldots,N. \tag{46}$$

It should be mentioned that because of K-positive definitness, the operator A has a bounded inverse operator: $\|A^{-1}\|\le 1/(\gamma\beta)$ and also the number $(Au,\ Ku)$ is real and the following property applies

$$(Au,Kv)=(Ku,Av)\qquad \forall u,v\in D(A).$$

On the basis of these properties on $D(A)$ we can introduce the scalar product $(u,v)_K=(Au,\ Ku)$, $v\in D(A)$ and, consequently, $D(A)$ after completion converts to the Hilbert space H_K with the norm

$$\|u\|_k-(u,u)_K^{1/2}.$$

Definition. The element $u\in H_K$ is referred to as the *generalized solution* of equation (45) if it satisfies the equality

$$(u,v)_K+(Tu,v)_K=(f_1,v)_K\qquad \forall v\in H_K. \tag{47}$$

where $T = A^{-1}B$, $f_1 = A^{-1}f$.

It is evident that if the element u satisfies (45), it is also a generalized

solution (the reverse, generally speaking, is not true).

The system (46) can now be written in the form

$$\sum_{i=1}^{N}[(\varphi_i,\varphi_j)_K+(B\varphi_i,K\varphi_j)]a_i=(f,K\varphi_j),\qquad j=1,\ldots,N. \tag{48}$$

The algorithm (48) may be regarded as the process of determination of the approximate generalized solution u_N.

Theorem 7. Let equation (45) *to have a unique generalized solution and the operator* $T = A^{-1}B$ *is completely continuous in* H_K.

Consequently:

1) *There is an integer* N_0 *such that at any* $N \geq N_0$ *system* (48) *has a unique solution* a_i;

2) *The approximate solution* u_N *converges in* H_K (*and also in* H) *to the solution of equation* (45).

4.3. Projection methods in the Hilbert and Banach spaces

4.3.1. The projection method in the Hilbert space

Let us examine the equation

$$Au = f,\qquad f\in H, \tag{49}$$

in the Hilbert space, where A, generally speaking, is an unbounded operator acting in H and having the bounded inverse operator A^{-1}. It is assumed that the domain of definition $D(A)$ and the range $R(A)$ are dense in H.

We now introduce the linearly independent system $\{\varphi_i\}$ in H. The corresponding N-dimensional subspaces, generated by $\{\varphi_i\}$ are denoted by M_N. It is assumed that the sequence $\{M_N\}$ has limiting density in H. We specify the sequence of projection operators P_N each of which maps H on the corresponding space M_N. It is also assumed that $||P_N|| \leq c$, $N = 1,2,\ldots$ (here P_N are not necessarily orthoprojectors, i.e., it is required that they only fullfill the properties $P_N^2 = P_N$, $P_N H = M_N$).

We also introduce a linearly independent system $\{\varphi_i\}$, $\varphi_i\in D(A)$. The subspaces, generated by $\{\varphi_i\}$, are denoted by H_N, and the span of the system $\{A\varphi_i\}$ by AH_N. It is assumed that the sequence of the subspaces $\{AH_N\}$ has limiting density in H and for any element $u\in H$

$$\varepsilon_N(u,N)=\inf_{u_N}\|u-u_N\|\to 0,\quad N\to\infty,\quad u_N\in AH_N.$$

The approximate solution of the problem (49) is represented in the form $u_N=\sum_{i=1}^{N}a_i\varphi_i$, where a_i is determined from the equation

$$P_N\ Au_N = P_N f. \tag{50}$$

Theorem 8. *Let it be that for any N and any element* $v\in AH_N$ *we have the inequality* $\tau||v||\leq||P_N v||$, *where the constant* $\tau > 0$ *does not depend on N. Consequently, at any N equation (50) has a unique solution* $u_N=\sum_{i=1}^{N}a_i\varphi_i$

and the residual $Au_N - f$ tends to zero as $N \to \infty$ and the estimates are valid

$$\varepsilon(f, N) \le \|Au_N - f\| \le (1 + c/\tau)\varepsilon(f, N),$$

where $\varepsilon(f,N) = \inf_{f_N} \|f - f_N\|$, $f_N \in AH_N$.

4.3.2. The Galerkin–Petrov method

We examine equation (49). The algorithm of the approximate solution of this equation maybe described as follows:

1) We specify two, generally speaking, different bases $\{\varphi_i\} \subset D(A)$, $\{\psi_i\} \subset H$;

2) The approximate solution u_N is found in the form $u_N = \sum_{i=1}^{N} a_i \varphi_i$;

3) The coefficients a_i are determined from the system of equations

$$(Au_N - f, \psi_i) = 0, \qquad i = 1, \ldots, N. \tag{51}$$

This method is also a particular case of the projection method formulated in the paragraph 4.3.1. Actually, let the operator P_N in (50) be the orthoprojector of the span M_N of the functions $\{\psi_i\}$. Consequently, equation (50) is equivalent to a system of equations $(Au_N - f, \psi_i) = 0$, $i = 1, \ldots, N$, i.e., the system (51). Thus, the Galerkin–Petrov method is a particular case of the projection method examined in the Hilbert space on the condition that P_N is the operator of orthogonal projection. Consequently, the theorem in the paragraph 4.3.1. remains also valid in this case.

We examine now a special set of basis functions $\{\psi_i\}$ in the Galerkin-Petrov method (in this case, the method is sometimes referred to as the *method of integration in sub-domains*).

Let us assume that $H = L_2(\Omega)$, where Ω is the domain of the m-dimensional Euclidean space, H_N is the span $\{\varphi_i\}$. The basis $\{\psi_i\}$ has here already a specific form. Ω is divided into N-sub-domains $\Omega_1, \ldots \Omega_N$, in such a manner that $\bigcup_{i=t}^{N} \Omega_i = \Omega_j \cap \Omega_i = 1$ at $i \ne j$. $\tilde{\psi}_k(x)$, $x \in \Omega$, denotes the characteristic function of the domain Ω_k: $\tilde{\psi}_k(x)$ is equal to 1 at $x \in \Omega_k$ and 0 at $x \notin \Omega_k$. We introduce the function $\psi_k(x) = (1/\sqrt{\mathrm{mes}(\Omega_k)})\tilde{\psi}_k(x)$, $k = 1, \ldots, N$, and assume that M_N is a subspace being the span of the system $\{\psi_k\}$. In this case, equation (51) is equivalent to the system:

$$\sum_{i=1}^{N} a_i \int_{\Omega_j} A\varphi_i dx = \int_{\Omega_j} f dx, \qquad j = 1, \ldots, N. \tag{52}$$

The convergence of the integration method in subdomains results from the theorem 8.

4.3.3. The projection method in the Banach space

Let E and F be Banach spaces (complex or real). We examine the equation

$$Au = f, \tag{53}$$

A is a linear (generally speaking, unbounded) operator with the domain

of definition $D(A)\subset E$ and the range of the values $R(A)\subset F$. *The projection method* of solving (53) may be described as follows. We specify two sequences $\{E_N\}$ and $\{F_N\}$: $E_N\subset D(A)\subset E$, $F_N\subset F$ (N = 1,2...), and also linear projection operators (projectors) P_N, projecting F on F_N: $P_N^2 = P_N$, $P_N f = F_N$ (N = 1,2,...). Equation (53) is substituted by the approximate equation

$$P_N A u_N = P_N f, \qquad u_N \in E_N. \tag{54}$$

Since the projector P_N has the form $P_N u = \sum_{k=1}^{N} l_k(u)\psi_i$, where $\psi_1,\ldots,\psi_N$ is the basis of the subspace F_N, and $l_k(u)$ are linear functionals bounded on F, then denoting by $\varphi_1,\varphi_2,\ldots,\varphi_N$ the basis in E_N, equation (54) is reduced to the system of linear algebraic equations

$$\sum_{k=1}^{N} l_j(A\varphi_k)a_k = l_j(f), \quad j = 1,2,\ldots,N; \tag{55}$$

When writing the equation in this form it is not necessary to indicate explicitly the subspace F_N and it is sufficient to show the functionals.

Determining u_N from equation (54) (or from equation (55)), it is treated as the approximate solution of equation (53).

It is said that the sequence of the subspaces $\{E_N\}$ has limiting density in E if for every $w\in E$ we have $P(w,E_N)\to 0$, $N\to\infty$, where $P(w,E_N) = \inf_{w_N\in E_N}\|w - w_N\|$.

The following theorem, proven by G.M. Vainikko, justifies the formulated algorithm.

Theorem 9. *Let the domain of definition $D(A)$ of operator A be dense in E,* and $R(A)$ in F, and let A transform $D(A)$ to $R(A)$ *one-to-one. Let the subspaces AE_N and F_N be closed in F, and the projectors P_N be bounded in relation to N:* $\|P_N\| \le C$ (N = 1,2,...).

Consequently, to ensure that at any $f\in F$ starting with some $N = n_0$, there is a unique solution u_N of equation (54) and to ensure that $\|Au_N - f\|\to 0$, $N\to\infty$, *it is necessary and sufficient to fullfill the following conditions:*

1) *The sequence of the subspaces AE_N has limiting density in F;*
2) *At $N \ge N_0$ the operator P_N transforms AE_N one-to-one to F_N;*
3) $\tau = \underline{\lim}_{N\to\infty}\tau_N > 0$ *where* $\tau_N = \inf_{w_N\in AE_N, \|w_N\|=1}\|P_N w\|$.

The rate of convergence when fulfilling the conditions 1)–3) is characterised by inequalities

$$P(f, AE_N) \le \|Au_N - f\| \le \left(1 + \frac{C}{\tau_N}\right)P(f, AE_N).$$

(*In the case in which the subspaces E_N, F_N have finite dimensions and their dimensions coincide, condition* 2) *is a consequence of condition* 3)).

4.3.4. The collocation method

Let in equation (53) operator A be a differential operator of the order s, $E = C^{(s)}(\Omega)$, $F = C(\Omega)$. We select a sequence of linearly independent functions $\varphi_1,\varphi_2,\ldots,\varphi_N$, satisfying all the boundary-value conditions of the problem. The span of $\varphi_1,\varphi_2,\ldots,\varphi_N$ is regarded as E_N, and let $u_N = \sum_{k=1}^{N} a_k\varphi_k$. In the domain Ω we now select N points $\xi_1,\xi_2,\ldots,\xi_N$ and set $l_j(u) = u(\xi_j)$. The system (55) has the following form in this case

$$\sum_{k=1}^{N} a_k (A\varphi_k)(\xi_j) = f(\xi_j), \qquad j = 1, 2, \ldots, N, \tag{56}$$

and the projection algorithm in the Banach space is referred to as the *collocation method*. To justify the collocation method, we may use the theorem 9. This method is used widely for the approximate solution of integral and differential equations.

4.4. Main concepts of the projection-grid methods

The natural and attractive feature is the concept of designing the algorithms of the approximate solution of mathematical physics problems which, on the one hand, would be of the variational or projection form and on the other hand, they would lead to a system of equations similar to those formed in difference methods (i.e. a small number of elements of the matrices of these systems would be of the non-zero type). These algorithms include the *projection-grid methods* (referred to also as the *finite element methods*).

In order to construct these algorithms, it is sufficient to use the functions with finite support (finite functions) in variational or projection methods as the basis functions $\{\varphi_i\}$, i.e. the functions differing from zero only in a small part of the region in which the required solution of the problem is determined. Thus, we assume that we examine the problem

$$-\frac{d^2u}{dx^2} + u = f(x), \qquad x \in (0,1), \qquad u(0) = u(1) = 0, \tag{57}$$

where $f \in L_2(0,1)$, which is reduced to the following variational problem:

$$J(u) = \inf_{v \in \overset{\circ}{W}{}_2^1(0,1)} J(v), \qquad \text{where} \qquad J(v) = \int_0^1 \left(\left(\frac{du}{dx} \right)^2 + u^2 - 2uf \right) dx.$$

We introduce, on [0,1], the grid $x_i = ih$, $i = 0, 1,\ldots, N$, $h = 1/N$, and functions of the type

$$\varphi_i(x) = \frac{1}{\sqrt{h}} \begin{cases} \dfrac{x - x_{i-1}}{h}; & x \in (x_{i-1}, x_i); \\ \dfrac{x_{i+1} - x}{h}; & x \in (x_{i;} x_{i+1}); \\ 0, & x \notin (x_{i-1}, x_{i+1}), \end{cases} \qquad i = 1, \ldots, N-1;$$

which are regarded as basis functions. We find the approximate solution in the form $u_N(x) = \sum_{i=1}^{N-1} a_i \varphi_i(x)$, where the coefficients are determined using a variational algorithm. In this case, this may be carried out on the basis of the conditions of minimisation of the functional $J(u_N)$, i.e., using the Ritz method in the space $\overset{\circ}{W}{}_2^1(0,1)$ (which is also the energy space in this case). Consequently, we obtain the following system of equations for $a_1,\ldots,a_{N-1}$

$$\sum_{j=1}^{N-1} A_{ij} a_j = f_i, \qquad i = 1, \ldots, N-1. \tag{58}$$

Taking into account the specific features of the selected basis functions, we can easily calculate the form of elements A_{ij}, i, j = 1,...,N–1:

$$A_{ij} = \begin{cases} \dfrac{2}{h^2} + \dfrac{4}{6}, & i = j, \\ -\dfrac{1}{h^2} + \dfrac{1}{6}, & j = i-1, i+1, \\ 0, & |j - i| > 1. \end{cases}$$

Thus, the application of the variational algorithm with the previously examined finite functions resulted in the situation in which the system of equation (58) is a system of some difference equations similar to a system formed in the difference method. The matrix of the system is in this case also tri-diagonal and, consequently, suitable for numerical solution (58). In addition to this, taking into account the fact that in constructing the approximate solution we have used the variational algorithm, the matrix $\hat{A}$ will be symmetric. In addition to this, it is positive definite:

$$\sum_{i,j=1}^{N-1} A_{ij} a_i a_j \geq \lambda_{\min} \sum_{i=1}^{N-1} a_i^2, \qquad \text{where} \qquad \lambda_{\min} = \frac{4\sin^2 \pi h/2}{h^2} > 0.$$

Thus, the properties of the positive definiteness and symmetricity of the operator of the problem are retained here in the application of the projection-grid method, and the *projection-grid algorithm has a number of useful properties of both the variational and difference method.*

We shall mention other attractive features of the projection-grid method.

For example, coefficients a_i in system (58) are often characterised by explicit interpretation of the meaning. In the examined problem, coefficient a_i is equal to the value of the approximate solution at the node x_i, multiplied by coefficient $\sqrt{h}$. In addition, it has been shown that the finite basis functions can often be easily 'adapted' to the geometry of the domain; this removes one of the difficulties formed in the difference method. In addition to this, we shall pay attention to the fact that if in the solution of the examined problem we have selected the projection algorithm and its basis functions in the appropriate manner, then the further process of constructing the solution of the problem takes place 'automatically' with the application of computers. These circumstances and a number of other circumstances determine the application of the projection-grid algorithms for solving greatly differing problems of mathematical physics – multidimensional problems in a domain with a complicated geometry of the boundaries, linear and non-linear problems, problems in hydrodynamics and aerodynamics, equations in electrodynamics and wave processes, and many others. In the majority of cases we retain the main concept of these methods based on the *application of the projection* (*including variational*) *methods using in them finite functions of different types* which are used widely in approximation theory.

5. METHODS OF INTEGRAL IDENTITIES

The method of integral identities (integro-interpolation methods, the balance method) belongs traditionally to different methods. The projection forms of the integral identities have been obtained and they make it possible to treat this method as one of the modifications of the projection method.

5.1. The main concepts of the method

The principle of the method of integral identities will be explained on the example of the equation

$$-\frac{d}{dx}p(x)\frac{du}{dx}+q(x)u=f(x), \quad a<x<b, \tag{59}$$

with some boundary-value conditions. The principle of the method of integral identities for solving this equation maybe described as follows: we introduce a grid $a = x_0 < x_{1/2} < x_1 < \ldots < x_{N-1/2} < x_N = b$ and equation (59) is integrated in the interval $(x_{i-1/2}, x_{i+1/2})$. Consequently, we obtain the integral identities:

$$-(W_{i+1/2}-W_{i-1/2})+\int_{x_{i-1/2}}^{x_{i+1/2}}(qu(x)-f(x))dx=0,$$

where $W_{i+1/2} = W(x_{i+1/2})$, $W(x) = p(x)\,(du/dx)(x)$. Subsequently, using the approximations of the derivatives and integrals included in these identities, and also the boundary-value conditions, we obtain the appropriate difference schemes.

However, this algorithm maybe reformulated as follows. We introduce a system $\{\psi_i(x)\}$ of step-like functions

$$\psi_i(x) = \begin{cases} 1 & \text{for} \quad x \in (x_{i+1/2}, x_{i-1/2}), \\ 0 & \text{for} \quad x \in (x_{i+1/2}, x_{i+1/2}). \end{cases}$$

Consequently, the resultant identities are nothing else but the result of projection in $L_2(a,b)$ of equation (59) on the system $\{\psi_i\}$, i.e. the identities maybe represented in the form of the system

$$-\left(\frac{d}{dx} p \frac{du}{dx}, \psi_i\right) + (qu, \psi_i) = (f, \psi_i).$$

This stage of construction of the approximate solution coincides with the stage in the projection-grid method, and the identities are the result of projection of the examined equation on some basis system.

Thus, *if some system of the basis functions* $\{\psi_i(x)\}$ *is general, then the integral identities may be determined by projecting the equations, describing the problem, on the given system*.

In the next stage of construction of the numerical solution of the problem, the integral identities maybe approximated by two methods: a) *either we calculate approximately the integrals using* quadrature rules and *so on*; or b) *find the approximate solution* u_h *in the form of an expansion using, generally speaking, the other basis functions* $\{\varphi_i(x)\}$. The second method makes it possible in many cases to interpret the method of integral identities as one of the modifications of the projection algorithm and use the theory of projection methods for justification of the process of solving the problem.

5.2. The method of Marchuk's integral identity

The numerical solution of ordinary differential equations is often carried out using the method of integral identities developed by G.I. Marchuk. The principle of the method maybe described as follows. We examine a boundary-value problem for a one-dimensional diffusion equation of the type

$$-\frac{d}{dx} p(x) \frac{du}{dx} + q(x)u = f(x), \quad x \in (a,b), \quad u(a) = u(b) = 0. \tag{60}$$

It is assumed that $p(x)$, $q(x) \in L_\infty(a,b)$, $f(x) \in L_2(a,b)$, $p(x) > 0$, $q(x) \geq 0$.

The method of integral identities used for solving the problem is based on the fact that, using (60), we obtain the identities

$$(u(x_k)-u(x_{k+1}))\left(\int_{x_k}^{x_{k+1}}\frac{dx}{p(x)}\right)^{-1}+(u(x_k)-u(x_{k-1}))\left(\int_{x_{k-1}}^{x_k}\frac{dx}{p(x)}\right)^{-1}+$$

$$+\int_{x_{k-1/2}}^{x_{k+1/2}}(qu-f)dx=-\left(\int_{x_k}^{x_{k-1}}\frac{dx}{p(x)}\right)^{-1}\int_{x_k}^{x_{k+1}}\frac{dx}{p(x)}\int_{x_{k+1/2}}^{x}(qu-f)d\xi+$$

$$+\left(\int_{x_{k-1}}^{x_k}\frac{dx}{p(x)}\right)^{-1}\int_{x_{k-1}}^{x_k}\frac{dx}{p(x)}\int_{x_{k-1/2}}^{x}(qu-f)d\xi,\quad k=1,\ldots,N-1, \tag{61}$$

where $a = x_0 < x_{1/2} < x_1 < x_{3/2} < \ldots < x_{N-3/2} < x_{N-1} < x_{N-1/2} < x_N = b$ is some set of the points. Using now some approximation of the integrals included in (61) we obtain the corresponding difference scheme.

The identities (61) were determined by G.I. Marchuk and used widely for constructing difference schemes. However, later it was found that (61) are equivalent to each of the following relationships:

$$(u(x_k)-u(x_{k+1}))\left(\int_{x_k}^{x_{k+1}}\frac{dx}{p(x)}\right)^{-1}+(u(x_k)-u(x_{k-1}))\left(\int_{x_{k-1}}^{x_k}\frac{dx}{p(x)}\right)^{-1}+$$
$$+(qu,Q_k)=(f,Q_k),\quad k=1,\ldots,N-1, \tag{62}$$

$$\left(p(x)\frac{du}{dx},\frac{dQ_k}{dx}\right)+(qu,Q_k)=(f,Q_k),\quad k=1,\ldots,N-1, \tag{63}$$

$$\left(p(x)\frac{du_I}{dx},\frac{dQ_k}{dx}\right)+(qu,Q_k)=(f,Q_k),\quad k=1,\ldots,N-1, \tag{64}$$

where $(\varphi,\psi)=\int_a^b\varphi\psi dx,\quad \|\varphi\|=(\varphi,\varphi)^{1/2}, u_I(x)$ is some interpolant of the function $u(x)$ which is such that $u_I(x_i) = u(x_0)$, $i = 0,1,\ldots, N$, and the derivative du_I/dx has the meaning. Function $Q_k(x)$ ($k = 1,2,\ldots,N-1$) has the form

$$Q_k(x)=\begin{cases}1-\int_x^{x_k}\frac{d\xi}{p(\xi)}\left(\int_{x_{k-1}}^{x_k}\frac{d\xi}{p(\xi)}\right)^{-1}, & x\in(x_{k-1},x_k),\\ 1-\int_{x_k}^{x}\frac{d\xi}{p(\xi)}\left(\int_{x_k}^{x_{k+1}}\frac{d\xi}{p(\xi)}\right)^{-1}, & x\in(x_k,x_{k+1}),\\ 0, & x\notin(x_{k-1},x_{k+1})\end{cases} \tag{65}$$

It should be mentioned that the relationship (63) is nothing else but the well-known equality used in the Bubnov–Galerkin method for obtaining the approximate solution using $\{Q_k(x)\}$ as basis functions. If we use the relationships (62), we obtain that in this case we can use discontinuous base functions $\{\varphi_i(x)\}$. Actually, let us assume that the functions $\varphi_i(x)$ are piecewise continuous with possible jumps of the first kind at the points which do not coincide with the nodes of the grid x_k, $k = 1,\ldots, N-1$. Therefore, each of these functions has a finite value in x_k and this means that their linear combination $u_h = \sum_{k=0}^{N} a_k \varphi_k(x)$ with arbitrary constants a_k will have a finite value at nodes x_j, i.e. $u_h(x_j) = \sum_{k=0}^{N} a_k \varphi_k(x_j) < \infty, j = 1,\ldots, N-1.$ Thus, the functions $\{\varphi_k(x)\}$ can be used to obtain the approximate solution using the relationships (62), although $\{\varphi_j(x)\}$ is a system of discontinuous functions.

Using the identities in the form of equation (64) shows that it is quite simple to examine the processes of convergence and, in particular, obtain estimates of the rate of convergence in uniform metrics.

Thus, the method of integral identities may be used as one of the projection algorithms, and relationships (62)–(64) maybe used together with the identity (61), with a sufficiently wide range of the basis functions for approximation $u(x)$.

5.3. Generalized formulation of the method of integral identities

We shall explain the method of integral identities in general formulation, obtaining identities by projecting on the system of certain basis functions.

5.3.1. Algorithm of constructing integral identities

In the Hilbert space, we examine the equation

$$Au + Bu = f, \qquad f \in H, \tag{66}$$

where A is a linear (in the general case unbounded) operator, acting in H with the domain of definition $D(A) \subset H$, dense in H and with the range of the values $R(A) \subseteq H$, B – linear, symmetric, bounded in H operator with $D(B) = H$, $R(B) \subseteq H$. It is assumed that the operators $A + B$ and B are H-defined: $((A+B)v, v) \geq \gamma_0^2 \|v\|^2$, $v \in D(A)$, $(Bv, v) \geq \gamma_1^2 \|v\|^2, v \in H, \gamma_i > 0,$ and also that equation (66) has a unique solution $u \in D(A)$ at the given function $f \in H$.

We introduce in H a system of linearly independent functions $\{\psi_i\}$ dense in H. The span of the system $\{\psi_i\}$ is denoted by $H_\psi^{(N)}$. We project the equation (66) orthogonally on $H_\psi^{(N)}$. Consequently, we obtain a system of integral indentities

$$A_i^{(N)} u + (Bu, \psi_i) = (f, \psi_i), \qquad i = 1,\ldots, N. \tag{67}$$

where $A_i^{(N)} u = (Au, \psi_i).$

It may appear that the operator $A^{(N)}$ defined as a system of the operators $A_i^{(N)}: A^{(N)} = \{A_i^{(N)}, i = 1,\ldots,N\}$ has the domain of definition that is wider in comparison with $D(A)$ and, consequently, $A^{(N)}$ can be extended, i.e. construct an operator $A^{(N)}$ with the domain of definition $D(A^{(N)})$, so that $D(A^{(N)}) \subset D(\overline{A}^{(N)})$ and $\overline{A}^{(N)}u = A^{(N)}u$ at $u \in D(A^{(N)})$. It should be mentioned that if $v \in D(A)$, then according to the definition of extension of the operator we have $\overline{A}^{(N)}v = A^{(N)}v = \{A_i^{(N)}v, i = 1,\ldots,N\} = \{(Av,\psi_i), i = 1,\ldots,N\}$. However, if $v \in D(\overline{A}^{(N)})$, but $v \notin D(A)$, then in the general case $\overline{A}^{(N)}$ cannot be regarded as a set of the operators $\{A_i^{(N)}v\} = \{(Av,\psi_i)\}$, and this can be carried out only through the system $\{\overline{A}_i^{(N)}\}$, for which the representations (Au,ψ_i) are no longer valid.

According to the assumption, $u \in D(A)$ and $\overline{A}_i^{(N)}u = A_i^{(N)}u$, and therefore the system (67) is equivalent to the system:

$$\overline{A}_i^{(N)}u + (Bu,\psi_i), \qquad i = 1,\ldots,N, \tag{68}$$

which is the main system of the integral identity.

To construct an approximate solution we can now use this system in particular. Consequently, it is possible to construct the approximate solution using the basis systems not belonging to $D(A)$, i.e. we can find the approximation $u_N = \sum_{i=1}^{N} a_i\varphi_i \approx u$, where the function φ_i belongs to only $D(\overline{A}^{(N)})$. The last method of constructing the approximate solution using φ_i is referred to as a *projection approach* in the method of integral indentities. However, we can also use the classic algorithm of approximation of the expression included in (68) with the help of different ratios and quadrature rulese. This approach is refered to as the difference approach.

5.3.2. The difference method of approximating the integral identities

We examine the difference approach to constructing the approximate solutions of the problem on the basis of integral identities (68). For this purpose, we introduce a grid in the domain of definition of the functions and determine grid functions: u_h – the projection of the exact solution on the grid, u^h – the vector of the approximate solution of the problem. It is assumed that the dimensions u_h and u^h coincide and are equal to N. Let it be that for the terms included in (68) we determine the approximations:

$$\overline{A}_i^{(N)}u = \sum_{j=1}^{N} \overline{A}_{ij}^{(N)}u_{h,j} + \varepsilon_i^{(1)}, \qquad (Bu,\psi_i) = \sum_{j=1}^{N} B_{ij}u_{h,j} + \varepsilon_k^{(2)},$$

where $\varepsilon_i^{(k)}$ are the approximation errors. We determine the matrices and vectors of the type:

$$\widehat{A}_h = (A_{ij}^{(N)}), \quad \widehat{B}_h = (B_{ij}),$$

$$f = ((f,\psi_1),\ldots,(f,\psi_n))^T, \quad \varepsilon^{(k)} = (\varepsilon_1^{(k)},\ldots,\varepsilon_N^{(k)})^T.$$

Consequently, the system of equations for determining the approximate grid solution u^h has the form

$$\widehat{A}_h u^h + \widehat{B}_h u^h = f. \tag{69}$$

We assume that we introduce the grid space $H_{1,h}$ – the space of the solutions with the norm $||\cdot||_{1,h}$ and the grid space H_h with the norm $||\cdot||_h$. In this case, the scheme (69) can be examined using the general theory of difference schemes which gives the following claims. If the system (69) has a unique solution at any N, and the appriori estimate holds, then when fulfilling the approximation conditions $||\varepsilon^{(1)}||_h \le \varepsilon_1(N) \to 0$, $||\varepsilon^{(2)}||_h \le \varepsilon_2(N) \to 0$, $N \to \infty$, the approximate solutions u^h converge to u_h at $N \to \infty$; the error estimate

$$||u^h - u_h||_{1,h} \le O\,(\varepsilon_1 + \varepsilon_2).$$

holds in this case.

It should be mentioned that in comparison with the conventional difference methods (when there is no preliminary projection of the initial equation on $H_\psi^{(N)}$) we carry out approximation of the integral identities (68) and, consequently, to justify the algorithm we may require a smaller number of restrictions on the smoothness of the solution of equation (66) in comparison with the difference methods. However, in the algorithm examined here one of the main difficulties of the difference methods remains: it is necessary to construct $A_{ij}^{(N)}, B_{ij}^{(N)}$, in order to obtain 'good' approximation without losing stability in this case.

It should also be mentioned that in approximation of the identities it may appear that $\varepsilon_i^{(1)} = 0$. In this case, there is the problem of obtaining the estimate $\varepsilon_i^{(2)}$ which is solved in most cases by a simple procedure and with weaker restrictions on the exact solution of the initial equation (66).

5.3.3. The projection method of approximating the integral identities

We introduce another system of the linear independent function $\{\varphi_i\}$ such that $\varphi_i \in D(\overline{A}_i^{(N)})$, $i = 1,\ldots,N$. The span $\{\varphi_i\}$ is denoted by $H_\varphi^{(N)}$. It is assumed that the sequence of the subspaces $\{H_\varphi^{(N)}\}$ has limiting density in H. Evidently, $H_\varphi^{(N)} \subset D(\overline{A}^{(N)})$ and the system $\{D(\overline{A}^{(N)})\}$ is dense in H. The projection operator $H_\varphi^{(N)}$ is denoted by $P_\varphi^{(N)}$. It is assumed that $D(P_\varphi^{(N)}) = H$. It should be mentioned that there are no restrictions in selection of $P_\varphi^{(N)}$ and it is only required that $P_\varphi^{(N)} v \in H_\varphi^{(N)}$ $\forall v \in H$. It is also assumed that the dimensions $H_\varphi^{(N)}$ and $H_\psi^{(N)}$ coincide.

We examine a general scheme of the projection approach to con-

structing the approximate solution. We shall find it in $H_{\varphi}^{(N)}$ in the form $u_N = \sum_{i=1}^{N} a_i \varphi_i$, where a_i is determined from the system of equations

$$\sum_{j=1}^{N} (\overline{A}_i^{(N)} \varphi_j + (B\varphi_j, \psi_i)) a_j = (f, \psi_i), \qquad i = 1, \ldots, N. \tag{70}$$

We formulate some conditions which, if fulfilled, result in the unique solvability of (70) and in the convergence of u_N to u at $N \to \infty$. It is assumed that the bases satisfy the condition of *uniform linear independence* i.e.

$$d_1 \|c\|_2 \le \left\| \sum_{i=1}^{N} c_i \varphi_i \right\| \le d_2 \|c\|_2, \qquad d_3 \|c\|_2 \le \left\| \sum_{i=1}^{N} c_i \psi_i \right\| \le d_4 \|c\|_2,$$

where $d_1, \ldots, d_4 > 0$ are constants independent of $c = (c_1, \ldots c_N)^N$, and

$$\|c\|_2 = \left(\sum_{i=1}^{N} c_i^2 \right)^{1/2}.$$

We shall also use the notation

$$(u, v)_B = (Bu, v), \qquad \|u\|_B = (u, u)_B^{1/2}.$$

We introduce the following conditions.

Condition 1. Matrix $\widehat{A} = (A_{ij})$ with the elements $A_{ij} = \overline{A}_i^{(N)} \varphi_j + (\varphi_i, \varphi_j)_B$ is positive definite and

$$\sum_{i=1}^{N} \sum_{i=1}^{N} A_{ij} c_i c_j \ge \sum_{i=1}^{N} \sum_{i=1}^{N} c_i c_j (\varphi_i, \varphi_j)_B.$$

Condition 2. The bases $\{\varphi_i\}, \{\psi_i\}$ satisfy the restriction

$$\sup_{c \ne 0} \frac{\left| \sum_{i,j=1}^{N} c_i c_j (\varphi_i, \varphi_j - \psi_j)_B \right|}{|c|_{A,N}^2} \le \theta < 1,$$

where the constant θ is independent of $c = (c_1, \ldots c_N)^T$ and N, and the norm $|c|_{A,N}$ has the form

$$|c|_{A,N} = \left[\sum_{i,j=1}^{N} c_i c_j (\overline{A}_i^{(N)} \varphi_j + (\varphi_i, \varphi_j)_B) \right]^{1/2}.$$

Condition 3. In $H_{\varphi}^{(N)}$ we can find a function such that

$$u_{\varphi} = \sum_{i=1}^{N} b_i(u) \varphi_i$$

with some constants b_i, and for any non-zero vector $c = (c_1, \ldots c_N)^T$,

$$\frac{\left|\sum_{i=1}^{N} c_i (A_i^{(N)}(u_\varphi - u) + (u_\varphi - u, \psi_i)_B)\right|}{|c|_{A,N}} \le \varepsilon_1(N) \to 0,$$

$$\|u - u_\varphi\|_B \le \varepsilon_1(N) \to 0, \quad N \to \infty.$$

Theorem 10. *If the conditions* 1–3 *are fulfilled, then:*

1) *System* (70) *has a unique solution a;*

2) *The apriori estimate*

$$|a|_{A,N} \le \frac{c\|f\|}{(1-\theta)},$$

is valid;

3) *Approximate solutions*

$$u_N = \sum_{i=1}^{N} a_i \varphi_i$$

converge to the exact solution as $N \to \infty$ *and the error estimates*

$$\left[\sum_{i,j=1}^{N} \overline{A}_i^{(N)} \varphi_j (b_i - a_i)(b_j - a_j) + \|u_\varphi - u_N\|_B^2\right]^{1/2} \le O\left(\frac{\varepsilon_1}{1-\theta}\right),$$

$$\|u - u_N\| \le O\left(\frac{\varepsilon_1 + \varepsilon_2}{1-\theta}\right).$$

are valid.

Comment 1. It may be that $A_i^{(N)} u_\varphi = \overline{A}_i^{(N)} u$, $i = 1, \ldots, N$. In this case, the approximation problem is reduced to the simpler problem of deriving the estimates

$$\left|\sum_{i=1}^{N} c_i (u_\varphi - u, \psi_i)\right| \le \varepsilon_1(N) |c|_{A,N}, \qquad \|u - u_\varphi\| B \le \varepsilon_2(N).$$

Comment 2. If $\varphi_i \in D(A)$, the examined algorithm coincides with the Galerkin-Petrov method. In this case, the error estimates have the form:

$$((A(u_\varphi - u_N), u_\varphi - u_N) + \|u_\varphi - u_N\|_B^2)^{1/2} \le O\left(\frac{\varepsilon_1}{1-\theta}\right),$$

$$\|u - u_\varphi\| \le O\left(\frac{\varepsilon_1 + \varepsilon_2}{1-\theta}\right).$$

It should be mentioned that here the term $(A(u_\varphi - u_N), u_\varphi - u_N)$ includes the function u_φ (and not the exact solution u).

5.4. Applications of the methods of integral identities in mathematical physics problems

5.4.1. The method of integral identities for the diffusion equation

We examine the application of identities (62)–(64) for the approximate solution of the problem (59) using two bases in this case: one – of step-like functions, being the example of basis functions discontinuous on (a,b); the second – of functions $\{Q_i(x)\}$, which are continuous on (a,b).

Let it be that $h_i = x_{i+1/2} - x_{i-1/2}$, $h = \max_i h_i$, and $\varphi_i(x)$ denotes the characteristic function of the interval $(x_{i-1/2}, x_{i+1/2})$ and $\varphi_0(x)$ – of the interval $(x_0, x_{1/2})$, $\varphi_N(x)$ – of the interval $(x_{N-1/2}, x_0)$. We accept $\{\varphi_i\}$ as the basis functions and find the approximate solution in the form $u^h(x) = \sum_{i=1}^{N} a_i\varphi_i(x)$, where it is assumed that $a_0 = a_N = 0$. Then $u^h(x) = \sum_{i=1}^{N-1} a_i\varphi_i(x)$, where $\{a_i\}$ is determined from the relationship (obtained on the basis of (62))

$$(u^h(x_h) - u^h(x_{k+1}))\left(\int_{x_k}^{x_{k+1}} \frac{dx}{p(x)}\right)^{-1} + (u^h(x_k) - u^h(x_{k-1}))\left(\int_{x_{k-1}}^{x_k} \frac{dx}{p(x)}\right)^{-1} + \\ +(qu^h, Q_k) = (f, Q_k), \quad k = 1,\ldots,N-1. \tag{71}$$

The system (71) in the matrix form reads

$$\hat{A}a = f,$$

where

$$a = (a_1,\ldots,a_{N-1})^T, \quad f = (f_1,\ldots,f_{N-1})^T, \quad \hat{A} = (A_{ij}), \quad f_i = (f, Q_i),$$

$$A_{ij} = (\varphi_j(x_i) - \varphi_j(x_{i+1}))\left(\int_{x_i}^{x_{i+1}} \frac{dx}{p(x)}\right)^{-1} + (\varphi_j(x_i) - \varphi_j(x_{i-1}))\left(\int_{x_{i-1}}^{x_i} \frac{dx}{p(x)}\right)^{-1} + \\ + \int_{x_{i-1}}^{x_{i+1}} \varphi_j(x)Q_i(x)q(x)dx, \quad i, j = 1,\ldots,N-1.$$

Computing A_{ij} and f_i and solving the system, we determine the coefficients $a_1,\ldots,a_{N-1}$, which can be used for constructing the piecewise constant representation of the approximate solution

$$u^h(x) = \sum_{i=1}^{N-1} a_i\varphi_i(x), \qquad a_0 = a_N = 0.$$

Assuming that $p(x)$, $q(x)$ are bounded functions, and $f(x) \in L_2(a,b)$, it maybe shown that at sufficiently small h the system (71) has a unique solution, and

$$\max_i \left|u(x_i) - u^h(x_i)\right| + (q(u-u^h), u-u^h)^{1/2} \le ch,$$

where c = const > 0.

We can construct the approximate solution also in the form

$$u^h(x) = \sum_{i=1}^{N-1} a_i Q_i(x),$$

where the unknown $\{a_i\}$ are determined from (71). It may be easily seen that, in this case, the algorithm coincides with the Bubnov–Galerkin method and, consequently, the corresponding results regarding the convergence of the methods are valid. However, using the specific features of the method of integral identities, we can simply determine the estimate of the error of the type

$$\max_i \left| u(x_i) - u^h(x_i) \right| + \\ + \left[\left(p(x) \frac{d}{dx}(u_I - u^h), \frac{d}{dx}(u_I - u^h) \right) + \left(q(u_I - u^h), u - u^h \right) \right]^{1/2} \le O(h^2).$$

which, generally speaking, does not follow from the theory of the Bubnov–Galerkin method.

5.4.2. The solution of degenerating equations

We examine a problem for an equation with degeneration

$$-\frac{d}{dx} x^\alpha p(x) \frac{du}{dx} + q(x)u = f(x), \qquad x \in (0,1), \qquad u(0) = u(1) = 0, \tag{72}$$

where $a > 0$, $p(x) \in L_\infty(0,1)$, $p(x) \in L_\infty(0,1)$, $f(x) \in L_2(0,1)$, $0 < p_0 \le p(x) \le p_1$, p_0, p_1 are constants, $q(x) \ge 0$.

When constructing approximate solutions we restrict ourselves to the simplest basis functions $\{\varphi_i\}$ (characteristic) and the case $0 < a < 1$ (weak degeneration).

We introduce the function

$$\begin{cases} 1 - \displaystyle\int_x^{x_k} \frac{d\xi}{\xi^\alpha p(\xi)} \left(\int_{x_{k-1}}^{x_k} \frac{d\xi}{\xi^\alpha p(\xi)} \right)^{-1}, & x \in (x_{k-1}, x_k), \\ 1 - \displaystyle\int_{x_k}^{x} \frac{d\xi}{\xi^\alpha p(\xi)} \left(\int_{x_k}^{x_{k-1}} \frac{d\xi}{\xi^\alpha p(\xi)} \right)^{-1}, & x \in (x_k, x_{k+1}), \\ 0, \quad x \notin (x_{k-1}, x_{k-1}), \end{cases}$$

$$k = 1, \ldots, N-1,$$

and using the well-known transformation we obtain the identities

$$(u(x_k)-u(x_{k+1}))\left(\int_{x_k}^{x_{k+1}}\frac{d\xi}{\xi^\alpha p(\xi)}\right)^{-1}+$$

$$+(u(x_k)-u(x_{k-1}))\left(\int_{x_{k-1}}^{x_k}\frac{d\xi}{\xi^\alpha p(\xi)}\right)^{-1}+(qu,Q_k)=(f,Q_k). \tag{73}$$

Let it be that as in paragraph 5.4.1, we introduce the characteristic function $\varphi_i(x)$, $i = 0, 1,\ldots,N$. We find the approximate solution in the form

$$u^h(x)=\sum_{i=0}^{N}a_i\varphi_i(x),$$

where it is accepted that $a_0 = a_N = 0$, and the remaining constants are determined from the system of equations

$$(u^h(x_k)-u^h(x_{k+1}))\left(\int_{x_k}^{x_{k-1}}\frac{d\xi}{\xi^\alpha p(\xi)}\right)^{-1}+(u^h(x_k)-u^h(x_{k-1}))\left(\int_{x_{k-1}}^{x_k}\frac{d\xi}{\xi^\alpha p(\xi)}\right)^{-1}+$$

$$+(qu^h,Q_k)=(f,Q_k),\quad k=1,\ldots,N-1. \tag{74}$$

This system can be presented in the matrix form

$$\hat{A}a=f,$$

where

$$\hat{A}=(A_{ij}),\qquad a=(a_1,\ldots,a_{N-1})^T,\quad f=(f_1,\ldots,f_{N-1})^T,\quad f_i=(f,Q_i),$$

$$A_{ij}=\begin{cases}-\left(\int_{x_i}^{x_{i+1}}\frac{d\xi}{\xi^\alpha P(\xi)}\right)^{-1}+(qQ_i,\varphi_{i+1}), & j=i+1,\\ -\left(\int_{x_{i-1}}^{x_i}\frac{d\xi}{\xi^\alpha P(\xi)}\right)^{-1}+(qQ_i,\varphi_{i-1}), & j=i-1,\\ \left(\int_{x_i}^{x_{i+1}}\frac{d\xi}{\xi^\alpha P(\xi)}\right)^{-1}+-\left(\int_{x_{i+1}}^{x_i}\frac{d\xi}{\xi^\alpha P(\xi)}\right)^{-1}+(qQ_i,\varphi_i), & j=i+1,\\ 0, & |j-i|>1.\end{cases}$$

The system (73) is uniquely solvable at sufficiently small h and the apriori estimate for u^h is valid:

$$\left(x^\alpha p(x)\frac{du_I^h}{dx},\frac{du_I^h}{dx}\right)+(qu^h,u^h)\le\frac{c\|f\|^2}{1-\varepsilon^2(h)},$$

where

$$\varepsilon^2(h) = O(h)\max_i \frac{x_{i+1/2}^{1-\alpha} - x_{i-1/2}^{1-\alpha}}{1-\alpha}, \quad x_{-1/2} = x_0 = 0, \quad x_{N+1/2} = x_N = 1,$$

and $u_I^h(x) \equiv \sum_{i=1}^{N-1} a_i Q_i(x)$ is the interpolant of the approximate solution: $u_I^h(x_k) = u^h(x_k), k = 1,\ldots,N-1.$ The error estimate has the form in this case

$$\max_i \left| u(x_i) - u^h(x_i) \right| + (q(u-u^h), u-u^h)^{1/2} \le c\frac{\varepsilon(h)\|f\|}{(1-\alpha)^{1/2}(1-\varepsilon(h))}.$$

Comment. The same procedure can be used to examine the case of the problem with strong degeneration where $\alpha \ge 1$ and the boundary-value condition has the form $u(1) = 0$. Here, we can introduce the grid $0 = x_{1/2} < x_1 < x_{2/3} < \ldots < x_{N-1/2} < x_N = 1$. Let

$$Q_1(x) = \begin{cases} 1, & x \in (x_{1/2}, x_1) \\ 1 - \displaystyle\int_{x_1}^{x} \frac{d\xi}{\xi p(\xi)} \left(\int_{x_1}^{x_2} \frac{d\xi}{\xi p(\xi)} \right)^{-1}, & x \in (x_1, x_2), \\ 0, & x \notin (x_{1/2}, x_2), \end{cases}$$

and remaining Q_i, $i = 2,\ldots, N-1$ are the same as previously. The further course of construction of the approximate solution $u^h(x) = \sum_{i=1}^{N-1} a_i \varphi_i(x)$, where $\varphi_i(x)$ is the characteristic function of the interval $(x_{i-1/2}, x_{i+1/2})$ at $i = 1,\ldots, N-1$, remains the same as previously.

5.4.3. The method of integral identities for eigenvalue problems

We examine a problem of finding numbers λ and non-zero functions $u(x)$ such that

$$-\frac{d}{dx} p(x) \frac{du}{dx} + q(x)u = \lambda u, \quad a < x < b, \quad u(a) = u(b) = 0, \quad \|u\|_{L_2(a,b)}^2 = 1, \tag{75}$$

where $p(x)$, $q(x)$ are the positive bounded functions. For approximate solution of the problem (75) we use the method of integral identities. For this purpose we introduce the grid $a = x_0 < x_{1/2} < x_1 < \ldots < x_N = b$ and relate, to each node x_i function $Q_i(x)$ of the type (65) and the function $\varphi_i(x)$ (generally speaking, differing from $Q_i(x)$) for which the conditions $\varphi_i(x_j) < \infty$, $j = 1,\ldots, N-1$ are satisfied. It is assumed that $\{\varphi_i(x)\}$ is the basis. We project the equation (75) in $L_2(a,b)$ on the function $\{Q_j(x)\}$. Consequently, we obtain identities

$$(u(x_i) - u(x_{i+1}))\left(\int_{x_i}^{x_{i+1}} \frac{dx}{p(x)} \right)^{-1} + (u(x_i) - u(x_{i-1}))\left(\int_{x_{i-1}}^{x_i} \frac{dx}{p(x)} \right)^{-1} + \\ + (qu, Q_i) = \lambda(u, Q_i), \quad i = 1,\ldots, N-1. \tag{76}$$

The approximation to the eigenfunction $u(x)$, corresponding to λ, will be

found in the form $u_h(x) = \sum_{i=1}^{N-1} a_i \varphi_i(x)$. Since the boundary-value conditions $u(a) = u(b) = 0$ are the main conditions, it is assumed that the basis functions $\{\varphi_i(x)\}$ in respect of which $u_h(x)$ is expanded, satisfy them. The coefficients $\{a_i\}$ are determined from the system

$$(u_h(x_i) - u_h(x_{i+1}))\left(\int_{x_i}^{x_{i+1}} \frac{dx}{p(x)}\right)^{-1} + (u_h(x_i) - u_h(x_{i-1}))\left(\int_{x_{i-1}}^{x_i} \frac{dx}{p(x)}\right)^{-1} + \\ +(qu_h, Q_i) = \lambda^h (u_h, Q_i), \qquad i = 1, \ldots, N-1, \tag{77}$$

or in the matrix form

$$\hat{L}a = \lambda^h \widehat{M} a, \tag{78}$$

where

$$a = (a_1, \ldots, a_{N-1})^T, \qquad \widehat{M} = (M_{ij}), \quad \hat{L} = (L_{ij}), \quad M_{ij} = (\varphi_i, Q_j),$$

$$L_{ji} = (\varphi_i(x_j) - \varphi_i(x_{j+1}))\left(\int_{x_j}^{x_{j+1}} \frac{dx}{p(x)}\right)^{-1} + (\varphi_i(x_j) - \varphi_i(x_{j-1}))\left(\int_{x_{j-1}}^{x_j} \frac{dx}{p(x)}\right)^{-1} + \\ +(q\varphi_i, Q_j), \qquad i, j = 1, \ldots, N-1.$$

Finding the eigenvalues $\lambda_i^h, i = 1, \ldots, N-1,$ the matrix problem (78) and the corresponding eigenvectors a_i using a suitable numerical method, we accept λ_i^h as approximations of some exact eigenvalues λ_i of the problem (75), and the functions $u_i^h(x) = \sum_{j=1}^{N-1} a_{ij} \varphi_j(x)$, constructed on the basis of the vectors a_i as approximations of the eigenfunctions $u_i(x)$. Let it be that in further consideration to simplify examination of the convergence of λ_i^h and λ_i it is assumed that $\varphi_i(x) = Q_i(x)$, $i = 1, \ldots, N-1$. In this case, the matrices $\hat{L}$, $\widehat{M}$ are symmetric and positive definite. Consequently, the approximate eigenvalues λ_i^h are real and positive.

Let it be that $\{u_1^h\}$ is the sequence of approximated eigenfunctions corresponding to $\{\lambda_1^h\}$ and governed by the normalization condition $\|u_1^h\| = 1$. Consequently, at $h \to 0$ the following estimates are valid:

$$\lambda_1 \le \lambda_1^h \le \lambda_1 + ch^2 (q_1 + \lambda_1)^2, \qquad \|u_1 - u_1^h\| c_{(a,b)} \le \frac{ch^2}{\lambda_2 - \lambda_1},$$

where c = const > 0, $q_1 \equiv ||q||_{L\infty(a,b)}$.

The method of integral identities is also used widely in solving the equations of higher orders, systems of equations, transfer equations, elliptical equations, equations of gas dynamics and a number of other equations of mathematical physics.

BIBLIOGRAPHIC COMMENTARY

Many concepts and approaches of computational mathematics may be found in textbooks [79–81]. The main numerical methods of solving a wide range of the problems of mathematical physics, formed in examination of physical and technical problems, have been described in [29] which also gives practical recommendations for the application of each method. Gradual explanation of the numerical methods has been published in [3] in which these methods are usually theoretically substantiated; special attention is given to the problem of optimization of algorithms. The methods of solving ordinary differential equations, equations in partial derivatives and integral equations are examined in detail in [40]. The book [60] contains explanation of the numerical methods of solving the problems of mathematical physics which in the process of solution are usually reduced to simpler problems permitting the application of algorithms in a computer; many advanced approaches to designing numerical algorithms are examined; a special chapter of the book is devoted to the review of main ideas and concepts in computational mathematics, a large number of references are given with systematization in respect of the main sections of computational mathematics. The monograph in [65] is concerned with different methods of increasing the accuracy of solution of difference and variational-difference schemes; theoretical substantiation of the proposed methods is illustrated by the numerical examples.

Many currently used projection and variational methods are described in [35,64,71,72]. The book [35] deals with the examination of the approximate methods of operator equations; special attention is given to systematic construction of the theory of projection methods in Hilbert and Banach spaces; the approximate methods of solving non-linear operator equations are also examined.

Chapter 6

SPLITTING METHODS

Keywords: evolution equations, Cauchy problem, difference schemes, approximation, stability, convergence, the sweep method, splitting methods, diffusion equation, heat conduction equation, Navier–Stokes equation, shallow water equations, dynamics of oceanic flows.

MAIN CONCEPTS AND DEFINITIONS

The Cauchy problem – the problem of the type $\frac{d\varphi}{dt} + A\varphi = f$, $t \in (0,T)$, $\varphi|_{t=0} = g$.

$\frac{d\varphi_h}{dt} + A_h\varphi_h = f_h, \varphi_h|_{t=0} = g_h$ – Approximation of the Cauchy problem in respect of the spatial variables.

$\tilde{L}_{h\tau}\tilde{\varphi}_{h\tau} = \tilde{f}_{h\tau}$ – Approximation of the Cauchy problem in respect of all variables (including the time variable $t \in [0,T]$).

$A = \sum_{\alpha=1}^{n} A_\alpha$ – 'splitting' of operator A into a family of operators $\{A_\alpha\}$ with a simpler structure than A.

Φ, F, ...,G – Hilbert spaces in which the initial problem is formulated.

Φ_h, F_h, ...,G_h – finite-dimension Hilbert spaces with the solutions of difference schemes at $t \in [0,T]$ (or at points $t_j = j\tau$, $j = 0,1,2,...$), τ is the step of the grid in respect of t.

Ω_h – the grid approximation of the domain $\Omega \subset \mathbb{R}^n$.

$\partial\Omega_h$ – the grid approximation of the boundary $\partial\Omega$.

$\|\cdot\|_x$, $(\cdot,\cdot)_x$ – the norm and scalar product in Hilbert space X.

1. INTRODUCTION

The splitting methods (the method of fractional steps) are based on the concept of the approximate reduction of the initial evolution problems with

complex operators to solving a sequence of problems with operators with a simpler structure which may be efficiently solved, for example, by the finite difference methods, the finite element methods and projection methods.

The concept of the solution of complex problems of mathematical physics by the splitting methods was proposed in the fifties and sixties of the twentieth century when solving one-dimensional problems using factorisation methods (V.S. Vladimirov, M.V. Keldysh, I.M. Gel'fand, O.V. Lokutsievskii, V.V. Rusanov, S.K. Godunov, A.A. Abramov, V.B. Andreev).

At the beginning of the sixties, Douglas, Pieceman and Rachford proposed the method of alternating directions based on the reduction of multi-dimensional problems to a sequence of one-dimensional problems with tri-diagonal matrices easily converted in a computer by the factorisation methods. This method has had a significant effect on the construction of the theory of related methods and of the entire group of methods which at present are referred to as *splitting methods* (the fractional step methods). The theory of the splitting methods and their application to solving complex applied problems of mathematical physics has been described in the studies by G.I. Marchuk, A.A. Samarskii, N.N. Yanenko, E.G. D'yakonov, J. Douglas, J. Gunn, G. Strang, A. Birkhof, R. Varga, D. Young and many other investigators [38,61,62,67,80,90,102].

The class of the splitting methods (the fractional step methods) is large: the methods of component splitting, two-cyclic methods of multicomponent splitting, the predictor–corrector method, the methods of alternating directions, and others.

The splitting methods are used widely for the approximate solutions of many applied problems of hydrodynamics, protection of environment, and the meteorological theory of the climate, which plays a significant role in modern society. In many cases, these methods are the only methods that can be used to solve these problems.

2. INFORMATION FROM THE THEORY OF EVOLUTION EQUATIONS AND DIFFERENCE SCHEMES

Many splitting methods are formulated in applications of non-stationary problems of mathematical physics, possibly approximate ones, obtained by the application of difference schemes and reduced to the Cauchy problem for evolution equations. Therefore, the main assumptions of the theory of the Cauchy problem for these equations are useful when examining and substantiating many of the splitting methods. Below, we represent some of these assumptions and also give the main concepts of the theory of difference schemes.

2.1. Evolution equations

2.1.1. The Cauchy problem

We examine, in Banach space X, the equation

$$\frac{d\varphi}{dt} + A\varphi = 0, \quad t \in (0,T) \tag{1}$$

with linear operator A, independent of t and having a domain of definition $D(A)$ dense everywhere in X.

The solution of the equation in the segments $[0,T]$ is the function $\varphi(t)$, satisfying the conditions:

1) the values of the function $\varphi(t)$ belong to $D(A)$ for all $t \in [0,T]$;

2) at every point $t \in [0,T]$ there is a *strong* derivative of the function $\varphi'(t)$, i.e.

$$\left\| \varphi'(t) - \frac{\varphi(t+\Delta t) - \varphi(t)}{\Delta t} \right\|_X \to 0, \quad \Delta t \to 0,$$

where $\|\cdot\| \equiv \|\cdot\|_x$;

3) equation (1) is satisfied at all $t \in [0,T]$.

Evidently, the solution $\varphi(t)$ of equation (1) is a function continuous on $[0,T]$, i.e. $\|\varphi(t) - \varphi(t_0)\| \to 0$ at $t \to t_0$ $\forall t_0 \in [0,T]$.

The Cauchy problem for equation (1) on $[0,T]$ is the problem of finding the solution of equation (1) in $[0,T]$ satisfying the initial condition

$$\varphi(0) = \varphi_0 \in D(A). \tag{2}$$

It is said that the Cauchy problem is formulated correctly on $[0,T]$ if: (I) at any $\varphi_0 \in D(A)$ the problem has a unique solution, and (II) this solution depends continuously on the initial data in the sense that from $\varphi_n(0) \to 0$ ($\varphi_n(0) \in D(A)$) for the corresponding solutions $\varphi_n(t)$ we obtain $\varphi_n(t) \to 0$ at every $t \in [0,T]$.

Comment. Because of the independence of t of operator A, the correctness of the Cauchy problem in some sections $[0,T]$ indicates its correctness in any section $[0,T_1]$($T_1>0$), i.e. correctness on the entire half-axis $[0,\infty)$.

We examine the operator $U(t)$, which links the element $\varphi_0 \in D(A)$ with the value of the solution $\varphi(t)$ of the Cauchy problem ($\varphi(0) = \varphi_0$) at the moment of time $t > 0$.

If the Cauchy problem is well-posed, operator $U(t)$ is defined on $D(A)$. Because of the linearity of equation (1) and the property (I) the operator is additive and homogeneous, and because of the property (II) it is continuous. Since $D(A)$ is dense in X, the operator $U(t)$ maybe extended in respect of continuity to the linear bounded operator, defined in the entire space X which is also denoted by $U(t)$.

It is said that the family of linear bounded operators $U(t)$, which depend on the parameter t ($0 < t < \infty$), is referred to as a *semi-group* if

$$U(t_1+t_2) = U(t_1)\, U(t_2) \qquad (0 < t_1,\ t_2 < \infty). \tag{3}$$

It maybe shown that the operators $U(t)$, generated by the well-posed problem (1), (2) form a semi-group.

We now examine the function $U(t)\varphi_0$ at any $\varphi_0 \in X$ and $t > 0$. Since $D(A)$

is dense in X, there is a sequence of elements $\varphi_0^{(n)} \in D(A)$ such that $\varphi_0^{(n)} \to \varphi_0$, and consequently $\varphi_n(t) = U(t)\,\varphi_0^{(n)} \to U(t)\varphi_0$ because of the boundedness of the operator $U(t)$. Thus, the function $U(t)\varphi_0$ is the limit of the sequence of solutions of equation (1) on $(0,\infty)$ and maybe referred to as a *generalized solution* of this equation.

If $U(t)\varphi_0$ is a generalized solution, then $\|U(t)\varphi_0\|$ is measurable (as the limit of sequence of continuous functions). The semi-group property of the operators $U(t)$ makes it possible to enhance this claim.

Lemma 1. *If the Cauchy problem for equation* (1) *is well-posed, then all generalized solutions of this equation are continuous on* $(0,\infty)$.

Theorem 1. *If the Cauchy problem for equation* (1) *is well-posed, then its solution is given by the formula*

$$\varphi(t) = U(t)\varphi_0 \quad (\varphi_0 \in D(A)) \tag{4}$$

where $U(t)$ *is the semi-group of the operators strongly continuous at* $t > 0$.

Theorem 2. *If the Cauchy problem for equation* (1) *is well-posed, then every generalized solution of the equation increases at infinity at the rate not greater than the exponent, and*

$$\lim_{t\to\infty} \frac{\ln\|U(t)\|}{t} = \omega < \infty, \tag{5}$$

where the number ω *is referred to as the type of semi-group* $U(t)$ *and the type of Cauchy problem* (1), (2).

Generally speaking, the group of the generalized solutions of equations (1), which are not solutions of the Cauchy problem, may contain differentiable functions. $\mathcal{D}$ denotes the set of elements φ_0 for which $U(t)\varphi_0$, additionally determined at zero as φ_0, and differentiable (from the right) in zero. On elements from $\mathcal{D}$ we determine the linear operator

$$\lim_{t\to+0} \frac{U(t)\varphi_0 - \varphi_0}{t} = U'(0)\varphi_0. \tag{6}$$

Operator $U'(0)$ is referred to as the *generating operator of the semi-group*.

Lemma 2. *If* $\varphi_0 \in \mathcal{D}$ *the generalized solution* $U(t)\varphi_0$ *has a continuous derivative at* $t > 0$.

Theorem 3. *If the Cauchy problem for equation* (1) is well-posed, then $D(A) \subset \mathcal{D}$, $U'(0)\varphi_0 = -A\varphi_0$ *at* $\varphi_0 \in D(A)$ *and operator* A *generating the well-posed Cauchy problem, may be extended to the generating operator* $U'(0)$ *of the strongly continuous semi-group* $U(t)$.

The following two concepts are important in the theory of the Cauchy problems. The well-posed Cauchy problem is referred to as uniformly well-posed if from $\varphi_n(0) \to 0$ it follows that $\varphi_n(t) \to 0$ uniformly in t in every finite segment $[0,T]$. The semi-group $U(t)$ *belongs to the class* C_0 if it is strongly

continuous at $t > 0$ and satisfies the condition $\lim_{t\to+0} U(t)\varphi=\varphi$ at any $\varphi\in X$.

Theorem 4. *For a uniform well-posed Cauchy problem at any* $\varphi\in X$ *we have* $\lim_{t\to+0} U(t)\varphi=\varphi$, i.e. *the semi-group* $U(t)$, *generated by the uniformly well-posed Cauchy problem, belongs to class* C_0.

Theorem 5. If the semi-group belongs to class C_0, the following estimates applies to it:

$$||U(t)||\le Me^{\omega t}$$

where $M = \sup_{0\le t\le 1}||U(t)||$.

If $||U(t)||\le 1$ $(0 \le t \le \infty)$ the semi-group is referred to as *compressing*.

Theorem 6. *If the semi-group belongs to class* C_0, *the domain of definition* $\mathcal{D}$ *of the generating operator* $U'(0)$ *is dense everywhere in space* X *and in addition to this, the set of the elements on which all the degrees of the operator* $U'(0)$ *are determined, is dense everywhere in* X.

Theorem 7. *If the semi-group belongs to class* C_0, *the generating operator of the semi-group is closed.*

Theorem 8. *If the Cauchy problem for equation (1) is uniformly well-posed, the closure of the operator* $-A$ *coincides with the operator* $U'(0)$.

Theorem 9. *To ensure that the problem* (1), (2), *where* A *is a closed operator, its uniformly well-posed, it is necessary and sufficient that* $-A$ *is the generating operator of the semi-group of class* C_0.

Thus, if we restrict ourselves to examining the equations with closed operators, then the class of equation (1) for which the Cauchy problem is uniformly well-posed, coincides with the class of the equations, for which the operator $-A$ is generating for the semi-group of class C_0.

2.1.2. *The nonhomogeneous evolution equation*

We examine the Cauchy problem for the nonhomogeneous equation of the type

$$\frac{d\varphi}{dt}+A\varphi=f(t), \qquad \varphi(0)=\varphi_0, \tag{7}$$

where $f(t)$ is the given continuous function with the values in X and $\varphi_0\in D(A)$.

Theorem 10. *If for equation* (1) *the Cauchy problem is uniformly well-posed, then the equation*

$$\varphi(t)=U(t)\varphi(0)+\int_0^t U(t-s)f(s)ds \tag{8}$$

gives the solution of the problem (7) *at* $\varphi\in$ D(A) *and function* $f(t)$, *satisfying*

one of the two conditions:

1) *Values of* $f(t)\in D(A)$ *and function* $Af(t)$ *are continuous*;
2) *Function* $f(t)$ *is continuously differentiable.*

Equation (8) at any $\varphi_0\in X$ and the continuous function $f(t)$ gives a continuous function which is referred to as the *generalized solution* of the Cauchy problem (7).

2.1.3. Evolution equations with bounded operators

We examine an important (in particular, in the theory and applied splitting methods) class of problems (1), (2) and (7), where operator A is bounded. In this case, it is evident that the Cauchy problem for the equation (1) is uniformly well-posed, and the semi-group $U(t)$ is presented in the form

$$U(t) = e^{-tA}, \tag{9}$$

where function e^{tB} determined by the series

$$e^{tB} = \sum_{n=0}^{\infty} \frac{1}{n!}(tB)^n \tag{10}$$

has the following properties

$$\begin{aligned} &e^{(t_1+t_2)B} = e^{t_1B}e^{t_2B}, \quad \frac{d}{dt}(e^{tB}) = Be^{tB}, \\ &e^{(A+B)t} - e^{At}e^{Bt} = (BA-AB)\frac{t^2}{2} + \ldots, \\ &e^{(A+B)t} = e^{At}e^{Bt}, \quad \text{if} \quad AB = BA. \end{aligned} \tag{11}$$

The solution of the problems (1), (2) and (7) in this case is given by the respective formulae

$$\varphi(t) = e^{-tA}\varphi_0, \tag{12}$$

$$\varphi(t) = e^{-tA}\varphi_0 + \int_0^t e^{(-t-s)A} f(s)ds. \tag{13}$$

We note another simple consequence of the representation (12) and properties (11). Let it be that the operator A is in the form of the sum of two commutative operators: $A=A_1+A_2$, where $A_1A_2=A_2A_1$. Then, according to (12) and the last property from (11) we have $\varphi(T) = e^{-TA_2}\tilde{\varphi}$, where $\tilde{\varphi} = e^{-TA_1}\varphi_0$. Thus, *to find the solution of the problem* (1), (2) (*in which* $A = A_1 + A_2$, $A_1A_2 = A_2A_1$) at $t = T$, *it is sufficient to solve consecutively the problems of the type*

$$\begin{aligned} &\frac{d\varphi_1}{dt} + A_1\varphi_1 = 0, \quad t\in[0,T], \quad \varphi_1(0) = \varphi_0, \\ &\frac{d\varphi_2}{dt} + A_2\varphi_2 = 0, \quad t\in[0,T], \quad \varphi_2(0) = \varphi_1(T) \end{aligned} \tag{14}$$

and accept $\varphi(T) = \varphi_2(T)$. It may be seen that if A_1, A_2 have a simpler structure than A, and we are interested in function $\varphi(T)$, then finding this function by solving the two problem (14) may be preferred to the determination of $\varphi(T)$ by finding the solution directly of the problems (1), (2).

To supplement the properties (11), we present two further formulae (playing a significant role in the theory of the splitting methods).

Let the bounded operator A be a sum of two positive definite operators A_1 and A_2:

$$A = A_1 + A_2, \quad (A_i\varphi, \varphi) \geq \gamma_i^2 \|\varphi\|^2, \quad i = 1, 2. \tag{15}$$

Consequently, the *Trotter formula* holds

$$e^{-tA} = \lim_{N\to\infty} \left\{ e^{-\frac{t}{N}A_2} e^{-\frac{t}{N}A_1} \right\}^N \tag{16}$$

and the similar *Chernov formula*:

$$e^{-tA} = \lim_{N\to\infty} \left\{ \left(I + \frac{t}{N} A_2 \right)^{-1} \left(I + \frac{t}{N} A_1 \right)^{-1} \right\}^N, \tag{17}$$

where I is the identical operator (which is often also denoted by E).

The formulae (16) and (17) can be used to find the solution of the problem (1), (2) at $t = T$. Actually, let us assume that we examine the problem (1), (2) in which the relationships (15) hold. Selecting N relatively large, we obtain that the function

$$\varphi^N(T) \equiv \left\{ e^{-\frac{T}{N}A_2} e^{-\frac{T}{N}A_1} \right\}^N \varphi_0$$

will approach $\varphi(T)=e^{-TA}\varphi_0$ (because of (16)). Finding $\varphi^N(T)$ is reduced to solving the sequence of the problems of the type (14).

However, if it is assumed that

$$\varphi^N(T) \equiv \left\{ \left(I + \frac{T}{N} A_2 \right)^{-1} \left(I + \frac{T}{N} A_1 \right)^{-1} \right\}^N \varphi_0,$$

then $\varphi^N(T)$ maybe regarded as an approximation of $\varphi(T)$ according to (17). In this case, $\varphi^N(T)$ can be determined as follows. We separate the interval $[0,T]$ into N intervals of equal length $\Delta t = T/N$ and determine in the recurrent manner the family of elements denoted by $\varphi^{n+i/2}$. These elements are determined successively for the increasing values and $n+i/2$ ($n = 0,\ldots, N-1$, $i = 1,2$). Starting the process of calculations $\varphi^0 \equiv \varphi_0$ and assuming that at the $(n+1)$-th step $\varphi^0, \ldots, \varphi^n$ are known, the elements $\varphi^{n+1/2}, \varphi^{n+1}$ are determined as the solution of the equations

$$\begin{aligned} &\frac{\varphi^{n+1/2} - \varphi^n}{\Delta t} + A_1\varphi^{n+1/2} = 0, \\ &\frac{\varphi^{n+1} - \varphi^{n+1/2}}{\Delta t} + A_1\varphi^{n+1} = 0. \end{aligned} \tag{18}$$

Since $\varphi^{n+1/2}=(I+\Delta t A_1)^{-1}\varphi^n$, $\varphi^{n+1}=(I+\Delta t A_2)^{-1}\varphi^{n+1/2}$, we conclude that $\varphi^N=\varphi^N(T)$. Thus, element φ^N is approximately equal at the moment $t = T$ to the solution $\varphi(t)$ of the problem (1), (2) and the determination of φ^N consists of the successive solution of the problem (18) which may be carried out sufficiently and economically if each of the operators A_1 and A_2 has a simple or 'special

structure' in comparison with operator A in the initial problem (1), (2). Consequently, *formulae* (16) *and* (17) *and algorithms for realization of these formulae by solving the problems of the type* (14) *or* (18) *form the base of the class of the splitting methods* (*the method of fractional steps*) *whose partial cases are the algorithms (14) and (18).*

2.2. Operator equations in finite-dimensional spaces

2.2.1. The evolution system

Let $X \equiv \mathrm{R}^N$ be the N-dimensional Euclidean space of vectors $\varphi = (\varphi_1,\ldots,\varphi_N)$ with some scalar product (φ,ψ) and the norm $||\varphi|| = (\varphi,\varphi)^{1/2}$. It should be mentioned that since in *any finite-dimensional space all the norms are equivalent*, the norm $||\varphi||$ is equivalent to the Euclidean norm

$$\|\varphi\|_2 \equiv |\varphi| = \left(\sum_{j=1}^{N} |\varphi_j|^2\right)^{1/2}.$$

We examine the Cauchy problem (7) where operator A is the matrix $A=\{a_{ij}\}$ with size $N \times N$ with the elements a_{ij}, $i, j = 1,\ldots,N$ independent of t. Consequently, the solution of the problem (7) does exist, it is unique and given by equation (13).

Let the matrix A be *non-degenerate*, i.e. there is A^{-1}, and vector f does not depend on t. Therefore, taking into account that $A^{-1}e^{-tA} = e^{-tA}A^{-1}$, from (13) we obtain the following expression for $\varphi(t)$:

$$\varphi(t) = e^{-tA}\varphi_0 + A^{-1}(1-e^{-tA})f. \tag{19}$$

It is assumed that the matrix is symmetric: $a_{ij} = a_{ji}$, $i, j = 1,\ldots,N$ (i.e. $A=A^T$), and positive definite: $(A\varphi,\varphi) \geq \alpha||\varphi||^2 \; \forall \; \varphi \in \mathrm{R}^N$, where $\alpha = \text{const}>0$. $\{\varphi^{(k)}\}$, $\{\lambda_k\}$ denote the eigenvectors and eigenvalues of A: $A\varphi^{(k)} = \lambda_k\varphi^{(k)}$, $k = 1, \ldots,N$. The eigenvectors of any symmetric matrix can be selected in such a manner that they form an orthonormal system in R^N. Therefore, it is assumed that

$$(\varphi^{(k)},\varphi^{(l)}) = \delta_{kl} \equiv \begin{cases} 1 & \text{at} \quad k = l, \\ 0 & \text{at} \quad k \neq l. \end{cases}$$

Q denotes a matrix whose columns are the eigenvectors of the matrix A. Consequently, it is evident that $Q^T Q = I$ is the unit matrix, i.e. the matrix whose extra-diagonal elements $(I)_{kl}$, $k \neq l$, are equal to zero, and elements of the main diagonal are equal to unity. (It should be mentioned that in many cases the unit matrix is also denoted by E). Consequently, $Q^T = Q^{-1}$, i.e. Q is an *orthogonal matrix*.

Taking into account the properties of the matrix Q and the eigenvectors $\{\varphi^{(l)}\}$, we easily obtain the following representation for A and e^{-tA}:

$$A = Q^T \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_N \end{bmatrix} Q, \quad e^{-tA} = Q^T \begin{bmatrix} e^{-t\lambda_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & e^{-t\lambda_N} \end{bmatrix} Q, \tag{20}$$

and equation (13) has the form

$$\varphi(t) = \sum_{k=1}^{N} \beta_k(t)\varphi^{(k)}, \tag{21}$$

where

$$\beta_k(t) = e^{-\lambda_k t}(\varphi_0, \varphi^{(k)}) + \int_0^t e^{-\lambda k(t-s)}(f(s), \varphi^{(k)})ds, \tag{22}$$

$$\varphi_0 = \sum_{k=1}^{N}(\varphi_0, \varphi^{(k)})\varphi^{(k)}, \quad f(t) = \sum_{k=1}^{N}(f(t), \varphi^{(k)})\varphi^{(k)},$$

i.e. the equation (21) is the representation of the solution of the system (7) obtained by the *method of eigenfunctions* (*vectors*).

If f is independent of t, then

$$\beta_k(t) = e^{-\lambda kt}(\varphi_0, \varphi^{(k)}) + \frac{(f, \varphi^{(k)})}{\lambda_k}(1 - e^{-\lambda_k t}), \quad k = 1, \ldots, N.$$

It should be mentioned that *all eigenvalues of any symmetric positive definite matrix are positive*. Therefore, in the case of the matrix A we have $\lambda_k \geq \alpha = \text{const} > 0$, $k = 1,\ldots,N$. Taking this into account, from (21), (22) we obtain

$$\lim_{t\to\infty} \beta_k(t) = \beta_k \equiv \frac{(f, \varphi^{(k)})}{\lambda_k}, \qquad \lim_{t\to\infty} \varphi(t) = \varphi \equiv \sum_{k=1}^{N} \beta_k \varphi^{(k)}, \tag{23}$$

and φ is the solution of the system of linear algerbriac equation

$$A\varphi = f \tag{24}$$

with a symmetric positive definite matrix A.

The following estimate holds for the difference

$$\|\varphi(t) - \varphi\| \leq \|\varphi_0 - \varphi\| e^{-\alpha t}, \tag{25}$$

and *regardless of the initial condition* φ_0 we have $\varphi(T) \to \varphi$ at $t = T \to \infty$. This fact plays a fundamental role in constructing approximate solutions of the system of linear algebraic equations with symmetric positive definite matrices by the *method of stationarisation and the splitting methods* (already examined as the methods of solving the systems of type (24)).

2.2.2. Stationarisation method

We need to find the solution

$$\varphi = A^{-1}\varphi_0 f$$

of the system (24) with a symmetric positive definite matrix A. This can be achieved, by for example, the *method of excluding unknown quantities* (the *Gauss method*), belonging to the group of *direct methods* of solving the systems of type (24), i.e. the methods which make it possible to find, at a finite number of operations (possibly a large number, if N is large), the exact (theoretically) solution $\varphi = A^{-1}f$. However, in practical calculations, there are

errors of different types (in definition of f, the errors of calculation in a computer, etc). Therefore, $\varphi = A^{-1}f$ is actually calculated approximately. Consequently, it is natural to make an assumption on the rationality of formulating the problem of constructing the knowingly approximate solution of the system (24) with any accuracy given in advance. The solution of this problem can be obtained using an entire group of iteration algorithms, many of which are based on the stationarisation methods. We explain the concept of this method with special reference to system (24).

Let A be symmetric and positive definite. On the basis of (24) it is concluded that at $t\to\infty$ the solution $\varphi(t)$ of the problem (7) converges in R^N to the solution of the problem (24), i.e. $||\varphi(t)-\varphi||\to 0$ at $t\to\infty$ (this claim is also valid in all general cases: infinite-dimensional spaces, non-linear and unbounded operators, etc, but these cases are not examined here). Consequently, for a sufficiently large $t = T$ and 'rational' φ_0 the error $||\varphi(T)=\varphi||$ can be made as small as required (and also positive). Thus, if we find the solution $\varphi(t)$ of the Cauchy problem (7) at $t = T$ – a sufficiently large number, then $\varphi(T)$ maybe regarded as the approximate solution of the system (24). Therefore, to obtain the approximate solution of the system (24) we can:

1) examine the evolution problem (7) (with an arbitrary but 'rational' initial element φ_0);

2) find the solution $\varphi(t)$ at $t = T$ – a sufficiently large number ($\varphi(T)$ can also be found approximately);

3) accept the element (T) or approximation to $\varphi(T)$ as approximate to φ.

The stages 1) – 3) form the principle of the method of stationarisation of the approximate solution of the system of the type (24) by examining the evolution system and constructing their solutions.

In turn, $\varphi(T)$ can be constructed using different methods. For example, it is assumed that $A = A_1 + A_2$, where A_1, A_2 are the positive definite matrices with a simple structure. If $A_1 A_2 = A_2 A_1$, construction of $\varphi(t)$ is reduced to a successive solution of the problem (14) with the matrices A_1, A_2 which are simpler than A. Another method of approximate construction of $\varphi(T)$ is given by equation (18). Thus, *the splitting methods* (14), (18) *are already regarded here as the methods of solving the systems of equations of type* (24).

2.3. Concepts and information from the theory of difference schemes

2.3.1. Approximation

We shall mention several concepts of the theory of finite difference methods used in subsequent considerations. We examine a stationary problem of mathematical physics in the operator form

$$A\varphi = f \quad \text{in} \quad \Omega, \quad \Omega\in \mathsf{R}^n, \quad a\varphi = g \quad \text{on } \partial\Omega \tag{26}$$

where A is a linear operator, $\varphi\in\Phi$, $f\in F$, here Φ, F are real spaces, whose elements are determined on $\Omega\cup\partial\Omega=\bar{\Omega}$ and Ω, respectively, a is the linear operator of the boundary condition, $g\in G$, where G is the real Hilbert space of the functions determined on $\partial\Omega$. For definiteness and simplifying notations, we assume here that the functions from Φ, F, G depend only on two

variables x and y (which can be regarded as spatial variables).

We construct a finite-dimensional approximation of problems (26), for example by the finite-difference method. For this purpose we examine a set of points (x_k, y_l), where k, l are arbitrary integers. The set of the points of this type is termed the *grid* and the points the *nodes of the grid*. The distance between the nodes of the grid is characterised by the number h – the *step of the grid* (naturally, this distance can also be estimated by two parameters h_x, h_y – the steps of the grid in respect of x and y, respectively). Ω_h denotes the set of the nodes of the grid approximating (in some sense) the set of the points of domain Ω, and $\partial\Omega_h$ denotes the set of the nodes approximating the boundary $\partial\Omega$. The functions whose definition domain is a grid will be referred to as *grid functions*. The set of the grid functions φ^h with the domain of definition Ω_h is denoted by Φ_h. Each function $\varphi\in\Phi$ can be related with the grid function $(\varphi)_h$ in accordance with the rule: the value $(\varphi)_h$ in the node (x_k, y_l) is equal to $\varphi(x_k, y_l)$ (of course, if the values $\{\varphi(x_k, y_l)\}$have any meaning). This correspondence is a linear operator acting from Φ in Φ_h; this operator is referred to as the projection of the function of φ on the grid. Function $\psi \equiv A\varphi$ can also be projected on the grid, assuming $(\psi)_h = (A\varphi)$. The correspondence between $(\varphi)_h$ and $(A\varphi)_h$ is a linear operator, determined on the grid functions.

We examine the problem in the finite dimensional space of the grid function

$$A^h\varphi^h = f^h \text{ in } \Omega_h, \quad a^h\varphi^h = g^h \text{ on } \partial\Omega_h \tag{27}$$

is the finite-difference analogue of the problem (26). Here A^h, a^h are linear operators which depend on the step of the grid h, $\varphi^h \in \Phi_h$, $f^h\in F_h$, $g^h\in G_h$, and Φ_h, F_h, G_h are the spaces of the real grid functions.

We introduce in Φ_h, f_h G_h the respective norms $\|\cdot\|_{\Phi_h}$, $\|\cdot\|_{F_h}$, $\|\cdot\|_{G_h}$. Let $(\cdot)_h$ be the notation of the linear operator which relates the element $\varphi\in\Phi$ with the element $(\varphi)_h \in \Phi_h$, such that $\lim_{h\to 0}||(\varphi)_h||_{\Phi_h} = ||(\varphi)||_\Phi$. We shall say that the problem (27) *approximates* the problem (26) with the order n on the solution φ, if there are such positive constants $\bar{h}$, M_1, M_2, that for all $h<\bar{h}$ the following inequalities are satisfied

$$\left\|A^h(\varphi)_h - f^h\right\|_{F_h} \le M_1 h^{n_1}, \quad \left\|a^h(\varphi)_h - g^h\right\|_{G_h} \le M_2 h^{n_2}$$

and $n = \min(n_1, n_2)$.

We shall also assume that the problem (26) has been reduced to problem (27). If the boundary condition from (27) is used to exclude the values of the solution in the boundary points of the domain $\Omega_h\cup\partial\Omega_h$, we obtain the equivalent problem

$$\tilde{A}^h\,\tilde{\varphi}^h = \tilde{f}^h. \tag{28}$$

The values of the solution at the boundary points can be determined from equation (27) after constructing the solution of the equation (28). In some

cases it is convenient to use the approximation problem in the form (26) and in some cases in the form (27). In the theory of difference schemes, we often use real Hilbert spaces of the grid functions with the norms $\left\|\varphi^h\right\|_{\Phi_h} = (\varphi^h, \varphi^h)^{1/2}_{\Phi_h}$, $\left\|f^h\right\|_{F_h} = (f_h, f_h)^{1/2}_{F_h}$, and so on. However, it should be mentioned that many of the introduced concepts (approximation, etc) can be transferred to the case of Banach spaces, and in a number of assertions and illustration examples we introduce the norms of grid functions which are not linked with the scalar product by the above-mentioned relationship.

We illustrate this on the example of the problem

$$-\Delta\varphi = f \quad \text{in} \quad \Omega, \quad \varphi = 0 \quad \text{on} \quad \partial\Omega, \tag{29}$$

where $\Omega=\{(x,y):0<x<1,\ 0<y<1\}$, $f=f(x,y)$ is a smooth function, $\Delta = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$. Let F be the Hilbert space of the real functions $L_2(\Omega)$ with the scalar product $(u,v) = \int_\Omega uv\, dx\, dy$ and the norm $||u|| = (u,u)^{1/2}$. Here Φ denotes the set of the functions continuous in $\bar{\Omega} = \Omega \cup \partial\Omega$ and having the first and second derivatives continuous in Ω with the same norm $||\cdot||_\Phi = ||\cdot||$ as in F. G is represented by the Hilbert space of the functions $L_2(\partial\Omega)$, determined on $\partial\Omega$, with the norm $\|g\|_{L_2(\partial\Omega)} = \left(\int_{\partial\Omega} |g|^2 d\Gamma\right)^{1/2}$. If the operators $A\varphi = -\Delta\varphi$, $a\varphi = \varphi|_{\partial\Omega}$ are introduced, the problem (29) can be represented in the form (26) at $g \equiv 0$.

We introduce the finite-dimensional approximation of the problem (29). For this purpose, the square $\bar{\Omega} = \Omega \cup \partial\Omega$ is covered by a grid with step h and uniform in respect of x and y. The nodes of the domain will be denoted by the indices k, l, where the index k ($0 \le k \le N$) corresponds to the point of division in respect of the coordinate x, and the index l ($0 \le l \le N$) in respect of y. We examine the following approximations:

$$\frac{\partial^2\varphi}{\partial x^2} \equiv \varphi_{xx} \cong \Delta_x \nabla_x (\varphi)_h, \qquad \frac{\partial^2\varphi}{\partial y^2} \equiv \varphi_{yy} \cong \Delta_y \nabla_y (\varphi)_h,$$

where Δ_x, Δ_y, ∇_x, ∇_y are the difference operators defined on the grid function φ^h (with the components φ^h_{kl}) in the following manner:

$$(\Delta_x \varphi^h)_{kl} = \frac{1}{h}\left(\varphi^h_{k+1,l} - \varphi^h_{kl}\right), \quad (\nabla_x \varphi^h)_{kl} = \frac{1}{h}\left(\varphi^h_{kl} - \varphi^h_{k-1,l}\right),$$

$$(\Delta_y \varphi^h)_{kl} = \frac{1}{h}\left(\varphi^h_{k,l+1} - \varphi^h_{kl}\right), \quad (\nabla_y \varphi^h)_{kl} = \frac{1}{h}\left(\varphi^h_{kl} - \varphi^h_{k,l-1}\right).$$

Consequently, the problem (29) can be approximated by a finite-difference problem

$$A^h\varphi^h \equiv -\Delta^h\varphi^h \equiv -[\Delta_x\nabla_x\varphi^h + \Delta_y\nabla_y\varphi^h] = f^h \ \text{ in } \Omega_h, \quad \varphi^h = 0 \ \text{ on } \partial\Omega_h,$$

where $\partial\Omega_h$ is the set of the nodes belonging to the boundary $\partial\Omega$, and Ω_h is the set of the nodes of the grid internal in Ω. Here

$$A^h = A_x + A_y, \qquad A_x = -\Delta_x\nabla_x, \qquad A_y = -\Delta_y\nabla_y,$$

i.e. A^h is the difference analogue of the operator $-\Delta$: $A^h \equiv -\Delta^h$, and φ^h and f^h are the vectors with the components φ^h_{kl} and f^h_{kl} and

$$(\Delta^h \varphi^h)_{kl} = \frac{1}{h^2}\left(\varphi^h_{k+1,l} + \varphi^h_{k-1,l} + \varphi^h_{k,l+1} + \varphi^h_{k,l-1} - 4\varphi^h_{kl}\right),$$

$$f^h_{kl} = \frac{1}{h^2}\int_{x_{k-1/2}}^{x_{k+1/2}}\int_{y_{l-1/2}}^{y_{l+1/2}} f\ dx\ dy, \quad x_{k\pm 1/2} = x_k \pm \frac{h}{2}, \quad y_{l\pm 1/2} = y_l \pm \frac{h}{2}.$$

In the schemes given here and below f^h_{kl} is taken as the averaging of the function $f(x,y)$ calculated from the previously given equation. (This, generally speaking, makes it possible to examine the difference schemes with the function $f(x,y)$ not being sufficiently smooth).

We examine space Φ_h. The domain of definition of the grid functions from Φ_h is represented by $\bar{\Omega}_h = \Omega_h \cup \partial\Omega_h = \{(x_k, y_l)\colon x_k = hk,\ y_l = hl\}$. The scalar product and the norm in Φ_h are defined from:

$$(\varphi^h, \psi^h)_{\Phi_h} = \sum_{k,l=0}^{N} h^2 \varphi^h_{kl}\psi^h_{kl}, \quad \left\|\varphi^h\right\|_{\Phi_h} = \left(\sum_{k,l=0}^{N} h^2 (\varphi^h_{kl})^2\right)^{1/2}.$$

As F_h we use the space of the grid functions defined on the set

$$\Omega^h = \{(x_k, y_l) : x_k = hk, y_l = hl, 1 \le k \le N-1,\ 1 \le l \le N-1\},$$

with the scalar product and the norm:

$$(\varphi^h, \psi^h)_{F_h} = \sum_{k,l=0}^{N-1} h^2 \varphi^h_{kl}\psi^h_{kl}, \quad \left\|\varphi^h\right\|_{F_h} = \left(\sum_{k,l=0}^{N-1} h^2 (\varphi^h_{kl})^2\right)^{1/2}.$$

In the same manner, we introduce the space G_h of the grid functions determined on $\partial\Omega_h$.

As $(\varphi)_h$ we use the vector whose components are the values of the function in the corresponding nodes of the grid. Consequently, using the Taylor series expansion of the functions $\varphi(x,y)$ and $f(x,y)$ we obtain

$$\left\|-\Delta^h(\varphi)_h - f^h\right\|_{F_h} \le M_1 h^2,$$

where M_1 = const < ∞. The approximation of the boundary conditions on $\partial\Omega_h$ in this example is carried out without errors. Taking into account the last estimate, this means that the examined finite-difference problem approximates the problem (29) with the second order on the solution of the problem (29) having bounded fourth derivatives.

It should be mentioned that if it is required that the grid functions from Φ_h satisfy the condition $\varphi^h\big|_{\partial\Omega_h} = 0$, then in this case the scalar products in Φ_h and F_h (and, therefore, the norms generated by them) coincide for these functions. We examine the following identities analogous to the first and second Green formulas:

$$-\sum_{k=1}^{N-1}(\Delta_x\nabla_x\varphi^h)_{kl}\,\psi^h_{kl}=\sum_{k=1}^{N}(\nabla_x\varphi^h)_{kl}(\nabla_x\psi^h)_{kl}$$

$$-\sum_{k=1}^{N-1}(\Delta_x\nabla_x\varphi^h)_{kl}\,\psi^h_{kl}=-\sum_{k=1}^{N}(\Delta_x\nabla_x\psi^h)_{kl}\varphi^h_{kl},\quad \phi^h,\psi^h\in\Phi_h.$$

Using them and identical identities for the sums in respect of index l, it may easily be shown that

$$(A^h\varphi^h,\psi^h)_{F_h}=(\varphi^h,A^h\psi^h)_{F_h},$$

$$(A^h\varphi^h,\varphi^h)_{F_h}=h^2\sum_{k,l=1}^{N}\left[\left(\left(\nabla_x\varphi^h\right)_{kl}\right)^2+\left(\left(\nabla_y\varphi^h\right)_{kl}\right)^2\right]>0$$

at $\varphi^h\not\equiv 0$, i.e. operator A^h on functions from Φ^h, satisfying the conditions $\varphi^h\big|_{\partial\Omega_h}=0$ is a self-adjoint, positive definite operator. It should also be mentioned that operator A^h is presented here in the form of the sum of the symmetric positive definite commutative operators A_x, A_y: $A^h = A_x+A_y$, $A_xA_y=A_yA_x$.

We now examine the problem of approximating the evolution equation

$$\begin{gathered}\frac{\partial\varphi}{\partial t}+A\varphi=f\quad\text{in}\quad\Omega\times\Omega_t,\quad\Omega_t=(0,T),\\ a\varphi=g\quad\text{on}\ \partial\Omega\times\Omega_t,\qquad\varphi=\varphi_0\quad\text{in}\quad\Omega\ \text{ at } t=0.\end{gathered}\tag{30}$$

The problem (30) will be approximated in two stages. Initially, the problem will be approximated in the domain $(\Omega_h\cup\partial\Omega_h)\times\Omega_t$ in respect of the spatial variables. Consequently, we obtain the equation, differential in respect of time and difference in respect of the spatial variables.

In the resultant differential-difference problem it is often easy to exclude solutions at the boundary points of the domain $(\Omega_h\cup\partial\Omega_h)\times\Omega_t$ using difference approximations of the boundary conditions. Assuming that this has been carried out, we obtain the evolution equation of the type

$$\frac{d\varphi^h}{dt}+\Lambda\varphi^h=f^h,\tag{31}$$

where $\Lambda\equiv\tilde{A}^h$, f^h and φ^h are the functions of time t. Equation (31) is the system of conventional differential equations for the components of the vector φ^h. To simplify considerations, the index h in the problem (31) can be omitted, assuming that (31) is a difference analogue in respect on the spatial variables of the initial problem of mathematical physics.

Taking these considerations into account, we examine the Cauchy problem

$$\frac{d\varphi}{dt}+\Lambda\varphi=f,\quad\varphi=g\quad\text{at}\quad t=0.\tag{32}$$

It is assumed that operator Λ is independent of time. We examine the simplest methods of approximating the problem (32) in respect of time. The most useful difference schemes at present are the schemes of the first and second order of approximation in respect of t.

One of them is the *explicit scheme of the first order of approximation on* the grid Ω_τ (i.e. on the set of the nodes $\{t_j\}$ in respect of variable t):

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\varphi^j = f^j, \quad \varphi^0 = g, \tag{33}$$

where $\tau = t_{j+1}-t_j$ and for f^j we can accept here $f^j = f(t_j)$.

The *implicit scheme of the first order of approximation* has the form

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\varphi^{j+1} = f^j, \quad \varphi^0 = g \tag{34}$$

at $f^j = f(t_{j+1})$. The schemes (33) and (34) are of the first order of approximation in respect of time. This may easily be seen by means of expansion using the Taylor equation in respect of time, assuming, for example, the existence of bounded derivatives (in respect of time) of the second order of the solution.

Solving the schemes (33) and (34) in relation to φ^{j+1} we obtain the recurrent relationship

$$\varphi^{j+1}=T\varphi^j+\tau Sf^j, \tag{35}$$

where T is the operator of the step, S is the operator of the source, determined in the following manner: for the scheme (33) $T = E-\tau\Lambda$, $S = E$; for the scheme (34) $T=(E+\tau\Lambda)^{-1}$, $S = T$.

The difference schemes of the type (35) for evolution equations are referred to as the *two-layer schemes*.

The scheme of the second order of approximation – the *Cranck–Nicholson scheme* – is used widely in the applications:

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\frac{\varphi^{j+1}+\varphi^j}{2} = f^j, \quad \varphi^0 = g, \tag{36}$$

where $f^j = f(t_{j+1/2})$. The scheme (36) can also be presented in the form (35) at

$$T=\left(E+\frac{\tau}{2}\Lambda\right)^{-1}\left(E-\frac{\tau}{2}\Lambda\right), \quad S=\left(E+\frac{\tau}{2}\Lambda\right)^{-1}.$$

In some cases, the difference equations (33), (34) and (36) can be written conveniently in the form of a system of two equations from which one approximates the equation itself in $\Omega_{h\tau}$, and the second – the boundary condition on $\partial\Omega_{h\tau}$. In this case, the difference analogue of the problem (30) has the form

$$\begin{gathered} L^{h\tau}\varphi^{h\tau} = f^{h\tau} \ \text{ in } \ \Omega_{h\tau}, \qquad l^{h\tau}\varphi^{h\tau} = g^{h\tau} \ \text{ on } \ \partial\Omega_{h\tau}, \\ \Omega_{h\tau} = \Omega_h\times\Omega_\tau, \quad \partial\Omega_{h\tau} = \Omega_h\times\{0\}\cup\partial\Omega_h\times\Omega_\tau, \end{gathered} \tag{37}$$

where the operators $L^{h\tau}$, $l^{h\tau}$ and $f^{h\tau}$, $g^{h\tau}$, satisfy the inequalities

$$\left\|L^{h\tau}(\varphi)_{h\tau} - f^{h\tau}\right\|F_{h\tau} \le M_1h^n + N_1\tau^p,$$

$$\left\|l^{h\tau}(\varphi)_{h\tau} - g^{h\tau}\right\|G_{h\tau} \le M_2h^n + N_2\tau^p.$$

In these inequalities $(\ldots)_{h\tau}$ is the operator of projection on the appropriate grid space.

Using the vector-functions determined on $\Omega_h\times\Omega_\tau$ (where Ω_τ is the set $\{t_j\}$), and new operators, (37) can also be written in the form

$$\tilde{L}^{h\tau}\tilde{\varphi}^{h\tau} = \tilde{f}^{h\tau}. \tag{38}$$

Thus, the evolution equation taking into account the boundary conditions and initial data can be approximately reduced to a problem of linear algebra (38) in finite-dimensional space.

2.3.2. Stability

We shall discuss another important concept of the theory of finite-difference methods – *stability*. For this purpose, we examine the problem

$$\frac{\partial\varphi}{\partial t} + A\varphi = f \quad \text{in} \quad \Omega\times\Omega_t, \quad \varphi = g \ \text{ at } \ t = 0, \tag{39}$$

which is approximated by the difference problem

$$\varphi^{j+1} = T\varphi^j + \tau S f^j \ \text{ on } \ \Omega_h\times\Omega_\tau, \quad \varphi^0 = g. \tag{40}$$

It will be assumed that the difference scheme (40) is stable if at any parameter h characterising the difference approximation and $j \le T/\tau$ we can write the following relationship

$$\left\|\varphi^j\right\|_{\Phi_h} \le C_1\|g\|_{G_h} + C_2\left\|f^{h\tau}\right\|_{F_{h\tau}}, \tag{41}$$

where the constants C_1 and C_2 are uniformly bounded on $0 \le t \le T$, and are independent of τ, h, g and f; G_h denotes the space to which g belongs in (40).

The determination of stability is closely linked with the concept of the correctness of a problem with a continuous argument. It maybe said that the stability establishes the continuous dependence of the solution on the initial data in the case of problems with a discrete argument. It may easily be seen that the determination of stability in the sense of fulfilling (41) already links the solution itself with the apriori information on the initial data of the problem.

It may easily be shown that the difference schemes (34), (36) are stable in the sense of (41) if $\lambda > 0$ and $\left\|\varphi^j\right\|_{\Phi_h} = \left[\sum_{k,l}\left|\varphi_{kl}^j\right|^2 h^2\right]^{1/2}$. This can be done on the basis of the estimate $\|(E+\tau\Lambda)^{-1}\| \le 1$ for $\Lambda \ge 0$, $\tau > 0$, resulting from relationships

$$\frac{\left\|(E+\tau\lambda)^{-1}\varphi\right\|^2_{\Phi_h}}{\|\varphi\|^2_{\Phi_h}} = \frac{\|\psi\|^2_{\Phi_h}}{\|(E+\tau\lambda)\psi\|^2_{\Phi_h}} =$$

$$= \frac{\|\psi\|^2_{\Phi_h}}{\|\psi\|^2_{\Phi_h} + \tau((\Lambda+\Lambda^*)\psi,\psi)_{\Phi_h} + \tau^2\|\Lambda\psi\|^2_{\Phi_h}} \le 1$$

and the following theorem which is used for analysis of the stability of many schemes.

Theorem 11 (Kellog theorem) *If the operator A, acting in the real Hilbert space* Φ*, is positive semi-definite, and the numerical parameter* σ *is not negative, then*

$$\left\|(E-\sigma A)(E+\sigma A)^{-1}\right\| \le 1.$$

Comment. At $A > 0$ and $\sigma > 0$, the following inequality holds in the theorem 11

$$\left\|(E-\sigma A)(E+\sigma A)^{-1}\right\| < 1.$$

When examining the scheme (33) it may easily be seen that the scheme is stable under the additional condition of the type $||T||<1$, which, for example is fulfilled if Λ is the symmetric positive definite matrix with the eigenumbers from the segment $[\alpha,\beta]$, and τ satisfies the relationships $0 < \tau < 2/\beta$.

When solving the difference analogues of the evolution problems of mathematical physics it is necessary to deal with the approximation in respect of both time (with step t) and space (with characteristics step h). This means that the transition operator $T = T(\tau,h)$ depends on both τ and h.

The problem of designing a stable algorithm at the given method of approximation is usually reduced to the establishment of the relationship between τ and h ensuring stability. If the difference scheme is stable at any values of τ and h, it is regarded as *absolutely stable*. However, if the scheme is stable only at a specific relationship between τ and h, it is referred to as *conditionally stable*. (Thus, the schemes (34), (36) are absolutely stable, and the scheme (33) is conditionally stable).

Comment. If the approximation of the evolutionary equation is examined in the spaces of grid functions, determined on $\Omega_h \times \Omega_\tau$, then the definition of stability can often be carried out conveniently in terms of the same spaces. Infact, let the difference problem have the form

$$L^{h\tau}\varphi^{h\tau} = f^{h\tau} \text{ in } \Omega_h \times \Omega_\tau, \quad l^{h\tau}\varphi^{h\tau} = g^{h\tau} \text{ on } \partial\Omega_h \times \Omega_\tau.$$

We introduce the stability criterion

$$\left\|\varphi^{h\tau}\right\|_{\Phi_{h\tau}} \le C_1 \left\|f^{h\tau}\right\|_{F_{h\tau}} + C_2 \left\|g^{h\tau}\right\|_{G_{h\tau}},$$

where C_1 and C_2 are constants independent of h,τ, $f^{h\tau}$, $g^{h\tau}$.

However, if the initial problem of mathematical physics is approximated using the difference equation in such a manner that the boundary and initial conditions are also taken into account in constructing this equation, the stability criterion can be conveniently introduced in the following form:

$$\left\|\varphi^{h\tau}\right\|_{\Phi_{h\tau}} \le C \left\|f^{h\tau}\right\|_{F_{h\tau}}.$$

2.3.3. Convergence

We formulate the main result of finite-difference algorithm – the *convergence theorem*. The examination of the convergence of the difference solution to the solution of the original problem for both stationary and evolution prob-

lems of mathematical physics is carried out on the basis of the same principles. Therefore, we formulate the theorem of convergence when examining the stationary problem (26) approximated by the difference scheme (27) (i.e. the system of equations approximating both the equation and the boundary condition from (26)).

Theorem 12. *We assume that:*

1) *the difference scheme* (27) *approximates the original problem* (26) *on solution* φ *with order* n;

2) A^h *and* a^h *are the linear operators;*

3) *the difference scheme* (27) *is stable in the sense of* (41), i.e. *there are positive constants* $\bar{h}$, C_1, C_2 such that for all $h < \bar{h}$, $f^h \in F_h$, $g^h \in G_h$, there is a unique solution φ^h of the problem (27) satisfying the inequality

$$\left\|\varphi^h\right\|_{\Phi_h} \le C_1 \left\|f^h\right\|_{F_h} + C_2 \left\|g^h\right\|_{G_h}.$$

Consequently, the solution of the difference problem φ^h *is converging to the solution* φ *of the original problem,* i.e.

$$\lim_{h\to 0} \left\|(\varphi)_h - \varphi^h\right\|_{\Phi_h} = 0,$$

and the following estimate of the convergence rate holds:

$$\left\|(\varphi)_h - \varphi^h\right\|_{\Phi_h} \le (C_1 M_1 + C_2 M_2) h^n,$$

where M_1 *and* M_2 *are the constants from approximation estimates*

Thus, using the finite difference method and approximating the original evolution problem in respect of all variables (with the exception of variable t), this problem can be reduced to solving a system of ordinary differential equations:

$$\frac{d\varphi}{dt} + A\varphi = f \quad \text{in} \quad \Omega_t, \quad \varphi = g \quad \text{at} \quad t = 0. \tag{42}$$

This system can now be solved using the splitting methods. (It should be mentioned that this approach of successive approximation of the problems is often used in practical calculations).

2.3.4. The sweep method

It is required to find a solution of the following system of *three-point equations*:

$$\begin{aligned} c_0 y_0 - b_0 y_1 &= f_0, & i &= 0, \\ -a_i y_i - 1 + c_i y_i - b_i y_i + 1 &= f_i, & 1 &\le i \le N-1, \\ -a_N y_N - 1 + c_N y_N &= f_N, & i &= N. \end{aligned}$$

The systems of this type form in three-point approximations of the boundary-value problems for ordinary differential equations of the second order, for example, by the finite difference methods, and also in application of difference schemes for equations with partial derivatives. In the latter case, it is usually necessary to solve not one but a series of such problems with different right-hand sides. Therefore, effective methods of solving systems of this type are required. One of these methods is the *sweep method* (*factorisation method*) which is given by the following formulae.

1) *The direct course of sweep* (we determine *sweeping coefficients* α_i and β_i):

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i}, \quad i = 1, 2, \ldots, N-1, \qquad \alpha_1 = \frac{b_0}{c_0},$$

$$\beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}, \quad i = 1, 2, \ldots, N, \qquad \beta_1 = \frac{f_0}{c_0}.$$

2) *The reverse course* (values y_i are determined):

$$y_i = \alpha_i\, y_{i+1} + \beta_{i+1}, \quad i = N-1, N-2, \ldots, 0, \quad y_N = \beta_{N+1}.$$

The conditions of correctness and stability of this method are formulated as follows: *let the coefficients of the system of three-point equations be real and satisfy the conditions* $|b_0| \geq 0$, $|a_N| \geq 0$, $|c_0| > 0$, $|c_N| > 0$, $|a_i| > 0, |b_i| > 0$, $i = 1,2,\ldots, N-1$, *and also* $|c_i| \geq |a_i| + |b_i| > 0$, $i = 1,2,\ldots, N-1, |c_0| \geq |b_0|, |c_N| \geq |a_N|$, *and the strict inequality is fulfilled in at least one of the latter relationships. Consequently, the following inequalities apply to the sweep method* $c_i - a_i\alpha_i \neq 0$, $|\alpha_i| \leq 1, i = 1,2,\ldots,N$, *guaranteeing correctness and stability of the method.*

When solving the system of difference equations, we use other typesd of the sweep method (*the method of counter sweep, the flow variant of the sweep method, the cyclic sweep method, the matrix sweep method, etc*).

3. SPLITTING METHODS

We examine algorithms used widely at present in solving the complex multi-dimensional problems of mathematical physics. To concentrate special attention to their formulation, the main object of examination will be represented by the evolution problem of mathematical physics

$$\frac{d\varphi}{dt} + A\varphi = f \quad \text{in} \quad \Omega \times \Omega_t, \quad \varphi = g \quad \text{in} \quad \Omega \quad \text{at} \quad t = 0$$

with a positive semidefinite operator $A \geq 0$ (i.e. $(A\varphi,\varphi) \geq 0$). It is assumed that the approximation of the initial problem in respect of all variables (with the exception of t) has already been carried out, Ω is the grid domain, A is a matrix, φ, f, g are the grid functions. The index of the parameter of the grid h is omitted to simplify considerations. The solution of the problem on $\partial\Omega$ satisfies the given boundary conditions and taking this into account, A, f are determined. Let it also be that the spaces of the grid functions Φ and F coincide and, if not said otherwise, they have the same norm

$$\|\varphi\|_\Phi = (\varphi,\varphi)_\Phi^{1/2}.$$

The examined problem may also be written in the form

$$\frac{d\varphi}{dt} + A\varphi = f \quad \text{in} \quad \Omega_t, \quad \varphi = g \quad \text{at} \quad t = 0, \tag{43}$$

which means that the first equation is examined in $\Omega \times \Omega_t$, and the second in $\Omega \times \{0\}$. In subsequent considerations, we use mainly the last form of the problem.

It should be mentioned that, if required, the equations described previously

may be regarded as equations of the initial problem without any preliminary approximation and the algorithms, formulated below, can also be related directly to this problem. However, the description of this algorithm should often be regarded as formal and the theoretical substantiation is considerably more difficult in this case.

3.1. The method of component splitting (the fractional step methods)

3.1.1. The splitting method based on implicit schemes of the first order of accuracy

We examine problem (43), where

$$A = \sum_{\alpha=1}^{n} A_\alpha, \qquad A_\alpha \geq 0, \qquad \alpha = 1, \ldots, N$$

(i.e. $\{A_\alpha\}$ are the positive semidefinite operators). Let operator A be independent of time. The splitting algorithm, based on using the implicit schemes of the first order of accuracy in respect of τ has the form

$$\frac{\varphi^{j+1/n} - \varphi^{j}}{\tau} + A_1 \varphi^{j+1/n} = 0,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\frac{\varphi^{j+1} - \varphi^{j+(n-1)/n}}{\tau} + A_n \varphi^{j+1} = f^{j}, \quad j = 0, 1, \ldots; \tag{44}$$

$$\varphi^{0} = g.$$

The scheme (44) is stable and has the first order of accuracy in respect of τ, and the following estimate is valid:

$$\left\| \varphi^{j+1} \right\| \leq \|g\| + j\tau \max_{j} \left\| f^{j} \right\|.$$

The realisation of (44) is based on the consecutive solution of the equations from (44). If the splitting of the operator A by the sum $\sum_{\alpha=1}^{n} A_\alpha$ is carried out in such a manner that the inversion of the operators $(E + A_\alpha)$ is simple (for example, A_α and $(E + A_\alpha)$ are tridiagonal or triangular matrices), then it is easy to find φ^{j+1} – the approximate solution of the problem corresponding to $t = t_{j+1}$.

The algorithm (44) permits the evident generalisation if A depends on time. In this case, in the cycle of calculations using the splitting scheme instead of A we must know the suitable difference approximation Λ^j of this operator in each interval $t_j \leq t \leq t_{j+1}$.

3.1.2. The method of component splitting based on the Cranck-Nicholson schemes

Let it be that $f \equiv 0$ in (43), and operator A has the form $A = \sum_{\alpha=1}^{n} A_\alpha$, $A \geq 0$, $n \geq 2$. In the interval $t_j \leq t \leq t_{j+1}$ we introduce the approximation Λ^j, Λ_α^j of the operators A, A_α so that $\Lambda^j = \sum_{\alpha=1}^{n} \Lambda_\alpha^j$, $\Lambda_\alpha^j \geq 0$. The multi-component splitting

scheme, constructed on the basis of the elementary Cranck–Nicholson schemes is represented by a system of equations

$$\left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+\alpha/n}=\left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+(\alpha-1)/n},$$
$$\alpha=1,2,\ldots,n;\quad j=0,1,\ldots; \tag{45}$$
$$\varphi^0=g.$$

When Λ_α^j are commutative and $\Lambda_\alpha^j = A_\alpha(t_{j+1/2})$, or $\Lambda_\alpha^j = (A(t_{j+1}) + A(t_j))/2$, the given system is absolutely stable and has the second order of approximation in respect of τ. For non-commutative operators Λ_α^j the scheme (45) generally speaking will be the scheme of the first order with accuracy in respect of τ.

The concept of confirmation of the formulated claims is typical of many splitting methods and consists of the following.

It should be mentioned that the system of equations (45) is reduced to a single equation of the type

$$\varphi^{j+1}=\prod_{\alpha=1}^{n}\left(E+\frac{\tau}{2}\Lambda_\alpha\right)^{-1}\left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^j,$$

from which on the basis of theorem 11 we have $\|\varphi^{j+1}\|\le\|\varphi^j\|\le\ldots\le\|g\|$. If operators Λ_α^j are *skew-symmetric*, $(\Lambda_\alpha^j\varphi,\varphi)=0$, then $\|\varphi^{j+1}\|=\|\varphi^j\|=\ldots=\|g\|$. Thus, scheme (45) is absolutely stable.

To determine the approximation order, we expand the expression τ (assuming that $(\tau/2)\|\Lambda_\alpha\|<1$)

$$T^j=\prod_{\alpha=1}^{n}\left(E+\frac{\tau}{2}\Lambda_\alpha^j\right)^{-1}\left(E-\frac{\tau}{2}\Lambda_\alpha^j\right)$$

in respect of the degrees of the small parameter τ. Since $T^j=\Pi_{\alpha=1}^n T_\alpha^j$, initially we expand operators T_α^j into a series:

$$T_\alpha^j=E-\tau\Lambda_\alpha^j+\left(\frac{\tau^2}{2}\right)(\Lambda_\alpha^j)^2+\ldots$$

Consequently, we obtain

$$T^j=E-\tau\Lambda^j+\frac{\tau^2}{2}\left[\left(\Lambda^j\right)^2+\sum_{\alpha=1}^{n}\sum_{\beta=\alpha+1}^{n}(\Lambda_\alpha^j\Lambda_\beta^j-\Lambda_\beta^j\Lambda_\alpha^j)\right]+O(\tau^3).$$

When the operators Λ_α^j are commutative, the expression below the sign of the double sum converts to zero. Therefore

$$T^j=E-\tau\Lambda^j+\frac{\tau^2}{2}(\Lambda^j)^2+O(\tau^3).$$

Comparing this expansion with the expansion of the operator

$$T^j=\left(E+\left(\frac{\tau}{2}\right)\Lambda^j\right)^{-1}\left(E-\left(\frac{\tau}{2}\right)\Lambda^j\right)$$

into a series in respect of the degrees $\tau\Lambda^j$, where Λ^j is determined by means of one

of the following relationships: $\Lambda^j = A(t_j) + (\tau/2)\, A'(t_j)$; $\Lambda^j = A((t_j + t_{j+1})/2)$; $\Lambda^j = (1/2)(A(t_{j+1}) + A(t_j))$, we confirm that the scheme (45) has the second order of approximation in respect of τ. If the operators Λ_α^j are non-commutative, the splitting scheme has only the first order of accuracy in respect of τ.

3.2. Methods of two-cyclic multi-component splitting

3.2.1. The method of two-cyclic multi-component splitting

We examine problem (43) where

$$A(t) = \sum_{\alpha=1}^{n} A_\alpha(t),\ A_\alpha(t) \ge 0.$$

We approximate $A_\alpha(t)$ not on $t_j \le t \le t_{j+1}$, but on a larger interval $t_{j-1} \le t \le t_{j+1}$. We set $\Lambda_\alpha^j = A_\alpha(t_j)$ and in this case there is no requirement on the commutation of the operators $\{\Lambda_\alpha\}$.

Other examples of selecting Λ_α^j are the following approximations:

$$\Lambda_\alpha^j = A_\alpha(t_j) + \frac{\tau}{2}\frac{\partial}{\partial t} A_\alpha(t_j), \qquad \Lambda_\alpha^j = \frac{1}{2}(A_\alpha(t_{j+1}) + A_\alpha(t_j)),$$

having the second order of approximation.

For the problem (43) in the interval $t_{j-1} \le t \le t_{j+1}$ the scheme of two-cyclic multi-component splitting has the form

$$\begin{gathered}
\left(E + \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-(n-1)/n} = \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-1}, \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\
\left(E + \frac{\tau}{2}\Lambda_n^j\right)\left(\varphi^j - \tau f^j\right) = \left(E - \frac{\tau}{2}\Lambda_n^j\right)\varphi^{j-1/n}, \\
\left(E + \frac{\tau}{2}\Lambda_n^j\right)\varphi^{j+1/n} = \left(E - \frac{\tau}{2}\Lambda_n^j\right)\left(\varphi^j + \tau f^j\right), \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\
\left(E + \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+1} = \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+(n-1)/n},
\end{gathered} \tag{46}$$

where

$$\Lambda_\alpha^j = A_\alpha(t_j).$$

Assuming the required smoothness, this scheme has the second order of approximation in respect of τ and is absolutely stable.

The multi-component system of equations (46) can be written in the equivalent form:

$$\left(E+\frac{\tau}{2}\Lambda_\alpha\right)\varphi^{j-(n+1-\alpha)/(n+1)}=\left(E-\frac{\tau}{2}\Lambda_\alpha\right)\varphi^{j-(n+1-\alpha+1)/(n+1)},$$

$$\alpha=1,2,\ldots,n,$$

$$\varphi^{j+1/(n+1)}=\varphi^{j-1/(n+1)}+2\tau f^{j},$$

$$\left(E+\frac{\tau}{2}\Lambda_{n-\alpha+2}\right)\varphi^{j+\alpha/(n+1)}=\left(E-\frac{\tau}{2}\Lambda_{n-\alpha+2}\right)\varphi^{j+(\alpha-1)/(n+1)},$$

$$\alpha=2,3,\ldots,n+1,$$

which may be preferred to the form (46).

3.2.2. Method of two-cyclic component splitting for quasi-linear problems

We examine an evolution problem with operator A which depends on time and on the solution of the problem

$$\frac{\partial\varphi}{\partial t}+A(t,\varphi)\varphi=0 \quad \text{in} \quad \Omega\times\Omega_t,$$

$$\varphi=g \quad \text{in} \quad \Omega \quad \text{at} \quad t=0.$$

As regards the operator $A(t,\varphi)$ we assume that it is non-negative, has the form

$$A(t,\varphi)=\sum_{\alpha=1}^{n}A_\alpha(t,\varphi), \qquad A_\alpha(t,\varphi)\geq 0,$$

and is sufficiently smooth. We also assume that the solution φ is also a sufficiently smooth function of time. In the interval $t_{j-1}\leq t\leq t_{j+1}$ we examine the splitting scheme

$$\begin{gathered}
\frac{\varphi^{j+1/n-1}-\varphi^{j-1}}{\tau}+A_1^j\,\frac{\varphi^{j+1/n-1}-\varphi^{j-1}}{2}=0,\\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\frac{\varphi^{j}-\varphi^{j-1/n}}{\tau}+A_n^j\,\frac{\varphi^{j}+\varphi^{j-1/n}}{2}=0,\\
\frac{\varphi^{j+1/n}-\varphi^{j}}{\tau}+A_n^j\,\frac{\varphi^{j+1/n}+\varphi^{j}}{2}=0,\\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\frac{\varphi^{j+1}-\varphi^{j+(n-1)/n}}{\tau}+A_1^j\,\frac{\varphi^{j+1}+\varphi^{j+(n-1)/n}}{2}=0,
\end{gathered}\tag{47}$$

where

$$A^j=A(t_j,\tilde{\varphi}^j), \quad \tilde{\varphi}^j=\tilde{\varphi}^{j-1}-\tau A(t_{j+1},\varphi^{j-1})\varphi^{j-1}, \quad \tau=t_j-t_{j-1}.$$

The methods described previously for linear operators which depend on time only may be used to prove that the splitting scheme (47) has the second order of approximation in respect of τ and is absolutely stable. Similarly, the splitting methods can be used for nonhomogeneous quasi-linear equations.

3.3. The splitting method with factorisation of operators

The concept of this group of the method may be described as follows. It is assumed that to solve the problem (43) we use a different scheme in respect of time written in the form

$$B\varphi^j = F^j, \quad j = 1,2,\ldots,$$

where

$$F^j = F^j(\varphi^{j-1}, \varphi^{j-2}, \ldots; \tau)$$

is the known function of τ, $\varphi^{j-1}, \varphi^{j-2}$, … (in the case of the two-layer scheme $F^j = F^j(\tau, \varphi^{j-1})$, and B is some operator constructed in such a manner that it permits (factorised) representation

$$B = B_1 B_2 \ldots B_p.$$

Here the operators B_α are the operators with a simpler structure than operator B, and can be efficiently inverted. The difference schemes permitting this representation of B, are referred to as *factorised schemes* or the *schemes of factorisation of the difference operator*.

The numerical realisation of solution of equation $B\varphi^j = F^j$ in every time layer can be carried out by consecutive solution of simpler equations:

$$B_1\varphi^{j+1/p} = F^j, \qquad B_2\varphi^{j+2/p} = \varphi^{j+1/p}, \qquad B_p\varphi^{j+1} = \varphi^{j+(p-1)1/p}.$$

If, for example, $p = 2$, and B_1, B_2 are triangular matrices, then these equations represent the known scheme of the explicit (running) counting.

3.3.1. The implicit splitting scheme with approximate factorisation of the operator

Let it be that for the problem (43) with the operator $A \geq 0$ and $f = 0$ we examine the implicit scheme of the type

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + \Lambda\varphi^{j+1} = 0, \quad j = 0,1,\ldots; \quad \varphi^0 = g,$$

where $\Lambda = \sum_{\alpha=1}^{n} \Lambda_\alpha$, $\Lambda_\alpha \geq 0$, which is written in the form $(E + \tau\Lambda)\varphi^{j+1} = \varphi^j$. We factorise the operator $(E + \tau\Lambda)$ approximately with the accuracy to the members of the order $O(\tau^2)$. For this purpose, we replace the operator $(E+\tau\Lambda)$ by the factorised operator

$$(E+\tau\Lambda_1)(E+\tau\Lambda_2)\ldots(E+\tau\Lambda_n) = E+\tau\Lambda+\tau^2 R,$$

where

$$R = \sum_{i<j} \Lambda_i\Lambda_j + \tau \sum_{i<j<k} \Lambda_i\Lambda_j\Lambda_k + \ldots + \tau^{n-2}\Lambda_1\ldots\Lambda_n.$$

Consequently, we obtain an implicit scheme with approximate factorisation of the operator

$$B\varphi^{j+1} = \varphi^j, \qquad j = 0,1,\ldots; \qquad \varphi^0 = g,$$

where $B = \Pi_{\alpha=1}^{n} B_\alpha, B_\alpha = E + \tau\Lambda_\alpha$.

These equations can be solved by successive solution of the equations

$$
\begin{gathered}
(E+\tau\Lambda_1)\varphi^{j+1/n}=\varphi^j,\\
(E+\tau\Lambda_1)\varphi^{j+2/n}=\varphi^{j+1/n},\\
\dots\dots\dots\dots\dots\dots\dots\dots\\
(E+\tau\Lambda_n)\varphi^{j+1}=\varphi^{j+(n-1)/n},\\
j=0,1,\dots;\\
\varphi^0=g.
\end{gathered}
\tag{48}
$$

If the operators Λ_α in the representation $\Lambda=\sum_{\alpha=1}^{n}\Lambda_\alpha$ have a simple structure and permit efficient conversion, it is also quite simple to solve also the system (48).

The scheme (48) has the first order of approximation in respect of τ. It is absolutely stable for $\Lambda_\alpha\geq 0$ in the metrics of the space of grid Φ_h because in this case $\|(E+\tau\Lambda_\alpha)^{-1}\|_{\Phi_h}\leq 1$.

3.3.2. The stabilisation method (the explicit–implicit schemes with approximate factorisation of the operator)

We examine a homogeneous problem (43) where $A=A_1+A_2$, $A\geq 0$, $f\equiv 0$, and the following two-layer scheme

$$
\left(E+\frac{\tau}{2}A_1\right)\left(E+\frac{\tau}{2}A_2\right)\frac{\varphi^{j+1}-\varphi^j}{\tau}+A\varphi^j=0,\quad j=0,1,\dots;\ \ \varphi^0=g. \tag{49}
$$

If $A_1\geq 0$, $A_2\geq 0$, then at sufficient smoothness of solution of the problem (43) this scheme has the second order of approximation in respect of τ and is absolutely stable, and the following estimate holds

$$
\left\|\varphi^{j+1}\right\|_{C_2}\leq\|g\|_{C_2},\quad j=0,1,\dots,
$$

where

$$
\|\psi\|_{C_2}=(C_2\psi,\psi)_\Phi^{1/2},\qquad C_2=\left(E+\frac{\tau}{2}A_2^*\right)\left(E+\frac{\tau}{2}A_2\right)
$$

The difference scheme (49) permits efficient realisation in a computer:

$$
F^j=A\varphi^j,\quad \left(E+\frac{\tau}{2}A_1\right)\xi^{j+1/2}=-F^j,
$$

$$
\left(E+\frac{\tau}{2}A_2\right)\xi^{j+1}=\xi^{j+1/2},\quad \varphi^{j+1}=\varphi^j+\tau\xi^{j+1}.
$$

Here $\xi^{j+1/2}$ and ξ^{j+1} are some auxiliary vectors reducing the problem (49) to a sequence of the simplest problems, the first and last equations of which are explicit relationships. This means that the operator should be inverted in the second and third equations which contain only the simplest operators A_1 and A_2.

We examine a heterogeneous problem (43), where $A=A_1+A_2$, $A_1\geq 0$, $A_2\geq 0$, $f\equiv 0$. The scheme of the stabilisation method is written in the form:

$$\left(E+\frac{\tau}{2}A_1\right)\left(E+\frac{\tau}{2}A_2\right)\frac{\varphi^{j+1}-\varphi^j}{\tau}+A\varphi^j=f^j,\quad \varphi^0=g,$$

where $f^j = f(t_{j+1/2})$. If the elements of the matrices A are independent of time, then at a sufficient smoothness of the solution and the function f of the problem (43) the given scheme is absolutely stable and approximate the initial problem with the second order of accuracy in respect of τ.

Comment. The scheme of the stabilisation method is obtained by introducing into the Cranck–Nicholson scheme (i.e. into the explicit–implicit scheme) of the operator

$$B=E+\left(\frac{\tau^2}{4}\right)A_1A_2,$$

and the operator of the problem then becomes factorisable. Therefore, the stabilisation scheme may also be referred to as the *explicit–implicit scheme with approximate factorisation of the operator*.

We formulate the stabilisation scheme for a nonhomogeneous problem (43) in which

$$A=\sum_{\alpha=1}^{n}A_\alpha,\quad n>2,\quad A_\alpha\geq 0,$$

and the operators A_α are independent of t. Under these assumptions, the stabilisation method may be represented in the form

$$\prod_{\alpha=1}^{n}\left(E+\frac{\tau}{2}A_\alpha\right)\frac{\varphi^{j+1}-\varphi^j}{\tau}+A\varphi^j=f^j,\quad \varphi^0=g, \tag{50}$$

where $f^j = f(t_{j+1/2})$. The realisation scheme of the algorithm has the following form:

$$F^j=-A\varphi^j+f^j,$$

$$\left(E+\frac{\tau}{2}A_1\right)\xi^{j+1/n}=F^j,$$

$$\left(E+\frac{\tau}{2}A_2\right)\xi^{j+2/n}=\xi^{j+1/n},$$

.....

$$\left(E+\frac{\tau}{2}A_n\right)\xi^{j+1}=\xi^{j+(n-1)/n},$$

$$\varphi^{j+1}=\varphi^j+\tau\xi^{j+1}.$$

At a sufficient smoothness the stabilisation method (43) has the second order approximation in respect of τ. The stability is ensured if the condition is satisfied: $||T|| \leq 1$, where T is the operator of the step determined by the equation

$$T=E-\tau\prod_{\alpha=n}^{1}\left(E+\frac{\tau}{2}A_\alpha\right)^{-1}A.$$

It should be noted that here the condition $A_\alpha \geq 0$ does not indicate stability in any form, as observed in the case $n = 2$. Therefore, the condition is $||T|| \leq 1$ is regarded here as an additional condition for ensuring the stability of the scheme (50).

Comment. N.N. Yanenko also obtained schemes of approximate factorisation, formed from multilayer systems

$$A_1\varphi^{j+1} + A_0\varphi^j + A_{-1}\varphi^{j-1} + \ldots + A_{-p+1}\varphi^{j-p+1} + f^j = 0,$$

which may form in the approximation of the equation of the type

$$B_1\frac{\partial\varphi}{\partial t} + B_2\frac{\partial^2\varphi}{\partial t^2} + \ldots + B_p\frac{\partial^P\varphi}{\partial t^P} + A\varphi = f,$$

where $B_1,\ldots B_2,\ldots\ldots,B_p$ are linear operators.

3.4. The predictor–corrector method

Another class of splitting methods – the predictor–corrector method (the scheme of approximation correction), like the scheme with factorisation of the operator, will be examined in application to matrix evolution equation (43) with the operator independent of t.

3.4.1. The predictor–corrector method. The case $A = A_1+A_2$.

The concept of the predictor–corrector method maybe described as follows. The entire interval $0 \leq t \leq T$ is divided into partial intervals and within each elementary interval $t_j \leq t \leq t_{j+1}$ problem (43) is solved in two approaches. Initially, using the scheme of the first order of accuracy and a relatively large 'reserve' of stability, we determine the approximate solution of the problem at the moment of time $t_{j+1/2}=t_j+\tau/2$ – this stage is usually referred to as a *predictor*. Subsequently, in the entire interval (t_j, t_{j+1}) we write the initial equation with the second order of approximation used as the *corrector*.

For a nonhomogeneous problem, the predictor–corrector method has the form

$$\begin{aligned}
&\frac{\varphi^{j+1/4} - \varphi^j}{\tau/2} + A_1\varphi^{j+1/4} = f^j,\\
&\frac{\varphi^{j+1/2} - \varphi^{j+1/4}}{\tau/2} + A_2\varphi^{j+1/2} = 0,\\
&\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^{j+1/2} = f^j,
\end{aligned} \tag{51}$$

where $f^j = f(t_{j+1/2})$. With this selection of f^j, the scheme approximates the initial problem with the second order in respect of τ and the following estimate holds

$$\left\|\varphi^j + \frac{\tau}{2}f^j\right\|_{C_1^{-1}} \leq \left\|g + \frac{\tau}{2}f^0\right\|_{C_1^{-1}} + \tau j\|f\|_{C_1^{-1}},$$

where

$$\|f\|_{C_1^{-1}} = \max\|f^j\|_{C_1^{-1}}, \quad C_1^{-1} = \left(E + \frac{\tau}{2}A_1^*\right)^{-1}\left(E + \frac{\tau}{2}A_1\right)^{-1},$$

$$\|\varphi\|_{C_1^{-1}} = (C_1^{-1}\varphi, \varphi)_\Phi^{1/2},$$

i.e. at $0 \le t_j \le T$ we again have the stability of the difference scheme. Thus, if $A_1 \ge 0$, $A_2 \ge 0$ and the elements of the matrix A_1, A_2 are independent of time, then at a sufficient smoothness of the solution and the right-hand side f of the problem (43), the difference scheme (51) is absolutely stable and enables us to obtain the solution of the second order of accuracy in respect of τ.

3.4.2. The predictor–corrector method. Case $A = \sum_{\alpha=1}^{n} A_\alpha$.

Let in (43) be $A = \sum_{\alpha=1}^{n} A_\alpha, A_\alpha \ge 0, n > 2.$ The scheme of the predictor–corrector method in this case has the form

$$\left(E + \frac{\tau}{2}A_1\right)\varphi^{j+1/(2n)} = \varphi^j + \frac{\tau}{2}f^j,$$

$$\left(E + \frac{\tau}{2}A_2\right)\varphi^{j+2/(2n)} = \varphi^{j+1/(2n)},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\left(E + \frac{\tau}{2}A_n\right)\varphi^{j+1/2} = \varphi^{j+(n-1)/(2n)}, \tag{52}$$

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^{j+1/2} = f^j,$$

where it is again assumed that $A_\alpha > 0$ and $f^j = f(t_{j+1/2})$. The system of equations (52) is reduced to a single equation of the type

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\prod_{\alpha=n}^{1}\left(E + \frac{\tau}{2}A_\alpha\right)^{-1}\left(\varphi^j + \frac{\tau}{2}f^j\right) = f^j, \quad \varphi^0 = g.$$

It follows from here that the predictor–corrector method at a sufficiently smooth solution has the second order of approximation in respect of τ.

The last equation maybe written in the form

$$\varphi^{j+1} = T\varphi^j + \frac{\tau}{2}(E + T)f^j,$$

where T is the step operator:

$$T = E - \tau A\prod_{\alpha=n}^{1}\varphi^{j+1}\left(E + \frac{\tau}{2}A_\alpha\right)^{-1}.$$

The requirement of countable stability in the final analysis is reduced to estimating the norm of the operator T. Unfortunately, again in this case the constructive condition $A_\alpha \ge 0$ does not make it possible to prove the stability of the scheme. When the operators A_α commute with each other and have the common basis, the condition $A_\alpha \ge 0$ indicates the stability of the investigated scheme.

3.5. The alternating-direction method and the method of the stabilising correction

3.5.1. The alternating-direction method

We examine the matrix homogeneous evolution problem (43) (i.e. at f=0), where $A=A_1+A_2$. The scheme of the method of alternating directions for the given problems has the form

$$\frac{\varphi^{j+1/2}-\varphi^{j}}{\tau}+\frac{1}{2}(A_1\varphi^{j+1/2}+A_2\varphi^{j})=0,$$
$$\frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau}+\frac{1}{2}(A_1\varphi^{j+1/2}+A_2\varphi^{j+1})=0, \tag{53}$$
$$j=0,1,2\ldots,\qquad \varphi^{0}=g.$$

In this form, the scheme was proposed by Peaceman, Rachford and Douglas when used for a parabolic problem with two spatial variables. The operator A_α is the difference approximation of the one-dimensional differential operator $-a^2\partial^2/\partial x_\alpha^2$. It should be mentioned that in this case the scheme (53) is symmetric, i.e. x_1 and x_2 in the scheme change the roles from the first fractional step to the second one (this also determines the name of the method). The solution of each equation in the parabolic problem is easily realised by the sweep method and, therefore, the scheme (53) is also referred to as the scheme of *longitudinal–transverse sweep*. If $\varphi^{j+1/2}$ is excluded from (53) we obtain

$$\frac{\varphi^{j+1}-\varphi^{j}}{\tau}+A\left(\frac{\varphi^{j+1}+\varphi^{j}}{2}\right)+\frac{\tau^2}{4}A_1A_2\left(\frac{\varphi^{j+1}-\varphi^{j}}{\tau}\right)=0.$$

Comparing this equation with the Cranck–Nicholson scheme, we conclude that it has a second order of approximation in respect of τ. Further, if we examine (53) where A_α is the three-point approximation of the operator $-a^2\partial^2/\partial x_\alpha^2 : A_\alpha=-a^2(\Delta_{x\alpha}\nabla_{x\alpha})/h_\alpha^2$, it may easily be established that the given scheme is absolutely stable. However, Yanenko showed that the scheme of the method of alternating directions is not suitable for the three-dimensional parabolic problem. Thus, it appears that in this case the scheme (at $A_\alpha=-a^2(\Delta_{x\alpha}\nabla_{x\alpha})/h_\alpha^2$)

$$\frac{\varphi^{j+1/3}-\varphi^{j}}{\tau}+\frac{1}{3}(A_1\varphi^{j+1/3}+A_2\varphi^{j}+A_3\varphi^{j})=0,$$
$$\frac{\varphi^{j+2/3}-\varphi^{j+1/3}}{\tau}+\frac{1}{3}(A_1\varphi^{j+1/3}+A_2\varphi^{j+2/3}+A_3\varphi^{j+1/3})=0,$$
$$\frac{\varphi^{j+1}-\varphi^{j+2/3}}{\tau}+\frac{1}{3}(A_1\varphi^{j+2/3}+A_2\varphi^{j+2/3}+A_3\varphi^{j+1})=0$$

is not absolutely stable. Therefore, in many problems it is preferred to use the schemes of the methods of stabilising correction (which together with the scheme of alternating directions are sometimes also referred to as *implicit schemes of alternating directions*).

3.5.2. The method of stabilising correction

If in (43) $A=A_1+A_2+A_3$, $f = 0$, the scheme of the stabilising correction of solving the three-dimensional heat conductivity equation has the form

$$\frac{\varphi^{j+1/3}-\varphi^{j}}{\tau}+A_1\varphi^{j+1/3}+A_2\varphi^{j}+A_3\varphi^{j}=0,$$

$$\frac{\varphi^{j+2/3}-\varphi^{j+1/3}}{\tau}+A_2(\varphi^{j+2/3}-\varphi^{j})=0,$$

$$\frac{\varphi^{j+1}-\varphi^{j+2/3}}{\tau}+A_3(\varphi^{j+1}-\varphi^{j})=0,$$

$$j=0,1,2,\ldots,\qquad \varphi^{0}=g.$$

Excluding $\varphi^{j+1/3}$, $\varphi^{j+2/3}$, we obtain the equation

$$\frac{\varphi^{j+1}-\varphi^{j}}{\tau}+A\varphi^{j+1}+\tau^2(A_1A_2+A_1A_3+A_2A_3)\left(\frac{\varphi^{j+1}-\varphi^{j}}{\tau}\right)+$$

$$+\tau^3A_1A_2A_3\left(\frac{\varphi^{j+1}-\varphi^{j}}{\tau}\right)=0.$$

It follows from here that this scheme has the first order of accuracy in respect of τ. Examining the scheme for the heat conductivity equation, we can easily see its absolute stability. In addition to this, the structure of the scheme is as follows: the first fractional step gives the complete approximation of the heat conductivity equation, the next fractional steps are correcting and are used for improving stability. Therefore, these schemes are referred to as the *schemes of stabilising correction* or *the schemes with a correction for stability*.

Comment. The alternating-direction methods, proposed by Peaceman, Douglas and Rachford are usually linked with the division of the operator A into one-dimensional operator A_α, and the sweep method is used in every fractional step for solving the equation. However, the application of the method to problems with three special variables encounters difficulties. On the other hand, the requirement on the one-dimensional form of the operators A_α may be ignored. Therefore, it is interesting to split the operator A into operators which would make it possible to realise efficiently the solution of the problem in its step and retain the main advantages of the alternating-direction method. Such a split includes the division of the matrix operator $A = A^*$ into two matrices A_1 and A_2 such that $A_1 = A_2^*$, $A_1+A_2=A$. If, after this we formally write the alternating-direction scheme, we obtain the schemes of the alternating–triangular method proposed and justified for a number of problems of mathematical physics by Samarskii and Il'in. In his studies, Il'in also proposed schemes generalising the alternating-direction method, in which A_1, A_2 are already arbitrary matrices (in particular, triangular) which are also referred to as the schemes of the *alternating operator method*.

3.6. Weak approximation method

3.6.1. The main system of problems

In some Hilbert space Φ we examine an abstract Cauchy problem (43), where $A = A(t)$ as a linear operator with the domain of definition dense in Φ and the range of values from Φ. In addition to this, the operator A is presented in the form of a sum $A = \sum_{i=1}^{n} A_i$ of linear operators $A_i(t)$, having the same domain of definition as A, and also $f = \sum_{i=1}^{n} f_i$. This problem is replaced by the following system

$$\begin{aligned}
&\frac{d\varphi_1}{dt} + A_1\varphi_1 = f_1(t), \quad t \in \theta_j \equiv (t_j, t_{j+1}], \qquad \phi_1(t_j) = v(t_j),\\
&\frac{d\varphi_2}{dt} + A_2\varphi_2 = f_2(t), \quad t \in \theta_j, \qquad \varphi_2(t_j) = \varphi_1(t_{j+1}),\\
&\qquad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&\frac{d\varphi_n}{dt} + A_n\varphi_n = f_n(t), \quad t \in \theta_j, \qquad \varphi_n(t_j) = \varphi_n(t_{j+1})
\end{aligned} \tag{54}$$

and we set

$$v(t_{j+1}) = \varphi_n(t_{j+1}), \quad j = 0, \dots, n-1, \quad v(0) = g.$$

Consequently, the process of determination of the approximate solution $v(t)$ of the initial problem in respect of time in the interval $[t_j, t_{j+1}]$ is reduced to a successive solution of each of equations (54) for $\varphi_i, i = \overline{1,n}$, in the given interval $[t_j, t_{j+1}]$. For operators commutative in pairs

$$A_i(t'), A_j(t''), \quad i, j = 1, \dots, n, \quad i \neq j, \quad t', t'' \in [0, T]$$

on the condition for the exact the solution of the problem (43) of the type $||A_i A_j \varphi|| \leq M < \infty$, we can establish that the problem (54) approximates (43) in the *overall sense*, i.e.

$$\left\| \sum_{i=1}^{n} \psi_i \right\| = 0(\tau),$$

where

$$\psi_i = f_i(t) - A_i(t)\varphi(t_{j+1}), \quad i > 1, \quad \psi_1 = f_1(t) - A_1(t)\varphi(t) - \frac{d\varphi}{dt}.$$

The system of problems (54) is the principle of the *method of weak approximation* of the problem (43).

3.6.2. Two-cyclic method of weak approximation

The problem (43) can be split into a system of Cauchy's problems, approximated with the order $O(\tau^2)$ of the problem (43), using the two-cyclic procedure. We present one of the possible schemes. In the interval $t_{j-1} < t \leq t_j$ we have

$$\frac{d\varphi_i}{dt} + A_i\varphi_i = 0, \quad i = 1,\ldots,n-1,$$
$$\frac{d\varphi_n}{dt} + A_n\varphi_n = f + \frac{\tau}{2}A_n f, \tag{55}$$

and in the interval $t_j < t \le t_{j+1}$

$$\frac{d\varphi_{n+1}}{dt} + A_n\varphi_{n+1} = f - \frac{\tau}{2}A_n f,$$
$$\frac{d\varphi_{n+i}}{dt} + A_{n-i+1}\varphi_{n+i} = 0, \quad i = 2,\ldots,n, \tag{56}$$

on the condition

$$\varphi_1(t_{j-1}) = v(t_{j-1}), \quad \varphi_{i+1}(t_{j-1}) = \varphi_i(t_j), \quad i = 1,\ldots,n-1,$$
$$\varphi_{n+1}(t_j) = \varphi_n(t_j), \quad \varphi_{k+1}(t_j) = \varphi_k(t_{j+1}), \quad k = n+1,\ldots,2n-1. \tag{57}$$

In this case, the approximation of the problem (43) using (55)–(57) is examined in the double interval $[t_{j-1}, t_{j+1}]$ and it is assumed that $v(t_{j+1}) = \varphi_{2n}(t_{j+1})$.

3.7. The splitting methods – iteration methods of solving stationary problems

3.7.1. The general concepts of the theory of iteration methods

The solution of many stationary problems with positive operators maybe regarded as a limiting, at $t\to\infty$, solution of the non-stationary problem. When solving stationary problems by the methods of asymptotic stationarisation, no attention is given to the intermediate values of the solution because they are of no interest, whereas when solving the non-stationary problems these intermediate values have physical meaning.

Let us assume that we have a system of linear algebraic equations (which are, for example, the result of approximation, by the finite-difference method, of some stationary mathematical physics problem):

$$A\varphi = f,$$

where $\varphi\in\Phi$, $f\in F$.

We also examine a non-stationary problem

$$\frac{d\psi}{dt} + A\psi = f, \quad \psi(0)=0.$$

Assuming that $A \equiv A^T > 0$, it was proven previously that $\lim_{t\to\infty}\psi = \varphi$. (If the operator of the stationary problem has the spectrum with an arbitrary structure, then in this case there cannot be such a simple and transparent relationship between the solutions ψ,φ).

The nonstationary problem for the new function ψ can be solved by a difference method in respect of t, for example

$$\frac{\psi^{j+1} - \psi^j}{\tau} + A\psi^j = f.$$

Consequently

$$\psi^{j+1} = \psi^j - \tau(A\psi^j - f).$$

If our task is to solve the stationary problem, then in the determination of the relationship between τ and $\beta(A)$, where $\beta \equiv \beta(A)$ is the maximum eigenvalue of A, we have $\lim_{t\to\infty}\psi^j = \varphi$. The parameter τ can be both dependent and independent of the number j which when solving the stationary problem should be regarded as the number of the step of the iteration process.

There is another special feature: in non-stationary problems to ensure the accuracy of solution, the values of τ should be sufficiently small, and in the stationary problems the optimum iteration parameters τ are selected from the condition of minimisation of the number of iterations and may have relatively large values.

The majority of iteration methods, used for solving linear systems, maybe combined by the equation

$$B_j \frac{\varphi^{j+1} - \varphi^j}{\tau_j} = -\alpha(A\varphi^j - f), \tag{58}$$

Where α is some positive number, $\{B_j\}$ is the sequence of non-degenerate matrices, and $\{\tau_j\}$ is the sequence of real parameters.

3.7.2. Iteration algorithms

For an efficient realization of the method (58), operator B_j should have a simpler structure in comparison with A. In many cases that are of interest for practice operator B_j has the form

$$B_j = \prod_{i=1}^{m}(E + \tau_j \tilde{B}_i),$$

where $\tilde{B}_i$ are some matrices of the same order N as the matrix A. These matrices are selected in such a manner as to ensure that the matrices $(E + \tau_j \tilde{B}_i)$ are easily inverted, i.e. the inversion of the entire matrix B_j is simpler than the conversion of the matrix A. In many cases, $\{\tilde{B}_i\}$ is selected taking into account

$$A = \sum_{k=1}^{n} A_k.$$

We initially set $n = m = 2$, $\tau_j = \tau$ and, returning to the splitting and alternating direction methods, described in the previous sections, we present the appropriate iteration algorithms of solution of the system $A\varphi = f$.

The alternating-direction method. At $\alpha = 2$, $\tilde{B}_i = A_i$ from (58) we have the algorithm of the type

$$(E + \tau_j A_1)(E + \tau_j A_2)\frac{\varphi^{j+1} - \varphi^j}{\tau_j} = -2(A\varphi_j - f),$$

which after simple transformation may also be written in ‘fractional steps’:

$$\frac{\varphi^{j+1/2}-\varphi^{j}}{\tau_j}+(A_1\varphi^{j+1/2}+A_2\varphi^{j})=f,$$

$$\frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau_j}+(A_1\varphi^{j+1/2}+A_2\varphi^{j+1})=f.$$

The stabilising correction method is obtained at $\alpha=1$, $\widetilde{B}_i=A_i$:

$$(E+\tau_j A_1)(E+\tau_j A_2)\frac{\varphi^{j+1}-\varphi^{j}}{\tau_j}=-(A\varphi^{j}-f).$$

It can be written in the form

$$\frac{\varphi^{j+1/2}-\varphi^{j}}{\tau_j}=-(A_1\varphi^{j+1/2}+A_2\varphi^{j})=f,\quad \frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau_j}+A_2(\varphi^{j+1}-\varphi^{j})=0.$$

The splitting method (the fractional steps method) for an arbitrary value $m = n \geq 2$ may be presented in the following form:

$$\frac{\varphi^{j+1/n}-\varphi^{j}}{\tau_j}+A_1(\varphi^{j+1/n}-\varphi^{j})=-\alpha_j\left(\sum_{k=1}^{n}A_k\varphi^{j}-f\right),$$

$$\frac{\varphi^{j+k/n}-\varphi^{j+(k-1)/n}}{\tau_j}+A_k(\varphi^{j+1/n}-\varphi^{j})=0,\quad k=2,\ldots,n,$$

or, similarly,

$$B_j\frac{\varphi^{j+1}-\varphi^{j}}{\tau_j}=-\alpha_j(A\varphi^{j}-f),$$

where $B_j=\Pi_{k=1}^{n}(E+\tau_j A_k)$ and τ_j,α_j are some iteration parameters. Other iteration algorithms may also be formulated in the same form.

The convergence of the formulated iteration algorithms is accelerated either by special selection of the parameters τ_j,α_j, or by using some accelerating procedure for these algorithms.

4. SPLITTING METHODS FOR APPLIED PROBLEMS OF MATHEMATICAL PHYSICS

The splitting methods are used widely for the numerical solution of different problems of mathematical physics. Two paths may be followed in this case. The first of them is based on approximating the initial problem in respect of spatial variables in the first stage of solution of the problem followed by the application of the splitting methods of approximating the problems in respect of the time variable, i.e. the splitting method is used for the system of ordinary differential equations.

In the second approach, the splitting method is eventually used for reducing the initial problem to a system of sub-problems with simpler operators which are then solved by standard numerical methods.

The advantages of the first method used for solving the problems by the

splitting methods is that there are no problems with the boundary conditions for sub-problems in the 'fractional steps'. However, the difficulties with the construction of suitable approximations in respect of spatial variables remain and become considerably greater in multidimensional problems.

The second approach is used widely at the present time for solving complex problems of mathematical physics. Since in this approach the operators of boundary-value sub-problems in the small (intermediate) step have a simpler structure, the construction of numerical approximations is far simpler and is often carried out by the well developed numerical methods. However, in this approach one of the difficulties is associated with selecting the boundary conditions for 'intermediate sub-problems'. It should be mentioned that the problems of selecting the boundary conditions for the problem with fractional steps and correct approximation of the boundary conditions from the initial formulation are generally typical of the splitting methods.

Below, we present some of the splitting methods for heat conduction, Navier–Stokes and shallow water equations, and also in the numerical modelling of sea and ocean flows.

4.1. Splitting methods of heat conduction equations

4.1.1. The fractional step method

We examine a three-dimensional problem for the heat conduction equation

$$\frac{\partial\varphi}{\partial t}-\Delta\varphi=0 \ \text{ in } \ \Omega\times\Omega_t,$$
$$\varphi=0 \ \text{ on } \ \partial\Omega, \quad \varphi=g \ \text{ in } \ \Omega \ \text{ at } \ t=0, \tag{59}$$

where $\Omega=\{(x,y,z)\colon 0< x,y,z < 1\}$, $\Omega_t = (0,T)$, g is a given function, $\Delta = \partial^2/\partial x^2+\partial^2/\partial y^2+\partial^2/\partial z^2$. After approximating (59) by the method of finite differences in respect of the variables x,y,z and taking into account the given boundary conditions (59), we obtain a matrix evolution problem

$$\frac{\partial\varphi}{\partial t}+A\varphi=0 \ \text{ in } \ \Omega_t,\ \varphi=g \ \text{ at } \ t=0$$

where (see also an example in paragraph 2.3.1)

$$A=A_x+A_y+A_z,$$

$A_x=-(\Delta_x\nabla_x)\equiv\Lambda_1$, $A_y=-(\Delta_y\nabla_y)\equiv\Lambda_2$, $A_z=-(\Delta_z\nabla_z)\equiv\Lambda_3$, and g, φ are the vectors, and in formation of $\varphi=\varphi(t)$ we take into account the given boundary value conditions. Operator A is examined in the Hilbert space $\Phi=F$ with the norm

$$\|\varphi\|_\Phi=\left(\sum_{k=1}^{N_x-1}\sum_{l=1}^{N_y-1}\sum_{p=1}^{N_z-1}h_xh_yh_z\varphi_{klp}^2\right)^{1/2}.$$

The scheme (44) when used for problem (59) has the form

$$\frac{\varphi^{j+1/3}-\varphi^{j}}{\tau}+\Lambda_1\varphi^{j+1/3}=0,$$
$$\frac{\varphi^{j+2/3}-\varphi^{j+1/3}}{\tau}+\Lambda_2\varphi^{j+2/3}=0, \tag{60}$$
$$\frac{\varphi^{j+1}-\varphi^{j+2/3}}{\tau}+\Lambda_3\varphi^{j+1}=0.$$

Each equation in (60) can be solved simply by the factorisation method. The scheme (60) is absolutely stable and has the first order of approximation in respect of τ and, therefore, the corresponding convergence theorem is valid in this case.

4.2.1. Locally one-dimensional schemes

If the operators A_α (or their approximations) in (43) are one-dimensional differential, the corresponding difference scheme is referred to as *locally one-dimensional*.

The theory of locally one-dimensional schemes for a number of differential equations was developed by Samarskii. We formulate such a scheme for a problem for the heat conduction equation

$$\frac{\partial\varphi}{\partial t}+A\varphi=f \quad \text{in} \quad \Omega\times\Omega_t,$$
$$\varphi\big|_{\partial\Omega}=\varphi_{(\Gamma)}(x,t), \qquad \varphi=g \quad \text{in} \quad \Omega \quad \text{at} \quad t=0, \tag{61}$$

where $A=\sum_{\alpha=1}^{n}A_\alpha,\ A_\alpha=-\partial^2/\partial x_\alpha^2,\ x=(x_1,\ldots,x_n)\in\Omega=\{0<x_\alpha<1,\alpha=1,\ldots,n\}$. It is assumed that problem (61) has a unique sufficiently smooth solution.

When $t\in\theta_j=\{t_j\le t\le t_{j+1}\}$, instead of (61) we find the sequence of equations

$$\frac{\partial\varphi_\alpha}{\partial t}+A_\alpha\varphi_\alpha=f_\alpha \quad x\in\Omega, \quad t\in\theta_j, \quad \alpha=1,\ldots,n, \tag{62}$$

with the initial conditions

$$\varphi_1(x,0)=g, \quad \varphi_1(x,t_j)=\varphi_n(x,t_j), \quad j=1,2,\ldots,$$
$$\varphi_\alpha(x,t_j)=\varphi_{\alpha-1}(x,t_{j+1}), \quad j=0,1,\ldots, \quad \alpha=2,3,\ldots,n,$$

where it is assumed that $\varphi(x,t_{j+1})=\varphi_n(x,t_{j+1})$. Here f_α are some functions such that $\sum_{\alpha=1}^{n}f_\alpha=f$. The boundary conditions for φ_α are given only for parts $\partial\Omega_\alpha$ of the boundary $\partial\Omega$, consisting of the faces $x_\alpha=0$ and $x_\alpha=1$. In Ω we introduce a uniform grid in respect of each variable with the step h. We determine the difference approximation of the operator A_α:$\Lambda_\alpha=-(\Delta_{x\alpha}\nabla_{x\alpha})/(h^2)$, $\alpha=1,\ldots,n$. Transferring from (62) to different approximations of the problems of spatial variables, where φ_α, f_α are vectors, Λ_α are the matrices, Ω is the grid domain, and $\partial\Omega_\alpha$ is the grid on the bounds $x_\alpha=0$ and $x_\alpha=1$, and carrying out the approximation in respect of t using the two-layer implicit scheme of the first order with accuracy in respect of τ, we obtain the locally one-dimensional scheme

$$\frac{\varphi_\alpha^{j+1}-\varphi_\alpha^{j}}{\tau}+\Lambda_\alpha\varphi_\alpha^{j+1}=f_\alpha(t_{j+1/2}),\quad \alpha=1,\ldots,n,\quad j=0,1,\ldots, \tag{63}$$

with the initial conditions

$$\begin{gathered}\varphi_1^0=g,\quad \varphi_1^j=\varphi_n^j\quad j=1,2,\ldots,\\ \varphi_\alpha^j=\varphi_{\alpha-1}^{j+1},\quad \alpha=2,3,\ldots,n,\quad j=0,1,\ldots,\end{gathered} \tag{64}$$

and the boundary conditions

$$\varphi_\alpha^{j+1}\Big|_{\partial\Omega_\alpha}=\varphi_{(\Gamma)}(t_{j+1}). \tag{65}$$

Each of the problems (63)–(65) at every fixed α is a one-dimensional first boundary-value problem and can be solved by the factorisation method.

To find the approximate values of the solutions of the initial problem in the layer t_{j+1} in respect of the data on the layer t_j it is necessary to solve successively n one-dimensional problems. The locally one-dimensional scheme (63)–(65) is stable in respect of the initial and boundary data and of the right-hand side in the metrics $\|\varphi\|_C=\max_{x_j\in\Omega}|\varphi_j|$, i.e., uniformly. However, if the solution of the problem (61) has a unique continuous, in $\Omega\times\Omega_t$, solution $\varphi=\varphi(x,t)$ and in $\Omega\times\Omega_t$ there exist derivatives

$$\frac{\partial^2\varphi}{\partial t^2},\quad \frac{\partial^4\varphi}{\partial x_\alpha^2\partial x_\beta^2},\quad \frac{\partial^3\varphi}{\partial t\partial x_\alpha^2},\quad \frac{\partial^2 f}{\partial x_\alpha^2},\quad 0\le\alpha,\ \beta\le n,$$

the scheme (63)–(65) uniformly converges at the rate $O(h^2+\tau)$, i.e. it has the first order of accuracy in respect of τ, therefore

$$\left\|\varphi^j-\varphi(t_j)\right\|_C\le M(h^2+\tau),\quad j=1,2,\ldots,$$

where M = const is independent of τ and h.

4.1.3. Alternating-direction schemes

It is required to solve the following boundary-value problems for the heat conduction equation without mixed derivatives:

$$\begin{gathered}\frac{\partial\varphi}{\partial t}+A\varphi=f\quad\text{in}\quad \Omega\times\Omega_t,\\ \varphi\big|_{\partial\Omega}=\varphi_{(\Gamma)}(x,t),\quad \varphi=g(x)\quad\text{in}\quad\Omega\quad\text{at}\quad t=0,\end{gathered} \tag{66}$$

where

$$A=\sum_{\alpha=1}^{2}A_\alpha,\qquad A_\alpha=-\frac{\partial}{\partial x_\alpha}k_\alpha\frac{\partial}{\partial x_\alpha}$$

$$x=(x_1,\ldots,x_2)\in\Omega=\{0<x_\alpha<1,\ \alpha=1,2\},\quad k_\alpha>0.$$

It is assumed that the problem (66) has a unique sufficiently smooth solution.

In $\overline{\Omega}=\Omega\cup\partial\Omega$ we construct the grid $\overline{\Omega}^h$ uniform in respect of x_α with the steps h_α. Operator A_α are replaced by different operators

$$\Lambda_\alpha:\Lambda_\alpha\varphi=-(\Delta_{x_\alpha}k_\alpha^h\nabla x_\alpha\varphi)/(h_\alpha^2).$$

In contrast to the case of constant coefficients, the operators Λ_α are positive and self-adjoint but not commutative. Instead of (66) we examine the problem approximating it

$$\frac{\partial\varphi}{\partial t}+\Lambda\varphi=f \quad \text{in} \quad \Omega^h, \quad \Lambda=\Lambda_1+\Lambda_2,$$
$$\varphi\Big|_{\partial\Omega^h}=\varphi^h_{(\Gamma)}, \quad \varphi=g^h \quad \text{in} \quad \Omega^h \quad \text{at} \quad t=0. \tag{67}$$

In the relationships (67) φ and f are vectors and Λ_1 and Λ_2 are matrices. It is assumed that the solution φ belongs to the space Φ of the grids functions. In the layer $t_j \le t \le t_{j+1}$ to solve the problem (67) we write the variable directions scheme

$$\frac{\varphi^{j+1/2}-\varphi^j}{\tau}+\frac{1}{2}\left(\Lambda_1^{j+1/2}\varphi^{j+1/2}\Lambda_2^j\varphi^j\right)=\frac{1}{2}f^j,$$
$$\frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau}+\frac{1}{2}\left(\Lambda_1^{j+1/2}\varphi^{j+1/2}\Lambda_2^{j+1}\varphi^{j+1}\right)=\frac{1}{2}f^j, \tag{68}$$
$$j=0,1,2,\ldots,$$

with the initial conditions

$$\varphi^0=g^h. \tag{69}$$

To equations (68), (69) it is necessary to add the difference boundary-value conditions which may be written, for example, in the form

$$\varphi^{j+1}\Big|_{\partial\Omega_1^h}=\varphi_{(\Gamma)}(t_{j+1}), \quad \varphi^{j+1/2}\Big|_{\partial\Omega_2^h}=\overline{\varphi}_{(\Gamma)}, \tag{70}$$

where $\partial\Omega_\alpha^h$ is the grid at the faces $x_\alpha = 0$ and $x_\alpha = 1$,

$$\overline{\varphi}_{(\Gamma)}=\frac{1}{2}[\varphi_{(\Gamma)}(t_{j+1})+\varphi_{(\Gamma)}(t_j)]+\frac{\tau}{4}\Lambda_2[\varphi_{(\Gamma)}(t_{j+1})-\varphi_{(\Gamma)}(t_j)].$$

Equation (68) is written in the form

$$\frac{2}{\tau}\varphi^{j+1/2}+\Lambda_1^{j+1/2}\varphi^{j+1/2}=F^j, \quad F^j=\frac{2}{\tau}\varphi^j-\Lambda_2^j\varphi^j+f^j,$$
$$\frac{2}{\tau}\varphi^{j+1}+\Lambda_1^{j+1}\varphi^{j+1}=F^{j+1/2}, \quad F^{j+1/2}=\frac{2}{\tau}\varphi^{j+1/2}-\Lambda_2^{j+1/2}\varphi^{j+1/2}+f^j. \tag{71}$$

Each of the problems (71) with the corresponding conditions (70) is a one-dimensional first boundary-value problem and can be solved by the factorisation method.

If the value of k_α^h at the node i_α is calculated, for example, using the equation $\left(k_\alpha^h\right)_{i_\alpha}=[k_\alpha(x_{i\alpha})+k_\alpha(x_{i_\alpha+1})]/2, \quad 1\le i_\alpha<I_\alpha$, the operator Λ_α approximates the operator A_α with the second order, i.e. $\Lambda_\alpha\varphi-A_\alpha\varphi=O(h_\alpha^2)$. Let $k_\alpha \equiv$ const. Operator Λ is self-adjoint and positive in Φ. We introduce the metrics

$$\|\varphi\|_\Lambda^2=\sum_{i_1=1}^{I_1-1}\sum_{i_2=1}^{I_2}(\Delta_{i_1}\varphi)^2h_1h_2+\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2-1}(\Delta_{i_2}\varphi)^2h_1h_2.$$

In this metrics the scheme (68)–(70) is stable in respect of the initial boundary-value conditions and the right-hand side.

Let the problem (66) have a unique solution $\varphi = \varphi(x,t)$ continuous in

$\Omega \times \Omega_t$ and in $\Omega \times \Omega_t$ there exist bounded derivatives

$$\frac{\partial^3\varphi}{\partial t^3}, \quad \frac{\partial^5\varphi}{\partial x_\alpha^2 \partial x_\beta^2 \partial t}, \quad \frac{\partial^4\varphi}{\partial x_\alpha^4}, \quad 0 \le \alpha,\ \beta \le 2.$$

Consequently, the scheme (68)–(70) converges in the grid norm with the $O(\tau^2+|h|^2)$, so that

$$\left\|\varphi^j - \varphi(t_j)\right\|_\Lambda \le M\left(|h|^2 + \tau^2\right),$$

where m = const is independent of τ and $|h|$.

4.2. Splitting methods for hydrodynamics problems

4.2.1. Splitting methods for Navier–Stokes equations

We examine a non-stationary problem for Navier–Stokes equations

$$\begin{gathered}\frac{\partial \mathbf{u}}{\partial t} - \nu\Delta\mathbf{u} + (\mathbf{u},\nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in} \quad \Omega\times(0,T),\\ \operatorname{div}\mathbf{u}=0 \quad \text{in} \quad \Omega\times(0,T),\\ \mathbf{u} = 0 \quad \text{on} \quad \Gamma\times[0,T],\\ \mathbf{u}(x,0) = \mathbf{u}_0(x) \quad \text{in} \quad \Omega,\end{gathered} \tag{72}$$

where $\Omega \subset \mathrm{R}^n$ is a bounded domain with the boundary Γ, ν = const>0, $\mathbf{u} = (u_1,\ldots,u_n)$ is the vector function (velocity vector), p is the scalar function (pressure), $\mathbf{f} \in (L_2(\Omega \times (0,T)))^n$ (almost for all $t\in(0,T)$ function $\mathbf{f}$ belongs to the closure of smooth vector-functions $\mathbf{v} = (v_1,\ldots,v_n)$ with a compact support in Ω such that div $\mathbf{v} \equiv \partial v_i/\partial x_i = 0$). It is assumed that $\mathbf{u}_0(x) \in (\overset{\circ}{W}{}_2^1)^n$, and also div $\mathbf{u}_0 = 0$.

We formulate some of the splitting schemes for (72).

The first scheme:

$$\frac{\mathbf{u}^{n+1/2} - \mathbf{u}}{\tau} - \nu\Delta\mathbf{u}^{n+1/2} + (\mathbf{u}^{n+1/2},\nabla)\mathbf{u}^{n+1/2} + \frac{1}{2}(\operatorname{div}\mathbf{u}^{n+1/2})\mathbf{u}^{n+1/2} = \mathbf{f}^n \quad \text{in} \quad \Omega,$$

$$\mathbf{u}^{n+1/2} = 0 \quad \text{on} \quad \Gamma,$$

$$\Delta p^{n+1} = \frac{1}{\tau}\operatorname{div}\mathbf{u}^{n+1/2} \quad \text{in} \quad \Omega, \quad \frac{\partial p^{n+1}}{\partial n} = 0 \quad \text{on} \quad \Gamma,$$

$$\mathbf{u}^{n+1} = \mathbf{u}^{n+1/2} - \tau\nabla p^{n+1} \quad \text{in} \quad \Omega, \quad n = 0,1,2,\ldots,N,$$

where

$$\mathbf{f}^n = \frac{1}{\tau}\int_{(n-1)/\tau}^{n\tau} \mathbf{f}(t,x)dt, \quad \tau = \frac{T}{N}.$$

Here the problem for $\mathbf{u}^{n+1/2}$ is the ‘conventional’ non-linear Dirichlet problem, and the problem for p^{n+1} is the Neuman problem for the Poisson equation.

It should be mentioned that the boundary conditions $\partial p/\partial n$ on Γ for the ‘true pressure’ $p(x,t)$ are not satisfied. The appearance of this presure for p^{n+1} is caused by the errors of approximation in the accepted splitting scheme.

The numerical solution of the problems for $\mathbf{u}^{n+1/2}$, p^{n+1} may be carried out by the method of the finite differences or finite elements using the approaches employed for solving the linear equations.

The second scheme (the alternating-direction scheme):

$$\frac{\mathbf{u}^{n+1/2}-\mathbf{u}^{n}}{\tau/2}-\nu\Delta\mathbf{u}^{n+1/2}+\nabla p^{n+1/2}+(\mathbf{u}^{n},\nabla)\mathbf{u}^{n}=\mathbf{f}^{n+1/2}\ \text{ in }\ \Omega,$$

$$\operatorname{div}\mathbf{u}^{n+1/2}=0\ \text{ in }\ \Omega,\quad \mathbf{u}^{n+1/2}=0\ \text{ on }\ \Gamma,$$

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^{n+1/2}}{\tau/2}-(\mathbf{u}^{n+1/2},\nabla)\mathbf{u}^{n+1}-\nu\Delta\mathbf{u}^{n+1/2}+\nabla p^{n+1/2}=\mathbf{f}^{n+1}\ \text{ on }\ \Omega,$$

where $\mathbf{u}^0=\mathbf{u}_0$. Here the problem for $\mathbf{u}^{n+1/2}$ is a linear problem – 'the generalised Stokes problem', and a large number of efficient algorithms have been developed for solving this problem, whereas the equation for $\mathbf{u}^{n+1}$ is of the first order and may be solved by the characteristics method.

The third scheme (the alternating-direction scheme with a weight):

$$\frac{\mathbf{u}^{n+1/2}-\mathbf{u}^{n}}{\tau/2}-\theta\nabla\nu\Delta\mathbf{u}^{n+1/2}+p^{n+1/2}-(1-\theta)\nu\Delta\mathbf{u}^{n}+(\mathbf{u}^{n},\nabla)\mathbf{u}^{n}=\mathbf{f}^{n+1/2},$$

$$\operatorname{div}\mathbf{u}^{n+1/2}=0\ \text{ in }\ \Omega,\quad \mathbf{u}^{n+1/2}=0\ \text{ on }\ \Gamma,$$

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^{n+1/2}}{\tau/2}-(1-\theta)\nu\Delta\mathbf{u}^{n+1}+(\mathbf{u}^{n+1},\nabla)\mathbf{u}^{n+1}-\theta\nu\nabla\mathbf{u}^{n+1/2}+$$

$$+\nabla p^{n+1/2}=\mathbf{f}^{n+1}\ \text{ in }\ \Omega,\quad \mathbf{u}^{n+1}=0\ \text{ on }\ \Gamma.$$

where $\mathbf{u}^0=\mathbf{u}_0$, and the 'weight' θ is selected from the interval (0,1).

Further realization of the scheme consists of the numerical solution of the generalized Stokes problem and the nonlinear Dirichlet problem.

These and a number of other splitting schemes for the Navier–Stokes equation were investigated and numerical algorithms were developed for application of these schemes in studies by A. Chorin, R. Teman, F. Harlow, R. Glovinskii, O.M. Belotserkovskii, V.A. Gushchin, V.V. Shchennikov, V.M. Koven', G.A. Tarnavskii, S.G. Chernyi, and many other investigators.

4.2.2. The fractional steps method for the shallow water equations

Let $\Omega\subset\mathbb{R}^2$ be a bounded domain (in some horizontal plane – 'the counting plane') with the boundary $\Gamma\equiv\partial\Omega$, $\mathbf{n}=(n_x,n_y)$ – the unit vector of the external normal to Γ, $\tau=(-n_y,n_x)$. Here $x, y\in\overline{\Omega}\equiv\Omega\cup\partial\Omega$ denotes spatial variables, and $t\in[0,T]$ is the time variable. It is assumed that $\Gamma=\Gamma_{\text{op}}\bigcup\Gamma_C\,(\Gamma_{\text{op}}\bigcap\Gamma_C=\varnothing)$, where $\Gamma_{\text{op}}=\bigcup_{m=1}^{M}\Gamma_{op,m}$ ($M<\infty$) is the *open part of the boundary* ('liquid boundary') whereas Γ_C is the *closed part of the boundary* ('solid boundary').

Let us assume that $\mathbf{v}(x,y,t)=(u,v)^T$ is the velocity vector, $\xi(x,y,t)$ is the level of the liquid in Ω in relation to the counting plane, $-h_0(x,y)$ is the depth

of the liquid below this plane, $h = \xi + h_0$. We examine the shallow water equations in the conservative form

$$\frac{\partial(h\mathbf{v})}{\partial t} + \nabla(h\mathbf{v}\mathbf{v}) - \nabla(\mu_1 h\nabla\mathbf{v}) + gh\nabla h = hF(\mathbf{v},h),$$
$$\frac{\partial h}{\partial t} + \nabla\cdot(h\mathbf{v}) = 0, \tag{73}$$

where

$$F(\mathbf{v},h) = f(\mathbf{v},h) - l\times\mathbf{v} - \frac{g|\mathbf{v}|\mathbf{v}}{hC^2},$$
$$f(\mathbf{v},h) = g\nabla h_0 + \frac{\mathbf{w}}{\rho h} - \frac{\nabla P_a}{\rho} \equiv (f_1, f_2)^T, \quad \nabla\cdot\mathbf{v} \equiv \mathrm{div}\mathbf{v},$$
$$l\times\mathbf{v} = (-lv, lu)^T,$$

g is the gravitational acceleration, l is the Coriolis parameter, C is the Chézy coefficient, $\mathbf{w} = (w_x, w_y)^T$ is the direction of the force of the wind, P_a is the atmospheric pressure on the free surface, ρ = const is the density of water. It should be mentioned that the effect of the variation of the depth of the liquid takes place if the term $\mathbf{F}(\mathbf{v},h)$ in (73) is given in the form

$$F(\mathbf{v},h) = f(\mathbf{v},h) - l\times\mathbf{v} - \frac{g|\mathbf{v}|\mathbf{v}}{K^2h^{1/3}},$$

where K is the Stricker coefficient.

The boundary conditions for $\mathbf{v}$, h are represented by

$$v_n = 0, \quad \mu_1 h\frac{\partial\mathbf{v}}{\partial n}\tau = 0 \quad \text{on } \Gamma_C\times(0,T),$$
$$\mu_1\left(h\frac{\partial v_n}{\partial n} + \frac{\partial h_{(\Gamma)}}{\partial t}\right) = 0, \quad \mu_1 h\frac{\partial v_\tau}{\partial n} = 0 \quad \text{on } \Gamma_{\text{op}}\times(0,T), \tag{74}$$
$$(|v_n| - v_n)(h_{(\Gamma)} - h) = 0 \quad \text{on } \Gamma_{\text{op}}\times(0,T),$$

where $v_n = (\mathbf{v},\mathbf{n})$, $v_\tau = (\mathbf{v},\tau)$, $h_{(\Gamma)}$ is a given function.

The initial condition for $\mathbf{v}, h$ is

$$\mathbf{v} = \mathbf{v}_{(0)}, \quad h = h_{(0)} \text{ at } t = 0, \quad (x,y)\in\Omega. \tag{75}$$

Equations (73)–(75) will be written in the *weak form*: to find

$$\varphi \equiv (\mathbf{v},h) \in X \equiv W\times W_2^1(\Omega) \quad \forall t,$$

in order to satisfy the equalities

$$\left(\frac{\partial}{\partial t}B(h)\varphi,\hat{\varphi}\right) + \sum_{i=1}^{3} a_i(\varphi;\varphi,\hat{\varphi}) = \sum_{i=1}^{3} f_i(\varphi,\hat{\varphi}) \quad \forall t\in(0,T),$$
$$\varphi = \varphi_{(0)} \quad \text{at} \quad t = 0, \quad \forall\hat{\varphi}\in W\times W_2^1(\Omega), \tag{76}$$

where we use the following notations

$$f_1(\bar{\varphi};\hat{\varphi}) = \frac{1}{2}\int_{\Gamma} h_{(\Gamma)}(\left|(\bar{\mathbf{v}},\mathbf{n})\right| - (\bar{\mathbf{v}},\mathbf{n}))\bar{\mathbf{v}}\hat{\mathbf{v}}\, d\Gamma,$$

$$f_2(\bar{\varphi};\hat{\varphi}) = \int_{\Omega}\left(\frac{\bar{h}}{\rho}\nabla(gh_0\rho - P_a) + \frac{\mathbf{w}}{\rho}\right)\hat{\mathbf{v}}d\Omega - \int_{\Gamma}\mu_1\frac{\partial h_{(\Gamma)}}{\partial t}(\hat{\mathbf{v}},\mathbf{n})d\Gamma,$$

$$f_3(\bar{\varphi};\hat{\varphi}) = \frac{1}{2}\int_{\Gamma} g(\left|(\bar{\mathbf{v}},\mathbf{n})\right| - (\bar{\mathbf{v}},\mathbf{n}))h_{(\Gamma)}\hat{h}\, d\Gamma,$$

$$\Gamma = \Gamma_0 \cup \Gamma_{\text{op,inf}} \cup \Gamma_{\text{op,out}}, \quad \Gamma_0 = \{x\in\Gamma : (\bar{\mathbf{v}}(x),\mathbf{n}(x)) = 0\}, \quad \alpha\in[0,2],$$

$$\Gamma_{\text{op,inf}} = \{x\in\Gamma : (\bar{\mathbf{v}}(x),\mathbf{n}(x)) < 0\}, \quad \Gamma_{\text{op,out}} = \{x\in\Gamma : (\bar{\mathbf{v}}(x),\mathbf{n}(x)) > 0\},$$

$$\varphi = (\mathbf{v},h)^T, \quad \hat{\varphi} = (\hat{v},\hat{h})^T = (\hat{u},\hat{v},\hat{h})^T, \quad \bar{\varphi} = (\bar{v},\bar{h})^T = (\bar{u},\bar{v},\bar{h})^T,$$

$$W = \left\{\hat{v} = \left(\hat{u},\hat{v}\right)^T : \hat{u},\hat{v}\in W_2^1(\Omega), (\hat{\mathbf{v}},\mathbf{n}) = 0 \text{ on } \Gamma_C, (\hat{\mathbf{v}},\tau) = 0 \text{ on } \Gamma_{\text{op}}\right\},$$

$$\left(\frac{\partial}{\partial t}B(\bar{h})\varphi,\hat{\varphi}\right) = \int_{\Omega}\left(\frac{\partial\bar{h}u}{\partial t}\hat{u} + \frac{\partial\bar{h}v}{\partial t}\hat{v} + g\frac{\partial h}{\partial t}\hat{h}\right)d\Omega,$$

$$a_1(\bar{\varphi};\varphi,\hat{\varphi}) = -\int_{\Omega}\bar{h}\mathbf{v}(\bar{\mathbf{v}},\nabla)\hat{\mathbf{v}}d\Omega + \frac{1}{2}\int_{\Gamma}\bar{h}((\bar{\mathbf{v}},\mathbf{n}) + \left|(\bar{\mathbf{v}},\mathbf{n})\right|)\mathbf{v}\hat{\mathbf{v}}d\Gamma -$$

$$-\frac{\alpha}{2}\int_{\Gamma}\nabla\cdot(\bar{h}\bar{\mathbf{v}})\mathbf{v}\hat{\mathbf{v}}d\Omega, \quad \alpha\in[-2,2],$$

$$a_2(\bar{\varphi};\varphi,\hat{\varphi}) = \int_{\Omega}\left(\mu_1\bar{h}\nabla\mathbf{v}\cdot\nabla\left(\hat{\mathbf{v}} + C(\bar{\varphi}) + \frac{\alpha}{2}\nabla\cdot(\bar{h}\bar{\mathbf{v}})\right)\mathbf{v}\hat{\mathbf{v}}\right)d\Omega, \quad C(\bar{\varphi}) = \frac{g\left|\mathbf{v}\right|}{C^2},$$

$$a_3(\bar{\varphi};\varphi,\hat{\varphi}) = \int_{\Omega}(g\bar{h}\nabla\cdot(h\hat{\mathbf{v}}) - g\bar{h}(\mathbf{v},\nabla)\hat{h} + l\hat{h}(u\hat{v} - v\hat{u}))d\Omega +$$

$$+\frac{1}{2}\int_{\Gamma} g(\left|(\bar{\mathbf{v}},\mathbf{n})\right| + (\bar{\mathbf{v}},\mathbf{n}))h\hat{h}\, d\Gamma.$$

Accepting that $\bar{\varphi}\equiv\varphi$, equalities (76) may be written in the operator form of the type

$$\frac{\partial}{\partial t}B(\bar{h})\varphi + \sum_{i=1}^{3}\Lambda_i(\bar{\varphi})\varphi = \sum_{i=1}^{3} f_i(\bar{\varphi}), \quad t\in(0,T), \quad \varphi(0) = \varphi_{(0)},$$

where the operator $\{\Lambda_i\}$ and the functions of the right-hand sides $\{f_i\}$ are determined using the following relationship:

$$(\Lambda_i(\bar{\varphi})\tilde{\varphi},\hat{\varphi}) \equiv a_i(\bar{\varphi};\tilde{\varphi},\hat{\varphi}), \quad (f_i,\tilde{\varphi}) \equiv f_i(\bar{\varphi},\hat{\varphi}) \quad \forall\tilde{\varphi},\hat{\varphi}\in X, \quad i = 1,2,3.$$

Let $\Delta t>0$ be the step in respect of the time variable, $t_j = j\Delta t$, (j = 0,1,…., j), and also φ^j, $\mathbf{v}^j$…be the values φ, $\mathbf{v}$, … at $t = t_j$. It is assumed that φ^{j-1} is known, $\varphi^0\equiv\varphi_{(0)}$. Consequently, the initial-boundary value problem for the shallow water equations maybe reformulated as follows: find $\varphi\in X(\forall t\in(t_j,t_{j+1}))$, satisfying the equations of the type

$$\frac{\partial}{\partial t}B(\bar{h})\varphi+\sum_{i=1}^{3}\Lambda_i(\bar{\varphi})\,\varphi=\sum_{i=1}^{3}f_i$$

$$t\in(t_j,t_{j+1}),\quad \varphi(t_j)=\varphi^j,\quad j=0,1,\ldots,J-1.$$

It may easily be verified that if $\alpha=1$, $\bar{h}_0\geq 0$ and $C(\varphi)+\nabla\cdot(\bar{h}\bar{\mathbf{v}})/2\geq 0$, Then $(\Lambda_i(\bar{\varphi})\varphi,\varphi)\geq 0\ \forall\varphi\in X$. These properties of the operators $\{\Lambda_i\}$ enable us to formulate the fractional steps scheme:

$$\begin{gathered}
\frac{\partial}{\partial t}B(\bar{h})\varphi^{j+1/3}+\Lambda_1(\bar{\varphi})\varphi^{j+1/3}=f_1(\bar{\varphi}),\quad t\in(t_j,t_{j+1}),\\
\varphi^{j+1/3}(t_j)=\varphi^j(t_j),\\
\frac{\partial}{\partial t}B(\bar{h})\varphi^{j+2/3}+\Lambda_2(\bar{\varphi})\varphi^{j+2/3}=f_2(\bar{\varphi}),\quad t\in(t_j,t_{j+1}),\\
\varphi^{j+2/3}(t_j)=\varphi^{j+1/3}(t_{j+1}),\\
\frac{\partial}{\partial t}B(\bar{h})\varphi^{j+1}+\Lambda_3(\bar{\varphi})\varphi^{j+1}=f_3(\bar{\varphi}),\quad t\in(t_j,t_{j+1}),\\
\varphi^{j+1}(t_j)=\varphi^{j+2/3}(t_{j+1}),\\
j=0,1,\ldots,J-1,\qquad \varphi^0(t_0)\equiv\varphi_{(0)},
\end{gathered}\tag{77}$$

where $\bar{\varphi}=(\bar{\mathbf{v}},\bar{h})^T$ is the approximation of φ on (t_j,t_{j+1}). Let $\bar{\varphi}=\varphi^j(t_j)$ on (t_j,t_{j+1}). In this case, the scheme (77) has the accuracy $O(\Delta t)$ in the entire interval $(0,T)$.

If the function $\varphi^{j+1/3}$, $\varphi^{j+2/3}$, φ^{j+1}, determined by means of (77) are sufficiently smooth, the scheme (77) maybe written in the terms of differential operators with the appropriate boundary conditions in the following form

$$\begin{gathered}
\frac{\partial\bar{h}\mathbf{v}^{j+1/3}}{\partial t}+\nabla\cdot(\bar{h}\bar{\mathbf{v}}\mathbf{v}^{j+1/3})-\frac{\alpha}{2}\nabla\cdot(\bar{h}\bar{\mathbf{v}})\mathbf{v}^{j+1/3}=0\ \text{ in }\ \Omega\times(t_j,t_{j+1}),\\
\bar{h}\mathbf{v}^{j+1/3}\Big|_{t_j}=\bar{h}\mathbf{v}^{j}\Big|_{t_j},\quad \mathbf{v}^{j+1/3}\cdot\mathbf{n}=0\ \text{ on }\ \Gamma_C\times(t_j,t_{j+1}),\\
\bar{h}\mathbf{v}^{j+1/3}=h_{(\Gamma)}\bar{\mathbf{v}}\ \text{ on }\ \Gamma_{\inf}(\bar{\mathbf{v}})\times(t_j,t_{j+1}),\\
\frac{\partial\bar{h}\mathbf{v}^{j+2/3}}{\partial t}+\nabla\cdot(\mu_1\bar{h}\nabla\mathbf{v}^{j+2/3})+\left(C(\bar{\varphi})+\frac{\alpha}{2}\nabla\cdot(\bar{h}\bar{\mathbf{v}})\right)\mathbf{v}^{j+2/3}=\\
=\frac{\bar{h}}{\rho}\nabla(gh_0\rho\text{-}P_a)+\frac{\mathbf{w}}{\rho}\ \text{ in }\ \Omega\times(t_j,t_{j+1}),\\
\bar{h}\mathbf{v}^{j+2/3}\Big|_{t_j}=\bar{h}\mathbf{v}^{j+1/3}\Big|_{t_{j+1}},\quad \mu_1\bar{h}\frac{\partial\mathbf{v}^{j+2/3}}{\partial n}\tau=0,\quad \mathbf{v}^{j+2/3}\cdot\mathbf{n}=0\ \text{ on }\ \Gamma_C\times(t_j,t_{j+1}),\\
\mu_1\left(\bar{h}\frac{\partial\mathbf{v}^{j+2/3}}{\partial n}\mathbf{n}+\frac{\partial h_{(\Gamma)}}{\partial t}\right)=0,\quad \mu_1\bar{h}\frac{\partial\mathbf{v}^{j+2/3}}{\partial n}\tau=0\ \text{ on }\ \Gamma_{\text{op}}\times(t_j,t_{j+1}),
\end{gathered}\tag{78}$$

$$\frac{\partial \bar{h}u^{j+1}}{\partial t} - l\bar{h}v^{j+1} + g\bar{h}\frac{\partial h^{j+1}}{\partial x} = 0, \quad \frac{\partial \bar{h}u^{j+1}}{\partial t} - l\bar{h}v^{j+1} + g\bar{h}\frac{\partial h^{j+1}}{\partial y} = 0$$

$$\text{in } \Omega\times(t_j, t_{j+1}),$$

$$g\frac{\partial h^{j+1}}{\partial t} + g\nabla\cdot\left(\bar{h}\mathbf{v}^{j+1}\right) = 0 \text{ in } \Omega\times(t_j, t_{j+1}),$$

$$\bar{h}\mathbf{v}^{j+1}\Big|_{t_j} = \bar{h}\mathbf{v}^{j+2/3}\Big|_{t_{j+1}}, \quad h^{j+1}\Big|_{t_j} = h^{j+2/3}\Big|_{t_{j+1}},$$

$$\bar{h}(\mathbf{v}^{j+1}\cdot\mathbf{n}) = h^{j+1}\left(\frac{\left|\bar{v}_n\right| + \bar{v}_n}{2}\right) - h_{(\Gamma)}\left(\frac{\left|\bar{v}_n\right| + \bar{v}_n}{2}\right) \text{ on } \Gamma_{\text{op}}\times(t_j, t_{j+1}),$$

$$\mathbf{v}^{j+1}\cdot\mathbf{n} = 0 \text{ on } \Gamma_C\times(t_j, t_{j+1}) \quad (j = 0, 1, 2, \ldots, J-1).$$

If we assume that $\alpha = 2$ and ignore the members including l, equation for $\mathbf{v}^{j+1/3}$, $\mathbf{v}^{j+1}$, h^{j+1} in (78) are reduced respectively to the following equations

$$\frac{\partial \mathbf{v}^{j+1/3}}{\partial t} + \left(\bar{\mathbf{v}}, \nabla\right)\mathbf{v}^{j+1/3} = 0 \text{ in } \Omega\times(t_j, t_{j+1}), \quad \mathbf{v}^{j+1/3}\Big|_{t_j} = \mathbf{v}^{j}\Big|_{t_j},$$

$$\mathbf{v} = \bar{\mathbf{v}} \text{ on } \Gamma_{inf}\times(t_j, t_{j+1}), \quad \mathbf{v}^{j+1/3}\cdot\mathbf{n} = 0 \text{ on } \Gamma_C\times(t_j, t_{j+1}),$$

$$h_{tt}^{j+1} - \nabla\cdot\left(\bar{h}\nabla h^{j+1}\right) = 0 \text{ in } \Omega\times(t_j, t_{j+1}),$$

$$h^{j+1}\Big|_{t_j} = h^{j+2/3}\Big|_{t_{j+1}}, \quad \frac{\partial h^{j+1}}{\partial t}\Big|_{t_j} = -\nabla\cdot\left(\bar{h}\mathbf{v}^{j+2/3}\right)\Big|_{t_{j+1}},$$

$$g\bar{h}\frac{\partial h^{j+1}}{\partial n} + \frac{\partial}{\partial t}\left(h^{j+1}\frac{\left|\bar{v}_n\right| + \bar{v}_n}{2} - h_{(\Gamma)}\frac{\left|\bar{v}_n\right| - \bar{v}_n}{2}\right) = 0 \text{ on } \Gamma_{\text{op}}\times(t_j, t_{j+1}), \quad (79)$$

$$g\bar{h}\frac{\partial h^{j+1}}{\partial n} = 0 \text{ on } \Gamma_C\times(t_j, t_{j+1}),$$

$$\frac{\partial \bar{h}\mathbf{v}^{j+1}}{\partial t} = -g\bar{h}\nabla h^{j+1} \text{ in } \Omega\times(t_j, t_{j+1}),$$

$$\bar{h}\mathbf{v}^{j+1}\Big|_{t_j} = \bar{h}\mathbf{v}^{j+2/3}\Big|_{t_{j+1}} \quad \mathbf{v}^{j+1}\cdot\mathbf{n} = 0 \text{ on } \Gamma_C\times(t_j, t_{j+1}),$$

and the problem for $\mathbf{v}^{j+2/3}$ has the same form as in (78) (at $\alpha = 2$). (Attention should be given to the fact that in the splitting schemes (78), (79) we also solve the problem of selecting the boundary conditions for problems of 'fractional steps').

The next stage of numerical solution of the initial problem is the solution of the initial-boundary problems from (78), (79). This can be carried out using the well-known different schemes in respect of t (explicit or implicit, splitting schemes, etc) and suitable approximations of this problem in respect of

spatial variables (the finite element method, the finite difference method, etc). On the basis of the results obtained using the theory of these schemes and approximations we can formulate appropriate claims for the entire numerical algorithm of solution of the shallow water equations.

The presented splitting schemes for the shallow water equations were proposed and investigated by Agoshkov and Saleri who also examined other types of boundary-value conditions.

4.3. Splitting methods for the model of dynamics of sea and ocean flows

4.3.1. The non-stationary model of dynamics of sea and ocean flows

We examined a system of equations of the ocean dynamics

$$\begin{gathered}
\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+w\frac{\partial u}{\partial z}-lv+\frac{1}{\rho_0}\frac{\partial p}{\partial x}=\mu\Delta u+\frac{\partial}{\partial z}\nu\frac{\partial u}{\partial z},\\
\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+w\frac{\partial v}{\partial z}-lv+\frac{1}{\rho_0}\frac{\partial p}{\partial y}=\mu\Delta v+\frac{\partial}{\partial z}\nu\frac{\partial v}{\partial z},\\
\frac{\partial p}{\partial z}=g\rho,\\
\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}=0,
\end{gathered}\tag{80}$$

$$\begin{gathered}
\frac{\partial T}{\partial t}+u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial y}+w\frac{\partial T}{\partial z}-\gamma_T w=\mu_T\Delta T+\frac{\partial}{\partial z}\nu_T\frac{\partial T}{\partial z},\\
\frac{\partial S}{\partial t}+u\frac{\partial S}{\partial x}+v\frac{\partial S}{\partial y}+w\frac{\partial S}{\partial z}-\gamma_S w=\mu_S\Delta S+\frac{\partial}{\partial z}\nu_S\frac{\partial S}{\partial z},\\
\rho=\alpha_T T+\alpha_S S.
\end{gathered}$$

The boundary and initial conditions for the system (80) are represented by the conditions

$$\begin{gathered}
\frac{\partial u}{\partial z}=\frac{-\tau_x}{\nu\rho_0},\quad \frac{\partial v}{\partial z}=\frac{-\tau_y}{\nu\rho_0},\quad \frac{\partial T}{\partial z}=f_3,\quad \frac{\partial S}{\partial z}=f_4,\quad w=0\quad \text{at}\quad z=0,\\
\frac{\partial T}{\partial z}=0,\quad \frac{\partial S}{\partial z}=0,\quad \frac{\partial u}{\partial z}=0,\quad \frac{\partial v}{\partial z}=0,\quad w=0\quad \text{at}\quad z=H,
\end{gathered}\tag{81}$$

$$\begin{gathered}
u=0,\quad v=0,\quad \frac{\partial T}{\partial n}=0,\quad \frac{\partial S}{\partial n}=0\quad \text{on}\quad \Sigma,\\
u=u^0,\quad v=v^0,\quad T=T^0,\quad S=S^0\quad \text{at}\quad t=t_j.
\end{gathered}\tag{82}$$

It should be mentioned that for the system (80) with the boundary and initial conditions (81), (82) (and also for a number of other formulations of

the boundary conditions), there is uniqueness of the classic solution.

From (80) we exclude the function ρ and linearize the resultant scheme of equations on the time period $t_{j-1} \le t \le t_{j+1}$; consequently, it may be written in the operator form:

$$B\frac{\partial \varphi}{\partial t} + A\varphi = 0, \quad B\varphi = BF \quad \text{at} \quad t = 0. \tag{83}$$

Here:

$$\varphi = \begin{bmatrix} u \\ v \\ w \\ p \\ T \\ S \end{bmatrix}, \qquad A = \begin{bmatrix} R_1 & -\rho_0 l & 0 & \frac{\partial}{\partial x} & 0 & 0 \\ \rho_0 l & R_2 & 0 & \frac{\partial}{\partial y} & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial}{\partial z} & -g\alpha_T E & -g\alpha_S E \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} & 0 & 0 & 0 \\ 0 & 0 & g\alpha_T E & 0 & R_3 & 0 \\ 0 & 0 & g\alpha_S E & 0 & 0 & R_4 \end{bmatrix}$$

$$B = \begin{bmatrix} \rho_0 E & 0 & 0 & 0 & 0 & 0 \\ \rho_0 l & \rho_0 E & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{g\alpha_T}{\gamma_T} E & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{g\alpha_T}{\gamma_T} E \end{bmatrix}, \qquad F = \begin{bmatrix} u^0 \\ v^0 \\ 0 \\ 0 \\ T^0 \\ S^0 \end{bmatrix},$$

$$R_i = D_0 + D_i, \quad D_0 = \mathrm{div}\mathbf{u}^{j\text{-}1},$$

$$D_1 = D_2 = -\rho_0 \left(\mu\Delta + \frac{\partial}{\partial z} v \frac{\partial}{\partial z} \right),$$

$$D_3 = -\frac{g\alpha_T}{\gamma_T} \left(\mu_T \Delta + \frac{\partial}{\partial z} v_T \frac{\partial}{\partial z} \right),$$

$$D_4 = -\frac{g\alpha_S}{\gamma_T} \left(\mu_S \Delta + \frac{\partial}{\partial z} v_S \frac{\partial}{\partial z} \right).$$

It is assumed that the solution of the problem (83) belongs to a set of continuously differentiated functions satisfying the boundary conditions (81). We introduce the scalar product by the relationship:

$$\left(\bar{a},\bar{b}\right)=\sum_{i=1}^{6}a_i\,b_i\,d\Omega,$$

where a_i and b_i are the components of the vector functions $\bar{a}$ and $\bar{b}$. It may be shown that on the functions from this subspace the operator A is non-negative.

4.3.2. The splitting method

We present the operator A in the form of the sum of two operators:

$$A_1=\begin{bmatrix} R_1 & 0 & 0 & 0 & 0 & 0\\ 0 & R_2 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & R_3 & 0\\ 0 & 0 & 0 & 0 & 0 & R_4 \end{bmatrix},$$

$$A_2=\begin{bmatrix} 0 & -\rho_0 lE & 0 & \dfrac{\partial}{\partial x} & 0 & 0\\ \rho_0 lE & 0 & 0 & \dfrac{\partial}{\partial y} & 0 & 0\\ 0 & 0 & 0 & \dfrac{\partial}{\partial z} & -g\alpha_T E & -g\alpha_S E\\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} & 0 & 0 & 0\\ 0 & 0 & g\alpha_T E & 0 & 0 & 0\\ 0 & 0 & g\alpha_S E & 0 & 0 & 0 \end{bmatrix},$$

here the first operator is positive semi-definite, the second one is *anti-Hermitian*: $A = A_1 + A_2$, $(A_1\varphi,\varphi)\geq 0$, $(A_2\varphi,\varphi) = 0$.

Now, we examine the entire time period $0 \leq t \leq T$ and divide it into equal intervals $t_{j-1} \leq t \leq t_j$ with the length $t_j - t_{j-1} = \tau$. In each expanded interval $t_{j-1} \leq t \leq t_{j+1}$ to solve the problem (83) we use the two-cyclic splitting method. If it is assumed that φ^{j-1} is the solution of problem (80)–(82) at the moment of time t_{j-1}, then in the interval $t_{j-1} \leq t \leq t_j$ we have

$$B\frac{\partial\varphi_1}{\partial t}+A_1\varphi_1=0,\quad B\varphi_1^{j-1}=B\varphi^{j-1}, \tag{84}$$

in the interval $t_{j-1} \leq t \leq t_j$ we have

$$B\frac{\partial\varphi_2}{\partial t}+A_2\varphi_2=0,\quad B\varphi_2^{j-1}=B\varphi_1^{j-1} \tag{85}$$

and in the interval $t_{j-1} \leq t \leq t_j$ we solve the problem

$$B\frac{\partial\varphi_3}{\partial t}+A_1\varphi_3=0,\quad B\varphi_3^{j}=B\varphi_2^{j+1}. \tag{86}$$

In the componentwise form, the problem (84) has the form

$$\frac{\partial u_1}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}u_1=\mu\Delta u_1+\frac{\partial}{\partial z}\nu_1\frac{\partial u_1}{\partial z},$$
$$\frac{\partial v_1}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}v_1=\mu\Delta v_1+\frac{\partial}{\partial z}\nu_1\frac{\partial v_1}{\partial z},$$
$$\frac{\partial T_1}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}T_1=\mu\Delta T_1+\frac{\partial}{\partial z}\nu_T\frac{\partial T_1}{\partial z},$$
$$\frac{\partial S_1}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}S_1=\mu\Delta S_1+\frac{\partial}{\partial z}\nu_S\frac{\partial S_1}{\partial z}$$

on the condition

$$\operatorname{div}\mathbf{u}^{j-1}=0,$$

and also at the boundary conditions (81) and initial data

$$u_1^{j-1}=u^{j-1},\quad v_1^{j-1}=v^{j-1},\quad T_1^{j-1}=T^{j-1},\quad S_1^{j-1}=S^{j-1};$$

in the interval $t_{j-1}\le t\le t_j$

$$\frac{\partial u_2}{\partial t}-lv_2+\frac{1}{\rho_0}\frac{\partial p_2}{\partial x}=0,\qquad \frac{\partial u_2}{\partial x}+\frac{\partial v_2}{\partial y}+\frac{\partial w_2}{\partial z}=0,$$
$$\frac{\partial v_2}{\partial t}-lv_2+\frac{1}{\rho_0}\frac{\partial p_2}{\partial y}=0,\qquad \frac{\partial T_2}{\partial t}+\gamma_T w_2=0,$$
$$\frac{\partial p_2}{\partial z}=g(\alpha_T T_2+\alpha_S S_2),\qquad \frac{\partial S_2}{\partial t}+\gamma_S w_2=0$$

on the condition

$$w_2=0\quad\text{at}\quad z=0,$$
$$w_2=0\quad\text{at}\quad z=0,$$
$$(u_2)_n=0\quad\text{on}\quad\Sigma$$

and the initial data

$$u_2^{j-1}=u_1^j,\quad v_2^{j-1}=v_1^j,\quad T_2^{j-1}=T_1^j,\quad S_2^{j-1}=S_1^j.$$

In the final stage of splitting ($t_{j-1}\le t\le t_j$) we obtain

$$\frac{\partial u_3}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}u_3=\mu_1\Delta u_3+\frac{\partial}{\partial z}\nu_1\frac{\partial u_3}{\partial z},$$
$$\frac{\partial v_3}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}v_3=\mu_2\Delta v_3+\frac{\partial}{\partial z}\nu_1\frac{\partial v_3}{\partial z},$$
$$\frac{\partial T_3}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}T_3=\mu_T\Delta T_3+\frac{\partial}{\partial z}\nu_T\frac{\partial T_3}{\partial z},$$
$$\frac{\partial S_3}{\partial t}+\operatorname{div}\mathbf{u}^{j-1}S_3=\mu_S\Delta S_3+\frac{\partial}{\partial z}\nu_S\frac{\partial S_3}{\partial z}$$

on the condition

$$\operatorname{div}\mathbf{u}^{j-1}=0,$$

and also at the boundary conditions (81) and the initial data

$$u_3^j = u_2^{j+1}, \quad v_3^j = v_2^{j+1}, \quad T_3^j = T_2^{j+1}, \quad S_3^j = S_2^{j+1}.$$

The solution of the sub-problem, included in all stages of the splitting methods, can be obtained by the method of finite differences or other suitable methods of computational mathematics.

BIBLIOGRAPHIC COMMENTARY

One of the first monographs dealing with the theory and applications of the splitting methods is the book published by Yanenko [102]. Systematic explanation of these splitting methods was published in a monograph by Marchuk [61], where the author examines the examples of application of these methods to different problems of mathematical physics, problems of hydrodynamics, oceanology and meteorology, and also presents a large number of literature references. The splitting method have been examined in considerable detail in [80].

The application of the splitting methods to problems of the Navier–Stokes equations and the theoretical justifications have been published in [90]. The authors of [67] examine different mathematical models of the dynamics of water in oceans and seas and describe the splitting methods for these cases. In an article by Agoskov and Saleri (Mathematical modelling, 1996, Volume 65, No.9, p.3–24) the authors describe a number of splitting schemes for the shallow water equations and solve the problem of boundary conditions for these schemes. Mathematical simulation of these problems, associated with the protection of environment, using the splitting methods, has been carried out in [62]. Other diversified applications of the splitting methods and a large number of references have been published in [61].

Chapter 7

METHODS FOR SOLVING NON-LINEAR EQUATIONS

Keywords: Variation of the functional, Gateaux differential, functional gradient, convex functional, monotonic operator, non-linear boundary-value problem, critical point of the functional, variational method, Ritz method, Newton–Kantorovich method, The Galerkin–Petrov method, perturbation method.

MAIN CONCEPTS AND DEFINITIONS

The variation of the functional – limit

$$Vf(u,h) = \lim_{t\to 0} \frac{f(u+th) - f(u)}{t},$$

where $f(u)$ is a non-linear functional given on the normalized space E, u, $h \in E$.

The Gateaux differential – variation $Df(u,h) = Vf(u,h)$ if it is linear in respect of h.

The Gateaux derivative – map $f(x)$ from the representation $Df(u,h) = f'(u)h$.

The gradient of the functional – the Gateaux derivative $f'(u)$ being a linear bounded functional.

The convex functional – the real differentiable functional $f(u)$ for which the following inequality is fulfilled at all u, u_0:

$$f(u) - f(u_0) - Df(u_0, u - u_0) \geq 0.$$

The monotonic operator – map F: $E \to E^*$, for which the inequality $\langle u-v, F(u)-F(v)\rangle \geq 0$ holds, where E is the normalized space, E^* is adjoint to it and $\langle .,. \rangle$ is the duality relation.

The *lower semi-continuous functional* $f(u)$ – the real functional $f(u)$ given in the normalized space E; it is lower semi-continuous at the point $u_0 \in E$ if for any sequence $u_n \in E$ such that $u_n \to u_0$, we have the inequality $f(u_0) \leq \underline{\lim}_n\ f(u_n)$.

The critical point of the functional – point u_0 at which the Gateaux differential converts to 0: $Df(u_0,h) = 0$.

The variational method – the method of examining non-linear equations based on reducing the non-linear equation to the problem of determining the critical points of some functional.

The minimising sequence – sequence u_n satisfying the condition $\lim_{n\to\infty} f(u_n) = d$, where $f(u)$ is a functional, and $d = \inf f(u)$.

The Ritz method – the method of finding the minimum of the functional by constructing approximations in the form of linear combinations of special basis functions.

The Newton–Kantorovich method – the iteration process $u_{n+1} = u_n - [F'(u_n)]^{-1}F(u_n)$ of solving the equation $F(u) = 0$.

The Galerkin–Petrovich method – the method of solving nonlinear equations by deriving an equation in the finite-dimensional subspace and constructing approximation in the form of linear combinations from this subspace.

The perturbation method – the method of solving nonlinear equations with a small parameter by constructing approximations in the form of a series in respect of the degrees of this parameter.

1. INTRODUCTION

In mathematical modelling of physical processes and phenomena, it is often necessary to solve the nonlinear problems of mathematical physics, especially well-known Monge–Ampere, Navier–Stokes, Kolmogorov– Petrovskii equations, and others. Together with the boundary and initial conditions, nonlinear equations lead to formulations of nonlinear boundary-value problems. These problems can be formulated in turn as the operator equations in functional spaces. Recently, powerful facilities in nonlinear functionalists have been developed for solving nonlinear operator equations. One of the main methods of examining nonlinear equations is the variational method in which the solution of the initial equation is reduced to the problem of finding critical points of some functional. An important role is also played by the methods of minimising sequences, especially the method of the steepest decent, Ritz method, the Newton–Kantorovich method. The Galerkin–Petrov method is one of the most widely used methods of examining and numerical solutions of nonlinear problems. In this method, the initial equation is projected onto the finite-dimensional sub-space, and approximations to the grid are found in (possibly) another sub-space. The classic perturbation method is used to find solutions of the nonlinear problems by expanding them in respect of a small parameter. These methods are used widely in solving nonlinear problems of mathematical physics.

In many cases, the description of physical phenomena and processes is carried out using differential or integro-differential equations which together with the set of the initial data form the mathematical model of the examined process.

However, the mathematical description of the physical phenomena often

includes some simplification. If all factors are taken into account in description, mathematically unsolved problems would form very often. Therefore, mathematical description is nothing else but the approximation of physical reality. The description, leading to a linear equation, is a sort of first approximation whose advantage is that it leads to mathematical problems solvable using the mathematical facilities available at the moment. A more accurate description of the physical phenomena would result in nonlinear equations. Thus, nonlinear description is the next approximation which makes it possible to examine additional factors. This should be illustrated on the following example. The Poisson equation is a partial case of the equation

$$-\frac{\partial}{\partial x}\left(k(x,y)\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(k(x,y)\frac{\partial u}{\partial y}\right)=f(x,y) \text{ in } \Omega,$$

where the given function $k = k(x,y)$ characterises, for example, the heat conductivity of the material at the point $(x,y)\in\Omega\subset\mathrm{R}^3$, and the function $u = u(x,y)$ is the temperature distribution. If the heat conductivity of the material is constant, we obtain the Poisson equation

$$-k\Delta u = f \text{ in } \Omega,$$

k = const, Δ is the Laplace operator. However, it is well known that the heat conductivity of the material may change not only with the position but also with the temperature to which the material is subjected, i.e. function k may depend on u and, in the final analysis, also on its derivatives $k=k\left(x,y,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\right)$. Consequently, the temperature distribution is described by the equation

$$-\frac{\partial}{\partial x}\left(k\left(x,y,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\right)\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(k\left(x,y,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\right)\frac{\partial u}{\partial y}\right)=f(x,y). \tag{1}$$

However, this equation is already nonlinear. The nonlinear equation (1) is a partial case of the equation

$$-\frac{\partial}{\partial x}\left(a_1(x,y;u,\mathrm{grad}u)\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(a_2(x,y;u,\mathrm{grad}u)\frac{\partial u}{\partial y}\right)+a_0 = f(x,y), \tag{2}$$

where $a_0=a_0(x,y;u,\mathrm{grad}u)$. If we set in equation (2)

$$a_1(x,y;\ \xi_0,\xi_1,\xi_2)=a_2(x,y;\ \xi_0,\xi_1,\xi_2)=k\ (x,y;\ \xi_0,\xi_1,\xi_2),\ a_0=0,$$

we obtain equation (1). In equation (2), $a_i=a_i(x,y;\ \xi_0,\xi_1,\xi_2,)$ are given functions determined for $(x,y)\in\Omega$ $\xi_0,\xi_1,\xi\in\mathrm{R}^3$, and $f = f(x,y)$ is the function given on Ω. The equation of type (2) is often encountered in practice.

Coupled with the boundary and initial conditions, nonlinear equations lead to formulations of the linear boundary-value problems. These problems can be formulated in turn as the operator equations in Banach spaces.

We examine, for example, problem (2) with the boundary-value condition $u|_{\partial\Omega}=0$. Let a_i, $i = 0,1,2,$ be sufficiently smooth functions. Let v denote the left-hand side of equation (2). Consequently, function $u\in C^2(\bar{\Omega})$ is related

with the function $v \in C^0(\bar{\Omega})$. This correspondence determines the operator A mapping the Banach space $C^2(\bar{\Omega})$ into the Banach space $C^0(\bar{\Omega})$. As the domain of definition of this operator we can use $D(A)=\{u \in C^2(\bar{\Omega}): u|_{\partial\Omega}=0\}$. Consequently, the examined boundary-value problem may be written in the form of the operator equation:

$$Au = f,$$

i.e. for the given function $f \in C^0(\bar{\Omega})$ it is required to find function $u \in D(A)$ such that $Au = f$. Thus, the nonlinear boundary-value problem can be reduced to the operator equation of the type $Au = f$.

It should be mentioned that the explicit formulae, giving the solution of the boundary-value problem for nonlinear differential equations, maybe determined only in exceptional cases. For these reasons, the approximate methods of examination and numerical solution of nonlinear problems are used widely. Recently, mathematicians have developed in the nonlinear functional analysis powerful facilities for solving these problems [7,9,10,17,36,42,50,77,94,99,116].

2. ELEMENTS OF NONLINEAR ANALYSIS

2.1. Continuity and differentiability of nonlinear mappings

2.1.1. Main definitions

Definition 1. Let E_x and E_y be the normalized spaces. Operator F: $E_x \to E_y$ is *weakly continuous* at point $u_0 \in E_x$ if from $u_n \to u_0$ get $F(u_n) \rightharpoonup F(u_0)$, where the symbol $\rightharpoonup$ denotes weak convergence.

Definition 2. Operator F: $E_x \to E_y$ is *compact* if it transforms any bounded set of the space E_x into the relatively compact set of spaces E_y.

Definition 3. Operator F: $E_x \to E_y$ is *completely continuous* if it is continuous and compact.

Definition 4. Operator F: $E_x \to E_y$ is referred to as *strongly continuous* if it converts any weakly converging sequence into a strongly converging one, i.e. $u_n \rightharpoonup u_0 \Rightarrow F(u_n) \to F(u_0)$.

Definition 5. Operator F: $E_x \to E_y$ is referred to as *coercive* if

$$\lim_{R\to\infty} \frac{\langle u, f(u)\rangle}{R} = +\infty \quad (R = \|u\|).$$

It should be mentioned that in reflexive spaces E_x from the strong continuity of the operator F: $E_x \to E_y$ follows that this operator is completely continuous. However, examples are known of nonlinear completely continuous mappings in L_2 which are not strongly continuous.

Theorem 1. *For linear operators, the complete continuity coincides with strong continuity.*

2.1.2. Derivative and gradient of the functional

Let $f(u)$ be a nonlinear functional given on the everywhere dense linear subspace $D(f)$ of the real normalized space E. It is assumed that at the point $u \in D(f)$ for all $h \in D(f)$ the following limit exists:

$$\lim_{t\to 0}\frac{f(u+th)-f(u)}{t}=Vf(u,h);$$

and it is referred to as the *variation of the functional f*. This variation is a homogeneous functional of h, because $Vf(u,\alpha h) = \alpha Vf(u,h)$, but it may not necessarily be an additive functional in respect of h. If the variation $Vf(u,h)$ is a function linear in respect of h, it is referred to as the *Gateaux differential* and we can write $Vf(u,h) = Df(u,h)$. This linear functional of h is written in the form $Df(u,h) = f'(u)h$ and it is said that $f'(u)$ is the *Gateaux derivative* of functional f at point u. Similarly, we introduce the concept of the Gateaux derivative of operator F acting from $D(F)\subset E_x$ into E_y.

If at the given fixed u the derivative $f'(u)$ is a bounded linear functional, i.e. $f'(u)\in E^*$, it is referred to as the gradient of functional f and denoted by grad $f(u)$. In this case, $Df(u,h) = \langle h, \text{grad } f(u)\rangle$. Here $\langle h, \text{grad } f(u)\rangle$ denotes the value of the continuous linear functional $f'(u) = \text{grad } f(u)$ on the vector $h\in D(f)$. Since $D(f)$ is dense in E, then because of continuity, the functional can be extended from $D(f)$ to all E. This extension is also denoted by grad $f(u)$. (Here u is the fixed vector from E). Thus, according to the definition of the gradient we have

$$\langle h, \text{grad} f(u)\rangle = \frac{d}{dt} f(u+th)\Big|_{t=0} = \lim_{t\to 0}\frac{f(u+th)-f(u)}{t},$$

where $\langle h, v\rangle$ is the value of the linear functional $v\in E^*$ on the vector $h\in E$. If grad $f(u)$ is determined for any $u\in \omega\subset E$, then grad $f(u)$ maps ω on some set from E^*.

If the functional $f(u)$ is Gateaux differentiable at every point of the open convex set ω of the normalized space E, then for any u, $u+h\in\omega$ there is such τ, $0<\tau<1$, that the *Lagrangre formula*

$$f(u+h)-f(u)=Df(u+\tau h,h),$$

is valid. In the case of the bounded Gateaux derivative we have

$$f(u+h)-f(u)=\langle h, \text{grad} f(u+\tau h)\rangle,$$

hence, in the case of boundedness of grad $f(u)$ on ω, i.e. if $\|\text{grad } f(u)\| \le c = \text{const}$, from the last formula we obtain the *Lipschitz inequality*

$$\left|f(u+h)-f(u)\right| \le \|\text{grad} f(u+\tau h)\|\,\|h\|,$$

i.e.

$$\left|f(u+h)-f(u)\right| \le c\|h\|, \quad c=\text{const}>0.$$

2.1.3. Differentiability according to Fréchet

Let be E_x and E_y be the normalized spaces and F be the mapping of the open set $\omega \subset E_x$ into E_y.

Definition 6. If at fixed $u \in \omega$ and all $h \in E_x$, for which $u+h \in \omega$, $F(u+h)-F(u) = g(u,h)+\omega(u,h)$, where $g(u,h)$ – the linear operator continuous in respect of h, and

$$\lim_{h \to 0} \frac{\omega(u,h)}{\|h\|} = 0,$$

then $g(u,h)$ is referred to as the *Fréchet differential* of the operator F at point u, and $\omega(u,h)$ is the residual of this differential. We write $g(u,h) = dF(u,h)$. We can also write $dF(u,h)=F'(u)h$, where $F'(u) \in \mathcal{L}(E_x,E_y)$ is referred to as the *Fréchet derivative* of the operator F at point u. Here $\mathcal{L}(E_x,E_y)$ is the space of linear continuous operators from E_x into E_y. If $E_y = \mathbb{R}^1$, then $F = f$ is a functional. In this case, we have $f(u+h)-f(u) = f'(u)h+\omega(u,h)$, where the Fréchet derivative $f'(u)$ as the bounded linear functional is grad $f(u)$.

The definition of the Gateaux and Fréchet differentials shows that if the mapping is Fréchet differentiable, then it is Gateaux differentiable and the derivatives coincide. The reverse claim is not always true. There are mappings which are Gateaux differentiable everywhere but not Fréchet differentiable. Taking this into account, the following theorem is of interest.

Theorem 2. *If the Gateaux derivative F' exists at some vicinity at the point u_0, is bounded and continuous at point u_0 in the sense of the norm of the space $\mathcal{L}(E_x,E_y)$, then at point u_0 there is a Fréchet derivative F' and $DF(u_0,h) = dF(u_0,h)$.*

2.1.4. Derivatives of high orders and Taylor series

Let X,Y be Banach spaces and $F_k(u_1,....u_k)$ is an operator which depends on k variables $u_1,...u_k \in X$, with the values in Y. The operator $F_k(u_1,...,u_k)$ is referred to as the *k-linear* operator if it is linear in respect of each of its arguments u_i with other arguments fixed. Further, the k-linear operator $F_k(u_1,...,u_k)$ is *bounded* if there is a constant $m \in \mathbb{R}$, $m > 0$ such that

$$\|F_k(u_1,...,u_k)\| \le m\|u_1\|...\|u_k\|.$$

The smallest of these constants is the *norm of* F_k and denoted by $||F_k||$. The operator $F_k(u_1,...,u_k)$ linear in respect of every argument is *symmetric* if its values do not change at any permutation of its arguments. Let it be that $F_k(u_1,...,u_k)$ is the k-linear symmetric operator. The operator $F_k(u,...,u)$ is the k-power operator and is denoted by $F_k u^k$. Further, the k-linear operator is written in the form $F_k u_1,...,u_k$.

Let the operator $F(u)$ be differentiable in the vicinity of S, and the differential $dF(u,h)$ is also differentiable at point u_0:

$$F'(u_0+g)h - F'(u_0)h = (Bg)h + \rho(g)h,$$

where $||\rho(g)||=o(||g||)$ as $g \to 0$. In this case, the operator

$$B = F''(u_0) \in \mathcal{L}(X, \mathcal{L}(X,Y))$$

is the *second derivative* of the operator $F(u)$ at point u_0. The definition of $F''(u_0)$ shows that $F''(u_0)hg$ is the bounded bilinear symmetric operator. Further,

$$d^2F(u_0;h) = d\left[dF(u,h);u_0,g\right]|_{g=h},$$

from which $d^2F(u_0;h) = F''(u_0)h^2$ is the quadratic or two-power operator obtained from the bilinear operator at $g=h$.

If $a^nF(u;h) = F^{(n)}(u)h^n$ is already determined, then assuming its differentiability at point u_0, it is assumed that

$$d^{n+1}F(u_0;h) = d\left[d^nF(u,h);u_0,g\right]\Big|_{g=h},$$

from which $d^{n+1}F(u_0;h)=F^{(n+1)}(u_0)h^{(n+1)}$. The operator $F^{(n)}(u_0)h_1,\ldots h_n$ is the n-linear symmetric operator, and $d^nF(u_0,h)$ is the n-power operator.

Let F_ku^k be the k-power operators from X in Y ($k\in N$), and $F_0\in Y$. We form a formal power series

$$\sum_{k=0}^{\infty} F_ku^k \ \ (F_0u^0 = F_0).$$

It is assumed that the numerical series

$$\sum_{k=0}^{\infty} \|F_k\|\|u\|^k,$$

majorising the examined power series, has the convergence radius $\overline{\rho}>0$. Consequently, in any ball S_ρ, where $\rho\in(0,\overline{\rho})$, the original power series converges absolutely and uniformly. Let $\overline{\rho}>0$, and $F(u)$ be the sum of the original power series, i.e.

$$\sum_{k=0}^{n} F_ku^k \to F(u)$$

as $n \to \infty$. Therefore, the operator $F(u)$ is an *analytical operator* at the point $u = 0$.

If the infinitely differentiable operator $F(u)$ is given, the power series

$$\sum_{k=0}^{\infty} \frac{F_k(0)}{k!}u^k$$

is referred to as the *Taylor series of* $F(u)$ at the point $u = 0$. Since the expansion of the operator $F(u)$ into a power series is unique, any power series of the analytical operator is its Taylor series.

2.2. Adjoint nonlinear operators

2.2.1. Adjoint nonlinear operators and their properties

Let E be the reflexive real Banach space and $F : E\to E^*$ is the non-linear mapping continuously Gateaux differentiable and, consequently, Frechét differentiable, i.e. $F'\in\mathcal{L}(E,E^*)$, and for each fixed $u\in E$

$$\|F'(u)-F'(v)\|_{L(E,E^*)} \to 0 \text{ as } \|u-v\| \to 0.$$

It is also assumed that $F(0) = 0$. From this set of the operators $\{F\}$ we separate the class $\mathcal{D}$ which is such that every $F\in\mathcal{D}$ is linked with the mapping G from the same set that is such that $\langle G'(u)v,w\rangle=\langle F'(u)v,w\rangle$ (for

all $u,v,w \in E$). When the last equality is fulfilled, we write [10] $G'(u)=(F'(u))^*$ and say that the G is *an operator adjoint with* F. Thus, the set $\mathcal{D}$ contains all operators continuously Gateaux differentiable, converting to zero at zero and having adjoint operators G with the same properties.

The adjoint operator, if it does exist, is unique. Because of the uniqueness of the adjoint operator G we can write $G = F^*$. For linear operators F and G the last equality coincides with the conventional definition of the adjoint operator.

We now examine the properties of adjoint operators.

1. If F has an adjoint operator, then F^* also has an adjoint operator, and $F^{**}\equiv(F^*)^*=F$. This property follows directly from the definition.

2. If the operator F and G have adjoint operators, then $F+G$ has an adjoint operator, and $(aF+bG)^*=aF^*+bG^*$ for any real numbers a,b.

3. If F has an adjoint operator, then $F^*(u) = \int_0^1 (F'(tu))^* u\,dt$ (for all $u \in E$).

4. If operator F has an adjoint operator, then $\langle F(u),u\rangle=\langle F^*(u),u\rangle$.

Other definitions of the adjoint operator maybe found in [116].

2.2.2. Symmetry and skew symmetry

Operator $F\in\mathcal{D}$ is called *symmetric* if $F = F^*$. If $F\in\mathcal{D}$ and $F^* = -F$, then F is *skew-symmetric*. From the properties 1° and 2° it follows directly that if $F \in \mathcal{D}$ then $(F+F^*)$ is a symmetric operator, and $(F-F^*)$ is a skew-symmetric operator.

Theorem 3. *Operator $F \in \mathcal{D}$ is symmetric iff it is strongly potential, i.e. there is a functional $f\in E^*$ differentiable according to Frechét that $F(u) = grad\ f(u)$.*

Theorem 4. *Operator $F\in \mathcal{D}$ is skew symmetric iff it is linear and for any $u\in E$ the equality $\langle F(u),u\rangle = 0$ is fulfilled.*

2.3. Convex functionals and monotonic operators

Definition 7. The real differentiable functional $f(u)$ given on an open convex set ω of the normalized space E, is referred to as *convex on* ω if for any $u,u_0\in\omega$ the inequality $f(u) = f(u_0)-Df(u_0, u-u_0) \geq 0$, is fulfilled, and it is strictly convex if the equality is possible only at $u = u_0$.

If the differential $Df(u,h)$ is bounded in respect of h, the last inequality has the form $f(u)-f(u_0)-\langle F(u_0), u-u_0)\rangle\geq 0$, where $F(u) = \text{grad}\ f(u)$. Another definition is also known.

Definition 8. If for all u_1, $u_2 \in\omega$ and $\lambda\in\omega$ (0,1) the inequality $f(\lambda u_1+$ $+(1-\lambda)u_2\leq\lambda f(u_1)+(1-\lambda)f(u_2)$ is fulfilled, then f is a *convex* functional.

For the functionals differentiable according to Gateaux, the definition 7 and 8 are equivalent.

Definition 9. The mapping F: $E \to E^*$ is referred to as *monotonic* on the set $\sigma \subset E$ if for any $u,v \in \sigma$ the inequality $\langle u-v, F(u)-F(v)\rangle \geq 0$ is satisfied, and it is strictly monotonic if the equality is fulfilled only at $u = v$.

Theorem 5. *For monotonicity* (*strict monotonicity*) *of* $F(u) = grad\ f(u)$ *on the open convex set* $\omega \in E$ *it is necessary and sufficient that the functional* f *is convex* (*strictly convex*).

Definition 10. The real functional f, given in the normalized space E is *lower semi-continuous* (*weakly semi-continuous*) *at point* u_0 if for any sequence $(u_n) \in E$ such that $u_n \to u_0 (u_n \rightharpoonup u_0)$ we have the inequality $f(u_0) \leq \underline{\lim}_n f(u_n)$.

Functional f is *lower semi-continuous* (*weakly semi-continuous*) on σ if it has this property at every point of σ. As for real functions, it may be shown that the sum of the lower semi-continuous (weakly semi-continuous) functionals represents a lower semi-continuous (weakly semi-continuous) functional.

Theorem 6. *Let the convex functional* f *differentiable according to Gateaux be given on the open convex set* ω *of the normalized space. Consequently, if* $Df(u,h)$ *is continuous in respect of* h, *then* f *is lower weakly semi-continuous on* ω.

We now present a criterion of lower weak semi-continuity of functionals.

Theorem 7. *For the functional* f *given in the normalized space, to be lower weakly semi-continuous, it is necessary and sufficient to fulfil the condition: for any real number* c, *the set* $E_c = \{u : f(u) \leq c\}$ *is sequentially weakly closed.*

Theorem 8. *For weakly lower semi-continuity of the convex functional, given in the Banach space it is necessary and sufficient that this functional is lower semi-continuous.*

The problem of lower semi-continuity is also associated with the concept of the reference functional.

Definition 11. The linear functional $v_0 \in E^*$ is referred to as the reference (or sub-gradient) to the functional f at the point $u_0 \in E$ if $f(u)-f(u_0) \geq \langle v_0,\ u-u_0\rangle$.

From this inequality we obtain a lower semi-continuity f. To formulate the next theorem, we need to use the following concept. Let F be the mapping of a normalized space in another normalized space. If there is a limit

$$V_+F(u,h) = \lim_{t \to +0} \frac{F(u+th) - F(u)}{t}$$

for and $h \in D(F)$, then $V_+F(u,h)$ is the variation of F at the point u in the direction h.

Theorem 9. *If E is the normalized space and the finite convex functional f given on the open convex set* $\omega \subset E$ has the variation $V_+F(u,h)$ *continuous in respect of h for any* $u \in \omega$, *then there is the reference functional at every point* $u \in \omega$.

2.4. Variational method of examining nonlinear equations.

2.4.1. Extreme and critical points of functionals

Let f be the real functional given in the normalized space E. Point $u_0 \in E$ is the extreme point of the functional f if in some vicinity $V(u_0)$ of this point one of the following inequalities is satisfied:

1) $f(u) \le f(u_0)$;

2) $f(u) \ge f(u_0)$ for all $u \in V(u_0)$.

If the second inequality holds for all $u \in E$, then u_0 is the point of the global minimum of the functional f. Further if f is Gateaux differentiable at point u_0, then when fulfilling the condition $Df(u_0,h)=0$ the point u_0 is the critical point of the functional f. Since the fact that the differential is equal to zero indictates that it is continuous in respect of h, then the last equality has the form $\langle \text{grad} f(u_0),h\rangle=0$, and since this equality holds for arbitrary $h \in E$, it may be said that u_0 is the critical point f if $\text{grad} f(u_0)=0$.

Theorem 10. *Let functional f be given in the domain* ω *of the normalized space E and* u_0 *is the internal point of the domain* ω, *in which the linear Gateaux differential exists.*

The following claims are valid in this case.

1. *For the point* u_0 *to be extreme, it is necessary that it is critical, i.e. that*

$$\text{grad } f(u_0) = 0. \tag{3}$$

2. *If additionally at some convex vicinity* $U(u_0)$ *of the point* u_0 *functional f is convex (or grad f(u) is the monotonic operator), then equality* (3) *is necessary and sufficient for the point* u_0 *to be the point of the minimum of functional f.*

Theorem 11. (*Generalized Weierstrasse theorem*). *If the finite lower weakly lower semi-continuous functional* σ *is given on the weakly bounded set in the reflexive Banach space E, then the functional is lower bounded and reaches its lower bound on* σ.

2.4.2. The theorems of existence of critical points

Theorem 12. (Elementary principle of the critical point). *Let* $K=\{u:u \in E, \|u\| \le r, r > 0\}$ *be the ball of the reflexive Banach space E and S is the surface of the ball.*

Therefore, if the lower weakly semi-continuous functional f is Gateaux differentiable on the open ball $\|u\|<r$ and $\inf_{\|u\|=r} f(u) > f(u^*)$, *where* $\|u^*\|<r$, *then at the point of the minimum* $u_0 \in K$ *of this functional* $grad f(u_0)=0$.

Theorem 13. *Let the real functional f, given in the real reflexive Banach space E, be Gateaux differentiable and grad f=F satisfy the conditions:*

1) *Function* $\langle F(tu),u\rangle$ *is continuous in respect of t in* [0,1] *at any* $u\in E$;

2) $\langle F(u+h)\text{-}F(u),h\rangle\geq 0$ (*at all u, h*$\in E$);

3) $\lim_{||u||\to\infty}(\langle F(u),u\rangle/||u||)=+\infty$, i.e. *F is coercive.*

Consequently, there exists a minimum point u_0 *of the functional f and grad* $f(u_0)=0$. *If in the condition* 2) *the equality is possible only at h*=0, i.e. *F is the strictly monotonic operator, then the minimum point of the functional is unique, and f in it has the global minimum.*

Theorem 14. *Let the real functional Gateaux differentiable f, given in the reflexive real Banach space, be lower weakly semi-continuous and satisfies on some sphere* $S=\{u:u\in E,||u||=R>0$ *the condition* $\langle F(u),u\rangle>0$.

Consequently, there is an internal point u_0 *of the ball* $||u||\leq R$ *in which F has the local minimum and, consequently,* grad$f$ (u_0) =0.

Theorem 15. *Let the monotonic potential operator* $F(u)$, *given in the reflexive real Banach space E, satisfy the condition* $\langle F(u),u\rangle\geq||u||\gamma(||u||)$, *where the function* $\gamma(t)$ *is integrable on the segment* [0,R] at R>0 *and*

$$\overline{\lim_{R\to+\infty}}\int_0^R\gamma(t)dt=c>0.$$

Consequently, potential f of the operator F has a minimum point. This minimum point is unique and f in it has the global minimum if F is the strictly monotonic operator.

2.4.3. Main concept of the variational method

Let Φ be a potential operator, i.e. there is a functional φ such that Φ = gradφ. The variational method of proving the existence of the solution of equation $\Phi(u)=0$ reduces to determination of the critical points of the functional φ. For example, if u_0 is the critical point of the function φ, i.e. grad$\varphi(u_0)=0$, then u_0 is the solution of the equation $\Phi(u)=0$. This is why the problem of the unconditional minimum of the functionals is interesting for non-linear equations with potential operators.

If Φ is not a potential operator, the following cases may occur.

1. Equation $\Phi(u)=0$ is replaced by the equivalent equation $\Psi(u)=0$, where Ψ is the potential operator.

2. We find the minimum of the functional $||\Phi(u)||=f(u)$.

2.4.4. The solvability of the equations with monotonic operators

The following claims [10] are proven using the variational methods.

Theorem 16. *Let it be that the potential monotonic operator F, given in the reflexive real Banach space E, satisfies the coercivity condition*

$$\lim_{\|u\|\to\infty}\frac{\langle F(u),u\rangle}{\|u\|}=+\infty,$$

and $\langle F(tu),u\rangle$ *is continuous in respect of* t *on* [0.1] *at any* $u\in E$. *Consequently, equation* $F(u) = v$ *has a solution at any* $v\in E^*$, *i.e.* F *is the surjective mapping* $E\to E^*$ (*mapping of* E *onto* E^*). *If* F *is the strictly monotonic operator, then it represents bijective (surjective and one-to-one) mapping* $E\to E^*$.

Theorem 17. *Let the potential monotonic operator* F, *given in the reflexive real Banach space* E, *satisfy the conditions*:

1) At any $u\in E$ there exists $\int_0^1 \langle F(tu,u)\rangle dt$;

2) $\langle F(u),u\rangle \geq ||u||\gamma(||u||)$, where $\gamma(t)$ *is a function integrable on the segment* $[0,R]$ *at any* $R>0$, *and*

$$\lim_{R\to+\infty}\frac{1}{R}\int_0^R \gamma(t)dt = +\infty.$$

Consequently, the mapping F: $E\to E^*$ *is surjective. If* F *is a strictly monotonic operator, then* F: $E\to E^*$ *is bijective.*

2.5 Minimising sequences

The constructive method of proving the existence of the unconditional minimum of the functional is the construction of the converging minimizing sequences which in turn may be used for constructing an approximate solution of the problem.

2.5.1. Minimizing sequences and their properties

Let E be the real vector space, f be the real functional, given E, and σ be some set on E.

Definition 12. Any sequence $(u_n)\subset\sigma$, satisfying the conditions $\lim_{n\to\infty} f(u_n)=d$, where $d=\inf_\sigma f(u)$, is referred to as minimizing (for f on σ). This definition is also valid when $d=f(u_0)$, where u_0 is the point of the conditional minimum f in relation to σ or the unconditional minimum f, i.e. the point of the local or global minimum f in relation to space E.

The problem of the convergence of the minimising sequences to the point of the unconditional minimum of the functional is important. When examining this problem, an important role is played by convex functionals characterized by some additional properties. If E is a finite-dimensional space, then owing to the fact that any closed ball in this space is compact, it is possible to claim: if the real function f grows along every beam, originating from point u_0, then $f(u)-f(u_0)$ has a monotonic minorant $c(t)$ for $t \geq 0$, $c(0) = 0$, which can be regarded as continuous, strictly increasing and satisfying the condition $f(u)-f(u_0)\geq c(||u-u_0||)$. Of course, if this inequality is also valid for the functional f, then u_0 is the point of the minimum $f(u)$, and any minimizing sequence converges to this point u_0. However, in infinitely dimensional spaces, there are strictly convex functionals having a minimum point u_0 and, consequently, growing along any beam originating from u_0 for which the inequality is not satisfied.

It is necessary to examine the problem of the conditions resulting in boundedness of any minimizing sequence. We shall present such a condition using the following definition.

Definition 13. The finite real functional f, given in the normalized space E, is referred to as increasing if for any number c the set $\{u:u\in E, f(u)\leq c\}$ is bounded.

Theorem 18. *If f is an increasing functional, then any minimising sequence of the functional is bounded.*

The last claim shows that if the increasing functional is given in the reflexive Banach space and $d = f(u_0)$, then from any minimizing sequence $d = \inf f = f(u_0)$ we can separate a subsequence weakly converging to u_0.

2.5.2. Correct formulation of the minimisation problem

The following definition has been proposed by A.N. Tikhonov.

Definition 14. The problem of minimization of the real functional, given on some subset of the normalized space, is formulated correctly if it is solvable, has a unique solution and any minimizing sequence converges to it in the sense of the norm of the space.

The sufficient condition of the correctness of the problem of minimization of functional will now be presented.

Theorem 19. *Let* $\gamma(t)$ *be the non-negative function integrable on* $[0,R]$ *at any* $R>0$ *and such that* $c(R)=\int_0^R \gamma(t)dt$ is increasing *and at some* R: $c(R)>R||F(0)||$.

Therefore, if the operator $F=\text{grad}f$, *given in the reflexive Banach space* E, *satisfies the condition: for any* $h, v\in E$ *the function* $\langle F(v+th),h\rangle$ *is integrable in respect of* t *on* $[0,1]$ *and* $\langle F(v+h)-\langle F(v),h\rangle\geq||h||\gamma(||h||)$, *the problem of minimization of the functional* f *is correctly formulated.*

Below, we examine some methods of constructing the minimizing sequences. Of the known methods of minimization of the non-linear functionals we examine the method of steepest descent, the Ritz method and the Newton method, because these methods are used widely for solving non-linear equations.

3. THE METHOD OF THE STEEPEST DESCENT

3.1. Non-linear equation and its variational formulation

Let F be a potential operator acting from the Banach space E to the adjoint space E^*. This means that there is a functional $f\in E^*$ such that $F=\text{grad}f$.

We formulate the problem of finding solution of the non-linear equation:

$$F(u) = 0. \tag{4}$$

As indicated by theorem 10, this problem is reduced to finding critical

point of the functional f. We formulate the problem of the unconditional minimum of the functional f, namely: find $u_0 \in E$ is such that

$$f(u_0) = \inf_{u \in E} f(u). \tag{5}$$

Formulation of the problem in this form is referred to as the variational formulation.

If operator f is not potential, then instead of (5) we examine the following problem of the minimization of the functional $||F(u)||$: find $u_0 \in E$ such that

$$\|F(u_0)\| = \inf_{u \in E} \|F(u)\|. \tag{6}$$

The variational formulations on the form of (5), (6) are often used for studying and numerical solution of the initial non-linear problem (4). Below, we use for this purpose the method of the steepest descent which makes it possible to find the approximate solutions of the variational formulations.

3.2. Main concept of the steepest descent methods

Let a real non-linear functional f be Gateaux differentiable and lower bounded. We set the $d = \inf_{u \in H}$ and F = grad f. We use the arbitrary vector $u_1 \in H$ and assume that $F(u_1) \neq 0$. Of course, if f is a strictly convex functional, it may have only a unique point on the minimum and its value in it will be equal to d; therefore, the requirement $F(u_1) \neq 0$ shows that u_1 is not the point of the minimum of f. We select the vector $h \in H$ in such a manner that its length is equal to $||h||=||F(u_1)||$ and explain how the direction h can be selected for the derivative

$$\frac{d}{dt} f(u_1 + th) = \langle F(u_1 + th), h \rangle$$

to have the lowest value at t=0, i.e. that h is the direction of the fastest decrease of $f(u)$ at the point u_1. For this purpose, we initially select the direction h in such a manner that $(F(u_1),h)$ has the highest value, and then change the sing of the vector h so that $(F(u_1),-h)$ has the lowest value. Since $(F(u_1), h\rangle \leq ||F(u_1)||\ ||h||=||F(u_1)||^2$ then $(F(u_1),-h)$ has the highest values only at $h = F(u_1)$ and the lowest value at $h = -F(u_1)$, i.e. when the direction h coincides with the direction of the anti-gradient f. We set $h_1 = -F(u_1)$ and examine the real function $\varphi(t) = f(u_1+th_1)$, $t \geq 0$. The function $\varphi(t)$ decreases in some right half-vicinity of the point t=0. Let it be that there is $\min\varphi(t)$ and t_1 is the lowest positive value of t for which $\varphi(t_1) = \min\varphi(t_1)$. We set $u_2=u_1+t_1h_1=u_1-t_1 \geq 0$. According to construction $f(u_2) = f(u_1+t_1h_1) < f(u_1)$. Assuming that $F(u_2) \neq 0$, we can repeat previous considerations. Thus, if $F(u_k) \neq 0$ is fulfilled for every k, we obtain the following process:

$$u_{n+1} = u_n - t_n F(u_n) \quad (n=1,2,3,...), \tag{7}$$

which is referred to as the process of the steepest descent or the process of the gradient descent. Although the sequence (7) is such that

$$f(u_{n+1}) < f(u_n) \quad (n = 1,2,3,...), \tag{8}$$

it may and may not be minimizing.

Definition 15. Any sequence $\{u_n\}$, for which the inequality (8) is fulfilled, is referred to as a relaxation sequence, and numbers t_n in (7) are referred to as relaxation multipliers.

It should be mentioned that even in the case of the Hilbert space there are difficulties when determining relaxation multipliers. They can be calculated efficiently only in the simplest cases. Consequently, the numbers t_n are replaced by the positive numbers ε_n which should be given either a priori or boundaries within which they can change arbitrarily should be specified for them. When t_n is replaced by ε_n (7) is referred to as the process of the descent type (gradient type).

We examine a more general case in which the real functional f, Gateaux differentiable and lower bounded is given in the reflexive real Banach space E. Let F=grad f and u_1 be an arbitrary vector from E such that $F(u_1)\neq 0$. We select vector $h_1 \in E$ such that $h_1 = -UF(u_1)$, where U is the operator acting from E^* into E and satisfying the conditions $||Uv||=||v||$, $\langle v, Uv\rangle=||v||^2$. Consequently, the process of the steepest descent has the form

$$u_{n+1} = u_n - t_n UF(u_n) \ \ (n=1,2,3,...),$$

where t_n is the lowest positive value of t for which $\varphi(t_n)$ = min $\varphi(t)$ at $\varphi(t) = f(u_n+th_n)$.

Usually t_n is not calculated and we examine the process

$$u_{n+1} = u_n - \varepsilon_n UF(u_n). \tag{9}$$

The restrictions imposed on selection of ε_n are such that the process of the type of the steepest descent converges.

3.3. Convergence of the method

The following claim holds

Lemma 1. *Let the real functional f, given in reflexive real Banach space E, be Gateaux differentiable and its gradient F satisfy the condition*

$$\langle F(u+h)-F(u),h\rangle \le M\|h\|^2 \text{ for } u,u+h\in D_r=\{u\in E:\|u\|\le r\}, \tag{10}$$

where $M = M(r)$ is the arbitrary positive increasing function on the half-axis $r \ge 0$ and the norm E^* of the space is Gateaux differentiable. Consequently, if $\varepsilon_n M_n \le 1/2$, where $M_n = max[1, M(R_n)]$, $R_n=||u_n||+||F(u_n)||$, the process (9) *is a relaxation process.*

It should be mentioned that to fulfil the inequality (10) it is sufficient that the operator F satisfies the Lipschitz condition

$$\|F(u+h)-F(u)\| \le M(r)\|h\|, \tag{11}$$

and the Lipschitz constant may also be the increasing function of $r = ||u||$.

Lemma 2. *Let the following conditions be satisfied:*

1) *E is the reflexive real Banach space, and the norm in E* is Gateaux differentiable;*

2) *The real functional f on E, Gateaux differentiable, is lower bounded and is increasing* (*or the set* E_1 *is bounded*), *and its gradient satisfies the Lipschitz condition* (11)*;*

3) *The relaxation multiplies* ε_n *satisfy inequalities* $1/4 \le \varepsilon_n M_n \le 1/2$.

Consequently, the iteration process (9) *will be a relaxation process and*

$$\lim_{n\to\infty} F(u_n) = 0.$$

It should be mentioned that in addition to the conditions of lemma 2 we require that $f(u)-d \le \|F(u)\|^{\alpha}$ ($\alpha>0$), where $d = \inf f(u)$, the sequence u_n will be minimizing.

Theorem 20. *Let the following conditions be satisfied:*

1) *E is the reflexive real Banach space, and the norm in E* is Gateaux differentiable;*

2) *The real function f given on E, is Gateaux differentiable and its gradient F has a property according to which* $\langle F(tu),h\rangle$ *is the function integrable in respect of* $t \in [0,1]$ *and satisfying the inequalities*

$$\|F(u+h)-F(u)\| \le M(r)\|h\|,\ u, u+h \in K_r,\ (K_r = \{u : u \in E, \|u\| \le r\}),$$

$$\langle F(u+h)-F(u), h\rangle \ge \|h\|\,\gamma(\|h\|),$$

where $M(r)$ *is the continuous increasing non-negative function given for* $r \ge 0$, *and* $\gamma(t)$, $t \ge 0$, *is the increasing continuous function such that* $\gamma(0)=0$, *and the function*

$$c(R) = \frac{1}{R}\int_0^R \gamma(w)\,dw$$

is increasing also, and at some R, the inequality $c(R) > \|F(0)\|$ *holds*;

3) $1/4 \le \varepsilon_n M_n \le 1/2$, $M_n = \max[1, M(R_n)]$, $R_n = \|u_n\| + \|F(u_n)\|$.

Consequently, the sequence $u_{n+1} = u_n - \varepsilon_n UF(u_n)$ is a relaxation sequence minimising and converging to the point of the global minimum of the functional f.

In the conditions of theorem 20 the method of the steepest descent (9) may be used to find the approximate solution of the problem (4). We stop the process (9) at $n = N$ and u_N will be treated as the *N-th approximation* to the exact solution u_0 of the problem. According to theorem (20) for any $\varepsilon > 0$ there is a number N such that $\|u_N - u_0\| < \varepsilon$. This means that using algorithm (9) we can approach the exact solution of the non-linear problem (4) with any accuracy specified in advance.

4. THE RITZ METHOD

Let F be the potential operator acting from the Banach space E into adjoint E^*, i.e. there is a functional $f \in E^*$ such as $F = \text{grad } f$.

It is assumed that we examine a problem of mathematical physics which reduces to the non-linear operator equation $F(u)=0$. As indicated by section 3.1. this problem may be formulated in the variational formulation (5): find $u_0 \in E$ such that $f(u_0) = \inf_{u\in E} f(u)$. In this section it is assumed that

E is a separable real normalized space, and f is a finite real functional given on E. The completeness of E is required only in the second half of the section. To minimize the functional f, if it is lower bounded E, we use the Ritz method. W. Ritz used his method to solve specific problems. Subsequently, his method was developed by S.G. Mikhlin and other authors.

The following definition is required when formulating different assumptions regarding the Ritz method.

Definition 16. Functional f is referred to as upper (lower) semi-continuous at the point $u_0 \in E$ if for every $\varepsilon>0$ there exists $\delta>0$ such that for $\|u-u_0\|<\delta$

$$f(u_0)-f(u)>-\varepsilon\,(f(u_0)-f(u)<\varepsilon).$$

Functional f is referred to as upper (lower) semi-continuous on the set $M \subset E$ if it is upper (lower) semi-continuous at every point $u \in M$.

4.1. Approximations and Ritz systems

Let f be a real functional lower bounded and given on the normalized space E. The Ritz method of minimization of functional f may be described as follows. Initially, we specify the so-called co-ordinate system in E, i.e. a linearly independent system of the vectors $\varphi_1, \varphi_2,\ldots,\varphi_n,\ldots$, and the set of all possible combinations of these vectors is dense in E. This is followed by constructing a sequence of finite-dimension sub-spaces $\{E_n\}$, where E_n is the n-dimensional space – the span of the vectors $\varphi_1, \varphi_2,\ldots,\varphi_n$. From the lower boundedness of f on E it follows that f is lower bounded on E_n. Let it be that $d_n = \inf_{u\in E_n} f(u)$. According to construction, $d_1 \geq d_2 \geq d_3 \geq \ldots \geq d_n \ldots$. It is assumed that at every n there is $u_n \in E_n$ such that $f(u_n) = d_n$. Consequently

$$u_n = \sum_{k=1}^{n} a_k \varphi_k, \tag{12}$$

where coefficients a_k depend on n. Vectors u_n are referred to as the Ritz approximations.

Let f be Gateaux differentiable on E and F = grad f. Consequently, f is Gateaux differentiable also on E_n, and for arbitrary vectors u, $h \in E_n$, i.e. for $u = \sum_{k=1}^{n} \alpha_k \varphi_k, h = \sum_{k=1}^{n} \beta_k \varphi_k$, where α_k and β_k are arbitrary, we have

$$\frac{d}{dt} f(u+th)\Big|_{t=0} = \left\langle F(u), h\right\rangle = \sum_{i=1}^{n} \beta_i \left\langle F\left(\sum_{k=1}^{n} \alpha_k \varphi_k\right), \varphi_i\right\rangle.$$

Therefore, according to theorem 10 we obtain that if u_n is the point of the global minimum f on E_n, then

$$\left\langle F\left(\sum_{k=1}^{n} a_k \varphi_k\right), \varphi_i\right\rangle = 0 \quad (i = 1, 2, \ldots, n). \tag{13}$$

System (13) determining the coefficient a_k is the Ritz system.

It should be mentioned that if f is convex, any solution of system (13) gives, using equation (12), the Ritz approximation, i.e. the point of the

global minimum of f on E_n.

Lemma 3. *If the convex functional f, given on the linear space (not necessarily normalized) has two different minimum points, then its values at these points coincide.*

4.2. Solvability of the Ritz systems

If the convex Gateaux differentiable functional is given on E, then to ensure that the vector u_n given by equation (12) represents the Ritz approximation, it is necessary and sufficient that the coefficients a_k satisfy the Ritz system (13). If f is a functional strictly convex on E, then this functional will also be strictly convex on $E_n \subset E$ and therefore the system (13) cannot have more than one solution.

The solution $(a_1, a_2,\ldots,a_n)$ of the system (13) may give, using equation (12) the critical point of f, i.e. the point at which the gradient of f converts to zero, and if this is only the point of the global minimum of f on E_n, the solution of the system (13) determines the Ritz approximation. However, if f is a convex functional, then any critical point of the functional will also be the point of the global minimum (see theorem 10).

The solvability of the Ritz system (13) can be examined assuming that the real functional f characterized by rapid increase, (i.e. $\lim_{||u||\to\infty} f(u)=+\infty$) and Gateaux differentiable and given on E, is lower semi-continuous at in every sub-space $E_n \subset E$.

Since f is a rapidly increasing functional, we find $r>0$ such that $f(u)>f(0)$, as soon as $||u||>r$. In E_n we examine a ball $K_r^n = \{u{:}u\in E_n,\ ||u||<r\}$. Because of lower semi-continuity of f on K_r^n (according to the well-known theorem of classic analysis), there is a point $u_0\in K_r^n$ at which f has the least value, i.e.

$$f(u_0) = \inf_{u\in K_r^n} f(u),\ u_0 = \sum_{k=1}^{n} a_k^{(0)}\varphi_k.$$

This point may not belong to the surface of the ball K_r^n because here $f(u)>f(0)$. However, outside the ball K_r^n we have $f(u) > f(0) \geq f(u_0)$. Consequently, $f(u_0) = \inf_{u\in E_n} f(u)$ and $a_1^{(0)}, a_2^{(0)},\ldots,a_n^{(0)}$ satisfy (13).

Lemma 4. *If the real rapidly increasing and Gateaux differentiable functional f given on E is lower semi-continuous in every sub-space $E_n\subset E$, the Ritz system* (13) *is solvable at any n.*

Lemma 5. *If the real functional f with the gradient F given on E and Gateaux differentiable is lower semi-continuous on every sub-space $E_n\subset E$, and at some $r > 0$*

$$\langle F(u),u\rangle>0 \quad \text{for} \quad ||u||=r, \tag{14}$$

then the Ritz system (13) *is solvable at any n.*

4.3. Convergence of the Ritz method

Let the Ritz approximations (12) for the functional f, given on E and lower bounded, exist at any n, and let $f(u)$ be upper semi-continuous. Let $d = \inf_{u \in E_n} f$ and $(u^{(n)}) \subset E$ be a minimising sequence satisfying the inequalities

$$f\left(u^{(n)}\right) \le d + 1/n \; (n = 1, 2, 3, \ldots).$$

Because of the completeness of the coordinate system $\varphi_1, \varphi_2, \ldots, \varphi_n$, every vector $u^{(n)}$ and the positive number δ_n are related to the vector $v^{(m)} \in E_m$ such that

$$v^{(m)} = \sum_{k=1}^{m} a_k^{(m)} \varphi_k \; (m = m(n) \ge n), \; \left\| u^{(n)} - v^{(m)} \right\| \le \delta_n.$$

Since f is upper semi-continuous, then δ_n can be selected so small that for the arbitrary vector $w \in E$ satisfying the inequality $||u^{(n)} - w|| \le \delta_n$, we have $f(u^{(n)}) - f(w) \ge -1/n$. Assuming that $w = v^{(m)}$ from here and from the previous we obtain that

$$f\left(v^{(m)}\right) \le f\left(u^{(n)}\right) + 1/n \le d + 2/n.$$

this inequality shows that $(v^{(m)})$ is the minimising sequence. However, since for the Ritz approximations (12) $f(u_m) = d = \inf_{u \in E_m} f(u)$, then $f(u_m) \le f(v^{(m)}) \le d + 2/n$. Taking into account that $f(u_m) \ge d$, we conclude that the $\lim_{n \to \infty} f(u_n) = d$.

It follows from here:

Lemma 6. *Let the Ritz approximations* (12) *for the functional f, given on E and lower bounded, exist at any n. Therefore, if f*(*u*) *is upper semi-continuous, then its Ritz approximations form a minimising sequence.*

In the conditions of lemma 6 the Ritz method can be used to find the approximate solution of the problem $F(u) = 0$. Vector u_n at $n=N$ will be regarded as the *N-th approximation* of the exact solution u_0 of the problem. According to lemma 6 for any $\varepsilon > 0$ there is number N such that $||u_N - u_0|| < \varepsilon$. This shows that using the Ritz method we can find the approximate solution u_N with any accuracy given in advance.

5. THE NEWTON–KANTOROVICH METHOD

5.1. Description of the Newton iteration process

I. Newton proposed the effective method of calculating solution of the equation

$$F(u) = 0 \tag{15}$$

for the case of function $F(u)$ with real values depending on the real variable u. Subsequently, the Newton method was transferred to the systems of equations (when $F(u) \in \mathrm{R}^m$) and subsequently generalized in studies of L.V. Kantorovich on equations in the Banach spaces.

Let $F(u)$ be a non-linear operator defined in the vicinity S of the solution u^* of equation (15) and continuously Fréchet differentiable on S. Let the operator $F'(u)$ be continuously invertible on S. The Newton iteration process

may be described as follows. We select the initial approximation $u_0 \in S$, quite close to u^*. Further approximations u_n, $n = 1,2,\ldots$, should be calculated from the equation

$$u_n = u_{n-1} - \left[F'(u_{n-1})\right]^{-1} F(u_{n-1}),\ n = 1,2,\ldots \tag{16}$$

At present, the Newton method is one of the most widely used computational methods. Its main advantage is (under specific assumptions) a very rapid convergence of the successive approximation (16) to the solution u^*. The method is also suitable for cases in which the equation (15) has several solutions.

Let u, $F(u) \in \mathbb{R}^m$. If $f_i(x_1,\ldots,x_m)$, $i=1,\ldots,n$, are the coordinate functions $F(u)$, the equation (15) is a short form of writing equations $f_i(x_1,\ldots,x_m) = 0$, $i=1,\ldots,m$.

The derivative $F'(u) = \left(\dfrac{\partial f_i(u)}{\partial x_j}\right)_{i,j=1}^{m}$ is the Jacobi matrix, and $[F'(u)]^{-1}$ is the inverse matrix. Thus, equation (16) represent the matrix form of the Newton iteration process in $\mathbb{R}^m$.

5.2. The convergence of the Newton iteration process

We present one of the most suitable variants of the theorems of the Newton method. In this case, the existence of the solution u^* is not assumed but it is proven. The problem of the uniqueness of the solution in the examined ball is not discussed here.

Theorem 21. *Let in the ball $S_r(u_0)$ the operator $F'(u)$ be differentiable, and its derivatives satisfy in this ball the Lipschitz condition with constant l. Let in $S_r(u)_0$ the operator $F'(u)$ be continuously invertible and there is a constant $m>0$ such that in $S_r(u_0)$*

$$||[F'(u)^{-1}]|| \leq m. \tag{17}$$

Let also $\|F(u_0)\| \leq \eta$.

Therefore, if $q = m^2 l\eta/2<1u$ and

$$r' = m\eta \sum_{k=0}^{+\infty} q^{2^k-1} < r, \tag{18}$$

the equation $F(u)=0$ has the solution $u^* \in \overline{S_{r'}(u_0)}$, to which the Newton iteration process, starting from u_0, converges. The rate of convergence of u_n to u^* is given by the inequality

$$\|u_n - u^*\| \leq m\eta \frac{q^{2^k-1}}{1-q^{2^n}}.$$

5.3. The modified Newton method

We examine the modification of the *Newton iteration process*:

$$u_{n+1} = u_n - \left[F'(u_0)\right]^{-1} F(u_n),\ n = 1, 2, \ldots \tag{19}$$

The advantage of equation (19) is that the calculations are simplified (the inverse operator is calculated only once). As shown below, the shortcoming of the equation (19) is that the rate of convergence decreases in comparison with the Newton iteration method.

Theorem 22. *Let the operator $F(u)$ be differentiable in the ball $S_r(u_0)$, and its derivatives satisfy in $S_r(u_0)$ the Lipschitz condition with constant l. Let $F'(u_0)$ be continuously invertible and $\|[F'(u_0)]^{-1}\| \le m$. Let it also be that $\|[F(u_0)]\| \le \eta$.*

Therefore, if $2m^2 l\eta < 1$ and

$$r' = \frac{1 - \sqrt{1 - 2m^2 l\eta}}{ml} < r, \tag{20}$$

the equation $F(u)=0$ has the solution $u^* \in \overline{S_{r'}(u_0)}$, to which the modified Newton iteration process (19), starting from u_0, converges. The rate of convergence of u_n to u^* is given by the inequality

$$\left\|u_n - u^*\right\| \le \frac{\left(1 - \sqrt{1 - 2m^2 l\eta}\right)^n}{\sqrt{1 - 2m^2 l\eta}} m\eta.$$

In the conditions of the theorems 21, 22, the Newton method or its modification can be used for the approximate solution of the non-linear equation $F(u)=0$. The process (16) will be stopped at $n = N$ and u_N will be referred to as the *N-th approximation* of the exact solution u^*. According to theorem 21 for any $\varepsilon > 0$ there is the number N (for example any N satisfying the inequality $m\eta \frac{q^{2^n - 1}}{1 - q^{2^n}} < \varepsilon$) such that $\|u_N - u^*\| < \varepsilon$. This means that using the iteration process (16) we can approach the exact solution of the non-linear problem $F(u)=0$ with any accuracy given in advance.

6. THE GALERKIN–PETROV METHOD FOR NON-LINEAR EQUATIONS

6.1. Approximations and Galerkin systems

Let E be a separable normalized space with the basis and F is the mapping of E into E^*. The Galerkin method of approximate solution of the equation $F(u)=0$ may be described as follows. Initially, in space E we specify the basis, i.e. the linearly independent system of the vectors φ_1, φ_2, φ_3, ..., having the properties that any vector from E is represented in the unique manner in the form $u = \sum_{k=1}^{\infty} \alpha_k(u)\, \varphi_k$. Using this basis, we construct the sequence of finite-dimensional spaces (E_n), where E_n is the n-dimensional span of the vectors $\varphi_1, \varphi_2, \ldots, \varphi_n$.

The Galerkin approximation of the solution of the equation $F(u) = 0$ is the vector $u_n \in E_n$, i.e.

$$u_n = \sum_{k=1}^{\infty} a_k \varphi_k, \tag{21}$$

satisfying the system of equations

$$\left\langle F\left(\sum_{k=1}^{\infty} a_k \varphi_k\right), \varphi_i \right\rangle = 0 \quad (i = 1, 2, ..., n). \tag{22}$$

This system whose solution determines the Galerkin approximation u_n, is the Galerkin system. It should be mentioned that regardless of the fact that the system (22) coincides with the system (13), the problem of its solvability is solved by a different procedure. In the fourth section of this chapter we used the properties of the functional f and when examining the system (22) we used the properties of the mapping F.

6.2. Relation to projection methods

Let P_n be the operator of projection of E on E_n, and P_n^* is the adjoint operator which, as is well known, projects E^* on the n-dimensional subspace E_n^*. The projection method of the approximate solution of equation $F(u) = 0$, where F is the mapping of E in E^*, may be expressed by the fact that the given equation is replaced by an equation in the finite-dimensional space

$$P^*_n F(P_n u) = 0, \tag{23}$$

and the solution of the last equation is referred to as the approximate solution of the initial equation. We shall show that equation (23) is equivalent to the system (22). In fact, if h is an arbitrary vector of E^*, then equation (23) is equivalent to the equation

$$0 = \langle P_n^* F(P_n u), h\rangle = \langle F(P_n u), P_n h\rangle =$$

$$= \left\langle F\left(\sum_{k=1}^{n} a_k \varphi_k\right), \sum_{i=1}^{n} \beta_i \varphi_i \right\rangle = \sum_{i=1}^{n} \beta_i \left\langle F\left(\sum_{k=1}^{n} a_k \varphi_k\right), \varphi_i \right\rangle.$$

This equation, because of the arbitrary nature of β_i, is equivalent to the system (22). Thus, the Galerkin method of solution of the equation $F(u)=0$ examined here coincides with the projection method of solving this equation.

It should also be mentioned that if P is the mapping of E_x in E_y, where E_x and E_y are the normalized space, then the projection method of solving the equation $P(u)=0$ may be described as follows. We specify two sequences of the sub-spaces $\left(E_x^{(n)}\right)$ and $\left(E_y^{(n)}\right)$ (where n is the indicator of the dimension), the union of these sub-spaces is dense in E_x and E_y, respectively, and also the sequences of the projectors (P_n) and (Q_n), $P_n E_x = \left(E_x^{(n)}\right)$, $Q_n E_y = \left(E_y^{(n)}\right)$. Subsequently, the equation $P(u)=0$ is replaced by the equation

$$Q_n P(P_n u) = 0, \tag{24}$$

whose solution is regarded as the approximate solution of the initial equation.

If $E_x = E_y = H$, where H is the Hilbert space, the projection method is

the Bubnov–Galerkin method; in the general case, the projection method is referred to as the Galerkin–Petrov method.

To conclude this section it should be mentioned that because of (24), if F is the mapping of E in E, the Galerkin approximation for the equation $F(u)=0$ are found from the equation in the finite-dimensional sub-space

$$P_n F\left(P_n u\right)=0. \tag{25}$$

6.3. Solvability of the Galerkin systems

Let $F:E\to E^*$ be a monotonic continuous operator, where E is the real separable normalized space satisfying the following condition of the sphere $||u||=r>0$

$$\langle F(u),u\rangle>0\ \left(\|u\|=r\right). \tag{26}$$

In these conditions we can show the solvability of the system (22). Because of the equivalence of the systems (22) and (23) it is sufficient to show that the equation (23) has a solution. We set $P_n u = w\in E_n$. Therefore, because of the monotonic nature of F and the finite dimension of E_n, $\Phi_n(w)=P_n^* F(w)$ is the continuous mapping of E_n in $E_n^*=P_n^* E^*$, i.e. the continuous mapping of the n-dimensional space in the n-dimensional space. Identifying E_n and E_n^*, we find that Φ_n is the continuous mapping of the n-dimensional space in itself. Further, since at $||w||=r$

$$\langle\Phi_n(w),w\rangle=\langle P_n^* F(w),w\rangle=\langle F(w),w\rangle>0,$$

on the ball $||w||=r$ we have $||\Phi_n(w)||>0$. It follows from this that equation $\Phi_n(w)=0$ has a solution belonging to the ball $||w||<r$.

Lemma 7. *Let the monotonic continuous operator F: $E\to E^*$, where E is the real separable normalized space, satisfy the condition* (26) *on the ball* $||u||=r>0$.

Consequently, the Galerkin system (22) *is solvable at any n and the Galerkin approximations u_n satisfy the inequality $||u_n||<r$.*

6.4. The convergence of the Galerkin–Petrov method

Here, we assume that u_n is the Galerkin approximation of the solution of the equation $F(u)=0$ where F is the mapping of the normalized space E in E^*. In this case, the approximations u_n are determined from the system (22) or equation (23).

Lemma 8. *Let the Galerkin approximation u_n, satisfying the systems* (22), *exist at any n, and $||u_n||\le r$. Consequently, if F is a bounded operator, the sequence $(F(u_n))$ converges E-weakly to zero.*

Let E be a separable normalized space with the basis φ_1, φ_2, φ_3, ...; (E_n) be a sequence of the sub-spaces, examined in paragraph 6.1; (P_n) be the sequence of the projectors ($P_n E = E_n$) and (P_n^*) be the sequence of adjoint projectors. As mentioned in paragraph 6.2, P_n^* projects E^* on the n-dimensional sub-space E_n^*.

Let h be an arbitrary vector of E^*, h_n $P_n^* h$ and $h^{(n)}=h-h_n$. It is said that

the sequence $\{E_n^*\}$ is limitary dense in E^* if for any $h^{(n)}\to 0$ at $n\to\infty$.

Lemma 9. *Let Galerkin approximations u_n, satisfying equation* (25), *exist at any n and* $||u_n||\le r$ = const. *Therefore, if $F:E\to E$ is a bounded operator and the sequence $\{E_n^*\}$ is limitary dense in E^*, then the sequence $(F(u_n))$ converges weakly to zero.*

We shall use the following definition.

Definition 17. The mapping $G:E\to E^n$ is referred to as uniformly monotonic if $\langle G(u)-G(v), u-v\rangle\ge||u-v||\gamma(||u-v||)$, where t $\gamma(t)$ is the increasing real function, converting to zero at zero.

The following theorem of convergence holds.

Theorem 23. *Let E be the reflexive real Banach space with the basis $\{\varphi_k\}$ and the continuous uniformly monotonic and bounded operator $F:E\to E^*$ satisfy the inequality $\langle F(u),u\rangle>0$ if $||u||\ge r>0$.*

Consequently, Galerkin approximations u_n exist at any n and converge to the unique solution u_0 of the equation $F(u) = 0$.

In the conditions of the theorem 23, the Galerkin–Petrov method may be used to find the approximate solution of the equation $F(u)=0$. The vector u_n at $n = N$ from (21) will be referred to as the *N-th approximation* of the exact solution u_0. According to theorem 23, for any $\varepsilon>0$ there is the number N such that $||u_N-u_0||<\varepsilon$. Thus, using the Galerkin–Petrov method, we can find the approximate solution u_N with any accuracy given in advance.

7. PERTURBATION METHOD

One of the powerful methods of solving non-linear problems of mathematical physics is the perturbation method or the method of the small parameter. The mathematical perturbation theory was formulated in studies by A. Poincaré and A.M. Lyapunov. In the form in which the perturbation algorithms are used in problems of eigenvalues, they were developed in studies by Rayleigh and Schrödinger. The mathematically strict theory of perturbations evidently started with studies of F. Rellich. The mathematical perturbation theory was developed further in studies by K.O. Friedrichs, T. Kato, N.N. Bogoluybov and Yu.A. Mitropol'skii, A.B. Vasil'eva and B.F. Butuzov, M.I. Vishik and L.A. Lyusternik, B. Sz. Nagy, J.L. Lions, S.A. Lomov, N.N. Moiseev, V.P. Maslov, V.A. Trenogrin, R.Bellman, A.N. Filatov, M.D. Van Dyke, and many others. These studies continued the development of the perturbation theory applied to a large number of mathematical physics problems. However, the joint concept of all these studies was usually the possibility of expanding the solution in respect of the small parameter and justification of the convergence of the resultant series to the exact solution of the problem [116].

7.1. Formulation of the perturbation algorithm

Let X and Y be the Hilbert spaces. It is assumed that X is imbedded densely

into Y and continuously. We examine a non-linear operator $\Phi(u,\varepsilon)$ acting from X into Y and dependent on the numerical parameter ε, where $\varepsilon \in [-\bar{\varepsilon}, \bar{\varepsilon}]$, $\bar{\varepsilon} > 0$. The domain of definition $D(\Phi)$ of this operator is assumed to be a linear set dense in X. Let operator Φ at any fixed ε have a continuous Gateaux derivative $\Phi'(u,\varepsilon) \equiv \partial\Phi/\partial u$ at every point $u \in D(\Phi)$, and Φ' is regarded as the operator from X in Y. It is also assumed that the domain of definition $D(\Phi')$ of the operator Φ' contains $D(\Phi)$.

We examine the equation

$$\Phi(U,\varepsilon) = 0 \tag{27}$$

which will be referred to as a perturbed problem. We fixed the element $U_0 \in D(\Phi)$, setting $f(\varepsilon) \equiv -\Phi(U_0,\varepsilon)$ and transfer from (27) to the equation

$$A(\varepsilon)u = f(\varepsilon) \tag{28}$$

where

$$A(u,\varepsilon) = \int_0^1 \Phi'(U_0 + tu, \varepsilon)\,dt, \quad u = U - U_0.$$

The operator $A(u,\varepsilon)$ acts from X in Y with the definition domain $D(A)=D(F)$. The adjoint operator has the form

$$A^*(u,\varepsilon) = \int_0^1 \left(\Phi'(U_0 + tu, \varepsilon)\right)^* dt, \quad u \in D(F).$$

this operator is the operator from Y^* into X^*; its definition domain is denoted by $D(A^*)$. The operator A^* is referred to as the adjoint operator corresponding to $\Phi(u,\varepsilon)$; it is one of the adjoint operators which can be introduced when examining equation (27).

In addition with (28), we examine the adjoint equation

$$A^*(u,\varepsilon)u^* = g(\varepsilon), \tag{29}$$

where the element $g(\varepsilon) \in X^*$ is analytical in respect of ε:

$$g(\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i g_i, \quad g_i = \frac{1}{i!}\frac{d^i g}{d\varepsilon^i}\bigg|_{\varepsilon=0}, \quad g_i \in X^*.$$

it will also be assumed that the operator $\Phi(U,\varepsilon)$ is analytical in respect of all its variables, and equation (28) has a unique solution, presented in the form of a series in respect of the powers of ε, converging at $|\varepsilon| < \bar{\varepsilon}, \bar{\varepsilon} > 0 : u = \sum_{i=0}^{\infty} \varepsilon^i u_i$. As U_0, we consider the solution of the equation

$$\Phi(U_0,0) = 0,$$

which will be referred to as the unperturbed problem. Therefore, $u_0=0$, and

$$u = \sum_{i=1}^{\infty} \varepsilon^i u_i \tag{30}$$

is the solution of the equation (28).

The perturbation algorithm for solving the problem (28) consists of consecutive finding corrections u_i in the expansion (29). To determine the type of equation for u_i (i=1,2,...) we expand $f(\varepsilon)$ into a series in respect of the powers of the parameter ε;

$$f(\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i f_i = \sum_{i=1}^{\infty} \varepsilon^i f_i,$$

where

$$f_0 = f(0) = 0, \qquad f_i = \frac{1}{i!}\frac{d^i f}{d\varepsilon^i}\bigg|_{\varepsilon=0}, \quad i = 1, 2, \ldots$$

In particular,

$$f_1 = -\frac{\partial}{\partial\varepsilon}\Phi\left(U_0\,\varepsilon\right)\Big|_{\varepsilon=0}, \; f_2 = -\frac{1}{2}\frac{\partial^2\Phi(U_0,\varepsilon)}{\partial\varepsilon^2}\bigg|_{\varepsilon=0}$$

($\frac{\partial\Phi}{\partial\varepsilon}$ and $\frac{\partial^2\Phi}{\partial\varepsilon^2}$ are the partial derivatives of $\Phi(U,\varepsilon)$, in respect of ε at fixed U). Substituting (30) and (28) and cancelling one in respect of ε, we obtain the equation

$$A\left(\sum_{i=1}^{\infty}\varepsilon^i u_i, \varepsilon\right)\left(\sum_{i=1}^{\infty}\varepsilon^{i-1}u_i\right) = \sum_{i=1}^{\infty}\varepsilon^{i-1} f_i. \tag{31}$$

From this at ε=0 we obtain an equation for u_1. Subsequently, differentiating consecutively equation (31) in respect of ε and assuming ε = 0, we obtain an infinite system of equations for determining u_i:

$$A_0 u_1 = f_1,$$

$$A_0 u_2 = f_2 - A_1(U_0, u_1)u_1,$$

$$A_0 u_3 = f_3 - A_1(U_0, u_1)u_2 - \frac{1}{2}A_2(U_0, u_1, u_2)u_1, \tag{32}$$

.....

where

$$A_0 = A(0,0) = \Phi'(U_0, 0),$$

$$A_1(U_0, u_1) = \frac{d}{d\varepsilon}A\left(\sum_{i=1}^{\infty}\varepsilon^i u_i, \varepsilon\right)\bigg|_{\varepsilon=0} = \int_0^1 \Phi''(U_0, tu_1, 0)\,dt + \frac{\partial}{\partial\varepsilon}\Phi(U_0, 0),$$

$$A_2(U_0, u_1, u_2) = \frac{d^2}{d\varepsilon^2}A\left(\sum_{i=1}^{\infty}\varepsilon^i u_i, \varepsilon\right)\Big|_{\varepsilon=0}.$$

Equations (32) form the principle of the perturbation algorithm for finding corrections u_i. Solving successively these equations, we obtain

$$U = U_0 + \sum_{i=1}^{\infty}\varepsilon^i u_i.$$

The element

$$U_{(N)} = U_0 + \sum_{i=0}^{N}\varepsilon^i u_i \tag{33}$$

is the approximation of the *N-th order for U*.

The same procedure is used to formulate the perturbation algorithm for solving the adjoint equation (29). Assuming that $u^* = \sum_{i=0}^{\infty}\varepsilon^i u_i^*$, equations

for finding the corrections u_i^* (i = 1,2,...) have the form

$$A_0^* U_0^* = g_0,\ A_0^* = \left(\Phi'(U_0,0)\right)^*,$$

$$A_0^* u_1^*, = g_1 - \frac{d}{d\varepsilon} A^* \Big|_{\varepsilon=0} u_0^*,$$

$$A_0^* u_2^* = g_2 - \frac{d}{d\varepsilon} A^* \Big|_{\varepsilon=0} u_1^* - \frac{1}{2}\frac{d^2 A^*}{d\varepsilon^2}\Big|_{\varepsilon=0} u_0^*, \tag{34}$$

.....

Solving the first N+2 equations we find the approximation of the N-th order to u^* using the formula

$$U_{(N)}^* = u_0^* + \sum_{i=1}^{N} \varepsilon^i u_i^*.$$

The perturbation algorithms of the type (32), (34) are referred to as the algorithms of the regular perturbations in the literature because they assume the presence in the solution of the problem of the analytical dependence with respect to the perturbation parameter.

7.2. Justification of the perturbation algorithms

Let X, Y be the Hilbert spaces, introduced previously, and the initial operator $\Phi(U,\varepsilon)$ is given by the equation

$$\Phi(U,\varepsilon) = AU+\varepsilon F(U) - f,$$

where $f\in Y$, A: $X\to Y$ is the linear closed operator with the definition domain $D(A)$, dense in X, and $F(U)$ is some non-linear operator acting from X into Y with the domain of definition $D(F) = D(A)$. Consequently, the perturbed problem (27) has the form

$$AU+\varepsilon F(U) = f, \tag{35}$$

The non-perturbed problem is obtained from (35) at ε=0:

$$AU_0 = f. \tag{36}$$

The following claim holds.

Theorem 24. *Let it be that*

1) *Operator A is continuously invertible and* $R(A)=Y$, i.e. *for any* $y\in S$ *there is the unique solution* $x\in D(A)$ *of the equation* $Ax = y$ *such that*

$$||x||_x \le c_0 ||x||_Y, \quad c_0 = \text{const}>0;$$

2) *The operator F satisfies the Lipschitz condition*

$$||F(u)-F(v)||_Y \le k||u-v||_X,\ k = \text{const}>0,\ \forall u,v\in D(F).$$

Therefore, under the condition $|\varepsilon| < \bar{\varepsilon}$, *where* $\bar{\varepsilon} = 1/(c_0 k)$, *the perturbed problem* (35) *has a unique solution* $U\in D(F)$.

If in addition to this, F is the analytical operator, the solution of the problem (35) *is presented in the form of a series*

$$U + U_0 + \sum_{i=1}^{\infty} \varepsilon^i u_i, \tag{37}$$

converging at $|\varepsilon| < \bar{\varepsilon}$ *where the function* u_i *can be calculated using the perturbation algorithm. The rate of convergence of the perturbation algorithm is determined by the equation*

$$\left\| U - U_{(N)} \right\|_X \le c \left| \frac{\varepsilon}{\varepsilon_0} \right|^{N+1}, \quad c = \text{const} > 0, \quad |\varepsilon| < \varepsilon_0,$$

where $U_{(N)}$ *is the approximation of the N-th order from* (33), $\varepsilon_0 < \bar{\varepsilon}$.

This theorem results from the well-known statements of non-linear analysis; its proof can be obtained, for example, as a consequence of the following theorem 25.

Condition 2) of theorem (34) is relatively strict and is rarely satisfied in practice. However, this condition is often satisfied in a ball. Consequently, we can prove an identical theorem where the condition on ε depends on the radius of the ball. For example, the following is valid

Theorem 25. Let it be that:

1) *Operator A is continuously invertible and* $R(A) = Y$, i.e. *for any* $y \in Y$ *there is a unique solution* $x \in D(A)$ *of the equation* $Ax = y$ *such that*

$$\|x\|_X \le c_0 \|y\|_Y, \quad c_0 = \text{const} > 0;$$

2) *For some* $R > 0$ *operator F satisfies the condition*

$$\|F(u) - F(v)\|_Y \le k \|u - v\|_X \ \forall u, v \in B(U_0, R),$$

where $B(U_0,R) = \{u \in D(F) : \|u - U_0\|_X \le R\}$, U_0 *is the solution of the non-perturbed problem* (36).

Consequently, at $|\varepsilon| < \bar{\varepsilon}$, *where*

$$\bar{\varepsilon} = \left[c_0 \left(k + \frac{1}{R} \|F(U_0)\|_Y \right) \right]^{-1},$$

the problem (35) *has a unique solution* $U \in D(F)$, *satisfying the condition* $\|U - U_0\|_X \le R$. *If, in addition to this, F is the analytical operator, the solution of the problem* (35) *is represented in the form of a series*

$$U = U_0 + \sum_{i=1}^{\infty} \varepsilon^i u_i, \tag{38}$$

converging at $|\varepsilon| < \bar{\varepsilon}$ *where the function* u_i *can be calculated using the perturbation algorithm. The rate of convergence of the algorithm of perturbation is determined by the formulae*

$$\left\| U - U_{(N)} \right\|_X \le c \left| \frac{\varepsilon}{\varepsilon_0} \right|^{N+1}, \quad c = \text{const} > 0, \quad |\varepsilon| < \varepsilon_0,$$

where $U_{(N)}$ *is the approximation of the N-th order from* (33), $\varepsilon_0 < \bar{\varepsilon}$.

If the constant k in the condition 2) of theorem 25 does not depend on R and this condition is fulfilled for any R, then (in transition to the limit $R \to \infty$) theorem 25 transforms to theorem 24.

Using the proof of the theorem 25, we obtain the claim on the solvability of the problem (35) also at $\varepsilon = 1$. Let $F(0) = 0$.

Theorem 26. *Let it be that:*

1) *The condition* 1) *of theorem* 25 *is satisfied*;

2) *for any* $R>0$, *the Lipschitz condition holds*

$$\left\|F(u)-F(v)\right\|_Y \le k\left\|u-v\right\|_X \quad \forall u,v\in B(U_0,R),$$

where $B(U_0,R)=\{u\in D(F),\ \|u-U_0\|_x\le R\}$, $k = k(U_0,R) = \text{const} > 0$;

3) *At* $R = \|U_0\|_X$ *the inequality* $c_0k<1/2$ *is satisfied.*

Consequently, the problem (35) at $\varepsilon=1$ *has a unique solution* $U\in D(F)$ *for which the estimate* $\|U\|_X \le c\|f\|_Y$, $c = \text{const} >0$ *is valid.*

7.3. Relation to the method of successive approximations

In the conditions of theorems 24, 25, the approximate solution of the problem (35) can be obtained using the perturbation algorithm. The equation for finding corrections u_i from (37), (38) has the following form:

$$\begin{aligned} AU_0 &= f, \\ Au_1 &= -F(U_0), \\ Au_2 &= -F'(U_0)u_1, \\ &\dots\dots\dots\dots \\ Au_i &= f_{i-1} = f_{i-1}(U_0,u_1\dots,u_{i-1}), \\ &\dots\dots\dots\dots \end{aligned} \tag{39}$$

where the right-hand sides f_{i-1} depend on the derivatives of the operator F at point U_0 to the $(i-1)$-th order. The rate of convergence of this algorithm in the conditions of the theorems 24 and 25 is determined from the formula

$$\left\|U-U_{(N)}\right\|_X \le c\left|\frac{\varepsilon}{\varepsilon_0}\right|^{N+1}, |\varepsilon|<\varepsilon_0<\bar{\varepsilon}, \tag{40}$$

where $U_{(N)} = U_0 + \sum_{i=1}^{N}\varepsilon^i u_i$ is the approximation of the N-th order to the exact solution U.

We also examine the method of successive approximations for solving the problem (35) in the form

$$AU^{N+1} = -\varepsilon F\left(U^N\right) + f \tag{41}$$

at the initial guess $U^0=U_0$, where U_0 is the solution of the non-perturbed problem (36). As indicated by the proof of the theorem 25, at $|\varepsilon|\le\bar{\varepsilon}$ the following estimate of the rate of convergence is valid

$$\left\|U-U^N\right\|_X \le c_1\frac{\left(K|\varepsilon|\right)^{N+1}}{1-K|\varepsilon|}, c_1 = \text{const} > 0,$$

and $K = c_0k < 1/\varepsilon_0$, where ε_0 is a constant from (40). Equation (40) shows that U^N and $U_{(N)}$ are approximations to U of the same order $O(|\varepsilon/\varepsilon_0|^{N+1})$. In addition to this, it may be shown that the approximation U^N is presented in the form

$$U^N = U_{(N)} + \sum_{i=N+1}^{\infty}\varepsilon^i U_i^{(N)}, \tag{42}$$

where $U_{(N)}$ is the approximation of the N-th order according to the perturbation algorithm (39). In fact, as indicated by the proof of the theorem 25, at fixed N the function U^N from (41) is analytical in respect of ε for $|\varepsilon|<\varepsilon_0$ and the expansion into the series $U^N=\sum_{i=0}^{\infty}\varepsilon^i U_i^{(N)}$ with some $U_i^{(N)} \in D(F)$ is justified. Substituting these expansion into (41), we find equations for $U_i^{(N)}$. Comparing these equations with equations (39), it is established successively that $U_0^{(N)}=U_0, U_1^{(N)}=u_1,\ldots,U_N^{(N)}=u_N$, because the corresponding equation for $U_i^{(N)}$ and u_i coincide. This also confirms the validity of equation (42). It follows from here that

Theorem 27. *Let conditions of the theorem* 24 *or* 25 *be satisfied and we examine the method of successive approximation* (41) *with initial approximation* (36). *Consequently, approximation* U^N *is given in the form*

$$U^N+U_{(N)}+\sum_{i=N+1}^{\infty}\varepsilon^i U_i^N,$$

where $U_{(N)}$ *is the approximation of the N-th order of the perturbation algorithm and the following estimate holds*

$$\left\|U^N-U_{(N)}\right\|_X \le c\left|\frac{\varepsilon}{\varepsilon_0}\right|^{N+1}, c=\text{const}>0, \varepsilon_0<\bar{\varepsilon}.$$

It should be mentioned that in some cases the method of successive approximations may be preferred for calculations in comparison with the perturbation algorithm [116].

8. APPLICATIONS TO SOME PROBLEM OF MATHEMATICAL PHYSICS

8.1. The perturbation method for a quasi-linear problem of non-stationary heat conduction

We examine the initial boundary-value problem for the quasi-linear equation of heat conduction of the type

$$C(T)\frac{\partial T}{\partial t}-div\left(L\,\text{grad}T\right)=f\left(t,\bar{x}\right), t\in\left(0,\mathcal{T}\right),\quad \bar{x}\in\Omega,$$
$$T\big|_{t=0}=T_0\left(\bar{x}\right), T\big|_{\gamma_1}=T_1\left(t\right), \frac{\partial T}{\partial \vec{n}}\Big|_{\gamma_2}=0, \tag{43}$$

where $\Omega\subset\mathbb{R}^m$ $(1\le m\le 3)$ is a bounded domain with the piecewise smooth boundary $\delta\Omega=\gamma_1\cup\gamma_2$, $T=T(t,\bar{x})$ is the unknown function of temperature. The coefficients of heat conduction $L=L(t,\bar{x})$ and heat capacity $C(T)$ and also the function $f(t,\bar{x})$, $T_0(\bar{x})$, $T_1(t)$ are assumed to be real and sufficiently smooth, $\bar{x}=(x_1,\ldots,x_m)^T\in\Omega$, $\vec{n}$ is the unit vector of the external normal to $\delta\Omega$, $\mathcal{T}<\infty$. (When $m=1$, where Ω is the segment with the ends γ_1 and γ_2, the role of $\frac{\partial T}{\partial \vec{n}}$ is played by the derivative $\frac{\partial T}{\partial x}$).

The problem of type (43) form when describing the process of propa-

gation of heat in the bounded domains (rods, plates, etc.) heated by heat flows at the boundaries, in the presence of also internal sources or sinks [32]. In a number of problems heat conduction coefficient $L(T)$ is given as a function of temperature. In this case, using the Kirchhoff substitution, the heat conduction equation can be easily reduced to the form of (43), where $L \equiv 1$.

Initially, we obtain a perturbation problem. Heat capacity is presented in the form

$$C(T) = D(T)R(T),$$

whee $R(T)=1+\beta T$, $\beta \in \mathrm{R}$. Using this substitution of variables, the equation of heat conductivity from (43) is written in the form

$$D(T)R(T)\frac{dT}{dR}\frac{\partial R}{\partial t} - \mathrm{div}\left(\mathrm{L}\frac{dT}{dR}\mathrm{grad}R\right) = f\left(t,\bar{x}\right). \tag{44}$$

Let $N \in \mathrm{R}$ be some constant such that $N/\beta>0$, which is determined below. Assuming that

$$\psi(T) = D(T)R(T)\frac{dT}{dR} - \frac{1}{N}$$

and substituting $D(T)R(T)\dfrac{dT}{dR}$ from (44) by $(1/N)+\varepsilon\psi(T)$, $0 \le \varepsilon \le 1$, from (43) and (44), we obtain the perturbation problem:

$$\left(\frac{1}{N}+\varepsilon\psi(T)\right)\frac{\partial R}{\partial t} - \frac{1}{\beta}\mathrm{div}\left(L\mathrm{grad}R\right) = f\left(t,\bar{x}\right), t \in \left(0,\mathcal{T}\right), \quad \bar{x} \in \Omega,$$

$$R\big|_{t=0} = f_0\left(\bar{x}\right) \equiv 1+\beta T_0\left(\bar{x}\right), \quad R\big|_{\gamma_1} = f_1\left(t\right) \equiv 1+\beta T_1\left(t\right), \quad \frac{\partial R}{\partial \vec{n}}\Big|_{\gamma_2} = 0,$$

or, in a different form

$$\begin{gathered}\frac{\partial R}{\partial t} - \frac{N}{\beta}\mathrm{div}\left(L\,\mathrm{grad}\,R\right) + \varepsilon\left(\frac{N}{\beta}C\left(\frac{R-1}{\beta}\right)-1\right)\frac{\partial R}{\partial t} = Nf\left(t,\bar{x}\right),\\ R\big|_{t=0} = f_0\left(\bar{x}\right), \quad R\big|_{\gamma_1} = f_1\left(t\right), \quad \frac{\partial R}{\partial \vec{n}}\Big|_{\gamma_2} = 0.\end{gathered} \tag{45}$$

At $\varepsilon=1$ from (45) we obtain the original problem (43) and the problem obtained from (45) at $\varepsilon=0$ will be referred to as non-perturbed.

Assuming the sufficient smoothness of the initial data, the non-perturbed problem has a unique solution R_0 satisfying the problem

$$\begin{gathered}\frac{\partial R_0}{\partial t} - \frac{N}{\beta}\mathrm{div}\left(L\mathrm{grad}R_0\right) = Nf\left(t,\bar{x}\right), \quad t \in \left(0,T\right), \quad \bar{x} \in \Omega,\\ R_0\big|_{t=0} = f_0\left(\bar{x}\right), \quad R_0\big|_{\gamma_1} = f_1\left(t\right), \quad \frac{\partial R_0}{\partial \vec{n}}\Big|_{\gamma 2} = 0.\end{gathered} \tag{46}$$

Subtracting (46) from (45), we can write the problem for the difference $\tilde{R} = R - R_0$ between the solutions of the perturbed and non-perturbed problems:

$$\frac{\partial \tilde{R}}{\partial t}-\frac{N}{\beta}\operatorname{div}\left(L\operatorname{grad}\tilde{R}\right)+\varepsilon F\left(\tilde{R}+R_0\right)=0,\quad t\in\left(0,\mathcal{T}\right),\quad \overline{x}\in\Omega,$$

$$\tilde{R}\big|_{t=0}=0,\quad \tilde{R}\big|_{\gamma 1}=0,\quad \left.\frac{\partial \tilde{R}}{\partial \vec{n}}\right|_{\gamma_2}=0, \tag{47}$$

where $F\left(R\right)=\left(\frac{N}{\beta}C\left(\frac{R-1}{\beta}\right)-1\right)\frac{\partial R}{\partial t}$.

We introduce specific restrictions on the function $L(t,\overline{x})$, $C(T)$. Let $L(t,\overline{x})$ do not depend on t, i.e. $L(t,\overline{x})$ = L $(\overline{x})$ and

$$0<L_0\le L\left(\overline{x}\right)\le L_1<\infty,\quad 0<C_0\le C\left(T\right)\le C_1<\infty, \tag{48}$$

where L_i, C_i = const, i = 1,2.

To write the problem in an operator form, we examine the space $H=L_2(\Omega)$ of the real-valued functions $u(\overline{x})$, square-integrable integrable according to Lebesgue on Ω, and the space $X=\{u(\overline{x})\in W_2^2(\Omega)$: $u\big|_{\gamma_1}=0\}$, where $W_2^2(\Omega)$ is the Sobolev space of functions from $L_2(\Omega)$ which have square-integrable first and second derivatives. We also examine $Y=L_2(0,\mathcal{T};H)$, $Y_1=L_2(0,\mathcal{T};X)$– the spaces of abstract functions $v(t)$ with the values in H, X and space $W=\{v\in Y_1: dv/dt\in Y\}$, $W_{\mathcal{T}}=\{w\in W: w|_{t=\mathcal{T}}=0\}$. It is assumed that the spaces H and Y are identified with their adjoints: $H\equiv H^*$, $Y^*=Y$, $(\cdot,\cdot)_{L_2(0,\mathcal{T};H)}\equiv(\cdot,\cdot)$.

We introduced a generalized form for the perturbation problem (47) in the form: find the function $\tilde{R}\in Y$ such that

$$-\left(\tilde{R},\frac{dw}{dt}\right)+\left(\tilde{R},Aw\right)+\varepsilon\left(F\left(\tilde{R}+R_0\right),w\right)=0\quad \forall w\in W_{\mathcal{T}}. \tag{49}$$

Here A is a linear operator acting in Y with the domain of definition $D(A)=Y_1$ and determined by the equation

$$AR=-\frac{N}{\beta}\operatorname{div}\left(L\operatorname{grad}R\right),\quad R\in Y_1,$$

and $F(R)$ is the operator defined by the equality

$$\left(F\left(R\right),w\right)=\left(R,\frac{dw}{dt}\right)-\frac{N}{\beta}\left(\tilde{C}\left(R\right),\frac{dw}{dt}\right),$$

where

$$w\in W_{\mathcal{T}},\quad \tilde{C}\left(R\right)=\int_0^R C\left(\frac{R'-1}{\beta}\right)dR'.$$

(It is assumed that the function $C(T)$ is defined at almost all $T\in(-\infty,+\infty)$ and $f_0=0$).

Lemma 10. *The operator F is bounded from Y into $W_{\mathcal{T}}^*$.*

Lemma 11. *Operator F at any point $R\in Y$ has a Gateaux derivative $F'(R)$ defined by the relationship*

$$\left(F'(R)v,w\right)=\left(v,\frac{dw}{dt}\right)-\frac{N}{\beta}\left(C\left(\frac{R-1}{\beta}\right)v,\frac{dw}{dt}\right),\quad v\in Y,\quad \forall w\in W_{\mathcal{T}}.$$

Operator $F'(R)$ *is bounded from* Y in $W^*_{\mathcal{T}}$, *and*

$$\left\|F'(R)v\right\|_{W^*_{\mathcal{T}}}\le k\left\|v\right\|_Y,$$

where

$$k=\sup_{t,\bar{x}}\left|1-\frac{N}{\beta}C\left(\frac{R-1}{\beta}\right)\right|=\max\left(\left|1-\frac{N}{\beta}C_0\right|,\left|1-\frac{N}{\beta}C_1\right|\right),$$

and constants C_0, C_1 are defined in (48).

Theorem 28. *Let* $R_0\in Y$ *be the solution of the non-perturbed problem* (46) *and the restrictions* (48) *be satisfied. Therefore, at* $0<\varepsilon<1/k$, *where*

$$k=\max\left(\left|1-\frac{N}{\beta}C_0\right|,\left|1-\frac{N}{\beta}C_1\right|\right),\tag{50}$$

the perturbation problem has a unique solution $\tilde{R}\in Y$ *in the sense of* (49). *If in addition to this the function* $C(T)$ *is analytical, then the solution* $\tilde{R}$ *of problem* (47) *is represented in the form of a series in respect of the powers of* ε:

$$\tilde{R}=\sum_{i=1}^{\infty}\varepsilon^i\tilde{R}_i,\quad \tilde{R}_i\in Y,\tag{51}$$

converging at $\varepsilon<1/k$, where the functions $\tilde{R}_i$ can be calculated using the perturbation algorithm.

Theorem 28 gives the actual justification of the perturbation method in application to problem (47). Formula (50) is a sufficient condition for ε, at which the perturbed problem has a unique solution presented in the form of a series (51). At $k<1$, i.e. at

$$\max\left(\left|1-\frac{N}{\beta}C_0\right|,\left|1-\frac{N}{\beta}C_1\right|\right)<1,\tag{52}$$

the theorem remains valid if we set $\varepsilon=1$.

Condition (52) can be used for selecting the constants N and β which have so far been regarded as arbitrary. If, for example, we set $\xi=N/\beta$ and $g(\xi)=\max(|1-C_0\xi|,|1-C_1\xi|)$, $\xi>0$, then the solution of the inequality $g(\xi)<1$, is

$$0<\xi<2/C_1.\tag{53}$$

This means that, selecting as N/β any number from the range (53), it may be asserted that at $\varepsilon=1$ the perturbation problem (47) has a unique solution presented in the form of a converging series, and it can be solved by the perturbation algorithm.

Substituting the series (51) into (47) and equating the terms at the same powers of ε, we obtain a system of equations for determining the corrections $\tilde{R}_i$

$$\frac{\partial \tilde{R}_1}{\partial t}-\frac{N}{\beta}div\left(L\,\mathrm{grad}\tilde{R}_1\right)=-\left(\frac{N}{\beta}C\left(\frac{R_0-1}{\beta}\right)-1\right)\frac{\partial R_0}{\partial t},$$

$$\tilde{R}_1\big|_{t=0}=0,\quad \tilde{R}_1\big|_{\gamma_1}=0,\quad \frac{\partial \tilde{R}_1}{\partial \vec{n}}\Big|_{\gamma_2}=0,$$

$$\frac{\partial \tilde{R}_2}{\partial t}-\frac{N}{\beta}\mathrm{div}\left(L\,\mathrm{grad}\tilde{R}_2\right)=-\left(\frac{N}{\beta}C\left(\frac{R_0-1}{\beta}\right)-1\right)\frac{\partial \tilde{R}_1}{\partial t}-\frac{N}{\beta_2}C'\tilde{R}_1\frac{\partial R_0}{\partial t},$$

$$\tilde{R}_2\big|_{t=0}=0,\quad \tilde{R}_2\big|_{\gamma_1}=0,\quad \frac{\partial \tilde{R}_2}{\partial \vec{n}}\Big|_{\gamma_2}=0$$

and so on. Computing N corrections $\tilde{R}_i$, i = 1,...,N, we find the approximation of the N-th order to $\tilde{R}$ using the equation:

$$\tilde{R}_{(N)}=R_0+\sum_{i=0}^{N}\varepsilon^i\tilde{R}_i,$$

where according to the theorem 28 $\left\|\tilde{R}-\tilde{R}_{(N)}\right\|_Y\le c\varepsilon^{n+1}$, c=const>0.

8.2. The Galerkin method for problems of dynamics of atmospheric processes

Let S be a sphere with radius r. We examine the problem of the two-dimensional equation of a barotropic atmosphere on a sphere in the form [23]

$$\frac{\partial\varphi}{\partial t}+v(-\Delta)^s\varphi+J\left(\Delta^{-1}\varphi,\varphi\right)=f\in(0,T),\quad \varphi(0)=u, \tag{54}$$

where

$$J(v,w)=\frac{1}{r^2}\left(\frac{\partial v}{\partial\lambda}\frac{\partial w}{\partial\mu}-\frac{\partial v}{\partial\mu}\frac{\partial w}{\partial\lambda}\right),$$

$$\varphi=\varphi\,(t,\lambda,\mu),\quad \mu=\sin\psi,\quad (\lambda,\psi)\in S,\quad 0\le\lambda\le 2\pi,\quad -\pi/2\le\psi\le\pi/2,$$

Δ is the Laplace-Beltrami operator on the sphere

$$\Delta=\frac{1}{r^2}\left(\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\lambda^2}+\frac{\partial}{\partial\mu}\left(1-\mu^2\right)\frac{\partial}{\partial\mu}\right). \tag{55}$$

Here $\varphi=\varphi(t,\lambda,\mu)$ is the function of the vortex, λ is the longitude, ψ is the latitude, $t\in[0,T]$, v = const>0, $s\ge1$, the member $v(-\Delta)^s\varphi$ describes the turbulent viscosity, $f(\lambda,t,\mu)$ is the external source of vorticity, $u=u(\lambda,\mu)$ is the function of the initial condition.

In this model, the atmosphere is treated as a layer of an incompressible fluid of constant density whose thickness is small in comparison with the horizontal scale of motion. Regardless of relative simplicity, this equation also takes into account the important dynamic processes as non-linear interaction and wave dispersion. The fine-scale motion of the atmosphere within this model is taken into account and by means of the turbulent member and the external source of vorticity.

When $s = 1$, equation (54) is derived from the conventional Navier-Stokes equations on the sphere. 'Artificial viscosity', when $s > 0$ is often used to prove the theorems of existence and uniqueness, and also in the numerical solution of the problem.

We introduce $H = \overset{\circ}{L}_2(S)$ – the Hilbert space of real-valued functions, defined on S, square-integrable and orthogonal to the constant, with the normal scalar product $(\cdot,\cdot)$ and the norm $\|\cdot\|=(\cdot,\cdot)^{1/2}$. The Laplace"Beltrami operator will be regarded as the operator acting from H into H with the domain of definition $D(A)=\{u\in H:\Delta u\in H\}$.

Using the powers of the Laplace operator, we can introduce the Sobolev spaces $\overset{\circ}{H}{}^{\gamma}(S)$, $\gamma\in R$, with the scalar product and the norm:

$$(u,v)_{\overset{\circ}{H}{}^{\gamma}(S)} = \left((-\Delta)^{\gamma/2} u,\quad (-\Delta)^{\gamma/2} v\right) = \sum_{j=1}^{\infty} \Lambda_j^{\gamma} u_j v_j,$$

$$\|u\|_{\overset{\circ}{H}{}^{\gamma}(S)} = (u,u)^{1/2}_{\overset{\circ}{H}{}^{\gamma}(S)} = \left\|(-\Delta)^{\gamma/2} u\right\| = \left(\sum_{j=1}^{\infty} \Lambda_j^{\gamma} u_j^2\right)^{1/2}, \tag{56}$$

where Λ_j are the eigenvalues of the operator $-\Delta$, corresponding to the eigenfunctions ω_j, $u_j = (u,\omega_j)$. Here $D(\Delta) = \overset{\circ}{H}{}^2(S)$, $H = \overset{\circ}{H}{}^0(S)$.

We define the operator A using the equation

$$A\varphi = v(-\Delta)^s \varphi; \tag{57}$$

it acts in H with the domain of definition $D(A)=\overset{\circ}{H}{}^{2s}(S)$. For $\gamma\in R$ we introduce the spaces $X^{\gamma} = \overset{\circ}{H}{}^{2\gamma s}$, $Y^{\gamma} = L_2\left(0,T;X^{\gamma}\right)$, and $W^{\gamma} = \left\{\varphi\in Y^{\gamma+1/2}:\dfrac{d\varphi}{dt}\in Y^{\gamma-1/2}\right\}$ with the norms

$$\|\varphi\|_{Y^{\gamma}} = \left(\int_0^T \|\varphi\|^2_{X^{\gamma}} dt\right)^{1/2}, \quad \|\varphi\|_{W^{\gamma}} = \left(\left\|\frac{d\varphi}{dt}\right\|^2_{Y^{\gamma-1/2}} + \|\varphi\|^2_{Y^{\gamma+1/2}}\right)^{1/2}.$$

For $\varphi\in W^{\gamma}$ we defined a non-linear operator $F(\varphi)$

$$F(\varphi) = J\left(\Delta^{-1}\varphi,\varphi\right).$$

It is well known that if $s > 1$, $\gamma \geq 1/(2s)$ or $0\leq\gamma\leq(s-1)/(2s)$ or $s>1$, then the operator F acts from $Y^{\gamma+1/2}$ into $Y^{\gamma-1/2}$ with the domain of definition $D(F) = W^{\gamma}$, it is continuously differentiable according to Frechét and

$$\left\|F'(\varphi)\right\|_{Y^{\gamma+1/2}\to Y^{\gamma-1/2}} \leq c_3 \|\varphi\|_{W^{\gamma}},\ c_3 = \text{const} > 0.$$

Taking this into account, the problem (54) can be written in the operator form

$$\frac{d\varphi}{dt} + A\varphi + F(\varphi) = f,\quad t\in(0,T),\quad \varphi(0) = u. \tag{58}$$

It is an abstract non-linear evolution equation. This problem can be also written in the form $A\varphi=f$, selecting as A the operator $A\varphi = \dfrac{d\varphi}{dt} + A\varphi + F(\varphi)$

acting from $Y^{\gamma+1/2}$ into $Y^{\gamma-1/2}$ with the domain of definition $D(A)=\{\varphi\in W^{\gamma}:\varphi(0)=u\}$.

It is well-known that at $f\in Y^{-1/2}=L_2(0,T;\overset{\circ}{H}{}^{-s})$, $s\geq 1$, $u\in \overset{\circ}{L}_2(S)$ there is a unique solution $\varphi\in W^0$ of the problem (58).

We examine the Galerkin method to solve the problem (54). As basis functions we examine the finite set of the eigenfunctions $\{\omega_j\}_{j=\overline{1,N}}$ of the Laplace–Beltrami operator. The approximate solution of equation (58) will be determined in the form

$$\varphi_N(t)=\sum_{j=1}^{N}\varphi_{Nj}(t)\omega_j. \tag{59}$$

Introducing the projection operator P_N using the equation

$$P_N\xi=\sum_{j=1}^{N}\xi_j\omega_j,\qquad \xi_j=(\xi,\omega_j)$$

and applying it to equation (58), we get

$$\frac{\partial\varphi_N}{\partial t}+P_N J\left(\Delta^{-1}\varphi_N,\varphi_N\right)=\nu\Delta\varphi_N+P_N f. \tag{60}$$

This gives the system of ordinary differential equations for determining the functions $\varphi_{Nj}(t)$

$$\begin{gathered}\varphi'_{Nj}+\nu\Lambda_j^s\varphi_{Nj}+\sum_{i,k=1}^{N}\left(J\left(\Delta^{-1}\omega_i,\omega_k\right),\omega_j\right)\varphi_{Nj}\varphi_{Nk}=\left(f,\omega_j\right),\\ j=1,\ldots,N.\quad \varphi_{Nj}(0)=\left(u,\omega_j\right).\end{gathered} \tag{61}$$

The existence of the solution of the system (61) follows from the theory of ordinary differential equations and a priori estimates. On the basis of solvability of the system (61) we establish the existence and uniqueness of the solution of the initial problem (54).

8.3. The Newton method in problems of variational data assimilation

At present, because of the investigations of global changes, it is important to examine the problem of obtaining and rational application of the results of measurements for retrospective analysis in various areas of knowledge. The mathematical model of the problem may be formulated as a problem of the collection and processing of multi-dimensional (including the dependence of the time and spatial variables) data representing one of the optimal control problems.

Let us assume that we examine some physical process whose mathematical model is written in the form of a non-linear evolution problem

$$\frac{\partial\varphi}{\partial t}=F(\varphi),\ t\in(0,T),\ \varphi\big|_{t=0}=u, \tag{62}$$

where $\varphi=\varphi(t)$ is the unknown function belonging for each t to the Hilbert space X, $u\in X$, F is the non-linear operator acting from X into X. Let $Y=L_2(0,T;X)$, $(\cdot,\cdot)_{L_2(0,T;X)}=(\cdot,\cdot)$, $\|\cdot\|=(\cdot,\cdot)^{1/2}$. We introduce the functional

$$S(u)=\frac{\alpha}{2}\|u-u_{\text{obs}}\|_X^2+\frac{1}{2}\int_0^T\|C\varphi-\varphi_{\text{obs}}\|_X^2\,dt, \tag{63}$$

where α = const ≥ 0, $u_{\text{obs}}\in X$, $\varphi_{\text{obs}}\in Y$ are the given functions (the results of observation), Y_{obs} is a sub-space of Y, $C{:}Y\to Y_{\text{obs}}$ is the linear operator.

We examine the following problem of data assimilation in order to restore the initial condition: find u and φ such that

$$\frac{\partial\varphi}{\partial t}=F(\varphi),\quad t\in(0,T),\quad \varphi|_{t=0}=u,\quad S(u)=\inf_v S(v). \tag{64}$$

The necessary optimality condition reduces the problem (64) to the system

$$\frac{\partial\varphi}{\partial t}=F(\varphi),\quad t\in(0,T),\quad \varphi\big|_{t=0}=u, \tag{65}$$

$$-\frac{\partial\varphi^*}{\partial t}-\left(F'(\varphi)\right)^*\varphi^*=-C^*\left(C\varphi-\varphi_{\text{obs}}\right),\ t\in(0,T),\quad \varphi^*\big|_{t=T}=0, \tag{66}$$

$$\alpha\left(u-u_{\text{obs}}\right)-\varphi^*\big|_{t=0}=0 \tag{67}$$

with the unknown φ,φ^*, u, where $(F'(\varphi))^*$ is the operator adjoint to the Frechét derivative of the operator F, and C^* is the operator adjoint to C.

Assuming that there is a solution of the problem (65)–(67), we examine the Newton method to find this solution.

The system (65)–(67) with three unknown φ,φ^*,u can be regarded as the operator equation of the type

$$\mathcal{F}(U)=0, \tag{68}$$

where $U=(\varphi,\varphi^*,u)$. To use the Newton method, we must calculate $\mathcal{F}'(U)$. It is assumed that the original operator F is differentiable twice continuously according to Frechét. Consequently, the Newton method

$$U_{n+1}=U_n-\left[\mathcal{F}'(U_n)\right]^{-1}\mathcal{F}(U_n),\ U_n=\left(\varphi_n,\varphi_n^*,u_n\right) \tag{69}$$

consists of the following steps.

1. We find $V_n=[\mathcal{F}'(U_n)]^{-1}\mathcal{F}'(U_n)$ as a solution of the problem $\mathcal{F}'(U_n)\,V_n=\mathcal{F}(U_n)$ at $V_n=(\psi_n,\psi_n^*,v_n)$

$$\frac{\partial\psi_n}{\partial t}-F'(\varphi_n)\psi_n=\frac{\partial\varphi_n}{\partial t}-F(\varphi_n),\ \psi_n\big|_{t=0}=v_n+\varphi_n\big|_{t=0}-u_n, \tag{70}$$

$$-\frac{\partial\psi_n^*}{\partial t}-\left(F'(\varphi_n)\right)^*\psi_n^*=p_1^n,\ \psi_n^*\big|_{t=T}=\varphi_n^*\big|_{t=T}, \tag{71}$$

$$\alpha v_n-\psi_n^*\big|_{t=0}=\alpha\left(u_n-u_{\text{obs}}\right)-\varphi_n^*\big|_{t=0}, \tag{72}$$

where $p_1^n=\left(F''(\varphi_n)\psi_n\right)^*\varphi_n^*-C^*C\psi_n-\dfrac{\partial\varphi_n^*}{\partial t}-\left(F'(\varphi_n)\right)^*\varphi_n^*+C^*\left(C\varphi_n-\varphi_{\text{obs}}\right)$.

2. Assume $U_{n+1}=U_n-V_n$, i.e.

$$\varphi_{n+1}=\varphi_n-\psi_n,\ \varphi_{n+1}^*=\varphi_n^*-\psi_n^*,\quad u_{n+1}=u_n-v_n. \tag{73}$$

Since $U_{n+1}=U_n-V_n$, then the two steps (70)–(73) can be reformulated as follows: at given $\varphi_n,\varphi_n^*,u_n$ find $\varphi_{n+1},\varphi_{n+1}^*,u_{n+1}$ such that

$$\frac{\partial\varphi_{n+1}}{\partial t}-F'(\varphi_n)\varphi_{n+1}=F(\varphi_n)-F'(\varphi_n)\varphi_n,\quad \varphi_{n+1}\big|_{t=0}=u_{n+1}, \tag{74}$$

$$-\frac{\partial\varphi^*_{n+1}}{\partial t}-\left(F'(\varphi_n)\right)^*\varphi^*_{n+1}=p_2^n,\ \varphi^*_{n+1}\big|_{t=T}=0, \tag{75}$$

$$\alpha\left(u_{n+1}-u_{\text{obs}}\right)-\varphi^*_{n+1}\big|_{t=0}=0, \tag{76}$$

where $p_2^n=\left(F''(\varphi_n)(\varphi_{n+1}-\varphi_n)\right)^*\varphi_n^*-C^*\left(C\varphi_{n+1}-\varphi_{\text{obs}}\right)$.

We fix the point $\varphi_0\in Y$, the real number $R>0$ and examine the ball $S_R(\varphi_0)=\{\varphi\in Y:\|\varphi-\varphi_0\|\le R\}$. It is assumed that the initial mathematical model satisfies at all $\varphi\in S_R(\varphi_0)$ the following conditions:

1) the solution of the problem

$$\frac{\partial\psi}{\partial t}-F'(\varphi)\psi=f,\quad \psi\big|_{t=0}=v$$

satisfies the inequality

$$\|\psi\|\le c_1\left(\|f\|+\|v\|_X\right),\quad c_1=c_1(R,\varphi_0)>0; \tag{77}$$

2) the solution of the adjoint problem

$$-\frac{\partial\psi^*}{\partial t}-\left(F'(\varphi)\right)^*\psi^*=p,\quad \psi^*\big|_{t=T}=g$$

satisfies

$$\left\|\psi^*\right\|+\left\|\psi^*\big|_{t=0}\right\|_X\le c_1^*\left(\|p\|+\|g\|_X\right),\quad c_1^*=c_1^*(R,\varphi_0)>0; \tag{78}$$

3) operator F is three time continuously differentiable according to Frechét, and

$$\left\|F''(\varphi)\right\|\le c_2,\quad \left\|F'''(\varphi)\right\|\le c_3,\quad c_k=c_k(R,\varphi_0)>0,\quad k=1,2. \tag{79}$$

Comment 1. For a bi-linear operator F constant c_2 does not depend on R, φ_0 and $c_3\equiv 0$.

We find the solution of the problem (65)–(67) in the ball

$$S_r=\left\{(\varphi,\varphi^*,u):\|\varphi-\varphi_0\|+\left\|\varphi^*\right\|+\|u-u_0\|\text{x}\le r\right\},$$

$$u_0\in X,\quad r=\min\left(c_2^{-1},R\right).$$

In the conditions of complete observation ($C\equiv E$), it holds

Theorem 29. Let $u_0\in X$, $\varphi_0\in Y$, $R>0$, $\varphi_0^*=0$

$$B\eta\left(1+\frac{B(c_2+c_3r)r}{2}\right)\le r, \tag{80}$$

where

$$\eta=\left\|\frac{\partial\varphi_0}{\partial t}-F(\varphi_0)\right\|+\left\|\varphi_0\big|_{t=0}-u_0\right\|_X+\|\varphi_0-\varphi_{\text{obs}}\|+\alpha\|u_0-u_{\text{obs}}\|_X,$$

$$B=\max(\beta_1,\beta_2,\beta_3),\quad \beta_1=\alpha^{-1}\left(2c_1c_1^*+2c_1^2c_1^*+4c_1^2c_1^{*2}\right)+c_1+2c_1c_1^*,$$

$$\beta_2=\alpha^{-1}\left(c_1^*+c_1c_1^*+2c_1c_1^{*2}\right)+c_1^*,\quad \beta_3=\alpha^{-1}\left(1+c_1+2c_1c_1^*\right).$$

Consequently, (65)–(67) *has a unique solution* φ, φ^*, u *in the ball* S_r. *Starting from* φ_0, φ_0^*, u_0, *the Newton method converges to* φ, φ^*, u. *The following estimate of the convergence rate holds*:

$$\|\varphi-\varphi_n\|+\|\varphi^*-\varphi_n^*\|+\|u-u_n\|_X \le B\eta\frac{(h/2)^{2^n-1}}{1-(h/2)^{2^n}}, \tag{81}$$

where $h = B^2\eta(c_2+c_3r)<2$.

Comment 2. Condition (80) holds at sufficiently small η, which means that φ_0, φ_0^*, u_0 is sufficiently close to the exact solution.

BIBLIOGRAPHIC COMMENTARY

The main definitions of non-linear functional analysis and its applications to solving non-equations are given in [10, 94]. The detailed explanation of the variational methods used for non-linear operator equations is presented in [7,9,17]. The applications of topological methods to examining non-linear differential and integral equations are examined in [77]. The fundamentals of the theory of cones and of the methods of examining equations, containing non-linearities, are presented in [36].

The book [42] is an introduction into the theorem of non-linear differential equations. The book [99] is concerned with the qualities of theory of non-linear parabolic equations with the explanation illustrated by examples of specific problem for different scientific areas – hydrodynamics, chemical kinetics, population genetics. The monograph of J.L. Lions is a classic book discussing the solution of non-linear differential equation [50].

Different aspects of the theory of adjoint equations and algorithms of perturbations in relation to applications in non-linear problems are presented in [116].

References

1 Arsenin V.Ya., Methods of mathematical physics and special functions, Nauka, Moscow (1984).
2 Bakhvalov N.S., Numerical methods, Nauka, Moscow (1973).
3 Bakhvalov N.S., Zhidkov N.P. and Kobel'kov G.M., Numerical methods, Nauka, Moscow (1987).
4 Bitsadze A.V., Equations of mathematical physics, Nauka, Moscow (1982).
5 Brelo M., Fundamentals of the classic theory of potential, IL, Moscow (1964).
6 Bogolyubov N.N. and Mitropol'skii Yu.A., Asymptotic methods in the theory of non-linear oscillations, Nauka, Moscow (1974).
7 Weinberg M.M., Variational methods of examining non-linear equations, Gostekhizdat, Moscow (1956).
8 Weinberg M.M. and Trenogin V.A., Theory of bifurcation of solutions of non-linear equations, Nauka, Moscow (1969).
9 Weinberg M.M., Variational methods and the method of monotonic operators in the theory of non-linear equations, Nauka, Moscow (1972).
10 Weinberg M.M., Functional analysis, Prosveshchenie, Moscow (1979).
11 Vishik M.I. and Lyusternik L.A., Some problems of perturbations of boundary-value problems with differential equations in partial derivatives, DAN SSSR, 129, No.6 (1959).
12 Vladimirov V.S., Mathematical problems of single-rate theory of transfer of particles, Trudy MIAN, No.61 (1961).
13 Vladimirov V.S., Equations of mathematical physics, Nauka, Moscow (1988).
14 Voevodin V.V. and Kuznetsov Yu.A., Matrices and calculations, Nauka, Moscow (1984).
15 Volkov E.A., Numerical methods, Nauka, Moscow (1982).
16 Voloshchuk V.M., Kinetic theory of coagulation, Gidrometeoizdat, Leningrad (1984).
17 Gaevsky Kh., Gröger K. and Zakharias K., Non-linear operator equations and operator differential equations, Mir, Moscow (1978).
18 Godunov S.K., Current aspects of linear algebra, Nauchnaya Kniga, Novosibirsk (1997).
19 Grinberg G.A., Selected problems of mathematical theory of electrical and magnetic phenomena, Izd-vo AN SSSR, Moscow (1948).
20 Gunter N.M., Potential theory and its application to the main problems of mathematical physics, Gostekhizdat, Moscow (1953).
21 Ditkin V.A. and Prudnikov A.P., Integral transformation and operational calculus, Nauka, Moscow (1974).
22 Dymnikov V.P., Computational methods in geophysical hydrodynamics, OVM AN SSSR, Moscow (1984).
23 Dymnikov V.P. Filatov A.N., Fundamentals of the mathematical theory on the climate, VINITI, Moscow (1987).
24 D'yakonov E.G., Difference methods of solving boundary value problems, Moscow State University, Moscow (1971).
25 Egorov Yu.V. and Shubin M.A., Linear differential equations with partial derivatives. Fundamentals of the classic theory. Itogi nauki i tekhniki. Current problems of mathematics. Fundamental directions, Vol. 30, VINITI, Moscow (1987).
26 Zaitsev V.F. and Polyanin A.D., A handbook of differential equations with partial derivatives. Exact solutions, International Educational Programme, Moscow (1996).
27 Zelenyak T.I., Quantitative theory of the boundary value problems for quasi-linear parabolic equations of the second order, NGU, Novosibirsk (1972).
28 Iosida K., Functional analysis, Mir, Moscow (1967).
29 Kalitkin N.N., Numerical methods, Nauka, Moscow (1978).
30 Kantorovich L.V. and Akilov G.P., Functional analysis in normalised spaces, Nauka,

Moscow (1977).
31 Kato T., Theory of perturbation of linear operators, Mir, Moscow (1972).
32 Kozdoba L.A., Methods of solving non-linear problems of heat conduction, Nauka, Moscow (1975).
33 Kolmogorov A.N. and Fomin S.V., Elements of the theory of functions and functional analysis, Nauka, Moscow (1981).
34 Korn G. and Korn T., A handbook of mathematics, Nauka, Moscow (1984).
35 Krasnosel'skii M.A., Vainikko G.M., Zabreiko P.P., Rutitskii Ya.B. and Stetsenko V.Ya., Approximate solutions of operator equations, Nauka, Moscow (1969).
36 Krasnosel'skii M.A., Positive solutions of operator equations, GTTL, Moscow (1962).
37 Collatz L., Eigenvalue problems, Nauka, Moscow (1968).
38 Krein S.G., Linear differential equations in Banach space, Nauka, Moscow (1967).
39 Krein S.G., Linear equations in Banach space, Nauka, Moscow (1971).
40 Krylov V.I. and Bobkov V.V., Computing methods, Vol. 2, Nauka, Moscow (1977).
41 Kupradze V.D., Boundary value problems of the theory of vibrations and integral equations, Gostekhizdat, Moscow and Leningrad (1950).
42 Kufner A. and Fuchik S., Non-linear differential equations, Nauka, Moscow (1980).
43 Ladyzhenskaya O.A., Mathematical problems of the dynamics of a viscous incompressible flow, Nauka, Moscow (1970).
44 Ladyzhenskaya O.A., Boundary-value problems of mathematical physics, Nauka, Moscow (1980).
45 Ladyzhenskaya O.A., Solonnikov V.A. and Ural'tseva N.N., Linear and quasi-linear parabolic equations, Nauka, Moscow (1967).
46 Ladyzhenskaya O.A. and Ural'tseva N.N., Linear and quasi-linear elliptic equations, (1973).
47 Landkof N.S., Fundamentals of the current potential theory, (1966).
48 Lebedev V.I., Functional analysis and numerical mathematics, VINITI, Moscow (1994).
49 Levin V.I. and Grosberg Yu.I., Differential equations of mathematical physics, GITTL, Leningrad (1951).
50 Lyons G.L., Some methods of solving non-linear boundary-value problems, Mir, Moscow (1972).
51 Lyons J.L. and Magenes E., Non-homogeneous boundary value problems and their applications, Mir, Moscow (1971).
52 Lomov S.A., Introduction into the general theory of singular perturbations, Nauka, Moscow (1981).
53 Lewins J., Importance. The adjoint function, Pergamon Press, New York (1965).
54 Lyusternik L.A. and Sobolev V.I., Elements of functional analysis, Nauka, Moscow (1965).
55 Lyapunov A.M., General problem of the stability of motion, Khar'k. mat. ob., Khar'kov (1892).
56 Lyapunov A.M., Collected works, Vol.2.,Gostekhizdat, Moscow and Leningrad (1956).
57 Manzhirov A.V. and Polyanin A.D., A handbook of integral equations. Solution methods, Faktorial Press, Moscow (2000).
58 Marchuk G.I., Methods of calculating nuclear reactors, Gidrometeoizdat, Leningrad (1974).
59 Marchuk G.I., Numerical solution of the problems of dynamics of atmosphere and oceans, Gidrometeoizdat, Leningrad (1974).
60 Marchuk G.I., Methods of numerical mathematics, Nauka, Moscow (1989).
61 Marchuk G.I., Splitting methods, Nauka, Moscow (1988).
62 Marchuk G.I., Mathematical modelling in the problem of the environment, Nauka, Moscow (1982).
63 Marchuk G.I., Adjoint equations and analysis of complex systems, Nauka, Moscow (1992).
64 Marchuk G.I. and Agoshkov V.I., Introduction into projection-grid methods, Nauka, Moscow (1981).
65 Marchuk G.I. and Shaidurov V.V., Increasing the accuracy of solutions of different schemes, Nauka, Moscow (1979).
66 Maslov V.P., Theory of perturbations and asymptotic methods, Moscow State University,

Moscow (1965).
67 Marchuk G.I. and Sarkisyan A.S. (eds), Mathematical models of circulation in oceans, Nauka, Novosibirsk (1980).
68 Mizokhata C., Theory of equations with partial derivatives, Mir, Moscow (1977).
69 Mikhailov V.P., Differential equations in partial derivatives, Nauka, Moscow (1983).
70 Mikhlin S.G., Lectures in mathematical physics, Nauka, Moscow (1968).
71 Mikhlin S.G., Variational methods in mathematical physics, Nauka, Moscow (1970).
72 Mikhlin S.G. and Smolitskii Kh.L., Approximate methods of solving differential and integral equations, Nauka, Moscow (1965).
73 Moiseev N.N., Asymptotic methods of non-linear mechanics, Nauka, Moscow (1981).
74 Nayfeh A.H., Perturbation methods, John Wiley, New York (1973).
75 Nikol'skii S.M., Approximation of functions of many variables and imbedding theorems, Nauka, Moscow (1969).
76 Nikiforov A.F. and Uvarov V.B., Special functions of mathematical physics, Nauka, Moscow (1984).
77 Nirenberg L., Lecturers in non-linear functional analysis, Mir, Moscow (1977).
78 Rozhdestvenskii B.L. and Yanenko N.N., Systems of quasi-linear equations, Nauka, Moscow (1968).
79 Ryaben'kii V.S., Introduction into numerical mathematics, Nauka, Moscow (1994).
80 Samarskii A.A., Theory of difference schemes, Nauka, Moscow (1982).
81 Samarskii A.A., Introduction into numerical methods, Nauka, Moscow (1982).
82 Samarskii A.A. and Gulin A.V., Numerical methods, Nauka, Moscow (1989).
83 Smirnov V.I., Lectures in higher mathematics, Vol. 4, Part 2, Nauka, Moscow (1981).
84 Sobolev S.L., Some applications of functional analysis in mathematical physics, Izd-vo LGU, Leningrad (1950).
85 Sobolev S.L., Equations of mathematical physics, Nauka, Moscow (1966).
86 Sologub V.S., Development of the theory of elliptical equations in the 18th and 19th Century, Naukova dumka, Kiev (1975).
87 Sretenskii L.N., Theory of the Newton potential, Gostekhizdat, Moscow and Leningrad (1946).
88 Steklov V.A., Main problems of mathematical physics, Nauka, Moscow (1983).
89 Sneddon I., Fourier transforms, IL, Moscow (1955).
90 Temam R., Navier-Stokes equations, Mir, Moscow (1981).
91 Tikhonov A.N. and Samarskii A.A., Equations of mathematical physics, Nauka, Moscow (1977).
92 Tikhonov A.N. and Samarskii A.A., Differential equations, Nauka, Moscow (1980).
93 Tranter K.J., Integral transforms in mathematical physics, Gostekhizdat, Moscow (1956).
94 Trenogin V.A., Functional analysis, Nauka, Moscow (1980).
95 Uspenskii S.V., Demidenko G.I. and Perepelkin V.T., Imbedding theorems and applications to differential equations, Nauka, Novosibirsk (1984).
96 Uflyand Ya.S., Integral transforms in the problems of elasticity theory, Izd-vo AN SSSR, Moscow and Leningrad (1963).
97 Wermer J., Potential theory, Mir, Moscow (1980).
98 Faddeev D.K. and Faddeeva V.N., Numerical methods of linear algebra, Fizmatgiz, Moscow and Leningrad (1963).
99 Henry D., Geometric theory of semi-linear parabolic equations, Springer, New York (1981).
100 Hille E. and Phillips R., Functional analysis and semi-groups, IL, Moscow (1962).
101 Shaidurov V.V., Multi-grid methods of finite elements, Nauka, Moscow (1989).
102 Yanenko N.N., Fractional-step methods of solving multi-dimensional problems of mathematical physics, Nauka, Novosibirsk (1967).
103 Agoshkov V.I., Boundary value problems for transport equations, Birkhauser, Basel (1998).
104 Ashyralyev A. and Sobolevskii P.E., Well-posedness of parabolic difference equations, Birkhauser, Basel (1994).
105 Bellman R., Perturbation techniques in mathematics, physics and engineering, Holt, New York (1964).

106 Bellman R. and Kalaba R.E., Quasilinearlisation and non-linear boundary-value problems, American Elsevier Publishing Company, New York (1965).
107 Bensoussan A., Lions J.L. and Papanicolau G., Asymptotic methods in periodic structures, Amsterdam, North Holland (1978).
108 Ciarlet P.G., Introduction to numerical linear algebra and optimisation, Cambridge University Press, Cambridge (1989).
109 Collatz L., Functional analysis and numerical mathematics, Academic Press, New York (1974).
110 Courant R. and Hilbert D., Methoden der mathematischen Physik, Springer, Berlin (1931).
111 Dubovskii P.B., Mathematical theory of coagulation, Seoul National University, Seoul, GARC-KOSFF (1994).
112 Dunford N. and Schwartz J.T., Linear operators, I, II, III, Wiley-Interscience, New York (1958).
113 Friedrichs K.O., Perturbation of spectra in Hilbert space, American Math. Society, Providence (1965).
114 Glowsinski R., Numerical methods for non-linear variational problems, Springer, New York (1984).
115 Lions J.L., Controllabilite exacte perturbations et stabilisation de systemes distribues, Masson, Paris (1988).
116 Marchuk G.I., Agoshkov V.I., Shutyaev V.P., Adjoint equations and peturbation algorithms in non-linear problems, CRC Press Inc., New York (1996).
117 Maslova N.B., Non-linear evolution equations, Kinetic approach, World Scientific, New York (1993).
118 Poincaré H., Les methods nouvelles de la mecanique celeste, Gauthier-Villar, Paris (1892).
119 Rayleigh L. and Stratt J.W., Theory of sound, McMillan, London (1926).
120 Rellich F., Storungthorie des Spekralzerlegung, Math. Ann., V. 117 (1936).
121 Rellich F., Perturbation theory of eigenvalue problems, Gordon and Breach Sci. Publ., New York (1969).
122 Samarskii A.A. and Vabishchevich P.N., Computational heat transfer, Wiley, Chichester (1995).
123 Schrödinger E., Quantisierung als Eigenwertproblem, Ann. Phys., 80 (1926).
124 Schwartz L., Theorie des Distributions, Hermann, Paris (1966).
125 Strang G. and Fix G.J., An analysis of the finite element method, Prentice-Hall, New York (1973).
126 Van Dyke M.D., Perturbation methods in fluid mechanics, Academic Press, New York (1964).
127 Whitham G.B., Linear and non-linear waves, John Wiley, New York (1974).

Index

A

adjoint nonlinear operators 279
alternating-direction method 252
approximation requirement 164
Arzelà–Ascoli theorem 6

B

Banach space 1, 5, 6, 206
Bessel inequality 12, 101
Bochner transform 131, 141
boundary of the domain 4
boundary-value conditions 33
boundary-value conditions of the first kind 98
boundary-value conditions of the third kind 98
boundary-value problem 33
Bubnov–Galerkin method 201

C

Cauchy problem 3, 33, 225
Cauchy–Bunyakovskii inequality 10, 46
Cauchy–Kovalevskii theorem 37
characteristic function 4
Chebyshef polynomial 107
Chernov formula 230
Chézy coefficient 264
class of functions 4
closed linear operator 16
coercivity condition 283
collocation method 208
compact set 5
completeness equation 12
convergence in respect to energy 16
convergence theorem 173, 240
convolution transform 131
Coriolis parameter 264
Courant method 196
Cranck–Nicholson scheme 238
cylindrical coordinates 58
cylindrical function 106

D

d'Alembert operator 2, 26
differential operator 15
diffusion length 150
dipole coefficient 67
direct methods 189
Dirichlet boundary condition 23
Dirichlet integral 2, 38
Dirichlet problem 38, 71
discrete Laplace transform 130
domain of definition of the equation 32
Dubois–Raymond lemma 19

E

eigen subspace 18
eigenfunction 97
eigenvalue problems 94
energetic space 16
energy method 46
energy space 192
equation of state 29
equations of continuity 29
Euclidean distance 3
Euclidean space 3, 11
Euler equation 104
Euler equation of motion 29
evolution equations 225

F

factorisation method 169
finite-difference methods 166
first Green formula 60
Fourier coefficient 101
Fourier map 132
Fourier series 10
Fourier transform 133
Fourier–Bessel integral 136
Fréchet derivative 278
Fréchet differential 278
Fredholm equation 2, 73
Fredholm integral equation 50

functional of the energy method 46

G

Galerkin approximation 294
Galerkin method 201
Galerkin–Petrov method 201
Gateaux derivative 277
Gauss hypergeometrical series 109
generalized derivatives 19
generalized Minkovskii inequality 9
generalized solutions 37
generating function 160
Gilbert transform 131
Green formula 21, 40
Green function 76, 84

H

Haar system 142
Hankel function 107
Hankel image 146
Hankel transform 131, 136
Hardy inequality 9
heat conductivity equation 27
Helmholtz equation 26
Hermite polynomial 108
Hermitian weakly polar kernel 54
Hilbert spaces 9
Hilbert transform 140
Hilbert–Schmidt orthogonalization 12
Hilbert–Schmidt theorem 54
Hölder function 1
Hölder inequality 9
Hölder space 7
hyperbolicity condition 32
hyperplane method 183

I

ill-posed problem 36
integral Green formula 60
internal point of the set 3
internal spherical functions 104
isometric space 5
isomorphous space 5

J

jump of the normal derivative 66

K

Kantorovich method 196
Kellog theorem 240
kernel 14
kinetic coagulation equation 159
Kontorovich–Lebedev transform 131, 138
Kummer function 109

L

Laguerre polynomial 108
Laguerre transform 131
Laplace equation 23
Laplace operator 2, 59
Laplace–Beltrami operator 105
law of oscillation 96
Lebesgue integral 7
Lebesgue space 7
Legendre polynomial 105
Legendre transform 131
linear functional 14
linear normalised space 5
linear operator 13
linear sets 5
Lipschitz inequality 277
Lipschits condition 4
logarithmic potential 63
logarithmic simple layer potential 66
longitudinal–transverse sweep 252
Lyapunov surface 65, 76

M

MacDonald function 138
Mathieu equation 109
Mathieu function 109
Maxwell equations 27
Mehler–Fock transform 131
Meller–Fock transform 139
Mellin transform 131
method of arbitrary lines 182
method of eigenfunctions 232
method of Gauss exclusion 169
method of integral identities 210
method of least squares 195
method of Marchuk's integral identity 211
method of stationarisation 232
method of two-cyclic multi-component splitting 245
method of weak approximation 254
Meyer transform 138
minimising sequence 190
mixed problem 33

moments method 204
multiplicity 17

N

Navier–Stokes equation 262
neighbourhood of the set 3
net method 166
Neumann condition 23
Neumann problem 34, 40, 48
Newton iteration process 291
Newton potential 61
nonhomogeneous evolution equation 228
normalized eigenfunction 100

O

one-dimensional wave equation 24
oriented surface 67
orthogonal basis 10
orthogonal system 10
orthonormalized system 10
Ostrogradskii–Gauss formula 60

P

Parseval equality 102
Parseval identity 134
Parseval–Steklov equality 12
periodicity conditions 99
piecewise smooth surface 4
Poincaré inequality 41
Poincaré–Perron method 80
point spectrum 18
Poisson equation 23
Poisson integral 73
potential of the vector field 57
predictor-corrector method 250
problem of a string 124
projection method 205
projection-grid method 208

Q

quadrature method 187

R

reflexive real Banach space 283
Reiss–Fischer theorem 11
retardation time 92
Riesz theorem 8, 41, 46
Ritz method 190, 289
Rodrig formula 106

S

Schmidt equation 54
Schwarz method 78
second Green formula 60
second Hankel function 82
simple eigenvalue 18, 97
simple layer potential 64
Sobolev classes 20
Sobolev space 1, 19
special functions 103
spherical coordinates 58
spherical function 103
splitting method 225, 242
stability 239
standing wave 96
steepest descent method 286
Stricker coefficient 264
Struve function 137
Sturm–Liouville operator 103
super-harmonic function 80
support 5
sweep method 169, 241
sweeping method 80
symmetric hyperbolic systems 32
symmetric kernel 54

T

The minimisation condition J(uN) leads to a system 195
three-dimensional wave equation 25
transfer equation 28
Trefftz method 197
Tricomi equation 31
Trotter formula 230
two-cyclic method of weak approximation 254
two-dimensional wave equation 25

V

variational formulation of problems 45
variational method 283
Volterra equation 89

W

wavelet integral transform 142
wavelet transform 131
weakly polar kernel 51

Weber transform 137
Weierstrass theorem 6, 11
Weierstrasse theorem 282
weight function 99

Y

Young inequality 9