TOMSK POLYTECHNIC UNIVERSITY

S.A. Lopatkin, V.I. Reizlin

ADDITIONAL CHAPTERS ON MATHEMATICS

Recommended for publishing as a study aid
by the Editorial Board of the Tomsk Polytechnic University

Tomsk Polytechnic University Publishing House
2008

UDC 51(075.8)
BBC 22.1я73
L89

Lopatkin S.A.
L89 Additional Chapters on Mathematics: study aid / S.A. Lopatkin, V.I. Reizlin. – Tomsk: TPU Publishing House, 2008. – 126 p.

ISBN 5-98298-315-2

The study aid contains computational methods, basics of edge problems, solution for ordinary differential and partial derivative equations as well as some issues on functional analysis.

The study aid is developed in the framework of Innovative Educational Programme of TPU on the direction "Power-saving, basic, special, and industrial discharge, radiation, and plasma-beam technologies". The manual is intended for training students majoring in the course 140200 "Power Electrical Engineering" and studying the master programme "Technology and Physics of High Voltage".

UDC 51(075.8)
BBC 22.1я73

Reviewer

Doctor of Phisics and Mathematics,
Professor of the Tomsk State University
A.I. Potekaev

ISBN 5-98298-315-2

1. INTRODUCTION

Men live in the world of models created by his own, the construction of which is not just a whim but the way to perceive the reality. We got used to it so much that regular modeling is considered to be natural. The models able to render qualitative characteristic of the studied process are called qualitative. More complicated models are quantitative ones with the help of which one may predict precise numerical values of phenomenon characteristics. Weather casting based on national superstitions is a qualitative one. Meanwhile, those used at weather casting stations and meant to predict numerical values of weather characteristics (such as temperature, wind velocity, etc.) refer to quantitative models. Mathematical model is the total of equations (algebraic, differential, integral), describing processes in the phenomenon being modeled. Each equation of such type is already a lower level process model on its own. Models of more complex phenomena are constructed, as a rule, from simple models. Mathematical modeling results in range of formulae (in case of successful equation analytical solution finding), which allow to calculate the characteristics of the process modeled and tables for their values (in case of the analytical solution is impossible to be found). Frequently, one comes across with the last case which implies some computational procedures so that to get the solution being tabled. These procedures are called numerical models. The feature of such models is the real possibility to get quite approximate solution. It is caused by computing machines discontinuity, the lack of storage, processing speed, etc. Our purpose is to learn how to research the algorithm for finding a model equation numerical solution.

By means of mathematics, motion of lower number particles systems is basically given through ordinary differential equations. In case the number of particles is greater, the separate particles motion observation is almost impossible. It becomes more convenient to consider particles system as continuum environment and characterize it by the average values, such as density, temperature at the point, etc.

Continuum environment mathematical models lead to partial derivative equations, to which the averages mentioned above are satisfactory. For example, temperature changes in motionless body is expressed by means of thermal conductivity equation

$$c(u,r,t)\cdot\frac{\partial u}{\partial t}=\operatorname{div}\left[k(u,r,t)\cdot\operatorname{grad}\ u\right]+q(u,r,t), \tag{1.1}$$

here u is temperature, c is thermal capacity, k is thermal conductivity coefficient and q is density of thermal sources.

The partial derivative equations are attached with problems on gas dynamics, thermal conductivity, radioactive transmission, neutron diffusion, theory of elasticity, electromagnetic fields, transport processes in gases, quantum mechanics and many others.

As for problems of physics the independent variables are presented by t – time, r – position data and some others (e. g. $\boldsymbol{v}$ – particles velocity in diffusion problems). It is demanded to find variation of $G(t, r, \boldsymbol{v}, \ldots)$ independent variables in some domain. The complete mathematical problem contains differential equation, as well as additional conditions to point out the only solution within its range for differential equation. The additional conditions are usually given on the boarder of G domain.

If time t is one of variables, more often a domain becomes

$$G(t,r,\ldots)=g(r,\ldots)\times[t_0,T], \tag{1.2}$$

The solution is found in some spatial domain $g(r, \ldots)$ on duration segment $t_0 \le t \le T$. Thus, the conditions given at $t = t_0$ are called initial and those given at the $\Gamma(r)$ border of $g(r)$ domain are boundary or edge data.

A problem with initial conditions only is called Cauchy problem. For example, one can set a problem for thermal conductivity equation in non bounded space with initial conditions as

$$u(r,t_0)=\mu\ (r). \tag{1.3}$$

If $\mu(r)$ is a bounded piecewise continuous function, the solution of the equations (1.1) and (1.3) is the only one in bounded functions group (when the equation coefficients are limited somehow).

Problem with initial and boundary conditions is a mixed edge or non stationary edge problem. The additional conditions for equation (1.1) may become as follows:

$$u(r,t_0)=\mu\ (r),\quad r\in g(r),\quad u(r,t)_\Gamma=\mu_1\ (r,t),\quad t_0\le t\le T. \tag{1.4}$$

Other boundary conditions are feasible for given equation, those containing normal derivative of solution.

There are problems in which $G(t, r)$ becomes quite different. The mere example is a problem with characteristics conditions appearing while there is studying of drying process, gas trapping and many other processes.

While researching established conditions or stationary processes (not depending on time) in continuum environment, some mathematical problems

independent from time are set. Their solution is searched for in $g(r)$ domain, meanwhile the additional conditions are boundary. These are edge problems.

In this tutorial we take into consideration only correctly identified problems, where for some class of initial and boundary data where a solution exists, the sole one, and dependent on these data continuously. One should presume the continuous dependency of solution on equation coefficients.

Most of problems mentioned above can be expressed by operator form. Accordingly, thermal conductivity equation $\partial_t T = \mu\Delta T + Q,$ wave equation $\partial_{tt} U = \alpha\partial_{xx} U,$ a diffusion equation of some value C with velocity v along the axis x $\partial_t C + \boldsymbol{v}\partial_x C = 0$ can be written as

$$(\partial_t - \mu\Delta)T = Q;$$
$$(\partial_{tt} - \alpha\partial_{xx})U = 0;$$
$$(\partial_t + \boldsymbol{v}\partial_x)C = 0.$$

All those in brackets are operators. In case of designating it as A, these equations become as follows:

$$\mathrm{A}_1 T = Q, \quad \mathrm{A}_1 = \partial_t - \mu\Delta;$$
$$\mathrm{A}_2 U = 0, \quad \mathrm{A}_2 = \partial_{tt} - \alpha\,\partial_{xx};$$
$$\mathrm{A}_3 C = 0, \quad \mathrm{A}_3 = \partial_t + \boldsymbol{v}\partial_x.$$

Operator equation $\mathrm{A}f = g$ is possible to be interpreted in different ways. So, the function f under A becomes a function g. Another interpretation is possible (this one is to be followed). Assuming there are two sets of F and G functions and f is an element of F set as g is an element of G set. The operator A represents the compliance between the sets F and G. Taking into account that the subset g from the domain G and the type of the operator A are known, we are to find the subset f from the domain F.

Linear operators properties and other necessary principles from linear algebra and functional analysis are given in Appendix.

2. SOLVING THE EDGE PROBLEMS FOR ORDINARY DIFFERENTIAL EQUATIONS AND SYSTEMS

An edge problem is the problem of finding a particular solution of the system

$$\frac{d}{dx}U_k(x) = f_k(x, U_1, U_2, ..., U_p),\ 1 \le k \le p \tag{2.1}$$

on the segment $a \le x \le b$, whereas additional conditions are imposed on $U_k(x)$ function values at more than one point of this segment.

Additional conditions may join the values of several functions to one another; then for a system (2.1) of order p they become as follows:

$$\begin{aligned} &\varphi_k(U_1(\xi_k), U_2(\xi_k), ..., U_p(\xi_k)) = \eta_k, \\ &1 \le k \le p, \quad a \le \xi_k \le b. \end{aligned} \tag{2.2}$$

Generally, there are many problems with more complicated conditions. Let's notice that p-order equation

$$y^{(p)}(x) = F(x, y(x), y'(x), ..., y^{(p-1)}(x)), \tag{2.3}$$

where $y^{(p)}(x)$ is the derivative of order k, $k = 0, 1, \ldots, p$, $y^{(0)}(x) = y(x)$, can be turned into a system of differential equations (2.1) by variables replacement:

$$\begin{aligned} &U_0(x) = y(x); \\ &U_1(x) = y'(x); \\ &\ldots \\ &U_{p-2}(x) = y^{(p-2)}(x); \\ &U_{p-1}(x) = y^{(p-1)}(x). \end{aligned} \tag{2.4}$$

Indeed, according to the replacement (2.4),

$$\begin{aligned} &U_0'(x) = y'(x) = U_1(x); \\ &U_1'(x) = y''(x) = U_2(x); \\ &\ldots \\ &U_{p-2}'(x) = y^{(p-1)}(x) = U_{(p-1)}(x), \end{aligned}$$

and the equation (2.3) becomes the system of formulae (2.1):

$$U_k'(x) = U_{k+1}(x), \quad k = 0,1,..., p-2;$$
$$U_{p-1}'(x) = F(x, U_0(x),..., U_{(p-1)}(x)).$$

Here the last equation is obtained by substitution of the equation (2.4) into (2.3).

An example of simple edge problem for second-order differential equation is the problem of searching for static sag $y(x)$ of loaded string with fixed ends:

$$y''(x) = -f(x), \quad a \le x \le b, \quad y(a) = y(b) = 0. \tag{2.5}$$

Here $f(x)$ is an external bending load for a unit string length divided by string elasticity.

It should be mentioned that the general edge problem (2.1) may:

- have no solutions;
- have one solution;
- have several or even infinitely many of them.

Examples:

1. An edge problem $y'' + y = 0, \quad y(0) = y(\pi) = 0$ has infinitely many solutions like y = C sin(x), where C is an arbitrary constant.
2. An edge problem $y'' + y = 0, \quad y(0) = 0, \quad y(b) = 1$ when $0 < b < \pi$, has the only solution $y_b = \dfrac{\sin x}{\cos b}$, and for $b = \pi$, there are no solutions at all.

Further, the solution of the edge problem supposed to be existing.

Let's consider in more details an important particular case, when differential equation and edge conditions are linear. That is a **linear edge problem**.

Linear differential equation of order n can be reduced to

$$L[y] = f(x), \tag{2.6}$$

where $L[y] = p_0(x)y^{(n)} + p_1(x)y^{(n-1)} + ... + p_n(x)y$.

It is usually supposed that $p_i(x)$ $(i = 0, 1,..., n)$ and $f(x)$ are known continuous functions on the given segment $[a, b]$.

To simplify, we assume edge conditions to contain two abscissas: $x_1 = a$ and $x_2 = b$ $(a < b)$, as the ends of $[a, b]$ segment. These conditions are called two-point edge conditions. The edge conditions are called linear ones if they have the form as

$$R_\nu[y] = \gamma_\nu, \quad \nu = 1,2,...,n, \tag{2.7}$$

where $R_\nu[y] = \sum_{k=0}^{n-1} [\alpha_k^{(\nu)} y^{(k)}(a) + \beta_k^{(\nu)} y^{(k)}(b)]$

and $\alpha_k^{(\nu)}$, $\beta_k^{(\nu)}$, γ_ν are given constants, moreover

$$\sum_{k=0}^{n-1}\left(\left|\alpha_k^{(\nu)}\right|+\left|\beta_k^{(\nu)}\right|\right)\neq 0$$

for $v = 1, 2, \ldots, n$.

For example, the edge conditions given above are linear.

Linear edge conditions are also presented by periodicity conditions, which for second-order differential equation become

$$y(a) = y(b),\ \ y'(a) = y'(b).$$

The linear edge problem is **homogeneous** if:

- firstly, $f(x) \equiv 0$ at $a \le x \le b$, that means the differential equation (2.6) is homogeneous;
- secondly, $\gamma_\nu = 0$; $\nu = 1, 2, \ldots, n$, that provides homogeneous edge conditions.

Otherwise, the problem stabed with the formulae (2.6)–(2.7) is **inhomogeneous**.

Example 1. Let's take the problem about deflection of a horizontal beam of length l, situated on two supports and being under distributed shear load with linear density of $q = q(x)$ (Fig. 1).

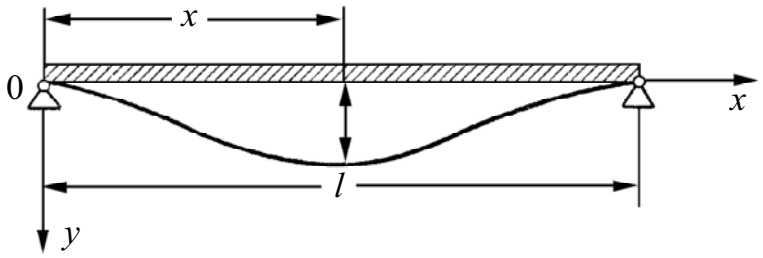


Fig. 1. Horizontal beam deflection problem

It is known from strength of materials that vertical deflection of homogeneous beam satisfies approximately to linear differential equation

$$[EI(x)y'']'' = q(x), \tag{2.8}$$

where $EI(x)$ is beam flexural stiffness, meanwhile deflecting moment M and shearing force Q are defined by the relations

$$M = EI(x)y''$$

and

$$Q = M' = [EI(x)y'']'.$$

Edge conditions depend on ways of beam-ends restraint. There are some basic cases:

1. The end is free. Moment of deflection M and shearing force Q both equal zero. Consequently, *the edge conditions for free beam end* are as follows:

$$y'' = 0\,;\ \ y''' = 0. \tag{2.9^a}$$

2. The end is supported hingedly. The deflection y and deflecting moment M both equal zero. Therefore, *the edge conditions for the hinged end* are as follows:

$$y = 0 \text{ and } y'' = 0. \tag{2.9^b}$$

3. The end is fixed toughly. The deflection y and the angle of turning $\varphi = \operatorname{arctg} y'$ both equal zero. In the view of this, *the edge conditions of tightly tailed end* are as follows:

$$y = 0 \text{ and } y' = 0. \tag{2.9^c}$$

Also, other more complicated cases for edge conditions are possible.

Obviously, the problems (2.8)–(2.9) are linear edge problems.

Example 2. Let beam flexural stiffness EI be constant. In this case, the equation (2.8) for deflection y is replaced by the following equation:

$$EIy^{IV} = q(x). \tag{2.10}$$

Supposing, the beam is fixed hingedly at the end $x = 0$ and fixed toughly at the end $x = l$. For the deflection y to be fulfilled it provides following edge (boundary) conditions:

$$\left.\begin{aligned} y(0) = 0,\ \ y''(0) = 0, \\ y(l) = 0,\ \ y'(l) = 0. \end{aligned}\right\} \tag{2.11}$$

The edge conditions (2.11) are linear and homogeneous.

The problems (2.10)–(2.11) are easy to be solved. Therefore, supposing for simplicity that load density is constant:

$$q(x) = p,$$

we have

$$EIy = \frac{px^4}{24} + c_1x^3 + c_2x^2 + c_3x + c_4.$$

From boundary conditions (2.11) it follows that

$$c_1 = -\frac{pl}{16},\ \ c_2 = 0,\ \ c_3 = \frac{pl^3}{48},\ \ c_4 = 0.$$

Thus, the solution is

$$y = \frac{p}{48EI}(2x^4 - 3lx^3 + l^3x).$$

This example makes evident that in case we can find the common solution for differential equation, solution of the two-point edge problem is of near the same difficulty as the problem with initial conditions. However,

if one common solution can't be found using regular means, the edge problem solution will lead to a new difficulty. It results from absence of initial point that could be an origin for constructing the solution with one of above-mentioned methods.

2.1. Shooting method

This is a numerical method, aimed at turning an edge problem to solution of *sequence* of Cauchy problems for the same system of differential equations. Let's discuss this method applied to the simplest problem for the system of two firs-order differential equations with general edge conditions:

$$\left.\begin{aligned} U'(x) &= f(x, U, V), \\ V'(x) &= g(x, U, V), \\ a <= x <= b, & \end{aligned}\right\} \qquad (2.15^{a})$$

$$\varphi\ (U(a),\ V(a)) = 0;\ \psi\ (U(b),\ V(b)) = 0\,. \qquad (2.15^{b})$$

Let's take an arbitrary value $U(a) = \eta$ and examine left edge condition as algebraic equation

$$\varphi\ (\eta,\ V(a)) = 0$$

and then find $V(a) = \xi(\eta)$ from it. After that, we use $U(a) = \eta,\ \ V(a) = \xi$ as initial conditions of Cauchy problem for the system (2.15^{a}) and integrate Cauchy problem by one of numerical methods. In this case we get solution $U(x,\eta),\ \ V(x,\eta)$ depending on η as on the parameter.

The value ξ is chosen in such a way that the solution found satisfy left edge condition (2.15^{b}). On the other hand, it probably doesn't satisfy right edge condition. On it's substitution, the left part of edge condition at b point, considered as function of parameter η

$$\bar{\psi}(\eta) = \psi(U(b,\eta), V(b,\eta))\,, \qquad (2.16)$$

isn't equal to zero.

It is necessary to change the parameter somehow until we get the value η for which $\bar{\psi}(\eta) \approx 0$ with required precision. Thus, the solution of the edge problem (2.15) turns to finding the root of equation

$$\bar{\psi}(\eta) = 0\,. \qquad (2.17)$$

The simplest method to find the root is **dichotomy** (bisection of the segment).

One makes test "shootings", so called calculations with hit-and-miss chosen values η_i until there will be $\bar{\psi}(\eta_i)$ values with different signs. The pair of such values η_i and η_{i+1} makes up a "bracket". Successive halving

of $[\eta_i, \eta_{i+1}]$ segment till the required precision achieved, we make a "zeroing in on" parameter η. That's why the method was called shooting.

Searching for each new value of function $\bar{\psi}(\eta)$ requires the numerical integration of system (2.15[a]), that is time taking.

The root of equation (2.17) is better to find by a faster method.

Let's try to use the Newton's method:

$$\eta_{i+1} = \eta_i - \frac{\bar{\psi}(\eta_i)}{\bar{\psi}'(\eta_i)}. \tag{2.18}$$

The calculation of the derivative $\bar{\psi}'(\eta_i)$ is difficult, so it is better to replace it with the difference ratio

$$\bar{\psi}'(\eta_i) \approx \frac{\bar{\psi}(\eta_i) - \bar{\psi}(\eta_{i-1})}{\eta_i - \eta_{i-1}}. \tag{2.19}$$

Substituting the formula (2.19) into (2.18), we get an iterative formula for **secant method**:

$$\eta_{i+1} = \eta_i - \frac{(\eta_i - \eta_{i-1})\bar{\psi}(\eta_i)}{\bar{\psi}(\eta_i) - \bar{\psi}(\eta_{i-1})}. \tag{2.20}$$

In this method the first two calculations are carried out with occasionally chosen values η_0 and η_1 close to each other. The next parameter values are calculated using the formula (2.20).

It needs to be mentioned that the method converges very quickly near the root. The convergence far from the root depends on the proper choice of a zero approximation.

Now let's consider a **linear** edge problem, for which the solution is very simple:

$$\left.\begin{aligned} &U'(x) = \alpha_1(x)\cdot U + \beta_1(x)\cdot V + \gamma_1(x), \\ &V'(x) = \alpha_2(x)\cdot U + \beta_2(x)\cdot V + \gamma_2(x), \\ &a <= x <= b, \end{aligned}\right\} \tag{2.21[a]}$$

$$p_1\cdot U(a) + q_1\cdot V(a) = t_1,\ \ p_2\cdot U(b) + q_2\cdot V(b) = t_2. \tag{2.21[b]}$$

Let's use well-known principle from the theory of differential equations. It provides that a common solution to *linear inhomogeneous system* equals the sum of its any *particular solution* and *common solution of the corresponding homogeneous system*.

Let's find a particular solution to *inhomogeneous* system (2.21[a]), putting $U(a) = \eta_0 = 0$ into the left-hand condition of the formulae (2.21[b]). We denote this solution as $U_0(x),\ \ V_0(x)$ and notice that $V_0(a) = \dfrac{t_1}{q_1}$.

Now, the corresponding *homogeneous* system is

$$\left.\begin{aligned} U'(x) &= \alpha_1(x)\cdot U + \beta_1(x)\cdot V, \\ V'(x) &= \alpha_2(x)\cdot U + \beta_2(x)\cdot V \end{aligned}\right\}$$

with *homogeneous* initial conditions

$$U(a) = \eta_1 = 1, \quad V(a) = -\frac{p_1}{q_1}.$$

Let's solve this Cauchy problem and express it through $U_1(x)$, $V_1(x)$. Consider functions $U(x) = U_0(x) + CU_1(x)$ and $V(x) = V_1(x) + CV_1(x)$. Evidently, these functions satisfy the edge condition at point *a*:

$$p_1\cdot U(a) + q_1\cdot V(a) = p_1\cdot(U_0(a) + C\cdot U_1(a)) + q_1\cdot(V_0(a) + C\cdot V_1(a)) =$$

$$p_1(0 + C\cdot 1) + q_1\left(\frac{t_1}{q_1} + C\left(-\frac{p_1}{q_1}\right)\right) = C\cdot p_1 + t_1 - C\cdot 1 p_1 = t_1.$$

That is why the common solution to inhomogeneous Cauchy problem, satisfying the left-hand edge condition (2.21^b), is presented by one-parameter set

$$U(x) = U_0(x) + C\cdot U_1(x),\ V(x) = V_0(x) + C\cdot V_1(x). \tag{2.22}$$

The value of parameter C is chosen to satisfy the right-hand edge condition (2.21^b):

$$C = -\frac{p_2\cdot U_0(b) \ + \ q_2\cdot V_0(b) - t_2}{p_2\cdot U_1(b) \ + \ q_2\cdot V_1(b)}. \tag{2.23}$$

Desired solution of the edge problem (2.21) is found with the formula (2.23).

Therefore, the solution to the linear edge problem needs only two "shootings", and auxiliary Cauchy problems are solved twice.

2.2. Method of finite differences or the mesh method

Let's consider the linear edge problem

$$\begin{aligned} &y'' + p(x)\cdot y' + q(x)\cdot y = f(x), \\ &a \le x \le b, \end{aligned} \tag{2.24}$$

$$\left.\begin{aligned} \alpha_0\cdot y(a) + \alpha_1\cdot y'(a) &= A, \\ \beta_0\cdot y(b) + \beta_1\cdot y'(b) &= B, \end{aligned}\right\} \tag{2.25}$$

$$(|\alpha_0| + |\alpha_1| \ne 0, \quad |\beta_0| + |\beta_1| \ne 0),$$

where $p(x)$, $q(x)$ and $f(x)$ are continuous on $[a; b]$.

Let's divide the segment $[a; b]$ into *n* equal parts of length or intervals

$$h = \frac{(b-a)}{n}.$$

The points of division $y_i = y_i(x_i),\ y_i' = y'(x_i),\ y_i'' = y''(x_i)$, where $x_i = x_0 + i \cdot h,\ \ i = 0,\ 1,...,\ n,\ x_0 = a,\ \ x_n = b,$ are called nodes and their arrangement is called a **mesh** (grid and lattice as well) on the segment [*a*; *b*]. The values of the desired function $y = y(x)$ at nodes $x_i = x_0 + i \cdot h$ and the derivatives $y' = y'(x),\ y'' = y''(x)$ we denote as $y_i = y_i(x_i),\ y_i' = y'(x_i),\ y_i'' = y''(x_i)$.

Let's introduce notation $p_i = p(x_i),\ q_i = q(x_i),\ f_i = f(x_i).$

Then substitute the derivatives for so called one-side finite difference ratio

$$\left.\begin{aligned} y_i' &\approx \frac{y_{i+1} - y_i}{h}, \\ y_i'' &\approx \frac{y_{i+1}' - y_i'}{h} = \frac{\dfrac{y_{i+2} - y_{i+1}}{h} - \dfrac{y_{i+1} - y_i}{h}}{h} = \frac{y_{i+2} - 2 \cdot y_{i+1} + y_i}{h^2}. \end{aligned}\right\} \qquad (2.26)$$

Equations (2.26) express approximately the values of derivatives in the internal points of interval [*a*, *b*].

For boundary points we put

$$y_0' = \frac{y_1 - y_0}{h},\ \ y_n' = -\frac{y_{n-1} - y_n}{h}. \qquad (2.27)$$

Using equations (2.26), the differential equation (2.24) for $x = x_i$, (*i* = 1, 2, …, *n* – 1) may be approximately replaced by the linear system of equations

$$\begin{aligned} &\frac{y_{i+2} - 2 \cdot y_{i+1} + y_i}{h^2} + p_i \cdot \frac{y_{i+1} - y_i}{h} + q_i \cdot y_i = f_i, \\ &i = 0,\ 1,...,\ n-2. \end{aligned} \qquad (2.28)$$

Besides, because of equations (2.27), the edge conditions (2.25) give two more additional equations:

$$\alpha_0 y_0 + \alpha_1 \frac{y_1 - y_0}{h} = A, \quad \beta_0 y_n + \beta_1 \frac{y_n - y_{n-1}}{h} = B. \qquad (2.29)$$

Thus, we have linear system of (*n* + 1) equations with (*n* + 1) unknown variables $y_0, y_1, ..., y_n$, being the values of desired function *y*(*x*) at mesh nodes. The system of equations (2.28), (2.29) substituting approximately the edge problem (2.24), (2.25) is commonly called the **difference scheme**. This system can be solved by any common numerical method. Nevertheless, the schemes (2.28), (2.29) are of specific kind and it may be solved by the specific method, usually referred to as **sweep method**. Specific character of the system is provided by its formulae content: equations include three neighboring unknown variables and the matrix of the system is three-diagonal.

Let' s rearrange the equations (2.28):

$$y_{i+2} + (-2 + h \cdot p_i) y_{i+1} + (1 - h \cdot p_i + h^2 q_i) y_i = f_i \cdot h^2 . \qquad (2.30)$$

Introducing notation $-2 + h \cdot p_i = m_i, 1 - h \cdot p_i + h^2 \cdot q_i = n_i,$ we get

$$y_{i+2} + m_i \cdot y_{i+1} + n_i \cdot y_i = f_i \cdot h^2, (i = 0, 1, \ldots, n - 2). \qquad (2.31)$$

The edge conditions may be written in the same way:

$$\alpha_0 \cdot y_0 + \alpha_1 \cdot \frac{y_1 - y_0}{h} = A, \qquad \beta_0 \cdot y_n + \beta_1 \frac{y_n - y_{n-1}}{h} = B . \qquad (2.32)$$

The sweep method's idea is as follows. Let's evaluate equation (2.31) with respect to y_{i+1}:

$$y_{i+1} = \frac{f_i}{m_i} h^2 - \frac{n_i}{m_i} y_i - \frac{1}{m_i} \cdot y_{i+2} . \qquad (2.33)$$

Supposing that the member containing y_i is excluded from equation with help of the total system (2.31), the equation can be written as

$$y_{i+1} = c_i \cdot (d_i - y_{i+2}), \qquad (2.34)$$

where c_i and d_i must be defined. Let's find the equations for these coefficients. For $i = 0$ from equation (2.33) and edge conditions (2.32) it follows that

$$y_1 = \frac{h^2}{m_0} \cdot f_0 - \frac{n_0}{m_0} \cdot y_0 - \frac{1}{m_0} y_2,$$

$$y_0 = \frac{\alpha_1 y_1 - A \cdot h}{\alpha_1 - \alpha_0 \cdot h}.$$

Excluding y_0 from these two equations we find

$$y_1 = \frac{f_0}{m_0} \cdot h^2 - \frac{n_0}{m_0} \cdot \frac{\alpha_1 \cdot y_1 - A \cdot h}{\alpha_1 - \alpha_0 \cdot h} - \frac{1}{m_0} \cdot y_2 .$$

Now, we evaluate y_1:

$$y_1 = \frac{\frac{n_0}{m_0} \cdot \frac{A \cdot h}{\alpha_1 - \alpha_0 \cdot h} + \frac{f_0}{m_0} \cdot h^2 - \frac{1}{m_0} \cdot y_2}{1 + \frac{n_0}{m_0} \cdot \frac{\alpha_1}{\alpha_1 - \alpha_0 \cdot h}} =$$

$$= \frac{\alpha_1 - \alpha_0 \cdot h}{m_0 \cdot (\alpha_1 - \alpha_0 \cdot h) + n_0 \cdot \alpha_1} \left(\frac{n_0 \cdot A \cdot h}{\alpha_1 - \alpha_0 \cdot h} + f_0 \cdot h^2 - y_2 \right). \qquad (2.35)$$

However, according to the equation (2.34)

$$y_1 = c_0(d_0 - y_2). \tag{2.36}$$

Comparing Eqs. (2.25) and (2.26), we get

$$\left.\begin{aligned} c_0 &= \frac{\alpha_1 - \alpha_0 \cdot h}{m_0 \cdot (\alpha_1 - \alpha_0 \cdot h) + n_0 \cdot \alpha_1}, \\ d_0 &= \frac{n_0 \cdot A \cdot h}{\alpha_1 - \alpha_0 h} + f_0 h^2. \end{aligned}\right\} \tag{2.37}$$

Let us supposing that $i > 0$, i. e. $i = 1, 2, \ldots, (n - 2)$. Evaluating y_i according to the equation (2.34), we have

$$y_i = c_{i-1} \cdot d_{i-1} - c_{i-1} \cdot y_{i+1}.$$

Substituting this into the equation (2.33), we get

$$y_{i+1} = \frac{f_i}{m_i} \cdot h^2 - \frac{n_i}{m_i}(c_{i-1} \cdot d_{i-1} - c_{i-1} \cdot y_{i+1}) - \frac{1}{m_i} y_{i+2}.$$

Solving the obtained equation with regard to y_{i+1}, we find

$$y_{i+1} = \frac{\frac{f_i}{m_i} h^2 - \frac{n_i}{m_i} \cdot c_{i-1} \cdot d_{i-1} - \frac{1}{m_i} \cdot y_{i+2}}{1 - \frac{n_i}{m_i} \cdot c_{i-1}},$$

or

$$y_{i+1} = \frac{1}{m_i - n_i c_{i-1}}(f_i h^2 - n_i c_{i-1} d_{i-1} - y_{i+2}). \tag{2.38}$$

Comparing equations (2.34) and (2.38), we get the recurrent formulae for coefficients c_i and d_i:

$$\left.\begin{aligned} c_i &= \frac{1}{m_i - n_i \cdot c_{i-1}}, \\ d_i &= f_i h^2 - n_i c_{i-1} d_{i-1}, \\ i &= 1, 2, \ldots, n - 2. \end{aligned}\right\} \tag{2.39}$$

Therefore, as c_0 and d_0 are already found from equation (2.37), the coefficients c_i and d_i to c_{n-2} and d_{n-2} may be obtained inclusively one after another by means of equation (2.39). These calculations are called **the direct sweep** of sweep method.

From equation (2.33) for $i = n - 2$ and the second edge condition (2.32) it goes

$$\left.\begin{aligned} y_{n-1} &= c_{n-2}(d_{n-2} - y_n), \\ \beta_0 y_n + \beta_1 \frac{y_n - y_{n-1}}{h} &= B. \end{aligned}\right\}$$

Solving this system with regard to y_n, we get

$$y_n = \frac{\beta_1 \cdot c_{n-2} \cdot d_{n-2} + B \cdot h}{\beta_1 \cdot (1 + c_{n-2}) + \beta_0 \cdot h}. \tag{2.40}$$

Now, using Eq. (2.34) and the first edge condition (2.32), we can find $y_{n-1},\ y_{n-2},\ldots,y_0$. It is **the inverse sweep** of sweep method.

Thus, we get the following chain:

$$\left.\begin{aligned} &y_{n-1} = c_{n-2}(d_{n-2} - y_n), \\ &y_{n-2} = c_{n-3}(d_{n-3} - y_{n-1}), \\ &\ldots \\ &y_1 = c_0(d_0 - y_2), \\ &y_0 = \frac{\alpha_1 y_1 - Ah}{\alpha_1 - \alpha_0 h}, \end{aligned}\right\} \quad \alpha_0 = 1,\ \ \alpha_1 = 0,\ \ \beta_0 = 1,\ \ \beta_1 = 0. \tag{2.41}$$

For the simplest edge conditions $y(a) = A,\ \ y(b) = B$, the equations for $c_0,\ \ d_0,\ \ y_0$, and y_n become simplified. Presuming in this case

$$\alpha_0 = 1,\ \ \alpha_1 = 0,\ \ \beta_0 = 1,\ \ \beta_1 = 0,$$

from Eqs. (2.37), (2.40), and (2.41) we have

$$c_0 = \frac{1}{m_0},\ \ d_0 = -n_0 A + f_0 h^2,$$

$$y_n = B,\ \ y_0 = A\ .$$

The considered approach turns the edge linear problem to the system of linear algebraic equations. There are three main questions:

1. Is there any solution to an algebraic system of (2.31) type?
2. How can this solution actually be found?
3. Does the difference solution converge to exact one at zeroing of interval?

We can prove that, if an edge problem is of the type

$$y'' + p(x)y = f(x),$$

$$y(a) = \alpha,\ \ y(b) = \beta,$$

and $p(x) > 0$, the solution of the system (2.31) and (2.32) does exist and is to be the only one. Practically, searching for a solution may be carried out by, for example, sweep method. The next theorem responds the third question.

Theorem. *In case of $p(x)$ and $f(x)$ are continuously differentiated twice, the difference solution corresponding to the scheme with replacement*

$$y_i' \approx \frac{y_{i+1} - y_i}{h}, \quad y_i'' \approx \frac{y_{i+2} - 2 \cdot y_{i+1} + y_i}{h^2},$$

converges eventually to the precise one with inaccuracy of $O(h)$ for $h \to 0$.

Thus, the schemes (2.28), (2.29) give an approximate solution to the edge problem, but its precision is low. It is caused by the low order of precision of approximation by derivative $y' \approx \frac{y_{i+1} - y_i}{h}$; inaccuracy of this approximation

$$r_i(h) = \frac{h}{2} \cdot y''(\xi), \quad x_i < \xi < x_{i+1}.$$

More precise difference scheme may be obtained when for transfer from linear edge problem to finite difference equations we use the central formulae of derivatives:

$$y_i' \approx \frac{y_{i+1} - y_{i-1}}{2h}, \tag{2.42}$$

$$y_i'' \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}, \tag{2.43}$$

$i = 1, 2, \ldots, n$.

Inaccuracy of equation (2.42) is expressed as

$$r_i(h) = -\frac{h^2}{6} \cdot y'''(\xi), \quad x_{i-1} < \xi < x_{i+1},$$

it means that this formula has the second order of accuracy with respect to the mesh interval h. Substituting the equations (2.42), (2.43) for (2.24), (2.25), after some rearrangements we get the following system:

$$\begin{cases} y_{i+1} + m_i \cdot y_i + n_i \cdot y_{i-1} = \dfrac{2 \cdot h^2}{2 + h \cdot p_i} \cdot f_i, \quad i = 1, 2, \ldots, n. \\ \alpha_0 \cdot y_0 + \alpha_1 \cdot \dfrac{y_1 - y_0}{h} = A, \\ \beta_0 \cdot y_n + \beta_1 \cdot \dfrac{y_{n+1} - y_{n-1}}{2 \cdot h} = B, \end{cases} \tag{2.44}$$

where

$$m_i = \frac{2 \cdot q_i \cdot h^2 - 4}{2 + h \cdot p_i}, \quad n_i = \frac{2 - h \cdot p_i}{2 + h \cdot p_i}.$$

The system (2.44) is three-diagonal again and may also be solved by the sweep method. Here the algorithm looks like the next. First, the coefficients

$$\left.\begin{aligned} c_1 &= \frac{\alpha_1 - \alpha_0 \cdot h}{m_1 \cdot (\alpha_1 - \alpha_0 \cdot h) + n_1 \cdot \alpha_1}, \\ d_1 &= \frac{2 \cdot f_1 h^2}{2 + p_1 \cdot h} + n_1 \frac{A \cdot h}{\alpha_1 - \alpha_0 \cdot h} \end{aligned}\right\} \tag{2.45}$$

should be found. Then the next coefficients c_i, d_i are defined by the recurrent equations as

$$\left.\begin{aligned} c_i &= \frac{1}{m_i - n_i \cdot c_{i-1}}, \\ d_i &= \frac{2 f_i \cdot h^2}{2 + h \cdot p_i} - n_i \cdot c_{i-1} \cdot d_{i-1}, \\ i &= 2,3,...,n. \end{aligned}\right\} \tag{2.46}$$

The inverse sweep starts with finding y_n:

$$y_n = \frac{2 \cdot B \cdot h - \beta_1 \cdot (d_n - c_{n-1} \cdot d_{n-1})}{2 \cdot \beta_0 \cdot h + \beta_1 \cdot (c_{n-1} - \frac{1}{c_n})}. \tag{2.47}$$

Then we find $y_n,...,y_1,y_0$ by the equations

$$y_i = c_i \cdot (d_i - y_{i+1}), \qquad i = n-1,\ n-2,...,1, \tag{2.48}$$

$$y_0 = \frac{A \cdot h - \alpha_1 \cdot y_1}{\alpha_0 \cdot h - \alpha_1}. \tag{2.49}$$

Concerning the scheme (2.44), we can prove that it has the one solution at

$$\max_{a \le x \le b} \left| p(x) \right| < \frac{2}{h}$$

and

$$q(x) \ \le 0,$$
$$a \le x \le b$$

and this solution can be found by abovementioned sweep method. Besides, for scheme (2.44) the next theorem takes place.

Theorem. *Let the solution to the boundary problems (2.44), (2.45) be the single and continuously differentiated on [a, b] until the fourth order of accuracy inclusively. If the following conditions are fulfilled*

$$\max_{a \le x \le b} |p(x)| < \frac{2}{h}, \quad q(x) \le 0, \ \alpha_0\alpha_1 \le 0, \ \beta_0\beta_1 \ge 0,$$

the scheme (2.44) will uniformly converge to the solution to the problem (2.24), (2.25) with inaccuracy of $O(h^2)$.

Let's notice that the conditions in theorems are sufficient but unnecessary. Hence, the violation of these conditions in practical numerical calculations doesn't cause worsening of calculation schemes.

2.3. Semi-analytical methods of edge problem solving

2.3.1. Collocation method

It needs to find the function $y = y(x)$ satisfying the linear differential equation

$$L(y(x)) \equiv y'' + p(x)y' + q(x)y = f(x) \tag{2.50}$$

and linear edge conditions

$$\left.\begin{aligned} \Gamma_a \equiv \alpha_0 y(a) + \alpha_1 y'(a) = A \\ \Gamma_b \equiv \beta_0 y(b) + \beta_1 y'(b) = B \end{aligned}\right\}, \tag{2.51}$$

where $|\alpha_0| + |\alpha_1| \ne 0, \ \ |\beta_0| + |\beta_1| \ne 0.$

Let's choose a set of linearly independent functions

$$U_0(x), U_1(x), ..., U_n(x), \tag{2.52}$$

and call this set the system of basic functions.

Let function $U_0(x)$ satisfy inhomogeneous edge conditions:

$$\Gamma_a(U_0) = A, \ \ \Gamma_b(U_0) = B; \tag{2.53}$$

and others functions satisfy corresponding homogeneous edge conditions:

$$\Gamma_a(U_i) = 0, \ \ \Gamma_b(U_i) = 0, \quad i = 1, \ 2, ..., \ n. \tag{2.54}$$

If the edge conditions (2.51) are homogeneous ($A = B = 0$), we can put $U_0(x) = 0$ and consider only system of functions $U_i(x), \ \ i = 1, 2, ..., n$.

Let's find an approximate solution to the edge problem (2.50), (2.51) as a linear combination of basic functions

$$y = U_0(x) + \sum_{i=1}^{n} c_i U_i(x). \tag{2.55}$$

In this case the function y satisfies the edge conditions (2.51). Indeed, because of linearity of edge conditions, we have

$$\Gamma_a(y) = \Gamma_a(U_0) + \sum_{i=1}^{n} c_i \Gamma_a(U_i) = A + \sum_{i=1}^{n} c_i \cdot 0 = A,$$

and also

$$\Gamma_b(y) = B.$$

Let's compose the function $R = L(y) - f(x)$. Substituting here Eq. (2.55) for y, we get

$$R(x,\ c_1,...,\ c_n) \equiv L(y) - f(x) = L(U_0) - f(x) + \sum_{i=1}^{n} c_i L(U_i). \qquad (2.56)$$

If for a set of coefficients c_i the equality

$$R(x, c_1, ..., c_n) \equiv 0 \ \text{ for } \ a \le x \le b$$

is satisfied, the function y is a precise solution to the edge problems (2.50), (2.51). However, generally it isn't easy to choose the proper functions U_i and coefficients c_i. That's why usually it is required that function $R(x, c_1, ..., c_n)$ turned into zero at the given system of points $x_1, x_2, ..., x_n$ for the interval [a, b]. These points are called the collocation points. The function R itself is called the **misclosure** of the equation (2.50). Evidently, at collocation points the differential equation (2.50) will be satisfied completely, and the misclosure here equals zero.

Thus, the collocation method brings us to the system of linear equations

$$\left.\begin{aligned} &R(x_1, c_1, ..., c_n) = 0; \\ &... \\ &R(x_n, c_1, ..., c_n) = 0. \end{aligned}\right\} \qquad (2.57)$$

From the system (2.57), in case of it's compatibility, we can get coefficients $c_1, ..., c_n$. After that, the equation (2.55) gives us an approximate solution to the edge problem.

Example. Solve the edge problem by collocation and mesh methods:

$$\left.\begin{aligned} &y'' + (1 + x^2) y + 1 = 0, \\ &y(-1) = 0, \ \ y(1) = 0. \end{aligned}\right\} \qquad (2.58)$$

1. Collocation method

Let's choose polynomials $U_n(x) = x^{2n-2}(1 - x^2), \ \ n = 1, \ 2, \ ...$
as basic functions. These polynomials satisfy the edge conditions $U_n(\pm 1) = 0$.

We take $x_{-1} = -\frac{1}{2}, \ \ x_0 = 0, \ \text{and} \ \ x_1 = +\frac{1}{2}$ as collocation points. Limiting our set with only two basic functions, we get

$$y = c_1(1 - x^2) + c_2(x^2 - x^4).$$

Let's find function $R \equiv L(y) - f(x)$:

$$R(x) = -2c_1 + c_2(2-12x^2) + (1+x^2)[c_1(1-x^2) + c_2(x^2 - x^4)] + 1 = \\ = 1 - c_1(1+x^4) + c_2(2-11x^2 - x^6). \tag{2.59}$$

At collocation points $x_0 = 0, \quad x_{\pm 1} = \pm\frac{1}{2}$ we get

$$R(x_0) = 0, \quad R(x_{\pm 1}) = 0.$$

Substituting here the formula (2.59), we find

$$\left.\begin{aligned} 1 - c_1 + 2c_2 &= 0, \\ 1 - \frac{17}{16}c_1 - \frac{49}{64}c_2 &= 0. \end{aligned}\right\} \tag{2.60}$$

Solving this system, we obtain the coefficients c_1 and c_2

$$c_1 = 0.957, \; c_2 = -0.022.$$

Consequently, an approximate solution is

$$y \approx 0.957(1-x^2) - 0.022\ (x^2 - x^4).$$

For example, for $x = 0$ we get $y(0) = 0.957$.

2. Mesh method

For rough solution, let's choose interval $h = 1/2$ (Fig. 2).

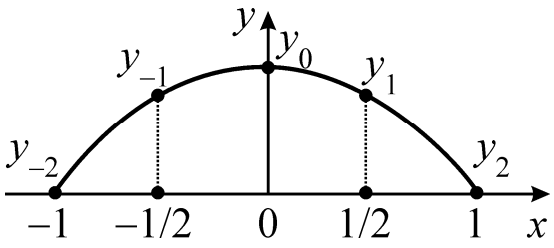


Fig. 2. Illustration to the mesh method

Supposing $x_{-2} = -1, \; x_{-1} = -1/2, \; x_0 = 0, \; x_1 = 1/2, \; x_2 = 1$, because of the equation symmetry and the edge conditions, we have

$$y_{-2} = y_2 = 0; \; y_{-1} = y_1. \tag{2.61}$$

Thus, it needs to find only two ordinates y_0 and y_1. Assuming $x = 0$ and using symmetrical equations for derivatives

$$y' = \frac{y_{i+1} - y_{i-1}}{2h}, \quad y'' = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2},$$

we get

$$\frac{y_{-1} - 2y_0 + y_1}{\frac{1}{4}} + y_0 = -1.$$

Similarly, for $x = 1/2$ (i. e. for $i = 1$) we get

$$\frac{y_0 - 2y_1 + y_2}{\frac{1}{4}} + (1 + \frac{1}{4})\, y_1 = -1.$$

Taking into account the equation (2.61), we find the system

$$\left.\begin{aligned} -7y_0 + 8y_1 &= -1, \\ 4y_0 - 6\frac{3}{4}y_1 &= -1. \end{aligned}\right\}$$

Solving the system, we get $y_0 = 0.967$ and $y_1 = 0.721$. Let's compare: collocation method gives $y_0 = 0.957$, mesh method gives $y_0 = 0.967$.

2.3.2. Galerkin's method

Let a differential equation be given with *linear* edge conditions:

$$L(y(x)) = f(x), \tag{2.62}$$

$$\left.\begin{aligned} \Gamma_a(y) \equiv \alpha_0 y(a) + \alpha_1 y'(a) = A, \\ \Gamma_b(y) \equiv \beta_0 y(b) + \beta_1 y'(b) = B. \end{aligned}\right\} \tag{2.63}$$

We find approximate solution to this edge problem as the summation:

$$y_n(x) = \varphi_0(x) + \sum_{k=1}^{n} a_k \varphi_k(x), \tag{2.64}$$

where $\varphi_0(x)$ is some continuous function satisfying *inhomogeneous* edge condition (2.63), and $\varphi_k(x)$, $k = 1,\ 2,\ldots\ (1 \le k < \infty)$ is a system of linearly independent functions satisfying *homogeneous* edge conditions

$$\Gamma_a(\varphi_k) = 0, \quad \Gamma_b(\varphi_k) = 0, \tag{2.65}$$

and besides, the functions $\varphi_k(x)$ for $1 \le k < \infty$ form a complete system in the class of functions $c_2[a, b]$ satisfying conditions (2.65).

Let's note that the completeness is understood in the following way.

Through G we denote the class of functions $y(x)$, belonging to $c_2[a, b]$ (being twice continuously differentiated on $[a, b]$) and satisfying the edge condition (2.65). It is said that the system of functions $\{\varphi_k(x)\}$ is complete in G-class, if for any $\varepsilon > 0$ and any function $y(x) \in G$ it is possible to point out such n and such parameters $a_1, a_2, \ldots, a_n$ that inequality

$$\left| y^{(i)}(x) - g_n^{(i)}(x) \right| < \varepsilon, \quad i = 0,\ 1,\ 2, \quad a \le x \le b,$$

where $g_n = \sum_{k=1}^{n} a_k \varphi_k(x)$, takes place.

It means that for any function $y(x) \in G$ there is always such function $g_n(x)$, that approximates the function $y(x)$ and its derivatives $y'(x)$ and $y''(x)$ on [a, b] to any degree of accuracy.

Let's prove that in case for some function $F(x)$ and complete system of functions $\varphi_k(x)$, the relation of orthogonality

$$\int_a^b F(x)\varphi_k(x)dx = 0 \quad \text{for} \quad 1 \le k \le \infty \tag{2.66}$$

is fulfilled, this function $F(x) \equiv 0$ on [a, b]. For proving we create the complete orthogonal system $\psi_k(x)$ from complete system $\varphi_k(x)$ using sequential orthogonalization

$$\varphi_k(x) = \sum_{m=1}^{k} c_{km}\psi_m(x),$$

where $c_{kk} \neq 0$, otherwise, $\varphi_k(x)$ could be linearly dependent. Expanding function $F(x)$ by new system, we have:

$$F(x) = \sum_{l=1}^{\infty} d_l\psi_l(x).$$

Substituting this expansion into the relation of orthogonality (2.66), we get the equality

$$0 = \int_a^b F(x)\varphi_k(x)dx = \int_a^b \left(\sum_{l=1}^{\infty} d_l\psi_l\right)\left(\sum_{m=1}^{k} c_{km}\psi_m\right)dx, \; k = 1, \; 2, \; ... \tag{2.67}$$

Let's calculate the last integral:

$$\int_a^b \sum_{l=1}^{\infty} d_l\psi_l \cdot \sum_{m=1}^{k} c_{km}\psi_m \cdot dx =$$

$$\int_a^b (d_1\psi_1 + d_2\psi_2 + ... + d_l\psi_l + ...)(c_{k1}\psi_1 + c_{k2}\psi_2 + ... + c_{kk}\psi_k)dx =$$

$$= \sum_{m=1}^{k} d_m c_{km}, \text{ as } \int_a^b \psi_m(x)\psi_l(x) = \begin{cases} 0, & m \neq l, \\ 1, & m = l. \end{cases}$$

Thus, the equation (2.67) becomes

$$\sum_{m=1}^{k} d_m c_{km} = 0, \quad k = 1, \; 2, ...$$

Taking $k = 1$, we get $d_1 c_{11} = 0$. So as $c_{11} \neq 0$, we have $d_1 = 0$. Taking $k = 2$, we get $d_2 = 0$ and so on, and so forth. Consequently, all coefficients d_l

equal zero in $F(x)$ expansion and hence $F(x)$ itself identically equals zero, which was to be proved.

Going back to the problems (2.62), (2.63) we see that did it happen to find function $y(x)$ satisfying condition (2.63) and $L(y(x)) - f(x)$ be orthogonal to $\varphi_k(x)$ for any $k \geq 1$, it would give $L(y(x)) = f(x)$, and the problem (2.62), (2.63) would be solved. If the orthoganality takes place only for $k \leq n$, it means that expansion of $L(y(x)) - f(x)$, according to the system $\varphi_k(x)$, contains d_{n+1} and higher coefficients, i. e. $L(y(x)) \approx f(x)$.

Galerkin's method provides the solving to the problems (2.62), (2.63) in the form of the formula (2.64), where it is demanded the orthoganality of $L(y(x)) - f(x)$ to functions of the complete system $\varphi_k(x)$ for $k = 1, 2, \ldots, n$, i. e.

$$\int_a^b \left[L(y_n(x)) - f(x) \right] \varphi_k(x) dx = 0, \quad 1 \leq k \leq n, \tag{2.68}$$

where

$$y_n(x) = \varphi_0(x) + \sum_{k=1}^{n} a_k \varphi_k(x).$$

It gives the algebraic system of equations for finding the coefficient a_k. Having found the coefficients from it, we get an approximate solution.

If the operator $L(U)$ is nonlinear, the system (2.68) is also nonlinear, and the solution is difficult to be found. If the operator $L(U)$ is linear, the system is also linear, and the problem can be solved for plenty of coefficients.

In Galerkin's method, function $\varphi_0(x)$ must satisfy the edge condition (2.63). That's why the function $\varphi_0(x)$ is of a kind $\varphi_0(x) = \alpha + \beta \cdot x$ may be chosen and the coefficients α, β are to be found as solution of the system

$$\left.\begin{aligned} \alpha_0(\alpha + \beta \cdot a) + \alpha_1 \beta = A, \\ \beta_0(\alpha + \beta \cdot b) + \beta_1 \beta = B. \end{aligned}\right\}$$

Analogously the functions $\varphi_k(x)$ are to be found. Let's, for instance, choose a complete system $\varphi_k(x)$ as consequent orders polynomials:

$$\varphi_k(x) = \sum_{i=0}^{k+1} c_{ik} \cdot x^i, \quad k = 1, \ldots, n.$$

Coefficients c_{ik} are to be found from *homogeneous* edge conditions (2.65)

$$\left.\begin{aligned} \alpha_0 \varphi_k(a) + \alpha_1 \varphi_k'(a) = 0, \\ \beta_0 \varphi_k(b) + \beta_1 \varphi_k'(b) = 0, \end{aligned}\right\} \tag{2.65^a}$$

for all values $k = 1,2, \ldots, n$.

So, $\varphi_1(x) = c_{01} + c_{02} \cdot x + c_{03} \cdot x^2$ for $k = 1$ and condition (2.65[a]) become:

$$\alpha_0(c_{01} + c_{02}a + c_{03}a^2) + \alpha_1(c_{02} + 2c_{03}a) = 0,$$

$$\beta_0(c_{01} + c_{02}b + c_{03}b^2) + \beta_1(c_{02} + 2c_{03}b) = 0.$$

In this system of two equations there are three unknown variables: c_{01}, c_{02}, c_{03}. One of them may be chosen freely, putting, for example, $c_{01} = 1$. Similarly, other coefficients c_{0k} may be found for $k = 2, \ldots, n$.

For simple conditions like $y(a) = A, \quad y(b) = B$ meaning $\alpha_0 = \beta_0 = 1$ and $\alpha_1 = \beta_1 = 0$ the functions $\varphi_k(x)$ are found as

$$\varphi_k(x) = (x-a)^k(x-b), \quad k = 1,\ 2,\ldots,\ n$$

or

$$\varphi_k(x) = (x-a)(x-b)^k, \quad k = 1,\ 2,\ldots,\ n.$$

Let's note that for a nonlinear edge condition such as $y'(a) = g(U(a))$, the linear combination (2.64) with arbitrary coefficients a_k won't satisfy this edge condition. Thus, Galerkin's method may be used only for linear edge conditions, but a nonlinear operator L is permitted there.

Example 1. Using Galerkin's method, find an approximate solution to the equation

$$y'' + xy' + y = 2x$$

with conditions

$$y(0) = 1, \quad y(1) = 0.$$

As a system of basic functions $U_k(x)$ we choose

$$\varphi_0(x) = 1 - x,$$

$$\varphi_k(x) = x^k(1-x), \quad k = 1,\ 2,\ldots \ .$$

Let's limit the system with only four functions φ_k, i. e. $k = 0, 1, 2, 3$.

The solution is to be found as

$$y = (1-x) + a_1x(1-x) + a_2x^2(1-x) + a_3x^3(1-x).$$

Let's find function $F(x)$. As $F(x) = L(y(x)) - f(x)$ and $L(y(x)) = y''(x) + xy' + y$, $f(x) = 2x$, we get

$$F(x) = 1 - 4x + a_1(-2 + 2x - 3x^2) + a_2(2 - 6x + 3x^2 - 4x^3) +$$

$$+a_3(6x - 12x^2 + 4x^3 - 5x^4).$$

Let's introduce orthoganality of $F(x)$ to $\varphi_k(x),\ k=1,\ 2,\ 3$. It gives the system

$$\left.\begin{aligned}
&\int_0^1 (x-x^2)F(x)dx=0,\\
&\int_0^1 (x^2-x^3)F(x)dx=0,\\
&\int_0^1 (x^3-x^4)F(x)dx=0.
\end{aligned}\right\}$$

Replacing $F(x)$ by the expression for this function and integrating, we get

$$\left.\begin{aligned}
&133a_1+63a_2+36a_3=-70,\\
&140a_1+108a_2+79a_3=-98,\\
&264a_1+252a_2+211a_3=-210.
\end{aligned}\right\}$$

The solution to this system are as follows:

$$a_1=-0.2090,\ \ a_2=-0.7894,\ \ a_3=0.2090.$$

Consequently, $y\approx(1-x)(1-0.2090\,x-0.7894\,x^2+0.2090\,x^3)$.

Example 2. Let's solve the problem $y''+y=-x,\ \ y(0)=y\left(\dfrac{\pi}{2}\right)=0.$

Let $\varphi_0(x)=0$ and choose the complete system of functions

$$\varphi_k(x)=x^k\left(\frac{\pi}{2}-x\right),\quad 1\le k<\infty.$$

Taking only $k=1$, we get

$$a_1\approx\frac{5\pi}{40-\pi^2}\approx 0.521.$$

If we take two members ($k=1, 2$), we get $a_1\approx 0.815,\ \ a_2\approx 0.377.$
The following table may be calculated for this problem:

x	$y_1(x)$	$y_2(x)$	precise solution $y(x)$
$\pi/8$	0.241	0.445	0.208
$\pi/4$	0.322	0.685	0.325
$3\pi/8$	0.241	0.582	0.273

3. NUMERICAL SOLUTION OF PARTIAL DERIVATIVE EQUATIONS

3.1. Difference schemes. Fundamental issues

Let independent variables x, y change at some area $\mathcal{D}$, limited by ***Γ-contour***. It is said that the second-order equation for function $u(x,y)$ is given in $\mathcal{D}$-area, if for any $\mathcal{D}$-area point the ratio takes place:

$$\begin{aligned} L(u) = a(x,y)\frac{\partial^2 u}{\partial x^2} + 2b(x,y)\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\frac{\partial^2 u}{\partial y^2} + \\ +2d(x,y)\frac{\partial u}{\partial x} + 2e(x,y)\frac{\partial u}{\partial y} + g(x,y)u = f(x,y), \end{aligned} \tag{3.1}$$

where $a(x,y)$, $b(x,y)$, … are the coefficients; $f(x, y)$ is an equation free member. These functions are known and considered as defined in closed area $\overline{\mathcal{D}} = \mathcal{D} + \Gamma$. Let's designate $\delta(x,y) = b^2 - ac$. An equation $L(u) = f$ is called **elliptical**, **parabolic** or **hyperbolic** in $\mathcal{D}$ if the conditions $\delta(x,y) < 0$, $\delta(x,y) = 0$, or $\delta(x,y) > 0$ are fulfilled respectively for all $(x,y) \in \mathcal{D}$.

Depending on the type of differential equation, boundary and initial conditions for this equation are given in different ways. Further we consider particular cases of Eq. (3.1):

- *the Poisson's equation (elliptical)*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f\left(x,y\right);$$

- *the equation of thermal conductivity (parabolic)*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t);$$

- *the wave equation (hyperbolic)*

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x,y).$$

3.1.1. Convergence, approximation and stability of difference schemes

Let u be the solution for differential equation

$$L(u) = f, \tag{3.2}$$

given in $\mathcal{D}$-area. Let's consider a set $\mathcal{D}_h = \{M_h\}$ containing isolated points M_h, belonging to the closed area $\overline{\mathcal{D}} = \mathcal{D} + \Gamma$. The quantity of points in $\mathcal{D}_h$ is to be characterized by the value h. The smaller is h, the more points are in $\mathcal{D}_h$. The set $\mathcal{D}_h$ is called the *mesh* and points $M_h \in \mathcal{D}_h$ are mesh *nodes*. Function defined for nodes is called the *mesh function*.

Let's denote as U the space of functions $u(x, y)$ continuous in $\mathcal{D}$. Let U_h is the space formed by the set of mesh functions $u_h(x,y)$ defined on $\mathcal{D}_h$. In the mesh method, the replacement U-space by U_h-space takes place.

Let $u(x,y)$ be the precise solution of equation (3.2). Then $u(x,y)$ belongs to U. The problem is to find the values of $u_h(x,y)$. Totally, these values form a table, in which the number of values is equal to number of point in $\mathcal{D}_h$. The problem formulated may be rarely solved precisely. Usually it is possible to calculate such mesh values $u^{(h)}$ that we may suppose: $u^{(h)} \approx u_h(x,y)$.

The values $u^{(h)}$ are called *approximate mesh values* of the solution $u(x,y)$. To calculate them, the system of numerical equations to be constructed, which we write as

$$L_h(u^{(h)}) = f^h, \tag{3.3}$$

where L_h is a linear operator, corresponding to operator L; $f^{(h)} \in F_h$. In case $f(x,y) \in F$, F_h is arranged from F in the same way as U^h is formed from U. Eq. (3.3) is called the **difference scheme**.

Let in linear spaces U_h and F_h the norms $\|\bullet\|_{U_h}$ and $\|\bullet\|_{F_h}$ respectively are introduced, which are mesh analogues to $\|\bullet\|_U$ and $\|\bullet\|_F$ norms in initial spaces.

The difference scheme (3.3) said to be convergent, if for $h \to 0$ the following condition is satisfied

$$\left\| u_h(x,y) - u^{(h)} \right\|_{U_h} \to 0.$$

If the condition

$$\left\| u_h(x,y) - u^{(h)} \right\|_{U_h} \le ch^s$$

is satisfied, where c is a constant independent on h and $s > 0$, it is said there is a convergence of s-order degree relatively to interval h.

It is considered that difference scheme (3.3) *approximates* the problem (3.2) to the solution $u(x, y)$, if

$$L_h(u_h(x,y)) = f^{(h)} + \delta f^{(h)} \quad \text{and} \quad \left\| \delta f^{(h)} \right\|_{F_h} \to 0 \quad \text{for} \quad h \to 0.$$

The value $\delta f^{(h)}$ is called *approximation inaccuracy* or *misclosure* of a difference scheme. If $\left\| \delta f^{(h)} \right\|_{F_h} \le Mh^{\sigma}$, where M is a constant independent on h and $\sigma > 0$, it is said that difference scheme (3.3) approximates the problem (3.2) to the solution $u(x, y)$ with inaccuracy of order σ relatively to the interval h.

The difference scheme (3.3) is called *stable*, if such $h_0 > 0$ exists that for all $h < h_0$ and for any $f^{(h)} \in F_h$, the following conditions are satisfied:

1) difference scheme (3.3) has the single solution;
2) $\left\| u^{(h)} \right\|_{U_h} \le M \left\| f^{(h)} \right\|_{F_h}$, where M is a constant independent on h and $f^{(h)}$.

In other words, the difference scheme is stable in case its solution depends continuously on input data. The stability characterizes the scheme sensitivity towards any inaccuracy; it is the intrinsic feature of difference scheme that does not depend on the initial problem, in contrary to convergence and approximation. There is a relation between convergence, approximation and stability, which sense is that convergence follows from approximation and stability as the next theorem states.

Theorem. *Let difference scheme* $L_h(u^{(h)}) = f^{(h)}$ *approximates the problem* $L(u) = f$ *to the solution* $u(x, y)$ *with order s relatively to h and it is stable. Then this scheme converges and the order of its convergence coincides with approximation order, i. e. the evaluation*

$$\left\| u_h(x,y) - u^{(h)} \right\|_{U_h} \le kh^{s}, \tag{3.4}$$

is correct, where k is a constant independent on h.

Prove. According to definition of approximation, we have

$$\left\| df^{(h)} \right\|_{F_h} \le ch^{s}, \text{ where } \delta f^{(h)} = L_h(u_h(x,y)) - f^{(h)}.$$

Let's denote $\varepsilon_h(x,y) = u_h(x,y) - u^{(h)}$. Because of linear character of L_h we have for $\varepsilon_h(x,y)$ the formula

$$L_h(\varepsilon_h(x,y)) = \delta f^{(h)}.$$

From the definition of stability we have

$$\left\|\varepsilon_h(x,y)\right\|_{U_h} \le M\left\|\delta \cdot f^{(h)}\right\|_{F_h} \le M(Ch^s) = Kh^s,$$

where $K = MC$. So the evaluation (3.4) is stated and the theorem is proved.

Usually the mesh method application is as follows:

1. At the beginning, a rule of choosing the mesh is stated, i. e. the way of replacing area $\mathcal{D}$ and contouring Γ by some mesh area. More often the rectangular and uniform mesh is used.
2. Then one or more difference schemes are stated and constructed. The condition of approximation is checked, and its order is estimated.
3. The stability of constructed difference schemes is proved. It is of the most important and complicated issues. If the difference scheme has approximation and stability, the convergence is assessed, according to the proved theorem.
4. Numerical solution of difference schemes is considered.

If the linear difference schemes it is a system of linear algebraic equations. The order of such systems might be huge.

3.2. Difference schemes for parabolic equations

3.2.1. The solution of Cauchy problem

Let's view the Cauchy problem for thermal conductivity

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \varphi(x,t),\ -\infty < x < +\infty,\ t > 0, \tag{3.5}$$

with a condition at the line $t = 0$

$$u(x,0) = \psi(x),\ -\infty < x < +\infty. \tag{3.6}$$

We need to find the function $u(x, t)$, which under $t > 0$ and $-\infty < x < +\infty$ would satisfy the equation (3.5), meanwhile, under $t = 0$ the condition (3.6).

Let's consider the problem (3.5) and (3.6), it has the only solution $u(x, t)$ in the upper halve plane. Let's also this solution and its derivatives are continuous:

$$\frac{\partial^{(i)} u}{\partial t^i},\ i=1, 2 \text{ and } \frac{\partial^{(k)} u}{\partial x^k},\ k=1, 2, 3, 4.$$

Write the problem (3.5), (3.6) as $L(u) = f$. For this, it is sufficient to put

$$L(u) = \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2}, & -\infty < x < +\infty,\ t > 0, \\ u(x,0), & -\infty < x < +\infty,\ t = 0, \end{cases}$$

$$f = \begin{cases} \varphi(x,t), & -\infty < x < +\infty, \ t > 0, \\ \psi(x), & -\infty < x < +\infty, \ t = 0. \end{cases}$$

Further we take t changing in $0 \le t \le T < \infty$. In the case considered

$$\mathcal{D} = \{-\infty < x < +\infty, \ 0 < t < T\},$$

Γ is the union of lines $t = 0$ and $t = T$.

Let's choose a triangular mesh and replace the area $\overline{\mathcal{D}} = \mathcal{D} + \Gamma$ by $\mathcal{D}_h$ mesh area. The set of nodes (x_m, t_n) is referred to the $\mathcal{D}_h$ area, where

$$x_m = mh, \ m = 0, \pm 1, \pm 2, \ldots, \ h > 0,$$

$$t_n = n\tau, \ n = 0, 1, \ldots, N, \ \tau > 0, \ N\tau \le T < (N+1)\tau.$$

Let's substitute the problem $L(u) = f$ with the difference scheme $L_h(u^{(h)}) = f^{(h)}$. Then we designate the precise solution of $L(u) = f$ at the node (x_m, t_n) as $u(x_m, t_n)$ and through u_m^n we mark the corresponding approximate solution. Thus, we have

$$L(u)\big|_{(x_m,\, t_n)} \equiv \begin{cases} \left.\dfrac{\partial u}{\partial t}\right|_{(x_m,\, t_n)} - \left.\dfrac{\partial^2 u}{\partial x^2}\right|_{(x_m,\, t_n)}, \\ m = 0, \ \pm 1, \ \pm 2, \ldots, \ n = 1, \ 2, \ldots, \ N, \\ u(x, \ 0)\big|_{(x_m,\, t_n)}; \end{cases}$$

$$f\big|_{(x_m,\, t_n)} \equiv \begin{cases} \varphi(x, \ t)\big|_{(x_m,\, t_n)}, \\ m = 0, \ \pm 1, \ldots, \ n = 1, \ 2, \ldots, \ N, \\ \psi(x)\big|_{(x_m,\, t_n)}. \end{cases}$$

So that to replace the expressions $\left.\dfrac{\partial u}{\partial t}\right|_{(x_m,\, t_n)}$ and $\left.\dfrac{\partial^2 u}{\partial x^2}\right|_{(x_m,\, t_n)}$, we use the formulae of numerical differentiating as follows:

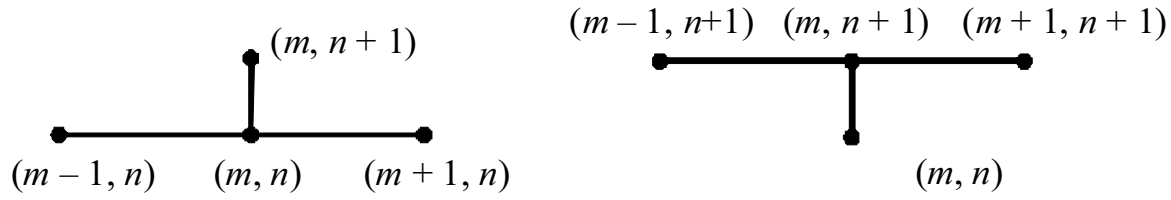
$$\left.\frac{\partial u}{\partial t}\right|_{(x_m,\, t_n)} = \frac{u(x_m, t_{n+1}) - u(x_m, t_n)}{\tau} - \frac{\tau}{2} \cdot \left.\frac{\partial^2 u}{\partial t^2}\right|_{(x_m,\, t_n^{(1)})}, \qquad (3.7)$$

$$\left.\frac{\partial u}{\partial t}\right|_{(x_m,t_n)} = \frac{u(x_m,t_n)-u(x_m,t_{n-1})}{\tau} + \frac{\tau}{2}\cdot\left.\frac{\partial^2 u}{\partial t^2}\right|_{(x_m,t_n^{(2)})}, \quad (3.8)$$

$$\left.\frac{\partial u}{\partial t}\right|_{(x_m,t_n)} = \frac{u(x_m,t_{n+1})-u(x_m,t_{n-1})}{2\tau} - \frac{\tau^2}{6}\cdot\left.\frac{\partial^3 u}{\partial t^3}\right|_{(x_m,t_n^{(3)})}, \quad (3.9)$$

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_m,t_n)} = \frac{u(x_{m+1},t_n)-2u(x_m,t_n)+u(x_{m-1},t_n)}{h^2} - \frac{h^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{(x_m^{(1)},t_n)}. \quad (3.10)$$

Let's some set of nodes, taken for substitution of the problem $L(u)=f$ at the node (x_m,t_n), of difference scheme $L_h(u^{(h)})=f^{(h)}$, be called a **pattern**. More often the following patterns are used (Fig. 3).



a) explicit two-layer pattern; *b) implicit two-layer pattern*

Fig. 3. Explicit and implicit patterns

Let's discuss the explicit two-layer pattern (Fig. 3*a*)

$$L(u)\big|_{(x_m,t_n)} \equiv$$

$$\equiv \begin{cases} \dfrac{u(x_m,t_{n+1})-u(x_m,t_n)}{\tau} - \dfrac{u(x_{m+1},t_n)-2u(x_m,t_n)+u(x_{m-1},t_n)}{h^2} + r_{mn}^{(1)}, \\ u(x_m,0)+0. \end{cases} \quad (3.11)$$

Here we used formulae (3.7), (3.10) and designated

$$r_{mn}^{(1)} = -\frac{\tau}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{(x_m,t_n^{(1)})} - \frac{h^2}{12}\cdot\left.\frac{\partial^4 u}{\partial x^4}\right|_{(x_m^{(1)},t_n)}.$$

Let's input the value of

$$f^{(h)} \equiv \begin{cases} \varphi(x_m,t_n), \\ \psi(x_m). \end{cases} \quad (3.12)$$

Based on formulae (3.11), (3.12), the difference scheme might be written for the problem $L(u)=f$:

$$L_h^{(1)}(u^{(h)}) = f^{(h)}, \tag{3.13}$$

where a difference operator $L_h^{(1)}$ is defined by the rule

$$L_h^{(1)}(u^{(h)}) \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \\ u_m^0, \\ m = 0, \ \pm 1, ..., \\ n = 0, \ 1, \ 2, ..., \ N-1. \end{cases}$$

Similarly, if we take the implicit two-layer pattern (Fig. 3*b*), the following difference scheme may be obtained:

$$L_h^{(2)}(u^{(h)}) = f^{(h)}, \tag{3.14}$$

where

$$L_h^{(2)}(u^{(h)}) \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2}, \\ u_m^0, \\ m = 0, \ \pm 1, ..., \\ n = 0, \ \ 1, \ \ 2, ..., \ N-1; \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \varphi(x_m, t_n), \\ \psi(x_m). \end{cases}$$

On the basis of formulae (3.11), (3.13) we have

$$L_h^{(1)}(u_h(x,y)) = f^{(h)} + \delta^{(1)} f^{(h)},$$

where

$$\delta^{(1)} f^{(h)} = \begin{cases} r_{mn}^{(1)}(h), \\ 0. \end{cases}$$

The same rule for the Eqs. (3.11), (3.10), (3.14)

$$L_h^{(2)}(u_h(x,y)) = f^{(h)} + \delta^{(2)} f^{(h)},$$

$$\delta^{(2)} f^{(h)} \equiv \begin{cases} r_{mn}^{(2)}(h) = \dfrac{\tau}{2} \dfrac{\partial^2 u}{\partial t^2}\Bigg|_{(x_m, t_n^{(2)})} - \dfrac{h^2}{12} \cdot \dfrac{\partial^4 u}{\partial x^4}\Bigg|_{(x_m^{(1)}, t_{n+1})}. \\ 0 \end{cases}$$

Let's define the approximation order of difference schemes (3.13), (3.14). As F_h, we take a linear set of all limited function pairs:

$$g^{(h)} = \begin{Bmatrix} \alpha_m^n \\ \beta_m \end{Bmatrix}.$$

The norm in F_h is to be defined by the rule

$$\left\| g^{(h)} \right\| = \max_{m,n} \left| \alpha_m^n \right| + \max_m \left| \beta_m \right|.$$

Let $\tau = rh^s$, where r and s are some positive numbers.

Supposing that the following bounds should be right for $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^4 u}{\partial x^4}$,

$$\max_{(x,t)\in D} \left| \frac{\partial^2 u}{\partial t^2} \right| \le M_2, \quad \max_{(x,t)\in D} \left| \frac{\partial^4 u}{\partial t^4} \right| \le M_4.$$

Then it is easy to obtain

$$\left\| \delta^{(1)} f^{(h)} \right\|_{F_h} = \max_{m,n} \left| r_{mn}^{(1)}(h) \right| \le \left(\frac{r}{2} M_2 + \frac{h^{2-S}}{12} M_4 \right) \cdot h^S, \qquad (3.15)$$

$$\left\| \delta^{(2)} f^{(h)} \right\|_{F_h} = \max_{m,n} \left| r_{mn}^{(2)}(h) \right| \le \left(\frac{r}{2} M_2 + \frac{h^{2-S}}{12} M_4 \right) \cdot h^S. \qquad (3.16)$$

For parabolic equations, as in case of scheme (3.15) can be taken $S = 2$, but as for scheme (3.14) we can take $S = 1$.

From formulae (3.15), (3.16) it becomes known that the difference schemes (3.13), (3.14) approximate the problem $L(u) = f$ with inaccuracy being relevant to h.

The difference scheme (3.13) allows to calculate the values u_m^0, $m = 0, \pm 1, \ldots$ on the first level by means of solution on the zero level, in other words, by the values u_m^1, $m = 0, \pm 1, \ldots$. It is sufficient to put $n = 0$ into the formula (3.13) and make calculations of recursive character. Then using the values u_m^1, it is possible to calculate u_m^2 the same way at $n = 1$. That's why this difference (3.13) is called explicit.

Difference scheme (3.14) has no such feature. Truly, if we put $n = 0$ into (3.14) in left part it will be a linear combination with values u_{m-1}^1, u_m^1, u_{m+1}^1, u_m^0; there will be values of known function $\varphi(x_m, 0)$ and $\psi(x_m)$ in the right part. To calculate the values on the first layer –

..., u^1_{-2}, u^1_{-1}, u^1_0, u^1_1, u^1_2,... – a linear equation discontinuous system is to be solved at first. That's why the scheme (3.14) is called implicit.

3.2.2. Stability of two-layer difference schemes

Let's define the norm in space u_h by the rule

$$\left\|u^{(h)}\right\|_{u_h} = \max_{m,n}\left|u^n_m\right|.$$

Let's view an explicit difference scheme (3.13). Exactly under which values r, $\tau = rh^2$ the stability of this scheme is possible.

To prove the stability, we are to outline the difference scheme, which will be solved in a single way at any

$$g^{(h)} = \begin{cases}\alpha^n_m \\ \beta_m\end{cases},\ g^{(h)} \in F_h$$

and the bound is $\left\|z^{(h)}\right\|_{U_h} \le M\left\|g^{(h)}\right\|_{F_h}$,

where M is consonant independent on h and $g^{(h)}$, and $L^{(1)}_h(z^{(h)}) = g^{(h)}$.

The difference scheme (3.13) is explicit and its single solution is evident.

Let's rewrite the formula $L^{(1)}_h(z^{(h)}) = g^{(h)}$ as

$$z^{n+1}_m = r(z^n_{m+1} + z^n_{m-1}) + (1-2r)z^n_m + \tau\alpha^n_m,\ z^0_m = \beta_m, \qquad (3.17)$$
$$m = 0,\ \pm 1,\ \pm 2,...,\ n = 0,\ 1,\ 2,...,\ N-1.$$

Let the condition be fulfilled:

$$1 - 2r \ge 0 \text{ or } r = \frac{\tau}{h^2} \le \frac{1}{2}. \qquad (3.18)$$

Then from the formula (3.17) we get

$$\max_m\left|z^{n+1}_m\right| \le r(\max_m\left|z^n_m\right| + \max_m\left|z^n_m\right|) + (1-2r)\max_m\left|z^n_m\right| + \tau\max_m\left|\alpha^n_m\right|,$$

or

$$\max_m\left|z^{n+1}_m\right| \le r(\max_m\left|z^n_m\right| + \tau\max_m\left|\alpha^n_m\right|. \qquad (3.19)$$

The inequality (3.19) means that at $\alpha^n_m \equiv 0$, $\max_m\left|z^{n+1}_m\right|$ doesn't exceed $\max_m\left|z^n_m\right|$, so that $\max_m\left|z^n_m\right|$ doesn't increase with n.

This property of homogeneous difference scheme is called as principle of maximum. Let's put $n = 0,\ 1,...,\ N-1$ into the formula (3.19). It gives

$$\max_m \left|z_m^1\right| \le \max_m \left|z_m^0\right| + \tau \max_{m,n} \left|\alpha_m^n\right|,$$
$$\max_m \left|z_m^2\right| \le \max_m \left|z_m^1\right| + \tau \max_{m,n} \left|\alpha_m^n\right|,$$
$$\dots$$
$$\max_m \left|z_m^N\right| \le \max_m \left|z_m^{N-1}\right| + \tau \max_{m,n} \left|\alpha_m^n\right|.$$

It should be noticed that $\max_{m,n} \left|\alpha_m^n\right|$ is the number independent from m and n. Having summed the previous inequalities and considering $z_m^0 = \beta_m,$ we get

$$\begin{aligned} &\max_m \left|z_m^N\right| \le \max_m \left|\beta_m\right| + N\tau \max_{m,n} \left|\alpha_m^n\right| \le \\ &\le \max_m \left|\beta_m\right| + T \max_{m,n} \left|\alpha_m^n\right| \le \\ &\le \max(1,T)(\max_{m,n} \left|\alpha_m^n\right| + \max_m \left|\beta_m\right|) = M \left\|g^{(h)}\right\|_{F_h}, \end{aligned} \tag{3.20}$$

where M is denoted

$$M = \max(1,T) = \begin{cases} 1, & \text{for } T < 1, \\ T, & \text{for } T \ge 1. \end{cases}$$

According to the formula (3.20)

$$\max_{m,n} \left|z_m^n\right| \le M \left\|g^{(h)}\right\|_{F_h} \text{ or } \left\|z^{(h)}\right\|_{U_h} \le M \left\|g^{(h)}\right\|_{F_h}.$$

Thus, the scheme (3.13) at the condition (3.18) is fulfilled is stable one. So, condition (3.18) is tough as it gives

$$\tau \le \frac{1}{2} h^2. \tag{3.21}$$

If we'd like to preserve stability, the small pitch at time τ should be chosen while calculating by the scheme (3.13). Let's turn to the difference scheme (3.14) corresponding to the pattern in the Fig. 4 and rewrite it as

$$\left.\begin{aligned} &-r(u_{m+1}^{n+1} + u_{m-1}^{n+1}) + (1+2r)u_m^{n+1} = u_m^n + \tau\varphi(x_m, t_n), \\ &u_m^0 = \psi(x_m), \\ &m = 0, \ \pm 1, \ \pm 2, \dots, \ n = 0, \ 1, \ 2, \dots, \ N-1, \ r = \frac{\tau}{h^2}. \end{aligned}\right\} \tag{3.22}$$

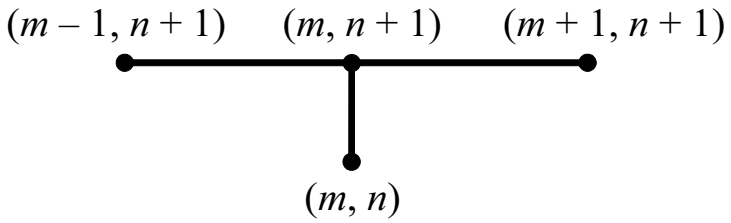


Fig. 4. Implicit two-layer pattern

What calculations should be carried out to find values u_m^1 (by the formula (3.22)) on the first-time layer with value u_m^0 on zero-time layer. Putting $n = 0$ into the Eq. (3.22), we have

$$\left.\begin{aligned} &-r(u_{m+1}^1 + u_{m-1}^1) + (1+2r)u_m^1 = u_m^0 + \tau\varphi(x_m, 0), \\ &u_m^0 = \psi(x_m), \\ &m = 0,\ \pm 1,\ \pm 2,\ldots\ . \end{aligned}\right\} \qquad (3.23)$$

These formulae are the continuum system of linear equations relevant to the unknown ..., u_{-2}^1, u_{-1}^1, u_0^1, u_1^1, u_2^1,...

It is difficult to solve such problems being time taking. So, difference schemes of the formula (3.14) are not convenient for Cauchy problem on continuous segments and is not widely used. If the segment of axis x, taken by Cauchy problem, is ended, that means $a \le x \le b$, $b - a \le K$, and additional limits to the solution of $u(x,t)$ are given on the lines $x = a$ and $x = b$, it provides the difference schemes of the Eq. (3.14) to be effective. Particularly, these schemes are absolutely stable at any $r = \tau / h^2 > 0$.

If on segments of these lines (mentioned above) the conditions $u(a,t) = \gamma_0(t)$, $u(b,t) = \gamma_1(t)$ are given, the system (3.23) changes:

$$\left.\begin{aligned} &-r(u_{m+1}^1 + u_{m-1}^1) + (1+2r)u_m^1 = \psi(x_m) + \tau\varphi(x_m, 0), \\ &u_m^1 = \gamma_0(t_1),\ \ u_M^1 = \gamma_1(t_1), \\ &m = 1,\ 2,\ldots,\ M-1,\ \ h = \frac{b-a}{M}. \end{aligned}\right\} \qquad (3.24)$$

The formula (3.24) presents the system ($M + 1$) of algebraic equations relevant to u_0^1, u_1^1,..., u_M^1. The matrix of these systems is three-diagonal and may be solved by the sweep method. Realization of implicit difference schemes requires more calculations for setting the solution on one-time layer. There may be few such layers, because of the limits absence to the ra-

tio τ/h^2. If we use explicit difference scheme, the recursive rule is used to find the solution on the next layer and it doesn't involve a lot of calculations. The number of time layers may be greater, it stands for limits $\frac{\tau}{h^2} \le \frac{1}{2}$.

Let's now consider the convergence of scheme (3.13), which approximates the problems (3.5), (3.6) with inaccuracy $O(r+h^2)$ is stable at $r \le 1/2$. Therefore scheme (3.13) is convergent by the theory on approximation and stability. Meanwhile inaccuracy of order $O(r+h^2)$ will be for an approximate solution.

3.3. Difference schemes for the equations of an elliptic type

These problems are to be considered by the equation of Poisson with constant coefficients.

3.3.1. Construction of difference approximation for the Poisson's equation

Let the Poisson's equation be in some area $\mathcal{D}$ with border Γ

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y). \qquad (3.25)$$

A rectangular mesh is chosen by the rule $(m,n) \in \Gamma_h$.

Then all nodes are transferred to the mesh area $\mathcal{D}_h$, which belong to the area $\overline{\mathcal{D}} = \mathcal{D} + \Gamma$ (Fig. 5).

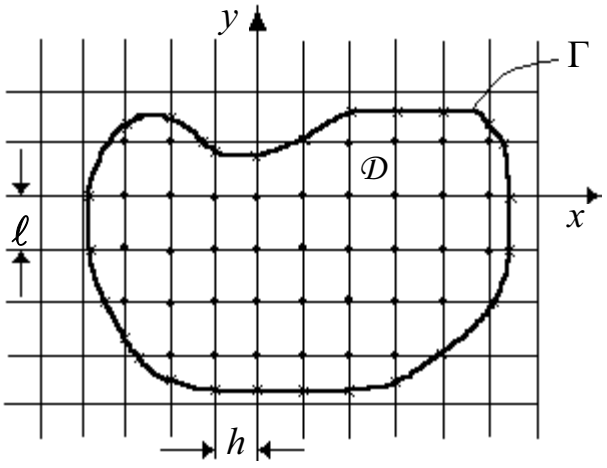


Fig. 5. Difference approximation construction

Let's take five-point pattern (Fig. 6).

Using the location of points at this pattern, let's divide area nodes into two categories: interior and boundary. The node (*m*, *n*) is taken as an interior,

if it by itself and four adjoining points of the pattern belong to area $\mathcal{D}_h$ (the nodes are marked by symbol °). Let's express the set of interior nodes through $\mathcal{D}_h^0$. Other nodes will be defined as boundary (marked asterisk "*"), and their set will be expressed through Γ_h.

So, $\mathcal{D}_h = \mathcal{D}_h^0 + \Gamma_h$.

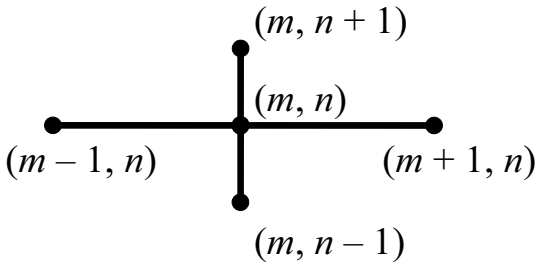


Fig. 6. Five-point pattern

It is clear that division of nodes from $\mathcal{D}_h$ into interior and boundary depends on chosen pattern.

The node $(m,n) \in \mathcal{D}_h^0$. The substitution of differential equations (3.25) by difference one will be fulfilled only in interior nodes.

We have

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_m,y_n)} + \left.\frac{\partial^2 u}{\partial y^2}\right|_{(x_m,y_n)} = f(x_m,y_n). \qquad (3.26)$$

Having used approximation of the second derivatives we get

$$\frac{u(x_{m+1},y_n) - 2u(x_m,y_n) + u(x_{m-1},y_n)}{h^2} - \frac{h^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{(x_m^{(1)},y_n)} +$$

$$\frac{u(x_m,y_{n+1}) - 2u(x_m,y_n) + u(x_m,y_{n-1})}{l^2} - \frac{l^2}{12}\left.\frac{\partial^4 u}{\partial y^4}\right|_{(x_m,y_n^{(1)})} = f(x_m,y_n), \quad (3.27)$$

$$(m,n) \in \mathcal{D}_h^0,$$

$$x_{m-1} < x_m^{(1)} < x_{m+1},\ y_{n-1} < y_n^{(1)} < y_{n+1}.$$

Let $\frac{\partial^4 u}{\partial x^4}$ and $\frac{\partial^4 u}{\partial y^4}$ be limited by an absolute value in $\overline{\mathcal{D}}$. Then in formulae (3.27) at sufficiently small h and l we can ignore members, containing h^2 and l^2 as a multipliers, we get the desired difference equation

$$L_h^{(1)}\ (u^{(h)}) = f^{(h)}, \tag{3.28}$$

where $L_h^{(1)}(u^{(h)}) \equiv \dfrac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{h^2} + \dfrac{u_{m,n+1} - 2u_{m,n} + u_{m,n-1}}{l^2}$,

$(m,n) \in \mathcal{D}_h^0$, $f^{(h)} \equiv f(x_m, y_n)$.

Being defined an equation residual we can get

$$L_h^{(1)}(u_h(x,y)) = f^{(h)} + \delta \cdot f^{(h)}, \tag{3.29}$$

where $u_h(x, y)$ is the exact solution at nodes:

$$\delta \cdot f^{(h)} \equiv \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}\bigg|_{(x_m^{(1)}, y_n)} + \frac{l^2}{12}\frac{\partial^4 u}{\partial y^4}\bigg|_{(x_m, y_n^{(1)})},\ (m,n) \in \mathcal{D}_h^0. \tag{3.30}$$

Under made suppositions relevantly to $\dfrac{\partial^4 u}{\partial x^4}$ and $\dfrac{\partial^4 u}{\partial y^4}$, as it goes from the formula (3.30), there is a bound

$$\left\| \delta \cdot f^{(h)} \right\|_{F_h} \le Mh^2. \tag{3.31}$$

Here M is a constant, independent from h and $l = \alpha \cdot h$.

The bound (3.31) shows that difference equation (3.28) approximates the equation (3.25) to the solution $u(x, y)$ with inaccuracy of order $O(h^2)$.

3.3.2. Different edge problems and approximation of edge conditions

There are three types of conditions that join to the equations of elliptical type (particularly, to the Poisson's equation (3.25)):

1) boundary conditions of the first type:

$$u_\Gamma = \varphi(M); \tag{3.32}$$

2) boundary conditions of the second type:

$$\frac{\partial u}{\partial n}\bigg|_\Gamma = \varphi(M), \tag{3.33}$$

$\dfrac{\partial u}{\partial n}$ – exterior normal derivative;

3) boundary conditions of the third type:

$$\left[\alpha(x,y)\frac{\partial u}{\partial n} + \beta(x,y)u\right]_\Gamma = \varphi(M), \tag{3.34}$$

α, β, φ are known functions.

If it is demanded to define the function $u(x, y)$, which satisfies equation (3.25) in $\mathcal{D}$-area, and on the boundary Γ it satisfies one of the edge conditions, the edge problem for elliptical equation is given.

The problems (3.25), (3.32) are called Dirichlet's problems,

Problems (3.25), (3.33) – Neuman's problems,

The problems (3.25), (3.34) are mixed boundary problem.

How boundary conditions of the first type are substituted by difference conditions (Fig. 6)? The boundary conditions are substituted by the conditions on the set of boundary nodes Γ_h.

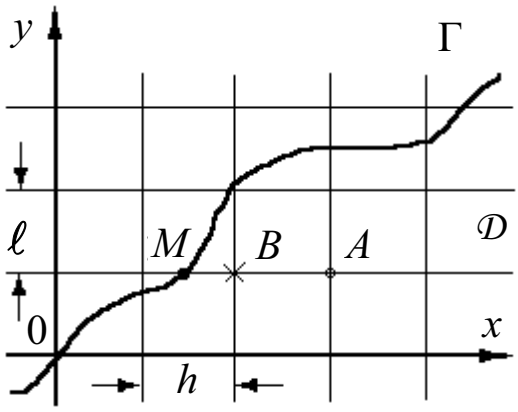


Fig. 6. Substitution of the boundary conditions of the first type by difference conditions

Let (m, n) be some node from Γ_h. Let's denote it as B; $(m + 1, n)$ is an interior node, the nearest to B at x direction. By letter E we define the point of the contour Γ, the nearest to B at x direction.

The coordinates of these points are as follows:

$$M(x_m - \delta, y_n), \ 0 < \delta < h, \ B(x_m, y_n), \ A(x_{m+1}, y_n).$$

By the condition (3.32) we have $u\big|_{\Gamma} = \varphi(M)$.

It allows for nodes $(m,n) \in \Gamma_h$ to put

$$u_{mn} = \varphi(M) = \varphi(x_m - \delta, y_n). \tag{3.35}$$

Let's find inaccuracy of the formula (3.35)

$$\varphi(M) = u(x_m - \delta, y_n) = u(x_m, y_n) - \delta \left.\frac{\partial u}{\partial x}\right|_{(x_m^{(1)}, y_n)}, \ x_m - \delta < x_m^{(1)} < x_m.$$

It means

$$u(x_m, y_n) - \varphi(M) = u(x_m, y_n) - u_{mn} = -\delta \left.\frac{\partial u}{\partial x}\right|_{(x_m^{(1)}, y_n)}.$$

The inaccuracy of the formula (3.35) has the first order relevantly to h supposing that $\delta = \alpha h$. If points M and B coincide, the formula (3.35) will be precise. The precision of u_{mn} at $(m,n) \in \Gamma_h$ may be increased by values $u(x, y)$ at A-point.

We have:

$$u(M) = u(x_m - \delta, y_n) = u(B) - \delta \left.\frac{\partial u}{\partial x}\right|_B + O(\delta^2), \tag{3.36}$$

$$u(A) = u(x_m + h, y_n) = u(B) + h \left.\frac{\partial u}{\partial x}\right|_B + O(h^2). \tag{3.37}$$

Having excluded $\left.\frac{\partial u}{\partial x}\right|_B$ from the Eq. (3.36) with the help of the formula (3.37), we get

$$u(B) = \frac{h\varphi(M) + u(A)\delta}{h + \delta} + O(h^2).$$

Taking away the value $O(h^2)$, we search for difference boundary condition that approximates the boundary condition (3.32) at the node $(m,n) \in \Gamma_h$ with inaccuracy $O(h^2)$:

$$u_{mn} = \frac{h\varphi(M) + u_{m+1,n}\delta}{h + \delta}. \tag{3.38}$$

Let's turn to the interchange of the second order boundary condition with difference equation (Fig. 7).

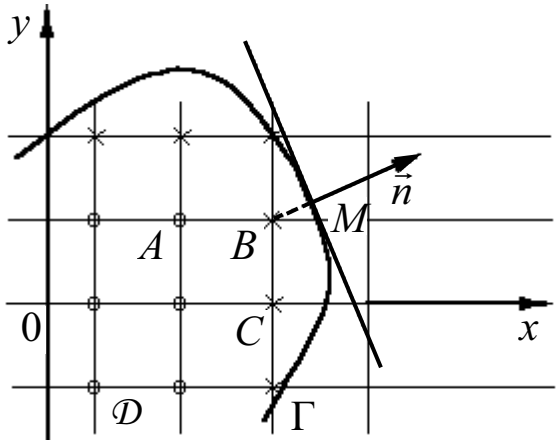


Fig. 7. Substitution of the second order boundary conditions by difference conditions

Let B be a boundary node with coordinates (x_m, y_n), M – the nearest point of Γ contour to B, A – an interior node with coordinates (x_{m-1}, y_n), C – a boundary node with coordinates (x_m, y_{n-1}), $\vec{n}$ – an exterior normal

to Γ at M point. Let's denote the angle between $\vec{n}$ and the axes Ox through α, and between $\vec{n}$ and Oy – through β. It is clear that $\beta = \frac{3\pi}{2} + \alpha$.

According to the definition we have

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial x}\cos\alpha + \frac{\partial u}{\partial y}\cos\beta = \frac{\partial u}{\partial x}\cos\alpha + \frac{\partial u}{\partial y}\sin\alpha .$$

Let's suppose that at point B the direction of normal is the same as at point M. As far as the distance between B and M is the value of order O(h), this supposition will be connected with taking in an accuracy of the same order O(h). It means that

$$\left.\frac{\partial u}{\partial n}\right|_M \approx \left.\frac{\partial u}{\partial n}\right|_B .$$

Finally we get

$$\frac{u(x_m, y_n) - u(x_{m-1}, y_n)}{h}\cos\alpha + \frac{u(x_m, y_n) - u(x_m, y_{n-1})}{l}\sin\alpha + O(h+l) = \varphi(M).$$

Using approximate mesh values we can find

$$\frac{u_{m,n} - u_{m-1,n}}{h}\cos\alpha + \frac{u_{m,n} - u_{m,n-1}}{l}\sin\alpha = \varphi(M). \qquad (3.39)$$

This formula is a difference approximation at the node $(m,n) \in \Gamma_h$ of the second order boundary condition with accuracy $O(h+l)$.

The expressions of (3.39) type must be written for all boundary nodes $u\big|_\Gamma = \varphi(M)$, afterwards, the difference boundary conditions will be got, approximating boundary conditions (3.33). The substitution of boundary conditions by the difference conditions might be too complicated particularly if the contour Γ has no simple form. The substitution of the third order boundary conditions can be fulfilled with the help of formulae (3.35), (3.37), (3.38).

3.3.3. The construction of difference scheme in case of Dirichlet's problem for Poisson's equation

Let the Poisson's equation be defined in a rectangular area $\mathcal{D} = \{0 < x < a\ ,\ \ 0 < y < b\}$:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y), \qquad (3.40)$$

and the Dirichlet's condition be on the border Γ of $\mathcal{D}$-area

$$u\big|_\Gamma = \varphi(M). \qquad (3.41)$$

Let the problem (3.40), (3.41) has the only solution $u(x, y)$ in area $\overline{\mathcal{D}} = \mathcal{D} + \Gamma$ and this solution has continuous derivatives $\frac{\partial^4 u}{\partial x^4}$ and $\frac{\partial^4 u}{\partial y^4}$ in $\mathcal{D}$.

Let's take a rectangular mesh

$$x_m = mh, \quad m = 0,\ 1,\ 2,...,\ M, \quad h = \frac{a}{M},$$

$$y_n = nl, \quad n = 0,\ 1,\ 2,...,\ N, \quad l = \frac{b}{N}.$$

Using formulae (3.28), (3.35) we write a difference scheme which approximates the problems (3.40), (3.41) with inaccuracy of order $O(h^2 + l^2)$:

$$L_h(u^{(h)}) = f^{(h)}, \tag{3.42}$$

where $L_h(u^{(h)}) \equiv \begin{cases} \dfrac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{h^2} + \dfrac{u_{m,n+1} - 2u_{m,n} + u_{m,n-1}}{l^2}, \\ m = 1,\ 2,...,\ M-1; \quad n = 1,\ 2,...,\ N-1, \\ u_{mn},\ (mh,\ nl) \in \Gamma_h; \end{cases}$

$$f^{(h)} \equiv \begin{cases} f(x_m, y_n), \ m = 1,\ 2,...,\ M-1; \ n = 1,\ 2,...,\ N-1, \\ \varphi(x_m, y_n), \ (mh, nl) \in \Gamma_h. \end{cases}$$

The difference scheme (3.42) is a linear algebraic equations system. The number of this system equation equals $(M-1)\times(N-1)$.

There is the same number of unknown quantities: u_{mn} at $m = 1, 2, \ldots, M-1$; $n = 1, 2, \ldots, N-1$.

Prove of the scheme (3.42) stability leads to two qualities:

1. The difference scheme $L_h(z^{(h)}) = g^{(h)}$, where $g^{(h)} \equiv \begin{cases} \alpha_{mn} \\ \beta_{mn} \end{cases}$ is an arbitrary element from F_h, is to be solved in a single way.
2. There is a bound $\left\| z^{(h)} \right\|_{U_h} \le C \left\| g^{(h)} \right\|_{F_h}$, where C is a constant independent of h and $g^{(h)}$.

Here the norms mentioned above are defined by the rule

$$\left\| z^{(h)} \right\|_{U_h} = \max_{m,n} \left| z_{m,n} \right|, \quad \left\| g^{(h)} \right\|_{F_h} = \max_{m,n} \left| \alpha_{mn} \right| + \max_{m,n} \left| \beta_{mn} \right|.$$

Let's introduce the designation $\Lambda_h(U^{(h)}) \equiv \Lambda_{xx}(U^{(h)}) + \Lambda_{yy}(U^{(h)})$.

Lemma 1. *Let* $\vartheta^{(h)} = \{\vartheta_{mn}\}$, $\vartheta_{mn} \neq const$ *be a mesh function, defined on* $\mathcal{D}_h = \mathcal{D}_h^0 + \Gamma_h$. *If the condition*

$$\Lambda_h(\vartheta^{(h)})\Big|_{(x_m,y_n)} \geq 0, \text{ where } (x_m, y_n) \in \mathcal{D}_h^0 \tag{3.43}$$

is fulfilled, $\vartheta^{(h)}$ *reaches its highest-range value on* $\mathcal{D}^{(h)}$ *at boundary points, that means on* Γ_h.

Prove. Let's take the opposite. Let $\vartheta^{(h)}$ have its highest-range value ϑ_{ij} at $\mathcal{D}_h^0$. We can consider even only of values $\vartheta_{i+1,j}$; $\vartheta_{i-1,j}$; $\vartheta_{i,j+1}$; $\vartheta_{i,j-1}$ exactly less than ϑ_{ij}. Then we get

$$\Lambda_{xx}(\vartheta^{(h)})\Big|_{(x_i,y_j)} = \frac{\vartheta_{i+1,j} - 2\vartheta_{ij} + \vartheta_{i-1,j}}{h^2} = \frac{(\vartheta_{i+1,j} - \vartheta_{ij}) + (\vartheta_{i-1,j} - \vartheta_{ij})}{h^2} \leq 0;$$

$$\Lambda_{yy}(\vartheta^{(h)})\Big|_{(x_i,y_j)} = \frac{\vartheta_{i,j+1} - 2\vartheta_{ij} + \vartheta_{i,j-1}}{l^2} = \frac{(\vartheta_{i,j+1} - \vartheta_{ij}) + (\vartheta_{i,j-1} - \vartheta_{ij})}{l^2} \leq 0.$$

One of them is fully negative in virtue of the supposition about value ϑ_{ij}. Finally we get

$$\Lambda_h(\vartheta^{(h)})\Big|_{(x_i,y_j)} = \Lambda_{xx}(\vartheta^{(h)})\Big|_{(x_i,y_j)} + \Lambda_{yy}(\vartheta^{(h)})\Big|_{(x_i,y_j)} < 0. \tag{3.44}$$

The formula (3.44) contradicts the condition of the lemma (3.43), and, consequently, our supposition is wrong.

Lemma 2. *Let* $\vartheta^{(h)} = \{\vartheta_{mn}\}$ *be a mesh function, defined on* $\mathcal{D}_h$, $\vartheta_{mn} \neq const$. *If the following condition is fulfilled:*

$$\Lambda_h(\vartheta^{(h)})\Big|_{(x_m,y_n)} = 0, \quad (x_m, y_n) \in \mathcal{D}_h^0. \tag{3.45}$$

Then $\vartheta^{(h)}$ *reaches its lowest-range value on* $\mathcal{D}_h$ *at the boundary points that is on* Γ_h.

The proof is the same.

Theorem (maximum principle). *Each solution of difference equation*

$$\Lambda_h(\vartheta^{(h)})\Big|_{(x_m,y_n)} = 0, \quad (x_m, y_n) \in \mathcal{D}_h^0$$

takes its highest and lowest-range value at some points of the boundary Γ_h.

The prove follows lemma 1 and 2.

Let's use the theorem to prove a single solution of the difference scheme

$$L_h(z^{(h)}) = g^{(h)} \tag{3.46}$$

at any $g^{(h)} \in F_h$. So, we consider a homogeneous difference scheme $L_h(z^{(h)}) = 0$ and show that it has only zero solution $z^{(h)} \equiv 0$. We write down the homogeneous difference scheme $L_h(z^{(h)}) = 0$ as

$$L_h(z^{(h)}) \equiv \begin{cases} \Lambda_h(z^{(h)})\Big|_{(x_m, y_n)} = 0, & (x_m, y_n) \in \mathcal{D}_h^0; \\ z^{(h)}\Big|_{(x_m, y_n)} = 0, & (x_m, y_n) \in \Gamma_h. \end{cases}$$

If $z^{(h)}$ doesn't equal zero identically, the value $z^{(h)}$ could have the highest-range and lowest-range values at points in view of maximum principle. On Γ_h $z^{(h)} \equiv 0$, that means even in $\mathcal{D}_h$ it will be the same $z^{(h)} \equiv 0$. But $z^{(h)}$ is the solution to the homogeneous system of algebraic equations. If this system has only a trivial solution, the determinant of the array equals zero. The difference scheme (3.46) is the system of algebraic linear equations with the same array. As its determinant is different, the scheme (3.46) has only one solution. The scheme (3.46) is solvable.

Now let's show that the scheme (3.46) has the property

$$\max_{m,n} |z_{mn}| \le C\Big(\max_{(m,n) \in \mathcal{D}_h^0} |\alpha_{mn}| + \max_{(m,n) \in \Gamma_h} |\beta_{mn}| \Big). \qquad (3.47)$$

If we find it out, the scheme stability (3.42) will be proved. The bound (3.47) is to be got by constructing the majorants for function $\left|z^{(h)}\right|$ by Guershgorin's rule.

Firstly, let $P(x, y)$ be a polynomial of the second order for two arguments: $P(x, y) = a_{00} + a_{10}x + a_{20}x^2 + a_{01}y + a_{02}y^2 + a_{11}xy$.

Then

$$\Lambda_h(P^{(h)})\Big|_{(x_m, y_n)} = 2(a_{20} + a_{02}). \qquad (3.48)$$

Actually, let's view

$$\Lambda_{xx}(P^{(h)})\Big|_{(x_m, y_n)} = \frac{1}{h^2}\Big[P_{m+1,j} - 2P_{m,j} + P_{m-1,j} \Big] =$$

$$\frac{1}{h^2}\Big[P(x_m + h, y_n) - 2P(x_m, y_n) + P(x_m - h, y_n) \Big] =$$

$$= \frac{1}{h^2} = [a_{00} + a_{10}(x_m + h) + a_{20}(x_m + h)^2 + a_{01}y_n + a_{02}y_n^2 +$$

$$+ a_{11}(x_m + h)y_n + a_{00} + a_{10}(x_m - h) + a_{20}(x_m - h)^2 + a_{01}y_n +$$

$$+ a_{02}y_n^2 + a_{11}(x_m - h)y_n - 2(a_{00} + a_{10}x_m + a_{20}x_m^2 + a_{01}y_n +$$

$$+ a_{02}y_n^2 + a_{11}x_m y_n)] = \frac{2a_{20}h^2}{h^2} = 2a_{20}.$$

Similarly we get $\Lambda_{yy}(P^{(h)})\Big|_{(x_m,y_n)} = 2a_{02}$.

The majorant $z(x,y)$ is to be defined by rule

$$z(x,y) = \frac{1}{4}\Big[(a^2+b^2)-(x^2+y^2)\Big] \max_{(m,n)\in\mathcal{D}_h^0} \left|\alpha_{mn}\right| + \max_{(m,n)\in\Gamma_h} \left|\beta_{mn}\right|.$$

Let's illustrate a geometrical sense of the function

$$S(x,y) = (a^2+b^2)-(x^2+y^2).$$

Let's view the Fig. 8 at which there is the area $\mathcal{D}$ with Γ contour.

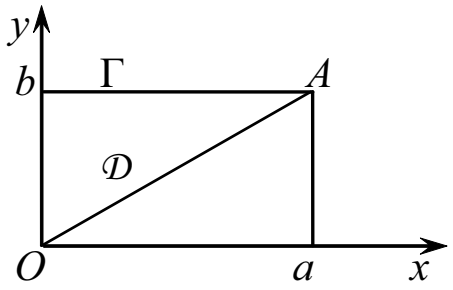


Fig. 8. Attached to the prove of maximum principle

The length of a diagonal OA equals $\sqrt{a^2+b^2}$, that's why the equation of the curve $S(x,y)=0$ is the equation of circle with the center at the origin of coordinates and the radius equals to $\left|OA\right| = \sqrt{a^2+b^2}$. So if $(x,y)\in\overline{\mathcal{D}} = \mathcal{D}+\Gamma$, then $S(x,y)\ge 0$, while $S(x,y)=0$ only for one point from $\mathcal{D}_h$, when $x = b$, $y = b$. This point doesn't belong to $\mathcal{D}_h$ and $S(x,y)\big|_{(x_m,y_n)\in\mathcal{D}_h} > 0$.

Let's show that

$$\left|z^{(h)}\right| \le Z^{(h)}, \tag{3.49}$$

or, in other words, $Z^{(h)}$ is a majorant for $z^{(h)}$.

Taking the sum of $\left(z^{(h)}+Z^{(h)}\right)$ on the set of Γ_h we get

$$\begin{aligned}\Big[z^{(h)}+Z^{(h)}\Big]\Big|_{(x_m,y_n)\in\Gamma_h} &= z^{(h)}\Big|_{\Gamma_h} + Z^{(h)}\Big|_{\Gamma_h} = \beta_{mn}\big|_{(x_m,y_n)\in\Gamma_h} + \\ &+\frac{1}{4}S(x,y)\big|_{(x_m,y_n)\in\Gamma_h} \cdot \max_{(m,n)\in\mathcal{D}_h^0}\left|\alpha_{mn}\right| + \max_{(m,n)\in\Gamma_h}\left|\beta_{mn}\right| \ge 0,\end{aligned} \tag{3.50}$$

as far as $\max\limits_{(m,n)\in\Gamma_h} |\beta_{mn}| \geq \beta_{mn}\big|_{(x_m,y_n)\in\Gamma_h}$.

Besides,

$$\Lambda_h(z^{(h)}) + Z^{(h)})\Big|_{\mathcal{D}_h^0} =$$

$$= \Lambda_h(z^{(h)})\Big|_{\mathcal{D}_h^0} + \Lambda_h(Z^{(h)})\Big|_{\mathcal{D}_h^0} = \alpha_{m,n} + (-\max_{(m,n)\notin\mathcal{D}_h^0} |\alpha_{mn}|) \leq 0. \quad (3.51)$$

In virtue of the second lemma it goes from (3.51) that function $(z^{(h)} + Z^{(h)})$ takes its lowest value at points Γ_h, but according to (3.50) on Γ_h the inequality $(z^{(h)} + Z^{(h)}) \geq 0$ is fulfilled. So everywhere in $\mathcal{D}_h$ there is

$$z^{(h)} + Z^{(h)} \geq 0. \quad (3.52)$$

If we consider the difference $(z^{(h)} - Z^{(h)})$ and use the first lemma, we get everywhere in $\mathcal{D}_h$

$$z^{(h)} - Z^{(h)} \leq 0. \quad (3.53)$$

From the Eqs. (3.52), (3.53) it follows

$$\left|z^{(h)}\right| \leq Z^{(h)} \quad (3.54)$$

for every $(x_m, y_n) \in \mathcal{D}_h$. From the bound (3.54) we find

$$|z_{mn}| \leq \frac{1}{4} S(x_m, y_n) \max_{(m,n)\in\mathcal{D}_h^0} |\alpha_{mn}| + \max_{(m,n)\in\Gamma_h} |\beta_{mn}| \leq$$

$$\leq \frac{1}{4}(a^2 + b^2) \max_{(m,n)\in\mathcal{D}_h^0} |\alpha_{mn}| + \max_{(m,n)\in\Gamma_h} |\beta_{mn}| \leq$$

$$\leq \max\left\{\frac{1}{4}(a^2 + b^2),\ 1\right\}\left[\max_{(m,n)\in\mathcal{D}_h^0} |\alpha_{mn}| + \max_{(m,n)\in\Gamma_h} |\beta_{mn}|\right] =$$

$$= C \cdot \left[\max_{(m,n)\in\mathcal{D}_h^0} |\alpha_{mn}| + \max_{(m,n)\in\Gamma_h} |\beta_{mn}|\right].$$

So, $\max\limits_{m,n}|z_{mn}| \le C\left[\max\limits_{(m,n)\in\mathcal{D}_h^0}|\alpha_{mn}| + \max\limits_{(m,n)\in\Gamma_h}|\beta_{mn}|\right]$.

It means that the bound (3.53) is defined, and, consequently, the stability of difference scheme (3.42) is proved.

The difference scheme (3.42) approximates the problems (3.40), (3.41) with inaccuracy of order $O(h^2)$ (we suppose that $\frac{h}{l} = const$). Besides, this scheme is stable. Finally the scheme is convergent, and the rate of its convergence is $O(h^2)$.

3.3.4. Matrix sweep method

Let's rewrite the difference scheme (3.42):

$$u_{m+1,n} - 2u_{mn} + u_{m-1,n} + \alpha(u_{m,n+1} - 2u_{mn} + u_{m,n-1}) = h^2 f(x_m, y_n), \quad (3.55)$$

$$m = 1,\ 2,...,\ M-1; \quad n = 1,\ 2,...,\ N-1;$$
$$u_{0n} = \varphi(0, y_n), \quad u_{Mn} = \varphi(a, y_n), \quad n = 1, 2, \dots, N-1;$$
$$u_{m0} = \varphi(x_m, 0), \quad u_{mN} = \varphi(x_m, b), \quad m = 1, 2, \dots, M-1;$$
$$\frac{h^2}{l^2} = \alpha > 0.$$

Let's be $M > N$. We introduce the following symbols:

$$u_m = \begin{Vmatrix} u_{m,1} \\ \vdots \\ u_{m,N-1} \end{Vmatrix}, \quad m = 0,...,M. \quad (3.56)$$

Let's put $n = 1, \dots, N-1$, into the formula (3.55), using (3.56), and write the system of equations (3.55).

In some way

$$\left.\begin{aligned} &u_{m+1} + Au_m + u_{m-1} = f_m, \qquad m = 1,\ ...,\ M-1, \\ &u_0 = \varphi_0, \quad u_M = \varphi_a, \end{aligned}\right\} \quad (3.57)$$

where A is a three-diagonal matrix of $(N-1)$ order

$$A = \begin{Vmatrix} -(2+2\alpha) & \alpha & 0 & \dots & 0 & 0 & 0 \\ \alpha & -(2+2\alpha) & \alpha & \dots & 0 & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & \dots & \alpha & -(2+2\alpha) & \alpha \\ 0 & 0 & 0 & \dots & 0 & \alpha & -(2+2\alpha) \end{Vmatrix};$$

$$f_m = \begin{Vmatrix} h^2 f(x_m, y_1) - \alpha\varphi(x_m, 0) \\ h^2 f(x_m, y_2) \\ \vdots \\ h^2 f(x_m, y_{N-2}) \\ h^2 f(x_m, y_{N-1}) - \alpha\varphi(x_m, b) \end{Vmatrix};$$

$$\varphi_0 = \begin{Vmatrix} \varphi(0, y_1) \\ \varphi(0, y_2) \\ \vdots \\ \varphi(0, y_{N-1}) \end{Vmatrix}, \ \varphi_a = \begin{Vmatrix} \varphi(a, y_1) \\ \varphi(a, y_2) \\ \vdots \\ \varphi(a, y_{N-1}) \end{Vmatrix}.$$

The problem (3.57) is identical to the problem having been solved with us by the matrix sweep method. Particularly, it has a vector form.

Let's state the algorithm of (3.57) problem solution, which is called the matrix sweep method.

1. By formula $R_{m+1} = -(A + R_m)^{-1}$, $m = 1, \dots, M-1$, supposing $R_1 = 0$, we calculate matrices $R_m = (R_{ij}^{(m)})$, $m = 1, 2, \dots, M$. Their order is $(N-1)\cdot(N-1)$.

   After that we place vector $S_1 = \varphi_0$, then by the formula

$$S_{m+1} = R_{m+1}(S_m - f_m), \quad m = 1, 2, \dots, M-1$$

we calculate vectors

$$S_m = \begin{Vmatrix} S_1^{(m)} \\ S_2^{(m)} \\ \vdots \\ S_{N-1}^{(m)} \end{Vmatrix}, \quad m = 1, 2, \dots, M.$$

2. Let's put $u_M = \varphi_a$. By the formula

$$u_{m-1} = R_m u_m + S_M, \quad m = M, M-1, \dots, 1,$$

calculate the desired values of the problem solution (3.57) $u_M, u_{M-1}, \dots, u_1, u_0$.

3.3.5. Iteration method of difference solution method for Dirichlet's problem

Let's express the values of u_{mn} from the scheme (3.42)

$$\left.\begin{aligned} &u_{mn} = \frac{1}{2(1+\alpha)}\Big[u_{m+1,n} + u_{m-1,n} + \alpha(u_{m,n+1} + u_{m,n-1}) - h^2 f_{m,n}\Big], \\ &m = 1,\ 2,...,\ M-1; \quad n = 1,\ 2,...,\ N-1; \\ &\alpha = \frac{h^2}{l^2}, \quad f_{m,n} = f(x_m, y_n). \end{aligned}\right\} \tag{3.58}$$

The values on the boundaries are known:

$$\left.\begin{aligned} &u_{0n} = \varphi(0, y_n), \quad u_{Mn} = \varphi(a, y_n), \quad n = 1,\ 2,...,\ N-1; \\ &u_{m0} = \varphi(x_m, 0), \quad u_{mN} = \varphi(x_m, b), \quad m = 1,\ 2,...,\ M-1. \end{aligned}\right\} \tag{3.59}$$

In equality (3.58) the value u_{mn} is expressed through four adjoining values u_{ij} by five-point pattern.

In the iteration solution method the values u_{mn} at all inner points of the area $\mathcal{D}_h$ it's supposed to be equal to some arbitrary values.

Often it is thought as

$$u_{mn}^{(0)} = 0, \ \ m = 1,\ 2,\ ...\ ,\ M-1; \ \ n = 1,\ 2,\ ...\ ,\ N-1.$$

Then with the help of formulae (3.58), (3.59) new values $u_{mn}^{(1)}$, then $u_{mn}^{(2)}$ (and any others) are calculated until the maximum deflection of mesh functions values on previous and current iterations becomes less by module, than some of given accuracy ε in advance. So the iterations are stopped when the following condition fulfilled:

$$MAX = \max_{m,n}\left|u_{mn}^{(l)} - u_{mn}^{(l-1)}\right| < \varepsilon. \tag{3.60}$$

3.4. Difference schemes for simple equations of hyperbolic type

A typical and the simplest equation of the hyperbolic type is a wave equation

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y). \tag{3.61}$$

With respect to this equation, we are going to view the following problems:

1. **Caushy problem**. In area $\mathcal{D} = \{y > 0, \quad -\infty < x < \infty\}$ it is to be found the function $U(x, y)$ satisfying the equation (3.61), and on the line $y = 0$ it satisfies the initial conditions

$$u(x,0)=\varphi(x), \qquad \left.\frac{\partial u}{\partial y}\right|_{y=0}=\psi(x). \tag{3.62}$$

2. **Mixed boundary problem**. In area $\mathcal{D}=\{y>0,\ \alpha<x<\beta\}$ it is to be found the function $u(x,y)$, which satisfies the equation (3.61). On the boundary Γ of $\mathcal{D}$-area at $y = 0$ it satisfies the initial conditions (3.62). At $x=\alpha,\ x=\beta$ it satisfies one of the boundary conditions:

a) conditions of the first type

$$u(\alpha,y)=\mu_1(y), \quad u(\beta,y)=\mu_2(y); \tag{3.63}$$

b) conditions of the second type

$$\left.\frac{\partial u}{\partial x}\right|_{x=\alpha}=\delta_1(y), \quad \left.\frac{\partial u}{\partial x}\right|_{x=\beta}=\delta_2(y); \tag{3.64}$$

c) conditions of the third type

$$\left.\begin{aligned}\left[\tau_1(y)\frac{\partial u}{\partial x}+\tau_2(y)u\right]_{x=\alpha}&=\omega_1(y),\\ \left[\delta_1(y)\frac{\partial u}{\partial x}+\delta_2(y)u\right]_{x=\beta}&=\omega_2(y).\end{aligned}\right\} \tag{3.65}$$

3.4.1. Solving Cauchy problem

Let's choose a triangular mesh, taking

$$x_m=mh, \quad m=0,\ \pm1,\ \dots;$$
$$y_n=nl, \quad n=0,\ 1,\ \dots.$$

Let's view three-layer pattern (Fig. 7)

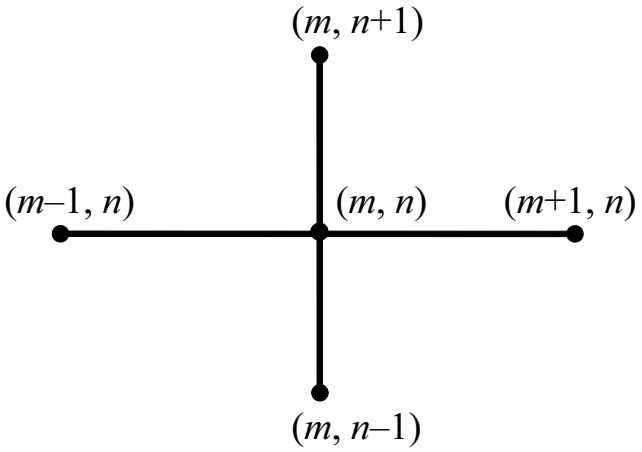


Fig. 7. Three-layer pattern

According to this pattern we define the set $\mathcal{D}_h^0$ of interior nodes and the set Γ_h of boundary nodes. We refer nodes on the line $y = 0$ to the set Γ_h, and nodes $(x_m, y_n) \in \mathcal{D}$ – to $\mathcal{D}_h^0$. The total mesh area $\mathcal{D}_h = \mathcal{D}_h^0 + \Gamma_h$ contains the nodes $(x_m, y_n) \in \mathcal{D} = \mathcal{D} + \Gamma$. Using the given pattern we get the difference scheme

$$L_h(u^{(h)}) = f^{(h)}, \tag{3.66}$$

where

$$L_h(u^{(h)}) = \begin{cases} \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} - \dfrac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{l^2}, \\ u_m^0, \\ \dfrac{u_m^1 - u_m^0}{l}, \\ m = 0, \ \pm 1, \ \ldots; \quad n = 1, \ 2, \ \ldots; \end{cases}$$

$$f^{(h)} = \begin{cases} f(x_m, y_n), \\ \varphi(x_m), \\ \psi(x_m), \\ m = 0, \ \pm 1, \ \ldots; \quad n = 1, \ 2, \ \ldots . \end{cases}$$

This scheme, that is easy to be proved, approximates the equation (3.61) with inaccuracy of order $O(h^2 + l^2)$, and the initial conditions with inaccuracy $O(l)$.

The order of initial conditions approximation can be increased. To this purpose we replace $\left.\dfrac{\partial U}{\partial y}\right|_{(x_m, 0)}$ by formula

$$\left.\frac{\partial u}{\partial y}\right|_{(x_m, 0)} = \frac{u(x_m, y_1) - u(x_m, y_{-1})}{2l} - \frac{l^2}{6}\left.\frac{\partial^3 u}{\partial y^3}\right|_{(x_m, y_0^{(1)})}, \tag{3.67}$$

where $y_{-1} = -l$, $-l < y_0^{(1)} < l$. From formula (3.67) we get a mesh condition, approximating the initial condition $\left.\dfrac{\partial u}{\partial y}\right|_{y=0} = \psi(x)$ with inaccuracy of order $O(l^2)$:

$$\frac{u_m^1 - u_m^{-1}}{2l} = \psi(x_m), \quad m = 0, \ \pm 1, \ \ldots . \tag{3.68}$$

The value u_m^{-1} may be excluded with the help of difference equation (3.66), having put into it $n = 0$. We have

$$\frac{u_{m+1}^0 - 2u_m^0 + u_{m-1}^0}{h^2} - \frac{u_m^1 - 2u_m^0 + u_m^{-1}}{l^2} = f(x_m, y_0).$$

That's why the Eq. (3.68) can be rearranged

$$\frac{u_m^1 - u_m^0}{l} = \psi(x_m) + \frac{1}{2} l \Lambda_{xx}(U_m^0) - \frac{1}{2} l f(x_m, y_0), \tag{3.69}$$

where $\Lambda_{xx}(u_m^0) \equiv \frac{u_{m+1}^0 - 2u_m^0 + u_{m-1}^0}{h^2}$.

So, instead of the difference scheme (3.66) another one can be written, approximating problems (3.61), (3.62) with inaccuracy of order $O(h^2 + l^2)$:

$$L_h(u^{(h)}) = f^{(h)}. \tag{3.70}$$

The operator L_h is calculated in the Eq. (3.66) but

$$f^{(h)} = \begin{cases} f(x_m, y_n), \\ \varphi(x_m), \\ \psi(x_m) + \frac{l}{2}(\Lambda_{xx}(u_m^0) - f(x_m, 0)), \\ m = 0, \ \pm 1, \ldots; \quad n = 1, \ 2, \ \ldots . \end{cases}$$

Let's demonstrate how to calculate values u_m^2 using values u_m^0 and u_m^1. By virtue of the formulae (3.66) and (3.70) we have

$$\begin{aligned} & u_m^0 = \varphi(x_m), \\ & u_m^1 = \varphi(x_m) + l(\psi(x_m) + \frac{1}{2} l \Lambda_{xx}(\varphi(x_m)) - \frac{1}{2} l f(x_m, 0)), \\ & m = 0, \ \pm 1, \ \ldots . \end{aligned} \tag{3.71}$$

The difference equation in the scheme (3.70) is to be rewritten as

$$\begin{aligned} & u_m^{n+1} = 2u_m^n + l^2 \Lambda_{xx}(u_m^n) - u_m^{n-1} - l^2 f(x_m, y_n), \\ & m = 0, \ \pm 1, \ \ldots; \quad n = 1, \ 2, \ \ldots . \end{aligned} \tag{3.72}$$

At $n = 1$ by the formula (3.72) we calculate values u_m^2, $m = 0, \ \pm 1, \ldots$; values u_m^0 and u_m^1 are known in the view of the Eq. (3.71). Then by the formula (3.72) at $n = 2$ we find values u_m^3 through known quantities: u_m^1 , u_m^2, and so on.

3.4.2. Solving mixed problem

Let the equation be solved

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y), \tag{3.73}$$

given in the area $\mathcal{D} = \{0 < y \le Y < \infty, \ \alpha < x < \beta\}$.

We are going to consider that the initial conditions are joined the equation (3.73)

$$u(x,0) = \varphi(x), \ \left.\frac{\partial u}{\partial y}\right|_{y=0} = \psi(x), \ \alpha \le x \le \beta \tag{3.74}$$

and also the boundary conditions of the third order

$$\begin{aligned} \left[\tau_1(y)\frac{\partial u}{\partial x} + \tau_2(y)u\right]_{x=\alpha} &= \omega_1(y), \\ \left[\delta_1(y)\frac{\partial u}{\partial x} + \delta_2(y)u\right]_{x=\beta} &= \omega_2(y). \end{aligned} \tag{3.75}$$

We choose the plain triangular mesh, taking

$$x_m = mh, \quad m = 0, 1, ..., M, \ h = \frac{b-\alpha}{M};$$

$$y_n = nl, \quad n = 0, 1, \ \dots, N, \ Nl \le Y < (N+1)\cdot l, \quad l > 0.$$

To substitute the equation (3.73) by difference one we use by five-points three-layer pattern. The mesh area $\mathcal{D}_h$ is to be divided into set of $\mathcal{D}_h^0$ interior nodes

$$\mathcal{D}_h^0 = \{(x_m, y_n), \ m = 1, \dots, M-1; \ n = 1, \dots, N-1\},$$

and a set of Γ_h boundary nodes

$$\begin{aligned} \Gamma_h = \{&(x_m, y_n), \ m = 0, \dots, M, \ n = 0; \\ &m = 0, \ n = 0, ..., N; \ m = M, \ n = 0, \dots, N\}. \end{aligned}$$

On the set of $\mathcal{D}_h^0$ the equation (3.73) and the initial conditions (3.74) are approximated by the difference scheme of (3.70) type. To substitute the boundary conditions on lines $x = \alpha$ and $x = \beta$, we use formulae of (3.68) type

$$\left.\begin{aligned} \tau_{1n}\frac{u_1^n - u_0^n}{h} + \tau_{2n}u_0^n &= \omega_{1n}, \\ \delta_{1n}\frac{u_M^n - u_{M-1}^n}{h} + \delta_{2n}u_M^n &= \omega_{2n}, \end{aligned}\right\} \tag{3.76}$$

$n = 0, 1, \dots, N,$

where $\tau_{1n} = \tau_1(y)$, $\tau_{2n} = \tau_2(y_n)$, and so on.

Difference condition (3.76) approximates the boundary conditions (3.75) with inaccuracy of order $O(h)$.

The numerical realization of the scheme (3.70) with the conditions (3.76) is carried over as well as numerical realization of the scheme (3.70). Firstly, using formulae of (3.71) type, we calculate the values on a zero-layer u_m^0, then values on the first-layer u_m^1. In both cases m changes in the range $0 \le m \le M$. Further, by the formula (3.72) at $n = 1$ we calculate $u_1^2,\ u_2^2,\ \ldots,\ u_{M-1}^2$.

To fulfill calculations on the second layer of values u_0^2 and u_M^2 we use difference boundary conditions (3.76) at $n = 2$. Similarly, by values u_M^1, u_M^2, $m = 0, \ldots, M$ the values $u_1^3, u_2^3, \ldots, u_{M-1}^3$ are calculated, and so on.

In conclusion, the following condition for stability of difference scheme $L_h\left(u^{(h)}\right) = f^h$ according to initial data is sufficient:

$$\frac{l^2}{h^2} \le \frac{1}{1+\varepsilon},\ \ \varepsilon > 0.$$

It is noticeable, implicit difference schemes are also viewed for equations of the hyperbolic type.

3.5. Method of finite elements (MFE)

3.5.1. General remarks

The basic idea of the finite elements method provides that any continuum value, such as temperature, pressure, etc., may be approximated by a discrete model, which is constructed on the set of piecewise continuous functions defined at the finite number of points in viewed sub-area.

Continuous value is unknown in advance, and we should find the values of this quantity at some interior points of the area.

While constructing a discrete model, it is carried out:

1. The finite number of points is fixed in the viewed area. They are called node points or **nodes**.
2. The value of continuous quantity at each node point is regarded as variable that must be defined.
3. The area of defining a continuous quantity is divided into finite number of sub areas called the **elements**. These elements have common node points and all together approximate the form of the area.
4. Continuous quantity is approximated on each element of the polynomial, which is defined with help of these quantities node values. The polynomial

is defined for each element. The polynomials are chosen so that to have the quantity continuous along the boundary of the elements. Polynomial connected with each element is called the **function of the element**.

5. Joining of the finite elements together into **ensemble**. Here the node values must be regulated so as to provide the best approximation to the real continuous distribution. This step leads to an algebraic system of linear equations relevantly to node values. This system is the model of desired continuous function.
6. Solution to the system has been found (finding out the node values).
7. Searching for value of desired quantity at any point of the area by the node values and elements functions.

The basic concept of MFE can be demonstrated by the uni-dimensional example of distribution of temperature in the pod (Fig. 9)

The continuous quantity $T(x)$ is to be considered. The area of definition is the segment OL along the axis x. There are five fixed and numbered points on the axis x. They are node points.

Let's consider the case if the values $T(x)$ are known at each node point: $T1, \ldots, T5$.

The division of area into elements can be carried out by two different methods. It is possible to limit each element by two adjoining node points, so making four elements. It is also possible to divide the area into two elements; each of these contains 3 nodes (Fig. 11).

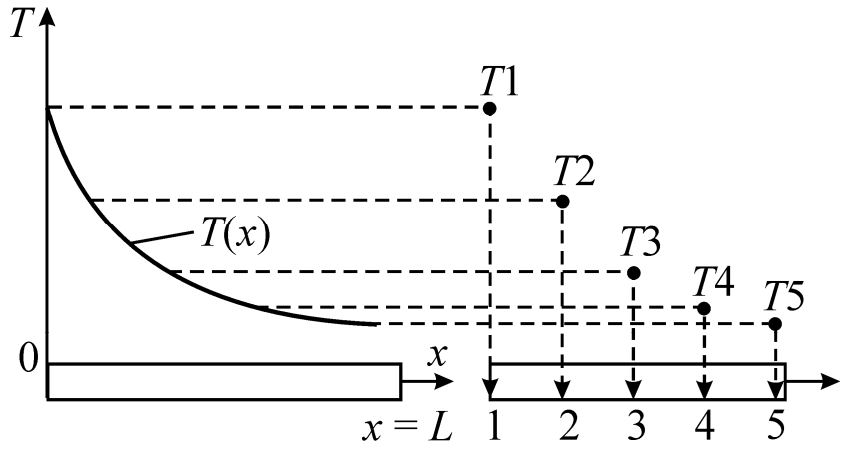


Fig. 9. Using of MFE for given distribution of temperature in uni-dimensional pod

Polynom (relevant to the element) is defined by the values $T(x)$ at node points. In case of area division into four elements (there are two nodes for each element), the function of the element is linear by x. The final approximation $T(x)$ will consist of four piecewise continuous functions (each one is defined on a separate element).

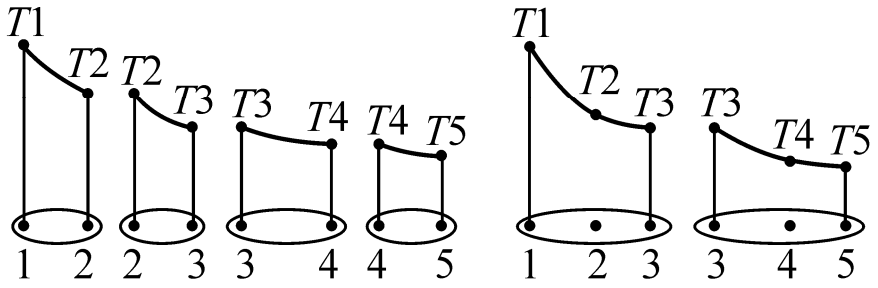


Fig. 10. Division of area into elements by two different methods

Another method of area division into two elements with three node points that gives the element function in the form of polynomial of the second order. In this case the final approximation $T(x)$ is the total of two piecewise continuous quadratic functions.

Constructing a discrete model of continuous quantity, defined in two- or three-dimensional area, the basic method of finite elements is used identically.

In two-dimensional case the elements are described by functions x, y. More often the elements are viewed in triangular or quadrangle form. Element functions are constructed as flat or curved surfaces (Fig. 11 *a*, *b*). Element functions are given as plane, if for this element minimal number of node points is given. For triangular element the number is 3, and for quadrangular element the number is 4.

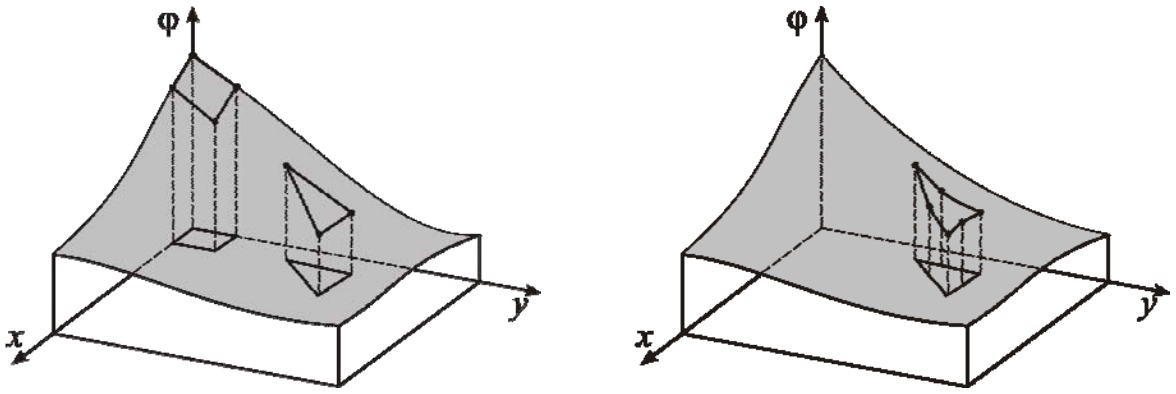


Fig. 11. Elements in a form of triangle (three nodes) and quadrangle (four nodes) (a) and in a form of triangle (six nodes) (b)

If used number of nodes is more than minimal, the element function is relevant to the curved surface.

Besides, an excessive number of nodes allows to consider elements with curved boundaries.

The final approximation of two-dimensional quantity $\varphi(x, y)$ is the set of piecewise continuous function, each of these is defined on a separate element with the help of values $\varphi(x, y)$ at the relevant node points.

3.5.2. Discretization of area and numbering of nodes

The division of unidimensional area into elements is brought to the division of a segment into shorter areas. The division of two-dimensional area is started from it's boundary to fulfill more precise approximation of boundary form. Then the division of interior area takes place. Particularly, the division of area into elements is fulfilled by several steps. At first, the area is divided into quite large subareas. The boundaries between them lie whereon there are changes of material quality, geometry, applied load, etc. Afterwards, each subarea is devided into elements. More often these elements are triangles as they are the simplest of two-dimensional elements in sense of analytical provision. The body is divided into quadrangular and triangular subareas that are subdivided into almost equilateral triangles.

Changes of finite elements in size on the boundaries of subareas are avoided.

Numbering of nodes is the next procedure at pointing out finite elements. The order of numbering is of vital importance as it has influence on the method effectiveness.

The array of algebraic linear equations, leading to MFE, is absolutely disperse matrix of a band structure. Non-zero elements of such matrix are situated in a parallel way relevantly to the principle diagonal. The whole number L, being the highest-range difference between numbers of non-zero elements in a line, is called a bandwidth. The less the width is, the fewer volume of matrix storage is needed for at fulfilling MFE on computer and the less time spending is for equation system solving.

The width of the band depends on number of orders for free nodes and on method of numbering nodes.

Number of freedom order is amount of unknown functions, defined at each node. For example, for two-dimensional problems on hydraulics at each node there are three variables to be defined: pressure and components of speed by axis x and y. On numbering nodes, it is preferably to use method, providing minimal difference between node numbers at each separate element. If we define the highest-range difference between node numbers over the whole area through R, but the number of freedom orders through Q, the bandwidth is

$$L = (R + 1) \cdot Q .$$

In some cases decreasing of R may be done by sequent numbering of node numbers at moving to the direction of smaller given area size.

At Fig. 12 there are two different methods of numbering nodes in an arbitrary area, divided into finite elements.

In the first method $R = 14$ and in the second $R = 6$.

The width of band for these methods at one node freedom order is 15 and 7, at two freedom orders it is 30 and 14 respectively. Rational numbering

in case *b* reduces the volume of Random access memory (RAM) twice in comparison to *a*.

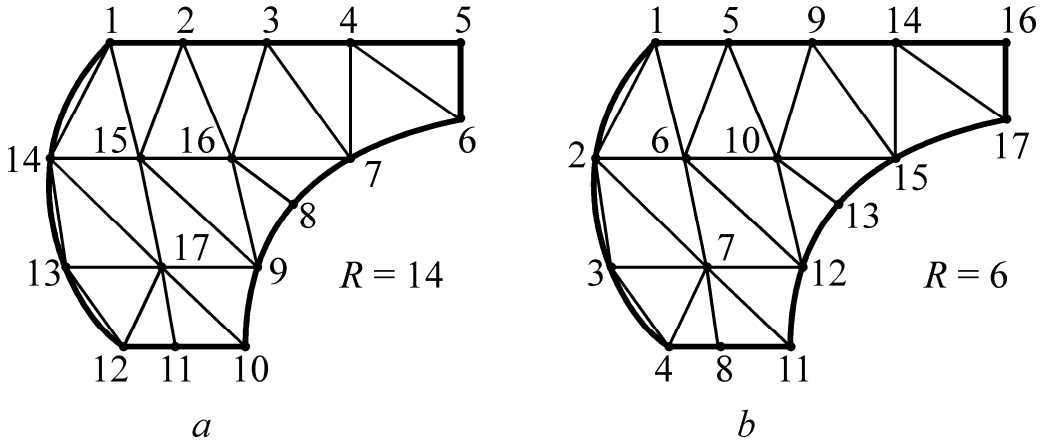


Fig. 12. Two different methods of node numbering

In MFE we number the elements too. It may be done randomly, as numbering of elements doesn't influence the calculations of the problem.

3.5.3. Linear interpolar polynomials

The classification of finite elements can be carried out in accordance with polynomials order (the function of these elements). Three groups of elements are taken:

- simplex elements,
- complex elements,
- multiplex elements.

Simplex elements are presented by first order polynomials and complex elements by higher order polynomials. In simplex element the number of nodes equals the dimension of area +1. In complex element there is a greater number of this quantity.

For multiplex elements the higher order polynomials are used. The boundaries of elements must be parallel to coordinates axis.

3.5.4. One-dimensional simplex element

One-dimensional simplex element is a straight-line segment L with two nodes, one at each end of the segment (Fig. 13). Nodes are given by indices *i* and *j*, node values are Φ_i, Φ_j.

Function of the element φ

$$\varphi = \alpha_1 + \alpha_2 x\,. \tag{3.77}$$

Coefficients α_1 and α_2 are easily to be defined.

At $x = X_i$, $\varphi = \Phi_i$, and the formula (3.77) gives $\Phi_i = \alpha_1 + \alpha_2 X_i$.

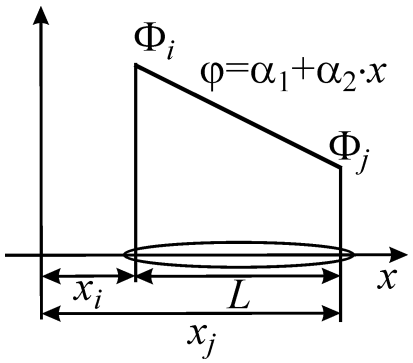


Fig. 13. One-dimensional simplex element

Identically $\Phi_j = \alpha_1 + \alpha_2 X_j$.

Solving two last equations relevantly to α_1 and α_2 we derive

$$\alpha_1 = \frac{\Phi_i \cdot X_j - \Phi_j \cdot X_i}{L}, \tag{3.78}$$

$$\alpha_2 = \frac{\Phi_j - \Phi_i}{L}. \tag{3.79}$$

Substituting the values α_1 and α_2 into the formula (3.77), we get expression for φ:

$$\varphi = \frac{\Phi_i \cdot X_j - \Phi_j \cdot X_i}{L} + \frac{\Phi_j - \Phi_i}{L} \cdot x, \tag{3.80}$$

that can be rewritten as

$$\varphi = \left(\frac{X_j - x}{L} \right) \cdot \Phi_i + \left(\frac{x - X_i}{L} \right) \cdot \Phi_j. \tag{3.80^a}$$

Linear functions from x in the formula (3.80) are called form functions or interpolar functions. We express them through N_i and N_j:

$$N_i = \frac{X_j - x}{L}, \; N_j = \frac{x - X_i}{L}. \tag{3.81}$$

Here indices i and j at N define the node, which form function refers to. Now the ratio (3.80) can be written in matrix form

$$\varphi = N_i \Phi_i + N_j \Phi_j = \left[N \right] \{ \Phi \}, \tag{3.82}$$

where $\left[N \right] = [N_i, N_j]$ is a matrix line and $\left\{ \Phi \right\} = \begin{Bmatrix} \Phi_i \\ \Phi_j \end{Bmatrix}$ is a column vector.

It is seen from the inequality (3.81) that the function N_i equals 1 at node of i-number and equals zero at node j-number. Identically, the function N_j equals zero i-number node and equals 1 at j-number node.

These values are particular for form function. They equal 1 at one definite node and equal zero at the rest of nodes.

3.5.5. Two-dimensional simplex element

Two-dimensional simplex element is demonstrated in the Fig. 14.

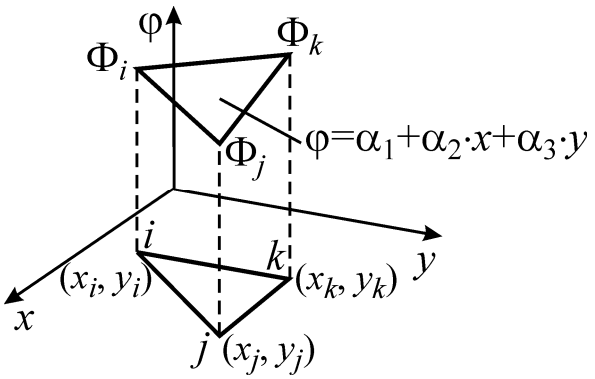


Fig. 14. Two-dimensional simplex element

Let's numerate nodes counter-clockwise. The interpolar polynomial is

$$\varphi = \alpha_1 + \alpha_2 \cdot x + \alpha_3 \cdot y\,. \tag{3.83}$$

Conditions at nodes i, j, k leads to the combined equations

$$\left.\begin{aligned} \Phi_i &= \alpha_1 + \alpha_2 \cdot X_i + \alpha_3 \cdot Y_i\ , \\ \Phi_j &= \alpha_1 + \alpha_2 \cdot X_j + \alpha_3 \cdot Y_j, \\ \Phi_k &= \alpha_1 + \alpha_2 \cdot X_k + \alpha_3 \cdot Y_k. \end{aligned}\right\} \tag{3.84}$$

The solution of this system is

$$\alpha_1 = \frac{1}{2\cdot A}[\left(X_j \cdot Y_k - X_k \cdot Y_j\right)\cdot \Phi_i + \left(X_k \cdot Y_i - X_i \cdot Y_k\right)\cdot \Phi_j +$$
$$\left(X_i \cdot Y_j - X_j \cdot Y_i\right)\cdot \Phi_k],$$
$$\alpha_2 = \frac{1}{2\cdot A}\left[\left(Y_j - Y_k\right)\cdot \Phi_i + \left(Y_k - Y_i\right)\cdot \Phi_j + \left(Y_i - Y_j\right)\cdot \Phi_k\right],$$
$$\alpha_3 = \frac{1}{2\cdot A}\left[\left(X_k - X_j\right)\cdot \Phi_i + \left(X_i - X_k\right)\cdot \Phi_j + \left(X_j - X_i\right)\cdot \Phi_k\right].$$

Here A is an area of the triangle ijk connected with the determinant of the system (3.84) in the following way:

$$2 \cdot A = \begin{vmatrix} 1 & X_i & Y_i \\ 1 & X_j & Y_j \\ 1 & X_k & Y_k \end{vmatrix}. \tag{3.85}$$

Substituting the values α_1, α_2, α_3 into the formula (3.83), we transform the expression for φ to make it similar to the Eq. (3.82)

$$\varphi = N_i \cdot \Phi_i + N_j \cdot \Phi_j + N_k \cdot \Phi_k, \tag{3.86}$$

where

$$N_i = \frac{1}{2 \cdot A}\left[a_i + b_i \cdot x + c_i \cdot y\right], \begin{cases} a_i = X_j \cdot Y_k - X_k \cdot Y_j, \\ b_i = Y_j - Y_k, \\ c_i = X_k - X_j, \end{cases} \tag{3.86^a}$$

$$N_j = \frac{1}{2 \cdot A}\left[a_j + b_j \cdot x + c_j \cdot y\right], \begin{cases} a_j = X_k \cdot Y_i - X_i \cdot Y_k, \\ b_j = Y_k - Y_i, \\ c_j = X_i - X_k, \end{cases} \tag{3.86^b}$$

$$N_k = \frac{1}{2 \cdot A}\left[a_k + b_k \cdot x + c_k \cdot y\right], \begin{cases} a_k = X_i \cdot Y_j - X_j \cdot Y_i, \\ b_k = Y_i - Y_j, \\ c_k = X_j - X_i, \end{cases} \tag{3.86^c}$$

It is easy to demonstrate that the value N_i at j-number node equals to 1, $N_i = 0$ at the second and third nodes, and at other points of the line constructed through these nodes.

3.5.6. Local system of coordinates

Getting of the system for node values of unknown quantities includes integrating of form function elements or their partial derivatives over the square. Integrating can be simplified, if we write interpolar ratios in the coordinate system, connected with the elements. This is the local system.

Let's view a triangular element, where the scalar quantity is as follows:

$$\varphi = N_i \Phi_i + N_j \Phi_j + N_k \Phi_k,$$

The form function is defined by the formulae (3.86^a), (3.86^b), (3.86^c).

Let's mark the beginning of the local system in the centre of the element (Fig. 15).

Let's write formulae of coordinate transformation:

$$\left.\begin{aligned} x &= \overline{X} + s, \\ y &= \overline{Y} + t. \end{aligned}\right\} \tag{3.87}$$

Here $\overline{X}$ and $\overline{Y}$ are coordinates of the triangle centre:

$$\overline{X} = \frac{X_i + X_j + X_k}{3},$$
$$\overline{Y} = \frac{Y_i + Y_j + Y_k}{3}. \tag{3.88}$$

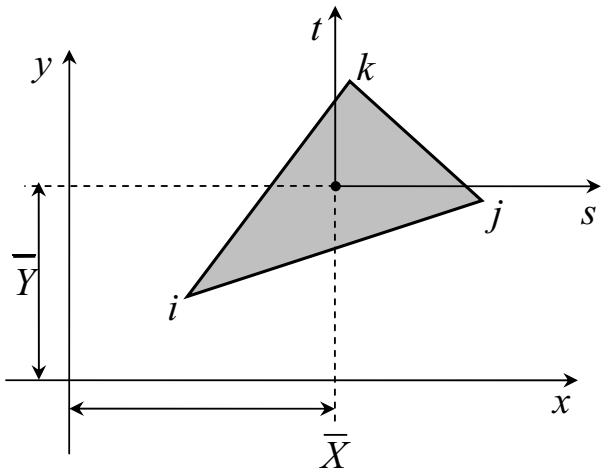


Fig. 15. Local coordinate system

Form function N_i after inserting (3.87) becomes

$$N_i = \frac{1}{2 \cdot A}\left[\left(a_i + b_i \cdot \overline{X} + c_i \cdot \overline{Y}\right) + b_i \cdot s + c_i \cdot t\right]. \tag{3.89}$$

Taking in consideration (3.86[a]) and the ratio (3.88), we can derive

$$a_i + b_i \cdot \overline{X} + c_i \cdot \overline{Y} = \frac{2 \cdot A}{3}.$$

So form function in the local system becomes

$$N_i = \frac{1}{2 \cdot A}\left[\frac{2 \cdot A}{3} + (Y_j - Y_k) \cdot s + (X_k - X_j) \cdot t\right]. \tag{3.90a}$$

We obtain identically

$$N_j = \frac{1}{2 \cdot A}\left[\frac{2 \cdot A}{3} + (Y_i - Y_k) \cdot s + (X_k - X_i) \cdot t\right], \tag{3.90b}$$

$$N_k = \frac{1}{2 \cdot A}\left[\frac{2 \cdot A}{3} + (Y_i - Y_j) \cdot s + (X_j - X_i) \cdot t\right]. \tag{3.90c}$$

Integral from a function given in the global coordinate system can be calculated in the local coordinate system with the help of ratio

$$\int_R f(x,y)dx\,dy = \int_{R^*} f\,\left[x\,(s,t), y\,(s,t)\right] \left| J \right|\, ds\,dt\,, \tag{3.91}$$

where R and R^* are old and new areas of integrating relevantly, $|J|$ is Jacobean module of coordinate system transforming, that equals the ratio of two squares in two systems of coordinates A_{xy}/A_{st}. As two systems are rectangular and the sizes of measurement coincide in there, it gives $|J| = 1$. Besides, forms of elements R and R^* remain.

So, the ratio (3.91) becomes

$$\int_R f(x,y)dxdy = \int_{R^*} f\ [x(s,t),\ y(s,t)\ ]\ dsdt\,. \tag{3.92}$$

The function $f(x, y)$ in the left part of this equality is a form function, expressed in the global coordinate system, whereas $f[x(s,t),\ y(s,t)]$ comply with element form function, given in the local system.

3.5.7. Two-dimensional L-coordinates

For triangular element the most frequent form is a coordinate system defined by three coordinates: L_1, L_2, L_3 (Fig. 16).

Each coordinate is a ratio of distance from the chosen point of the triangle to one of its sides s and height h, dropped to one side from the opposite top. The coordinates L_i change in the range of 0 to 1. The coordinates L_1, L_2, L_3 are called L-coordinates. Their values give relevant quantities of triangle squares, into which the element is divided (Fig. 17).

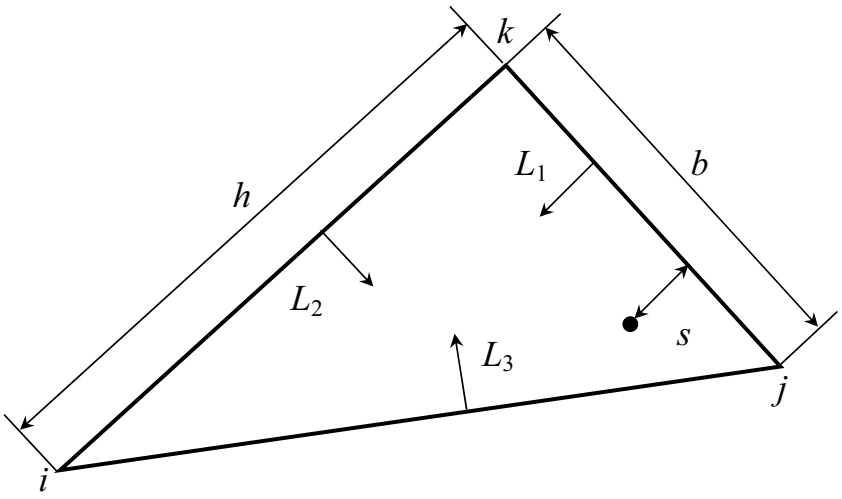


Fig. 16. L-coordinates for a triangle element

L-coordinates of point B are squares of triangles in the Fig. 17. The square A_t of the triangle ijk is given as

$$A_t = \frac{b \cdot h}{2}. \tag{3.93}$$

The square A_1 of the hatched triangle (Bjk) equals

$$A_1 = \frac{b \cdot s}{2}. \tag{3.94}$$

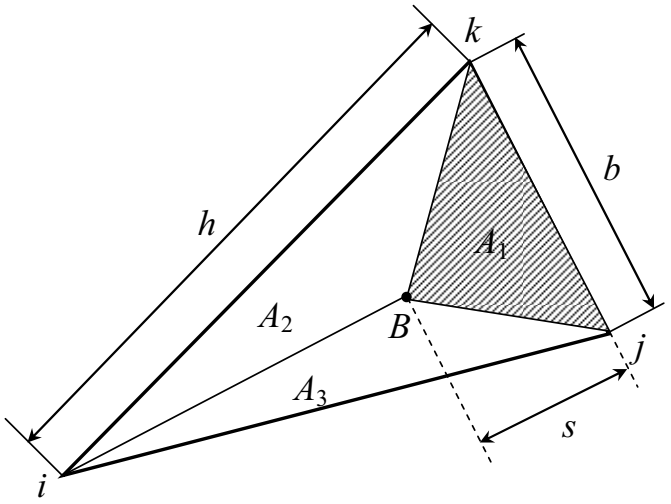


Fig. 17. Geometrical interpretation of L-coordinates

Therefore,

$$\frac{A_1}{A_t} = \frac{s}{h} = L_1. \tag{3.95}$$

Identically

$$L_2 = \frac{A_2}{A_t};\ L_3 = \frac{A_3}{A_t}. \tag{3.96}$$

As $A_1 + A_2 + A_3 = A_t$, we have

$$L_1 + L_2 + L_3 = 1. \tag{3.97}$$

The coordinates L_1, L_2, L_3 are form functions for the simplex element of the triangle:

$$N_i = L_1, \quad N_j = L_2, \quad N_k = L_3. \tag{3.98}$$

As it goes from Fig. 17

$$L_1 = \begin{cases} 1 & \text{at the nodes with the number } i, \\ 0 & \text{at the nodes } j \text{ and } k. \end{cases}$$

Such ratios are the same for L_2 and L_3. Besides, the formula (3.97) states that at arbitrary point of function element the forms sum is always equal to 1.

The advantage of *L*-coordinates is existence of integral formulae, which simplify calculation of integrals along the sides of the element sides and its square:

$$\int_L L_1^a L_2^b dL = \frac{a!b!}{(a+b+1)!} \cdot L \tag{3.99}$$

(L is a distance between two nodes of a taken side);

$$\int_A L_1^a L_2^b L_3^c dA = \frac{a!b!c!}{(a+b+c+2)!} \cdot 2A. \qquad (3.100)$$

Using the ratio (3.100) can be shown while calculating the integral

$$\int_A N_i \cdot N_j dA,$$

where N_i and N_j are functions of x and y. The integral over square of the element is transformed as

$$\int_A N_i \cdot N_j dA = \int_A L_1^1 \cdot L_2^1 \cdot L_3^0 dA = \frac{1!1!0!}{(1+1+0+2)!} \cdot 2A = \frac{2A}{4!} = \frac{A}{12}.$$

3.5.8. Aggregation of elements into ensemble

Interpolar polynomial for each element is

$$\varphi^{(e)} = [N^{(e)}] \cdot \{\Phi\} = N_i^{(e)} \cdot \Phi_i + N_j^{(e)} \cdot \Phi_j + N_k^{(e)} \cdot \Phi_k, \qquad (3.101)$$

where index (e) presents an arbitrary element.

The method of including element into area can be demonstrated by an example of simple five-element configuration (Fig. 18).

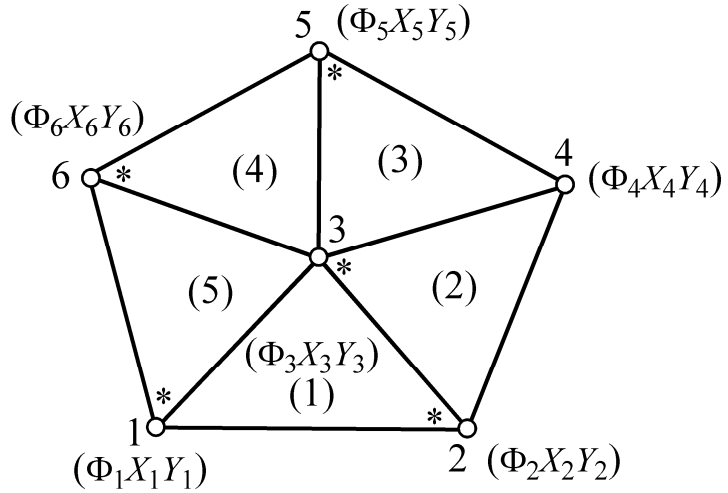


Fig. 18. Five-element configuration

Nodes are numbered from one to six. Quantities – Φ_1, Φ_2, Φ_3, Φ_4, Φ_5, Φ_6 – present global orders of freedom. Coordinate of nodes (X_β, Y_β), $\beta = 1, \ldots, 6$ are supposed to be known. Numbers of element are given in the round brackets.

To number element nodes the indexes i, j, k given above can be used as soon as the first node is defined in each element. At the Fig. 18 i-node in each element is markes by an asterisk (*).

Fixing of node *i* allows to write following equalities for different elements: element one:

$$i = 2, j = 3, k = 1; \qquad (3.102^{a})$$

element two:

$$i = 3, j = 2, k = 4; \qquad (3.102^{b})$$

element three:

$$i = 5, j = 3, k = 4; \qquad (3.102^{c})$$

element four:

$$i = 6, j = 3, k = 5; \qquad (3.102^{d})$$

element five:

$$i = 1, j = 3, k = 6. \qquad (3.102^{e})$$

These ratios allow to include the element into the area, as they put element indexes *i*, *j*, *k* to the global numbers of nodes relevantly. This process fixes coordinates of element nodes.

Values of the indexes *i*, *j*, *k* can be put into the formula (3.101). This gives the following equations:

$$\left.\begin{aligned}
\varphi^{(1)} &= N_2^{(1)} \cdot \Phi_2 + N_3^{(1)} \cdot \Phi_3 + N_1^{(1)} \cdot \Phi_1; \\
\varphi^{(2)} &= N_3^{(2)} \cdot \Phi_3 + N_2^{(2)} \cdot \Phi_2 + N_4^{(2)} \cdot \Phi_4; \\
\varphi^{(3)} &= N_5^{(3)} \cdot \Phi_5 + N_3^{(3)} \cdot \Phi_3 + N_4^{(3)} \cdot \Phi_4; \\
\varphi^{(4)} &= N_6^{(4)} \cdot \Phi_6 + N_3^{(4)} \cdot \Phi_3 + N_5^{(4)} \cdot \Phi_5; \\
\varphi^{(5)} &= N_1^{(5)} \cdot \Phi_1 + N_3^{(5)} \cdot \Phi_3 + N_6^{(5)} \cdot \Phi_6.
\end{aligned}\right\} \qquad (3.103)$$

Form functions are the multipliers at nodes values in the formula (3.103) and they are defined by putting numerical values of *i*, *j*, *k* into equations of form functions.

So function $N_k^{(e)}$ is given as

$$\left.\begin{aligned}
&N_k^{(e)} = \frac{1}{2A^{(e)}}\left[a_k^{(e)} + b_k^{(e)} \cdot x + c_k^{(e)} \cdot y\right], \\
&\text{where} \\
&a_n^{(e)} = X_i \cdot Y_j - X_j \cdot Y_i, \\
&b_k^{(e)} = Y_i - Y_j, \\
&c_k^{(e)} = X_j - X_i.
\end{aligned}\right\} \qquad (3.104)$$

For the 5th element $i = 1, j = 3, k = 6$ that gives

$$\left.\begin{aligned} N_6^{(5)} &= \frac{1}{2A^{(5)}}\left[a_6^{(5)} + b_6^{(5)} \cdot x + c_6^{(5)} \cdot y\right], \\ a_6^{(5)} &= X_1 \cdot Y_3 - X_3 \cdot Y_1, \\ b_6^{(5)} &= Y_1 - Y_3, \\ c_6^{(5)} &= X_3 - X_1. \end{aligned}\right\} \qquad (3.105)$$

Functions of form $N_6^{(4)}$ and $N_6^{(5)}$ in (3.103) are different values, even if the quatities of $A^{(4)}$ and $A^{(5)}$ are equal.

There are constants of the expression $N_6^{(4)}$:

$$a_6^{(4)} = X_3 Y_5 - X_5 Y_3,$$
$$b_6^{(4)} = Y_3 - Y_5,$$
$$c_6^{(4)} = X_5 - X_3.$$

It makes clear that $N_6^4 \neq N_6^{(5)}$.

With the help of the equalities (3.103) the finite elements are united into an ensemble, and interpolar functions are expressed through global node values and global coordinates, which are introduced instead of arbitrary i, j, k.

3.5.9. Finding the equations for element with the help of Galerkin's method

If we take differential equations

$$Lu - f = 0$$

and approximate solution is to be found as

$$\bar{u} = \sum N_i u_i,$$

we will have

$$L\bar{u} - f = \varepsilon,$$

where ε is an error, as the solution $\bar{u}$ is approximate.

It is necessary to make ε a small quantity. In Galerkin's method it is done with the help of ratios of orthogonality for each basic function N_i:

$$\int_R N_i \cdot \varepsilon \, dR = 0.$$

This equality means that basic functions must be orthogonal to the error by area R.

The usage of this method in the framework of MFE gives the following equations:

$$\int_R N_в \cdot L(\varphi)\, dR = 0, \quad \beta = i, j, k, \ \ldots\ , \tag{3.106}$$

where φ is a desired quantity, which is approximated by the expression

$$\varphi = [N_i, N_j, N_k, \ldots]\{\Phi\}, \tag{3.107}$$

and $L(\varphi)$ is the left part of the differential equation $L(\varphi) = 0$, which is to be solved.

3.5.10. Example. Calculation of one-dimensional temperature field in a homogeneous rod

Supposing there is a rod with length L and square of cross-section S. One end of the rod is fixed and it is provided with heat flow q of given intensity (Fig. 19).

At the free end of the rod there is convection heat exchange with the environment.

Coefficient of heat exchange is α, and temperature of environment is T_0.

Along the side surface the rod is heat-insulted.

The temperature field in the rod is given by the equation of heat conductivity:

$$\lambda \cdot \frac{d^2T}{dx^2} = 0\,. \tag{3.108}$$

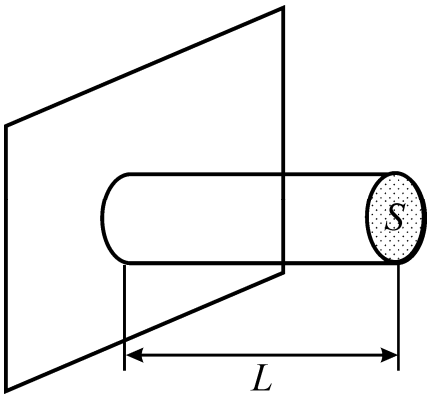


Fig. 19. Illustration to the example (3.5.10)

The edge conditions are

$$\lambda \cdot \frac{dT}{dx} + q = 0 \ \text{ for } x = 0, \tag{3.109^a}$$

$$\lambda \cdot \frac{dT}{dx} + \alpha \cdot \left(T - T_0\right) = 0 \ \text{ for } x = L. \tag{3.109^b}$$

Here λ is the coefficient of heat-conductivity, and α is the coefficient of heat transmission.

Let divide the rod into two finite elements and define the length of each through $L^{(e)}$, $e = 1, 2$.

Having used Galerkin's method to the equation (3.108), we derive

$$\int_V [N]^T \lambda \cdot \frac{d^2T}{dx^2}\, dV = 0, \tag{3.110}$$

where $[N]^T$ is a column vector, got by transposing the line $[N]$ from form function of one-dimensional simplex element (3.81).

Let's put into the formula (3.110) a formula for differentiation of the multiplication:

$$\int_V [N]^T \lambda \cdot \frac{d^2T}{dx^2} dV =$$

$$\int_V \frac{d}{dx}\left([N]^T \lambda \cdot \frac{dT}{dx} \right) dV - \int_V \frac{d}{dx}\left([N]^T \right) \lambda \cdot \frac{dT}{dx} dV. \tag{3.111}$$

Interpolar function T is piecewise-linear one, therefore integrals in the Eq. (3.111) can be given as the sum of corresponding integrals for separate elements.

So the 2nd integral in the Eq. (3.111) can be presented as

$$\int_V \frac{d[N]^T}{dx} \lambda \frac{dT}{dx} dV = \sum_{e=1}^{2} \int_{V^{(e)}} \frac{d\left[N^{(e)}\right]^T}{dx} \lambda^{(e)} \frac{dT^{(e)}}{dx} dV^{(e)}. \tag{3.112}$$

Let's calculate the integrals in (3.112), referring to the separate elements

$$\frac{d\left[N^{(e)}\right]^T}{dx} = \frac{d}{dx} \begin{bmatrix} \dfrac{X_j - x}{L^{(e)}} \\ \dfrac{x - X_i}{L^{(e)}} \end{bmatrix} = \frac{1}{L^{(e)}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \tag{3.113}$$

$$\frac{dT^{(e)}}{dx} = \frac{d}{dx}\left(\left[N^{(e)} \right] \cdot \{T^{(e)}\} \right) = \frac{1}{L^{(e)}} [-1,\ 1] \begin{Bmatrix} T_i \\ T_j \end{Bmatrix}. \tag{3.114}$$

We have

$$\int_{V^{(e)}} \frac{d\left[N^{(e)}\right]}{dx} \lambda^{(e)} \frac{dT^{(e)}}{dx} dV^{(e)} = \int_{V^{(e)}} \frac{\lambda^{(e)}}{L^{(e)} L^{(e)}} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} T_i \\ T_j \end{Bmatrix} dV^{(e)} =$$

$$= \frac{S^{(e)} \lambda^{(e)}}{L^{(e)}} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} T_i \\ T_j \end{Bmatrix}. \tag{3.115}$$

The 1st integral on the basis of Ostrogradsky–Gauss theorem is transformed as

$$\int_V \frac{d}{dx}\left([N]^T \cdot \lambda \cdot \frac{dT}{dx} \right) dV = \int_S [N]^T \cdot \lambda \cdot \frac{dT}{dx} \cdot l_x dS, \qquad (3.116)$$

where $l_x \dfrac{dT}{dx} = \dfrac{dT}{dn}$, and n is an exterior normal to the viewed surface.

Under the edge conditions (3.109[a]) at the point $x = 0$ at the 1st element the integral (3.116) will be the following:

$$\int_S \left[N^{(1)} \right]^T \lambda^{(1)} \frac{dT^{(1)}}{dn} dS =$$

$$= \int_S \begin{bmatrix} \dfrac{x_2}{L^{(1)}} \\ -\dfrac{x_1}{L^{(1)}} \end{bmatrix} (-q)\, dS = \int_S \begin{bmatrix} -q \\ 0 \end{bmatrix} dS = \begin{bmatrix} -qS_1 \\ 0 \end{bmatrix}. \qquad (3.117)$$

Under the edge conditions (3.109[b]) at the point $x = L$ for the 2nd element the integral (3.111) will be the following

$$\int_S \left[N^{(2)} \right]^T \lambda^{(2)} \cdot \frac{dT^{(2)}}{dx} dS = \int_S \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot \left[-\alpha \left(T_3 - T_0 \right) \right] dS =$$

$$= S_3 \begin{bmatrix} 0 \\ 1 \end{bmatrix} (-\alpha T_3 + \alpha T_0). \qquad (3.118)$$

Here S_1, S_3 are the left and right cross-sections of the rod.

Taking into consideration the fact there are matrixes under the integrals (3.112), (3.117), (3.118), we are to find that at summation the lines of these matrixes must be summed, which respond to the same nodes.

Having summed the expression (3.115) for the 1st and the 2nd elements and the expressions (3.117), (3.118), and having equated the sum to zero, we get the combined equations

$$\begin{bmatrix} -1 & -c_1 & 0 \\ -c_1 & c_1 + c_2 & -c_2 \\ 0 & -c_2 & c_2 + \alpha \cdot S_3 \end{bmatrix} \cdot \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} q \cdot S_1 \\ 0 \\ -\alpha \cdot S_3 \cdot T_0 \end{bmatrix} = 0. \qquad (3.119)$$

Here we have $c_1 = \dfrac{S^{(1)} \cdot \lambda^{(1)}}{L^{(1)}}$, $c_2 = \dfrac{S^{(2)} \cdot \lambda^{(2)}}{L^{(2)}}$.

The system (3.119) defines the node values T_1, T_2, T_3 at the final procedure MFE is the solution of algebraic linear equation system

3.5.11. Two-dimensional equations of the field theory

The range of problems on physics and engineering are described by the equation

$$L(\varphi) = \lambda \cdot \left(\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} \right) + Q = 0 . \qquad (3.120)$$

In particular, the equation (3.120) is used to solve problems about liquid current, heat-transmission, torsion of tight pod, search for electrostatic field, etc.

Let's use Galerkin's method to solve to this equation in the problem on heat transmission.

In this case λ is the coefficient of heat conductivity, φ is temperature, Q is the source of heat inside the body ($Q > 0$, if the hit is transmitted to the body). For example, in some part of the boundary there is convection heat exchange that gives the edge condition. Here h is the coefficient of heat transmission:

$$\lambda \frac{\partial \varphi}{\partial n} + h \cdot (T - T_0) = 0 . \qquad (3.120^{a})$$

The insertion of (3.120) into (3.106) gives

$$\int_V [N]^T \cdot \left(\lambda \cdot \frac{\partial^2 \varphi}{\partial x^2} + \lambda \cdot \frac{\partial^2 \varphi}{\partial y^2} + Q \right) dV = 0 . \qquad (3.121)$$

First of all, it is needed to transform the equation (3.121): the way it could contain only the 1st derivatives.

Using the multiplication differentiation formulae we find

$$[N]^T \frac{\partial^2 \varphi}{\partial x^2} = \frac{\partial}{\partial x} \left([N]^T \frac{\partial \varphi}{\partial x} \right) - \frac{\partial}{\partial x} [N]^T \cdot \frac{\partial \varphi}{\partial x} . \qquad (3.122)$$

The 1st item in integral (3.121) gives

$$\int_V [N]^T \frac{\partial^2 \varphi}{\partial x^2} dV = \int_V \frac{\partial}{\partial x} \left([N]^T \frac{\partial \varphi}{\partial x} \right) - \int_V \frac{\partial}{\partial x} [N]^T \frac{\partial \varphi}{\partial x} dV . \qquad (3.123)$$

By Ostrogradsky–Gauss theorem we have

$$\int_V \frac{\partial}{\partial x} \left([N]^T \frac{\partial \varphi}{\partial x} \right) dV = \int_S [N]^T \frac{\partial \varphi}{\partial x} \cdot l_x dS . \qquad (3.124)$$

Identically with the member containing $\frac{\partial^2 \varphi}{\partial y^2}$ and supposing that $dV = tdA, \quad dS = tdL$, the equation (3.121) becomes

$$\int_L [N]^T \lambda \left(\frac{\partial \varphi}{\partial x} l_x + \frac{\partial \varphi}{\partial y} l_y \right) dL -$$

$$- \int_A \lambda \left(\frac{\partial [N]^T}{d x} \cdot \frac{\partial \varphi}{\partial x} + \frac{\partial [N]^T}{\partial y} \cdot \frac{\partial \varphi}{\partial y} - [N]^T Q \right) dA = 0. \qquad (3.125)$$

The thickness of the element is $t = 1$. The surface integral (3.125) can be expressed through the derivative $\frac{\partial \varphi}{\partial n}$, as a result we have

$$\int_A \lambda \left(\frac{\partial [N]^T}{\partial x} \cdot \frac{\partial \varphi}{\partial x} + \frac{\partial [N]^T}{\partial y} \cdot \frac{\partial \varphi}{\partial y} \right) dA -$$

$$- \int_A [N]^T Q \cdot dA - \int_L \lambda [N]^T \cdot \frac{\partial \varphi}{\partial n} \cdot dL = 0. \qquad (3.126)$$

Evidently, if $\frac{\partial \varphi}{\partial n}$ becomes zero on the boundary, the 3rd integral disappears.

Interpolar function φ is piecewise linear, therefore the integrals (3.126) can be given as a summation of relevant integrals for separate elements:

$$\sum_{e=1}^{R} \left\{ \lambda \int_{A^{(e)}} \left(\frac{\partial \left[N^{(e)} \right]^T}{\partial x} \cdot \frac{\partial \varphi^{(e)}}{\partial x} + \frac{\partial \left[N^{(e)} \right]^T}{\partial y} \cdot \frac{\partial \varphi^{(e)}}{\partial y} \right) dA \; - \right.$$

$$\left. - \int_{A^{(e)}} \left[N^{(e)} \right]^T Q^{(e)} dA - \int_{L^{(e)}} \lambda \left[N^{(e)} \right]^T \frac{\partial \varphi^{(e)}}{\partial n} dL \right\} = 0. \qquad (3.127)$$

Unknown functions $\varphi^{(e)}$ in the equation (3.127) are defined by the ratio

$$\varphi^{(e)} = [N^{(e)}] \{\Phi\}. \qquad (3.128)$$

Let's find the integrals for separate elements. Meanwhile, we are going to drop the upper index (e) in all symbols of element matrixes except the case when it is necessary to distinguish two elements.

Let's view the 1st integral in (3.127). Taking (3.128) we derive

$$I = \lambda \int_A \left(\frac{\partial [N]^T}{\partial x} \cdot \frac{\partial \varphi}{\partial x} + \frac{\partial [N]^T}{\partial y} \cdot \frac{\partial \varphi}{\partial y} \right) dA =$$

$$= \int_A \lambda \left(\frac{\partial [N]^T}{\partial x} \cdot \frac{\partial [N]}{\partial x} + \frac{\partial [N]^T}{\partial y} \cdot \frac{\partial [N]}{\partial y} \right) dA \cdot \{\Phi\}. \qquad (3.129)$$

Introducing symbols

$$[\mathrm{B}] = \begin{bmatrix} \dfrac{\partial [N]}{\partial x} \\ \\ \dfrac{\partial [N]}{\partial y} \end{bmatrix}, \tag{3.130}$$

we rewrite this integral as

$$I = \int_A \lambda \cdot [B]^T \cdot [B] dA \cdot \{\Phi\}. \tag{3.131}$$

The unknown function φ in the equation (3.131) is defined by the formula

$$\varphi = [N]\{\Phi\},$$

therefore,

$$\frac{\partial \varphi}{\partial x} = \frac{\partial}{\partial x}\left([N]\{\Phi\}\right), \ \frac{\partial \varphi}{\partial y} = \frac{\partial}{\partial y}\left([N]\{\Phi\}\right).$$

Inserting this formulae into the 1th intergral in (3.126), we have

$$\int_A \left(\frac{\partial [N]^T}{\partial x} \cdot \frac{\partial [N]}{\partial x} + \frac{\partial [N]^T}{\partial y} \cdot \frac{\partial [N]}{\partial y} \right) dA \cdot \{\Phi\}. \tag{3.132}$$

Let's take the 3rd integral into equation.

Supposing we have the range of triangular elements along the vertical boundary (Fig. 20) along this boundary:

$$\int_L [N]^T \frac{\partial \varphi}{\partial n} dL. \tag{3.133}$$

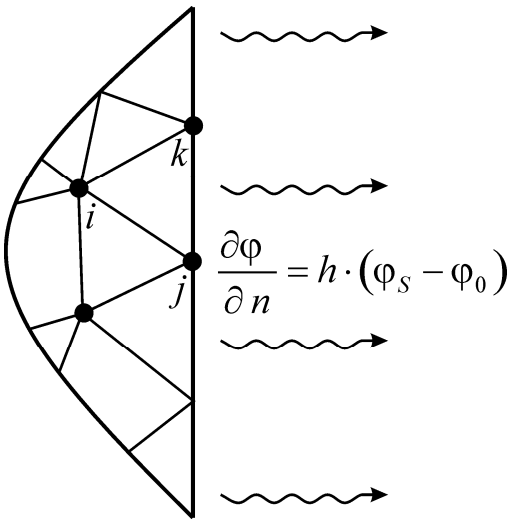


Fig. 20. Referents to the derivation of the 3rd integral in the equation (3.126)

The thermal current along the boundary is relevant to the heat loss, caused by convection heat exchange, and presented by the value

$$\frac{\partial \varphi}{\partial n} = h \cdot \left(\varphi_S - \varphi_0\right), \tag{3.134}$$

where φ_S is boundary temperature of body, and φ_0 is temperature of environment.

Temperature inside the element

$$\varphi = N_i \cdot \Phi_i + N_j \cdot \Phi_j + N_k \cdot \Phi_k, \tag{3.135}$$

and we have four points of the surface at L-coordinates

$$\varphi_s = 0 \cdot \Phi_i + L_2 \cdot \Phi_j + L_3 \cdot \Phi_k, \tag{3.136}$$

as far as along the taken boundary $L_1 = 0$

For heat flow we get

$$\frac{\partial \varphi}{\partial n} = h \cdot \left(\varphi_S - \varphi_0\right) = h \cdot \begin{bmatrix} 0 & L_2 & L_3 \end{bmatrix} \cdot \begin{Bmatrix} \Phi_i \\ \Phi_j \\ \Phi_k \end{Bmatrix} - h \cdot \varphi_0. \tag{3.137}$$

Inserting (3.137) into (3.133) gives

$$\int_L [N]^T \frac{\partial \varphi}{\partial n} dL = h \cdot \int_L [N]^T \cdot [N] \cdot \{\Phi\} dL - \int_L [N]^T \cdot h \cdot \varphi_0 dL, \tag{3.138}$$

where $[N] = [0\ L_2\ L_3]$. Let's fulfill integration with L-coordinates, taking so as form function for linear triangle element really is as follows:

$$N_\beta = \frac{1}{2 \cdot A}\left(a_\beta + b_\beta \cdot x + c_\beta \cdot y\right), \quad \beta = i, j, k. \tag{3.139}$$

Then

$$\frac{\partial [N]}{\partial x} = \frac{1}{2 \cdot A} \cdot \begin{bmatrix} b_i, b_j, b_k \end{bmatrix}, \quad \frac{\partial [N]}{\partial y} = \frac{1}{2 \cdot A} \cdot \begin{bmatrix} c_i, c_j, c_k \end{bmatrix},$$

that gives

$$[B] = \frac{1}{2 \cdot A} \begin{bmatrix} b_i & b_j & b_k \\ c_i & c_j & c_k \end{bmatrix}. \tag{3.140}$$

Inserting (1.140) into (3.138), we find

$$\begin{aligned} I &= \int_A \frac{\lambda}{4 \cdot A^2} \cdot \begin{bmatrix} b_i & c_i \\ b_j & c_j \\ b_k & c_k \end{bmatrix} \cdot \begin{bmatrix} b_i & b_j & b_k \\ c_i & c_j & c_k \end{bmatrix} dA \{\Phi\} = \\ &= \frac{\lambda}{4 \cdot A} \left\{ \begin{bmatrix} b_i b_i & b_i b_j & b_i b_k \\ b_j b_i & b_j b_j & b_j b_k \\ b_k b_i & b_k b_j & b_k b_k \end{bmatrix} + \begin{bmatrix} c_i c_i & c_i c_j & c_i c_k \\ c_j c_i & c_j c_j & c_j c_k \\ c_k c_i & c_k c_j & c_k c_k \end{bmatrix} \right\} \{\Phi\}. \end{aligned} \tag{3.141}$$

Let's view the 3rd integral in (3.137)

$$I_k = \int_L \lambda \left[N\right]^T \frac{\partial \varphi}{\partial n} dL. \tag{3.142}$$

This integral is different from zero on the side of the triangle, on which $\frac{\partial \varphi}{\partial n} \neq 0$. The same thing is on the boundary side, which heat loss goes through, caused by the convection heat exchange. This current is presented by the quantity $h\left(\varphi - \varphi_0\right)$ on the boundary side

$$\frac{\partial \varphi}{\partial n} = h \cdot \left(\varphi - \varphi_0\right). \tag{3.143}$$

Taking (3.143) for the integral (3.142), we have

$$I_k = \int_L h\left[N^T\right]\left(\varphi - \varphi_0\right) dL = \int_L h\left[N^T\right]\left(\left[N \right]\{\Phi\} - \varphi_0\right) dL, \tag{3.144}$$

where integral is calculated along the side of the triangle, which the heat exchange takes place (Fig. 21).

Firstly, let's calculate the integral:

$$\int_L h \cdot \left[N\right]^T \cdot \left[N\right] dL = h \cdot \int_L \begin{bmatrix} N_i \cdot N_i & N_i \cdot N_j & N_i \cdot N_k \\ N_j \cdot N_i & N_j \cdot N_j & N_j \cdot N_k \\ N_k \cdot N_i & N_k \cdot N_j & N_k \cdot N_k \end{bmatrix} dL. \tag{3.145}$$

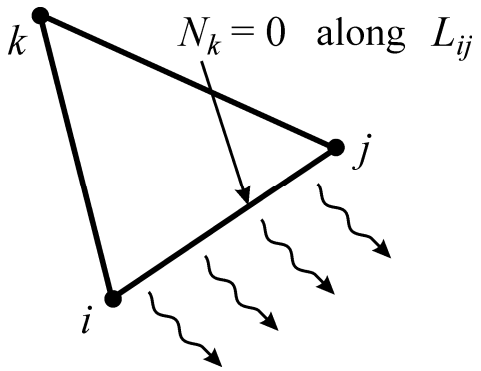


Fig. 21. Convection heat exchange through the side ij

Let the side between nodes *i* and *j* be under influence of convection $N_k = 0$ along this side and the integral becomes

$$\int_L h \cdot \left[N\right]^T \cdot \left[N\right] dL = h \cdot \int_L \begin{bmatrix} N_i \cdot N_i & N_i \cdot N_j & 0 \\ N_j \cdot N_i & N_j \cdot N_j & 0 \\ 0 & 0 & 0 \end{bmatrix} dL. \tag{3.146}$$

Let's use L-coordinates and put

$$L_1 = N_i,\ L_2 = N_j,\ L_3 = N_k = 0.$$

Using the formula (3.99), we get for the side ij

$$\int_L h \cdot [N]^T \cdot [N] dL = \frac{hL}{6} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{3.147}$$

Identical formulae can be got for the sides jk and ki.

Let's find the integral

$$\int_L h[N^T]\varphi_0 dL. \tag{3.148}$$

Fulfilling the calculations for the side ij and using L-coordinates, we get

$$\int_L h \cdot \varphi_0 \cdot \begin{bmatrix} L_1 \\ L_2 \\ 0 \end{bmatrix} dL = \frac{h \cdot \varphi_0 \cdot L}{2} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}. \tag{3.149}$$

The last integral in (3.127) is

$$\int_A [N]^T \cdot Q\ dA. \tag{3.150}$$

Supposing Q to be a constant one in the element, we have

$$\int_A [N]^T \cdot Q\ dA = Q \cdot \int_A \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} dA = \frac{Q \cdot A}{3} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{3.151}$$

So all integrals in (3.137) for each element are calculated.

It is needed to calculate matrixes for each element and to unite them into an ensemble. It leads to the system of algebraic linear equations.

Obtaining such a system and its solving is a complicated procedure. As a rule, a computer is used to fulfill them.

APPENDIX
ELEMENTS OF FUNCTIONAL ANALYSIS

1. Transformations

Let's consider two sets *F* and *G*, containing arbitrary elements. Let there is a rule A, according to which the only element *g* from the set *G* is taken for any element *f* from the set *F* relevantly. Thus, we state that a transformation of the set *F* into *G* is given (Fig. 22).

The following symbols are to be used:

$\mathrm{A}: F \to G$ A is the transformation of *F* into *G*.

$f \mapsto g,\ f \in F,\ g \in G$ the element *f* from *F* is transformed into *g* from *G*.

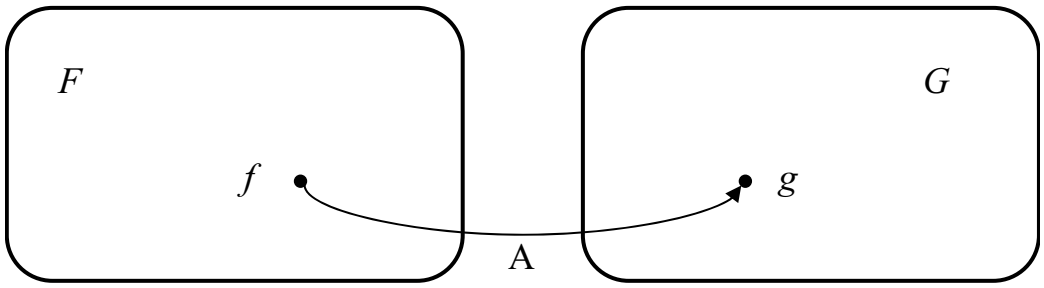


Fig. 22. Transformation A of the element $f \in F$ *into the element* $g = \mathrm{A}f$ *; the element g = Af is an image of element f*

As far as transformation A is for any element from *f*, we say that A is defined on the set *F*, and the set *F* is called the domain of A and signify it as *D*(A). In this case *D*(A) = *F*. If $f \in D(\mathrm{A})$, A*f* belongs to *G* and is called an image of the element *f*. Identically the image of set is defined. Let *P* be a subset of *D*(A). The set of images of all these elements from *P* make the image of *P* set. The image of a domain is called the set of transformation A values and is signified as *R*(A). If $Q \subset R(\mathrm{A})$, the set of the elements $f \in D(\mathrm{A})$, so that $\mathrm{A}f \in Q$ is called a preimage of the set *Q* and is signified as $\mathrm{A}^{-1}Q$. Symbol A^{-1} is not, generally speaking, transformation, because according to the statement there are several elements $f_1,\ f_2,\ldots,\ f_n \in D(\mathrm{A})$ which are able to have one and the same image $g \in R(\mathrm{A})$, therefore, the element *g* has the set $\{f_1, f_2, ..., f_n\} \subset D(\mathrm{A})$ as its pre-image. In case when each element $g \in R(\mathrm{A})$

has a pre-image containing one element $\mathrm{A}^{-1}g \in D(\mathrm{A})$, A is called one-one transformation. The A^{-1} is another one-one transformation called inverse to A. Some transformations have special names:

- if $D(\mathrm{A}) \subset \mathbb{R}^n$ and $R(\mathrm{A}) \subset \mathbb{R}^1$, such transformation is called the function of *n*-variables (by dimension);
- if transformation is defined on the function set, and $R(\mathrm{A}) \subset \mathbb{R}^1$, such transformation is a functional one;
- if transformation is defined on the function set with values in functions, it is called the operator;
- if there are two transformations A and B giving $\mathrm{A}: M \to N$ and $\mathrm{B}: N \to P$, we can define another transformation $\mathrm{C}: M \to P$, called the composition of A and B transformation and is defined through $\mathrm{B} \circ \mathrm{A}$. If $m \in M$, A transforms *m* into the element $\mathrm{A}m \in N$, meanwhile B turns it into element $\mathrm{B}(Am) \in P$. Thus, $(\mathrm{B} \circ \mathrm{A})m = \mathrm{B}(\mathrm{A}m)$.

2. Vector space

Let's find out what we can say about those sets between each element the transformation A establishes conformity. Let's view plane. Let's choose some point, call it zero and signify it by 0. Afterwards, to any point of the plane we can connect a vector (as it is presented at school: a directed segment with an arrow going from point 0 to any point of the plane). The set of points can be interpreted as the set of vectors, having common start from point 0. This interpretation is one-one transformation of points set to the set of coplanar vectors going from point 0. Let two points *p* and *q* lie at the same line with point 0 (or similarly two vectors *p* and *q* lie on the same line). Let us be able to measure the length. We signify the length of vector by l. If $l_p / l_q = \alpha$, it provides that $p = \alpha q$, when *p* and *q* lie on the one side from point 0, and $p = -\alpha q$, when they are lying on different sides (Fig. 23 *a*). Thus, we have defined multiplication of vector by number. Further, let *p* and *q* be two arbitrary vectors. Let's find their sum *r* as a vector, directed along the diagonal of the parallelogram, constructed of these vectors with the length equal to the length of the diagonal, what means $r = p + q$ (Fig. 23 *b*).

It is important to understand that methods of finding αq and $p + q$ were so-to-say defined by us. The set of point by itself doesn't provide the method to define αq and $p + q$. We can (if necessary) define these operations in a different way and even to call differently (there no interior reasons to call vector *r* a summation but not by multiplication). The way we defined the multiplication by number, and summation provides respect for traditions. Multiplication by num-

ber and the sum of vectors are examples of transformation, which were mentioned above. The 1st one transforms the plane into itself: some point of the plane is transformed into point of the same plane. The 2nd one transforms any pair of vectors (element of the domain presents any pair of vector by itself) into a vector: there is the 3rd point of this plane, which is put relevantly to any pair of points. The defined transformations have the range of features. At first there is commutativity and associativity of addition and multiplication by number:

1. $p+q=q+p$,
2. $p+(q+r)=(p+q)+r$,
3. $\alpha(p+q)=\alpha p+\alpha q$,
4. $(\alpha+\beta)p=\alpha p+\beta p$,
5. $(\alpha\beta)p=\alpha(\beta p)$.

Here α,β are numbers and p, q, and r are vectors. Further the zero vector is relevant to point 0, for which it is right $p+0=p$.

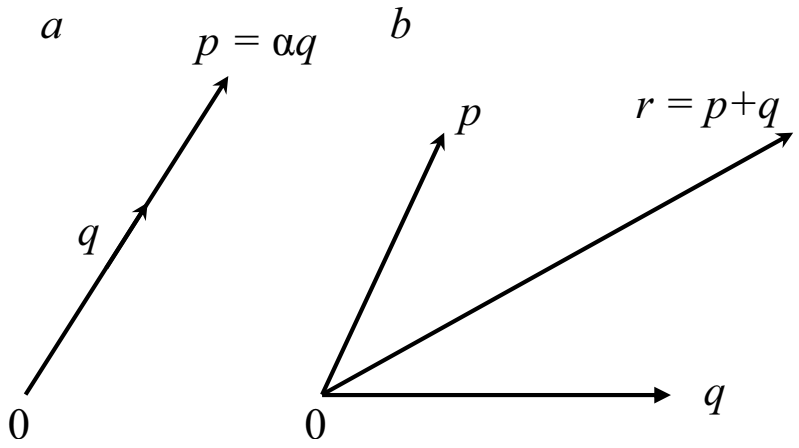


Fig. 23. Operations with vectors

Besides, for any vector p there is always vector q $p+q=0$, and it is actually, expressed through p. If vector p is multiplied by 1, it will finally transform into itself (the length and direction remain the same). The set, for which elements there is addition and multiplication by number defined, having the pointed features, is called vector space. It is particular that a vector, an element of vector space, can be presented not only by the point of the plane (or an arrow), and also by object of any origin (as we will see it further – number, function, operator, etc.). It is necessary to define addition and multiplication by number, possessing the features mentioned above. Let's take all those given above by the following way. Let V be some not empty set and f, g, h be some of its elements. This set is called *vector* (*or linear*) *space*, if there is mentioned rule, according to which the 3rd element from V respectively called the sum of elements, is put in

compliance with two elements of V. The same is about the rule according to which to any element from V and any number (generally, complex number) we put an element from V respectivety called multiplication of element by number. Both these rules obey the following axioms:

1. $f + g = g + f$ is the commutative law;
2. $(f + g) + h = f + (g + h)$ is the associative law;
3. There is an element 0 called zero, at which $f \cdot 0 = 0$;
4. For any f there is an opposite element $-f$, at which $f + (-f) = 0$;
5. $1 \cdot f = f$;
6. $\alpha(f + g) = \alpha f + \alpha g$;
7. $(\alpha + \beta) f = \alpha f + \alpha f$;
8. $(\alpha\beta) f = \alpha(\beta f)$.

In axioms 5–8 $1, \alpha, \beta$ are numbers. The elements $f, g, h, \ldots \in V$ are called the points (or the vectors).

Example 1. $\mathbb{R}^1$ is the set of substantial numbers. Fulfilling the axioms 1–8 for defined by usual way addition and multiplication is not difficult to be proved. So $\mathbb{R}^1$ is the vector space, the points and vector of which are presented by substation numbers. If we "place" all substation numbers on the line (to choose zero point, to connect p with number α, if the distance from 0 to p equals α), the vectors can be presented as arrows directed from point 0 to point p.

Example 2. $\mathbb{R}^n$ is the set, an element of which is ordered totality from n-numbers $(x^1, x^2, \ldots, x^n)$ (the sign above x is not a power but index). The number x^i will be called i-component of the element. Let's define the addition of elements $\mathbb{R}^n$ and their multiplication by number component.

If $f = \left(f^1, f^2, ..., f^n\right)$ and $g = \left(g^1, g^2, ..., g^n\right)$ are elements of $\mathbb{R}^n$, and α is the number, thus

$$f + g = \left(f^1 + g^1, f^2 + g^2, ..., f^n + g^n\right)$$

and

$$\alpha f = \left(\alpha f^1, \alpha f^2, ..., \alpha f^n\right).$$

The element (0, 0, … 0) is called zero element. The axioms 1–8 are again easily checked, so that the set $\mathbb{R}^n$ is also vector space.

Let's make addition to the example 2. Let P, Q be two arbitrary sets, consisting of elements p_i and q_i respectively. A new set can be formed, the elements of which will be possible ordered pairs (p_i and q_i). This new set is called

a direct product of P and Q and it is signified through $P \times Q$. Let V and W be vector spaces. The direct product $V \times W$ may be turned into vector space, if addition and multiplication by number are to be defined by the following way:

$$(f,g)+(p,q)=(f+p,g+q),$$

$$\alpha(f,g)=(\alpha f,\alpha g),$$

for $f,p \in V;\ g,q \in W;\ (f,g),(p,q) \in V \times W,$ and α are substantial or complex numbers. It is clear, the space $\mathbb{R}^n$ can be interpreted as a direct product n of vector spaces $\mathbb{R}^1$:

$$\mathbb{R}^n = \underbrace{\mathbb{R}^1 \times \mathbb{R}^1 \times \ldots \times \mathbb{R}^1}_{n}.$$

Example 3. $\mathbb{C}$ is a set of complex numbers $(\alpha + i\beta)$, where α,β are sustention numbers and i is an imaginary unit. Addition and multiplication by number is to be defined by the following way:

$$(\alpha + i\beta) + (\gamma + i\delta) = (\alpha + \gamma) + i(\beta + \delta),$$

$$\gamma(\alpha + i\beta) = (\gamma\alpha) + i(\gamma\beta).$$

The element $(0 + i0)$ is called the zero element. Axioms 1–8 are also fulfilled here, it means that $\mathbb{C}$ is also vector space.

Example 4. The set $n \times n$ of matrixes will be also vector space, if the sum of matrixes and multiplication of matrix by number, as it is done in linear algebra that is component by component. A zero element of this space will be a zero matrix, all elements of which equal zero.

The number of examples can be increased.

If some subset S of vector space V arranges the vector space itself, it will be called sub space of vector space V. For example, any plane going through point 0 at $\mathbb{R}^3$ is subspace $\mathbb{R}^3$, as far as it is itself vector space $\mathbb{R}^2$ itself. Identically any line going through point 0 is subspace $\mathbb{R}^3$. Besides, the given line is subspace of those planes $\mathbb{R}^2$, in which the line lies.

The sum of non-zero vectors products is called a linear combination of vectors $f, g, h\ldots$

$$\alpha f + \beta g + \gamma h + \ldots$$

It is evident, if V is vector space, it contains any linear combination of its elements (a linear combination is a vector). Vector which is a linear combination of some other vectors is called the linear dependent. If it can't be presented as linear combination of mentioned vectors unit, it is linear independent from them. If we chose any vector f in $\mathbb{R}^1$, not equal to zero, all other vectors

are called linear dependent on it and they can be written as αf, where α is the number. In vector space $\mathbb{R}^2$ it is different. Having chosen non-zero vector f, we can't state that other vectors will be linear dependent on it, as vectors linear dependent on f will lie on the line going through point 0 and f. Two vectors not lying on the same line is enough for other vector to be linear dependent on them. The totality of non-zero vectors f, g… from some linear space is called linear independent, if there's no such unit of numbers α, β, … giving

$$\alpha f + \beta g + ... = 0.$$

For arbitrary set of vectors the maximum number n of linear independent vectors is called its dimension. So the set of points on the line is one-dimensional, and the set of points on the plane is two-dimensional. If there's no such maximum number (the number of linear independent vectors is larger than any number n given in advance), the set is called infinite dimensional, in the contrary case, it is called infinite dimensional.

3. Basis of vector space

Let's view some vector space V, and let n be its dimension. It means that in V it is possible to choose n linear independent vectors (there are different method to choose). Any $(n + 1)$ – number vector will be surely linear dependent on them; it can be written as a linear combination of the 1st linear independent vectors. In other words, if $f_i \in V$, $i = 1,...,n$ are chosen by us n linear independent vectors (at which $\sum \alpha_i f_i \neq 0$ for non-zero unit $\alpha_i \in \mathbb{R}^1$), in this case for any another vector $g \in V$ there are always real numbers α_i, β, such as

$$\sum_{i=1}^{n} \alpha_i f_i + \beta g = 0. \tag{4.1}$$

So, having chosen n linear independent elements $\{f_i\}_{i=1}^n$, we can write another element from V. In fast, from (4.1) we have

$$g = -\sum_{i=1}^{n} \frac{\alpha_i}{\beta} f_i = \sum_i \gamma_i f_i . \tag{4.2}$$

This set $\{f_i\}$ of linear independent elements is called basis of vector space V. Numbers $\{\gamma_i\}$ are called components of vector g in basis $\{f_i\}$. Pointing out basis is important, as there can be a lot of basis and in each new basis the components of one and the same vector g will be different. We are going to enumerate basic vectors by a lower index, for example e_i, and components of the vector in this basis will be signified by the same

letter (of each the vector itself with upper index). Thus, for element f we have a division by basis $\{e_i\}_{i=1}^{n}$:

$$f = \sum_{i=1}^{n} f^i e_i . \tag{4.3}$$

4. Coordinate transformation

From all mentioned above, each element f from n-dimensional vector space V can be given by set of numbers $\{f^i\}$, responding to the chosen basis $\{e^i\}$. Each vector $f \in V$ is put in accordance with the set of n-numbers $\left(f^1, f^2, ...\right)$, $f^i \in \mathbb{R}^1$ that provides the transformation φ, turning any vector from vector space V into vector $\mathbb{R}^n$, or $\varphi : V \to \mathbb{R}^n$. Such transformation is called coordinate, and the unit $\{f^i\}$ is called coordinates of vector f (or point) related to the transformation φ. It is necessary to mention transformations as well as basis. Different transformations give a different coordinate system. Let the transformation φ put the point f in compliance with the unit coordinates $\{f^i\}$ and the transformation ψ be put in accordance with f the unit of coordinates $\{g^i\}$ (Fig. 24).

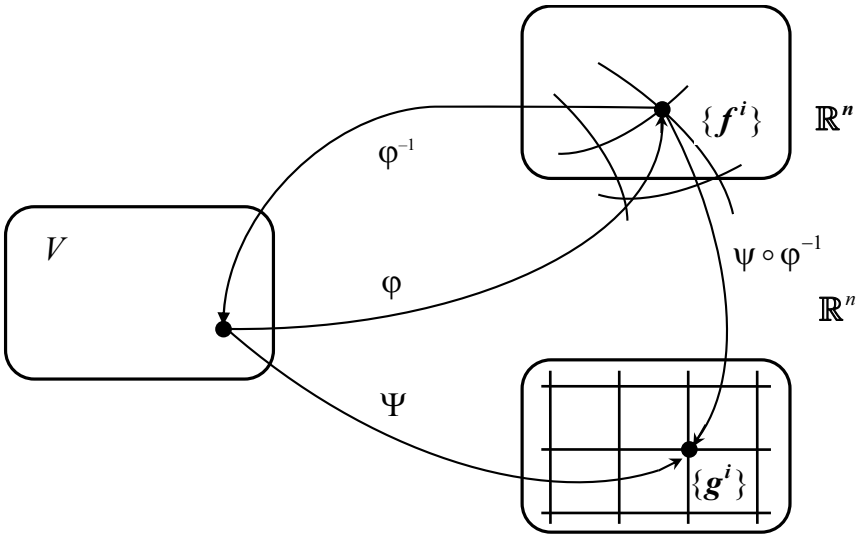


Fig. 24. Two coordinate transformations φ, and ψ transforms point f of vector space V into different points $\mathbb{R}^n$. *One-one coordinate transformations φ, and ψ gives transformation of coordinates* $\psi \circ \varphi^{-1} : \mathbb{R}^n \to \mathbb{R}^n$

As far as different points $\{f^i\}$ and $\{g^i\}$ from $\mathbb{R}^n$ are images of one and the same point $f \in V$, there must be a connection between them. In other

words, there must be a connection between different coordinate system $\{f^i\}$ and $\{g^i\}$. We are going to consider only one-one coordinate transformations. As transformation φ is one-for-one, it has inverse transformation φ^{-1}, which transforms point $\{f^i\} \in \mathbb{R}^n$ into point $f \in V$ the point *f* is transferred to the point $\{g^i\} \in \mathbb{R}^n$ by the transformation ψ.

So compositions $\psi \circ \varphi^{-1}$ are the transformation $\mathbb{R}^n \mapsto \mathbb{R}^n$ or

$$\psi(\varphi^{-1}\{f^i\}) = \{g^i\}. \tag{4.4}$$

As a result we have functional ratios, defining the transformations of coordinates

$$\begin{aligned} g^1 &= g^1(f^1, f^2, ..., f^n), \\ g^2 &= g^2(f^1, f^2, ..., f^n), \\ &\cdots \\ g^n &= g^n(f^1, f^2, ..., f^n). \end{aligned} \tag{4.5}$$

Let *V* be the plane. Let's transform it on $\mathbb{R}^2$ the following way. Let's chose point 0 in the plane with which we connect zero vector $\mathbb{R}^2$, exactly (0, 0). Let's construct the line at any direction through point 0. It will be called abscissa axis, and to its each point we put vector from $\mathbb{R}^2$ according to the type (α, 0), where α is a substation number. Let's construct through point 0 another line perpendicular to the 1st and call it the axis of ordinates. Then let's connect vectors from $\mathbb{R}^2$ of type (α, β), where β is substation number with it identically. To other points of the plane we put accordingly the vectors from $\mathbb{R}^2$ of type (α, β), if perpendiculars derived from point on the axis cross this axis respectevely at the points (α, 0) and (0, β). So the coordinate transformation is constructed: each element from *V* is transformed into an element from $\mathbb{R}^2$ (Fig. 25) such coordinates of plane point are called Cartesian coordinates.

Let's transform the plane into $\mathbb{R}^2$ in a different way. We start to construct this transformation as well as Cartesian one, including the construction of the abscissa axis. In new transformation, it will be called polar axis. Further, to each plane point we put the vector (ρ, ϑ) accordingly, and if this point lies on the cross of circumference of radius ρ with the centre at point 0 (this circumference crosses the polar axis at point (ρ, 0)) with a half line, going from point 0 at the angle ϑ to the polar axis.

Angle is measured anticlockwise and called the polar angle.

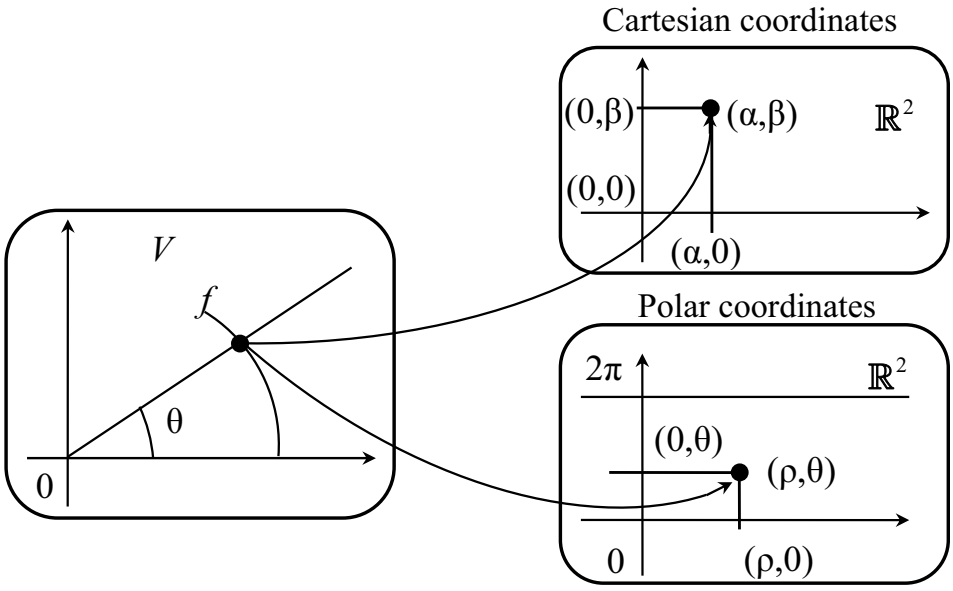


Fig. 25. Constructions of coordinate transformations

In the 1st case the plane is transformed onto all $\mathbb{R}^2$, and in the 2nd – onto subset consisting of vectors (ρ,ϑ): $\rho\in[0,+\infty)$, $\vartheta\in[0,2\pi]$.

Now we can construct ratios (4.5), and the dependency ρ and ϑ on α and β. It is evident,

$$\rho=\rho(\alpha,\beta)=\sqrt{\alpha^2+\beta^2},\quad \vartheta=\vartheta(\alpha,\beta)=\arctan\frac{\beta}{\alpha}$$

and

$$\alpha=\alpha(\rho,\vartheta)=\rho\cos\vartheta,\quad \beta=\beta(\rho,\vartheta)=\rho\sin\vartheta.$$

The rule putting to each point f some number accordingly (value of function at this point) will be called the function given on V.

So coordinates $\{f^i\}$ are functions in vector space V: to each point $f\in V$ there are n-coordinates accordingly and their values change from point to point.

5. Metrics and norm

We have already used the notion of distance without mentioning the way of measuring. Let's view this issue in detail. If we took only point on the line in the plane or in a 3-dimensional space, everything could be clear. Now points are presented by elements of vector spaces. As far as the plane is a particular case of vector space, the distance between points of this space is identical to the distance between points in the plane. Firstly, the distance is not negative. Secondly, it depends on mutual location of points and doesn't depend on their location relevant to zero. Thirdly, it is not important to know how to measure distance: from the 1st point to the 2nd or vice versa. Fourthly, if points coincide, the distance between them equals

zero. Finally, if we view distance between three points, any of them doesn't overcome the sum of two others (the length of a triangle side doesn't overcome the sum of two lengths of two other sides). The same features belong to the distance between points of any vector space or even just arbitrary set (in the plane the distance between point has the sense, if we even consider only one part of it). Let X be an arbitrary set and $f, g, h \in X$. We put to each pair f, g not negative number $d(f, g)$ accordingly so that for any f, g, h from X it is right that

1. $d(f, g) > 0$, for $f \neq g$,
2. $d(f, f) = 0$,
3. $d(f, g) = d(g, f)$,
4. $d(f, g) \leq d(f, h) + d(h, g)$.

It is evident that $d(f, g)$ is the function transformation defined at any pair of vectors from X with values in numbers (in $\mathbb{R}^1$). Such function is called the metrics to X. And X itself provided with metrics is metrical space. Pay attention that in order to make any set metrical space there we should introduce metrics. It doesn't become vector space, because the addition of elements and their multiplication by number is not defined in it. In this case when we work with vector space V, we can use one bold point 0 and introduce a quite definite notion of a vector norm, in fact, the distance from the element to 0. As here we suppose some rule, putting a sustention number to the point which is the distance to point 0, we have function transforming V into $\mathbb{R}^1$ to signify the norms a special sign is used $\|\cdot\|$, the norm of vector f is signified as $\|f\|$. Let's write the definition to the norm for more general complex case we need later. Let V be vector space and $f \in V$ is the norm of vector f is called not negative new metrical function $\|f\|$ defined in V, such that for any $f, g \in V$ and $\alpha \in \mathbb{R}^1$ the following conditions are fulfilled:

1) $\|f\| > 0,\ f \neq 0$,
2) $\|0\| = 0$,
3) $\|\alpha f\| = |\alpha| \cdot \|f\|$,
4) $\|f + g\| \leq \|f\| + \|g\|$ (triangle inequality).

Changing the element g into $(h - f)$ and the following re-symbolizing it is easy to get other variants of triangle inequality. They can be presented as

$$\left|\|f\| - \|g\|\right| \leq \|f + g\| \leq \|f\| + \|g\|,$$

$$\left|\|f\| - \|g\|\right| \leq \|f - g\| \leq \|f\| + \|g\|.$$

Linear space provided with the norm is called the normed vector space. If we know the distance to 0 (norm) for each point from V, it is easy to meas-

ure the distance between points from V that means to get metrics. The distance between two points f and g can be the norm of their distance:

$$d(f,g)=\|f-g\|. \tag{4.6}$$

The definition of the norm given above doesn't present it in the only way. Very often we can introduce several norms for one and the same vector space. Here the spaces found are considered to be different:

1. Linear space $x\in\mathbb{R}^1$ becomes normed, if the norm $\|x\|$ of the element $x\in\mathbb{R}^1$ is its module $\|x\|$. It is evident that this definition in $\mathbb{R}^1$ is correct.
2. The vector norm $x=\left(x^1,x^2,...,x^n\right)$ can be introduced by different ways in the space $\mathbb{R}^n$. The used norms are the following:
   a) octahedral norm or norm $\|\cdot\|_1$:

$$\|x\|_1=\sum_{i=1}^{n}\left|x^i\right|; \tag{4.7}$$

   b) spherical (euclidean) norm or norm $\|\cdot\|_2$:

$$\|x\|_2=\left(\sum_{i=1}^{n}\left|x^i\right|^2\right)^{\frac{1}{2}}; \tag{4.8}$$

   c) norms $\|\cdot\|_p$, where p is a natural number (norms $\|\cdot\|_1$, $\|\cdot\|_2$ are particular cases of norms $\|\cdot\|_p$):

$$\|x\|_p=\left(\sum_{i=1}^{n}\left|x^i\right|^p\right)^{\frac{1}{p}}; \tag{4.9}$$

   d) cubic norm $\|\cdot\|_\infty$:

$$\|x\|_\infty=\max_i\left|x^i\right|. \tag{4.10}$$

3. Visible presentation of these norms is given by the set of the elements $x\in\mathbb{R}^n$, for which $\|x\|=1$, or so called unit sphere. It is demonstrated by the Fig. 26.
4. The vector spaces are very important in our course, elements of which are functions (functions will be points of vectors of the given space).

Let's view the set $F\left(\left[a,b\right]\right)$ of sustention functions defined on the segment $\left[a,b\right]$. Let's take $f,g \in F$ and $\alpha \in \mathbb{R}^1$. Let's define new functions $(f+g)$ and αf taking for all $x \in \left[a,b\right]$:

$$(f+g)(x) = f(x) + g(x),$$
$$(\alpha f)(x) = \alpha f(x).$$

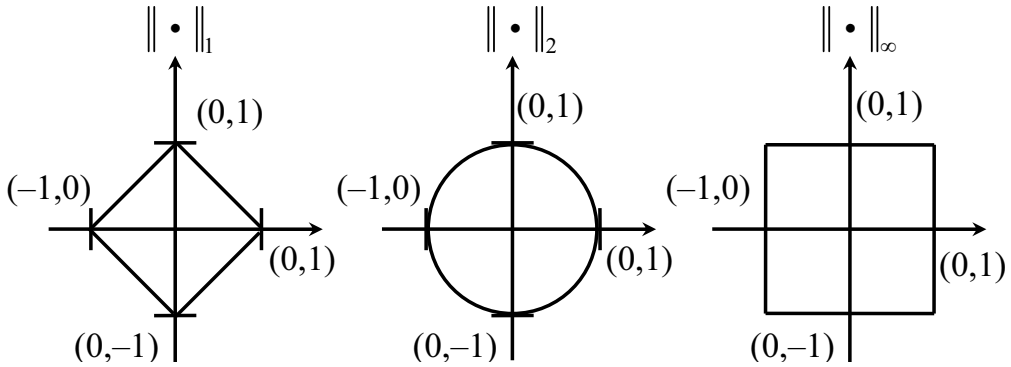


Fig. 26. The view of unit sphere to $\mathbb{R}^2$ *for different norms*

In other words, the value of function sum equals the sum of values for function-items at the same point. It is identical for αf. The axiom of vector space is fulfilled as a rule; such generally linear spaces of functions are not viewed. The sub-set F is studied, which in their turn form linear spaces, for example, the space $C^0([a, b])$ of continuous functions defined on $\left[a,b\right]$, or the space $C^k([a, b])$ of functions having k as limited continuous derivative. Functional linear spaces are infinitely dimensional. These spaces can be normed by constructing a relevant norm. So $C^0([a,b])$ are often provided with so-called the uniform norm

$$\left\|f\right\| = \max_{x\in[a,b]} \left|f(x)\right|. \tag{4.11}$$

It is clear, this is an infinitely dimensional norm $\left\|\cdot\right\|_\infty$. Another possible norm is a norm $\left\|\cdot\right\|_1$:

$$\left\|f\right\|_1 = \int_a^b \left|f(x)\right| dx. \tag{4.12}$$

There are other variants of norms. The uniform norm can be given visible interpretation. Let M be the set of functions, which gives $\left\|f(x)\right\| < \alpha$ for all $f(x) \in M$ and $\alpha > 0$, $\alpha \in \mathbb{R}^1$. All f from M must belong to to the band $\pm\alpha$ relative to the abscissa axis (Fig. 27).

If P is the set of function f, for which the distance from the given function g doesn't overcome β, it gives $\|f-g\| \le \beta$. All changes of f must be concluded in the band with width 2β, including function g. For norm $\|\cdot\|_1$ it is not right as only the integral is limited, whilst the value of function at separate points are possible not to satisfy this limit. Thus, the limitation of point $\|f\| \le \alpha$ is more definite than $\|f\|_1 \le \alpha$, and from the first we have the second, but not in another way.

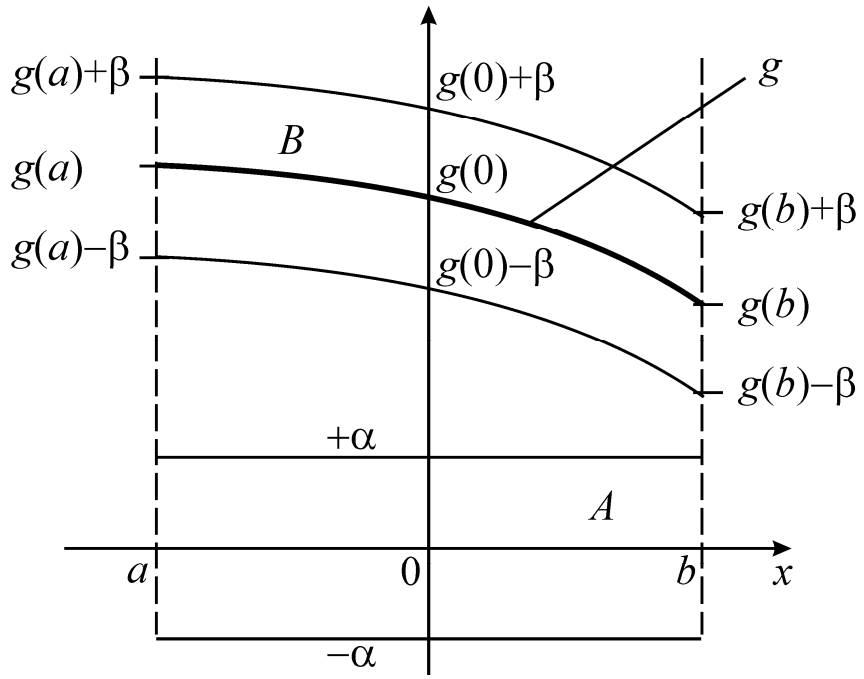


Fig. 27. The oscillations of function f, for which $\|f\| < \alpha$, must occur in the band A; functions f, the distance of which doesn't overcome $\beta > 0$ (that is $\|f-g\| \le \beta$), change in the band B

6. Banach space

At solving the equation as $Af = g$ very often the interational procedures are used. They state that with the help of some algorithm by chosen initial approximation f_0 the sequence $\{f_n\}$ is made up, each following member of which must be more precise approximation of solution f. The sequence of approximations $\{f_n\}$ must converge to the precise solution f. If functions f_n are to be interpreted as vectors (points) of vector space, there will be a need in definition of convergence of vector sequence. It is known that a numerical sequence $\{a_n\}_{n=1}^{\infty}$ converges to some number α, when the difference module $|a_n - a|$ goes to zero at n tending to ∞. If members of the sequence $\{f_n\}$ be-

long to an arbitrary set X, it is impossible to use the convergence defined for numerical sequences. If we introduce X metrics or norm, we will be able to define the convergence $\{f_n\}$ as well as it was done for the numbers. Let's view some linear space V with metrics $d(f,g)$, $f,g \in V$ (if V is normed, $d(f,g) = \|f-g\|$). Let $\{f_n\}_{n=1}^{\infty}$ is the sequence of points from V, i. e. countable subset of V (as it is numbered). If there is such point $f \in V$, that with increase of n the distance between f_n and f reduces within the limits to zero

$$\lim_{n\to\infty} d(f_n, f) = 0, \text{ or } \lim_{n\to\infty} \|f_n - f\| = 0,$$

the sequence $\{f_n\}$ will be convergent, and the element f is a sequence limit $\{f_n\}$. It is represented in the following way:

$$\lim_{n\to\infty} f_n = f, \text{ or } f_n \to f.$$

To appreciate whether the sequence is convergent, we must know the limit, and be sure that the distance from sequence points goes to zero till the limit. It is not convenient, as the limit is usually unknown, and we have only sequence members. It is clear that with the increase of points f_i and f_j size the distance between them goes to zero. The increasing numbers i and j are not necessary to be connected. Indeed, let $f_i, f_j \in \{f_n\}$ and $f_n \to f$, so that according to the triangle inequality we have

$$d(f_i, f_j) \le d(f_i, f) + d(f_j, f).$$

If $i \to \infty$ (independently on i) and $j \to \infty$, $d(f_i, f) \to 0$ and $d(f_j, f) \to 0$, we have $d(f_i, f_j) \to 0$. The sequence $\{f_n\}$ with the feature $d(f_i, f_j) \to 0$ at $i, j \to \infty$ is called fundamental or Cauchy sequence. So each convergent sequence is Cauchy sequence. What about the reverse? It is proved as right for the finite-dimensional. As for the non-finite dimensional (the set of functions are non-finite dimensional as a rule) case it is not so. The example is as follows. Let $V = C_0([-1, 1])$ be the space of non-finite functions defined on the segment $[-1, 1]$ with a norm $\|f\|_1 = \int_{-1}^{1} |f(x)|\, dx$. Let's consider the function sequence $\{f_n\}$ defined as

$$f_n = \begin{cases} 1, & x \in [-1, 0], \\ 1 - nx, & x \in (0, \frac{1}{n}), \\ 0, & x \in [\frac{1}{n}, 1]. \end{cases} \qquad (4.13)$$

The first few functions are demonstrated in the Fig. 28. Surely, these functions belong to the viewed normalized linear space: they are continuous at each point of function definition. Besides, it is easy to check that

$$\left\| f_i - f_j \right\|_1 = \int_{-1}^{1} \left| f_i(x) - f_j(x) \right| dx \to 0 .$$

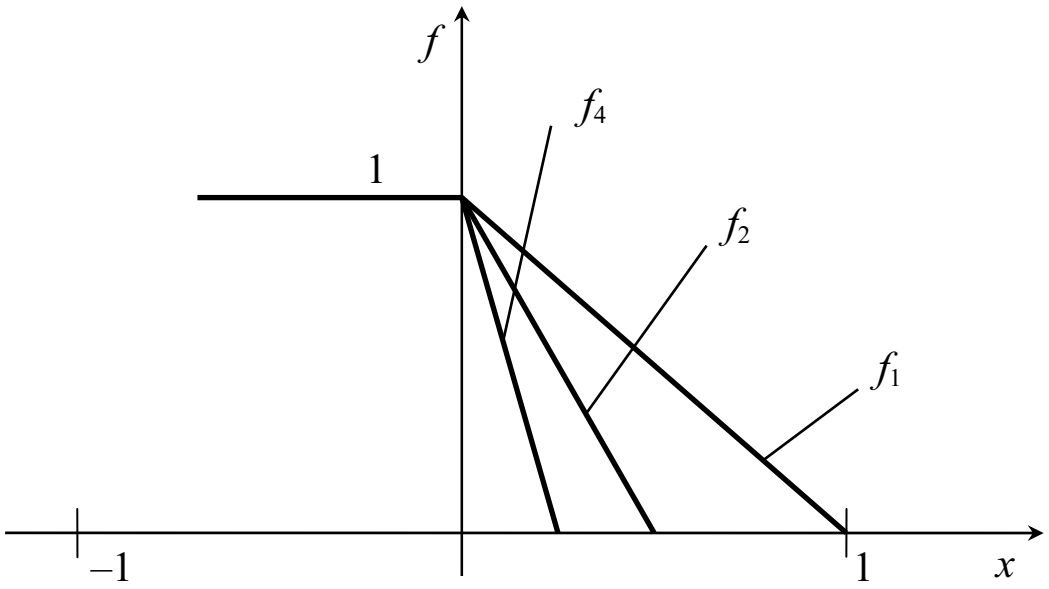


Fig. 28. Functions defined by the expression (4.13)

Indeed let's view the module of difference equation $\left| f_i - f_j \right|,\ j > i$. The norm $\left\| f_i - f_j \right\|_1$ is the square of the triangle, base of which equals $\frac{1}{i}$, and the height is defined by difference $\left| 1 - \frac{i}{j} \right|$. With *i* increasing, the base of the triangle goes to zero, consequently, the norm $\left\| f_i - f_j \right\|_1$ goes to zero too.

So the sequence of function $\{f_n\}$ is the sequence of Cauchy. As far as $\frac{1}{n} \to 0$, the limit $\{f_n\}$ is a function

$$f_\infty(x) = \begin{cases} 1, & x \in [-1, 0], \\ 0, & x \in [0, 1], \end{cases}$$

having the breach at point 0 not belonging to *V*. Thus, we have constructed Cauchy sequence, not converging to any point of the viewed space of continuous functions.

Let's consider the same function sequence, but in space W different only by norm: let W has a uniform norm $\|f\| = \max_x |f(x)|$.

It is easy to understand

$$\|f_i - f_j\| = \max|(i-j)x| = |1 - i/j|.$$

As far as i and j go to the infinity independently, their ratio can be any like, as the result, their norm doesn't go to 0. For example, if $i = 2j$, then $\|f_i - f_j\| \to 1$. So in the space W sequence $\{f_n\}$ is not Cauchy sequence and doesn't converge. Let's view the similar sequence of functions $\{g_n\}$ (Fig. 29) to be sure there are no sequences in W:

$$g_n(x) = \begin{cases} \dfrac{1-nx}{1+n}, & x \in [-1,\ \dfrac{1}{n}], \\ 0, & x \in [\dfrac{1}{n},\ 1]. \end{cases} \tag{4.14}$$

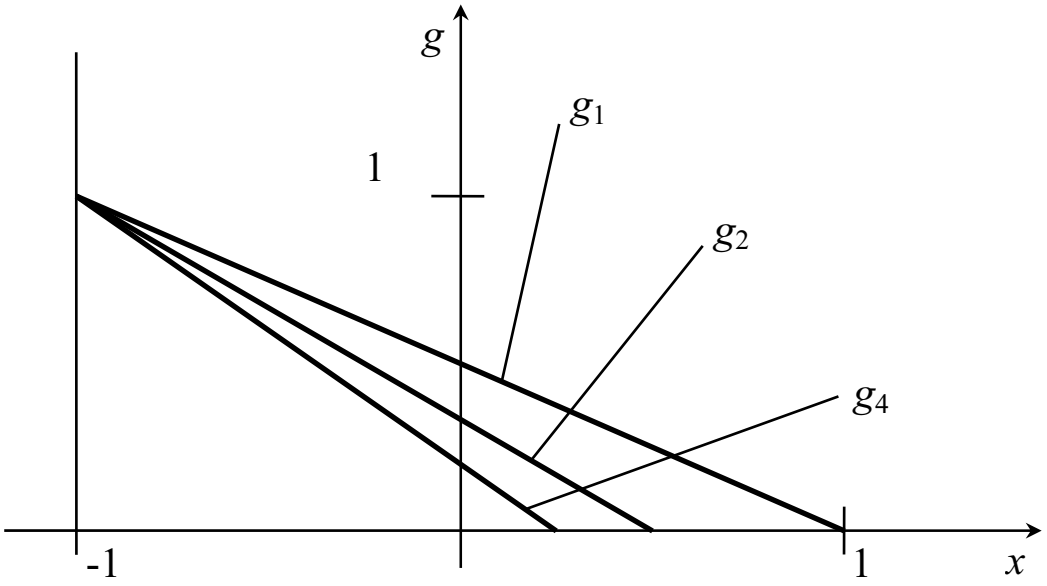


Fig. 29. Functions defined by the expression (4.14)

Simple calculations give $\|g_i - g\| = \left|\dfrac{1 - i/j}{1+i}\right| \to 0$.

Thus, the sequence $\{g_n\}$ is Cauchy sequence and goes to the continuous function:

$$g(x) = \begin{cases} -x, & x \in [-1, 0], \\ 0, & x \in [0, 1]. \end{cases}$$

Functions g_n belong to the space *V*, where the sequence of Cauchy is also formed.

If any Cauchy problem converge by the norm (the limit is the element of the same space) in the normalized vector space, it will be called fully normalized vector space (these spaces we are to define by *B*). A completeness of some subset *S* of normalized vector space *V* may be stated. If $V = \mathbb{R}^1$, each segment $S = [a,b]$ is complete as so as any Cauchy sequence of its points converge to some point *S*. Vice versa, the interval *P* = (*a*; *b*) is not complete, as there are sequences going to *a* (for example, $\{a + \frac{1}{n}\}$), and point *a* doesn't belong to *P*. The set containing all its limit points is called the closed set. It is a sphere with surface and a circle with boundary.

Statement. The subset *S* of Banach space *B* is complete when it is closed in *B*.

Let *S* be some subset *V*. If it is not closed (doesn't contain all its limit points), a new subset may be defined $\overline{S} \subset V$, constructing by adding to *S* all its limit points, and is called the closing of *S*. So, if *S* is closed, $S = \overline{S}$. The term completeness is easy to be presented: the set is complete when there is a lack of points in it to make any Cauchy sequence convergent. If vector space given one norm is complete, and given another norm – it doesn't, the question arises what norms do not disturb the completeness of vector space. The answer is the equivalent norm. Two norms $\| \cdot \|_a$ and $\| \cdot \|_b$ in the vector space *V* are equivalent, if the numbers $c_1 > 0$ and $c_2 > 0$ are found, at which for any $f \in V$

$$c_1 \|f\|_a \le \|f\|_b \le c_2 \|f\|_a. \tag{4.15}$$

Similarly, we may limit the norm $\|f\|_a$:

$$\frac{1}{c_2}\|f\|_b \le \|f\|_a \le \frac{1}{c_1}\|f\|_b$$

or in different symbols

$$k_1 \|f\|_b \le \|f\|_a \le k_2 \|f\|_b,\ k_1, k_2 = const > 0.$$

If norms $\| \cdot \|_a$ and $\| \cdot \|_b$ are equivalents, from the convergence by norm $\| \cdot \|_a$ follows the convergence by norm $\| \cdot \|_b$, and vice versa. Indeed let $\|f_n - f\|_a \to 0$ at $n \to \infty$. Then it goes from the formula (4.15) that numerical sequence $\|f_n - f\|_b$ is majorated by sequence going to 0 and converging to 0 as well.

7. Hilbert space

One of methods to introduce norms in Banach space is to give a scalar (or interior) product in it. A scalar product is a numerical function of two arguments $(\cdot,\cdot)$. It puts a number to each pair of vectors accordingly. The scalar product must satisfy the following conditions. Let V be vector space, $f,g,h \in V$ and $\alpha \in \mathbb{R}^1$, then:

1. $(f,f) \geq 0, \quad (f,f)=0 \Leftrightarrow f=0$,
2. $(f,g+h)=(f,g)+(f,h)$ – associative,
3. $(f,g)=(g,f)$ – symmetry,
4. $(\alpha f,g)=\alpha(f,g)$ – homogeneity.

Vector space V with the inserted product into it is called pre-Hilbert one or Cartesian.

Examples

1. In $\mathbb{R}^n$ (which elements are ordered sets of numbers $f=(f^1,...,f^n)$, $g=(g^1,...,g^n)$) a scalar product is defined by

$$(f,g)=f^1g^1+f^2g^2+...+f^ng^n.$$

All axioms of scalar product are easy to be checked.

2. Let $V=C^0([0,1])$ be the set of continuous functions defined on [0, 1]. A scalar product can be as follows:

$$(f,g)=\int_0^1 f(x)g(x)dx.$$

Having found the scalar product (having obtained), we can introduce a norm:

$$\left\| f=\sqrt{(f,f)} \right\|. \qquad (4.16)$$

The fact that it is a norm is easy to be proved. The first 3 features are satisfied. So there is a triangle inequality to be checked. We take it as

$$\|f+g\|^2 \leq (\|f\|+\|g\|)^2.$$

For any f and g from the very space we have

$$\|f+g\|^2=(f+g,f+g)=\|f\|^2+(f,g)+(g,f)+\|g\|^2=$$

$$=\|f\|^2+2(f,g)+\|g\|^2 \leq \|f\|^2+2|(f,g)|+\|g\|^2.$$

If it becomes possible to demonstrate that $|(f,g)| \leq \|f\| \cdot \|g\|$, the triangle inequality will turn out to be right. This ratio is really fulfilled. Let's take the norm of the vector $f+\alpha g$, where α is a number. It is evident that

$$0 \le \|f + \alpha g\|^2 = \|f\|^2 + 2(f, \alpha g) + \|\alpha g\|^2 =$$
$$= \|f\|^2 + 2\alpha(f, g) + \alpha^2 \|g\|^2.$$

This implyies

$$-2\alpha(f, g) \le \|f\|^2 + \alpha^2 \|g\|^{2\cdot}. \tag{4.17}$$

As far as α is an arbitrary number, we take

$$\alpha = -\frac{(f, g)\|f\|}{|(f, g)| \cdot \|g\|},$$

and, inserting it into the Eq. (4.17), we get

$$2\frac{(f, g)^2 \|f\|}{|(f, g)| \cdot \|g\|} \le \|f\|^2 + \frac{(f, g)^2 \|f\|^2}{|(f, g)^2| \cdot \|g\|^2} \|g\|^2,$$

or

$$|(f, g)| \le \|f\| \cdot \|g\|. \tag{4.18}$$

The equality (4.18) is called the equality of Cauchy–Bunyakovsky. It allows to get

$$\|f + g\|^2 \le \|f\|^2 + 2|(f, g)| + \|g\|^2 \le (\|f\| + \|g\|)^2,$$

or

$$\|f + g\| \le \|f\| + \|g\|.$$

So any pre-Hilbert space may be normalized, if the norm is defined by the equality (4.16). If the pre-Hilbert space is complete by the norm (4.16), it is called Hilbert one (we define such spaces by the letter *H*). In other words, Banach space is called Hilbert one in which the norm is defined through scalar product. Initially, it is needed to define the scalar product. Not every Banach space may be done Hilbert. There is one important result: if in each space for any vectors *f*, *g*, the rule of parallelogram is fulfilled

$$\|f + g\|^2 + \|f - g\|^2 = 2(\|f\|^2 + \|g\|^2).$$

In this case we can introduce the scalar product into it (to make the system Hilbert one), defining it by the formula

$$(f, g) = \frac{1}{4}(\|f + g\|^2 - \|f - g\|^2).$$

8. Orthogonality and the theories of Fourier

From Cauchy–Bunyakovsky inequality (4.18) it is stated that function $\frac{(f,g)}{\|f\|\cdot\|g\|}$ doesn't overcome by 1 module. This allows to equal it with cosine of angle φ between vectors *f* and *g*, and also to write the scalar product as $(f,g)=\|f\|\cdot\|g\|\cos\varphi$.

Numbers $f_g \equiv \|f\|\cos(\varphi)$ and $g_f \equiv \|g\|\cos(\varphi)$ are called the projection *f* on *g* and *g* on *f* accordingly. In case when the norm of one of the vectors, for example *g*, equals one, we have $(f,g)=\|f\|\cos\varphi = f_g$. It means that the scalar product of the vector *f* by the unit vector *g* equals the projection *f* on *g*. If $(f,g)=0$ and $\|f\|\neq 0,\ \|g\|\neq 0$, it follows $\varphi=\pm\frac{\pi}{2}$. In this case the vectors are orthogonal that is signified by $f\perp g$. If there is a sequence of the non-zero vectors $\{f_n\}$ and for each pair of the vectors f_i, f_j the equality $(f_i,g_j)=0,\ \ i\neq j$ is right, such sequence is orthogonal. It is easy to show that an orthogonal vector is linear independent, i. e.

$$\alpha_1 f_1+\alpha_2 f_2+...+\alpha_n f_n=0,$$

only if $\alpha_1=\alpha_2=...\alpha_n=0$. Let $\{f_i\}_{i=1}^n$ be an orthogonal vector and suppose they are linear dependent, i. e. $\sum\alpha_i f_i=0$ is zero vector. Let's take any vector $f_j\in\{f_i\}_{i=1}^n$ and multiply it by both parts of the equality in a scalar way

$$(f_j,\sum\alpha_i f_i)=(f_j,0)=0.$$

At the same time from orthogonality of vectors it is a provided formula

$$(f_j,\sum_i\alpha_i f_i)=\sum_i\alpha_i(f_j,f_i)=\alpha_j\left\|f_j\right\|^2=0.$$

Supposing that a sequence $\{f_i\}_{i=1}^n$ doesn't contain a zero vector, consequently $\left\|f_j\right\|\neq 0$, from where $\alpha_j=0$. The same is right for any other element from $\{f_i\}_{i=1}^n$, that is all $\alpha_i=0$.

If the sequence is $\{f_i\}$, it gives

$$(f_i,f_j)=\delta_{ij}=\begin{cases}0, & i\neq j,\\ 1, & i=j\end{cases} \qquad (4.19)$$

and is called orthonormalized. In this case norms of all vectors f_i equal one. We are to signify such system of vectors as $\{e_i\}$.

Procedure, making possible to make up an orthonormalized set from the countable set of linear independent vectors, is called the process of orthogonality of Grama–Schmidta and prescribes the following. Let in Hilbert space H the sequence of linear independent vectors $\{h_i\}_{i=1}^n$ be given. As far as a norm of orthonormalized vector equals one, to construct the first vector e_1 of orthonormalized sequence $\{e_i\}_{i=1}^n$, it is sufficient to equal it to

$$e_1 = \frac{h}{\|h_1\|}.$$

The next vector e_2 must be orthoganalized e_1. We are to make up the linear combination g_2 being orthogonal to e_1 from elements e_1 and h_2. Let $g_2 = h_2 - \alpha_{21}e_1$, where α_{21} is a number, then

$$(e_1, g_2) = (e_1, h_2 - \alpha_{21}e_1) = (e_1, h_2) - \alpha_{21}\|e_1\|^2 = (e_1, h_2) - \alpha_{21} = 0.$$

From this $\alpha_{21} = (e_1, h_2)$ or $g_2 = h_2 - (h_2, e_1)e_1$. We chose vector e_2 equal to $e_2 = \frac{g_2}{\|g_2\|}$. Identically we construct vector $e_3 = \frac{g_3}{\|g_3\|}$, where $g_3 = h_3 - \alpha_{32}e_2 - \alpha_{31}e_1$, and coefficients α_{31} and α_{32} are found from the condition of orthoganality g_3 by vectors e_1 and e_2. It gives

$$(g_3, e_1) = (h_3, e_1) - \alpha_{32}(e_2, e_1) - \alpha_{31}(e_1, e_1) = (h_3, e_1) - \alpha_{31} = 0,$$
$$(g_3, e_2) = (h_3, e_2) - \alpha_{32}(e_2, e_2) - \alpha_{31}(e_1, e_2) = (h_3, e_2) - \alpha_{32} = 0,$$

or $\alpha_{31} = (h_3, e_1)$, $\alpha_{32} = (h_3, e_2)$, and so on. All elements $\{e_i\}$ are orthonormalized and linear independent.

Having orthonormalized sequence $\{e_i\}_{i=1}^{\infty}$, we can write another vector f from Hilbert space H as a series by e_i:

$$f = \sum_{i=1}^{\infty} f^i e_i, \tag{4.20}$$

where f^i are some numbers. The coefficients f^i are easy to be found using $\{e_i\}$. Let's multiply both parts of equality scalar (4.20) by some vector e_j:

$$(f, e_j) = \left(\sum f^i e_i, e_j = \sum f^i (e_i, e_j) \right)$$

and take (4.19). We obtain that number f^j equals a projection f to a unit vector e_j:

$$f^j = (f, e_j). \tag{4.21}$$

The series (4.20), coefficients of which can be found according to the formula (4.21), is called the series of Fourier for *f*, and coefficients f^i are the coefficients of Fourier. Any segment of this series (its partial sum) has a particular feature of the best approximation, precisely: let us approximate some element $f \in H$ by function $\widehat{f}$, which is a linear combination *m* of vectors e_i:

$$f \approx \widehat{f} = \sum_{i=1}^{m} \alpha_i e_i, \tag{4.22}$$

where $\alpha_i \in \mathbb{R}^1$ are numbers.

The best approximation is the one for which

$$\left\| f - \widehat{f} \right\| = \min.$$

The norm of difference $\left\| f - \widehat{f} \right\|$ is a function $\alpha_i, i = 1,...,m$ and gets the minimum in case the coefficients α_i are the coefficients of Fourier and are fond by formula (4.21). In other words, the linear combination (4.22) presents the segment of this series. Actually,

$$\left\| f - \widehat{f} \right\|^2 = \left\| f - \sum \alpha_i e_i \right\|^2 = \left(f - f - \sum \alpha_i e_i, f \right).$$

Calculating the scalar product we get

$$\left\| f - \widehat{f} \right\|^2 = \|f\|^2 - 2\left(f, \sum \alpha_i, e_i \right) + \left(\sum \alpha_i, e_i, \sum \alpha_i, e_i \right).$$

The second member equals $2\sum \alpha_i (f, e_i) = 2\sum \alpha_i f^i$, where f^i are coefficients of Fourier calculated by (4.21), and the third in virtue of orthonormalization $\{e_i\}$ equals $\sum \alpha_i^2$. Further we have

$$\left\| f - \widehat{f} \right\|^2 = \|f\|^2 + \sum \alpha_i^2 - 2\sum \alpha_i f^i = \|f\|^2 + \sum (\alpha_i^2 - 2\alpha_i f^i).$$

Let's add and subtract the value $(f^i)^2$ from brackets. Then the second item may be taken as a square of difference

$$\left\| f - \widehat{f} \right\|^2 = \|f\|^2 - \sum (f^i)^2 + \sum (\alpha_i - f^i)^2.$$

This expression is evidently minimal at $a_i = f^i$, that has been to be proved. Thus, we get

$$\left\| f - \widehat{f} \right\| = \min \;\Rightarrow\; \left\| f - \widehat{f} \right\|^2 = \|f\|^2 - \sum (f^i)^2.$$

So as $\left\| f - f \right\|^2 \geq 0$, from the previous equality it goes

$$\sum_{i=1}^{m} (f^i)^2 \leq \left\| f \right\|^2 .$$

This inequality is right at any m and as $\left\| f \right\|^2$ are independent from m, the series of $\sum_{i=1}^{\infty} (f^i)^2$ converge, and we have the inequality of Bessel:

$$\sum_{i=1}^{\infty} (f^i)^2 \leq \left\| f \right\|^2 . \tag{4.23}$$

Geometrically it means that the sum of squares of projections of the vector f on orthogonal directions do not overcome the length square of the vector itself (non-finite dimensional analogue to Pythagorean theorem). If S is some subset of Hilbert space H that is $S \subset H$, the elements from H, orthogonal to all vectors from S, make up the set called the orthogonal addition to S (is signified as $S^{\perp}$); eventually $S \subset H^{\perp}$. The simple evident example in $\mathbb{R}^3$ is a line in perpendicular plane to it (Fig. 30): any vector on the line S is orthogonal to each vector lying on the perpendicular plane $S^{\perp}$ and vice versa.

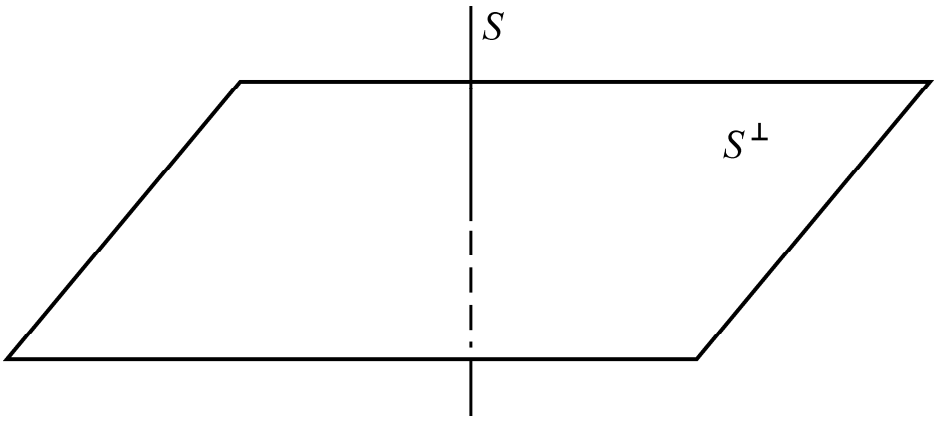


Fig. 30. The example of the set and its orthogonal addition

9. Basis of Hilbert space

In n-dimensional vector space any totality n of linear independent elements is a basis. Each vector can be presented unambiguously as a linear combination of basic vectors. Identically in the case of non-finite dimensional space V, we are to call the sequence of linear independent vectors $\{e_i\}_{i=1}^{\infty}$, $e_i \in V$ as basis, if any element f from V may be given unambiguously as convergent series

$$f = \sum_{i=1}^{\infty} f^i e_i .$$

The requirement of unambiguity states that the components of division f^i of zero vector equal zero.

The simple example of non-finite dimensional space is the analogue of space $\mathbb{R}^n$ – the space ℓ, the elements of which are non-finite sequences of numbers $x = (x^1, x^2, ..., x^n, ...)$. If we are to provide this space with Cartesian norm (to view the Banach space ℓ_2), the elements $e_1 = (1,0,0,...)$, $e_2 = (0,1,0,...)$ make basis. Any vector $x \in \ell_2$ can be written unambiguously as $\sum_{i=1}^{\infty} x^i e_i$, converging to *x* by the norm $\lim_{n\to\infty} \left\| x - \sum_{i=1}^{n} x^i e_i \right\|_2 = 0.$

As for the space ℓ_2 everything is quite simple, as each vector $x \in \ell_2$ is characterized by countable unit of number. In general case the separation property is very important. Let's take the set of real numbers or the space $\mathbb{R}^1$ we can point out the sub-set of whole numbers of $\mathbb{Z}$ and the sub-set of rational numbers $\mathbb{Q}$. Other numbers are irrational. Let's calculate an approximate value $\tilde{\alpha}$ of some number $\alpha \in {}^1$. Approximation will be good, if the error (difference module) doesn't overcome some value ε given in advance. If $\varepsilon \geq \frac{1}{2}$, it is evident that α may be approximated by numbers both from $\mathbb{Z}$ and $\mathbb{Q}$. However, while ε may be an arbitrary small, the set $\mathbb{Z}$, probably, is not sufficient. The set $\mathbb{Q}$ suits still very well. It happens because the whole numbers are situated on the numerical axis very rarely. It is said that $\mathbb{Z}$ is non-dense in $\mathbb{R}^1$. Is there the value of ε for $\mathbb{Q}$, for which rational numbers have ceased to approximate arbitrary sustention number? Actually there is not. Between any ration numbers situated at any closed distance there is always another rational number. We can make up such sequence of rational numbers, which will get to any substantion number (both rational and irrational). So the set $\mathbb{Q}$ is different from $\mathbb{R}^1$, because it doesn't contain all its limit points. In other words, $\mathbb{R}^1$ is a closing of $\mathbb{Q}$. In this case the set $\mathbb{Q}$ is said to be dense in the set $\mathbb{R}^1$. The definition of dense set is as follows. Let $S_1 \subset S_2 \subset B$, where *B* is Banach space. The set S_1 is called dense in S_2 if closing $\overline{S_1}$ in S_2 coincide with S_2. The presence of dense countable set $\mathbb{Q}$ in the set $\mathbb{R}^1$ allows to approximate any number from $\mathbb{R}^1$, using only numbers from $\mathbb{Q}$. Banach's space *B*, containing countable dense set is called separation

space. In other words, if B is separable, we have the set $S \subset B$, $S = \{f_n\}_{n=1}^{\infty}$, which for any $f \in B$ and $\varepsilon > 0$ gives $f_n \in S$ satisfying the inequality $\|f - f_n\| < \varepsilon$. There are some examples:

1. Generalizing the case $\mathbb{R}^1$, in which the set of ration numbers $\mathbb{Q}$ is dense, we view vector space $\mathbb{R}^n$. It also contains a countable dense set. It is evident, it will be the set of elements with rational coordinates.
2. The space $C^0([a,b])$ of the continuous functions defined on the segment $[a,b]$ with uniform norm is also separable, as the set of polynomials with the ration coefficients is dense in it.

If B is non-countable but separable and contains the dense sub-set S, any element B can be precisely approximated by the element of the countable set S as much as possible. This feature allows choosing a convenient basis in non-finite dimensional basis. The sequence of the vectors $\{e_i\}_{i=1}^{\infty}$ from the vector space V makes up basis of this space, if any element $f \in V$ can be presented in a single way as convergent by the serious norm $f = \sum_i f^i e_i$.

On the other hand, if we have the sequence of the non-zero linear independent vectors $\{g_i\}$, $g_i \in V$, all possible finite linear combinations (containing the finite number of members) of the vector form $S \subseteq V$. If $S = V$, the elements $\{g_i\}$ arrange the basis V. There are two questions:

1. In which case $S = V$?
2. How to choose the elements so that they could be linear independent?

In order to fulfill the equality $S = V$, the sequence of the vectors $\{g_i\}$ must be complete. It means completeness of vector system is mentioned, not of the vector space. In both cases the notion of completeness is that there must be a sufficient number of elements to fulfill some conditions. For the vector space there must be as many elements so that the limit of any elements sequence could be the element of this space (in our case the sequence must contain as many elements as to have $S = V$). We are going to view the feature of competence in detail a bit later. If $\{g_i\}_{i=1}^{\infty}$ generates the space S ($S = V$), V has countable dimension. It means that it is the space of ℓ type, elements of which are non-finite unites of numbers. If the dimension V is non-countable, $S \neq V$ and S can be dense in V. Then any vector $f \in V$ may be written as endless series:

$$f = \sum_i f^i g_i .$$

The answer to this question is connected with a feature of minimality of sequence. It is evident that, if $\{g_i\}$ is complete, the addition of new elements doesn't deprive it of this feature. Vectors become linear independent and cannot be basis. Minimal set of vector $\{g_i\}$, remaining the feature of competence, and will be linear independent and candidate for basis space *V*.

Features of minimality and completeness are established easily for orthogonal vectors. We have already proved above that the orthogonal vectors are linear independent, consequently, the orthogonal system is minimal. As for the terms for the orthogonal vectors, the completeness is defined in the following way. Let *H* be Hilbert space and $e_i \in H$ .The sequence is complete, if orthogonal addition to it equals 0 (it is the set with the only zero element). In other words, if $\{e_i\}$ is a complete system of vectors, there is no vectors from *H* different from zero, orthogonal to it, as far as from $(f, e_j) = 0$ for all $e_j \in \{e_i\}$. There is $f = 0$.

Let $H = \mathbb{R}^3$ and a vector $e_1 \in H$. It doesn't form a complete system of vectors. The subset, given by it, is $\mathbb{R}^1 \neq H$. As the orthogonal addition is non-zero and mean a plane, going through zero perpendicular to e_1. (Fig. 31, *a*). Let's choose in this plane the element e_2 and arrange a sequence $\{e_1, e_2\}$. It is incomplete and arranged space $\mathbb{R}^2$, the plane; the orthogonal addition is not 0 but a line perpendicular to this plane. (Fig. 31, *b*). Finally, when we choose the vector e_3 on this line, the sequence $\{e_1, e_2, e_3\}$ becomes complete and bare the space $\mathbb{R}^3 = H$; the orthogonal addition to it becomes 0 (Fig. 31, *c*).

The complete orthonormalized sequence of the vectors $\{e_i\}_{i=1}^{\infty}$ will be the basis of the space, if any element $f \in H$ is written as convergent to *f* of Fourier series $f = \sum_{i=1}^{\infty} f^i e_i$.

As the coefficients f^i are defined by the equality $f^i = (f, e_i)$ unambiguously, we must prove that, if $\{e_i\}$ is a complete orthoganalized sequence, the series $\sum (f, e_i) e_i$ converges to $f \in H$. Actually, as *H* is Hilbert space, it is completed by $\sqrt{(f, f)}$ norm (i. e. the limit of any Cauchy sequence of elements from *H* is itself an element from *H*). In other words, for any $\alpha_1, \alpha_2, ..., \alpha_m$ and any *m* there is an element $f \in H$, such that

$$\left\| f - \sum_{i=1}^{m} \alpha_i e_i \right\| = \| f - s_m \| < \varepsilon,$$

where ε is any given positive number (the sequence of elements getting to f is arranged not by the elements $\alpha_i e_i$ but by the partial sums $s_m = \sum_{i=1}^{m} \alpha_i e_i$, which are the elements from H). According to the best approximation of Fourier series, we have

$$\varepsilon > \left\| f - \sum_{i=1}^{m} \alpha_i e_i \right\| \geq \left\| f - \sum_{i=1}^{m} (f, e_i) e_i \right\|.$$

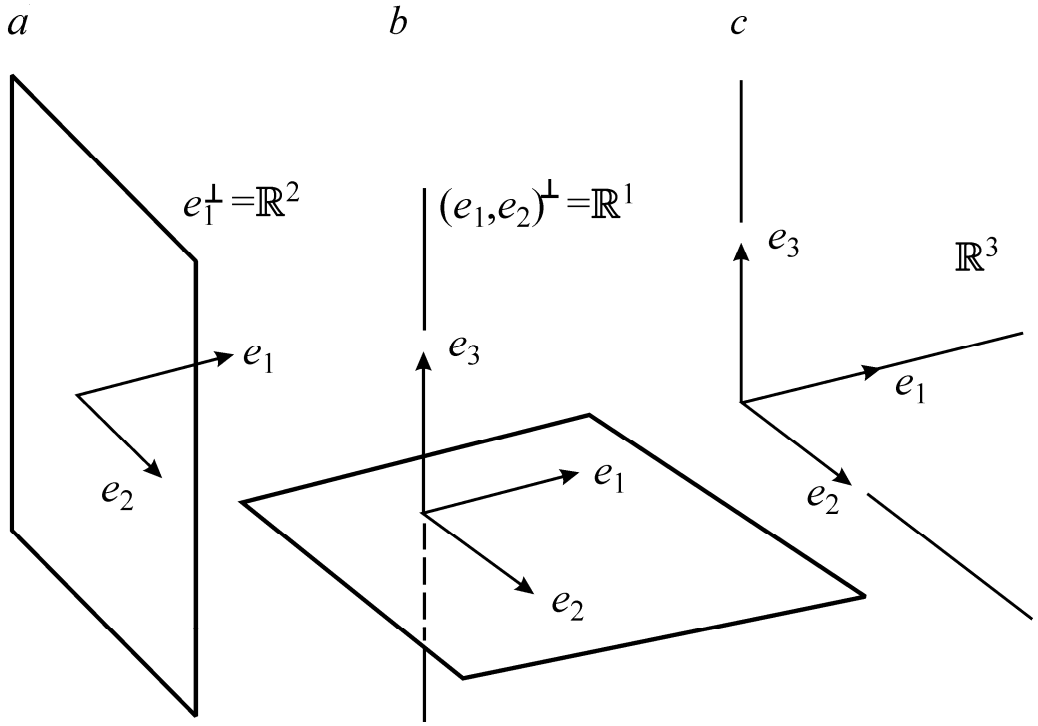


Fig. 31. Construction of complete system of orthogonal vectors in $\mathbb{R}^3$

Taking the limit by $m \to \infty$, at arbitrary ε we get

$$f = \sum_{i=1}^{\infty} (f, e_i) e_i .$$

So Fourier series don't converge to the element $f \in H$, therefore the sequence $\{e_i\}$ is the basis. As the sequence $\{e_i\}$ is orthonormalized, the basis arranged by it is called orthonormalized. Finally, it is possible to show that separable Hilbert space H has orthonormalized basis. Indeed, if H is separable, it contains a countable dense set $\{f_n\}$. Let's orthoganalize $\{f_n\}$ with the help of Gram–Schmidt process, dropping f_n, linear dependent on arranged orthonormalized vectors e_i. The found orthornormalized set $\{e_i\}$ can be finite,

then H is also finite-dimensional. It is complete in any case, as linear span arranged by vectors $\{f_n\}$ coincides with linear span arranged by the vectors $\{e_n\}$ (we dropped only linear dependent vectors); $\{f_n\}$ is dense in H. The vector orthogonal to all vectors from $\{e_n\}$ is a zero vector. Thus, the orthonormalized set $\{e_n\}$ is complete that arranges orthonormalized basis of separable Hilbert space H. Existence of the basis makes Hilbert space similar (that is isomorphic) to Euclidean space $\mathbb{R}^n$ (if H is finite-dimensional) or to the space ℓ_2 of finite sequences of numbers with Euclidean norm (if H is non-finite dimensional).

10. Linear operations

Let's consider the equation

$$\mathrm{A}f = g, \tag{4.24}$$

where f and g are the elements of the linear spaces V and W accordingly; A is the transformation of space V into space W. If V and W are the spaces of functions, the transformation A is called the operator. We are to fulfill the following: having $g \in W$ and the operator A, we define $f \in V$ if it is possible.

Let's specify the terms and definitions. Let V and W be linear spaces, A is an operator defined as $D(\mathrm{A}) \subset V$ with values in $R(\mathrm{A}) \subset W$. The operator A is one-for-one (injective or enclosure), it provides that there is only one relevant element $f \in D(\mathrm{A})$ for any $g \in R(\mathrm{A})$. If $R(\mathrm{A}) = W$ (the area of values is presented by the whole W-space), the operator A is surjective, or it is called superposition. If it is also one-for-one, it is bijective.

Let's view some examples of transformation $\mathrm{A}: \mathbb{R}^1 \to \mathbb{R}^2$:

1. $y = \mathrm{A}x \equiv x$. It is clear that $R(\mathrm{A}) = \mathbb{R}^1$, and A is injective, consequently, A is a bijection.
2. $y = \mathrm{A}x \equiv x^2$. The area of values $R(\mathrm{A}) = \{y : y \geq 0\}$, consequently, A is not surjective. It is also not injective, as it transforms the points $\pm x$ into point x^2.
3. $y = \mathrm{A}x \equiv x(x^2 - 1)$. The operator is surjective: $R(\mathrm{A}) = \mathbb{R}^1$ but not injective, it is not bijective as well, because points $0, \pm 1$ are transformed into point 0.
4. $y = \mathrm{A}x \equiv \exp(x)$. $R(\mathrm{A}) = \{y : y \geq 0\}$ – the operator is not surjective but it is one-for-one (injective).

Many operators are continuous. The operator is continuous at point f_0, if it transforms any sequence $\{f_n\}$, convergent to f_0, it is transformed into the sequence $\{\mathrm{A}f_n\}$, convergent to $\mathrm{A}f_0$, that is:

$$\mathrm{A} \lim f_n = \lim \mathrm{A}f_n.$$

If the operator is continuous at each point $f \in D(\mathrm{A})$, it is called continuous one. The simplest function to start their studies is a linear function. Identically the simplest operator is a linear operator (let's signify it as L), the very operator that makes the linear combination of elements

$$h = \alpha f + \beta g, \quad \alpha,\beta \in \mathbb{R}^1, \quad f, g, h \in D(\mathrm{L})$$

be transformed into linear combination

$$\mathrm{L}h = \mathrm{L}(\alpha f + \beta g) = \alpha \mathrm{L}f + \beta \mathrm{L}g, \quad \mathrm{L}f,\ \mathrm{L}g,\ \mathrm{L}h \in R(\mathrm{L}).$$

Clearly, $R(\mathrm{L})$ and $D(\mathrm{L})$ must be linear spaces to make the sense for definition itself. Let's view some examples of linear operators.

1. The particular problem is solving the system of linear algebraic equations

$$\sum_{j=1}^{n} a_{ij} x_j = b_i .$$

Comes to $\mathrm{L}f = g$, if we mark with L a matrix which elements are coefficients a_{ij}, with f element $(x_1, x_2, ..., x_n)$, and with g element $(b_1, b_2, ..., b_n)$. In that case L is an operator which reflects $\mathbb{R}^n$ to $\mathbb{R}^n$. It is linear and uninterrupted.

2. Consider Fredwholm's equation as an example of an integral operator.

$$f(x) - \int_a^b k(x,y) f(y) dy = g(x),$$

where $f, g \in C^0([a,b]); \quad k \in C^0([a,b] \times [a,b])$, k and g are given functions. If we define the operator L by the equation

$$\mathrm{L}f = \int_a^b k(x,y) f(y) dy,$$

then L is a linear operator $\mathrm{L}: C^0([a,b]) \to k \in C^1([a,b])$ and input equation will look like

$$f - \mathrm{L}f = g.$$

Let's illustrate that the operator is uninterrupted in the area of continuous functions with uniform norm. Let $\{f_n\}_{n=1}^{\infty}$ is so that $f_n \to f \in C^0([a,b])$. Consider the norm of difference

$$\left\| \mathrm{L}f_n - \mathrm{L}f \right\| = \left\| \int_a^b k(x,y)(f_n(y) - f(y)) dy \right\| = \max_x \left| \int_a^b k(x,y)(f_n(y) - f(y)) dy \right| \le$$

$$\le \max_x \int_a^b \left| k(x,y) \cdot \left| (f_n(y) - f(y)) dy \right| \right|.$$

It is clear that the integral at the last expression tends to zero when $f_n \to f$.

3. Among different differentiation operators, consider the operator of differentiation first.

$$\mathrm{L}f \equiv f'.$$

It is clear that the operator is defined only on the set of functions which have a derivative. It is linear but not uninterrupted and it is very to prove that. And to prove that we need to find one element for which continuity does not exist. Lets take the consecution $\{f_n\}_{n=1}^{\infty}$, where $f_n = \frac{1}{n}\sin(nx)$. In the area of continuous functions, defined on the $[a,b]$ with uniform norm, this consecution tends to zero. The consecution $\mathrm{L}f_n \equiv f_n' = \cos(nx)$ does not agrees, i. e. the continuity of the operator of differentiation at point $\frac{1}{n}\sin(nx)$ breaks.

4. Consider the differential equation

$$\alpha_0 f''(x) + \alpha_1 f'(x) + \alpha_2 f(x) = g(x),$$

where $x \in [a,b]$, $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}^1$. Let $g \in C^0([a,b])$, then f must belong to a set of functions, which have a second derivative, i. e. $C^2([a,b])$. Now we can consider a linear differentiation operator $\mathrm{L}: C^2([a,b]) \to C^0([a,b])$, where $\mathrm{L}f \equiv \alpha_0 f'' + \alpha_1 f' + \alpha_2 f$. With all this going on the equation takes the form of $\mathrm{L}f = g$.

5. Consider, at last, two special operators. Let $f \in V$, V is vector space. The operator I is so that $\mathrm{I}f = f$, $\forall f \in V$ is called a unit (identity) operator ($\mathrm{I}: V \to V$, $f \mapsto f$). Let W be one more vector space (probably concurrent with V). The operator O which reflects arbitrary elements $f \in V$ to zero element of the area W, i. e. $\mathrm{O}f = 0$, is called a zero operator $\mathrm{O}: V \to W$, $f \mapsto 0$).

Suppose, an operator equation is given

$$\mathrm{L}f = g, \quad f \in D(\mathrm{L}) \subset V, \quad g \in R(\mathrm{L}) \subset W,$$

where V and W are vector spaces. Formally, to solve the equation we need to find the operator L^{-1} (if is really possible) so that

$$\mathrm{L}^{-1}(\mathrm{L}f) = \mathrm{I}f = f \quad \text{or} \quad \mathrm{L}^{-1} \cdot \mathrm{L} = \mathrm{I}.$$

If such an operator does really exist, then it is called inverse operator. Which properties the operator L should have to make it possible to build a reverse operator L^{-1}?

Consider the following variants.

1. The operator L is injective or one-for-one. In that case L^{-1} does exist, it is a linear operator out of $R(\mathrm{L})$ to $D(\mathrm{L})$. The equation $\mathrm{L}f = g$ admits a solution for $g \in R(\mathrm{L})$ and does not admit a solution for $g \notin R(\mathrm{L})$.
2. The operator L is one-to-one. It is clear that L^{-1} exists for any $g \in W$ and there is only one solution for any $g \in W$.
3. The operator L is not injective, i. e. it is not one-for-one. In that case it is not possible to build an operator L^{-1} because there is at least one element $g \in R(\mathrm{L})$ for which there is more than one solution.

11. Linear operator matrix

If a linear operator is defined in the area with a basis then it is possible to record it as a matrix (finite matrix or infinite matrix, depending on the rank of space which it influences). In fact, let H be a separable Hilbert space, $\{e_i\}$ is orthonormal basis in it, and L is linear operator, $\mathrm{L}: H \to H$. The vector $f \in H$ is transformed by the operator L into some vector $g \in H$. In a similar manner the measuring vectors $e_i \in H$ are transformed into some vectors $\ell_i \in H$. As $\{e_i\}$ is a basis, the images ℓ_i of the measuring vectors e_i could be written through the own components in that basis:

$$\mathrm{L}e_i = \ell_i = \sum \ell_i^j e_j, \quad \ell_i^j \in \mathbb{R}^1, \tag{4.25}$$

where numbers ℓ_i^j are vector ℓ_i components in the basis $\{e_j\}$, i. e. $\ell_i = (\ell_i^1, \ell_i^2, ..., \ell_i^j, ...)$. If we put numbers in the form of a table, where i is a number of a line, and j is a number of a column then we have a squarte matrix Λ which is called operator matrix L in the basis $\{e_i\}$. The vectors f and g in the basis look like

$$f = \sum f^i e_i, \quad \mathrm{g} = \sum g^i e_i, \quad f^i, g^i \in \mathbb{R}^1, \tag{4.26}$$

where f^i, g^i are components of the vectors f and g in the basis $\{e_i\}$. Using a linearity of the operator and the equations (4.25), (4.26) we can write down the equation $\mathrm{L}f = g$ in a so-called *component form.*

$$\mathrm{L}f = \mathrm{L}\sum_i f^i e_i = \sum_i f^i \mathrm{L}e_i = \sum_i f^i \left(\sum_j \ell_i^j e_j \right) = \sum_j \left(\sum_i \ell_i^j f^i \right) e_j = \sum_k g^k e_k.$$

In order to find the expression of the nth components of vector g through the components of vector f and matrix Λ of operator L, it is sufficient to multiply both parts of the last sequent equality scalar by e_n. As far as basis is orthonormalized, we have

$$\sum_j \left(\sum_i \ell_i^j f^i \right) \delta_{jn} = \sum_k g^k \delta_{kn} \Rightarrow \sum_i \ell_i^n f^i = g^n.$$

Identically we can arrange the equations for some other components of vector g. As a result we have the system of algebraic linear equations

$$\begin{cases} \ell_1^1 f^1 + \ell_2^1 f^2 + \ell_3^1 f^3 + ... = g^1, \\ \ell_1^2 f^1 + \ell_2^2 f^2 + \ell_3^2 f^3 + ... = g^2, \\ \quad . \quad . \quad . \quad . \quad . \quad . \end{cases}$$

Introducing not incline symbols for vectors (in notions of linear algebra) $\mathrm{f} = (f^1, f^2, ...)^T$, $\mathrm{g} = (g^1, g^2, ...)^T$, where the upper index T signifies transposition. Finally we have a component form

$$\Lambda \mathrm{f} = \mathrm{g}.$$

So, the initial problem was converged to the problem of linear algebra.

Let's take $\mathrm{L}f \equiv f''(x)$, $f \in \mathbb{C}^2([-\pi, \pi])$. As basis one we choose vectors $\{e^{inx}\}_{n=1}^{\infty}$ (here i is an imaginary unit, $i^2 = -1$). Let's write the matrix of operator L. For this we are to find the element $\ell_n = \mathrm{L}e_n$.

$$\ell_n = \mathrm{L}e_n = (e^{inx})'' = -n^2 e^{inx}.$$

As $\ell_n = \sum_j \ell_n^j e_j$, it is clear, that

$$\ell_n^j = \begin{cases} -n^2, & n = j, \\ 0, & n \neq j. \end{cases}$$

So, the matrix of operator is a diagonal one.

12. Convergence method

A vigorous means of solving $\mathrm{L}f = g$ equation is the method, based on development of L_n^{-1} operators' sequence, which approach L^{-1} operator. This method is named the **convergence method**. Actually, L^{-1} operator is interpreted as a limit of certain sequence of operators. The sequence develops

while solving the exercise. Let read into these intuitive considerations; it requires defining of notion of operators' convergence and their proximity degree. One of the approaches is based on building of structure of Banach space operators and using of this banach space characteristics. Let V and W be vector spaces, L, L_1 and L_2 be linear operators, L, L_1, $L_2 : V \to W$. Now define $(\mathrm{L}_1 + \mathrm{L}_2)$ operator named ***sum of operators*** L_1 and L_2, such as

$$(\mathrm{L}_1 + \mathrm{L}_2) f = \mathrm{L}_1 f + \mathrm{L}_2 f, \quad \forall f \in V$$

and $(\alpha\mathrm{L})$, $\alpha \in \mathbb{R}^1$ operator, named ***product of operator by number***, such as

$$(\alpha\mathrm{L}) f = \alpha(\mathrm{L}f), \quad \forall f \in V, \quad \forall \alpha \in \mathbb{R}^1.$$

Denote zero operator as O, and received vector space of linear operators via $\mathcal{L}(V,W)$.

Provided that operators map V into V, then vector space of these operators may be denoted via $\mathcal{L}(V)$. The fact that L operator from $\mathcal{L}(V,W)$ has a backward operator, may be recorded as existence of $L^{-1} \in \mathcal{L}(V,W)$ operator (pay attention to the sequence order of V and W spaces). Now, when we have built the vector space of linear operators (which means that we can consider linear operator as a vector or a point of this space), it is naturally to try to introduce a norm to have an opportunity to estimate the proximity of points to one another. Now try to determine the norm of the operator, integrating the notion of the vector uniform norm (function). Consider $L : D(\mathrm{L}) \to R(\mathrm{L})$ operator, let $D(\mathrm{L}) \subset B$, $R(\mathrm{L}) \subset C$, where B and C are banach spaces (we take banach spaces here since it is necessary for the vectors of $D(\mathrm{L})$ and $R(\mathrm{L})$ to posess the norm). L operator maps f point from $D(\mathrm{L})$ into Lf point from $R(\mathrm{L})$. Name this operator a ***bound*** one, provided that $m < \infty$ number exists that

$$\|\mathrm{L}f\|_C \le m\|f\|_B, \quad \forall f \in D(\mathrm{L}) \tag{4.27}$$

(the first norm is in C vector space, the second norm is in B vector space). Provided that L operator is not limited in $D(\mathrm{L})$, then it is named ***unlimited***. It is obviously that for each $f \in D(\mathrm{L})$ point a certain m minimum value exists; (4.27) inequality is correct for this value (it is clear that if L operator is limited). The peak value of such m, when (4.27) inequality is correct in all $f \in D(\mathrm{L})$ points, we name L ***norm of operator*** and denote as $\|\mathrm{L}\|$.

In case of the linear operators the problem of their limitation is solved by the following theorem.

Theorem. *Linear operator L from B into C is limited in* $D(\mathrm{L})$ *if and only if it is continuous.*

Then by $\mathcal{L}(B,C)$ we imply the vector space of limited linear operators.

Consider an example. Let $\mathrm{L}:\mathbb{R}^n \to \mathbb{R}^n$ is *n*-dimensional linear operator, which, as we already know, in $\{e_j\}$ basis may be specified as a matrix

$$\mathrm{L}f = \sum_j \sum_i \ell_i^j f^i e_j.$$

Suppose that $\mathbb{R}^n$ possesses a certain norm $\|\cdot\|_\infty$, then

$$\|\mathrm{L}f\|_\infty = \max_j \left| \sum_{i=1}^n \ell_i^j f^i \right| \le \max_j \sum_i \left|\ell_i^j\right| \cdot \left|f^i\right| \le$$

$$\left(\max_j \sum_i \left|\ell_i^j\right| \right) (\max_i \left|f^i\right|) = m\|f\|_\infty,$$

where $m = \max_j \sum_i \left|\ell_i^j\right|$. Hence $\|\mathrm{L}\| \le m$. Disclose that $\|\mathrm{L}\| = m$. Try to work out such *f*, for which $\|\mathrm{L}f\|_\infty = m\|f\|_\infty$. As evident from *m* definition, there exist a certain *k* integer, when $m = \sum_i \left|\ell_i^k\right|$. As *f* take a vector with a norm equal to one and with such coordinates $\ell_i^k \Big/ \left|\ell_i^k\right|$. Then inequality chain turns into chain of equations

$$\|\mathrm{L}f\|_\infty = \max_j \left| \sum_i l_i^j f^i \right| = \left| \sum_i l_i^j f^i \right| = \sum_i \left| l_i^k \right| = m\|f\|_\infty,$$

and we have a required result.

The norm of operator introduced beyond specifies in $\mathcal{L}(B,C)$ a Banach space structure. Now disclose that $\mathcal{L}(B,C)$ space is complete by operator norm. Take Cauchy operators consequence $L_n \in \mathcal{L}(B,C)$ and disclose that the limit of this consequence is also situated in $\mathcal{L}(B,C)$.

1. As L_n is Cauchy consequence, then $\|\mathrm{L}_n - \mathrm{L}_m\| \to 0$ at $n,m \to \infty$, hence the consequence $\{\mathrm{L}_n f\}$, where $f \in B$ and $\mathrm{L}_n f \in C$, also forms Cauchy consequence since

$$\|\mathrm{L}_n f - \mathrm{L}_m f\| \le \|\mathrm{L}_n - \mathrm{L}_m\| \cdot \|f\|.$$

But if *C* is complete, then $\mathrm{L}_n f \to g \in C$.

2. Let $\mathrm{L}f = \lim\limits_{n\to\infty} \mathrm{L}_n f = g$. It is obviously that L is a linear operator. Make certain that it is limited. Indeed,

$$\|\mathrm{L}f\| = \lim_{n\to\infty} \|\mathrm{L}_n f\| \le \lim_{n\to\infty} \|\mathrm{L}_n\| \cdot \|f\|.$$

But since

$$\|\mathrm{L}_n - \mathrm{L}_m\| \ge \|\|\mathrm{L}_n\| - \|\mathrm{L}_m\|\| \text{ and } \|\mathrm{L}_n - \mathrm{L}_m\| \to 0,$$

then $\|\mathrm{L}_n\|$ also forms Cauchy consequence in $\mathbb{R}^1$. Denote $\lim\limits_n \|\mathrm{L}_n\| = \alpha$, then

$$\|\mathrm{L}f\| \le \alpha \|f\|.$$

3. Disclose that $\mathrm{L}_n \to \mathrm{L}$, i. e. that

$$\|\mathrm{L}_n - \mathrm{L}\| \to 0.$$

As $\{\mathrm{L}_n\}$ is Cauchy consequence, then for any $\varepsilon > 0$ N number may be found, such number that if $m,n > N$, then $\|\mathrm{L}_n - \mathrm{L}_m\| < \varepsilon$. Then for $m,n > N$ we have $\|\mathrm{L}_n f - \mathrm{L}_m f\| \le \varepsilon \|f\|$ and

$$\|\mathrm{L}_n f - \mathrm{L}f\| = \lim_{m\to\infty} \|\mathrm{L}_n f - \mathrm{L}_m f\| \le \varepsilon \|f\|,$$

which is correct for any $f \in B$. Hence $\|\mathrm{L}_n - \mathrm{L}\| \le \varepsilon$, and because ε is an arbitrary value, then

$$\lim_{n\to\infty} \|\mathrm{L}_n - \mathrm{L}\| = 0.$$

Let's turn to backward operator construction with the help of the convergence method. The idea of the method is quite simple. Introduce L operator as

$$\mathrm{L} = \mathrm{I} - \mathrm{M},$$

where I is an identity operator, and M is a certain linear operator, then

$$\mathrm{L}f \equiv (\mathrm{I} - \mathrm{M})f = g,$$

or

$$f = g + \mathrm{M}f. \tag{4.28}$$

To determine f let's do the following integration. As f zero approximation let's take g point, i. e.

$$f_0 = g.$$

Then, substituting zero approximation in the (4.28) right-hand member, we have the first approximation

$$f_1 = g + \mathrm{M}f_0.$$

Continuing this operation we have an opportunity to get any n approximation in the following *recurrence* formula

$$f_n = g + \mathrm{M}f_{n-1},\ f_0 = g,\ \ n \geq 1. \tag{4.29}$$

Resolve the formula (4.29) by substituting the previous approximation:

$$f_n = g + \mathrm{M}f_{n-1} = g + \mathrm{M}(g + \mathrm{M}f_{n-2}) =$$

$$= (\mathrm{I} + \mathrm{M})g + \mathrm{M}(\mathrm{M}f_{n-2}). \tag{4.30}$$

We get some new operator in the last expression that we will denote as *the grade of operator.* This is the particular case of *the product of operators*. If $\mathrm{L}: V \to U,\ \mathrm{N}: U \to W$ are the linear operators and V, U, W are vector spaces then the *product of operators* NL is presented by the operator $\mathrm{P}: V \to W$, making the element $\boldsymbol{v} \in V$ relevant to the element

$$\boldsymbol{w} = \mathrm{N}(\mathrm{L}, \boldsymbol{v}) \in W.$$

The applicable domain of $D(P)$ is the set of elements of $\boldsymbol{v} \in D(\mathrm{L})$ such as $\mathrm{L}\boldsymbol{v} \in D(\mathrm{N})$.

If L and N are bounded operators and V, U, W are normalized spaces, then operator $\mathrm{P} = \mathrm{NL}$ is also limited and

$$\|\mathrm{P}\| \leq \|\mathrm{N}\| \cdot \|\mathrm{L}\|. \tag{4.31}$$

In fact, for any $\boldsymbol{v} \in D(\mathrm{L})$, and thus $\boldsymbol{v} \in D(\mathrm{P})$ by definition of the operator norm we have

$$\|\mathrm{P}\boldsymbol{v}\| \leq \|\mathrm{P}\| \cdot \|\boldsymbol{v}\|,\ \ \|\mathrm{L}\boldsymbol{v}\| \leq \|\mathrm{L}\| \cdot \|\boldsymbol{v}\|.$$

In a similar manner, for any $u \in D(\mathrm{N})$

$$\|\mathrm{N}u\| \leq \|\mathrm{N}\| \cdot \|u\|.$$

Since $\mathrm{L}\boldsymbol{v} \in D(\mathrm{N})$, then we obviously get

$$\|\mathrm{P}\boldsymbol{v}\| = \|\mathrm{N}(\mathrm{L}\boldsymbol{v})\| \leq \|\mathrm{N}\| \cdot \|\mathrm{L}\boldsymbol{v}\| \leq (\|\mathrm{N}\| \cdot \|\mathrm{L}\|)\|\boldsymbol{v}\|.$$

Hence, the inequation (4.31) follows according to the definition of the operator norm. If $\mathrm{N} = \mathrm{L}$, then the product NN we will identify as ***the grade of operator*** and denote N^2. We can identify just as any operator integral power. It is reasonable to consider $\mathrm{N}^0 = \mathrm{I}$. Thus, the expression (4.30) for f_n we can write through the grade of the operator M:

$$f_n = (\mathrm{I} + \mathrm{M})g + \mathrm{M}^2 f_{n-2} = (\mathrm{I} + \mathrm{M})g + \mathrm{M}^2(g + \mathrm{M}f_{n-3}) =$$

$$= (\mathrm{I} + \mathrm{M} + \mathrm{M}^2)g + \mathrm{M}^3 f_{n-3} = \ldots = \left(\mathrm{I} + \sum_{i=1}^{n-1} \mathrm{M}^i\right) g + \mathrm{M}^n f_0,$$

or taking in to the account that $f_0 = g$, and $\mathrm{M}^0 = \mathrm{I}$,

$$f_n = \sum_{i=1}^{n-1} \mathrm{M}^i g.$$

We await that $\lim\limits_{n\to\infty} f_n = f$ and

$$f = \sum_{i=0}^{\infty} \mathrm{M}^i g. \tag{4.32}$$

If to the operator L the backward operator L^{-1} exists then $f = \mathrm{L}^{-1} g$ and this means

$$\mathrm{L}^{-1} = (\mathrm{I} - \mathrm{M})^{-1} = \sum_{i=0}^{\infty} \mathrm{M}^i . \tag{4.33}$$

All these, however, are the formal constructions as we have no confidence in that the series (4.32) converges. Let's see, on what conditions this happens that is in what case the convergence method works. Let B is the banach space and $L \in \mathcal{L}(B)$, and this means that $\mathrm{M} \in \mathcal{L}(B)$. There is a need to clarify when $\mathrm{L}^{-1} = (\mathrm{I} - \mathrm{M})^{-1}$ exists or in other words $(\mathrm{I} - \mathrm{M})^{-1} \in \mathcal{L}(B)$. The latter denotes that the operator L has the bounded backward operator that is $\left\|\mathrm{L}^{-1}\right\| < \infty$. From the formula (4.33) we have

$$\left\|\mathrm{L}^{-1}\right\| = \left\|(\mathrm{I} - \mathrm{M})^{-1}\right\| = \left\|\sum_n \mathrm{M}^n\right\| \le \sum_n \left\|\mathrm{M}^n\right\| \le \sum_n \|\mathrm{M}\|^n < \infty.$$

The series $\sum\limits_n \|\mathrm{M}\|^n$ is already numerical (that is to say, the geometric series), relatively to which it is known that it comes together if $\|\mathrm{M}\| < 1$. This is the condition for convergence of the series (4.32). It is easy to establish, that at $\|\mathrm{M}\| < 1$ the consequence $\{f_n\}$ has the bound *f*. From the equation

$$f = g + \mathrm{M}f \quad \text{and} \quad a_n = g + \mathrm{M}f_{n-1}$$

it follows that

$$\|f_n - f\| = \|\mathrm{M}f_{n-1} - \mathrm{M}f\| \le$$

$$\le \|\mathrm{M}\| \cdot \|f_{n-1} - f\| \le \ldots \le \|\mathrm{M}\|^n \cdot \|f_0 - f\|. \tag{4.34}$$

Whence for $\|\mathrm{M}\| < 1$ it follows that

$$\lim_{n\to\infty}\left\|f_n - f = 0\right\| \quad \text{или} \quad \lim_{n\to\infty} f_n = f.$$

The inequality (4.34) characterizes the iterative procedure convergence rate of the convergence method. But this assessment should be reformulate since *f* is unknown. From the equation

$$(\mathrm{I} - \mathrm{M})f = g$$

it runs out that

$$\left\|f_0 - f\right\| = \left\|f_0 - g - \mathrm{M}f\right\| = \left\|f_0 - g - \mathrm{M}f_0 + \mathrm{M}f_0 - \mathrm{M}f\right\| \le$$
$$\left\|f_0 - g - \mathrm{M}f_0\right\| + \left\|\mathrm{M}\right\| \cdot \left\|f_0 - f\right\| = \left\|\mathrm{L}f_0 - g\right\| + \left\|\mathrm{M}\right\| \cdot \left\|f_0 - f\right\|.$$

Whence

$$\left\|f_0 - f\right\| \le \frac{1}{1 - \left\|\mathrm{M}\right\|}\left\|\mathrm{L}f_0 - g\right\|.$$

Substituting this result in (4.34) we finally get

$$\left\|f_n - f\right\| \le \frac{\left\|\mathrm{M}\right\|^n}{1 - \left\|\mathrm{M}\right\|}\left\|\mathrm{L}f_0 - g\right\|. \qquad (4.35)$$

The value $\left\|\mathrm{L}f_0 - g\right\|$ in the inequality (4.35) is the norm of zero approximation residual and it is easy to calculate it. That is why if you know $\left\|\mathrm{M}\right\|$ this is worth nothing to calculate the number of iterations that is necessary to achieve the given precision.

It would be interesting to examine the action of the operator M and the condition $\left\|\mathrm{M}\right\| < 1$ from the point of geometry. Let L, and consequently M reflect banach space inside: $B \to B$ (otherwise we cannot put in the operator grade). Then $f_0, f_1, ..., f_n, ... \in B$ are the points of one and the same space *B*. The distance between two neighboring points of the consequence $\{f_n\}$ is determined by their norm of difference $\left\|f_n - f_{n-1}\right\|$. Let's see how is this distance changing with *n* rising or what is the same under the influence of the operator M. Express the distance between the points f_{n+1} and f_n through the distance between the points f_n and f_{n-1}. As

$$f_n = g + \mathrm{M}f_{n-1}, \;\; f_{n+1} = g + \mathrm{M}f_n,$$

we get

$$\left\|f_{n+1} - f_n\right\| = \left\|\mathrm{M}f_n - \mathrm{M}f_{n-1}\right\| \le \left\|\mathrm{M}\right\| \cdot \left\|f_n - f_{n-1}\right\|.$$

If $\left\|M\right\| < 1$, then it is obvious that the distance between two consequent points contracting with *n* rising, so the points are approaching. At $\left\|M\right\| \ge 1$ the

distances between the neighboring points can rise. If $\|M\| < 1$, then the operator M is denoted as the contraction operator. The precise definition is as follows. It is said that the operator A satisfies the Lipshic's conditions on $D(\mathrm{A})$ with the Lipshic's constant q, if it exists such $q < \infty$, that

$$\|\mathrm{A}f - \mathrm{A}g\| \le q\|f - g\|, \ \forall f, g \in D(\mathrm{A}).$$

If $q < 1$, then the operator A is denoted as the contraction operator. This term is very evident, what is confirmed by geometry illustration of M operator action. In conclusion, we consider the application of method of sequence approximations to solving the system of algebraic linear equations and limits, coming from $\|M\| < 1$. Let the following equation be solved

$$\mathrm{A}\mathrm{f} = \mathrm{g}, \tag{4.36}$$

where the operator $\mathrm{A}: \mathbb{R}^n \to \mathbb{R}^n$, usually written as square matrix $n \times n$, and $\mathrm{f}, \mathrm{g} \in \mathbb{R}^n$ are vectors of unknown values and right parts. Let's take A as (I–M), where I is a unit matrix, and M is some new matrix (operator). Let a_{ij} are elements of matrix A, then numbers $(\delta_{ij} - a_{ij})$ are elements of matrix M (δ_{ij} is Chroneker's symbol). The equation (4.36) is

$$(\mathrm{I} - \mathrm{M})\mathrm{f} = \mathrm{g}.$$

If $\|M\| < 1$, we can arrange the international procedure

$$\mathrm{f}_{n+1} = \mathrm{g} + \mathrm{M}\mathrm{f}_n$$

and calculate $\mathrm{f} = \lim\limits_{n\to\infty} \mathrm{f}_n$. This is well known method to solve the system of linear equations. It is Jacoby's method. We are to find out what the limits for elements of matrix A are to be implied to by requirement $\|M\| < 1$. We are going to work with norm $\| \cdot \|_\infty$, then

$$\|\mathrm{M}\| = \max_j \sum_{i=1}^{n} |\delta_{ij} - a_{ij}|.$$

As far as $\|M\| < 1$, $\sum\limits_{i=1}^{n} |\delta_{ij} - a_{ij}| < 1$ for any j is

$$\sum_{i=1}^{n} |\delta_{ij} - a_{ij}| = \sum_{i \ne j} |a_{ij}| + |1 - a_{jj}| < 1.$$

We have

$$\sum_{i \neq j} \left| a_{ij} \right| < 1 - \left| 1 - a_{jj} \right| < \left| a_{ij} \right|,$$

or each diagonal element by module must be more than the sum of other elements in given column. This is a well known condition of Jacob's method convergence condition of diagonal dominance of matrix A.

13. Spectral radius of operator

Method of sequent approximations works in case when $\|M\| < 1$ or the operator M is an operator of contraction. If $\|M\| > 1$, is this method applicable or not? It is not necessary. If the choice was not successful and there could be another more appropriate norm to make our operator the operator of contraction. The norm should be searched for among the sets of equivalent norms. Referring to the equivalent norms, the convergent sequences remain convergent, as for closed set will be closed, open will be open and so on. If operator M in norm $\| \cdot \|$ satisfied Lipshits condition, then in the equivalent norm $\| \cdot \|_*$ it will satisfy it.

The final decision about convergence of the method of successive approximations can be made only if the equivalent norm is found, where operator M is contraction operator, or if minimum possible equivalent norm is made, and M is still not a contraction operator. Minimum norm is closely connected with so-called *spectral radius* of the operator. The procedure of the operator L^{-1} existence criterion establishment was the following: we wrote down the series of intensifying each other inequalities until we got geometric progression. Here everything became obvious and, unfortunately, very rough. It is possible to obtain more precise results if not going beyond this:

$$\left\| \mathrm{L}^{-1} \right\| = \left\| (\mathrm{I} - \mathrm{M})^{-1} \right\| = \left\| \sum_n \mathrm{M}^n \right\| \leq \sum_n \left\| \mathrm{M}^n \right\|. \qquad (4.37)$$

Let us implement the root test to the last series and calculate the limit

$$r = \lim_{n \to \infty} \sqrt[n]{\left\| \mathrm{M}^n \right\|}.$$

r is called the *spectral radius* of operator M.

Since $\left\| \mathrm{M}^n \right\| \leq \left\| \mathrm{M} \right\|^n$, it is obvious that $r \leq \left\| \mathrm{M} \right\|$. If $r < 1$, then series (4.37) converge, so operator L^{-1} exists; $r > 1$, then series (4.37) diverge and operator L do not have backward operator. If $r = 1$ we can't say anything about series convergence.

Knowing the spectral radius r we can make the equivalent norm in space B at which the norm of the linear operator M is arbitrary close to its spectral radius. Now it is easy to formulate the convergence criterion of the method of successive approximations in the terms of the operator M spectral radius, namely: if spectral radius $r(\mathrm{M}) < 1$, then successive approximations converge to the equation solution. Since we can make the equivalent norm $\|\mathrm{M}\|_*$, in such a way that

$$\|\mathrm{M}\|_* \le r + \varepsilon, \quad \forall \varepsilon > 0,$$

then we can rewrite the obtained estimation of the rate of convergence (4.35) in the following way

$$\|f_n - f\| \le \frac{(r(\mathrm{M}) + \varepsilon)^n}{1 - (r(\mathrm{M}) + \varepsilon)} \|\mathrm{L}f_0 - g\|.$$

If for the successive approximations the following inequation is fair

$$\|f_n - f\| \le cq^n, \quad c = const,$$

then they say that approximations f_n converge to f at rate of geometric progression with q ratio. Thus, for the linear equation

$$\mathrm{L}f \equiv (\mathrm{I} - \mathrm{M})f = g$$

successive approximations converge to the equation solution at rate of geometric progression, which ratio is arbitrary close to the spectral radius $r(\mathrm{M})$. All formulated criteria and estimations presuppose the fact that we know how to calculate the operator norm or its spectral radius. However we have yet only such a non-constructive definition of the norm as maximum value of the constant m, at which for any $f \in B$

$$\|\mathrm{M}f\| \le m\|f\|,$$

or in other words

$$\|\mathrm{M}\| = \max_{f \in B} \frac{\|\mathrm{M}f\|}{\|f\|}.$$

Let's consider the equation

$$\mathrm{M}f = \lambda f, \tag{4.38}$$

where λ is some number (actually complex number). If we take the norm from the right and the left part we will get:

$$\|\mathrm{M}f\| = |\lambda| \cdot \|f\| \le \|\mathrm{M}\| \cdot \|f\|.$$

Hence, $\|\mathrm{M}\|$ is no less the maximum $|\lambda|$, which satisfy the equation (4.38). Thus, the task of the operator norm estimation is to find λ values, at which the equation (4.38) or the following equation

$$(\lambda I - \mathrm{M}) f = 0 \qquad (4.39)$$

has nontrivial, i. e. non-zero, solution. The later means that at these λ operator $(\lambda I - \mathrm{M})$ cease to be injective and, consequently, there is no backward operator $(\lambda I - M)^{-1}$. Indeed, if $f \neq 0$ is a solution for (4.39), then α*f* is also a solution for (4.39) for any real or complex α. As a result one-oneness is disturbed. The theory turns to be simpler in complex case, so we will consider $\lambda \in \mathbb{C}$. Those λ values, at which $(\lambda I - \mathrm{M})^{-1} \in \mathcal{L}(B)$, i. e. $(\lambda I - \mathrm{M})^{-1}$ is linear bounded operator, are called *regular* and form so-called *resolvent set* ρ(M) of M operator, and the operator $\mathrm{R}_\lambda = (\lambda \mathrm{I} - \mathrm{M})^{-1}$ itself for $\lambda \in \rho(\mathrm{M})$ is called *resolvent* of M operator. All other λ which do not enter the resolvent set are called operator M spectrum and are marked in the following way: σ(M). Thus, complex plane is divided into two parts: resolvent set ρ(M) and spectrum σ(M). λ values, at which the equation (4.38) have a solution, are called *proper values of linear operator* M. Vectors *f* being a solution for (4.38) (accurate to constant factor) at some proper λ value are called *proper vectors of operator* M, responding to the given λ. If *B* is function space, then proper values are often called *proper functions*. It is not difficult to show that set of proper vectors for some λ forms vector space. At the same time vector space is a subset of space *B*. Thus, proper vectors responding to some proper λ value form vector subset in *B,* which is called *proper subset* responding to λ. Set σp(M) of all proper values of operator M is called its *poin*t (or *discrete*) spectrum. For finite-dimensional operators spectrum coincides with point spectrum. If M is an infinite-dimensional operator, then at some $\lambda \in \sigma(\mathrm{M})$ operator $(\lambda I - \mathrm{M})^{-1}$ can exist but it will be unbounded. If with all this domain of operator $(\lambda I - \mathrm{M})^{-1}$ is dense in *B* (and this is carried out, as a rule), then such λ form so-called *continuous spectrum* $\sigma_c(\mathrm{M})$ of operator M. Above defined resolvent can be considered as a reflection $\mathbb{C} \to \mathcal{L}(B)$ of complex plane on vector space $\mathcal{L}(B)$, i. e. operator-valued function of complex argument. Then spectrum σ(M) is scar set of function R_λ. The task (4.39) to find proper values of operator M is called *proper value problem* or *spectral problem*.

Examples:

1. Let operator $\mathrm{M}: C^0([a,b]) \to C^0([a,b])$ is defined by the formula

$$\mathrm{M}f(x) = \alpha(x) f(x),$$

where $\alpha(x)$ is some given continuous function. Then

$$(\lambda \mathrm{I} - \mathrm{M}) f(x) = (\lambda - \alpha(x)) f(x),$$

or

$$(\lambda \mathrm{I} - \mathrm{M})^{-1} = \frac{1}{\lambda - \alpha(x)}(x).$$

All λ turn to be the spectrum of operator M, for which $\lambda - \alpha(x) = 0$ at some $x \in [a,b]$, i. e. all values of function $\alpha(x)$. Since $\alpha(x) \in C^0$, then spectrum is continuous. There is no point spectrum because there are no proper values.

2. Let $\mathrm{M} : \mathbb{R}^n \to \mathbb{R}^n$ represents by $n \times n$ matrix. Then its spectrum is pure point consisting of finite number of proper values. There is no continuous spectrum, because if $(\lambda I - \mathrm{M})^{-1}$ exists, it is bounded.

REFERENCES

1. Ascher U. M., Mattheij R.M.M., Russell R. D. Numerical solution for boundary value problems for ordinary differential equations, Prentice-Hall. 1988.
2. Arnold V. I. Mathematical Methods of Classical Mechanics, 2nd ed. – New York: Springer, 1989.
3. Meyer G. M., Continuous orthonormalization for boundary value problems // J. Comput. Phys., 62, 248–262 (1986).
4. Schmitt K. and Thompson R. C. Nonlinear Analysis and Differential Equations. An Introduction. – Utah State University, 2004.
5. Allgower E. and Georg K. Numerical Continuation Methods, An Introduction. – New York: Springer-Verlag, 1990.
6. Hirsch M. W. and Smale S. Differential Equations, Dynamical Systems, and Linear Algebra. – San Diego: Academic Press, 1989.
7. Coppel W. Stability and Asymptotic Behavior of Differential Equations. – Boston: Heath, 1965.
8. Dang H. and Schmitt K., Existence and Uniqueness Theorems for Nonlinear Boundary Value Problems // Rocky Mtn. J. Math., 24, 77–91 (1994).
9. Peitgen H. O. and Schmitt K., Global Analysis of Two-Parameter Elliptic Eigenvalue Problems // Trans. Amer. Math. Soc., 283, 57–95 (1984).
10. Royden H. L. Real Analysis, 3rd ed. – New York: Macmillan Publishing Co., 1988.
11. Hartman P. Ordinary Differential Equations. – Boston: Birkhauser, 1982.
12. Achenbach J. Wave Propagation in Elastic Solids.– North Holland, 1973.
13. Gurtin M. E. An Introduction to Continuum Mechanics. – Academic Press, 1981.
14. Bryant R. L., Griffiths P. A., Grossman D. A. Exterior Differential Systems and Euler-Lagrange Partial Differential Equations. – Chicago, IL: University of Chicago Press, 2003. – 205 p.
15. Robert L. Bryant, Phillip A. Griffiths, and Lucas Hsu, Hyperbolic Exterior Differential Systems and Their Conservation Laws, I,II // Selecta Math. (N.S.) 1 (1995).
16. Christodoulou D. Global Solutions of Nonlinear Hyperbolic Equations for Small Initial Data // Comm. Pure Appl. Math. 39 (1986).
17. Chicone C. Ordinary Differential Equations with Applications. – New York: Springer, 1999.
18. Robinson C. Dynamical Systems: Stability, Symbolic Dynamics, and Chaos. – Boca Raton: CRC Press, 1995.
19. Perko L. Differential Equations and Dynamical Systems, 2nd ed. – New York: Springer, 1996.

20. Bakhvalov N. S., Zhidkov N. P., Kobel'kov G. M. Calculus of Approximations. – [in Russian] M.: Nauka, 1987. – 600 p.
21. Babenko K. I. Basis of Numeric Analysis. – M.: Nauka, 1986. – 374 p.
22. Volkov E. A. Calculus of Approximations. – M.: Nauka, 1987. – 248 p.
23. Demidovich B. P., Maron I. A., Shuvalova E. Z. Numerical Methods of Analysis. – M.: GIFL, 1963. – 400 p.
24. Zenkevich O., Morgan K. Finite Elements and Approximation. – M.: Mir, 1986. – 149 p.
25. Kalitkin N. N. Calculus of Approximations. – M.: Nauka, 1978. – 512 p.
26. Kolmogorov A. N., Fomin S. V. Elements of Function Theory and Functional Analysis. – M.: Nauka, 1968. – 250 p.
27. Rikhtmayer R. Principles of Modern Mathematical Physics. Ch.1. – M.: Mir, 1982. – 312 p.
28. Turchak L. I. Basis of Calculus of Approximations. – M.: Nauka, 1987. – 250 p.
29. http://www.arxiv.org/PS_cache/math/pdf/0207/0207039v1.pdf.
30. http://www.mat.univie.ac.at/~gerald/ftp/book-ode/ode.pdf.

CONTENTS

Educational Edition

Томский политехнический университет

ЛОПАТКИН Сергей Анатольевич

РЕЙЗЛИН Валерий Израилевич

ДОПОЛНИТЕЛЬНЫЕ ГЛАВЫ МАТЕМАТИКИ

Учебное пособие

Издательство Томского политехнического университета, 2008

На английском языке

Science editor
Doctor of Physics and Mathematics,
Professor — *V.V. Lopatin*

Typesetting — *K.S. Chechel'nitskaya*

Cover design — *O.Yu. Arshinova*
O.A. Dmitriev

Signed for the press 22.12.2008. Format 60x84/16. Paper “Snegurochka”.
Print XEROX. Arbitrary printer’s sheet 7.33. Publisher's signature 6.63.
Order 896. Size of print run 200.

Tomsk Polytechnic University
Quality management system
of Tomsk Polytechnic University was certified by
NATIONAL QUALITY ASSURANCE on ISO 9001:2000

TPU PUBLISHING HOUSE. 30, Lenina Ave, Tomsk, 634050, Russia