1. BASIC CONCEPTS

Nowadays, numerical methods are a powerful mathematical tool for solving many scientific and technical problems. This is due to both the impossibility to get an exact analytical solution in the vast majority of cases and the rapid development of computers. Despite the existence of numerous application software packages designed to solve problems in various areas of engineering, scientists, engineers, and technicians must also understand the essence of the basic numerical methods and algorithms used, since often the interpretation of calculation results is non-trivial and requires special knowledge regarding the features of the methods used.

1.1. ERROR STRUCTURE

The use of numerical methods inevitably leads to the emergence of various types of computational and algorithmical errors. In this regard, it is important to understand the error structure when solving specific problems. There are absolute and relative errors.

Definition. Suppose a is an exact, generally speaking, unknown numerical value of a certain variable, and $\bar{a}$ is a known approximate numerical value of this variable, then the number

$$\Delta(a) = |a - \bar{a}|$$

is called the *absolute error* of the number a, and the ratio of the measurement absolute error to the modulus of the reference value of this variable

$$\delta(a) = \frac{\Delta(a)}{|a|}$$

is called the *relative error*.

Not only the exact numerical value, but also any approximate numerical value of this variable can act as a reference value. The relative error is dimensionless; its numerical value can be specified, for example, in percent. It can be shown that absolute errors are summed in addition and subtraction, while relative errors – in division and multiplication.

There are four sources of errors gained as a result of the numerical solution to a problem:

1) physical and mathematical models of the object under study;
2) initial data included in the problem description;
3) approximation of the numerical method used for the solution;
4) rounding errors, arising in the process of calculation using a computer.

The first two sources of errors lead to a so-called inherent error. This error may be present even if the accurate solution to the formulated problem is found.

A method error arises from the fact that the exact mathematical operator and the initial data, in particular, the initial and boundary conditions, are replaced by approximate ones according to certain rules. So, derivatives are replaced by their difference analogues, integrals – by sums, functions – by special polynomials. Also, when solving many problems, infinite iteration processes are built, which naturally stop after a finite number of iterations.

Usually, it is possible to estimate and control method errors. Such estimation for some methods will be presented below. The method error should be chosen so that it is no more than one order lower than the inherent error.

A rounding error (or computational error) occurs due to the fact that calculations are carried out with a finite number of significant digits, i.e. rounding is performed during calculations.

Definition. Significant digits in the record of an approximate number are as follows:

1) all nonzero digits;
2) zeros contained between nonzero digits;
3) zeros at the end of the number, which are representatives of the decimal places preserved during rounding.

Rounding is carried out according to the following rules:

1) if there is a digit smaller than five in the highest order to be discarded, then the contents of the stored orders do not change;

2) if there is a digit greater than five in the highest order to be discarded, then one (1) is added to the lowest order to be stored;

3) if there is a digit five in the highest order to be discarded and there are nonzero digits among other discarded digits, then one (1) is added to the lowest order to be stored;

4) if there is a digit five in the highest order to be discarded and all the discarded digits are zeros, then the lowest order to be stored remains unchanged, if it is even, and one (1) is added to the lowest order to be stored, if it is odd (the even-digit rule).

The even-digit rule should provide compensation for error signs. Obviously, an error that occurs during rounding does not exceed half of the lowest order digit left.

Example. Number 2.74 could be obtained by rounding any numbers from 2.735000… to 2.745000…, the difference between which and the rounded number does not exceed 0.005.

Significant digits are underlined in the following example.

Example. 2.305; 0.0357; 0.001123; 0.035299879 ≈ 0.035300.

Rounding the number 0.035299879 to six decimal places results in the number 0.035300, in which the last two zeros are significant. If we discard these zeros, then the obtained number 0.0353 is not equivalent to the number 0.035300 – the approximate value of the number 0.035299879, since the errors of the specified approximate numbers differ (0.00005 and 0.0000005).

Re-rounding should not be carried out, as it can lead to an increase in the error.

Example. The number 2.7346 is rounded to 2.73, but double rounding leads first to the number 2.735 and then to the number 2.74, which is not only further from the initial value than 2.73 but also its error of 0.0054 exceeds the allowable value 0.005.

It is easy to see that the absolute error is characterized by the number of correct digits after the decimal point, and the relative error – by the number of correct significant digits.

Definition. The first *n* significant digits in the record of an approximate number are called *correct* in the narrow sense, if the absolute error of the number does not exceed half of the one (1) of the order corresponding to the *n*-th significant digit, counting from left to right.

Example. Determine correct digits of the approximate value 2.721 of the number *e*, if it is known that $e = 2.718281828\ldots$. Since $|2.721 - e| < 0.003 < 0.005$, only the first three digits 2.72 are correct, and the last digit can be discarded. Suppose that $x = 1.10253 \pm 0.00009$. Then, only the first four significant digits of *x* are correct, since $0.00009 < 0.0005$, and digits 5 and 3 do not meet the definition.

So, it is easy to show that the order of the last correct significant digit of the number with its known absolute error Δ can be presented by the following formula: $p = \left[\log_{10}(2\Delta)\right] + 1$.

Since number is usually recorded on modern computers with at least 10–12 decimal places, the error of a single rounding $\Delta = 10^{-10} \div 10^{-12}$ is generally negligible as compared to the inherent error and the method error. Although billions of operations are performed when solving large problems and it can be assumed that rounding errors can noticeably accumulate, however, since they are random in nature, their mutual compensation usually occurs. Nevertheless, special algorithms are often built; in particular, these are iterative algorithms that have low sensitivity to rounding errors.

1.2. PROBLEM CORRECTNESS

When solving basic problems using numerical methods, it is necessary to know some input (original) data – initial, boundary values of the desired function, coefficients, and right-hand sides of equations, etc. Obviously, it is important for

the researcher to know whether there is a solution to the postulated problem, whether it is unique, and to what extent it depends on the input data.

Definition. It is said (following Hadamard) that a mathematical problem is formulated *correctly* if the following takes place:

1) it can be solved with any valid input data;

2) it always has only one solution;

3) its solution is continuously dependent on the input data in some reasonable topology, i.e. a small change in the input data corresponds to a small change in the solution (in this case, it is said that the problem is *stable*).

The problem is considered to be incorrectly formulated if it does not have at least one of the abovementioned properties, for example, if its solution is unstable with respect to the input data, i.e. a small change in it can correspond to a large change in the solution.

It is known that a numerical integration problem is a correct problem, and a numerical differentiation problem is an incorrect problem. A classic example of an incorrect problem is a Cauchy problem for Laplace's equation. This incorrectness of the original problem is also manifested when it is solved using numerical methods.

Nowadays, there are also developed methods for solving incorrect problems. The so-called regularization methods are among them. They reduce the solution of the original problem to the solution of an auxiliary one close to it with some small parameter ε, so that if $\varepsilon \to 0$, the solution of the auxiliary problem must tend to the solution of the original problem. For some numerical methods below, the conditions for correctness and stability will be formulated.

1.3. MATRIX PROPERTIES

Let's recall some information from linear algebra that will be needed later. Let's consider the rectangular matrix:

$$A = \begin{bmatrix} a_{11}\ a_{12} \ldots a_{1n} \\ \ldots\ \ldots \ldots \ldots \\ a_{m1}\, a_{m2} \ldots a_{mn} \end{bmatrix}.$$

Two matrices $A \equiv \left[a_{ij} \right]$ and $B \equiv \left[b_{ij} \right]$ of dimension $m \times n$ are equal to each other if $a_{ij} = b_{ij}$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The sum of the two $m \times n$ matrices *A* and *B* is the $m \times n$ matrix:

$A + B \equiv \left[a_{ij} \right] + \left[b_{ij} \right] \equiv \left[a_{ij} + b_{ij} \right]$.

The product of the matrix *A* multiplied by the scalar α is the $m \times n$ matrix:

$\alpha A \equiv \alpha \left[a_{ij} \right] = \left[\alpha a_{ij} \right]$.

The product of the $m \times n$ matrix *A* multiplied by the $n \times r$ matrix *B* is the $m \times r$ matrix *C*:

$C = AB \equiv \left[a_{ij} \right] \left[b_{jk} \right] \equiv \left[c_{ik} \right]$, where $c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}$.

So, the element c_{ik} of the matrix *C* is the sum of the products of the *i*-th row of the matrix *A* multiplied by the corresponding elements of the *k*-th column of the matrix *B*, and the number of columns of the matrix *A* must be equal to the number of rows of the matrix *B*. The existence of the product *AB* does not at all imply the existence of the product *BA*. There are both *AB* and *BA* for square matrices ($m = n$) of the same order, but generally speaking, $AB \neq BA$.

Definition. The *determinant* of the square matrix will be presented as $\det A$:

$$\det A = \begin{vmatrix} a_{11}\ a_{12} \ldots a_{1n} \\ \ldots\ \ldots \ldots \ldots \\ a_{n1}\, a_{n2} \ldots a_{nn} \end{vmatrix}.$$

The determinant of the square matrix with n^2 (real or complex) numbers (elements a_{ij}) is the following sum:

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} M_{ij} ,$$

where M_{ij} is an additional minor to the element a_{ij}. This formula is called the *i*-row expansion. Also, a similar expansion by any column is valid, as a result of which the determinant of the matrix does not change when transposed.

Definition. The *k*-th order *minor* of the matrix *A* is the *k*-th order determinant composed of the elements that are at the intersection of *k* rows and *k* columns of the matrix *A*. The *rank* of the matrix *A* is such a number *r* that all minors of the order $r+1$ and higher are equal to zero. An *additional minor* of *n*-th order to the element a_{ij} of the matrix *A* is the $n-1$ order determinant composed of all elements of the matrix *A* that are not at the intersection of the *i*-th row and *j*-th column.

The determinant of the matrix of the first order $A=[a_{11}]$ is equal to a_{11}.

The determinant of the matrix of the second order $A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ is equal to:

$$\det A=\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}=(-1)^{1+1}a_{11}a_{22}+(-1)^{1+2}a_{12}a_{21}=a_{11}a_{22}-a_{12}a_{21}.$$

Example. Let's find the determinant of the matrix $A=\begin{pmatrix} 1 & 2 & 3 \\ 3 & -2 & 4 \\ -1 & 0 & 2 \end{pmatrix}$.

Let's use expansion by the first row:

$$\det A=\begin{vmatrix} 1 & 2 & 3 \\ 3 & -2 & 4 \\ -1 & 0 & 2 \end{vmatrix}=(-1)^{1+1}\cdot 1\cdot\begin{vmatrix} -2 & 4 \\ 0 & 2 \end{vmatrix}+(-1)^{1+2}\cdot 2\cdot\begin{vmatrix} 3 & 4 \\ -1 & 2 \end{vmatrix}+(-1)^{1+3}\cdot 3\cdot\begin{vmatrix} 3 & -2 \\ -1 & 0 \end{vmatrix}=$$

$$=1\cdot(-4)-2\cdot 10+3\cdot(-2)=-30.$$

Definition. The expression $(-1)^{i+j}M_{ij}$ is called a *cofactor* of the *k*-th order of the matrix *A*.

An important special case of a square matrix is a *diagonal* matrix:

$$A=\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}.$$

If $a_{11}=a_{22}=\dots=a_{nn}=1$, such a matrix is called a *unity* matrix and is denoted by *E*. Another special case of a square matrix is a *tridiagonal* matrix:

$$A=\begin{bmatrix} a_{11} & a_{12} & 0 & \dots & 0 & 0 \\ a_{21} & a_{22} & a_{23} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & 0 & \dots & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

Such matrices are often met when solving differential equations.

A matrix is called *lower triangular* if all its elements located above the main diagonal are equal to zero:

$$A=\begin{bmatrix} a_{11} & 0 & \dots & 0 & 0 \\ a_{21} & a_{22} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

The *upper triangular* matrix is defined similarly.

A square matrix is called *symmetric* if its elements satisfy the relation $a_{ij}=a_{ji}$, $i,j=1,\dots,n$. If we swap rows with columns in the matrix, we get the *transposed* matrix A^T. Obviously, x^T means the row vector. The square matrix *A* is symmetric if $A=A^T$.

Definition. The matrix A^{-1} is called the *inverse* of the matrix *A* if $AA^{-1}=A^{-1}A=E$. A necessary and sufficient condition for the existence of the inverse matrix A^{-1} is the condition $\det A\neq 0$. It is said that, in this case, the matrix *A* is improper or nondegenerate. There are no inverse matrices for non-square matrices and degenerate matrices.

The matrix that is the inverse of the matrix *A* can be presented as follows:

$$A^{-1}=\frac{adj(A)}{\det A},$$

where *adj*(*A*) is the *adjoint* matrix (the matrix composed of cofactors for the corresponding elements of the transposed matrix).

Example. Let's find the inverse matrix for $A=\begin{pmatrix} -1 & 2 \\ 3 & -5 \end{pmatrix}$.

$$A^T=\begin{pmatrix} -1 & 3 \\ 2 & -5 \end{pmatrix}.\ A^{-1}=\frac{\begin{pmatrix} (-1)^{1+1}\cdot(-5) & (-1)^{1+2}\cdot 2 \\ (-1)^{2+1}\cdot 3 & (-1)^{2+2}\cdot(-1) \end{pmatrix}}{5-6}=\frac{\begin{pmatrix} -5 & -2 \\ -3 & -1 \end{pmatrix}}{-1}=\begin{pmatrix} 5 & 2 \\ 3 & 1 \end{pmatrix}.$$

1.4. NORMS OF VECTORS AND MATRICES

In mathematics, sets with elements that are numbers, vectors, matrices or functions are often considered. The sets themselves are usually normalized linear spaces, since the operations of addition of elements and their multiplication by a number are defined in them, and the norm of each element is introduced. The concepts of the norms of vectors and matrices are needed to study the convergence of numerical methods for solving problems in linear algebra.

Definition. The *norm of vector* $x=(x_1, x_2,\ldots, x_n)^T$ (denoted as $\|x\|$) in the *n*-dimensional real space of vectors $x\in R^n$ is a real nonnegative number calculated using the components of the vector and satisfying the following conditions:

1) $\|x\|\ge 0$, and $\|x\|=0$ if and only if $x=\vartheta$;

2) $\|\alpha\cdot x\|=|\alpha|\cdot\|x\|$ for any $\alpha\in R$;

3) $\|x+y\|\le\|x\|+\|y\|$ (triangle inequality).

The most common norms of vectors are the following ones:

$$\|x\|_1=\sum_{i=1}^{n}|x_i|,$$

$$\|x\|_2=\sqrt{\sum_{i=1}^{n}x_i^2}=\sqrt{(x,x)}\ \text{(Euclidean norm)},$$

$$\|x\|_p=\left[\sum_{i=1}^{n}|x_i|^p\right]^{\frac{1}{p}}\ (p\ge 3),$$

$$\|x\|_c=\max_i|x_i|\ \text{(in the limit when } p\to\infty).$$

Obviously, $\|x\|=|x|$ on the set of real numbers.

There is a relation between different norms:

$$\|x\|_1\ge\|x\|_2\ge\|x\|_p\ge\|x\|_c.$$

Example. The vector $x=(1,-2,3)^T$ generates the following norms:

$$\|x\|_1=1+2+3=6,\ \|x\|_2=\sqrt{1+4+9}=\sqrt{14}\approx 3.742,$$

$$\|x\|_3=\sqrt[3]{1+8+27}=\sqrt[3]{36}\approx 3.302,\ \|x\|_4=\sqrt[4]{1+16+81}=\sqrt[4]{98}\approx 3.146,$$

$$\|x\|_{10}=\sqrt[10]{1+1024+59049}=\sqrt[10]{60074}\approx 3.005,\|x\|_c=\max(1,2,3)=3.$$

Definition. The *norm of matrix* $A_{n\times n}$ (denoted as $\|A\|$) with real elements in the space of matrices is a real nonnegative number calculated using the elements of the matrix and having the following properties:

1) $\|A\| \geq 0$, and $\|A\| = 0$ if and only if $A = \vartheta$;

2) $\|\alpha \cdot A\| = |\alpha| \cdot \|A\|$ for any $\alpha \in R$;

3) $\|A + B\| \leq \|A\| + \|B\|$ (triangle inequality for addition);

4) $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ (triangle inequality for multiplication).

As it can be seen from the last property (if vector *x* is used as the matrix *B*), the norm of the matrices must be consistent with the norm of the vectors. This consistency is provided by the following relation:

$\|A \cdot x\| \leq \|A\| \cdot \|x\|$.

In the space of square $n \times n$ matrices, the following norms are most commonly used (consistent with the corresponding norms of vectors):

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|,$$

$$\|A\|_2 = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2},$$

$$\|A\|_c = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

Norm $\|A\|_1$ is the maximum number among the sums of modulus of matrix column elements, norm $\|A\|_c$ is the maximum number among the sums of modulus of matrix row elements, norm $\|A\|_2$ is spherical or Euclidean. Let's note that norm $\|A\|_c$ is consistent with all the norms of vectors stated above.

Let's introduce the concept of the limit of vectors and matrices.

Definition. Let's consider the sequence of vectors $x^{(1)}, x^{(2)}, \ldots$ with components $x_1^{(1)}, \ldots, x_n^{(1)}, \; x_1^{(2)}, \ldots, x_n^{(2)}, \ldots$. If there are limits:

$$x_i = \lim_{k\to\infty} x_i^{(k)}, \; i = 1, 2, \ldots, n,$$

then it is said that the vector x with components $x_1, x_2, \ldots, x_n$ is the *limit of the sequence* $x^{(1)}, x^{(2)}, \ldots$. The limit of the sequence of matrices is defined similarly.

A necessary and sufficient condition for the convergence of vectors $x^{(k)}$ to x is condition $\|x^{(k)} - x\| \to 0$, and $\|x^{(k)}\| \to \|x\|$. A similar statement is true for matrices.

Convergence in one of the norms leads to convergence in the others.

Example. Calculate different norms $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_c$ for matrix $A = \begin{pmatrix} -1 & 2 \\ 3 & -5 \end{pmatrix}$ and vector $b = \begin{pmatrix} 3 \\ -4 \end{pmatrix}$. Check the fulfilment of the condition of consistency of norms $\|Ax\| \le \|A\|\|x\|$ for various combinations of norms.

Let's calculate the corresponding norms:

$\|b\|_1 = |3| + |-4| = 7$,

$\|b\|_2 = (3^2 + (-4)^2)^{1/2} = 5$,

$\|b\|_c = \max(|3|, |-4|) = 4$;

$\|A\|_1 = \max(|-1| + |3|,\ |2| + |-5|) = 7$,

$\|A\|_2 = (|(-1)|^2 + 3^2 + 2^2 + (-5)^2)^{1/2} = \sqrt{39} \approx 6.245$,

$\|A\|_c = \max(|-1| + |2|,\ |3| + |-5|) = 8$.

To check the consistency condition, let's calculate different norms of the vector $c = Ab = \begin{pmatrix} -11 \\ 29 \end{pmatrix}$:

$\|c\|_1 = |-11| + |29| = 40$,

$\|c\|_2 = ((-11)^2 + 29^2)^{1/2} = \sqrt{962} \approx 31.016$,

$\|c\|_c = \max(|-11|, |29|) = 29$.

It is easy to verify that the consistency condition is fulfilled for the corresponding norms:

$\|c\|_1 = 40 \le \|A\|_1 \|b\|_1 = 7 \cdot 7 = 49$,

$\|c\|_2 = \sqrt{962} \le \|A\|_2 \|b\|_2 = \sqrt{39} \cdot 5 = \sqrt{975}$,

$\|c\|_c = 29 \le \|A\|_c \|b\|_c = 8 \cdot 4 = 32$.

Moreover, it is known that the matrix norm $\|A\|_c$ is consistent with all norms of vectors introduced above. In this example, it is confirmed by the fulfilment of the following inequalities:

$$\|c\|_1 = 40 \le \|A\|_c \|b\|_1 = 8 \cdot 7 = 56,$$

$$\|c\|_2 = \sqrt{962} \le \|A\|_c \|b\|_2 = 8 \cdot 5 = 40.$$

At the same time the use of a number of other combinations of norms of matrix and vector leads, in this case, to a violation of the consistency condition:

$$\|c\|_c = 29 > \|A\|_1 \|b\|_c = 7 \cdot 4 = 28,$$

$$\|c\|_c = 29 > \|A\|_2 \|b\|_c = \sqrt{39} \cdot 4.$$

This example clearly illustrates the importance of using consistent matrix and vector norms.