Numerical analysis

Aleksei Savelev

Harbin Institute of Technology

December 15, 2024

Contents

1 Calculation of matrix eigenvalues 4
1.1 Characteristic equation of a matrix 4
1.2 The Power Method 7
1.3 Eigenvalues of a Symmetric Tridiagonal Matrix 13
1.4 QR Method 14
2 The initial value problem for the first-order ordinary differential equation . 15
2.1 Introduction 15
2.2 Solution by Taylor's series 16
2.3 Picard's method of successive approximations 18
2.4 Euler's method 20
2.4.1 Error Estimates for the Euler Method 21
2.4.2 Modified Euler's Method 23
2.5 Runge-Kutta Methods 24
3 Systems of first-order ordinary differential equations. Higher-order differential equations 29
3.1 Cubic spline method 29
3.2 Boundary-value problems 31
3.3 Finite-difference Method 31
3.4 Galerkin's Method 37
4 Numerical solution of Partial Differential Equations. 41
4.1 Introduction 41
4.2 Finite-Difference approximations to derivatives 42
4.3 Heat Equation in One Dimension 45
4.3.1 Finite-difference approximation 45
4.4 Wave equation 50
5 Partial differential equations of elliptic type. Poisson's equation. 54
5.1 Solution of Laplace's equation 54
5.1.1 Jacobi's Method 55
5.1.2 Gauss-Seidel Method 55
5.1.3 Successive Over Relaxation (SOR) Method 55
5.1.4 ADI Method 62

6 The Finite Element Method . . . 67
6.1 Introduction . . . 67
6.1.1 Functionals . . . 68
6.1.2 Base Functions . . . 72
6.2 Methods of approximation . . . 73
6.2.1 Rayleigh-Ritz Method . . . 73
6.2.2 Galerkin's Method . . . 79
6.3 Application to two-dimensional problem . . . 79
6.4 Finite Element Method . . . 80
6.4.1 Finite Element Method for One-dimensional Problems . . . 82
7 Error analysis . . . 90
7.1 ERRORS AND THEIR COMPUTATIONS . . . 90
7.2 Absolute, relative and percentage errors . . . 91
7.3 A GENERAL ERROR FORMULA . . . 95
7.4 ERROR IN A SERIES APPROXIMATION . . . 97
8 Monte Carlo Method . . . 104
8.1 General idea of the method . . . 104
8.1.1 The origin of the Monte Carlo method. . . . 104
8.1.2 Example. . . . 104
8.1.3 Two features of the Monte Carlo method. . . . 105
8.1.4 Problems solved by the Monte Carlo method. . . . 106
8.1.5 More about the example. . . . 106
8.2 Random variables . . . 107
8.2.1 Discrete random variables. . . . 108
8.2.2 Continuous random variables. . . . 110
8.2.3 Normal random variables. . . . 113
8.2.4 The central limit theorem of probability theory. . . . 114
8.2.5 The general scheme of the Monte Carlo method. . . . 115
8.3 Getting random variables on a computer . . . 116
8.3.1 Tables of random numbers. . . . 116
8.3.2 Random number generators. . . . 118
8.3.3 Pseudo-random numbers. . . . 118
8.4 Transformations of random variables . . . 120
8.4.1 Playing a discrete random variable. . . . 120
8.4.2 Playing a continuous random variable. . . . 121
8.4.3 The Neumann method for playing a continuous random variable. . . . 124
8.4.4 About playing normal quantities. . . . 124
8.4.5 Again about the example from paragraph 8.1.2. . . . 125
8.5 Calculation of the queuing system . . . 126
8.5.1 Description of the task. . . . 126

8.5.2 The simplest flow of applications. 126
8.5.3 The calculation scheme. 127
8.5.4 More complex tasks. 129
8.6 Calculation of product quality and reliability 129
8.6.1 The simplest quality calculation scheme. 129
8.6.2 Examples of reliability calculation. 131
8.6.3 Further possibilities of the method. 132
8.7 Calculation of the passage of neutrons through the plate 134
8.7.1 Setting the task. 134
8.7.2 A calculation scheme by modeling true trajectories. 136
8.7.3 Calculation scheme using weights replacing absorption. . . 139
8.8 Calculation of a certain integral 140
8.8.1 The calculation method. 140
8.8.2 How to choose a calculation scheme. 141
8.8.3 A numerical example. 142
8.8.4 About error evaluation. 144

1 Calculation of matrix eigenvalues

1.1 Characteristic equation of a matrix

Let A be a square matrix of order n with elements a_{ij}. We wish to find a column vector X and a constant λ such that

$$AX = \lambda X \tag{1.1}$$

In Eq. (1.1), λ is called the *eigenvalue* and X is called the corresponding *eigenvector*.

The matrix Eq. (1.1), when written out in full, represents a set of homogeneous linear equations:

$$\left.\begin{aligned} (a_{11}-\lambda)x_1 + a_{12}x_2 + \cdots + a_{1n}\,x_n &= 0 \\ a_{21}x_1 + (a_{22}-\lambda)x_2 + \cdots + a_{2n}\,x_n &= 0 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + (a_{nn}-\lambda)x_n &= 0. \end{aligned}\right\} \tag{1.2}$$

A nontrivial solution exists only when the coefficient determinant in (1.2) vanishes. Hence, we have

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn}-\lambda \end{vmatrix} = 0. \tag{1.3}$$

This equation, called the *characteristic equation* of the matrix A, is a polynomial equation of degree n in λ, the polynomial being called the *characteristic-polynomial* of A. If the roots of Eq. (1.3) be given by λ_i $(i = 1, 2, \dots, n)$, then for each value of λ_i, there exists a corresponding X_i such that

$$AX_i = \lambda_i X_i. \tag{1.4}$$

The eigenvalues λ_i may be either *distinct* (i.e. all different) or *repeated*. The evaluation of eigenvectors in the case of the repeated roots is a much involved process and will not be attempted here. The set of all eigenvalues, λ_i, of a matrix A is called the *spectrum* of A and the largest of $|\lambda_i|$ is called the *spectral radius* of A.

The eigenvalues are obtained by solving Eq. (1.3). This method, which is demonstrated in Example 1.1, is unsuitable for matrices of higher order and better methods must be applied. For symmetric matrices, in particular, several methods are available and a recent method, known as Householder's method, is described in a subsequent section.

In some practical applications, only the numerically largest eigenvalue and the corresponding eigenvector are required, and in Example 1.2, we will describe an iterative method to compute the largest eigenvalue. This method is easy of application and also well-suited for machine computations.

Example 1.1. Find the eigenvalues and eigenvectors of the matrix:

$$A = \begin{bmatrix} 5 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 5 \end{bmatrix}.$$

The characteristic equation of this matrix is given by

$$\begin{bmatrix} 5-\lambda & 0 & 1 \\ 0 & -2-\lambda & 0 \\ 1 & 0 & 5-\lambda \end{bmatrix} = 0,$$

which gives $\lambda_1 = -2, \lambda_2 = 4$ and $\lambda_3 = 6$. The corresponding eigenvectors are obtained thus

(i) $\lambda_1 = -2$. Let the eigenvector be

$$X_1 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Then we have:

$$A \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = -2 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

which gives the equations

$$7x_1 + x_3 = 0 \quad \text{and} \quad x_1 + 7x_3 = 0.$$

The solution is $x_1 = x_3 = 0$ with x_2 arbitrary. In particular, we take $x_2 = 1$ and the eigenvector is

$$X_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

(ii) $\lambda_2 = 4$. With

$$X_2 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

as the eigenvector, the equations are

$$x_1 + x_3 = 0 \quad \text{and} \quad -6x_2 = 0,$$

from which we obtain

$$x_1 = -x_3 \quad \text{and} \quad x_2 = 0.$$

We choose, in particular, $x_1 = \dfrac{1}{\sqrt{(2)}}$ and $x_3 = -\dfrac{1}{\sqrt{(2)}}$ so that $x_1^2 + x_2^2 + x_3^2 = 1$. The eigenvector in this way is said be *normalized*. We, therefore, have

$$X_2 = \begin{bmatrix} 1/\sqrt{(2)} \\ 0 \\ -1/\sqrt{(2)} \end{bmatrix}.$$

(iii) $\lambda_3 = 6$. If

$$X_3 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

is the required eigenvector, then the equations are

$$\begin{aligned} -x_1 + x_3 &= 0 \\ -8x_2 &= 0 \\ x_1 - x_3 &= 0, \end{aligned}$$

which give $x_1 = x_3$ and $x_2 = 0$.

Choosing $x_1 = x_3 = \frac{1}{\sqrt{(2)}}$, the normalized eigenvector is given by

$$X_3 = \begin{bmatrix} 1/\sqrt{(2)} \\ 0 \\ 1/\sqrt{(2)} \end{bmatrix}.$$

Example 1.2. Determine the largest eigenvalue and the corresponding eigenvalue of the matrix

$$A = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Let the initial eigenvector be

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = X^{(0)}, \text{ say.}$$

Then we have

$$AX^{(0)} = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = X^{(1)}, \text{ say.}$$

Hence an approximate eigenvalue is 1 and an approximate eigenvector is $X^{(1)}$. Hence we have

$$AX^{(1)} = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 7 \\ 3 \\ 0 \end{bmatrix} = 3 \begin{bmatrix} 2.3 \\ 1 \\ 0 \end{bmatrix}$$

from which we see that

$$X^{(2)} = \begin{bmatrix} 2.3 \\ 1 \\ 0 \end{bmatrix}$$

and that an approximate eigenvalue is 3.
Repeating the above procedure, we successively obtain

$$4 \begin{bmatrix} 2.1 \\ 1.1 \\ 0 \end{bmatrix}; \quad 4 \begin{bmatrix} 2.2 \\ 1.1 \\ 0 \end{bmatrix}; \quad 4.4 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}; \quad 4 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}; \quad 4 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}.$$

It follows that the largest eigenvalue is 4 and the corresponding eigenvector is

$$\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}.$$

1.2 The Power Method

Power method is normally used to determine the largest eigenvalue (in magnitude) and the corresponding eigenvector of the Eq. (1.1).

Let $\lambda_1, \lambda_2, ..., \lambda_n$ be the eigenvalues of A such that

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \tag{1.5}$$

and further assume that the corresponding eigenvectors $v_1, v_2, \ldots, v_n$ forms a basis for $\mathbb{R}^n$. Therefore, any vector $v \in \mathbb{R}^n$ can be written as

$$v = c_1 v_1 + c_2 v_2 + \cdots + c_n v_n.$$

Premultiplying by A and substituting $Av_i = \lambda_i v_i,\ i = 1, \ldots, n$, we get

$$\begin{aligned} Av =& c_1 \lambda_1 v_1 + \cdots + c_n \lambda_n v_n \\ =& \lambda_1 \left(c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right) v_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1} \right) v_n \right). \end{aligned}$$

Premultiplying by A again and simpying, we get

$$\begin{aligned} A^2v =& \lambda_1^2 \left(c_1v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^2 v_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^2 v_n \right) \\ & \cdots \\ & \cdots \\ & \cdots \\ A^kv =& \lambda_1^k \left(c_1v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n \right). \end{aligned}$$

Using the assumption (1.5), we can see that

$$|\frac{\lambda_k}{\lambda_1}| < 1, \ k = 2, \ldots, n.$$

Therefore, we have

$$\lim_{k\to\infty} \frac{A^kv}{\lambda_1^k} = c_1v_1. \tag{1.6}$$

For $c_1 \neq 0$, the RHS is a scalar multiple of the eigenvector. Also, from the above expression for A^kv, we get

$$\lim_{k\to\infty} \frac{(A^{k+1}v)_i}{(A^kv)_i} = \lambda_1, \tag{1.7}$$

where i denotes a component of the corresponding vectors.

The power method is based by results (1.6) and (1.7).

Algorithm 1 The Power Method

Choose an arbitary initial guess $x^{(0)}$

1: **for** $k = 1, 2, \ldots$ **do**
2: Compute $y^{(k)} = Ax^{(k-1)}$
3: Take $\mu_k = y_i^{(k)}$, where $||y^{(k)}||_\infty = |y_i^{(k)}|$,
4: Set $x^{(k)} = \frac{y^{(k)}}{\mu_k}$.
5: If $||x^{(k-1)} - x^{(k)}||_\infty > \epsilon$, go to step 1.

For some pre-assigned positive quantity ϵ.

Let us now study the convergence of this method.

Theorem 1.1 (Power Method)**.**

Let A be an non-singular $n \times n$ matrix with following conditions:

I *A has n linearly independent eigenvectors, v_i, $i = 1, \ldots, n$.*

II *The eigenvalues* λ_i *satisfy*

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|.$$

III *The vector* $x^{(0)} \in \mathbb{R}^n$ *is such that*

$$x^{(0)} = \sum_{j=1}^{n} c_j v_j, \quad c_1 \neq 0.$$

Then the power method converges in the sense that there exists constants C_1 *and* C_2 *such that*

$$||x^{(k)} - Kv_1|| \leq C_1 \left|\frac{\lambda_2}{\lambda_1}\right|^k, \quad \textit{for some } K \neq 0$$

and

$$||\lambda_1 - \mu_k|| \leq C_1 |\frac{\lambda_2}{\lambda_1}|^k.$$

Proof. From the definition of $x^{(k)}$, we have

$$x^{(k)} = \frac{Ax^{(k-1)}}{\mu_k} = \frac{Ay^{(k-1)}}{\mu_k\mu_{k-1}} = \frac{AAx^{(k-2)}}{\mu_k\mu_{k-1}} = \frac{A^2x^{(k-2)}}{\mu_k\mu_{k-1}} = \cdots = \frac{A^kx^{(0)}}{\mu_k\mu_{k-1}\ldots\mu_1}.$$

Therefore, we have

$$x^{(k)} = m_k A^k x^{(0)},$$

where $m_k = 1/(\mu_1\mu_2\ldots\mu_k)$. But, $x^{(0)} = \sum_{j=1}^{n} c_j v_j, \quad c_1 \neq 0$. Therefore

$$x^{(k)} = m^k\lambda_1^k \left(c_1 v_1 + \sum_{j=2}^{n} c_j \left(\frac{\lambda_j}{\lambda_1}\right)^k v_j \right).$$

Taking maximum norm on both sides and noting that $||x_\infty^{(k)} = 1||$, we get

$$1 = |m^k\lambda_1^k| \left\| c_1 v_1 + \sum_{j=2}^{n} c_j \left(\frac{\lambda_j}{\lambda_1}\right)^k v_j \right\|_\infty.$$

This implies on taking limit,

$$|\lim_{k\to\infty} m_k\lambda_1^k| = \frac{1}{|c_1|\|v_1\|_\infty} < \infty.$$

This is equivalent to

$$\lim_{k\to\infty} m_k\lambda_1^k = \pm\frac{1}{c_1\|v_1\|_\infty} < \infty.$$

Finally,

$$\lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} m_k\lambda_1^k \cdot c_1 v_1 = Kv_1.$$

Moreover,

$$\|x^{(k)} - K_k v_1\|_\infty = \left\| m_k\lambda_1^k \sum_{j=2}^{n} c_j \left(\frac{\lambda_j}{\lambda_1}\right)^k v_j \right\|_\infty \leq C\left|\frac{\lambda_2}{\lambda_1}\right|^k.$$

For eigenvalue,

$$\mu_k x^{(k)} = y^{(k)}.$$

Therefore,

$$\mu_k = \frac{y_i^{(k)}}{x_i^{(k)}} = \frac{(Ax^{(k-1)})_i}{(x^{(k)})_i}.$$

Taking limit, we have

$$\lim_{k\to\infty} \mu_k = \frac{A(Kv_1)_i}{K(v_1)_i} = \frac{\lambda(v_1)_i}{(v_1)_i} = \lambda_1,$$

which gives the desired result. □

Example 1.3. Consider the matrix

$$A = \begin{bmatrix} 3 & 0 & 0 \\ -4 & 6 & 2 \\ 16 & -15 & -5 \end{bmatrix}.$$

The eigenvalue of this matrix are $\lambda_1 = 3, \lambda_2 = 1$ and $\lambda_3 = 0$. The corresponding eigenvectors are $X_1 = (1, 0, 2)^T, X_2 = (0, 2, -5)^T$ and $X_3 = (0, 1, -3)^T$.
Initial Guess 1: Let us take $x_0 = (1, 0.5, 0.25)^T$. The power method gives the following:

Iteration No: 1

$$y_1 = Ax_0 = (3.000000, -0.500000, 7.250000)^T$$
$$\mu_1 = 7.250000$$
$$x_1 = \frac{y_1}{\mu_1} = (0.413793, -0.068966, 1.000000)^T$$

Iteration No: 2

$$y_2 = Ax_1 = (1.241379, -0.068966, 2.655172)^T$$
$$\mu_2 = 2.655172$$
$$x_2 = \frac{y_2}{\mu_2} = (0.467532, -0.025974, 1.000000)^T$$

Iteration No: 3

$$y_3 = Ax_2 = (1.402597, -0.025974, 2.870130)^T$$
$$\mu_3 = 2.870130$$
$$x_3 = \frac{y_3}{\mu_3} = (0.488688, -0.009050, 1.000000)^T$$

Iteration No: 4

$$y_4 = Ax_3 = (1.466063, -0.009050, 2.954751)^T$$
$$\mu_4 = 2.954751$$
$$x_4 = \frac{y_4}{\mu_4} = (0.496172, -0.003063, 1.000000)^T$$

Iteration No: 5

$$y_5 = Ax_4 = (1.488515, -0.003063, 2.984686)^T$$
$$\mu_5 = 2.984686$$
$$x_5 = \frac{y_5}{\mu_5} = (0.498717, -0.001026, 1.000000)^T$$

Iteration No: 6

$$y_6 = Ax_5 = (1.496152, -0.001026, 2.994869)^T$$
$$\mu_6 = 2.994869$$
$$x_6 = \frac{y_6}{\mu_6} = (0.499572, -0.000343, 1.000000)^T$$

Iteration No: 7

$$y_7 = Ax_6 = (1.498715, -0.000343, 2.998287)^T$$
$$\mu_7 = 2.998287$$
$$x_7 = \frac{y_7}{\mu_7} = (0.499857, -0.000114, 1.000000)^T$$

Iteration No: 8

$$y_8 = Ax_7 = (1.499571, -0.000114, 2.999429)^T$$
$$\mu_8 = 2.999429$$
$$x_8 = \frac{y_8}{\mu_8} = (0.499952, -0.000038, 1.000000)^T$$

Iteration No: 9

$$y_9 = Ax_8 = (1.499857, -0.000038, 2.999809)^T$$
$$\mu_9 = 2.999809$$
$$x_9 = \frac{y_9}{\mu_9} = (0.499984, -0.000013, 1.000000)^T$$

Iteration No: 10

$$y_{10} = Ax_9 = (1.499952, -0.000013, 2.999936)^T$$
$$\mu_{10} = 2.999936$$
$$x_{10} = \frac{y_{10}}{\mu_{10}} = (0.499995, -0.000004, 1.000000)^T$$

Initial Guess 2: Let us take $x_0 = (0, 0.5, 0.25)^T$. The power method gives the following:
Iteration No: 1

$$y_1 = Ax_0 = (0.000000, 3.500000, -8.750000)^T$$
$$\mu_1 = 8.750000$$
$$x_1 = \frac{y_1}{\mu_1} = (0.000000, 0.400000, -1.000000)^T$$

Iteration No: 2

$$y_2 = Ax_1 = (0.000000, 0.400000, -1.000000)^T$$
$$\mu_2 = 1.000000$$
$$x_2 = \frac{y_2}{\mu_2} = (0.000000, 0.400000, -1.000000)^T$$

Iteration No: 3

$$y_3 = Ax_2 = (0.000000, 0.400000, -1.000000)^T$$
$$\mu_3 = 1.000000$$
$$x_3 = \frac{y_3}{\mu_3} = (0.000000, 0.400000, -1.000000)^T$$

Iteration No: 4

$$y_4 = Ax_3 = (0.000000, 0.400000, -1.000000)^T$$
$$\mu_4 = 1.000000$$
$$x_4 = \frac{y_4}{\mu_4} = (0.000000, 0.400000, -1.000000)^T$$

Note that in the second initial guess, the first coordinate is zero and therefore, c_1 in the power method is zero. This makes the iteration to converge to λ_2, which is the next dominant eigenvalue. □

1.3 Eigenvalues of a Symmetric Tridiagonal Matrix

Since symmetric matrices can be reduced to symmetric tridiagonal matrices, the determination of eigenvalues of a symmetric tridiagonal matrix is of particular interest. Let

$$A_1 = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{12} & a_{22} & a_{23} \\ 0 & a_{23} & a_{33} \end{bmatrix}. \tag{1.8}$$

To obtain the eigenvalues of A_1, we form determinant equation

$$|A_1| = \begin{vmatrix} a_{11}-\lambda & a_{12} & 0 \\ a_{12} & a_{22}-\lambda & a_{23} \\ 0 & a_{23} & a_{33}-\lambda \end{vmatrix} = 0.$$

Suppose that the above equation is written in the form

$$\phi_3(\lambda) = 0. \tag{1.9}$$

Expanding the determinant in terms of the third row, we obtain

$$\begin{aligned} \phi_3(\lambda) =& (a_{33}-\lambda)\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{12} & a_{22}-\lambda \end{vmatrix} - a_{23}\begin{vmatrix} a_{11}-\lambda & 0 \\ a_{12} & a_{23} \end{vmatrix} \\ =& (a_{33}-\lambda)\phi_2(\lambda) - a_{23}(a_{11}-\lambda)a_{23} \\ =& (a_{33}-\lambda)\phi_2(\lambda) - a_{23}^2\phi_1(\lambda) \\ =& 0. \end{aligned} \tag{1.10}$$

We, thus, obtain the recursion formula

$$\phi_0(\lambda) = 1$$

$$\phi_1(\lambda) = a_{11}-\lambda \tag{1.11}$$

$$= (a_{11}-\lambda)\phi_0(\lambda) \tag{1.12}$$

$$\begin{aligned} \phi_2(\lambda) &= \begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{12} & a_{22}-\lambda \end{vmatrix} \\ &= (a_{11}-\lambda)(a_{22}-\lambda) - a_{12}^2 \\ &= \phi_1(\lambda)(a_{22}-\lambda) - a_{12}^2\phi_0(\lambda) \end{aligned} \tag{1.13}$$

$$\phi_3(\lambda) = \phi_2(\lambda)(a_{33}-\lambda) - a_{23}^2\phi_1(\lambda). \tag{1.14}$$

In general, if

$$\phi_k(\lambda) = \begin{vmatrix} a_{11}-\lambda & a_{12} & 0 & \dots & 0 \\ a_{12} & a_{22}-\lambda & a_{23} & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & a_{k-1,k} & a_{kk}-\lambda \end{vmatrix}, \quad (2 \le k \le n), \tag{1.15}$$

then the recursion formula is

$$\phi_k(\lambda) = (a_{kk} - \lambda)\phi_{k-1}(\lambda) - a^2_{k-1,k}\phi_{k-2}(\lambda), \qquad (2 \le k \le n). \tag{1.16}$$

The equation $\phi_k(\lambda) = 0$ is the characteristic equation and can be solved by one of the methods described in Section 1.1. We might therefore consider the problem as solved, but we would like to remark that the sequence $\{\phi_k(\lambda), 0 \le k \le n\}$ has special properties which make it a *Sturm sequence* and from these properties one can isolate the eigenvalues of A_1. Once the eigenvalues have been isolated, one of the methods of Section 1.1 can be used to calculate the roots rapidly. The theory of Sturm sequence will not be discussed here, but the interested reader may refer to Henrici [2] for details. When the eigenvalues of the tridiagonal matrix are known its eigenvectors can be calculated by the general method of solving a homogeneous system.

1.4 QR Method

This is the most efficient and widely used general method for the computation of all the eigenvalues of general nonsymmetric matrix. Originally, due to J.G.F. Francis, the method is quite complicated and, therefore, only a brief presentation is given below.

Let $A_1 = A$ be the given matrix. Suppose that A_1 is factorized into form

$$A_1 = Q_1 R_1 \tag{1.17}$$

where Q_1 is an orthogonal matrix and R_1 is an upper triangular matrix.

Therefore,

$$Q_1^{-1} A_1 = R_1 \tag{1.18}$$

The essential feature of this method is to find orthogonal matrices $P_1 P_2 \dots P_{n-1}$ such that

$$P_{n-1} P_{n-2} \dots P_2 P_1 A_1 = R_1 \tag{1.19}$$

The matrices P are of the form $I - 2VV^T$ such that $P_1 A_1$ will contain zeros below the diagonals in its first column, $P_2 \quad P_1 A_1$ will contain zeros in its second column below the diagonal, and so on. If we carry out this procedure with each column of A_1, then the final result will be R_1, which is an upper triangular matrix. The sequence $\{A_i\}$ converges either to triangular matrix or to near triangular matrix. In either case, the eigenvalues can be computed easily.

2 The initial value problem for the first-order ordinary differential equation

2.1 Introduction

Many problems in science and engineering can be reduced to the problem of solving differential equations satisfying certain given conditions. The analytical methods of solution, with which the reader is assumed to be familiar, can be applied to solve only a selected class of differential equations. Those equations which govern physical systems do not possess, in general closed-form solutions, and hence recourse must be made to numerical methods for solving such differential equations.

To describe various numerical methods for the solution of ordinary differential equations, we consider the general first order differential equation

$$\frac{dy}{dx} = f(x, y) \tag{2.1a}$$

with the initial condition,

$$y(x_0) = y_0 \tag{2.1b}$$

and illustrate the theory with respect to this equation. The methods so developed, in general, be applied to the solution of first-order equations, and will yield the solution in one of the two forms:

(i) A series for y in terms of powers of x, from which the value of y can be obtained by direct substitution.

(ii) A set of tabulated values of x and y.

The methods of Taylor and Picard belong to class (i), whereas those of Euler, Runge-Kutta, Adams-Bashforth, etc., belongs to class (ii). These latter methods are called *step-by-step* methods or *marching* methods because the values of y are computed by short ahead for equal intervals h of the independent variable. In the methods of Euler and Runge-Kutta, the interval length h should be kept small and hence these methods can be applied for tabulating y over a limited range only. If, however, the function values are desired over wider range, the methods due to Adams-Bashforth, Adams-Moulton, Milne, etc., may be used. These methods use finite-differences and require 'starting values' which are usually obtained by Taylor's series or Runge-Kutta methods.

It is well-known that a differential equation of the nth order will have n arbitrary constants in its general solution. In order to compute the numerical solution of such equation, we therefore need n conditions. Problems in which all the initial conditions are specified at the *initial* point only are called *initial value problems*. For example, the problem defined by Eqs. (2.1) is an *initial value problem*. On the other hand, in problems

involving second-and higher-order differential equations, we may prescribe the conditions at two or more points. Such problems are called *boundary value problems*.

We shall first describe methods for solving initial value problems of the type (2.1), and at the end of the chapter we will outline methods for solving boundary value problems for second-order differential equations.

2.2 Solution by Taylor's series

We consider the differential equation

$$y' = f(x, y) \tag{2.1a}$$

with the initial condition

$$y(x_0) = y_0 \tag{2.1b}$$

If $y(x)$ is the exact solution of Eq. (2.1), then the Taylor's series for $y(x)$ around $x = x_0$ is given by

$$y(x) = y_0 + (x - x_0)\, y_0' + \frac{(x - x_0)^2}{2!} y_0'' + \cdots \tag{2.2}$$

If the values of $y_0', y_0'', \ldots$ are known, then Eq. (2.2) gives a power series for y. Using the formula for total derivatives, we can write

$$y'' = f' = f_x + y' f_y = f_x + f f_y$$

where the suffixes denote partial derivatives with respect to the variable concerned. Similarly, we obtain

$$\begin{aligned} y''' = f'' &= f_{xx} + f_{xy} f + f\,(f_{yx} + f_{yy} f) + f_y\,(f_x + f_y f) \\ &= f_{xx} + 2 f f_{xy} + f^2 f_{yy} + f_x f_y + f f_y^2 \end{aligned}$$

and other higher derivatives of y. The method can easily be extended to simultaneous and higher-order differential equations.

Example 2.1 From the Taylor series for $y(x)$, find $y(0.1)$ correct to four decimal places if $y(x)$ satisfies

$$y' = x - y^2 \quad \text{and} \quad y(0) = 1$$

The Taylor series for $y(x)$ is given by

$$y(x) = 1 + x y_0' + \frac{x^2}{2} y_0'' + \frac{x^3}{6} y_0''' + \frac{x^4}{24} y_0^{\mathrm{iv}} + \frac{x^5}{120} y_0^{\mathrm{v}} + \cdots$$

The derivatives $y_0', y_0'', \ldots$ etc. are obtained thus:

$$\begin{array}{ll} y'(x) = x - y^2 & y'_0 = -1 \\ y''(x) = 1 - 2yy' & y''_0 = 3 \\ y'''(x) = -2yy'' - 2y'^2 & y'''_0 = -8 \\ y^{\text{iv}}(x) = -2yy''' - 6y'y'' & y^{\text{iv}}_0 = 34 \\ y^{\text{v}}(x) = -2yy^{\text{iv}} - 8y'y''' - 6y''^2 & y^{\text{v}}_0 = -186 \end{array}$$

Using these values, the Taylor series becomes

$$y(x) = 1 - x + \frac{3}{2}x^2 - \frac{4}{3}x^3 + \frac{17}{12}x^4 - \frac{31}{20}x^5 + \cdots$$

To obtain the value of $y(0.1)$ correct to four decimal places, it is found that the terms up to x^4 should be considered, and we have $y(0.1) = 0.9138$.

Suppose that we wish to find the range of values of x for which the above series, truncated after the term containing x^4, can be used to compute the values of y correct to four decimal places. We need only to write

$$\frac{31}{20}x^5 \leq 0.00005 \quad \text{or} \quad x \leq 0.126$$

Example 2.2 Given the differential equation

$$y'' - xy' - y = 0$$

with the conditions $y(0) = 1$ and $y'(0) = 0$, use Taylor's series method to determine the value of $y(0.1)$.

We have $y(x) = 1$ and $y'(x) = 0$ when $x = 0$. The given differential equation is

$$y''(x) = xy'(x) + y(x) \tag{i}$$

Hence $y''(0) = y(0) = 1$. Successive differentiation of (i) gives

$$y'''(x) = xy''(x) + y'(x) + y'(x) = xy''(x) + 2y'(x) \tag{ii}$$

$$y^{\text{iv}}(x) = xy'''(x) + y''(x) + 2y''(x) = xy'''(x) + 3y''(x) \tag{iii}$$

$$y^{\text{v}}(x) = xy^{\text{iv}}(x) + y'''(x) + 3y'''(x) = xy^{\text{iv}}(x) + 4y'''(x) \tag{iv}$$

$$y^{\text{vi}}(x) = xy^{\text{v}}(x) + y^{\text{iv}}(x) + 4y^{\text{iv}}(x) = xy^{\text{v}}(x) + 5y^{\text{iv}}(x) \tag{v}$$

and similarly for higher derivatives. Putting $x = 0$ in (ii) to (v), we obtain

$$y'''(0) = 2y'(0) = 0, \quad y^{\text{iv}}(0) = 3y''(0) = 3, \quad y^{\text{v}}(0) = 0, \quad y^{\text{vi}}(0) = 5.$$

By Taylor's series, we have

$$\begin{aligned} y(x) =& y(0) + xy'(0) + \frac{x^2}{2}y''(0) + \frac{x^3}{6}y'''(0) + \frac{x^4}{24}y^{\text{iv}}(0) \\ &+ \frac{x^5}{120}y^{\text{v}}(0) + \frac{x^6}{720}y^{\text{vi}}(0) + \cdots \end{aligned}$$

Hence

$$\begin{aligned} y(0.1) &= 1 + \frac{(0.1)^2}{2} + \frac{(0.1)^4}{24}(3) + \frac{(0.1)^6}{720}(5) + \cdots \\ &= 1 + 0.005 + 0.0000125, \text{ neglecting the last term} \\ &= 1.0050125, \text{ correct to seven decimal places.} \end{aligned}$$

2.3 Picard's method of successive approximations

Integrating the differential equation given in Eq. (2.1), we obtain

$$y = y_0 + \int_{x_0}^{x} f(x, y)dx \tag{2.3}$$

Equation (2.3), in which the unknown function y appears under the integral sign, is called an integral equation. Such an equation can be solved by the method of successive approximations in which the first approximation to y is obtained by putting y_0 for y on right side of Eq. (2.3), and we write

$$y^{(1)} = y_0 + \int_{x_0}^{x} f\,(x, y_0)\,dx$$

The integral on the right can now be solved and the resulting $y^{(1)}$ is substituted for y in the integrand of Eq. (2.3) to obtain the second approximation $y^{(2)}$:

$$y^{(2)} = y_0 + \int_{x_0}^{x} f\left(x, y^{(1)}\right) dx$$

Proceeding in this way, we obtain $y^{(3)}, y^{(4)}, \ldots, y^{(n-1)}$ and $y^{(n)}$, where

$$y^{(n)} = y_0 + \int_{x_0}^{x} f\left(x, y^{(n-1)}\right) dx \quad \text{with } y^{(0)} = y_0 \tag{2.4}$$

Hence this method yields a sequence of approximations $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$ and it can be proved (see, for example, the book by Levy and Baggot) that if the function $f(x, y)$ is bounded in some region about the point (x_0, y_0) and if $f(x, y)$ satisfies the Lipschitz condition, viz.,

$$|f(x, y) - f(x, \bar{y})| \leq K|y - \bar{y}| \quad K \text{ being a constant} \tag{2.5}$$

then the sequence $y^{(1)}, y^{(2)}, \ldots$ converges to the solution of Eq. (2.1).

Example 2.3 Solve the equation $y' = x + y^2$, subject to the condition $y = 1$ when $x = 0$.

We start with $y^{(0)} = 1$ and obtain

$$y^{(1)} = 1 + \int_0^x (x+1)dx = 1 + x + \frac{1}{2}x^2$$

Then the second approximation is

$$\begin{aligned} y^{(2)} &= 1 + \int_0^x \left[x + \left(1 + x + \frac{1}{2}x^2 \right)^2 \right] dx \\ &= 1 + x + \frac{3}{2}x^2 + \frac{2}{3}x^3 + \frac{1}{4}x^4 + \frac{1}{20}x^5 \end{aligned}$$

It is obvious that the integrations might become more and more difficult as we proceed to higher approximations.

Example 2.4 Given the differential equation

$$\frac{dy}{dx} = \frac{x^2}{y^2 + 1}$$

with the initial condition $y = 0$ when $x = 0$, use Picard's method to obtain y for $x = 0.25, 0.5$ and 1.0 correct to three decimal places.

We have

$$y = \int_0^x \frac{x^2}{y^2 + 1} dx$$

Setting $y^{(0)} = 0$, we obtain

$$y^{(1)} = \int_0^x x^2 dx = \frac{1}{3}x^3$$

and

$$y^{(2)} = \int_0^x \frac{x^2}{(1/9)x^6 + 1} dx = arctg\left(\frac{1}{3}x^3 \right) = \frac{1}{3}x^3 - \frac{1}{81}x^9 + \cdots$$

so that $y^{(1)}$ and $y^{(2)}$ agree to the first term, viz., $(1/3)x^3$. To find the range of values of x so that the series with the term $(1/3)x^3$ alone will give the result correct to three decimal places, we put

$$\frac{1}{81}x^9 \leq 0.0005$$

which yields

$$x \leq 0.7$$

Hence

$$y(0.25) = \frac{1}{3}(0.25)^3 = 0.005$$
$$y(0.5) = \frac{1}{3}(0.5)^3 = 0.042$$
$$y(1.0) = \frac{1}{3} - \frac{1}{81} = 0.321$$

2.4 Euler's method

We have so far discussed the methods which yield the solution of a differential equation in the form of a power series. We will now describe the methods which give the solution in the form of a set of tabulated values.

Suppose that we wish to solve the Eqs. (2.1) for values of y at $x = x_r = x_0 + rh (r = 1, 2, \ldots)$. Integrating Eq. (2.1), we obtain

$$y_1 = y_0 + \int_{x_0}^{x_1} f(x, y) dx. \tag{2.6}$$

Assuming that $f(x, y) = f(x_0, y_0)$ in $x_0 \leq x \leq x_1$, this gives Euler's formula

$$y_1 \approx y_0 + hf(x_0, y_0). \tag{2.7a}$$

Similarly for the range $x_1 \leq x \leq x_2$, we have

$$y_2 = y_1 + \int_{x_1}^{x_2} f(x, y) dx.$$

Substituting $f(x_1, y_1)$ for $f(x, y)$ in $x_1 \leq x \leq x_2$ we obtain

$$y_2 \approx y_1 + hf(x_1, y_1). \tag{2.7b}$$

Proceeding in this way, we obtain the general formula

$$y_{n+1} = y_n + hf(x_n, y_n), \quad n = 0, 1, 2, \ldots \tag{2.8}$$

The process is very slow and to obtain reasonable accuracy with Euler's method, we need to take a smaller value for h. Because of this restriction on h, the method is unsuitable for practical use and a modification of it, known as the *modified Euler method*, which gives more accurate results, will be described in Section 2.4.2.

Example 2.5. To illustrate Euler's method, we consider the differential equation $y' = -y$ with the condition $y(0) = 1$.
Successive application of Eq. (2.8) with $h = 0.01$ gives

$$\begin{aligned} y(0.01) &= 1 + 0.01(-1) = 0.99 \\ y(0.02) &= 0.99 + 0.01(-0.99) = 0.9801 \\ y(0.03) &= 0.9801 + 0.01(-0.9801) = 0.9703 \\ y(0.04) &= 0.9703 + 0.01(-0.9703) = 0.9606 \end{aligned}$$

The exact solution is $y = e^{-x}$ and from this the value at $x = 0.04$ is 0.9608.

2.4.1 Error Estimates for the Euler Method

Let the true solution of the differential equation at $x = x_n$ be $y(x_n)$ and also let the approximate solution be y_n. Now, expanding $y(x_{n+1})$ by Taylor's series, we get

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \cdots \\ &= y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(\tau_n), \quad \text{where } x_n \le \tau_n \le x_{n+1}. \end{aligned} \tag{2.9}$$

We usually encounter two types of errors in the solution of differential equations. These are (i) local errors, and (ii) rounding errors. The local error is the result of replacing the given differential equation by means of the equation

$$y_{n+1} = y_n + hy'_n.$$

This error is given by

$$L_{n+1} = -\frac{1}{2}h^2 y''(\tau_n). \tag{2.10}$$

The total error is then defined by

$$e_n = y_n - y(x_n). \tag{2.11}$$

Since y_0 is exact, it follows that $e_0 = 0$.
Neglecting the rounding error, we write the total solution error as

$$\begin{aligned} e_{n+1} &= y_{n+1} - y(x_{n+1}) \\ &= y_n + hy'_n - \left[y(x_n) + hy'(x_n) - L_{n+1}\right] \\ &= e_n + h\left[f(x_n, y_n) - y'(x_n)\right] + L_{n+1}. \\ \Rightarrow \quad e_{n+1} &= e_n + h\left[f(x_n, y_n) - f(x_n, y(x_n))\right] + L_{n+1}. \end{aligned}$$

By mean value theorem, we write

$$f(x_n, y_n) - f(x_n, y(x_n)) = [y_n - y(x_n)] \frac{\partial f}{\partial y}(x_n, \xi_n), \quad y(x_n) \le \xi_n \le y_n.$$

Hence, we have

$$e_{n+1} = e_n [1 + h f_y(x_n, \xi_n)] + L_{n+1}. \tag{2.12}$$

Since $e_0 = 0$, we obtain successively:

$$e_1 = L_1; \quad e_2 = [1 + h f_y(x_1, \xi_1)] L_1 + L_2;$$
$$e_3 = [1 + h f_y(x_2, \xi_2)] [1 + h f_y(x_1, \xi_1)] (L_1 + L_2) + L_3; \text{ etc.}$$

See the book by Isaacson and Keller [3] for more details.

Example 2.6. We consider, again, the differential equation $y' = -y$ with the condition $y(0) = 1$, which we have solved by Euler's method in Example 2.5.

Choosing $h = 0.01$, we have

$$1 + h f_y(x_n, \xi_n) = 1 + 0.01(-1) = 0.99,$$

and

$$L_{n+1} = -\frac{1}{2} h^2 y''(\rho_n) = -0.00005 y(\rho_n).$$

In this problem, $y(\rho_n) \le y(x_n)$, since y' is negative. Hence we successively obtain

$$|L_1| \le 0.00005 = 5 \times 10^{-5},$$
$$|L_2| \le (0.00005)(0.99) < 5 \times 10^{-5},$$
$$|L_3| \le (0.00005)(0.9801) < 5 \times 10^{-5},$$

and so on. For computing the total solution error, we need an estimate of the rounding error. If we neglect the rounding error, i.e., if we set

$$R_{n+1} = 0,$$

then using the above bounds, we obtain from Eq. (2.12) the estimates

$$\begin{aligned}
e_0 &= 0, \\
|e_1| &\le 5 \times 10^{-5} \\
|e_2| &\le 0.99 e_1 + 5 \times 10^{-5} < 10^{-4} \\
|e_3| &\le 0.99 e_2 + 5 \times 10^{-5} < 10^{-4} + 5 \times 10^{-5} \\
|e_4| &\le 0.99 e_3 + 5 \times 10^{-5} < 10^{-4} + 10^{-4} = 2 \times 10^{-4} = 0.0002 \\
&\vdots
\end{aligned}$$

It can be verified that the estimate for e_4 agrees with the actual error in the value of $y(0.04)$ obtained in Example 2.5.

2.4.2 Modified Euler's Method

Instead of approximating $f(x, y)$ by $f(x_0, y_0)$ in Eq. (2.6), we now approximate the integral given in Eq. (2.6) by means of trapezoidal rule to obtain

$$y_1 = y_0 + \frac{h}{2}\left[f(x_0, y_0) + f\left(x_1, y_1^{(0)}\right)\right] \tag{2.13}$$

We thus obtain the iteration formula

$$y_1^{(n+1)} = y_0 + \frac{h}{2}\left[f(x_0, y_0) + f\left(x_1, y_1^{(n)}\right)\right], \quad n = 0, 1, 2, \ldots \tag{2.14}$$

where $y_1^{(n)}$ is the nth approximation to y_1. The iteration formula (2.14) can be started by choosing $y_1^{(0)}$ from Euler's formula:

$$y_1^{(0)} = y_0 + hf(x_0, y_0).$$

Example 2.7. Determine the value of y when $x = 0.1$ given that

$$y(0) = 1 \quad \text{and} \quad y' = x^2 + y$$

We take $h = 0.05$. With $x_0 = 0$ and $y_0 = 1.0$, we have $f(x_0, y_0) = 1.0$. Hence Euler's formula gives

$$y_1^{(0)} = 1 + 0.05(1) = 1.05.$$

Further, $x_1 = 0.05$ and $f\left(x_1, y_1^{(0)}\right) = 1.0525$. The average of $f(x_0, y_0)$ and $f\left(x_1, y_1^{(0)}\right)$ is 1.0262. The value of $y_1^{(1)}$ can therefore be computed by using Eq. (2.14) and we obtain

$$y_1^{(1)} = 1.0513.$$

Repeating the procedure, we obtain $y_1^{(2)} = 1.0513$. Hence we take $y_1 = 1.0513$, which is correct to four decimal places.

Next, with $x_1 = 0.05, y_1 = 1.0513$ and $h = 0.05$, we continue the procedure to obtain y_2, i.e., the value of y when $x = 0.1$. The results are

$$y_2^{(0)} = 1.1040, \quad y_2^{(1)} = 1.1055, \quad y_2^{(2)} = 1.1055.$$

Hence we conclude that the value of y when $x = 0.1$ is 1.1055.

2.5 Runge-Kutta Methods

As already mentioned, Euler's method is less efficient in practical problems since it requires h to be small for obtaining reasonable accuracy. The Runge-Kutta methods are designed to give greater accuracy and they possess the advantage of requiring only the function values at some selected points on the subinterval.

If we substitute $y_1 = y_0 + hf\,(x_0, y_0)$ on the right side of Eq. (2.13), we obtain

$$y_1 = y_0 + \frac{h}{2}\left[f_0 + f\,(x_0 + h, y_0 + hf_0)\right],$$

where $f_0 = f\,(x_0, y_0)$. If we now set

$$k_1 = hf_0 \quad \text{and} \quad k_2 = hf\,(x_0 + h, y_0 + k_1)$$

then the above equation becomes

$$y_1 = y_0 + \frac{1}{2}\,(k_1 + k_2)\,, \tag{2.15}$$

which is the *second-order Runge-Kutta* formula. The error in this formula can be shown to be of order h^3 by expanding both sides by Taylor's series. Thus, the left side gives

$$y_0 + hy_0' + \frac{h^2}{2}y_0'' + \frac{h^3}{6}y_0''' + \cdots$$

and on the right side

$$k_2 = hf\,(x_0 + h, y_0 + hf_0) = h\left[f_0 + h\frac{\partial f}{\partial x_0} + hf_0\frac{\partial f}{\partial y_0} + \mathrm{O}\left(h^2\right)\right].$$

Since

$$\frac{df(x,y)}{dx} = \frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y},$$

we obtain

$$k_2 = h\left[f_0 + hf_0' + \mathrm{O}\left(h^2\right)\right] = hf_0 + h^2 f_0' + \mathrm{O}\left(h^3\right),$$

so that the right side of Eq. (2.15) gives

$$\begin{aligned} y_0 + \frac{1}{2}\left[hf_0 + hf_0 + h^2 f_0' + \mathrm{O}\left(h^3\right)\right] &= y_0 + hf_0 + \frac{1}{2}h^2 f_0' + \mathrm{O}\left(h^3\right) \\ &= y_0 + hy_0' + \frac{h^2}{2}y_0'' + \mathrm{O}\left(h^3\right). \end{aligned}$$

It therefore follows that the Taylor series expansions of both sides of Eq. (2.15) agree up to terms of order h^2, which means that the error in this formula is of order h^3.

More generally, if we set

$$y_1 = y_0 + W_1 k_1 + W_2 k_2 \tag{2.16a}$$

where

$$\left.\begin{aligned} k_1 &= h f_0 \\ k_2 &= h f\left(x_0 + \alpha_0 h, y_0 + \beta_0 k_1\right) \end{aligned}\right\} \tag{2.16b}$$

then the Taylor series expansions of both sides of the last equation in (2.16a) gives the identity

$$\begin{aligned} y_0 + h f_0 + \frac{h^2}{2}\left(\frac{\partial f}{\partial x} + f_0 \frac{\partial f}{\partial y}\right) + \mathrm{O}\left(h^3\right) = y_0 + \left(W_1 + W_2\right) h f_0 \\ + W_2 h^2 \left(\alpha_0 \frac{\partial f}{\partial x} + \beta_0 f_0 \frac{\partial f}{\partial y}\right) + \mathrm{O}\left(h^3\right). \end{aligned}$$

Equating the coefficients of $f(x, y)$ and its derivatives on both sides, we obtain the relations

$$W_1 + W_2 = 1, \quad W_2 \alpha_0 = \frac{1}{2}, \quad W_2 \beta_0 = \frac{1}{2}. \tag{2.17}$$

Clearly, $\alpha_0 = \beta_0$ and if α_0 is assigned any value arbitrarily, then the remaining parameters can be determined uniquely. If we set, for example, $\alpha_0 = \beta_0 = 1$, then we immediately obtain $W_1 = W_2 = 1/2$, which gives formula (2.15).

It follows, therefore, that there are several second-order Runge-Kutta formulae and that formulae (2.16) and (2.17) constitute just one of several such formulae.

Higher-order Runge-Kutta formulae exist, of which we mention only the *fourth-order formula* defined by

$$y_1 = y_0 + W_1 k_1 + W_2 k_2 + W_3 k_3 + W_4 k_4 \tag{2.18a}$$

where

$$\left.\begin{aligned} k_1 &= h f\left(x_0, y_0\right) \\ k_2 &= h f\left(x_0 + \alpha_0 h, y_0 + \beta_0 k_1\right) \\ k_3 &= h f\left(x_0 + \alpha_1 h, y_0 + \beta_1 k_1 + v_1 k_2\right) \\ k_4 &= h f\left(x_0 + \alpha_2 h, y_0 + \beta_2 k_1 + v_2 k_2 + \delta_1 k_3\right), \end{aligned}\right\} \tag{2.18b}$$

where the parameters have to be determined by expanding both sides of the first equation of (2.18a) by Taylor's series and securing agreement of terms up to and including those containing h^4. The choice of the parameters is, again, arbitrary and we have therefore several fourth-order Runge-Kutta formulae. If, for example, we set

$$\left.\begin{array}{lll} \alpha_0 = \beta_0 = \frac{1}{2}, & \alpha_1 = \frac{1}{2}, & \alpha_2 = 1, \\ \beta_1 = \frac{1}{2}(\sqrt{2}-1), & \beta_2 = 0 & \\ v_1 = 1 - \frac{1}{\sqrt{2}}, & v_2 = -\frac{1}{\sqrt{2}}, & \delta_1 = 1 + \frac{1}{\sqrt{2}}, \\ W_1 = W_4 = \frac{1}{6}, & W_2 = \frac{1}{3}\left(1 - \frac{1}{\sqrt{2}}\right), & W_3 = \frac{1}{3}\left(1 + \frac{1}{\sqrt{2}}\right), \end{array}\right\} \tag{2.19}$$

we obtain the method of Gill, whereas the choice

$$\left.\begin{array}{ll} \alpha_0 = \alpha_1 = \frac{1}{2}, & \beta_0 = v_1 = \frac{1}{2} \\ \beta_1 = \beta_2 = v_2 = 0, & \alpha_2 = \delta_1 = 1 \\ W_1 = W_4 = \frac{1}{6}, & W_2 = W_3 = \frac{2}{6} \end{array}\right\} \tag{2.20}$$

leads to the fourth-order Runge-Kutta formula, the most commonly used one in practice:

$$y_1 = y_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \tag{2.21a}$$

where

$$\left.\begin{array}{l} k_1 = hf\left(x_0, y_0\right) \\ k_2 = hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right) \\ k_3 = hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_2\right) \\ k_4 = hf\left(x_0 + h, y_0 + k_3\right) \end{array}\right\} \tag{2.21b}$$

in which the error is of order h^5. Complete derivation of the formula is exceedingly complicated, and the interested reader is referred to the book by Levy and Baggott [4]. We illustrate here the use of the fourth-order formula by means of examples.

Example 2.8. Given $dy/dx = y - x$ where $y(0) = 2$, find $y(0.1)$ and $y(0.2)$ correct to four decimal places.

(i) *Runge-Kutta second-order formula*: With $h = 0.1$, we find $k_1 = 0.2$ and $k_2 = 0.21$. Hence

$$y_1 = y(0.1) = 2 + \frac{1}{2}(0.41) = 2.2050.$$

To determine $y_2 = y(0.2)$, we note that $x_0 = 0.1$ and $y_0 = 2.2050$. Hence, $k_1 = 0.1(2.105) = 0.2105$ and $k_2 = 0.1(2.4155 - 0.2) = 0.22155$.

It follows that

$$y_2 = 2.2050 + \frac{1}{2}(0.2105 + 0.22155) = 2.4210.$$

Proceeding in a similar way, we obtain

$$y_3 = y(0.3) = 2.6492 \quad \text{and} \quad y_4 = y(0.4) = 2.8909$$

We next choose $h = 0.2$ and compute $y(0.2)$ and $y(0.4)$ directly. With $h = 0.2$, $x_0 = 0$ and $y_0 = 2$, we obtain $k_1 = 0.4$ and $k_2 = 0.44$ and hence $y(0.2) = 2.4200$. Similarly, we obtain $y(0.4) = 2.8880$.

From the analytical solution $y = x + 1 + e^x$, the exact values of $y(0.2)$ and $y(0.4)$ are respectively 2.4214 and 2.8918. To study the order of convergence of this method, we tabulate the values as follows:

x	Computed y	Exact y	Difference	Ratio
0.2	$h = 0.1 : 2.4210$	2.4214	0.0004	
	$h = 0.2 : 2.4200$		0.0014	3.5
0.4	$h = 0.1 : 2.8909$	2.8918	0.0009	
	$h = 0.2 : 2.8880$		0.0038	4.2

It follows that the method has an h^2-order of convergence.

(ii) *Runge-Kutta fourth-order formula*: To determine $y(0.1)$, we have $x_0 = 0, y_0 = 2$ and $h = 0.1$. We then obtain

$$\begin{aligned} k_1 &= 0.2, \\ k_2 &= 0.205, \\ k_3 &= 0.20525, \\ k_4 &= 0.21053. \end{aligned}$$

Hence

$$y(0.1) = 2 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = 2.2052.$$

Proceeding similarly, we obtain $y(0.2) = 2.4214$.

Example 2.9. Given $dy/dx = 1 + y^2$, where $y = 0$ when $x = 0$, find $y(0.2)$, $y(0.4)$ and $y(0.6)$.

We take $h = 0.2$. With $x_0 = y_0 = 0$, we obtain from (2.21a) and (2.21b),

$$
\begin{aligned}
k_1 &= 0.2, \\
k_2 &= 0.2(1.01) = 0.202, \\
k_3 &= 0.2(1 + 0.010201) = 0.20204, \\
k_4 &= 0.2(1 + 0.040820) = 0.20816,
\end{aligned}
$$

and

$$
y(0.2) = 0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = 0.2027,
$$

which is correct to four decimal places.

To compute $y(0.4)$, we take $x_0 = 0.2, y_0 = 0.2027$ and $h = 0.2$. With these values, Eqs. (2.21a) and (2.21b) give

$$
\begin{aligned}
k_1 &= 0.2\left[1 + (0.2027)^2\right] = 0.2082, \\
k_2 &= 0.2\left[1 + (0.3068)^2\right] = 0.2188, \\
k_3 &= 0.2\left[1 + (0.3121)^2\right] = 0.2195, \\
k_4 &= 0.2\left[1 + (0.4222)^2\right] = 0.2356,
\end{aligned}
$$

and

$$
y(0.4) = 0.2027 + 0.2201 = 0.4228,
$$

correct to four decimal places.

Finally, taking $x_0 = 0.4, y_0 = 0.4228$ and $h = 0.2$, and proceeding as above, we obtain $y(0.6) = 0.6841$.

Example 2.10. We consider the initial value problem $y' = 3x + y/2$ with the condition $y(0) = 1$.

The following table gives the values of $y(0.2)$ by different methods, the exact value being 1.16722193 . It is seen that the *fourth-order Runge-Kutta* method gives the accurate value for $h = 0.05$.

Method	h	*Computed value*
Euler	0.2	1.100 000 00
	0.1	1.132 500 00
	0.05	1.149 567 58
Modified Euler	0.2	1.100 000 00
	0.1	1.150 000 00
	0.05	1.162 862 42
Fourth-order Runge-Kutta	0.2	1.167 220 83
	0.1	1.167 221 86
	0.05	1.167 221 93

3 Systems of first-order ordinary differential equations. Higher-order differential equations

3.1 Cubic spline method

The governing equations of a cubic spline have been discussed detail previously, where the cubic spline function has been obtained in terms of its second derivatives, M_i. In certain applications, e.g. the solution of initial-value problems, it would be convenient to use governing equations in terms of its first derivatives, i.e., m_i. Using Hermite's interpolation formula, it would not be difficult to derive the following formula for cubic spline $s(x)$ in $x_{i-1} \le x \le x_i$ in terms of its first derivatives $s'(x_i) = m_i$:

$$\begin{aligned} s(x) =& m_{i-1}\frac{(x_i - x)^2(x - x_{i-1})}{h^2} - m_i\frac{(x - x_{i-1})^2(x_i - x)}{h^2} \\ &+ y_{i-1}\frac{(x_i - x)^2[2(x - x_{i-1}) + h]}{h^3} + y_i\frac{(x - x_{i-1})^2[2(x_i - x) + h]}{h^3}, \end{aligned} \tag{3.1}$$

where $h = x_i - x_{i-1}$. Differentiating Eq. (3.1) with respect to x and simplifying we obtain

$$\begin{aligned} s'(x) =& \frac{m_{i-1}}{h^2}(x_i - x)(2x_{i-1} + x_i - 3x) - \frac{m_i}{h^2}(x - x_{i-1})(x_{i-1} + 2x_i - 3x) \\ &+ \frac{6(y_i - y_{i-1})}{h^3}(x - x_{i-1})(x_i - x). \end{aligned} \tag{3.2}$$

Again,

$$\begin{aligned} s''(x) =& -\frac{2m_{i-1}}{h^2}(x_{i-1} + 2x_i - 3x) - \frac{2m_i}{h^2}(2x_{i-1} + x_i - 3x) \\ &+ \frac{6(y_i - y_{i-1})}{h^3}(x_{i-1} + x_i - 2x), \end{aligned} \tag{3.3}$$

which gives

$$\begin{aligned} s''(x) =& \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(y_i - y_{i-1}) \\ =& \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(s_i - s_{i-1}). \end{aligned} \tag{3.4}$$

If we now consider the initial-value problem

$$\frac{dy}{dx} = f(x, y) \tag{3.5a}$$

and

$$y(x_0) = y_0 \tag{3.5b}$$

then from Eq. (3.5a), we obtain

$$\frac{d^2y}{dx^2} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial x},$$

or

$$\begin{aligned} y''(x_i) =& f_x(x_i, y_i) + f_y(x_i, y_i) f(x_i, y_i) \\ =& f_x(x_i, s_i) + f_y(x_i, s_i) f(x_i, s_i). \end{aligned} \tag{3.6}$$

Equating Eqs. (3.4) and (3.6), we obtain

$$\frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(s_i - s_{i-1}) = f_x(x_i, s_i) + f_y(x_i, s_i) f(x_i, s_i) \tag{3.7}$$

from which s_i can be computed. Substitution in Eq. (3.1) gives the required solution.

The following example demonstrates the usefulness of the spline method.

Example 3.1. We consider again the initial-value problem defined by

$$y' = 3x + \frac{1}{2}y, \qquad y(0) = 1, \tag{i}$$

whose exact solution is given by

$$y = 13e^{x/2} - 6x - 12 \tag{ii}$$

We take, for simplicity, $n = 2$, i.e. $h = 0.5$ and compute the value of $y(0.5)$. Here $f(x, y) = 3x + y/2$ and therefore we have $f_x = 3$ and $f_y = 1/2$. Also,

$$f(x_i, s_i) = 3x_i + \frac{1}{2}s_i.$$

Hence, Eq. (3.7) gives

$$4m_0 + 8m_1 - 24(s_1 - s_0) = 3 + \frac{1}{2}\left(\frac{3}{2} + \frac{1}{2}s_1\right)$$

and

$$4m_1 + 8m_2 - 24(s_2 - s_1) = 3 + \frac{1}{2}\left(\frac{3}{2} + \frac{1}{2}s_2\right).$$

Since $m_0 = 1/2, m_1 = 3/2 + s_1/2$ and $m_2 = 3 + s_2/2$, the above equations give on simplification

$$s_1 = 1.691358 \quad \text{and} \quad s_2 = 3.430879.$$

The errors in these solutions are given by 0.000972 and 0.002497, respectively. It can be shown that, under certain conditions, the spline method gives $\mathrm{O}(h^4)$ convergence and compares well with multi-step Milne's method. For details, the reader is referred to Patricio [5].

3.2 Boundary-value problems

Some simple examples of two-point linear boundary-value problems are:

(a) $$y''(x) + f(x)y'(x) + g(x)y(x) = r(x) \tag{3.8}$$

with the boundary conditions

$$y(x_0) = a \qquad \text{and} \qquad y(x_n) = b \tag{3.9}$$

(b) $$y^{iv} = p(x)y(x) + q(x) \tag{3.10}$$

with

$$y(x_0) = y'(x_0) = A \qquad \text{and} \qquad y(x_n) = y'(x_n) = B \tag{3.11}$$

Problems of the type (b), which involve the four-order differential equation, are much involved and will not be discussed here. There exist many methods of solving second-order boundary-value problems of type (a). Of these, the finite difference method is a popular one and will be described in Section 3.3. Finally, in Section 3.4 we discuss method based on the application of weighted residuals.

3.3 Finite-difference Method

The finite-difference method for the solution of two-point boundary value problem consists in replacing the derivatives occurring in the differential equation (and in the boundary conditions as well) by means of their finite-difference approximations and then solving the resulting linear system of equations by a standard procedure.

To obtain the appropriate finite-difference approximations to the derivatives, we proceed as follows.

Expanding $y(x + h)$ in Taylor's series, we have

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \frac{h^3}{6}y'''(x) + \dots \tag{3.12}$$

from which we obtain

$$y'(x) = \frac{y(x + h) - y(x)}{h} - \frac{h}{2}y''(x) - \dots$$

Thus we have

$$y'(x) = \frac{y(x + h) - y(x)}{h} + O(h) \tag{3.13}$$

which is forward difference approximation for $y'(x)$. Similarly, expansion of $y(x - h)$ in Taylor's series gives

$$y(x - h) = y(x) - hy'(x) + \frac{h^2}{2}y''(x) - \frac{h^3}{6}y'''(x) + \dots \tag{3.14}$$

from which we obtain

$$y'(x) = \frac{y(x) - y(x-h)}{h} + O(h) \tag{3.15}$$

which is the backward difference approximation for $y'(x)$.

A central difference approximation for $y'(x)$ can be obtained by subtracting Eq. (3.14) from Eq. (3.12). We thus have

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + O(h^2). \tag{3.16}$$

It is clean that Eq. (3.16) is a better approximation to $y'(x)$ then either Eq. (3.13) or Eq. (3.15). Again, adding Eq. (3.12) and Eq. (3.14), we get an approximation for $y''(x)$

$$y''(x) = \frac{y(x-h) - 2y(x) + y(x+h)}{h^2} + O(h^2). \tag{3.17}$$

In a similar manner, it is possible to derive finite-difference approximations to higher derivatives.

To solve the boundary-value problem defined by (3.8) and (3.9), we divide the range $[x_0,\ x_n]$ into n equal subintervals of width h so that

$$x_i = x_0 + ih, \quad i = 1,\ 2,\ \ldots,\ n.$$

The corresponding values of y at these points are denoted by

$$y(x_i) = y_i = y(x_0 + ih), \quad i = 0,\ 1,\ 2,\ \ldots,\ n.$$

From Eqs. (3.16) and (3.17), values of $y'(x)$ and $y''(x)$ at the point $x = x_i$ can now be written as

$$y_i' = \frac{y_{i+1} - y_{i-1}}{2h} + O(h^2)$$

and

$$y_1'' = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + O(h^2).$$

Satisfying the differential equation at the point $x = x_i$, we get

$$y_i'' + f_i y_i' + g_i y_i = r_i$$

Substituting the expressions for y_i' and y_i'', this gives

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + f_i \frac{y_{i+1} - y_{i-1}}{2h} + g_i y_i = r_i, \qquad i = 1, 2, \ldots, n-1,$$

where $y_i = y(x_i), g_i = g(x_i)$, etc.

Multiplying through by h^2 and simplifying, we obtain

$$\left(1-\frac{h}{2}f_i\right)y_{i-1}+\left(-2+g_ih^2\right)y_i+\left(1+\frac{h}{2}f_i\right)y_{i+1}=r_ih^2, \qquad i=1,\ 2,\ \ldots,\ n-1 \tag{3.18}$$

with

$$y_0=a \quad \text{and} \quad y_n=b \tag{3.19}$$

Equation (3.18) with the conditions (3.19) comprise a tridiagonal system which can be solved by the next method.

Solution of Tridiagonal Systems

Consider the system of equations defined by

$$\left.\begin{aligned} b_1u_1+c_1u_2&=d_1\\ a_2u_1+b_2u_2+c_2u_3&=d_2\\ a_3u_2+b_3u_3+c_3u_4&=d_3\\ &\vdots\\ a_nu_{n-1}+b_nu_n&=d_n. \end{aligned}\right\} \tag{*.1}$$

The matrix of coefficients is

$$A=\begin{bmatrix} b_1 & c_1 & 0 & 0 & \cdots & \cdots & \cdots & 0\\ a_2 & b_2 & c_2 & 0 & \cdots & \cdots & \cdots & 0\\ 0 & a_3 & b_3 & c_3 & \cdots & \cdots & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & \cdots & 0 & a_{n-1} & b_{n-1} & c_{n-1}\\ 0 & 0 & 0 & \cdots & 0 & 0 & a_n & b_n \end{bmatrix} \tag{*.2}$$

Matrices of the type, given in Eq. (*.2), called the tridiagonal matrices, occur frequently in the solution of ordinary and partial differential equations by finite difference methods. The method of factorization described earlier can be conveniently applied to solve the system (*.1). For example, for a (3×3) matrix we have

$$\begin{bmatrix} b_1 & c_1 & 0\\ a_2 & b_2 & c_2\\ 0 & a_3 & b_3 \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0\\ l_{21} & 1 & 0\\ 0 & l_{32} & 1 \end{bmatrix}\begin{bmatrix} b_1 & c_1 & 0\\ 0 & u_{22} & c_2\\ 0 & 0 & u_{33} \end{bmatrix}$$

This matrix equation gives

$$\left.\begin{aligned} l_{21}b_1&=a_2, \quad & l_{21}c_1+u_{22}&=b_2\\ l_{32}u_{22}&=a_3, \quad & l_{32}c_2+u_{33}&=b_3 \end{aligned}\right\} \tag{*.3}$$

From these four equations, we can compute l_{21}, u_{22}, l_{32} and u_{33} and these values are stored in the locations occupied by a_2, b_2, a_3 and b_3, respectively. These computations can be achieved by the following statements:

$$\begin{aligned} &\text{Do } i = 2(1)N \\ &a(i) = a(i)/b(i-1) \\ &b(i) = b(i) - a(i)c(i-1) \end{aligned}$$

Next i

When the decomposition is complete, forward and back substitutions give the required solution. This algorithm is due to Thomas and possesses all the advantages of the *LU* decomposition.

Return to boundary value problem

The solution of this tridiagonal system constitutes an approximate solution of the boundary value problem defined by Eqs. (3.18) and (3.19).

To estimate the error in the numerical solution, we define the *local truncation error,* τ, by

$$\tau = \left(\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - y_i''\right) + f_i\left(\frac{y_{i+1} - y_{i-1}}{2h} - y_i'\right).$$

Expanding y_{i-1} and y_{i+1} by Taylor's series and simplifying, the above gives

$$\tau = \frac{h^2}{12}(y_i^{iv} + 2f_i y_i''') + O(h^4). \tag{3.20}$$

Thus, the finite difference approximation defined by Eq. (3.18) has second-order accuracy for functions with continuous fourth derivatives on $[x_0, x_n]$. Further, it follows that $\tau \to 0$ as $h \to 0$, implying that greater accuracy in the result can be achieved by using a smaller value of h. In such a case, of course, more computational effort would be required since the number of equations become larger.

An easier way to improve accuracy is to employ *Richardson's deferred approach to the limit,* assuming that the $O(h^2)$ error is proportional to h^2. This means that the error has the form

$$y(x_i) - y_i = h^2 e(x_i) + O(h^4) \tag{3.21}$$

For extrapolation to the limit, we solve Eq. (3.18) twice, with the interval lengths h and $h/2$ respectively. Let the corresponding solutions of Eq.(3.18) be denoted by $y_i(h)$ and $y_i(h/2)$. For a point x_i common to both, we therefore have

$$y(x_i) - y_i(h) = h^2 e(x_i) + O(h^4) \tag{3.22a}$$

and

$$y(x_i) - y_i\left(\frac{h}{2}\right) = \frac{h^2}{4} e(x_i) + O(h^4) \tag{3.22b}$$

from which we obtain

$$y(x_i) = \frac{4y_i(h/2) - y_i(h)}{3}. \tag{3.23}$$

We have explained the method with simple boundary conditions (3.19) where the function values on the boundary are prescribed. In many applied problems, however, derivative boundary conditions may be prescribed, and this requires a modification of the procedures described above. The following examples illustrate the application of the finite-difference method.

Example 3.2. A boundary-value problem is defined by

$$y'' + y + 1 = 0, \quad 0 \le x \le 1$$

where

$$y(0) = 0 \quad \text{and} \quad y(1) = 0.$$

With $h = 0.5$, use the finite-difference method to determine the value of $y(0.5)$.

This example was considered by Bickley [1]. Its exact solution is given by

$$y(x) = \cos x + \frac{1 - \cos 1}{\sin 1} \sin x - 1,$$

from which, we obtain

$$y(0.5) = 0.139493927.$$

Here $nh = 1$. The differential equation is approximated as

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + y_i + 1 = 0$$

and this gives after simplification

$$y_{i-1} - \left(2 - h^2\right) y_i + y_{i+1} = -h^2, \qquad i = 1,\ 2,\ \ldots,\ n-1$$

which together with the boundary conditions $y_0 = 0$ and $y_n = 0$, comprises a system of $(n+1)$ equations for the $(n+1)$ unknowns $y_0, y_1, \ldots, y_n$.

Choosing $h = 1/2$ (i.e. $n = 2$), the above system becomes

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

With $y_0 = y_2 = 0$, this gives

$$y_1 = y(0.5) = \frac{1}{7} = 0.142857142\ldots$$

Comparison with the exact solution given above shows that the error in the computed solution is 0.00336.

On the other hand, if we choose $h = 1/4$ (i.e. $n = 4$), we obtain the three equations:

$$\begin{aligned} y_0 - \frac{31}{16}y_1 + y_2 &= -\frac{1}{16} \\ y_1 - \frac{31}{16}y_2 + y_3 &= -\frac{1}{16} \\ y_2 - \frac{31}{16}y_3 + y_4 &= -\frac{1}{16}, \end{aligned}$$

where $y_0 = y_4 = 0$. Solving the system we obtain

$$y_2 = y(0.5) = \frac{63}{449} = 0.140311804,$$

the error in which is 0.00082. Since the ratio of the two errors is about 4, it follows that the order of convergence is h^2.

These results show that the accuracy obtained by the finite-difference method depends upon the width of the subinterval chosen and also on the order of the approximations. As h is reduced, the accuracy increases but the number of equations to be solved also increases.

Example 3.3. Solve the boundary-value problem

$$\frac{d^2y}{dx^2} - y = 0$$

with

$$y(0) = 0 \quad \text{and} \quad y(2) = 3.62686.$$

The exact solution of this problem is $y = \sinh x$. The finite-difference approximation is given by

$$\frac{1}{h^2}\left(y_{i-1} - 2y_i + y_{i+1}\right) = y_i. \tag{i}$$

We subdivide the interval $[0, 2]$ into four equal parts so that $h = 0.5$. Let the values of y at the five points be y_0, y_1, y_2, y_3 and y_4. We are given that

$$y_0 = 0 \quad \text{and} \quad y_4 = 3.62686.$$

Writing the difference equations at the three interval points (which are the unknowns), we obtain

$$\left.\begin{aligned} 4\left(y_0 - 2y_1 + y_2\right) &= y_1 \\ 4\left(y_1 - 2y_2 + y_3\right) &= y_2 \\ 4\left(y_2 - 2y_3 + y_4\right) &= y_3, \end{aligned}\right\} \tag{ii}$$

respectively. Substituting for y_0 and y_4 and rearranging, we get the system

$$\left.\begin{aligned} -9y_1 + 4y_2 &= 0 \\ 4y_1 - 9y_2 + 4y_3 &= 0 \\ 4y_2 - 9y_3 &= -14.50744. \end{aligned}\right\} \tag{iii}$$

The solution of (iii) is given in the table below.

x	*Computed value of* y	*Exact value* $y = \sinh x$	Error
0.5	0.52635	0.52110	0.00525
1.0	1.18428	1.17520	0.00908
1.5	2.13829	2.12928	0.00901

It is possible to obtain a better approximation for the value of $y(1.0)$ by extrapolation to the limit. For this we divide the interval $[0, 2]$ into two subintervals with $h = 1.0$. The difference equation at the single unknown point y_1 is given by

$$y_0 - 2y_1 + y_2 = y_1$$

Using the values of y_0 and y_2, we obtain

$$y_1 = 1.20895.$$

Hence Eq. (3.23) gives

$$y(1.0) = \frac{4(1.18428) - 1.20895}{3} = 1.17606,$$

which is a better approximation since the error is now reduced to 0.00086.

3.4 Galerkin's Method

This method, also called the *weighted residual* method, uses *trial functions* (or approximating functions) which satisfy the boundary conditions of the problem. The trial function is substituted in the given differential equation and the result is called the *residual*. The integral of the product of this residual and a weighted function, taken over the domain, is then set to zero which yields a system of equations for the unknown parameters in the trial functions.

Let the boundary value problem be defined by

$$y'' + p(x)y' + q(x)y = f(x) \quad a < x < b \tag{3.24}$$

with the boundary conditions

$$\left.\begin{aligned} p_0 y(a) + q_0 y'(a) &= r_0 \\ p_1 y(b) + q_1 y'(b) &= r_1 \end{aligned}\right\} \tag{3.25}$$

Let the approximate solution be given by

$$t(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x), \tag{3.26}$$

where $\phi_i(x)$ are called *base functions*. Substituting for $t(x)$ in Eq. (3.24), we obtain a residual. Denoting this residual by $R(t)$, we obtain

$$R(t) = t'' + p(x)t' + q(x)t - f(x). \tag{3.27}$$

Usually the base functions $\phi_i(x)$ are chosen as weight functions. We, therefore, have

$$I = \int_a^b \phi_i(x)R(t)dx = 0, \tag{3.28}$$

which yields a system of equations for the parameters α_i. When α_i are known, $t(x)$ can be calculated from Eq. (3.26).

Example 3.4. Solve the boundary value problem defined by

$$y'' + y + x = 0, \quad 0 < x < 1$$

with the conditions

$$y(0) = y(1) = 0.$$

Let

$$t(x) = \alpha_1 \phi_1(x).$$

Since both the boundary conditions must be satisfied by $t(x)$, we choose

$$\phi_1(x) = x(1-x).$$

Substituting for $t(x)$ in the given differential equation, we obtain

$$R(t) = t'' + t + x.$$

Hence we have

$$\begin{aligned} I &= \int_0^1 \left(t'' + t + x\right) \alpha_1 x(1-x)dx = 0 \\ \Rightarrow \quad &\int_0^1 \left(t'' + t + x\right) x(1-x)dx = 0 \end{aligned} \tag{i}$$

Now,

$$\int_0^1 t''x(1-x)dx = \left[t'x(1-x)\right]_0^1 - \int_0^1 t'(1-2x)dx,$$

on integrating by parts.

$$\begin{aligned} &= -\int_0^1 t'(1-2x)dx, \text{ since the first term vanishes.} \\ &= -\left[\{t(1-2x)\}_0^1 - \int_0^1 t(-2)dx\right] \\ &= -2\int_0^1 tdx, \text{ since } t = 0 \text{ at } x = 0 \text{ and } x = 1. \end{aligned}$$

Hence (i) simplifies to

$$\begin{aligned} &-2\int_0^1 tdx + \int_0^1 tx(1-x)dx + \int_0^1 x^2(1-x)dx = 0 \\ \Rightarrow \qquad &-2\int_0^1 \alpha_1 x(1-x)dx + \int_0^1 \alpha_1 x^2(1-x)^2 dx + \left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 = 0 \\ \Rightarrow \qquad &\alpha_1 = \frac{5}{18} = 0.2778, \text{ an simplification.} \end{aligned}$$

Then a first approximation to the solution is

$$y(0.5) = \frac{5}{18}(0.5)(0.5) = 0.06944.$$

The exact solution to the given boundary value problem is

$$y(x) = \frac{\sin x}{\sin 1} - x,$$

which means that our solution has an error of 0.0003.

The above approximation can be improved by assuming that

$$t(x) = \alpha_1 x(1-x) + \alpha_2 x^2(1-x).$$

Proceeding as above, we obtain

$$\alpha_1 = 0.1924 \text{ and } \alpha_2 = 0.1707.$$

It is clear that by adding more terms to $t(x)$, we can obtain the result to the desired accuracy.

4 Numerical solution of Partial Differential Equations.

4.1 Introduction

Partial differential equations occur in many branches of applied mathematics, for example, hydrodynamics, elasticity, quantum mechanics and electromagnetic theory. The analytical treatment of these equations is a rather involved process since it requires application of advanced mathematical techniques. On the other hand, it is generally easier to produce sufficiently approximate solutions by simple and efficient numerical methods. There exist several numerical for the solution of partial differential equations; for example, finite difference methods, spline methods, finite element methods, integral equation methods, etc. Of these, only the finite difference methods have become popular and are more gainfully employed than others. In this chapter, we discuss these methods, very briefly, and apply them to solve simple problems.

The general second order linear partial differential equation is of the form

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu = G,$$

which can be written as

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G \tag{4.1}$$

where $A,\ B,\ C,\ D,\ E,\ F$ and G are all functions of x and y. Equations of form (4.1) can be classified with respect to sign of the discriminant

$$\Delta = B^2 - 4AC, \tag{4.2}$$

where Δ is computed at any point in the (x, y) plane. Equation (4.1) is said to be *elliptic, parabolic* or *hyperbolic* according as $\Delta < 0, \Delta = 0, \Delta > 0$.

For example,

$$u_{xx} + u_{yy} = 0 \quad \text{(Laplace equation) is elliptic} \tag{4.3}$$

$$u_{xx} - u_{yy} = 0 \quad \text{(Wave equation) is hyperbolic} \tag{4.4}$$

$$u_t = u_{xx} \quad \text{(heat conduction equation) is parabolic} \tag{4.5}$$

In the study of partial differential equations, usually three types of problems arise:

(i) *Dirichlet's Problem.* Given a continuous function f of the boundary C of a region R, it is required to find function $u(x, y)$, satisfying the Laplace equation in R, i.e.m to find $u(x, y)$ such that

$$\left.\begin{aligned} u_{xx} + u_{yy} &= 0 \text{ in } R, \\ \text{and} \qquad u &= f \text{ on } C. \end{aligned}\right\} \tag{4.6}$$

(ii) *Cauchy's Problem.*

$$\left.\begin{aligned} u_{tt} - u_{xx} &= 0 \text{ for } t > 0, \\ u(x,0) &= f(x) \\ \text{and} \quad \tfrac{\partial u(x,0)}{\partial t} &= g(x) \end{aligned}\right\} \tag{4.7}$$

where $f(x)$ and $g(x)$ are arbitrary.

(iii)

$$\left.\begin{aligned} u_t &= u_{xx} \text{ for } t > 0, \\ \text{and } u(x,0) &= f(x) \end{aligned}\right\} \tag{4.8}$$

In partial differential equations, the form of the equation is always associated with a particular type of boundary conditions. In this case, the problem is said to be *well-defined* (or well-posed). The problems defined in Eqs. (4.6) to (4.8) are well-posed. If, however, we associate Laplace equation with Cauchy boundary conditions, the problem is said to be *ill-posed*. Thus, the problem defined by

$$\left.\begin{aligned} u_{xx} + u_{yy} &= 0 \\ u(x,0) &= f(x) \\ \text{and} \quad u_y(x,0) &= g(x) \end{aligned}\right\} \tag{4.9}$$

is an ill-posed problem.

4.2 Finite-Difference approximations to derivatives

Let the (x, y) plane be divided into a network of rectangles of sides $\Delta x = h$ and $\Delta y = k$ by drawing the sets of lines

$$\begin{aligned} x &= ih, \quad i = 0, 1, 2, \ldots \\ y &= jk, \quad j = 0, 1, 2, \ldots \end{aligned}$$

The points of intersection of these families of lines are called *mesh* points, *lattice* points or *grid* points. Then, we have

$$u_x = \frac{u_{i+1,j} - u_{i,j}}{h} + O(h) \tag{4.10}$$

$$= \frac{u_{i,j} - u_{i-1,j}}{h} + O(h) \tag{4.11}$$

$$= \frac{u_{i+1,j} - u_{i-1,j}}{2h} + O(h^2) \tag{4.12}$$

and

$$u_{xx} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + O(h^2) \tag{4.13}$$

where

$$u_{i,j} = u(ih, jk) = u(x, y)$$

Similarly, we have the approximations

$$u_y = \frac{u_{i,j+1} - u_{i,j}}{k} + O(k) \tag{4.14}$$

$$= \frac{u_{i,j} - u_{i,j-1}}{k} + O(k) \tag{4.15}$$

$$= \frac{u_{i,j+1} - u_{i,j-1}}{2k} + O(k^2) \tag{4.16}$$

and

$$u_{yy} = \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{k^2} + O(k^2) \tag{4.17}$$

We can now obtain the *finite-difference analogues* of partial differential equations by replacing the derivatives in any equations by their corresponding difference approximations given above. Thus, the Laplace equation in two dimensions, namely

$$u_{xx} + u_{yy} = 0$$

has its finite-difference analogue

$$\frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + \frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = 0. \tag{4.18}$$

If $h = k$, this gives

$$u_{i,j} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}), \tag{4.19}$$

which shows that the value of u at any point is the mean of its values at the four neighbouring points. This is called the *standard five-point formula* [see Fig. 4.1(a)], and is written

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0 \tag{4.20}$$

By expanding the terms on the right side of Eq. (4.19) by Taylor's series, it can be shown that

$$\begin{aligned} u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} &= h^2(u_{xx} + u_{yy}) - \frac{1}{6}h^4 u_{xxyy} + O(h^6) \\ &= -\frac{1}{6}h^4 u_{xxyy} + O(h^6) \end{aligned} \tag{4.21}$$

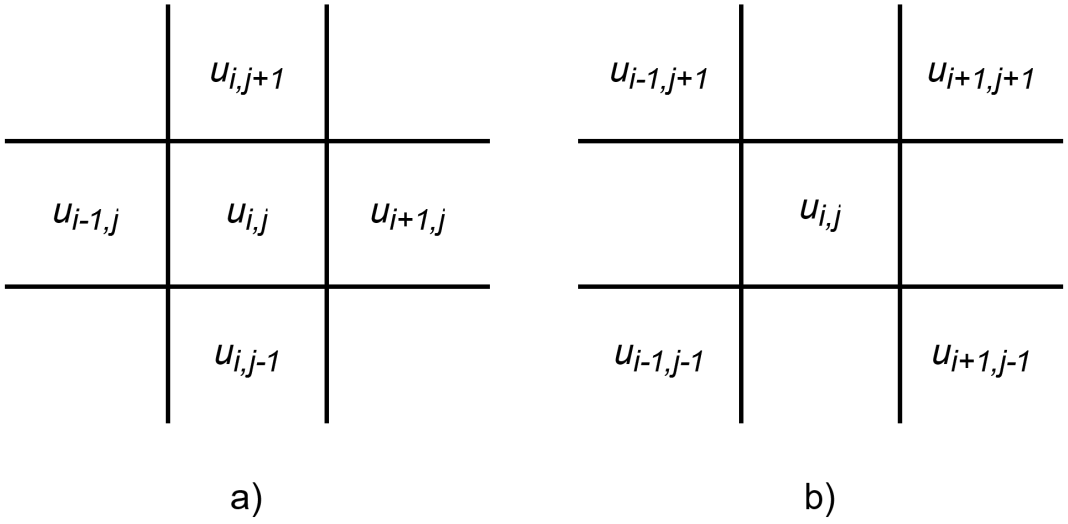


Figure 4.1: Approximations to Laplace's equation.

Instead of formula given in Eq. (4.19), we may also use the formula

$$u_{i,j} = \frac{1}{4}(u_{i-1,j-1} + u_{i+1,j-1} + u_{i+1,j+1} + u_{i-1,j+1}) \tag{4.22}$$

which uses the function values at the diagonal points [see Fig. 4.1(b)], and is therefore called the *diagonal five-point formula*. This is perfectly valid since it is well-known that the Laplace equation remains invariant when the coordinate axes are rotated through 45°. Expanding the terms on the right side of Eq. (4.22) by Taylor's series, it can be shown that

$$u_{i-1,j-1} + u_{i+1,j-1} + u_{i+1,j+1} + u_{i-1,j+1} - 4u_{i,j} = \frac{2}{3}h^4 u_{xxyy} + O(h^6) \tag{4.23}$$

Neglecting terms of the order h^6, it follows from Eqs. (4.21) and (4.23) that the error in the diagonal formula is four times that in the standard formula. Hence, in all computations we should prefer to use the standard five-point formula, whenever possible.

Eliminating the term containing h^4 from both Eqs. (4.21) and (4.23), we obtain the *nine-point formula*

$$\begin{aligned} u_{i-1,j-1} + u_{i+1,j-1} + u_{i+1,j+1} + u_{i-1,j+1} + \\ + 4(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}) - 20u_{i,j} = 0 \end{aligned} \tag{4.24}$$

It is clean that the error in this formula is of order h^6. In a similar manner, the finite-difference analogues of Eqs. (4.4) and (4.5) can be obtained.

In this chapter we consider those partial differential equations which can be replaced by the finite-difference analogues, or *difference equations,* are then used as approximations to the concerned partial differential equations. Our general procedure is, therefore, to replace the partial differential equation by a finite-difference analogue and then obtain the solution at the mesh points.

4.3 Heat Equation in One Dimension

The heat equation in one dimension is a typical parabolic partial differential equation and is a time variable problem. If we consider a long thin insulated rod and equate the amount of heat absorbed to the difference between the amount of heat entering a small element and that leaving the element in time Δt, we obtain the partial differential equation

$$\frac{\partial u}{\partial t} = \alpha^2 \frac{\partial^2 u}{\partial x^2} \tag{4.25}$$

where

$$\alpha^2 = \frac{k}{s\rho} \tag{4.26}$$

In Eq. (4.26), k is the coefficient of conductivity of the material, ρ is its density and s is its specific heat. Analytical solutions of Eq. (4.25), obtained by the method of separation of variables are given by

$$\left.\begin{aligned} u(x,t) &= e^{-p^2\alpha^2 t}(c_1\ cos\ px + c_2\ sin\ px) \\ u(x,t) &= e^{p^2\alpha^2 t}\ (c_2 e^{px} + c_3 e^{-px}) \end{aligned}\right\} \tag{4.27}$$

From Eq. (4.27), the appropriate form of solution should be chosen depending upon the boundary conditions given. It is clear that to solve Eq. (4.25), we need one initial condition and two boundary conditions. In the sequel, we shall discuss the *finite difference approximation* to this equation.

4.3.1 Finite-difference approximation

We divide the (x,t) plane into smaller rectangles by means of the sets lines

$$\begin{aligned} x = ih, \quad i &= 0,\ 1,\ 2,\ \ldots \\ t = kl, \quad j &= 0,\ 1,\ 2,\ \ldots \end{aligned}$$

where $h = \Delta x$ and $l = \Delta t$. Denoting $u(ih,\ kl) = u_i^k$, we have

$$\frac{\partial u}{\partial t} \approx \frac{u_i^{k+1} - u_i^k}{l} \tag{4.28}$$

and

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{1}{h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) \tag{4.29}$$

Equation (4.25) is replaced by the finite difference analogue

$$\frac{u_i^{k+1} - u_i^k}{l} = \alpha^2 \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2},$$

which simplifies to

$$u_i^{k+1} = \lambda u_{i-1}^k + u_{i+1}^k + (1 - 2\lambda)u_i^k, \tag{4.30}$$

where

$$\lambda = \frac{\alpha^2 l}{h^2}. \tag{4.31}$$

In Eq. (4.30), u_i^{k+1} is expressed *explicitly* in terms of u_{i-1}^k, u_{i+1}^k and u_i^k. Hence it is called the *explicit* formula for the solution of one-dimensional heat equation. It can be shown that Eq. (4.30) is valid only for $0 \leq \lambda \leq \frac{1}{2}$, which is called the *stability condition* for the explicit formula.

If we set $\lambda = \frac{1}{2}$ in Eq. (4.30), we obtain the simple formula

$$u_i^{k+1} = \frac{1}{2}(u_{i-1}^k + u_{i+1}^k), \tag{4.32}$$

which is called *Bender-Schmidt recurrence formula.* It is clear that Eqs. (4.30) and (4.32) have limited application because of restriction on the values of λ. A formula which does not have any restrictions on λ is that due to Crank and Nicolson. In Eq. (4.25), if we replace $\frac{\partial^2 u}{\partial x^2}$ by the average of its finite difference approximations on the kth and $(k+1)$th time levels, we obtain

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{2h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k + u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)$$

Hence, Eq. (4.25) is approximated by

$$\frac{u_i^{k+1} - u_i^k}{l} = \frac{\alpha^2}{2h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k + u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)$$

which simplifies to

$$-\lambda u_{i-1}^{k+1} + (2 + 2\lambda)u_i^{k+1} - \lambda u_{i+1}^{k+1} = \lambda u_{i-1}^k + (2 - 2\lambda)u_i^k + \lambda u_{i+1}^k \tag{4.33}$$

On the left side of Eq. (4.33) we have three unknowns and on the right side, all are known quantities. This is called *Crank-Nicolson formula* for the one-dimensional heat equation and it is an *implicit formula*.

It is convergent *for all finite values of* λ. If there are N internal mesh points on each time row, then Eq. (4.33) gives N simultaneous equations for the N unknowns. In a similar way, values of u on all time rows can be calculated.

Example 4.1. Use the Bender-Schmidt formula to solve the heat conduction problem

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2}$$

with the conditions $u(x,0) = 4x - x^2$ and $u(0,t) = u(4,t) = 0$.

Setting $h = 1$, we see that $l = 1$ when $\lambda = \dfrac{1}{2}$.
Now, the initial values are

$$\begin{gathered} u(0,0) = 0,\ u(1,0) = 3, \\ u(2,0) = 4,\ u(3,0) = 3 \\ \text{and} \qquad u(4,0) = 0. \end{gathered}$$

Further, $u(0,t) = u(4,t) = 0$.

For $l = 1$, Bender-Schmidt formula gives

$$\begin{aligned} u_1^1 &= \frac{1}{2}(0+4) = 2, \\ u_2^1 &= \frac{1}{2}(3+3) = 3, \\ u_3^1 &= \frac{1}{2}(4+0) = 2. \end{aligned}$$

Similarly, for $l = 2$, we obtain

$$\begin{aligned} u_1^2 &= \frac{1}{2}(0+3) = 1.5, \\ u_2^2 &= \frac{1}{2}(2+2) = 2, \\ u_3^2 &= \frac{1}{2}(3+0) = 1.5. \end{aligned}$$

Continuing in this way, we obtain

$$\begin{aligned} &u_1^3 = 1, && u_2^3 = 1.5, && u_3^3 = 1, \\ &u_1^4 = 0.75, && u_2^4 = 1, && u_3^4 = 0.75, \\ &u_1^5 = 0.5, && u_2^5 = 0.75, && u_3^5 = 0.5, \quad \text{and so on.} \end{aligned} \tag{4.34}$$

Example 4.2. Solve the heat conduction problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the conditions $u(x,0) = \sin \pi x,\ 0 \le x \le 1$, and $u(0,t) = u(1,t) = 0$. Use Bender-Schmidt's and Crank-Nicolson formulae to compute the value of $u(0.6,\ 0.04)$ and compare the results with the exact value.

The exact solution of this problem is given by

$$u(x,t) = e^{-\pi^2 t} \sin \pi x,$$

so that the exact value of $u(0.6,\ 0.04)$ is 0.6408.

(a) *Bender-Schmidt formula*

Let $h = 0.2$. Then $l = \lambda h^2 = \frac{1}{2}(0.04) = 0.02$.
The initial values of u are

$$u_0^0 = 0, \qquad u_1^0 = 0.5878, \qquad u_2^0 = 0.9510,$$

$$u_3^0 = 0.9510, \qquad u_4^0 = 0.5878, \qquad u_5^0 = 0.$$

Then Bender-Schmidt formula gives

$$u_1^1 = \frac{1}{2}(0.9510) = 0.4755, \qquad u_2^1 = \frac{1}{2}(0.5878 + 0.9510) = 0.7694,$$

$$u_3^1 = 0.7694, \qquad u_4^1 = 0.4755.$$

Also,

$$u_1^2 = \frac{1}{2}(0.7694) = 0.3847, \qquad u_2^2 = 0.62245,$$

$$u_3^2 = 0.62245, \qquad u_4^2 = 0.3847.$$

Therefore, $u(0.6,\ 0.04) = u_3^2 = 0.6224$, the error in which is 0.0184.

(b) *Crank-Nicolson formula*

Let $h = 0.2$ and $l = 0.04$, so that $\lambda = 1$.
For $\lambda = 1$, Crank-Nicolson formula becomes

$$-u_{i-1}^{k+1} + 4u_i^{k+1} - u_{i+1}^{k+1} = u_{i-1}^k + u_{i+1}^k \qquad \text{(i)}$$

Putting $k = 0$ in (i), we obtain

$$-u_{i-1}^1 + 4u_i^1 - u_{i+1}^1 = u_{i-1}^0 + u_{i+1}^0$$

Corresponding to $i = 1$, 2, 3, and 4, we obtain the four equations

$$\begin{aligned} 4u_1^1 - u_2^1 &= 0.9510 \\ -u_1^1 + 4u_2^1 - u_3^1 &= 1.5388 \\ -u_2^1 + 4u_3^1 - u_4^1 &= 1.5388 \\ -u_3^1 + 4u_4^1 &= 0.9510 \end{aligned}$$

By symmetry, we have

$$u_1^1 = u_4^1 \quad \text{and} \quad u_2^1 = u_3^1$$

Solving the above system, we obtain

$$u_2^1 = u_3^1 = 0.6460$$

Hence, $u(0.6, 0.04) \approx 0.6460$, the error in which is 0.0052.

Example 4.3. Solve the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the conditions

$$u(x, 0) = 0, \quad u(0, t) = 0 \quad \text{and} \quad u(1, t) = t.$$

Using Crank-Nicolson scheme, find the value of $u(\frac{1}{2}, \frac{1}{8})$ taking successively (i) $h = \frac{1}{2}, l = \frac{1}{8}$, (ii) $h = \frac{1}{4}, l = \frac{1}{8}$. Compare the results obtained with the exact value of $u(\frac{1}{2}, \frac{1}{8}) = 0.01878$.

(i) $h = \frac{1}{2}, l = \frac{1}{8}$. Then $\lambda = \frac{1}{2}$.

Crank-Nicolson scheme gives

$$-u_{i-1}^{k+1} + 6u_i^{k+1} - u_{i+1}^{k+1} = u_{i-1}^k + 2u_i^k + u_{i+1}^k.$$

Setting $k = 0$ and $i = 1$ in the above equation, we obtain

$$-u_0^1 + 6u_1^1 - u_2^1 = u_0^0 + 2u_1^0 + u_2^0.$$

$$u_1^1 = \frac{1}{48}, \quad \text{since } u_0^1 = 0 \text{ and } u_2^1 = \frac{1}{8}.$$

$$= 0.02083 \text{ (error} = 0.00205).$$

(ii) $h = \frac{1}{4}, l = \frac{1}{8}$. Then $\lambda = 2$.

Therefore, Crank-Nicolson scheme gives

$$-u_{i-1}^{k+1} + 3u_i^{k+1} - u_{i+1}^{k+1} = u_{i-1}^k - u_i^k + u_{i+1}^k.$$

With $k = 0$, we obtain

$$-u_{i-1}^1 + 3u_i^1 - u_{i+1}^1 = u_{i-1}^0 - u_i^0 + u_{i+1}^0,\ i = 1, 2, 3.$$

Corresponding to $i = 1, 2$ and 3, we obtain three equations

$$\begin{aligned} -3u_1^1 - u_2^1 &= 0 \\ u_1^1 - 3u_2^1 + u_3^1 &= 0 \\ u_2^1 - 3u_3^1 &= -\frac{1}{8}. \end{aligned}$$

Solving the above system, we obtain

$$u_2^1 = u\left(\frac{1}{2}, \frac{1}{8}\right) \approx \frac{1}{56} = 0.01786 \text{ (error} = 0.00092).$$

4.4 Wave equation

The wave equation is defined by the boundary value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial t^2} \tag{4.35}$$

with the boundary conditions

$$\left.\begin{aligned} u(x, 0) &= f(x) \\ u_t(x, 0) &= \phi(x) \\ u(0, t) &= \psi_1(t) \\ u(1, t) &= \psi_2(t) \end{aligned}\right\} \tag{4.36}$$

for $0 \le t \le T$. This equation is of hyperbolic type and models the transverse vibrations of a stretched string. As earlier, we use the following difference approximations for the derivatives

$$u_{xx} = \frac{1}{h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) + O(h^2) \tag{4.37}$$

and

$$u_{tt} = \frac{1}{l^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) + O(l^2) \tag{4.38}$$

where $x = ih, t = kl$, and $u(x, t) = u(ih, kl) = u_i^k$. Further, $u_t(x, t)$ is approximated by

$$u_t(x, t) = \frac{u_i^{k+1} - u_i^{k-1}}{2l} + O(l^2) \tag{4.39}$$

Substituting from Eqs. (4.36) and (4.37) in Eq. (4.35), we obtain

$$\frac{1}{l^2}\left(u_i^{k-1} - 2u_i^k + u_i^{k+1}\right) = \frac{c^2}{h^2}\left(u_{i-1}^k - u_i^k + u_{i+1}^k\right).$$

Setting $\alpha = \dfrac{cl}{h}$ in the above and rearranging the terms, we have

$$u_i^{k+1} = -u_i^{k-1} + \alpha^2(u_{i-1}^k + u_{i+1}^k) + 2(1-\alpha^2)u_i^k. \tag{4.40}$$

Equation (4.40) shows that the function values at the kth and $(k-1)$th time levels are required to determine those at the $(k+1)$th time level. Such difference schemes are called *three level difference schemes* compared to the two level schemes derived in the parabolic case.[1] Formula (4.40) holds good if $\alpha < 1$, which is the condition for stability.

There exist implicit finite difference schemes for the equation given by Eq. (4.35). Two such schemes are

$$\frac{u_i^{k-1} - 2u_i^k + u_i^{k+1}}{l^2} = \frac{c^2}{2h^2}\left(u_{i-1}^{k-1} - 2u_i^{k-1} + u_{i+1}^{k-1} + u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right) \tag{4.41}$$

and

$$\frac{u_i^{k-1} - 2u_i^k + u_i^{k+1}}{l^2} = \frac{c^2}{4h^2}\left[\left(u_{i-1}^{k-1} - 2u_i^{k-1} + u_{i+1}^{k-1}\right) + 2\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) + \left(u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)\right] \tag{4.42}$$

Equations (4.41) and (4.42) hold good for all values of $\dfrac{cl}{h}$. The use of formula given in Eq. (4.40) is demonstrated in the following examples.

Example 4.4. Solve the equation $u_{tt} = u_{xx}$ subject to the following conditions

$$u(0,t) = 0, \quad u(1,t) = 0, \quad t > 0$$

and

$$\frac{\partial u}{\partial t}(x,0) = 0, \; u(x,0) = \sin^3(\pi x), \quad 0 \le x \le 1.$$

This problem has an exact solution given by

$$u(x,t) = \frac{3}{4}\sin \pi x \cos \pi t - \frac{1}{4}\sin 3\pi x \cos 3\pi t.$$

Let $h = 0.25$ and $l = 0.2$. Then $\alpha = 0.8 < 1$. The given conditions are

$$u_0^k = 0, u_4^k = 0, u_i^0 = \sin^3(\pi i h), \quad i = 1, 2, 3, 4.$$

[1] A three level scheme for solving parabolic equations in one dimension may be found in Sastry [7]

Also,

$$\frac{\partial u(x,0)}{\partial t} = 0 \Rightarrow u_i^1 - u_i^{-1} = 0$$
$$\Rightarrow u_i^{-1} = u_i^1.$$

With $\alpha = 0.8$, the explicit formula becomes

$$u_i^{k+1} = -u_i^{k-1} + 0.64\left(u_{i-1}^k + u_{i+1}^k\right) + 2(0.36)u_i^k.$$

Setting $k = 0$, the above gives

$$u_i^1 = -u_i^{-1} + 0.64\left(u_{i-1}^0 + u_{i+1}^0\right) + 0.72u_i^0.$$
$$\Rightarrow u_i^1 = 0.32\left(u_{i-1}^0 + u_{i+1}^0\right) + 0.36u_i^0, \quad \text{since } u_i^{-1} = u_i^1.$$

Therefore,

$$\begin{aligned} u_1^1 &= 0.32\left(u_0^0 + u_2^0\right) + 0.36u_1^0 \\ &= 0.32(0+1) + 0.36(0.3537) \\ &= 0.4473 \text{ (error} = 0.0365) \end{aligned}$$

Similarly

$$u_2^1 = 0.5867 \text{ (error} = 0.0571)$$

and

$$u_3^1 = 0.4473 \text{ (error} = 0.0365).$$

The computations can be continued for $k = 1, 2 \ldots$

Example 4.5. Solve the boundary value problem defined by $u_{tt} = 4u_{xx}$ subject to conditions.

$$u(0,t) = 0 = u(4,t),\ u_t(x,0) = 0,\ u(x,0) = 4x - x^2.$$

Let

$$h = 1 \text{ and } \alpha = 1 \text{ so that } l = 0.5.$$

We have

$$u_0^k = u_4^k = 0 \text{ for all } k.$$

Since

$$u_t(x,0) = 0, \text{ we obtain } u_i^{-1} = u_i^1.$$

Further,

$$u(x,0) = 4x - x^2$$
$$\Rightarrow u_i^0 = 4i - i^2, \text{ since } h = 1.$$

Then,

$$u_0^0 = 0,\ u_1^0 = 3,\ u_2^0 = 4,\ u_3^0 = 3 \text{ and } u_4^0 = 0.$$

For $\alpha = 1$, the explicit scheme becomes

$$u_i^{k+1} = -u_i^{k-1} + u_{i-1}^k + u_{i+1}^k \tag{i}$$

Now, for $k = 0$, Eq. (i) gives

$$u_i^1 = -u_i^{-1} + u_{i-1}^0 + u_{i+1}^0$$
$$\Rightarrow u_i^1 = \frac{1}{2}\left(u_{i-1}^0 + u_{i+1}^0\right), \text{ since } u_i^{-1} = u_i^1.$$

Hence

$$u_1^1 = \frac{1}{2}\left(u_0^0 + u_2^0\right) = 2, \quad u_2^1 = \frac{1}{2}(3+3) = 3, \quad u_3^1 = \frac{1}{2}(4+0) = 2.$$

Similarly, we obtain

$$u_1^2 = 0, \quad u_2^2 = 0, \quad u_3^2 = 0,$$

and the succeeding time rows can be built up.

5 Partial differential equations of elliptic type. Poisson's equation.

5.1 Solution of Laplace's equation

We wish to solve Laplace's equation

$$u_{xx} + u_{yy} = 0 \tag{5.1}$$

in bounded region R with boundary C. As in Dirichlet's problem, let the value of u be specified everywhere on C. For simplicity, let R be a square region so that it can be divided into a network of a small squares of side h. Let the values of $u(x, y)$ on the boundary C be given by c_i and let the interior mesh points and the boundary points be as in Fig 5.1.

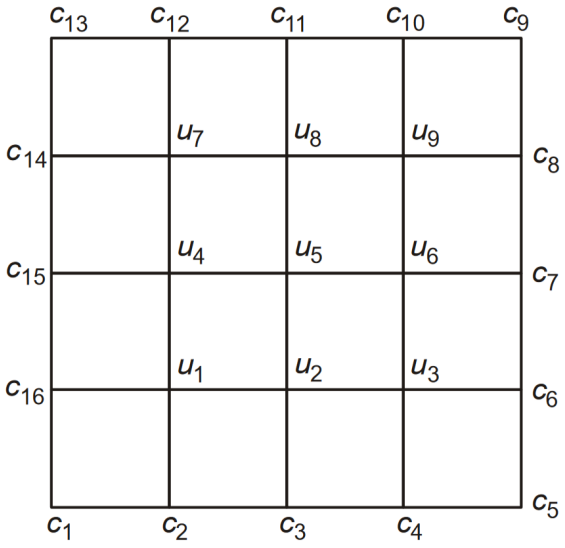


Figure 5.1: Interior mesh points and boundary points.

Then, as shown in the previous section, Eq. (5.1) can be replaced by either the standard five-point formula, viz. Eq. (4.20); or the diagonal five-point formula given in Eq. (4.22). The approximate funciton values at the interior mesh points can now be computed according to the scheme: we first use the diagonal five-point formula Eq. (4.22) and compute $u_5,\ u_7,\ u_9,\ u_1$ and u_3 in this order. Thus, we obtain

$$u_5 = \frac{1}{4}(c_1 + c_5 + c_9 + c_{13}); \qquad u_7 = \frac{1}{4}(c_{15} + u_5 + c_{11} + c_{13});$$
$$u_9 = \frac{1}{4}(u_5 + c_7 + c_9 + c_{11}); \qquad u_1 = \frac{1}{4}(c_1 + c_3 + u_5 + c_{15});$$
$$u_3 = \frac{1}{4}(c_3 + c_5 + c_7 + u_5).$$

We then compute, in the order, the remaining quantities, viz., $u_8,\ u_4,\ u_6$ and u_2 by the *standard five-point formula* (4.20). Thus, we have

$$u_8 = \frac{1}{4}(u_5 + u_9 + c_{11} + u_7); \qquad u_4 = \frac{1}{4}(u_1 + u_5 + u_7 + c_{15});$$
$$u_6 = \frac{1}{4}(u_3 + c_7 + u_9 + u_5); \qquad u_2 = \frac{1}{4}(c_3 + u_3 + u_5 + u_1).$$

When once all the $u_i,\ (i = 1, 2, 3, \ldots, 9)$ are computed, their accuracy can be improved by any of the iterative methods described below.

5.1.1 Jacobi's Method

Let $u_{i,j}^{(n)}$ denotes the nth iterative value of $u_{i,j}$. An iterative procedure to solve Eq. (4.20) is

$$u_{i,j}^{(n+1)} = \frac{1}{4}[u_{i-1,j}^{(n)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n)} + u_{i,j+1}^{(n)}] \tag{5.2}$$

for the interior mesh points. This is called the *point Jacobi method.*

5.1.2 Gauss-Seidel Method

The method uses the latest iterative values available and scans the mesh points systematically from left to right along successive rows. The iterative formula is:

$$u_{i,j}^{(n+1)} = \frac{1}{4}[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)}] \tag{5.3}$$

It can be shown that the Gauss-Seidel scheme converges twice fast as the Jacobi scheme. This method is also referred to as *Liebmann's method.*

5.1.3 Successive Over Relaxation (SOR) Method

Equation (5.3) can be written as

$$\begin{aligned} u_{i,j}^{(n+1)} &= u_{i,j}^{(n)} + \frac{1}{4}\left[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)} - 4u_{i,j}^{(n)}\right] \\ &= u_{i,j}^{(n)} + \frac{1}{4}R_{i,j} \end{aligned}$$

which shows that $(1/4)R_{i,j}$ is the change in the value of $u_{i,j}$ for one Gauss-Seidel iteration. In the SOR method, a larger change than this is given to $u_{i,j}^{(n)}$, and the iteration formula is written as

$$u_{i,j}^{(n+1)} = u_{i,j}^{(n)} + \frac{1}{4}\omega R_{i,j}$$
$$= \frac{1}{4}\omega\left[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)}\right] + (1-\omega)u_{i,j}^{(n)}$$
$$= \omega u_{i,j}^{(n+1)} + (1-\omega)u_{i,j}^{(n)} \tag{5.4}$$

The rate of convergence of Eq. (5.4) depends on the choice of ω, which is called the *accelerating factor* and lies between 1 and 2.

The percentage error in the value u_{ij} is given by

$$|\varepsilon_{ij}| = \left|\frac{u_{i,j}^{(n+1)} - u_{i,j}^{(n)}}{u_{ij}^{(n+1)}}\right| \times 100\% \tag{5.5}$$

It was shown by B.A. Carré that for $\omega = 1.875$, the rate of convergence of Eq. (5.4) is twice as fast as that when $\omega = 1$, and for $\omega = 1.9$, the rate of convergence is 40 times greater than that when $\omega = 1$. In general, however, it is difficult to estimate the best value of ω. The following examples illustrate the methods of solution.

Example 5.1. Solve Laplace's equation for the square region shown in Fig. 5.2, the boundary values being as indicated.

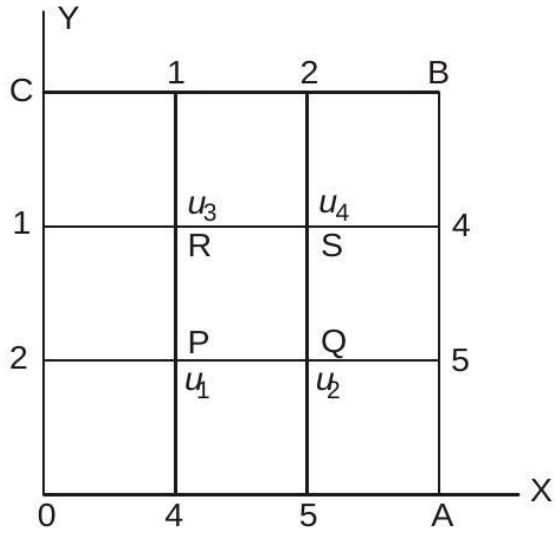


Figure 5.2

It is seen from the figure that the boundary values are symmetric about the diagonal AC. Hence, $u_1 = u_4$ and we need find only u_1, u_2 and u_3. The standard five-point formula applied at the point P gives

$$u_2 + u_3 + 2 + 4 - 4u_1 = 0.$$

Hence we have

$$u_1 = \frac{1}{4}\left(u_2 + u_3 + 6\right).$$

The iteration formula is therefore

$$u_1^{(n+1)} = \frac{1}{4}\left[u_2^{(n)} + u_3^{(n)} + 6\right].$$

Similarly, the iteration formulae at the points Q and R are given by

$$u_2^{(n+1)} = \frac{1}{2}u_1^{(n+1)} + \frac{5}{2},$$

and

$$u_3^{(n+1)} = \frac{1}{2}u_1^{(n+1)} + \frac{1}{2}.$$

For the first iteration, let $u_2 = 5$ (since it is nearer to the value $u = 5$), and $u_3^{(0)} = 1$. Hence

$$u_1^{(1)} = \frac{1}{4}(5 + 1 + 6) = 3,$$
$$u_2^{(1)} = \frac{1}{2}(3) + \frac{5}{2} = 4,$$
$$u_3^{(1)} = \frac{1}{2}(3) + \frac{1}{2} = 2.$$

For the second iteration, we have

$$u_1^{(2)} = \frac{1}{4}(4 + 2 + 6) = 3,$$
$$u_2^{(2)} = \frac{1}{2}(3) + \frac{5}{2} = 4,$$

and

$$u_3^{(2)} = \frac{1}{2}(3) + \frac{1}{2} = 2.$$

Since the values are unchanged, we conclude that $u_1 = 3, u_2 = 4, u_3 = 2$ and $u_4 = 3$.

Example 5.2. Solve the equation $u_{xx} + u_{yy} = 0$ in the domain of Fig. 5.3, below by (a) Jacobi's method, (b) Gauss-Seidel's method, and (c) SOR method.

(a) To start *Jacobi's iteration process*, we obtain the approximate values of u_1, u_2, u_3 and u_4 as follows:

$$u_1^{(1)} = \frac{1}{4}(0 + 0 + 0 + 1) = 0.25;$$
$$u_2^{(1)} = \frac{1}{4}(0 + 0 + 0 + 1) = 0.25;$$
$$u_3^{(1)} = \frac{1}{4}(1 + 1 + 0 + 0) = 0.5;$$
$$u_4^{(1)} = \frac{1}{4}(1 + 1 + 0 + 0) = 0.5.$$

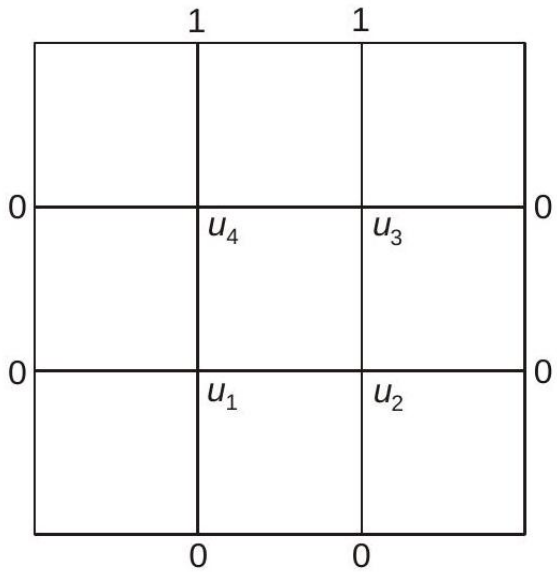


Figure 5.3

The iterations have been continued using Eq. (5.2), and seven successive iterates are given below:

u_1	u_2	u_3	u_4
0.1875	0.1875	0.4375	0.4375
0.15625	0.15625	0.40625	0.40625
0.14062	0.14062	0.39062	0.39062
0.13281	0.13281	0.38281	0.38281
0.12891	0.12891	0.37891	0.37891
0.12695	0.12695	0.37695	0.37695
0.12598	0.12598	0.37598	0.37598

(b) *Gauss-Seidel method:* Five successive iterates are given below:

u_1	u_2	u_3	u_4
0.25	0.3125	0.5625	0.46875
0.21875	0.17187	0.42187	0.39844
0.14844	0.13672	0.38672	0.38086
0.13086	0.12793	0.37793	0.37646
0.12646	0.12573	0.37573	0.37537

(c) *SOR method:* With $\omega = 1.1$, three successive iterates obtained by using Eq. (5.4) are given below.

u_1	u_2	u_3	u_4
0.275	0.35062	0.35062	0.35062
0.16534	0.10683	0.38183	0.37432
0.11785	0.12181	0.37216	0.37341

Example 5.3. Solve Laplace's equation for Fig. 5.4 given below:

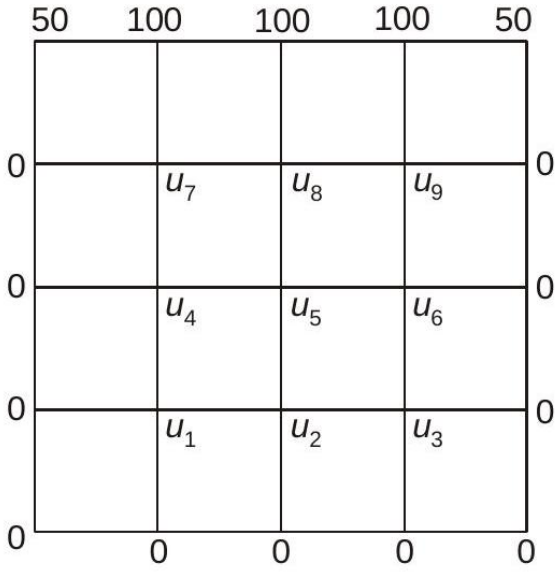


Figure 5.4

We first compute the quantities u_5, u_7, u_9, u_1 and u_3 by using the diagonal five-point formula given in Eq. (4.22). Thus, we obtain

$$u_5^{(1)} = 25.00; \quad u_7^{(1)} = 42.75; \qquad u_9^{(1)} = 43.75;$$
$$u_1^{(1)} = 6.25; \quad u_3^{(1)} = 6.25.$$

We now compute u_8, u_4, u_6 and u_2 successively by using the standard fivepoint formula given in Eq. (4.20)

$$u_8^{(1)} = 53.12; \quad u_4^{(1)} = 18.75;$$
$$u_6^{(1)} = 18.75; \quad u_2^{(1)} = 9.38.$$

We have thus obtained the first approximations of all the nine mesh points and we can now use one of the iterative formulae given in Section 5. We give below the first-four iterates obtained by using the Gauss-Seidel formula:

u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
7.03	9.57	7.08	18.94	25.10	18.98	43.02	52.97	42.99
7.13	9.83	7.20	18.81	25.15	18.84	42.94	52.77	42.90
7.16	9.88	7.18	18.81	25.08	18.79	42.89	52.72	42.88
7.17	9.86	7.16	18.78	25.04	18.77	42.88	52.70	42.87

Example 5.4. Solve the Poisson equation

$$u_{xx} + u_{yy} = -10\left(x^2 + y^2 + 10\right).$$

in the domain of Fig. 5.5. Let the values of u at the four grid points, $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ be

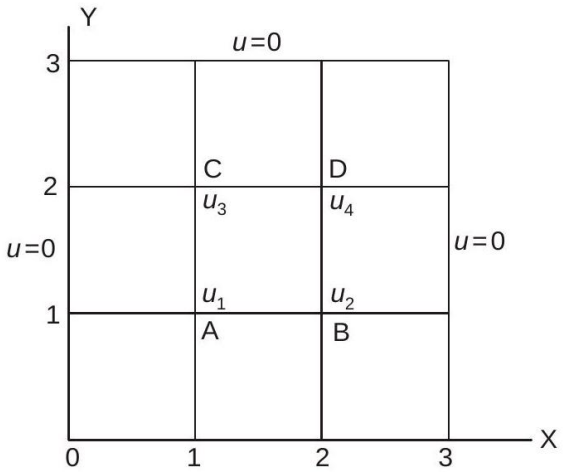


Figure 5.5

u_1, u_2, u_3, u_4, respectively. Let the grid points be defined by $x = ih, y = jh$, where $h = 1$, $i, j = 0, 1, 2, 3$. At the point $\mathrm{A}, i = 1, j = 1$. The standard five-point formula applied at the point A gives

$$u_2 + u_3 + 0 + 0 - 4u_1 = -10(1 + 1 + 10)$$

i.e.,

$$u_1 = \frac{1}{4}\left(u_2 + u_3 + 120\right) \tag{i}$$

Again, the standard five-point formula applied at the point B gives

$$u_1 + u_4 + 0 + 0 - 4u_2 = -10(4 + 1 + 10)$$

i.e.,

$$u_2 = \frac{1}{4}(u_1 + u_4 + 150) \qquad \text{(ii)}$$

Similarly, the standard five-point formula applied at the points C and D gives, respectively

$$u_3 = \frac{1}{4}(u_1 + u_4 + 150) \qquad \text{(iii)}$$

and

$$u_4 = \frac{1}{4}(u_2 + u_3 + 180) \qquad \text{(iv)}$$

From (ii) and (iii), it is seen that $u_2 = u_3$ and so we need to find only u_1, u_2 and u_4 from (i), (ii) and (iv). The iteration formulae are therefore given by

$$\begin{aligned}
u_1^{(n+1)} &= \frac{1}{2}u_2^{(n)} + 30 \\
u_2^{(n+1)} &= \frac{1}{4}\left[u_1^{(n+1)} + u_4^{(n)} + 150\right] \\
u_4^{(n+1)} &= \frac{1}{2}u_2^{(n+1)} + 45
\end{aligned}$$

For the first iteration, we assume that $u_2^{(0)} = u_4^{(0)} = 0$. Hence we obtain

$$\begin{aligned}
u_1^{(1)} &= 30, \\
u_2^{(1)} &= \frac{1}{4}(30 + 0 + 150) = 45 \\
u_4^{(1)} &= \frac{1}{2}(45) + 45 = 67.5.
\end{aligned}$$

For the second iteration, we have

$$\begin{aligned}
u_1^{(2)} &= \frac{1}{2}u_2^{(1)} + 30 = \frac{1}{2}(45) + 30 = 52.5 \\
u_2^{(2)} &= \frac{1}{4}\left[u_1^{(2)} + u_4^{(1)} + 150\right] = \frac{1}{4}[52.5 + 67.5 + 150] = 67.5 \\
u_4^{(2)} &= \frac{1}{2}\left[u_2^{(2)}\right] + 45 = 78.75.
\end{aligned}$$

For the third iteration, we obtain

$$\begin{aligned}
u_1^{(3)} &= \frac{1}{2}u_2^{(2)} + 30 = \frac{1}{2}(67.5) + 30 = 63.75 \\
u_2^{(3)} &= \frac{1}{4}\left[u_1^{(3)} + u_4^{(2)} + 150\right] = \frac{1}{4}[63.75 + 78.75 + 150] = 73.125 \\
u_4^{(3)} &= \frac{1}{2}u_2^{(3)} + 45 = \frac{1}{2}(73.125) + 45 = 81.5625.
\end{aligned}$$

The fourth iteration gives

$$u_1^{(4)} = \frac{1}{2}u_2^{(3)} + 30 = \frac{1}{2}(73.125) + 30 = 66.5625$$
$$u_2^{(4)} = \frac{1}{4}\left[u_1^{(4)} + u_4^{(3)} + 150\right] = \frac{1}{4}[66.5625 + 81.5625 + 150] = 74.53125$$
$$u_4^{(4)} = \frac{1}{2}u_2^{(4)} + 45 = \frac{1}{2}(74.53125) + 45 = 82.2656.$$

For the fifth iteration, we obtain

$$u_1^{(5)} = \frac{1}{2}u_2^{(4)} + 30 = \frac{1}{2}(74.53125) + 30 = 67.2656$$
$$u_2^{(5)} = \frac{1}{4}\left[u_1^{(5)} + u_4^{(4)} + 150\right] = \frac{1}{4}[67.2656 + 82.2656 + 150] = 74.8828$$
$$u_4^{(5)} = \frac{1}{2}u_2^{(5)} + 45 = \frac{1}{2}(74.8828) + 45 = 82.4414.$$

The sixth iteration gives

$$u_1^{(6)} = \frac{1}{2}u_2^{(5)} + 30 = \frac{1}{2}(74.8828) + 30 = 67.4414$$
$$u_2^{(6)} = \frac{1}{4}\left[u_1^{(6)} + u_4^{(5)} + 150\right] = \frac{1}{4}[67.4414 + 82.4414 + 150] = 74.9707$$
$$u_4^{(6)} = \frac{1}{2}u_2^{(6)} + 45 = \frac{1}{2}(74.9707) + 45 = 82.4854.$$

From the last two iterates, we conclude that

$$u_1 = 67, \quad u_2 = u_3 = 75, \quad \text{and} \quad u_4 = 83.$$

5.1.4 ADI Method

This is an efficient method for the numerical solution of elliptic partial differential equations and was proposed by Peaceman and Rachford. It is quite general but, for easy understanding, we demonstrate its applicability with reference to the Laplace equation in two dimensions. For more details, the reader is referred to Isaacson and Keller [3].

We consider Laplace's equation in two dimensions, viz.,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{5.6}$$

and the standard five-point formula

$$u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j} = 0 \tag{4.20}$$

The use of formula given in Eq. (4.20) involves the solution of a system of algebraic equations, whose coefficient matrix, for $n = 6$, is of the form

$$A = \begin{bmatrix} -4 & 1 & 0 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 \\ 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & 0 & 1 & -4 \end{bmatrix} \tag{5.7}$$

The general form of such a system is given by

$$B = \begin{bmatrix} T & I & & & 0 & \\ I & T & I & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & & I & T & I \\ & & & & I & T \end{bmatrix}, \tag{5.8}$$

where T is a tridiagonal matrix of the form

$$T = \begin{bmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{bmatrix} \tag{5.9}$$

System A is called a *block tridiagonal* system and such systems are solved by Gaussian elimination or, in the case of large systems, by Gauss-Seidel iterations. But tridiagonal systems of the type of Eq. (5.9) are much easier to solve than block tridiagonal systems. Hence the question arises as to whether we can obtain directly tridiagonal systems in the numerical solution of Laplace's equation. Peaceman and Rachford showed that this is possible and their method of procedure, called the *alternating direction implicit* method (or the ADI method) is described below.

We rearrange Eq. (4.20) in either of two ways:

$$u_{i-1,j} - 4u_{i,j} + u_{i+1,j} = -u_{i,j-1} - u_{i,j+1} \tag{5.10}$$

or

$$u_{i,j-1} - 4u_{i,j} + u_{i,j+1} = -u_{i-1,j} - u_{i+1,j} \tag{5.11}$$

The ADI is an *iteration* method and Eqs. (5.10) and (5.11) are used as iteration formulae

$$u_{i-1,j}^{(r+1)} - 4u_{i,j}^{(r+1)} + u_{i+1,j}^{(r+1)} = -u_{i,j-1}^{(r)} - u_{i,j+1}^{(r)} \tag{5.12}$$

and

$$u_{ij-1}^{(r+2)} - 4u_{i,j}^{(r+2)} + u_{ij+1}^{(r+2)} = -u_{i-1,j}^{(r+1)} - u_{i+1,j}^{(r+1)} \tag{5.13}$$

Equation (5.12) is used to compute function values at all internal mesh points along rows and Eq. (5.13) those along columns. For $j = 1, 2, 3, \ldots,, n-1$, Eq. (5.12) yields a tridiagonal system of equations and can easily be solved. Similarly, for $i = 1, 2, 3, \ldots, n-1$, Eq. (5.13) also yields a tridiagonal system of equations.

In the ADI method, formulae (5.12) and (5.13) are used alternately. For example, for the first row $j = 1$, and Eq. (5.12) gives

$$u_{i-1,1}^{(r+1)} - 4u_{i,1}^{(r+1)} + u_{i+1,1}^{(r+1)} = -u_{i,0}^{(r)} - u_{i,2}^{(r)}, \quad (i = 1, 2, 3, \ldots, n-1) \tag{5.14}$$

Together with the boundary conditions, Eq. (5.14) represents a tridiagonal system of equations and are easily solved for $u_{i,1}^{(r+1)}$. We next put $j = 2$ and obtain the values of $u_{i,2}^{(r+1)}$ on the second row. The process is repeated for all the rows, viz. up to $j = n-1$. We next alternate the direction, i.e. we use Eq. (5.13) to compute $u_{i,j}^{(r+2)}$. It is easy to see that at every stage we will be solving a tridiagonal system of equations. Example 5.5 demonstrates the method of solution.

Example 5.5. Solve Laplace's equation, $u_{xx} + u_{yy} = 0$, in the domain of Fig. 5.6 (see Example 5.2).

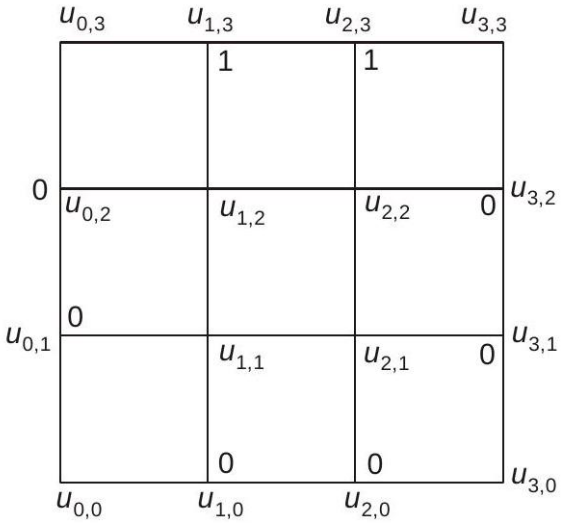


Figure 5.6

To apply formulae given in Eqs. (5.12) and (5.13), we relabel the internal mesh points, as in Fig. 5.6.

To start the iterations, we set $r = 0$. For the first row, $j = 1$. Then, Eq. (5.12) gives

$$u^{(1)}_{i-1,1} - 4u^{(1)}_{i,1} + u^{(1)}_{i+1,1} = -u^{(0)}_{i,0} - u^{(0)}_{i,2}. \tag{i}$$

With $i = 1$ and $i = 2$, this gives two equations

$$u^{(1)}_{0,1} - 4u^{(1)}_{1,1} + u^{(1)}_{2,1} = -u^{(0)}_{1,0} - u^{(0)}_{1,2}$$

and

$$u^{(1)}_{1,1} - 4u^{(1)}_{2,1} + u^{(1)}_{3,1} = -u^{(0)}_{2,0} - u^{(0)}_{2,2}.$$

Substituting the boundary values and assuming that $u^{(0)}_{1,2} = 1$ and $u^{(0)}_{2,2} = 1$, the above equations yield

$$u^{(1)}_{1,1} = u^{(1)}_{2,1} = \frac{1}{3} = 0.3333.$$

For computing the function values on the second row, we set $j = 2$ in (5.12) to obtain

$$u^{(1)}_{i-1,2} - 4u^{(1)}_{i,2} + u^{(1)}_{i+1,2} = -u^{(0)}_{i,1} - u^{(0)}_{i,3} \tag{ii}$$

With $i = 1$ and $i = 2$, Eq. (ii) gives

$$u^{(1)}_{0,2} - 4u^{(1)}_{1,2} + u^{(1)}_{2,2} = -u^{(0)}_{1,1} - u^{(0)}_{1,3}$$

and

$$u^{(1)}_{1,2} - 4u^{(1)}_{2,2} + u^{(1)}_{3,2} = -u^{(0)}_{2,1} - u^{(0)}_{2,3}.$$

Substituting the boundary values and solving the above, we obtain

$$u^{(1)}_{1,2} = u^{(1)}_{2,2} = \frac{1}{3} = 0.3333.$$

Having completed the computations on the two rows, we now alternate the direction and compute the function values on the columns, starting with the first one. For this, we use Eq. (5.13) with $r = 0$. Setting $i = 1$, Eq. (5.13) becomes

$$u^{(2)}_{1,j-1} - 4u^{(2)}_{1,j} + u^{(2)}_{1,j+1} = -u^{(1)}_{0,j} - u^{(1)}_{2,j} \tag{iii}$$

Putting $j = 1$ and $j = 2$ in the above, we obtain the equations

$$u^{(2)}_{1,0} - 4u^{(2)}_{1,1} + u^{(2)}_{1,2} = -u^{(1)}_{0,1} - u^{(1)}_{2,1}$$

and

$$u^{(2)}_{1,1} - 4u^{(2)}_{1,2} + u^{(2)}_{1,3} = -u^{(1)}_{0,2} - u^{(1)}_{2,2}.$$

Substituting the boundary values and solving the above equations, we obtain

$$u_{1,1}^{(2)} = \frac{8}{45} = 0.1778 \quad \text{and} \quad u_{1,2}^{(2)} = \frac{17}{45} = 0.3778$$

To compute the values on the second column, we now set $i = 2$ in Eq. (5.13)

$$u_{2,j-1}^{(2)} - 4u_{2,j}^{(2)} + u_{2,j+1}^{(2)} = -u_{1,j}^{(1)} - u_{3,j}^{(1)} \tag{iv}$$

Putting $j = 1$ and $j = 2$ in the above, we obtain the equations

$$u_{2,0}^{(2)} - 4u_{2,1}^{(2)} + u_{2,2}^{(2)} = -u_{1,1}^{(1)} - u_{3,1}^{(1)}$$

and

$$u_{2,1}^{(2)} - 4u_{2,2}^{(2)} + u_{2,3}^{(2)} = -u_{1,2}^{(1)} - u_{3,2}^{(1)}$$

Substituting the boundary values in the above two equations and solving them, we obtain

$$u_{2,1}^{(2)} = 0.1778 \quad \text{and} \quad u_{2,2}^{(2)} = 0.3778$$

The iterations are continued to improve the function values obtained first on the rows, then on the columns, and so on. You can continue these computations for the next iteration.

6 The Finite Element Method

6.1 Introduction

In Chapters 3, 4 and 5 we discussed finite difference methods for the solution of boundary-value problems defined by ordinary and partial differential equations. We now describe another class of methods for the solution of such problems, known as the *finite element methods*. A full discussion of these methods is outside the scope of this course. We give here only a brief presentation so as to enable to know that such methods exist. The discussion includes an elementary formulation of the method with simple applications to ordinary differential equations.

The basic idea behind the finite element method is to replace a continuous function by means of piecewise polynomials. Such an approximation, called the *piecewise polynomial approximation*, will be discussed in Section 6.1.2 We already know about the importance of polynomial approximations in numerical analysis. These are used in the numerical solution of practical problems where the exact functions are difficult to obtain or cumbersome to use. The idea of piecewise polynomial approximation is also not new, since the cubic spline already discussed, belongs to this class of polynomials.

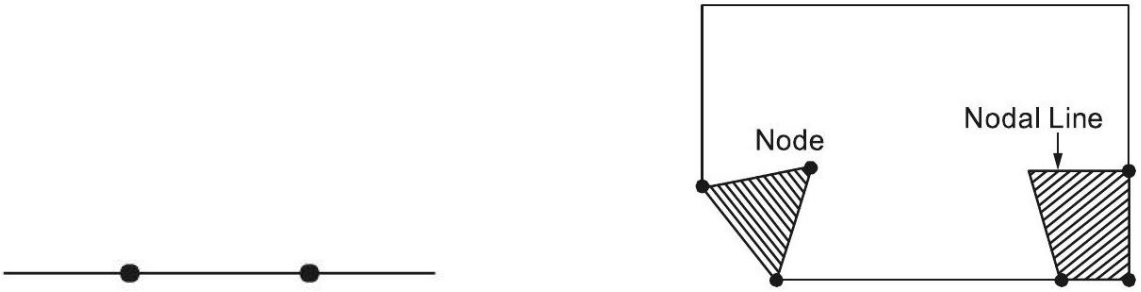
In engineering applications, several approximate methods of solution are used, e.g. the method of least squares, method of collocation, etc. In Section 6.2, we discuss two important methods of approximation, viz., the Rayleigh-Ritz method and the Galerkin technique. Rayleigh developed the method to solve certain vibration problems and Ritz provided a mathematical basis for it and also applied it to more general problems. Whereas the Rayleigh-Ritz method is based on the existence of a *functional* (see Section 6.1.1), the Galerkin technique uses the governing equations of the problem and minimizes the error of the approximate solution. The latter does not require a functional. A disadvantage of both these methods is that higher-order polynomials have to be used to obtain reasonable accuracy.

The finite element method, described now, is one of the most important numerical applications of the Rayleigh-Ritz and Galerkin methods. Its mathematical software is quite popular and used extensively in the solution of many practical problems of engineering and applied science. In the finite element method, the domain of integration is subdivided into a number of smaller regions called *elements* and over each of these elements the continuous function is approximated by a suitable piecewise polynomial. To obtain a better approximation one need not use higher-order polynomials but only use a finer subdivision, i.e. increase the number of elements

In practice, several types of elements are in use, the type used being largely dependent upon the geometrical shape of the region under consideration. In two-dimensional problems, the elements used are triangles, rectangles and quadrilaterals. For three-dimensional problems, tetrahedra, hexahedra and parallelopiped elements are used. Since our attempt in this chapter is only to introduce the finite element method, we restrict its application

to the solution of simple one-dimensional problems (see Section 6.4.1)

Examples of typical finite elements are shown in Fig. 6.1



(a) Line element

(b) Triangular and Quadrilateral elements

Figure 6.1: Typical finite elements.

6.1.1 Functionals

The concept of a functional is required to understand the Rayleigh-Ritz method, which will be discussed in the next section. This concept arises in the study of variational principles, which occur widely in physical and other problems. Mathematically, a variational principle consists in determining the extreme value of the integral of a typical function, say $f\left(x, y, y'\right)$. Here the integrand is a function of the coordinates and their derivatives and the integration is performed over a region. Consider, for example, the integral defined by

$$I(y) = \int_a^b f\left(x, y, y'\right) dx, \tag{6.1}$$

where $y(x)$ satisfies the boundary conditions $y(a) = y(b) = 0$.

The integrand f is integrated over the one-dimensional domain $x.I$ is said to be a functional and is defined as a function which transforms a function y into a real number, the value of the definite integral in Eq.(6.1). From calculus of variations we know that a necessary condition for $I(y)$ to have an extremum is that $y(x)$ must satisfy the Euler-Lagrange differential equation

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0. \tag{6.2}$$[1]

Similarly, for functionals of the form

$$I(y) = \int_a^b f\left(x, y, y', y''\right) dx \tag{6.3}$$

[1]For example, see Sastry [8].

the Euler-Lagrange equation takes the form

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial f}{\partial y''}\right) = 0. \tag{6.4}$$

The Euler-Lagrange equation (6.2) has several solutions and the one which satisfies the given boundary conditions is selected. Thus, one determines the functional so that it takes on an extremum value from a set of permissible functions. This is the central problem of a variational principle. An important point here is that an extremum may not exist. In other words, a variational principle may exist, but an extremum may not exist. Furthermore, not all differential equations have a variational principle. These difficulties are serious and therefore impose limitations on the application of the variational principle to the solution of engineering problems.

Many problems arising in physics and engineering are modelled by boundary-value problems and initial boundary-value problems. Frequently, these equations are equivalent to the problem of the minimization of a functional which can be interpreted in terms of the total energy of the given system. In any physical situation, therefore, the functional is obtained from a consideration of the total energy explicitly. Mathematically, however, it would be useful to be able to determine the functional from the governing differential equation itself. This is illustrated below with an example.

Example 6.1. Find the functional for the boundary-value problem defined by

$$\frac{d^2y}{dx^2} = f(x) \tag{i}$$

and

$$y(a) = y(b) = 0 \tag{ii}$$

We have

$$\begin{aligned}
\delta \int_a^b f y dx &= \int_a^b f \delta y dx \\
&= \int_a^b \frac{d^2y}{dx^2} \delta y dx, \text{ since } f(x) = \frac{d^2y}{dx^2} \\
&= \left[\frac{dy}{dx} \delta y \right]_a^b - \int_a^b \frac{dy}{dx} \frac{d}{dx} (\delta y) dx, \text{ on integrating by parts} \\
&= - \int_a^b \frac{dy}{dx} \frac{d}{dx} (\delta y) dx, \text{ since } \delta y(a) = \delta y(b) = 0 \\
&= - \int_a^b \frac{dy}{dx} \delta \left(\frac{dy}{dx} \right) dx, \text{ since } \frac{d}{dx}(\delta y) = \delta \left(\frac{dy}{dx} \right) \\
&= - \int_a^b \frac{1}{2} \delta \left(\frac{dy}{dx} \right)^2 dx \\
&= -\delta \int_a^b \frac{1}{2} \left(\frac{dy}{dx} \right)^2 dx
\end{aligned}$$

Hence

$$\delta \int_a^b \left[f y + \frac{1}{2} \left(\frac{dy}{dx} \right)^2 \right] dx = 0.$$

It follows that a unique solution of the problem (i) to (ii) exists at a minimum value of the integral defined by

$$I(v) = \int_a^b \left[f v + \frac{1}{2} \left(\frac{dv}{dx} \right)^2 \right] dx \tag{iii}$$

A quicker way of finding the functional of a boundary value problem is the following (See Reddy [6]).

Let $v(x)$ be a function satisfying the essential boundary conditions, viz. $v(a) = v(b) = 0$. Multiply the differential equation written in the form

$$-y'' + f(x) = 0$$

by v and integrate with respect to x. We then obtain

$$\begin{aligned}
0 &= - \int_a^b v y'' dx + \int_a^b v f dx \\
&= \left[-v y' \right]_a^b + \int_a^b v' y' dx + \int_a^b v f dx \\
&= \int_a^b \left(v' y' + v f \right) dx
\end{aligned}$$

Finally, substitute $y = v$ in the above and multiply the *bilinear* terms by 1/2. We then obtain the required functional

$$I(v) = \int_a^b \left[\frac{1}{2}v'^2 + vf\right] dx,$$

which is the same as Eq. (iii) obtained earlier.

By definition, therefore, the integral in (iii) represents the required functional of the problem. In a similar way, functionals of other boundary-value and initial boundary-value problems can be derived.

It is outside the scope of this course to deal extensively with the determination of functionals corresponding to boundary-value problems. We list below some familiar boundary-value problems with their associated functionals and these would be useful in understanding the problems discussed in this chapter.

(i)
$$\frac{d^2y}{dx^2} = f(x), y(a) = y(b) = 0 \tag{6.5}$$

$$I(v) = \int_a^b v\left(2f - v''\right) dx. \tag{6.6}$$

(ii)
$$\frac{d^2y}{dx^2} + ky = x^2, 0 < x < 1; y(0) = 0, \left(\frac{dy}{dx}\right)_{x=1} = 1 \tag{6.7}$$

$$I(v) = \frac{1}{2}\int_0^1 \left[\left(\frac{dv}{dx}\right)^2 - kv^2 + 2vx^2\right] dx - v(1). \tag{6.8}$$

(iii)
$$x^2y'' + 2xy' = f(x), y(a) = y(b) = 0 \tag{6.9}$$

$$I(v) = \int_a^b v\left[2f - \frac{d}{dx}\left(x^2y'\right)\right] dx. \tag{6.10}$$

(iv) $\nabla^2 u = 0, u = 0$ on the boundary C of R. (6.11)

$$I(v) = \iint_R \frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dxdy. \tag{6.12}$$

(v) $\nabla^2 u = -f, u = 0$ on the boundary C of R. (6.13)

$$I(v) = \iint_R \left\{\frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] - uf\right\} dxdy. \tag{6.14}$$

(vi)
$$\left.\begin{aligned} EI\frac{d^4y}{dx^4} + ky = f(x),\ 0 < x < l \\ y = 0 = \frac{d^2y}{dx^2} \text{ at } x = 0, l \end{aligned}\right\} \tag{6.15}$$

$$I(v) = \frac{1}{2}\int_0^l \left[EI\left(\frac{d^2v}{dx^2}\right)^2 + kv^2 - 2vf\right] dx. \tag{6.16}$$

6.1.2 Base Functions

Suppose we wish to approximate a real-valued function $f(x)$ over a finite interval $[a, b]$. A usual approach is to divide $[a, b]$ into a number of subintervals $[x_i, x_{i+1}]\,, i = 0, 1, 2, \ldots, n-1$, where $x_0 = a$ and $x_n = b$, and to interpolate linearly between the values of $f(x)$ at the end points of each subinterval. In $[x_i, x_{i+1}]$, the linear approximating function is given by

$$l_i(x) = \frac{1}{h_i}\left[(x_{i+1} - x)\, f_i + (x - x_i)\, f_{i+1}\right], \tag{6.17}$$

where $h_i = x_{i+1} - x_i$. From this, we construct the piecewise linear interpolating function over $[x_0, x_n]$ by the formula

$$P(x) = \sum_{i=0}^{n} \phi_i(x) f_i \tag{6.18}$$

where

$$\left.\begin{aligned} \phi_0(x) &= \begin{cases} (x_1 - x)\,/h_0, & x_0 \le x \le x_1, \\ 0, & x_1 \le x \le x_n, \end{cases} \\ \phi_i(x) &= \begin{cases} (x - x_{i-1})\,/h_{i-1}, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)\,/h_i, & x_i \le x \le x_{i+1}, \\ 0, & x \ge x_{i+1} \end{cases} \\ \phi_n(x) &= \begin{cases} 0, & x_0 \le x \le x_{n-1}, \\ (x - x_{n-1})\,/h_{n-1} & x_{n-1} \le x \le x_n. \end{cases} \end{aligned}\right\} \tag{6.19}$$

The functions $\phi_i(x), i = 1, 2, \ldots, n$ are called *base functions* or *shape functions*. It is easily seen that the base functions $\phi_i(x)$ are identically zero except for the range $[x_{i-1}, x_{i+1}]$ with $\phi_i\,(x_i) = 1$.

Other types of base functions such as piecewise Hermite polynomials, cubic splines, etc., are also used in the literature but these will not be considered in our course.

6.2 Methods of approximation

In this section we discuss two methods of approximation, viz. the *Rayleigh-Ritz* and *Galerkin* methods. As mentioned earlier, the former method is based on the existence of a functional which is then minimized. The second technique is due to Galerkin who proposed it as an error minimization method. It belongs to a wider class of methods called *weighted residual methods*. An advantage of the Galerkin method is that it works with the governing equations of the problem and does not require a functional.

Both the methods have a common feature in that they seek an approximate solution in the form of a linear combination of base functions. Nevertheless, they differ from each other in choosing the base functions.

6.2.1 Rayleigh-Ritz Method

In this method we do not obtain the actual minimum but only an approximate solution as nearer the actual solution as the base functions allow. To obtain a good approximation, therefore, the choice of the base functions is important and to improve the approximation, the number of base functions should be increased.

We explain this method by considering second-order boundary-value problem defined by

$$y'' + p(x)y + q(x) = 0, \quad y(a) = y(b) = 0. \tag{6.20}$$

The functional for the above problem is given by

$$I(v) = \int_a^b \left[\left(\frac{dv}{dx} \right)^2 - pv^2 - 2qv \right] dx = 0. \tag{6.21}$$

From the definition of the functional we know that if $y(x)$, the solution of Eq. (6.20), is substituted in Eq. (6.21), then the integral I will be minimum. Since we do not know the solution of Eq. (6.20), we try with an approximate solution and determine the parameters of the approximation so that the integral is minimum. This is the central idea of the Rayleigh-Ritz method. Now, let

$$v(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x) \tag{6.22}$$

be an approximate solution where the base functions, $\phi_i(x)$, are linearly independent and satisfy the boundary conditions given in Eq. (6.20), i.e. let

$$\phi_i(a) = 0 \quad \text{and} \quad \phi_i(b) = 0. \tag{6.23}$$

Substituting for v in Eq. (6.21), we obtain

$$I\left(\alpha_1, \alpha_2, \ldots \alpha_n\right) = \int_a^b \left\{ \left[\frac{d}{dx} \sum \alpha_i \phi_i(x) \right]^2 - p \left[\sum \alpha_i \phi_i(x) \right]^2 - 2q \sum \alpha_i \phi_i(x) \right\} dx = 0. \tag{6.24}$$

For minimum, we have

$$\frac{\partial I}{\partial \alpha_1}\delta\alpha_1 + \frac{\partial I}{\partial \alpha_2}\delta\alpha_2 + \cdots + \frac{\partial I}{\partial \alpha_n}\delta\alpha_n = 0. \tag{6.25}$$

Since the $\delta\alpha_i$ are arbitrary, Eq. (6.25) gives

$$\frac{\partial I}{\partial \alpha_i} = 0, \quad i = 1, 2, \ldots, n. \tag{6.26}$$

If I is a quadratic function of y and dy/dx, then Eq. (6.26) will be linear in α_i and can be solved easily.

We state, without proof, that the Rayleigh-Ritz method converges to the actual solution of the problem provided that the functions ϕ_i are linearly independent and satisfy at least the essential boundary conditions of the problem. The following examples illustrate the method of procedure.

Example 6.2. We consider the two-point boundary-value problem defined by

$$y'' + x = 0, \quad 0 < x < 1, \quad y(0) = y(1) = 0. \tag{i}$$

From Eq. (6.6), we have

$$I(v) = \int_0^1 v\left(-2x - v''\right) dx = -\int_0^1 v\left(2x + v''\right) dx. \tag{ii}$$

Let

$$v(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x) \tag{iii}$$

where

$$\phi_i(0) = \phi_i(1) = 0 \quad \text{for all } i. \tag{iv}$$

Substituting (iii) in (ii), we obtain

$$I(v) = -\int_0^1 \left[\sum_{i=1}^{n} \alpha_i \phi_i(x)\right] \left[2x + \sum_{j=1}^{n} \alpha_j \phi_j''(x)\right] dx. \tag{v}$$

For convenience, we set

$$p_i = \int_0^1 x\phi_i(x)dx \tag{vi}$$

and

$$\begin{aligned} q_{ij} &= \int_0^1 \phi_i(x)\phi_j''(x)dx \\ &= \left[\phi_i(x)\phi_j'(x)\right]_0^1 - \int_0^1 \phi_i'(x)\phi_j'(x)dx, \text{ on integrating by parts} \\ &= -\int_0^1 \phi_i'(x)\phi_j'(x)dx, \end{aligned} \tag{vii}$$

using boundary conditions (iv).

Then Eq. (v) becomes

$$I(v) = -2\sum_{i=1}^{n} \alpha_i p_i - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j q_{ij}$$

Hence $\partial I/\partial \alpha_i = 0$ gives

$$2p_i + 2\sum_{j=1}^{n} \alpha_j q_{ij} = 0, \quad (i = 1, 2, \ldots, n). \tag{viii}$$

We wish to find an approximate solution with $n = 2$ and we therefore choose $\phi_1(x) = x(1-x)$ and $\phi_2(x) = x^2(1-x)$, so that the boundary conditions (iv) are satisfied.

Now, from (vi), we have

$$p_1 = \int_0^1 x^2(1-x)dx = \frac{1}{12}$$

and

$$p_2 = \int_0^1 x^3(1-x)dx = \frac{1}{20}.$$

Also, $\phi_1'(x) = 1 - 2x$ and $\phi_2'(x) = 2x - 3x^2$. Equation (vii) gives

$$\begin{aligned}
q_{11} &= -\int_0^1 (1-2x)^2\, dx = -\frac{1}{3} \\
q_{12} &= -\int_0^1 (1-2x)\left(2x - 3x^2\right) dx = -\frac{1}{6} = q_{21}, \text{ by symmetry} \\
q_{22} &= -\int_0^1 \left(2x - 3x^2\right)^2 dx = -\frac{2}{15}.
\end{aligned}$$

Equations (viii) now give

$$4\alpha_1 + 2\alpha_2 = 1 \quad \text{and} \quad 10\alpha_1 + 8\alpha_2 = 3,$$

whose solution is $\alpha_1 = \alpha_2 = 1/6$. Hence

$$v(x) = \frac{1}{6}x(1-x) + \frac{1}{6}x^2(1-x) = \frac{1}{6}x\left(1 - x^2\right).$$

It can be verified that this is the exact solution of the problem (i).

Example 6.3. Solve the boundary-value problem defined by

$$y'' + y = -x, \quad 0 < x < 1 \tag{i}$$

with

$$y(0) = y(1) = 0 \tag{ii}$$

The exact solution of the problem (i) and (ii) is given by

$$y(x) = \frac{\sin x}{\sin 1} - x. \tag{iii}$$

To find the approximate solution by the Rayleigh-Ritz method, we take the functional in the form

$$I(v) = \int_0^1 \left(vv'' + v^2 + 2vx\right) dx. \tag{iv}$$

Let an approximate solution be given by

$$v(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x), \tag{v}$$

where

$$\phi_i(0) = \phi_i(1) = 0 \text{ for all } i. \tag{vi}$$

Substituting for v in (iv), we obtain

$$I(v) = \int_0^1 \left[\sum_{i=1}^{n} \alpha_i \phi_i(x) \sum_{j=1}^{n} \alpha_j \phi_j''(x) + \sum_{i=1}^{n} \alpha_i \phi_i(x) \sum_{j=1}^{n} \alpha_j \phi_j(x) + 2x \sum_{i=1}^{n} \alpha_i \phi_i(x)\right] dx \tag{vii}$$

As in the previous example, we let

$$p_i = \int_0^1 x\phi_i(x)dx \tag{viii}$$

and

$$q_{ij} = \int_0^1 \phi_i(x)\phi_j''(x)dx = -\int_0^1 \phi_i'(x)\phi_j'(x)dx. \tag{ix}$$

Further, let

$$r_{ij} = \int_0^1 \phi_i(x)\phi_j(x)dx. \tag{x}$$

Equation (vii) now becomes

$$I(v) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j q_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j r_{ij} + 2\sum_{i=1}^{n} \alpha_i p_i \tag{xi}$$

For minimum, we, therefore, have

$$\frac{\partial I}{\partial \alpha_i} = 2\sum_{j=1}^{n} \alpha_j q_{ij} + 2\sum_{j=1}^{n} \alpha_j r_{ij} + 2p_i = 0,$$

which simplifies to

$$\sum_{j=1}^{n} \alpha_j \left(q_{ij} + r_{ij}\right) = -p_i, \quad i = 1, 2, \ldots, n. \tag{xii}$$

To obtain an approximate solution, we take $n = 2$. Then, Eq. (xii) becomes

$$\left.\begin{aligned} \alpha_1 \left(q_{11} + r_{11}\right) + \alpha_2 \left(q_{12} + r_{12}\right) &= -p_1 \\ \alpha_1 \left(q_{21} + r_{21}\right) + \alpha_2 \left(q_{22} + r_{22}\right) &= -p_2 \end{aligned}\right\} \tag{xiii}$$

Choosing $\phi_1(x) = x(1-x)$ and $\phi_2(x) = x^2(1-x)$, we then obtain

$$\begin{aligned} p_1 &= \int_0^1 x^2(1-x)dx = \frac{1}{12}; \\ p_2 &= \int_0^1 x^3(1-x)dx = \frac{1}{20}; \\ q_{11} &= -\int_0^1 (1-2x)^2 dx = -\frac{1}{3}; \\ q_{12} &= -\int_0^1 (1-2x)\left(2x - 3x^2\right) dx = -\frac{1}{6}; \\ q_{22} &= -\int_0^1 \left(2x - 3x^2\right)^2 dx = -\frac{2}{15}; \\ r_{11} &= \int_0^1 x^2(1-x)^2 dx = \frac{1}{30}; \\ r_{12} &= \int_0^1 x^3(1-x)^2 dx = \frac{1}{60}; \\ r_{22} &= \int_0^1 x^4(1-x)^2 dx = \frac{1}{105}. \end{aligned}$$

Equation (xiii) now give

$$\begin{aligned} 18\alpha_1 + 9\alpha_2 &= 5 \\ 63\alpha_1 + 52\alpha_2 &= 21. \end{aligned}$$

Solving, we obtain

$$\alpha_1 = 0.1924 \quad \text{and} \quad \alpha_2 = 0.1707.$$

Hence the approximation is given by

$$y = x(1-x)(0.1924 + 0.1707x).$$

Example 6.4. Solve the boundary-value problem defined by

$$y'' - x = 0 \tag{i}$$

and

$$y(0) = 0, y'(1) = -\frac{1}{2} \tag{ii}$$

by the Rayleigh-Ritz method.

In this case, one of the boundary conditions is essential and the other natural. Also, the exact solution of the problem is given by

$$y(x) = \frac{x^3}{6} - x. \tag{iii}$$

The functional for this problem is given by

$$I(v) = \int_0^1 \left(v'^2 + 2vx\right) dx + v(1). \tag{iv}$$

Let

$$v(x) = \alpha_1 x + \alpha_2 x^2 \tag{v}$$

be an approximate solution so that $v(x)$ satisfies the essential boundary condition, viz., $v(0) = 0$. Then $v'(x) = \alpha_1 + 2\alpha_2 x$ and Eq. (iv) gives

$$I(v) = \int_0^1 \left[(\alpha_1 + 2\alpha_2 x)^2 + 2x\left(\alpha_1 x + \alpha_2 x^2\right)\right] dx + \alpha_1 + \alpha_2 \tag{vi}$$

Hence,

$$\left.\begin{array}{l} \frac{\partial I}{\partial \alpha_1} = 0 = \int_0^1 \left[2\left(\alpha_1 + 2\alpha_2 x\right) + 2x^2\right] dx + 1 \\ \frac{\partial I}{\partial \alpha_2} = 0 = \int_0^1 \left[2\left(\alpha_1 + 2\alpha_2 x\right) 2x + 2x^3\right] dx + 1. \end{array}\right\} \tag{vii}$$

Simplification gives the two equations

$$\alpha_1 + \alpha_2 = -\frac{5}{6} \quad \text{and} \quad \alpha_1 + \frac{4}{3}\alpha_2 = -\frac{3}{4}, \tag{viii}$$

whose solution is $\alpha_1 = -13/12$ and $\alpha_2 = 1/4$.
The approximate solution is given by

$$y^{(1)} = -\frac{13}{12}x + \frac{1}{4}x^2.$$

The student should compare this with the exact solution.

6.2.2 Galerkin's Method

The Rayleigh-Ritz method discussed in Section 6.2.1 is a powerful technique for the solution of boundary-value problems. It has, however, the disadvantage of requiring the existence of a functional which is not always possible to obtain. In fact, not all differential equations have a variational principle. Most engineering problems are expressed in terms of certain governing equations and boundary conditions, and not in terms of a functional. Galerkin's method belongs to a wider class of methods called the *weighted residual methods*. In this method, an approximating function called the *trial function* (which satisfies all the boundary conditions) is substituted in the given differential equation and the result is called the *residual* (the result will not be zero since we have substituted an approximating function). The residual is then weighted and the integral of the product, taken over the domain, is then set to zero. It can be shown that if the Euler-Lagrange equation corresponding to a functional coincides with the differential equation of the problem, then both the Rayleigh-Ritz and Galerkin methods yield the same system of equations.

See Section 3.4 for application of this method to solve two-point boundary value problems.

6.3 Application to two-dimensional problem

The application of the Rayleigh-Ritz and Galerkin methods to two-dimensional problems, although straightforward, is more complicated because of the increase in the number of parameters to be determined. We illustrate the application of Ritz method with an example.

Example 6.5. We consider Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = k, \quad 0 \le x, y \le 1 \tag{i}$$

with $u = 0$ on the boundary C of the region S.
The functional for the above problem is given by

$$I(v) = \iint_S v \left(2k - \frac{\partial^2 v}{\partial x^2} - \frac{\partial^2 v}{\partial y^2}\right) dxdy, \tag{ii}$$

where v vanishes on the boundary C. Let

$$v(x, y) = \alpha xy(x-1)(y-1) \tag{iii}$$

be a first approximation to u. Clearly, v satisfies the boundary conditions, i.e. $v = 0$ on the boundary C. The derivatives are given by

$$\left.\begin{array}{ll} \dfrac{\partial v}{\partial x} = \alpha y(y-1)(2x-1); & \dfrac{\partial v}{\partial y} = \alpha x(x-1)(2y-1); \\ \dfrac{\partial^2 v}{\partial x^2} = 2\alpha y(y-1); & \dfrac{\partial^2 v}{\partial y^2} = 2\alpha x(x-1). \end{array}\right\} \tag{iv}$$

Substituting for v in Eq. (ii), we obtain

$$I(v) = \int_0^1 \int_0^1 \alpha xy(x-1)(y-1)[2k - 2\alpha y(y-1) - 2\alpha x(x-1)]dxdy. \qquad \text{(v)}$$

Let

$$\left.\begin{aligned} a &= \int_0^1 \int_0^1 xy(x-1)(y-1)dxdy = \frac{1}{36} \\ b &= \int_0^1 \int_0^1 xy^2(x-1)(y-1)^2 dxdy = -\frac{1}{180} \\ c &= \int_0^1 \int_0^1 x^2 y(x-1)^2(y-1)dxdy = -\frac{1}{180}. \end{aligned}\right\} \qquad \text{(vi)}$$

Equation (v) now simplifies to

$$I(v) = 2k\alpha a - 2\alpha^2 b - 2\alpha^2 c.$$

Hence

$$\frac{\partial I}{\partial \alpha} = 0 = 2ka - 4\alpha b - 4\alpha c.$$

Thus

$$\alpha = \frac{ak}{2(b+c)} = -\frac{5}{4}k, \quad \text{using (vi).}$$

It follows that the required approximation for u is given by

$$u \approx v = -\frac{5}{4}kxy(x-1)(y-1).$$

The student should verify that the Galerkin method gives the same solution as above.

6.4 Finite Element Method

The Rayleigh-Ritz and Galerkin methods, discussed in the previous sections, cannot be applied directly for obtaining the global approximate solutions of engineering problems. An important reason for this is the difficulty associated with the choice of trial functions (satisfying the boundary conditions) particularly for complicated boundaries. This means that the application is restricted to problems with a simple geometry. Another reason is that very high-order polynomials have to be used to obtain global solutions with a reasonable accuracy. In the finite element method, the ideas of both the Rayleigh-Ritz and Galerkin methods are used in such a way that the above mentioned difficulties are overcome.

In the finite element method the region of interest is subdivided into a finite number of subregions, called the *elements*, and over each element the variational formulation of the given differential equation is constructed using simple functions for approximations.

The individual elements are then assembled and the equations for the whole problem are formed by a piecewise application of the variational method. For better accuracy it will not be necessary to increase the order of the functions used, but it would be sufficient to use a finer mesh. In this way, the difficulties encountered in the direct application of the variational methods are overcome. The basic steps involved in the finite element method are as follows:

(i) *Discretization:* The given domain is divided into a number of finite elements. The points of intersection are called *nodes*. The nodes and the elements are both numbered.

(ii) *Derivation of element equations:* For the given differential equation, a variational formulation is constructed over a typical element. The element equations are obtained by substituting a typical dependent variable, say

$$u = \sum_{i=1}^{n} u_i \psi_i$$

into the variational formulation. After choosing ψ_i, the interpolation functions, the element matrices are computed.

(iii) *Assembly:* The next step is the assembly of the element equations so that the total solution is continuous. When this is done, the entire system takes the matrix form

$$K\mathbf{u}' = \mathbf{F}'$$

where K = assemblage property matrix, and $\mathbf{u}'$ and $\mathbf{F}'$ are column vectors containing unknowns and external forces.

(iv) *Boundary conditions:* The above system of equations is modified using the boundary conditions of the problem.

(v) *Solution of the equations:* After incorporating the boundary conditions, the system is solved by any standard technique, for example, the LU decomposition.

The preceding steps are quite general but they are common to most finite element approaches. In the following section, these steps are elaborated and explained with an example of one-dimensional problem. Since the two-dimensional problems are modelled by partial differential equations, their finite element analysis is more complicated and are therefore not considered here.

6.4.1 Finite Element Method for One-dimensional Problems

We consider the two-point boundary value problem defined by

$$\frac{d^2y}{dx^2} = -f(x), \quad 0 < x < 1 \tag{6.27}$$

with the boundary conditions

$$y(0) = 0, \quad \left[\frac{dy}{dx}\right]_{x=1} = 0 \tag{6.28}$$

The basic steps involved in the finite element method are now elaborated and explained (see Reddy [6]):

Step 1 (Discretization of the region): In the present problem, the region of interest is the x-axis from $x = 0$ to $x = 1$. Suppose that this is divided into a set of subintervals, called *elements*, of unequal length, in general. The intersection points are called *nodes*. Let these be given by $x_0, x_1, x_2, \ldots, x_{n-1}, x_n$, where $x_0 = 0$ and $x_n = 1$. The elements are numbered as ①, ②, ③, ..., ⓝ, a typical element being the eth element of length h_e from node $e-1$ to node e. Let x_{e-1} and x_e be the values of x at the nodes $e-1$ and e, and let $y^{(e-1)}$ and $y^{(e)}$ be the values of y at these nodes, respectively. In general, $y^{(e)}$ satisfies the condition that *outside* e.

$$y^{(e)}(x) = 0 \quad \text{for all elements } e. \tag{6.29}$$

For example, in Fig. 6.2 $y^{(e)}(x_3)$ is nonzero whereas $y^{(e)}(x_4) = 0$.

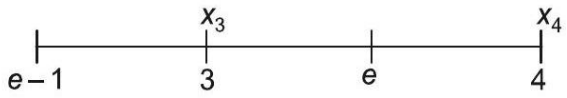


Figure 6.2: Typical eth element.

Using Eq. (6.29), it follows that the global approximate solution, $y(x)$, can be written as

$$y(x) = \sum_e y^{(e)}(x), \tag{6.30}$$

where the summation is taken over all the elements.

This completes the discretization process and in the next step, we choose a particular element e and formulate a variational principle for it.

Step 2 (Variational formulation over the element e): From Eq. (6.27), we obtain

$$\int_{x_{e-1}}^{x_e} v\frac{d^2y}{dx^2}dx = -\int_{x_{e-1}}^{x_e} vf\,dx$$

which is written as

$$\begin{aligned} 0 &= \int_{x_{e-1}}^{x_e} \left[v\frac{d^2y}{dx^2} + vf \right] dx \\ &= \left[v\frac{dy}{dx} \right]_{x_{e-1}}^{x_e} - \int_{x_{e-1}}^{x_e} v'\frac{dy}{dx}dx + \int_{x_{e-1}}^{x_e} vf dx \\ &= -\int_{x_{e-1}}^{x_e} \left[v'y' - vf \right] dx + v\left(x_e\right) D_2^{(e)} + v\left(x_{e-1}\right) D_1^{(e)}, \end{aligned} \tag{6.31}$$

where

$$D_1^{(e)} = \left[-\frac{dy}{dx} \right]_{x_{e-1}} \quad \text{and} \quad D_2^{(e)} = \left[\frac{dy}{dx} \right]_{x_e}. \tag{6.32}$$

In the next step, we use a variational method to approximate Eq. (6.31). We demonstrate this by using the Rayleigh-Ritz method.

Step 3 (Rayleigh-Ritz approximation over the element e): Let $y_e(x)$ be an approximation to $y(x)$ over the element e, so that

$$y_e(x) = \sum_{j=1}^{n} \alpha_j^{(e)} \phi_j(x), \tag{6.33}$$

where the α_j are parameters to be determined and $\phi_j(x)$ are approximation functions to be chosen. Substituting Eq. (6.33) in Eq. (6.31), we obtain

$$\begin{aligned} &\sum_{j=1}^{n} \alpha_j^{(e)} \left[\int_{x_{e-1}}^{x_e} \phi_j'(x)\phi_i'(x)dx \right] \\ &\quad = \int_{x_{e-1}}^{x_e} f\phi_i(x)dx + \phi_i\left(x_e\right) D_2^{(e)} + \phi_i\left(x_{e-1}\right) D_1^{(e)}, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{6.34}$$

Equation (6.34) can be written in the matrix form

$$K_{ij}^{(e)} \alpha_j^{(e)} = F_i^{(e)}, \tag{6.35}$$

where K_{ij} and F_i are called the *stiffness matrix* and *force vector* respectively, and are given by

$$K_{ij}^{(e)} = \int_{x_{e-1}}^{x_e} \phi_i'(x)\phi_j'(x)dx \tag{6.36}$$

and

$$F_i^{(e)} = \int_{x_{e-1}}^{x_e} f\phi_i(x)dx + \phi_i\left(x_e\right) D_2^{(e)} + \phi_i\left(x_{e-1}\right) D_1^{(e)}. \tag{6.37}$$

In the Rayleigh-Ritz and Galerkin methods, the system of equations is obtained in terms of the arbitrary parameters α_j. In the finite element method, on the other hand, the unknown

values of the dependent variable y at the nodes are taken as parameters. This is done in the following way. Let

$$y(x) = \alpha_1 + \alpha_2 x \tag{6.38}$$

be an approximation in the element e. We have

$$\left.\begin{aligned} y\left(x_{e-1}\right) &= \alpha_1 + \alpha_2 x_{e-1} = y_1^{(e)} \\ y\left(x_e\right) &= \alpha_1 + \alpha_2 x_e = y_2^{(e)}. \end{aligned}\right\} \tag{6.39}$$

Solving the equations given in Eq. (6.39), we obtain

$$\alpha_1 = \frac{y_1^{(e)} x_e - y_2^{(e)} x_{e-1}}{x_e - x_{e-1}} \tag{6.40}$$

and

$$\alpha_2 = \frac{y_2^{(e)} - y_1^{(e)}}{x_e - x_{e-1}}. \tag{6.41}$$

Equation (6.38) now becomes

$$\begin{aligned} y(x) &= \frac{y_1^{(e)} x_e - y_2^{(e)} x_{e-1}}{x_e - x_{e-1}} + \frac{y_2^{(e)} - y_1^{(e)}}{x_e - x_{e-1}} x \\ &= \frac{x_e - x}{x_e - x_{e-1}} y_1^{(e)} + \frac{x - x_{e-1}}{x_e - x_{e-1}} y_2^{(e)} \\ &= \sum_{i=1}^{2} y_i^{(e)} \phi_i^{(e)}(x) \end{aligned} \tag{6.42}$$

where

$$\phi_1^{(e)}(x) = \frac{x_e - x}{x_e - x_{e-1}} \quad \text{and} \quad \phi_2^{(e)}(x) = \frac{x - x_{e-1}}{x_e - x_{e-1}}. \tag{6.43}$$

With $x_1 = x_{e-1}$ and $x_2 = x_e$, the functions $\phi_i^{(e)}$ have the property

$$\phi_i^{(e)}\left(x_j\right) = \begin{cases} 0, & i \neq j \\ 1, & i = j. \end{cases} \tag{6.44}$$

Instead of Eq. (6.35), we now have

$$K_{ij}^{(e)} y_j^{(e)} = F_i^{(e)}, \tag{6.45}$$

where $K_{ij}^{(e)}$ and $F_i^{(e)}$ are given by Eq. (6.36) and (6.37).

With the choice of $\phi_i^{(e)}(x)$ as in Eq. (6.43), we now demonstrate the computation of $K^{(e)}$ and $F^{(e)}$. In particular, we choose $f = 2$. With $h_e = x_e - x_{e-1}$, we obtain

$$\frac{d\phi_1^{(e)}}{dx} = -\frac{1}{h_e} \quad \text{and} \quad \frac{d\phi_2^{(e)}}{dx} = \frac{1}{h_e} \tag{6.46}$$

where

$$\left.\begin{aligned} K_{11} &= \int_{x_{e-1}}^{x_e} \left(-\frac{1}{h_e}\right)^2 dx = \frac{1}{h_e} \\ K_{12} &= \int_{x_{e-1}}^{x_e} -\frac{1}{h_e^2} dx = -\frac{1}{h_e} = K_{21} \\ K_{22} &= \int_{x_{e-1}}^{x_e} \frac{1}{h_e^2} dx = \frac{1}{h_e} \end{aligned}\right\} \tag{6.47}$$

and

$$\left.\begin{aligned} F_1^{(e)} &= 2\int_{x_{e-1}}^{x_e} \frac{x_e - x}{h_e} dx + D_1^{(e)} = h_e + D_1^{(e)} \\ F_2^{(e)} &= 2\int_{x_{e-1}}^{x_e} \frac{x - x_{e-1}}{h_e} dx + D_2^{(e)} = h_e + D_2^{(e)}. \end{aligned}\right\} \tag{6.48}$$

As a particular case, we consider the following example.

Example 6.6. We consider the following problem defined by

$$\frac{d^2y}{dx^2} = -2, \quad 0 < x < 1, \quad y(0) = 0, y'(1) = 0. \tag{i}$$

The exact solution of the above problem is given by

$$y(x) = 2x - x^2 \tag{ii}$$

Comparison with Eq. (6.27) shows that $f(x) = 2$.

(a) To demonstrate the steps involved in the finite element solution, we divide $[0, 1]$ into two equal subintervals with $h_e = 1/2$. From Eqs. (6.43) and (6.44), we obtain the equations for both elements.

(i) $e = 1 : x_{e-1} = 0, x_e = 1/2$,

$$K^{(1)} = \begin{bmatrix} 2 & -2 \\ \\ -2 & 2 \end{bmatrix}, \quad F^{(1)} = \begin{bmatrix} \frac{1}{2} + D_1^{(1)} \\ \\ \frac{1}{2} + D_2^{(1)} \end{bmatrix}.$$

(ii) $e = 2 : x_{e-1} = 1/2, x_e = 1$

$$K^{(2)} = \begin{bmatrix} 2 & -2 \\ \\ -2 & 2 \end{bmatrix}, \quad F^{(2)} = \begin{bmatrix} \frac{1}{2} + D_1^{(2)} \\ \\ \frac{1}{2} + D_2^{(2)} \end{bmatrix}.$$

Having determined the equations for each element, these have to be assembled now to determine the global approximations. This will be the next step in the finite element solution.

Step 4 (Assembly of element equations): We shall explain this step with reference to the two elements obtained in Example 6.6. In this case, the two elements are connected at the node 2. Since the function $y(x)$ is continuous, it follows that y_2 of element 1 should be the same as y_1 of element 2. For the two elements of Example 6.6, the correspondence can be expressed mathematically as follows:

$$y_1^{(1)} = Y_1, \quad y_2^{(1)} = Y_2 = y_1^{(2)}, \quad y_2^{(2)} = Y_3.$$

In the finite element analysis, such relations are usually called *interelement continuity conditions*.

Using the above relations, the global finite element model of the given boundary value problem is

$$\begin{bmatrix} 2 & -2 & 0 \\ -2 & 2+2 & -2 \\ 0 & -2 & 2 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1/2 + D_1^{(1)} \\ 1 + D_2^{(1)} + D_2^{(2)} \\ 1/2 + D_2^{(2)} \end{bmatrix}.$$

The next step is the imposition of boundary conditions.

Step 5 (Imposition of boundary conditions): The homogeneous boundary condition gives $Y_1 = 0$. Then, we obtain the equations:

$$4Y_2 - 2Y_3 = 1, \quad -2Y_2 + 2Y_3 = \frac{1}{2}$$

since $D_2^{(1)}$ and $D_2^{(2)}$ cancel each other and $D_2^{(2)} = 0$ is the natural boundary condition. The solution of this system is given by

$$Y_2 = \frac{3}{4} \quad \text{and} \quad Y_3 = 1$$

Finally, the approximate solution throughout the interval $[0, 1]$ can now be found using the formula

$$y(x) = \sum_{i=1}^{2} Y_i \phi_i^{(e)}(x) = \begin{cases} Y_1\phi_1^{(1)}(x) + Y_2\phi_2^{(1)}(x), & 0 \le x \le \dfrac{1}{2} \\ Y_2\phi_1^{(2)}(x) + Y_3\phi_2^{(2)}(x), & \dfrac{1}{2} \le x \le 1 \end{cases}$$

$$= \begin{cases} \dfrac{3}{2}x, & 0 \le x \le \dfrac{1}{2} \\ \dfrac{x+1}{2}, & \dfrac{1}{2} \le x \le 1 \end{cases}$$

on substitutions and simplification.

From the above, we obtain the approximate value of $y(1/4) \approx 3/8 = 0.375$, whereas its exact value $= 2(1/4) - 1/16 = 7/16$.

(b) To improve the accuracy, we now consider four elements of length $1/4$. In this case, the element matrices become

(i) $e = 1 : x_{e-1} = 0, x_e = 1/4$

$$K_{11}^{(1)} = 4, \quad K_{12}^{(1)} = K_{21}^{(1)} = -4, \quad K_{22}^{(1)} = 4$$

$$F_1^{(1)} = \frac{1}{4} + D_1^{(1)}, \quad F_2^{(1)} = \frac{1}{4} + D_2^{(1)}$$

$$K^{(1)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \quad F^{(1)} = \begin{bmatrix} 1/4 + D_1^{(1)} \\ 1/4 + D_2^{(1)} \end{bmatrix}$$

(ii) $e = 2 : x_{e-1} = 1/4, x_e = 1/2$

$$K^{(2)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \quad F^{(2)} = \begin{bmatrix} 1/4 + D_1^{(2)} \\ 1/4 + D_2^{(2)} \end{bmatrix}$$

(iii) $e = 3 : x_{e-1} = 1/2, x_e = 3/4$

$$K^{(3)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \quad F^{(3)} = \begin{bmatrix} 1/4 + D_1^{(3)} \\ 1/4 + D_2^{(3)} \end{bmatrix}$$

(iv) $e = 4 : x_{e-1} = 3/4, x_e = 1$

$$K^{(4)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \quad F^{(4)} = \begin{bmatrix} 1/4 + D_1^{(4)} \\ 1/4 + D_2^{(4)} \end{bmatrix}$$

To avoid confusion, we now write down the complete system for each element

$$e=1\begin{bmatrix}4&-4&0&0&0\\-4&4&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{bmatrix}\begin{bmatrix}Y_1\\Y_2\\Y_3\\Y_4\\Y_5\end{bmatrix}=\begin{bmatrix}1/4+D_1^{(1)}\\1/4+D_2^{(1)}\\0\\0\\0\end{bmatrix}$$

$$e=2\begin{bmatrix}0&0&0&0&0\\0&4&-4&0&0\\0&-4&4&0&0\\0&0&0&0&0\\0&0&0&0&0\end{bmatrix}\begin{bmatrix}Y_1\\Y_2\\Y_3\\Y_4\\Y_5\end{bmatrix}=\begin{bmatrix}0\\1/4+D_1^{(2)}\\1/4+D_2^{(2)}\\0\\0\end{bmatrix}$$

$$e=3\begin{bmatrix}0&0&0&0&0\\0&0&0&0&0\\0&0&4&-4&0\\0&0&-4&4&0\\0&0&0&0&0\end{bmatrix}\begin{bmatrix}Y_1\\Y_2\\Y_3\\Y_4\\Y_5\end{bmatrix}=\begin{bmatrix}0\\0\\1/4+D_1^{(3)}\\1/4+D_2^{(3)}\\0\end{bmatrix}$$

$$e=4\begin{bmatrix}0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&4&-4\\0&0&0&-4&4\end{bmatrix}\begin{bmatrix}Y_1\\Y_2\\Y_3\\Y_4\\Y_5\end{bmatrix}=\begin{bmatrix}0\\0\\0\\D_1^{(4)}\\D_2^{(4)}\end{bmatrix}$$

Adding up the above, we obtain

$$\begin{bmatrix}4&-4&0&0&0\\-4&4+4&-4&0&0\\0&-4&4+4&-4&0\\0&0&-4&4+4&-4\\0&0&0&-4&4\end{bmatrix}\begin{bmatrix}Y_1\\Y_2\\Y_3\\Y_4\\Y_5\end{bmatrix}=\begin{bmatrix}1/4+D_1^{(1)}\\1/2+D_2^{(1)}+D_1^{(2)}\\1/2+D_2^{(2)}+D_1^{(3)}\\1/2+D_2^{(3)}+D_1^{(4)}\\1/4+D_2^{(4)}\end{bmatrix}$$

By boundary condition, we have $Y_1=0$.
The system now becomes

$$\begin{aligned}8Y_2-4Y_3&=1/2\\-4Y_2+8Y_3-4Y_4&=1/2\\-4Y_3+8Y_4-4Y_5&=1/2\\-4Y_4+4Y_5&=1/4+D_2^{(4)}\end{aligned}$$

But $D_2^{(4)} = y'(1) = 0$.
The solution of the system is

$$Y_2 = \frac{7}{16}, \quad Y_3 = \frac{3}{4}, \quad Y_4 = \frac{15}{16}, \quad Y_5 = 1.$$

Then the approximate solution valid for $[0, 1]$ is

$$y(x) = \begin{cases} \dfrac{7x}{4}, & 0 \le x \le \dfrac{1}{4} \\ \dfrac{10x+1}{8}, & \dfrac{1}{4} \le x \le \dfrac{1}{2} \\ \dfrac{6x+3}{8}, & \dfrac{1}{2} \le x \le \dfrac{3}{4} \\ \dfrac{x+3}{4}, & \dfrac{3}{4} \le x \le 1. \end{cases}$$

From the above, we obtain $y(1/4) = 7/16$, which is the same as the exact solution.

7 Error analysis

7.1 ERRORS AND THEIR COMPUTATIONS

There are two kinds of numbers, exact and approximate numbers. Examples of exact numbers are $1, 2, 3, \ldots, 1/2, 3/2, \ldots, \sqrt{2}, \pi$, e, etc., written in this manner. Approximate numbers are those that represent the numbers to a certain degree of accuracy. Thus, an approximate value of π is 3.1416 , or if we desire a better approximation, it is 3.14159265. But we cannot write the exact value of π.

The digits that are used to express a number are called significant digits or significant figures. Thus, the numbers $3.1416, 0.66667$ and 4.0687 contain five significant digits each. The number 0.00023 has, however, only two significant digits, viz., 2 and 3, since the zeros serve only to fix the position of the decimal point. Similarly, the numbers $0.00145, 0.000145$ and 0.0000145 all have three significant digits. In case of ambiguity, the scientific notation should be used. For example, in the number 25,600, the number of significant figures is uncertain, whereas the numbers $2.56 \times 10^4, 2.560 \times 10^4$ and 2.5600×10^4 have three, four and five significant digits, respectively.

In numerical computations, we come across numbers which have large number of digits and it will be necessary to cut them to a usuable number of figures. This process is called rounding off. It is usual to round-off numbers according to the following rule:

To round-off a number to n significant digits, discard all digits to the right of the nth digit, and if this discarded number is
(a) less than half a unit in the nth place, leave the nth digit unaltered;
(b) greater than half a unit in the nth place, increase the nth digit by unity;
(c) exactly half a unit in the nth place, increase the nth digit by unity if it is odd; otherwise, leave it unchanged.
The number thus rounded-off is said to be correct to n significant figures.

Example 7.1 The numbers given below are rounded-off to four significant figures:

1.6583	to	1.658
30.0567	to	30.06
0.859378	to	0.8594
3.14159	to	3.142

In hand computations, the round-off error can be reduced by carrying out the computations to more significant figures at each step of the computation. A useful rule is: at each step of the computation, retain at least one more significant figure than that given in the data, perform the last operation and then round-off. However, most computers allow more number of significant figures than are usually required in engineering computations. Thus, there are computers which allow a precision of seven significant figures in the range of about 10^{-38} to 10^{39}. Arithmetic carried out with this precision is called single precision arithmetic, and several computers implement double precision arithmetic, which could be

used in problems requiring greater accuracy. Usually, the double precision arithmetic is carried out to 15 decimals with a range of about 10^{-308} to 10^{308}. In MATLAB, there is a provision to use double precision arithmetic.

In addition to the round-off error discussed above, there is another type of error which can be caused by using approximate formulae in computations, —such as the one that arises when a truncated infinite series is used. This type of error is called truncation error and its study is naturally associated with the problem of convergence. Truncation error in a problem can be evaluated and we are often required to make it as small as possible. Sections 7.4 and 7.5 will be devoted to a discussion of these errors.

7.2 Absolute, relative and percentage errors

Absolute error is the numerical difference between the true value of a quantity and its approximate value. Thus, if X is the true value of a quantity and X_1 is its approximate value, then the absolute error E_{A} is given by

$$E_{\mathrm{A}} = X - X_1 = \delta X \tag{7.1}$$

The relative error E_{R} is defined by

$$E_{\mathrm{R}} = \frac{E_{\mathrm{A}}}{X} = \frac{\delta X}{X} \tag{7.2}$$

and the percentage error (E_{P}) by

$$E_{\mathrm{P}} = 100 E_{\mathrm{R}} \tag{7.3}$$

Let ΔX be a number such that

$$|X_1 - X| \leq \Delta X \tag{7.4}$$

Then ΔX is an upper limit on the magnitude of the absolute error and is said to measure absolute accuracy. Similarly, the quantity

$$\frac{\Delta X}{|X|} \approx \frac{\Delta X}{|X_1|}$$

measures the relative accuracy.

It is easy to deduce that if two numbers are added or subtracted, then the magnitude of the absolute error in the result is the sum of the magnitudes of the absolute errors in the two numbers. More generally, if $E_{\mathrm{A}}^1, E_{\mathrm{A}}^2, \ldots, E_{\mathrm{A}}^n$ are the absolute errors in n numbers, then the magnitude of the absolute error in their sum is given by

$$\left|E_{\mathrm{A}}^1\right| + \left|E_{\mathrm{A}}^2\right| + \cdots + |E_{\mathrm{A}}^n|$$

Note: While adding up several numbers of different absolute accuracies, the following procedure may be adopted:
(i) Isolate the number with the greatest absolute error,
(ii) Round-off all other numbers retaining in them one digit more than in the isolated number,
(iii) Add up, and
(iv) Round-off the sum by discarding one digit.

To find the absolute error, E_{A}, in a product of two numbers a and b, we write $E_{\mathrm{A}} = \left(a + E_{\mathrm{A}}^1\right)\left(b + E_{\mathrm{A}}^2\right) - ab$, where E_{A}^1 and E_{A}^2 are the absolute errors in a and b respectively. Thus,

$$\begin{aligned} E_{\mathrm{A}} &= aE_{\mathrm{A}}^2 + bE_{\mathrm{A}}^1 + E_{\mathrm{A}}^1 E_{\mathrm{A}}^2 \\ &= bE_{\mathrm{A}}^1 + aE_{\mathrm{A}}^2, \text{ approximately} \end{aligned} \tag{7.5}$$

Similarly, the absolute error in the quotient a/b is given by

$$\begin{aligned} \frac{a + E_{\mathrm{A}}^1}{b + E_{\mathrm{A}}^2} - \frac{a}{b} &= \frac{bE_{\mathrm{A}}^1 - aE_{\mathrm{A}}^2}{b\left(b + E_{\mathrm{A}}^2\right)} \\ &= \frac{bE_{\mathrm{A}}^1 - aE_{\mathrm{A}}^2}{b^2\left(1 + E_{\mathrm{A}}^2/b\right)} \\ &= \frac{bE_{\mathrm{A}}^1 - aE_{\mathrm{A}}^2}{b^2}, \text{ assuming that } E_{\mathrm{A}}^2/b \text{ is small in comparison with 1} \\ &= \frac{a}{b}\left(\frac{E_{\mathrm{A}}^1}{a} - \frac{E_{\mathrm{A}}^2}{b}\right) \end{aligned} \tag{7.6}$$

Example 7.2

If the number X is rounded to N decimal places, then

$$\Delta X = \frac{1}{2}\left(10^{-N}\right)$$

If $X = 0.51$ and is correct to 2 decimal places, then $\Delta X = 0.005$, and the percentage accuracy is given by $\frac{0.005}{0.51} \times 100 = 0.98\%$.

Example 7.3

An approximate value of π is given by $X_1 = 22/7 = 3.1428571$ and its true value is $X = 3.1415926$. Find the absolute and relative errors. We have

$$E_{\mathrm{A}} = X - X_1 = -0.0012645$$

and

$$E_{\mathrm{R}} = \frac{-0.0012645}{3.1415926} = -0.000402$$

Example 7.4 Three approximate values of the number 1/3 are given as $0.30, 0.33$ and 0.34 . Which of these three values is the best approximation?

We have

$$\left|\frac{1}{3} - 0.30\right| = \frac{1}{30}$$
$$\left|\frac{1}{3} - 0.33\right| = \frac{0.01}{3} = \frac{1}{300}$$
$$\left|\frac{1}{3} - 0.34\right| = \frac{0.02}{3} = \frac{1}{150}$$

It follows that 0.33 is the best approximation for 1/3.

Example 7.5 Find the relative error of the number 8.6 if both of its digits are correct.

Here

$$E_{\mathrm{A}} = 0.05$$

Hence

$$E_{\mathrm{R}} = \frac{0.05}{8.6} = 0.0058$$

Example 7.6

Evaluate the sum $S = \sqrt{3} + \sqrt{5} + \sqrt{7}$ to 4 significant digits and find its absolute and relative errors.

We have

$$\sqrt{3} = 1.732, \sqrt{5} = 2.236 \text{ and } \sqrt{7} = 2.646$$

Hence $S = 6.614$. Then

$$E_{\mathrm{A}} = 0.0005 + 0.0005 + 0.0005 = 0.0015$$

The total absolute error shows that the sum is correct to 3 significant figures only. Hence we take $S = 6.61$ and then

$$E_{\mathrm{R}} = \frac{0.0015}{6.61} = 0.0002$$

Example 7.7

Sum the following numbers:

$0.1532, 15.45, 0.000354, 305.1, 8.12, 143.3, 0.0212, 0.643$ and 0.1734, where in each of which all the given digits are correct.

Here we have two numbers which have the greatest absolute error. These are 305.1 and 143.3, and the absolute error in both these numbers is 0.05 . Hence, we round-off all the other numbers to two decimal digits. These are:

$$0.15, 15.45, 0.00, 8.12, 0.02, 0.64 \text{ and } 0.17$$

The sum S is given by

$$\begin{aligned} S &= 305.1 + 143.3 + 0.15 + 15.45 + 0.00 + 8.12 + 0.02 + 0.64 + 0.17 \\ &= 472.95 \\ &= 473 \end{aligned}$$

To determine the absolute error, we note that the first-two numbers have each an absolute error of 0.05 and the remaining seven numbers have an absolute error of 0.005 each. Thus, the absolute error in all the 9 numbers is

$$\begin{aligned} E_{\text{A}} &= 2(0.05) + 7(0.005) \\ &= 0.1 + 0.035 \\ &= 0.135 \\ &= 0.14 \end{aligned}$$

In addition to the above absolute error, we have to take into account the rounding error in the above and this is 0.01 . Hence the total absolute error is $S = 0.14 + 0.01 = 0.15$. Thus,

$$S = 472.95 \pm 0.15$$

Example 7.8
Find the difference

$$\sqrt{6.37} - \sqrt{6.36}$$

to three significant figures.
We have

$$\sqrt{6.37} = 2.523885893$$

and

$$\sqrt{6.36} = 2.521904043$$

Therefore, $\sqrt{6.37} - \sqrt{6.36} = 0.001981850$

$= 0.00198$, correct to three significant figures.

Alternatively, we have

$$\begin{aligned}\sqrt{6.37}-\sqrt{6.36} &= \frac{6.37-6.36}{\sqrt{6.37}+\sqrt{6.36}} \\ &= \frac{0.01}{2.524+2.522} \\ &= 0.198\times 10^{-2}, \text{ which is the same result as obtained before.}\end{aligned}$$

Example 7.9

Two numbers are given as 2.5 and 48.289, both of which being correct to the significant figures given. Find their product.

Here the number 2.5 is the one with the greatest absolute error. Hence, we round-off the second number to three significant digits, i.e., 48.3. The required product is given by

$$\begin{aligned}P &= 48.3\times 2.5 \\ &= 1.2\times 10^{2}\end{aligned}$$

In the product, we retained only two significant digits, since one of the given numbers, viz. 2.5 , contained only two significant digits.

7.3 A GENERAL ERROR FORMULA

We now derive a general formula for the error committed in using a certain formula or a functional relation. Let

$$u = f(x, y, z) \tag{7.7}$$

and let the errors in x, y, z be $\Delta x, \Delta y$ and Δz, respectively. Then the error Δu in u is given by

$$u + \Delta u = f(x + \Delta x, y + \Delta y, z + \Delta z) \tag{7.8}$$

Expanding the right-side of Eq. (7.8) by Taylor's series, we obtain

$$\begin{aligned}u + \Delta u =& f(x, y, z) + \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z \\ &+ \text{ terms involving higher powers of } \Delta x, \Delta y \text{ and } \Delta z\end{aligned} \tag{7.9}$$

Assuming that the errors $\Delta x, \Delta y, \Delta z$ are small, their higher powers can be neglected and Eq. (7.9) becomes

$$\Delta u = \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z \tag{7.10}$$

The relative error in u is then given by

$$E_{\mathrm{R}} = \frac{\Delta u}{u} = \frac{\partial u}{\partial x}\frac{\Delta x}{u} + \frac{\partial u}{\partial y}\frac{\Delta y}{u} + \frac{\partial u}{\partial z}\frac{\Delta z}{u} \tag{7.11}$$

Example 7.10
Find the value of

$$s = \frac{a^2\sqrt{b}}{c^3},$$

where $a = 6.54 \pm 0.01, b = 48.64 \pm 0.02$, and $c = 13.5 \pm 0.03$.
Also, find the relative error in the result.

We have

$$a^2 = 42.7716, \sqrt{b} = 6.9742 \text{ and } c^3 = 2460.375$$

Therefore,

$$\begin{aligned} s &= \frac{42.7716 \times 6.9742}{2460.375} = 0.12124\ldots \\ &= 0.121 \end{aligned}$$

Also,

$$\log s = 2\log a + \frac{1}{2}\log b - 3\log c$$

$$\begin{aligned} \Rightarrow \left|\frac{\Delta s}{s}\right| \leq 2\left|\frac{\Delta a}{a}\right| + \frac{1}{2}\left|\frac{\Delta b}{b}\right| + 3\left|\frac{\Delta c}{c}\right| &= 2\left(\frac{0.01}{6.54}\right) + \frac{1}{2}\left(\frac{0.02}{48.64}\right) + 3\left(\frac{0.03}{13.5}\right) \\ &= 0.009931 \end{aligned}$$

Example 7.11
Given that

$$u = \frac{5xy^2}{z^3}$$

find the relative error at $x = y = z = 1$ when the errors in each of x, y, z is 0.001 .
We have

$$\frac{\partial u}{\partial x} = \frac{5y^2}{z^3}, \frac{\partial u}{\partial y} = \frac{10xy}{z^3} \text{ and } \frac{\partial u}{\partial z} = -\frac{15xy^2}{z^4}$$

Then

$$\Delta u = \frac{5y^2}{z^3}\Delta x + \frac{10xy}{z^3}\Delta y - \frac{15xy^2}{z^4}\Delta z.$$

In general, the errors $\Delta x, \Delta y$ and Δz may be positive or negative. Hence, we take the absolute values of the terms on the right side. We then obtain

$$(\Delta u)_{\max} = \left|\frac{5y^2}{z^3}\Delta x\right| + \left|\frac{10xy}{z^3}\Delta y\right| + \left|\frac{15xy^2}{z^4}\Delta z\right|$$

but $\Delta x = \Delta y = \Delta z = 0.001$ and $x = y = z = 1$. Then, the relative maximum error $(E_{\mathrm{R}})_{\max}$ is given by

$$\begin{aligned}(E_{\mathrm{R}})_{\max} &= \frac{(\Delta u)_{\max}}{u} \\ &= \frac{0.03}{5} = 0.006\end{aligned}$$

7.4 ERROR IN A SERIES APPROXIMATION

The truncated error committed in a series approximation can be evaluated by using Taylor's series. If x_i and x_{i+1} are two successive values of x, then we have

$$f(x_{i+1}) = f(x_i) + (x_{i+1} - x_i) f'(x_i) + \cdots + \frac{(x_{i+1} - x_i)^n}{n!} f^{(n)}(x_i) + R_{n+1}(x_{i+1}), \quad (7.12)$$

where

$$R_{n+1}(x_{i+1}) = \frac{(x_{i+1} - x_i)^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \quad x_i < \xi < x_{i+1} \quad (7.13)$$

In Eq. (7.12), the last term, $R_{n+1}(x_{i+1})$, is called the remainder term which, for a convergent series, tends to zero as $n \to \infty$. Thus, if $f(x_{i+1})$ is approximated by the first-n terms of the series given in Eq. (7.12), then the maximum error committed by using this approximation (called the nth order approximation) is given by the remainder term $R_{n+1}(x_{i+1})$. Conversely, if the accuracy required is specified in advance, then it would be possible to find n, the number of terms, such that the finite series yields the required accuracy.

Defining the interval length,

$$x_{i+1} - x_i = h, \quad (7.14)$$

Equation (7.12) may be written as

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2!} f''(x_i) + \cdots + \frac{h^n}{n!} f^{(n)}(x_i) + \mathrm{O}\left(h^{n+1}\right), \quad (7.15)$$

where $\mathrm{O}\left(h^{n+1}\right)$ means that the truncation error is of the order of h^{n+1}, i.e., it is proportional to h^{n+1}. The meaning of this statement will be made clearer now.

Let the series be truncated after the first term. This gives the zero-order approximation:

$$f\left(x_{i+1}\right) = f\left(x_i\right) + \mathrm{O}(h) \tag{7.16}$$

which means that halving the interval length h will also halve the error in the approximate solution. Similarly, the first-order Taylor series approximation is given by

$$f\left(x_{i+1}\right) = f\left(x_i\right) + hf'\left(x_i\right) + \mathrm{O}\left(h^2\right) \tag{7.17}$$

which means that halving the interval length, h will quarter the error in the approximation. In such a case we say that approximation has a second-order of convergence. We illustrate these facts through numerical examples.

Example 7.12

Evaluate $f(1)$ using Taylor's series for $f(x)$, where

$$f(x) = x^3 - 3x^2 + 5x - 10$$

It is easily seen that $f(1) = -7$ but it will be instructive to see how the Taylor series approximations of orders 0 to 3 improve the accuracy of $f(1)$ gradually.

Let $h = 1, x_i = 0$ and $x_{i+1} = 1$. We then require $f\left(x_{i+1}\right)$. The derivatives of $f(x)$ are given by

$$f'(x) = 3x^2 - 6x + 5, f''(x) = 6x - 6, f'''(x) = 6$$

$f^{\mathrm{iv}}\left(x\right)$ and higher derivatives being all zero. Hence

$$f'\left(x_i\right) = f'(0) = 5, \quad f''\left(x_i\right) = f''(0) = -6, \quad f'''(0) = 6$$

Also,

$$f\left(x_i\right) = f(0) = -10$$

Hence, Taylor's series gives

$$f\left(x_{i+1}\right) = f\left(x_i\right) + hf'\left(x_i\right) + \frac{h^2}{2}f''\left(x_i\right) + \frac{h^3}{6}f'''\left(x_i\right) \tag{i}$$

From Eq. (i), the zero-order approximation is given by

$$f\left(x_{i+1}\right) = f\left(x_i\right) + \mathrm{O}(h) \tag{ii}$$

and, therefore,

$$f(1) = f(0) + O(h) \approx -10,$$

the error in which is $-7+10$, i.e., 3 units.
For the first approximation, we have

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \mathrm{O}\left(h^2\right) \tag{iii}$$

and, therefore,

$$f(1) = -10 + 5 + \mathrm{O}\left(h^2\right) \approx -5$$

the error in which is $-7+5$, i.e., -2 units.
Again, the second-order Taylor approximation is given by

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2}f''(x_i) + \mathrm{O}\left(h^3\right) \tag{iv}$$

and, therefore,

$$f(1) = -10 + 5 + \frac{1}{2}(-6) + O\left(h^3\right) \approx -8$$

in which the error is $-7+8$, i.e., 1 unit.
Finally, the third-order Taylor series approximation is given by

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2}f''(x_i) + \frac{h^3}{6}f'''(x_i), \tag{v}$$

and, therefore,

$$\begin{aligned} f(1) &= f(0) + hf'(x_0) + \frac{h^2}{2}f''(x_0) + \frac{h^3}{6}f'''(x_0) \\ &\approx -10 + 5 + \frac{1}{2}(-6) + \frac{1}{6}(6) \\ &= -7 \end{aligned}$$

which is the exact value of $f(1)$.
This example demonstrates that if the given function is a polynomial of third degree, then its third-order Taylor series approximation gives exact results.

Example 7.13 Given $f(x) = \sin x$, construct the Taylor series approximations of orders 0 to 7 at $x = \pi/3$ and state their absolute errors.

Let $x_{i+1} = \pi/3$ and $x_i = \pi/6$ so that $h = \pi/3 - \pi/6 = \pi/6$. We then have

$$\begin{aligned} f\left(\frac{\pi}{3}\right) =& f\left(\frac{\pi}{6}\right) + hf'\left(\frac{\pi}{6}\right) + \frac{h^2}{2}f''\left(\frac{\pi}{6}\right) + \frac{h^3}{6}f'''\left(\frac{\pi}{6}\right) + \frac{h^4}{24}f^{\mathrm{iv}}\left(\frac{\pi}{6}\right) \\ &+ \frac{h^5}{120}f^{\mathrm{v}}\left(\frac{\pi}{6}\right) + \frac{h^6}{720}f^{\mathrm{vi}}\left(\frac{\pi}{6}\right) + \frac{h^7}{5040}f^{\mathrm{vii}}\left(\frac{\pi}{6}\right) + \mathrm{O}\left(h^8\right) \end{aligned} \tag{i}$$

Since $f(x) = \sin x$, Eq. (i) becomes:

$$\begin{aligned}\sin\left(\frac{\pi}{3}\right) \approx & \sin\left(\frac{\pi}{6}\right) + \frac{\pi}{6}\cos\left(\frac{\pi}{6}\right) + \frac{1}{2}\left(\frac{\pi}{6}\right)^2\left(-\sin\frac{\pi}{6}\right) + \frac{1}{6}\left(\frac{\pi}{6}\right)^3\left(-\cos\frac{\pi}{6}\right) \\ & + \frac{1}{24}\left(\frac{\pi}{6}\right)^4\left(\sin\frac{\pi}{6}\right) + \frac{1}{120}\left(\frac{\pi}{6}\right)^5\left(\cos\frac{\pi}{6}\right) + \frac{1}{720}\left(\frac{\pi}{6}\right)^6\left(-\sin\frac{\pi}{6}\right) \\ & + \frac{1}{5040}\left(\frac{\pi}{6}\right)^7\left(-\cos\frac{\pi}{6}\right) \\ = & 0.5 + \frac{\pi}{12}\sqrt{3} - \frac{1}{4}\frac{\pi^2}{36} - \frac{\sqrt{3}}{12}\left(\frac{\pi}{6}\right)^3 + \frac{1}{48}\left(\frac{\pi}{6}\right)^4 + \frac{\sqrt{3}}{240}\left(\frac{\pi}{6}\right)^5 - \frac{1}{1440}\left(\frac{\pi}{6}\right)^6 \\ & - \frac{\sqrt{3}}{10080}\left(\frac{\pi}{6}\right)^7\end{aligned}$$

The different orders of approximation can now be evaluated successively. Thus, the zero-order approximation is 0.5; the first-order approximation is $0.5+\pi\sqrt{3}/12$, i.e., 0.953449841; and the second-order approximation is

$$0.5 + \frac{\pi\sqrt{3}}{12} - \frac{\pi^2}{144}$$

which simplifies to 0.884910921. Similarly, the successive approximations are evaluated and the respective absolute errors can be calculated since the exact value of $\sin(\pi/3)$ is 0.866025403. Table 1.1 gives the approximate values of $\sin(\pi/3)$ for the orders 0 to 7 as also the absolute errors in these approximations. The results show that the error decreases with an increase in the order of approximation.

Table 7.1 Taylor's Series Approximations of $f(x) = \sin x$

Order of approximation	Computed value of $\sin \pi/3$	Absolute error
0	0.5	0.366025403
1	0.953449841	0.087424438
2	0.884910921	0.018885518
3	0.864191613	0.00183379
4	0.865757474	0.000267929
5	0.86604149	0.000016087
6	0.86602718	0.000001777
7	0.866025326	0.000000077

We next demonstrate the effect of halving the interval length on any approximate value. For this, we consider the first-order approximation in the form:

$$f(x+h) = f(x) + hf'(x) + E(h) \tag{ii}$$

where $E(h)$ is the absolute error of the first-order approximation with interval h. Taking $f(x) = \sin x$ and $x = \pi/6$, we obtain

$$\sin\left(\frac{\pi}{6}+h\right)=\sin\frac{\pi}{6}+h\cos\frac{\pi}{6}+E(h) \tag{iii}$$

Putting $h=\pi/6$ in (iii), we get

$$\sin\frac{\pi}{3}=0.5+\frac{\pi\sqrt{3}}{12}+E(h)=0.953449841+E(h)$$

Since $\sin(\pi/3)=0.866025403$, the above equation gives

$$E(h)=-0.087424438$$

Now, let the interval be halved so that we now take $h=\pi/12$. Then, (iii) gives:

$$\sin\left(\frac{\pi}{6}+\frac{\pi}{12}\right)=0.5+\frac{\pi}{12}\frac{\sqrt{3}}{2}+E\left(\frac{h}{2}\right) \tag{iv}$$

where $E(h/2)$ is the absolute error with interval length $h/2$. Since

$$\sin\left(\frac{\pi}{6}+\frac{\pi}{12}\right)=\sin\frac{\pi}{4}=\frac{1}{\sqrt{2}}$$

Equation (iv) gives

$$E\left(\frac{h}{2}\right)=\frac{1}{\sqrt{2}}-0.5-\frac{\pi\sqrt{3}}{24}=-0.019618139$$

and then

$$\frac{E(h)}{E(h/2)}=4.45630633$$

In a similar way, we obtain the values

$$\frac{E(h/2)}{E(h/4)}=4.263856931$$

and

$$\frac{E(h/4)}{E(h/8)}=4.141353027$$

The h^2-order of convergence is quite revealing in the above results.

Example 7.14

The Maclaurin expansion for e^x is given by

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^{n-1}}{(n-1)!}+\frac{x^n}{n!}e^{\xi},\quad 0<\xi<x$$

We shall find n, the number of terms, such that their sum yields the value of e^x correct to 8 decimal places at $x = 1$.

Clearly, the error term (i.e. the remainder term) is $(x^n/n!)\, e^{\xi}$, so that at $\xi = x$, this gives the maximum absolute error, and hence the maximum relative error is $x^n/n\,!$ For an 8 decimal accuracy at $x = 1$, we must have

$$\frac{1}{n!} < \frac{1}{2}\left(10^{-8}\right)$$

which gives $n = 12$. Thus, we need to take 12 terms of the exponential series in order that its sum is correct to 8 decimal places.

Example 7.15

Derive the series

$$\log_e \frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right)$$

and use it to compute the value of $\log_e(1.2)$, correct to seven decimal places. If, instead, the series for $\log_e(1+x)$ is used, how many terms must be taken to obtain the same accuracy for $\log_e(1.2)$?

We have

$$\log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \cdots \tag{i}$$

and

$$\log_e(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \frac{x^5}{5} - \cdots \tag{ii}$$

Therefore,

$$\log_e \frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right) \tag{iii}$$

Putting $x = \frac{1}{11}$ in Eq. (iii), we obtain

$$\begin{aligned}\log_e 1.2 &= 2\left[\frac{1}{11} + \frac{1}{3(11)^3} + \frac{1}{5(11)^5} + \cdots\right] \\ &= 2\left[\frac{1}{11} + \frac{1}{3(11)^3} + \frac{1}{5(11)^5}\right], \text{ since } \frac{1}{7(11)^7} = 7.33 \times 10^{-9}.\end{aligned}$$

Hence we obtain

$$\begin{aligned}\log_e 1.2 &= 2[0.09090909 + 0.00025044 + 0.00000124] \\ &= 0.1823216\end{aligned}$$

On the other hand, if we use series (i), we have

$$\begin{aligned} & \left|\frac{x^n}{n}\right| < 2 \times 10^{-7} \\ \Rightarrow \quad & \frac{(1.2)^n}{n} < 2 \times 10^{-7} \\ \Rightarrow \quad & n > 9. \end{aligned}$$

Thus, 9 terms of the series (i) have to be taken in order to obtain a seven decimal accuracy.

8 Monte Carlo Method

8.1 General idea of the method

The Monte Carlo method is a numerical method for solving mathematical problems by modeling random variables.

8.1.1 The origin of the Monte Carlo method.

The birth date of the Monte Carlo method is generally considered to be 1949, when an article entitled "The Monte Carlo method" appeared). The creators of this method are considered to be American mathematicians J. Neumann and S. Ulam. In the Soviet Union, the first articles on the Monte Carlo method were published in 1955-1956).

Curiously, the theoretical basis of the method has been known for a long time. Moreover, some statistical tasks were sometimes calculated using random samples, that is, in fact, using the Monte Carlo method. However, before the advent of electronic computers, this method could not find any wide application, because manually modeling random variables is a very time-consuming job. Thus, the emergence of the Monte Carlo method as a very universal numerical method became possible only thanks to the advent of computers.

The very name "Monte Carlo" comes from the city of Monte Carlo in the Principality of Monaco, famous for its gambling house. The fact is that one of the simplest mechanical devices for obtaining random variables is... roulette. This will be discussed in Chapter 3. Here, perhaps, it is worth answering the frequently asked question: "Does the Monte Carlo method help you win at roulette?" No, it does not help. And he doesn't even do it.

8.1.2 Example.

In order to make it more clear what will be discussed, let's look at a very simple example. Suppose we need to calculate the area of a flat figure S. It can be a completely arbitrary shape with a curved border, defined graphically or analytically, connected or consisting of several pieces. Let this be the figure shown in Fig. 1, and assume that it is all located inside a single square.

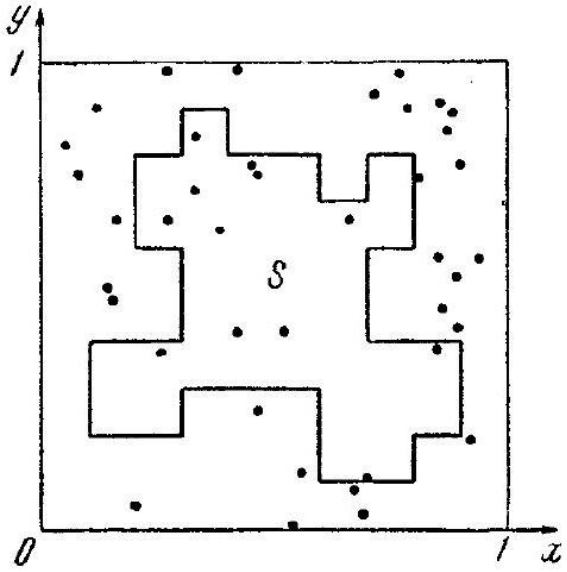


Рис. 1.

Let's choose N random points in the square. Denote by N' the number of points that fall inside S. Geometrically, it is obvious that the area of S is approximately equal to the ratio N'/N. The larger the N, the greater the accuracy of this estimate.

In the example shown in Figure 1, $N = 40$ points are selected. Of these, $N' = 12$ points ended up inside S. The ratio of $N'/N = 12/40 = 0.3$, while the true area of S is 0.35.

8.1.3 Two features of the Monte Carlo method.

The first feature of the method is the simple structure of the computational algorithm. As a rule, a program is compiled to carry out one random test (in the example from paragraph 8.1.2, you need to select a random point in the square and check whether it belongs to S). Then this test is repeated N times, and each experiment does not depend on all the others, and the results of all experiments are averaged.

Therefore, sometimes the Monte Carlo method is called the statistical test method.

The second feature of the method: the calculation error is usually proportional to $\sqrt{D/N}$, where D is some constant, and N is the number of tests. From this formula, it can be seen that in order to reduce the error by 10 times (in other words, to get another correct decimal place in the answer), you need to increase N (that is, the amount of work) by 100 times.

It is clear that it is impossible to achieve high accuracy in this way. Therefore, it is usually said that the Monte Carlo method is especially effective in solving those problems in which the result is needed with low accuracy (5-10%).

However, the same problem can be solved by different variants of the Monte Carlo method, which are answered by different values of D. In many tasks, it is possible to significantly increase accuracy by choosing a calculation method that corresponds to a significantly lower value of D.

8.1.4 Problems solved by the Monte Carlo method.

First, the Monte Carlo method allows you to simulate any process that is influenced by random factors. Secondly, for many mathematical problems that are not related to any randomness, you can artificially come up with a probabilistic model (and even more than one) that allows you to solve these problems. As a matter of fact, this was done in the example from paragraph 8.1.2.

Thus, we can talk about the Monte Carlo method as a universal method for solving mathematical problems.

It is especially interesting that in some cases it is advantageous to abandon the simulation of a true random process and instead use an artificial model.

8.1.5 More about the example.

Let's go back to the example from paragraph 8.1.2. For the calculation, we had to choose random points in a unit square. How do I do this physically?

Let's imagine such an experiment. Fig. 1 (on an enlarged scale) with the figure S and a square hung on the wall as a target. The shooter, who is at some distance from the wall, shoots N times, aiming at the center of the square. Of course, all bullets will not fall exactly in the center: they will pierce N random points on the target. Is it possible to estimate the area of S from these points?

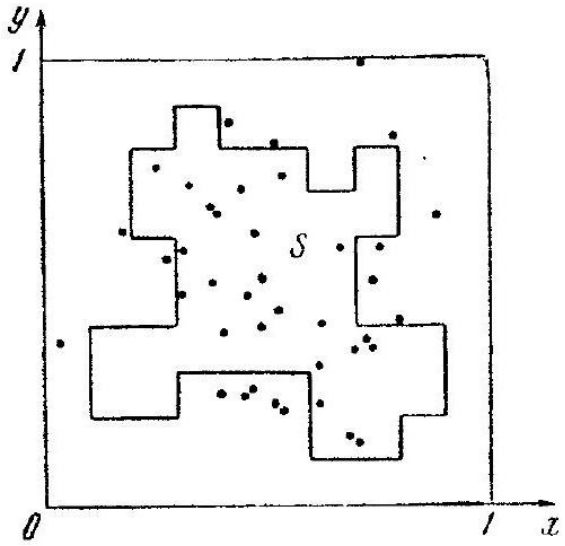


Fig. 2.

The result of such an experiment is shown in Fig. 2. In this experiment, $N = 40, N' = 24$ and the ratio $N'/N = 0.60$, which exceeds the true value of the area of (0.35) by almost two times. However, it is already clear that with a very highly qualified shooter, the result of the experience will be very bad, since almost all bullets will fall near the center and hit S.

It is not difficult to realize that our method of calculating the area will be valid only when random points are not "just random", but also "uniformly scattered" throughout the square. To give these words an accurate meaning, it is necessary to get acquainted with the definition of random variables and some of their properties. We will discuss these issues now.

MODELING OF RANDOM VARIABLES

8.2 Random variables

The words "random variable" in the ordinary sense are used when they want to emphasize that it is not known what the specific value of this value will be. Moreover, sometimes these words hide simply ignorance of what this value is.

However, the mathematician uses the same words "random variable", putting in them a very definite positive meaning. Indeed, he says, we do not know what value this value will take in this particular case, but we know what values it can take, and we know what the probabilities of certain values are. Based on these data, we cannot accurately predict the result of a single test associated with this random variable, but we can very reliably predict the totality of the results of a large number of tests. The more trials (as they say, the more statistics), the more accurate our predictions will be.

So, to set a random variable, you need to specify what values it can take and what are the probabilities of these values.

8.2.1 Discrete random variables.

A random variable ξ is called discrete if it can take a discrete set of values $(x_1, x_2, \ldots, x_n)$.

The discrete random variable ξ is defined by the table

$$\xi = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix}, \tag{T}$$

where $x_1, x_2, \ldots, x_n$ - the possible values of ξ, and $p_1, p_2, \ldots, p_n$ - are the corresponding probabilities. More precisely, the probability that the random variable ξ will take the value x_i (denote it by $P\{\xi = x_i\}$) equal to p_i :

$$P\{\xi = x_i\} = p_i.$$

The table (T) is called the random variable distribution.

The numbers $x_1, x_2, \ldots, x_n$ can be, generally speaking, any. However, the probabilities $p_1, p_2, \ldots, p_n$ must satisfy two conditions:

a) all p_i are positive:

$$p_i > 0 \tag{1}$$

b) the sum of all p_i is equal to 1 :

$$p_1 + p_2 + \ldots + p_n = 1 \tag{2}$$

The last condition means that ξ must take one of the values in each case $x_1, x_2, \ldots, x_n$.

The mathematical expectation of a random variable ξ is a number

$$\mathbf{M}_\xi = \sum_{i=1}^{n} x_i p_i \tag{3}$$

To find out the physical meaning of this value, we write it down in the following form:

$$\mathbf{M}\xi = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}$$

From this it can be seen that $\mathbf{M}\xi$ is the average value of ξ, and the more likely values of x_i are included in the sum with large weights.

Note the basic properties of mathematical expectation: if c is some kind of non-random variable, then

$$\mathbf{M}(\xi + c) = \mathbf{M}\xi + c \tag{4}$$
$$\mathbf{M}(c\xi) = c\mathbf{M}\xi \tag{5}$$

if ξ and η are any two random variables, then

$$\mathbf{M}(\xi + \eta) = \mathbf{M}\xi + \mathbf{M}\eta \tag{6}$$

The variance of a random variable ξ is called a number

$$\mathbf{D}\xi = \mathbf{M}\left[(\xi - \mathbf{M}\xi)^2\right] \tag{7}$$

Therefore, the variance of $\mathbf{D}\xi$ is the mathematical expectation of the square of the deviation of a random variable ξ from its average value $\mathbf{M}\xi$. Obviously, always $\mathbf{D}\xi > 0$.

Mathematical expectation and variance are the most important numerical characteristics of the random variable ξ. What is their practical significance?

If we observe the value of ξ many times and get the values $\xi_1, \xi_2, \ldots, \xi_N$ (each of which will be equal to one of the numbers $x_1, x_2, \ldots, x_n$), then the arithmetic mean of these values will be close to $\mathbf{M}\xi$

$$\frac{1}{N}\left(\xi_1 + \xi_2 + \ldots + \xi_N\right) \approx \mathbf{M}\xi \tag{8}$$

A variance of $\mathbf{D}\xi$ characterizes the spread of these values around the average $\mathbf{M}\xi$.

Formula (7) for the variance can be transformed using formulas (4) - (6):
$\mathbf{D}\xi = \mathbf{M}\left[\xi^2 - 2\mathbf{M}\xi \cdot \dot{\xi} + (\mathbf{M}\xi)^2\right] = \mathbf{M}\left(\xi^2\right) - 2\mathbf{M}\xi \cdot \mathbf{M}\xi + (\mathbf{M}\xi)^2$, where from

$$\mathbf{D}\xi = \mathbf{M}\left(\xi^2\right) - (\mathbf{M}\xi)^2 \tag{9}$$

Calculating the variance using formula (9) is usually easier than using formula (7).

Note the main properties of the variance: if c is some kind of non-random variable, then

$$\mathbf{D}(\xi + c) = \mathbf{D}\xi \tag{10}$$
$$\mathbf{D}(c\xi) = c^2\mathbf{D}\xi. \tag{11}$$

The concept of independence of random variables plays an important role in probability theory. In fact, this is a rather complex concept, but in the simplest cases it is very obvious.

The following relations are valid for independent random variables ξ and η:

$$\mathbf{M}(\xi\eta) = \mathbf{M}\xi \cdot \mathbf{M}\eta \tag{12}$$
$$\mathbf{D}(\xi + \eta) = \mathbf{D}\xi + \mathbf{D}\eta. \tag{13}$$

Example. Consider a random variable x with a distribution

$$x = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}$$

Obviously, the realization of this value can be considered the number of points that fall on the dice: any value is equally likely. Let's calculate the mathematical expectation and the variance of x. According to the formula (3)

$$\mathbf{M}x = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3,5$$

According to the formula (9)

$$\mathbf{D}x = \mathbf{M}\left(x^2\right) - (\mathbf{M}x)^2 = \frac{1}{6}\left(1^2 + 2^2 + \ldots + 6^2\right) - (3,5)^2 = 2,917$$

Example.
Consider a random variable θ with a distribution

$$\theta = \begin{pmatrix} 3 & 4 \\ 1/2 & 1/2 \end{pmatrix}.$$

The realization of this value can be considered a game of eagle with the condition that, for example, an eagle is 3 points and a grid is 4 points. In this case

$$\mathbf{M}\theta = 0,5 \cdot 3 + 0,5 \cdot 4 = 3,5;$$
$$\mathbf{D}\theta = 0,5\left(3^2 + 4^2\right) - (3,5)^2 = 0,25.$$

We see that $\mathbf{M}\theta = \mathbf{M}x$, but $\mathbf{D}\theta < \mathbf{D}x$. This could have been easily foreseen, since the values of θ can differ from 3.5 only by ± 0.5, while in the values of x the spread can reach $\pm 2,5$.

8.2.2 Continuous random variables.

Suppose that there is a certain amount of radium on the plane at the origin. When each radium atom decays, a α particle flies out of it. Its direction will be characterized by the angle ψ (Fig. 3). Since theoretically and practically any departure directions are possible, this random variable can take any value from 0 to 2π.

We will call a random variable ξ continuous if it can take any value from a certain interval (a, b).

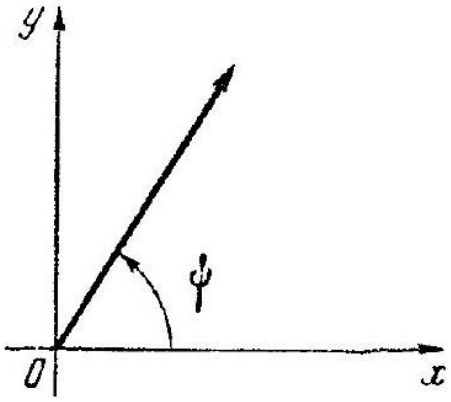


Fig. 3.

A continuous random variable ξ is defined by specifying an interval (a, b) containing the possible values of this value and a function $p(x)$, which is called the probability density of the random variable ξ (or the density of the distribution ξ).

The physical meaning of $p(x)$ follows: let (a', b') be an arbitrary interval contained in (a, b) (that is, $a \leqslant a', b' \leqslant b$). Then the probability that ξ will be in the range (a', b') is equal to the integral

$$P\{a' < \xi < b'\} = \int_{a'}^{b'} p(x)dx \tag{14}$$

In Fig. 4, the shaded area is equal to the value of the integral (14).

The set of values of ξ can be any interval. Even the case of $a = -\infty$ is possible, as well as $b = \infty$.

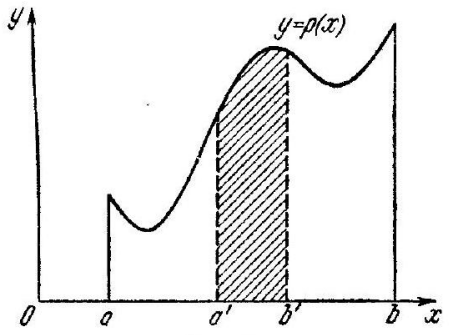


Рис. 4.

However, the density of $p(x)$ must satisfy two conditions similar to conditions (1) and (2) for discrete quantities:

a) the density of $p(x)$ is positive:

$$p(x) > 0 \tag{15}$$

b) the integral of the density $p(x)$ over the entire interval (a, b) is equal to 1:

$$\int_a^b p(x)dx = 1 \tag{16}$$

The mathematical expectation of a continuous random variable is called a number

$$\mathbf{M}\xi = \int_a^b xp(x)dx \tag{17}$$

The meaning of this characteristic is the same as in the case of a discrete random variable. In fact, since

$$\mathbf{M}\xi = \frac{\int_a^b xp(x)dx}{\int_a^b p(x)dx},$$

then it is easy to see that this is the average value of ξ: after all, the value of ξ can be any number x from the interval (a, b), which is included in the integral with weight p(x).

Everything stated in clause 2.1 from formula (4) to formula (13) inclusive is also valid for continuous random variables: both the definition of variance (7), and formula (9) for calculating it, and all the properties of $\mathbf{M}\xi$ and $\mathbf{D}\xi$. We will not repeat them.

We will only indicate one more formula for the mathematical expectation of a random function. Let the random variable ξ still have a probability density of $p(x)$. Let's choose an arbitrary continuous function $f(x)$ and consider the random variable $\eta = f(\xi)$, which is sometimes called a random function. It can be proved that

$$\mathbf{M}f(\xi) = \int_a^b f(x)p(x)dx \tag{18}$$

We emphasize that, generally speaking, $\mathbf{M}f(\xi) \neq f(\mathbf{M}\xi)$.

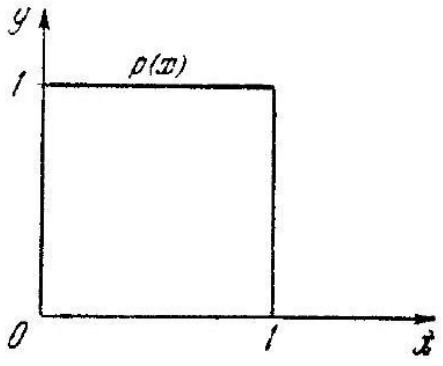


Fig. 5.

The random variable γ, defined in the range $(0,1)$ and having a density $p(x) = 1$, is called uniformly distributed in (0, 1) (Fig. 5).

In fact, whatever interval (a', b') inside $(0,1)$ we take, the probability that γ will fall into (a', b'), equal to

$$\int_{a'}^{b'} p(x)dx = b' - a'$$

that is, the length of this interval. In particular, if we divide $(0,1)$ into any number of intervals of equal length, then the probability of hitting γ in any of these intervals will be the same.

It is easy to calculate that

$$\mathbf{M}\gamma = \int_0^1 xp(x)dx = \int_0^1 xdx = \frac{1}{2}$$
$$\mathbf{D}\gamma = \int_0^1 x^2 p(x)dx - (\mathbf{M}\gamma)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

We will meet with the random variable γ more than once in the future.

8.2.3 Normal random variables.

A normal (or Gaussian) random variable is a random variable ζ defined on the entire axis $(-\infty, \infty)$ and having a density of

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}} \tag{19}$$

where a and $\sigma > 0$ are numeric parameters.

The parameter a does not affect the shape of the curve $p(x)$: changing it only leads to a shift of the curve along the x axis. However, when σ changes, the shape of the curve changes. Indeed, it is easy to see that

$$\max p(x) = p(a) = \frac{1}{\sigma\sqrt{2\pi}}$$

If you decrease σ, then $\max p(x)$ will increase. However, the entire area under the curve $p(x)$ is equal to 1 by condition (16). Therefore, the curve will stretch upwards in the vicinity of $x = a$, but decrease for all sufficiently large values of x. Figure 6 shows two normal densities corresponding to $a = 0, \sigma = 1$ and $a = 0, \sigma = 0.5$. (Another normal density is shown in Fig. 21)

It can be proved that

$$\mathbf{M}\zeta = a, \quad \mathbf{D}\zeta = \sigma^2.$$

Normal random variables are very common in the study of a wide variety of issues by their nature. The reason for this will be discussed below. For example, the measurement error δ is usually a normal random variable.

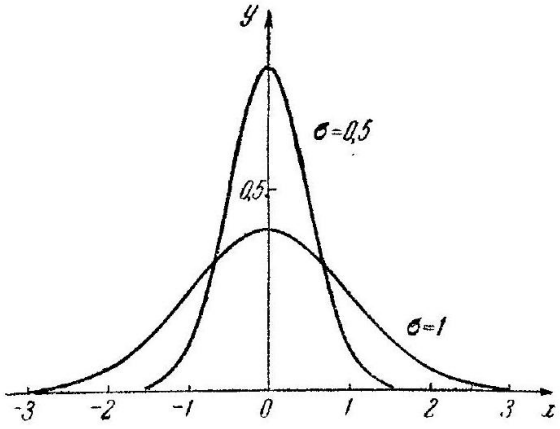


Fig. 6.

If there is no systematic measurement error, then $a = \mathbf{M}\delta = 0$. And the value $\sigma = \sqrt{\mathbf{D}\delta}$, called the mean quadratic error, characterizes the error of the measurement method.

The three sigma rule. It is not difficult to calculate that, whatever a and σ are in (19),

$$\int_{a-3\sigma}^{a+3\sigma} p(x)dx = 0,997.$$

It follows from (14) that

$$P\{a - 3\sigma < \zeta < a + 3\sigma\} = 0,997 \tag{20}$$

The probability of 0.997 is so close to 1 that sometimes the last formula is interpreted as follows: in one test, it is almost impossible to obtain a value of ζ that differs from $\mathbf{M}\zeta$ by more than 3σ.

8.2.4 The central limit theorem of probability theory.

This remarkable theorem was first formulated by Laplace. Many outstanding mathematicians, including P. L. Chebyshev, A. A. Markov and A.M. Lyapunov, were engaged in generalizing this theorem. Proving it is quite difficult.

Consider N identical independent random variables $\xi_1, \xi_2, \ldots, \xi_N$. In other words, the probability densities of these quantities coincide. Consequently, their mathematical expectations and variances are also the same.

Let's denote

$$\mathbf{M}\xi_1 = \mathbf{M}\xi_2 = \ldots = \mathbf{M}\xi_N = m$$
$$\mathbf{D}\xi_1 = \mathbf{D}\xi_2 = \ldots = \mathbf{D}\xi_N = b^2.$$

Denote by ρ_N the sum of all these quantities:

$$\rho_N = \xi_1 + \xi_2 + \ldots + \xi_N$$

It follows from formulas (6) and (13) that

$$\mathbf{M}\rho_N = \mathbf{M}\left(\xi_1 + \xi_2 + \ldots + \xi_N\right) = Nm$$
$$\mathbf{D}\rho_N = \mathbf{D}\left(\xi_1 + \xi_2 + \ldots + \xi_N\right) = Nb^2.$$

Now consider a normal random variable ζ_{N} with the same parameters: $a = Nm, \sigma^2 = Nb^2$.

The theorem. The density of the sum ρ_N approaches the density of the normal value ζ_N when $N \to \infty$:

$$\lim_{N\to\infty} |p_{\rho_N}(x) - p_{\zeta_N}(x)| = 0$$

The physical meaning of this theorem is obvious: the sum of ρ_{N} of a large number of identical random variables is approximately normal ($p_{\rho_N}(x) \approx p_{\zeta_N}(x)$).

In fact, this theorem holds true under much broader conditions: all terms $\xi_{\mathrm{I}}, \xi_2, \ldots, \xi_N$ do not have to be the same and independent; it is only essential that individual terms do not play too big role in the sum.

It is this theorem that explains why normal random variables are so common in nature. Indeed, every time we encounter the combined effects of a large number of insignificant random factors, the resulting random variable turns out to be normal.

For example, the deviation of an artillery shell from the target almost always turns out to be a normal random variable, since it depends on meteorological conditions in different parts of the trajectory, and on many other factors.

8.2.5 The general scheme of the Monte Carlo method.

Let's say that we need to calculate some unknown value m. Let's try to come up with such a random variable ξ that $\mathbf{M}\xi = m$. Let at the same time $\mathbf{D}\xi = b^2$.

Consider N random variables $\xi_1, \xi_2, \ldots, \xi_N$ whose distributions coincide with the distribution of ξ. If N is large enough, then according to the theorem from clause 2.4, the distribution of the sum $\rho_N = \xi_1 + \xi_2 + \ldots + \xi_N$ will be approximately normal with the parameters $a = Nm, \sigma^2 = Nb^2$, It follows from (20) that

$$P\left\{Nm - 3b\sqrt{N} < \rho_N < Nm + 3b\sqrt{N}\right\} \approx 0,997$$

If we divide the inequality in the curly bracket by N, we get an equivalent inequality and its probability remains the same:

$$P\left\{m - \frac{3b}{\sqrt{N}} < \frac{\rho_N}{N} < m + \frac{3b}{\sqrt{N}}\right\} \approx 0,997.$$

We will rewrite the last ratio in a slightly different form:

$$P\left\{\left|\frac{1}{N}\sum_{j=1}^{N}\xi_j - m\right| < \frac{3b}{\sqrt{N}}\right\} \approx 0,997 \tag{21}$$

This is an extremely important relation for the Monte Carlo method. It gives us both the method of calculating m and the error estimate.

In fact, we will find N values of the random variable ξ. It can be seen from (21) that the average arithmetic value of these values will be approximately equal to m. With a high probability, the error of such an approximation does not exceed the value of $3b/\sqrt{N}$. Obviously, this error tends to zero with growth N.

8.3 Getting random variables on a computer

The very formulation of the question "obtaining random variables on a computer" sometimes causes confusion: after all, everything that a machine does must be pre-programmed; where can randomness come from?

There are indeed some difficulties in this matter, but they relate more to philosophy, so we will not dwell on them.

Just in case, let's just note that the random hams discussed in section 2 are ideal mathematical concepts. The question of whether it is possible to describe any natural phenomenon with their help is solved by experience. Such a description is always approximate. Moreover, a random variable that quite satisfactorily describes some physical quantity in one range of phenomena may turn out to be a poor characteristic of the same quantity in the study of other phenomena.

In the same way, a road that can be considered straight on the map of the country (an ideal mathematical straight line "without width") becomes a strip with bends on a large-scale plan of a settlement...

There are usually three ways to obtain random variables: tables of random numbers, random number generators, and the pseudorandom number method.

8.3.1 Tables of random numbers.

Let's do the following experiment. Let's write the numbers $0, 1, 2, \ldots, 9$ on ten identical pieces of paper. Let's put these pieces of paper in the cap, mix them up and take out one

piece of paper from there, returning it back each time and mixing all the pieces of paper again. The figures obtained in this way will be written in the form of a table.

Such a table is called a table of random numbers, although it would be more correct to call it a table of random numbers. You can enter it into the computer's storage device. And during the calculation, when we need the value of a random variable with a distribution

$$\begin{pmatrix} 0 & 1 & 2 & \dots & 9 \\ 0,1 & 0,1 & 0,1 & \dots & 0,1 \end{pmatrix} \tag{22}$$

we will take the next figure from this table.

The largest published random number table contains 1,000,000 digits. Of course, it was compiled with the help of more modern technology than the hat: a special roulette was designed using electronics. The simplest scheme of such a roulette is shown in Fig. 7 (the rotating disk stops abruptly and the number indicated by the fixed arrow is selected).

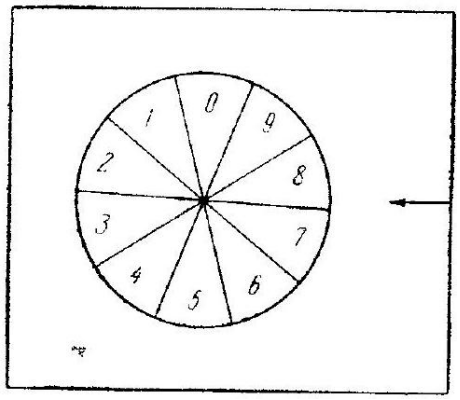


Fig. 7.

It should be noted that making a good table of random numbers is not as easy as it might seem. Any real physical device generates random variables with a distribution slightly different from the ideal distribution (22). In addition, mistakes may occur during the experiment (for example, one of the pieces of paper may stick to the lining for a while). Therefore, the compiled tables are carefully checked using special statistical tests: whether certain properties of a group of numbers contradict the hypothesis that these numbers are values of a random variable (22).

Here is one of the simplest tests. Consider a table containing N digits. Let's say the number of zeros in this table is v_0, the number of ones is v_1, the number of twos is v_2, etc. Calculate the amount

$$\sum_{i=0}^{9} (v_i - 0,1N)^2$$

Probability theory allows you to predict within what limits this amount may lie: it should not be too large (since the mathematical expectation of each of v_i is $0.1N$), but it should not be too small (since this would mean a "too regular" distribution of values).

Random number tables are used only for manual Monte Carlo calculations.

8.3.2 Random number generators.

It would seem that the roulette mentioned in 3.1 can be connected to a computer and generate random numbers as needed. However, any mechanical device will be too slow for a computer. Therefore, noise in electronic lamps is most often used as random variable generators: if for some fixed period of time Δt the noise level exceeds a given threshold an even number of times, then zero is recorded, and if an odd number of times, then one is recorded.

At first glance, this is a very convenient way. Let m of such generators work in parallel, work all the time and send random zeros and ones to all binary digits of a special cell. Each clock cycle is one m-bit number. At any time of the account, you can access this cell and take from there the value of the random variable γ, evenly distributed in the interval (0,1). Of course, the value is approximate, written in the form of m-bit binary fraction $0, \alpha_{(1)}\alpha_{(2)} \ldots \ldots \alpha_{(m)}$, where each of the values is $\alpha_{(i)}$ simulates a random variable with a distribution

$$\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$$

However, this method is not free from disadvantages either. Firstly, it is difficult to check the "quality" of the numbers produced. Checks have to be done periodically, since due to any malfunctions, the so-called "distribution drift" may occur (that is, zeros and ones in any of the digits will not appear equally often). Secondly, usually all calculations on a computer are performed twice to eliminate the possibility of accidental failure. But it is impossible to reproduce the same random numbers if you do not remember them in the course of counting, And if you remember them, then we come back to the case of tables.

8.3.3 Pseudo-random numbers.

Since the "quality" of the random numbers used is checked using special tests, you don't have to be interested in how these numbers are obtained: as long as they satisfy the accepted test system. You can even try to calculate them using a given formula. But, of course, it must be a very tricky formula.

Numbers obtained by some formula and simulating the values of a random variable γ are called pseudo-random numbers. The word "simulating" means that these numbers satisfy a number of tests as if they were the values of this random variable.

The first algorithm for obtaining pseudorandom numbers was proposed by J. Neiman. It is called the mid-squares method. Let's explain it with an example.

Let's give a 4-digit integer $n_1 = 9876$. Let's square it. We get, generally speaking, an 8-digit number $n_1^2 = 97535376$. Let's choose four middle digits from this number and denote $n_2 = 5353$.

Then we square n_2 ($n_2^2 = 28654609$) and again, we will extract 4 average numbers. We will get $n_3 = 6546$.

Further, $n_3^2 = 42850116$, $n_4 = 8501$; $n_4^2 = 72267001$, $n_5 = 2670$; $n_5^2 = 07128900$, $n_6 = 1289$ . .

It was suggested to use the values of u as values $0,9876$; $0,5353$; $0,6546$; $0,8501$; $0,2670$; $0,1289$ etc.

But this algorithm did not pay off: it turned out to be more than necessary for small values. Therefore, different researchers have developed other algorithms. Some of them use the features of specific computers. As an example, let's consider one of these algorithms used on the Strela computer.

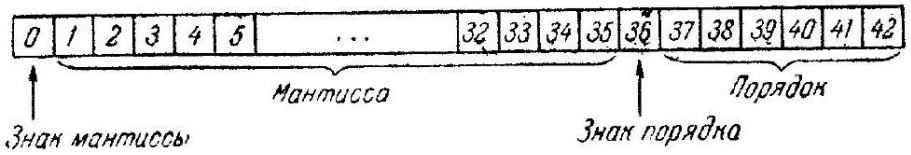


Fig. 8.

Example. The arrow is a three-address floating-point computer. The cell in which the number x is written consists of 43 binary digits (Fig. 8). The machine works with binary numbers in normalized form: $x = \pm q \cdot 2^{\pm p}$, where p is the order of the number, q is the mantissa. The jth digit of the cell may contain zero or one; let's denote this value ε_j. Then

$$q = \frac{\varepsilon_1}{2^1} + \frac{\varepsilon_2}{2^2} + \ldots + \frac{\varepsilon_{35}}{2^{35}}, \quad p = \varepsilon_{37}2^5 + \varepsilon_{38}2^4 + \ldots + \varepsilon_{42}2^0$$

The normalization condition $0.5 \leqslant q < 1$ means that ε_1 is always equal to 1. The "+" sign is represented by zero, the "-" sign is represented by one.

The number γ_{k+1} is obtained from the number γ_k by three operations:

1. multiply γ_k by a large constant, usually 10^{17};
2. the image of the product $10^{17}\gamma_k$ is shifted 7 digits to the left (so that the first 7 digits of the product disappear, and zeros appear in the digits from the 36th to the 42nd);
3. we take the absolute value of the resulting number (in this case, the number is normalized); this will be γ_{k+1}.

If you start with $\gamma_0 = 1$, then this process produces more than 80,000 different numbers γ_k, then a period occurs in the sequence, and the numbers begin to repeat. Various checks of the first 50,000 numbers gave quite satisfactory results. These numbers have been repeatedly used to solve a wide variety of problems.

The advantages of the pseudorandom number method are quite obvious. Firstly, it takes only a few simple operations to get each number, so that the speed of generating random numbers is of the same order as the speed of computer operation. Secondly, the

program occupies only a few cells of the drive, And thirdly, any of the numbers γ_k can be easily reproduced. Fourth, you only need to check the "quality" of such a sequence once, then you can safely use it many times when calculating similar tasks.

The only drawback of the method is the limited "stock" of pseudorandom numbers. However, there are ways to get much more numbers. In particular, you can change the initial numbers γ_0.

The vast majority of Monte Carlo calculations are currently performed using pseudo-random numbers.

8.4 Transformations of random variables

When solving various problems, it is necessary to simulate various random variables. In the early stages of using the Monte Carlo method, some calculators tried to build their own roulette to find each random variable. For example, to find the values of a random variable with a distribution

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ 0,5 & 0,25 & 0,125 & 0,125 \end{pmatrix}$$

you can use the roulette shown in Figure 9, which acts exactly like the roulette shown in Figure 7, but with uneven divisions proportional to p_i.

However, this turned out to be completely unnecessary: the values of any random variable can be obtained by converting the values of one (so to speak, "standard") random variable. Usually, the role of such value is played by a random variable γ, evenly distributed in $(0, 1)$. We already know how to get the values of γ.

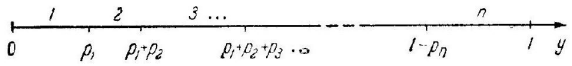


Fig. 9.

Let's agree that the process of finding the value of any random variable ξ by converting one or more values of γ is called playing a random variable ξ.

8.4.1 Playing a discrete random variable.

Let's say that we need to get the values of a random variable ξ with a distribution

$$\xi = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

Consider the interval $0 < y < 1$ and divide it into n intervals whose lengths are $p_1, p_2, \dots, p_n$. The coordinates of the division points will obviously be $y = p_1$, $y = p_1 + p_2$, $y = p_1 + p_2 + p_3$, ..., $y = p_1 + p_2 + \dots + p_{n-1}$.

The resulting intervals are numbered with the numbers $1, 2, \ldots$ ..., n (Fig. 10). This concludes the preparations for the drawing of ξ. Every time we need to "put an experiment" and play the value of ξ, we will select the value of γ and build a point $y = \gamma$. If this point falls into the interval with the number i, then we assume that $\xi = x_i$ (in this experiment).

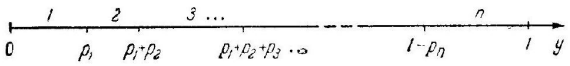


Fig. 10.

It is quite easy to prove the legality of such a procedure. Indeed, since the random variable γ is uniformly distributed in $(0, 1)$, the probability that γ will be in a certain interval is equal to the length of this interval. Means,

$$P\{0 < \gamma < p_1\} = p_1,$$
$$P\{p_1 < \gamma < p_1 + p_2\} = p_2,$$
$$P\{p_1 + p_2 + \ldots + p_{n-1} < \gamma < 1\} = p_n.$$

According to our procedure, $\xi = x_i$ when

$$p_1 + p_2 + \ldots + p_{i-1} < \gamma < p_1 + p_2 + \ldots + p_i$$

and the probability of this is p_i.

Of course, you can do without Fig. 10 on a computer. Suppose that the numbers $x_1, x_2, \ldots, x_n$ are located in the storage cells in a row, and the probabilities are $p_1, p_1 + p_2, p_1 + p_2 + : + p_3, \ldots, 1$-also in a row.

Example. Play 10 values of a random variable θ with a distribution

$$\theta = \begin{pmatrix} 3 & 4 \\ 0,58 & 0,42 \end{pmatrix}$$

As the values of γ, we select pairs of digits from the table multiplied by $0,01$. So, $\gamma = 0,86; 0,51; 0,59; 0,07; 0,95;\ 0,66; 0,15; 0,56; 0,64; 0,34$.

Obviously, according to our scheme, the values of γ less than 0.58 correspond to the values of $\theta = 3$, and the values of $\gamma \geqslant 0.58$ correspond to the values of $\theta = 4$. Therefore, we get the values: $\theta = 4; 3; 4; 3; 4; 4; 3; 3; 4; 3$.

Note that the numbering order of the values $x_1, x_2, \ldots\ldots, x_n$ in the distribution of ξ can be arbitrary, but it must be fixed before the start of the draw.

8.4.2 Playing a continuous random variable.

Now let's assume that we need to get the values of a random variable ξ distributed in the range (a, b) with a density of $p(x)$.

Let's prove that the values of ξ can be found from the equation

$$\int_a^{\xi} p(x)dx = \gamma \tag{23}$$

that is, by choosing the next value of γ, you need to solve equation (23) and find the next value of ξ.

To prove this, consider the function (Fig. 12)

$$y = \int_a^x p(x)dx$$

It follows from the general properties of density (15) and (16) that

$$y(a) = 0, \quad y(b) = 1$$

and the derivative

$$y'(x) = p(x) > 0$$

This means that the function $y(x)$ increases monotonously from 0 to 1. And any line $y = \gamma$, where $0 < \gamma < 1$, intersects the graph $y = y(x)$ at a single point, the abscissa value of which we take for ξ. Thus, equation (23) always has one and only one solution.

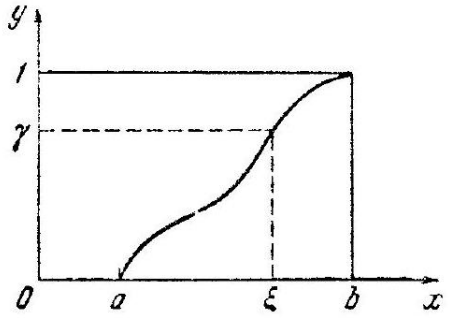


Fig. 12.

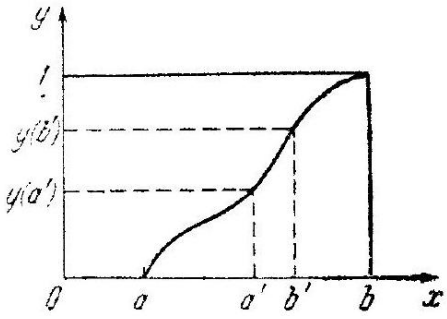


Fig. 13.

Now let's choose an arbitrary interval (a', b') contained inside (a, b). The points of this interval

$$a' < x < b'$$

the ordinates of the curve $y = y(x)$ satisfying the inequality correspond

$$y\left(a'\right) < y < y\left(b'\right)$$

Therefore, if ξ belongs to the interval $a' < x < b'$, then γ belongs to the interval $y\left(a'\right) < y < y\left(b'\right)$, and vice versa (Fig. 13). Means,

$$P\left\{a' < \xi < b'\right\} = P\left\{y\left(a'\right) < \gamma < y\left(b'\right)\right\}$$

Since γ is evenly distributed in $(0, 1)$, then

$$P\left\{y\left(a'\right) < y < y\left(b'\right)\right] = y\left(b'\right) - y\left(a'\right) = \int_{a'}^{b'} p(x)dx$$

So,

$$P\left\{a' < \S < b'\right\} = \int_{a'}^{b'} p(x)dx,$$

and this just means that the random variable ξ, which is the root of equation (23), has a probability density of $p(x)$.

Example. A random variable η is called uniformly distributed in the interval (a, b) if its density is constant in this interval:

$$p(x) = \frac{1}{b-a} \quad \text{for} \quad a < x < b.$$

To play the values of $\boldsymbol{\eta}$, let's make up the equation (23):

$$\int_a^1 \frac{dx}{b-a} = \gamma$$

The integral is easily calculated:

$$\frac{\eta - a}{b-a} = \gamma$$

This gives an explicit expression for η :

$$\eta = a + \gamma(b-a). \tag{24}$$

8.4.3 The Neumann method for playing a continuous random variable.

It may turn out that it is very difficult to solve equation (23) with respect to ξ. For example, in the case when the integral of $p(x)$ is not expressed in terms of elementary functions or when the density of $p(x)$ is given graphically.

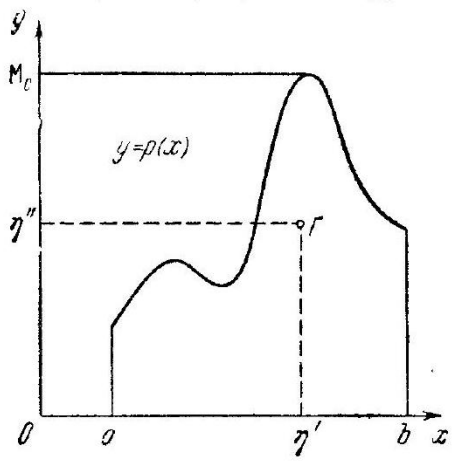


Fig. 14.

Suppose that the random variable ξ is defined on a finite interval (a, b) and its density is limited (Fig. 14):

$$p(x) \leqslant M_0$$

The value of ξ can be played as follows:

1. select two values γ' and γ'' of a random variable γ and construct a random point $\Gamma\,(\eta'; \eta'')$ with coordinates

$$\eta' = a + \gamma'(b - a), \quad \eta'' = \gamma'' M_0$$

2. if the point Γ lies under the curve $y = p(x)$, then we assume $\xi = \eta'$, if the point Γ lies above the curve $y = p(x)$, then the pair (γ', γ'') discard and select a new pair of values (γ', γ'').

8.4.4 About playing normal quantities.

There are many very diverse techniques for playing various random variables. We will not dwell on them here. They are usually used only when the techniques of paragraphs 4.2 and 4.3 are ineffective.

In particular, such a case holds for the normal random variable ζ, since the equation

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\zeta} e^{-\frac{x^2}{2}} dx = \gamma$$

it is explicitly undecidable, and the range of possible values of ζ is infinite.

The tables contain the values (already played) of the normal random variable ζ with the mathematical expectation $\mathbf{M}_\zeta = 0$ and with the variance $\mathbf{D}\zeta = 1$. It is not difficult to prove that a random variable

$$\zeta' = a + \sigma\zeta \tag{25}$$

it will also be normal, and it follows from (10) and (11) that

$$\mathbf{M}\xi' = a, \quad \mathbf{D}\xi' = \sigma^2$$

Thus, formula (25) allows using the table to play out the values of any normal values.

8.4.5 Again about the example from paragraph 8.1.2.

Now we can explain how random points were selected in Fig. 1 and 2. In Fig. 1, points with coordinates are plotted

$$x = \gamma', \quad y = \gamma''$$

The values of γ' and γ'' were calculated using five digits from the table A : $x_1 = 0,86515$; $y_1 = 0,90795$; $x_2 = 0,66155$; $y_2 = 0,66434$ etc.

It can be proved that since the abcissa and ordinates of our points are independent, the probability of such a point hitting any area inside the square is equal to the area of this area. In other words, the points are evenly distributed in a square.

Figure 2 shows the points with coordinates

$$x = 0,5 + 0,2\zeta', \quad y = 0,5 + 0,2\zeta''$$

moreover, the values of ζ' and ζ'' were selected from table B in a row:

$$x_1 = 0,5 + 0,2 \cdot 0,2005, \quad y_1 = 0,5 + 0,2 \cdot 1,1922$$
$$x_2 = 0,5 + 0,2(-0,0077), \ldots$$

One of the points outside the square is discarded.

It follows from formula (25) that the abscissae and ordinates of these points are normal random variables with averages $a = 0.5$ and with variances $\sigma^2 = 0.04$.

EXAMPLES OF USING THE MONTE CARLO METHOD

8.5 Calculation of the queuing system

8.5.1 Description of the task.

Consider one of the simplest queuing systems. This system consists of n lines (or channels, or service points), each of which can "serve consumers". The system receives applications, and the moments of their receipt are random. Each request is sent to line number 1. If at the time of receipt of the kth application (let's call it T_k) this line is free, then it starts servicing the application, which lasts t_3 minutes (t_3 is the busy time of the line). If line number 1 is occupied at the moment T_k, the request is instantly transferred to line number 2. Etc...

Finally, if all n lines are busy at the moment T_k, then the system issues a failure.

Is it necessary to determine how many (on average) applications the system will serve during T and how many refusals it will give?

It is clear that tasks of this type are encountered in the study of the organization of work of any enterprises, and not only consumer service enterprises. In some very special cases, it is possible to find analytical solutions. However, in difficult cases (they will be discussed below), the Monte Carlo method turns out to be the only calculation method.

8.5.2 The simplest flow of applications.

The first question that arises when considering such a system is: what is the flow of incoming applications? This issue is solved by experience, through sufficiently long-term monitoring of applications. The study of application flows in various conditions allowed us to identify some fairly common cases.

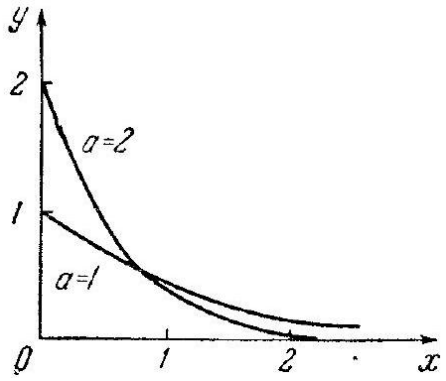


Fig. 15.

The simplest flow (or Poisson flow) is called such a flow of applications when the time interval τ between two consecutive applications is a random variable distributed in the interval $(0, \infty)$ with density

$$p(x) = ae^{-ax} \tag{26}$$

The law (26) is also called the exponential distribution (Fig. 15). It is easy to calculate the mathematical expectation of τ:

$$\mathbf{M}\tau = \int_0^\infty xp(x)dx = \int_0^\infty xae^{-ax}dx$$

After integration in parts ($u = x, dv = ae^{-Ax}dx$) we get

$$\mathbf{M}\tau = \left[-xe^{-ax}\right]_0^\infty + \int_0^\infty e^{-ax}dx = \left[-\frac{e^{-ax}}{a}\right]_0^\infty = \frac{1}{a}$$

Parameter a is called the application flow density.

The formula for drawing τ is easily obtained from equation (23), which in our case will be written as follows:

$$\int_0^\tau ae^{-ax}dx = \gamma$$

Having calculated the integral on the left, we obtain the ratio

$$1 - e^{-a\tau} = \gamma$$

where from

$$\tau = -\frac{1}{a}\ln(1-\gamma)$$

However, the value of $1 - \gamma$ is distributed in exactly the same way as γ, and therefore you can use the formula instead of the last formula

$$\tau = -\frac{1}{a}\ln\gamma. \tag{27}$$

8.5.3 The calculation scheme.

So, let's consider the operation of the system from clause 8.5.1 in the case of the simplest flow of applications.

We will match each line with a cell of the internal storage of the computer, into which we will record the moment when this line is released. Let's denote the moment of the release of the ith line by t_i. For the initial moment of calculation, we will select the moment of receipt of the first application $T_1 = 0$. We can assume that at this moment all t_i are equal to T_1 : all lines are free. The end time of the calculation is $T_{\text{fin}} = T_1 + T$.

The first request is received on line number 1. This means that this line will be occupied during t_3. Therefore, we must replace t_1 with a new value $(t_1)_{\text{new}} = T_1 + t_3$, add one to the counter of completed applications and proceed to the consideration of the second application.

Let's assume that k applications have already been reviewed. Then you need to play the moment of receipt of $(k+1)The$th application. To do this, select the next value of γ and use formula (27) to calculate the next value of $\tau = \tau_k$. And then we calculate the moment of receipt

$$T_{k+1} = T_k + \tau_k$$

Is the first line available at this moment? To establish this, you need to check the condition

$$t_1 \leqslant T_{k+1}. \tag{28}$$

If this condition is met, it means that by the time T_{i+1} the line has already been released and can serve this request. We have to replace t_1 with $T_{h+1} + t_3$, add one to the counter of completed applications and proceed to the next application.

If condition (28) is not met, it means that the first line is occupied at the moment T_{k+1}. Then we check if the second line is free:

$$t_2 \leqslant T_{k+1}? \tag{29}$$

If condition (29) is met, then we replace t_2 with $T_{k+1} + t_3$, add one to the counter of completed applications and proceed to the next application.

If condition (29) is not met, then proceed to checking the condition

$$t_3 \leqslant T_{k+1}$$

It may turn out that for all i from 1 to n

$$t_i > T_{k+1}$$

that is, all lines are occupied at the moment T_{h+1}. Then you need to add one to the bounce counter and then proceed to consider the next application.

Every time, having calculated T_{h+1}, it is necessary to check the end condition of the experiment

$$T_{k+1} > T_{\text{fin}}$$

When this condition is fulfilled, the experience ends. The counter of completed applications and the bounce counter will contain the numbers μ_{done} n μ_{canc} .

This experience is repeated N times (using different methods). And the results of all experiments are averaged:

$$\mathbf{M}\mu_{\text{done}} \approx \frac{1}{N}\sum_{j=1}^{N}\mu_{\text{done }(j)},$$

$$\mathbf{M}\mu_{\text{canc}} \approx \frac{1}{N}\sum_{j=1}^{N}\mu_{\text{canc }(j)}$$

where $\mu_{\text{done }(j)}$ and $\mu_{\text{canc }(j)}$ - values μ_{done} and μ_{canc}, obtained in the ith experience.

8.5.4 More complex tasks.

It is very easy to show that the same method allows you to calculate incomparably more complex systems. For example, the value of t_3 may not be constant, but random and different for different lines (which corresponds to different equipment or different qualifications of maintenance personnel). The calculation scheme will remain basically the same, but the values of t_3 will have to be played each time, and the playing formula for each line will be different.

We can consider the so-called waiting systems, in which a refusal is not issued immediately: the application is stored for some time t_{n} (the time the application stays in the system), and if during this time any line is released, it will serve this application.

We can consider systems in which the next application is accepted by the line that was released first. You can take into account the accidental failure of individual lines and the random repair time of each of them. It is possible to take into account changes in the density of the flow of applications over time. And much, much more.

Of course, calculations of such systems are not given in vain. To get results that have practical value, you need to choose a good model. To do this, you have to carefully study the actual application flows, time the work of individual nodes, etc.

In general, it is necessary to know the probabilistic laws of the functioning of individual parts of the system. Then the Monte Carlo method allows you to calculate the probabilistic laws of the entire system, no matter how complex it may be.

Such calculation methods are extremely useful in enterprise planning: instead of an expensive (and sometimes simply impossible) experiment in kind, we can experiment on a computer, modeling different options for organizing work or using equipment.

8.6 Calculation of product quality and reliability

8.6.1 The simplest quality calculation scheme.

Consider a product S consisting of some (maybe a large) number of elements. For example, if S is an electrical device, then its elements can be resistances $\left(R_{(k)}\right)$, capacitances $\left(C_{(k)}\right)$, lamps, etc. Suppose that the quality of the product is determined by the value one output parameter U, which can be calculated by knowing the parameters of all elements

$$U = f\left(R_{(1)}, R_{(2)}, \ldots; C_{(1)}, C_{(2)}, \ldots; \ldots\right) \tag{30}$$

If, for example, U is the voltage at the working section of an electric circuit, then according to Ohm's laws, you can make equations for the circuit and, solving them, find U.

In reality, however, the parameters of the elements are not exactly equal to the specified values. For example, the resistance can be anything in the range from 20.9 to 23.1 K

The question arises: how will the deviations of the parameters of all elements from the nominal ones affect the value of U?

You can try to estimate the limits of the U change by selecting the "worst" parameter values for all elements. However, it is not always known which set of parameters will be the "worst". In addition, if the number of elements is large, then such an estimate will be greatly overestimated: in fact, it is unlikely that all the parameters at the same time turn out to be the worst.

Therefore, it is more reasonable to consider the parameters of all elements and the value U itself as random variables and try to estimate the mathematical expectation $\mathbf{M}U$ and the variance $\mathbf{D}U$. The value of $\mathbf{M}U-$ is the average value of U for the entire batch of products, and the value of $\mathbf{D}U$ shows what deviations U from $\mathbf{M}U$ will occur in practice.

Recall (this was stated in paragraph 2.2) that

$$\mathbf{M}U \neq f\left(\mathbf{M}R_{(1)}, \mathbf{M}R_{(2)}, \ldots; \mathbf{M}C_{(1)}, \mathbf{M}C_{(2)}, \ldots; \ldots\right)$$

It is impossible to calculate analytically the distribution of U for a small complex function f. Sometimes this can be done experimentally by viewing a large batch of finished products. But this is not always possible, and never at the design stage.

Let's try to apply the Monte Carlo method. To do this, you need to: a) know the probabilistic characteristics of all elements, b) know the function f (more precisely, be able to calculate the value of U for any fixed values $R_{(1)},\ R_{(2)}, \ldots; C_{(1)}, C_{(2)}, \ldots; \ldots$).

The probabilistic distribution of parameters for each individual element can be obtained experimentally by viewing a large batch of such elements. Very often this distribution turns out to be normal. Therefore, many researchers do the following: for example, they consider the resistance of an element to be a normal random variable ρ with the mathematical expectation $\mathbf{M}\rho = 22$ and with $3\sigma = 1.1$ (recall that in one experiment the value of ρ deviates from $\mathbf{M}\rho$ is more than 3σ, almost impossible, see (20)).

The calculation scheme turns out to be very simple: the parameter value is played for each element; then the value of U is calculated using the formula (30). Having repeated this experiment N times and obtained the values $U_4, U_2, \ldots, U_N$, we can approximately assume that

$$\mathbf{M}U \approx \frac{1}{N}\sum_{j=1}^{N} U_j$$

$$\mathbf{D}U \approx \frac{1}{N-1}\left[\sum_{j=1}^{N}(U_j)^2 - \frac{1}{N}\left(\sum_{j=1}^{N} U_j\right)^2\right].$$

For large N, the multiplier $1/(N-1) can be replaced in the last formula$ by $1/N$, and then this formula will turn out to be a simple consequence of formulas (8) and (9). In mathematical statistics, it is proved that for small N it is better to keep the multiplier $1/(N-1)$.

8.6.2 Examples of reliability calculation.

Let's say we want to estimate the average uptime of the product, assuming that the uptime characteristics of each of the elements are known.

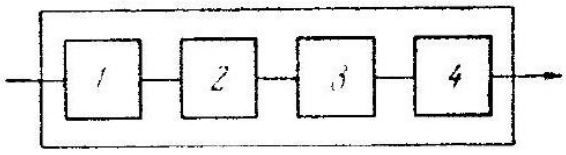


Fig. 16.

If we assume that the uptime of each element $t_{(k)}$ is a fixed value, then it will not be difficult to calculate the uptime t of the product. For example, for the product schematically shown in Fig. 16, in which the failure of one element entails the failure of the entire product,

$$t = \min\left(t_{(1)}; t_{(2)}; t_{(3)}; t_{(4)}\right) \tag{31}$$

And for the product schematically shown in Fig. 19, in which one of the elements is duplicated,

$$t = \min\left[t_{(1)}; t_{(2)}; \max\left(t_{(3)}; t_{(4)}\right); t_{(5)}\right] \tag{32}$$

since if, for example, element No. 3 fails, the product will continue to work on one element No. 4.

In fact, the uptime of any element is a random variable $\mathbf{\Theta}_{(k)}$. When we say that the service life of an electric bulb is 1000 hours, then this is only the average value of $\mathbf{M}\Theta$ of the value Θ : everyone knows that one bulb burns out faster, and the other (exactly the same) burns longer.

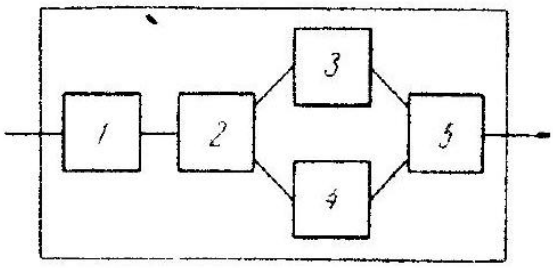


Fig. 17.

If the densities of distributions $\Theta_{(k)}$ are known for each of the elements of the product, then $\mathbf{M}\Theta$ can be calculated using the Monte Carlo method in exactly the same way as it was done in clause 6.1. In fact, for each element, you can play the value of $\Theta_{(k)}$, let it be $t_{(k)}$. Then, using the corresponding formula (31) or (32), the value of t of the random variable Θ can be calculated. Having repeated this experience quite a few (N) times, we can assume that

$$\mathbf{M}\Theta \approx \frac{1}{N}\sum_{j=1}^{N} t_j$$

where t_j is the value of t obtained in the j experiment.

It should be noted that the question of the distribution of the lifetime of Θ_k for individual elements is not so simple: for the most durable elements, the organization of the experiment is difficult, since you need to wait until a lot of elements fail.

8.6.3 Further possibilities of the method.

The examples given show that the methodology for calculating the quality of the designed products is very simple in theory. You need to know the probabilistic characteristics of all the elements of the product and be able to calculate the value of interest to us as a function of the parameters of these elements. Then the randomness of the parameters can be taken into account by modeling.

When modeling, you can get much more useful information, not just the mathematical expectation and variance of the value we are interested in. Let's say, for example, we got a large number of N values of $U_1, U_2, \ldots, U_N$ of the random variable U. Based on these values, an approximate distribution density of U can be constructed. This question relates, in essence, to statistics, since we are talking about processing the results of experiments (only they were conducted on a computer). Therefore, we will limit ourselves to a specific example.

Let's say that we got only $N = 120$ values of $U_1, U_2, \ldots, U_{120}$ of the random variable U, and they are all enclosed within

$$1 < U_j < 6,5.$$

Let's divide the interval $1 < x < 6.5$ into 11 (any number, not too large and not too small) equal intervals of length $\Delta x = 0.5$ and count how many values of U_j fell into each interval. The numbers of hits are shown in Fig. 20.

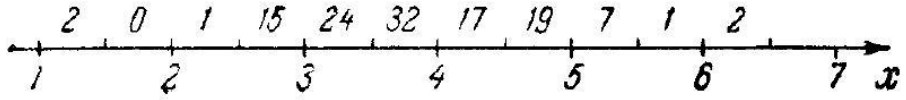


Рис. 20.

The frequency of hits in any interval is obtained by dividing the number of hits by $N = 120$. In our example, the frequencies are equal: 0,017; 0; 0,008; 0,12; 0,20; 0,27; 0,14; 0,16; 0,06; 0,008; 0,017.

Let's build a rectangle over each of the partition intervals, the area of which is equal to the frequency of hitting U_j in this interval (Fig. 21). In other words, the height of each rectangle is equal to the frequency divided by Δx. The resulting step line is called a histogram.

The histogram serves as an approximation to the unknown density of the random variable U. Therefore, for example, the area of the histogram enclosed between $x = 2.5$ and $x = 5.5$ gives us an approximate probability value

$$P\{2,5 < U < 5,5\} \approx 0,94$$

Therefore, based on the calculation (experience), it can be assumed that with a probability approximately equal to 0.94, the value of U is enclosed in the range of $2.5 < U < 5.5$.

In Fig. 21, for comparison, the density of a normal random variable ξ' with parameters $a = 3.85$, $\sigma = 0.88$ is plotted. If we use this density to calculate the probability that ζ' is enclosed in the range $2.5 < \zeta' < 5.5$, then we get a fairly close value of 0.91.

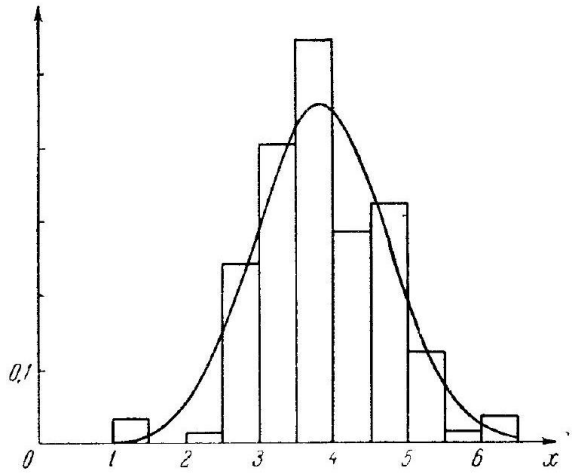


Fig. 21.

8.7 Calculation of the passage of neutrons through the plate

The probabilistic laws of interaction of a single elementary particle (neutron, photon, meson, etc.) with matter are known. It is usually necessary to find the macroscopic characteristics of processes involving a huge number of such particles: densities, fluxes, etc. This situation is similar to the one we encountered in §5 and 6, and is very convenient for using the Monte Carlo method.

Perhaps the Monte Carlo method is most often used in neutron physics. We will consider the simplest version of the problem of the passage of neutrons through a plate.

8.7.1 Setting the task.

Let a neutron flux with energy E_0 fall on a homogeneous infinite plate $0 \leqslant x \leqslant h$. The angle of incidence is 90°. When colliding with atoms of the substance that makes up the plate, neutrons can elastically disperse or be absorbed. Let's assume for simplicity that the neutron energy does not change during scattering and any direction of the neutron "rebound" from the atom is equally likely (the latter is sometimes true in substances with heavy atoms). In Fig. 22 shows the various destinies of neutrons: a neutron (a) passes through the plate, a neutron (b) is absorbed, a neutron (c) is reflected by the plate.

It is required to calculate the probability of neutrons passing through the plate p^+, the probability of neutron reflection by the plate p^- and the probability of neutron absorption in the plate p^0.

In this case, the interaction of neutrons with matter is characterized by two constants Σ_c and Σ_s, which are called the absorption cross section and the scattering cross section. The indexes $\boldsymbol{c}$ and s are the first letters of the English words capture and scattering.

The sum of these sections is called the total section

$$\Sigma = \Sigma_c + \Sigma_s.$$

The physical meaning of the cross sections is as follows: when a neutron collides with an atom of matter, the probability of absorption is Σ_c/Σ, and the probability of scattering isΣ_s/Σ.

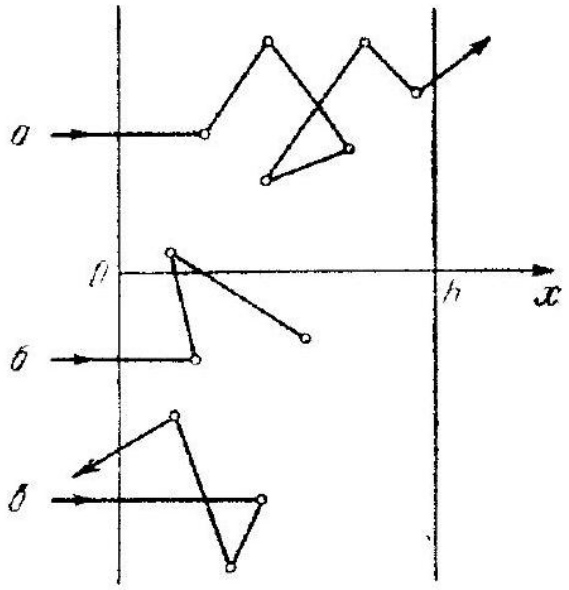


Fig. 22.

The free path length of a neutron λ (that is, the length of the path from collision to collision) is a random variable. It can take any positive values with probability density

$$p(x) = \Sigma e^{-\Sigma x}.$$

It is easy to see that this density of λ coincides with the density (26) of the random variable τ for the simplest flow of applications. By analogy with clause 5.2, we can immediately write an expression for the average free run length

$$\mathbf{M}\lambda = 1/\Sigma$$

and the formula for playing λ :

$$\lambda = -\frac{1}{\Sigma}\ln\gamma.$$

It remains to be seen how to choose a random neutron direction after scattering. Since the problem is symmetric with respect to the x axis, the direction is completely determined

by one angle φ between the direction of the neutron velocity and the Ox axis. It can be proved that the requirement of equal probability of any direction in this case is equivalent to the requirement that the cosine of this angle $\mu = \cos\varphi$ be evenly distributed in the interval (-1.1). From formula (24), for $a = -1, b = 1$, the formula for playing follows μ :

$$\mu = 2\gamma - 1$$

8.7.2 A calculation scheme by modeling true trajectories.

Suppose that the neutron experienced the kth scattering inside the plate at the point with the abscissa x_k and after that began to move in the direction of μ_k.

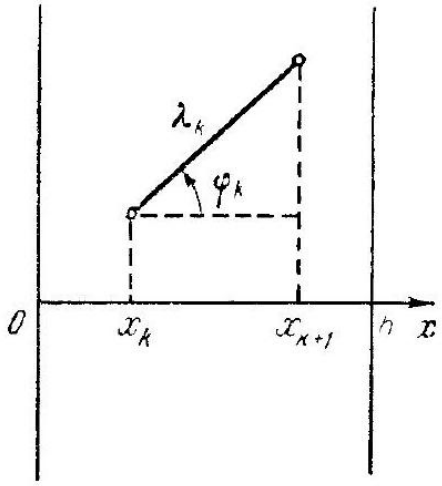


Fig. 23.

Let's play the length of the free run

$$\lambda_k = -(1/\Sigma)\ln\gamma$$

and calculate the abscissa of the next collision (Fig. 23)

$$x_{k+1} = x_k + \lambda_k\mu_k$$

Let's check the condition of passage through the plate:

$$x_{k+1} > h.$$

If this condition is met, then the count of the neutron trajectory ends and one is added to the counter of the passed particles. Otherwise, we check the reflection condition:

$$x_{k+1} < 0$$

If this condition is met, the trajectory count ends and a unit is added to the counter of reflected particles. If this is not conditionally fulfilled, that is, $0 \leqslant x_{k+1} \leqslant h$, then the neutron has experienced $(k+1)The$th collision is inside the plate, and it is necessary to play out the "fate" of the neutron in a collision.

According to clause 4.1, we select the next value of γ and check the absorption condition:

$$\gamma < \Sigma_c/\Sigma$$

If the last inequality is satisfied, then the trajectory count ends and one is added to the counter of absorbed particles. Otherwise, we believe that the neutron has experienced scattering at the point with the abscissa x_{h+1}. Then we play a new direction of the neutron velocity

$$\mu_{k+1} = 2\gamma - 1$$

and then we repeat the whole cycle again (but, of course, with different values of γ).

All γ are written without indexes, since it means that each value of γ is used only once. To calculate one link of the trajectory, three values of γ are needed.

Initial values for each trajectory:

$$x_0 = 0, \quad \mu_0 = 1$$

After N trajectories are counted, it turns out that N^+ neutrons passed through the plate, N^- neutrons were reflected from it, and N^0 neutrons were absorbed. Obviously, the probabilities sought are approximately equal to the relations

$$p^+ \approx \frac{N^+}{N}, \quad p^- \approx \frac{N^-}{N}, \quad p^0 \approx \frac{N^0}{N}.$$

Figure 24 shows a flowchart of the program for calculating this task. The index j is the trajectory number, the index k is the collision number (along the trajectory).

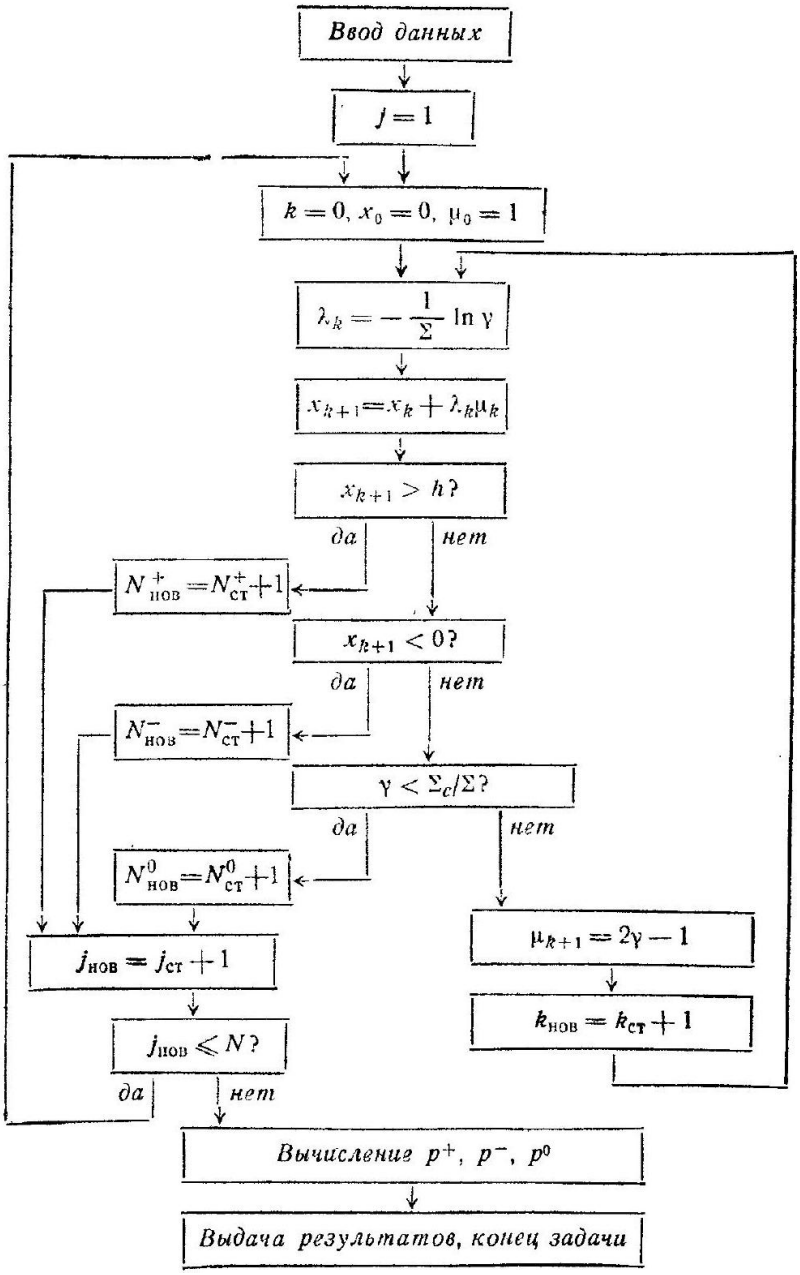


Fig. 24.

This calculation method, although very natural, is imperfect. In particular, it is difficult to calculate the probability of p^+ using this method when it is very small. And such a case just has to be faced when calculating radiation protection.

There are more "tricky" versions of the Monte Carlo method that allow you to calculate

such cases. Let's briefly focus on one of the simplest calculation options using the so-called "scales".

8.7.3 Calculation scheme using weights replacing absorption.

Consider the same problem of neutron passage. Suppose that a "packet" consisting of a large number w_o of identical neutrons is moving along the same trajectory. When colliding at the point with the abscissa x_1 the number of neutrons absorbed from the "package" is on average $w_0\,(\Sigma_c/\Sigma)$, and the number of neutrons that have experienced scattering is on average equal to $w_0\,(\Sigma_s/\Sigma)$.

Add the value $w_0\,(\Sigma_c/\Sigma)$ to the counter of absorbed particles, and we will follow the movement of the scattered "package" further, assuming that the entire remaining "package" has scattered in one direction.

All the account formulas given in clause 7.2 remain the same. Only with each collision will the number of neutrons in the "package" decrease

$$w_{k+1} = w_k\,(\Sigma_s/\Sigma)\,,$$

since the part of the "package" containing $w_n\,(\Sigma_c/\Sigma)$ neutrons, will be absorbed. And the trajectory now cannot end in absorption.

The value of w_k is usually called the weight of a neutron, and instead of talking about a "package" consisting of w_k neutrons, they talk about one neutron with a weight of w_k. The initial weight of w_0 is usually assumed to be 1. This does not contradict our reasoning about the "big package", because it is easy to see that all w_k obtained by calculating one trajectory contain w_0 by a common multiplier.

The flowchart of the program for implementing such a calculation is shown in Fig. 25. Obviously, it is no more complicated than the flowchart shown in Fig. 24. However, it can be proved that the calculation of p^+ by the method of this paragraph is always more profitable than the calculation by the method of paragraph 7.2.

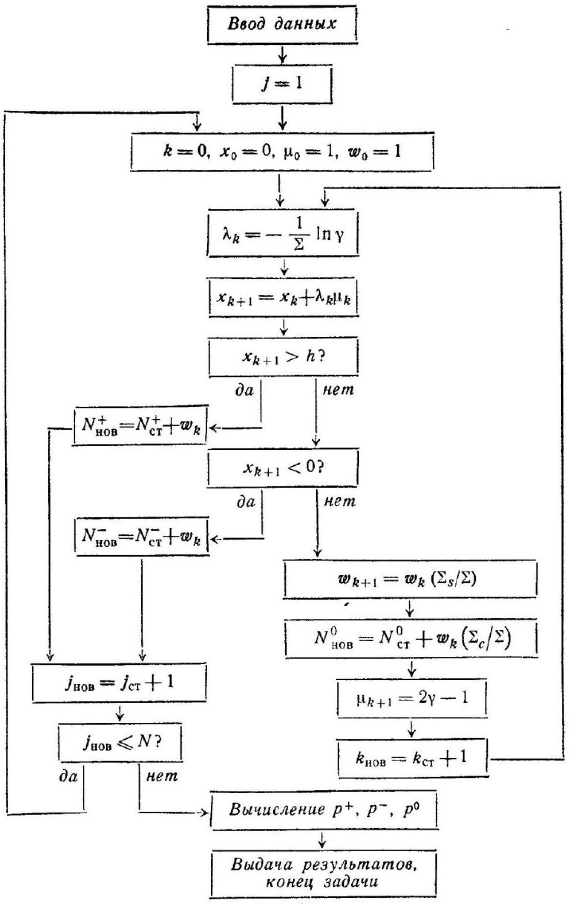


Fig. 25.

8.8 Calculation of a certain integral

The problems discussed in §5, 6, 7 were probabilistic in nature, and using the Monte Carlo method to solve them seemed quite natural. Here we consider a simple mathematical problem: the approximate calculation of a certain integral.

Since calculating certain integrals is equivalent to calculating areas, we could use the method of Example 1.2. Instead, we will present another, more effective method that allows us to build different probabilistic models for solving this problem using the Monte Carlo method, and indicate how to choose "better" models among them.

8.8.1 The calculation method.

Consider the function $g(x)$ given on the interval $a < x < b$. It is required to approximate the integral

$$I = \int_a^b g(x)dx \tag{33}$$

Let's choose an arbitrary distribution density $p_{\text{g}}(x)$ defined on the interval (a, b) (that is, an arbitrary function $p_\xi(x)$ satisfying conditions (15) and (16)).

Along with the random variable ξ defined in the range (a, b) with density $p_\varepsilon(x)$, we will need a random variable

$$\eta = g(\xi)/p_\xi(\xi)$$

According to (18)

$$\mathbf{M}\eta = \int_a^b \left[g(x)/p_\xi(x)\right] p_\xi(x)dx = I$$

Now let's consider N identical random variables $\eta_1, \eta_2, \ldots, \eta_N$ and apply the central limit theorem of clause 2.4 to their sum. Formula (21) in this case is written as follows:

$$P\left\{\left|\frac{1}{N}\sum_{j=1}^{N}\eta_j - I\right| < 3\sqrt{\frac{D_\eta}{N}}\right\} \approx 0,997 \tag{34}$$

The latter relation means that if we choose N values of $\xi_1, \xi_2, \ldots, \xi_N$, then with a sufficiently large N

$$\frac{1}{N}\sum_{j=1}^{N}\frac{g\left(\xi_j\right)}{p_\xi\left(\xi_j\right)} \approx I \tag{35}$$

It also shows that, with a very high probability, the approximation error (35) does not exceed $3\sqrt{\overline{\text{D}}/N}$.

8.8.2 How to choose a calculation scheme.

We have seen that any random variable ξ defined in the range (a, b) can be used to calculate the integral (33). Anyway

$$\mathbf{M}\eta = \mathbf{M}\left[g(\xi)/p_\xi(\xi)\right] = I$$

However, the variance of $\mathbf{D}\eta$, and with it the error estimate of formula (35), depend on which value of ξ we use. Really,

$$\mathbf{D}\eta = \mathbf{M}\left(\eta^2\right) - I^2 = \int_a^b \left[g^2(x)/p_{\tilde{\text{j}}}(x)\right] dx - I^2$$

It can be proved that this expression will be minimal when $p_\xi(x)$ is proportional to $|g(x)|$.

Of course, it is impossible to choose very complex $p_g(x)$, since the procedure for playing the values of ξ will become very time-consuming. But you can follow this recommendation when choosing $p_\xi(x)$ (see example section 8.3).

In practice, integrals of the form (33) are not calculated using the Monte Carlo method: there are more accurate methods for this - quadrature formulas. However, when switching to multiple integrals, the situation changes: quadrature formulas become very complex, and the Monte Carlo method remains almost unchanged.

8.8.3 A numerical example.

Let's calculate the integral approximately

$$I = \int_0^{\pi/2} \sin x dx$$

The exact value of this integral is known:

$$\int_0^{\pi/2} \sin x dx = [-\cos x]_0^{\pi/2} = 1$$

We use two different random variables ξ to calculate: with constant density $p_\xi(x) \equiv 2/\pi$ (that is, ξ is uniformly distributed in the range $(0, \pi/2)$) and with linear density $p_\xi(x) = 8x/\pi^2$. Both of these densities, together with the integral function $\sin x$, are plotted in Fig. 26. From this figure it can be seen that the linear density is more consistent with the recommendation of paragraph 8.2 that the proportionality of $p_\xi(x)$ and $\sin x$ is desirable. Therefore, we must expect that the second method of calculation will give the best result.

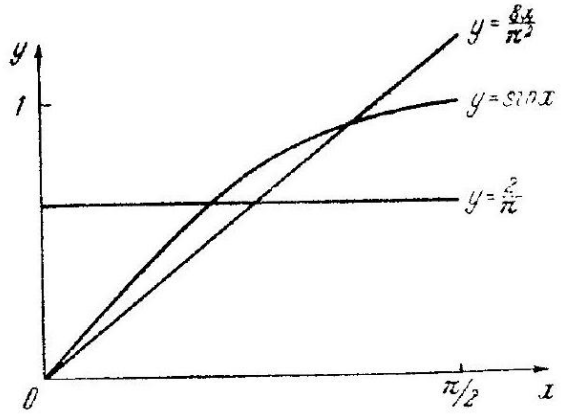


Fig. 26.

(a) Let $p_\xi(x) \equiv 2/\pi$ on the interval $(0, \pi/2)$. The formula for playing ξ can be obtained from formula (24) for $a = 0$ and $b = \pi/2$:

$$\xi = \frac{\pi}{2}\gamma$$

And the formula (35) will take the form

$$I \approx \frac{\pi}{2N} \sum_{j=1}^{N} \sin \xi j$$

Let $N = 10$. As the values of γ, we use triples of numbers from Table A (multiplied by 0.001). The intermediate results are summarized in Table 1.

Table 1

j	1	2	3	4	5	6	7	8	9	10
γ_f	0,865	0,159	0,079	0,566	0,155	0,664	0,345	0,655	0,812	0,332
ξ_f	1,359	0,250	0,124	0,889	0,243	1,043	0,542	1,029	1,275	0,521
$\sin \xi f$	0,978	0,247	0,124	0,776	0,241	0,864	0,516	0,857	0,957	0,498

The result of the calculation:

$$I \approx 0,952$$

(b) Now let's say $p_\xi(x) = 8x/\pi^2$. To play ξ, we use equation (23)

$$\int_0^\xi \left(8x/\pi^2\right) dx = y$$

from where, after simple calculations, we get

$$\xi = \frac{\pi}{2}\sqrt{\gamma}$$

Formula (35) will take the following form:

$$I \approx \frac{\pi^2}{8N} \sum_{j=1}^{N} \frac{\sin \xi_j}{\xi_j}$$

Let $N = 10$. The numbers γ will be the same as in the case (a). The intermediate results are summarized in Table 2.

Table 2

j	1	2	3	4	5	6	7	8	9	10
γ_j	0,865	0,159	0,079	0,566	0,155	0,664	0,345	0,655	0,812	0,332
ξ_j	1,461	0,626	0,442	1,182	0,618	1,280	0,923	1,271	1,415	0,905
$\frac{\sin \xi_j}{\xi_j}$	0,680	0,936	0,968	0,783	0,937	0,748	0,863	0,751	0,698	0,868

The result of the calculation:

$$I \approx 1,016$$

As we expected, the second method of calculation gave a more accurate result.

8.8.4 About error evaluation.

In clause 8.8.2, it was already noted that the absolute error in calculating the integral I according to formula (35) practically cannot exceed the value of $3\sqrt{\mathrm{D}/N}$. However, in reality, the error, as a rule, turns out to be noticeably less than this value. Therefore, in practice, another value is often used to characterize the error - the so-called probable error

$$\delta_{\text{prob}} = 0,675\sqrt{\mathrm{D}\eta/N}.$$

The actual absolute error depends on the random numbers used in the calculation and may be 2-3 times larger or several times smaller than δ_{prob}. So δ_{prob} does not give us the error boundary, but its order.

Bibliography

[1] W.G. Bickley. Piecewise cubic interpolation and two-points boundary value problems. *Computer Journal*, 11:206, 1968.

[2] P. Henciri. *Applied and Computational Complex Analysis*. John Wiley & Sons, 1974.

[3] E Isaacson and H.B. Keller. *Analysis of Numerical Methods*. John Wiley & Sons, 1966.

[4] H. Leyve and E.A. Baggott. *Numerical Solution of Differential Equations*. Dover, 1950.

[5] F. Patricio. Cubic spline functions and inital-value problems. *BIT*, 18:343, 1978.

[6] J.N. Reddy. *An Introduction to the Finite Element*. Book Co., Singapore, 1985.

[7] S.S. Sastry. Finite difference approximations to one-dimensional parabolic equations using cubic spline technique. *J. Comp. and App. Math.*, 2:23, 1976.

[8] S.S. Sastry. *Engineering Mathematics*. Prentice-Hall of India, 2004.