САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ПРИОРИТЕТНЫЙ
НАЦИОНАЛЬНЫЙ ПРОЕКТ
"ОБРАЗОВАНИЕ"

Проект «Инновационная образовательная среда в классическом университете»

Пилотный проект № 22 «Разработка и внедрение
инновационной образовательной программы «Прикладные математика и физика»»

Физический факультет

Кафедра вычислительной физики

И. В. Андронов, С. Ю. Славянов

ЧИСЛЕННЫЕ МЕТОДЫ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Учебно-методическое пособие

.

.

Санкт Петербург
2007 г.

- Рецензент: зав. кафедрой вычислительной физики, д.ф.м.н. Яковлев С. Л.

- Печатается по решению методической комиссии физического факультета СПбГУ.

- Рекомендовано Ученым советом физического факультета СПбГУ.

ЧИСЛЕННЫЕ МЕТОДЫ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ. – СПб., 2007

В учебно-методическом пособии рассмотрены численные методы, применяемые для решения задачи Коши, краевой задачи и задачи Штурма-Лиувилля, а также некоторые методы решения интегральных уравнений. Пособие предназначено для студентов 4-7-го курсов, аспирантов, соискателей и других обучающихся…

Chapter 1

Examples of differential equations

1.1 Physical pendulum

Our primary goal would be to present examples of typical ordinary differential equations as they appear in various physical theories. The first example is the second order nonlinear differential equation describing a physical pendulum.

$$\frac{d^2\phi}{dt^2} = -\frac{g}{l}\sin(\phi). \tag{1.1}$$

Here $\phi = \phi(t)$, t is time, ϕ is the angle of the pendulum with the vertical axis, l is the length of the pendulum and g is the acceleration due to gravity. For this equation initial data (Cauchy data) should be posed. For instance

$$\begin{cases} \phi(t)|_{t=0} = \phi_0, \\ \left.\dfrac{d\phi}{dt}\right|_{t=0} = \phi_0'. \end{cases}$$

The particular feature of this equation is that we can integrate it once obtaining so called first integral of motion. New variables are introduced:

$$\frac{d\phi}{dt} = \dot{\phi} = \psi$$

is accepted as the dependent variable and ϕ is accepted as the independent variable. Then (1.1) can be rewritten as

$$\frac{1}{2}\frac{d\dot{\phi}^2}{d\phi} = -\frac{g}{l}\sin(\phi).$$

Integrating this equation we arrive to the first order equation

$$\frac{1}{2}\dot{\phi}^2 - \frac{g}{l}\cos(\phi) = E.$$

This first order integral of motion corresponds to conservation of energy E at the pendulum oscillations. The derived equation after returning to old variables is simplified to

$$\frac{d\phi}{dt} = \sqrt{E + 2\frac{g}{l}\cos(\phi)},$$

which in its turn can be easily integrated

$$t = \int_{\phi_0}^{\phi} \frac{1}{\sqrt{E + 2(g/l)\cos(\phi)}} d\phi. \tag{1.2}$$

Unfortunately the integral at the right hand side of (1.2) can not be presented in terms of elementary functions and is known as an elliptic integral. In the studied simple case the problem of numeric solution of the given differential equation is reduced to numeric quadratures which can be performed by Newton-Cotes numeric integration.

1.2 Normal waves in the ocean waveguide

Another example gives a theory of the long-range propagation of sound waves generated by a monochromatic point source in a sound channel in the ocean (ocean waveguide). It is an important problem in the underwater acoustics. This channel appears because the velocity of sound propagation has a minimum at a certain depth. This problem can be studied under various assumptions about the structure of the water layer and the ocean bottom. Here the simplest case is considered when the properties of the water layer are independent of horizontal coordinates and the bottom is a uniform nonelastic half-space in which waves propagate with a constant velocity. Then the field of a sound wave $u(r), r = (x, y, z)$, is the solution of the problem

$$\Delta u + \frac{\omega^2}{c^2} u = -\delta(\mathbf{r} - \mathbf{r}_0), \qquad 0 < z < H, \tag{1.3}$$

$$\Delta u + \frac{\omega^2}{c_H^2} u = 0, \qquad z > H,$$

$$\left.\frac{\partial u}{\partial z}\right|_{z=H-0} = \frac{1}{\kappa}\left.\frac{\partial u}{\partial z}\right|_{z=H+0}.$$

Here $\delta(\mathbf{r}-\mathbf{r}_0)$ is the delta function that represents a point source at $\mathbf{r}_0$, $\mathbf{r}_0 = (0,0,z_0)$, H is the depth of the ocean, ω is the frequency, $c(z)$ is the sound velocity in a water layer, c_H the sound velocity in the soil, $c_H > c(z)$, and κ is the density of the soil (the water density is assumed to be unity).

The problem can be solved by separating the variables in the cylindrical coordinate system z, ρ, ϕ. Neglecting the decay of the wave field as a result of a propagation along the horizontal coordinates, the amplitude of the acoustic field is represented as a sum of so-called normal modes

$$u(z) = -\sum_{m=1}^{M} \frac{i}{4} H_0^{(1)}(k\mu_n\rho) A_m \psi_m(z). \tag{1.4}$$

The solution contains the factor $(i/4)H_0^{(1)}$ that represents a cylindric wave generated by a point source and the factor $\psi_m(z)$ that describes the dependence of this wave amplitude on depth of the observation point. Equations (1.3) imply that $\psi_m(z)$ are the eigenfunctions of the homogeneous boundary value problem for the vertical coordinate z:

$$\begin{cases} \psi''(z) + p^2(n^2(z) - \mu^2)\psi(z) = 0, \\ \psi(0) = 0, \\ \psi'(H) + \dfrac{p}{\kappa}\sqrt{\mu^2 - n_H^2}\,\psi(H) = 0. \end{cases} \tag{1.5}$$

Here $p = \omega d/c*$ is a large dimensionless parameter of the problem, $c* = \min c(z)$, d is a characteristic scale of the coordinates. The eigenvalues μ_m^2 of the problem (1.5) determine the phase velocities c_m of normal modes $\psi_m(z)$ as

$$c_m = \frac{\omega d}{\mu_m p}.$$

The coefficients A_m of the normal mode expansion (1.3) can be found from the position of the source

$$A_m = \frac{\psi_m(z_0)}{dN_m^2}.$$

The constants N_m^2 are the normalization factors of the functions $\psi_m(z)$ extended to the semi-axis $[0, \infty)$. Since for $z > H$ the functions $\psi_m(z)$ can be

written explicitly, we get for N_m^2

$$N_m^2 = \int_0^H \psi_m^2(z)dz + \frac{\psi_m^2(H)}{2p\sqrt{\mu^2 - n_H^2}}.$$

The primary task in the problem under consideration is to calculate asymptotics of phase velocities c_m and eigenfunctions $\psi_m(z)$ for large values of p. The calculations have two peculiarities. First, for a given position of the source and the observation point, that is, for given z and z_0, the terms in expansion (1.4) for which $c_m < c(z_0) - \delta$ or $c_m < c(z) - \delta$ should be omitted. For such terms one of the factors, either $\psi_m(z)$ or $\psi_m(z_0)$, is exponentially small (the value of δ is determined by desired accuracy of calculations). Second, the total field $u(r)$ is a result of interference of many normal modes, and an important role in summing their contributions is played by the differences in phase velocities. The accuracy of calculating the functions $\psi_m(z)$ might therefore be lower than the accuracy with which phase velocities are computed.

1.3 Waves in the shallow water

Around hundred years ago a famous nonlinear partial differential equation (which now is known as Korteveg-de-Vries (KdV) equation) was proposed in order to describe the propagation of waves in shallow water. This equation can be written as

$$\psi_{xxx} - 6\psi\psi_x + 3\psi_t = 0, \tag{1.6}$$

where $\psi = \psi(t, x)$. Particular solutions of (1.6) are functions

$$\psi = -\frac{3c}{2}\frac{1}{\cosh^2\left(\frac{\sqrt{3c}}{2}(x - ct + \delta)\right)},$$

where c and δ are constants.

Another way to solve (1.6) is possible. New variables are introduced

$$\psi(t, x) = t^{-2/3}(w'(z) + w^2(z))$$

and

$$z = xt^{1/3}$$

with the following differentiation rules

$$\frac{\partial}{\partial x} = t^{-1/3}\frac{d}{dz}, \quad \frac{\partial}{\partial t} = -3\frac{z^2}{x^3}\frac{d}{dz},$$
$$\psi_{xxx} = t^{-5/3}(w' + w^2), \quad -6\psi\psi_x = -6t^{-5/3}(w' + w^2)(w' + w^2)',$$
$$-3\psi_t = t^{-5/3}(-2(w' + w^2) - z(w' + w^2)').$$

They induce reduction of the KdV equation to the ordinary differential equation

$$(w' + w^2)''' - 6(w' + w^2)(w' + w^2)' - 2(w' + w^2) - z(w' + w^2)',$$

which in its turn can be reduced to a second order equation

$$v'' - 2v^3 - zv = 0. \tag{1.7}$$

This equation belongs to the set of the so-called *Painleve equations* which enjoys the so-called *Painleve property.*

1.4 Painleve equations

Suppose that an n-th order nonlinear ODE

$$w^n(z) = G(w^{n-1}, \dots, w, z) \tag{1.8}$$

is studied with the function $G(w^{n-1}, \dots, w, z)$ having good analytical properties in its variables which we specify later. Solutions of this equation as functions of the independent complex variable z may have singularities determined by the analytical behavior of the function $G(w^{n-1}, \dots, w, z)$ only (without taking into account initial data for the solution). These singularities are called *fixed* singularities. On the contrary those singularities of solutions which can not be predicted by the coefficients of the equation and which change their position if initial data for the solution change are called *movable* singularities. The movable singularity may be a pole of the solution, an essential singularity, a branch point (both algebraic and transcendental).

Examples

The equation

$$w''(z) + (w')^2(z) = 0 \tag{1.9}$$

has the solution

$$w(z) = \ln(z + c_1) + c_2. \tag{1.10}$$

Hence, eq. (1.9) has a movable singularity which is a transcendent branch point.

Linear equations which may be considered as a special case of nonlinear equations have no movable singularities.

In a series of articles started by P. Painlevé the following problem has been solved. Consider the second-order nonlinear ODE of the form

$$q''(t) = F(t, q, q'), \tag{1.11}$$

where $F(t, q, q')$ is a rational function in its arguments[1] The question arises: when solutions of such an equation have no movable (depending on initial data) critical points. By critical points are meant branch points and essential singularities. In this case movable singularities may be only poles of the solution. Absence of the movable critical points is known as the Painlevé property. All equations (namely, 50 of them) of the form (1.8) possessing the Painlevé property have been found. Among these are many solvable in terms of elementary or other known functions (e.g. elliptic functions). But several equations did not simplify to the known equations. Nowadays they usually are called Painlevé equations denoted by P^{VI}, P^{V}, P^{IV}, P^{III}, P^{II} and P^{I}. Although the method of investigation proposed by Painlevé is sufficiently simple the practical calculations involve considering many special cases and therefore are rather laborious. Since the original studies on the problem have to our mind more historical interest than the practical meaning we give here a list of Painlevé equations. Standard Painlevé class equations are the following

$$\begin{aligned} P_z^{VI} y(z) = y'' - \frac{1}{2}\left(\frac{1}{y} + \frac{1}{y-1} + \frac{1}{y-z}\right)(y')^2 + \left(\frac{1}{z} + \frac{1}{z-1} + \frac{1}{y-z)}\right) y' \\ - \frac{y(y-1)(y-z)}{z^2(z-1)^2}\left(\alpha + \frac{\beta z}{y^2} + \frac{\gamma(z-1)}{(y-1)^2} + \frac{\delta z(z-1)}{(y-z)^2}\right) = 0, \end{aligned} \tag{1.12}$$

[1] More often F is considered even to be a second-order polynomial in q'.

$$P_z^V y(z) = y'' - \left(\frac{1}{2y} + \frac{1}{y-1}\right)(y')^2 + \frac{1}{z}y' - \frac{(y-1)^2}{z^2}\left(\alpha y + \frac{\beta}{y}\right) + \frac{\gamma y}{z} + \frac{\delta y(y+1)}{y-1} = 0, \quad (1.13)$$

$$P_z^{IV} y(z) = y'' - \frac{1}{2y}(y')^2 + \frac{3}{2}y^3 - 4z^2 - 2(z^2-\alpha)y - \frac{\beta}{y} = 0, \quad (1.14)$$

$$P_z^{III} y(z) = y'' - \frac{1}{y}(y')^2 + \frac{1}{z}y' - \frac{1}{z}(\alpha y^2 + \beta) - \gamma y^3 - \frac{\delta}{y} = 0, \quad (1.15)$$

$$P_z^{II} y(z) = y'' - y^3 - zy - \alpha = 0, \quad (1.16)$$

$$P_z^{I} y(z) = y'' - 6y^2 - z = 0. \quad (1.17)$$

Above the conventional notation for Painlevé equations was used. The derived equation (1.7) coincides with P^{II} at particular value of a parameter $\alpha = 0$.

One of the main advantages of Painlevé equations in comparison to other nonlinear equations is that beyond local solutions global solutions of these equations can be constructed with prescribed asymptotic behavior at certain points of independent variable in complex plane. This fact is due to the isomonodromic deformation theory which recently has strongly influenced the theory of special functions.

1.5 Prüfer transform

Suppose that a second order ODE of the form

$$y''(x) + p^2 Q(x,\lambda)y = 0 \quad (1.18)$$

is considered with appropriate boundary conditions, say

$$y(0) = y(L) = 0.$$

It is assumed that

$$Q(x,\lambda) \geq 0 \quad \text{on } [0,L]$$

resulting in the fact that all solutions of (1.18) are oscillating. The parameter λ plays the role of eigen value parameter. The parameter p is supposed to be large. This causes difficulties in numerical solution of (1.18).

The principal idea of Prüfer transform is to present the solution with the help of two auxiliary function — the amplitude and the phase. The fine structure of the solution could be various. Here we shall focus on one of possible strategies. Consider a comparison equation with constant coefficient.

$$w''(\xi) + w(\xi) = 0. \tag{1.19}$$

Those solutions $w(\xi)$ are chosen linearly independent of its derivative (for example, $w = \sin(\xi)$). Solutions of (1.18) are sought in the form

$$y = A(x)w(p\tau(x) + h(x)), \tag{1.20}$$

and in addition it is required that the amplitude $A(x)$ and the phases $\tau(x)$ and $h(x)$ satisfy first order equations. The function $\tau(x)$ is called the main term of the phase and $h(x)$ is the correction term of the phase. For the derivative $y(x)$ we have

$$y' = A'w + Aw'p\tau' + Aw'h'.$$

The next differentiation will produce the second derivative A''. To avoid this we choose the correction term of the phase $h(x)$ such that it nullifies the coefficient of A', that is

$$A'w + Aw'h' = 0,$$

which results in

$$\frac{A'}{A} = -\frac{w'h'}{w}. \tag{1.21}$$

Note, that the functions A and h both have an oscillating character.

Further differentiation yields

$$y'' = A'w'p\tau' + Aw'p\tau'' + Aw''p\tau'h' + Aw''p^2(\tau')^2,$$

or after use of (1.18) and (1.19)

$$A'w'p\tau' + Aw'p\tau''Aw''p^2(\tau')^2 + Aw''p\tau'h' + p^2QAw = 0. \tag{1.22}$$

First we require that the sum of the terms with the highest power of p be zero

$$(\tau')^2 = Q, \tag{1.23}$$

or

$$\tau' = \sqrt{Q}$$

with

$$\tau'' = \frac{Q'}{2\sqrt{Q}}.$$

This gives the expression for the main part of the phase, namely the function $\tau(x)$

$$\tau = \int \sqrt{Q} dx. \tag{1.24}$$

After manipulation with (1.22)

$$\begin{aligned} A'w'p\tau' + Aw'p\tau'' + Aw''p\tau'h' = 0, \\ -\frac{(w')^2h'}{w}\tau' + w'\tau'' - \tau'h'w = 0, \end{aligned}$$

the equation for the correction $h(p,x)$ is obtained

$$h'\tau' \left((w')^2 + w^2\right) = ww'\tau''. \tag{1.25}$$

Suppose that the solution w is chosen as

$$w = \sin(\xi), \qquad w' = \cos(\xi).$$

Then it holds

$$(w')^2 + w^2 = 1,$$

and the equation for $h(x,p)$ reduces to

$$h' = ww'\frac{\tau''}{\tau'},$$

which simplifies to

$$h' = ww'\frac{Q'}{2Q} = \frac{Q'}{4Q}\sin(2p\tau + 2h).$$

Returning to the equation for the amplitude we obtain the equation

$$\frac{A'}{A} = -(w')^2 \frac{Q'}{2Q}$$

with the solution

$$A = A_0 \exp\left(-\int \left[\cos^2(p\tau + h)\frac{Q'}{2Q}\right] dx\right) = A_0 \exp\left(-\int \left[1 + \cos(2p\tau + 2h))\frac{Q'}{4Q}\right] dx\right). \quad (1.26)$$

It means that the Prüfer transform reduces solution of the initial linear equation (1.18) to two explicit integrations and solution of a nonlinear first order equation for h

$$\begin{cases} \tau = \int \sqrt{Q} dx, \\ A = A_0 \exp\left(-\int \left[1 + \cos(2p\tau + 2h))\frac{Q'}{4Q}\right] dx\right), \\ h' = \frac{Q'}{4Q}\sin(2p\tau + 2h). \end{cases}$$

The latter is solved by methods typical for nonlinear equations considered in the next chapter.

Chapter 2

Numerical methods for initial-value problems

2.1 Introduction

2.1.1 Initial-value problem

As we already seen many practical problems of physics can be reduced to the solution of ordinary differential equations. General n-th order differential equation can be written as

$$F\left(x, y(x), y'(x), \ldots, y^{(n)}(x)\right) = 0. \tag{2.1}$$

Defining the vector-function $\mathbf{z}(x)$ as

$$y(x) = z_0(x), \quad y'(x) = z_1(x), \quad \ldots, \quad y^{(n-1)}(x) = z_{n-1}(x), \tag{2.2}$$

equation (2.1) can be written as

$$F\left(x, z_0(x), z_1(x), \ldots, z_{n-1}(x), z'_{n-1}(x)\right) = 0.$$

Combining it with (2.2) yields the system of first-order differential equations

$$\begin{cases} F\left(x, z_0(x), z_1(x), \ldots, z_{n-1}(x), z'_{n-1}(x)\right) = 0, \\ z'_{n-2}(x) = z_{n-1}(x), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ z'_1(x) = z_0(x). \end{cases} \tag{2.3}$$

Evidently, equation (2.1) is equivalent to system (2.3). Reduction of equation (2.1) to the system of first-order differential equations is always possible, but is not unique.

Equation (2.1) or system (2.3) should be complimented with some additional conditions that fix a unique solution. It is well known that for the differential equation of the n-th order one needs to specify n conditions. In this chapter we consider only such cases when these conditions specify the values of the unknown function $y(x)$ and its derivatives $y'(x)$, ... , $y^{(n-1)}(x)$ at a single value x_0 of the argument. The problems which are originated by such conditions are called initial-value problems for differential equations or Cauchy problems. The cases when the additional conditions specify relations between the values of the unknown function and/or its derivatives at different values of the argument are discussed in the next chapter.

It is convenient to transform equation (2.1) to the form resolved with respect to the highest order derivative

$$y^{(n)}(x) = f\left(x, y(x), y'(x), \dots, y^{(n-1)}(x)\right) \tag{2.4}$$

and correspondingly to rewrite system (2.3) in the form

$$\mathbf{z}'(x) = \mathbf{F}\left(x, \mathbf{z}(x)\right), \tag{2.5}$$

where $\mathbf{z} = (z_0, z_1, \dots, z_{n-1})^T$, and the vector-function $\mathbf{F}$ has the following components

$$\mathbf{F}_j = z_j(x), \qquad j = 0, 1, \dots, n-2,$$
$$\mathbf{F}_{n-1} = f\left(x, z_0(x), z_1(x), \dots, z_{n-1}(x)\right).$$

Note, that sometimes it is not possible to rewrite equation (2.1) in the form (2.4) because it requires analytic solution of a nonlinear in the general case equation

$$F(x, z_0, \dots, z_{n-1}, z_n) = 0$$

for z_n.

Nevertheless most numerical methods exploit the form (2.4) of the differential equation or the form (2.5) for the system of first-order equations.

2.1.2 Regularization

The important property of an initial-value problem for the differential equation (2.4) is its stability with respect to small variation of initial data or

right-hand side of the equation. The problem is called ill-conditioned if small variations of the right-hand side cause large variations of the solution. Consider the example of ill-conditioned problem

$$\begin{cases} y'(x) = y(x) - x, \\ y(0) = 1. \end{cases} \tag{2.6}$$

The general solution of this differential equations is

$$y(x) = 1 + x + ce^x,$$

where c is an arbitrary constant. Its value is obtained from the initial condition. In the case of problem (2.6) $c = 0$. Thus $y(100) = 101$. However if one changes a little the initial condition, say replace it by $y(0) = 0.9999$ then the solution will become

$$y(x) = 1 + x - 0.0001e^x,$$

and $y(100) \approx -0.2710^{40}$.

Evidently that due to round-off errors the exponential will be added to the solution $y = 1+x$ and will cause large deviation of the numerical solution from the true one.

Thus before the application of a particular numerical method to the given initial-value problem one needs to check if the problem is well-conditioned and when necessary perform such its transformations to achieve solution stability. We require that for any solution of the differential equation the following estimate holds

$$\|\mathbf{z}(x)\| \leq \|\mathbf{z}(x_0)\|, \qquad x > x_0. \tag{2.7}$$

In this case the problem is called stable.

If the problem is not stable, then standard numerical methods are inapplicable and one needs to develop special numerical schemes. Some approaches are discussed in the section 2.6.2.

2.1.3 Numerical schemes

Let us first assume that the system (2.5) consists of one first order equation and can be resolved with respect to the derivative. So, let the Cauchy problem to be of the form

$$\begin{cases} y'(x) = f\left(x, y(x)\right), \\ y(x_0) = y_0. \end{cases} \tag{2.8}$$

It is required to find the function $y(x)$ for $x > x_0$. Two approaches are possible. In the first approach the function $y(x)$ can be searched in some analytic form, for example as a series. In the second approach the values of the function $y(x)$ are searched only at some nodal set $\{x_k\}_{k=0,1,\ldots,N}$. The methods of the first type are conventionally called continuous and the methods of the second type — discrete or finite-difference.

Continuous methods are discussed in more details in chapter 3 where boundary-value problems are considered. Here we present only the method of series which allows a truncated Taylor series

$$y(x) = y(x_0) + y'(x_0)(x - x_0) + \ldots + \frac{1}{m!} f^{(m)}(x_0)(x - x_0)^m. \qquad (2.9)$$

for the solution to be found.

The zero-th order term $y(x_0)$ is given by the initial value

$$y(x_0) = y_0.$$

Substituting this value to the differential equation, one finds

$$y'(x_0) = f(x_0, y_0).$$

In order to find higher order terms of the series (2.9) one differentiates the equation in the problem (2.8). In particular

$$y''(x_0) = \frac{\partial f(x_0, y_0)}{\partial x} + \frac{\partial f(x_0, y_0)}{\partial y} y'(x_0) = \frac{\partial f(x_0, y_0)}{\partial x} + \frac{\partial f(x_0, y_0)}{\partial y} f(x_0, y_0),$$

$$y^{(3)}(x_0) = \frac{\partial^2 f(x_0, y_0)}{\partial x^2} + 2\frac{\partial^2 f(x_0, y_0)}{\partial x \partial y} f(x_0, y_0) + \frac{\partial^2 f(x_0, y_0)}{\partial y^2} f(x_0, y_0)^2 +$$
$$+ \frac{\partial f(x_0, y_0)}{\partial y} \frac{\partial f(x_0, y_0)}{\partial x} + \left(\frac{\partial f(x_0, y_0)}{\partial y} \right)^2 f(x_0, y_0), \ldots$$

Each differentiation makes the formulae more cumbersome which results both in the increase of time needed for computation and (which is more essential) in possible loss of accuracy.

Besides finiteness of the radius of convergency of the series for the function puts additional restrictions to the applicability of the method of series. In order not to have cumbersome formulae one is forced to deal with small number of terms. This naturally diminishes the domain of x where the approximation (2.9) can be used.

The method of series can be modified. In a neighborhood of the initial point x_0 the truncated series (2.9) can be constructed. This approximation for the function $y(x)$ has acceptable accuracy only in a vicinity of x_0. Let a point x_1 be chosen in that vicinity and $y(x_1) = y_1$ be computed by the formula (2.9). Using that data the problem (2.8) can be reformulated as the initial-value problem with the initial data given at the point x_1. Again, the truncated series similar to (2.9) can be constructed in a neighborhood of the point x_1. In some vicinity of x_1 this approximation is acceptable and by choosing new point x_2 and computing $y(x_2)$, the initial problem (2.8) can be rewritten with the initial data specified at $x = x_2$. Repeating the steps of the procedure several times allows the desired value x_k of the argument x to be reached. The procedure results in a step-wise polynomial approximation of the solution $y(x)$ on the interval $x_0 < x < x_k$.

Note, that if only the principal order coefficients are stored, one gets a discrete approximation of the solution.

In order to use finite-difference methods, let the mesh x_k, $k = 1, 2, 3, \ldots$ to be introduced, and let the values of the function $y(x)$ at the nodes to be searched. Denoting $y(x_k) = y_k$, the computational rules of the form

$$y_{k+1} = Y_1(y_{k-q+1}, y_{k-q+2}, \ldots, y_k) \tag{2.10}$$

or

$$y_{k+1} = Y_2(y_{k-q+1}, y_{k-q+2}, \ldots, y_k, y_{k+1}) \tag{2.11}$$

can be introduced. The numerical method defined by the formula (2.10) is called q-step explicit method and the method defined by the formula (2.11) is q-step implicit method.

In the following sections we discuss one-step explicit methods, and multi-step explicit and implicit schemes.

2.1.4 Accuracy and stability

Every computational rule (2.10) or (2.11) is characterized by some local discretization error δ. The total discretization error is accumulated when the computational rule is used many times. Formulae (2.10) or (2.11) are usually taken in such a way, that the approximate equality $y(x_{k+1}) = y_{k+1}$ becomes the identity if the function $y(x)$ is any polynomial of some degree m. This can be achieved for example by the following method. Specify the function

Y in the form of expression containing parameters a_j, $j = 1, 2, \ldots, L$. Then decompose the right-hand side of (2.10) in a vicinity of point $x = x_{k+1}$ in Taylor series in powers of the step h up to h^m and demand the zero-order term to coincide with the value of the function $y(x)$ at the point $x = x_{k+1}$ and higher order terms to be equal to zero. These requirements yield a system of equations for the parameters a_j. This system is usually nonlinear and its solvability should be examined for each particular type of functions Y. If one manages to find such parameters a_j that satisfy the system and do not depend on the function $f(x, y)$, then the local discretization error of the computational rule (2.10) will be of order $O(h^{m+1})$. In order to find the value of the function $y(x)$ at a given point x, the computational rule should be applied approximately x/h times. If at each step the local discretization error is of order $O(h^{m+1})$, then on the whole interval the discretization error can increase up to the order $O(h^m)$.

Besides the error, as it was already remarked, there is another important characteristics of the computational rule, namely its stability. Let the problem of stability be studied taking as an example the equation

$$y'(x) = -Ay(x), \qquad A = \text{const}, \quad \text{Re}(A) > 0, \qquad y(0) = y_0. \tag{2.12}$$

The solution of this equation is the decreasing exponential

$$y(x) = y_0 \exp(-Ax)$$

and the stability criteria is satisfied for this solution. That means that for any $x \geq 0$ the estimate

$$|y(x)| \leq |y_0| \tag{2.13}$$

holds. Note, that for a general differential equation the solution is stable if

$$\text{Re}\left(\frac{\partial f}{\partial y}\right) \leq 0.$$

For the numerical scheme the condition similar to (2.13) is required, namely

$$|y_k| \leq |y_0|, \qquad k = 1, 2, 3, \ldots \tag{2.14}$$

Different types of stability of numerical schemes are distinguished. The method is called conditionally stable if one needs to set some restrictions

to the step h in order inequalities (2.14) to be satisfied. The method is called α-stable if no restriction to the step is needed if the value of A lies in the sector of the complex plane $\mathrm{Re}(A) > \tan(\alpha)|\mathrm{Im}(A)|$. The $\pi/2$-stable method is called absolutely stable.

2.2 One-step methods

One-step methods can be written in the form

$$y_{k+1} = Y(y_k).$$

The accuracy and stability of these methods depend on the function Y.

2.2.1 Euler method

The simplest one-step method is Euler method

$$y_{k+1} = y_k + f(x_k, y_k)h. \tag{2.15}$$

Here $h = x_{k+1} - x_k$ is the step. Euler method is a modified step-wise method of series with $m = 1$. This method can be also considered as based on the quadrature formula of left rectangles applied to the integral in the formula

$$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} f(t, y(t))\, dt.$$

The accuracy of Euler method is of low order, the local discretization error is of the second order by the step h and the total discretization error is of the first order, that is

$$y(x_{k+1}) - y_{k+1} = O(h^2), \qquad y(x) - y_{k(x)} = O(h).$$

Euler method is used as a starting component in more complicated schemes.

To analyse the stability of Euler method it is needed to be applied to problem (2.12). This yields

$$y_{k+1} = (1 - Ah)y_k.$$

It is evident that the stability requirement (2.14) is fulfilled if $h \leq 2\mathrm{Re}(A)/|A|^2$. If this condition is violated, it results in the infinite growth of $|y_k|$ for $k \to \infty$. Thus Euler method is conditionally stable with the stability condition

$$h \leq 2\mathrm{Re}\left(\frac{\partial f}{\partial y}\right)\left|\frac{\partial f}{\partial y}\right|^{-2}. \tag{2.16}$$

In the case of real equations this condition is reduced to

$$h \leq 2\left(\frac{\partial f}{\partial y}\right)^{-1}.$$

2.2.2 Runge-Kutta methods

Runge-Kutta methods are based on the idea of increasing the accuracy of the computational rule without differentiating the right-hand side of differential equation (2.8). The general scheme of y_k computation according to Ruge-Kutta method of order m can be presented in the following form

$$\begin{aligned}
&S_1 = f(x_k, y_k),\\
&S_2 = f(x_k + \alpha_1 h, y_k + h\beta_{11} S_1),\\
&S_3 = f(x_k + \alpha_2 h, y_k + h\beta_{21} S_1 + h\beta_{22} S_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&S_m = f\left(x_k + \alpha_m h, y_k + h\textstyle\sum_{j=1}^{m-1} \beta_{mj} S_j\right),\\
&y_{k+1} = y_k + h\textstyle\sum_{i=1}^{m} \gamma_i S_i.
\end{aligned}$$

The coefficients α_j, β_{ij} and γ_i are chosen such to achieve the highest posible order of the accuracy.

Consider the case $m = 2$ and illustrate the procedure of finding the coefficients α_1, β_{11}, γ_1 and γ_2. We start with the general form of the computational rule

$$\begin{aligned}
y_{k+1} = \; & y_k + h\Big(\gamma_1 f(x_k, y_k) + \\
& + \gamma_2 f\left(x_k + \alpha_1 h, y_k + \beta_{11} h f(x_k, y_k)\right)\Big).
\end{aligned} \tag{2.17}$$

The conditions for the coefficients presented in formula (2.17) are originated from the requirement that (2.17) coincides with Taylor series for the function $y(x)$ in as many terms as possible. It is easy to derive the decomposition of

the right-hand side of formula (2.17) into series by powers of the step h. Suppressing the arguments x_k, y_k of the function f and its derivatives, it can be written as

$$y_{k+1} = y_k + (\gamma_1 + \gamma_2)fh + \gamma_2 \left(\frac{\partial f}{\partial x}\alpha_1 + \frac{\partial f}{\partial y}\beta_{11} f \right) h^2 +$$

$$+ \gamma_2 \left(\frac{\partial^2 f}{\partial x^2}\alpha_1^2 + \frac{\partial^2 f}{\partial x \partial y}\alpha_1\beta_{11} f + \frac{\partial^2}{\partial y^2}\beta_{11}^2 f^2 \right) h^3 + \dots$$

Equating coefficients at powers of the step h in this formula and in (2.9), we find that coefficients at h^0 coincide automatically. Equating coefficients at h^1 yields the equation

$$\gamma_1 + \gamma_2 = 1.$$

At h^2 there are terms with different combinations of the function f and its derivatives, namely: $\partial f/\partial x$ and $f\ \partial f/\partial y$. As the method should work for differential equations with arbitrary right-hand side and the coefficients of the Runge-Kutta method should be independent of f, we equate coefficients at these combinations separately. That gives two more equations

$$\gamma_2\alpha_1 = 1/2, \qquad \gamma_2\beta_{11} = 1/2.$$

Hence there are 3 equations for 4 unknowns

$$\begin{cases} \gamma_1 + \gamma_2 = 1, \\ \gamma_2\alpha_1 = 1/2, \\ \gamma_2\beta_{11} = 1/2. \end{cases} \tag{2.18}$$

The remaining arbitrariness as it can be shown is insufficient to equate terms at h^3. Thus method (2.17) corresponding to any solution of system (2.18) has the third order for local discretization error and the second order for the total discretization error. Two variants of the second order Runge-Kutta method are usually used

$$S_2 = f(x_k + h, y_k + hS_1), \qquad y_{k+1} = y_k + \frac{S_1 + S_2}{2}h \tag{2.19}$$

and

$$S_2 = f\left(x_k + \frac{h}{2}, y_k + \frac{h}{2}S_1\right), \qquad y_{k+1} = y_k + S_2 h. \tag{2.20}$$

Formula (2.19) is the analog of trapezoid quadrature, and formula (2.20) corresponds to the mean rectangles. The approximate values of the function $y(x)$ required for both computational rules are computed by Euler method.

The application of (2.19) and (2.20) to problem (2.12) allows the stability conditions to be found. Schemes (2.19) and (2.20) for (2.12) are reduced to

$$y_{k+1} = \left(1 - \frac{h}{2}(A + A(1 - Ah))\right) y_k = \left(1 - Ah + \frac{1}{2}(Ah)^2\right) y_k$$

and

$$y_{k+1} = \left(1 - Ah\left(1 - A\frac{h}{2}\right)\right) = \left(1 - Ah + \frac{1}{2}(Ah)^2\right) y_k$$

correspondingly. Thus in both cases the schemes are conditionally stable with the same requirement (2.16) as for Euler method.

The derivation of the systems for the parameters of Runge-Kutta methods of higher orders is more cumbersome, but these systems can be derived and their solutions can be found. Only some examples of the 3-rd and of the 4-th order methods are presented here. One variant of the 3-rd order Runge-Kutta method corresponds to Simpson quadrature formula

$$\begin{aligned} &S_2 = f\left(x_k + \frac{1}{2}h, y_k + \frac{1}{2}hS_1\right), \quad S_3 = f(x_k + h, y_k - hS_1 + 2hS_2), \\ &y_{k+1} = y_k + \frac{h}{6}\Big(S_1 + 4S_2 + S_3\Big). \end{aligned} \tag{2.21}$$

Another computational scheme is

$$S_2 = f\left(x_k + \frac{1}{3}h, y_k + \frac{1}{3}hS_1\right), \quad S_3 = f\left(x_k + \frac{2}{3}h, y_k + \frac{2}{3}hS_1\right),$$

$$y_{k+1} = y_k + \frac{h}{4}\Big(S_1 + 3S_3\Big). \tag{2.22}$$

Another analog of Simpson formula

$$S_2 = f\left(x_k + \frac{1}{2}h, y_k + \frac{1}{2}hS_1\right), \quad S_3 = f\left(x_k + \frac{1}{2}h, y_k + \frac{1}{2}hS_2\right),$$

$$S_4 = f\left(x_k + h, y_k + hS_3\right),$$

$$y_{k+1} = y_k + \frac{h}{6}\Big(S_1 + 2S_2 + 2S_3 + S_4\Big) \tag{2.23}$$

and the analog of the method of "three eights"

$$S_2 = f\left(x_k + \frac{1}{3}h, y_k + \frac{1}{3}hS_1\right),$$

$$S_3 = f\left(x_k + \frac{2}{3}h, y_k - \frac{1}{3}hS_1 + hS_2\right),$$

$$S_4 = f\left(x_k + h, y_k + h(S_1 - S_2 + S_3)\right),$$

$$y_{k+1} = y_k + \frac{h}{8}\Big(S_1 + 3S_2 + 3S_3 + S_4\Big). \tag{2.24}$$

are the most popular among the 4-th order Runge-Kutta methods.

The stability conditions for these schemes are

$$h \le C\mathrm{Re}\left(\frac{\partial f}{\partial y}\right)\left|\frac{\partial f}{\partial y}\right|^{-2}, \tag{2.25}$$

where the coefficient C depends on the method: for (2.21) $C \approx 2.51$, for (2.22) $C = 2$, for (2.23) and (2.24) $C \approx 2.79$.

For Runge-Kutta methods with $m \le 4$ the local discretization error has the order $m + 1$. And there remain arbitrariness in the choice of the coefficients on the computational formula. For $m \ge 5$ the situation is different/ Now the number of coefficients appears insufficient to increase the accuracy. The local discretization error of order $O(h^6)$ can be achieved only in the methods with $m = 6$. One of the 6-th order Runge-Kutta methods is the following

$$S_2 = f\left(x_k + \frac{1}{3}h, y_k + \frac{1}{3}hS_1\right),$$

$$S_3 = f\left(x_k + \frac{2}{5}h, y_k + \frac{4}{25}hS_1 + \frac{6}{25}hS_2\right),$$

$$S_4 = f\left(x_k + h, y_k + \frac{1}{4}hS_1 - 3hS_2 + \frac{15}{4}hS_3\right),$$

$$S_5 = f\left(x_k + \frac{2}{3}h, y_k + \frac{2}{27}hS_1 + \frac{10}{9}hS_2 - \frac{50}{81}hS_3 + \frac{8}{81}hS_4\right),$$

$$S_6 = f\left(x_k + \frac{4}{5}h, y_k + \frac{2}{25}hS_1 + \frac{13}{25}hS_2 + \frac{2}{15}hS_3 + \frac{8}{25}hS_4\right),$$

$$y_{k+1} = y_k + \frac{h}{192}\Big(23S_1 + 125S_3 - 81S_5 + 125S_6\Big).$$

This example shows that with the increase of the degree of accuracy the formulae become much more complicated. Besides if the function $f(x, y)$ lacks smoothness, then the error can not be made smaller than the order of the last continuous derivative of f. The stability condition also becomes more restrictive. For the above scheme the constant C in condition (2.25) is approximately equal to 0.93.

Note that the results obtained according to Runge-Kutta methods do not provide information on the errors of discretization. In practice one needs to redo the computations twice with different steps h. For example one performs one step of computations with the step h to get approximate value $y^{(1)}$ and performs two steps of computations with the step $h/2$ to get another approximation $y^{(2)}$ for the same value of the unknown function $y(x)$. Then, knowing the order of accuracy of the method one can estimate the error comparing these two approximations. The approximations $y^{(1)}$ and $y^{(2)}$ can be also used to increase the accuracy. For example, let the discretization error of the scheme has the leading term

$$y(x_{k+1}) - y_{k+1} = Ch^n + O(h^{n+1})$$

with C depending on the derivatives of the function $y(x)$ and on the point x. Assuming this dependence sufficiently smooth, one can suggest the approximate value

$$y_{k+1} = \frac{2^{n-1}y^{(1)} - y^{(2)}}{2^{n-1} - 1}$$

for which the order of accuracy is one degree higher.

2.2.3 Methods of quadrature formulae

When using Runge-Kutta or other methods one needs to choose the step h. If the step is too large, then the discretization error will be large, if the step is too small, then the computational costs will increase. As it was already mentioned in order to estimate the errors in a particular numerical procedure and by this to check if the step size is chosen correctly one needs to do the computations twice. From this point of view it is convenient to suggest such

numerical schemes that allow the local discretization error to be estimated via some intermediate values used in the scheme. Some of such methods are considered below.

By integrating the differential equation in (2.8) can be reduced to

$$y_{k+1} = y_k + \int_{x_k}^{x_{k+1}} f(x, y(x))\, dx = y_k + h \int_0^1 f\Big(x_k + ht, y(x_k + ht)\Big) dt.$$

One can use a certain quadrature formula

$$\int_0^1 \Phi(t)\, dt \approx \sum_{i=0}^{N} W_i \Phi(t_i) \tag{2.26}$$

to calculate the integral, which yields

$$y_{k+1} \approx y_k + h \sum_{i=0}^{N} W_i f\left(x_k + ht_i, y(x_k + ht_i)\right). \tag{2.27}$$

The right-hand side of (2.27) contains values of the function $f(x, y)$ at $y = y(x_k + ht_i)$. These values (with the only exception of $N = 0$, $t_0 = 0$) are unknown. However due to the multiplier h before the sum in (2.27) it is sufficient to use lower order approximations for them. If the quadrature formula (2.26) has the accuracy of order $O(h^n)$, then the approximations for $y(x_k + ht_i)$ can be taken with the accuracy of order $O(h^{m-1})$. This will not reduce the total accuracy of formula (2.27). In order to find these approximate values one can use a formula similar to (2.27) based on a quadrature formula of lower accuracy

$$\int_0^1 \Phi(t)\, dt \approx \sum_{j=0}^{M} V_j \Phi(s_j).$$

This yields

$$y(x_k + ht_i) = y_k + ht_i \sum_{j=0}^{M} V_i f\left(x_k + ht_i s_j, y(x_k + ht_i s_j)\right).$$

The right-hand side can again contain unknown values of the function $y(x)$. They can be replaced with approximations having accuracy $O(h^{n-2})$. These approximations are also obtained by a formula similar to (2.27). Repeating these steps several times one comes to the formulae of the type (2.27) in the right-hand side of which it is sufficient to use approximations for $y(x)$ with the discretization error of order $O(h^2)$. Such approximations are given by Euler method (2.15).

Consider some examples of the method described above. Let $N = 0$ and the algebraic accuracy M of the quadrature formula (2.26) be equal to one. The most popular formulae of this type are the formula based on left rectangles quadrature formula which results in Euler method (2.15) and the formula of right rectangles which yields implicit method

$$y_{k+1} = y_k + hf(x_{k+1}, y_{k+1}). \tag{2.28}$$

Note that the computational formula (2.28) is absolutely stable. Indeed in the case of equation (2.12) one has

$$y_{k+1} = y_k - Ahy_{k+1},$$

where from

$$y_{k+1} = \frac{1}{1 - Ah} y_k,$$

and for any positive h (remind that $\mathrm{Re}(A) \geq 0$) one gets the estimate (2.14).

In a general case the computations according to the formula (2.28) require to solve a nonlinear equation with respect to y_{k+1}. For that an initial approximation $y_k^{(0)}$ is needed. Then it is iterated

$$y_{k+1}^{(j)} = y_k + hf\Big(x_{k+1}, y_{k+1}^{(j-1)}\Big), \qquad j = 1, 2, \ldots, J. \tag{2.29}$$

The initial approximation can be obtained by Euler method. Combining two computational rules with only one iteration (2.29) yields the formula

$$y_{k+1} = y_k + hf\Big(x_{k+1}, y_k + hf(x_k, y_k)\Big). \tag{2.30}$$

Note that the results obtained by Euler method and method (2.30) provide two-side approximations of the solution in the leading order by h. One can

apply the procedure described at the end of the previous section to increase the order of the accuracy. This results in Euler-Cauchy method

$$y_{k+1} = y_k + \frac{h}{2}\Big(f(x_k, y_k) + f\big(x_{k+1}, y_k + hf(x_k, y_k)\big)\Big) \tag{2.31}$$

which has the local discretization error of the 3-rd order.

Analysis of stability shows that the scheme (2.30) is stable if the condition (2.25) with $C = 1$ is satisfied and the stability condition for the scheme (2.31) is the same as for Euler method.

Let now $M = 2$, $N = 0$. The quadrature formula with one node which has algebraic accuracy equal to two ($M = 2$), that is the highest possible accuracy is the formula of mean rectangles. It leeds to the computational rule

$$\begin{aligned} y_{k+1/2} &= y_k + \tfrac{h}{2} f(x_k, y_k), \\ y_{k+1} &= y_k + h f(x_{k+1/2}, y_{k+1/2}). \end{aligned} \tag{2.32}$$

Here the notation $x_{k+1/2} = x_k + h/2$ is introduced for simplicity. Analogous notations are used below.

Consider the method with $M = 2$, $N = 1$, based on trapezoid quadrature formula. One has implicit method

$$y_{k+1} = y_k + \frac{h}{2}\Big(f(x_k, y_k) + f(x_{k+1}, y_{k+1})\Big).$$

Similar to the implicit method of rectangles (2.28) this formula is usually combined with Euler method

$$\begin{aligned} y^{[2]}_{k+1} &= y_k + h f(x_k, y_k), \\ y^{[3]}_{k+1} &= y_k + \tfrac{h}{2}\Big(f(x_k, y_k) + f(x_{k+1}, y^{[2]}_{k+1})\Big). \end{aligned} \tag{2.33}$$

By the upper indices in square brackets we denote here and below the order of local discretization error, that is

$$y^{[2]}_k = y(x_k) + O(h^2), \qquad y^{[3]}_k = y(x_k) + O(h^3).$$

Thus, method (2.33) gives two approximations for $y(x_k)$. Their difference allows the accuracy of computations to be estimated and by this to check if the step h is correctly chosen.

Note also that compared to rule (2.32) based on Gauss quadrature formula, rule (2.33) appears mere effective. Indeed, to perform one step of computations, that is to pass from $x = x_k$ to $x = x_{k+1}$ according to rule (2.32) the right-hand side of the equation should be calculated two times. When using the rule (2.33) the right-hand side should be computed two times, too. However, the value $f(x_{k+1}, y^{[2]}_{k+1})$ can be used at the next step if scheme (2.33) is modified to

$$\begin{aligned} y^{[2]}_{k+1} &= y_k + hf(x_k, y^{[2]}_k), \\ y^{[3]}_{k+1} &= y_k + \tfrac{h}{2}\Big(f(x_k, y^{[2]}_k) + f(x_{k+1}, y^{[2]}_{k+1})\Big). \end{aligned} \tag{2.34}$$

The accuracy of scheme (2.34) is a bit lower than that of the scheme (2.33), but has the same order $O(h^3)$.

The stability conditions for schemes (2.32), (2.33) and (2.34) are the same as for Euler method.

For $M = 3$, $N = 1$ one can use Gauss-Rado quadrature formula

$$\int_0^1 f(x)\,dx \approx \frac{1}{4}\Big(f(0) + 3f(2/3)\Big),$$

and mean rectangles formula for the computation of $y^{[3]}_{k+2/3}$. This gives the scheme

$$\begin{aligned} y^{[2]}_{k+1/3} &= y_k + \tfrac{h}{3} f(x_k, y_k), \\ y^{[3]}_{k+2/3} &= y_k + \tfrac{2h}{3} f(x_{k+1/3}, y^{[2]}_{k+1/3}), \\ y^{[4]}_{k+1} &= y_k + \tfrac{h}{4} f(x_k, y_k) + \tfrac{3h}{4} f(x_{k+2/3}, y^{[3]}_{k+2/3}), \end{aligned}$$

equivalent to the 3-rd order Runge-Kutta method (2.21).

One more example with $M = 3$, $N = 2$, is based on Simpson method

$$\begin{aligned} y^{[2]}_{k+1/2} &= y_k + \tfrac{h}{2} f(x_k, y_k), \\ y^{[3]}_{k+1/2} &= y_k + \tfrac{h}{4}\Big(f(x_k, y_k) + f(x_{k+1/2}, y^{[2]}_{k+1/2})\Big), \\ y^{[3]}_{k+1} &= y_k + hf(x_{k+1/2}, y^{[2]}_{k+1/2}), \\ y^{[4]}_{k+1} &= y_k + \tfrac{h}{6}\Big(f(x_k, y_k) + 4f(x_{k+1/2}, y^{[3]}_{k+1/2}) + f(x_{k+1}, y^{[3]}_{k+1})\Big). \end{aligned} \tag{2.35}$$

One step of computations requires the right-hand side of the differential equation to be calculated 4 times. Taking into account that the value $f(x_{k+1}, y^{[3]}_{k+1})$

can be used at the next step, one can conclude that the effectiveness of that method is the same as of Runge-Kutta method (2.21). However compared to Runge-Kutta method the method (2.35) allows the local discretization error to be estimated. Besides the stability of the method (2.35) is a bit better, since the coefficient C in the inequality (2.25) for this method is approximately equal to 2.9.

2.3 Multi-step methods

The advantage of one-step methods is in the possibility to apply the computational scheme starting from the very first step. In order to apply methods (2.10) with $q > 1$ one needs first to find several values y_j, $j = 1, 2, \ldots, q-1$ by some one-step method. However the use of values y_{k-1}, y_{k-2}, ... in the computational rule allows the accuracy of the method to be increased without additional computation of the right-hand side of the differential equation.

2.3.1 Extrapolation Adams methods

Multi-step Adams methods can be explicit or implicit. First, consider explicit methods. The general form of explicit Adams methods is as follows

$$y_{k+1} = y_k + h \sum_{i=0}^{q-1} \alpha_i f(x_{k-i}, y_{k-i}). \tag{2.36}$$

The computational rule (2.36) is based on the computation of the integral in the formula

$$y_{k+1} = y_k + \int_{x_k}^{x_{k+1}} f(x, y(x))\, dx \tag{2.37}$$

according to the approximate formula in which the integrand is replaced by the polynomial interpolating it at the points x_k, x_{k-1}, ... , x_{k-q+1}. Therefore formulae (2.36) are also called extrapolation formulae. The coefficients α_i are uniquely determined by the integration of the interpolating polynomial

$$\alpha_i = \int_{x_k}^{x_{k+1}} L_i(x)\, dx. \tag{2.38}$$

Here

$$L_i(x) = \prod_{j \neq i} \frac{x - x_{k-j}}{x_{k-i} - x_{k-j}}$$

are Lagrange polynomials.

Note that the coefficients α_i differ from the weights in the usual quadrature formulae because the integration in (2.38) is carried out not along the interval of interpolation $[x_{k-q+1}, x_k]$, but outside of it.

When using the rule (2.36) the right-hand side of the differential equation is computed only once at a step. Nevertheless the accuracy can be of sufficiently high order by h. This is the main advantage of multi-step methods compared to one-step methods such as Runge-Kutta methods. The disadvantages are twofold. Firstly, as it was already noted, one needs to know q values of the function in the previous nodes. Secondly, Adams methods do not allow to change the step of integration so easily because the coefficients α_i depend on the positions of nodes. The simplest case when the step h is constant is considered below.

The method with $q = 2$ is

$$y_{k+1} = y_k + \frac{h}{2}\Big(3f_k - f_{k-1}\Big). \tag{2.39}$$

Here $f(x_\ell, y_\ell)$ is denoted as f_ℓ. The local discretization error of (2.39) is given by the formula

$$y(x_{k+1}) - y_{k+1} = \frac{5}{12} y'''(x_k) h^3 + O(h^4).$$

Consider the stability property of method (2.39). In the case of differential equation (2.12) it gives

$$y_{k+1} = y_k - \frac{Ah}{2}(3y_k - y_{k-1}).$$

Substituting here $y_k = \lambda^k$ yields characteristic equation

$$\lambda^2 - \left(1 - \frac{3}{2}Ah\right)\lambda - \frac{Ah}{2} = 0. \tag{2.40}$$

For any positive Ah the discriminant $D = 1 - ah + \frac{9}{4}(Ah)^2$ of this equation is positive. Hence, the zeros λ_1 and λ_2 are real and $\lambda_1 \neq \lambda_2$. Scheme (2.39)

is stable if $|\lambda_1| \leq 1$ and $|\lambda_2| \leq 1$. It is not difficult to check that for real A these conditions are satisfied if

$$\left|\frac{3}{2}Ah - 1\right| \leq 1 - \frac{Ah}{2}$$

or $Ah \leq 1$. Thus the scheme (2.39) is conditionally stable with restriction (2.25) with $C = 1$. Compared to the second order Runge-Kutta methods the stability condition sets stronger restrictions to the step h.

For $q = 3$ one gets the formula

$$y_{k+1} = y_k + \frac{h}{12}\Big(23f_k - 16f_{k-1} + 5f_{k-2}\Big),$$

with the local discretization error

$$y(x_{k+1}) - y_{k+1} = \frac{3}{8}y^{(4)}(x_k)h^4 + O(h^5),$$

and for $q = 4$ the formula is

$$y_{k+1} = y_k + \frac{h}{24}\Big(55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}\Big) \tag{2.41}$$

with local discretization error

$$y(x_{k+1}) - y_{k+1} = \frac{251}{720}y^{(5)}(x_k)h^5 + O(h^6).$$

The characteristic equations for these methods are

$$\lambda^3 - \left(1 - \frac{23}{12}Ah\right)\lambda^2 - \frac{4}{3}Ah\lambda + \frac{5}{12} = 0$$

and

$$\lambda^4 - \left(1 - \frac{55}{24}Ah\right)\lambda^3 - \frac{59}{24}Ah\lambda^2 + \frac{37}{24}Ah\lambda - \frac{3}{8}Ah.$$

The conditions of stability are $Ah \leq 0.545$ for the three-step method and $Ah \leq 0.3$ for the four-step method.

2.3.2 Interpolation Adams methods

Implicit Adams methods (also called interpolation schemes) have the form

$$y_{k+1} = y_k + h \sum_{i=0}^{q-1} \beta_i f(x_{k+1-i}, y_{k+1-i}). \tag{2.42}$$

The coefficients β_i are computed by integrating the interpolation polynomials constructed on the nodes x_{k+1}, x_k, ... , x_{k-q+2}.

Formula (2.42) contains the unknown value y_{k+1} in the right-hand side and requires solving a nonlinear equation. The method of simple iterations

$$y_{k+1}^{(j)} = y_k + h \sum_{i=0}^{q-2} \beta_{i+1} f(x_{k-i}, y_{k-i}) + h\beta_0 f(x_{k+1}, y_{k+1}^{(j-1)}), \qquad j = 1, 2, \ldots$$

appears to be the most natural for that. This process converges because the derivative of the right-hand side of (2.42) by y_{k+1} is small due to the multiplier h.

To start this procedure one needs the initial approximation $y_{k+1}^{(0)}$ which can be obtained with the help of the explicit Adams method of the corresponding order. Such combinations of explicit and implicit methods are known as predictor-corrector method.

Some implicit Adams schemes are presented below. For $q = 2$, one has the scheme of trapezoid method

$$y_{k+1} = y_k + \frac{h}{2}\Big(f(x_{k+1}, y_{k+1}) + f_k\Big). \tag{2.43}$$

The local discretization error can be written as

$$y(x_{k+1}) - y_{k+1} = -\frac{1}{12} y'''(x_k) h^3 + O(h^4).$$

For $q = 3$ one has the formula

$$y_{k+1} = y_k + \frac{h}{12}\Big(5f(x_{k+1}, y_{k+1}) + 8f_k - f_{k-1}\Big) \tag{2.44}$$

with the local discretization error

$$y(x_{k+1}) - y_{k+1} = -\frac{1}{24} y^{(4)}(x_k) h^4 + O(h^5),$$

and for $q = 4$ the formula is

$$y_{k+1} = y_k + \frac{h}{24}\Big(9f(x_{k+1}, y_{k+1}) + 19f_k - 5f_{k-1} + f_{k-2}\Big) \tag{2.45}$$

and its local discretization error is

$$y(x_k) - y_k = -\frac{19}{720}y^{(5)}(x_{k-1})h^5 + O(h^6).$$

Comparing these formulae with the corresponding explicit schemes one can note that the local discretization errors have the same order, but the coefficients in the estimates are smaller. Another advantage of implicit schemes is in their better stability. The scheme (2.43) is absolutely stable. Indeed in the case of equation (2.12) one has

$$y_{k+1} = y_k - \frac{Ah}{2}\Big(y_{k+1} + y_k\Big),$$

where from

$$y_{k+1} = \frac{1 - Ah/2}{1 + Ah/2}y_k.$$

For the stability of the method the absolute value of the fraction in the right-hand side of this formula should not exceed unity which is the case for any positive $\mathrm{Re}(A)h$.

In fact method (2.43) is the most accurate absolutely stable linear method.

The characteristic equation for method (2.44) is

$$\left(1 + \frac{5}{12}Ah\right)\lambda^2 - \left(1 - \frac{2}{3}Ah\right)\lambda + \frac{Ah}{12} = 0.$$

Its zeros can be easily found

$$\lambda_{1,2} = \frac{6 - 4Ah \pm \sqrt{36 - 36Ah + 21(Ah)^2}}{12 + 5Ah}$$

For any $Ah > 0$ the zero λ_1 is positive and is less than unity. The zero λ_2 is negative and monotonously decreases with the step. Solving the inequality $\lambda_2 > -1$ one can find the maximal step. It holds

$$18 + Ah \geq \sqrt{36 - 36Ah + 21(Ah)^2},$$

which yields $Ah \leq 6$. Thus the method (2.44) is conditionally stable with the restriction

$$h \leq 6 \left(\frac{\partial f}{\partial y} \right)^{-1}.$$

Similar analysis of the characteristic equation

$$\left(1 + \frac{3}{8}Ah\right)\lambda^3 - \left(1 - \frac{19}{24}Ah\right)\lambda^2 - \frac{5}{24}Ah\lambda + \frac{Ah}{24} = 0$$

for the four-step method (2.45) allows the stability condition to be found as

$$h \leq 3 \left(\frac{\partial f}{\partial y} \right)^{-1}.$$

2.3.3 BDF methods

The described above Adams methods exploit the idea of replacing the integrand in the formula (2.37) with interpolation polynomial. In the backward differentiation formula (BDF) methods the solution $y(x)$ is interpolated itself. Namely, in the BDF method of order n one performs polynomial interpolation of the known data $\{x_{k-n+1}, y_{k-n+1}\}$, ... , $\{x_k, y_k\}$ and unknown value y_{k+1} of the function $y(x)$ at the point x_{k+1}. The constructed interpolation polynomial $P_n(x)$ is substituted into the differential equation and the unknown value y_{k+1} is found from the requirement that this equation is satisfied at the point $x = x_{k+1}$, that is

$$\frac{\partial P_n(x_{k+1})}{\partial x} = f\left(x_{k+1}, P_n(x_{k+1})\right).$$

We present some examples. In the first order method

$$P_1 = y_k + \frac{y_{k+1} - y_k}{h}(x - x_k),$$

which yields the equation

$$y_{k+1} - y_k = hf(x_{k+1}, y_{k+1}) \tag{2.46}$$

for the unknown y_{k+1}.

In the second order method the expression

$$\frac{\partial P_2(x_{k+1})}{\partial x} = \frac{1}{h}\left(\frac{1}{2}y_{k-1} - 2y_k + \frac{3}{2}y_{k+1}\right).$$

for the derivative of the interpolation polynomial is used. In that case the equation for y_{k+1} is

$$\frac{1}{2}y_{k-1} - 2y_k + \frac{3}{2}y_{k+1} = hf(x_{k+1}, y_{k+1}). \tag{2.47}$$

Finally the third-order BDF method is

$$\frac{11}{6}y_{k-2} - 3y_{k-1} + \frac{3}{2}y_k - \frac{1}{3}y_{k+1} = hf(x_{k+1}, y_{k+1}). \tag{2.48}$$

Equation (2.46) already appeared (see (2.28)) as the result of application of right rectangles quadrature formula. As it was shown this scheme is absolutely stable. It can be also shown that scheme (2.47) is absolutely stable.

2.4 Systems of equations and equations of the second and higher orders

Any differential equation of n-th order can be reduced to a system of n differential equations of the first order. All the above methods can be applied to such systems. Nevertheless the specific features of systems derived from higher order differential equations allow some times to suggest more effective algorithms. Moreover, the specifics of the higher order differential equation can be also taken into account. For example, effective methods are developed for those second order differential equations which do not contain the first order derivative.

We illustrate the specifics of the initial-value problems for higher order differential equations taking as an example the equation

$$y''(x) = f\big(x, y(x), y'(x)\big) \tag{2.49}$$

with the initial data

$$y(x_0) = y_0, \qquad y'(x_0) = y'_0.$$

In order to pass from the argument x_k to x_{k+1} one needs to compute the value of the function y_{k+1} and the value of its derivative y'_{k+1}. We rewrite

equation (2.49) once and twice integrated from x_k to x_{k+1}

$$y'(x_{k+1}) = y'_k + h \int_0^1 f\big(x_k + ht, y(x_k + ht), y'(x_k + ht)\big)\, dt, \tag{2.50}$$

$$y(x_{k+1}) = y_k + hy'_k + h^2 \int_0^1 (1-s) f\big(x_k + hs, y(x_k + hs), y'(x_k + hs)\big)\, ds. \tag{2.51}$$

Applying quadrature formulae one gets

$$y'_{k+1} = y'_k + h \sum_{i=0}^{N} W_i f\big(x_k + ht_i, y(x_k + ht_i), z(x_k + ht_i)\big), \tag{2.52}$$

$$y_{k+1} = y_k + hy'_k + h^2 \sum_{i=0}^{N'} W'_i (1 - s_i) f\big(x_k + hs_i, y(x_k + hs_i), z(x_k + hs_i)\big). \tag{2.53}$$

Note that the quadrature formulae used for computing the integrals in (2.50) and (2.51) can be taken different. If the nodes coincide, $t_i = s_i$, the number of calculations of the right-hand side of the differential equation is reduced. One can also use the fact that the integral in (2.51) is multiplied by h^2 and the integral in (2.50) is multiplied by h. Hence the quadrature formula for the integral in (2.51) can have a lower by one order of accuracy than the quadrature formula for the integral in (2.50). Finally note that the multiplier $1 - s$ in (2.51) can be treated as the weight. Denoting $W_i(1 - s_i) = V_i$ and requesting the algebraic accuracy of (2.52) and (2.53) to be equal to some number m, one gets the system of nonlinear equations for the quantities t_i, W_i and s_i, V_i

$$\sum_{i=0}^{N} W_i t_i^j = \frac{1}{j+1}, \qquad j = 1, 2, 3, \ldots, m-1,$$

$$\sum_{i=0}^{N'} V_i s_i^j = \frac{1}{j(j+1)}, \qquad j = 1, 2, 3, \ldots, m-2.$$

In order this system to be solvable it is necessary that $m \leq 2N + 2$ and $m \leq 2N' + 3$.

For example, in the case of $m = 2$ one can use the method of trapezoids and the method of left rectangles. Changing the sequence of the formulae one has

$$\begin{aligned} y_{k+1} &= y_k + hy'_k + \tfrac{1}{2}h^2 f_k, \\ y'_{k+1} &= y'_{k+1} + \tfrac{h}{2}\left(f_k + f(x_{k+1}, y_{k+1}, y'_k + hf_k)\right). \end{aligned} \tag{2.54}$$

Here $f_k = f(x_k, y_k, y'_k)$. Both formulae in (2.54) have the local discretization error of order $O(h^3)$. Adding the formula

$$y_{k+1} = y_k + hy'_k + \frac{h^2}{6}\Big(2f_k + f(x_{k+1}, y_{k+1}, y'_k + hf_k)\Big), \tag{2.55}$$

which does not require computing the right-hand side of the equation, yields the method of the 4-th order. Formulae (2.54) and (2.55) also allow the discretization error to be estimated by comparing the approximations

$$y'^{[2]}_{k+1} = y'_k + hf_k, \qquad y^{[3]}_{k+1} = y_k + hy'_k + \frac{1}{2}h^2 f_k$$

with more accurate ones

$$y'^{[3]}_{k+1} = y'_{k+1} + \frac{h}{2}\Big(f_k + f(x_{k+1}, y_{k+1}, y'_k + hf_k)\Big),$$

$$y^{[4]}_{k+1} = y_k + hy'_k + \frac{h^2}{6}\Big(2f_k + f(x_{k+1}, y_{k+1}, y'_k + hf_k\Big).$$

Consider now the case of differential equation not containing the first order derivative

$$y''(x) = f\big(x, y(x)\big). \tag{2.56}$$

Note that any linear equation of the second order can be transformed to the form (2.56). The usual methods for the equations of the type (2.56) are the explicit method

$$y_{k+1} - 2y_k + y_{k-1} = h^2 \sum_{i=0}^{N} \alpha_i f(x_{k-i}, y_{k-i})$$

and the implicit method

$$y_{k+1} - 2y_k + y_{k-1} = h^2 \sum_{i=0}^{N} \beta_i f(x_{k+1-i}, y_{k+1-i}).$$

The coefficients α_i and β_i are derived from the requirement that the difference of the left-hand sides and the right-hand sides of these equations be of the highest possible power of the step h.

The simplest explicit method is Verlet scheme

$$y_{k+1} - 2y_k + y_{k-1} = h^2 f(x_k, y_k)$$

widely used in molecular dynamics simulation.

One of the implicit methods of the 4-th order of accuracy is the Numerov method

$$y_{k+1} - 2y_k + y_{k-1} = \frac{h^2}{12}\Big(f(x_{k-1}, y_{k-1}) + 10f(x_k, y_k) + f(x_{k+1}, y_{k+1})\Big). \tag{2.57}$$

(See also chapter 3).

2.5 Stiff systems

Some systems of differential equations appear specifically difficult for numerical solution. Usually this is the case when the processes described by the equations in the system have significantly different scales of variation. Most of the methods described above require too small step h and as a sequence too large amount of computations. The development of effective algorithms for numerical solution of such systems is an actual problem in the theory of numerical methods.

Consider the example of the second order differential equation

$$y''(x) + 1001y'(x) + 1000y(x) = 0.$$

The general solution of this equation is

$$y(x) = C_1 e^{-x} + C_2 e^{-1000x}.$$

Let the initial condition be $y(0) = 1$, $y'(0) = -1$. Then the solution is

$$y(x) = e^{-x}.$$

Now try to find this solution numerically. Use for example the 4-th order Runge-Kutta method (2.23). Rewrite the equation in the form of the system

$$\begin{cases} z_1' = z_2, \\ z_2' = -1001z_2 - 1000z_1. \end{cases} \tag{2.58}$$

The stability condition for the method (2.23) applied to the system (2.58) is

$$h \leq 0.00279.$$

Indeed, Runge-Kutta method with the step $h = 0.0025$ gives a good approximation, but it requires 400 steps and 1200 computations of the right-hand side of the system in order to find $y(1)$. Though the solution is slowly varying the increase of the step even to $h = 0.003$ causes divergence of the solution. Thus the condition of stability is defined by the most quick function e^{-1000x} in the general solution of the equation independently on if it is presented or not in the solution of the initial-value problem. Similar difficulties appear when other single-step or multi-step methods are used.

2.5.1 Implicit trapezoids rule

One of the most satisfactory tool to solve system (2.58) is the use of implicit method (2.43) which is stable for any step h. For the system (2.58) the method can be written as

$$\begin{cases} z_1(x_{k+1}) = z_1(x_k) + \dfrac{h}{2}\Big(z_2(x_k) + z_2(x_{k+1})\Big), \\ z_2(x_{k+1}) = z_2(x_k) - \dfrac{h}{2}\Big(1001z_2(x_k) + 1000z_1(x_k) \\ \qquad\qquad + 1001z_2(x_{k+1}) + 1000z_1(x_{k+1})\Big) \end{cases}$$

In the general case the system of algebraic equations at each step requires application of some iterative method. However in our case the solution can be found explicitly

$$\begin{cases} z_1(x_{k+1}) = \dfrac{2 + 1001h - 500h^2}{2 + 1001h + 500h^2} z_1(x_k) + \dfrac{2h}{2 + 1001h + 500h^2} z_2(x_k), \\ z_2(x_{k+1}) = -\dfrac{2000h}{2 + 1001h + 500h^2} z_1(x_k) + \dfrac{2 - 1001h - 500h^2}{2 + 1001h + 500h^2} z_2(x_k) \end{cases}$$

Choosing the step $h = 0.1$, one gets

$$z_1(0.1) = 0.9047619048, \qquad z_2(0.1) = -0.9047619048,$$

which is a good approximation to the exact solution

$$y(0.1) = 0.9048374180, \qquad y'(0.1) = -0.9048374180.$$

Repeating the computations, after 10 steps one finds $z_1(1) = 0.3675725425$ which is rather close to the exact value $y(1) = e^{-1} = 0.3678794412$.

The implicit trapezoid method can be similarly applied to other stiff systems. If the equations that appear at each step of the procedure can not be solved explicitly, one needs to use iterative methods. Usually (for not stiff systems) it is the method of simple iterations. The method of simple iterations converges if the Jacobian

$$J_{ij} = \frac{\partial f_i(x, \mathbf{z})}{\partial z_j}.$$

satisfies the condition

$$\frac{h}{2}\|J\| < 1. \tag{2.59}$$

In the case of stiff problems the Jacobian J contains large elements and the condition (2.59) significantly shrinks the step h.

To overcome this difficulty there is a useful modification of implicit trapesoids method. This modification is in the use of Newton method. Then no difficulties with convergence arising from large negative derivatives $\partial f_i/\partial z_j$ are encountered. However, each iteration requires computation of the Jacobian.

2.5.2 Implicit Runge-Kutta methods

We already remarked that the implicit trapezoid methods (2.43) is the most accurate among absolutely stable linear methods. The increase of accuracy is possible if nonlinear schemes are used. We consider only the methods of Runge-Kutta type. In the section 2.2.2 explicit formulae for the quantities S_i were used. In order to increase the stability of the scheme, let us use the formulae

$$S_i = f(x_k + \alpha_i h, y_k + h\sum_{j=1}^{m} \beta_{ij} S_j, \qquad i = 1, 2, \ldots, m.$$

That is, finding S_i requires now solution of the system of nonlinear equations. If $\beta_{ij} = 0$ for $j > i$ this system is triangular and can be approximately linearized with one Newton iteration

$$S_i = \left[1 - h\beta_{ii} f_y \left(x_k + \alpha_i h, y_k + h \sum_{j=1}^{i-1} \beta_{ij} S_j\right)\right]^{-1} \times \\ \times f\left(x_k + \alpha_i h, y_k + h \sum_{j=1}^{i-1} \beta_{ij} S_j\right). \tag{2.60}$$

The value y_{k+1} is computed by the formula

$$y_{k+1} = y_k + h \sum_{i=1}^{m} \gamma_i S_i \tag{2.61}$$

as in the conventional Runge-Kutta methods. The methods, based on the formulae (2.60) and (2.61) are called linearized semi-explicit Runge-Kutta methods or Rozenbrok methods.

Consider the example of the method with $m = 2$. In that case there are 7 coefficients that should be chosen to achieve highest possible accuracy. As in the case of conventional Runge-Kutta methods we equate coefficients in the series decompositions of the right-hand side of formula (2.61) and in Taylor series for the solution $y(x_k+h)$. The leading order coefficients coincide automatically. Equating coefficients at first powers of h yields equation

$$\gamma_1 + \gamma_2 = 1.$$

At h^2 there are terms containing f_x and $f_y f$. Equating coefficients at these terms separately yields the equations

$$\gamma_1\alpha_1 + \gamma_2\alpha_2 = 1/2,$$
$$\gamma_1\beta_{11} + \gamma_2\beta_{21} + \gamma_2\beta_{22} = 1/2.$$

At the next order there are combinations f_{xx}, $f_x f_y$, $f_{yy} f^2$, $f_{xy} f$ and $f_y^2 f$, which yields equations

$$\gamma_1\alpha_1^2 + \gamma_2\alpha_2^2 = 1/3,$$
$$\gamma_1\alpha_1\beta_{11} + \gamma_1\alpha_1\beta_{21} + \gamma_2\alpha_2\beta_{22} = 1/6,$$
$$\gamma_2\beta_{21}^2 + 2\gamma_2\beta_{21}\beta_{22} = 1/3,$$
$$\gamma_1\alpha_1\beta_{11} + \gamma_2\alpha_2\beta_{21} + \gamma_2\alpha_2\beta_{22} = 1/3,$$
$$\gamma_1\beta_{11}^2 + \gamma_2\beta_{11}\beta_{21} + \gamma_2\beta_{21}\beta_{22} + \gamma_2\beta_{22}^2 = 1/6.$$

Though there are 8 equations and there are only 7 parameters, there exist two solutions

$$\begin{array}{llll} \gamma_1 = \frac{1}{2} \pm \frac{i}{6}\sqrt{3}, & \gamma_2 = \frac{1}{2} \mp \frac{i}{6}\sqrt{3}, & & \\ \alpha_1 = \frac{1}{4} \pm \frac{i}{12}\sqrt{3}, & \alpha_2 = \frac{3}{4} \pm \frac{i}{12}\sqrt{3}, & & \\ \beta_{11} = \frac{1}{4} \pm \frac{i}{12}\sqrt{3}, & \beta_{21} = \frac{1}{2} \pm \frac{i}{6}\sqrt{3}, & \beta_{22} = \frac{1}{4} \mp \frac{i}{12}\sqrt{3}. & \end{array}$$

The disadvantage of such approach is in the use of complex calculus. However, as it can be easily checked, the method is absolutely stable. Indeed for (2.12) it reduces to

$$y_{k+1} = \frac{1 - \frac{1}{2}Ah + \frac{1}{12}(Ah)^2}{1 + \frac{1}{2}Ah + \frac{1}{12}(Ah)^2}\, y_k.$$

2.6 Special cases

2.6.1 Singular points

Differential equations for which the right-hand side or its derivative of some order can be infinite at some points of the interval present additional difficulty for numerical solving. Let the initial point x_0 be singular, that is, $f(x_0, y_0)$ does not exist. Evidently that no explicit method can be applied to such initial-value problem. Although the use of some implicit schemes is possible, they give wrong results.

There are three main approaches to finding solutions of a differential equation with singular initial point

1. Change of variables that eliminates the singularity.

2. Construction of an approximate analytic solution in a small vicinity of the singular point and passage to another initial point with the help of this solution.

3. Development of special computational schemes that take into account the specifics of the right-hand side of the differential equation.

Let us illustrate the above approaches taking as an example the following problem

$$\begin{cases} y'(x) = \frac{1}{2\sqrt{x}} - y^2(x), \\ y(0) = 0. \end{cases} \tag{2.62}$$

In the first approach one can make the substitution $t = \sqrt{x}$ and denote $y(x) = Y(t)$. Then the problem (2.62) is transformed to

$$\begin{cases} Y'(t) = 1 - 2tY^2(t), \\ Y(0) = 0, \end{cases}$$

which has no singularity.

In the second approach one can apply for example Picar method. Setting the initial approximation as $y_0(x) \equiv 0$ the problem (2.62) can be rewritten in the form of iterative rule

$$y_{j+1}(x) = \int_0^x \left(\frac{1}{2\sqrt{s}} - y_j^2(s) \right) ds.$$

It determines

$$y_1(x) = \sqrt{x}, \quad y_2(x) = \sqrt{x} - \frac{x^2}{2}, \quad y_3(x) = \sqrt{x} - \frac{x^2}{2} + \frac{2}{7}x^3\sqrt{x} - \frac{x^5}{20}.$$

The above approximations give

$$y_1(0.1) \approx 0.3162, \quad y_2(0.1) \approx 0.3112, \quad y_3(0.1) \approx 0.3113,$$

which allows the initial data $y(0) = 0$ in the problem (2.62) to be moved to the point $x = 0.1$.

In the third approach one can construct a special scheme for equation (2.62). For that the equation should be rewritten in the integral form

$$y_{k+1} = y_k + \int_{x_k}^{x_{k+1}} \left(\frac{1}{2\sqrt{s}} - y^2(s) \right) ds.$$

The first term can be integrated analytically and left rectangles quadrature formula can be applied to the second term. This results in the explicit scheme similar to Euler method

$$y_{k+1} = y_k + \sqrt{x_{k+1}} - \sqrt{x_k} - hy_k^2.$$

If the right-hand side of the equation has singularities at the internal points of the interval $[x_0, x]$, then since in general $f(x, y)$ depends on the unknown solution $y(x)$ it is impossible to say in advance at what points there will be singularities. Due to this fact it is preferable to apply the approach of special schemes adequate to such problems.

2.6.2 Special schemes

The methods of special schemes construction are based on the study and use of the properties of general solutions for a class of similar equations. The main idea is in the choice of appropriate representation for the solution. For example, one can replace some functions of the argument x in the initial differential equation by constants. This enables to transform the equation

$$y'(x) = f\big(x, y(x)\big)$$

to some other equation

$$u'(x) = g\big(x, u(x)\big)$$

that has explicit solution $u(x)$.

Further several approaches are possible. Either to consider the ratio $v(x) = y(x)/u(x)$, the difference $w(x) = y(x) - u(x)$ or some other more complicated combination of $y(x)$ and $u(x)$ and derive for this function a new initial-value problem. The equation for $v(x)$ is

$$v'(x) = \frac{f\big(x, v(x)u(x)\big) - v(x)g\big(x, u(x)\big)}{u(x)}$$

and for $w(x)$ it is

$$w'(x) = f\big(x, w(x) + u(x)\big) - g\big(x, u(x)\big).$$

If $u(x)$ sufficiently well approximates the solution $y(x)$, then the new initial-value problem for $v(x)$ or $w(x)$ appears to be more appropriate for numeric solution.

In the other approach the auxiliary solution $u(x)$ is used on a small interval of the argument x in a similar way as in the step-wise variant of the method of series the truncated Taylor series was used. We illustrate it taking the example of the equation

$$y'(x) = -y^2(x) - \rho(x).$$

If $\rho(x) = const$ the equation

$$u(x)' = -u^2(x) - r^2$$

has explicit solution

$$u(x) = -r \tan(rx + C).$$

The constant C is found from the condition $u(x_k) = u_k$ and further $u(x_{k+1})$ is determined. This results in the scheme

$$\operatorname{arctg}\left(\frac{u_{k+1}}{r}\right) - \operatorname{arctg}\left(\frac{u_k}{r}\right) = -hr.$$

To rewrite this scheme for the function $y(x)$ one takes $r = \sqrt{\rho(x_{k+1/2})}$ at each step,

$$\operatorname{arctg}\left(\frac{y_{k+1}}{\sqrt{\rho_{k+1/2}}}\right) - \operatorname{arctg}\left(\frac{y_k}{\sqrt{\rho_{k+1/2}}}\right) = -h\sqrt{\rho_{k+1/2}}.$$

Solving this equation for y_{k+1}, yields the special explicit scheme

$$y_{k+1} = \sqrt{\rho_{k+1/2}}\frac{y_k - \sqrt{\rho_{k+1/2}}\tan\left(h\sqrt{\rho_{k+1/2}}\right)}{\sqrt{\rho_{k+1/2}}\tan\left(h\sqrt{\rho_{k+1/2}}\right)}. \tag{2.63}$$

It appears that scheme (2.63) is valid even if the stability condition is violated and the exact solution has poles.

2.6.3 Equations which are not resolved for the highest order derivative

Consider finally equations (2.1) that can not be represented in the form (2.4). Let it be a first order equation

$$F\big(x, y(x), y'(x)\big) = 0. \tag{2.64}$$

We can formally write

$$y'(x) = f\big(x, y(x)\big),$$

where for every x and y the value of the function f is defined as the solution of the equation

$$F\big(x, y, f(x, y)\big) = 0. \tag{2.65}$$

For solving equation (2.64) one can use explicit methods in which the values of $f(x, y)$ are determined as solutions of nonlinear equation (2.65) with the use of some iterative method. The value obtained by interpolation can serve as the initial approximation for $f(x, y)$. This usually reduces the number of iterations.

When using implicit schemes, for example of the form (2.42), it is needed to solve the system

$$\begin{cases} y_{k+1} = y_k + h \sum_{i=0}^{q-1} \beta_i f_{k+1-i}, \\ F(x_{k+1}, y_{k+1}, f_{k+1}) = 0 \end{cases}$$

for unknowns y_{k+1} and f_{k+1}.

Note also that by increasing the order of an equation it is possible to derive from the differential equation (2.1) a differential equation in the form (2.4). Differentiation of equation (2.64) yields

$$F_x(x, y, y') + F_y(x, y, y')y' + F_{y'}(x, y, y')y'' = 0,$$

which can be resolved for the second order derivative

$$y''(x) = -\frac{F_x(x, y, y') + F_y(x, y, y')y'}{F_{y'}(x, y, y')}. \tag{2.66}$$

Chapter 3

Numerical methods for boundary-value problems

3.1 Introduction

3.1.1 Boundary-value problem formulation

For the second and higher order differential equations boundary-value problems and spectral problems can be formulated. Here mainly linear problems are considered, that is the differential equation and the boundary conditions are assumed to be linear. In the general form the second order linear boundary value problem can be writen as

$$\begin{cases} y''(x) + p(x)y'(x) + q(x)y(x) = f(x), & a < x < b, \\ \alpha_0 y(a) + \alpha_1 y'(a) = A, \\ \beta_0 y(b) + \beta_1 y'(b) = B. \end{cases} \tag{3.1}$$

Here $p(x)$, $q(x)$ and $f(x)$ are given functions and α_0, α_1, β_0, β_1, A and B are given constants.

Note that if the solution $y(x)$ should be found not only on the interval $[a, b]$, but also in its exterior, then after solving the boundary-value problem, that is after determining $y(x)$ for $x \in [a, b]$, the computation of $y(x)$ for $x < a$ and for $x > b$ is performed by solving the corresponding Cauchy problems.

Equation (3.1) can be conveniently represented in the form not containing the first order derivative. This can be achieved for example by introducing

the new unknown $u(x)$ by the formula

$$y(x) = \exp\left(-\frac{1}{2}\int^{x} p(t)\,dt\right) u(x).$$

The function $u(x)$ satisfies the equation

$$u''(x) + Q(x)u(x) = F(x), \tag{3.2}$$

where

$$Q(x) = q(x) - \frac{1}{2}p'(x) - \frac{1}{4}p^2(x), \qquad F(x) = \exp\left(\frac{1}{2}\int^{x} p(t)\,dt\right) f(x).$$

The boundary conditions formulated at the points $x = a$ and $x = b$ can be of three types: the conditions of the first kind if $\alpha_1 = \beta_1 = 0$; the conditions of the second kind if $\alpha_0 = \beta_0 = 0$; and the conditions of the third kind or mixed boundary conditions with all coefficients different from zero. Note also that by subtracting some function that satisfies the inhomogeneous boundary conditions problem (3.1) can be transformed to the boundary-value problem with $A = B = 0$.

3.1.2 Numerical methods

Most of numerical methods applied to boundary-value problems are based on the same ideas that are used for initial-value problems. The unknown function can be represented in the form of an expansion in some basic set of functions or it can be replaced by its values at the nodes of a mesh.

The methods can be assorted into the following groups

- the methods that convert the problem to Cauchy problems (shooting method and the method of differential sweep),
- the methods that allow the approximate values of the unknown function to be found at a discrete set of points (finite difference methods),
- the methods of collocations in which the solution is represented in the form of decomposition in some basis and the equation is satisfied at a set of collocation points,

- the projective methods, in which the coefficients of the decomposition are found from the requirement that the residual is orthogonal to some basis,
- the variational methods that reduce the problem to the minimization of some functional.

3.2 Shooting method and its generalizations

3.2.1 Shooting method

One of the ideas that can be exploited in numerical schemes for boundary-value problems is in the replacement of the boundary-value condition at the point $x = b$ by another condition at the point $x = a$. Consider instead of the problem (3.1) the initial value problem

$$\begin{cases} y''(x) + p(x)y'(x) + q(x)y(x) = f(x), & a < x < b, \\ \alpha_0 y(a) + \alpha_1 y'(a) = A, \\ \gamma_0 y(a) + \gamma_1 y'(a) = \xi. \end{cases} \tag{3.3}$$

Its solution can be found with the use of computational schemes described in the previous chapter. Indeed, solving the system of equations

$$\begin{cases} \alpha_0 y(a) + \alpha_1 y'(a) = A, \\ \gamma_0 y(a) + \gamma_1 y'(a) = \xi, \end{cases}$$

one can find the initial data

$$y(a) = y_0, \qquad y'(a) = y'_0$$

for the differential equation. The quantities y_0 and y'_0 depend on the parameter ξ, called the shooting parameter. Therefore the solution will also depend on ξ. Performing computations up to $x = b$ one can compute the residual for the second boundary condition

$$r(\xi) = \beta_0 y(b) + \beta_1 y'(b) - B.$$

By the choice of the shooting parameter it is possible to make the residual sufficiently small. Hence, the boundary-value problem (3.1) is reduced to solution of the equation

$$r(\xi) = 0. \tag{3.4}$$

When the parameter ξ is found (usually by bisection method), the solution of the boundary-value problem (3.1) coincides with the solution of the initial value problem (3.3).

3.2.2 Differential sweep method

The differential sweep method is developed by A. A. Abramov for linear problems. We consider it for the second order boundary-value problem. Rewrite the differential equation in the form of the system

$$\begin{cases} y_1' + a_{11}y_1 + a_{12}y_2 = f_1, \\ y_2' + a_{21}y_1 + a_{22}y_2 = f_2. \end{cases} \tag{3.5}$$

Here $A = \{a_{ij}\}_{i,j=1,2}$ is a given matrix-function and $F = (f_1, f_2)^T$ is a given vector-function (symbol ^{T} denotes transposition).

Rewrite the boundary conditions in the form

$$\begin{cases} b_{11}y_1(a) + b_{12}y_2(a) = g_1, \\ b_{21}y_1(b) + b_{22}y_2(b) = g_2. \end{cases} \tag{3.6}$$

The main idea of the method is in converting the first condition in (3.6) into some condition for the values of the unknown functions at the point $x = b$. Then the values $y_1(b)$ and $y_2(b)$ can be obtained from this new condition and the second condition in (3.6). This reduces the problem to a Cauchy problem which should be solved in inverse direction from $x = b$ to $x = a$.

Let us write a relation

$$\mathbf{u}(x)\mathbf{y}(x) = z(x). \tag{3.7}$$

Here functions $\mathbf{u} = (u_1, u_2)$ and $z(x)$ are arbitrary at this step. Let $u_1(a) = b_{11}$, $u_2(a) = b_{12}$, $z(a) = g_1$, then relation (3.7) for $x = a$ coincides with the first boundary condition in (3.6).

Our goal is to find such functions u_1, u_2 and z that the relation (3.7) is satisfied for arbitrary x. Differentiating relation (3.7) and using the differential equation (3.5), one gets

$$\Big(\mathbf{u}' - \mathbf{u}A\Big)\mathbf{y} = z' - \mathbf{u}\mathbf{f}.$$

If one requires the identities

$$\mathbf{u}' = \mathbf{u}A \tag{3.8}$$

and

$$z' = \mathbf{u}\mathbf{f} \tag{3.9}$$

to be satisfied, then relation (3.7) is satisfied on the entire interval $a < x < b$ for any vector-function $\mathbf{y}$ which satisfies the differential equation (3.5) and the boundary condition at the point $x = a$.

In this way, solving the Cauchy problems for equations (3.8) and (3.9) and finding the values $\mathbf{u}(b)$ and $z(b)$, one gets the relation for $y_1(b)$ and $y_2(b)$. Combining it with the second boundary condition, yields the system

$$\begin{cases} u_1(b)y_1(b) + u_2(b)y_2(b) = z(b), \\ b_{21}y_1(b) + b_{22}y_2(b) = g_2. \end{cases} \tag{3.10}$$

This gives the Cauchy problem for differential equation (3.5).

In practice the use of the described above scheme meets difficulties originated by round-off error accumulation resulting in ill conditioned relations (3.7) especially in the case when solutions rapidly increase. This defect is avoided in the following modification. The vector function $\mathbf{u}$ is represented in the form of the product

$$\mathbf{u} = Q\mathbf{v}.$$

The function $Q(x)$ is assumed to be equal to a unity at the point $x = a$. By differentiating (3.8) one gets

$$\mathbf{v}' + S\mathbf{v} - \mathbf{v}A = 0,$$

where

$$S = Q'/Q.$$

Now the function Q can be chosen such that the norm of the vector $\mathbf{v}$ remains constant. If

$$S = \frac{1}{\mathbf{v}\mathbf{v}^T}\mathbf{v}A\mathbf{v}^T, \tag{3.11}$$

then

$$\left(\mathbf{v}\mathbf{v}^T\right)' = \mathbf{v}'\mathbf{v}^T + \mathbf{v}\left(\mathbf{v}^T\right)' =$$

$$= S\mathbf{v}\mathbf{v}^T - \mathbf{v}A\mathbf{v}^T + \mathbf{v}\mathbf{v}^T S^T - \mathbf{v}A^T\mathbf{v}^T = 0.$$

Addressing to (3.11) one finally gets for $\mathbf{v}(x)$

$$\mathbf{v}' = \mathbf{v}A - \frac{1}{\mathbf{v}\mathbf{v}^T}\mathbf{v}A\mathbf{v}^T\mathbf{v}. \tag{3.12}$$

Now the equation for the right-hand side of the relation

$$\mathbf{v}(x)\mathbf{y}(x) = h(x) \tag{3.13}$$

can be derived. Differentiating and replacing the derivatives of $\mathbf{v}$ and $\mathbf{y}$ with the help of (3.5) and (3.12), it is easy to get the differential equation

$$h' + \frac{1}{\mathbf{v}\mathbf{v}^T}\mathbf{v}A\mathbf{v}^T h = \mathbf{v}\mathbf{f} \tag{3.14}$$

for $h(x)$.

Comparative to equations (3.8) and (3.9), equations (3.12) and (3.14) appear superior for numerical realization of the differential sweep method.

Note that the idea of differential sweep method can be also applied to the boundary-value problems for systems of larger dimension.

3.3 Finite difference methods

For the numerical solution of a Cauchy problem, the differential equation is replaced by a set of computational formulae for step by step calculation of the values $y_1, y_2, \ldots, y_N$. Finite difference methods for the boundary-value problem (3.1) are based on the idea to consider these computational formulae as the system of equations for the unknown values of the function $y(x)$ at the nodes. Compared to Cauchy problems, when the step can be chosen adapted for achieving the required accuracy, in finite difference methods the mesh $\{x_i\}_{i=0}^N$ should be chosen in advance. We consider only equidistant meshes with a step h.

Besides the accuracy those systems are preferred which can be easily solved, for example systems with three-diagonal matrices.

3.3.1 Simple finite difference method

In the simplest finite difference method the derivatives in the differential equation are replaced with finite difference approximations

$$y''(x_k) = \frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + O(h^2), \tag{3.15}$$

$$y'(x_k) = \frac{y_{k+1} - y_{k-1}}{2h} + O(h^2). \tag{3.16}$$

This yields

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + p(x_k)\frac{y_{k+1} - y_{k-1}}{2h} + q(x_k)y_k = f(x_k) + O(h^2). \tag{3.17}$$

Equations (3.17) can be written for $k = 1, 2, \ldots, N-1$. Two more equations are taken from the boundary conditions.

Note that equations (3.17) have the accuracy of order $O(h^2)$. If the boundary conditions contain derivatives $y'(a)$ and/or $y'(b)$ then, since one can not use symmetric approximations (3.16) for the derivatives, simple finite difference approximations

$$y'(a) = \frac{y_1 - y_0}{h} + O(h), \qquad y'(b) = \frac{y_N - y_{N-1}}{h} + O(h)$$

reduce the accuracy by one order. To write the approximations of order $O(h^2)$ one needs to use the differential equation. For example, in order to get the desired approximation for $y'(a)$ one writes Taylor series

$$y(a+h) = y(a) + y'(a)h + \frac{1}{2}y''(a)h^2 + O(h^3)$$

and substitutes the second derivative $y''(a)$ from the differential equation

$$y(a+h) = y(a)\left(1 - \frac{1}{2}q(a)h^2\right) + y'(a)h\left(1 - \frac{1}{2}p(a)\right) + \frac{1}{2}f(a)h^2 + O(h^3).$$

Taking into account that $y(a) = y_0$ and $y(a+h) = y_1$ one finds

$$y'(a) = \frac{y_1 - y_0\left(1 - \frac{1}{2}q(a)h^2\right) - \frac{1}{2}f(a)h^2}{h\left(1 - \frac{1}{2}p(a)h\right)} + O(h^2) =$$

$$= \frac{y_1 - y_0}{h}\left(1 + \frac{p(a)h}{2} + \frac{p(a)^2h^2}{4}\right) + y_0\frac{2q(a)h}{2} - \frac{f(a)h}{2} + O(h^2). \tag{3.18}$$

The similar approximation can be derived for the derivative $y'(b)$.

3.3.2 Numerov method

Compared to simple finite difference method, Numerov method allows the accuracy of the scheme to be increased by two orders. For that the differential equation should be rewritten in the form without the first order derivative

$$y''(x) + q(x)y(x) = f(x). \tag{3.19}$$

Let the boundary conditions

$$y(a) = y_a, \qquad y(b) = y_b$$

be given.

One can write analogously to (3.15)

$$y^{(4)}(x_k) = \frac{y''(x_{k+1}) - 2y''(x_k) + y''(x_{k-1})}{h^2} + O(h^2).$$

Substituting the approximations for the 4-th order derivatives in the two Taylor formulae

$$y(x_{k\pm1}) = y_k \pm y'(x_k)h + \frac{1}{2}y''(y_k)h^2 \pm \frac{1}{6}y'''(x_k)h^3 + \frac{1}{24}y^{(4)}(x_k)h^4 \pm$$

$$\pm\frac{1}{120}y^{(5)}(x_k)h^5 + O(h^6)$$

and summation term by term yields

$$y_{k+1} + y_{k-1} = 2y_k + \frac{h^2}{12}y''(x_{k+1}) + \frac{5}{6}h^2y''(x_k) + \frac{h^2}{12}y''(x_{k-1}) + O(h^6).$$

Now it is possible to exclude the second order derivatives with the help of the differential equation. Simple manipulations result in the difference equation

$$\left(1 + \frac{1}{12}h^2q_{k+1}\right)y_{k+1} - \left(2 - \frac{5}{6}h^2q_k\right)y_k + \left(1 + \frac{1}{12}h^2q_{k-1}\right)y_{k-1} =$$

$$= \frac{1}{12}f_{k+1} + \frac{5}{6}f_k + \frac{1}{12}f_{k-1} + O(h^6) \tag{3.20}$$

of Numerov method.

The matrices of the simple finite difference method and of Numerov method are three-diagonal and the systems can be solved with sweep method. One can prove that for $q(x) \leq 0$ these systems are nondegenerate for any step h.

Numerov method can be also applied to a wider class of equations. Namely the right-hand side in (3.19) can be dependent of y, that is the equation can be nonlinear (quasi-linear).

It is worth noting that further increase of accuracy of the finite difference scheme requires either to increase the number of diagonals in the matrix, or to involve derivatives of functions $p(x)$ and $f(x)$.

3.4 Spline-collocation method

In some cases it is insufficient to find only the values of the unknown function at the nodes. Then one can interpolate the data $\{x_k, y_k\}$ obtained by a finite difference method. The spline-collocation method combines the process of data computation and the process of cubic spline interpolation in one procedure.

Consider the problem

$$\begin{cases} y'' + qy = f, \\ y(a) = A, \\ y(b) = B. \end{cases}$$

Let $\Delta = \{x_i\}_{i=0}^N$, $a = x_0 < x_1 < \ldots < x_N = b$ be an arbitrary mesh. We shall search the approximation of the solution in the form of a spline $S \in S_3^1(\Delta)$. The dimension of the space of cubic splines $S_3^1(\Delta)$ is equal to $N + 3$. That means to define a spline one needs $N + 3$ conditions to be set. Two of these conditions are the boundary conditions, the rest appear from the requirement that the differential equation is satisfied at a set of points $x = \xi_k$, which are called the collocation points, that is

$$S''(\xi_k) + q(\xi_k)S(\xi_k) = f(\xi_k), \qquad k = 1, 2, \ldots, N. \tag{3.21}$$

Note that it is not possible to choose the collocation point arbitrary. In particular it is not allowed to have more than 3 collocation points on an interval $\Delta_i = [x_i, x_{i+1}]$, because otherwise the spline on this interval will be completely defined by the data at these points, that is it will not depend on the boundary conditions. It is also evident that one can not take collocation points at jumps of the functions $p(x)$ and $f(x)$ if any exist.

Below we shall consider the case when collocation points coincide with the nodes of the mesh, that is $\xi_i = x_i$.

Let the following notations

$$S(x_i) = u_i, \qquad S''(x_i) = M_i$$

be introduced. If the values u_i are known, then the values M_i can be found from the system of equations

$$\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = \frac{6}{h_{i-1} + h_i} \left(\frac{u_{i+1} - u_i}{h_i} - \frac{u_i - u_{i-1}}{h_{i-1}} \right), \qquad (3.22)$$

(here $\lambda_i = h_i/(h_i + h_{i-1})$, $\mu_i = h_{i-1}/(h_i + h_{i-1})$) which should be completed with boundary conditions. After that the spline is given by the formulae

$$S(x) = u_i(1-t) + u_{i+1}t - \frac{h_i^2}{6} t(1-t)\Big(M_i(2-t) + (1+t)M_{i+1}\Big) \qquad (3.23)$$

written in local variable t which on each interval is defined by the formula

$$t = \frac{x - x_i}{h_i}.$$

Using the differential equation which according to our requirements is satisfied for the spline $S(x)$ at the nodes, the values M_i can be excluded from system (3.22). Multiplying by $\frac{1}{6}h_{i-1}h_i$, one finds

$$\lambda_i \left(1 + \frac{h_{i-1}^2}{6} q_{i-1}\right) u_{i-1} - \left(1 - \frac{h_i h_{i-1}}{3} q_i\right) u_i + \mu_i \left(1 + \frac{h_i^2}{6} q_{i+1}\right) u_{i+1} =$$

$$= \frac{h_{i-1}h_i}{6} \left(\mu_i f_{i-1} + 2f_i + \lambda_i f_{i+1}\right). \qquad (3.24)$$

Together with the boundary conditions $u_0 = A$, $u_N = B$ system (3.24) allows the values u_i and further the spline itself to be found.

Note that for $q \le 0$ by choosing the steps h_i satisfying

$$h_{i-1}^2 \max(-q_{i-1}, -q_i) \le 6 \qquad (3.25)$$

the system with diagonal predominance appears. The error estimate is given by the theorem

Theorem 1 *Let $q(x) \le Q < 0$ and the inequalities (3.25) are satisfied, then for the solution $y(x)$ from $C^2([a,b])$ the error is estimated as follows*

$$\max_{x\in[a,b]} |S(x) - y(x)| = O\left(\max_{i=1,N} h_i^2\right).$$

Proof
Let us introduce the spline $\tilde{S}(x)$ interpolating the table of values of the exact solution $y(x)$ at the nodes of the mesh Δ and satisfying the conditions $\tilde{S}'(a) = y'(a)$, $\tilde{S}'(b) = y'(b)$. It is known that the interpolation error is of order $O(h^4)$. Therefore

$$|S(x) - y(x)| \le |S(x) - \tilde{S}(x)| + |\tilde{S}(x) - y(x)| \le |S(x) - \tilde{S}(x)| + O(h^4). \tag{3.26}$$

The $S(x)$ and $\tilde{S}(x)$ are constructed by the explicit formulae that differ only in the replacement of u_i by y_i and M_i by $\tilde{S}''(x_i)$. From these formulae it is easy to derive the estimate

$$|S(x) - \tilde{S}(x)| \le \max_{i=0,N} |u_i - y_i| + \frac{h^2}{8} \max_{i=0,N} |M_i - \tilde{S}''(x_i)|. \tag{3.27}$$

Noting that the second derivative is approximated by the spline with the error of order $O(h^2)$ one concludes that the spline $\tilde{S}$ satisfies the differential equation with the error of order $O(h^2)$, that is

$$\tilde{S}''(x_i) + q_i \tilde{S}(x_i) = f_i + O(h^2). \tag{3.28}$$

Then

$$|M_i - \tilde{S}''(x_i)| \le q_i |u_i - y_i| + O(h^2).$$

Substituting this estimate into (3.27) yields

$$|S(x) - \tilde{S}(x)| \le \left(1 + \frac{h^2}{8} \max_{x\in[a,b]} q\right) \max_{i=0,N} |u_i - y_i| + O(h^4). \tag{3.29}$$

Hence, to prove the theorem it is needed to derive the estimate for the difference $u_i - y_i$. For this purpose the system (3.24) is rewritten in matrix form

$$A\mathbf{u} = \mathbf{d}.$$

Replacing equations (3.21) by (3.28) one can analogously get for $\mathbf{y} = (y_0, y_1, \ldots, y_N)^T$ the system of equations

$$A\mathbf{y} = \mathbf{d} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = O(h^4)$. For the difference $u_i - y_i$ this yields

$$A(u_i - y_i) = \boldsymbol{\varepsilon}.$$

For the systems with diagonal predominance the solution is estimated as

$$\max_{i=0,N} |u_i - y_i| \leq \max_{i=0,N} \frac{|\varepsilon_i|}{r_i},$$

where r_i is the diagonal predominance factor for the i-th row. In the case of system (3.24) these factors are

$$r_i = 1 - \frac{h_i h_{i-1}}{3} q_i - \lambda_i \left(1 + \frac{h_{i-1}^2}{6} q_{i-1}\right) - \mu_i \left(1 + \frac{h_i^2}{6} q_{i+1}\right) =$$

$$= -\frac{h_i h_{i-1}}{3} q_i - \lambda_i \frac{h_{i-1}^2}{6} q_{i-1} - \mu_i \frac{h_i^2}{6} q_{i+1} \leq -\frac{h_i h_{i-1}}{6} Q.$$

Therefore, $|u_i - y_i| = O(h^2)$, which together with (3.29) and (3.26) concludes the proof.

Note that the restriction $q(x) \leq Q < 0$ can be weakened to $q(x) \leq 0$. It is often possible to exclude condition (3.25), in particular the diagonal predominance takes place for any mesh if $q(x) \equiv const$.

3.5 The method of moments

In the above discussed methods the differential equation was replaced by a set of conditions meaning that it is satisfied at a set of points called the collocation points. However this is not the only method of passing from an equation satisfied on an interval to a finite set of equations. Let $u(x)$ be an approximate solution of the boundary-value problem (3.1). Consider the residual

$$r_u(x) \equiv u''(x) + p(x)u'(x) + q(x)u(x) - f(x).$$

If the solution is exact, then the residual is identically zero

$$r_y \equiv y''(x) + p(x)y'(x) + q(x)y(x) - f(x) = 0 \qquad (3.30)$$

In the method of moments the condition $r(x) \equiv 0$ is replaced by the requirements that the projections of the residual on a given set of functions $\psi_j(x)$ is equal to zero, that is

$$\int_a^b r_u(x)\psi_j(x)\,dx = 0. \tag{3.31}$$

Note that if the functions $\psi_j(x)$ constitute a basis in $L_2([a,b])$, then conditions (3.31) for $j = 1, 2, \ldots, J$ mean that the first J coefficients in the decomposition of the residual $r(x)$ in the basis $\psi_j(x)$ are equal to zero.

Consider the method of moments in more details. Let us first exclude the right-hand sides in the boundary conditions, that is transform the boundary-value problem (3.1) to the similar problem with the homogeneous boundary conditions. This can be done by subtracting a function that satisfies to the boundary conditions in (3.1), but does not satisfy the differential equation. Let

$$y(x) = y_0(x) + U(x),$$

where $y_0(x)$ satisfies

$$\begin{cases} \alpha_0 y_0(a) + \alpha_1 y_0'(a) = A, \\ \beta_0 y_0(b) + \beta_1 y_0'(b) = B. \end{cases}$$

Then the function $U(x)$ should be the solution of the problem

$$\begin{cases} U''(x) + p(x)U'(x) + q(x)U(x) = g(x), \\ \alpha_0 U(a) + \alpha_1 U'(a) = 0, \\ \beta_0 U(b) + \beta_1 U'(b) = 0, \end{cases} \tag{3.32}$$

where

$$g(x) = f(x) - \left(y_0''(x) + p(x)y_0'(x) + q(x)y_0(x)\right).$$

Consider a system of sufficiently smooth functions $\varphi_k(x)$, $k = 1, 2, \ldots$. Let these functions satisfy the boundary conditions from (3.32). Then any linear combination of these functions

$$u_N(x) = \sum_{k=1}^{N} C_k \varphi_k(x), \tag{3.33}$$

also satisfies the boundary conditions of the problem (3.32). Let the solution be searched in the form of such a combination. In order the functions $u_N(x)$ can approach the solution $U(x)$ when $N \to \infty$ one requires the system $\{\varphi_k(x)\}$ to be complete. (Actually it is sufficient to require that only the solution $U(x)$ can be decomposed by the system $\{\varphi_k(x)\}$).

Let the coefficients C_k of decomposition (3.33) be chosen so that conditions (3.31) are satisfied. Using the expressions for the residual these conditions can be written in the form of the linear system

$$\sum_{k=1}^{N} C_k A_{jk} = B_j, \qquad j = 1, 2, 3, \ldots, N.$$

Here

$$A_{jk} \int_a^b \psi_j(x) \left(\varphi_k''(x) + p(x)\varphi_k'(x) + q(x)\varphi_k(x)\right) dx, \qquad B_j = \int_a^b \psi_j(x) g(x)\, dx.$$

The case with $\varphi_j = \psi_j$ is called Galerkin method.

3.6 Variational methods

3.6.1 Least squares method

When using variational approach the boundary-value problem is replaced by some minimization problem. Two main approaches are possible: the least squares method being in the minimization of the residual norm $|r(x)|$, and Ritz method based on energetic ideas.

We start with a more universal least squares method. As in the method of moments the problem is transformed to the case of homogeneous boundary conditions and the solution is represented in the form (3.33)

$$u_N(x) = \sum_{k=1}^{N} C_k \varphi_k(x),$$

The coefficients C_k of this decomposition are found from the requirement that the norm of the residual is minimal, that is the minimum of

$$\|r_u(x)\|^2 \equiv \int_a^b |r_u(x)|^2 dx$$

is searched. After substitution of representation (3.33) for $u(x)$ we get the function of the variables C_1, C_2, ... , C_N

$$F(C_1, C_2, \dots, C_N) = \int_a^b \left| \left(\sum_{k=1}^{N} C_k \mathcal{L}\varphi_k(x) \right) - g(x) \right|^2 dx. \tag{3.34}$$

Here we denoted by $\mathcal{L}$ the differential operator in the equation, that is $\mathcal{L} = d^2/dx^2 + p(x)d/dx + q(x)$.

Thus the boundary-value problem for the differential equation is reduced to minimizing the function (3.34), which can be done by standard methods of minimization.

Note that the least squares method can be also applied to nonlinear problems and moreover with some simple modifications to ill-posed problems.

In our case of linear differential equation the function $F(C_1, C_2, \dots, C_N)$ is quadratic and its minimization is reduced to solving the system of linear equations

$$\frac{\partial F}{\partial C_k} = 0, \qquad k = 1, 2, \dots, N. \tag{3.35}$$

Calculating the derivatives yields explicit form of system (3.35)

$$\sum_{j=1,N} A_{jk} C_j = b_k, \qquad k = 1, 2, \dots, N,$$

where

$$A_{jk} = \mathrm{Re} \left(\int_a^b \mathcal{L}\varphi_j(x)\ \mathcal{L}\varphi_k(x)\ dx \right), \quad b_k = \mathrm{Re} \left(\int_a^b \mathcal{L}\varphi_k(x)\ g(x)\ dx \right).$$

One can chose in particular the basic splines $B_k^n(x)$ for the functions $\varphi_k(x)$. Due to the finiteness property of basic splines this yields band-type matrix A_{jk}. With the increase of the degree n of the spline the number of nonzero diagonals increases. On the contrary if n decreases, the number of diagonals also decreases. However it is not possible to diminish the degree n to zero because when computing the elements A_{jk} one needs to integrate the second derivatives of $B_k^n(x)$. This is possible only if $B_k^n \in C^1([a, b])$ which means that n can not be less than two.

The supports of splines $B_k^2(x)$ consist of three intervals of the mesh Δ. Hence, the matrix elements A_{kj} with $|j-k| \le 2$ are different from zero and the matrix of system (3.35) has 5 nonzero diagonals. The use of basic splines in the Ritz method allows three-diagonal system to be obtained.

3.6.2 Ritz method

Ritz method can be applied only to sufficiently "good" problems. Namely, it requires the operator $\mathcal{L}$ of the boundary-value problem to be positive, that is it should be symmetric and for any differentiable function $u(x)$ that satisfies the boundary conditions the scalar product

$$\langle u, \mathcal{L}u \rangle = \int_a^b u(x)\mathcal{L}u(x)\,dx$$

be non-negative. For example the operator of the problem

$$\begin{cases} \mathcal{L}y(x) \equiv -(p(x)y'(x))' + q(x)y(x) = f(x), & a < x < b, \\ y(a) = 0, \qquad y(b) = 0, \end{cases} \tag{3.36}$$

where p and q are positive-valued functions, is positive. Indeed, integrating by parts yields

$$\langle u, \mathcal{L}u \rangle = \int_a^b u(x)\left(-(p(x)u'(x))' + q(x)u(x)\right)dx.$$

$$= -\,upu'\Big|_a^b + \int_a^b \left(p(x)|u'(x)|^2 + q(x)|u(x)|^2\right)dx.$$

The non-integral terms disappear due to the boundary condition and the integral is non-negative because the functions p and q are positive-valued.

The following theorem is the cornerstone of Ritz method.

Theorem 2 *The minimization problem for the functional*

$$J(u) = \langle \mathcal{L}u, u \rangle - 2\langle f, u \rangle. \tag{3.37}$$

on the set of functions that satisfy the boundary conditions is equivalent to the boundary-value problem (3.36).

That means that the function which minimizes the functional J is the solution of problem (3.36) and vice a versa, the solution of the problem (3.36) minimizes functional (3.37).

We shall prove the above theorem for a boundary value problem for a general positive operator. Let the function $u(x)$ satisfies the boundary conditions and minimizes the functional J. Let a differential function $v(x)$ also satisfy the boundary conditions and z be any number. Consider the difference

$$J(u+zv) - J(u) = z\left(\langle \mathcal{L}u, v\rangle + \langle \mathcal{L}v, u\rangle - 2\langle f, v\rangle\right) + z^2\langle \mathcal{L}v, v\rangle. \tag{3.38}$$

As u minimizes the functional on the set of differential functions satisfying the boundary conditions and the function $u(x) + zv(x)$ belongs to this set for any z, the inequality

$$J(u+zv) \geq J(u) \tag{3.39}$$

holds. For sufficiently small $|z|$ the first (linear in z) term evidently dominates in (3.38). Evidently if

$$\langle \mathcal{L}u, v\rangle + \langle \mathcal{L}v, u\rangle - 2\langle f, v\rangle \neq 0,$$

one can choose z of appropriate sign, so that inequality (3.39) is violated. Therefore,

$$\langle \mathcal{L}u, v\rangle + \langle \mathcal{L}v, u\rangle - 2\langle f, v\rangle = 0. \tag{3.40}$$

Then

$$J(u+zv) - J(u) = z^2\langle \mathcal{L}v, v\rangle \geq 0$$

because the operator $\mathcal{L}$ is positive.

Now integrate by parts in the second term in (3.40)

$$\langle \mathcal{L}v, u\rangle = \langle v, \mathcal{L}u\rangle = \langle \mathcal{L}u, v\rangle.$$

Then equality (3.40) can be written in the form

$$2\langle \mathcal{L}u - f, v\rangle = 0.$$

This equality is satisfied for any (admissible) function v, which is possible only if

$$\mathcal{L}u - f = 0,$$

which is equivalent to the equation in the boundary-value problem.

We showed that the function that minimizes the functional J satisfies the boundary-value problem. Now we show the opposite, that the solution of the boundary-value problem minimizes the functional. Let u be the solution of the boundary value problem. Consider any other differential function w that satisfies the boundary conditions and calculate the difference

$$J(w)-J(u) = (\langle \mathcal{L}u, w-u\rangle + \langle \mathcal{L}(w-u), u\rangle - 2\langle f, w-u\rangle) + \langle \mathcal{L}(w-u), w-u\rangle.$$

Integrating by parts in the second term and taking into account that $\mathcal{L}u = f$, we see that the expression in brackets is equal to zero. Thus,

$$J(w) - J(u) = \langle \mathcal{L}(w-u), w-u\rangle,$$

which is non-negative due to the positive property of the operator $\mathcal{L}$. This concludes the proof.

The main method of finding the minimum of functional (3.37) is the construction of minimization sequence (or Ritz sequence). Assume an infinite set $\{\varphi_k(x)\}_{k=1}^{\infty}$ of functions $\varphi_k(x)$, that satisfy the boundary conditions, are sufficiently smooth for the application of the differential operator and all together form a complete system. Then the minimization sequence $\{U_n(x)\}_{n=1}^{\infty}$ can be constructed in the form

$$U_n(x) = \sum_{k=1}^{n} a_{kn}\varphi_k(x).$$

The coefficients a_{kn} are chosen such that the value of functional (3.37) on every U_n is minimal, that is

$$J(U_n) = \min_{a_{kn}} \left(\left\langle \mathcal{L}\sum_{k=1}^{n} a_{kn}\varphi_k(x), \sum_{k=1}^{n} a_{kn}\varphi_k(x) \right\rangle - 2\left\langle f, \sum_{k=1}^{n} a_{kn}\varphi_k(x) \right\rangle \right). \tag{3.41}$$

For every fixed n the function of variables a_{kn} that should be minimized is quadratic

$$\sum_{k=1}^{n}\sum_{j=1}^{n} a_{kn}a_{jn}\langle \mathcal{L}\varphi_k, \varphi_j\rangle - 2\sum_{k=1}^{n} a_{kn}\langle f, \varphi_k\rangle. \tag{3.42}$$

Note that in the scalar product $\langle \mathcal{L}\varphi_k, \varphi_j \rangle$ one can perform integration by parts, which in the case of problem (3.36) gives

$$\langle \mathcal{L}\varphi_k, \varphi_j \rangle = \int_a^b \left(p(x)\varphi_k'(x)\varphi_j'(x) + q(x)\varphi_k(x)\varphi_j(x) \right) dx \equiv A_{kj}.$$

This reduces the requirements of smoothness of the functions $\varphi_k(x)$ to $\varphi_k \in W_2^1(p)$, where $W_2^1(p)$ is the space of functions that have generalized derivative of the first order square-integrable on the interval $[a, b]$ with the weight $p(x)$.

The minimization problem (3.42) can be reduced to the system of linear equations

$$\sum_{j=1}^{n} A_{kj} a_{jn} = B_k, \qquad k = 1, 2, \ldots, n. \tag{3.43}$$

Here

$$B_k = \int_a^b f(x)\varphi_k(x)\, dx.$$

The system (3.43) is called Ritz system.

Compared to the least squares method one can use basic splines of the first degree in Ritz method applied to the boundary-value problem (3.36). This yields the system of equations with three-diagonal matrix.

3.7 Sturm-Liouville problem

Consider the following problem

$$\begin{cases} -(p(x)\psi'(x))' + q(x)\psi(x) = \lambda\psi(x), & a < x < b, \\ \psi(a) = 0, \qquad \psi(b) = 0 \end{cases} \tag{3.44}$$

for the parameter λ and function $\psi(x)$.

One can also take Neumann or mixed type boundary conditions.

The problem (3.44) looks similar to a boundary-value problem, but there is an additional parameter λ. If this parameter is given, then one deals with a homogeneous boundary-value problem for the function $\psi(x)$. Evidently that $\psi(x) \equiv 0$ solves that problem. However for some special values of

the parameter λ, called eigen-values λ_k, there exist nonzero solutions $\psi_k(x)$, called eigen-functions. Eigen-functions are defined up to a multiplier and to reduce this arbitrariness one can pose a normalization for example in the form

$$\int_a^b |\psi(x)|^2 dx = 1. \tag{3.45}$$

(Then the eigenfunctions are defined up to a multiplier ± 1.)

The problem of finding eigen-numbers λ_k and eigen-functions $\psi_k(x)$ of (3.44) is called the *Sturm-Liouville problem*. A Sturm-Liouville problem may appear as an independent problem or as a part of a more complicated one, for example, in variables separation method for partial differential equation.

There are several methods for solving Sturm-Liouville problem (3.44). Some of them are described below.

3.7.1 Shooting method

Similar to shooting method for boundary-value problems the main idea is in transforming the problem to Cauchy problem. One can set the second condition at the point $x = a$ as

$$\psi'(a) = 1$$

because the solution $\psi(x)$ of the problem (3.44) is defined up to a multiplier.

Then one gets the Cauchy problem

$$\begin{cases} -(p(x)\psi'(x))' + (q(x) - \lambda)\,\psi(x) = 0, & a < x < b, \\ \psi(a) = 0, \qquad \psi'(a) = 1 \end{cases} \tag{3.46}$$

for the function $\psi(x)$. Its solution for a given value of the spectral parameter λ can be obtained by any computational scheme of chapter 2, for example by Runge-Kutta method. Then the problem is reduced to solving the equation

$$\psi(b, \lambda) = 0 \tag{3.47}$$

with respect to λ.

Equation (3.47) has infinite set of solutions and if one needs to find all eigen-values of problem (3.44) on a given interval, two problems may appear. Firstly, one needs to separate the solutions and secondly, one needs to check that no solution is missed. To overcome these difficulties an analytic analysis of a particular problem or some physical considerations can be useful.

3.7.2 Finite difference method

Application of finite difference approximation of the differential operators in the differential equation yields spectral problems of linear algebra. If equation (3.44) is transformed to

$$y''(x) + P(x)y'(x) + (Q(x) - \lambda)\, y(x) = 0,$$

then the algebraic system of the simple finite difference method can be easily derived from equations (3.17), if one takes $f(x_k) = 0$, $p(x_k) = P(x_k)$ and $q(x_k) = Q(x_k) - \lambda$. Then one gets the spectral problem for the matrix

$$\begin{pmatrix} -\frac{2}{h^2} + Q(x_1) & \frac{1}{h^2} - \frac{P(x_1)}{h} & 0 & \dots & 0 \\ \frac{1}{h^2} + \frac{P(x_2)}{h} & -\frac{2}{h^2} + Q(x_2) & \frac{1}{h^2} - \frac{P(x_2)}{h} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \frac{1}{h^2} + \frac{P(x_k)}{h} & -\frac{2}{h^2} + Q(x_k) & \frac{1}{h^2} - \frac{P(x_k)}{h} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \frac{1}{h^2} + \frac{P(x_{N-1})}{h} & -\frac{2}{h^2} + Q(x_{N-1}) \end{pmatrix}.$$

To use Numerov method, one transforms the equation to the form without the first order derivative, that is

$$y''(x) + (Q(x) - \lambda)\, y(x) = 0.$$

Then replacing q in equations (3.20) by $Q - \lambda$ yields the spectral problem for the pencil

$$\mathbf{A} = \lambda \mathbf{B},$$

where $\mathbf{A}$ and $\mathbf{B}$ are three-diagonal matrices with the elements

$$a_k^k = -2 + \frac{5}{6}h^2 Q(x_k), \quad a_k^{k-1} = 1 + \frac{h^2 Q(x_{k-1})}{12}, \quad a_k^{k+1} = 1 + \frac{h^2 Q(x_{k+1})}{12},$$

$$b_k^k = \frac{5}{6}, \qquad b_k^{k-1} = b_k^{k+1} = \frac{1}{12}.$$

The eigen-numbers of these spectral problems approximate the first eigen-numbers of Sturm-Liouville problem.

3.7.3 Variational methods

The idea of variational methods (least squares method and Ritz method) is similar to that described in the case of boundary-value problems and leads to spectral problems of linear algebra.

3.7.4 The abstract Newton method

When solving the Sturm-Liouville problem by the methods described above the problem is reduced to spectral problem of linear algebra and then this problem is solved by the corresponding methods of linear algebra. Another approach is also possible. The abstract (or operator) Newton method reduces the Sturm-Liouville problem to a sequence of boundary-value problems.

Consider the abstract Newton method first in the general form. Suppose the nonlinear operator equation

$$\mathcal{F}(\mathbf{x}) = \mathbf{y} \tag{3.48}$$

should be solved. Here $\mathcal{F}$ is the operator acting from a linear space X to the linear space Y defined on some domain $D(\mathcal{F})$. The element $\mathbf{y} \in Y$ is given, the element $\mathbf{x} \in D(\mathcal{F})$ should be found.

In the simplest case when $X = Y = \mathbf{R}$, equation (3.48) is a nonlinear algebraic equation and its solution can be found by the usual Newton method as the limit of the following iterations

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\mathcal{F}(\mathbf{x}_k)}{\mathcal{F}'(\mathbf{x}_k)}, \tag{3.49}$$

where $\mathcal{F}'$ is the derivative of the function $\mathcal{F}$.

In a little more complicated case when $X = Y = \mathbf{R}^n$, equation (3.48) represents the system

$$\begin{cases} f_1(\mathbf{x}) = y_1, \\ f_2(\mathbf{x}) = y_2, \\ \dots \\ f_n(\mathbf{x}) = y_n. \end{cases} \qquad \text{or} \qquad \mathbf{f}(\mathbf{x}) = \mathbf{y}.$$

Here $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$. The Newton method in that case is in the iterative process

$$\mathbf{x}_{k+1} = \mathbf{x}_k - J^{-1}(\mathbf{x}_k)\mathbf{f}(\mathbf{x}_k),$$

where $J(\mathbf{x})$ is the Jacobi matrix

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \frac{\partial f_n(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_n(\mathbf{x})}{\partial x_n} \end{pmatrix}.$$

In the abstract case the iterations are given by the formulae

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\mathcal{B}(\mathbf{x}_k)\right)^{-1} \mathcal{F}(\mathbf{x}_k)$$

or

$$\mathcal{B}(\mathbf{x}_k)\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) = -\mathcal{F}(\mathbf{x}_k) \tag{3.50}$$

where the operator $\mathcal{B}(\mathbf{x})$ is the operator-derivative of $\mathcal{F}$ on the element $\mathbf{x}$. The operator $\mathbf{F}$ is differentiable on the element $\mathbf{x}$ if there exists a linear operator $\mathcal{B}(\mathbf{x})$ acting from the space X to the space Y, such that for any element $\mathbf{z} \in X$

$$\frac{\|\mathcal{F}(\mathbf{x}+\mathbf{z}) - \mathcal{F}(\mathbf{x}) - \mathcal{B}(\mathbf{x})\mathbf{z}\|}{\|\mathbf{z}\|} \underset{\|\mathbf{z}\|\to 0}{\longrightarrow} 0.$$

Let us now turn back to Sturm-Liouville problem (3.44). The unknowns are the eigen-value λ and the eigen-function $\psi(x)$ corresponding to that eigen-value. Let $\mathbf{x}$ consist of these unknowns

$$\mathbf{x} = \begin{pmatrix} \psi(x) \\ \lambda \end{pmatrix}.$$

The problem (3.44) in the space of square-integrable functions ($\psi \in L_2([a,b])$) is self-adjoint and its eigen-values belong to the real space $\mathbb{R}$. Thus, the elements $\mathbf{x}$ belong to the space being the product of $L_2([a,b])$ and $\mathbb{R}$,

$$X = L_2([a,b]) \otimes \mathbb{R}.$$

Let the operator $\mathcal{F}$ be defined on such elements $\mathbf{x}$ that have twice differentiable first component vanishing at the end points of the interval $[a,b]$. Let the action of the operator be defined by the formula

$$\mathcal{F}(\mathbf{x}) = \begin{pmatrix} -(p(x)\psi'(x))' + q(x)\psi(x) - \lambda\psi(x) \\ \int\limits_a^b |\psi(x)|^2 dx - 1 \end{pmatrix}.$$

The defined above operator $\mathcal{F}$ is not linear, it acts in the space X, that is $Y = X$.

Consider the problem

$$\mathcal{F}(\mathbf{x}) = \mathbf{0}. \tag{3.51}$$

The equation in the first component of (3.51) coincides with the differential equation of (3.44). The boundary conditions are taken into account by the domain of the operator $\mathcal{F}$. The equation in the second component of (3.51) is the normalization condition (3.45). Therefore the operator equation (3.51) is equivalent to Sturm-Liouville problem (3.44) with the normalization condition for the eigen-functions.

To perform the iterations by formula (3.50) one needs to define the operator derivative of $\mathcal{F}$. For that consider the element $\mathcal{F}(\mathbf{x}+\mathbf{z})$

$$\mathcal{F}(\mathbf{x}+\mathbf{z}) = \begin{pmatrix} -\left(p\left(\psi'+\zeta'\right)\right)' + q\left(\psi+\zeta\right) - (\lambda+\mu)(\psi+\zeta) \\ \int\limits_a^b \left(\psi(x)+\zeta(x)\right)^2 dx - 1 \end{pmatrix}.$$

Here

$$\mathbf{z} = \begin{pmatrix} \zeta(x) \\ \mu \end{pmatrix}.$$

Neglecting quadratic in $\mathbf{z}$ terms gives

$$\mathcal{F}(\mathbf{x}+\mathbf{z}) = \mathcal{F}(\mathbf{x}) + \begin{pmatrix} -\left(p(x)\zeta'(x)\right)' + \left(q(x)-\lambda\right)\zeta(x) - \mu\psi(x) \\ 2\int\limits_a^b \psi(x)\zeta(x)dx \end{pmatrix} + O(\|\mathbf{z}\|^2). \tag{3.52}$$

The second term in (3.52) gives the expressions for $\mathcal{B}(\mathbf{x})\mathbf{z}$, where $\mathcal{B}(\mathbf{x})$ is the operator derivative of $\mathcal{F}$ on the element $\mathbf{x}$.

Thus, at every step of the iterative procedure one needs to solve the following linear problems

$$\begin{cases} -(p\zeta_k')' + (q-\lambda_k)\zeta_k - \mu_k\psi_k = (p\psi_k')' - (q-\lambda_k)\psi_k, \\ \zeta(a) = \zeta(b) = 0, \\ 2\int\limits_a^b \psi_k(x)\zeta_k(x)dx = 1 - \|\psi_k\|^2 \end{cases} \tag{3.53}$$

for $\mathbf{z}_k = (\zeta_k(x), \mu_k)^T$, and then set

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{z}_k \qquad \text{or} \qquad \begin{pmatrix} \psi_{k+1}(x) \\ \lambda_{k+1} \end{pmatrix} = \begin{pmatrix} \psi_k(x) \\ \lambda_k \end{pmatrix} + \begin{pmatrix} \zeta_k(x) \\ \mu_k \end{pmatrix}.$$

Problem (3.53) can be reduced to two usual boundary-value problems, namely let the functions $\xi(x)$ and $\eta(x)$ be solutions of the following boundary-value

problems

$$\begin{cases} -(p\xi_k')' + (q-\lambda_k)\xi_k = (p\psi_k')' - (q-\lambda_k)\psi_k, \\ \xi_k(a) = \xi_k(b) = 0, \end{cases} \tag{3.54}$$

and

$$\begin{cases} -(p\eta_k')' + (q-\lambda_k)\eta_k = \psi_k, \\ \eta_k(a) = \eta_k(b) = 0. \end{cases}$$

Then the function $\zeta_k(x) = \xi_k(x) + \mu_k\eta_k(x)$ satisfies the differential equation in (3.53) and the boundary conditions for any value of μ_k. The required value of μ_k can be found from the second equation in (3.53)

$$\mu_k = \frac{1 - \int\limits_a^b \Big(\psi_k^2(x) + 2\psi_k(x)\xi_k(x)\Big)dx}{2\int\limits_a^b \psi_k(x)\eta_k(x)dx}. \tag{3.55}$$

Thus, the operator Newton method reduces the Sturm-Liouville problem to a sequence of boundary-value problems. The convergence of the process is quadratic as in the usual Newton method. However one needs a good initial approximation. To make the method more robust the right-hand side of (3.54) is sometimes multiplied by factor t, $0 < t < 1$. The additional instability arises when you are searching several eigenvalues and the process converges to already known solution. Suppose that eigen value $\lambda^{(1)}$ and eigen function $\psi^{(1)}(x)$ are known and we a searching for $\lambda = \lambda^{(2)}$, $\psi(x) = \psi^{(2)}(x)$. To avoid convergence of iterations to $\lambda^{(1)}$, $\psi^{(1)}(x)$ one can modify the process by adding the orthogonality condition

$$\int\limits_a^b \psi(x)\psi^{(1)}(x)dx = 0,$$

which, in fact, is the property of the eigen functions of Sturm–Liouville problem.

Minimization of the functional constructed accoding to the method of Lagrange multipliers

$$I(\psi) = \int\limits_a^b \Big(p(x)\left(\psi'(x)\right)^2 + (q(x)-\lambda)y^2(x)\Big)\,dx + 2\varepsilon\int\limits_a^b \psi(x)\psi^{(1)}(x)dx,$$

yields the problem

$$\begin{cases} -(p\psi')' + (q(x) - \lambda)\,\psi(x) + \varepsilon\psi^{(1)}(x) = 0, \\ \psi(a) = 0, \qquad \psi(b) = 0, \\ \int\limits_a^b |\psi(x)|^2\,dx - 1 = 0, \\ \int\limits_a^b \psi(x)\psi^{(1)}(x)dx = 0. \end{cases}$$

Then besides $\xi_k(x)$ and $\eta_k(x)$ each step of iterations requires finding solution $\theta_k(x)$ of the boundary value problem

$$\begin{cases} -(p\eta_k')' + (q - \lambda_k)\eta_k = \psi^{(1)}, \\ \eta_k(a) = \eta_k(b) = 0. \end{cases}$$

After that new approximation is computed as

$$\mathbf{x}_{k+1} = \left(\begin{pmatrix} \psi_{k+1}(x) \\ \lambda_{k+1} \end{pmatrix}\right) = \left(\begin{pmatrix} \psi_k(x) + \xi_k(x) + \mu_k\eta_k(x) - \varepsilon_k\theta_k(x) \\ \lambda_k + \mu_k \end{pmatrix}\right),$$

where coefficients μ_k and ε_k are solutions of the system that appears as linearization of normalization and orthogonality conditions

$$\begin{cases} \mu_k \int\limits_a^b \eta_k(x)\psi^{(1)}(x)dx - \varepsilon_k \int\limits_a^b \theta_k(x)\psi^{(1)}(x)dx = \\ \qquad\qquad = -\int\limits_a^b \Big(\psi_k(x) + \xi_k(x)\Big)\psi^{(1)}(x)dx, \\ \mu_k \int\limits_a^b \eta_k(x)\psi_k(x)dx - \varepsilon_k \int\limits_a^b \theta_k(x)\psi_k(x)dx = \\ \qquad\qquad = \frac{1}{2} - \frac{1}{2}\int\limits_a^b \Big(\psi_k(x) + 2\xi_k\Big)\psi_k(x)dx. \end{cases}$$

3.8 Some generalizations

In this chapter we have considered spectral problems for the second order linear differential equations. All the methods described in this chapter allow generalizations to higher order equations and also to nonlinear problems. The complifications are in the increase of dimension of the system of algebraic equations to which the problem is reduced by this or that method. In the case of nonlinear problems these algebraic systems will be nonlinear and their solution may be a complicated problem. If the problem is almost linear, then no difficulties usually appear, but if the problem is strongly nonlinear, then the iterative process may diverge.

For strongly nonlinear problems one may pass to a parametric set of problems. We illustrate this method taking as an example the boundary-value problem consisting of the system of differential equations

$$\mathbf{z}'(x) = \mathbf{f}\Big(x, \mathbf{z}(x)\Big) \tag{3.56}$$

and the system of boundary conditions

$$\mathbf{h}\Big(z_1(a), z_2(a), \dots, z_n(a), z_1(b), z_2(b), \dots, z_n(b)\Big) = 0. \tag{3.57}$$

Let us introduce a p and define functions $\mathbf{F}(x, \mathbf{z}, p)$ and $\mathbf{H}\left(\mathbf{z}(a), \mathbf{z}(b), p\right)$, such that $\mathbf{F}(x, \mathbf{z}, 1) = \mathbf{f}(x, \mathbf{z})$ and $\mathbf{H}\left(\mathbf{z}(a), \mathbf{z}(b), 1\right) = \mathbf{h}\left(\mathbf{z}(a), \mathbf{z}(b)\right)$, and for $p = 0$ the problem

$$\begin{cases} \mathbf{Z}'(x) = \mathbf{F}\Big(x, \mathbf{z}(x), p\Big), \\ \mathbf{H}\Big(Z_1(a), Z_2(a), \dots, Z_n(a), Z_1(b), Z_2(b), \dots, Z_n(b), p\Big) = 0 \end{cases} \tag{3.58}$$

is simple.

Then computing the solution of problem (3.58) for $p = 0$ one gets the vector-function $\mathbf{Z}(x, 0)$. This vector-function can be used as the initial approximation for problem (3.58) with some fixed sufficiently small parameter $p = p_1$. Again, computing the solution of this problem $\mathbf{Z}(x, p_1)$ one can use it as the initial approximation to the solution corresponding to a bit larger parameter $p = p_2$. Repeating the process and increasing the parameter p to $p = 1$ one finds the solution of the boundary-value problem (3.56), (3.57).

Chapter 4

Numerical methods for integral equations

4.1 Types of integral equations

The integral equation is the equation that contains the unknown function under the integration sign. We restrict ourselves to equations for the unknowns depending on one scalar argument and the integration is carried out along an interval on the real axis. A large class of one-dimensional integral equations can be presented in the form

$$y(x) = \int_a^b K\left(x, t, y(t)\right) dt, \qquad a \leq x \leq b. \tag{4.1}$$

Here $K(x, t, z)$ is a given function (the kernel), and the function $y(x)$ is the unknown. The equation of the form (4.1) are called *Uryson equations.*

If the kernel K is such that $K(x, t, z)$ does not depend on z for $t > x$ equation (4.1) can be rewritten in the form

$$y(x) = \int_a^x K\left(x, t, y(t)\right) dt + f(x), \tag{4.2}$$

where

$$f(x) = \int_x^b K(x, t, *) dt.$$

In this case the values of the function $y(x)$ depend only on its values $y(t)$ for $t < x$. Such equations are called *Volterra equations of the second kind.* Cauchy problems for differential equations can be reduced to such equations. Let for example the differential equation takes the form

$$y'(x) = f(x, y(x)).$$

Integrating reduces it to the equation

$$y(x) = y(a) + \int_a^x f(t, y(t))\, dt,$$

which is a particular case of (4.2).

Linear integral equations are mostly studied. The equations of the form

$$y(x) + \int_a^b K(x,t)y(t)dt = f(x) \tag{4.3}$$

with bounded (or weakly singular) kernel $K(x,t)$ are called *Fredholm equations of the second kind.*

In the case when the first term is absent

$$\int_a^b K(x,t)y(t)dt = f(x)$$

one deals with *Fredholm equation of the first kind* which is an ill-posed problem and requires regularization.

One can also formulate a spectral problem for the second kind Fredholm equations

$$\int_a^b K(x,t)y(t)dt = \lambda y(x),$$

being in finding such values of the spectral parameter λ, for which nontrivial solution $y(x)$ exists. Some kind of normalization condition is required to determine $y(x)$ uniquely.

Special type kernels of Fredholm integral equations that can be represented in the form

$$K(x,t) = \sum_{j=1}^{M} a_j(x) b_j(t),$$

are called degenerate kernels. Integral equations with degenerate kernels are equivalent to systems of M linear algebraic equations (we suppose that the functions a_j and b_j are known). Indeed, substituting the expression for the degenerate kernel into the integral equation (4.3) and changing the order of integration and summation reduces the equation to

$$y(x) + \sum_{j=1}^{M} B_j a_j(x) = f(x), \qquad B_j = \int_a^b b_j(t) y(t)\, dt.$$

Thus the solution $y(x)$ can be represented in the form of the decomposition

$$y(x) = f(x) - \sum_{k=1}^{M} B_k a_k(x).$$

Substituting this decomposition into the expression for B_j, yields the system of algebraic equations

$$B_j - \sum_{k=1}^{M} B_k \int_a^b b_j(t) b_k(t)\, dt = \int_a^b b_j(t) f(t)\, dt, \qquad j = 1, 2, \ldots, M. \tag{4.4}$$

System (4.4) is equivalent to the integral equation with degenerate kernel.

4.2 Method of iterations

Consider equation (4.1). The kernel $K(x,t,z)$ may weakly depend on the third argument, in a way that the inequality

$$\left\| \int_a^b \left(K(x,t,z_1(t)) - K(x,t,z_2(t)) \right) dt \right\| \le C \| z_1(t) - z_2(t) \| \tag{4.5}$$

holds with $C \leq 1$. In this case one can apply the method of simple iterations to the integral equation (4.1). This results in the following recurrent computations

$$y_0(x) \equiv 0, \qquad y_k(x) = \int_a^b K(x,t,y_{k-1}(t))dt, \quad k = 1,2,\ldots$$

The sequence of functions $y_k(x)$ converges to the solution $y(x)$ of the integral equation (4.1) under the supposition that condition (4.5) holds. Indeed, consider the difference $y_k(x) - y(x)$. Substituting here the expression for $y_k(x)$ and taking into account that $y(x)$ satisfies to the integral equation, one gets

$$y_k(x) - y(x) = \int_a^b \left(K(x,t,y_{k-1}(t)) - K(x,t,y(t))\right) dt.$$

Due to the inequality (4.5) the norm satisfyes

$$\|y_k(x) - y(x)\| \leq C\|y_{k-1}(x) - y(x)\|.$$

Repeating the same estimates for y_{k-1}, y_{k-2} ... , one finally finds

$$\|y_k(x) - y(x)\| \leq C^k\|y(x)\|.$$

At $C < 1$ the sequence $y_k(x)$ converges to $y(x)$.

For Volterra equations the sequence $y_k(x)$ converges under the following weaker condition

$$|K(x,t,z_1) - K(x,t,z_2)| \leq C|z_1 - z_2|, \tag{4.6}$$

where C is any finite constant. Similarly to the above, consider the difference $y_k(x) - y(x)$ and with the use of recurrent rule get

$$|y_k(x) - y(x)| = \left| \int_a^x \left(K(x,t,y_{k-1}(t)) - K(x,t,y(t))\right) dt \right|.$$

Now we use the fact that the absolute value of the integral is not greater than the integral of the absolute value and apply estimate (4.6). This gives the estimate

$$|y_k(x) - y(x)| \leq C \int_a^x |y_{k-1}(t) - y(t)|\, dt.$$

Using the recurrent formula once again yields

$$|y_k(x) - y(x)| \le C^2 \int_a^x \int_a^t |y_{k-2}(s) - y(s)|\, ds\, dt.$$

Changing the order of integration and computing the integral by t one finds

$$|y_k(x) - y(x)| \le C^2 \int_a^x (x-s)\, |y_{k-2}(s) - y(s)|\, ds$$

Repeating the similar derivations one gets

$$\begin{aligned} |y_k(x) - y(x)| &\le C^3 \int_a^x \frac{(x-s)^2}{2} |y_{k-3}(s) - y(s)|\, ds \\ &\le C^4 \int_a^x \frac{(x-s)^3}{6} |y_{k-4}(s) - y(s)|\, ds \\ &\le \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\le C^k \int_a^x \frac{(x-s)^{k-1}}{(k-1)!} |y(s)|\, ds. \end{aligned}$$

The last inequality states that with the increase of k the error $|y_k(x) - y(x)|$ tends to zero not slower than

$$C \frac{C^{k-1}(x-a)^{k-1}}{(k-1)!}.$$

Thus the sequence y_k converges to the solution exponentially.

Note that condition (4.6) is satisfied for any linear integral equations with bounded kernel. Also it holds for integral equations with kernels differentiable by the third argument and having bounded derivative.

For the equations other than Volterra equations the kernel should be small for the convergence of the method. If condition (4.5) is not satisfied, other methods are used.

4.3 Method of quadrature formulae

In the method of quadrature formulae the unknown function is replaced by a table of its values at a set of points which simultaneously play the role of the nodes of integration. If the function $y(x)$ is known then the integral in the right-hand side of (4.1) can be computed with the help of a quadrature formula. In the general form the quadrature formula can be presented as follows

$$\int_a^b f(t)dt \approx \sum_{j=1}^{N} w_j f(x_j). \tag{4.7}$$

Here w_j are the weights and x_j are the nodes. Using formula (4.7) one can replace the integral in the right-hand side of (4.1) with quadrature sum

$$y(x) = \sum_{j=1}^{N} w_j K\left(x, x_j, y(x_j)\right).$$

Now if only the values in the nodes are taken into account, that is the above equation is taken only for $x = x_k$, then the approximations y_j for the values $y(x_j)$ can be found from the system

$$\begin{cases} y_1 = \sum_{j=1}^{N} w_j K\left(x_1, x_j, y_j\right), \\ y_2 = \sum_{j=1}^{N} w_j K\left(x_2, x_j, y_j\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_N = \sum_{j=1}^{N} w_j K\left(x_N, x_j, y_j\right). \end{cases}$$

4.4 Collocation method

We shall consider only linear equations (4.3). In collocation method as well as in Bubnov-Galerkin method, discussed in the next section, the unknown function is searched in the form of the decomposition in a system of some functions $\varphi_j(x)$. This system should satisfy the following criteria:

1. functions $\varphi_j(x)$ should be integrable,
2. for any finite N the functions $\varphi_1(x), \varphi_2(x), \ldots, \varphi_N(x)$ should be linear independent,
3. any function can be approximated by the functions $\varphi_j(x)$ with arbitrary small error, that is the system $\{\varphi_j(x)\}$ should be complete.

Let us choose a finite number N and search for the solution $y(x)$ in the form

$$y(x) = \sum_{j=1}^{N} c_j \varphi_j(x) \tag{4.8}$$

with yet unknown coefficients c_j. Substituting representation (4.7) into integral equation (4.3) yields

$$\sum_{j=1}^{N} c_j \left(\varphi_j(x) + \int_a^b K(x,t)\varphi_j(t)\,dt \right) = f(x). \tag{4.9}$$

The integral equation and consequently the above equality should be satisfied for all values x from the interval $[a,b]$. Evidently that with the choice of the finite set of parameters c_j this can not be achieved in the general case. In the collocation method one requires equalities (4.9) to be satisfied at a set of preliminary chosen points x_k, which are called the *collocation points*. In total these equalities form the system of linear algebraic equations

$$\begin{cases} \displaystyle\sum_{j=1}^{N} c_j \left(\varphi_j(x_1) + \int_a^b K(x_1,t)\varphi_j(t)\,dt \right) = f(x_1), \\ \displaystyle\sum_{j=1}^{N} c_j \left(\varphi_j(x_2) + \int_a^b K(x_2,t)\varphi_j(t)\,dt \right) = f(x_2), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \displaystyle\sum_{j=1}^{N} c_j \left(\varphi_j(x_M) + \int_a^b K(x_M,t)\varphi_j(t)\,dt \right) = f(x_M). \end{cases}$$

For the solvability and uniqueness it is needed that the number of collocation points coincides with the number of unknown coefficients in decomposition (4.8). Due to the conditions posed on the functions $\varphi_j(x)$ this is also sufficient.

4.5 Bubnov-Galerkin method

Two systems of functions $\{\varphi_j(x)\}$ and $\psi_k(x)$ are used in Bubnov-Galerkin method. The first is called the coordinate system and is used in representation (4.8) for the solution similarly to collocation method. The other is used to project the residual

$$r(x) = \sum_{j=1}^{N} c_j \left(\varphi_j(x) + \int_a^b K(x,t)\varphi_j(t)\, dt \right) - f(x)$$

of equation (4.9). That is the system for the coefficients c_j is found from the requirement that the residual $r(x)$ is orthogonal to the first N functions of the system $\{\psi_k(x)\}$. The system $\{\psi_k(x)\}$ should be also linearly independent and complete. The system of equations takes the form

$$\sum_{j=1}^{N} c_j A_{jk} = B_k, \qquad k = 1, 2, \ldots, N,$$

where

$$A_{jk} = \int_a^b \int_a^b K(x,t)\varphi_j(t)\psi_k(x)\, dt\, dx, \quad B_k = \int_a^b f(x)\psi_k(x0\, dx.$$

Note that the matrix of that system is more complicated than in collocation methods and in the method of quadrature formulae. However, if the systems $\{\varphi_j(x)\}$ and $\{\psi_k(x)\}$ are appropriately chosen, the system that should be solved to get the solution with a given accuracy appears of less size than in the other methods.

Indeed, let the kernel $K(x,t)$ in (4.3) be represented in the form of a converging series

$$K(x,t) = \sum_{j=1}^{\infty} a_j(x) b_j(t). \tag{4.10}$$

Extracting a finite sum one can get the representation of the kernel in the form of the sum of the degenerate kernel $K_0(x,t)$ and the remaining series $K_1(x,t)$, that is

$$K(x,t) = K_0(x,t) + K_1(x,t),$$

where

$$K_0(x,t) = \sum_{j=1}^{M} a_j(x) b_j(t), \qquad K_1(x,t) = \sum_{j=M+1}^{\infty} a_j(x) b_j(t).$$

For sufficiently large M the kernel $K_1(x,t)$ will be small due to the convergence of the series (4.10).

If one neglects the correction $K_1(x,t)$ to the kernel, then the integral equation is replaced by the integral equation with degenerate kernel which is equivalent to a system of algebraic equations (4.4). Let the solution of this system be written as

$$B_j = \sum_{k=1}^{M} A_{jk} \int_a^b b_j(t) f(t)\, dt. \tag{4.11}$$

If one does not neglect the correction, then the solution can be searched in the form

$$y(x) = y^{(0)}(x) + y^{(1)}(x), \tag{4.12}$$

where

$$y^{(0)}(x) = f(x) - \sum_{j=1}^{M} B_j a_j(x),$$

and $y^{(1)}(x)$ is some reminder. Temporally assume the function $y^{(1)}(x)$ to be known. Then, substitution of representation (4.12) into the integral in (4.3) yields

$$y^{(0)}(x) + y^{(1)}(x) = f(x) - \sum_{j=1}^{M} B_j a_j(x) - \int_a^b K_1(x,t) \left(y^{(0)}(t) + y^{(1)}(t) \right) dt, \tag{4.13}$$

where

$$B_j = \int_a^b b_j(t) \left(y^{(0)}(t) + y^{(1)}(t) \right) dt.$$

After substituting the expression for $y^{(0)}(x)$ one gets the system for the coefficients B_j

$$B_j + \sum_{k=1}^{M} B_k \int_a^b b_j(t)a_k(t)\,dt = \int_a^b b_j(t)f(t)\,dt + \int_a^b b_j(t)y^{(1)}(t)\,dt.$$

Represent the solution of that system in the form similar to (4.11), namely

$$B_j = \sum_{k=1}^{M} A_{jk} \left(\int_a^b b_k(s)f(s)\,ds + \int_a^b b_k(s)y^{(1)}(s)\,ds \right).$$

Now substitute the above expressions into equation (4.13) and obtain

$$\begin{aligned} y^{(1)}(x) &= \int_a^b K_1(x,t) \left(f(t) - \sum_{j=1}^{M} B_j a_j(t) + y^{(1)}(t) \right) dt \\ &= F(x) + \int_a^b \tilde{K}(x,t)y^{(1)}(t)\,dt, \end{aligned} \tag{4.14}$$

where

$$F(x) = \int_a^b K_1(x,t)f(t)\,dt - \sum_{j=1}^{M} \int_a^b K_1(x,t)a_j(t)\,dt \sum_{k=1}^{M} A_{jk} \int_a^b b_k(s)f(s)\,ds,$$

$$\tilde{K}(x,t) = K_1(x,t) - \sum_{j=1}^{M} \sum_{k=1}^{M} A_{jk} \int_a^b K_1(x,s)a_j(s)\,ds b_k(t).$$

Formula (4.14) can be treated as an integral equation for the function $y^{(1)}(x)$. The kernel of this equation is proportional to small $K_1(x,t)$ and if condition (4.5) is satisfied for (4.14) it can be solved by iterations method.

Chapter 5

Program packages for numerical solution of ODE

5.1 Packages for broad range of application

First the well-known packages for broad range of application are discussed.

5.1.1 Maple 10

The package Maple is developed by Wateloo company. It includes the function *dsolve* with several parameters which computes the numerical solution of an initial value problem for linear and nonlinear ODEs as well as systems of ODEs. Maple also includes an additional package *DEtools* which gives more tools for plotting the results, for preliminary transformations of the equation etc. The boundary value problem can be solved, for instance by use of finite difference method and further application of the numerical package *LinearAlgebra*. Another possibility is proposed by Maple Power Tools a set of additional programs attached to the major package of Maple. It can be found on the web-site of Waterloo.

5.1.2 Mathematica 5.1

In this package produced by Wolfram Research company we also find the function *NDSolve* with several parameters for solution of initial value problems.

5.1.3 COMSOL MultiPhysics

In this package (previously known as FEMLAB) the finite elements method is used for solution of different boundary value problems.

5.1.4 NAG (Numerical Algorithm Group)

This package was developed mostly for UNIX platforms. It includes several programs for solution of ODEs. The original FORTRAN codes of earlier versions of them can be found. However, the interface facilities are very poor.

5.2 Specific ODE's solvers

The list of specific ODE's solvers is larger. First we mention the package BARSIC SLEIGN2 developed by Monachov, Matveeva and Kernitskii. It is characterized by friendly interface and enables to solve Sturm-Liouville problems. Solution of the Sturm-Liouvelle problems is based on Pruefer transform to amplitude-phase variables and the package SLEIGN2 taken as a basic source. The package SLEIGN2 was coded by Everitt, Zettle, Hinton and Baily. The package SLEDGE coded by Pruess and Fulton also gives a tool for solution Sturm-Lioville problems but the mathematical background is different. On the intervals where coefficients of the equation are taken as constants the explicit solutions are presented in the form of elementary functions and further a matching procedure is used. In the Laboratory of Informational Technologies of the Joint Institute of Nuclear Research several programs of numerical solution of Sturm-Liouville problems depending on a parameter have been developed on the basis of the abstract Newton method. The authors of these programs are Pusinina and Pusinin.

As a tool for solution of initial value problem the library ODEPACK by Hindmarsh can be proposed. It can be used for both stiff and nonstiff problems. The friendly interface for programs consisting ODEPACK was also developed by Monachov, Matveeva and Kernitskii.

Bibliography

[1] M. T. Heath, "Scientific Computing. An Introductory Survey", 2nd Ed McGraw Hill 2002.

[2] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, "Numerical Recipes in C", Cambridge University Press, 1996.

[3] R. Schilling, S. Harris, "Applied Numerical Methods for Engineers". Brooks/Cole Publishing, 2000.

[4] W. Cheney, D. Kincaid, "Numerical Mathemetics and Computing", (4th Ed.) ITP Books, 1999.

[5] A. Quarteroni, R. Sacco, F. Saleri, "Numerical Mathematics", Springer Verlag, 2000.

[6] C. Moler, "Numerical Computing with MATLAB", SIAM books 2004.

[7] G. Borse, "Numerical Methods with MATLAB", ITP Books, 1997.