NUMERICAL ANALYSIS

Josef Dalík

Department of Mathematics, Brno University of Technology
Žižkova 17, 662 37 Brno, Czech Republic
e–mail: dalik.j@fce.vutbr.cz

Contents

1 Introduction

1.1 What is a numerical analysis

Today, the use of mathematical tools for the solution of problems is non–replaceable not only in traditional areas like natural or technical sciences, but also in economy, medicine, social sciences and public administration.

Among a large amount of various definitions of this subject, we propose the following one: *Numerical analysis is a science whose aim is to make the achievements of "pure mathematics" applicable for the solution of practical problems.*

According to this definition, numerical analysis is an area characterized by the purpose and not by the theoretical tools which it is using. The development has shown that there exists a limited number of basic tools such that the most tools of numerical analysis are suitable combinations of them. In this book, we concentrate on explanation of these basic tools.

A general description of the use of numerical analysis may be as follows.

1. *Given problem:* To get some "objective knowledge" about a concrete natural, technical or social process.

Example 1. Determine the flow of a liquid or a gas, the process of a chemical reaction, deformation of a body, behaviour of a social group, ...

2. (To create a new or choose an existing) *mathematical model:* An exact mathematical description of essential relations determining the given process.

Example 2. Initial– or boundary–value problem for differential equations, system of equalities or inequalities, a functional, a system of linear or non–linear algebraic equations, a definite integral, ...

The models are studied by various branches of mathematics which provide essential information like existence and/or uniqueness of equations and/or

their qualitative properties. Exceptionally, mostly in more or less trivial cases, they provide a way how to find solutions of the models.

This is where the numerical analysis begins: We give up the search after "exact solutions" and concentrate on "approximate solutions (on approximations)" and on algorithms for their computation. The result is a so–called

3. *Numerical problem* or *discretization* characterized by a uniquely determined

- finite set of *input data*
- finite set of *output data*
- finite sequence of steps transforming the input data to the output data

4. Computations of concrete output data for concrete given input data

A *numerical method* is a way how to relate discretizations to models, simplify them and how to transform the given input data to the output data. *Numerical analysis* is a branch of mathematics whose aim is to propose new numerical methods, to study the quality of the old ones and to propose suitable structures of the input data (preprocessing) and of the output data (postprocessing). Usually, the quality of a numerical method is determined by its

- effectivity,
- accurracy,
- robustness.

Numerical analysis is based on linear algebra, calculus, and functional analysis. Moreover, an important part of numerical analysis consists of complexity of algorithms.

1.2 What is an error

If we solve a practical problem by a numerical method, we have

a) real problem given. We find

b) mathematical model, approximate it by

c) discretization and arrive at

d) the result

In each of the steps a) – d), another type of error appears: We call the difference between the exact solution of

$$\begin{cases} \text{a) and b) an } \textit{error of the model} \\ \text{b) and c) a } \textit{discretization error } (\textit{error of the numerical method}) \\ \text{c) and d) a } \textit{truncation error} \end{cases}$$

The common meaning of errors, i. e. mistakes (leeps, oversights) have also to be taken into account. The amount of them decreases whenever the time–schedule of work is planned reasonably, people work under good conditons, there are good relations between collaborators etc.

1.3 Some basic principles of the numerical analysis

a) *Iteration.* As the name suggests, it consists of repeating the same pattern of computation with the aim to improve the accurracy of the previous approximation. Iterative techniques are used to find roots of equations, solutions of systems of linear and non–linear equations, and solutions of differential equations. As an illustration, let us solve the equation

$$x = F(x)$$

for x real and F continuous on $\Re$. Solution by iteration consists in choosing an x_0 (starting value) and computing

$$\begin{aligned} x_1 &= F(x_0) \\ x_2 &= F(x_1) \\ &\vdots \\ x_{n+1} &= F(x_n) \\ &\vdots \end{aligned}$$

If this (theoretically) infinite *iterative sequence* (sequence of consecutive approximations) has a limit $\hat{x}$ then

$$\hat{x} = \lim_{n\to\infty} F(x_n) = F(\lim_{n\to\infty} x_n) = F(\hat{x}),$$

so that the limit $\hat{x}$ is the solution of our equation.

Example 3. (A quick computation of a square root) For $c > 0$ and $x \neq 0$ real, we have

$$\begin{aligned} x^2 &= c \\ 0 &= -x^2 + c \\ 0 &= (-x + \frac{c}{x})\frac{1}{2} \\ x &= (x + \frac{c}{x})\frac{1}{2} \equiv F(x). \end{aligned}$$

The iteration for $\sqrt{2}$ ($c = 2$), i. e.

$$x_0 = 1.5, \quad x_{n+1} = (x_n + \frac{2}{x_n})\frac{1}{2} :$$

n	x_n
0	1.5
1	1.4167
2	1.414216

and $\sqrt{2} = 1.4142141$.

b) *Local approximation of a complicated function by a linear function*

Let us consider the equation $f(x) = 0$. Graphically, we look after an intersection of the curve $y = f(x)$ with $y = 0$. Assume we know a starting point x_0 approximating the solution $\hat{x}$. We substitute the curve $y = f(x)$ by its tangent line in the point $[x_0, f(x_0)]$ and, instead of the original problem, we find the intersection of this tangent line with $y = 0$:

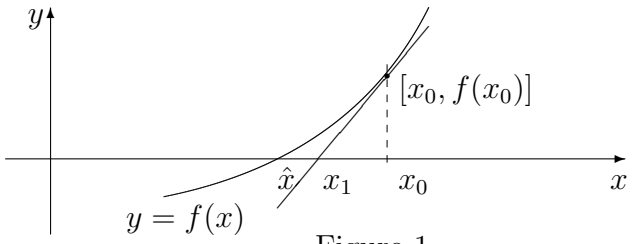


Figure 1

In combination with iteration, we obtain the popular Newton method:

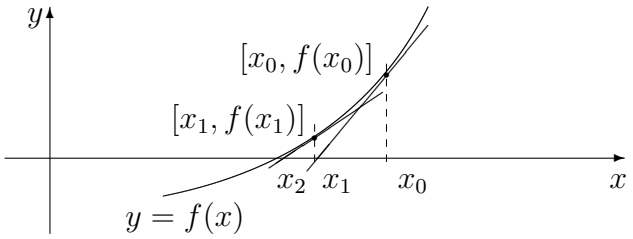


Figure 2

If we use a local approximation by secant for the approximation of the integral

$$I = \int_a^b f(x)\,dx, \quad \text{i. e.}$$

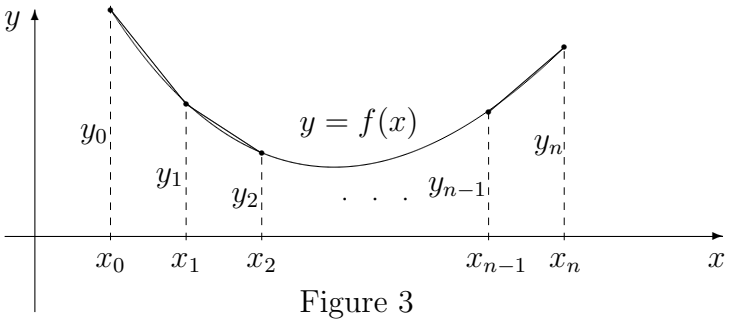


Figure 3

we obtain the numerical problem to compute the sum

$$T(h) = \frac{h}{2}\sum_{i=1}^{n}(y_{i-1} + y_i).$$

This numerical method is called a *Trapezoidal rule*. Later we show that the error $T(h) - I$ is proportional to h^2. The decrease of h means the increase of the amount of computations. If we want to gain a more exact approximation more effectively, we can use one of the following two important ideas:

i) We approximate the integrand $y(x)$ locally by a polynomial of a higher degree.

ii) We use the Trapezoidal rule with two different values of h and apply extrapolation.

We illustrate the extrapolation only: If we compute $T(h)$ and $T(2h)$ for example, we obtain

$$I - T(h) \approx K \cdot h^2, \quad I - T(2h) \approx K \cdot (2h)^2.$$

Then

$$4(I - T(h)) \approx I - T(2h)$$

and we obtain

$$I \approx (4T(h) - T(2h))/3.$$

Example 4. Compute the integral $\int_{10}^{12} f(x)\,dx$ for $f(x) = x^3$ and $f(x) = x^4$ by the Trapezoidal rule. Improve the approximations by extrapolation.

	$f(x) = x^3$	$f(x) - x^4$
$y(10)$	1 000	10 000
$y(11)$	1 331	14 641
$y(12)$	1 728	20 736
$T(2)$	2 728	30 736
$T(1)$	2 695	30 009
$(4T(1) - T(2))/3$	2 684	29 766.8
exact value	2 684	29 766.4

Very often, the kernel of the model consists of one or more ordinary or partial differential equations. To find effective and accurate methods for approximate solutions of differential problems is thus one of the most important aims of numerical analysis.

Example 5. (Derivation of an initial–value problem, I.V.P.)

Physical law: *The velocity of decay of a radioactive material is proportional to its mass.*

If we denote by $y(t)$ the mass of radioactive material at the time t then we have

$$\begin{array}{ll} \frac{y(t+h)-y(t)}{h} & \text{average velocity of decay} \\ y'(t) = \lim_{h\to 0} \frac{y(t+h)-y(t)}{h} & \text{velocity of decay at time } t \end{array}$$

The mathematical form of the physical law is now

$$y'(t) = -k\,y(t).$$

Here $k > 0$ is a constant characterizing the concrete radioactive material. By separation of variables we obtain

$$\begin{aligned} \frac{y'}{y} &= -k \\ \ln|y| &= -k\,t + C \\ y(t) &= Ce^{-k\,t} \quad (C \in \Re \text{ is arbitrary}) \end{aligned}$$

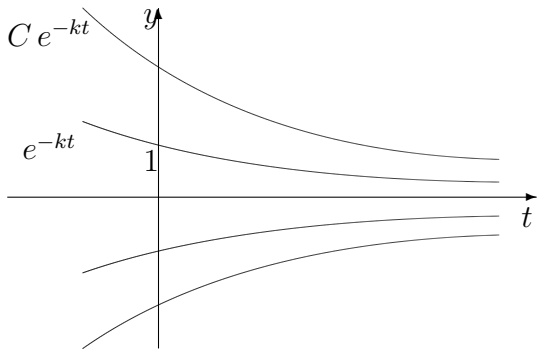


Figure 4

For unicity, it is sufficient to determine the value of mass $y(t)$ for one value of time t (to add the *initial condition*). We obtain the I. V. P.

$$y'(t) = -k\,y(t), \quad y(a) = y_0.$$

This I. V. P. has an exact solution

$$y(t) = y_0 e^{-k(t-a)}.$$

General form of the I. V. P. is

$$y'(x) = f(x, y), \quad y(0) = c.$$

This equation tells us that the slope of $y(x)$ is $f(x, y(x))$ in any point x.

The most simple numerical method (the Euler method) consists in

- choosing a *discretization step* $h > 0$
- taking $x_0 = 0, x_1 = h, x_2 = 2h, \ldots$
- substituting the derivative y' by a constant between any two consecutive points:

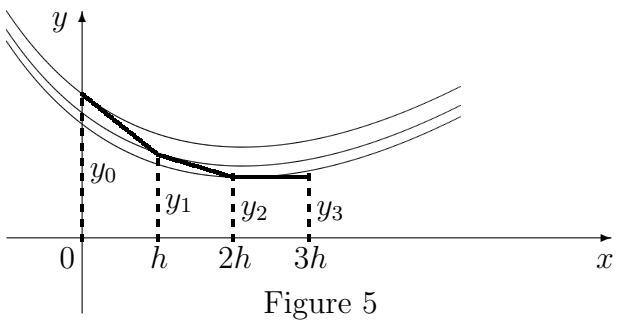


Figure 5

The Euler polygon connects the points $[0, y_0]$, $[h, y_1]$, $[2h, y_2]$, ... and

$$y_0 = c, \quad \frac{y_{i+1} - y_i}{h} = f(x_i, y_i),$$

so that

$$y_0 = c, \quad y_{i+1} = y_i + h\, f(x_i, y_i), i = 0, 1, \ldots$$

Example 6. Approximate the solution of the I. V. P.

$$y' = -0.5y \ \text{ in } \ (0, 1), \quad y(0) = 1.$$

By the Euler method with steps $h = 0.5$ and $h = 0.25$.

For $h = 0.5$ we have $y_0 = 1$ and $y_{i+1} = y_i + 0.5(-0.5y_i) = 0.75y_i$ for $i = 0, 1$:

i	x_i	y_i
0	0	1
1	0.5	0.75
2	1	0.5625

In the case of $h = 0.25$ we have $\overline{y}_0 = 1$ and $\overline{y}_{i+1} = \overline{y}_i + 0.25(-0.5\overline{y}_i) = 0.875\overline{y}_i$ for $i = 0, 1, 2, 3$:

i	$\overline{x}_i$	$\overline{y}_i$
0	0	1
1	0.25	0.875
2	0.5	0.765625
3	0.75	0.669922
4	1	0.586182

Exact solution $y(x) = e^{-0.5x}$ has exact values $y(0.5) = 0.778801$ and $y(1) = 0.6065307$ in the points from the rough mesh. Comparison of the errors in these points from the following table says that the order of error is proportional to h.

i	x_i	$y_i - y(x_i)$	$\overline{y}_{2i} - y(\overline{x}_{2i})$
1	0.5	-0.028801	-0.013176
2	1	-0.044031	-0.020349

In many applications, variable x means time and the differential equation represents a law controlling the changes of the system being considered. Our approximate solution then means a *numerical simulation* of these time–changes.

2 Error analysis

Definition 1. Let $\tilde{x}$ be the approximation of the exact value x. Then we denote by

$$E_x = x - \tilde{x}$$

the *(absolute) error*, by

$$R_x = \frac{E_x}{x}$$

the *relative error*, by

$$\varepsilon(x) \ : \quad |E_x| < \varepsilon(x)$$

an *error estimate* and by

$$\delta(x) \ : \ |R_x| < \delta(x)$$

a *relative error estimate*.

Remark 1. We of course have

$$|E_x| < \varepsilon(x) \Longleftrightarrow x \in (\tilde{x} - \varepsilon(x), \tilde{x} + \varepsilon(x)) \equiv x = \tilde{x} \pm \varepsilon(x).$$

Example 7. If we approximate the number $x = 2.7182818$ by $\tilde{x} = 2.72$ then we have

$$E_x = x - \tilde{x} = -0.001728, \quad R_x = \frac{E_x}{x} = -0.000635294.$$

In the memories of computers, numbers are stored in the following two basic ways:

a) *floating point mode:*

$$\tilde{x} = \pm .d_1 d_2 \dots d_k \cdot 10^e$$

is a decimal *floating point* representation of x with k digits significant whenever

$$|x - \tilde{x}| < 5 \cdot 10^{e-k-1}.$$

In the memory,

- the sign $\pm$,
- the *mantissa* $.d_1 d_2 \dots d_k \in [.10 \dots 0, .99 \dots 9]$

– the integer *exponent* $e \in [-L, L]$

are saved.

In this mode, non–zero numbers x such that

$$|x| < 0.1 \cdot 10^{-L} \quad \text{and} \quad |x| > 0.99\ldots9 \cdot 10^{L}$$

cannot be represented. We speak about an *underflow* and *overflow* respectively.

b) *fixed–point mode:* Numers are saved with the same number n of digits to the right from the decimal point.

Example 8. Let $x = 21.4976$. Then

a) The floating point representation with $k = 5$ is

$$\tilde{x} = +0.21498 \cdot 10^{2}.$$

b) The fixed–point representation with $n = 0$ is

$$\tilde{x} = 21.$$

We can see that even the above computer representations $\tilde{x}$ are approximations of the exact values x. Now, we use the Taylor theorem for the development of basic laws characterizing the way in which basic arithmetic operations change the error of this approximation. We first evaluate an arbitrary smooth function

$$f(x, y)$$

under the assumption that instead of x, y, we know the approximations

$$\tilde{x} = x + \Delta x, \quad \tilde{y} = y + \Delta y.$$

We have

$$\begin{aligned} f(\tilde{x}, \tilde{y}) &= f(x, y) + \Delta x \frac{\partial f}{\partial x}(x, y) + \Delta y \frac{\partial f}{\partial y}(x, y) \\ &+ \frac{1}{2}\left((\Delta x)^2 \frac{\partial^2 f}{\partial x^2} + 2\Delta x \Delta y \frac{\partial^2 f}{\partial x \partial y} + (\Delta y)^2 \frac{\partial^2 f}{\partial y^2} \right)(x, y) + \ldots \end{aligned}$$

If we assume the values $(\Delta x)^2$, $\Delta x \Delta y$ and $(\Delta y)^2$ to be too small, we obtain

$$|\Delta f(x, y)| = |f(\tilde{x}, \tilde{y}) - f(x, y)| \leq |\Delta x| \left| \frac{\partial f}{\partial x}(x, y) \right| + |\Delta y| \left| \frac{\partial f}{\partial y}(x, y) \right|$$

and

$$\begin{aligned}\left|\frac{\Delta f(x,y)}{f(x,y)}\right| &\doteq \left|\frac{x}{f(x,y)}\frac{\partial f(x,y)}{\partial x}\frac{\Delta x}{x}+\frac{y}{f(x,y)}\frac{\partial f(x,y)}{\partial y}\frac{\Delta y}{y}\right| \\ &\leq \left|\frac{x}{f(x,y)}\frac{\partial f(x,y)}{\partial x}\right|\left|\frac{\Delta x}{x}\right|+\left|\frac{y}{f(x,y)}\frac{\partial f(x,y)}{\partial y}\right|\left|\frac{\Delta y}{y}\right|.\end{aligned}$$

Example 9. We derive the following formulas for absolute and relative errors of results of basic arithmetic operations by putting $f(x,y) = x \pm y$, $f(x,y) = x \cdot y$ and $f(x,y) = x/y$.

1)

$$\begin{aligned}\Delta(x \pm y) &\doteq \Delta x \pm \Delta y \\ \frac{\Delta(x \pm y)}{x \pm y} &\doteq \frac{x}{x \pm y}\frac{\Delta x}{x} \pm \frac{y}{x \pm y}\frac{\Delta y}{y}\end{aligned}$$

2)

$$\begin{aligned}\Delta(xy) &\doteq y\Delta x + x\Delta y \\ \frac{\Delta(xy)}{xy} &\doteq \frac{\Delta x}{x}+\frac{\Delta y}{y}\end{aligned}$$

3)

$$\begin{aligned}\Delta\left(\frac{x}{y}\right) &\doteq \frac{1}{y}\Delta x - \frac{x}{y^2}\Delta y \\ \frac{\Delta\left(\frac{x}{y}\right)}{\frac{x}{y}} &\doteq \frac{\Delta x}{x}-\frac{\Delta y}{y}\end{aligned}$$

Example 10. (Loss of significant digits) If we solve the quadratic equation

$$x^2 - 56x + 9 = 0$$

by the well–known formula using the fixed–point mode with $n = 3$, we obtain

$$\begin{aligned}x_{1,2} &= \frac{56 \pm \sqrt{56^2-4}}{2} = 28 \pm \sqrt{28^2-1} \\ &= 28 \pm 27.982 = \begin{cases} 55.982 \\ 0.018 \pm 0.0005 \end{cases}\end{aligned}$$

The first root has 5 significant digits, but the second root has 2 significant digits only. One can see immediately that the reason for this loss of 3 significant digits is in the subtraction of two numbers of approximately equal value. This difficulty appears quite often in computations. We suggest the following two remedies.

a) $\sqrt{784} - \sqrt{783} = \frac{784-783}{\sqrt{784}+\sqrt{783}} \doteq 0.017863 \pm 5 \cdot 10^{-7}$

b) For $f(x) = \sqrt{x}$ and $x = 784$, we obtain by the Taylor theorem

$$f(x-1) = f(x) - f'(x) + \tfrac{1}{2} f''(x)$$
$$f(x) - f(x-1) = \tfrac{1}{56} + \tfrac{1}{8 \cdot 28 \cdot 784} = 0.017863 \pm 5 \cdot 10^{-7}$$

Definition . Algorithms for which the cumulative effect of truncation errors is limited are called stable.

Example 11. Evaluate the integrals

$$y_i = \int_0^1 \frac{x^i}{x+1} dx \quad \text{for} \quad i = 0, 1, \ldots, 6.$$

Round off to 4 decimals.

For effectivity of evaluations, we use the following recursive relation:

$$y_i + 10 y_{i-1} = \int_0^1 \frac{x^i + 10x^{i-1}}{x+10} dx = \int_0^1 x^{i-1} dx = \frac{1}{i}$$

If we use the fact that $y_0 = [\ln|x+1|]_0^1 = \ln 1.1 \doteq 0.0953$ and, by means of the above formula in the form $y_i = 1/i - 10 y_{i-1}$, we obtain

$$\begin{aligned}
y_1 &= 1 - 10 y_0 = 1 - 0.953 = 0.0470 \\
y_2 &= \frac{1}{2} - 10 y_1 = 0.5 - 0.47 = 0.0300 \\
y_3 &= \frac{1}{3} - 0.3 = 0.0333 \\
y_4 &= \frac{1}{4} - 0.333 = -0.083 \\
&\vdots
\end{aligned}$$

As all the values of y_i have to be positive obviously, value of y_4 is a nonsense. This algorithm is non–stable for the following reason: At the beginning, we have substituted the exact value $y_0 = \ln 1.1 = 0.09531018$ by its approximation $\tilde{y}_0 = 0.0953$ with the error $E_0 = 0.00001018$. If we neglect other sources

of error, we see that this error E_0 is multiplied by 10 in every step. Hence its values are approximately $E_1 = 0.0001018$, $E_2 = 0.001018$, $E_3 = 0.01018$, $E_4 = 0.1018$, ... We can see that, starting with E_4, these errors dominate.

If we use our recursive relation in the "opposite direction", i. e.

$$y_{i-1} = 0.1(\frac{1}{i} - y_i),$$

the error broth by y_i decreases 10 times in every step. By putting $y_7 = 0$, we obtain

$$\begin{aligned} y_6 &= \frac{1}{70} = 0.0143 \\ y_5 &= 0.1\left(\frac{1}{6} - y_6\right) = 0.0152 \\ y_4 &= 0.1\left(\frac{1}{5} - y_5\right) = 0.0185 \\ y_3 &= 0.1\left(\frac{1}{4} - y_4\right) = 0.0232 \\ y_2 &= 0.0310 \\ y_1 &= 0.0469 \\ y_0 &= 0.0953 \end{aligned}$$

which are exact for all indexes less that 5.

3 The non–linear equation $f(x) = 0$

Problem. Let f be a real function defined on an interval $I \subseteq \Re$. Find a number $x \in I$ such that

$$f(x) = 0. \tag{1}$$

The number x satisfying (1) is called a *root* of (1). Equation (1) is said to be

$$\begin{cases} \textit{algebraic} \text{ whenever } f(x) \text{ is a polynomial} \\ \textit{transcendental} \text{ otherwise} \end{cases}$$

Example 12. Find approximations of all roots of the equation

$$f(x) \equiv 10 \sin x - x - 5 = 0$$

from the interval $(0, \pi)$.

Graphical method consists in splitting the function $f(x)$ into two functions whose graph can be drawn schematically. In our case, we have

$$f(x) = 0 \Longleftrightarrow 10 \sin x = x + 5.$$

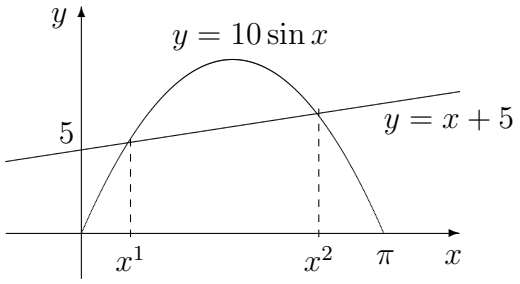


Figure 6

From this illustration, we can see that there are two roots of our equation, $x^1 \doteq 0.5$ and $x^2 \doteq 2.4$. We can obtain more accurate information from the following table.

x	$f(x)$
0.5	-0.7057
0.7	0.7422
2.4	-0.6454
2.2	0.8850

From the values of f in the points 0.5 and 0.7, we conclude that $x^1 \in (0.5, 0.7)$ and the values in the remaining two points tell us that $x^2 \in (2.2, 2.4)$. We have used the following important statement.

Theorem 1. If a function f is continuous on a closed interval $[a, b]$, i. e. $f \in C[a, b]$, and

$$f(a) \cdot f(b) < 0$$

then there exists $x \in (a, b)$ such that $f(x) = 0$.

The following two methods use Theorem 1 repeatedly. For

$$f \in C[a_0, b_0], \quad f(a_0) \cdot f(b_0) < 0$$

given, both construct intervals

$$[a_0, b_0] \supset [a_1, b_1] \supset \ldots \supset [a_n, b_n] \supset \ldots$$

such that

$$f(a_n) \cdot f(b_n) < 0 \quad \text{for} \quad n = 1, 2, \ldots$$

The new interval $[a_n, b_n]$ is constructed in the following way: A point $s_n \in (a_{n-1}, b_{n-1})$ is chosen. As $f(a_{n-1}, b_{n-1}) < 0$, exactly one of the following conditions 1 – 3 is valid.

1. $f(a_{n-1}) \cdot f(s_n) < 0$. In this case we put $a_n = a_{n-1}$, $b_n = s_n$.
2. $f(s_n) \cdot f(b_{n-1}) < 0$. In this case we put $a_n = s_n$, $b_n = b_{n-1}$.
3. $f(s_n) = 0$. In this case, the computation stops.

The method of *bisection* computes

$$s_n = \frac{1}{2}(a_{n-1} + b_{n-1})$$

and, for $\varepsilon > 0$ given, it stops whenever $(b_n - a_n)/2 < \varepsilon$. Then $x \doteq s_{n+1}$.

Velocity of convergence of bisection: At the beginning we have

$$x = s_1 \pm \varepsilon_0 \quad \text{for} \quad \varepsilon_0 = \frac{b_0 - a_0}{2}$$

and after n steps we have

$$x = s_{n+1} \pm \varepsilon_n \quad \text{for} \quad \varepsilon_n = \frac{b_n - a_n}{2}$$

As

$$\varepsilon_n = \frac{b_n - a_n}{2} = \frac{b_{n-1} - a_{n-1}}{4} = \ldots = \frac{b_0 - a_0}{2^{n+1}},$$

we have $\varepsilon_n = \varepsilon_0/2^n$ for $n = 1, 2, \ldots$

Example 2. How many steps decrease the error estimate 10 times?
n steps decrease the error–estimate 2^n times
the least n such that

$$10 \leq 2^n$$

is $n = 4$ ($10 \doteq 2^{3.3}$), so that 4 steps (3.3 steps) are needed.

Example 3. Approximate the root x^1 from Example 1 with an error less than 10^{-3}.

$$x^1 \in (0.5, 0.7) \Longrightarrow x^1 = 0.6 \pm 0.1.$$

Hence we have $d_0 = 0.1$ and we search the least n such that

$$\frac{0.1}{2^n} < 10^{-3}.$$

This is equivalent to $100 < 2^n$ and this is equivalent to $7 \leq n$. That is why we need at least 7 steps of bisection.

i	a_{i-1}	sgn$f(a_{i-1})$	b_{i-1}	sgn$f(b_{i-1})$	s_i	sgn$f(s_i)$
1	0.5	-	0.7	+	0.6	+
2	0.5	-	0.6	+	0.55	-
3	0.55	-	0.6	+	0.575	-
4	0.575	-	0.6	+	0.5875	-
5	0.5875	-	0.6	+	0.59375	+
6	0.5875	-	0.59375	+	0.590625	-
7	0.590625	-	0.59375	+	0.5921875	-
8	0.5921875	-	0.59375	+	0.59296875	

Hence we have $x^1 = 0.59296875 \pm 0.001$ and more exactly $x^1 = 0.59296875 \pm 0.00078125$.

The *regula falsi* method computes

$$s_n = a_{n-1} - \frac{(a_{n-1} - b_{n-1}) f(a_{n-1})}{(f(a_{n-1}) - f(b_{n-1}))}$$

and, for $\delta > 0$ given, the method stops whenever $|f(s_n)| < \delta$. Then $x \doteq s_n$.

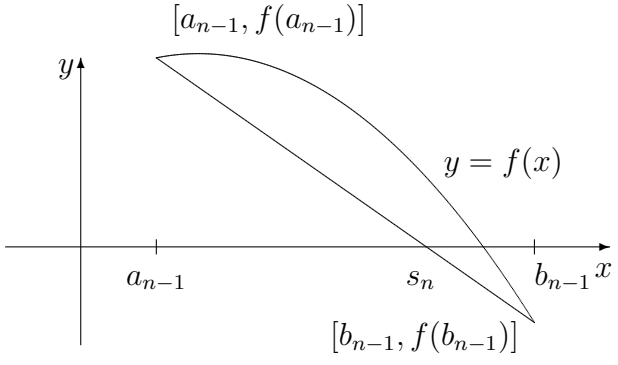


Figure 7

In the following Fig. 8 we illustrate situation in which this method terminates incorrectly and Fig. 9 illustrates a situation in which the effectivity of the regula falsi method is much worse than that of bisection.

Figure 8

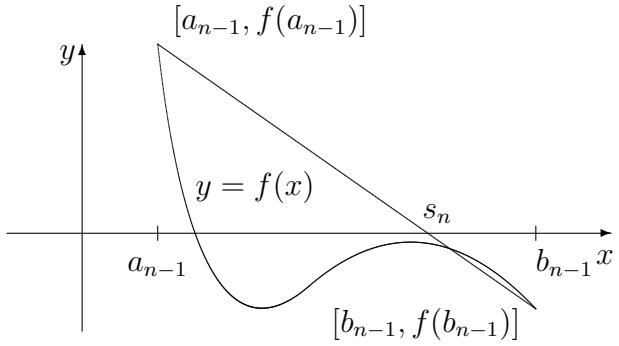
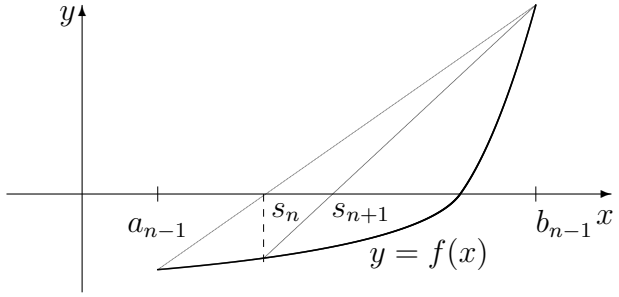


Figure 9



Example 4. Find the root of the equation $10 \sin x - x - 5 = 0$ from the interval $(0.5, 0.7)$ with accurracy $\delta = 0.0002$.

The solution is presented in the following table.

n	a_{n-1}	b_{n-1}	$f(a_{n-1})$	$f(b_{n-1})$	s_n	$\lvert f(s_n)\rvert$
1	0.5	0.7	-0.705745	é.742177	0.597484	0.028156
2	0.5	0.597484	-0.705745	0.028156	0.593744	0.000938
3	0.5	0.593744	-0.705745	0.000938	0.59361957	0.000031

We can see that $x^1 \doteq 0.59361957$.

4 Iteration

We repeat that the first stem in solving an equation $f(x) = 0$ consists in rewriting the given equation in the equivalent form $x = F(x)$. In what follows, we investigate equations of this form.

Definition . Let X be a non–empty set and $F : X \longrightarrow X$ be a map. An element $x \in X$ is called a *fixed point* of the map F whenever $x = F(x)$.

ITERATION: Starting point (*initial approximation*) $x_0 \in X$ is choosen

and the *consecutive approximations* $x_1, x_2, \ldots$ are computed by the formula

$$x_{i+1} = F(x_i)$$

for $i = 0, 1, \ldots$

If the map F is continuous and if $\lim_{i \longrightarrow \infty} x_i = x$ then we already know that $x = F(x)$. In order to make this consideration exact, we first have to answer the question what does it mean that F is continuous and that $\lim_{i \longrightarrow \infty} x_i = x$.

Definition . Let be $X \neq \emptyset$ and $d : X \times X \longrightarrow \Re$. If

D1 $d(x, y) \geq 0$, and $d(x, y) = 0 \Longleftrightarrow x = y$,

D2 $d(x, y) = d(y, x)$,

D3 $d(x, y) \leq d(x, z) + d(z, y)$

then we call X, more exactly the ordered pair (X, d) a *metric space*, elements $x \in X$ *points*, the map d a *metric* and the value $d(x, y)$ a *distance* between x and y.

Definition . Let X be a metric space, $(x_i)_0^\infty \subseteq X$ and let $x \in X$. We put

$$x = \lim_{i \longrightarrow \infty} x_i$$

whenever

$$d(x_i, x) \longrightarrow 0 \quad \text{as} \quad i \longrightarrow \infty.$$

A more precise formulation of this condition is the following:

$$\forall \varepsilon > 0 \exists i_0 > 0 \ : \ d(x_i, x) < \varepsilon \forall i > i_0.$$

Theorem 1. Every sequence in a metric space has at most one limit.
Proof. If we admit $x = \lim_{i \longrightarrow \infty} x_i$ and $y = \lim_{i \longrightarrow \infty} x_i \Longrightarrow$

$$0 \leq d(x, y) \leq d(y, x_i) + d(y, x_i) \longrightarrow 0$$

as $i \longrightarrow \infty \Longrightarrow x = y$.

Definition . Let X be a metric space and $(x_i)_0^\infty \subseteq X$. We say that $(x_i)_0^\infty$ is a *Cauchy (fundamental) sequence* whenever

$$d(x_i, x_j) \longrightarrow 0 \quad \text{as} \quad i \longrightarrow \infty, j \longrightarrow \infty$$

$$[\forall\varepsilon\exists i_0 \ : \ d(x_i, x_{i+p}) < \varepsilon\forall i > i_0, p > 0.]$$

Theorem 2. Every convergent sequence in a metric space is fundamental.

Proof. If $\lim_{i\longrightarrow\infty} x_i = x$ then $\forall\varepsilon > 0\exists i_0 \ : \ d(x_i, x) < \frac{1}{2}\varepsilon\forall i > i_0$. If $i > i_0$ and $p > 0$ then

$$d(x_i, x_{i+p}) \leq d(x_i, x) + d(x_{i+p}, x) < \varepsilon.$$

There exist Cauchy sequences which are not convergent.

Definition . A metric space X is said to be *complete* if every Cauchy sequence in X is convergent.

Definition . Let be X a metric space, $F : X \longrightarrow X$ and $\alpha \in [0, 1)$. The map F is called a *contraction* on X *with coefficient* α whenever

$$d(F(x), F(y)) \leq \alpha d(x, y)\forall x, y \in X.$$

Observe that any contraction is continuous.

Theorem 3. (The Banach Fixed–point Theorem) Let be X a complete metric space, F a contraction on X with coefficient α, x_0 an arbitrary point in X and $(x_i)_0^\infty$ the related iterative sequence. Then

a) there exists a unique fixed–point $\hat{x}$ of F in X,

b) $\hat{x} = \lim_{i\longrightarrow\infty} x_i$,

c) $d(x_i, \hat{x}) \leq \alpha^i d(x_0, \hat{x})$ for $i = 1, 2, \ldots$,

d) $d(x_i, \hat{x}) \leq \frac{\alpha^i}{1-\alpha} d(x_0, x_1)$ for $i = 1, 2, \ldots$.

Remark . The statement

a) declares the existence and unicity of a fixed–point,

b) describes the way how to approach $\hat{x}$,

c) says that x_i is the closer to $\hat{x}$ the

- smaller the distance $d(x_0, \hat{x})$
- smaller α

- bigger the index i

d) is a practically applicable error–estimate.

Proof. 1. *Uniqueness.* If $u = F(u)$ and $v = F(v)$ then $d(u,v) = d(F(u),F(v)) \leq \alpha d(u,v)$ and, consequently, $(1-\alpha)d(u,v) \leq 0$ and we obtain $d(u,v) \leq 0$, so that $d(x,y) = 0$ and $u = v$ due to D1.

2. $d(x_i, x_{i+1}) \leq \alpha^i d(x_0, x_1) \forall i = 1, 2, \ldots$: $d(x_i, x_{i+1}) = d(F(x_{i-1}), F(x_i)) \leq \alpha d(x_{i-1}, x_i) \leq \ldots \leq \alpha^i d(x_0, x_1)$.

3. $d(x_i, x_{i+p}) \leq d(x_i, x_{i+1}) + d(x_{i+1}, x_{i+2}) + \ldots + d(x_{i+p-1}, x_{i+p}) \leq (\alpha^i + \alpha^{i+1} + \ldots + \alpha^{i+p-1}) d(x_0, x_1) \leq \alpha^i (1 + \alpha + \ldots) d(x_0, x_1)) = \frac{\alpha^i}{1-\alpha} d(x_0, x_1)$.

4. *Existence.* If $\varepsilon > 0$ is arbitrary then there exists $i_0 > 0$ such that $\frac{\alpha^{i_0}}{1-\alpha} d(x_0, x_1) < \varepsilon$. Then $d(x_i, x_{i+p}) < \varepsilon$ for all $i > i_0, p > 0$ by 3. Hence $(x_i)_0^\infty$ is a Cauchy sequence. As X is a complete metric space, there exists $\hat{x} = \lim_{i \longrightarrow \infty}$. We already know that then $\hat{x} = F(\hat{x})$.

5. $d(x_i, \hat{x}) \leq \alpha^i d(x_0, \hat{x})$: $d(x_i, \hat{x}) = d(F(x_{i-1}), F(\hat{x})) \leq \alpha d(x_{i-1}, \hat{x}) \leq \ldots \leq \alpha^i d(x_0, \hat{x})$.

6. $d(x_i, \hat{x}) \leq \frac{\alpha^i}{1-\alpha} d(x_0, x_1)$: If $p > 0$ then $d(x_i, \hat{x}) \leq d(x_p, x_{i+p}) + d(x_{i+p}, \hat{x}) \leq \frac{\alpha^i}{1-\alpha} d(x_0, x_1) + d(x_{i+p}, \hat{x}) \longrightarrow \frac{\alpha^i}{1-\alpha} d(x_0, x_1)$ as $p \longrightarrow \infty$.

4.1 Examples of metric spaces

1. $\mathbf{E}_1 = (\Re, d)$ and $d(x,y) = |x - y|$. Verification of the axioms D1 – D3 is very simple. For example D3:

$$d(x,y) = |x-y| = |(x-z) + (z-y)| \leq |x-z| + |z-y| = d(x,z) + d(y,x).$$

2. $(Re^n, d/infty)$ with $d_\infty(x,y) = \max_{1 \leq i \leq n} |x_i - y_i|$. We again verify D3 only:

$$\begin{aligned} d_\infty(x,y) &= \max_{1 \leq i \leq n} |x_i - y_i| = |x_j - y_j \\ &\leq |x_j - z_j| + |z_j - y_j| \leq d(x,z) + d(z,y) \end{aligned}$$

3. $(\Re^n, d_1)$ with $d_1(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$. Verify D1 – D3.

4. $\mathbf{E}^n = (\Re^n, d_2)$ with $d_2(x,y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$. Verification of D3: We have $d_2(x,y) = \sqrt{\sum (x_i - y_i)^2}$ and $d_2(x,z) + d_2(z,y) = \sqrt{\sum (x_i - z_i)} + \sqrt{\sum (z_i - y_i)^2}$. If we put $a_i = x_i - z_i$ and $b_i = z_i - y_i$ then

$$D3 \iff d_2(x,y)^2 \leq (d_2(x,z) + d_2(z,y))^2$$

$$\iff \sum(a_i+b_i)^2 \le \left(\sqrt{\sum a_i^2}+\sqrt{\sum b_i^2}\right)^2$$

$$\iff \sum a_i^2 + 2\sum a_i b_i + \sum b_i^2 \le \sum a_i^2 + 2\sqrt{\sum a_i^2}\cdot\sqrt{\sum b_i^2} + \sum b_i^2$$

$$\iff \sum_{i=1}^{n} a_i b_i \le \sqrt{\sum_{i=1}^{n} a_i^2 \cdot \sum_{i=1}^{n} b_i^2}.$$

We show later that the last inequality is valid.

5. $(C[a,b], d_\infty)$ with $d_\infty(f,g) = \max_{a\le x\le b}|f(x)-g(x)|$. Verify that the axioms D1, D2, D3 are valid.

6. $(C[a,b], d_2)$ with $d_2(f,g) = \sqrt{\int_a^b [f(x)-g(x)]^2\,dx}$. Verify D1, D2 and show that

$$D3 \iff \int_a^b f(x)\cdot g(x)dx \le \sqrt{\int_a^b f^2(x)dx \cdot \int_a^b g^2(x)dx}$$

as in 5.

7. $(L_2(a,b), d_2)$ with $L_2(a,b) = \{f \mid \text{the Lebesgue integral } \int_a^b f^2(x)dx \text{ exists and is finite}\}$ is a metric space.

Remark . The metric spaces 1–5 are complete. Metric space 6 is not complete and the metric space 7 is complete.

Example . The sequence $(f_n(x))_{n=1}^{\infty}$ such that

$$f_n(x) = \begin{cases} -1 & 0 \le x \le -\frac{1}{n} \\ nx & -\frac{1}{n} \le x \le \frac{1}{n} \\ 1 & \frac{1}{n} \le x \le 1 \end{cases}, n = 1,2,\ldots$$

is a Cauchy sequence in $(C[-1,1], d_2)$ which has no limit in $(C[-1,1], d_2)$.

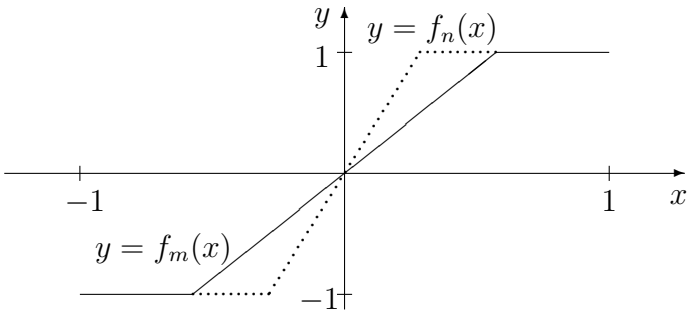


Figure 10

One can verify by computation that

$$\int_{-1}^{1} (f_m(x)-f_n(x))^2\,dx = \frac{m^2+mn+n^2}{3mn^2} < \frac{6n^2}{3mn^2} = \frac{2}{m} \longrightarrow 0$$

as $m, n \longrightarrow \infty$. It is easy to see that for

$$f(x) = \begin{cases} -1 & -1 \le x < 0 \\ 1 & 0 < x \le 1 \end{cases},$$

we have $d_2(f_n, f) \longrightarrow 0$ as $n \longrightarrow \infty$, i. e.

$$f(x) = \lim_{n \longrightarrow \infty} f_n(x) \ \text{ in } \ L_2(-1, 1).$$

As $f \notin C[-1, 1]$, the sequence $(f_n)_1^\infty$ has no limit in $(C[a, b], d_2)$.

Exercise. Decide whether $(f_n)_1^\infty$ has a limit in $(C[a, b], d_\infty)$.

Example . The sequence $\left(\frac{1}{n}\right)_{n=1}^\infty$ has a limit 0 in $\mathbf{E}_1 \Longrightarrow \left(\frac{1}{n}\right)$ is a Cauchy sequence in $((0, 1], d)$. But $\left(\frac{1}{n}\right)_1^\infty$ has no limit in $((0, 1], d)$.

4.2 Solution of $f(x) = 0$ by iteration

Let I be an arbitrary interval in $\Re$. Then the metric space (I, d) is complete if and only if I is closed.

Theorem 4. Let $\hat{x}$ be a fixed–point of a real function F. Let $\delta > 0$ and $\alpha \in [0, 1)$ satisfy

$$|F'(x)| \le \alpha \ \forall x \in (\hat{x} - \delta, \hat{x} + \delta).$$

Then F is a contraction with coefficient α in the metric space (I, d) for $I = [\hat{x} - \delta, \hat{x} + \delta]$.

Proof. a) If $x, y \in I$ are arbitrary then $F(x), F(y) \in I$ and

$$\begin{aligned} d(F(x), F(y)) &= |F(x) - F(y)| = |F'(\xi)(x - y)| \\ &\le \alpha |x - y| = \alpha d(x, y). \end{aligned}$$

The point ξ is situated between x and y by the Mean Value Theorem.

b) Let us consider the map $F : I \longrightarrow I$. Then $x \in I$ if and only if $|x - \hat{x}| \le \delta$ and we obtain $|F(x) - \hat{x}| = |F(x) - F(\hat{x})| \le \alpha |x - \hat{x}| \le \alpha\delta \le \delta$. That is why $F(x) \in I$.

Example . Find all roots of the equation $f(x) \equiv e^{-2x} + x - 3 = 0$ by iteration. Put $x = x_{i+1}$ whenever $|x_{i+1} - x_i| < 0.1 \cdot 10^{-4}$.

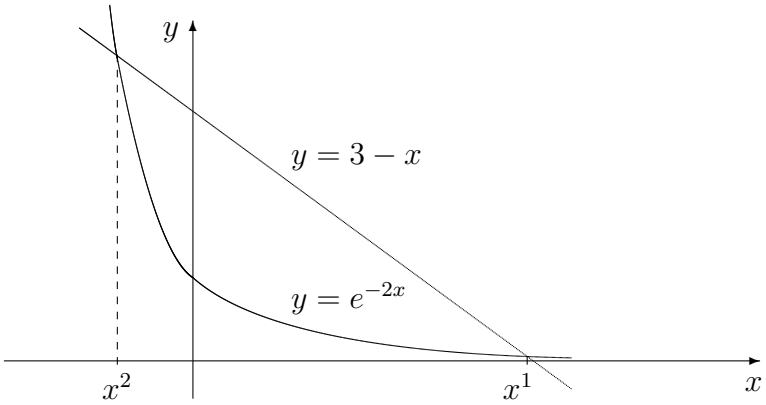


Figure 11

As $f(x) = 0 \iff e^{-2x} = 3 - x$, we can see from Fig. 11 that there exist two roots $x^1 \doteq 3$ and $x^2 \doteq -1$.

a) $f(x) = 0 \iff x = 3 - e^{-2x} \implies F_1(x) = 3 - e^{-2x}$. As

$$|F_1'(x)| = 2e^{-2x} < 1 \iff e^{-2x} < 0.5 \iff x > -0.5 \ln 0.5 \doteq 0.3466,$$

we use $F_1(x)$ for the computation of x^1, so that we put $x_0 = 3$ and compute $x_{i+1} = 3 - e^{-2x_i}$ for $i = 0, 1, \ldots$ with the results listed in the following table.

i	x_i
0	3.0
1	2.997521
2	2.9975095
3	2.9975095

We can see that $x^1 \doteq 2.9975095$.

b) $f(x) = 0 \iff -2x = \ln(3 - x) \iff x = -\frac{1}{2}\ln(3 - x) \equiv F_2(x)$. As

$$F_2'(x) = \frac{1}{6 - 2x} < \frac{1}{6}$$

for $x < 0$, F_2 is a contraction in $(-\infty, 0]$ and we use the iteration $x_0 = -1, x_{i+1} = -0.5 \ln(3 - x_i)$, $i = 0, 1, \ldots$ for the following approximation of the root x^2.

i	x_i
0	-1.
1	-0.653239
2	-0.653239
3	-0.647807
4	-0.647063
5	-0.646961
6	-0.646947
7	-0.646945
8	-0.646945

We can see that $x^2 \doteq -0.646945$.

4.3 Improvement by extrapolation (Aitken Δ^2 process)

If we solve the equation $x = F(x)$ by iteration and x_i is close to the exact solution x, we have

$$\begin{aligned} e_{i+1} &= x - x_{i+1} = F(x) - F(x_i) = F'(\xi)(x - x_i) \\ &\doteq F'(x)e_i \quad \text{and analogically} \\ e_{i+2} &\doteq F'(x)e_{i+1} \end{aligned}$$

If we eliminate $F'(x)$ (we use extrapolation), then we obtain

$$\frac{e_{i+1}}{e_i} \doteq \frac{e_{i+2}}{e_{i+1}} \implies \frac{x - x_{i+1}}{x - x_i} \doteq \frac{x - x_{i+2}}{x - x_{i+1}}.$$

If we express x, we obtain

$$x \doteq x_{i+2} - \frac{(x_{i+2} - x_{i+1})^2}{x_{i+2} - 2x_{i+1} + x_i} x_{i+2} - \frac{(\Delta x_{i+1})^2}{\Delta^2 x_i}.$$

Example . In the previous example, we obtain

$$x \doteq x_3 - \frac{(x_3 - x_2)^2}{x_3 - 2x_2 + x_1} = -0.646951 \text{ !!}$$

4.4 Steffensen's method

is based on a systematic use of the Aitken Δ^2 process in the following way: x_0 is choosen and for $i = 0, 1, \ldots$

$$x_{3i+1} = F(x_{3i}), \quad x_{3i+2} = F(x_{3i+1})$$
$$x_{3i+3} = x_{3i+2} - \frac{(\Delta x_{3i+1})^2}{\Delta^2 x_{3i}}$$

4.5 The Newton method (linearization)

Let us assume that x_i is close to the root x of the equation $f(x) = 0$. Then we have

$$0 = f(x) = f(x_i) + f'(x_i)(x - x_i) + \frac{f''(\xi)}{2}(x - x_i)^2. \tag{2}$$

If we divide (2) by $f'(x_i)$ (we assume $f'(x_i) \neq 0\,!$) and express x, we obtain

$$x = x_i - \frac{f(x_i)}{f'(x_i)} - K(x - x_i)^2 \quad \text{for} \quad K = \frac{f''(\xi)}{2f'(x_i)}. \tag{3}$$

If we neglect the rightmost term and substitute x by x_{i+1}, we obtain

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad \text{(one step of the Newton method)} \tag{4}$$

Example . Let us use the Newton method for an approximation of the root x^2 from the previous example. Put $x_0 = -1$ for comparison. Hence we compute

$$x_0 = -1, \quad x_{i+1} = x_i - \frac{3 - e^{-2x_i}(1 + 2x_i)}{1 - 2e^{-2x_i}}$$

with the results summarized in the following table.

i	x_i
0	-1.
1	-0.7540
2	-0.65896501
3	-0.64711038
4	-0.64694493
5	-0.64694490
6	-0.64694490

From this table it is apparent that in every step, the number of valid digits is multiplied by two approximately. The following error analysis justifies this fact.

4.5.1 Geometric construction of x_{i+1}

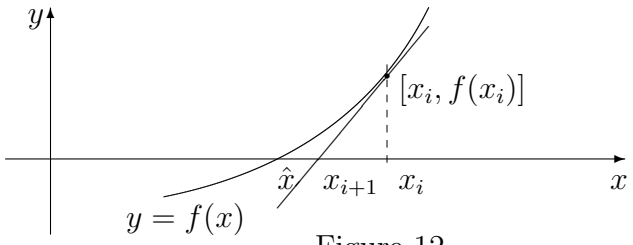


Figure 12

Instead of the non–linear equation $f(x) = 0$, we solve the "linearized equation"

$$y \doteq f(x_i) + f'(x_i)(x - x_i) = 0$$

and, if we put $x = x_{i+1}$, we obtain

$$f(x_i) + f'(x_i)(x_{i+1} - x_i) = 0 \tag{5}$$

4.5.2 Error analysis

If we subtract (5) from (2), we obtain

$$|x - x_{i+1}| = |K|\,|x - x_i|^2.$$

For comparison, bisection gives us

$$\varepsilon_{i+1} \leq \frac{1}{2}\varepsilon_i$$

and iteration gives us

$$|x - x_{i+1}| \leq \alpha\,|x - x_i|$$

for some $\alpha < 1$. Hence the Newton method is essentially more efficient than bisection or iteration under the assumption that the error $x - x_i$ is small enough. The Newton method has the following the following drawbacks:

i) it converges locally

ii) it requires higher smoothness of the function f (existence of first derivative)

iii) in each step, it requires to evaluate both $f(x_i)$ and $f'(x_i)$.

Definition . We say that an iterative method is of order r whenever the error estimates ε_i satisfy

$$\varepsilon_{i+1} \leq C \cdot (\varepsilon_i)^r$$

and C is a bounded function of r in some right neighbourhood of $r = 0$.

We have already seen that the Newton method is of order 2 and both bisection and iteration are of order 1.

4.5.3 Fourier conditions

are the following sufficient conditions for the convergence of the Newton method.

a) $f \in C^2[a, b]$ and $f(a) \cdot f(b) < 0$

b) f', f'' do not change their sign in $[a, b]$, $f'(x) \neq 0 \; \forall x \in [a, b]$

c) $x_0 = \begin{cases} a & \text{if} \quad f(a) \cdot f''(a) > 0 \\ b & \text{if} \quad f(b) \cdot f''(b) > 0 \end{cases}$

We do not verify this fact because of the following obvious geometric meaning of it.

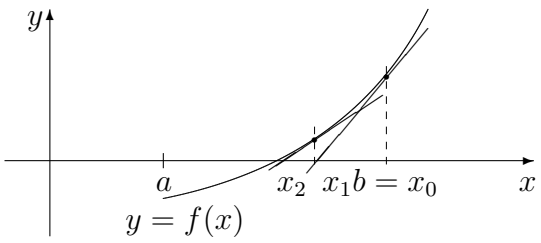


Figure 13

4.5.4 Modifications of the Newton method

a) The *secant method* consists in the approximation

$$f'(x_i) \doteq \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}.$$

After inserting into (4), we obtain one step of the secant method

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}$$

We can see that in each step of this method, the value $f(x_i)$ has to be evaluated only instead of $f(x_i)$ and $f'(x_i)$ in the Newton method. It is known that this method is of the order $r = 1.618 = (1 + \sqrt{5})/2$.

b) The *second modification* of the Newton method consists in the approximation

$$f'(x_i) \doteq \frac{f(x_i + f(x_i)) - f(x_i)}{f(x_i)},$$

so that

$$x_{i+1} = x_i - \frac{f^2(x_i)}{f(x_i + f(x_i)) - f(x_i)}.$$

The order of convergence of this method is $r = 2$.

5 Vector and matrix norms

A matrix norm (and a vector norm as a special case) is a characteristics of the sizes of the matrix entries. It is used to measure errors in matrix computations. Hence we need to understand how to compute and manipulate with them.

Definition . A *norm* on a real linear space $\mathcal{L} = (\mathcal{L}, +, \cdot)$ is a function $\|\cdot\| : \mathcal{L} \longrightarrow \Re$ satisfying the following axioms N1, N2, N3.

N1 $\|x\| \geq 0$ and $\|x\| = 0 \Longleftrightarrow x = o$

N2 $\|\alpha x\| = |\alpha| \, \|x\|$

N3 $\|x + y\| \leq \|x\| + \|y\|$ (triangular inequality)

In $\|\cdot\|$ is a norm on the linear space $(\Re^n, +, \cdot)$ then we speak about a *vector norm* and if it is a norm on the linear space $(\mathcal{M}_n, +, \cdot)$ of *square matrices of order n* then we speak about a *matrix norm*.

Example . We will work with the following concrete examples

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

of vector norms on $\Re^n$. More generally,

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

for all $p \in [1, \infty)$. As a limit for $p \longrightarrow \infty$, we obtain the norm

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|.$$

Example . Analogically, the most important norms on the linear space $(L_2(a,b), +, \cdot)$ are of the form

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}.$$

Remark. If $\|\cdot\|$ is a norm on a linear space $\mathcal{L}$ then the map $d : \mathcal{L} \longrightarrow \Re$, $d(x,y) = \|x - y\|$ is a metric on $\mathcal{L}$. Prove this simple but very important relation as an exercise.

Theorem 1. For any two norms $\|\cdot\|_a$, $\|\cdot\|_b$ on $\Re^n$ there exist positive constants c_1, c_2 such that

$$c_1 \|x\|_a \le \|x\|_b \le c_2 \|x\|_a$$

for all $x \in \Re^n$.

Example .

a) $\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2$

b) $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty$

a) $\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty$

Proof of a). We first show that $2|ab| \le a^2 + b^2 \;\; \forall a, b \in \Re$:

$$\begin{array}{rcl} 0 \le (a+b)^2 & \iff & -2ab \le a^2 + b^2 \\ 0 \le (a-b)^2 & \iff & 2ab \le a^2 + b^2 \end{array} \iff 2|ab| \le a^2 + b^2$$

Proof of a).

$$\begin{aligned} \|x\|_1^2 &= (|x_1| + \ldots + |x_n|)^2 = x_1^2 + \ldots + x_n^2 + 2 \sum_{1 \le i < j \le n} |x_i x_j| \\ &\le x_1^2 + \ldots + x_n^2 + 2 \sum_{1 \le i < j \le n} (x_i^2 + x_j^2) = n(x_1^2 + \ldots + x_n^2) = n\|x\|_2^2 \end{aligned}$$

and $\|x\|_2 \le \|x\|_1$ is obvious.

Verify b), c).

Definition . If $\|\cdot\|$ is a norm on $\Re^n$ then we define the *operator norm* on $\mathcal{M}_n$ by

$$\|A\| = \max_{x\in\Re^n, x\neq o} \frac{\|Ax\|}{\|x\|} = \max_{x\in\Re^n, \|x\|=1} \|Ax\|.$$

Theorem 2. Every operator norm on $\mathcal{M}_n$ is a matrix norm satisfying the following *consistency conditions*

N4 $\|Ax\| \leq \|A\|\,\|x\|$,

N5 $\|AB\| \leq \|A\|\,\|B\|$.

Proof. N1: $\|A\| \geq 0$ is obvious and $\|A\| = 0 \iff \|Ax\| = 0\ \forall x:\ \|x\| = 1 \iff A = O$.

N2: $\|cA\| = \max_{\|x\|=1} \|cAx\| = |c| \max_{\|x\|=1} \|Ax\| = |c|\,\|A\|$.

N3: $\|A+B\| = \max_{\|x\|=1} \|(A+B)x\| \leq \|A\| + \|B\|$.

N4: $\|Ax\| \leq \|A\|\,\|x\|$ is obvious for $x = o$. If $x \neq o$ then $\frac{\|Ax\|}{\|x\|} \leq \|A\|$ by definition.

N5: $\|A\cdot B\| = \max_{\|x\|=1} \|A\cdot Bx\| \leq \max_{\|x\|=1} \|A\|\cdot\|Bx\| = \|A\|\cdot\|B\|$.

Examples.

a) $\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$ is the operator norm of $\|x\|_\infty$. (See Scheid, 1.36)

b) $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|$ is the operator norm of $\|x\|_1$:

$$\begin{aligned}
\|x\|_1 &= |x_1| + |x_2| + \ldots + |x_n| = 1 \implies \\
\|Ax\|_1 &= \left|\sum_{j=1}^n a_{1j}x_j\right| + \left|\sum_{j=1}^n a_{2j}x_j\right| + \ldots + \left|\sum_{j=1}^n a_{nj}x_j\right| \\
&\leq |x_1| \sum_{i=1}^n |a_{i1}| + |x_2| \sum_{i=1}^n |a_{i2}| + \ldots + |x_n| \sum_{i=1}^n |a_{in}| \\
&\leq \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}| \implies \|A\|_1 \leq \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|.
\end{aligned}$$

Conversely, for $j = 1, \ldots, n$, we define x^j by

$$x_i^j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Then $\|x^j\|_1 = 1$ and $\|Ax^j\|_1 = |a_{1j}| + |a_{2j}| + \ldots + |a_{nj}| \leq \|A\|_1 \implies \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}| \leq \|A\|_1$.

Remark 1. If $\|\cdot\|$ is an operator norm and I is a unit matrix then

$$\|I\| = \max_{x \neq o} \frac{\|Ix\|}{\|x\|} = \max_{x \neq o} \frac{\|x\|}{\|x\|} = 1.$$

Remark 2. a) $\|\cdot\|_F : \mathcal{M}_n \longrightarrow \Re$, $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2}$ satisfies N1–N5 for the vector norm $\|\cdot\|_2$, but $\|I\|_F = \sqrt{n} \neq 1$, so that $\|\cdot\|_F$ is no operator norm due to Remark 1.

b) $\|\cdot\| : \mathcal{M}_n \longrightarrow \Re$, $\|A\| = \max_{1 \leq i,j \leq n} |a_{ij}|$ satisfies N1–N3, but neither N4 nor N5: If

$$A = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{then} \quad Ax = \begin{bmatrix} n \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and $\|Ax\| = n$, $\|A\| = \|x\| = 1$.

Definition . Let $\mathcal{L} = (\mathcal{L}, +, \cdot)$ be a real linear space. A map $\langle \cdot, \cdot \rangle : \mathcal{L} \times \mathcal{L} \longrightarrow \Re$ such that

S1 $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Longleftrightarrow x = o$

S2 $\langle x, y \rangle = \langle y, x \rangle$

S3 $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$

S4 $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$

is called a *scalar product* (on $\mathcal{L}$.)

Example . a) $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ is a scalar product on $\Re^n$.

b) $\langle f, g \rangle = \int_a^b f(x) g(x) dx$ is a scalar product on $L_2(a, b)$.

Theorem 1. (The Schwarz inequality) Let $\langle \cdot, \cdot \rangle$ ve a scalar product on a linear space $\mathcal{L}$. Then

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle \quad \forall x, y \in \mathcal{L}. \tag{6}$$

Proof. a) If $x = o$ then (1) is valid.

b) Let $x \neq o$. Then for any $\alpha \in \Re$,

$$\begin{aligned} 0 &\leq \langle \alpha x + y, \alpha x + y \rangle = \alpha^2 \langle x, x \rangle + 2\alpha \langle x, y \rangle + \langle y, y \rangle \\ &= \langle x, x \rangle \left[\alpha^2 + 2\alpha \frac{\langle x, y \rangle}{\langle x, x \rangle} + \frac{\langle x, y \rangle^2}{\langle x, x \rangle^2} \right] + \langle y, y \rangle - \frac{\langle x, y \rangle^2}{\langle x, x \rangle} \\ &= \langle x, x \rangle \left[\alpha + \frac{\langle x, y \rangle}{\langle x, x \rangle} \right]^2 + \langle y, y \rangle - \frac{\langle x, y \rangle^2}{\langle x, x \rangle} \end{aligned}$$

For $\alpha = -\frac{\langle x,y \rangle}{\langle x,x \rangle}$, we obtain

$$0 \leq \frac{\langle y, y \rangle \cdot \langle x, x \rangle - \langle x, y \rangle^2}{\langle x, x \rangle} \Longleftrightarrow \langle x, y \rangle^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle.$$

Remark . a) (1)$\Longleftrightarrow$ $|\langle x, y \rangle| \leq \sqrt{\langle x, y \rangle} \sqrt{\langle y, y \rangle}$.

b) $\langle x, y \rangle^2 = \langle x, x \rangle \cdot \langle y, y \rangle \Longleftrightarrow x = o$ or $y + \alpha x = o$ for some $\alpha \Longleftrightarrow x, y$ are linearly independent: If $x \neq o$ then

$$0 = \langle \alpha x + y, \alpha x + y \rangle \Longleftrightarrow \alpha x + y = o \Longleftrightarrow \alpha + \frac{\langle x, y \rangle}{\langle x, x \rangle} = o.$$

Theorem 2. If $\langle \cdot, \cdot \rangle$ is a scalar product on $\Longleftrightarrow \mathcal{L}$ then

$$\|x\| = \sqrt{\langle x, x \rangle}$$

is a norm on $\mathcal{L}$.

Proof. N1, N2 follow by S1, S2, S4. N3:

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle = \langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle \\ &\leq \|x\|^2 + 2\|x\| \, \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2. \end{aligned}$$

Remark . a) $\|x\|_2$ in $\Re^n$ and $\|f\|_2$ in $L_2(a, b)$ are constructed as in Theorem 2.

b) The inequalities from Ex. 4, 6 of metric spaces (Section 4) are valid due to (1).

6 Direct methods for systems of linear equations

We solve the problem to find the vector x satisfying

$$Ax = b \tag{7}$$

under the assumption that A is a regular matrix. (Then of course there exists a unique solution x.)

1 a) *Back substitution* is a method for the system

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \implies x_1 = \frac{b_1 - a_{12}x_2 - \ldots - a_{1n}x_n}{a_{11}} \\
a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \implies x_2 = \frac{b_2 - a_{23}x_3 - \ldots - a_{2n}x_n}{a_{22}} \qquad (8) \\
&\vdots \\
a_{n-1,n-1}x_{n-1} + a_{n,n-1}x_n &= b_{n-1} \implies x_{n-1} = \frac{b_{n-1} - a_{n,n-1}x_n}{a_{n-1,n-1}} \\
a_{nn}x_n &= b_n \implies x_n = \frac{b_n}{a_{nn}}
\end{aligned}$$

An essential characteristics of this procedure is the number of operations * and /:

$$1 + \sum_{i=1}^{n-1}(n-i) + 1 = n + (n-1) + \ldots + 1 = \frac{n(n+1)}{2} \approx \frac{n^2}{2}.$$

1 b) *Elimination* is a transformation of the system (7) to the system of the form (8) by "*addition of a multiple of one equation to another equation*".

Example .

$$\begin{aligned}
x_1 + 4x_2 + 3x_3 &= 1 \mid \cdot m_{21} = (-2) \mid \cdot m_{31} = (-1) \\
2x_1 + 5x_2 + 4x_3 &= 4 \\
x_1 - 3x_2 - 2x_3 &= 5 \\
-3x_2 - 2x_3 &= 2 \mid \cdot m_{32} = (-\frac{-7}{-3}) \\
-7x_2 - 5x_3 &= 4 \\
-\frac{1}{3}x_3 &= -\frac{2}{3}
\end{aligned}$$

The coefficients m_{ij} are called the *multipliers*. They appear during the process of elimination.

A general description of the extended matrix of coefficients in the *phase* i of elimination :

$$
\begin{array}{ccccccc|c}
a_{11} & a_{12} & & & & a_{1n} & b_1 \\
 & a_{22}^{(1)} & & & & a_{2n}^{(1)} & b_2^{(1)} \\
 & & \ddots & & & \vdots & \vdots \\
 & & & a_{ii}^{(i-1)} & & a_{in}^{(i-1)} & b_i^{(i-1)} \\
 & & & \vdots & & \vdots & \vdots \\
 & & & a_{ji}^{(i-1)} & & a_{jn}^{(i-1)} & b_j^{(i-1)} \\
 & & & \vdots & & \vdots & \vdots \\
 & & & a_{ni}^{(i-1)} & & a_{nn}^{i-1} & b_n^{(i-1)}
\end{array}
$$

The following fragment of a PASCAL – like code is a program of Gauss elimination:

for i from 1 to $n-1$ do
  for j from $i+1$ to n do
    $m_{ji} := -\frac{a_{ji}}{a_{ii}}$
    $b_j := b_j + m_{ji} \cdot b_i$
    for k from i to n do
      $a_{jk} := a_{jk} + m_{ji} a_{ik}$
    end do
  end do
end do

From this program it is apparent that the number of operations * and / in the process of elimination is the following:

$$
\begin{aligned}
\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (2 + n - i + 1) &= \sum_{i=1}^{n-1} (n-i)(3+n-i) \\
&= 3[(n-1) + (n-2) + \ldots + 1] \\
&+ (n-1)^2 + (n-2)^2 + \ldots + 1 \\
&= \frac{3}{2} n(n-1) + \frac{2n^3 - 3n^2 + n}{6} \approx \frac{n^3}{3}.
\end{aligned}
$$

Now, we can describe the complete process of *Gauss elimination:*

Step 1. Elimination. (Approximately $n^3/3$ operations)

Step 2. Back substitution. (Approximately $n^2/2$ operations)

In order to formulate the necessary and sufficient conditions for a successful application of Gauss elimination, we define the following symbols.

Definition . We put $A^{(1)} = [a_{11}]$, $A^{(2)} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$,...,$A^{(n)} = A$.

Theorem 1. The algorithm of Gauss elimination solves the system (7) successfully if and only if $|A^{(i)}| \neq 0$ for $i = 1, \ldots, n$.

In the following example, we can see that even if the assumptions of Theorem 1 are fulfilled, real computations lead to incorrect results.

Example . Solve the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 0.0001x_2 + x_3 &= 1 \\ x_2 + x_3 &= 0 \end{aligned}$$

Round off to 4 significant digits.

The result of elimination is the following system with upper triangular matrix:

$$\begin{array}{cccc|c} 1 & 1 & 1 & & 1 \\ & 0.0001 & 1 & & 1 \\ & & -9999 & & -10000 \end{array}$$

and the back substitution gives us the approximate solution

$$\tilde{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{while the exact solution is} \quad x = \begin{bmatrix} 1 \\ -\frac{10000}{9999} \\ \frac{10000}{9999} \end{bmatrix}.$$

Explanation: The roundoff error $x_3 - \tilde{x}_3 = 1/9999$ is divided by 0.0001, i. e. multiplied by 10 000 during the computation of $\tilde{x}_2$ in the back substitution. Hence the error increases essentially because of dividing the error by very small pivot 0.0001. This is the reason why modifications of the Gauss elimination are used frequently. One of them is presented in the following.

6.1 Gauss elimination with pivoting

In every phase $i = 1, 2, \ldots, n-1$, we first find $|a_{ji}^{(i-1)}| = \max_{i \leq k \leq n} |a_{ki}^{(i-1)}|$ and, if $j \neq i$, then exchange of equation i with equation j. See Fig. 15.

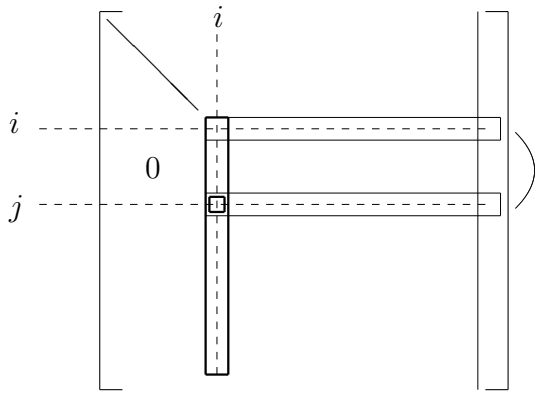


Figure 15

Remark . Gauss elimination with pivoting solves problem (7) for all regular matrices A and for all right–hand sides b.

6.2 LU–decomposition of matrices

Definition . A lower triangular matrix L with units in the main diagonal and an upper triangular matrix U create an LU–decomposition of $A \in \mathcal{M}_n$ whenever

$$A = L \cdot U.$$

Theorem 2. Let $A \in \mathcal{M}_n$ be such that $|A^{(k)}| \neq 0$ for $k = 1, \ldots, n$. Then $A = L \cdot U$ for

$$L = \begin{bmatrix} 1 & & & \\ -m_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ -m_{n1} & -m_{n2} & \ldots & 1 \end{bmatrix}, \quad U = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n-1)} \end{bmatrix}.$$

Sketch of the proof in the case $n = 3$: Verify consecutively for the matrices

$$M_1 = \begin{bmatrix} 1 & & \\ m_{21} & 1 & \\ m_{31} & & 1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & & \\ & 1 & \\ & m_{32} & 1 \end{bmatrix}$$

that

$$M_1 A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ & a_{22}^{(1)} & a_{23}^{(1)} \\ & a_{32}^{(1)} & a_{33}^{(1)} \end{bmatrix},$$

$M_2M_1A = U$ and $A = M_1^{-1}M_2^{-1}U$,

$$M_2^{-1} = \begin{bmatrix} 1 & & \\ & 1 & \\ & -m_{32} & 1 \end{bmatrix}, \quad M_1^{-1} = \begin{bmatrix} 1 & & \\ -m_{21} & 1 & \\ -m_{31} & & 1 \end{bmatrix}$$

and

$$M_2^{-1} \cdot M_1^{-1} = \begin{bmatrix} 1 & & \\ -m_{21} & 1 & \\ -m_{31} & -m_{32} & 1 \end{bmatrix}.$$

Hence it is sufficient to put $L = M_2^{-1} \cdot M_1^{-1}$.

Remark . If we solve the system $Ax = b$ and we know the LU–decomposition $A = L \cdot U$ then $Ax = b \iff L \cdot Ux = b$ and if we put $y = Ux$, we can solve the problem $Ax = b$ in the following two steps

1. $Ly = b$
2. $Ux = y$

with complexity approximately $n^2/2 + n^2/2$ which is essentially smaller than the complexity $n^3/3 + n^2/2$ of the Gauss elimination.

6.3 Matrix inversion

a) If a matrix $A \in \mathcal{M}_n$ is regular then a matrix X is inverse to A if and only if

$$AX = I \iff$$

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1n} \\ x_{21} & \dots & x_{2j} & \dots & x_{2n} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad (9)$$

If we denote by x^j and e^j the j–th column of the matrix X and of the unit matrix I then (9) is equivalent to the following n systems of equations

$$Ax^j = e^j \quad \text{for} \quad j = 1, \dots, n.$$

Example . If $A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 5 & 4 \\ 1 & -3 & -2 \end{bmatrix}$ then we find A^{-1} by solving the 3 systems from (9) by the following elimination.

1	4	3	1	0	0
2	5	4	0	1	0
1	-3	-2	0	0	1
	-3	-2	-2	1	0
	-7	-5	-1	0	1
		$-\frac{1}{3}$	$\frac{11}{3}$	$-\frac{7}{3}$	1

By performing the back substitution three times, we obtain the inverse matrix

$$A^{-1} = \begin{bmatrix} 2 & -1 & 1 \\ 8 & -5 & 2 \\ -11 & 7 & -3 \end{bmatrix}.$$

Roughly, the complexity of this algorithm is th complexity of one elimination and of n back substitutions. Hence the number of operations * and / is $n^3/3 + n \cdot n^2/2 \approx n^3$.

b) The *Jordan method* is a modification of the method a) consisting in transforming the matrix A not only to the upper triangular form but to the unit matrix. Then the n systems of equations are solved, so that in the place of the original right–hand sides, the inverse matrix appears.

Example . We continue the elimination from the previous example by transforming the upper triangular matrix on the left to the unit matrix.

1	4	3	1	0	0
	-3	-2	-2	1	0
		$-\frac{1}{3}$	$\frac{11}{3}$	$-\frac{7}{3}$	1
1	4	0	34	-21	9
	-3	0	-24	15	-6
		1	-11	7	-3
1	0	0	2	-1	1
	1	0	8	-5	2
		1	-11	7	-3

6.4 Special matrices

6.4.1 Symmetric positive definite (s. p. d.) matrices

Definition . A matrix $A \in \mathcal{M}_n$ is called s. p. d. whenever

$$A = A^\top \text{ and } x^\top A x > 0 \ \ \forall x \neq o.$$

Theorem 3. Let the matrix $A \in \mathcal{M}_n$ be symmetric. Then we have

a) A is s. p. d. $\Longleftrightarrow$ the pivots $a_{11}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}$ are positive.

b) $A_k = \begin{bmatrix} a_{k+1,k+1}^{(k)} & \ldots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots \\ a_{n,k+1}^{(k)} & \ldots & a_{n,n}^{(k)} \end{bmatrix}$ is symmetric for $k = 1, \ldots, n-1$.

Theorem 3 says that if A is s. p. d. then the Gauss elimination is successful for $Ax = b$ and, in all A_k, the elements in and above the main diagonal can be computed only.

Theorem 4. If $A \in \mathcal{M}_n$ is s. p. d. then there exists an upper triangular matrix L with positive entries in the main diagonal s. t.

$$A = L^\top \cdot L.$$

This representation is called the *Choleski decomposition* of A.

By consecutive comparisons of the elements of the matrices on the right and left sides of the following identity for the indices $(1,1), (1,2), \ldots, (1,n)$,

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & & & \\ l_{12} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{1n} & l_{2n} & \ldots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} & \ldots & l_{1n} \\ & l_{22} & \ldots & l_{2n} \\ & & \ddots & \vdots \\ & & & l_{nn} \end{bmatrix},$$

$(2,2), \ldots, (2,n), \ldots, (n,n)$, we obtain the following construction of the matrix L can be derived.

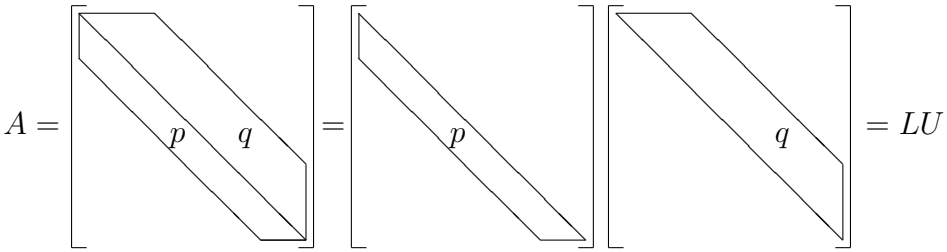
1. $l_{11} = \sqrt{a_{11}}$ and $l_{1j} = a_{1j}/l_{11}$ for $j = 2, \ldots, n$.
2. For $i = 2, \ldots, n-1 : l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ki}^2}$ and $l_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} l_{ki} l_{kj}\right) / l_{ii}$ for $j = i+1, \ldots, n$.
3. $l_{nn} = \sqrt{a_{nn} - \sum_{k=1}^{n-1} l_{kn}^2}$.

6.4.2 Band matrices

Definition . We say that $A \in \mathcal{M}_n$ is a *band matrix* if there exist $p \geq 0$, $q \geq 0$ s. t.

$$j < i - p \implies a_{ij} = 0 \text{ and } j > i + q \implies a_{ij} = 0$$

for $i = 1, \ldots, n$. Then the integer $p + 1 + q$ is called the *band breath*. The following scheme illustrates the band matrix A and the "preservation of the band structure in LU–decomposition".



6.5 Condition number of a matrix

Let us solve the problem (7) under the condition that the matrix $A \in \mathcal{M}_n$ is regular. We study the dependence of the error e_x:

$$x = \tilde{x} + e_x \tag{10}$$

on the error e_b:

$$b = \tilde{b} + e_b \tag{11}$$

under the assumptions that the entries of matrix A are exact and we solve the system (7) exactly. Thus, instead of (7), we solve

$$A\tilde{x} = \tilde{b}. \tag{12}$$

Then we obtain by (7), (10 – 12) that $A(\tilde{x} + e_x) = \tilde{b} + e_b$ and then

$$Ae_x = e_b. \tag{13}$$

If $\|\cdot\|$ denotes a vector norm as well as the related matrix operator norm then, due to N4, (13) gives us

$$\|\tilde{b}\| \leq \|A\| \, \|\tilde{x}\|$$

and (14) leads to

$$e_x = A^{-1}e_b \implies \|e_x\| \leq \|A^{-1}\| \, \|e_b\|.$$

By these two inequalities we obtain

$$\|e_x\| \, \|\tilde{b}\| \leq \|A\| \, \|A^{-1}\| \, \|e_b\| \, \|\tilde{x}\|.$$

If we divide this inequality by $\|x\|\,\|\tilde{b}\|$, we obtain

$$\frac{\|e_x\|}{\|\tilde{x}\|} \leq \|A\|\,\|A^{-1}\|\frac{\|e_b\|}{\|\tilde{b}\|}. \tag{14}$$

This inequality says that the relative error of the solution on the left–hand side is bounded by the product of the *condition number*

$$C(A) = \|A\|\,\|A^{-1}\|$$

of A and the relative error of the right–hand side.

Remark . Observe that the condition number $C(I)$ of the unit matrix I is equal to 1.

Definition . If $C(A) \gg 1$ then we say that the matrix A is *ill–conditioned.*

Remark . Analogically as we have derived the upper estimate of the influence of the right–side errors on the relative error of the solution, in Scheid 26.18 a lower estimate of this error is derived and in 26.19, the influence of the errors in coefficients of the matrix A is analysed.

Example . Let us solve the problem

$$\begin{aligned} x_1 + 0.7x_2 &= 1.69 \\ 0.7x_1 + 0.5x_2 &= 1.21 \end{aligned}$$

with $A = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 0.5 \end{bmatrix}$ and $b = \begin{bmatrix} 1.69 \\ 1.21 \end{bmatrix}$ by putting $\tilde{b} = \begin{bmatrix} 1.7 \\ 1.2 \end{bmatrix}$. Then $e_b = \begin{bmatrix} -0.01 \\ 0.01 \end{bmatrix}$, so that $\|e_b\|_\infty = 0.01$. If we solve the systems $A\tilde{x} = \tilde{b}$ and $Ax = b$ simultaneously, we obtain

1	0.7	1.7	1.69
0.7	0.5	1.2	1.21
	0.01	0.01	0.027

and $\tilde{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $x = \begin{bmatrix} -0.2 \\ 2.7 \end{bmatrix}$. Hence we have

$$e_x = \begin{bmatrix} -1.2 \\ 1.7 \end{bmatrix},$$

so that $\|e_x\|_\infty = 1.7$. As

$$A^{-1} = \begin{bmatrix} 50 & -70 \\ -70 & 100 \end{bmatrix},$$

we have $\|A\|_\infty = 1.7$, $\|A^{-1}\|_\infty = 170$. By inserting into (14), we obtain $1.7 \leq 1.7$. In this case, the inequality gives us an exact value, so that the right–hand side of (14) is not too large.

7 Eigenvalues and eigenvectors of matrices

7.1 Introduction

Every matrix $A \in \mathcal{M}_n$ represents a map $x \in \Re^n \mapsto Ax \in \Re^n$. This map is linear, i. e.

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay.$$

Definition . Let $A \in \mathcal{M}_n$. If a number λ (in general, $\lambda \in \mathcal{C}$) and a vector $x \neq o$ satisfy

$$Ax = \lambda x \tag{15}$$

then we call λ an *eigenvalue* and x an *eigenvector* of the matrix A.

Remark . Due to linearity, (15) is equivalent to

$$(A - \lambda I)x = o \tag{16}$$

saying that λ is an eigenvalue of A if and only if

$$|A - \lambda I| = 0. \tag{17}$$

This *characteristic equation* of the matrix A is a polynomial of degree n in λ. Hence there exist exactly n eigenvalues of A (real and complex, including multiplicity).

Definition . If $A \in \mathcal{M}_n$ then we call the set S of eigenvalues of A the *spectrum* of A and we we call

$$\varrho(A) = \max_{\lambda \in S} |\lambda|$$

the *spectral radius* of A.

Remark . For the operator norm of the norm $\|\cdot\|_2$, we have $\|A\|_2 = \varrho(A^\top A)$.

Example . Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

a) λ is an eigenvalue of A if and only if

$$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 1 & \\ 1 & 4-\lambda & 1 \\ & 1 & 2-\lambda \end{vmatrix} = (2-\lambda)(\lambda - 3 - \sqrt{3})(\lambda - 3 + \sqrt{3}) = 0$$

whenever $\lambda_1 = 2, \lambda_2 = 3 + \sqrt{3}, \lambda_3 = 3 - \sqrt{3}$.

b) x^i is an eigenvector of A related to λ_i if and only if

$$(A - \lambda_i I) \cdot x^i = o.$$

This system has the following form for $i = 1$:

$$\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{array}$$

and all solutions of the form $x_2 = 0, x_1 + x_3 = 0$. Hence we can choose $x^1 = [1, 0, 1]^\top$. In the same way we can choose $x^2 = [1, 1 + \sqrt{3}, 1]^\top$ and $x^3 = [1, 1 - \sqrt{3}, 1]^\top$.

It is interesting that in this case, all the eigenvalues are positive, real and the eigenvectors are mutually orthogonal.

Remark . The knowledge of an eigenvalue and of the related eigenvector is practically equivalent. Namely, if we know an eigenvalue λ of the matrix A then the related eigenvector x is any non–zero solution of (16). If we know the eigenvector x then λ satisfies

$$\begin{aligned} Ax &= \lambda x \\ x^\top Ax &= x^\top \lambda x = \lambda \cdot \|x\|_2^2 \\ \lambda &= \frac{x^\top Ax}{\|x\|_2^2}. \end{aligned}$$

This fraction expressing the eigenvalue by means of the eigenvector is called the *Rayleigh quotient.*

Theorem 1. (Gerschgorin) If $\lambda \in \mathcal{C}$ is an eigenvalue of $A \in \top\mathcal{M}_n$ then there exists an endex i such that

$$|\lambda - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

Proof. Let x_i be the component of the eigenvector x of A related to the eigenvalue λ of the largest absolute value. Then the j–th component of the vector $Ax - \lambda x = o$ is

$$(a_i i - \lambda)x_i + \sum_{j\neq i} a_{ij}x_j - 0.$$

Then

$$|\lambda - a_{ii}|\,|x_i| \leq \sum_{j\neq i} |a_{ij}|\,|x_j|$$

and we obtain

$$|\lambda - a_{ii}| \leq \sum_{j\neq i} |a_{ij}|.$$

Theorem 2. Let $A \in \mathcal{M}_n$, $x \neq o$ and $\lambda \in \mathcal{C}$ satisfy

$$Ax = \lambda x.$$

Then the following statements a) – d) are valid.

a) $A(cx) = \lambda(cx) \quad \forall c \in \Re$

b) $(A - cI)x = (\lambda - c)x \quad \forall c \in \Re$

c) $A^k x = \lambda^k x$ for $k = 2, 3, \ldots$

d) $A^{-1}x = \frac{1}{\lambda}x$ for $\lambda \neq 0$ if and only if A is regular.

Proof of d): $Ax = \lambda x \implies A^{-1}(Ax) = A^{-1}\lambda x \implies \frac{1}{\lambda}x = A^{-1}x$.

Theorem 3. If $\|\cdot\|$ is a consistent matrix norm then

$$\varrho(A) \leq \|A\| \quad \forall A \in \mathcal{M}_n.$$

Proof. Let $\varrho(A) = |\lambda|$ and $Ax = \lambda x$. Then

$$\|A\|\,\|x\| \geq \|Ax\| = \|\lambda x\| = \varrho(A)\|x\| \implies \|A\| \geq \varrho(A).$$

Theorem 4. If $A \in \mathcal{M}_n$ is s. p. d. then all eigenvalues of A are positive.
Proof. If $Ax = \lambda x$ then $\lambda = \frac{x^\top Ax}{\|x\|_2^2} > 0$.

Theorem 5. If $A \in \mathcal{M}_n$ is symmetric then

a) all eigenvalues of A are real and

b) eigenvectors related to mutually different eigenvectors are mutually orthogonal.

Proof of b). Assume that $Au = \lambda u$, $Av = \mu v$ and $\lambda \neq \mu$. Then we have consecutively

$$\begin{aligned} v^\top Au &= \lambda v^\top u \ \ |^\top \\ u^\top A^\top v &= \lambda u^\top v \\ u^\top Av &= \lambda u^\top v \\ u^\top Av &= \mu u^\top v \\ \mu\langle u, v\rangle &= \lambda\langle u, v\rangle \ \text{ and } \ \lambda \neq \mu \\ \langle u, v\rangle &= 0 \end{aligned}$$

Theorem 6. To every symmetric matrix $A \in \mathcal{M}_n$ there exists an orthogonal system of n eigenvectors.

7.2 The power method

ASSUMPTIONS:

1. The eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix A are real and

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

2. The related eigenvectors $x^1, \ldots, x^n$ are orthogonal.

BASIC IDEA: Consider $z^0 \in \Re^n$ and compute $z^1 = Az^0$,

$$z^2 = Az^1, \ldots, z^k = Az^{k-1} = A^k z^0, \ldots$$

Due to assumption 2, there exist $c_1, \ldots, c_n$ such that

$$z^0 = c_1 x^1 + \ldots + c_n x^n$$

and

$$\begin{aligned} z^k &= c_1 A^k x^1 + c_2 A^k x^2 + \ldots + c_n A^k x^n \\ &= c_1(\lambda_1)^k x^1 + c_2(\lambda_2)^k x^2 + \ldots + c_n(\lambda_n)^k x^n \\ &= (\lambda_1)^k \left[c_c x^1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k x^2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^k x^n \right] \end{aligned}$$

As $|\lambda_i| < |\lambda_1|$, $\left(\frac{\lambda_i}{\lambda_1}\right)^k \longrightarrow 0$ as $k \longrightarrow \infty$ for $i = 2, \ldots, n$. Hence $z^k \doteq (\lambda_1)^k c_1 x^1$ is an approximation of x^1 and

$$\sigma_k = \frac{(z^k)^\top A z^k}{\|z^k\|_2^2} \doteq \lambda_1 \quad \text{for } k \text{ large.}$$

Remark . a) Theoretically, if $c_1 = 0$ then $z^0 \perp x^1$ and also $z^k \perp x^1$ for $k = 2, 3, \ldots$. If $c_1 \doteq 0$ then z^k converges to the eigenvector λ_1 very slowly.

b) The following two dangers appear:

$$\begin{array}{lll} |\lambda_1| > 1 & \Longrightarrow & \|z^k\|_2 \longrightarrow \infty \text{ as } k \longrightarrow \infty \quad \text{overflow} \\ |\lambda_1| < 1 & \Longrightarrow & \|z^k\|_2 \longrightarrow 0 \text{ as } k \longrightarrow \infty \quad \text{underflow} \end{array}$$

Of course, from a) we conclude that it is reasonable to choose z^0 as close to x^1 as possible. Due to b), it is wise to "normalize" every new vector z^k: We put

$$y^k = \frac{1}{\|z^k\|_2} z^k.$$

Then we have $\|y^k\|_2 = 1$ and

$$\sigma_k = (y^k)^\top z^{k+1} = (y^k)^\top A y^k = \frac{(z^k)^\top A z^k}{\|z^k\|_2^2} \doteq \lambda_1.$$

Stop criterion: We approximate λ_1 by σ_k whenever $|\sigma_k - \sigma_{k-1}| < \varepsilon$ for some given small positive number ε.

THE ALGORITHM OF THE POWER METHOD:

Input data: $A \in \mathcal{M}_n$, $z^0 \in \Re^n$, $\varepsilon > 0$.

Step 1. $y^0 = \frac{1}{\|z^0\|} z^0$, $z^1 = A y^0$, $\sigma_1 = \langle y^0, z^1 \rangle$, $\sigma_0 = 1e6$.

Step 2. For $k = 1, 2, \ldots$ while $|\sigma_k - \sigma_{k-1}| \geq \varepsilon$ do

$$y^k = \frac{1}{\|z^k\|} z^0, \quad z^{k+1} = A y^k, \quad \sigma_{k+1} = \langle y^k, z^{k+1} \rangle$$

Step 3. $y = \frac{1}{\|z^{k+1}\|_2} z^{k+1}$.

Output data: $\sigma_k \doteq \lambda_1$, $y \doteq x^1$.

Example . $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}$, $z^0 - [2, 0, 2]^\top$, $\varepsilon = 0.0005$. The computation is presented in the following table

i	z_1^i	z_2^i	z_3^i	y_1^i	y_2^i	y_3^i	σ_i
0	2	0	2	0.7071	0	0.7071	
1	1.4142	1.4142	1.4142	0.5774	0.5774	0.5774	2.8284
2	1.7321	3.4641	1.7321	0.4082	0.8165	0.4082	4.0000
3	1.6330	4.0825	1.6330	0.3482	0.8704	0.3482	4.6667
4	1.5460	4.1964	1.5460	0.3267	0.8868	0.3267	4.7273
5	1.5403	4.2020	1.5403	0.3255	0.8877	0.3255	4.7317
6	1.5387	4.2020	1.5387	0.3252	0.8880	0.3252	4.7320

Remark . (Modifications of the power method)

a) *Approximation of the least eigenvalue.* The matrix A^{-1} has the eigenvalues $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$ and eigenvectors $x^1, \ldots, x^n$. If $\frac{1}{|\lambda_1|} \leq \ldots < \frac{1}{|\lambda_n|}$ then the power method, applied to A^{-1}, gives us an approximation of the eigenvalue $1/\lambda_n$ and of the eigenvector x^n. To avoid complicated computation of the inverse A^{-1}, the multiples $z^{i+1} = A^{-1}y^i$ are computed by solving the equivalent system of equations $Az^{i+1} = y^i$ and in this situation an application of the LU–decomposition of A is effective. Instead of $Az^{i+1} = y^i$, we solve the following two systems of equations 1, 2 with triangular matrices:

1. $Lw = y^i$,
2. $Uz^{i+1} = w$.

b) *Approximation of an aigenvalue nearest to the given value c:* We apply the power method to the matrix $(A-cI)^{-1}$. Indeed, λ is a nearest eigenvalue to $c \Longleftrightarrow \lambda - c$ is the least eigenvalue of $A - cI \Longleftrightarrow \frac{1}{\lambda - c}$ is the largest eigenvalue of $(A - cI)^{-1}$.

8 Iterative methods for linear systems

8.1 Basic notions

Our problem is to solve the system (7).

ASSUMPTIONS: $A \in \mathcal{M}_n$ is symmetric positive definite. (Then all eigenvalues of A are positive.)

AGREEMENT: $0 < \lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues and the corresponding eigenvectors $x^1, \ldots, x^n$ are mutually orthogonal, $\|x^i\| = 1$ for $i = 1, \ldots, n$.

Definition . We put $J(x) = \frac{1}{2}x^\top Ax - x^\top b$.

Motivation in the simple case $n = 1$: $A - [a]$ and A is s. p. d. if and only if $a > 0$. The problem (7) means to find $x \in \Re$ such that $ax = b$ for

$a, b \in \Re$ given. This is equivalent to the problem to find $x \in \Re$ such that the expression $\frac{1}{2}ax^2 - bx$ is minimal.

Theorem 1. (Main theorem) If $A \in \mathcal{M}_n$ is symmetric positive definite and $A\hat{x} = b$ then the following statements a), b), c) are valid.

a) $J(\hat{x}) = -\frac{1}{2}\hat{x}^\top A\hat{x}$.

b) $J(x) = \frac{1}{2}(x - \hat{x})^\top A(x - \hat{x}) + J(\hat{x})$.

c) $J(\hat{x}) < J(x) \quad \forall x \neq \hat{x}$.

Proof of a).

$$J(\hat{x}) \;=\; \frac{1}{2}\hat{x}^\top A\hat{x} - \hat{x}^\top A\hat{x} = -\frac{1}{2}\hat{x}^\top A\hat{x}.$$

Proof of b).

$$\begin{aligned} J(x) &= \frac{1}{2}x^\top Ax - x^\top A\hat{x} = \frac{1}{2}\left(x^\top Ax - x^\top A\hat{x}\right) - \frac{1}{2}x^\top A\hat{x} \\ &= \frac{1}{2}x^\top A(x - \hat{x}) - \frac{1}{2}(x - \hat{x})^\top A\hat{x} - \frac{1}{2}\hat{x}^\top A\hat{x} \\ &= \frac{1}{2}(x - \hat{x})^\top A^\top x - \frac{1}{2}(x - \hat{x})^\top A\hat{x} + J(\hat{x}) \\ &= \frac{1}{2}(x - \hat{x})^\top A(x - \hat{x}) + J(\hat{x}). \end{aligned}$$

The statement c) follows by b) and by the positivity of A immediately.

For a given constant $K \in \Re$ and for a given index i, we search after a constant c such that $s = cx^i$, x^i a unit eigenvector of A satisfies $\frac{1}{2}s^\top As = K$. As $Ax^i = \lambda_i x^i$, we obtain $\frac{c^2}{2}\lambda_i = K$. This is equivalent to $c = \sqrt{\frac{2K}{\lambda_i}}$. This value characterizes the length of the half–axis of the level–ellipse in the direction of x^i. For example, the relation between the largest and least half–axes is

$$\frac{\frac{1}{\sqrt{\lambda_1}}}{\frac{1}{\sqrt{\lambda_n}}} = \sqrt{\frac{\lambda_n}{\lambda_1}} \leq \sqrt{\|A\|\,\|A^{-1}\|} = \sqrt{C(A)}.$$

As

$$J(x) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij} - \sum_{i=1}^{n} x_i b_i,$$

we have

$$\partial J(x)/\partial x_p = \sum_{j=1}^{n} x_j a_{pj} - b_p \quad \text{for} \quad p = 1, \ldots, n$$

and we can see that the following statement is valid.

Theorem 2. $\operatorname{grad} J(x) = Ax - b$.

8.2 Jacobi iteration

We choose the starting approximation $x^{(0)} \in \Re^n$ and, for $x^{(k)}$ given, we construct the the p–th component y of $x^{(k+1)}$ $(p = 1, \ldots, n)$ as a value such that

$$J_p(y) = J(x_1^{(k)}, \ldots, x_{p-1}^{(k)}, y, x_{p+1}^{(k)}, \ldots, x_n^{(k)})$$

is minimal. This is equivalent to

$$\begin{aligned} \frac{\partial J_p}{\partial y} = 0 \quad &\Longleftrightarrow \quad y a_{pp} + \sum_{j \neq p} a_{pj} x_j^{(k)} - b_p = 0 \\ &\Longleftrightarrow \quad x_p^{(k+1)} \equiv y = \frac{1}{a_{pp}} \left(b_p - \sum_{j \neq p} a_{pj} x_j^{(k)} \right) \text{ for } p = 1, \ldots, n \end{aligned} \tag{18}$$

Example . Let us solve the following system of linear equations by the Jacobi iteration. Choose $x^{(0)} = o$.

$$\begin{aligned} 2x_1 - x_2 &= \frac{1}{3} \\ -x_1 + 2x_2 - x_3 &= 1 \\ -x_2 + 2x_3 &= -\frac{1}{3} \end{aligned}$$

The formulas (18) for one step of Jacobi iteration are of the following form

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{2}\left(\qquad x_2^{(k)} + \qquad \frac{1}{3} \right) \\ x_2^{(k+1)} &= \frac{1}{2}\left(x_1^{(k)} + \qquad x_3^{(k)} + 1 \right) \\ x_3^{(k+1)} &= \frac{1}{2}\left(\qquad x_2^{(k)} - \qquad \frac{1}{3} \right) \end{aligned}$$

and the sketch of computation can be found in the following table.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0	0	0
1	0.1667	0.5	-0.1667
2	0.4167	0.5	0.0833
⋮	⋮	⋮	⋮
25	0.6666	0.9999	0.3333
26	0.6666	0.9999	0.3333

Stop criterion: For $\varepsilon > 0$ given, we put $\tilde{x} = x^{(k+1)}$ whenever $\|x^{(k+1)-x^{(k)}}\| < \varepsilon$. Here $\|\cdot\|$ is an arbitrary consistent norm.

We can see from (18) that

$$x^{(k+1)} = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & 0 \end{bmatrix} x^{(k)} + \begin{bmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{bmatrix} \equiv Cx^{(k)} + d \equiv F(x^{(k)})$$

is the matrix form of one step (18) of the Jacobi iteration. We use this form in the following theorem.

Theorem 3. If $C \in \mathcal{M}_n$ satisfies $\|C\| < 1$ for some consistent norm then the iterative method

$$x^{(k+1)} = Cx^{(k)} + d$$

converges in $(\Re^n, d)$, $d(x, y) = \|x - y\|$, for any $x^{(0)} \in \Re^n$.

Proof. The map $F(x) = Cx + d$ is a contraction with coefficient $\alpha = \|c\|$ in the complete metric space $(\Re^n, d)$:

$$\begin{aligned} d(F(x), F(y)) &= \|F(x) - F(y)\| = \|Cx + d - (Cy + d)\| \\ &= \|C(x - y)\| \le \|C\|\, d(x, y). \end{aligned}$$

The statement follows by the Fixed Point Theorem.

Definition . A matrix $A \in \mathcal{M}_n$ is said to be *strongly diagonally dominant* whenever

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad \text{for} \quad i = 1, \dots, n.$$

Theorem 4. The Jacobi method converges for all systems of linear equations with a strongly diagonally dominant matrix.

Proof. For the system $Ax = b$,

$$x^{(k+1)} = Cx^{(k)} + d \quad \text{with} \quad C = (c_{ij}), \quad c_{ij} = \begin{cases} 0 & i = j \\ -\frac{a_{ij}}{a_{ii}} & i \neq j \end{cases} \tag{19}$$

is the Jacobi iteration. Then

$$\sum_{j=1}^{n} |c_{ij}| = \frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}| < 1$$

for $i = 1, \dots, n$, so that $\|C\|_\infty < 1$ and (19) converges due to Theorem 3.

8.3 Gauss–Seidel iteration

Let us assume that the k–th iteration $x^{(k)}$ and the components $x_1^{(k+1)}, \ldots, x_{p-1}^{(k+1)}$ are known for some index $p, 1 < p \leq n$. We find $x_p^{(k+1)}$ as that value of y which minimizes the function

$$\tilde{J}_p(y) = J(x_1^{(k+1)}, \ldots, x_{p-1}^{(k+1)}, y, x_{p+1}^{(k)}, \ldots, x_n^{(k)}).$$

Then, as in Section 8.2, we obtain the *Gauss–Seidel formula* for $y = x_p^{(k+1)}$.

$$x_p^{(k+1)} \equiv y = \frac{1}{a_{pp}} \left(- \sum_{j<p} a_{pj} x_j^{(k+1)} - \sum_{j>p} a_{pj} x_j^{(k)} + b_p \right) \tag{20}$$

Example . Let us solve the system of linear equations

$$\begin{aligned} 2x_1 - x_2 &= \frac{1}{3} \\ -x_1 + 2x_2 - x_3 &= 1 \\ -x_2 + 2x_3 &= -\frac{1}{3} \end{aligned}$$

by the Gauss–Seidel iteration. Choose $x^{(0)} = o$.

The formulas (19) for one step of Gauss–Seidel iteration are of the following form

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{2} \left(\qquad x_2^{(k)} + \qquad \frac{1}{3} \right) \\ x_2^{(k+1)} &= \frac{1}{2} \left(x_1^{(k+1)} + \qquad x_3^{(k)} + 1 \right) \\ x_3^{(k+1)} &= \frac{1}{2} \left(\qquad x_2^{(k+1)} - \qquad \frac{1}{3} \right) \end{aligned}$$

and the sketch of computation can be found in the following table.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0	0	0
1	0.1667	0.5833	0.1256
2	0.4583	0.7917	0.2294
⋮	⋮	⋮	⋮
14	0.6666	1.0000	0.3333
15	0.6666	1.0000	0.3333

The stop criterion is the same as in the Jacobi method.

Remark . Both Jacobi and Gauss–Seidel method converges for all s. p. d. matrices. In these cases, Gauss–Seidel is non–essentially more efficient.

8.4 Relaxation

is given by the relation

$$x_p^{(k+1)} = x_p^{(k)} + \frac{\omega}{a_{pp}} \left(b_p - \sum_{j<p} a_{pj} x_j^{(k+1)} - \sum_{j \geq p} a_{pj} x_j^{(k)} \right) \quad \text{for} \quad p = 1, \ldots, n \quad (21)$$

If we put $\omega = 1$ then we obtain the Gauss–Seidel formula where in the brackets the difference $x_p^{(k+1)} - x_p^{(k)}$ appears. Hence relaxation is a modification of the Gauss–Seidel method in which the difference $x_p^{(k+1)} - x_p^{(k)}$ is multiplied by a suitable *relaxation parameter* ω. It is well–known that for convergence, $\omega \in (0, 2)$ is necessary. If $\omega < 1$ [$\omega > 1$] then we speak about *subrelaxation* [*superrelaxation*]. The method issensitive w. r. to the value of the relaxation parameter.

Example . Solve the system of linear equations from the previous example by relaxatioin with $\omega = 1.2$. We have

$$\begin{aligned} x_1^{(k+1)} &= x_1^{(k)} + \frac{1.2}{2} \left(-2x_1^{(k)} + x_2^{(k)} + \frac{1}{3} \right) \\ x_2^{(k+1)} &= x_2^{(k)} + \frac{1.2}{2} \left(x_1^{(k+1)} - 2x_2^{(k)} + x_3^{(k)} + 1 \right) \\ x_3^{(k+1)} &= x_3^{(k)} + \frac{1.2}{2} \left(x_2^{(k+1)} - 2x_3^{(k)} - \frac{1}{3} \right) \end{aligned}$$

and the results of computations are summarized in the following table.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0	0	0
1	0.2	0.6	-0.2
2	0.52	0.672	0.2432
⋮	⋮	⋮	⋮
7	0.6667	1.0001	0.3334
8	0.6667	1.0001	0.3334

The stop criterion is the same as in the Jacobi method.

8.5 The steepest descent method

The subsequent methods for the minimization of

$$J(x) = \frac{1}{2} x^\top A x - x^\top b$$

(under the assumption that A is s. p. d.) are of the following form:

For given approximation $x^{(k)}$ and direction vector $v^{(k)}$, we compute $\alpha^k \in \Re$, so that

$$J(x^{(k)} + \alpha^k v^{(k)}) \leq J(x^{(k)} + \alpha v^{(k)}) \quad \forall \alpha \in \Re.$$

Then we put

$$x^{(k+1)}) = x^{(k)} + \alpha^k v^{(k)}.$$

We can derive an explicit formula for α^k in the following way (we omit the upper index (k)):

$$\begin{aligned} \tilde{J}(\alpha) &\equiv J(x + \alpha v) = \frac{1}{2}(x + \alpha v)^\top A(x + \alpha v) - (x + \alpha v)^\top b \\ &= \frac{1}{2}x^\top Ax - x^\top b + \alpha \left[\frac{1}{2}v^\top Ax + \frac{1}{2}x^\top Av - v^\top b\right] + \frac{\alpha^2}{2}v^\top Av. \end{aligned}$$

As $\frac{1}{2}v^\top Ax + \frac{1}{2}x^\top Av = v^\top Ax$ by symmetry of the matrix A, we have

$$\frac{d\tilde{J}}{d\alpha} = x^\top(Ax - b) + \alpha v^\top Av = 0$$

if and only if

$$\alpha = \frac{v^\top r}{v^\top Av} \quad \text{for} \quad r = b - Ax.$$

Hence we put

$$\alpha^k = \frac{(v^{(k)})^\top r^{(k)}}{(v^{(k)})^\top Av^{(k)}}, \quad \text{where} \quad r^{(k)} = b - Ax^{(k)} \tag{22}$$

is called a k–th *residuum*.

In the *steepest descent method,*

$$v^{(k)} = -\operatorname{grad} J(x^{(k)}) = b - Ax^{(k)} = r^{(k)}$$

is the direction of the most intensive decrease of values of J. Hence we have

$$\begin{aligned} &x^{(0)} \quad \text{is choosen} \\ &x^{(k+1)} = x^{(k)} + \alpha^k r^{(k)} \quad \text{for} \quad \alpha^k = \frac{r^{(k)\top} r^{(k)}}{r^{(k)\top} Ar^{(k)}}, \quad k = 0, 1, \ldots \end{aligned}$$

Remark . An important characteristics of the complexity of the steepest descent method is that the number of steps corresponds to the condition number $C(A)$.

Example . Approximate the solution of the system of equations

$$\begin{bmatrix} 2 & -1 & \\ -1 & 2 & -1 \\ & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 1 \\ \frac{1}{3} \end{bmatrix}$$

by the steepest descent method. Choose $\varepsilon = 0.00005$.

The computation is summarized in the following table.

α^k	k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$r_1^{(k)}$	$r_2^{(k)}$	$r_3^{(k)}$
0.5	0	0	0	0	1/3	1	-1/3
0.5	1	0.1667	0.5	-0.1667	0.5	0	0.5
0.5	2	0.4167	0.5	0.0833	0.5	0	0.5
0.5	3	0.4167	0.75	0.0833	0	0.5	0
⋮	⋮						⋮
0.5	27	0.6666	0.9998	0.3333	0	0.0001	0
0.5	28	0.6666	0.9998	0.3333	0.0001	0	0.0001

8.6 The heavy ball methods

The trajectory of a heavy ball on a surface in a gravitational field is not in the direction of steepest descent. It depends on the "old" direction, too. I. e.

$$v^{(k)} = r^{(k)} + \beta^{k-1} v^{(k-1)}, \qquad (v^{(-1)} = o)$$

with $\beta^{k-1} \geq 0$ suitably choosen. See Fig. 16.

Figure 16

8.7 The conjugate gradient method

Let us first relate a scalar product to every s. p. d. matrix.

Definition . Let A be a s. p. d. matrix. For $x, y \in \Re^n$, we put

$$\langle x, y \rangle_A = x^\top A y.$$

Theorem 1. $\langle \cdot, \cdot \rangle_A$ is a scalar product for every s. p. d. matrix A.

Proof. The properties

S1 $\langle x, x \rangle_A \geq 0$ and $\langle x, x \rangle_A = 0 \iff x = o$,

S2 $\langle x, y \rangle_A = \langle y, x \rangle_A$ and

S3 $\langle x, \alpha y + \beta z \rangle_A = \alpha \langle x, y \rangle_A + \beta \langle x, z \rangle_A$

are easy to verify.

Definition . Vectors x, y are said to be *conjugate* whenever $\langle x, y \rangle_A = 0$.

The following *conjugate gradient method* is a heavy ball method with β^{k-1} such that $v^{(k)}, v^{(k-1)}$ are conjugate:

$$x^{(0)} \text{ is choosen}$$
$$x^{(k+1)} = x^{(k)} + \alpha^k v^{(k)}, \quad \alpha^k = \frac{v^{(k)\top} r^{(k)}}{\langle v^{(k)}, v^{(k)} \rangle_A} \quad \text{for} \quad k = 1, 2, \ldots$$

Exercise. Verify that the vectors $v^{(k)}$ and $v^{(k+1)}$ are conjugate.

Example . By the conjugate gradient method solve the following system of linear equations.

$$\begin{bmatrix} 2 & -1 & \\ -1 & 2 & -1 \\ & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 1 \\ \frac{1}{3} \end{bmatrix}$$

The summary of the computations appears in the following two tables.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$r_1^{(k)}$	$r_2^{(k)}$	$r_3^{(k)}$
0	0	0	0	1/3	1	-1/3
1	0.1667	0.5	-0.1667	0.5	0	0.5
2	0.7051	0.8462	0.1410	-0.2308	0.1538	0.2308
3	0.6667	1.0000	0.3333	0	0	0

k	$v_1^{(k)}$	$v_2^{(k)}$	$v_3^{(k)}$	α^k	β^k
0	1/3	1	-1/3	0.5	0.4091
1	0.6364	0.4091	0.3636	0.8461	0.2604
2	-0.0651	0.2604	0.3254	0.5909	0
3	0	0	0	0.3530	

This result illustrates the general fact that the conjugate gradient method gives us an exact solution after n steps. But, for large n, the iteration stops much earlier usually. The necessary number of iterations corresponds to $\sqrt{C(A)}$.

9 Methods for systems of non–linear equations

Definition . For all $a \in \mathbf{E}_n$ and $\delta > 0$, we denote by $O_\delta(a)$ the open ball

$$\{x \in \mathbf{E_n} \,|\, \mathbf{d_2}(\mathbf{a}, \mathbf{x}) < \delta\}$$

with centre a and radius δ.

Definition . Let $\Omega \subseteq \mathbf{E_n}$ and $a \in \mathbf{E_n}$. We call a

a) an *inner point* of Ω whenever

$$\exists \delta > 0 : \; O_\delta(a) \subseteq \Omega.$$

b) a *boundary point* of Ω whenever for every $\delta > 0$,$O_\delta(a)$ contains both points from Ω different from a and points from $\mathbf{E_n} - \mathbf{\Omega}$ different from a. We denote by Γ_Ω the set of boundary points of Ω (the *boundary* of Ω).

Definition . Let $\Omega \subseteq \mathbf{E_n}$. We say that Ω is

a) an *open set* whenever

$$\Omega \cap \Gamma_\Omega = \emptyset$$

(All points in Ω are inner points of Ω.)

b) a *closed set* whenever

$$\Gamma_\Omega \subseteq \Omega.$$

c) *compact* whenever Ω is closed and bounded.

PROBLEM. Find a point $x = [x_1, x_2]$ satisfying

$$\begin{aligned} f_1(x) &= 0 \\ f_2(x) &= 0 \end{aligned} \tag{23}$$

for functions f_1, f_2 continuous on a domain $\Omega_0 \subseteq \mathbf{E_2}$.

In spite of systems of linear equations, the question concerning existence and/or uniqueness of solutions of this problem is not solved in general. The following example illustrates that this question is rather complicated.

Example . The problem

$$\begin{aligned} x^2 - y + a &= 0 \\ -x + y^2 + a &= 0 \end{aligned}$$

has

a) One solution for $a = 0.25$,

b) Two solutions for $a = 0$,

c) Four solutions for $a = -1$.

9.1 Iteration

We find functions F_1, F_2 : (23) is equivalent to

$$\begin{aligned} x_1 &= F_1(x_1, x_2) \\ x_2 &= F_2(x_1, x_2) \end{aligned} \tag{24}$$

on some closed domain $\Omega \subseteq \Omega_0$. If we use the vector noptation for (24), we obtain

$$x = F(x) \quad \text{for} \quad F(x) = [F_1(x), F_2(x)] \tag{25}$$

and we use the iteration

$$x^{(0)} \in \Omega \ \text{ is choosen, } \ x^{(i+1)} = F(x^{(i)}) \ \text{ for } \ i = 0, 1, \ldots$$

Stop criterion:
For $\varepsilon > 0$ given, $x \doteq x^{(i+1)}$ whenever $\|x^{(i+1)} - x^{(i)}\| < \varepsilon$.

Theorem 1. Problem (25) has exactly one solution on Ω whenever

a) $F : \Omega \longrightarrow \Omega$ and

b) $\exists \alpha < 1 : \max_{x \in \Omega} \left(\left| \frac{\partial F_i}{\partial x_1}(x) \right| + \left| \frac{\partial F_i}{\partial x_2}(x) \right| \right) \leq \alpha$.

Proof. The map F is a contraction on Ω with coefficient α: Let $x, y \in \Omega$, $x \neq y$. Then $\overline{xy} \subseteq \Omega$ and

$$x(t) = [x_1 + t(y_1 - x_1), x_2 + t(y_2 - x_2)], \quad t \in [0, 1],$$

its parametrisation. Let us put

$$\varphi_i(t) = F_i(x_1 + t(y_1 - x_1), x_2 + t(y_2 - x_2))$$

for $i = 1, 2$. Then there exists $\xi \in (0, 1)$ such that

$$\begin{aligned} |F_i(y) - F_i(x)| &= |\varphi_i(1) - \varphi_i(0)| = |\varphi'(\xi)| \\ &= \left| \frac{\partial F_i}{\partial x_1}(x(\xi))(y_1 - x_1) + \frac{\partial F_i}{\partial x_2}(x(\xi))(y_2 - x_2) \right| \\ &\leq \alpha \max(|y_1 - x_1|, |y_2 - x_2|) = \alpha \|y - x\|_\infty. \end{aligned}$$

for all points $x, y \in \Omega$. Hence $\|F(y) - F(x)\|_\infty \leq \alpha \|y - x\|_\infty$. An application of the Fixed Point Theorem to the contraction F on the complete metric space $(\Re^2, d_\infty)$ gives us Theorem 1.

Example . It is apparent from Fig. 16 that the system of equations

$$\begin{aligned} f_1(x) \equiv x_1 x_2 - x_2 - 1 = 0 \\ f_2(x) \equiv x_1^2 - x_2^2 - 1 = 0 \end{aligned}$$

has exactly two roots $x^1 \doteq [1.8, 1.1]$, $x^2 \doteq [-1.1, -0.6]$.

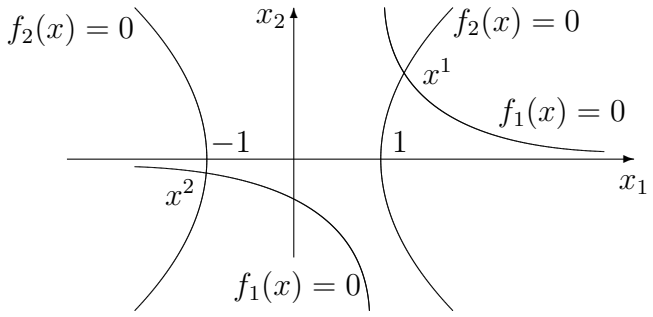


Figure 17

It is easy to see that this system of equations is equivalent to

$$\begin{aligned} x_1 = \sqrt{x_2^2 + 1} \equiv F_1(x) \\ x_2 = \sqrt{x_1 + \frac{x_2 - 1}{x_1}} \equiv F_2(x) \end{aligned} .$$

Then

$$J_F(x) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0 & \frac{x_2}{\sqrt{x_2^2+1}} \\ \frac{x_1^2 - x_2 + 1}{2\sqrt{x_1^3(x_1^2 + x_2 - 1)}} & \frac{1}{2\sqrt{x_1(x_1^2 + x_2 - 1)}} \end{bmatrix}$$

and, as

$$J_F(1.8, 1.1) = \begin{bmatrix} 0 & 0.740 \\ 0.356 & 0.204 \end{bmatrix},$$

we have $\|J_F(1.8, 1.1)\|_\infty = 0.74$. We conclude by Theorem 1 that the map F is a contractioin on some open ball with centre $[1.8, 1.1]$. The first six iterations

according to the formulas $x^{(0)} = [1.8, 1.1]$, $x^{(i+1)} = F(x^i)$ are presented in the following table.

i	$x_1^{(i)}$	$x_2^{(i)}$
0	1.8	1.1
1	1.4866	1.3622
2	1.6898	1.3154
3	1.6523	1.3698
4	1.6960	1.3697
5	1.7070	1.3864
6	1.7094	1.3905

9.2 The Newton method

Assume that a known approximation $x^{(k)}$ is near to the exact solution x of (23). If we approximate the zeros $f_1(x), f_2(x)$ by the Taylor polynomial of degree 1 around $x^{(k)}$, we obtain

$$\begin{array}{l} f_1(x^{(k)}) + \frac{\partial f_1}{\partial x_1}(x^{(k)})(x_1 - x_1^{(k)}) + \frac{\partial f_1}{\partial x_2}(x^{(k)})(x_2 - x_2^{(k)}) \doteq 0 \\ f_2(x^{(k)}) + \frac{\partial f_2}{\partial x_1}(x^{(k)})(x_1 - x_1^{(k)}) + \frac{\partial f_2}{\partial x_2}(x^{(k)})(x_2 - x_2^{(k)}) \doteq 0 \end{array} \tag{26}$$

The differences between left and right–hand sides in (26) correspond to

$$(x_1 - x_1^{(k)})^2, \quad (x_1 - x_1^{(k)})(x_2 - x_2^{(k)}), \quad (x_2 - x_2^{(k)})^2. \tag{27}$$

If we substitute x_1 by $x_1^{(k+1)}$, x_2 by $x_2^{(k+1)}$, we obtain one step of the *Newton method*

$$J(x^{(k)}) \begin{bmatrix} x_1^{(k+1)} - x_1^{(k)} \\ x_2^{(k+1)} - x_2^{(k)} \end{bmatrix} = - \begin{bmatrix} f_1(x^{(k)}) \\ f_2(x^{(k)}) \end{bmatrix} \tag{28}$$

for

$$J(x^{(k)}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} (x^{(k)}).$$

The Newton method for systems of non–linear equations has similar properties as the Newton method for one non–linear equation. It converges very quickly, but only locally, i. e. under the assumption that the initial approximation $x^{(0)}$ is close to the exact solution.

Example . Approximate the root x^1 of the system from the previous example by means of the Newton method.

As $\begin{array}{l} f_1(x) \equiv x_1 x_2 - x_2 - 1 = 0 \\ f_2(x) \equiv x_1^2 - x_2^2 - 1 = 0 \end{array}$, we have

$$J(x) = \begin{bmatrix} x_2 & x_1 - 1 \\ 2x_1 & -2x_2 \end{bmatrix}$$

and we put $x^{(0)} = [1.8, 1.1]^\top$, the system of equations

$$J(x^{(k)})(x^{(k+1)} - x^{(k)}) = - \begin{bmatrix} f_1(x^{(k)}) \\ f_2(x^{(k)}) \end{bmatrix}$$

attains the following form for $k = 0$:

1.1	0.8	0.12
3.6	-2.2	-1.03

By solving this system, we obtain $x_1^{(1)} - x_1^{(0)} = -0.10566$ and $x_2^{(1)} - x_2^{(0)} = 0.29528$. Then $x^{(1)} = [1.69434, 1.39528]^\top$. For $k = 1$, the linearized system is of the form

1.39528	0.69434	0.03120
3.38868	-2.70957	0.07603

A summary of the iterations appears in the following table.

k	$x_1^{(k)}$	$x_2^{(k)}$
0	1.8	1.1
1	1.69434	1.39528
2	1.716728	1.395226
3	1.716673	1.395337
4	1.716673	1.395337

10 Approximation of functions

10.1 Function spaces

Definition . A non–empty set $\mathcal{F}$ of (real) functions with the same domain is called a *function space* whenever

$$f, g \in \mathcal{F} \implies \alpha f + \beta g \in \mathcal{F} \quad \forall \alpha, \beta \in \Re$$

Examples.

a) $L_2(a, b)$ is a function space: $f, g \in L_2(a, b)$ is equivalent to $\int_a^b f^2(x)dx$ and $\int_a^b g^2(x)dx$ exist and are finite. Then, for any $\alpha, \beta \in \Re$, we have

$$\begin{aligned} \int_a^b (\alpha f + \beta g)^2 dx &= \alpha^2 \int_a^b f^2(x)dx + \beta^2 \int_a^b g^2(x)dx + 2\alpha\beta \int_a^b f(x)g(x)dx \\ &\leq \alpha^2 \int_a^b f^2(x)dx + \beta^2 \int_a^b g^2(x)dx \\ &+ 2|\alpha\beta| \sqrt{\int_a^b f^2(x)dx} \sqrt{\int_a^b g^2(x)dx} \end{aligned}$$

due to the Schwarz inequality. We can see that the integral is finite, so that $\alpha f + \beta g \in L_2(a,b)$.

b) $C[a,b]$ is a function space,

c) $C^k[a,b] = \{f \in C[a,b] \,|\, f, f', \ldots, f^{(k)} \in C[a,b]\}$ is a function space for $k = 1, 2, \ldots$

d) $C^\infty[a,b] = \bigcap_{k=1}^{\infty} C^{(k)}[a,b]$ is a function space.

Remark . The functions $1, x, \ldots, x^k$ are linearly independent in $C^\infty[a,b]$ for all k. That is why $C^\infty[a,b]$ has no finite basis. In this case we say that the dimension of $C^\infty[a,b]$ (and of all the function spaces from a) – d) as well) is *infinite.*

Definition . For arbitrary functions $f_1, \ldots, f_n$ from a function space $\mathcal{F}$, we put

$$\text{span}(f_1, \ldots, f_n) = \{\alpha_1 f_1 + \ldots + \alpha_n f_n \,|\, \alpha_1, \ldots, \alpha_n \in \Re\}.$$

Theorem 1. $\text{span}(f_1, \ldots, f_n)$ is the least function space containing the functions $f_1, \ldots, f_n$. Functions $f_1, \ldots, f_n$ create a basis of $\text{span}(f_1, \ldots, f_n)$ if and only if $f_1, \ldots, f_n$ are linearly independent.

Definition . We put $\mathcal{P}^k = \text{span}(1, x, \ldots, x^k)$ for $k = 0, 1, \ldots$

10.2 Polynomial interpolation

Definition . Real numbers $x_0, x_1, \ldots, x_n$ are called *nodes* whenever $x_i \neq x_j$ for all $i \neq j$.

PROBLEM. (Lagrange interpolation) For given nodes $x_0, x_1, \ldots, x_n$ and values $y_0, y_1, \ldots, y_n$, find a polynomial $P \in \mathcal{P}^n$ such that

$$P(x_i) = y_i \quad \text{for} \quad i = 0, 1, \ldots, n. \tag{29}$$

We call the polynomial P the *interpolation polynomial (interpolant).*

Theorem 2. For every nodes $x_0, x_1, \ldots, x_n$ and values $y_0, y_1, \ldots, y_n$ there exists a unique interpolation polynomial $P \in \mathcal{P}^n$.

Proof of existence of P:

I. The Lagrange form of P: We put

$$P(x) = y_0 L_0(x) + y_1 L_1(x) + \ldots + y_n L_n(x), \tag{30}$$

where $L_i \in \mathcal{P}^n$, $L_i(x_j) = \begin{cases} 1 & j-i \\ 0 & j \neq i \end{cases}$ It is easy to see that

$$L_i(x) = \frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}$$

does satisfy all the above requirements.

Proof of unicity of P: If $P, Q \in \mathcal{P}^n$ satisfy (29) then $(P-Q)(x_i) = 0$ for $i = 0, 1, \dots, n$ and $P - Q \in \mathcal{P}^n$. Then $P - Q$ is a zero polynomial, so that $P = Q$.

Example . Find the interpolant for the data

x_i	-1	1	2	3
y_i	-6	-2	-3	2

We put

$$\begin{aligned} L_0(x) &= \frac{(x-1)(x-2)(x-3)}{(-2)(-3)(-4)} = -\frac{1}{24}(x-1)(x-2)(x-3) \\ L_1(x) &= \frac{(x+1)(x-2)(x-3)}{2(-1)(-2)} = \frac{1}{4}(x+1)(x-2)(x-3) \\ L_2(x) &= -\frac{1}{3}(x+1)(x-1)(x-2) \\ L_3(x) &= \frac{1}{8}(x+1)(x-1)(x-2) \end{aligned}$$

and we have $P(x) = -6L_0(x) - 2L_1(x) - 3L_2(x) + 2L_3(x)$.

Remark . It is easy to see that the evaluation of the interpolant in Lagrange form requires $2n^2 + 2n$ operations of multiplication.

II. The *Newton form* of P: We find coefficients $a_0, a_1, \dots, a_n$ such that

$$P(x) = a_0 + a_1(x-x_0) + \dots + a_n(x-x_0)(x-x_1)\dots(x-x_{n-1}) \qquad (31)$$

satisfies conditions (29). For $i = 0, 1, \dots, n$, we consecutively obtain

$$\begin{aligned} a_0 &= y_0 \\ a_0 + a_1(x_1-x_0) &= y_1 \\ a_0 + a_1(x_2-x_0)(x_2-x_1) &= y_2 \\ &\vdots \\ a_0 + a_1(x_n-x_0) + \dots + a_n(x_n-x_0)\dots(x_n-x_{n-1}) &= y_n \end{aligned}$$

The solution of this system of equations with lower triangular matrix can be described in the following recursive way. We can see immediately that

$$a_0 = y_0, \quad a_1 = \frac{y_1 - y_0}{x_1 - x_0}$$

and it can be proved that

$$a_i = y(x_0, x_1, \ldots, x_i) \ \text{ for } \ i = 1, 2, \ldots, n,$$

where the expressions

$$\begin{aligned}
y(x_i, x_{i+1}) &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \ \text{ for } \ i = 0, \ldots, n-1, \\
y(x_i, x_{i+1}, x_{i+2}) &= \frac{y(x_{i+1}, x_{i+2}) - y(x_i, x_{i+1})}{x_{i+2} - x_i} \ \text{ for } \ i = 0, \ldots, n-2 \\
&\vdots \\
y(x_0, \ldots, x_n) &= \frac{y(x_1, \ldots, x_n) - y(x_0, \ldots, x_{n-1})}{x_n - x_0}
\end{aligned}$$

are called the *divided differences* of the first, second,..., n–th order consecutively. Hence the *Newton interpolation polynomial* P is of the form

$$P(x) = y_0 + y(x_0, x_1)(x - x_0) + \ldots + y(x_0, x_1, \ldots, x_n)(x - x_0)\ldots(x - x_{n-1}). \tag{32}$$

Example . Determine the Newton interpolation polynomial for the data

x_i	-1	1	2	3
y_i	-6	-2	-3	2

We compute the necessary divided differences mechanically by filling in the following table according to the recursive formula.

x_i	y_i	$y(x_i, x_{i+1})$	$y(x_i, x_{i+1}, x_{i+2})$	$y(x_0, x_1, x_2, x_3)$
-1	-6	2	1	-0.25
1	-2	5	0	
2	-3	5		
3	2			

Hence we have

$$P(x) = -6 + 2(x+1) + (x+1)(x-1) - 0.25(x+1)(x-1)(x-2)$$

Remark . Construction of the Newton form of interpolation polynomial is effective and, with this form, basic computations are effective, too. For example, evaluation of the Newton polynomials by the so–called generalized Horner scheme is of the same complexity as the Horner scheme for polynomials in standard for m. For example, evaluation the generalized Horner scheme for the preceding polynomial P consists in the following:

$$P(x) = -6 + (x+1)\,(2 + (x-1)\,(1 - 0.25(x-2)))$$

You can see that the number on multiplications is equal to the degree of the polynomial as in the case of the Horner scheme. Contrary to the Lagrange form, modifications of the Newton form are very simple. For example, add the node $x_4 = 0$ and value $y_4 = -4$ into the above table.

Definition . We call the nodes $x_0, x_1, \ldots, x_n$ *equidistant* whenever there exists a *step* $h > 0$ such that $x_i = x_0 + ih$ for $i = 1, \ldots, n$.

In the case of equidistant nodes, the Newton interpolation polynomial attains the following more simple form.

Definition . We call the expressions

$$\begin{aligned}
\Delta y_i &= y_{i+1} - y_i, \quad i = 0, \ldots, n-1 \\
\Delta^2 y_i &= \Delta y_{i+1} - \Delta y_i, \quad i = 0, \ldots, n-2 \\
&\vdots \\
\Delta^n y_0 &= \Delta^{n-1} y_1 - \Delta^{n-1} y_0
\end{aligned}$$

the first, second,..., n-th *differences*, respectively.

A comparison with the divided differences gives us

$$\begin{aligned}
y(x_i, x_{i+1}) &= \frac{\Delta y_i}{h} \quad , i = 0, \ldots, n-1 \\
y(x_i, x_{i+1}, x_{i+2}) &= \frac{\Delta^2 y_i}{2! h^2} \quad , i = 0, \ldots, n-2 \\
&\vdots \\
y(x_0, \ldots, x_n) &= \frac{\Delta^n y_0}{n! h^n}
\end{aligned}$$

and we obtain the following *Newton interpolation polynomial for equidistant nodes*:

$$P(x) = y_0 + \frac{\Delta y_i}{h}(x - x_0) + \ldots + \frac{\Delta^n y_0}{n! h^n}(x - x_0) \ldots (x - x_{n-1})$$

The differences can be computed by filling in an analogical triangular table as in the case of divided differences.

The error of approximation of a smooth function by its interpolation polynomial is characterized in the following theorem.

Theorem 3. Let $f \in C^{n+1}[a,b]$ and $P \in \mathcal{P}^n$ be the interpolant of f in the nodes $a \leq x_0 < x_1 < \ldots < x_n \leq b$. Then for every $x \in [a,b]$ there exists $\xi \in (a,b)$ such that

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x) \tag{33}$$

with $\omega(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$.

Proof. a) We use the Rolle Theorem saying that if $\varphi \in C^1[\alpha,\beta]$ and $\varphi(\alpha) = 0 = \varphi(\beta)$ then $\varphi'(\xi) = 0$ for some $\xi \in (\alpha,\beta)$.

b) Let $x \in [a,b]$, $x \notin \{x_0, x_1, \ldots, x_n\}$ be arbitrary. Then the function

$$g(t) = P(t) - f(t) - \frac{\omega(t)}{\omega(x)}(P(x) - f(x))$$

has $n+2$ roots $x, x_0, x_1, \ldots, x_n$ and $g \in C^{n+1}[a,b]$. Then, by repeated application of the Rolle Theorem, we consecutively obtain

$$\begin{array}{cc} g' & \text{has } n+1 \text{ roots in } (a,b) \\ g'' & \text{has } n \text{ roots in } (a,b) \\ \vdots & \\ g^{(n+1)} & \text{has 1 root } \xi \text{ in } (a,b). \end{array}$$

As $P^{(n+1)} = 0$ and $\omega^{(n+1)} = (n+1)!$, we have

$$0 = g^{(n+1)}(\xi) = -f^{(n+1)}(\xi) - \frac{(n+1)!}{\omega(x)}(P(x) - f(x)).$$

But then

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x). \tag{34}$$

Remark . (The *Runge phenomenon*) If the nodes $x_0, x_1, \ldots, x_n$ are equidistant and n is large then the values of $\omega(x)$ for x near to the point a or b are essentially greater than the values of $\omega(x)$ for x near to the centre of $[a,b]$. This determines similar relation between the values of the error $f(x) - P(x)$.

There exist various ways how to overcome the Runge phenomenon. Now, we present a suitable choice of non–equidistant nodes $x_0, x_1, \ldots, x_n$, so that maxima of $\omega(x)$ in the intervals (x_i, x_{i+1}) are the same.

Definition . The polynomials $T_0(x) = 1$, $T_1(x) = x$, and

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k = 2, 3, \ldots$$

are called *Chebyshev polynomials.*

Obviously, we have

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ T_5(x) &= 16x^5 - 20x^3 + 5x \\ T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1 \end{aligned}$$

and so on. These examples illustrate the validity of the following theorem.

Theorem 4. The following statements a) – e) are valid.

a) The degree of $T_n(x)$ is equal to n and the coefficient at x^n in $T_n(x)$ is 2^{n-1} for $n = 1, 2, \ldots$

b) $T_{2m}(x) = T_{2m}(-x)$ and $T_{2m+1}(x) = -T_{2m+1}(-x)$ for $m = 0, 1, \ldots$ and for all $x \in \Re$.

c) $T_n(x) = \cos(n \arccos x)$ for all $x \in [-1, 1]$ and for $n = 0, 1, \ldots$

d) For every $n > 0$, $T_n(x)$ has n distinct roots

$$x_k = \cos\left(\frac{(2k+1)\pi}{2n}\right) \quad \text{for} \quad k = 0, 1, \ldots, n-1$$

The points x_k are called *Chebyshev nodes.*

e) $\max_{-1 \le x \le 1} |T_n(x)| \le 1$ for $n = 0, 1, \ldots$

Proof. The statements a), b) follow directly from definition.

Proof of c). If $k \ge 2$ then

$$\cos(k\Theta) = \cos(2\Theta)\cos((k-2)\Theta) - \sin(2\Theta)\sin((k-2)\Theta).$$

Then, by means of the identities

$$\cos(2\Theta) = \cos^2\Theta - \sin^2\Theta = 2\cos^2\Theta - 1, \quad \sin(2\Theta) = 2\cos\Theta\sin\Theta,$$

we obtain

$$\begin{aligned}\cos(k\Theta) &= 2\cos\Theta\,[\cos\Theta\cos((k-2)\Theta) - \sin\Theta\sin((k-2)\Theta)] \\ &\quad - \cos((k-2)\Theta) \\ &= 2\cos\Theta\cos((k-1)\Theta) - \cos((k-2)\Theta)\end{aligned}$$

for all $x \in [-1,1]$. If we put

$$\Theta = \arccos x \Longleftrightarrow \cos\Theta = x \quad \text{for} \quad x \in [-1,1]$$

then we obtain

$$\cos(k\arccos x) = 2x\cos((k-1)\arccos x) - \cos((k-2)\arccos x) \qquad (35)$$

for all $x \in [-1,1]$. Now we have

$$T_0(x) = 1 = \cos(0\arccos x), \quad T_1(x) = x = \cos(1\arccos x)$$

and, by induction, if $T_k(x) = \cos(k\arccos x)$ for $k = 2,\ldots,n-1$ then

$$\begin{aligned}T_n(x) &= 2xT_{n-1}(x) - T_{n-2}(x) \\ &= 2x\cos((n-1)\arccos x) - \cos((n-2)\arccos x) \\ &= \cos(n\arccos x)\end{aligned}$$

due to (35).

Proof of d). $\cos(n\arccos x) = 0$ for $x \in [0,\pi]$

$$\begin{aligned}&\Longleftrightarrow \quad n\arccos x = \frac{(2k+1)\pi}{2} \quad \text{for} \quad \frac{(2k+1)\pi}{2} \in [0,n\pi] \\ &\Longleftrightarrow \quad \arccos x = \frac{2k+1}{2n}\pi \quad \text{for} \quad \frac{2k+1}{2n}\pi \in [0,\pi] \\ &\Longleftrightarrow \quad x = \cos\left(\frac{2k+1}{2n}\pi\right) \quad \text{for} \quad \frac{2k+1}{2n} \in [0,1].\end{aligned}$$

The statement e) follows by c) immediately.

For $n \geq 1$, the Chebyshev polynomial $T_n(x)$ is of degree n, its roots are just the Chebyshev nodes $x_0, x_1, \ldots, x_n$ and the coefficient at x^n is 2^{n-1}. The polynomial $\omega(x) = \omega_C(x)$ related to the Chebyshev nodes also is of degree n, its roots are just the Chebyshev nodes and the coefficient at x^n is 1. Hence we have $2^{1-n}T_n(x) = \omega_C(x)$ and, due to Theorem 3 e), $\max_{-1\leq x\leq 1}\omega_C(x) = 2^{1-n}$. The following theorem says that this maximum is the least one among all functions $\omega(x)$ related to all choices of $n+1$ nodes in $[-1,1]$.

Theorem 5. For a fixed $n > 0$, among all polynomials $\omega(x)$ related to all possible choices of the nodes $x_0, x_1, \ldots, x_n \in [-1, 1]$, the polynomial $\omega_C(x) = 2^{1-n} T_n(x)$ is a unique polynomial satisfying

$$2^{1-n} = \max_{-1 \le x \le 1} |\omega_C(x)| \le \max_{-1 \le x \le 1} |\omega(x)|.$$

10.3 Cubic splines

If the interval $[a, b]$ is long and the number of nodes $a = x_0 < x_1 < \ldots < x_n = b$ is large, instead of one interpolation polynomial of a high degree, we compose the interpolant on $[a, b]$ by (generally mutually different) polynomials of low degree on $[x_{i-1}, x_i]$ for $i = 1, \ldots, n$.

If the low degree is equal to one, we speak about *linear splines.* At this moment, we only tough this important class of interpolants by illustrating the solution of the problem of Lagrange interpolation by the linear spline in Fig. 18.

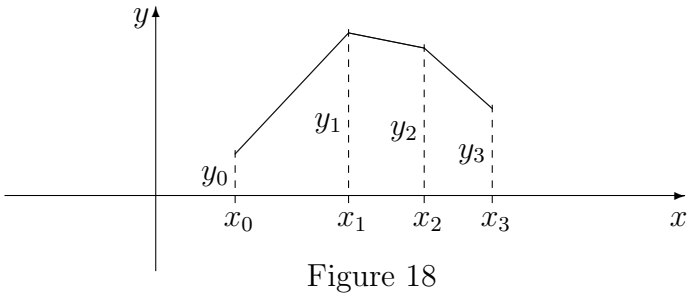


Figure 18

We devote more attention to the so–called cubic splines.

Definition . Let $a = x_0 < x_1 < \ldots < x_n = b$ for some $n > 1$. A *cubic spline* with nodes $x_0, \ldots, x_n$ is every function s on $[a, b]$ such that

a) $s(x) = s_i(x)$ for $x_{i-1} \le x \le x_i$ and $s_i \in \mathcal{P}^3$ for $i = 1, \ldots, n$.

b) $s \in C^2[a, b]$.

Condition b) is equivalent to

$$\begin{aligned} s_i(x) &= s_{i+1}(x) \\ s_i'(x) &= s_{i+1}'(x) \\ s_i''(x) &= s_{i+1}''(x) \end{aligned}$$

for $i = 1, \ldots, n-1$.

PROBLEM. For given nodes $a = x_0 < x_1 < \ldots < x_n = b$ $(n > 1)$ and values $y_0, y_1, \ldots, y_n$, find a cubic spline $s(x)$ with nodes $x_0, x_1, \ldots, x_n$ such

that

$$s(x_i) = y_i \quad \text{for} \quad i = 0, 1, \ldots, n. \tag{36}$$

If $y_i = f(x_i)$ then we call s an *interpolation cubic spline* of f in $x_0, x_1, \ldots, x_n$.

Theorem 6. Let us put $b_i = x_i - x_{i-1}$ for $i = 1, \ldots, n$. If s is a cubic spline in $x_0, x_1, \ldots, x_n$ satisfying (36) then, for $i = 1, \ldots, n$, we have

$$\begin{aligned} s_i(x) &= C_{i-1}\frac{(x_i - x)^3}{6h_i} + C_i\frac{(x - x_{i-1})^3}{6h_i} \\ &+ \left(y_{i-1} - \frac{C_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(y_i - \frac{C_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i} \end{aligned} \tag{37}$$

and the parameters $C_0, C_1, \ldots, C_n$ satisfy the equations

$$\frac{h_k}{6}C_{k-1} + \frac{h_k + h_{k+1}}{3}C_k + \frac{h_{k+1}}{6}C_{k+1} = \frac{y_{k+1} - y_k}{h_{k+1}} - \frac{y_k - y_{k-1}}{h_k}$$

for $k = 1, \ldots, n-1$.

Proof. If $s(x)$ is a cubic spline with nodes $x_0, x_1, \ldots, x_n$ then $s''(x)$ is a linear spline with nodes $x_0, x_1, \ldots, x_n$ Hence there exist values $C_0, C_1, \ldots, C_n$ such that

$$s''(x_i) = C_i \quad \text{for} \quad i = 0, 1, \ldots, n$$

and

$$s_i''(x) = C_{i-1}\frac{x_i - x}{h_i} + C_i\frac{x - x_{i-1}}{h_i} \quad \text{for} \quad x \in [x_{i-1}, x_i]. \tag{38}$$

Here $i = 0, 1, \ldots, n$. Due to (38), we have $s'' \in C[a, b]$. Integrating (38) two–times, we obtain

$$s_i(x) = C_{i-1}\frac{(x_i - x)^3}{6h_i} + C_i\frac{(x - x_{i-1})^3}{6h_i} + \alpha_i(x_i - x) + \beta_i(x - x_{i-1}),$$

where $\alpha_i, \beta_i \in \Re$ are arbitrary integration constants. If we require $s_i(x_{i-1}) = y_{i-1}$ and $s_i(x_i) = y_i$, we obtain (37). Thus, we have proved $s \in C[a, b]$ and (36). Differentiating (37), we obtain

$$s_i'(x) = -C_{i-1}\frac{(x_i - x)^2}{2h_i} + C_i\frac{(x - x_{i-1})^2}{2h_i} + \frac{y_i - y_{i-1}}{h_i} - \frac{C_i - C_{i-1}}{6}h_i.$$

As

$$\begin{aligned} s_i'(x_{i-1}) &= -C_{i-1}\frac{h_i}{3} - C_i\frac{h_i}{6} + \frac{y_i - y_{i-1}}{h_i}, \\ s_i'(x_i) &= C_i\frac{h_i}{3} + C_{i-1}\frac{h_i}{6} + \frac{y_i - y_{i-1}}{h_i}, \end{aligned}$$

we obtain that $s_{i+1}(x_i) = s_i'(x_i)$ is equivalent to the equation

$$C_{i-1}\frac{h_i}{6} + C_i\frac{h_i + h_{i+1}}{3} + C_{i+1}\frac{h_{i+1}}{6} = \frac{y_i - y_{i-1}}{h_i} - \frac{y_{i+1} - y_i}{h_{i+1}} \tag{39}$$

for $i = 1, \ldots, n-1$. If (39) is valid then $s' \in C[a, b]$.

Remark . 1. As (39) is a system of $n-1$ equations in $n+1$ unknowns $C_0, C_1, \ldots, C_n$, the spline s is not determined uniquely. Most often, the equations

$$C_0 = 0 = C_n \tag{40}$$

are added. Cubic splines satisfying (40), i. e.

$$s''(a) = 0 = s''(b),$$

are called natural.

2. With $C_0 = 0 = C_n$, (39) is a system of linear equations with strongly diagonally dominated three–diagonal matrix.

The following statement says that the $L_2 - norm$ of the second derivative of the interpolation cubic spline (the curvature of the cubic spline) is the least among all interpolants with the same values in the same nodes. This excludes the Runge phenomenon and is the main reason why the cubic splines are so popular.

Theorem 7. Let $n > 1$, $a = x_0 < x_1 < \ldots < x_n = b$ and $f \in C^2[a, b]$. If s is a natural cubic spline with nodes $x_0, x_1, \ldots, x_n$ and $s(x_i) = f(x_i)$ for $i = 0, 1, \ldots, n$ then

$$\int_a^b s''^2(x)\,dx \le \int_a^b f''^2(x)\,dx.$$

Proof. Integrating by parts, we obtain

$$\int_a^b (f'' - s'')s''dx = [(f' - s') \cdot s'']_a^b - \int_a^b (f' - s')s'''dx.$$

As $s''(a) = 0 = s''(b)$ and s''' is a constant $K_i = s_i'''$ on the interval (x_{i-1}, x_i) for $i = 1, \ldots, n$, we have

$$\int_a^b (f' - s')s'''dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} (f' - s')K_i dx = \sum_{i=1}^n K_i[(f - s)(x)]_{x_{i-1}}^{x_i} = 0,$$

we obtain

$$\int_a^b (f'' - s'')s''dx = 0.$$

Then

$$\begin{aligned}\int_a^b f''^2 dx &= \int_a^b (f''-s'')^2 dx + 2\int_a^b (f''-s'')s'' dx + \int_a^b s''^2 dx \\ &= \int_a^b (f''-s'')^2 dx + \int_a^b s''^2 dx \\ &\geq \int_a^b s''^2 dx.\end{aligned}$$

10.4 Hermite interpolation

PROBLEM (Hermite interpolation – osculation) For given nodes $a = x_0 < x_1 < \ldots < x_n = b$ and for a given function $f \in C^1[a,b]$, find a polynomial $H \in \mathcal{P}^{2n+1}$ such that

$$H(x_i) = f(x_i) \text{ and } H'(x_i) = f'(x_i) \text{ for } i = 0, 1, \ldots, n. \tag{41}$$

Remark . There exists a unique *Hermite interpolation polynomial* $H \in \mathcal{P}^{2n+1}$ satisfying (41).

Construction I. We search coefficients $a_0, \ldots, a_{2n+1} \in \Re$ such that

$$H(x) = a_0 + a_1 x + \ldots + a_{2n+1}x^{2n+1}, \quad H'(x) = a_1 + 2a_2 x + \ldots + a_{2n+1}x^{2n} \tag{42}$$

satisfy the conditions (41). We find the values of $a_0, a_1, \ldots, a_{2n+1}$ by solving the $2n+2$ equations which we obtain by putting the forms (42) into (41).

Example . (Standard form of H) Find the Hermite interpolation polynomial of the function f in the nodes from the following table.

i	x_i	$f(x_i)$	$f'(x_i)$
0	-1	2	1
1	1	0	1

As $n = 1$, we have $H(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \in \mathcal{P}^3$ and $H'(x) = a_1 + 2a_2 x + 3a_3 x^2$. If we insert the values of $x_0 = -1$ and $x_1 = 1$ into (41), we obtain the following system of four linear equations for the unknown coefficients $a_0, \ldots, a_3$.

$$\begin{aligned} a_0 - a_1 + a_2 - a_3 &= 2 \\ a_1 - 2a_2 + 3a_3 &= 1 \\ a_0 + a_1 + a_2 + a_3 &= 0 \\ a_1 + 2a_2 + 3a_3 &= 1 \end{aligned}$$

It is easy to see that this system has the solution $a_0 = 1, a_1 = -2, a_2 = 0, a_3 = 1$, so that $H(x) = 1 - 2x + x^3$.

Construction II. (Generalized Newton form) As we have two values for every node, we insert every node two times into the table. Then we have to evaluate the divided difference $f(x_i, x_i)$ which has no sense. If we evaluate this divided difference as a limit of $f(x_i, x)$ as x approaches x_i, we obtain

$$f(x_i, x_i) = \lim_{x \longrightarrow x_i} f(x_i, x) = \lim_{x \longrightarrow x_i} \frac{f(x) - f(x_i)}{x - x_i} = f'(x_i).$$

Hence we use the known values $f'(x_i)$ for $f(x_i, x_i)$ and the Hermite interpolation polynomial is of the form

$$\begin{aligned} H(x) &= f(x_0) + f(x_0, x_0)(x - x_0) + f(x_0, x_0, x_1)(x - x_0)^2 \\ &\quad + f(x_0, x_0, \ldots, x_n, x_n)(x - x_0)^2 \ldots (x - x_{n-1})^2 (x - x_n) \end{aligned}$$

Example . Express the Hermite interpolation polynomial in the generalized Newton form for the data from the following table.

i	x_i	$f(x_i)$	$f'(x_i)$
0	-1	2	1
1	1	0	1
2	2	1	3

The generalized Newton scheme is apparent from the following table

x_i	$f(x_i)$	...				
-1	2	1	-1	1	$-\frac{4}{9}$	$\frac{11}{27}$
-1	2	-1	1	$-\frac{1}{3}$	$\frac{7}{9}$	
1	0	1	0	2		
1	0	1	2			
2	1	3				
2	1					

and the resulting Hermite interpolation polynomial is

$$\begin{aligned} H(x) &= 2 + x + 1 - (x+1)^2 + (x+1)^2(x-1) \\ &- \frac{4}{9}(x+1)^2(x-1)^2 + \frac{11}{27}(x+1)^2(x-1)^2(x-2) \end{aligned}$$

The following type of interpolant is widely used because of simplicity of construction and of some further useful properties.

Definition . A *Hermite cubic spline* with nodes $a = x_0 < x_1 < \ldots < x_n = b$ is any function such that

a) $s = s_i \in \mathcal{P}^3$ on $[x_{i-1}, x_i]$ for $i = 1, \ldots, n$

b) $s \in C^1[a, b]$.

If $f \in C^1[a, b]$ and s satisfies

$$s(x_i) = f(x_i), \quad s'(x_i) = f'(x_i) \quad \text{for} \quad i = 0, 1, \ldots, n$$

then we say that s is a *Hermite interpolation cubic spline* of f in the nodes $x_0, x_1, \ldots, x_n$.

Remark . For $i = 1, \ldots, n$, the cubic polynomial s_i is determined by

$$\begin{array}{ll} s_i(x_{i-1}) = f(x_{i-1}) & s_i'(x_{i-1}) = f'(x_{i-1}) \\ s_i(x_i) = f(x_i) & s_i'(x_i) = f'(x_i) \end{array}$$

uniquely. We can use the constructions of the Hermite interpolation polynomial.

Example . Find the Hermite cubic spline for the data from the following table.

i	x_i	$f(x_i)$	$f'(x_i)$
0	-1	2	1
1	1	0	1
2	2	1	3

It is easy to compute by means of Construction I for example that

$$s(x) = \begin{cases} s_1(x) = 2 + (x+1) - (x+1)^2 + (x+1)^2(x-1) & \text{for} \quad x \in [-1, 1] \\ s_2(x) = x - 1 + (x-1)^2(x-2) & \text{for} \quad x \in [1, 2] \end{cases}$$

10.5 The least squares method (LSM)

This is another widely used general idea due to Gauss.

Definition . Let $\mathcal{L}$ be a normed vector space, $\mathcal{F}$ be a finite–dimensional subspace of $\mathcal{L}$ and f be an arbitrary element from $\mathcal{L}$. We say that $\tilde{f} \in \mathcal{F}$ is the *best approximation* of f (from $\mathcal{F}$) whenever

$$\|f - \tilde{f}\| \leq \|f - g\| \quad \text{for all} \quad g \in \mathcal{F}. \tag{43}$$

In this section we assume that there exists a scalar product $\langle \cdot, \cdot \rangle$ on $\mathcal{L}$ such that

$$\|x\| = \sqrt{\langle x, x \rangle} \quad \forall x \in \mathcal{L}.$$

Let $\varphi_1, \ldots, \varphi_n$ be a basis of $\mathcal{F}$. Then (43) is equivalent to

$$\begin{aligned} \|f - \tilde{f}\|^2 &\leq \|f - g\|^2 \quad \forall g \in \mathcal{F}, \\ \langle f - \tilde{f}, f - \tilde{f} \rangle &\leq \langle f - g, f - g \rangle \quad \forall g \in \mathcal{F} \end{aligned}$$

consecutively. Hence there exist coefficients $\tilde{x}_1, \ldots, \tilde{x}_n$ such that $\tilde{f} = \tilde{x}_1\varphi_1 + \ldots + \tilde{x}_n\varphi_n$. If we insert this form of $\tilde{f}$ into the last inequality then we obtain the problem to find $\tilde{x}_1, \ldots, \tilde{x}_n \in \Re$ such that the value of

$$F(\tilde{x}_1, \ldots, \tilde{x}_n) \equiv \langle f - \tilde{x}_1\varphi_1 - \ldots - \tilde{x}_n\varphi_n, f - \tilde{x}_1\varphi_1 - \ldots - \tilde{x}_n\varphi_n \rangle$$

is minimal. Necessary conditions for a minimum of F are

$$\frac{\partial F}{\partial x_i} = 0 \quad \text{for} \quad i = 1, \ldots, n.$$

As

$$F(x_1, \ldots . x_n) = \langle f, f \rangle - 2 \sum_{k=1}^{n} x_k \langle f, \varphi_k \rangle + \sum_{k=1}^{n} \sum_{m=1}^{n} x_k x_m \langle \varphi_k, \varphi_m \rangle,$$

we have

$$\frac{1}{2} \frac{\partial F}{\partial x_i} = -\langle f, \varphi_i \rangle + \sum_{j=1}^{n} x_j \langle \varphi_j, \varphi_i \rangle = 0 \quad \text{for} \quad i = 1, \ldots, n.$$

These equations have the following matrix form

$$\begin{bmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_2, \varphi_1 \rangle & \ldots & \langle \varphi_n, \varphi_1 \rangle \\ \langle \varphi_1, \varphi_2 \rangle & \langle \varphi_2, \varphi_2 \rangle & \ldots & \langle \varphi_n, \varphi_2 \rangle \\ \vdots & \vdots & & \vdots \\ \langle \varphi_1, \varphi_n \rangle & \langle \varphi_2, \varphi_n \rangle & \ldots & \langle \varphi_n, \varphi_n \rangle \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \langle f, \varphi_1 \rangle \\ \langle f, \varphi_2 \rangle \\ \vdots \\ \langle f, \varphi_n \rangle \end{bmatrix}. \tag{44}$$

They are called *normal equations*.

Theorem 8. The normal equations have exactly one solution $[\tilde{x}_1, \ldots, \tilde{x}_n]^\top$ and

$$\tilde{f} = \tilde{x}_1\varphi_1 + \ldots + \tilde{x}_n\varphi_n$$

is the best approximation of f in the space $\mathcal{F}$.

Proof. a) The matrix M of the system (44) is symmetric positive definite: M is symmetric obviously. For $x \in \Re^n$, $x \neq o$ denote

$$f_x = x_1\varphi_1 + \ldots + x_n\varphi_n.$$

Then $f_x \neq o$ because $x \neq o$ and $\varphi_1, \ldots, \varphi_n$ are linearly independent. Then $Mx = [\langle f_x, \varphi_1\rangle, \ldots, \langle f_x, \varphi_n\rangle]^\top$ and $x^\top Mx = \langle f_x, f_x\rangle = \|f_x\|^2 > 0$.

Due to a), the system (44) has exactly one solution $\tilde{x}$.

b) $\langle \tilde{f} - f, g\rangle = 0 \ \ \forall g \in \mathcal{F}$: It is easy to see that (44) is equivalent to

$$\langle \tilde{f} - f, \varphi_i\rangle = 0 \ \text{ for } \ i = 1, \ldots, n. \tag{45}$$

Now (45) and the linearity of the scalar product give us b).

c) $\tilde{f}$ is the best approximation of f: By means of b), we have

$$\begin{aligned} \|f - g\|^2 &= \|(f - \tilde{f}) + (\tilde{f} - g)\|^2 \\ &= \langle (f - \tilde{f}) + (\tilde{f} - g), (f - \tilde{f}) + (\tilde{f} - g)\rangle \\ &= \langle f - \tilde{f}, f - \tilde{f}\rangle + 2\langle f - \tilde{f}, \tilde{f} - g\rangle + \langle \tilde{f} - g, \tilde{f} - g\rangle \\ &= \|f - \tilde{f}\|^2 + \|\tilde{f} - g\|^2 \geq \|f - \tilde{f}\|^2 \end{aligned}$$

Definition . If the normed space $\mathcal{L}$ is a subspace of $\left\{ \begin{array}{c} \Re^n \\ L_2(a,b) \end{array} \right\}$ then we speak about a $\left\{ \begin{array}{c} \text{discrete} \\ \text{continuous} \end{array} \right\}$ least squares method.

Now, we present typical applications of the LSM. In Example 1, we use the LSM for an apprixomate solution of an overdetermined system of linear equations, Example 2 is a continuous LSM used for an approximation of a function and in Example 3, which is an approximation of function by the discrete LSM, we show that certain problems have to be reformulated before an application of the LSM.

Example 1. In order to determine the heights x_A, x_B, x_C of the points A, B, C, the following six heights or differences of heights have been measured:

$$\begin{array}{rrrcl} x_A & & & = & 1 \\ -x_A & & +x_C & = & 1 \\ & x_B & & = & 2 \\ & -x_B & +x_C & = & 2 \\ & & x_C & = & 3 \\ -x_A & +x_B & & = & 1 \end{array}$$

This system of 6 equation in 3 unknowns is equivalent to the problem

$$x_A \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} + x_B \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 1 \end{bmatrix} + x_C \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ 1 \end{bmatrix}.$$

Hence, if we denote by $\varphi_A, \varphi_B, \varphi_C$ the vector multiplied by x_A, x_B, x_C, consecutively, and by f the vector on the right–hand side of this system, we obviously cannot expect that we find an exact solution. In the solution by the least squares method, we search a vector $\tilde{f} = x_A\varphi_A + x_B\varphi_B + x_C\varphi_C$, so that $\|f - \tilde{f}\|_2$ is minimal. The normal equations are of the form

$$\begin{array}{rrr|r} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & 1 \\ -1 & -1 & 3 & 6 \end{array}$$

and its solution is $x_A = 1.25, x_B = 1.75, x_C = 3$. These are the approximations of the heights of the points A, B, C.

Example 2. Find the best approximation of $\sin x \in L_2(0, \pi/2)$ in the space $\mathrm{span}(1, x)$.

We put $\varphi_1 = 1$, $\varphi_2 = x$ and $f(x) = \sin x$. Then we

$$\langle \varphi_1, \varphi_1 \rangle = \int_0^{\frac{\pi}{2}} dx = \tfrac{\pi}{2}, \quad \langle \varphi_2, \varphi_1 \rangle = \int_0^{\frac{\pi}{2}} x\, dx = \tfrac{\pi^2}{8}, \quad \langle f, \varphi_1 \rangle = \int_0^{\frac{\pi}{2}} \sin x\, dx = 1$$
$$\langle \varphi_2, \varphi_2 \rangle = \int_0^{\frac{\pi}{2}} x^2\, dx = \tfrac{\pi^3}{24}, \quad \langle f, \varphi_2 \rangle = \int_0^{\frac{\pi}{2}} x \sin x\, dx = 1,$$

so that the normal equations are of the form

$$\begin{array}{cc|c} \frac{\pi}{2} & \frac{\pi^2}{8} & 1 \\ \frac{\pi^2}{8} & \frac{\pi^3}{24} & 1 \end{array}$$

and have the solution $\tilde{x}_1 = 8(\pi - 3)/\pi^2 \doteq 0.11477$, $\tilde{x}_2 = 24(4 - \pi)/\pi^3 = 0.66444$, so that the LSM approximation of $\sin x$ from $\mathrm{span}(1, x)$ is the function $\tilde{f} = 0.11477 + 0.66444x$.

Example 3. Comet Tentax moves on an elliptic orbit. In polar coordinates, the following positions of Tentax have been measured:

α	48^o	67^o	83^o	108^o	126^o
r	2.70	2.00	1.61	1.20	1.02

The Kepler law says that

$$r \equiv r(\alpha) = \frac{p}{1 - e\cos\alpha}, \tag{46}$$

where the unknown parameter p characterizes the size of the ellipse and e is its eccentricity.

In the LSM, we approximate a given function f by a function $\tilde{f}$ which is a linear combination $\tilde{x}_1\varphi_1 + \ldots + \tilde{x}_n\varphi_n$ of given functions $\varphi_1, \ldots, \varphi_n$ with unknown parameters $\tilde{x}_1, \ldots, \tilde{x}_n$.

In this problem, the unknown parameters $\tilde{x}_1 = p, \tilde{x}_2 = e$ do not create any linear combination with given functions. But, if we multiply both sides of (46) by $1 - e\cos\alpha$, we obtain

$$r = p + e\,r\cos\alpha \tag{47}$$

and we interpret this identity as a problem to approximate a function r given in the table as linear combination of the constant function $\tilde{\varphi}_1 = 1$ and the given function $\tilde{\varphi}_2 = r\cos\alpha$. The table gives us values of the functions $\tilde{\varphi}_1, \tilde{\varphi}_2$ in the five arguments α from the table. For this reason, we do not work in any function space, but in the vector space $\Re^5$. We approximate the vector

$$f = \begin{bmatrix} 2.70 \\ 2.00 \\ 1.61 \\ 1.20 \\ 1.02 \end{bmatrix}$$

as a linear combination of the vectors

$$\varphi_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \varphi_2 = \begin{bmatrix} 1.81 \\ 0.78 \\ 0.20 \\ -0.37 \\ -0.60 \end{bmatrix}$$

The resulting normal equations are

$$\begin{array}{cc|c} 5 & 1.82 & 8.53 \\ 1.82 & 3.43 & 5.10 \end{array}$$

and its solution is $p = 1.44, e = 0.72$. Hence the Kepler law for this comet is approximately

$$r(\alpha) = \frac{1.44}{1 - 0.72\cos\alpha}.$$

10.6 The (discrete) min–max approximations

Definition . Let the nodes $x_1, \ldots, x_N$ and values $y_1, \ldots, y_N$ (points $[x_1, y_1]$, $\ldots, [x_N, y_N]$) be arbitrary. For a polynomial $p \in \mathcal{P}^n$, we put

$$h_i = p(x_i) - y_i \quad \text{and} \quad H = \max_{1 \le i \le N} |h_i|.$$

We say that p is a *min–max (Chebyshev) approximation (polynomial)* whenever H is the smallest possible.

Definition . We say that $p \in \mathcal{P}^n$ is the *min–max polynomial* if and only if p has the following *equal–error property:* There exists $H \in \Re$ such that

$$p(x_i) - y_i = \pm H$$

in at least $n + 2$ points with alternating sign.

We prove the following general statement in the case $n = 1$ only.

Theorem 9. For every positive integer n there exists a unique min–max polynomial $p \in \mathcal{P}^n$ for all points $[x_1, y_1], \ldots, [x_N, y_N]$ such that $N \ge n + 2$.

Theorem 10. For any points $[x_1, y_1], [x_2, y_2], [x_3, y_3]$ there is exactly one $p \in \mathcal{P}^1$ such that

$$p(x_i) - y_i = \pm E \quad \text{for} \quad i = 1, 2, 3$$

with alternating sign.

Proof. a) Let $p(x) = Mx + B$, $p_i = Mx_i + B = y_i + h_i$ for $i = 1, 2, 3$. Then

$$\begin{aligned} (x_3 - x_2)p_1 - (x_3 - x_1)p_2 + (x_2 - x_1)p_3 &= \\ (x_3 - x_2)(Mx_1 + B) - (x_3 - x_1)(Mx_2 + B) + (x_2 - x_1)(Mx_3 + B) &= \\ M[(x_3 - x_2)x_1 - (x_3 - x_1)x_2 + (x_2 - x_1)x_3] &+ \\ B[(x_3 - x_2) - (x_3 - x_1) + (x_2 - x_1)] &= 0. \end{aligned}$$

If we put $\beta_1 = x_3 - x_2, \beta_2 = x_3 - x_1$ and $\beta_3 = x_2 - x_1$ then we have proved

$$\beta_1 p_1 - \beta_2 p_2 + \beta_3 p_3 = 0 \quad \forall p \in \mathcal{P}^1 \tag{48}$$

b) If $x_1 < x_2 < x_3$, i. e. $\beta_1 > 0, \beta_2 > 0, \beta_3 > 0$, then there exists a unique line with

$$h_1 = h, h_2 = -h, h_3 = h :$$

If $p_1 = y_1 + h, p_2 = y_2 - h, p_3 = y_3 + h$ then

$$\beta_1(y_1 + h) - \beta_2(y_2 - h) + \beta_3(y_3 + h) = 0$$

by (48) and this equality gives us

$$h = -\frac{\beta_1 y_1 - \beta_2 y_2 + \beta_3 y_3}{\beta_1 + \beta_2 + \beta_3}. \tag{49}$$

Hence there exists at most one line passing through $P_1 = [x_1, y_1 + h]$, $P_2 = [x_2, y_2 - h]$, $P_3 = [x_3, y_3 + h]$. By means of (48), we can prove that the slopes

$$\frac{y_2 - y_1 - 2h}{x_2 - x_1} \text{ of } P_1P_2 \text{ and } \frac{y_3 - y_2 + 2h}{x_3 - x_2} \text{ of } P_2P_3$$

are the same. (Prove as an exercise.) Hence there is exactly one $p \in \mathcal{P}^1$ satisfying b).

Definition . The polynomial p from Theorem 9 is said to be an equal–error line for the points $[x_1, y_1], [x_2, y_2], [x_3, y_3]$.

Example . Find the equal–error line for the points $[0, 0], [1, 0.5], [3, 3]$. As $x_1 = 0, x_2 = 1, x_3 = 3$, we have $\beta_1 = 2, \beta_2 = 3, \beta_3 = 1$. Then

$$h = -\frac{2 \cdot 0 - 3 \cdot 0.5 + 1 \cdot 3}{6} = -\frac{1}{4}$$

and our line passes through the points $[0, -\frac{1}{4}], [1, \frac{3}{4}], [3, \frac{11}{4}]$. See Fig. 19.

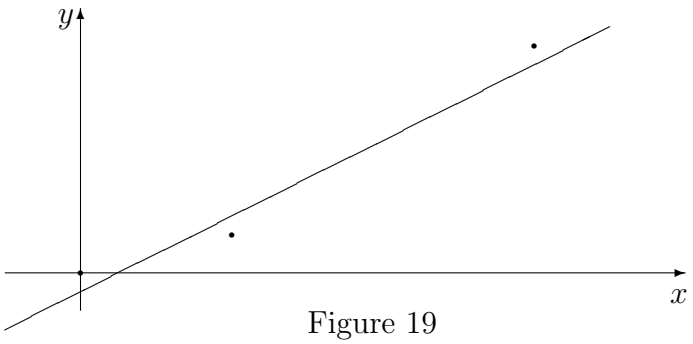


Figure 19

Theorem 11. The equal–error line from Th. 9 is the unique min–max line for $[x_1, y_1], [x_2, y_2], [x_3, y_3]$.

Proof. a) The errors of the equal–error line are $h, -h, h$. Let h_1, h_2, h_3 be the errors of another line $y = p(x)$ and put $H = \max(|h_1|, |h_2|, |h_3|)$. Then $p_1 = y_1 + h_1, p_2 = y_2 + h_2, p_3 = y_3 + h_3$ and by (49), (48),

$$h = -\frac{\beta_1(p_1 - h_1) - \beta_2(p_2 - h_2) + \beta_3(p_3 - h_3)}{\beta_1 + \beta_2 + \beta_3} = \frac{\beta_1 h_1 - \beta_2 h_2 + \beta_3 h_3}{\beta_1 + \beta_2 + \beta_3}. \tag{50}$$

As $\beta_i > 0$, the absolute value of the right–hand side increases if we replace h_1, h_2, h_3 by $H, -H, H$, respectively. We get $h \leq H$.

b) The min–max line is unique: $|h| = H$ if and only if $|h_i| = H$ by a) and the signs alternate. Then the given line is the Chebyshev polynomial.

Now, we can describe the *exchange method* for given N–tuple $[x_1, y_1], \ldots, [x_N, y_N]$ of points:

Step 1. If $x_1 < x_2 < x_3$ then choose the triple $[x_1, y_1], [x_2, y_2], [x_3, y_3]$. If the assumption is not valid, choose any other triple with this property.

Step 2. Find the Chebyshev line p and the error h of the given triple.

Step 3. Compute the errors $p(x_i) - y_i$ of all the points given and denote by H the largest absolute value of them.

Step 4. (Exchange step) Add a data point such that $|h_i| = H$to the old triple and discard one of the former points such that the new triple has errors with alternating signes. Return to Step 2.

We prove:

I. After a finite number of the exchange steps, $|h| = H$: Every new triple has errors $|h|, |h|, H$ with sign alternating and $|h| < H$. The new min–max line has errors $h^*, -h^*, h^*$ with

$$h^* = \frac{\beta_1 h - \beta_2 h + \beta_3 H}{\beta_1 + \beta_2 + \beta_3}$$

(see (50)), so that

$$|h^*| = \frac{\beta_1 |h| + \beta_2 |h| + \beta_3 H}{\beta_1 + \beta_2 + \beta_3} > |h|$$

as $H > |h|$. Hence no triple is choosen twice and procedure terminates.

II. The last Chebyshev line (such that $|h| = H$) is the min–max line of the given N–tuple of points: The errors $h_1, \ldots, h_N$ of any other line satisfy $|h| \leq \max |h_i|$ for h_i belonging to the last triple, so that $|h| \leq \max_{1 \leq i \leq N} |h_i|$.

Example . For the data from the following table

x_i	1	6	14	16	25	29
y_i	2	3	5	4	8	14

find the min–max polynomial.

Let us choose the first triple $[1,2],[6,3],[14,5]$. Then $p_1(x) = 1.6923 + 0.2308x$ and the errors are

i	1	2	3	4	5	6
$p_1(x_i) - y_i$	-0.07	0.07	-0.07	1.38	-0.54	-5.62

The result of the exchange step is the new triple $[1,2],[6,3],[29,14]$ and we obtain $p_2(x) = 1 + 0.4286x$ with the errors

i	1	2	3	4	5	6
$p_2(x_i) - y_i$	-0.57	0.57	2	3.86	3.71	-0.57

The exchange step gives us the triple $[1,2],[16,4],[29,14]$ and we obtain $p_3(x) = -0.6429 + 0.4286x$ with the errors

i	1	2	3	4	5	6
$p_3(x_i) - y_i$	-2.21	-1.07	0.35	2.21	2.07	-2.21

As in this case $H = |h| = 2.21$, $p_2(x)$ is the min–max polynomial for this set of points.

The exchange method can be used for polynomials of any positive degree n in an obvious manner. For an illustration of this modification, consider the following example.

Example . Find the min–max parabola for the data

i	1	2	3	4	5
x_i	-2	-1	0	1	2
y_i	2	1	0	1	2

For the initial quadruple related to $i = 1,2,3,4$, we construct the quadratic polynomial $p_1(x) = a + bx + cx^2$ satisfying $p_1(x_i) - y_i = \pm h$ alternately, i. e.

$$\begin{array}{rrrrrr} a & -2b & +4c & -2 & = & h \\ a & -b & +c & -1 & = & -h \\ a & & & & = & h \\ a & +b & +c & -1 & = & -h \end{array}$$

This system of linear equations has the solution

$$a = \frac{1}{4}, \quad b = 0, \quad c = \frac{1}{2}, \quad d = \frac{1}{4},$$

so that $p_1(x) = \frac{1}{4} + \frac{1}{2}x^2$. As

i	1	2	3	4	5
$p_1(x_i) - y_i$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$

we have $H = \frac{1}{4} = |h|$. Then $p_1(x)$ is the min–max polynomial.

Exercise. Starting with p_1 from the previous example, find the min–max parabola for $y = |x|$ at the points $x = -3, -2, -1, 0, 1, 2, 3$.

11 Numerical differentiation

PROBLEM. Approximate the values of the derivatives $f'(x)$, $f''(x)$ of a sufficiently smooth function f in a fixed point x by means of the values $f(x-h)$, $f(x)$, $f(x+h)$ for some discretization step $h > 0$.

If we approximate the function $f(t)$ by the interpolation polynomial

a) $P_+(t) = f(x) + \frac{f(x+h)-f(x)}{h}(t-x)$ then we obtain

$$f'(x) \doteq \frac{dP_+}{dt}(x) = \frac{f(x+h)-f(x)}{h} = f(x, x+h).$$

b) $P_-(t) = f(x-h) + \frac{f(x)-f(x-h)}{h}(t-x+h)$ then

$$f'(x) \doteq \frac{dP_-}{dt}(x) = \frac{f(x)-f(x-h)}{h} = f(x-h, x).$$

c) $P_2(t) = f(x-h)+f(x-h,x)(t-x+h)+f(x-h,x,x+h)(t-x+h)(t-x)$ then we can easily verify that

$$f'(x) \doteq \frac{dP_2}{dt}(x) = \frac{f(x+h)-f(x-h)}{2h}.$$

The characterizations of the accuracy of these approximations are presented in the following theorem.

Theorem 1. The following statements a) – d) are valid.

a) If $f \in C^2[x, x+h]$ then

$$f'(x) = \frac{f(x+h)-f(x)}{h} - \frac{h}{2}f''(\xi)$$

for some $\xi \in (x, x+h)$.

b) If $f \in C^2[x-h, x]$ then

$$f'(x) = \frac{f(x)-f(x-h)}{h} + \frac{h}{2}f''(\xi)$$

for some $\xi \in (x-h, x)$.

c) If $f \in C^3[x-h, x+h]$ then

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6} f'''(\xi)$$

for some $\xi \in (x-h, x+h)$.

d) If $f \in C^4[x-h, x+h]$ then

$$f''(x) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2} - \frac{h^2}{12} f^{(4)}(\xi)$$

for some $\xi \in (x-h, x+h)$.

Instead of a complete proof, we illustrate the standard method of proving statements of this kind only.

Proof of a). Due to the Taylor Theorem, we have $f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(\xi)h^2$ for a suitable $\xi \in (x, x+h)$ and the statement a) follows.

Proof of c). Again, by means of the Taylor Theorem, we obtain

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{6}f'''(\zeta)h^3 \\ f(x-h) &= f(x) - f'(x)h + \frac{1}{2}f''(x)h^2 - \frac{1}{6}f'''(\eta)h^3 \end{aligned}$$

By subtracting the second equation from the first one and by dividing the difference by two, we obtain

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{1}{6}h^2 \left[\frac{f'''(\zeta) + f'''(\eta)}{2} \right].$$

The Mean Value Theorem says that there exists a ξ between the points ζ, η such that the expression in the rectangular brackets is equal to $f'''(\xi)$. $\square$

Remark . The approximations from Theorem 1 are not applicable to h extremely small because of essential influence of round–off errors. It is possible to obtain accurate approximations of $f'(x)$ or $f''(x)$ without any use of extremely small discretization step h by extrapolation. For example, if we put

$$D_2(h) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}$$

then Theorem 1 d) tells us, roughly, that

$$\begin{aligned} f''(x) &\doteq D_2(h) + Ch^2 \\ f''(x) &\doteq D_2(2h) + C(2h)^2 \end{aligned}$$

for a suitable C. If we multiply the first equation by 4, the second by (-1), add them and divide by 3, we obtain

$$f''(x) \doteq \frac{4D_2(h) - D_2(2h)}{3}.$$

This approximation of $f''(x)$ is essentially more accurate than that one from Theorem 1 c).

12 Numerical integration

Our aim is to approximate the value of

$$I \equiv I(f) = \int_a^b f(x)\,dx \tag{51}$$

for a given function f on a given interval $[a, b]$. In this section, we briefly mention the most simple and most important tools for the solution of this problem.

12.1 Rectangular, trapezoidal and Simpson rules

Definition . In the following considerations, we denote by s the midpoint of the interval $[a, b]$, i. e. we put

$$s = \frac{a+b}{2}.$$

If we substitute $f(x)$ in (51) by
a) the constant $P_0(x) = f(s)$, we obtain the following *rectangular rule*

$$I = (b-a)f(s) + e_R(f).$$

b) the linear polynomial $P_1(x) = f(a) + f(a,b)(x-a)$, we obtain the following *trapezoidal rule*

$$I = \frac{b-a}{2}\,[f(a) + f(b)] + e_T(f).$$

c) the quadratic polynomial $P_2(x) = f(a) + f(a,s)(x-a) + f(a,s,b)(x-a)(x-s)$, we obtain the following *Simpson rule*

$$I = \frac{b-a}{6}\,[f(a) + 4f(s) + f(b)] + e_S(f).$$

Theorem 2. We have the following formulas for the errors e_R, e_T, e_S.
a) If $f \in C^2[a,b]$ then there exists $\xi \in (a,b)$ such that

$$e_R(f) = \frac{1}{24} f''(\xi)(b-a)^3.$$

b) If $f \in C^2[a,b]$ then there exists $\xi \in (a,b)$ such that

$$e_T(f) = -\frac{1}{12} f''(\xi)(b-a)^3.$$

c) If $f \in C^4[a,b]$ then there exists $\xi \in (a,b)$ such that

$$e_S(f) = -\frac{1}{24} f^{(4)}(\xi) \left(\frac{b-a}{2}\right)^5.$$

Proof of a). By the Taylor theorem, we have

$$f(x) = f(s) + f'(s)(x-s) + \frac{1}{2} f''(\xi)(x-s)^2$$

for a suitable $\xi \in (a,b)$. Then

$$\begin{aligned} I = \int_a^b f(x)dx &= \int_a^b \left[f(s) + f'(s)(x-s) + \frac{1}{2} f''(\xi)(x-s)^2 \right] dx \\ &= f(s)(b-a) + f'(s) \left[\frac{(x-s)^2}{2} \right]_a^b + \frac{1}{2} f''(\xi) \left[\frac{(x-s)^3}{3} \right]_a^b \\ &= f(s)(b-a) + \frac{1}{2} f''(\xi) \left[\frac{(b-s)^3}{3} - \frac{(a-s)^3}{3} \right] \\ &= f(s)(b-a) + f''(\xi) \frac{(b-a)^3}{24}. \end{aligned}$$

Proof of b). We obtain the statement b) by expressing $f(x)$ in the form

$$f(x) = f(a) + f(a,b)(x-a) + \frac{1}{2} f''(\xi)(x-a)(x-b)$$

from Theorem 3, Section 10.3. in (51).

Proof of c). We obtain c) by expressing $f(x)$ in the form

$$f(x) = f(a) + f(a,s)(x-a) + f(a,s,b)(x-a)(x-s) + \frac{1}{6} f'''(\xi)(x-a)(x-s)(x-b)$$

from Theorem 3, Section 10.3. in (51). $\square$

Theorem 2 tells us that the above rules a) – c) are accurate for short intervals $[a, b]$ only. If this interval is long, we divide it by equidistant nodes $a = x_0 < x_1 < \ldots < x_n = b$ with step $h = (b - a)/n$. As

$$I = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx, \tag{52}$$

we can use our rules a) – c) for the integrals $\int_{x_{i-1}}^{x_i} f(x)dx$ on the subintervals. We obtain the following *composite rules* d), e), f).

d) If we apply the rectangular rule a) to the integrals from (52), we obtain

$$\begin{aligned} I &= \sum_{i=1}^{n} hf\left(x_{i-1} + \frac{h}{2}\right) + E_R(f) \\ &= h\left[f\left(x_0 + \frac{h}{2}\right) + f\left(x_1 + \frac{h}{2}\right) + \ldots + f\left(x_{n-1} + \frac{h}{2}\right)\right] + E_R(f), \\ E_R(f) &= \frac{1}{24}h^3\left[f''(\xi_1) + f''(\xi_2) + \ldots + f''(\xi_n)\right] = \frac{1}{24}h^2(b-a)f''(\xi) \end{aligned}$$

for a suitable $\xi \in (a, b)$.

e) If we apply the trapezoidal rule b) to the integrals from (52), we obtain

$$\begin{aligned} I &= \sum_{i=1}^{n} \frac{h}{2}\left(f(x_{i-1}) + f(x_i)\right) + E_T(f) \\ &= \frac{h}{2}\left[f(x_0) + 2f(x_1) + \ldots + 2f(x_{n-1}) + f(x_n)\right] + E_T(f), \\ E_T(f) &= -\frac{1}{12}(b-a)h^2 f''(\xi) \end{aligned}$$

for a suitable $\xi \in (a, b)$.

f) If $n = 2k$ is even and we apply the Simpson rule c) to the integrals from x_{2i-2} to x_{2i}, $i = 1, \ldots, k$, we obtain

$$\begin{aligned} I &= \sum_{i=1}^{k} \frac{2h}{6}\left(f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right) + E_S(f) \\ &= \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \ldots + 4f(x_{n-1}) + f(x_n)\right] + E_S(f), \\ E_S(f) &= -\frac{1}{180}(b-a)h^4 f^{(4)}(\xi) \end{aligned}$$

for a suitable $\xi \in (a, b)$.

Definition . A rule for an approximation of $I(f)$ is of *algebraic order* n if n is the largest integer such that the rule calculates the integral $I(P)$ exactly for all $P \in \mathcal{P}^n$.

Example . The rectangular and trapezoidal rules are of algebraic order 1 and the Simpson rule is of algebraic order 3.

12.2 Gauss quadrature

Until now, our rules for the numerical integration give us an approximation of $I(f)$ as linear combinations of values of f in given equidistant nodes with suitable coefficients.

The *Gauss quadrature formulas* approximate the value $I(f)$ by a linear combination

$$c_1 f(x_1) + \ldots + c_n f(x_n)$$

with coefficients $c_1, \ldots, c_n$ and nodes $x_1, \ldots, x_n$ such that the algebraic order of the rule is the largest possible.

For every positive integer n there exists a Gauss formula of algebraic order $2n-1$. Moreover Gauss quadrature is optimal in the following sense. No formula for numerical integration using the values of integrand in n nodes has a bigger algebraic order.

The Gauss formula for $n = 1$ is just the rectangular rule. For $n = 2$, the Gauss formula means

$$I(f) = \frac{b-a}{2}\left[f\left(s - \frac{\sqrt{3}}{6}(b-a)\right) + f\left(s + \frac{\sqrt{3}}{6}(b-a)\right)\right] + \frac{(b-a)^5}{4320} f^{(4)}(\xi)$$

for a suitable $\xi \in (a,b)$ and for $n = 3$, the Gauss formula is of the form

$$\begin{aligned} I(f) &= \frac{(b-a)}{18}\left[5f\left(s - \sqrt{\frac{3}{20}}(b-a)\right) + 8f(s) + 5f\left(s + \sqrt{\frac{3}{20}}(b-a)\right)\right] \\ &+ \frac{(b-a)^7}{2016000} f^{(6)}(\xi). \end{aligned}$$

Here, ξ is a suitable point from the open interval (a,b).

Example . The exact value of the integral $\int_{-1}^{3} e^{-\frac{x^2}{2}} dx$ is 2.051912405. Approximate this value by the rectangular, trapezoidal, Simpson and Gauss rules with the number of nodes $n = 1, 2, 3, 4$.

The solution is summarized in the following table.

	$n=1$	$n=2$	$n=3$	$n=4$
rectangular rule	2.6475	2.1406	2.0896	2.0728
trapezoidal rule	0.9265	1.8801	1.9775	2.0104
Simpson rule		1.9929		2.0538
Gauss quadrature	2.6475	1. .9929	2.0555	

Remark . The trapezoidal rule seems to be the worst among the rules explained in this text. Its essential advantage consists in the possibility to increase its accuracy recursively in a very efficient way.

13 Numerical approximation of the initial–value problem for the ordinary differential equations (ODE)

We search a function $y(x)$ satisfying the ODE

$$y' = f(x, y(x)) \quad \text{for} \quad x \in (a,b) \quad \text{with the } \textit{initial condition} \quad y(a) = c. \tag{53}$$

Definition . We say that the function $f(x,y)$ satisfies the *Lipschitz condition* if there exists $L > 0$ such that

$$|f(x,y) - f(x,z)| \le L|y-z| \quad \forall x \in [a,b] \; \forall y \in \Re.$$

Theorem 1. If $f(x,y)$ satisfies the Lipschitz condition and $x_0 \in [a,b]$, $c_0 \in \Re$ are arbitrary then the problem

$$y' = f(x,y) \quad \text{in} \quad (a,b), \quad y(x_0) = c_0$$

has a unique solution $y(x)$ for $x \in (a,b)$.

Remark . If $\partial f/\partial y$ is bounded in the strip $\Omega = [a,b] \times \Re$ then $f(x,y)$ satisfies the Lipschitz condition with the constant $L = \max_\Omega |\partial f/\partial y|$. Indeed, by the Mean Value Theorem, we have

$$|f(x,y) - f(x,z)| = \left| \frac{\partial f}{\partial y}(x,\xi)(y-z) \right| \le L|y-z|$$

for some ξ between y and z.

All the subsequent numerical methods are based on the following *discretization of problem (53):*

We divide the interval $[a, b]$ by equidistant nodes $a = x_0 < x_1 < \dots$ with step $h > 0$, put $y_0 = c$ and, consecutively, according to a certain formula, calculate the approximations y_i for $i = 1, 2, \dots$ as long as $x_i \in [a, b]$.

The properties of the resulting approximations depend on the following properties of the used method essentially.

Definition . A numerical method for the problem (53) is called

a) the *k–step method* whenever the formula for the approximation y_{i+1} depends on the k preceding approximations $y_i, y_{i-1}, \dots, y_{i-k+1}$.

b) the *l–point method* whenever the formula for y_{i+1} requires to evaluate the function f in l different points.

Definition . A *(global) error* of the approximation y_i is the number $e_i = y(x_i) - y_i$ for $i = 0, 1, \dots$

Definition . We say that a numerical method for the problem (53) is of *order* p whenever $|e_i| \leq c(x_i, h)h^p$ for $i = 1, 2, \dots$ and for a function c bounded for $h \in (0, h_0)$ with h_0 a small positive number.

13.1 One–step methods

a) The *Euler method*: For $i = 0, 1, \dots$, (53) gives us $y'(x_i) = f(x_i, y(x_i))$. If we approximate

$$\begin{aligned} y'(x_i) &\doteq \frac{y(x_{i+1}) - y(x_i)}{h} \doteq \frac{y_{i+1} - y_i}{h} \\ f(x_i, y(x_i)) &\doteq f(x_i, y_i) \end{aligned}$$

we obtain

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i).$$

Then the *Euler method formula:*

$$\begin{aligned} y_0 &= c \\ y_{i+1} &= y_i + hf(x_i, y_i) \quad \text{for} \quad i = 0, 1, \dots \end{aligned}$$

The geometric meaning is illustrated in the following Fig. 21.

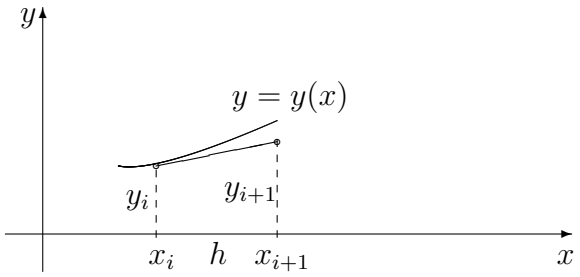


Figure 21

Theorem 2. The Euler method is of order 1.

Example 1. Find the numerical solution of the initial–value problem

$$y' = 2xy \ \text{ in } \ (0,1), \ \ y(0) = 1$$

with the step $h_1 = 0.25$ and $h_2 = 0.125$.

In the case $h = h_1$, the formula is of the form

$$\begin{aligned} y_0 &= 1 \\ y_{i+1} &= y_i + 0.25 \cdot 2x_i y_i = y_i(1 + 0.5x_i) \end{aligned}$$

The resulting Euler approximations $y_0, \ldots, y_4$ appear in the following table.

i	x_i	y_i
0	0	1
1	0.25	1
2	0.5	1.125
3	0.75	1.40625
4	1	1.933594

In the case $h = h_2$, the formula is of the form

$$\begin{aligned} \overline{y}_0 &= 1 \\ \overline{y}_{i+1} &= \overline{y}_i + 0.125 \cdot 2\overline{x}_i\overline{y}_i = \overline{y}_i(1 + 0.25\overline{x}_i) \end{aligned}$$

The resulting Euler approximations $\overline{y}_0, \ldots, \overline{y}_4$ appear in the following table.

i	$\overline{x}_i$	$\overline{y}_i$	$y(x_i)$
0	0	1	1
1	0.125	1	
2	0.25	1.03125	11.0645
3	0.375	1.09570	
4	0.5	1.19843	1.28408
5	0.625	1.34823	
6	0.75	1.55889	1.755053
7	0.875	1.85118	
8	1	2.25613	2.71828

Comparison of the global errors in the common nodes from the two preceding tables illustrates the first order of the Euler method. But this method can be improved by extrapolation on the basis of the following statement.

Theorem 3. If the exact solution $y(x)$ of the problem (53) is in $C^2[a, b]$ and $y(x, h)$ are the values of the Euler method approximation of y by step h then we have

$$y(x) = y(x, h) + c_1(x)h + c_2(x, h)h^2 \tag{54}$$

and $c_2(x, h)$ is bounded for $h \in (0, h_0)$ for some h_0 small and positive.

If we approximate the solution of problem (53) by the Euler method with two different steps h and kh with $k > 0, k \neq 1$, we obtain (54) and

$$y(x) = y(x, kh) + c_1(x)kh + c_2(x, kh)k^2h^2. \tag{55}$$

If we multiply the identity (54) by k and subtract (55), we obtain

$$y(x)(k-1) = ky(x, h) - y(x, kh) + kc_2(x, h)h^2 - c_2(x, kh)k^2h^2.$$

If we divide this identity by $k - 1$ and put

$$C(x, h) = \frac{k(c_2(x, h) - kc_2(x, kh))}{(k-1)},$$

we obtain

$$y(x) = \frac{ky(x, h) - y(x, kh)}{k-1} + C(x, h)h^2,$$

which gives us an approximation of $y(x)$ of order 2.

In the last example, we can put $k = 2$ and $h = 0.125$ and, for $i = 0, 1, 2, 3$, we obtain $y(x_i) \doteq 2y(x_i, 0.125) - y(x_i, 0.25) = 2\overline{y}_{2i} - y_i \equiv y_i^{extr}$. These values are summarized in the following table.

x_i	y_i^{extr}
0	1
0.25	1.0625
0.5	1.27185
0.75	1.71153
0.75	2.57866

Comparison with the values of the exact solution and with the approximations from the two preceding tables illustrates an essential increase of accurracy gained by extrapolation.

b) *Modification of the Euler method* (the *Heun method*)
is an improvement of the Euler method based on the observation that, if the graph of the exact solution $y(x)$ near to x_i is above the tangent line then the difference $k_1 \equiv y_{i+1} - y_i = hf(x_i, y_i)$ is smaller than the difference $y(x_i + 1) - y_i$ and the difference $k_2 = y_{i+2} - y_{i+1} = hf(x_i + h, y_i + k_1)$ is too large. See the illustration of this case in Fig. 22.

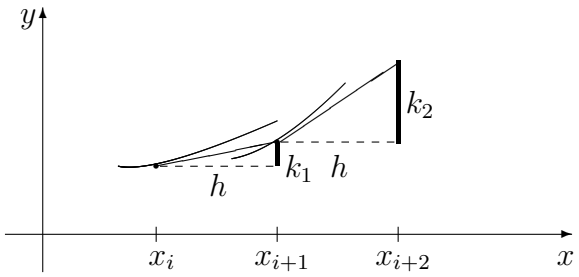


Figure 22

For this reason, this method approximates the difference $y(x_{i+1}) - y_i$ by the arithmetic mean $(k_1 + k_2)/2$. Hence the formula of the *modification of the Euler method* is the following.

$$\begin{aligned} y_0 &= c \\ k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + h, y_i + k_1) \\ y_{i+1} &= y_i + \frac{1}{2}(k_1 + k_2) \quad \text{for} \quad i = 0, 1, \ldots \end{aligned}$$

of course, this is a 1–step and 2–point method.

Theorem 4. The modification of the Euler method is of order 2.

c) An application of a general methodology for the development of new 1–step methods due to Runge and Kutta gave us the following *classical Runge–Kutta method:*

$$\begin{aligned} y_0 &= c \\ k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}) \\ k_3 &= hf(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}) \\ k_4 &= hf(x_i + h, y_i + k_3) \\ y_{i+1} &= y_i + (k_1 + 2k_2 + 2k_3 + k_4)/6 \quad \text{for} \quad i = 0, 1, \ldots \end{aligned}$$

Theorem 5. The classical Runge–Kutta method is of order 4.

Remark . We can see that this classical Runge–Kutta method is a 1–step 4–point method. There exist 1–step 3–point methods of order 3. Hence it seems that a suitable addition of one point increases the order of the method by 1. It is interesting that this linear increase of order stops at the number 4. It has been proved that there exist no 1–step 5–point methods of order 5. For this order, at least 6 points are necessary.

As an illustration of the effect of increase of order, we calculate approximations of the solution of the problem from the last examples by the methods c), d).

Example . Approximate the solution of the problem

$$y' = 2xy \quad \text{in} \quad (0, 1), \quad y(0) = 1$$

by

a) the Modification of the Euler method and

b) the classical Runge–Kutta method

with step $h = 0.5$.

a):

i	x_i	y_i	k_1	k_2	e_i
0	0	1	0	0.5	0
1	0.5	1.25	0.625	1.875	0.034025
2	1	2.5			0.2183

b):

i	x_i	y_i	k_1	k_2	k_3	k_4	e_i
0	0	1	0	0.25	0.281125	0.640625	0
1	0.5	1.283854	0.641927	1.203613	1.414245	2.6981	0.00017125
2	1	2.713145					0.005137

13.2 Multistep methods

a) The *rectangular method*:

$$(53) \implies y'(x_i) = f(x_i, y(x_i))$$

If we approximate

$$\begin{aligned} y'(x_i) &\doteq \frac{y(x_{i+1}) - y(x_{i-1})}{2h} \doteq \frac{y_{i+1} - y_{i-1}}{2h} \\ f(x_i, y(x_i)) &\doteq f(x_i, y_i), \end{aligned}$$

we obtain

$$\frac{y_{i+1} - y_{i-1}}{2h} = f(x_i, y_i),$$

so that the formula of the *rectangular method* is

$$\begin{aligned} y_0 &= c \\ y_1 &= \\ y_{i+1} &= y_{i-1} + 2hf(x_i, y_i) \quad \text{for} \quad i = 1, 2, \ldots \end{aligned}$$

Of course, this is a 2–step, 1–point method.

Theorem 6. The rectangular method is of order 2.

Remark 1. As the formula of this method is not applicable for the evaluation of y_1, it is possible to use any 1–step method for y_1. Even the Euler method does not disturbe the order 2 of the rectangular method.

Remark 2. As the rectangular method is a 1–step method, its complexity is on the level of the most simple Euler method. It seems as if we would gain an accurracy of order 2 gratis. The following example illustrates a sense in which the price for this gain is the loss of stability.

Example . Approximate the problem

$$y' = x - y - 3 \quad \text{in} \quad (0, 1), \quad y(0) = 1$$

by the rectangular method with step $h = 0.25$. The exact solution is $y(x) = 5e^{-x} + x - 4$.

Using the Euler method for y_1, we have

$$\begin{aligned} y_0 &= 1 \\ y_1 &= y_0 + 0.25(x_0 - y_0 - 3) \\ y_{i+1} &= y_{i-1} + 0.5(x_i - y_i - 3) \quad \text{for} \quad i = 1, 2, 3 \end{aligned}$$

and the numerical results computed together with the values of the exact solution are in the following table.

i	x_i	y_i	$y(x_i)$
0	0	1	1
1	0.25	0	0.144004
2	0.5	-0.375	-0.467347
3	0.75	-1.0625	-0.888167
4	1	-0.96875	-1.160603

As the following Fig. 23 illustrates, the signs of the global errors alternate with a tendency to oscillation.

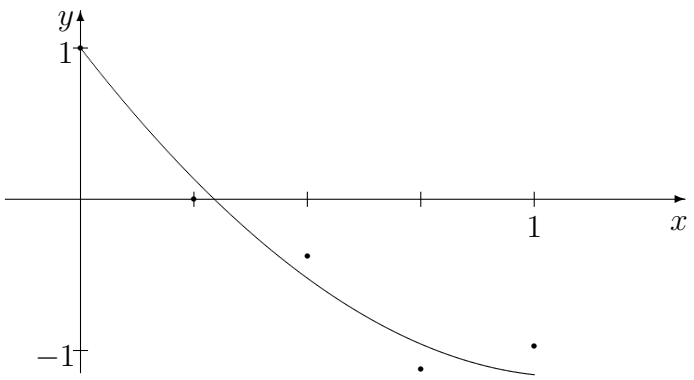


Figure 23

This tendency is removed by the following elementary postprocessing of the approximations calculated by the rectangular method using the Euler method for y_1 due to W. B. Gragg.

b) *Modification of the rectangular method*

$$\begin{aligned} y_0 &= c \\ y_1 &= y_0 + hf(x_0, y_0) \\ y_{i+1} &= y_{i-1} + 2hf(x_i, y_i) \quad \text{for} \quad i = 1, 2, \ldots \end{aligned}$$

Then

$$\overline{y}_i = \frac{y_{i-1} + y_i + hf(x_i, y_i)}{2} \quad \text{for} \quad i = 2, 3, \ldots$$

approximate $y(x_i)$ essentially better than y_i. By means of the results from the preceding example, we can see that

$$\overline{y}_4 = \frac{-1.0625 - 0.96875 + 0.25(-1.03125)}{2} = -1.144531\,!$$

13.3 Implicit methods

All the formulas of the methods presented, we compute

$$y_{i+1} = \mathcal{V}(y_{i-1}, y_i, x_i, f, h),$$

for a certain function $\mathcal{V}$, so that we obtain y_{i+1} by a simple evaluation. Such methods are called *explicit*. The situation becomes more complicated when the formula is of the form

$$y_{i+1} = \mathcal{W}(y_{i-1}, y_i, y_{i+1}, x_i, f, h).$$

Such methods are called *implicit*.

a) *Implicit (backward) Euler method* uses the discretization of the equation (53) of the form

$$\frac{y_{i+1} - y_i}{h} = f(x_{i+1}, y_{i+1}).$$

The resulting formula is

$$\begin{aligned} y_0 &= c \\ y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}) &\quad \text{for} \quad i = 0, 1, \ldots \end{aligned}$$

The geometric meaning illustrates the following Fig. 24.

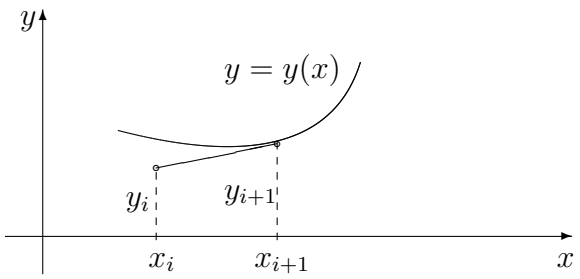


Figure 24

Example . Approximate the initial–value problem

$$y' = 2xy \ \text{ in } \ (0,1), \ \ y(0) = 1$$

by the implicit Euler method with step $h = 0.25$. The formula is of the form

$$\begin{aligned} y_0 &= 1 \\ y_{i+1} &= y_i + 0.5x_{i+1}y_{i+1}, \ i = 0,1,2,3. \end{aligned}$$

Fortunately, in this case, the formula can be "solved with respect to y_{i+1}" with the result

$$\begin{aligned} y_0 &= 1 \\ y_{i+1} &= \frac{y_i}{1 - 0.5x_{i+1}}, \ i = 0,1,2,3. \end{aligned}$$

The computed approximate values are included into the following table.

i	x_i	y_i	$y(x_i)$
0	0	1	1
1	0.25	1.14286	1.0645
2	0.5	1.52381	1.2840
3	0.75	2.43810	1.7550
4	1	4.87619	2.7183

b) The *trapezoidal method*(*Crank–Nicolson*): If we integrate both sides of (53) over $[x_i, x_{i+1}]$, we obtain

$$y(x_{i+1}) - y(x_i) = \int_{x_i}^{x_{i+1}} y'(x)\,dx = \int_{x_i}^{x_{i+1}} f(x, y(x))\,dx.$$

If we approximate $y(x_i)$ by y_i, $y(x_{i+1})$ by y_{i+1} and $\int_{x_i}^{x_{i+1}} f(x)\,dx$ by $h/2[f(x_i, y_i) + f(x_{i+1}, y_{i+1})]$, we obtain the following formula of the *trapezoidal method*:

$$\begin{aligned} y_0 &= c \\ y_{i+1} &= y_i + \frac{h}{2}[f(x_i, y_i) + f(x_{i+1}, y_{i+1})] \end{aligned}$$

Theorem 7. The implicit Euler method is of order 1 and the trapezoidal method is of order 2.

Remark . The trapezoidal method is the only α–method, i. e. method of the form

$$y_{i+1} = y_i + h[\alpha f(x_i, y_i) + (1 - \alpha) f(x_{i+1}, y_{i+1})] \ \text{ for } \ i = 0, 1, \ldots$$

for $\alpha \in [0,1]$, which is of order 2. Clearly, if we put

$\alpha = 1$, we obtain the Euler method,
$\alpha = 0$, we obtain the implicit Euler method and
$\alpha = 1/2$, we obtain the trapezoidal rule.

Example . Let us approximate the standard initial–value problem

$$y' = 2xy \ \text{ in } \ (0,1), \ \ y(0) = 1$$

by the trapezoidal rule with step $h = 0.25$.

In this case, the formula attains the form

$$\begin{aligned} y_0 &= 1, \\ y_{i+1} &= y_i + 0.25(x_i y_i + x_{i+1} y_{i+1}), \end{aligned}$$

after simplification

$$\begin{aligned} y_0 &= 1, \\ y_{i+1} &= y_i \frac{1 + x_i/4}{1 - x_{i+1}/4}. \end{aligned}$$

The resulting approximate values are inluded in the following table.

i	x_i	y_i
0	0	1
1	0.25	1.0667
2	0.5	1.2952
3	0.75	1.7934
4	1	2.8396

13.4 Stability of the numerical methods for the initial–value problems

Definition . The approximate solution $y_0, y_1, \ldots$ of the problem

$$y' = f(x,y) \ \text{ in } \ (a,\infty), \ \ y(a) = c \tag{56}$$

is said to be *non–stable* whenever any error in the computation of y_i inserts components into the approximations $y_{i+1}, y_{i+2}, \ldots$ whose size increases, so that it debases the whole computation essentially.

Example . The problem

$$y' = y - \frac{1}{x^2} - \frac{1}{x} - 1 \ \text{ in } \ (1,\infty), \quad y(1) = 2$$

has an exact solution $y(x) = \frac{1}{x} + 1$ and the general solution of the equation is $y(x) = Ae^x + \frac{1}{x} + 1$. The truncation errors and the error of the numerical method insert the term Ae^x with $A \neq 0$ into the exact solution and its influence increases as x increases. This non–stability is due to the given problem. We are interested in non–stability brought by the numerical method or by too large discretization step.

The stability of a method can be tested by its application to the following *trial problem*

$$y' = \lambda y \ \text{ in } \ (0,\infty), \quad y(0) = 1 \tag{57}$$

Explanation. Assume that, during the solution of a given ODE $y' = f(x,y)$, we have computed the value $y(x) + e(x)$ instead of the exact value $y(x)$. Every numerical method (with certain error which we neglect) requires that the ODE is fulfilled for the value $y(x) + e(x)$, i. e. we assume

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

in the correct case and

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

in the disturbed case. But the last problem is of the following equivalent forms, obtained by the Mean Value Theorem:

$$\begin{aligned} y' + e' &= f(x,y) + [f(x,y+e) - f(x,y)] \\ e' &= \frac{\partial f}{\partial y}(x,\xi) \cdot e \end{aligned}$$

If we put $\lambda \approx \partial f / \partial y(x,\xi)$, we obtain the equation from the problem (57). This consideration gives us the following interpretation of the problem (57):

$$\begin{array}{ll} y(x) & \text{error of the approximation} \\ \lambda \approx \frac{\partial f}{\partial y} & \text{a characteristics of the given problem} \end{array}$$

Definition . We say that a numerical method is *absolutely stable* for a given $\hat{h}$ whenever

$$y_i \longrightarrow 0 \ \text{ as } \ i \longrightarrow \infty$$

for the approximate solutions y_i obtained by a numerical method for the problem (57) with step h satisfying

$$\hat{h} = h\lambda.$$

The set of $\hat{h}$, for which the given problem is absolutely stable create the *interval of absolute stability.* Intervals of absolute stability for some of the basic methods can be found in the following table.

method	order	interval of absolute stability
Euler	1	(-2,0)
classical Runge–Kutta	4	(-2.78,0)
rectangular	2	$\{0\}$
implicit Euler	1	$(-\infty, 0) \cup (2, \infty)$
trapezoidal	2	$(-\infty, 0)$

Example . Approximate the solution of the initial–value problem

$$y' = -10(y-1) \text{ in } (0,1), \quad y(0) = 2$$

by the Euler, classical Runge–Kutta and trapezoidal method with the step a) $h = 0.25$ and b) $h = 0.2$. It is easy to see that this problem has an exact solution

$$y(x) = 1 + e^{-10x}.$$

a) $h = 0.25$:

i	x_i	y_i^{Eul}	y_i^{RK}	y_i^{trap}	$y(x_i)$
0	0	2	2	2	2
1	0.25	-0.5	1.64844	0.88889	1.08208
2	0.5	3.25	1.42047	1.01235	1.00055
3	0.75	-2.375	1.27265	0.99863	1.00055
4	1	6.0625	1.17680	1.00015	1.00005

In Fig. 25, we compare the non–stable Euler approximation $y^{Eul}(x)$ with the exact solution $y(x)$.

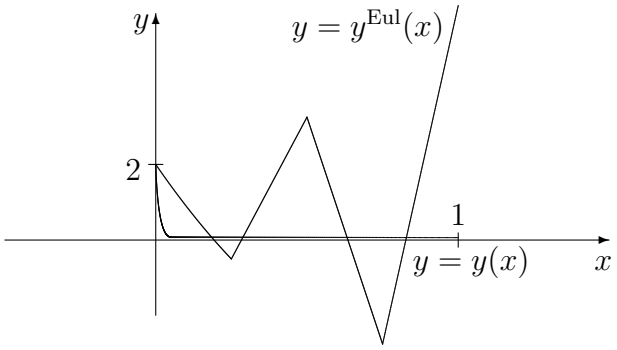


Figure 25

b) $h = 0.2$:

i	x_i	y_i^{Eul}	y_i^{RK}	y_i^{trap}
0	0	2	2	2
1	0.2	0	1.33333	1
2	0.4	2	1.11111	1
3	0.6	0	1.03704	1
4	0.8	2	1.01235	1
5	1	0	1.00412	1

In Fig. 26, the Euler approximation $y^{Eul}(x)$ related to the value of $\hat{\lambda}$ on the boundary of the interval of absolute stability is compared to the exact solution $y(x)$.

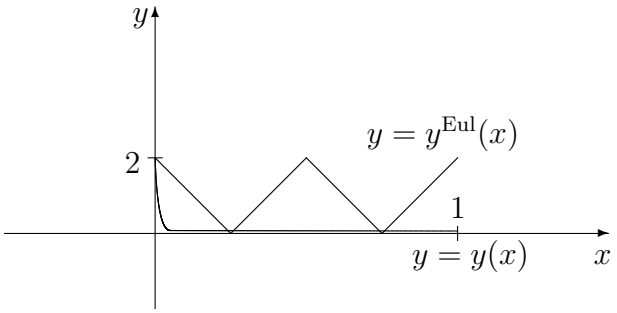


Figure 26

13.5 The initial–value problem for the systems of ODE of order one and of higher orders

In this section, we formulate the general initial–value problem for systems of ODE of order one, illustrate the way in which all the above–presented numerical methods can be used for this problem and construct an equivalent initial–value problem for any initial–value problem for systems of ODE of arbitrary order.

General formulation of the initial–value problem for systems of ODE of order one:

Definition . We search functions $y_1(x), y_2(x), \ldots, y_n(x)$ satisfying

$$\begin{aligned}
y_1' &= f_1(x, y_1(x), \ldots, y_n(x)) \quad \text{for} \quad x \in (a,b), \quad y_1(a) = c_1 \\
y_2' &= f_2(x, y_1(x), \ldots, y_n(x)) \quad \text{for} \quad x \in (a,b), \quad y_2(a) = c_2 \quad (58) \\
&\vdots \\
y_n' &= f_n(x, y_1(x), \ldots, y_n(x)) \quad \text{for} \quad x \in (a,b), \quad y_n(a) = c_n
\end{aligned}$$

If we put

$$y(x) = \begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{bmatrix}, \quad f(x,y) = \begin{bmatrix} f_1(x,y) \\ f_2(x,y) \\ \vdots \\ f_n(x,y) \end{bmatrix}, \quad \text{and} \quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix},$$

then (58) is equivalent to

$$\vec{y}'(x) = \vec{f}(x, \vec{y}(x)) \quad \text{for} \quad x \in (a,b), \quad \vec{y}(a) = \vec{c} \tag{59}$$

As for the problem (53), we formulate sufficient conditions for existence and uniqueness of the exact solution of (59).

Definition . A vector function $\vec{f}(x, \vec{y})$ satisfies the Lipschitz condition for $x \in \langle a, b\rangle$ and $\vec{y} \in \Re^n$ if there exists a constant $L > 0$ with the property

$$\|\vec{f}(x, \vec{y}) - \vec{f}(x, \vec{z})\| \leq L\|\vec{y} - \vec{z}\|$$

for all $x \in \langle a, b\rangle$, $\vec{y}, \vec{z} \in \Re^n$. Here $\|\cdot\|$ is an arbitrary norm on $\Re^n$.

Theorem 8. If $\vec{f}(x, \vec{y})$ satisfies the Lipschitz condition for $x \in \langle a, b\rangle$, $\vec{y} \in \Re^n$ then the problem

$$\vec{y}'(x) = \vec{f}(x, \vec{y}), \quad \vec{y}(x_0) = \vec{y}_0$$

has a unique solution $\vec{y}(x)$ in $\langle a, b\rangle$ for all $x_0 \in \langle a, b\rangle$, $\vec{y}_0 \in \Re^n$.

Remark . Every initial–value problem for a system of ODE of higher orders can be transformed to an equivalent initial–value problem for a system of ODE of order one.

We illustrate this statement by the following example.

Example . Let us consider the following initial–value porblem for one ODE of order four.

$$\begin{aligned} &y^{(4)} = -y' + x^2 y''^2 - y^2 \quad \text{in} \quad (0,1), \\ &y(0) = 0, y'(0) = 1, y''(0) = -1, y'''(0) = 3. \end{aligned}$$

If we denote the derivatives y, y', y'', y''' as new functions

$$y_1 = y, y_2 = y', y_3 = y'', y_4 = y'''$$

then the original problem is equivalent to the following initial–value problem

$$\begin{array}{ll} y_1' = y_2 & \\ y_2' = y_3 & \\ y_3' = y_4 & \text{in } (0,1), \\ y_4' = -y_2 + x^2 y_3^2 - y_1^2 & \end{array} \qquad \begin{array}{l} y_1(0) = 0 \\ y_2(0) = 1 \\ y_3(0) = -1 \\ y_4(0) = 3 \end{array}$$

Remark . Each of the above–mentioned numerical methods for the initial–value problem (53) can be used for the problem (59). In the formulas for the approximate solution, we just write

$$\begin{array}{lll} \vec{c} & \text{instead of} & c \\ \vec{y_i} & \text{instead of} & y_i \\ \vec{f}(x_i, \vec{y_i}) & \text{instead of} & f(x_i, y_i) \\ \vec{k_1} & \text{instead of} & k_1 \\ \vdots & \vdots & \vdots \end{array}$$

14 Numerical approximations of the boundary–value problems for the ODE of order two

14.1 Formulation of the boundary–value problem

For given $a^2 > 0$ and $p, q, f \in C\langle 0, l\rangle$ $(q \geq o)$, find $y \in C^2\langle 0, l\rangle$ such that

$$-a^2 y'' + py' + qy = f \quad \text{for} \quad x \in (0, l) \tag{60}$$

and

$$\begin{array}{rcl} y(x) & = & c \ \text{ Dirichlet boundary condition} \quad \text{or} \\ y'(x) & = & d \ \text{ Neumann boundary condition} \quad \text{or} \\ \alpha y(x) + \beta y'(x) & = & \gamma \ \text{ Newton boundary condition} \quad (\alpha \neq 0 \neq \beta) \end{array}$$

for $x = 0$ and $x = l$.

14.2 Physical meaning

There exists a lot of various physical meanings of this problem. We mention two of them.

$y(x)$	temperature [concentration of alloy]
a^2	heat conductivity [diffusion coefficient]
$p(x)$	velocity of flow
$f(x)$	intensity of sorces of heat [of alloy]
$q(x)$	absorption coefficient

The term $-a^2y'(x)$ means the intensity of flow in the positive direction of the x–axis. Then we can see that the Neumann boundary condition determines the intensity of flow through the boundary and the Newton condition means the heat– [alloy–]transfer condition. The following special case of the Newton condition for $x = l$ concerning heat–flow (here $c > 0$) means that the intensity of heat–flow is proportional to the difference between the temperature of the interval in the boundary point l and the "external" temperature near to the end–point l:

$$-a^2y'(l) = c(y(l) - y_{ext})$$

14.3 Existence of the exact solution

Theorem 1. Let us assume that $p(x)$ does not change the sign in the inteval $\langle 0, l\rangle$. If the Dirichlet boundary condition is given in the point $\left\{\begin{array}{lll} 0 & \text{for} & p \geq o \\ l & \text{for} & p \leq o \end{array}\right\}$, then the problem (60) has a unique solution.

14.4 The standard finite difference method

For the discretisation of the problem (60), consider equidistant nodes $0 = x_0 < x_1 < \ldots < x_n = l$ with the *step* $h = l/n$ and denote by y_i that approximation of $y(x_i)$ which we finally compute for $i = 0, 1, \ldots, n$.

For the inner nodes, i. e. for $i = 1, 2, \ldots, n-1$, (60) gives us

$$-a^2y''(x_i) + p_iy'(x_i) + q_iy(x_i) = f_i.$$

(Here $\varphi_i = \varphi(x_i)$.) If we substitute

$$\begin{array}{lll} y_i & \text{for} & y(x_i), \\ \frac{y_{i+1}-y_{i-1}}{2h} & \text{for} & y'(x_i), \\ \frac{y_{i-1}-2y_i+y_{i+1}}{h^2} & \text{for} & y''(x_i), \end{array}$$

we obtain the following $n-1$ equations

$$-a^2\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + p_i\frac{y_{i+1} - y_{i-1}}{2h} + q_iy_i = f_i,$$

equivalent to

$$-\left(a^2 + \frac{hp_i}{2}\right)y_{i-1} + (2a^2 + h^2q_i)y_i - \left(a^2 - \frac{hp_i}{2}\right)y_{i+1} = h^2f_i \qquad (61)$$

for the unknowns $y_0, y_1, \ldots, y_n$. Due to Theorem 1, Section 11, the errors we introduce by our substitutions correspond to h^2.

We obtain the remaining two equations by discretization of the boundary conditions.

a) The Dirichlet boundary conditions:

$$y(0) = c_1 \Longrightarrow y_0 = c_1 \quad y(l) = c_2 \Longrightarrow y_n = c_2$$

b) The Neumann boundary conditions:

$$y'(0) = d_1 \Longrightarrow \frac{y_1 - y_0}{h} = d_1 \quad y'(l) = d_2 \Longrightarrow \frac{y_n - y_{n-1}}{h} = d_2$$

According to Section 11, Theorem 1 a), b), the errors are proportional to h which is much worse than the errors introduced in the remaining $n-1$ equations. That is why we often approximate the Neumann conditions in the following more sophisticated way: We introduce fictitious "approximate values" y_{-1}, y_{n+1} of the non–existing values of the solution $y(-h), y(l+h)$. Then we can approximate the boundary condition $y'(0) = d_1$ and the equation (60) by the equations

$$\begin{gathered} \tfrac{y_1 - y_{-1}}{2h} = d_1, \\ -\left(a^2 + \tfrac{hp_0}{2}\right) y_{-1} + (2a^2 + h^2 q_0) y_0 - \left(a^2 - \tfrac{hp_0}{2}\right) y_1 = h^2 f_0. \end{gathered}$$

By elimination of y_{-1} from these two equations, we obtain a discretization of $y'(0) = d_1$ with an error corresponding to h^2.

The discretizations of the Neumann condition $y'(l) = d_2$ and of the Newton boundary conditions can be found in analogical two manners.

Definition . The matrix of the resulting system of linear equations (we eliminate y_0 and/or y_n whenever the corresponding boundary condition is of Dirichlet type) is called the *discretization matrix*.

Example 1. Approximate the solution of the problem

$$-0.2y'' + y' = 1 \ \text{ in } \ (0,1), \quad y(0) = 1, \quad y(1) + 0.2y'(1) = 0.5$$

with step $h = 0.2$.

On the basis of the physical meanings of the coefficients

$a^2 = 0.2, p = 1, q = 0, f = 1$ and of the heat transfer condition $-0.2y'(1) = y(1) - 0.5$,

we can roughly illustrate the exact solution in Fig. 27.

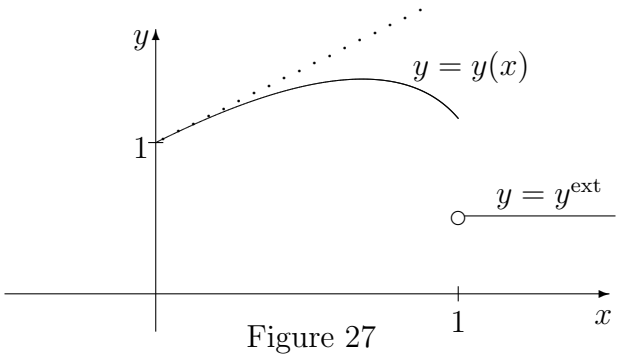


Figure 27

For $h = 0.2$, we have $n = 5$, $x_i = 0.2i$, $i = 0, 1, \ldots, 5$ and we can find the following approximate equations:

$$\begin{aligned}
i = 0 \quad &\Longrightarrow \quad y_0 = 1 \\
i = 1, \ldots, 4 \quad &\Longrightarrow \quad -0.2\frac{y_{i-1} - 2y_i + y_{i+1}}{0.04} + \frac{y_{i+1} - y_{i-1}}{0.4} = 1 \\
&\Longleftrightarrow \quad -0.3y_{i-1} + 0.4y_i - 0.1y_{i+1} = 0.04 \\
i = 5 \quad &\Longrightarrow \quad y_5 + 0.2\frac{y_6 - y_4}{0.4} = 0.5 \ \text{ and } \ -0.3y_4 + 0.4yy_5 - 0.1y_6 = 0.04 \\
&\Longrightarrow \quad -0.4y_4 + 0.6yy_5 = 0.14
\end{aligned}$$

The matrix form of the remaining system is

$$\left(\begin{array}{ccccc|c}
0.4 & -0.1 & & & & 0.34 \\
-0.3 & 0.4 & -0.1 & & & 0.04 \\
 & -0.3 & 0.4 & -0.1 & & 0.04 \\
 & & -0.3 & 0.4 & -0.1 & 0.04 \\
 & & & -0.4 & 0.6 & 0.14
\end{array}\right)$$

In the following table, the solution of this system is compared with the exact values of the exact solution $y(x)$.

i	y_i	$y(x_i)$
1	1.1950	1.1901
2	1.3760	1.3633
3	1.5219	1.4903
4	1.5597	1.4920
5	1.2731	1.1529

We can see that $\max_{i \le i \le 5} |y(x_i) - y_i| = 0.1202$.

If we put $\overline{h} = 0.1$ then we obtain $\overline{n} = 10$, $\overline{x}_i = 0.1i$ and by denoting $\overline{y}_i$ the computed approximate values of $y(\overline{x}_i)$ for $i = 0, 1, \ldots, 10$, we obtain

$\max_{i\le i\le 10} |y(\overline{x}_i) - \overline{y}_i| = 0.0271$. This is an indication that the following remark is valid.

Remark . The standard finite difference method is of order 2.

Example 2. Approximate the solution of the problem

$$-y'' + 15y' = 1 \ \text{ in } \ (0,1), \ \ y(0) = 1, \ \ y(1) = 0$$

with step $h = 0.2$.

Then we have $n = 5$, $x_i = 0.2i$ for $i = 0, 1, \ldots, 5$, $y_0 = 1, y_5 = 0$ and, for $i = 1, \ldots, 4$, the discretizations

$$-2.5y_{i-1} + 2y_i + y_{i+1} = 0.04$$

After substitution of the values of y_0 and y_5, we obtain the following matrix form of the resulting system of equations.

$$\left(\begin{array}{cccc|c} 2 & 0.5 & & & 2.54 \\ -2.5 & 2 & 0.5 & & 0.04 \\ & -2.5 & 2 & 0.5 & 0.04 \\ & & -2.5 & 2 & 0.04 \end{array}\right)$$

with the following resulting values

i	y_i	$y(x_i)$
1	1.0113	1.0133
2	1.0349	1.0267
3	0.9970	1.0374
4	1.2662	1.0002

The comparison with the values of the exact solution from the previous table shows that the approximate solution is debased by oscillations. This fact can be observed in the following Fig. 28, too.

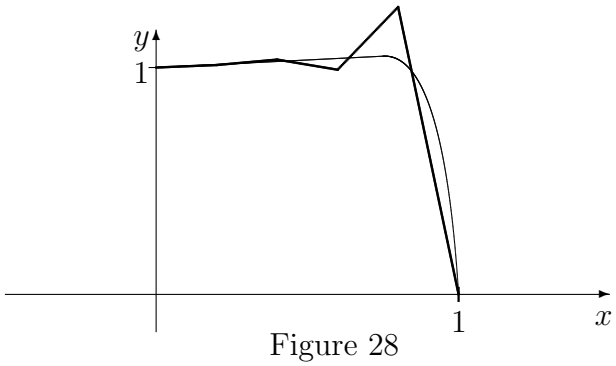


Figure 28

Now, we formulate sufficient conditions for the stability of the approximate solution which are used very successfully in many practical situations.

Definition . A square matrix is said to be *monotone* whenever A is regular and $A^{-1} \geq 0$ (this means that all entries of A^{-1} are non–negative).

Theorem 2. The discretization matrix A of any problem from Theorem 1 is monotone if and only if $i \neq j \Longrightarrow a_{ij} \leq 0$.

Example . It is easy to see that the discretization matrix from Example 1 is monotone. Then, for $i = 2, 3, 4$, we have

$$y_i = \frac{1}{0.4}(0.3y_{i-1} + 0.1y_{i+1} + 0.04),$$

so that, in essential, y_i is a weighted average of y_{i-1}, y_{i+1} with positive coefficients. That is why the resulting approximation is stable.

Remark . We obtain by Theorem 2 and by (63) that the discretization matrix A is monotone whenever

$$a^2 + \frac{hp_i}{2} \geq 0 \ \text{ and } \ a^2 - \frac{hp_i}{2} \geq 0 \iff a^2 \geq \left|\frac{hp_i}{2}\right|$$

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

The classical elementary stabilizations of the standard finite difference method are

a) The *artificial diffusion method*: If there exists an index i such that $a^2 < |hp_i/2|$ then we substitute the coefficient a^2 by the least $\overline{a}^2$, so that $\overline{a}^2 \geq \max_i |hp_i/2|$.

b) *Upwind*: Instead of the approximation $(y_{i+1} - y_{i-1})/2$, we approximate $y'(x_i)$ by the difference quotient $\begin{array}{lll} (y_i - y_{i-1})/h & \text{for} & p > 0 \\ (y_{i+1} - y_i)/h & \text{for} & p < 0 \end{array}$.

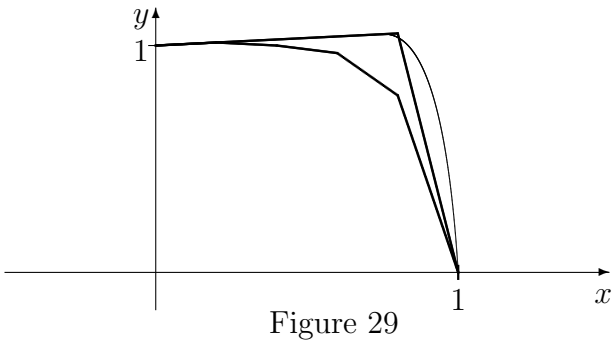


Figure 29

In Fig. 29, approximations of the exact solution (red graph) by the artificial diffusion method (blue graph) and by upwind (green graph) are illustrated. Although the modifications a), b) are stable, their accuracy is of order 1, much worse than the order 2 of the standard finite difference method. This is one of the reasons why a big lot of new numerical methods have been proposed for the stabilization of numerical approximation which are of high accuracy at the same time.