Numerical methods

Peeter Oja

February 27, 2018

Contents

Introduction

While investigating nature, computational methods are used according to the following scheme:

1) an object, a phenomenon or a process is investigated. Some of the more important ones: heat spreading in a body, electric current in semiconductors, infiltration of a liquid through soil, weather, sea waves, nuclear reactor, burning (more generally: a chemical reaction), the price of stocks on the stock exchange.

2) a mathematical model is formed. These are, e.g., function, equation (system of equations), differential equation (including ordinary or partial) and initial value problem or boundary value problem, integral equation, random process, optimization problem. Usually the laws of nature have been relied on here.

3) the mathematical model is explored: properties of the function, the existence and uniqueness of the solutions of the equations, properties of the solutions. This is the part that the classical disciplines – algebra, calculus, geometry, probability theory – deal with.

4) the true solution (function) is found. In practical applications it is almost always found approximately. Computers are used for this.

Numerical methods deal with the 4. step. Our goal is to answer the question: Which is the best way to carry out step 4?

In this course the main topics are:

1) solving equations;

2) solving systems of equations;

3) approximating functions;

4) approximating definite integrals.

Before going to the main topics we take a brief look at errors.

§1. A brief overview of errors

Errors do not mean anything bad, they are an inevitability instead. The existence of an error means that the solution, that has been found practically from the real world, differs from the absolutely exact solution.

Mistakes on the other hand are bad: a mistake occurs when something that can be done correctly has been done wrongly.

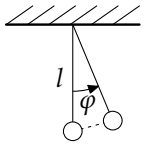
1.1 Types of errors

1) **Errors of mathematical formulation or errors of model.** This comes from the fact that mathematical models (equations, systems of equations) describe real phenomena approximately.

   Example. The differential equation

   $$l\frac{\mathrm{d}^2\varphi}{\mathrm{d}t^2} + k\frac{\mathrm{d}\varphi}{\mathrm{d}t} + g\sin\varphi = 0$$

   describes the oscillation of a pendulum. The notation is the following:

   t time;
   φ angle of deviation from vertical position;
   l length of the pendulum;
   k friction coefficient of air;
   g acceleration of gravity.

   The unknown is usually the function $\varphi = \varphi(t)$.

   While forming the equation it is assumed that the gravitational field has parallel field lines, but in reality they are centrosymmetric; the resistance force $k\frac{\mathrm{d}\varphi}{\mathrm{d}t}$ depends proportionately on the velocity ($\frac{\mathrm{d}\varphi}{\mathrm{d}t}$ is proportional to tangential velocity), but that only takes place at low speeds. Parameters l, k and g need to be measured and that can only be done approximately. When the oscillation is small, we have $\sin\varphi \approx \varphi$. Such a substitution simplifies the equation, because then we get a linear equation with constant coefficients, but that is additional approximation.

2) **Errors of method.** Methods are divided into exact and approximate ones. A method is called exact if it gives the exact solution after exact performing a finite number of arithmetic operations. Otherwise the method is called approximate. An example of an exact method is solving a system of linear equations using determinants. This course deals with approximate methods. It turns out that approximate methods have much bigger practical significance than exact methods.

3) **Round-off errors.** For finding the solution a computer is usually used, but any computer can operate only with a finite number of rational numbers. Therefore we cannot get by without rounding real numbers (or the real and imaginary part of a complex number).

An error of mathematical formulation is called *unconditional* error, error of method and round-off error is called *conditional* error. An unconditional error cannot be changed, it is determined by the model. A conditional error can be modified to be however small (usually it is accompanied by growing costs), but it is not reasonable to make it much smaller than the unconditional error. For this we must have at least some idea what is the magnitude of the unconditional error.

1.2 Absolute and relative error of a number

Consider a number a, which is approximately equal to number A. Call number A exact; in practical applications it is usually impossible to find the value of A. The actual error is $\Delta a = A - a$, which can also be expressed with the equality $A = a + \Delta a$.

Definition. The *absolute error* of an approximate number a is any number $\Delta > 0$ that satisfies the inequality $|A - a| \leqslant \Delta$, i.e., $|\Delta a| \leqslant \Delta$.

It is clear that the absolute error is not uniquely determined. E.g., if $A = \pi$, $a = 3.14$, then $\Delta = 1$, $\Delta = 0.0016$ or $\Delta = 0.001593$.

The inequality $|A - a| \leqslant \Delta$ is equivalent to the inequalities $-\Delta \leqslant A - a \leqslant \Delta$ or $a - \Delta \leqslant A \leqslant a + \Delta$. This situation is also denoted by $A = a \pm \Delta$, meaning still that $A \in [a - \Delta, a + \Delta]$.

The actual relative error is $\delta a = \frac{\Delta a}{a}$.

Definition. The *relative error* of an approximate number a is any number $\delta > 0$ that satisfies the inequality $|\frac{\Delta a}{a}| \leqslant \delta$, i.e., $|\delta a| \leqslant \delta$.

If Δ is known, we may take $\delta = \frac{\Delta}{|a|}$. If δ is known, then $\Delta = \delta \cdot |a|$ suits.

The absolute error has the same unit as a or A. The relative error δ has no unit, it is often given in percentages.

It is clear from the context that the requirements $A \neq 0$ and $a \neq 0$ are assumed.

1.3 Finding the error of the value of a function

Consider a function $u = u(x_1, \ldots, x_n)$, approximate numbers $x_1, \ldots, x_n$ as the approximations of exact numbers $X_1, \ldots, X_n$, the absolute errors $\Delta_1, \ldots, \Delta_n$ of numbers $x_1, \ldots, x_n$. We find $u(x_1, \ldots, x_n)$ and ask the question: what is the absolute error Δ_u of this number, also, what is its relative error δ_u.

Assume that function u is differentiable. We know that $X_i = x_i + \Delta x_i$, $|\Delta x_i| \leqslant \Delta_i$, $i = 1, \ldots, n$. Now

$$\begin{aligned}\Delta u &= u(X_1, \ldots, X_n) - u(x_1, \ldots, x_n) = \\ &= u(x_1 + \Delta x_1, \ldots, x_n + \Delta x_n) - u(x_1, \ldots, x_n) = \\ &= \sum_{i=1}^{n} \frac{\partial u}{\partial x_1}(x_1, \ldots, x_n)\Delta x_i + R,\end{aligned}$$

where the last equation is the Taylor's or Lagrange's formula, which holds for every differentiable u, R is the remainder.

Assume additionally, that the numbers Δ_i are small. Then the numbers Δx_i are also small and since u is differentiable then R is also small. If we ignore the remainder R, then we obtain

$$|\Delta u| \approx \left| \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \Delta x_i \right| \leqslant \sum_{i=1}^{n} \left| \frac{\partial u}{\partial x_i} \right| |\Delta x_i| \leqslant \sum_{i=1}^{n} \left| \frac{\partial u}{\partial x_i} \right| \Delta_i.$$

We take

$$\Delta_u \eqsim \sum_{i=1}^{n} \left| \frac{\partial u}{\partial x_i} \right| \Delta_i,$$

whereas the approximation sign means here that we may estimate down the absolute error by $|R|$. Additionally we obtain

$$\delta_u = \frac{1}{|u|}\Delta_u \eqsim \frac{1}{|u|} \sum_{i=1}^{n} \left| \frac{\partial u}{\partial x_i} \right| \Delta_i = \sum_{i=1}^{n} \left| \frac{\partial}{\partial x_i} \ln u \right| \Delta_i,$$

i.e.,

$$\delta_u \eqsim \sum_{i=1}^{n} \left| \frac{\partial}{\partial x_i} \ln u \right| \Delta_i.$$

We have solved the problem of finding the absolute and relative error at computing the value of a function approximately, but this is natural taking into account the generality of the situation.

Example. Consider $u = x_1 + \ldots + x_n$, then $\frac{\partial u}{\partial x_i} = 1$ and $\Delta_u = \Delta_1 + \ldots + \Delta_n$.

Example. Consider $u = x_1 - x_2$, then $\left|\frac{\partial u}{\partial x_i}\right| = 1$ and $\Delta_u = \Delta_1 + \Delta_2$.

Exercise 1. Prove that the formulae of absolute error for addition and subtraction are exact (without approximation unlike the general case) and that they hold for any size of errors (not just for small ones).

Exercise 2. Prove that if the summands have the same sign, then $\delta \leqslant \max\limits_{1 \leqslant i \leqslant n} \delta_i$, i.e., the relative error of the sum does not exceed the biggest relative error of the summands. This claim also holds for any size of errors.

Exercise 3. Consider $S = ab$, approximate numbers $a = 3$ and $b = 4$ with relative errors $\Delta_a = 2$ and $\Delta_b = 3$. Find the relative error of the product Δ_S. Take into account that in this case the absolute errors of the arguments are not small.

1.4 Finding the errors of the arguments of a function

Consider the function $u(x_1, \ldots, x_n)$, approximate numbers $x_1, \ldots, x_n$ and Δ_u. We need to find the absolute errors $\Delta_1, \ldots, \Delta_n$ of arguments $x_1, \ldots, x_n$ or the relative errors $\delta_1, \ldots, \delta_n$. If $n \geqslant 2$ then there is no unique way for finding those errors. Sometimes the errors are known for some of the arguments. The following variants are used

1) $\Delta_1 = \ldots = \Delta_n$,

2) $\delta_1 = \ldots = \delta_n$,

they determine the errors of the arguments on the basis of the formula for evaluating the overall error Δ_u (or δ_u).

Example. The floor of the room has sizes 3 m and 4 m. How precisely these should be measured to compute the area of the floor with the accuracy of at least 0.01 m^2?

Make the natural assumption that the room is rectangular. Then the area is computable by the formula $S = ab$, where $a = 3$ and $b = 4$. Whereas the error $\Delta_S = 0.01$ has also been given. Now $\Delta_S = a\Delta_b + b\Delta_a =$

$(a+b)\Delta$, as we assume that $\Delta = \Delta_a = \Delta_b$ and we may use the formula for finding Δ_S because the errors are small. Then

$$\Delta = \frac{\Delta_S}{a+b} = \frac{0.01}{3+4} = 0.0014\ldots \approx 0.001 \text{ m} = 1 \text{ mm}.$$

1.5 Rounding of errors

Error of the value of a function is to be rounded up, errors of the arguments are to be rounded down. The explanation of this principle is the following. Keeping in mind the notation from the previous two paragraphs, we see that the relation between the errors of arguments and the error of function value is following:

$$\begin{aligned} X_i \in &[x_i - \Delta_i, x_i + \Delta_i],\ i = 1, \ldots, n \implies \\ &\implies u(X_1, \ldots, X_n) \in [u(x_1, \ldots, x_n) - \Delta_u, u(x_1, \ldots, x_n) + \Delta_u]. \end{aligned}$$

If the numbers Δ_i increase, the implication fails to hold; the implication does not hold either if Δ_u decreases.

I Solving equations

Introduction

Most well known equations are algebraic equations

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n = 0, \qquad a_0 \neq 0.$$

Such an equation can be solved by radicals (using roots) for the cases $n = 1, \ldots, 4$. For $n \geqslant 5$ it is generally impossible. We will look mainly equations

$$f(x) = 0,$$

where f is any function. A great majority of the equations appearing in practical applications are solved approximately. For this, iteration methods are mostly used. These are as follows: a collection of initial values $x_0, \ldots, x_k$ is given or chosen (there may also be only one initial value). Then, step by step, next approximations are found using previous ones

$$x_0, \ldots, x_k \to x_{k+1} \to x_{k+2} \to \ldots .$$

So a sequence x_n is found by a method of iteration.

When dealing with iteration methods one must always study the convergence, i.e., one must answer the question: does the sequence x_n converge to the solution? If it does not, then there is no reason to use this sequence in order to find an approximate solution.

Example. Bisection method. Consider a function $f\colon [a, b] \to \mathbb{R}$, which is continuous and $f(a)f(b) < 0$. It is known that under such assumptions there exists a point $x^* \in (a, b)$ such that $f(x^*) = 0$, i.e., x^* is the solution to equation $f(x) = 0$. We take $x_0 = a$, $x_1 = b$, $x_2 = \frac{x_0+x_1}{2}$ and check whether $f(x_0)f(x_2) < 0$ or $f(x_2)f(x_1) < 0$. In the first case we take $x_3 = \frac{x_0+x_2}{2}$, in the second case $x_3 = \frac{x_1+x_2}{2}$. Then we continue by dividing into halves the interval in which the function f changes its sign. With this method the estimate

$$|x_n - x^*| \leqslant \frac{b-a}{2^{n-1}} \to 0$$

holds as $n \to \infty$. It means that the error ($\frac{b-a}{2^{n-1}}$ is the absolute error, $|x_n - x^*|$ is the absolute value of the actual error) decreases in geometric progression with the common ratio $\frac{1}{2}$. In practice this is considered slow.

§1. Ordinary iteration method

1.1 Description of the method and convergence theorem

Consider the equation

$$x = g(x). \tag{1.1}$$

In the ordinary iteration method, one initial value x_0 is needed, after that we find

$$x_{n+1} = g(x_n), \quad n = 0, 1, \dots .$$

Theorem 1 (Convergence Theorem)**.** *Assume that*

1) $g\colon [a,b] \to [a,b]$, *i.e.*, $x \in [a,b] \implies g(x) \in [a,b]$,

2) g *is a contraction on interval* $[a,b]$, *i.e.*, $\exists q < 1$ *such that*

$$|g(x_1) - g(x_2)| \leqslant q|x_1 - x_2| \quad \forall x_1, x_2 \in [a,b], x_1 \neq x_2.$$

Then equation (1.1) has exactly one solution x^* *in interval* $[a,b]$. *For every* $x_0 \in [a,b]$ *it holds* $x_n \to x^*$ *with the estimate*

$$|x_n - x^*| \leqslant \frac{q^n}{1-q}|x_0 - x_1|. \tag{1.2}$$

Proof. Choose arbitrarily $x_0 \in [a,b]$ and form the sequence

$$x_{n+1} = g(x_n), \quad n = 0, 1, \dots .$$

Then

$$x_0 \in [a,b] \implies x_1 = g(x_0) \in [a,b] \implies x_2 = g(x_1) \in [a,b] \implies \dots ,$$

i.e., $x_n \in [a,b]$ for every n. Show that x_n is a Cauchy sequence. Firstly

$$\begin{aligned} |x_n - x_{n+1}| = |g(x_{n-1}) - g(x_n)| &\leqslant q|x_{n-1} - x_n| \leqslant \\ &\leqslant q^2|x_{n-2} - x_{n-1}| \leqslant \dots \leqslant q^n|x_0 - x_1|. \end{aligned}$$

Then

$$|x_n - x_{n+p}| \leqslant |x_n - x_{n+1}| + |x_{n+1} - x_{n+2}| + \ldots + |x_{n+p-1} - x_{n+p}| \leqslant$$
$$\leqslant q^n|x_0 - x_1| + \ldots + q^{n+p-1}|x_0 - x_1| \leqslant$$
$$\leqslant \left(\sum_{k=n}^{\infty} q^k\right)|x_0 - x_1| = q^n(1 + q + \ldots)|x_0 - x_1| =$$
$$= \frac{q^n}{1-q}|x_0 - x_1| \to 0, \qquad (2')$$

as $n \to \infty$ independently of the index p. With that it has been shown that the sequence x_n is fundamental.

Every Cauchy sequence, consisting of real numbers, converges, therefore $\exists x^* \in \mathbb{R}$ such that $x_n \to x^*$. Since $[a, b]$ is closed, we have $x^* \in [a, b]$. From the equalities $x_{n+1} = g(x_n)$, going to the limit $n \to \infty$, we obtain that $x^* = g(x^*)$, because g is continuous (every contractive mapping is continuous).

Show the uniqueness of the solution. Having $x^*, x^{**} \in [a, b]$ such that $x^* = g(x^*)$ and $x^{**} = g(x^{**})$, we obtain that

$$|x^* - x^{**}| = |g(x^*) - g(x^{**})| \leqslant q|x^* - x^{**}|.$$

In general, $|x^* - x^{**}| = 0$ or $|x^* - x^{**}| > 0$. Last case gives $q|x^* - x^{**}| < |x^* - x^{**}|$ which contradicts to the inequality established above. Thus, $|x^* - x^{**}| = 0$ or $x^* = x^{**}$.

Now only the inequality (1.2) should be established, but this follows from the inequality (2′), if we go to the limit $p \to \infty$. Then $x_{n+p} \to x^*$, $x_n - x_{n+p} \to x_n - x^*$ and $|x_n - x_{n+p}| \to |x_n - x^*|$. ■

Corollary 2 (From the proof of Convergence Theorem)**.** *The theorem remains valid if the interval $[a, b]$ is replaced by the set of all real numbers $\mathbb{R}$ or any half-line $[a, \infty)$ or $(-\infty, b]$.*

Corollary 3. *In Convergence Theorem and in Corollary 2 the requirement of contraction 2) may be replaced by the assumption that g is differentiable and*

$$|g'(x)| \leqslant q < 1 \quad \forall x \in [a, b] \text{ (or } \forall x \in \mathbb{R}, [a, \infty), (-\infty, b]).$$

Proof. Note that from the assumptions made,

$$g(x_1) - g(x_2) = g'(\xi)(x_1 - x_2), \quad \xi \in (x_1, x_2).$$

Also, if $x_1, x_2 \in [a, b]$, then $\xi \in [a, b]$ and g is a contraction since $|g'(\xi)| \leqslant q < 1$. ■

Exercise 4. Let the equation $x = g(x)$ have a solution x^* and g be a contraction on an interval $(x^* - \delta, x^* + \delta)$, $\delta > 0$. Prove that if $x_0 \in (x^* - \delta, x^* + \delta)$, then the ordinary iteration method converges to the solution x^*.

Exercise 5. Let the equation $x = g(x)$ have a solution $x^* \in [a, b]$ and $0 \leqslant g'(x) \leqslant q < 1 \quad \forall x \in [a, b]$. Prove that at every $x_0 \in [a, b]$ the ordinary iteration method converges to the solution x^*.

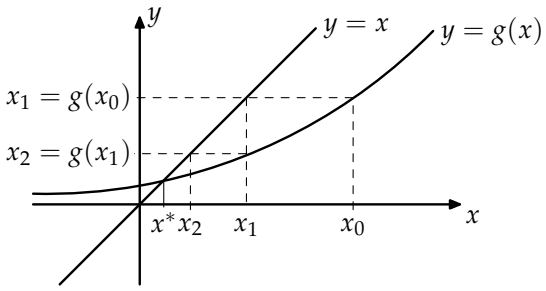
Exercise 6. Find an example of a function $g\colon \mathbb{R} \to \mathbb{R}$, where $|g(x_1) - g(x_2)| < |x_1 - x_2| \ \forall x_1, x_2 \in \mathbb{R}$, $x_1 \neq x_2$, while the equation $x = g(x)$ does not have a solution.

Exercise 7. The same as exercise 6, but the set $\mathbb{R}$ is replaced with some half-line $[a, \infty)$.

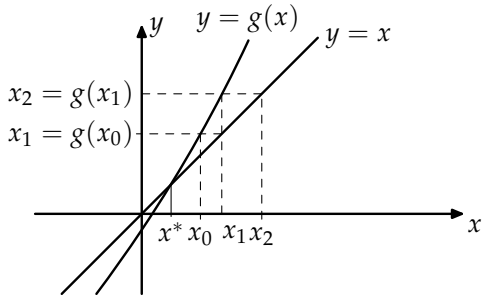
Remark. If $g\colon [a, b] \to [a, b]$ and $|g(x_1) - g(x_2)| < |x_1 - x_2| \ \forall x_1, x_2 \in [a, b]$, $x_1 \neq x_2$, then there exists $x^* \in [a, b]$ such that $x^* = g(x^*)$.

1.2 Geometric interpretation

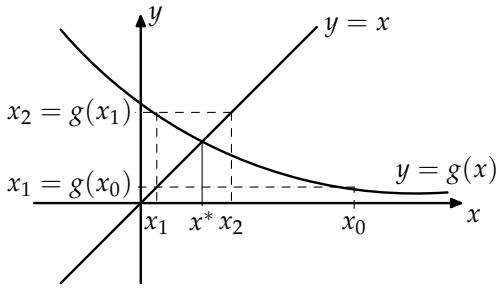
Here we present figures that show how the ordinary iteration method behaves.



In this figure $0 < g'(x) < 1$, the method converges.



In this figure $g'(x) > 1$, the method does not converge.



In this figure $-1 < g'(x) < 0$, the method converges.

Exercise 8. Draw a figure describing the case $g'(x) < -1$.

1.3 Behaviour of the ordinary iteration method near the solution

Assume that the function g is continuously differentiable. Let x^* be a solution of equation (1.1), i.e., $x^* = g(x^*)$. Let x_n be a sequence of iterations that has been found according to the rule $x_{n+1} = g(x_n)$, $n = 0, 1, \ldots$. Then according to Lagrange's formula,

$$x_{n+1} - x^* = g(x_n) - g(x^*) = g'(\xi_n)(x_n - x^*),$$

where either $\xi_n \in (x_n, x^*)$ or $\xi_n \in (x^*, x_n)$ if $x_n < x^*$ or $x_n > x^*$, respectively. If $x_n \approx x^*$ (i.e., if $|x_n - x^*|$ is small), then $\xi_n \approx x^*$ and $g'(\xi_n) \approx g'(x^*)$, because g' is continuous. Therefore

$$x_{n+1} - x^* \approx g'(x^*)(x_n - x^*).$$

The essence of this relation is that near the solution x^* the error $x_n - x^*$ behaves approximately like the geometric progression with the common ratio $g'(x^*)$.

If $0 < |g'(x^*)| < 1$ then (at least if we start form a small enough neighbourhood of the solution) the ordinary iteration method converges with the rate of geometric progression having the common ratio $g'(x^*)$. If $g'(x^*) = 0$, then the ordinary iteration method converges faster than any geometric progression, which means that for any $q > 0$ it holds the convergence $\frac{|x_n - x^*|}{q^n} \to 0$ as $n \to \infty$ (here we consider arbitrarily small values of q).

Exercise 9. Prove that if $g'(x^*) = 0$ then for any $q > 0$ it holds $\dfrac{|x_n - x^*|}{q^n} \to 0$ as $n \to \infty$.

If $|g'(x^*)| > 1$ then the ordinary iteration method does not converge to the solution.

In any case, the sign of $g'(\xi_n)$ shows whether the approximations x_n and x_{n+1} lie on one side of the solution x^* or on different sides. The figures in the previous paragraph also illustrate that claim.

1.4 Examples

We need to solve the equation $x^2 - a = 0$, i.e., we need to find the square root of number a. We may assume that $a > 0$. This equation is not in the form (1.1), we shall present two ways of converting it to that form.

1) Put the equation in the form $x = \frac{a}{x}$, meaning that $g(x) = \frac{a}{x}$. Then the iteration formula is $x_{n+1} = \frac{a}{x_n}$. Here $g'(x) = -\frac{a}{x^2}$ and if $(x^*)^2 = a$, then $g'(x^*) = -\frac{a}{(x^*)^2} = -1$. According to the argumentation given in the previous paragraph, we cannot decide whether this sequence converges. After more extensive research, that we shall not show here, we can affirm, that for the case $a > 1$ the sequence diverges, and converges if $a < 1$, but more slowly than any geometric progression, which means that for any $q \in (0,1)$ it holds $\frac{|x_n - x^*|}{q^n} \to \infty$ as $n \to \infty$ (here we are considering the case when q is arbitrarily close to 1).

2) Write the equation $x = \frac{a}{x}$ or $2x = x + \frac{a}{x}$ in the form $x = \frac{1}{2}(x + \frac{a}{x})$, i.e., here $g(x) = \frac{1}{2}(x + \frac{a}{x})$ and the iteration is done by the formula $x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n})$. In this case $g'(x) = \frac{1}{2}(1 - \frac{a}{x^2})$ and $g'(x^*) = 0$, which means that this iteration method converges faster than any geometric progression.

Exercise 10. Prove that if g is m times differentiable, $g^{(m)}$ is bounded and $g'(x^*) = 0, \ldots, g^{(m-1)}(x^*) = 0$, then the estimate

$$|x_{n+1} - x^*| \leqslant \text{const}\,|x_n - x^*|^m \tag{1.3}$$

holds in the ordinary method of iteration.

Convergence with estimate (1.3) is called *m-th order convergence*. If $m = 2$ then it is called *quadratic convergence* and if $m = 3$ then *cubic convergence*. More generally, if the estimation $|x_{n+1} - x^*| \leqslant \text{const}\,|x_n - x^*|^\alpha$, $\alpha > 1$ holds, then we may speak about *order convergence*, this is faster than any geometrical progression.

Exercise 11. With the help of exercise 10 prove that the method $x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n})$ for finding a square root has quadratic convergence.

§2. Newton's method

2.1 Description of the method

Consider the equation

$$f(x) = 0, \tag{1.4}$$

with this equation we also assume that f is differentiable. In the Newton's method, one initial value is given, next approximations are found according to the formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \dots .$$

In this method a step can be carried out if $f'(x_n) \neq 0$.

Newton's method can be viewed as a special case of the ordinary iteration method, where the equation has been transformed to the form

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

where g is defined in those points of the domain of f in which $f'(x) \neq 0$. For that reason, all the results that we obtained for the ordinary iteration method, can be applied here.

Exercise 12. Prove that if f is twice continuously differentiable, $f(x^*) = 0$ (i.e., x^* is the solution of equation (1.4)) and $f'(x^*) \neq 0$, then the corresponding function g has property $g'(x^*) = 0$, i.e., Newton's method converges faster than any geometric progression.

Exercise 13. Prove that if f is three times continuously differentiable with $f(x^*) = 0$, $f'(x^*) \neq 0$, then Newton's method has quadratic convergence. Use exercise 10 from the previous paragraph.

Assume that, while solving equation (1.4), x_n has been found (here x_0 may also be considered as x_n). Then, according to Taylor's formula, we see that

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + R(x_n; x).$$

The equation (1.4) is equivalent to the equation

$$f(x_n) + f'(x_n)(x - x_n) + R(x_n; x) = 0.$$

If we omit the remainder $R(x_n; x)$, then we obtain the (linear) equation

$$f(x_n) + f'(x_n)(x - x_n) = 0,$$

the solution of this equation is

$$x_n - \frac{f(x_n)}{f'(x_n)} = x_{n+1}.$$

So, at every step of Newton's method the initial equation is replaced by its linearization, every step consists of solving a linear equation. Newton's method is one of linearization methods.

2.2 Rate of convergence of Newton's method

In the previous paragraph, many results that were given (as exercises) apply to the rate of convergence of Newton's method. Here we show that many of the assumptions in those exercises can be weakened.

Assume that f is continuously differentiable, $f(x^*) = 0$ and $f'(x^*) \neq 0$ (solution x^* has multiplicity one). Denote $\varepsilon_n = x_n - x^*$. From the formula of Newton's method, we obtain

$$\varepsilon_{n+1} = \varepsilon_n - \frac{f(x_n)}{f'(x_n)}.$$

Using Taylor's expansion, we obtain

$$f(x_n) = f(x^*) + f'(\xi_n)(x_n - x^*) = f'(\xi_n)\varepsilon_n,$$

where $\xi_n \in (x_n, x^*)$ or $\xi_n \in (x^*, x_n)$. Now

$$\varepsilon_{n+1} = \varepsilon_n - \frac{f'(\xi_n)}{f'(x_n)}\varepsilon_n = \left(1 - \frac{f'(\xi_n)}{f'(x_n)}\right)\varepsilon_n.$$

If $x_n \to x^*$ then $\xi_n \to x^*$, $f'(\xi_n) \to f'(x^*) \neq 0$, $f'(x_n) \to f'(x^*)$, hence $1 - \frac{f'(\xi_n)}{f'(x_n)} \to 0$. This gives $|\varepsilon_{n+1}| = q_n|\varepsilon_n|$, where $q_n \to 0$, which means that Newton's method converges faster than any geometric progression. To emphasize the importance of this result, we present it as a proposition.

Proposition 4. *If f is continuously differentiable and the solution to* (1.4) *has multiplicity one, then Newton's method converges faster than any geometric progression.*

Assume that f' satisfies Lipschitz condition, i.e., there exists a number L such that $|f'(x) - f'(y)| \leqslant L|x - y|$. Then

$$\left|1 - \frac{f'(\xi_n)}{f'(x_n)}\right| = \frac{|f'(x_n) - f'(\xi_n)|}{|f'(x_n)|} \leqslant \frac{L|x_n - \xi_n|}{|f'(x_n)|} \leqslant \frac{L|\varepsilon_n|}{\frac{1}{2}|f'(x^*)|} = \text{const}\,|\varepsilon_n|,$$

because $|x_n - \xi_n| \leqslant |x_n - x^*| = |\varepsilon_n|$ and due to the convergence $f'(x_n) \to f'(x^*)$ we get $|f'(x_n)| \leqslant \frac{1}{2}|f'(x^*)|$ if x_n is close enough to the solution x^*. Therefore $|\varepsilon_{n+1}| \leqslant \text{const}\,|\varepsilon_n|^2$ and we formulate this result as a proposition as well.

Proposition 5. *If f' satisfies Lipschitz condition, then Newton's method has quadratic convergence in the case of solution having multiplicity one.*

Note that f' satisfies Lipschitz condition, if f'' is continuous (more generally, f'' is bounded), because $f'(x) - f'(y) = f''(\xi)(x - y)$, $\xi \in (x, y)$.

2.3 Rate of convergence of Newton's method in the case of multiple solution

Let f be m times continuously differentiable and $f(x^*) = 0, f'(x^*) = 0, \ldots, f^{(m-1)}(x^*) = 0, f^{(m)}(x^*) \neq 0$. In this case we say that the solution x^* has *multiplicity* m. In the equality

$$\varepsilon_{n+1} = \varepsilon_n - \frac{f(x_n)}{f'(x_n)},$$

which we obtained from the calculation rule of Newton's method, we expand $f(x_n)$ and $f'(x_n)$ according to Taylor's formula in the point x^*, then

$$\begin{aligned} f(x_n) &= f(x^*) + f'(x^*)(x_n - x^*) + \frac{f''(x^*)}{2!}(x_n - x^*)^2 + \ldots + \\ &\quad + \frac{f^{(m)}(\xi_n)}{m!}(x_n - x^*)^m = \frac{f^{(m)}(\xi_n)}{m!}\varepsilon_n^m, \quad \xi_n \in (x_m, x^*), \\ f'(x_n) &= f'(x^*) + f''(x^*)(x_n - x^*) + \ldots + \frac{f^{(m)}(\eta_n)}{(m-1)!}(x_n - x^*)^{m-1} = \\ &= \frac{f^{(m)}(\eta_n)}{(m-1)!}\varepsilon_n^{m-1}, \quad \eta_n \in (x_n, x^*). \end{aligned}$$

Now

$$\varepsilon_{n+1} = \varepsilon_n - \frac{f^{(m)}(\xi_n)(m-1)!}{m!f^{(m)}(\eta_n)}\varepsilon_n = \left(1 - \frac{1}{m}\frac{f^{(m)}(\xi_n)}{f^{(m)}(\eta_n)}\right)\varepsilon_n.$$

If $x_n \to x^*$, then $\xi_n \to x^*$, $\eta_n \to x^*$, $f^{(m)}(\xi_n) \to f^{(m)}(x^*) \neq 0$, $f^{(m)}(\eta_n) \to f^{(m)}(x^*)$, thus $\varepsilon_{n+1} = q_n\varepsilon_n$, where $q_n \to 1 - \frac{1}{m}$ (for instance, if $m = 2$ then $q_n \to \frac{1}{2}$, if $m = 3$ then $q_n \to \frac{2}{3}$). The greater is m, the slower is the convergence.

Proposition 6. *If f is m times continuously differentiable and the solution of equation* (1.4) *has multiplicity m, then Newton's method converges as fast as a geometric progression with common ratio* $1 - \frac{1}{m}$.

To improve this situation Newton–Schröder's method is used

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \ldots,$$

where m is the multiplicity of the solution.

Exercise 14. Show that if $f^{(m)}$ is continuous, then Newton–Schröder's method converges faster than any geometric progression, and if $f^{(m)}$ satisfies Lipschitz condition, then Newton–Schröder's method has quadratic convergence.

It is natural to ask that how do we find the multiplicity in Newton–Schröder's method from the initial equation (1.4), i.e., from the function f, if we do not know the solution x^* (this will not be found, but only its approximation) and we cannot find the derivatives at x^*.

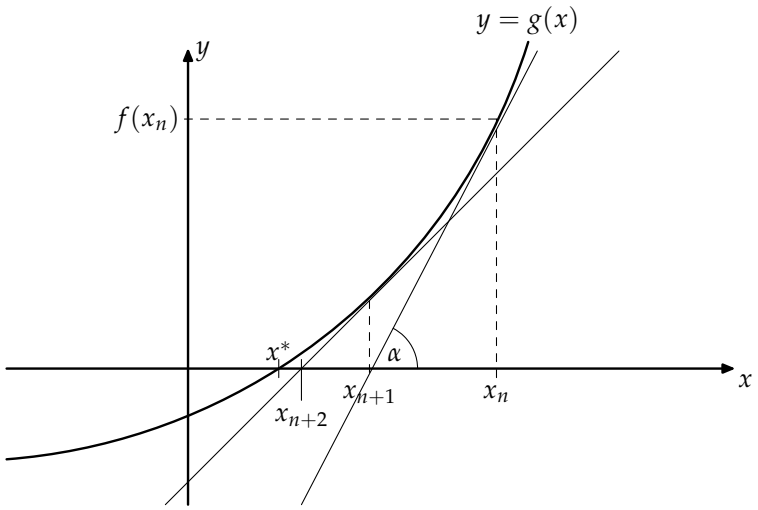
Exercise 15. Prove that if solution has multiplicity m and sequence x_n is found by Newton's method, then

$$\frac{x_{n+1} - x_n}{x_n - x_{n-1}} \to 1 - \frac{1}{m} \quad \text{and} \quad \frac{x_n - x_{n-1}}{-x_{n+1} + 2x_n - x_{n-1}} \to m.$$

Therefore, we may begin solving the equation with Newton's method and if after some steps the multiplicity of the solution is known, then we may continue with Newton–Schröder's method.

2.4 Geometric interpretation of Newton's method

In this paragraph we represent Newton's method geometrically. Look at the figure



At the point x_n we move vertically till the graph, there we draw the tangent. Let us show that the point where the tangent intersects the x-axis gives us x_{n+1}. We have

$$\frac{f(x_n)}{x_n - x_{n+1}} = \tan\alpha = f'(x_n),$$

where the first equality comes from a right-angled triangle, and the second equality comes from the geometric interpretation of derivative. From the established equality (if we omit the middle term $\tan\alpha$) we obtain the equality $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$. In the presented figure $f'(x) > 0$, $f''(x) > 0$, $x_n > x^*$.

Exercise 16. Draw figures for the cases

1) $f'(x) > 0$, $f''(x) < 0$;

2) $f'(x) < 0$, $f''(x) > 0$;

3) $f'(x) < 0$, $f''(x) < 0$.

For all the cases consider the situations $x_n > x^*$ and $x_n < x^*$. Draw three consecutive approximations x_n, x_{n+1} and x_{n+2} in the figures.

Exercise 17. Find a geometric figure, where

a) the equation $f(x) = 0$ has a solution, but Newton's method cannot be applied, because $f'(x_n) = 0$;

b) the equation $f(x) = 0$ has a solution and all members of the sequence x_n can be found using Newton's method, but the sequence x_n is unbounded and therefore does not converge;

c) the equation $f(x) = 0$ has a solution and all members of the sequence x_n can be found using Newton's method, the sequence x_n is bounded, but does not converge.

Due to the geometric interpretation, Newton's method is sometimes called *the method of tangents*.

2.5 Modified Newton's method

In modified Newton's method, the computational formula for solving equation (1.4) is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \quad n = 0, 1, \dots ,$$

i.e., on every step the derivative $f'(x_0)$ at initial value x_0 is used. If we take $g(x) = x - \frac{f(x)}{f'(x_0)}$, then we see that the modified Newton's method is a special case of the ordinary method of iteration. At that case $g'(x) = 1 - \frac{f'(x)}{f'(x_0)}$ and $g'(x^*) = 1 - \frac{f'(x^*)}{f'(x_0)}$. Generally $f'(x^*) \neq f'(x_0)$, which yields $g'(x^*) \neq 0$ and the modified Newton's method converges with the rate of a geometric progression, under the assumption that it converges at all.

2.6 Examples

1) Consider the equation $f(x) = x^2 - a = 0$. Then $f'(x) = 2x$, Newton's method is

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{2x_n^2 - x_n^2 + a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

That is familiar from the paragraph about the ordinary method of iteration, we know that it has quadratic convergence. It converges for any initial value x_0.

2) Consider the equation $f(x) = x^3 - a = 0$, then $f'(x) = 3x^2$ and the calculation formula is

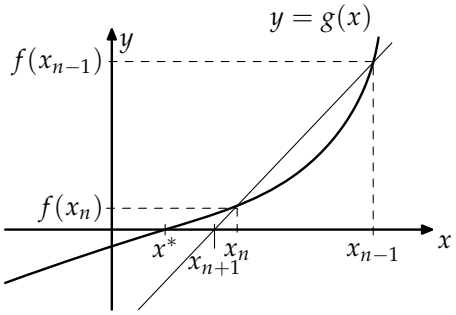
$$x_{n+1} = x_n - \frac{x_n^3 - a}{3x_n^2} = \frac{1}{3}\left(2x_n + \frac{a}{x_n^2}\right).$$

Exercise 18. How many initial values are in Example 2), and what do they look like, where Newton's method does not converge?
Hint: study the geometric interpretation.

§3. Other iteration methods

3.1 Secant method

Consider the equation $f(x) = 0$. Let $x_0, x_1, x_0 \neq x_1$ be given initial values. We connect the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$ by a straight line. Let the first coordinate of the intersection point of this line and the x-axis be x_2. Then we repeat the same procedure with points x_1 and x_2 and then so on. Let us derive the computational formula, where we find x_{n+1} using similarily the approximations x_{n-1} and x_n.



Using right-angled triangles, we obtain

$$\frac{f(x_{n-1})}{f(x_n)} = \frac{x_{n-1} - x_{n+1}}{x_n - x_{n+1}}.$$

Hence

$$\begin{aligned} f(x_{n-1})(x_n - x_{n+1}) &= f(x_n)(x_{n-1} - x_{n+1}), \\ (f(x_n) - f(x_{n-1}))x_{n+1} &= x_{n-1}f(x_n) - x_n f(x_{n-1}), \\ x_{n+1} &= \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}, \end{aligned}$$

this last formula can be used to find x_{n+1}. If we write $\pm x_n f(x_n)$ into the numerator, we obtain

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n).$$

If f is differentiable, then according to Lagrange's formula $f(x_n) - f(x_{n-1}) = f'(\xi_n)(x_n - x_{n-1})$, $\xi_n \in (x_n, x_{n-1})$, therefore

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(\xi_n)}.$$

The last formula shows that the secant method is similar to Newton's formula, where in the secant method there is $f'(\xi_n)$ instead of $f'(x_n)$. However, the secant method is not a particular case of the ordinary iteration method, because, on every step, two previous approximations are used to find the next approximation.

3.2 Rate of convergence of secant method

In this paragraph we will establish the rate of convergence of the secant method, assuming that the method converges.

We assume that function f is smooth enough and $f(x^*) = 0$, $f'(x^*) \neq 0$, $f''(x^*) \neq 0$ (hence the solution x^* has multiplicity one). Additionally assume that in the secant method $x_n \to x^*$ at least with the rate of some geometric progression, i.e., $|x_{n+1} - x^*| \leqslant q|x_n - x^*|$, $q < 1$. The aim is to find out, which is the actual rate of convergence.

We begin with the equality

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}.$$

If we subtract x^* from both sides of the equality and take into account that $\varepsilon_n = x_n - x^*$, we obtain

$$\varepsilon_{n+1} = \frac{\varepsilon_{n-1}f(x_n) - \varepsilon_n f(x_{n-1})}{f(x_n) - f(x_{n-1})},$$

which we shall study as the main relation between errors. In the numerator of the right hand side of the equality, we use Taylor's expansions

$$f(x_n) = f(x^*) + f'(x^*)(x_n - x^*) + \frac{1}{2}f''(x^*)(x_n - x^*)^2 + \frac{1}{6}f'''(\xi_n)(x_n - x^*)^3,$$

$$f(x_{n-1}) = f(x^*) + f'(x^*)\varepsilon_{n-1} + \frac{1}{2}f''(x^*)\varepsilon_{n-1}^2 + \frac{1}{6}f'''(\xi_{n-1})\varepsilon_{n-1}^3,$$

where $\xi_n \in (x_n, x^*)$ and $\xi_{n-1} \in (x_{n-1}, x^*)$. Then in the main relation

$$\text{numerator} = \frac{1}{2}f''(x^*)\varepsilon_{n-1}\varepsilon_n(\varepsilon_n - \varepsilon_{n-1}) + \frac{1}{6}f'''(\xi_n)\varepsilon_{n-1}\varepsilon_n(\varepsilon_n^2 - \varepsilon_{n-1}^2) + \\ + \frac{1}{6}\varepsilon_{n-1}\varepsilon_n(f'''(\xi_n) - f'''(\xi_{n-1}))\varepsilon_{n-1}^2.$$

We verify, that the first term is main, i.e., the other terms converge faster. Compared to the first term, in the second one $\varepsilon_n^2 - \varepsilon_{n-1}^2 = (\varepsilon_n - \varepsilon_{n-1})(\varepsilon_n + \varepsilon_{n-1})$ has the factor $\varepsilon_n + \varepsilon_{n-1} \to 0$ and $f'''(\xi_n)$ is bounded if f is smooth enough. In the last term $f'''(\xi_n) - f'''(\xi_{n-1}) = f^{\mathrm{IV}}(\hat{\xi}_n)(\xi_n - \xi_{n-1})$, $\hat{\xi}_n \in (\xi_n, \xi_{n-1})$ and again due to f being smooth, $f^{\mathrm{IV}}(\hat{\xi}_n)$ is bounded. If x_{n-1} and x_n are on different sides of the solution x^*, e.g., $x_{n-1} < x^* < x_n$, then $x_{n-1} < \xi_{n-1} < x^* < \xi_n < x_n$ giving $|\xi_n - \xi_{n-1}| \leqslant |x_{n-1} - x_n| = |\varepsilon_{n-1} - \varepsilon_n|$ and the faster convergence of the last term compared to the first term has been proved. But if x_{n-1} and x_n are on the same side of the solution x^*, e.g., $x^* < x_n < x_{n-1}$, then $|\xi_n - \xi_{n-1}| \leqslant |x_{n-1} - x^*|$ since $x^* < \xi_{n-1} < x_{n-1}$ and $x^* < \xi_n < x_n$. However,

$$|x_{n-1} - x^*| \leqslant |x_{n-1} - x_n| + |x_n - x^*| \leqslant |x_{n-1} - x_n| + q|x_{n-1} - x^*|,$$
$$(1-q)|x_{n-1} - x^*| \leqslant |x_{n-1} - x_n|,$$

from which we obtain

$$|\xi_n - \xi_{n-1}| \leqslant \frac{1}{1-q}|x_{n-1} - x_n| = \frac{1}{1-q}|\varepsilon_{n-1} - \varepsilon_n|$$

and also in this case we have shown that last term converges faster than first one. So, in the main relation numerator $\approx \frac{1}{2}f''(x^*)\varepsilon_{n-1}\varepsilon_n(\varepsilon_n - \varepsilon_{n-1})$. We proceed similarly with the denominator in the right hand side of the main relation of errors, expanding

$$f(x_n) = f(x^*) + f'(x^*)\varepsilon_n + \frac{1}{2}f''(\eta_n)\varepsilon_n^2, \quad \eta_n \in (x_n, x^*),$$
$$f(x_{n-1}) = f(x^*) + f'(x^*)\varepsilon_{n-1} + \frac{1}{2}f''(\eta_{n-1})\varepsilon_{n-1}^2, \quad \eta_{n-1} \in (x_{n-1}, x^*).$$

Then

$$\text{denominator} = f'(x^*)(\varepsilon_n - \varepsilon_{n-1}) + \frac{1}{2}f''(\eta_n)(\varepsilon_n^2 - \varepsilon_{n-1}^2) +$$
$$+ \frac{1}{2}(f''(\eta_n) - f''(\eta_{n-1}))\varepsilon_{n-1}^2.$$

Here also second term converges faster than first one. In addition, $f''(\eta_n) - f''(\eta_{n-1}) = f'''(\hat{\eta}_n)(\eta_n - \eta_{n-1})$, $\hat{\eta}_n \in (\eta_{n-1}, \eta_n)$ and $|\eta_n - \eta_{n-1}| \leqslant \frac{1}{1-q}|\varepsilon_n - \varepsilon_{n-1}|$. Therefore in the main relation the denominator $\approx f'(x^*)(\varepsilon_n - \varepsilon_{n-1})$. All in all

$$\varepsilon_{n+1} \approx \frac{\frac{1}{2}f''(x^*)\varepsilon_n\varepsilon_{n-1}}{f'(x^*)} = a\varepsilon_n\varepsilon_{n-1},$$

where we have denoted $a = \frac{f''(x^*)}{2f'(x^*)}$.

We are looking for a solution of the equation $\varepsilon_{n+1} = a\varepsilon_n\varepsilon_{n-1}$, where $\varepsilon_n \to 0$ and $\varepsilon_{n+1} = b\varepsilon_n^\alpha$. We are mostly interested in the order of the convergence α. Then $\varepsilon_n = b\varepsilon_{n-1}^\alpha$ and $\varepsilon_{n+1} = b\varepsilon_n^\alpha = b(b\varepsilon_{n-1}^\alpha)^\alpha = b^{1+\alpha}\varepsilon_{n-1}^{\alpha^2}$. On the other hand, $\varepsilon_{n+1} = a\varepsilon_n\varepsilon_{n-1} = a(b\varepsilon_{n-1}^\alpha)\varepsilon_{n-1} = ab\varepsilon_{n-1}^{\alpha+1}$, which leads to the equation $\alpha^2 = \alpha + 1$ or $\alpha^2 - \alpha - 1 = 0$. This equation has the solutions $\alpha = \frac{1}{2} \pm \frac{\sqrt{5}}{2}$. Solution $\alpha = \frac{1}{2} - \frac{\sqrt{5}}{2} < 0$ does not allow convergence $\varepsilon_n \to 0$, therefore we are left with a suitable solution $\alpha = \frac{1+\sqrt{5}}{2} = 1.618\ldots$ and convergence $\varepsilon_n \to 0$ takes place with the estimate

$$|x_{n+1} - x^*| \leqslant \text{const}\,|x_n - x^*|^{1.618\ldots}.$$

3.3 Steffensen's method

If we apply the ordinary iteration method to the equation $x = g(x)$ and, e.g., $|g'(x^*)| = 0.99$, then we need many iteration steps to get reasonable accuracy. Steffensen's method is one way to increase the convergence rate in such case. There are several ways to describe this method.

1) To solve the equation $x = g(x)$, we begin with the initial value x_0 and find $x_1 = g(x_0)$, then $x_2 = g(x_1)$ and

$$\tilde{x}_2 = x_2 - \frac{(x_2 - x_1)^2}{x_2 - 2x_1 + x_0} = \frac{x_0x_2 - x_1^2}{x_2 - 2x_1 - x_0}.$$

We continue with the approximation $\tilde{x}_2$ and find $x_3 = g(\tilde{x}_2)$, $x_4 = g(x_3)$, then we find $\tilde{x}_4$ similarly to $\tilde{x}_2$ and so on. In general, if $\tilde{x}_n$ has been found (n is even), then $x_{n+1} = g(\tilde{x}_n)$, $x_{n+2} = g(x_{n+1})$ and

$$\tilde{x}_{n+2} = x_{n+2} - \frac{(x_{n+2} - x_{n+1})^2}{x_{n+2} - 2x_{n+1} + \tilde{x}_n}.$$

The previous formula for finding $\tilde{x}_{n+2}$ is called *Aitken's transform*. So we may describe the method as follows:

$$x_0 \xrightarrow{\text{o. it.}} x_1 \xrightarrow{\text{o. it.}} x_2 \xrightarrow{\text{A. t.}} \tilde{x}_2 \xrightarrow{\text{o. it.}} x_3 \xrightarrow{\text{o. it.}} x_4 \xrightarrow{\text{A. t.}} \tilde{x}_4 \longrightarrow \ldots .$$

2) We find $x_1 = g(x_0)$ and apply the secant method at points x_0 and x_1

to the equation $f(x) \equiv x - g(x) = 0$. Then

$$\begin{aligned} f(x_0) &= x_0 - g(x_0) = x_0 - x_1, \\ f(x_1) &= x_1 - g(x_1) = x_1 - x_2, \\ \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} &= \frac{x_0(x_1 - x_2) - x_1(x_0 - x_1)}{x_1 - x_2 - (x_0 - x_1)} = \\ &= \frac{-x_0 x_2 + x_1^2}{-x_2 + 2x_1 - x_0} = \tilde{x}_2. \end{aligned}$$

Therefore we may present Steffensen's method as

$$x_0 \underset{x=g(x)}{\overset{\text{o. it.}}{\longrightarrow}} x_1 \underset{x-g(x)=0}{\overset{\text{s. m.}}{\longrightarrow}} \tilde{x}_2 \underset{x=g(x)}{\overset{\text{o. it.}}{\longrightarrow}} x_3 \underset{x-g(x)=0}{\overset{\text{s. m.}}{\longrightarrow}} \tilde{x}_4 \longrightarrow \dots .$$

3) Steffensen's method may also be considered as the ordinary iteration method applied to the equation $x = \varphi(x)$, where

$$\varphi(x) = \frac{x g(g(x)) - (g(x))^2}{g(g(x)) - 2g(x) + x},$$

i.e.,

$$x_0 \underset{x=\varphi(x)}{\overset{\text{o. it.}}{\longrightarrow}} \tilde{x}_2 \underset{x=\varphi(x)}{\overset{\text{o. it.}}{\longrightarrow}} \tilde{x}_4 \longrightarrow \dots .$$

We study the rate of convergence of Steffensen's method by using results about the ordinary iteration method and the third description of the method.

Assume that g is continuously differentiable, $x^* = g(x^*)$, $g'(x^*) \neq 0$, $g'(x^*) \neq 1$. Previously given formula does not define $\varphi(x^*)$ as for $x = x^*$ the value of the denominator is 0. Define $\varphi(x^*) = x^*$. Let us find $\varphi'(x^*) = \lim\limits_{x \to x^*} \dfrac{\varphi(x) - \varphi(x^*)}{x - x^*}$. According to Lagrange's formula,

$$g(g(x)) - g(x) = g'(\xi)(g(x) - x), \qquad \xi \in (x, g(x)),$$

and if $x \to x^*$, then $g(x) \to g(x^*) = x^*$, hence $\xi \to x^*$. Therefore

$$\begin{aligned} \varphi(x) &= \frac{x g(g(x)) - x g(x) + x g(x) - (g(x))^2}{g(g(x)) - g(x) - (g(x) - x)} = \\ &= \frac{x g'(\xi)(g(x) - x) - g(x)(g(x) - x)}{(g'(\xi) - 1)(g(x) - x)} = \\ &= \frac{x g'(\xi) - g(x)}{g'(\xi) - 1}, \end{aligned}$$

and to obtain the last equality, we use the fact $g(x) \neq 0$ for $x \neq x^*$, which holds if x is in a small enough neighbourhood of x^* taking into account $g'(x^*) \neq 0$. Using the expansion $g(x) = g(x^*) + g'(\xi_1)(x - x^*)$, $\xi_1 \in (x, x^*)$ where we also substitute $g(x^*)$ with x^*, we get

$$\begin{aligned}\varphi(x) - \varphi(x^*) &= \frac{xg'(\xi) - g(x)}{g'(\xi) - 1} - x^* = \\ &= \frac{xg'(\xi) - x^* - g'(\xi_1)(x - x^*) - x^*g'(\xi) + x^*}{g'(\xi) - 1} = \\ &= \frac{(g'(\xi) - g'(\xi_1))(x - x^*)}{g'(\xi) - 1}.\end{aligned}$$

Now

$$\lim_{x \to x^*} \frac{\varphi(x) - \varphi(x^*)}{x - x^*} = \lim_{x \to x^*} \frac{g'(\xi) - g'(\xi_1)}{g'(\xi) - 1} = \frac{g'(x^*) - g'(x^*)}{g'(x^*) - 1} = 0,$$

which means that $\varphi'(x^*) = 0$ and under such assumptions Steffensen's method converges faster than any geometric progression.

Exercise 19. Show, using exercise 10, that if g is smooth enough, then, under the assumptions done above, Steffensen's method has quadratic convergence.

3.4 Müller's method

Consider the equation $f(x) = 0$. Let x_0, x_1, x_2 be given initial values that are pairwise different, i.e., $x_0 \neq x_1$, $x_1 \neq x_2$, $x_2 \neq x_0$. There exists exactly one polynomial $P(x) = c_0 + c_1 x + c_2 x^2$, which satisfies the conditions $P(x_i) = f(x_i)$, $i = 0, 1, 2$, or, its graph, a quadratic parabola, goes through the points $(x_i, f(x_i))$, $i = 0, 1, 2$. We see it by checking, that the system $c_0 + c_1 x_i + c_2 x_i^2 = f(x_i)$, $i = 0, 1, 2$, to determine the coefficients c_0, c_1, c_2 has non-zero determinant. Denote this polynomial P_{012} and solve the quadratic equation $P_{012}(x) = 0$. Let that solution, which is closer to the number x_2, be approximation x_3. We repeat this process with approximations x_1, x_2, x_3, by finding the polynomial P_{123} and solving the quadratic equation $P_{123}(x) = 0$ to determine x_4 and so on. Such process of finding the sequence x_n is called *Müller's method* or *method of parabolas*. For comparison, recall that in the secant method a straight line (the graph of a first degree polynomial) was put through two points on the graph of function f.

It can be proved, that if the equation $f(x) = 0$ has a solution having multiplicity one, then the rate of convergence of Müller's method can be described by the estimate $|x_{n+1} - x^*| \leqslant \text{const}\,|x_n - x^*|^{1.84\ldots}$; if the solution has multiplicity 2, then $|x_{n+1} - x^*| \leqslant \text{const}\,|x_n - x^*|^{1.23\ldots}$.

3.5 Higher order iteration methods

Consider the solution of the equation $f(x) = 0$ with some iteration method. Let x_n be found. Recall that in Newton's method we expanded

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + R_1(x),$$

then we dropped the remainder R_1 and solving the equation $f(x_n) + f'(x_n)(x - x_n) = 0$ we obtained

$$x - x_n = -\frac{f(x_n)}{f'(x_n)} \quad \text{and} \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Take a longer expansion

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(x_n)}{2}(x - x_n)^2 + R_2(x),$$

abandon the remainder R_2 and obtain the equation

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

Here we have the following possibilities:

1) solve the quadratic equation and take as x_{n+1} its solution, which is closer to the approximation x_n.

   Exercise 20. Assuming, that the solution has multiplicity one, derive the algorithm for finding the approximation x_{n+1} (determine, which root of the quadratic equation must be taken), if x_n is close enough to the solution.

2) in the quadratic equation, substitute $(x - x_n)^2 = \left(\frac{f(x_n)}{f'(x_n)}\right)^2$, which is obtained from Newton's method. From the remaining linear equation, we obtain

   $$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f''(x_n)}{2f'(x_n)}\left(\frac{f(x_n)}{f'(x_n)}\right)^2$$

   as a solution. This method is called *Euler–Chebyshev's method*.

3) in the quadratic equation, replace the quadratic term as $(x-x_n)^2 = -\frac{f(x_n)}{f'(x_n)}(x-x_n)$, i.e., only one factor in the term $(x-x_n)^2$ is replaced according to Newton's method. The solution of the obtained linear equation is

$$x_{n+1} = x_n - \frac{f(x_n)f'(x_n)}{(f'(x_n))^2 - \frac{1}{2}f''(x_n)f(x_n)}.$$

Such a method is called *Halley's method*.

Exercise 21. Prove that both, Euler–Chebyshev's and Halley's methods have cubic convergence, if the solution has multiplicity one and f'' satisfies Lipschitz condition.

In practice these methods are almost not used, because they need the calculation of second derivative and it is usually impossible to do, e.g., usually it cannot be found from data obtained from an experiment. In addition if $\varepsilon_n = |x_n - x^*|$, then at cubic convergence from $\varepsilon_n \sim 10^{-1}$ we have $\varepsilon_{n+1} \sim 10^{-3}$, $\varepsilon_{n+2} \sim 10^{-9}$, but at quadratic convergence we have $\varepsilon_{n+1} \sim 10^{-2}$, $\varepsilon_{n+2} \sim 10^{-4}$, $\varepsilon_{n+3} \sim 10^{-8}$, i.e., usually the necessary accuracy can be obtained at quadratic convergence by adding only one more step, keeping in mind the fact, that before achieving the accuracy $\varepsilon_n \sim 10^{-1}$ cubic convergence has no advantage in rate of convergence.

II Solving systems of equations

Consider the system of equations

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0, \\ \ldots\ldots\ldots\ldots\ldots \\ f_n(x_1, \ldots, x_n) = 0, \end{cases}$$

where $f_1, \ldots, f_n$ are given functions, $x_1, \ldots, x_n$ are unknown numbers. Denote $x = (x_1, \ldots, x_n)$, $F(x) = (f_1(x), \ldots, f_n(x))$, then we may write the given system as

$$F(x) = 0,$$

where $F\colon \mathbb{R}^n \to \mathbb{R}^n$ or $F\colon \mathcal{D} \to \mathbb{R}^n$, $\mathcal{D} \subset \mathbb{R}^n$. Consider also the system $x = G(x)$, which, written by the equations, is

$$\begin{cases} x_1 = g_1(x_1, \ldots, x_n), \\ \ldots\ldots\ldots\ldots\ldots \\ x_n = g_n(x_1, \ldots, x_n), \end{cases}$$

i.e., here $G(x) = (g_1(x), \ldots, g_n(x))$. The solution of such systems is a vector $x^* \in \mathbb{R}^n$, where $F(x^*) = 0$ or $x^* = G(x^*)$, which means, that x^* satisfies every equation in the system.

In an iteration method a sequence of vectors x^m, $m = 0, 1, \ldots$, $x^m = (x_1^m, \ldots, x_n^m) \in \mathbb{R}^n$, is found. Consider some norm $\|\cdot\|$ in the space $\mathbb{R}^n$. We say that $x^m \to x^*$, if $\|x^m - x^*\| \to 0$ as $m \to \infty$. Actually, $\|x^m - x^*\| \to 0$ is equivalent to $x_i^m \to x_i^*$, $i = 1, \ldots, n$, as $m \to \infty$, regardless of the norm we use. Most common norms in the space $\mathbb{R}^n$ are

$$\begin{aligned} \|x\|_2 &= (x_1^2 + \ldots + x_n^2)^{\frac{1}{2}}, \\ \|x\|_1 &= |x_1| + \ldots + |x_n|, \\ \|x\|_\infty &= \max_{1 \leqslant i \leqslant n} |x_i|, \end{aligned}$$

also $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$, $1 \leqslant p < \infty$. With it, for every $x \in \mathbb{R}^n$, $\lim\limits_{p\to\infty} \|x\|_p = \|x\|_\infty$. In the case $n = 1$ every considered norm is $|x|$.

§1. Ordinary iteration method

Consider the system $x = G(x)$. In the ordinary iteration method an initial value $x^0 \in \mathbb{R}^n$ is needed, next approximations are found as $x^{m+1} = G(x^m)$, $m = 0, 1, \ldots$, which is coordinatewise

$$\begin{cases} x_1^{m+1} = g_1(x_1^m, \ldots, x_n^m), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{m+1} = g_n(x_1^m, \ldots, x_n^m). \end{cases}$$

Let $\|\cdot\|$ be a norm in space $\mathbb{R}^n$. A *closed ball* is a set $B = \overline{B}(a, r) = \{x : \|x - a\| \leqslant r\}$, the number r is the *radius* of the ball, $a \in \mathbb{R}^n$ is the *centre* of the ball. In the case $n = 1$, $\{x : |x - a| \leqslant r\} = [a - r, a + r]$ is a closed interval.

Theorem 7 (Convergence theorem for the ordinary iteration method)**.** *If 1) $G\colon B \to B$ and 2) $\exists q < 1$ such, that $\|G(x) - G(y)\| \leqslant q\|x - y\|$ for every $x, y \in B$, $x \neq y$, then the equation $x = G(x)$ has exactly one solution x^* in the ball B, for every initial value $x^0 \in B$ the ordinary iteration method converges to this solution ($x^m \to x^*$) with the estimate*

$$\|x^m - x^*\| \leqslant \frac{q^m}{1 - q}\|x^0 - x^1\|.$$

The proof of this theorem is absolutely similar to the proof in the case $n = 1$, if we replace $|\cdot|$ by the norm $\|\cdot\|$.

Additional results

1. Instead of the assumptions made in the theorem, we may also assume that $\|G(x) - G(y)\| \leqslant q\|x - y\|$ for every $x, y \in \mathbb{R}^n$, $x \neq y$ (of course, then $G\colon \mathbb{R}^n \to \mathbb{R}^n$). For justification, note that, in this case everywhere in the proof of the theorem we have to replace the ball B by space $\mathbb{R}^n$.

2. Assume that G is a contraction 2) as in the theorem, but instead of $G\colon B \to B$, we pose the condition $\|G(a) - a\| \leqslant (1-q)r$ (this is a restriction to a shift of the centre a of the ball B). For justification let us verify that under these assumptions $G\colon B \to B$. If $x \in B$, i.e., $\|x - a\| \leqslant r$, then $\|G(x) - a\| \leqslant \|G(x) - G(a)\| + \|G(a) - a\| \leqslant q\|x - a\| + (1-q)r \leqslant r$, hence $G(x) \in B$.

In the following we shall try to find sufficient conditions for G to be a contraction, keeping in mind the conditions obtained for the case $n = 1$ using the derivative. The differentiability of function G (we do not give here the definition of it) is equivalent to the differentiability of the functions g_i, $i = 1, \ldots, n$ (as the differentiability of a function of n variables). Then the derivative of the function G is

$$G'(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \ldots & \frac{\partial g_1}{\partial x_n}(x) \\ \ldots\ldots\ldots\ldots\ldots\ldots & & \\ \frac{\partial g_n}{\partial x_1}(x) & \ldots & \frac{\partial g_n}{\partial x_n}(x) \end{pmatrix}.$$

Note that the differentiability of functions g_i is not equivalent to the existence of partial derivatives $\frac{\partial g_i}{\partial x_j}$, $i, j = 1, \ldots, n$. In the case of $n = 1$, we used Lagrange's formula, but here, when $n \geqslant 2$, the equality $G(x) - G(y) = G'(\xi)(x - y)$ does not hold, but *Lagrange's mean value estimate*

$$\|G(x) - G(y)\| \leqslant \sup_{0<\lambda<1} \|G'(\lambda x + (1-\lambda)y)\| \|x - y\|$$

holds instead. From this, we obtain the general result: if $\|G'(x)\| \leqslant q < 1$ for every $x \in B$, then G is a contraction in the ball B. For justification, note that, if $x, y \in B$, $x \neq y$, then for every $\lambda \in (0,1)$ we have $\lambda x + (1-\lambda)y \in B$:

$$\begin{aligned} x, y \in B \implies & \|x - a\| \leqslant r,\ \|y - a\| \leqslant r \implies \\ \implies & \|\lambda x + (1-\lambda)y - a\| = \|\lambda(x - a) + (1-\lambda)(y - a)\| \leqslant \\ & \leqslant \lambda\|x - a\| + (1-\lambda)\|y - a\| \leqslant \lambda r + (1-\lambda)r = r, \end{aligned}$$

i.e.,

$$\lambda x + (1-\lambda)y \in B.$$

To check the condition $\|G'(x)\| \leqslant q < 1$, we have to explain, what is the norm of a matrix and how to find it.

If A is a matrix, then we define $\|A\| = \sup_{\|x\| \leqslant 1} \|Ax\|$. As we can see, the norm of a vector has been used in two places, therefore, e.g., we can

consider $\|A\|_{p\to q} = \sup\limits_{\|x\|_p\leqslant 1} \|Ax\|_q$, where $1 \leqslant p, q \leqslant \infty$. If $A = (a_{ij})$, then

$$\|A\|_{\infty\to\infty} = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|,$$

$$\|A\|_{1\to 1} = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|,$$

$$\|A\|_{2\to 2} \leqslant \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}.$$

According to the statements above, we obtain the following results:

1) if $\max\limits_{1\leqslant i\leqslant n} \sum\limits_{j=1}^{n} \left| \dfrac{\partial g_i}{\partial x_j}(x) \right| \leqslant q < 1 \quad \forall x \in B = \{x\colon \|x-a\|_\infty \leqslant r\}$,
then G is a contraction in the ball B with respect to the ∞-norm;

2) if $\max\limits_{1\leqslant j\leqslant n} \sum\limits_{i=1}^{n} \left| \dfrac{\partial g_i}{\partial x_j}(x) \right| \leqslant q < 1 \quad \forall x \in B = \{x\colon \|x-a\|_1 \leqslant r\}$,
then G is a contraction in the ball B with respect to the 1-norm;

3) if $\sum\limits_{i,j=1}^{n} \left(\dfrac{\partial g_i}{\partial x_j}(x) \right)^2 \leqslant q < 1 \quad \forall x \in B = \{x\colon \|x-a\|_2 \leqslant r\}$,
then G is a contraction in the ball B with respect to the 2-norm.

§2. Seidel's method

Consider the system $x = G(x)$ or

$$\begin{cases} x_1 = g_1(x_1, \ldots, x_n), \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = g_n(x_1, \ldots, x_n). \end{cases}$$

Seidel's method requires one initial value x^0. Let us describe the transition $x^m \to x^{m+1}$. If $x^m = (x_1^m, \ldots, x_n^m)$, then $x^{m+1} = (x_1^{m+1}, \ldots, x_n^{m+1})$ will be found by the formulae

$$\begin{aligned} x_1^{m+1} &= g_1(x_1^m, x_2^m, \ldots, x_n^m), \\ x_2^{m+1} &= g_2(x_1^{m+1}, x_2^m, \ldots, x_n^m), \\ &\ldots\ldots \\ x_n^{m+1} &= g_n(x_1^{m+1}, \ldots, x_{n-1}^{m+1}, x_n^m). \end{aligned}$$

Thus, the components x_i^{m+1} of the approximation x^{m+1} are found in such a way that the components with a smaller index x_j^{m+1}, $j = 1, \dots, i-1$, having already been found, are used to find x_i^{m+1} by the corresponding equation.

Theorem 8. *Let $\|G(x) - G(y)\|_\infty \leqslant q\|x - y\|_\infty$ $(q < 1)$ for every $x, y \in B$, $x \neq y$, where $B = \{x\colon \|x - a\|_\infty \leqslant r\}$ and $\|G(a) - a\|_\infty \leqslant (1-q)r$. Then the system $x = G(x)$ has exactly one solution x^* in the ball B, Seidel's method converges to this solution for every $x^0 \in B$ with the estimate*

$$\|x^m - x^*\|_\infty \leqslant \frac{q^m}{1-q}\|x^0 - x^1\|_\infty.$$

Proof. We do not need to prove the existence and uniqueness of the solution, because in the theorem given in the previous paragraph, the existence and uniqueness of the solution has been shown under the assumptions of this theorem (let us point out that the existence and uniqueness of the solution is a property of the system and does not depend on the method used to solve it). To prove this theorem we now only need to prove the estimation of error.

Denote $x^{m,i} = (x_1^{m+1}, \dots, x_{i-1}^{m+1}, x_i^m, \dots, x_n^m)$, here $x^{m,1} = x^m$, $x^{m,n+1} = x^{m+1} = x^{m+1,1}$. Then the calculation formula for Seidel's method is

$$x_i^{m+1} = g_i(x^{m,i}), \quad i = 1, \dots, n.$$

We show that $x^{m,i} \in B$ for all m, i. For this it is sufficient to solve the following exercise which contains a more general result.

Exercise 22. Let $G\colon B \to B$, $B = \{x\colon \|x - a\|_\infty \leqslant r\}$ and $x^{m,i}$ be vectors appearing at solving the system $x = G(x)$ by Seidel's method. Prove that if $x^0 \in B$, then $x^{m,i} \in B$ for every m, i.

As the next step we formulate:

Exercise 23. Prove that under the assumptions of the theorem, the inequality $\|x^{m+1} - x^*\|_\infty \leqslant q\|x^m - x^*\|_\infty$ holds.

From the previous exercise we obtain $\|x^m - x^*\|_\infty \leqslant q\|x^{m-1} - x^*\|_\infty \leqslant \dots \leqslant q^m\|x^0 - x^*\|_\infty$, from which the convergence already follows. To obtain the estimate of error given in the theorem, we estimate

$$\|x^0 - x^*\|_\infty \leqslant \|x^0 - x^1\|_\infty + \|x^1 - x^*\|_\infty \leqslant \|x^0 - x^1\|_\infty + q\|x^0 - x^*\|_\infty,$$

from which

$$(1-q)\|x^0 - x^*\|_\infty \leqslant \|x^0 - x^1\|_\infty$$

or

$$\|x^0 - x^*\|_\infty \leqslant \tfrac{1}{1-q}\|x^0 - x^1\|_\infty$$

and we use it in the last estimate of $\|x^m - x^*\|_\infty$. ■

Note that in the theorem, we may assume $\|G(x) - G(y)\|_\infty \leqslant q\|x - y\|_\infty$ for every $x, y \in \mathbb{R}^n$, $x \neq y$, then all assertions still hold in the form, where the ball B has been replaced by the space $\mathbb{R}^n$.

It is natural to ask about the comparison of the ordinary iteration method and Seidel's method: which of them is the better? We will answer the question about the convergence later. From the aspect of carrying out the calculations, the ordinary iteration method has one advantage: the components of the approximation x^{m+1} can be calculated from the components of x^m at the same time with so called *parallel computations*, but the Seidel's method lacks this option.

§3. Newton's method

We are familiar with this method for equations, but it can also be used for solving systems of equations. Newton's method can be viewed as the ordinary iteration method, but its basic idea is different: the system is replaced by an approximate linear system.

Consider the system

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x_1, \ldots, x_n) = 0, \end{cases} \tag{2.1}$$

which can be written as $F(x) = 0$, where, as before, $x = (x_1, \ldots, x_n)$, $F(x) = (f_1(x), \ldots, f_n(x))$. Let F be differentiable, it is equivalent for functions $f_1, \ldots, f_n$ to be differentiable. For Newton's method one initial value x^0 is needed, the following approximations are found as

$$x^{m+1} = x^m - (F'(x^m))^{-1}F(x^m), \quad m = 0, 1, \ldots, \tag{2.2}$$

where

$$F'(x) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(x) & \ldots & \dfrac{\partial f_1}{\partial x_n}(x) \\ \ldots\ldots\ldots\ldots\ldots\ldots & & \\ \dfrac{\partial f_n}{\partial x_1}(x) & \ldots & \dfrac{\partial f_n}{\partial x_n}(x) \end{pmatrix},$$

in the formula (2.2) $(F'(x^m))^{-1}$ is the inverse matrix. Here the step of transition (2.2) is equivalent to solving the linear system

$$F'(x^m)x^{m+1} = F'(x^m)x^m - F(x^m)$$

(x^{m+1} is unknown, all the rest are known). The step can be performed, if $(F'(x^m))^{-1}$ exists, i.e., $\det F'(x^m) \neq 0$.

Theorem 9. *Let F be continuously differentiable in some neighbourhood of the solution x^* of system* (2.1) *and let $(F'(x^*))^{-1}$ exist. Then Newton's method is implementable in the neighbourhood of x^* and the method converges to the solution x^* faster than any geometric progression, if the initial value x^0 has been chosen close enough to x^*.*

Theorem 10. *If, additionally to the assumptions of theorem 9, we assume that in some neighbourhood of x^**

$$\|F'(x) - F'(y)\| \leqslant L\|x - y\|,$$

then

$$\|x^{m+1} - x^*\| \leqslant \text{const}\, \|x^m - x^*\|^2.$$

As can be seen, these theorems directly generalise the results obtained for the method for equations. We will not prove these theorems, because the basic scheme of these proofs is the same as for the method for equations, but to completely understand all the details, quite severe results from functional analysis must be used. We will only note that to satisfy Lipschitz condition in theorem 10 it suffices that the functions $\frac{\partial^2 f_i}{\partial x_j \partial x_k}$ are continuous for all i, j, k in a neighbourhood of the solution x^*.

§4. Solving linear systems of equations

4.1 About the importance of iteration methods

Consider a linear system

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{cases}$$

Use notation

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix},$$

then we may write the system as $Ax = b$.

This system has exactly one solution if and only if $\det A \neq 0$. If $\det A = 0$, then there may be no solution or there may be infinitely many solutions.

From linear algebra, Cramer's rule (solving linear systems using determinants) is known, which can be implemented if $\det A \neq 0$. A widely used method is Gaussian elimination, while applying this method, there is no difference if $\det A \neq 0$ or $\det A = 0$.

Now we find out, how many calculations are needed for the Gaussian elimination method. We count all the multiplications and divisions (these are long operations), we shall not take into account the additions and subtractions (short operations). Assume that $\det A \neq 0$.

Firstly the 1. equation is divided by the number a_{11}, for this n divisions are performed: $\frac{a_{12}}{a_{11}}, \dots, \frac{a_{1n}}{a_{11}}, \frac{b_1}{a_{11}}$. Then the new 1. equation has been multiplied by $-a_{21}$ and added to the 2. equation, the 1. equation has been multiplied by $-a_{31}$ and added to the 3. equation and so on, this takes $(n-1)n$ multiplications. As a result of these operations, the 1. column is in the form $(1\ 0\ \dots\ 0)^T$, n^2 multiplications and divisions have been performed. Then the same is done with system, which is smaller by one dimension, having the new equations with numbers from 2 to n. While repeating these calculations, a situation is achieved, where the main diagonal consists of numbers 1 and there are only zeros below that, the number of calculations performed is

$$n^2 + (n-1)^2 + \dots + 1 = \frac{n(n+1)(2n+1)}{6}.$$

Now the numbers above the main diagonal are eliminated, e.g., the last column needs $n-1$ multiplications, because only the free term is multiplied by the corresponding numbers. To eliminate the numbers above the main diagonal, we need

$$(n-1) + (n-2) + \dots + 1 = \frac{(n-1)n}{2}$$

multiplications. For the whole Gaussian elimination method, the number of calculations needed is

$$\frac{n(n+1)(2n+1)}{6} + \frac{(n-1)n}{2} \sim \frac{n^3}{3},$$

if we consider the process $n \to \infty$. The sign $\sim$ means here, that the quotient of its both sides converges to the number 1 as $n \to \infty$, about such a situation is said, that the *order* of the operations is $\frac{n^3}{3}$.

Now we look at the number of operations in Cramer's method, in which $n+1$ determinants must be calculated and then n divisions of these determinants must be performed. Every determinant can be calculated by transforming it to the triangular form and then multiplying the numbers on the main diagonal. For this method it takes $\sim \frac{n^3}{3}$ operations per determinant, as a whole it takes $\sim \frac{n^4}{4}$ for the method.

We note that to calculate n-th order determinant by multiplying elements of one row with minors takes at least $n!$ multiplications. But, e.g., $30! > 10^{30}$, which means, that this scheme for calculating determinants cannot be used in the case of a little larger systems of linear equations.

A step of an iteration method usually requires an application of a matrix to a vector, which has n^2 multiplications for a system of n equations. If for example $n = 1000$ (in practice this is not a very large system), then an iteration method that converges at some dozen steps is considerably faster than the Gaussian elimination method.

In addition to the advantage in the quantity of calculations, there is another great reason, why iteration methods are needed for solving linear systems of equations. It is the existence of rounding errors, which are inevitable at using computers. Gaussian elimination and especially Cramer's method are sensitive to rounding errors, but iteration methods eliminate the influence of rounding errors made in previous steps, because nothing is lost, if due to rounding errors one extra step of iteration is made, compared to the number of steps at calculation with exact numbers.

4.2 Ordinary iteration method, conditions of convergence

Consider the system

$$\begin{cases} x_1 = b_{11}x_1 + \ldots + b_{1n}x_n + b_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = b_{n1}x_1 + \ldots + b_{nn}x_n + b_n, \end{cases}$$

or $x = Bx + b$, where $x = (x_j)$, $B = (b_{ij})$, $b = (b_i)$ are corresponding vectors and the matrix similar to the ones used in the previous section. The ordinary iteration method requires one initial value x^0, further we find

$$x^{m+1} = Bx^m + b, \quad m = 0, 1, \ldots .$$

The ordinary iteration method for solving linear system is a special case of the more general method which we considered earlier, here $G(x) = Bx + b$. We know a sufficient condition for convergence: for some $q < 1$ we have $\|G(x) - G(y)\| \leqslant q\|x - y\|$ $\forall x, y \in \mathbb{R}^n$, $x \neq y$. Then $G(x) - G(y) = Bx + b - (By + b) = B(x - y)$, therefore

$$\begin{aligned} \|G(x) - G(y)\| \leqslant q\|x - y\| \quad \forall x, y \in \mathbb{R}^n \iff \\ \|B(x - y)\| \leqslant q\|x - y\| \quad \forall x, y \in \mathbb{R}^n \iff \\ \|Bx\| \leqslant q\|x\| \quad \forall x \in \mathbb{R}^n \iff \\ \|B\| \leqslant q, \end{aligned}$$

where the last equivalence is an elementary result from functional analysis. With this we have obtained the following result.

Proposition 11. *If $\|B\| < 1$, then the ordinary iteration method converges for any initial value to the unique solution of the system $x = Bx + b$.*

We showed earlier how to find the norms of matrices corresponding to specific norm in space $\mathbb{R}^n$. Considering this, we have shown that the following theorem holds.

Theorem 12. *If $\max\limits_{1\leqslant i\leqslant n} \sum\limits_{j=1}^{n} |b_{ij}| < 1$ or $\max\limits_{1\leqslant j\leqslant n} \sum\limits_{i=1}^{n} |b_{ij}| < 1$ or $\sum\limits_{i,j=1} b_{ij}^2 < 1$, then the ordinary iteration method converges for any initial value to the unique solution of the system $x = Bx + b$.*

Let us point out, that the given conditions guarantee the unique solvability of the system.

Consider a $n \times n$ matrix A. Number λ is called an *eigenvalue* of matrix A, if there exists a vector $x \neq 0$ such, that $Ax = \lambda x$. Relying on linear algebra, it can be directly verified, that λ is an eigenvalue of A if and only if $\det(A - \lambda I) = 0$, i.e.,

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

This equation is called the *characteristic equation* of the matrix A, its degree is n and, as an algebraic equation, it has, counted with multiplicity, exactly n solutions. The set of all eigenvalues of matrix A is called its *spectrum* and is denoted $\sigma(A)$.

Theorem 13. *The ordinary iteration method for solving the system $x = Bx + b$ converges to the unique solution for any initial value if and only if the moduli of all eigenvalues of matrix B are smaller than 1 ($\lambda \in \sigma(B) \implies |\lambda| < 1$).*

Geometrically this necessary and sufficient condition means, that the spectrum of matrix B is contained inside the unit circle of the complex plane.

We do not give the proof of this theorem here, it relies on fundamental results of the theory of matrices, e.g., Jordan or Schur normal form of a matrix could be used.

Exercise 24. Prove that if $\|B\| < 1$, then $|\lambda| < 1$ for every $\lambda \in \sigma(B)$.

So the claim given in exercise 24 links together the conditions given in theorems 12 and 13.

Although, there exists a necessary and sufficient condition to detect the convergence of the ordinary iteration method, it is not effective in practice, because the problem of finding the eigenvalues of a matrix is considerably more complicated than the problem of solving a system of linear equations. Therefore the sufficient conditions given in theorem 12 are important in practice.

4.3 Jacobi method

Consider the system $Ax = b$. Denote the matrix containing only the elements of the (main) diagonal of A by

$$D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}$$

and $R = A - D$, then $A = D + R$. We write the system $Ax = b$ equivalently as $(D + R)x = b$ or $Dx = -Rx + b$.

Assume that the (main) diagonal of matrix A is such that $a_{ii} \neq 0$, $i = 1, \ldots, n$. Then

$$D^{-1} = \begin{pmatrix} a_{11}^{-1} & & 0 \\ & \ddots & \\ 0 & & a_{nn}^{-1}, \end{pmatrix}$$

as the direct computation gives $DD^{-1} = I$. Therefore $Ax = b$ is equivalently representable as $x = -D^{-1}Rx + D^{-1}b$. Let us apply here the ordinary iteration method

$$x^{m+1} = -D^{-1}Rx^m + D^{-1}b.$$

Such two-part operation (expression of the diagonal + ordinary iteration method) is called *Jacobi method* for solving the system $Ax = b$.

Expressing the diagonal by the equations means that the system

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \qquad i = 1, \ldots, n,$$

or

$$a_{ii}x_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j = b_i, \qquad i = 1, \ldots, n,$$

is converted to the form

$$x_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(-\frac{a_{ij}}{a_{ii}}\right) x_j + \frac{b_i}{a_{ii}}, \qquad i = 1, \ldots, n,$$

i.e., $x = Bx + D^{-1}b$, where in the matrix $B = (b_{ij})$ the elements are in the form $b_{ij} = -\frac{a_{ij}}{a_{ii}}$, $j \neq i$, $b_{ii} = 0$.

Let us present some notions.

It is said that the (main) diangonal of matrix A is *dominant in rows* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \qquad i = 1, \ldots, n\,.$$

Such dominance means, that absolute values of diagonal elements are greater than the sum of absolute values of other elements in the same row. This requirement must be satisfied for each row. Note that sometimes such a situation may be achieved by changing the order of the equations in the system.

It is said that the (main) diagonal of matrix A is *dominant in columns* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}|, \qquad i = 1, \ldots, n\,.$$

Here this condition means, that absolute values of diagonal elements are greater than the sum of absolute values of other elements in the same column. Such a situation may sometimes be achieved by changing the order of the unknowns in the system.

If matrix A has diagonal dominance in rows or columns, then $a_{ii} \neq 0$, $i = 1, \ldots, n$, therefore Jacobi method is applicable to the system $Ax = b$.

Let matrix A have diagonal dominance in rows. Then, after expressing the diagonal, in the obtained matrix B

$$\|B\|_{\infty\to\infty} = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |b_{ij}| = \max_{1\leqslant i\leqslant n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} = \max_{1\leqslant i\leqslant n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1\\ j\neq i}} |a_{ij}| <$$

$$< \max_{1\leqslant i\leqslant n} \frac{|a_{ii}|}{|a_{ii}|} = 1.$$

Thus, if in A diagonal is dominant in rows, then Jacobi method converges. In this case, the condition $\|B\|_{\infty\to\infty} < 1$ also guarantees, that the system has exactly one solution or $\det A \neq 0$.

Exercise 25. Prove without using the theory of iteration methods, that if A has diagonal dominance in rows, then $\det A \neq 0$.

What can be said, if A has diagonal dominance in columns?

If A has diagonal dominance in columns, then the transposed matrix A^T has diagonal dominance in rows and therefore $\det A^T \neq 0$. As $\det A^T = \det A$, it holds $\det A \neq 0$.

Example. Consider the matrix $A = \begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix}$ having diagonal dominance in columns. Expressing the diagonal gives us the matrix $B = \begin{pmatrix} 0 & -2 \\ 0 & 0 \end{pmatrix}$. Estimate the norm of matrix B by using the previously considered traditional p-norms, $1 \leqslant p \leqslant \infty$, in space $\mathbb{R}^2$. Let $\overline{x} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, then $\|\overline{x}\|_p = 1, 1 \leqslant p \leqslant \infty$. At the same time $B\overline{x} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}$ and $\|B\overline{x}\|_p = 2$ for every p-norm. We get

$$\|B\|_{p\to p} = \sup_{\|x\|_p\leqslant 1} \|Bx\|_p \geqslant \|B\overline{x}\|_p = 2.$$

This example shows, that the diagonal dominance in columns of matrix A is not sufficient to assert, that for some p-norm it holds $\|B\|_{p\to p} < 1$.

Warning: we do not assert that there does not exist such a norm in space $\mathbb{R}^2$ which gives in the corresponding matrix norm that $\|B\| < 1$.

Exercise 26. Prove that if A has diagonal dominance in columns, then all the eigenvalues of the matrix $B = -D^{-1}R$ (obtained by expressing the diagonal of A) are smaller than 1 by modulus ($\lambda \in \sigma(B) \implies |\lambda| < 1$). *Hint:* prove that if $|\lambda| \geqslant 1$ and $\lambda \in \sigma(B)$, then $\det(\lambda D + R) = 0$, which is impossible if A has diagonal dominance.

According to the last exercise we may say, that if matrix A has diagonal dominance in columns, then Jacobi method converges.

4.4 Seidel's method

Consider the system $x = Bx + b$, which was written by equations in section 4.2. In Seidel's method an initial value is chosen and the transition from approximation $x^m = (a_1^m, \ldots, x_n^m)$ to the next approximation $x^{m+1} = (x_1^{m+1}, \ldots, x_n^{m+1})$ is carried out as follows:

$$\begin{cases} x_1^{m+1} = b_{11}x_1^m + b_{12}x_2^m + \ldots + b_{1n}x_n^m + b_1, \\ x_2^{m+1} = b_{21}x_1^{m+1} + b_{22}x_2^m + \ldots + b_{2n}x_n^m + b_2, \\ \ldots\ldots\ldots \\ x_1^{m+1} = b_{n1}x_1^{m+1} + \ldots + b_{n,n-1}x_{n-1}^{m+1} + b_{nn}x_n^m + b_n. \end{cases}$$

Let $B = L + D + U$, where D has been obtained from the diagonal of matrix A as in section 4.3,

$$L = \begin{pmatrix} 0 & \ldots\ldots\ldots & 0 \\ b_{21} & 0 \quad \ldots & 0 \\ \ldots\ldots\ldots\ldots\ldots & & \\ b_{n1} & \ldots \quad b_{n,n-1} & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & b_{12} & \ldots & b_{1n} \\ 0 & 0 & \ldots & b_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ 0 & \ldots\ldots\ldots & & 0 \end{pmatrix}$$

(matrix L contains the elements of A that are below its diagonal, matrix U contains the elements of A that are above its diagonal). Then we may present Seidel's method as

$$x^{m+1} = Lx^{m+1} + (D + U)x^m + b,$$

or

$$(I - L)x^{m+1} = (D + U)x^m + b.$$

Here

$$\det(I - L) = \det \begin{pmatrix} 1 & 0 & \ldots & 0 \\ & & & \\ (-b_{ij}) & & & 1 \end{pmatrix} = 1,$$

therefore $(I-L)^{-1}$ exists. The step of Seidel's method can be written as

$$x^{m+1} = (I-L)^{-1}(D+U)x^m + (I-L)^{-1}b,$$

which means, that Seidel's method can be considered as the ordinary iteration method for solving the transformed system $x = (I-L)^{-1}(D+U)x + (I-L)^{-1}b$. We have a necessary and sufficient condition for the ordinary iteration method to converge to the unique solution of the system. From this we obtain here, that Seidel's method converges if and only if the solutions of the equation

$$\det((I-L)^{-1}(D+U) - \lambda I) = 0$$

are smaller than 1 by modulus. Thereby

$$\begin{aligned}\det((I-L)^{-1}(D+U) - \lambda I) = 0 &\iff \det(D+U-\lambda(I-L)) = 0 \iff \\ &\iff \det(\lambda L + D + U - \lambda I) = 0.\end{aligned}$$

With this we proved the next theorem.

Theorem 14. *Seidel's method for solving the system $x = Bx + b$ converges if and only if all solutions of the equation* $\det(\lambda L + D + U - \lambda I) = 0$ *are smaller than 1 by modulus (taking in view that $B = L + D + U$).*

The equation given in the previous theorem is

$$\begin{vmatrix} b_{11}-\lambda & b_{12} & \dots & b_{1n} \\ \lambda b_{21} & b_{22}-\lambda & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda b_{n1} & \dots & \lambda b_{n,n-1} & b_{nn}-\lambda \end{vmatrix} = 0.$$

Recall that for the ordinary iteration method we had requirements for the solutions of the characteristic equation

$$\begin{vmatrix} b_{11}-\lambda & \dots & b_{1n} \\ \dots & \dots & \dots \\ b_{n1} & \dots & b_{nn}-\lambda \end{vmatrix} = 0.$$

Keeping in view these conditions, we may answer the question about the relation between the regions of convergence of the ordinary iteration method and Seidel's method.

Exercise 27. Prove that if in the system $x = Bx + b$ we take the 2×2 matrix B as

1) $B = \begin{pmatrix} 2 & -1.5 \\ 1.5 & -1 \end{pmatrix}$,
then the ordinary iteration method converges, but Seidel's method does not;

2) $B = \begin{pmatrix} 2 & -2 \\ 1 & -0.1 \end{pmatrix}$,
then Seidel's method converges, but the ordinary iteration method does not.

As for the ordinary iteration method, for Seidel's method there are also the conditions to check immediately not requiring the solving of equations similar to the characteristic equation. We present them in the next result.

Theorem 15. *If* $\max\limits_{1\leqslant i\leqslant n}\sum\limits_{j=1}^{n}|b_{ij}| < 1$ *or* $\max\limits_{1\leqslant j\leqslant n}\sum\limits_{i=1}^{n}|b_{ij}| < 1$, *then Seidel's method converges.*

Proof. Note that the first condition gives $\|B\|_{\infty\to\infty} < 1$, which means, that $G(x) = Bx + b$ is a contraction in ∞-norm in space $\mathbb{R}^n$ and we may apply the convergence theorem for general nonlinear systems. About the second condition we formulate an exercise.

Exercise 28. Prove that if $\max\limits_{1\leqslant j\leqslant n}\sum\limits_{i=1}^{n}|b_{ij}| < 1$, then the number λ with $|\lambda| \geqslant 1$, cannot be a solution of the equation $\det(\lambda L + D + U - \lambda I) = 0$.
Hint: show that if $|\lambda| \geqslant 1$, then the matrix $\lambda L + D + U - \lambda I$ has diagonal dominance in columns.

■

4.5 Gauss–Seidel method

Consider the system of equations $Ax = b$. Gauss–Seidel method is expressing the diagonal of the matrix A and solving the obtained system using Seidel's method.

Let us represent the matrix A as $A = R_L + D + R_U$ where the diagonal matrix D has the same meaning as before, R_L and R_U are the matrices containing the elements of A that are either below or above the diagonal, respectively. Assume that $a_{ii} \neq 0$, $i = 1, \ldots, n$, then the matrix D^{-1} exists.

Transform the system $Ax = b$ as

$$\begin{aligned} Ax = b &\iff (R_L + D + R_U)x = b \iff \\ &\iff Dx = -R_L x - R_U x + b \iff \\ &\iff x = -D^{-1}R_L x - D^{-1}R_U x + D^{-1}b. \end{aligned}$$

Direct calculations show that multiplication by the diagonal matrix D^{-1} does not change the form of matrices $-R_L$ and $-R_U$ (does not create any new non-zero elements), therefore all possibly non-zero elements of matrices $-D^{-1}R_L$ and $-D^{-1}R_U$ are either below or above the diagonal, respectively. Application of Seidel's method to the last system looks like

$$x^{m+1} = -D^{-1}R_L x^{m+1} - D^{-1}R_U x^m + D^{-1}b, \qquad m = 0, 1, \dots ,$$

starting from the initial value x^0.

Exercise 29. Show that the observed step of iteration can be written in the form

$$x^{m+1} = -(I + D^{-1}R_L)^{-1}D^{-1}R_U x^m + (I + D^{-1}R_L)^{-1}D^{-1}b$$

or

$$x^{m+1} = -(D + R_L)^{-1}R_U x^m + (D + R_L)^{-1}b.$$

Exercise 30. Prove that if the matrix A has diagonal dominance (in rows or columns) then Gauss–Seidel method converges.

4.6 Richardson's method

Consider the system $Ax = b$. In Richardson's method we begin with the initial value x^0 and calculate the following approximations as

$$x^{m+1} = x^m + \omega(Ax^m - b), \quad m = 0, 1, \dots ,$$

where $\omega \in \mathbb{C}$, $\omega \neq 0$, is fixed. This method may be viewed as the ordinary iteration method, which has been applied to the system

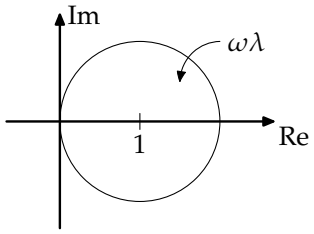
$$x = (I - \omega A)x + \omega b.$$

We know that convergence depends only on the matrix $B = I - \omega A$, the necessary and sufficient condition is that if $\lambda \in \sigma(B)$, then $|\lambda| < 1$.

Note, that

$$\begin{aligned}\lambda \in \sigma(A) &\iff \exists x \neq 0\colon Ax = \lambda x \iff \\ &\iff \exists x \neq 0\colon -\omega Ax = -\omega\lambda x \iff \\ &\iff \exists x \neq 0\colon x - \omega Ax = x - \omega\lambda x \iff \\ &\iff \exists x \neq 0\colon (I - \omega A)x = (1 - \omega\lambda)x.\end{aligned}$$

Therefore $\lambda \in \sigma(A)$ if and only if $1 - \omega\lambda \in \sigma(B)$. Thus, Richardson's method converges if and only if $|1 - \omega\lambda| < 1$, i.e., $|\omega\lambda - 1| < 1$ for every $\lambda \in \sigma(A)$. The set $\{z \in \mathbb{C} \mid |z - 1| < 1\}$ is a disc with centre 1 and radius 1, the necessary and sufficient condition says that all the numbers $\omega\lambda$, $\lambda \in \sigma(A)$, must be inside this disc.



It is natural to ask whether it is possible (or when it is possible) to find $\omega \in \mathbb{C}$ so that $|\omega\lambda - 1| < 1$ for every $\lambda \in \sigma(A)$? As $\omega \neq 0$, we have

$$|\omega\lambda - 1| < 1 \iff \left|\lambda - \frac{1}{\omega}\right| < \frac{1}{|\omega|}$$

and the necessary and sufficient condition of convergence is that there exists $\omega \in \mathbb{C}$ such, that

$$\sigma(A) \subset \left\{\lambda \in \mathbb{C} \,\middle|\, \left|\lambda - \frac{1}{\omega}\right| < \frac{1}{|\omega|}\right\},$$

which is a disc with centre $\frac{1}{\omega}$ and radius $\frac{1}{|\omega|}$ and therefore the boundary of this disc passes through the point 0. As $\frac{1}{\omega}$ can be an arbitrary non-zero complex number, we have obtained the following result.

Proposition 16. *The number $\omega \in \mathbb{C}$, $\omega \neq 0$, which is suitable for Richardson's method to converge, exists if and only if all the eigenvalues of the matrix A lie inside some circle that passes through the point 0.*

Corollary 17. *If $\sigma(A) \subset (0, \infty)$, then there exists an ω such that Richardson's method converges. The suitable ω is such that $0 < \omega < \frac{2}{\lambda_{\max}}$, where $\lambda_{\max}$ is the greatest eigenvalue of matrix A.*

Proof. Let $\sigma(A) \subset (0,\infty)$. If we choose ω such that $0 < \omega < \frac{2}{\lambda_{\max}}$ and $\lambda \in \sigma(A)$, then $0 < \omega\lambda < \frac{2\lambda}{\lambda_{\max}} \leqslant 2$, i.e., $0 < \omega\lambda < 2$ or $-1 < \omega\lambda - 1 < 1$, this yields that $|\omega\lambda - 1| < 1$. ■

Corollary 18. *If $\sigma(A) \subset (0,\infty)$ and $0 < \omega < \frac{2}{\|A\|}$, then Richardson's method converges.*

Proof. Earlier we actually used the fact that $|\lambda| \leqslant \|A\|$ for every $\lambda \in \sigma(A)$. Under the assumptions made, $\lambda > 0$ for every $\lambda \in \sigma(A)$ and therefore $\frac{2}{\|A\|} \leqslant \frac{2}{\lambda_{\max}}$. Hence if $0 < \omega < \frac{2}{\|A\|}$, then $0 < \omega < \frac{2}{\lambda_{\max}}$ and we only need to use corollary 17. ■

The importance of Corollary 18 is in the fact that finding the eigenvalues of matrix A may be considerably more complicated than calculation of its norm.

Corollary 19. *If matrix A is positive definite, then there exists $\omega > 0$ such that Richardson's method converges.*

Proof. Positive definiteness of matrix A means that $(Ax, x) > 0$ for every $x \neq 0$. But then we obtain from $Ax = \lambda x$, $x \neq 0$, that $(Ax, x) = \lambda(x, x) > 0$ and $(x, x) > 0$, therefore $\lambda > 0$, i.e., $\sigma(A) \subset (0,\infty)$. ■

Consider the case where none of the eigenvalues of the matrix A in the system $Ax = b$, are in any circle that passes through the zero point, but A is regular ($\det A \neq 0$). Then we can solve the equivalent system $A^TAx = A^Tb$, where A^T is the transposed matrix. These systems are equivalent, because $\det A^T = \det A \neq 0$ and A^T is regular or invertible. Matrix A^TA is positive definite, because if $x \neq 0$, then $Ax \neq 0$ and $(A^TAx, x) = (Ax, Ax) > 0$. According to Corollary 19 it is possible to find ω so, that Richardson's method

$$x^{m+1} = x^m - \omega(A^TAx^m - A^Tb)$$

converges. At it there is no need to calculate the product A^TA (it takes n^3 operations), but on each step, we may find Ax^m and then $A^T(Ax^m)$, which requires $2n^2$ operations on each step. The term A^Tb is calculated only once, we do not need to calculate it again on each step.

An extensive modern field of research is finding methods for solving the system $Ax = b$ by the iteration method

$$x^{m+1} = x^m - M_m(Ax^m - b),$$

where M_m is a sequence of $n \times n$ matrices. As a special case of those methods, we considered Richardson's method here, where $M_m = \omega I$, and the choice $M_m = \omega A^T$.

III Function approximation

§1. Interpolation problem

1.1 Problem formulation

Let us have given real numbers $x_0, \ldots, x_n$, where $x_i \neq x_j$ if $i \neq j$, and let the corresponding numbers be $f_0, \ldots, f_n$. It may be that $f_i = f(x_i)$ for a certain function f, but in practice the numbers f_i are results of measurement or data from experiments, which express some real dependence within limits of measurement errors. If there is a need for function values based on the argument values different from $x_0, \ldots, x_n$, then it is often proceeded as follows: find a function φ so that $\varphi(x_i) = f_i$, $i = 0, \ldots, n$, then use $\varphi(x)$, $x \neq x_i$. Function φ is called the *interpolant*, the points $x_0, \ldots, x_n$ are the *interpolation knots*, requirements $\varphi(x_i) = f_i$, $i = 0, \ldots, n$, the *interpolation conditions*. The interpolants are usually functions that are easy to operate with, e.g., polynomials, trigonometric polynomials, rational functions, splines (including piecewise polynomials). It is usually known whether the function f that is being interpolated, is continuous, continuously differentiable a number of times, or analytical.

In a more general interpolation problem it is necessary to find φ such that

$$\varphi^{(j)}(x_i) = f_{ij}, \qquad i = 0, \ldots, n, \qquad j = 0, \ldots, k_i,$$

where it is possible that, e.g., $f_{ij} = f^{(j)}(x_i)$ with an usually unknown function f. This is the *interpolation problem with multiple knots*, the number $k_i + 1$ is the multiplicity of the knot x_i. If there are no derivatives in the formulation, then every knot is simple.

1.2 Existence and uniqueness of the interpolant

Consider the situation where every knot $x_0, \ldots, x_n$ is simple. Assume that the so called *coordinate functions* $\psi_0, \ldots, \psi_m$ are given. Our aim is to find

the interpolant

$$\varphi_m = c_0\psi_0 + \ldots + c_m\psi_m,$$

where $c_0, \ldots, c_m$ will be determined from the interpolation conditions. This problem can be called a *linear interpolation problem*, because the interpolation conditions $\varphi_m(x_i) = f_i$, $i = 0, \ldots, n$, give a system of linear equations, namely

$$c_0\psi_0(x_i) + \ldots + c_m\psi_m(x_i) = f_i, \qquad i = 0, \ldots, n,$$

for finding the coefficients c_i. For the unique solvability of the system with arbitrary numbers f_i, it is necessary that $m = n$. Therefore we consider the system

$$c_0\psi_0(x_i) + \ldots + c_n\psi_n(x_i) = f_i,\ i = 0, \ldots, n.$$

This system is uniquely solvable for arbitrary numbers f_i, $i = 0, \ldots, n$, if and only if

$$\begin{vmatrix} \psi_0(x_0) & \ldots & \psi_n(x_0) \\ \ldots\ldots\ldots\ldots\ldots\ldots & & \\ \psi_0(x_n) & \ldots & \psi_n(x_n) \end{vmatrix} \neq 0.$$

Consider the case, where $\psi_j(x) = x^j$, therefore

$$\varphi(x) = c_0 + c_1x + \ldots + c_nx^n.$$

In this case we speak about *interpolation polynomial*. Then the corresponding determinant is the Vandermonde determinant

$$\begin{vmatrix} 1 & x_0 & {x_0}^2 & \ldots & {x_0}^n \\ 1 & x_1 & {x_1}^2 & \ldots & {x_1}^n \\ \ldots\ldots\ldots\ldots\ldots\ldots & & & & \\ 1 & x_n & {x_n}^2 & \ldots & {x_n}^n \end{vmatrix} = \prod_{i>j}(x_i - x_j) =$$

$$= (x_n - x_0)\ldots(x_n - x_{n-1})\cdot$$
$$\cdot (x_{n-1} - x_0)\ldots(x_{n-1} - x_{n-2})\cdot$$
$$\ldots\ldots$$
$$\cdot (x_1 - x_0) \neq 0.$$

With that we proved the following

Proposition 20. *For arbitrary numbers $f_0, \ldots, f_n$, there exists exactly one polynomial $P_n(x) = c_0 + c_1 x + \ldots + c_n x^n$ such that $P_n(x_i) = f_i$, $i = 0, \ldots, n$.*

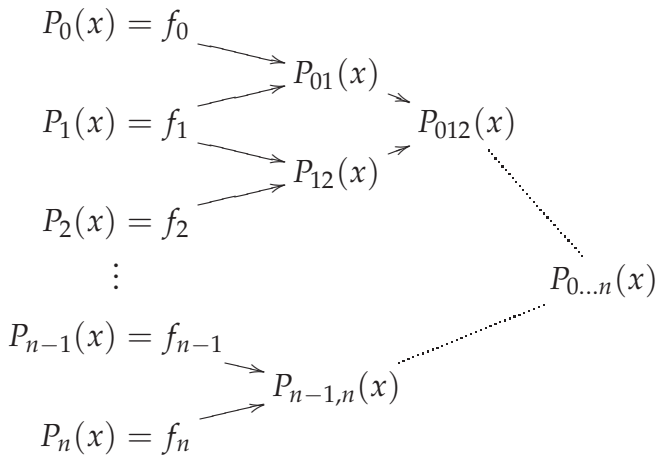
In the following we will look at different methods for finding such polynomials. One such method is described in the following exercise.

Exercise 31. Let it be given $x_0, \ldots, x_n$, $x_i \neq x_j$, if $i \neq j$, and $f_0, \ldots, f_n$. Let $P_{i,\ldots,i+k}$ be a polynomial, whose degree does not exceed k, and $P_{i,\ldots,i+k}(x_j) = f_j$, $j = i, \ldots, i+k$. Prove, that

$$P_{i,\ldots,i+k}(x) = \frac{(x - x_i)P_{i+1,\ldots,i+k}(x) + (x_{i+k} - x)P_{i,\ldots,i+k-1}(x)}{x_{i+k} - x_i},$$

if $P_j(x) = f_j$ for every x and j.

This exercise is the basis of Neville's scheme or algorithm for finding the interpolation polynomial stated in Proposition 20.



1.3 Lagrange fundamental polynomials

Let it be given $x_0, \ldots, x_n$, $x_i \neq x_j$, if $i \neq j$. Fix $i \in \{0, \ldots, n\}$. We already know that there exists exactly one polynomial ℓ_{ni} of degree at most n such that

$$\ell_{ni}(x_j) = \delta_{ij} = \begin{cases} 1, & \text{if } j = i, \\ 0, & \text{if } j \in \{0, \ldots, n\} \setminus \{i\}. \end{cases}$$

The knots $x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ are the zeroes of the polynomial ℓ_{ni}, therefore

$$\ell_{ni}(x) = a_i(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n),$$

where a_i is a constant, because if it was a polynomial of the first or higher degree, then the degree of ℓ_{ni} would exceed n. Besides the condition of zeroes we also have $\ell_{ni}(x_i) = 1$, i.e.,

$$a_i(x_i - x_0)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n) = 1,$$

from which we get

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

or

$$\ell_{ni}(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}.$$

Exercise 32. Let $\omega_n(x) = (x - x_0)\dots(x - x_n)$. Show that

$$\ell_{ni}(x) = \frac{\omega_n(x)}{(x - x_i)\omega_n'(x_i)}.$$

The polynomials ℓ_{ni}, $i = 0, \dots, n$, are called the *Lagrange fundamental polynomials*. They are uniquely defined by the knots $x_0, \dots, x_n$, and therefore it can be said that they are the Lagrange fundamental polynomials corresponding to the knots $x_0, \dots, x_n$.

Exercise 33. Prove that ℓ_{ni}, $i = 0, \dots, n$, are linearly independent.

1.4 Lagrange's interpolation formula

Let it be given knots $x_0, \dots, x_n$, $x_i \neq x_j$, if $i \neq j$, and the corresponding numbers $f_0, \dots, f_n$. We assert that then the polynomial

$$P_n(x) = \sum_{i=0}^{n} f_i \ell_{ni}(x) = f_0\ell_{n0}(x) + \dots + f_n\ell_{nn}(x),$$

is an interpolation polynomial, which satisfies the interpolation conditions $P_n(x_i) = f_i$, $i = 0, \dots, n$. To justify this we first notice that because ℓ_{ni} are polynomials of degree n, then the degree of their linear combination P_n does not exceed n. Additionally

$$P_n(x_j) = \sum_{i=0}^{n} f_i \ell_{ni}(x_j) = f_j, \qquad j = 0, \dots, n,$$

as $\ell_{ni}(x_j) = \delta_{ij}$.

By the previous expressions of ℓ_{ni} we can write

$$P_n(x) = \sum_{i=0}^{n} f_i \frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)} = \\ = \sum_{i=0}^{n} f_i \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j} = \sum_{i=0}^{n} f_i \frac{\omega_n(x)}{(x-x_i)\omega_n'(x_i)}.$$

This formula is called *Lagrange's interpolation formula*.

Additional remarks

1) The interpolation polynomial does not change if we change the order of the knots in Lagrange's interpolation formula.

   Note for justification that if we change the order of the knots, then the order of the summands in the Lagrange's formula changes, but the summands themselves do not. E.g., summand

   $$f_i \frac{\omega_n(x)}{(x-x_i)\omega_n'(x_i)}$$

   is defined by the knot set $x_0, \dots, x_n$, and the index i, because ω_n does not change if we change the order of the knots.

2) Lagrange fundamental polynomials $\ell_{n0}, \dots, \ell_{nn}$ form the basis of the set $\mathcal{P}_n$ of all polynomials of degree at most n, since $\dim \mathcal{P}_n = n+1$, and Lagrange fundamental polynomials are linearly independent. For every $P \in \mathcal{P}_n$

   $$P(x) = \sum_{i=0}^{n} P(x_i) l_{ni}(x),$$

   which comes from the fact that P and $\sum_{i=0}^{n} P(x_i) l_{ni}$ both satisfy the same interpolation conditions, their degree does not exceed n, but the interpolation polynomial is unique.

1.5 Divided differences

Let it be given the arguments $x_0, \ldots, x_n$, $x_i \neq x_j$ if $i \neq j$, and $f(x_0), \ldots, f(x_n)$. We define the *divided differences of the first order* as

$$f(x_i, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \qquad i \neq j.$$

Values $f(x_i)$ are called the *divided differences of the order* 0. The *divided differences of the second order* are defined as

$$f(x_i, x_j, x_k) = \frac{f(x_j, x_k) - f(x_i, x_j)}{x_k - x_i}, \quad i \neq j, \quad i \neq k, \quad j \neq k.$$

In general, the *divided differences of the kth order* are defined as

$$f(x_{i_0}, x_{i_1}, \ldots, x_{i_k}) = \frac{f(x_{i_1}, \ldots, x_{i_k}) - f(x_{i_0}, \ldots, x_{i_{k-1}})}{x_{i_k} - x_{i_0}}.$$

Remark. It is not important for the data $f(x_i)$ to be values of some function f, simply numbers $f_0, \ldots, f_n$ may be given. (For given $f_0, \ldots, f_n$ there always exists a function f for which $f_i = f(x_i)$, $i = 0, \ldots, n$). Besides the notation given above for divided differences, also the symbols f_{ij}, f_{ijk}, $f_{0\ldots k}$ are used.

Proposition 21. *It holds*

$$f(x_0, x_1, \ldots, x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n))} = \\ = \sum_{i=0}^{n} \frac{f(x_i)}{\prod\limits_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j)} = \sum_{i=0}^{n} \frac{f(x_i)}{w_n'(x_i)}. \tag{3.1}$$

Proof. For easier understanding, the equality (3.1) is written for knots $x_0, \ldots, x_n$, but it certainly holds for any set of pairwise different knots.

Let us prove the equality (3.1) with the use of induction on the number of knots.

For $n = 1$,

$$f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}.$$

Assume that the equality (3.1) holds for k knots, and let us show that it holds for $k+1$ knots. From the definiton and the assumption we get that

$$f(x_0, x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_k) - f(x_0, \ldots, x_{k-1})}{x_k - x_0} =$$
$$= \frac{1}{x_k - x_0}\left(\frac{f(x_1)}{\prod_{j=2}^{k}(x_1 - x_j)} + \cdots + \frac{f(x_k)}{\prod_{j=1}^{k-1}(x_k - x_j)} - \right.$$
$$\left. - \frac{f(x_0)}{\prod_{j=1}^{k-1}(x_0 - x_j)} - \cdots - \frac{f(x_{k-1})}{\prod_{j=0}^{k-2}(x_{k-1} - x_j)} \right).$$

We see that the coefficients of $f(x_0)$ and $f(x_k)$ are appropriate. Let $i \in \{1, \ldots, k-1\}$. Then the members containing $f(x_i)$ give

$$\frac{1}{x_k - x_0}\left(\frac{f(x_i)}{\prod_{\substack{j=1 \\ j\neq i}}^{k}(x_i - x_j)} - \frac{f(x_i)}{\prod_{\substack{j=0 \\ j\neq i}}^{k-1}(x_i - x_j)} \right) =$$
$$= \frac{f(x_i)}{\prod_{\substack{j=1 \\ j\neq i}}^{k-1}} \cdot \frac{1}{x_k - x_0}\left(\frac{1}{x_i - x_k} - \frac{1}{x_i - x_0} \right) =$$
$$= \frac{f(x_i)}{\prod_{\substack{j=0 \\ j\neq i}}^{k}(x_i - x_j)}.$$

■

Corollary 22.

$$(f_1 + f_2)(x_0, \ldots, x_n) = f_1(x_0, \ldots, x_n) + f_2(x_0, \ldots, x_n).$$

Corollary 23.

$$(cf)(x_0, \ldots, x_n) = cf(x_0, \ldots, x_n).$$

Corollary 24. *Divided differences are symmetric with respect to arguments.*

The first two corollaries mean that the divided differences are linear (additive and homogeneous) with respect to the functions to take the differences. For the justification of the third corollary note that if we change the order of the knots in the divided differences, then only the order of summands in the right hand side of the formula (3.1) changes, but the sum itself remains unchanged.

1.6 Newton's interpolation formula

Let it be given knots $x_0, \ldots, x_n$ and corresponding values $f(x_0), \ldots, f(x_n)$.

Proposition 25. *The polynomial of degree n which satisfies the interpolation conditions $P_n(x_i) = f(x_i)$, $i = 0, \ldots, n$, is*

$$P_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1) + \cdots + \\ + f(x_0, \ldots, x_n)(x - x_0) \ldots (x - x_{n-1}).$$

Proof. It is clear that the degree of P_n does not exceed n, hence it is only necessary to prove the validity of the interpolation conditions. We prove this by induction on n.

If $n = 0$ then $P_0(x) = f(x_0)$ for every x, therefore $P_0(x_0) = f(x_0)$. Assume now that the representation of the interpolating polynomial holds for $n - 1$, which means that for

$$P_{n-1}(x) = f(x_0) + \cdots + f(x_0, \ldots, x_{n-1})(x - x_0) \ldots (x - x_{n-2})$$

the conditions

$$P_{n-1}(x_i) = f(x_i),\ i = 0, \ldots, n - 1,$$

are satisfied. Consider the polynomial

$$P_n(x) = P_{n-1}(x) + f(x_0, \ldots, x_n)(x - x_0) \ldots (x - x_{n-1}).$$

Then

$$P_n(x_i) = P_{n-1}(x_i) = f(x_i), \qquad i = 0, \ldots, n - 1.$$

In addition,

$$P_n(x_n) = P_{n-1}(x_n) + (x_n - x_0)\dots(x_n - x_{n-1}) \sum_{i=0}^{n} \frac{f(x_i)}{\prod\limits_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j)} =$$

$$= P_{n-1}(x_n) + \sum_{i=0}^{n-1} \frac{(x_n - x_0)\dots(x_n - x_i)\dots(x_n - x_{n-1})}{\prod\limits_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j)} f(x_i) +$$

$$+ \frac{(x_n - x_0)\dots(x_n - x_{n-1}}{(x_n - x_0)\dots(x_n - x_{n-1})} f(x_n) = f(x_n),$$

because

$$\frac{(x_n - x_0)\dots(x_n - x_i)\dots(x_n - x_{n-1})}{\prod\limits_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j)} = -\prod_{\substack{j=0 \\ j\neq i}}^{n-1} \frac{x_n - x_j}{x_i - x_j} = -\ell_{n-1,i}(x_n),$$

and from the Lagrange formula

$$-\sum_{i=0}^{n-1} \ell_{n-1,i}(x_n) f(x_i) = -P_{n-1}(x_n).$$

■

The calculations of divided differences in the Newton's interpolation formula could be implemented using the triangle scheme

$$\begin{array}{cc|cc|c} x_0 & f(x_0) & & & \\ & & f(x_0,x_1) & & \\ x_1 & f(x_1) & & f(x_0,x_1,x_2) & \\ & & f(x_1,x_2) & & \\ x_2 & f(x_2) & & & \\ \vdots & & & & f(x_0,x_1,\dots,x_n), \\ x_{n-1} & f(x_{n-1}) & & & \\ & & f(x_{n-1},x_n) & & \\ x_n & f(x_n) & & & \end{array}$$

from which the interpolation formula only uses the first elements of each column.

Exercise 34. Find how many multiplications and divisions are necessary to calculate $P_n(x)$, $x \neq x_i$, $i = 0, \dots, n$, when using the Lagrange interpolation formula. Same question for the Newton's formula.

1.7 Remainder term of interpolation formula

Denote $R_n(x) = f(x) - P_n(x)$, where f is the function to interpolate, and P_n is the interpolation polynomial. Hence $f(x) = P_n(x) + R_n(x)$, where the term R_n is the *remainder term*. We see that $R_n(x_i) = f(x_i) - P_n(x_i) = 0$, $i = 0, \ldots, n$. We cannot say anything else about the remainder term without any additional assumptions, it may be arbitrary function whose value in the knots is 0.

Assume that $x_0, \ldots, x_n \in [a, b]$ and $f \in C^{n+1}[a, b]$. Fix $x \in [a, b]$, consider at first the case $x \neq x_i$, $i = 0, \ldots, n$. Consider the auxiliary function $\varphi(z) = f(z) - P_n(z) - K\omega_n(z) = R_n(z) - K\omega_n(z)$, where K is a constant and $\omega_n(z) = (z - x_0) \ldots (z - x_n)$. We see that $\varphi(x_i) = 0$, $i = 0, \ldots, n$ for every K. Let K be such that $\varphi(x) = 0$, which means that $R_n(x) - K\omega_n(x) = 0$ or $K = \dfrac{R_n(x)}{\omega_n(x)}$. The function φ has $n + 2$ different zeroes in the interval $[a, b]$. According to Rolle's theorem, the function φ' has at least $n + 1$ different zeroes in the interval (a, b), and they are located between the zeroes of function φ. Analogically we get that φ'' has n different zeroes until we finally have that $\varphi^{(n+1)}$ has at least one zero ξ in the interval (a, b), $\varphi^{(n+1)}(\xi) = 0$. Now we differentiate the function φ $n + 1$ times and we get $\varphi^{(n+1)}(z) = f^{(n+1)}(z) - K(n + 1)!$, from which $\varphi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \dfrac{R_n(x)}{\omega_n(x)}(n + 1)! = 0$, or $R_n(x) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x)$. Of course we note that ξ depends on the chosen number x, which means that it is more in details to write $\xi(x)$ instead of ξ. If $x = x_i$ then $\omega_n(x) = 0$ and $R_n(x) = 0$, and hence in this case the same formula for $R_n(x)$ is valid, where ξ can be taken arbitrary.

It is natural to ask whether the function $x \to f^{(n+1)}(\xi(x))$ is continuous? Is it a certain number of times continuously differentiable?

Exercise 35. Prove that $\lim\limits_{x \to x_i} f^{(n+1)}(\xi(x))$ exists. Suggestion: use L'Hospital's rule.

Exercise 36. Prove that there is a $\xi_i \in [a, b]$ such that

$$f^{(n+1)}(\xi_i) = \lim_{x \to x_i} f^{(n+1)}(\xi(x)).$$

Exercise 37. Prove that the function $x \to f^{(n+1)}(\xi(x))$ is continuously differentiable.

Since $f^{(n+1)}$ is continuous, it is also bounded in the interval $[a,b]$. Let $M_{n+1} = \max\limits_{a\leqslant x\leqslant b} |f^{(n+1)}(x)|$. Then $|R_n(x)| \leqslant \dfrac{M_{n+1}}{(n+1)!}|\omega_n(x)|$, which allows to estimate the accuracy of interpolation.

Exercise 38. Prove that if $f \in C^{n+1}[a,b]$ and $x_i \to x$, $i = 1,\dots,n$, then the limit of the interpolation formula gives the Taylor's formula

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \dots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \\ + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}.$$

Let $x \neq x_i$, $i = 0,\dots,n$. From the formula (3.1) we get that

$$f(x,x_0,\dots,x_n) = \frac{f(x)}{(x-x_0)\dots(x-x_n)} + \\ + \frac{f(x_0)}{(x_0-x)(x_0-x_1)\dots(x_0-x_n)} + \dots + \frac{f(x_n)}{(x_n-x)(x_n-x_0)\dots(x_n-x_{n-1})}.$$

Having expressed $f(x)$ from this equality, we get

$$f(x) = f(x_0)\frac{(x-x_1)\dots(x-x_n)}{(x_0-x_1)\dots(x_0-x_n)} + \dots + \\ + f(x_n)\frac{(x-x_0)\dots(x-x_{n-1})}{(x_n-x_0)\dots(x_n-x_{n-1})} + \\ + f(x,x_0,\dots,x_n)\omega_n(x).$$

In the part with fractions we recognize the interpolation polynomial in Lagrange form. Therefore the rest is the remainder term, i.e.,

$$R_n(x) = f(x,x_0,\dots,x_n)\omega_n(x).$$

Let us point out that the remainder term in such a form does not require the function f to be smooth or continuous. Under the assumption $f \in C^{n+1}[a,b]$ we have already obtained above that

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x).$$

The comparison of the two remainder terms gives us that

$$f(x,x_0,\dots,x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \qquad \xi \in (a,b).$$

We formulate this result as follows.

Theorem 26 (representation theorem for divided differences). *If $f \in C^n[a,b]$ and $x_0, \dots, x_n \in [a,b]$, then*

$$f(x_0, \dots, x_n) = \frac{f^{(n)}(\xi)}{n!}, \qquad \xi \in (a,b).$$

For particular case $n = 1$ we get well known Lagrange formula

$$f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f'(\xi).$$

.

1.8 On convergence of interpolation process

Let it be given an interval $[a,b]$ and a knot system $x_{ni} \in [a,b]$, $n = 0, 1, \dots$, $i = 0, \dots, n$, which we write in triangular form

$$\begin{array}{l} x_{00} \\ x_{10}, x_{11} \\ x_{20}, x_{21}, x_{22} \\ \dots\dots\dots\dots\dots \end{array}$$

Consider a function $f\colon [a,b] \to \mathbb{R}$. For each n, construct the interpolation polynomial P_n having the degree not exceeding n, and satisfying conditions $P_n(x_{ni}) = f(x_{ni})$, $i = 0, \dots, n$. This gives us the sequence of interpolation polynomials P_n, $n = 0, 1, \dots$. It is natural to ask whether

$$\max_{a \leqslant x \leqslant b} |P_n(x) - f(x)| \to 0, \qquad \text{as } n \to \infty?$$

Theorem 27 (Faber, 1914). *For arbitrary knot system $x_{ni} \in [a,b]$, $n = 0, 1, \dots$, $i = 0, \dots, n$, there exists a function $f \in C[a,b]$ such that it does not take place $\max\limits_{a \leqslant x \leqslant b} |P_n(x) - f(x)| \to 0$, as $n \to \infty$. In fact, there is a function f for which $\max\limits_{a \leqslant x \leqslant b} |P_n(x) - f(x)| \to \infty$ as $n \to \infty$.*

Additional result. *For each knot system $\max\limits_{a \leqslant x \leqslant b} \sum\limits_{i=0}^{n} |\ell_{ni}(x)| \geqslant c \ln n$, where c is a positive constant.*

Basing on this result, let us analyse the influence of errors in data at interpolation. Assume that instead of exact values $f(x_{ni})$ we have values f_{ni} for which

$$|f_{ni} - f(x_{ni})| \leqslant \varepsilon.$$

With the help of f_{ni} construct the interpolation polynomials

$$\tilde{P}_n(x) = \sum_{i=0}^{n} f_{ni}\ell_{ni}(x),$$

which are, in general, different from the polynomials

$$P_n(x) = \sum_{i=0}^{n} f(x_{ni})\ell_{ni}(x).$$

Then

$$\begin{aligned}\max_{a\leqslant x\leqslant b}|\tilde{P}_n(x) - P_n(x)| &= \max_{a\leqslant x\leqslant b}|\sum_{i=0}^{n}(f_{ni} - f(x_{ni}))\ell_{ni}(x)| \leqslant \\ &\leqslant \max_{a\leqslant x\leqslant b}\sum_{i=0}^{n}|f_{ni} - f(x_{ni})||\ell_{ni}(x)| \leqslant \\ &\leqslant \varepsilon \max_{a\leqslant x\leqslant b}\sum_{i=0}^{n}|\ell_{ni}(x)| \to \infty, \qquad \text{as } n \to \infty.\end{aligned}$$

In our estimates the inequalities can be equalities because

$$\max_{a\leqslant x\leqslant b}\sum_{i=0}^{n}|\ell_{ni}(x)| = \sum_{i=0}^{n}|\ell_{ni}(x_0)|$$

for some $x_0 \in [a, b]$, and depending on the signs of $\ell_{ni}(x_0)$ it can happen that

$$f_{ni} - f(x_{ni}) = \pm\varepsilon$$

so that

$$\begin{aligned}\max_{a\leqslant x\leqslant b}|\sum_{i=0}^{n}(f_{ni} - f(x_{ni}))\ell_{ni}(x)| &\geqslant |\sum_{i=0}^{n}(f_{ni} - f(x_{ni}))\ell_{ni}(x_0)| = \\ &= \sum_{i=0}^{n}|f_{ni} - f(x_{ni})||\ell_{ni}(x_0)| = \varepsilon\sum_{i=0}^{n}|\ell_{ni}(x_0)|.\end{aligned}$$

From this discussion we conclude that polynomial interpolation is unstable with respect to the errors in data as the degree of polynomials increases.

1.9 Interpolation of functions of several variables

Let us consider the differences which do not appear at interpolation of one variable functions. These phenomena appear even when we interpolate two variable functions.

Let it be given pairwise distinct points $(x_0, y_0), \ldots, (x_n, y_n)$ in the plane, let us also have $f_0, \ldots, f_n$, which may be the values of some function f, i.e., $f_i = f(x_i, y_i)$, $i = 0, \ldots, n$. It is required to to find a polynomial P_m of degree at most m satisfying $P_m(x_i, y_i) = f_i$, $i = 0, \ldots, n$. The general form of polynomial P_m is

$$\begin{aligned} P_m(x,y) = {} & c_{00} + c_{10}x + c_{20}x^2 + \ldots + c_{m0}x^m + \\ & + c_{01}y + c_{11}xy + c_{21}x^2y + \ldots + c_{m-1,1}x^{m-1}y + \\ & \ldots \\ & + c_{0,m-1}y^{m-1} + c_{1,m-1}xy^{m-1} + \\ & + c_{0m}y^m. \end{aligned}$$

The interpolation conditions give a linear system to determine the coefficients c_{ij}. There are in total $(m+1) + m + \ldots + 1 = \dfrac{(m+1)(m+2)}{2}$ coefficients, and $n+1$ interpolation conditions or equations. For the unique solvability of the system for any f_i, it is necessary that $n+1 = \dfrac{(m+1)(m+2)}{2}$. If $m = 0$, then $n = 0$ (1 knot), if $m = 1$, then $n = 2$ (3 knots), if $m = 2$ then $n = 5$ (6 knots), if $m = 3$ then $n = 9$ (10 knots), and so on. Thus for the unique determination of the interpolation polynomial the number of knots cannot be arbitrary.

Next, examine the system's determinant. If $m = 1$ and $n = 2$, then for the unique solvability of the system

$$c_{00} + c_{10}x_i + c_{01}y_i = f_i, \qquad i = 0, 1, 2,$$

the necessary and sufficient condition is that

$$\begin{vmatrix} 1 & x_0 & y_0 \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix} \neq 0.$$

The condition

$$\begin{vmatrix} 1 & x_0 & y_0 \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix} = 0$$

is equivalent to the condition that the linear homogeneous system

$$\begin{pmatrix} 1 & x_0 & y_0 \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = 0$$

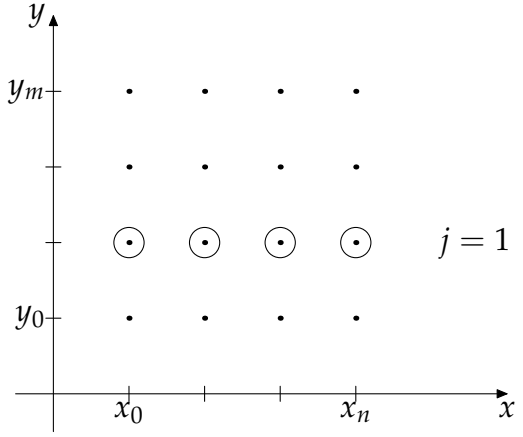
has a nontrivial solution: $|a_1| + |a_2| + |a_3| \neq 0$ or

$$a_1 + a_2 x_i + a_3 y_i = 0, \qquad i = 0, 1, 2, \qquad |a_1| + |a_2| + |a_3| \neq 0.$$

Geometrically this means that the points (x_i, y_i), $i = 0, 1, 2$, are on the straight line $a_1 + a_2 x + a_3 y = 0$. This discussion tells us that for the unique solvability of an interpolation problem with three knots (x_i, y_i), $i = 0, 1, 2$, it is necessary and sufficient for the three points (x_i, y_i) not to be on the same straight line. Analogically it is possible to show that for the unique solvability of an interpolation problem with six knots (x_i, y_i), $i = 0, \ldots, 5$, it is necessary and sufficient for the knots not to be on the same second order curve (ellipse, parabola, hyperbola, two straight lines), a ten knot problem's knots cannot be on the same third order curve and so on. In comparison to the one variable situation here we see another difference. Even though the number of knots is appropriate, they cannot be located arbitrarily.

The third problem is that it is not possible to get good expressions for a remainder term as it is for an one variable situation because Rolle's theorem does not apply here.

Let us now examine a different way of two variable interpolation with a special knot placement. Assume that it is a given rectangular grid of knots (x_i, y_j), $i = 0, \ldots, n$, $j = 0, \ldots, m$.



It is also given numbers f_{ij}, $i = 0, \ldots, n$, $j = 0, \ldots, m$. We solve the interpolation problem in the following way. Fix $j \in \{0, \ldots, m\}$ and find

one variable polynomial P_{nj} of degree at most n such that

$$P_{nj}(x_i) = f_{ij}, \qquad i = 0, \ldots, n.$$

Proceed so for every j and as a result we get the polynomials $P_{n0}, \ldots, P_{nm}$ of variable x. Then find the two variable polynomial $P_{nm} = P_{nm}(x, y)$ of degree at most m with respect to y such that

$$P_{nm}(x, y_j) = P_{nj}(x), \qquad j = 0, \ldots, m,$$

where for every x the different $P_{nj}(x)$ are used as the values of the function at interpolation. The obtained two variable polynomial $P_{nm}(x, y)$ is at most $n + m$ degree polynomial (for a fixed x the degree with respect to y does not exceed m and for a fixed y the degree with respect to x does not exceed n). At the same time $P_{nm}(x_i, y_j) = P_{nj}(x_i) = f_{ij}$ for all i and j. We cannot assert, though, that $P_{nm} = P_{nm}(x, y)$ is a polynomial of minimal degree satisfying interpolation conditions.

It is possible to interpolate in the other order, first by fixing $i \in \{0, \ldots, n\}$, then finding $\tilde{P}_{im}$ such that $\tilde{P}_{im}(y_j) = f_{ij}$, $j = 0, \ldots, m$. Doing so for every index i and then interpolating with respect to x we get a two variable polynomial $\tilde{P}_{nm} = \tilde{P}_{nm}(x, y)$ by using $\tilde{P}_{im}(y)$ as the function values, which means that

$$\tilde{P}_{nm}(x_i, y) = \tilde{P}_{im}(y), \qquad i = 0, \ldots, n.$$

Exercise 39. Prove that $\tilde{P}_{nm}(x, y) = P_{nm}(x, y)$. Suggestion: use the Lagrange's interpolation formula.

§2. Function approximation by least squares method

Let us recall some circumstances at interpolation with polynomials. In practice it is common to have experimental data $f_0, \ldots, f_n$ with errors. A lot of experiments or measurements are done to minimize the influence of random errors. But we know that if n increases then the sequence of interpolation polynomials P_n may not converge to the function f to interpolate. Recall also that enlarging n the influence of errors in data increases. To compensate these troubles the least squares method for function approximation is used.

2.1 Solving linear systems using least squares method

Consider the system

$$\begin{cases} a_{11}x_1 + \ldots + a_{1n}x_n = f_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + \ldots + a_{mn}x_n = f_n, \end{cases} \tag{3.2}$$

where, in general, $m \neq n$. E.g., it can happen that $m > n$.

Using the notation

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix}, \; x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \; f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix},$$

we write the system (3.2) as

$$Ax = f.$$

Generally, this system need not have a solution. For $x \in \mathbb{R}^n$ let

$$\begin{aligned} r_i(x) &= f_i - \sum_{j=1}^{n} a_{ij}x_j, \\ R(x) &= f - Ax = (r_1(x), \ldots, r_m(x))^T, \\ \|R(x)\|^2 = \|f - Ax\|^2 &= \sum_{i=1}^{m} r_i^2(x) = \sum_{i=1}^{m} \left(f_i - \sum_{j=1}^{n} a_{ij}x_j \right)^2 \end{aligned}$$

where $\|\cdot\|$ is Euclidean norm or 2-norm.

Definition. A vector $x \in \mathbb{R}^n$ is called the least squares solution of the system (3.2) if

$$\|R(x)\|^2 = \min_{y \in \mathbb{R}^n} \|R(y)\|^2.$$

We see that a least squares solution x is ordinary solution if and only if $\|R(x)\| = 0$.

The function $x \to \|R(x)\|^2 = g(x_1, \ldots, x_n)$ is a n variable differentiable function because it is a polynomial. If x is its minimum point, then $\dfrac{\partial g}{\partial x_k}(x) = 0, k = 1, \ldots, n$, which means that in the minimum point x

$$\frac{\partial g}{\partial x_k}(x) = \sum_{i=1}^{m} 2 \left(f_i - \sum_{j=1}^{n} a_{ij}x_j \right) (-a_{ik}) = 0, \qquad k = 1, \ldots, n,$$

i.e.

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}a_{ik}x_j = \sum_{i=1}^{m} a_{ik}f_i, \qquad k = 1, \ldots, n$$

or

$$\sum_{j=1}^{n}\left(\sum_{i=1}^{m} a_{ik}a_{ij}\right)x_j = \sum_{i=1}^{m} a_{ik}f_i, \qquad k = 1, \ldots, n.$$

This set of equalities can be written as

$$A^T Ax = A^T f, \tag{3.3}$$

where

$$A^T = \begin{pmatrix} a_{11} & \ldots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \ldots & a_{mn} \end{pmatrix}$$

is transposed matrix. System (3.2) is called the *normal system of equations*, its matrix $A^T A$ is a $n \times n$ matrix because A is a $m \times n$ matrix and A^T is a $n \times m$ matrix. The free term of system (3.2) is a vector having n components.

Above we saw that the least squares solution of system (3.2) is the solution of system (3.3). Now we will show that every solution of system (3.3) is a least squares solution of problem (3.2). Let x be a solution of system (3.3). Take an arbitrary $y \in \mathbb{R}^n$. Then

$$\begin{aligned} \|f - Ay\|^2 &= \|f - Ax + Ax - Ay\|^2 \\ &= (f - Ax + A(x-y), f - Ax + A(x-y)) = \\ &= \|f - Ax\|^2 + 2(A(x-y), f - Ax) + \|A(x-y)\|^2 = \\ &= \|f - Ax\|^2 + 2(x-y, A^T(f - Ax)) + \|A(x-y]\|^2 \geqslant \\ &\geqslant \|f - Ax\|^2, \end{aligned}$$

because $A^T(f - Ax) = 0$ and $\|A(x-y)\|^2 \geqslant 0$. Thus the finding a least squares solution of the system (3.2) is equivalent to the solution of the normal system of equations.

Exercise 40. Prove that the normal system of equations always has a solution. Suggestion: prove that $\operatorname{ran} A^T A = \operatorname{ran} A^T$, where

$$\operatorname{ran} A = \{Ax |\ x \in \mathbb{R}^n\}$$

for a $m \times n$ matrix A.

Theorem 28. *The normal system of equations (3.3) is uniquely solvable if and only if the columns of matrix A are linearly independent.*

Proof. Let

$$a_1 = \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix}, \ldots, a_n = \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix},$$

then for a $y \in \mathbb{R}^n$

$$Ay = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \ldots\ldots\ldots\ldots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} =$$
$$= \begin{pmatrix} a_{11}y_1 + \ldots + a_{1n}y_n \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}y_1 + \ldots + a_{mn}y_n \end{pmatrix} = y_1a_1 + \ldots + y_na_n.$$

Now

$a_1, \ldots, a_n$ are linearly independent $\Longleftrightarrow$
$\Longleftrightarrow \{y_1a_1 + \ldots + y_na_n = 0 \Rightarrow y_1 = \ldots = y_n = 0\} \Longleftrightarrow$
$\Longleftrightarrow \{Ay = 0 \Rightarrow y = 0\} \Longleftrightarrow \{A^TAy = 0 \Rightarrow y = 0\} \Longleftrightarrow$
$\Longleftrightarrow$ the system (3.3) is uniquely solvable.

■

In the proof we used the assertion of the next

Exercise 41. Prove that, if $A^TAy = 0$, then $Ay = 0$.

Clearly if $Ay = 0$, then $A^TAy = 0$. Hence there holds the equality

$$\ker A^TA = \ker A,$$

where

$$\ker A = \{x \in \mathbb{R} |\ Ax = 0\}.$$

This equality could be used in one of the possible solutions of exercise 40.

2.2 Function approximation by least squares method

Let us have given knots $x_0, \ldots, x_m$ and the corresponding numbers $f_0, \ldots, f_m$ (e.g., the values of a function in the knots or their approximations). Consider the approximate function $\varphi(x) = \sum_{j=0}^{n} c_j \varphi_j(x)$, where the coordinate functions $\varphi_0, \ldots, \varphi_n$ are given. If $m = n$ and the coefficients c_j are found from the system

$$\varphi(x_i) = f_i \qquad i = 0, \ldots, m,$$

or, in other words,

$$\sum_{j=0}^{n} c_j \varphi_j(x_i) = f_i, \qquad i = 0, \ldots, m, \tag{3.4}$$

then an interpolant is obtained. If the system (3.4) fails to have a solution (which is natural for $m > n$), then solving the system (3.4) with respect to the unknowns c_j using the least squares method, we are talking about the least squares approximation.

In this situation the matrix of the system (3.4) is

$$A = \begin{pmatrix} \varphi_0(x_0) & \ldots & \varphi_n(x_0) \\ \vdots & \ddots & \vdots \\ \varphi_0(x_m) & \ldots & \varphi_n(x_m) \end{pmatrix},$$

which means that $a_{ij} = \varphi_j(x_i)$. Thus the normal system of equations is

$$\sum_{j=0}^{n} \left(\sum_{i=0}^{m} \varphi_k(x_i) \varphi_j(x_i) \right) c_j = \sum_{i=0}^{m} \varphi_k(x_i) f_i, \qquad k = 0, \ldots, n.$$

Here, the normal system of equations is uniquely solvable if and only if the columns of matrix A,

$$\begin{pmatrix} \varphi_j(x_0) \\ \vdots \\ \varphi_j(x_m) \end{pmatrix}, \qquad j = 0, \ldots, n,$$

are linearly independent.

Note that at least squares approximation, it is allowed that there are equal numbers among the knots $x_0, \ldots, x_m$, e.g., a number (10 or 100) measurements are done for every distinct of others value x_i. It may be that $x_i = x_j$, but $f_i \neq f_j$.

2.3 Example: least squares approximation using polynomials

Choose $\varphi_0(x) = 1$, $\varphi_1(x) = x, \ldots, \varphi_n(x) = x^n$, therefore the approximant is a polynomial $P_n(x) = c_0 + c_1 x + \ldots + c_n x^n$. Here the coefficients of the system (3.4) are $\varphi_j(x_i) = x_i^j$, the normal system of equations is

$$\sum_{j=0}^{n} \left(\sum_{i=0}^{m} {x_i}^{k+j} \right) c_j = \sum_{i=0}^{m} {x_i}^k f_i, \qquad k = 0, \ldots, n. \tag{3.5}$$

Theorem 29. *The system (3.5) is uniquely solvable if and only if there are at least $n+1$ pairwise distinct knots among $x_0, \ldots, x_m$.*

Proof. Let us rely on Theorem 28 about the unique solvability of a normal system of equations. For this it is necessary and sufficient that the columns of the initial system of equations

$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} x_0 \\ \vdots \\ x_m \end{pmatrix}, \ldots, \begin{pmatrix} {x_0}^n \\ \vdots \\ {x_m}^n \end{pmatrix}$$

are linearly independent. This takes place if and only if the rank of matrix

$$A = \begin{pmatrix} 1 & x_0 & \ldots & {x_0}^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & \ldots & {x_m}^n \end{pmatrix}$$

$r(A)$ is equal to the number $n+1$ of columns. If there are $n+1$ pairwise distinct knots then there exists a nonzero minor of degree $n+1$ (Vandermonde determinant), and therefore $r(A) = n+1$. If it is not possible to find $n+1$ pairwise distinct knots, then, in every minor of degree $n+1$ there are at least two equal rows, and the minor itself is equal to 0, which means that $r(A) < n+1$. ∎

In section 2 we considered the least squares approximation for a function which is a linear combination of coordinate functions. Let us now consider an essentially more general problem.

Let it be given the knots $x_0, \ldots, x_m$, not necessarily pairwise distinct, and the values $f_0, \ldots, f_m$. Additionally, let it be possible that $x_i = x_j$ but $f_i \neq f_j$. The approximating function has the form $\varphi(x, c_0, \ldots, c_n)$, where the dependence (in general, nonlinear) on parameters c_i is known. Usually

in practice a particular object to study determines this dependence. The unknown parameters c_i have to be determined from the condition, that the expression

$$\sum_{i=0}^{m}(\varphi(x_i, c_0, \ldots, c_n) - f_i)^2$$

has to be minimal, where $(c_0, \ldots, c_n) \in D \subset \mathbb{R}^{n+1}$ (it can also be that $D = \mathbb{R}^{n+1}$). In general case, this is a complicated problem and the methods used to find the suitable solution belong to the field of optimization.

§3. Numerical differentiation

Numerical differentiation stands for finding derivatives of functions using a finite number of values. Thus the initial situation is the same as at interpolation: given are knots $x_0, \ldots, x_n$ and corresponding values $f_0, \ldots, f_n$, now instead of finding the function we find its derivative or derivatives of higher order. It is clear that the derivatives can also be restored only approximately.

One possible way to get numerical differentiation formulae is to use interpolation formulae. From the differentiation of the interpolation formula

$$f(x) = \varphi(x) + R(x)$$

we get

$$f'(x) = \varphi'(x) + R'(x),$$

and instead of $f'(x)$ we can use $\varphi'(x)$. Similarly

$$f^{(k)}(x) = \varphi^{(k)}(x) + R^{(k)}(x)$$

and instead of $f^{(k)}(x)$ we can use $\varphi^{(k)}(x)$. However, the fact that $R(x)$ is small need not mean that $R'(x), \ldots, R^{(k)}(x)$ are small.

3.1 Numerical differentiation formulae for equidistantly distributed knots

Consider the situation where the knots are $x_i = x_0 + ih$, $i = 0, \ldots, n$. We will treat finding derivatives in these knots. Formulae with better properties are such where the differentiation in a knot x_m uses the knots symmetrically around x_m. Therefore if we use the knots $x_0, \ldots, x_n$ to find the

derivative in knot x_m, then n is an even number and $n = 2m$. Occasionally it is inevitable also to use formulae where the knots are replaced asymmetrically around the knot at which the derivative is calculated, e.g., for reasons connected with the availability of the values f_i. In the following we consider the more important formulae commonly used in practice, for which we fix the notation $f_i = f(x_i)$. For $n = 2$ they are

$$f'(x_1) = \frac{1}{2h}(f_2 - f_0) - \frac{h^2}{6}f'''(\xi), \qquad \xi \in [x_0, x_2],$$
$$f''(x_1) = \frac{1}{h^2}(f_0 - 2f_1 + f_2) - \frac{h^2}{12}f^{(4)}(\xi),$$

and for $n = 4$

$$f'(x_2) = \frac{1}{12h}(f_0 - 8f_1 + 8f_3 - f_4) + \frac{h^4}{30}f^{(5)}(\xi), \qquad \xi \in [x_0, x_4],$$
$$f'''(x_2) = \frac{1}{2h^3}(-f_0 + 2f_1 - 2f_3 + f_4) - \frac{h^2}{4}f^{(5)}(\xi).$$

Among them let us derive the first formula in the form

$$f'(x) = \frac{1}{2h}(f(x+h) - f(x-h)) - \frac{h^2}{6}f'''(\xi).$$

With the help of Taylor expansion we find

$$\frac{1}{2h}(f(x+h) - f(x-h)) = \frac{1}{2h}\left(f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(\xi_1) - \right.$$
$$\left. - \left(f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(\xi_2)\right)\right) =$$
$$= f'(x) + \frac{h^2}{12}(f'''(\xi_1) + f'''(\xi_2)),$$

which can be done if $f \in C^3[x-h, x+h]$. Under the same assumption we get

$$2 \min_{x-h \leqslant z \leqslant x+h} f'''(z) \leqslant f'''(\xi_1) + f'''(\xi_2) \leqslant 2 \max_{x-h \leqslant z \leqslant x+h} f'''(z),$$

or

$$\min_{x-h \leqslant z \leqslant x+h} f'''(z) \leqslant \frac{1}{2}(f'''(\xi_1) + f'''(\xi_2)) \leqslant \max_{x-h \leqslant z \leqslant x+h} f'''(z).$$

The continuous function f''' takes all the values between its minimum and maximum, thus there exists $\xi \in [x-h, x+h]$ such that

$$f'''(\xi) = \frac{1}{2}(f'''(\xi_1) + f'''(\xi_2)).$$

Taking this into account, we have deduced the first of the formulae.

Exercise 42. Deduce three other numerical differentiation formulae under the assumptions $f \in C^4$ or $f \in C^5$, respectively.

3.2 Influence of errors at numerical differentiation

At numerical differentiation (just like at interpolation) the unconditional error stands for inaccuracies in function values f_i, caused, e.g., by errors in measurements or experiment data. The conditional error, though, is the size of the remainder term or its estimate. Here we see the following phenomenon: when decreasing the remainder term, the influence of the unconditional error increases. We explain this in an example even though this phenomenon takes place in any numerical differentiation formula.

From the Taylor expansion

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2}f''(\xi)$$

we get the numerical differentiation formula

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} - \frac{h}{2}f''(\xi).$$

Assume that $f \in C^2[x_0, x_0+\delta]$, $|f''(x)| \leqslant M$, $f_i = f(x_i)$, $i = 0, 1$. One finds $\widetilde{f}_0 = f_0 \pm \varepsilon$, $\widetilde{f}_1 = f_1 \pm \varepsilon$, and calculates $\dfrac{\widetilde{f}_1 - \widetilde{f}_0}{h}$ as the approximation of $f'(x_0)$. Then

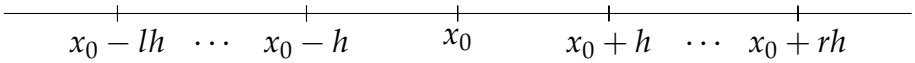
$$\begin{aligned}\left|\frac{\widetilde{f}_1 - \widetilde{f}_0}{h} - f'(x_0)\right| &= \left|\frac{\widetilde{f}_1 - \widetilde{f}_0}{h} - \frac{f_1 - f_0}{h} + \frac{f''(\xi)h}{2}\right| = \\ &= \left|\frac{\widetilde{f}_1 - f_1}{h} - \frac{\widetilde{f}_0 - f_0}{h} + \frac{f''(\xi)h}{2}\right| \leqslant \\ &\leqslant \frac{2\varepsilon}{h} + \frac{M}{2}h = g(h).\end{aligned}$$

If $h \to 0$, then $\frac{M}{2}h \to 0$ (the estimate of the remainder term or the conditional error decreases), but $\frac{2\varepsilon}{h} \to \infty$ (the influence of the unconditional error ε increases).

Studying the function g gives that $g'(h) = -\frac{2\varepsilon}{h^2} + \frac{M}{2}$, and $g'(h) = 0$ yields that $h = 2\sqrt{\frac{\varepsilon}{M}}$. At it $g''(h) = \frac{4\varepsilon}{h^3} > 0$, which means that value $g(h)$ is minimal for $h = 2\sqrt{\frac{\varepsilon}{M}}$. In practice, this may mean that data f_i are obtained with too small step and some of them should be dropped.

3.3 Convergence of numerical differentiation formulae

Consider a set of equidistant knots $x_0 - lh, \ldots, x_0 + rh$, where $l, r \geqslant 0$.



Let it be given the numerical differentiation formula

$$f^{(k)}(x_0) = \frac{1}{h^k} \sum_{i=-l}^{r} b_i f(x_0 + ih) + R_h(f). \tag{3.6}$$

Take k, l, r, and the coefficients b_i as fixed. We call the first part $\frac{1}{h^k} \sum_{i=-l}^{r} b_i f(x_0 + ih)$ in the formula (3.6) the *difference expression*.

Definition. We say that the formula (3.6) or the difference expression in it converges, if

$$\frac{1}{h^k} \sum_{i=-l}^{r} b_i f(x_0 + ih) \to f^{(k)}(x_0) \quad (\text{or } R_h(f) \to 0)$$

in the process of $h \to 0$ for every function $f \in C^k[x_0 - \delta, x_0 + \delta]$.

Let us point out the fact that the convergence is not considered with adding knots but, instead, the distance between them decreases. At it the knots in use are replaced according to the given stencil.

Let us introduce the *characteristic function* of the difference expression.

$$\chi(z) = \sum_{i=-l}^{r} b_i z^i,$$

where $\chi\colon \mathbb{C} \to \mathbb{C}$ or $\chi\colon \mathbb{C} \setminus \{0\} \to \mathbb{C}$. It turns out that many properties of the formula (3.6) depend on the behaviour of this function.

Theorem 30. *For the convergence of the numerical differentiation formula (3.6) it is necessary and sufficient that*

$$\chi(1) = 0,\ \chi'(1) = 0, \dots,\ \chi^{(k-1)}(1) = 0,\ \chi^{(k)}(1) = k! \tag{3.7}$$

Proof. Use Taylor expansions in the difference expression

$$\frac{1}{h^k}\sum_{i=-l}^{r} b_i f(x_0 + ih) = \frac{1}{h^k}\sum_{i=-l}^{r} b_i \Big(f(x_0) + f'(x_0)ih + \\ + \frac{f''(x_0)}{2} i^2 h^2 + \dots + \frac{f^{(k)}(x_0)}{k!} i^k h^k + \alpha_i \Big),$$

where $\frac{\alpha_i}{h^k} \to 0$ in the process $h \to 0$, if $f \in C^k$. Consider equalities

$$\begin{cases} \sum\limits_{i=-l}^{r} b_i = 0 \quad \text{(coefficient of } f(x_0)\text{)}, \\ \sum\limits_{i=-l}^{r} b_i i = 0 \quad \text{(coefficient of } f'(x_0)\text{)}, \\ \sum\limits_{i=-l}^{r} b_i i^2 = 0 \quad \text{(coefficient of } f''(x_0)\text{)}, \\ \dots\dots \\ \sum\limits_{i=-l}^{r} b_i i^{k-1} = 0, \\ \sum\limits_{i=-l}^{r} b_i \frac{i^k}{k!} = 1 \text{ or } \sum\limits_{i=-l}^{r} b_i i^k = k! \end{cases} \tag{3.8}$$

Taking into consideration that $\sum\limits_{i=-l}^{r} b_i \frac{\alpha_i}{h^k} \to 0$ as $h \to 0$, we get that if (3.8) holds then (3.6) converges or $R_h(f) \to 0$ for every $f \in C^k$.

For the converse, assume that (3.6) converges, i.e., $R_h(f) \to 0$ for every $f \in C^k$ in the process $h \to 0$. Use a test function f for which

$$f(x_0) = 1,\ f'(x_0) = 0, \dots,\ f^{(k)}(x_0) = 0.$$

From this we get the first equality in (3.8). If we take f such that

$$f(x_0) = 0,\ f'(x_0) = 1,\ f''(x_0) = 0, \dots,\ f^{(k)}(x_0) = 0,$$

then we get the second equality in (3.8). If we choose f such that

$$f(x_0) = 0, \ldots, \ f^{(k-1)}(x_0) = 0, \ f^{(k)}(x_0) = 1,$$

then we get the last equality in (3.8). The suitable test functions are, for example, $f(x) = \dfrac{(x - x_0)^j}{j!}$, $j = 0, 1, \ldots, k$.

With the discussion so far we have shown that (3.6) converges if and only if the equalities (3.8) hold. The equalities in the theorem's assertion are

$$\begin{cases} \chi(1) = \sum\limits_{i=-l}^{r} b_i = 0, \\ \chi'(1) = \sum\limits_{i=-l}^{r} b_i i = 0, \\ \chi''(1) = \sum\limits_{i=-l}^{r} b_i i(i-1) = 0, \\ \ldots \\ \chi^{(k-1)}(1) = \sum\limits_{i=-l}^{r} b_i i(i-1)\ldots(i-(k-2)) = 0, \\ \chi^{(k)}(1) = \sum\limits_{i=-l}^{r} b_i i(i-1)\ldots(i-(k-1)) = k! \end{cases} \tag{3.9}$$

The first two equalities are the same in (3.8) and (3.9). Taking into consideration the second equality we get that the third equalities are also equivalent. Proceeding in a similar way we see the equivalence of (3.8) and (3.9). ■

Exercise 43. Prove that formula (3.6) converges if and only if its characteristic function can be expressed as $\chi(z) = z^{-l}(z-1)^k Q(z)$, where Q is a polynomial and $Q(1) = 1$.

As examples let us consider the following formulae

1) $f'(x) = \dfrac{1}{h}(f(x+h) - f(x)) - \dfrac{h}{2}f''(\xi)$;

2) $f'(x) = \dfrac{1}{2h}(f(x+h) - f(x-h)) - \dfrac{h^2}{6}f'''(\xi)$;

3) $f''(x) = \dfrac{1}{h^2}(f(x+h) - 2f(x) + f(x-h)) - \dfrac{h^2}{12}f^{(4)}(\xi)$.

The remainder terms can be represented in such a form if $f \in C^2$, $f \in C^3$ or $f \in C^4$, respectively.

The characteristic function in the first formula is $\chi(z) = z - 1$, then

$$\chi(1) = 0, \quad \chi'(z) = 1, \quad \chi'(1) = 1!.$$

In the second formula

$$\chi(z) = \frac{1}{2}\left(z - \frac{1}{z}\right), \quad \chi(1) = 0, \quad \chi'(z) = \frac{1}{2}\left(1 + \frac{1}{z^2}\right), \quad \chi'(1) = 1!.$$

In the third formula

$$\chi(z) = z - 2 + \frac{1}{z},$$
$$\chi(1) = 0, \quad \chi'(z) = 1 - \frac{1}{z^2}, \quad \chi'(1) = 0, \quad \chi''(z) = \frac{2}{z^3}, \quad \chi'''(1) = 2!.$$

Thus, if $f \in C^1$, then the difference expression in the first and in the second formula converges to the derivative, but if $f \in C^2$, then in the third formula the difference expression converges to the second derivative as $h \to 0$.

We are able to judge about the rate of convergence from the remainder terms. We say that the formula (3.6) converges at the rate h^m, if $|R_h(f)| \leqslant ch^m$ for a sufficiently smooth enough function f. For the first formula the rate of convergence is h, for the second and third it is h^2. The rate of convergence for the formula (3.6) can be found using longer Taylor expansions.

IV Numerical integration (approximate calculation of definite integrals)

Introduction

Assume that one has to calculate the definite integral

$$\int_a^b f(x)\,\mathrm{d}x.$$

If it is possible to find the primitive function F, i.e., $F'(x) = f(x)$, then we can use the Newton–Leibniz formula

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

Sometimes it is not possible to find it as an elementary function. For example, $\int e^{x^2}\,\mathrm{d}x$ is not an elementary function (from this follows that unlike the values of an elementary function, this function's values are not so easy to find, e.g., in a computer). We are also unable to use the Newton–Leibniz formula if we only know a finite number of values of function f, e.g., from experimental data or measurement results. In this case approximation methods are used. If a finite number of values of function f is used to find the integral, then the corresponding formulae are called *quadrature formulae*. If a finite number of values of function f is used to find multiple integrals $\int_\Omega f(x)\,\mathrm{d}x$, $\Omega \subset \mathbb{R}^n$, then the corresponding formulae are called *cubature formulae*. We do not consider cubature formulae in this course. Sometimes, for any n, these formulae are called quadrature formulae.

Quite well known are the following quadrature formulae

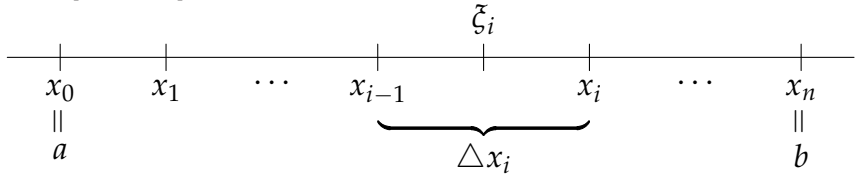
$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} A_i f(x_i),$$

here the argument values x_i are called knots of the quadrature formula, numbers A_i coefficients of the quadrature formula, the expression $\sum_{i=1}^{n} A_i f(x_i)$ the quadrature sum. It is natural to assume that $x_i \in [a, b]$.

Let f be an arbitrary Riemann integrable function. Based on the definition of integral we get

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{\max \Delta x_i \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i,$$

where $a = x_0 < x_1 < \ldots < x_n = b$, $\Delta x_i = x_i - x_{i-1}$, $i = 1, \ldots, n$, $\xi_i \in [x_{i-1}, x_i]$.



From this we obtain a large amount of quadrature formulae

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} f(\xi_i)\Delta x_i,$$

where ξ_i are knots of the quadrature formula, Δx_i the coefficients, and the integral sum is taken as quadrature sum.

This method cannot be used when the integral is improper, which means that either the function f or the domain of integration is unbounded. In this case the following quadrature formulae are used:

$$\int_a^b p(x)f(x)\,\mathrm{d}x = \sum_{i=1}^{n} A_i f(x_i) + R_n(f).$$

Here the following new notions appear: the remainder $R_n(f)$, the weight function p, properties of which reflect the improperness of the integral or the singularities of the function to be integrated, and f is a function with good properties (smooth, bounded).

To develop the theory the weight function is required to satisfy

1° $p(x) \geqslant 0$, $x \in [a, b]$, there exists $\int_a^b p(x)\,\mathrm{d}x > 0$,

2° $\int_a^b p(x)x^k\,\mathrm{d}x$, $k = 1, 2, \ldots$, exist.

These assumptions yield that the integral $\int_a^b p(x)P(x)\,\mathrm{d}x$ exists for every polynomial P. For example, in an interval $[a, b]$ the weight functions

$$p(x) = (x-a)^\alpha (b-x)^\beta, \qquad \alpha, \beta > -1,$$

are used, in the domain $[0, \infty)$ the weight functions

$$p(x) = x^\alpha e^{-x}, \qquad \alpha > -1,$$

are used, and in the domain $(-\infty, \infty)$ the weight functions $p(x) = e^{-x^2}$ or $p(x) = e^{-|x|}$ are used.

Let us look at an example where we separate the weight function. Write

$$\int_{-1}^{1} \frac{\mathrm{d}x}{\sqrt{1-x^6}} = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} \frac{1}{\sqrt{1+x^2+x^4}}\,\mathrm{d}x,$$

from which for $[a, b] = [-1, 1]$

$$p(x) = \frac{1}{\sqrt{1-x^2}} = (1+x)^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}} = (x-(-1))^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}},$$

i.e., $\alpha = \beta = -\frac{1}{2}$, $f(x) = \frac{1}{\sqrt{1+x^2+x^4}}$, and the function f is bounded and smooth.

§1. Interpolatory quadrature rules

Definition. The quadrature formula

$$\int_a^b p(x)f(x)\,\mathrm{d}x = \sum_{i=0}^{n} A_i f(x_i) + R_n(f) \tag{4.1}$$

is said to be of the interpolation type or interpolatory quadrature rule, if its quadrature sum is the integral of the interpolation polynomial with the knots x_i and the weight p. Thus

$$\sum_{i=0}^{n} A_i f(x_i) = \int_a^b p(x)P_n(x)\,\mathrm{d}x,$$

where $P_n(x_i) = f(x_i)$, $i = 0, \ldots, n$, and the degree of P_n does not exceed n.

We have $P_n(x) = \sum_{i=0}^{n} f(x_i)\ell_{ni}(x)$ from the Lagrange formula, therefore

$$\sum_{i=0}^{n} A_i f(x_i) = \int_a^b p(x) \left(\sum_{i=0}^{n} f(x_i)\ell_{ni}(x) \right) \mathrm{d}x = \\ = \sum_{i=0}^{n} \left(\int_a^b p(x)\ell_{ni}(x)\,\mathrm{d}x \right) f(x_i)$$

for every function f if and only if $A_i = \int_a^b p(x)\ell_{ni}(x)\,\mathrm{d}x$. With this it is proved the next

Proposition 31. *The quadrature formula (4.1) is interpolatory quadrature rule if and only if its coefficients are* $A_i = \int_a^b p(x)\ell_{ni}(x)\,\mathrm{d}x$.

Thus the coefficients in interpolatory quadrature rule are uniquely defined if the knots are given (of course, we keep in view that a, b and p are also fixed).

From the interpolation formula $f(x) = P_n(x) + R_n(x)$ we get

$$\int_a^b p(x)f(x)\,\mathrm{d}x = \int_a^b p(x)P_n(x)\,\mathrm{d}x + \int_a^b p(x)R_n(x)\,\mathrm{d}x,$$

hence $\int_a^b p(x)P_n(x)\,\mathrm{d}x = \sum_{i=0}^{n} A_i f(x_i)$ if and only if

$$R_n(f) = \int_a^b p(x)R_n(x)\,\mathrm{d}x.$$

This yields

Proposition 32. *The quadrature formula (4.1) is interpolatory quadrature rule if and only if its remainder term can be expressed by the interpolation formula's remainder term in the form* $R_n(f) = \int_a^b p(x)R_n(x)\,\mathrm{d}x$.

A quadrature formula is called exact for a function f if, for the function f, the integral and the quadrature sum are equal or $R_n(f)$ is equal to zero.

Theorem 33. *Quadrature formula (4.1) is interpolatory quadrature rule if and only if it is exact for all polynomials of degree at most n.*

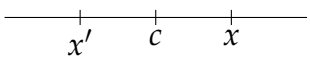
Proof. If (4.1) is exact for all polynomials of degree at most n, then it is also exact for Lagrange fundamental polynomials ℓ_{ni}. Therefore

$$\int_a^b p(x)\ell_{ni}(x)\,\mathrm{d}x = \sum_{j=0}^{n} A_j \ell_{ni}(x_j) = A_i,$$

because $\ell_{ni}(x_j) = \delta_{ij}$. With this we have shown that (4.1) is of the interpolation type.

Now assume that (4.1) is interpolatory quadrature rule. Take an arbitrary polynomial P of degree at most n. It is its own interpolation polynomial because its degree is at most n, it satisfies the interpolation conditions, and the interpolation polynomial is uniquely defined. Therefore the interpolation formula's remainder term $R_n(x) = 0$ and the quadrature formula's remainder term $R_n(P) = \int_a^b p(x)R_n(x)\,\mathrm{d}x = 0$, which means that (4.1) is exact for the polynomial P. ∎

Recall that a function f is an *even function*, if $f(-x) = f(x)$, and it is an *odd function*, if $f(-x) = -f(x)$ for every x from its domain of definition D, which itself has to be symmetric with respect to the point 0: if $x \in D$, then $-x \in D$. More generally, a function f is an even function with respect to a point c, if $f(x) = f(x')$ for all x and x' which are symmetric with respect to the point c: $x - c = c - x'$ or $x' = 2c - x$.



A natural assumption here is that the domain of definition D of function f is symmetric with respect to the point c, which means that $2c - x \in D$ for $x \in D$. Similarly we define an odd function with respect to a point.

Exercise 44. Prove that if in a quadrature formula of interpolation type the weight function p is even with respect to centre $c = \dfrac{a+b}{2}$ of the domain of integration, and the knots are replaced symmetrically with respect to c, then the coefficients corresponding to the symmetrical knots are equal, i.e., if $x_i - c = c - x_j$, then $A_i = A_j$.

Note that if $f \in C^{n+1}[a,b]$, then the interpolation formula's remainder term has the form $R_n(x) = \dfrac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega_n(x)$, and in the interpolatory quadrature rule the remainder term takes the form

$$R_n(f) = \frac{1}{(n+1)!}\int_a^b p(x)f^{(n+1)}(\xi(x))\omega_n(x)\,\mathrm{d}x.$$

§2. Newton–Cotes' formulae

Newton–Cotes' formulae are characterized by the following data:

1) they are interpolatory quadrature rules,
2) domain of integration is bounded, i.e., $a, b \in \mathbb{R}$,
3) $p(x) = 1$ for every $x \in [a,b]$,
4) $x_i = a + ih$, $i = 0, \ldots, n$, $h = \dfrac{b-a}{n}$.

Thus we are free to choose only a, b and n in the Newton–Cotes' formulae.

2.1 Properties of the coefficients in Newton–Cotes' formulae

Newton–Cotes' formulae have the form

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n} A_i f(x_i) + R_n(f).$$

We get that $A_i = \int_a^b \ell_{ni}(x)\,\mathrm{d}x$ because they are of the interpolation type. If we take into consideration the form of ℓ_{ni} then

$$A_i = \int_a^b \frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}\,\mathrm{d}x.$$

Consider the change of variable $x = a + th$, then $\mathrm{d}x = h\,\mathrm{d}t$,

$$x - x_j = a + th - (a + jh) = (t-j)h,$$
$$x_k - x_j = a + kh - (a + jh) = (k-j)h,$$

and the limits of integration a and b are replaced by the values 0 and n, respectively. With this we calculate

$$\begin{aligned} A_i &= h \int_0^n \frac{t(t-1)\dots(t-(i-1))(t-(i+1))\dots(t-n)}{i(i-1)\dots 1\cdot(-1)\dots(i-n)}\,\mathrm{d}t = \\ &= \frac{b-a}{n}\frac{(-1)^{n-i}}{i!(n-i)!}\int_0^n t(t-1)\dots(t-(i-1))(t-(i+1))\dots(t-n)\,\mathrm{d}t = \\ &= (b-a)B_i, \end{aligned}$$

where the numbers B_i are not dependent on the integration domain $[a,b]$, but they do depend on the value n, which is why we write $B_i = B_{ni}$ if we do not fixe n and it is important to emphasize the dependence. Therefore the Newton–Cotes' formulae can be written in the form

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)\sum_{i=0}^{n} B_i f(x_i) + R_n(f).$$

We present some of the properties of the coefficients B_i:

1) $\sum_{i=0}^{n} B_{ni} = 1$, we get this if we take $f(x) \equiv 1$. This function is a polynomial of degree 0 and thus the quadrature formula is exact;

2) $B_{ni} = B_{n,n-i}$ from Exercise 44 in the previous section;

3) $\sum_{i=0}^{n} |B_{ni}| \to \infty$, which we will not prove here.

Properties 1) and 3) allow us to claim that when n increases negative coefficients appear, the first one appears when $n = 8$, and when $n \geqslant 10$ they always appear. Property 3) also means that the influence of inaccuracies in the data increase as n increases. Let us show it. Assume that instead of the values $f(x_i)$ there are found $\widetilde{f}_i$ such that

$$\left|\widetilde{f}_i - f(x_i)\right| \leqslant \varepsilon.$$

Instead of calculating the quadrature sum $\sum_{i=0}^{n} B_{ni} f(x_i)$ we calculate $\sum_{i=0}^{n} B_{ni}\widetilde{f}_i$.

Then

$$\left|\sum_{i=0}^{n} B_{ni}\widetilde{f}_i - \sum_{i=0}^{n} B_{ni}f(x_i)\right| = \left|\sum_{i=0}^{n} B_{ni}\left(\widetilde{f}_i - f(x_i)\right)\right| \leqslant$$
$$\leqslant \sum_{i=0}^{n} |B_{ni}|\left|\widetilde{f}_i - f(x_i)\right| \leqslant$$
$$\leqslant \sum_{i=0}^{n} |B_{ni}|\varepsilon \to \infty,$$

as $n \to \infty$.

2.2 Remainder term of Newton–Cotes' formulae

At the end of the previous part we showed how to express in the interpolatory quadrature rule the remainder term for a smooth function. Using this representation in Newton–Cotes' formulae we get

$$R_n(f) = \frac{1}{(n+1)!}\int_a^b f^{(n+1)}(\xi(x))\omega_n(x)\,\mathrm{d}x,$$

if $f \in C^{n+1}[a,b]$. If we denote $M_{n+1} = \max\limits_{a\leqslant x\leqslant b}|f^{(n)}(x)|$ then we get the estimate

$$|R_n(f)| \leqslant \frac{1}{(n+1)!}M_{n+1}\int_a^b |\omega_n(x)|\,\mathrm{d}x,$$

but this is a considerable overestimation because every time ω_n passes a knot, it changes the sign (this discussion is also valid for other interpolatory quadrature rules which, in general, contain a weight function p). It turns out that we are able to get a much better form of the remainder term in Newton–Cotes' formulae.

We need the mean value theorem known from integral calculus. It holds that

$$\int_a^b f(x)g(x)\,\mathrm{d}x = f(\xi)\int_a^b g(x)\,\mathrm{d}x, \qquad \xi \in (a,b),$$

if f is continuous, g is integrable and keeps the sign (i.e., $g(x) \geqslant 0$ for all $x \in [a,b]$ or $g(x) \leqslant 0$ for all $x \in [a,b]$). For the more known particular case

$g(x) = 1$ for all $x \in [a, b]$ it holds that

$$\int_a^b f(x)\,\mathrm{d}x = f(\xi)(b-a).$$

Let it be $n = 1$ in Newton–Cotes' formula and $f \in C^2[a, b]$. Then

$$\begin{aligned} R_1(f) &= \frac{1}{2!}\int_a^b f''(\xi(x))(x-a)(x-b)\,\mathrm{d}x = \\ &= \frac{1}{2}f''(\xi)\int_a^b (x-a)(x-ib)\,\mathrm{d}x = \\ &= -\frac{(b-a)^3}{12}f''(\xi), \end{aligned}$$

because at interpolation we have seen the continuity of the function

$$x \to f''(\xi(x)),$$

and $(x-a)(x-b) \leqslant 0$ for all $x \in [a, b]$.

In a more general case it is possible to prove (it is based on the mean value theorem but the proof is quite technical), that if n is even, then

$$R_n(f) = \frac{f^{(n+2)}(\xi)}{(n+2)!}\int_a^b (x-c)\omega_n(x)\,\mathrm{d}x,$$

where $c \in \mathbb{R}$ is arbitrary (as an explanation note that $\int_a^b \omega_n(x)dx = 0$, because here ω_n is an odd function with respect to $c = \frac{a+b}{2}$); if n is odd, then

$$R_n(f) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\int_a^b \omega_n(x)\,\mathrm{d}x.$$

Of course these remainder term expressions are only valid if we assume that f has the required smoothness: $f \in C^{n+2}[a, b]$ for n even, $f \in C^{n+1}[a, b]$ for n odd.

These results give us that

$$R_2(f) = \frac{f^{(4)}(\xi)}{4!}\int_a^b (x-a)(x-c)^2(x-b)\,\mathrm{d}x =$$
$$= -\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi), \qquad \text{here we take } c = \frac{a+b}{2},$$
$$R_3(f) = -\frac{3}{80}\left(\frac{b-a}{3}\right)^5 f^{(4)}(\xi).$$

§3. Trapezoidal rule, Simpson's rule, Newton's $\frac{3}{8}$ rule, rectangular rule

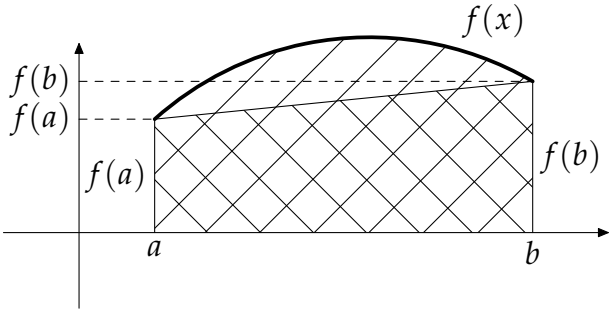
1. Consider Newton–Cotes' formula in the case $n = 1$. We only need to find the coefficients. They satisfy

$$B_0 + B_1 = 1,$$
$$B_0 = B_1,$$

from which $B_0 = B_1 = \frac{1}{2}$. The quadrature formula is

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)\frac{f(a)+f(b)}{2} - \frac{(b-a)^3}{12}f''(\xi), \qquad \xi \in (a,b),$$

where the remainder term is such if $f \in C^2[a,b]$. The calculation by this formula means geometrically that the area $\int_a^b f(x)\,\mathrm{d}x$ under the graph of function f is replaced by the quadrature sum, being the area of a trapezoid.



It is clear that the error here can be quite large. Because of that we proceed as follows. Let us divide interval $[a,b]$ into parts with the length $h = \dfrac{b-a}{n}$, the subintervals have endpoints $x_i = a + ih$, $i = 0,\ldots,n$. Express

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x,$$

and apply on each subinterval the trapezoidal rule

$$\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x = \frac{h}{2}(f(x_{i-1}) + f(x_i)) - \frac{h^3}{12}f''(\xi_i), \qquad \xi_i \in (x_{i-1}, x_i).$$

In sum we have the formula

$$\int_a^b f(x)dx = \frac{h}{2}\sum_{i=1}^{n}(f(x_{i-1}) + f(x_i)) - \frac{h^3}{12}\sum_{i=1}^{n} f''(\xi_i).$$

At it

$$n \min_{1\leqslant i\leqslant n} f''(\xi_i) \leqslant \sum_{i=1}^{n} f''(\xi_i) \leqslant n \max_{1\leqslant i\leqslant n} f''(\xi_i),$$

from which

$$\begin{aligned}\min_{\xi_1\leqslant x\leqslant \xi_n} f''(x) \leqslant \min_{1\leqslant i\leqslant n} f''(\xi_i) \leqslant \frac{1}{n}\sum_{i=1}^{n} f''(\xi_i) \leqslant \\ \leqslant \max_{1\leqslant i\leqslant n} f''(\xi_i) \leqslant \max_{\xi_1\leqslant x\leqslant \xi_n} f''(x).\end{aligned}$$

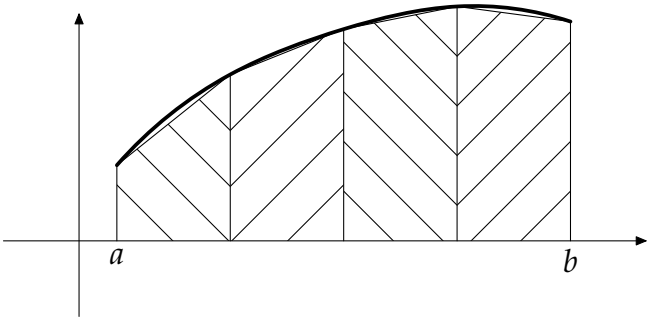
If f'' is continuous, then it attains all the values between its minimum and maximum, and, consequently, there exists $\xi \in (a,b)$ such that $f''(\xi) = \dfrac{1}{n}\sum_{i=1}^{n} f''(\xi_i)$ or $\sum_{i=1}^{n} f''(\xi_i) = nf''(\xi)$. As the result of such averaging having general character the remainder term takes the form

$$R_n(f) = -\frac{nh^3}{12}f''(\xi) = -\frac{(b-a)^3}{12n^2}f''(\xi).$$

The quadrature formula is (here $f_i = f(x_i)$)

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2n}(f_0 + 2f_1 + \ldots + 2f_{n-1} + f_n) - \frac{(b-a)^3}{12n^2}f''(\xi).$$

Geometrically this means that at calculation by this formula the actual area is replaced by a sum of areas of trapezoids, which is illustrated in the next figure in the case $n = 4$.



In the figure $f''(x) < 0$ for all $x \in [a,b]$, thus $-\frac{(b-a)^3}{12n^2}f''(\xi) > 0$ and the integral is greater than the quadrature sum.

If it is necessary to distinguish, then for $n = 1$ the term *simple formula* or the *elementary formula* is used, and for $n \geqslant 2$ it is the term *composite formula* or the *generalized formula*.

2. Take $n = 2$ in the Newton–Cotes' formula, then $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$. We are acquainted with the remainder term, we only have to find the coefficients. The properties of the coefficients give us that

$$\begin{aligned} B_0 + B_1 + B_2 &= 1, \\ B_0 &= B_2. \end{aligned}$$

The coefficients B_i do not depend on the interval $[a,b]$. To simplify the calculations we take the interval $[a,b] = [0,2]$ and the function

$$f(x) = x^2,$$

for which the formula is exact. Then

$$\int_0^2 x^2\,\mathrm{d}x = \frac{8}{3} = 2(B_0 \cdot 0 + B_1 \cdot 1 + B_2 \cdot 4)$$

or

$$B_1 + 4B_2 = \frac{4}{3}.$$

Taking into account the equality $B_1 + 2B_2 = 1$ we get

$$B_0 = B_2 = \frac{1}{6}, \qquad B_1 = \frac{4}{6}.$$

Thus the Newton–Cotes' formula for $n = 2$ is

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) - \\ - \frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi).$$

This formula is called Simpson's rule. Analogically to the trapezoidal rule divide interval $[a, b]$ into n parts, where now let n be an even number, but still $h = \dfrac{b-a}{n}$. Write the integral

$$\int_a^b f(x)\,\mathrm{d}x$$

as a sum of integrals over the intervals $[a, a+2h]$, $[a+2h, a+4h]$,..., $[b-2h, b]$, and apply Simpson's rule to integrals on subintervals. The result of this is

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{\frac{n}{2}} \frac{2h}{6}\left(f(x_{2i-2} + 4f(x_{2i-1}) + f(x_{2i})\right) - \sum_{i=1}^{\frac{n}{2}} \frac{h^5}{90} f^{(4)}(\xi_i).$$

Analogically to the trapezoidal rule case, we now average the values of $f^{(4)}$ and therefore

$$\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\xi_i) = \frac{n}{2} f^{(4)}(\xi), \qquad \xi \in (a, b).$$

In total we have the formula

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{3n}\left(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 4f_{n-1} + f_n\right) - \\ - \frac{(b-a)^5}{180n^4} f^{(4)}(\xi).$$

3. Take $n = 3$ in the Newton–Cotes' formula. We get the coefficients

$$B_0 = B_3 = \frac{1}{8}, \qquad B_1 = B_2 = \frac{3}{8}.$$

Exercise 45. Show how to find the coefficients $B_0, \ldots, B_3$ in the Newton–Cotes' formula for $n = 3$.

Denoting $h = \dfrac{b-a}{3}$ we get that

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{8}\big(f(a) + 3f(a+h) + 3f(a+2h) + f(b)\big) - \frac{3}{80}h^5 f^{(4)}(\xi),$$

which is called the Newton's $\frac{3}{8}$ rule.

Proceeding as earlier, we take a number n which is multiple of 3, then denote $h = \dfrac{b-a}{n}$, and finally decompose the integral $\int_a^b f(x)\,\mathrm{d}x$ into the sum of integrals on intervals $[a, a+3h]$, $[a+3h, a+6h]$, ..., $[b-3h, b]$. Then apply Newton's $\frac{3}{8}$ rule to each integral on subinterval and average the sum of the values of $f^{(4)}$ in the remainder terms. As the result we get

$$\int_a^b f(x)\,\mathrm{d}x = \frac{3(b-a)}{8n}\left(f_0 + 3f_1 + 3f_2 + 2f_3 + 3f_4 + \ldots + 3f_{n-1} + f_n\right) - \\ - \frac{(b-a)^5}{80n^4} f^{(4)}(\xi).$$

Comparing the remainder terms of Simpson's rule and Newton's $\frac{3}{8}$ rule, e.g., for $n = 12$, then it is reasonable to use Simpson's rule. But if $n = 9$, then only Newton's $\frac{3}{8}$ rule could be used.

From the remainder terms of considered formulae we conclude that in the process $h \to 0$ or $n \to \infty$ the quadratic sum in the trapezoidal rule converges to the integral if $f \in C^2[a,b]$, as well in Simpson's rule and Newton's $\frac{3}{8}$ rule if $f \in C^4[a,b]$. Actually, this convergence holds

for all Riemann integrable functions, e.g., in the trapezoidal rule

$$\begin{aligned}&\frac{h}{2}\left(f_0+2f_1+\cdots+2f_{n-1}+f_n\right)=\\&\quad=\frac{1}{2}h\left(f_0+f_1+\cdots+f_{n-1}\right)+\frac{1}{2}h\left(f_1+f_2+\cdots+f_n\right)\to\\&\quad\to\frac{1}{2}\int_a^b f(x)\,\mathrm{d}x+\frac{1}{2}\int_a^b f(x)\,\mathrm{d}x=\int_a^b f(x)\,\mathrm{d}x,\end{aligned}$$

because we divided the quadrature sum into two summands in which both are integral sums with coefficient $\frac{1}{2}$, in the first sum function values are taken at left endpoints of the subintervals, in the second sum at right endpoints.

Exercise 46. Prove that for all Riemann integrable functions Simpson's rule and Newton's $\frac{3}{8}$ rule converge.

4. Let us consider the quadrature formula

$$\int_a^b f(x)\,\mathrm{d}x=A_0f(x_0)+R_0(f),$$

where only one knot x_0 is used. Determine the coefficient A_0 and the knot x_0 in a way that the formula is exact for polynomials of highest possible degree. If $f(x)=1$ for every $x\in[a,b]$, then the exactness gives that

$$\int_a^b f(x)\,\mathrm{d}x=b-a=A_0\cdot 1,$$

therefore $A_0=b-a$. If $f(x)=x$ for every $x\in[a,b]$, then

$$\int_a^b x\,\mathrm{d}x=\frac{x^2}{2}\Big|_a^b=\frac{b^2-a^2}{2}=\frac{(b-a)(b+a)}{2},$$

the quadrature sum is $(b-a)x_0$, and the equality of the quadrature sum and integral gives us that $x_0=\dfrac{a+b}{2}$.

Assume that $f\in C^2[a,b]$, and let us try to find an appropriate form for the remainder term. With the help of Taylor expansion we get

that

$$R_0(f) = \int_a^b f(x)\,\mathrm{d}x - (b-a)f(x_0) =$$
$$= \int_a^b \left(f(x_0) + f'(x_0)(x-x_0) + \frac{f''(\xi(x))}{2}(x-x_0)^2 \right) \mathrm{d}x - (b-a)f(x_0) =$$
$$= \frac{1}{2}\int_a^b f''(\xi(x))(x-x_0)^2\,\mathrm{d}x,$$

because $\int_a^b (x-x_0)\,\mathrm{d}x = 0$. In the following transformations we use an assertion which we formulate in a more general way.

Exercise 47. Prove that in the Taylor expansion

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \ldots + \frac{f^{(n)}(\xi(x))}{n!}(x-x_0)^n$$

the function $x \to f^{(n)}(\xi(x))$ is continuous if $f \in C^n[x_0-\delta, x_0+\delta]$ for some $\delta > 0$.

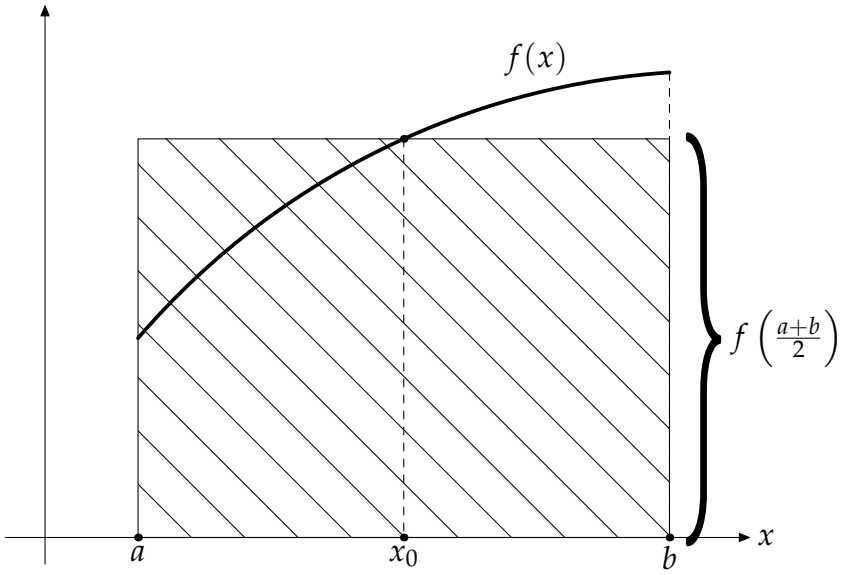
Basing on the mean value theorem of integral calculus and Exercise 47, we get that

$$R_0(f) = \frac{1}{2}f''(\xi)\int_a^b (x-x_0)^2\,\mathrm{d}x = \frac{(b-a)^3}{24}f''(\xi).$$

We now have the formula

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{24}f''(\xi),$$

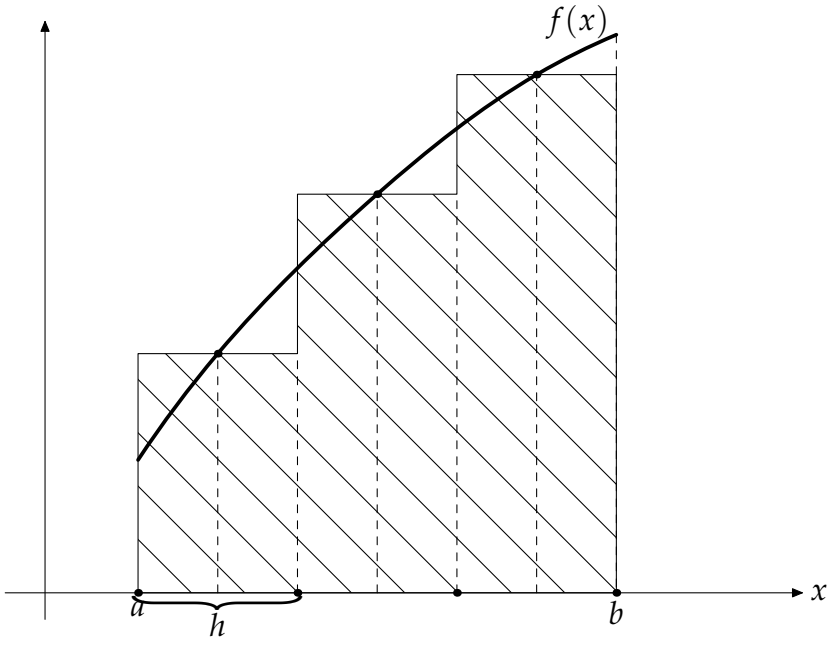
which is called the rectangular rule. Its name comes from the fact that geometrically the quadratic sum is the area of a rectangle which replaces the integral.



Here the composite rule is

$$\int_a^b f(x)\,\mathrm{d}x = h\left(f(a+\frac{h}{2}) + f(a+\frac{3}{2}h) + \ldots + f(b-\frac{h}{2})\right) + \frac{(b-a)^3}{24n^2}f''(\xi).$$

Geometrically this means that the quadrature sum of composite rule is the sum of rectangle areas.



It is clear that here the quadrature sum converges to the integral of every Riemann integrable function, because the quadrature sum itself is an integral sum.

Exercise 48. Are the trapezoidal rule, Simpson's rule, Newton's $\frac{3}{8}$ rule, rectangular rule (composite rules) of the interpolation type?

§4. Main part of remainder term

Consider the quadrature formula

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n} A_i f(x_i) + R_h(f),$$

where $x_i = a + ih$, $i = 0, \ldots, n$, $h = \dfrac{b-a}{n}$, $A_i = A_{ni}$, and keep in view the process, where n and h change so that $n \to \infty$ and $h \to 0$.

Definition. If for any function f smooth enough it holds

$$R_h(f) = K(f)h^q + \varrho_h(f),$$

where $\varrho_h(f) = O(h^{q+1})$, $K(f)$ is a constant which does not depend on h, and there exists a smooth function f such that $K(f) \neq 0$, then the part $K(f)h^q$ is called the *main part of the remainder term*.

As an example let us find the main part of the remainder term in trapezoidal rule. Above we saw that

$$R_h(f) = -\sum_{i=1}^{n} \frac{h^3}{12} f''(\xi_i),$$

where $\xi_i \in (x_{i-1}, x_i)$. Let us use here the Taylor expansion

$$f''(\xi_i) = f''\left(x_{i-1} + \frac{h}{2}\right) + f'''(\eta_i)\left(\xi_i - \left(x_{i-1} + \frac{h}{2}\right)\right),$$

where $\eta_i \in \left(x_{i-1} + \dfrac{h}{2}, \xi_i\right)$ or $\eta_i \in \left(\xi_i, x_{i-1} + \dfrac{h}{2}\right)$. Then

$$R_h(f) = -\sum_{i=1}^{n} \frac{h^3}{12}\left(f''\left(x_{i-1} + \frac{h}{2}\right) + f'''(\eta_i)\left(\xi_i - \left(x_{i-1} + \frac{h}{2}\right)\right)\right),$$

and in the representation

$$\sum_{i=1}^{n} \frac{h^3}{12} f''\left(x_{i-1} + \frac{h}{2}\right) = \frac{h^2}{12} \sum_{i=1}^{n} h f''\left(x_{i-1} + \frac{h}{2}\right),$$

we recognize the quadrature sum of rectangular rule (with the coefficient $\frac{h^2}{12}$). Therefore

$$\sum_{i=1}^{n} h f''\left(x_{i-1} + \frac{h}{2}\right) = \int_a^b f''(x)\,\mathrm{d}x - \frac{b-a}{24} h^2 f^{(4)}(\xi),$$

and if, for example, $f \in C^4[a,b]$, we get from the part of $R_h(f)$

$$-\sum_{i=1}^{n} \frac{h^3}{12} f''\left(x_{i-1} + \frac{h}{2}\right) = -\frac{h^2}{12} \int_a^b f''(x)\,\mathrm{d}x + O(h^4).$$

In the rest of $R_h(f)$ we have $|\xi_i - (x_{i-1} + \frac{h}{2})| \leqslant \frac{h}{2}$, and the whole rest part is of order $O(h^3)$. Thus $R_h(f) = -\frac{h^2}{12} \int_a^b f''(x)\,\mathrm{d}x + O(h^3)$, which means that in the trapezoidal rule $q = 2$ and

$$K(f) = -\frac{1}{12} \int_a^b f''(x)\,\mathrm{d}x = \frac{1}{12}(f'(a) - f'(b)).$$

Exercise 49. Find main parts of the remainder terms in Simpson's rule and Newton's $\frac{3}{8}$ rule.

The notion of remainder term's main part can also be used for the rectangular formula, where, in general, we consider the formulae

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} A_i f(x_i) + R_h(f),$$

$h = \frac{b-a}{n}$, $x_i = x_1 + (i-1)h$, $i = 2, \ldots, n$, with $x_i \in [a,b]$, $i = 1, \ldots, n$, and $A_i = A_{ni}$.

Exercise 50. Find the main part of remainder term in the rectangular formula.

§5. Runge's method

Consider the situation from the previous section where we separated the main part in a remainder term as follows:

$$R_h = Kh^q + \varrho_h, \qquad \varrho_h = O(h^{q+1}).$$

Let us use the symbol I for the integral, and let the quadrature sum be I_h at step h. Then $I = I_h + R_h$.

Assume that for finding the same integral we use a quadrature formula with two different steps, h and H, which of course means that the number of knots is different. In this situation

$$\begin{aligned} R_h &= Kh^q + \varrho_h = I - I_h, \\ R_H &= KH^q + \varrho_H = I - I_H \end{aligned}$$

and by subtraction we get

$$\begin{aligned} &K(H^q - h^q) + \varrho_H - \varrho_h = I_h - I_H, \\ &K = \frac{I_h - I_H}{H^q - h^q} + \frac{\varrho_h - \varrho_H}{H^q - h^q}. \end{aligned}$$

Using this we have

$$R_h = Kh^q + \varrho_h = \frac{I_h - I_H}{\left(\frac{H}{h}\right)^q - 1} + \frac{\varrho_h - \varrho_H}{\left(\frac{H}{h}\right)^q - 1} + \varrho_h.$$

Look at the situation, where $H = kh$, $k = \text{const} > 1$. Then

$$R_h = \frac{I_h - I_H}{k^q - 1} + \frac{\varrho_h - \varrho_H}{k^q - 1} + \varrho_h.$$

At it $\varrho_H = O(H^{q+1}) = O((kh)^{q+1}) = O(h^{q+1})$, thus

$$\frac{\varrho_h - \varrho_H}{k^q - 1} + \varrho_h = O(h^{q+1})$$

and

$$R_h = \frac{I_h - I_H}{k^q - 1} + O(h^{q+1}).$$

Here we see that the approximate value of the remainder term is $\dfrac{I_h - I_H}{k^q - 1}$ because it is of the same order as the main part Kh^q. The most common case is $k = 2$, where the approximate value of the remainder term

is $\frac{I_h - I_{2h}}{2^q - 1}$. For example, for the trapezoidal rule $q = 2$ and $R_h \approx \frac{I_h - I_{2h}}{3}$, for Simpson's rule and Newton's $\frac{3}{8}$ rule $q = 4$ and $R_h \approx \frac{I_h - I_{2h}}{15}$. Let us emphasize that by Runge's method we find an approximate value of remainder term, which may be either positive or negative.