3316321 Numerical analysis

Timo Erkama
LaTeX-translation
Marko Lamminsalo
University of Eastern Finland
December 20, 2010

Contents

1 Error analysis

1.1 Errors in numerical computations

In practice, computations consist of finite steps and decimals; the results are therefore *approximations*.

Let a denote an exact value and $\tilde{a}$ an approximated value of a. The difference

$$\varepsilon = \tilde{a} - a$$

is called the (absolute) *error* in the approximation $\tilde{a}$. In other words,

$$\tilde{a} = a + \varepsilon,$$

i.e. the approximation is the sum of exact value and error.

Example.

$$\begin{aligned} \tilde{a} = 10.5, \quad a = 10.2 \quad &\Rightarrow \varepsilon = 0.3 \\ \tilde{a} = 1.60, \quad a = 1.82 \quad &\Rightarrow \varepsilon = -0.22 \end{aligned}$$

The *relative error* ε_r in approximation $\tilde{a}$ is

$$\varepsilon_r = \frac{\varepsilon}{a} = \frac{\tilde{a} - a}{a} = \frac{\text{error}}{\text{exact value}} \quad (a \neq 0)$$

Clearly

$$\varepsilon_r \approx \frac{\varepsilon}{\tilde{a}}$$

if $|\varepsilon|$ is much smaller than $|\tilde{a}|$. The number $\gamma = a - \tilde{a} = -\varepsilon$ is called *correction*, so that

$$a = \tilde{a} + \gamma.$$

The *upper bound of the error* is a number β such that

$$|\tilde{a} - a| \leq \beta, \quad \text{i.e. } |\varepsilon| \leq \beta.$$

Sources of error

1. Idealisations in mathematical models
2. Errors in the input data

3. Truncation errors (arise when an infinite process is replaced by a finite one)

4. Rounding errors

In addition, also errors caused by carelessness are to be considered.

1.2 Floating point numbers

In the decimal representation of a real number infinite numbers are usually needed. Calculators and computers, however, can only handle finite number sequences. Therefore in computations real numbers are replaced with *floating point numbers*, in which the number of significant digits is a constant depending on the machine. For this purpose each real number must first be *rounded* using the following rules:

A decimal number is rounded to n decimals by

- removing all decimals on the right side of the n:th decimal
- if the removed number is $> \frac{1}{2} \cdot 10^{-n}$, the n:th decimal is raised by 1
- if the removed number is $= \frac{1}{2} \cdot 10^{-n}$, the n:th decimal is raised only if it is *odd*

Example. Rounding to 3 decimals:

$$\begin{aligned} 0.4711 &\approx 0.471 \\ 0.4716 &\approx 0.472 \\ 0.4715 &\approx 0.472 \\ 0.4705 &\approx 0.470 \end{aligned}$$

The last rule is an attempt to eliminate the systematic rounding error. Raising decimal occurs "as often" as leaving it unchanged.

The *rounding error* occurring in rounding a number to n decimals has an absolute value at most $\frac{1}{2} \cdot 10^{-n}$.

The numbers rounded in the example above have 3 significant digits. We say that the approximated value of a real number has n *significant digits*, if n is

the largest positive integer such that the absolute value of the absolute error is at most the product of the unit of the first non-zero number and $5 \cdot 10^{-n}$.

If e.g. the exact value $a = 1.1996$ has an approximation $\tilde{a} = 1.200$, the approximated value $\tilde{a}$ has 4 significant digits.

Similarly, the approximation is said to have n *correct decimals*, if the absolute error is at most $\frac{1}{2} \cdot 10^{-n}$.

The number of significant digits describes the relative accuracy and the number of correct decimals describes absolute accuracy.

Example.

(a)	$\tilde{a} = 47.11,$	$\lvert\varepsilon\rvert \leq 0.5 \cdot 10^{-2}$	2 corr. dec., 4 sig. digits
(b)	$\tilde{a} = 0.0047110,$	$\lvert\varepsilon\rvert \leq 0.5 \cdot 10^{-7}$	7 corr. dec., 5 sig. digits
(c)	$\tilde{a} = 4710 \cdot 10^2,$	$\lvert\varepsilon\rvert \leq 0.5 \cdot 10^{2}$	0 corr. dec., 4 sig. digits
(d)	$\tilde{a} = 47100,$	$\lvert\varepsilon\rvert \leq 0.5 \cdot 10^{2}$	0 corr. dec., 3 sig. digits

The number of significant digits:

(a)	$0.5 \cdot 10^{-2} = 10 \cdot 5 \cdot 10^{-n}$	$\Leftrightarrow$	$n = 4$
(b)	$0.5 \cdot 10^{-7} = 10^{-3} \cdot 5 \cdot 10^{-n}$	$\Leftrightarrow$	$n = 5$
(c)	$0.5 \cdot 10^{2} = 10^{5} \cdot 5 \cdot 10^{-n}$	$\Leftrightarrow$	$n = 4$

The decimal number system is a position system with base 10. Most computers use a position system with another base $\beta \geq 2$, e.g. $\beta = 2$ or $\beta = 16$. In such position systen any real number can be written as

$$(\pm d_n d_{n-1} \ldots d_2 d_1 d_0 . d_{-1} d_{-2} \ldots)_\beta$$

where $d_n, d_{n-1}, \ldots$ are integers between 0 and $\beta - 1$. The value of such a number is

$$d_n \beta^n + d_{n-1} \beta^{n-1} + \cdots + d_2 \beta^2 + d_1 \beta^1 + d_0 \beta^0 + d_{-1} \beta^{-1} + d_{-2} \beta^{-2} + \cdots .$$

For example $\pi = 3.1415 \ldots = 3 \cdot 10^0 + 1 \cdot 10^{-1} + 4 \cdot 10^{-2} + 1 \cdot 10^{-3} + \cdots$.

Example.

$$
\begin{aligned}
(760)_8 &= 7 \cdot 8^2 + 6 \cdot 8^1 + 0 \cdot 8^0 = (496)_{10} \\
(101.101)_2 &= 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} \\
&= (5.625)_{10} \\
(0.333)_{10} &= 3 \cdot 10^{-1} + 3 \cdot 10^{-2} + 3 \cdot 10^{-3} + \cdots = \frac{1}{3} \\
\frac{1}{5} &= (0.2)_{10} = (0.00110011\ldots)_2 \\
\frac{1}{5} &= 2^q + r
\end{aligned}
$$

In computer numbers are replaced with floating point numbers so that each number takes a fixed amount of memory. If the number system has base β, each nonzero real number can be expressed in the form

$$X = M \cdot \beta^e$$

where e is an integer and

$$M = \pm D_0.D_1D_2D_3\ldots$$

$$0 \leq D_i \leq \beta - 1$$

$$D_0 \neq 0.$$

In floating point system M is replaced with a finite number sequence

$$m = \pm d_0.d_1d_2\ldots d_t,$$

m has a finite number (t+1) of elements. The number (floating point number) stored to the computer is then

$$x = m \cdot \beta^e \quad \text{or} \quad x = \beta^{e+1}$$

The latter is gained when e.g. $D_0 = D_1 = D_2 = \ldots = \beta - 1$. Here m is called *mantissa* and e *exponent*. Since $0 \leq d_i \leq \beta - 1$ and $d_0 \neq 0$, every non-zero floating point number is normalized so that $1 \leq |m| < \beta$.

The amount of storage that is reserved for the exponent e determines the range of numbers that can be represented. The limits of e can be written

$$L \leq e \leq U$$

where L and U are negative and positive integers, respectively. If the result of a computation is a floating point number with $e > U$, a so called *overflow* occurs and the computer issues an error signal. The corresponding error with $e < L$ is called *underflow* an it does not usually terminate the process.

1.3 Numerical stability

An algorithm associated to a numerical problem is complete description of those finitely many operations needed to replace the solution of the problem from the initial values.

An algorithm is *stable* if the associated truncation and rounding errors have only a small effect to the output. Otherwise the algorithm is called *unstable*. Such *numerical instability* can often be avoided by choosing a better algorithm.

Mathematical instability is a property of the mathematical model and cannot be improved by choosing a different numerical algorithm; the problem is then called *ill-conditioned* (*häiriöaltis* tai *pahanlaatuinen*).

Example. Find the roots of equations

$$\text{(a) } x^2 - 4x + 2 = 0 \quad \text{and} \quad \text{(b) } x^2 - 40x + 2 = 0 \tag{1}$$

by using 4 significant digits in the computations.

The roots of a second degree equation $ax^2 + bx + c = 0$ can be obtained from the formulas

$$x_1 = \frac{1}{2a}\left(-b + \sqrt{b^2 - 4ac}\right), \quad x_2 = \frac{1}{2a}\left(-b - \sqrt{b^2 - 4ac}\right). \tag{2}$$

Since $x_1 x_2 = \frac{c}{a}$, alternative formulas are

$$x_1 \text{ as before}, \quad x_2 = \frac{c}{ax_1}. \tag{3}$$

Applying (2) to equation (1a) yields

$$x = 2 \pm \sqrt{2} = 2.000 \pm 1.414 \;\Rightarrow\; \begin{cases} x_1 = 3.414 \\ x_2 = 0.586 \end{cases}$$

Formula (3):

$$x_1 = 3.414, \quad x_2 = \frac{2.000}{3.414} = 0.5858.$$

Error in the last expression is $\leq 10^{-4}$.

Applying (2) to equation (1b) yields

$$x = 20 \pm \sqrt{398} = 20.00 \pm 19.95 \;\Rightarrow\; \begin{cases} x_1 = 39.95 \\ x_2 = 0.05 \end{cases}$$

Formula (3):

$$x_1 = 39.95, \quad x_2 = \frac{2.000}{39.95} = 0.05006.$$

Error in the last expression is $\leq 10^{-5}$.

1.4 Differential calculus in error analysis

Let's compute the value of the function $f(x) = \frac{1}{x^2}$ at $x = 0.015$ but first we round this value to 2 decimals. The error will be then

$$f(0.02) - f(0.015) = \frac{1}{(0.02)^2} - \frac{1}{(0.015)^2} = -1944.$$

A small perturbation (0.005) in the value of x causes a large error in the output. The problem is *ill-conditioned*.

A measure for ill-condition could be

$$\left| \frac{\text{error in the output}}{\text{error in the input}} \right|$$

If f is a continously differentiable function, according to the mean value theorem

$$f(x+\varepsilon) - f(x) = f'(\xi)\varepsilon,$$

where ξ is between x and $x + \varepsilon$.

$$\frac{|f(x+\varepsilon) - f(x)|}{|\varepsilon|} \leq \max |f'(\xi)|$$

$$|f(x+\varepsilon) - f(x)| \leq |\varepsilon| \max |f'(\xi)| \tag{1}$$

For functions of several variables the corresponding result is as follows: Denote $f = f(x_1, x_2, \ldots, x_n)$ the exact value at $(x_1, x_2, \ldots, x_n)$ and let $\tilde{f} = f(x_1 + \varepsilon_1, x_2 + \varepsilon_2, \ldots, x_n + \varepsilon_n)$ be the approximate value. Then

$$\tilde{f} - f = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x + \theta\varepsilon)\varepsilon_k,$$

where $0 < \theta < 1$, $x = (x_1, \ldots, x_n)$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$.

$$\Rightarrow \quad |\tilde{f} - f| \leq \sum_{k=1}^{n} |\varepsilon_k| \max \left| \frac{\partial f}{\partial x_k} \right| \tag{2}$$

where the maxima are computed on the segment joining x and $x + \varepsilon$.

Example 1. Estimate the error on computing

$$f(x_1, x_2, x_3) = \frac{x_1}{x_2^2 + x_3^2}$$

at (1.0,1.0,1.0) where the coordinates are given with 2 significant digits.

The error ε_k in the value of the coordinate x_k has an absolute value ≤ 0.05. $(1 \leq k \leq 3)$.

$$\frac{\partial f}{\partial x_1} = \frac{1}{x_2^2 + x_3^2}, \quad \frac{\partial f}{\partial x_2} = \frac{-2x_2x_1}{(x_2^2 + x_3^2)^2}, \quad \frac{\partial f}{\partial x_3} = \frac{-2x_3x_1}{(x_2^2 + x_3^2)^2}$$

$$\begin{aligned}
\max \left| \frac{\partial f}{\partial x_1} \right| &\leq \frac{1}{0.95^2 + 0.95^2} = 0.56 \\
\max \left| \frac{\partial f}{\partial x_2} \right| &\leq \frac{2 \cdot 1.05 \cdot 1.05}{(0.95^2 + 0.95^2)^2} = 0.68 \\
\max \left| \frac{\partial f}{\partial x_3} \right| &\leq \frac{2 \cdot 1.05 \cdot 1.05}{(0.95^2 + 0.95^2)^2} = 0.68
\end{aligned}$$

$$\begin{aligned}
(2) \ \Rightarrow \quad |\tilde{f} - f| &= |f(1 + \varepsilon_1, 1 + \varepsilon_2, 1 + \varepsilon_3) - f(1, 1, 1)| \\
&\leq 0.05 \cdot 0.56 + 0.05 \cdot 0.68 + 0.05 \cdot 0.68 = 0.096.
\end{aligned}$$

Example 2. $f = x_1 + x_2 + \ldots + x_n$

$$\frac{\partial f}{\partial x_k} = 1$$

$$(2) \Rightarrow |\tilde{f} - f| \leq \sum_{k=1}^{n} |\varepsilon_k|$$

In addition absolute errors are added.

Suppose that $n = 1000$ and $|\varepsilon_k| \leq 0.5 \cdot 10^{-5}$

$$(2) \Rightarrow \ |\tilde{f} - f| \leq 1000 \cdot 0.5 \cdot 10^{-5} = 0.5 \cdot 10^{-2}.$$

The actual error is probably much smaller. The upper bound $0.5 \cdot 10^{-2}$ is attained only if all ε_k are either positive or negative and have absolute value $= 0.5 \cdot 10^{-5}$.

Example 3. $f = x_1 - x_2$

$$\left|\frac{\partial f}{\partial x_1}\right| = \left|\frac{\partial f}{\partial x_1}\right| = 1$$

Absolute error:

$$|\tilde{f} - f| \leq |\varepsilon_1| + |\varepsilon_2|$$

Relative error:

$$\frac{|\tilde{f} - f|}{|f|} \leq \frac{|\varepsilon_1| + |\varepsilon_2|}{|x_1 - x_2|}$$

If $x_1 = 0.5763 \pm 0.5 \cdot 10^{-4}$ and $x_2 = 0.5765 \pm 0.5 \cdot 10^{-4}$, we get

$$\frac{|\tilde{f} - f|}{|f|} \leq \frac{10^{-4}}{10^{-4}} = 1 = 100\%.$$

The loss of accuracy occurring in the substraction of two almost equal numbers is called *cancellation*. Such loss of accuracy can often be avoided by reformulation into a mathematically equivalent expression.

Example 4.

$$f = x_1^{m_1} \cdot x_2^{m_2} \cdots x_n^{m_n} = \left(\frac{f}{x_k^{m_k}}\right) x_k^{m_k}$$

$$\left|\frac{\partial f}{\partial x_k}\right| = \left(\frac{f}{x_k^{m_k}}\right) m_k x_k^{m_{k-1}} = \frac{m_k}{x_k} f$$

The upper bound of the absolute error:

$$(2) \Rightarrow \ |\tilde{f} - f| \leq \sum_{k=1}^{n} |\varepsilon_k| \max \left|\frac{m_k}{x_k} f\right|$$

The upper bound of the relative error:

$$\frac{|\tilde{f} - f|}{|f|} \leq \sum_{k=1}^{n} |\varepsilon_k||m_k| \frac{\max f/x_k}{|f|} \approx \sum_{k=1}^{n} |m_k| \left|\frac{\varepsilon_k}{x_k}\right|$$

In multiplication and division upper bounds of relative errors are added.

Example 5. Suppose that x and y are numbers satisfying $5 \leq x \leq 10$ and $1 \leq y \leq 2$. Suppose that x has 3 and y has 4 correct decimals. Estimate the absolute error in $e^{\sin(xy)}$.

$$f(x, y) = e^{\sin(xy)}$$

$$(2) \Rightarrow |\tilde{f} - f| \leq |\varepsilon_x| \max \left|\frac{\partial f}{\partial x}\right| + |\varepsilon_y| \max \left|\frac{\partial f}{\partial y}\right|$$

Here $|\varepsilon_x| \leq \frac{1}{2} \cdot 10^{-3}$ and $|\varepsilon_y| \leq \frac{1}{2} \cdot 10^{-4}$.

$$\frac{\partial f}{\partial x} = y \cos(xy) e^{\sin(xy)}; \qquad \frac{\partial f}{\partial y} = x \cos(xy) e^{\sin(xy)}$$

When $5 \leq x \leq 10$ and $1 \leq y \leq 2$, we get

$$\left|\frac{\partial f}{\partial x}\right| \leq 2 \cdot 1 \cdot e^1 = 2e \quad \text{and} \quad \left|\frac{\partial f}{\partial y}\right| \leq 10 \cdot 1 \cdot e^1 = 10e,$$

and therefore

$$|\tilde{f} - f| \leq \frac{1}{2} \cdot 10^{-3} \cdot 2e + \frac{1}{2} \cdot 10^{-4} \cdot 10e = 1.5 \cdot e \cdot 10^{-3} < 0.005.$$

Example 6. In the sum

$$\sum_{k=1}^{N} \frac{1}{\sqrt{1 + cosx_k}}$$

each x_k is given with 3 correct decimals and $0 \leq x_k \leq 1 \; \forall \; k$. How large can N be if we require that the absolute error in the sum does not exceed 10^{-2}?

The error in one individual term is at most

$$|\varepsilon| \max |f'(\xi)|$$

where

$$f(x) = \frac{1}{\sqrt{1 + cosx_k}}, \quad 0 \leq \xi \leq 1, \quad |\varepsilon| \leq \frac{1}{2} \cdot 10^{-3}$$

The total error is obtained by multilying this with N. Thus N can be at most so large that the inequality

$$N \cdot \frac{1}{2} \cdot 10^{-3} \max_{0 \leq \xi \leq 1} |f'(\xi)| \leq 10^{-2}$$

holds. Solving N from the inequality yields

$$N \leq \frac{10}{\frac{1}{2} \max_{0 \leq \xi \leq 1} |f'(\xi)|} \leq \frac{10}{\frac{1}{2} \cdot 0.2201} = 90.9,$$

where

$$\max_{0 \leq x \leq 1} |f'(x)| = \max_{0 \leq x \leq 1} \left| \frac{\sin x}{2(1 + \cos x)^{\frac{3}{2}}} \right| \leq \frac{\sin 1}{2(1 + \cos 1)^{\frac{3}{2}}} = \frac{0.8415}{2(1.5403)^{\frac{3}{2}}} = 0.2201.$$

The largest integer less than 90.9 is 90 and thus the answer is $N = 90$.

1.5 Statistical error analysis

In the case of a very large number of similar operations as in Examples 2 and 6 the value error bounds can be very pessimistic. Sometimes one could get a more realistic error bound by using statistical analysis.

For example, if we multiply 1000 numbers each of which has 3 significant digits, then in each individual number the relative error is at most 0.5 ‰, but he relative error of the product is ≤ 50 %. A statistical analysis shows however that within 68 % probability the total error is $< 3.2\%$.

2 Numerical integration

2.1 Trapezoidal rule and Simpson's formula

We wish to compute (numerically)

$$\int_a^b f(x)dx.$$

If the integrand $f(x)$ is derivative of a known function F so that $f(x) = F'(x)$ we get the analytic solution

$$F(b) - F(a) = \int_a^b F'(x)dx = \int_a^b f(x)dx. \tag{1}$$

Geometrically, we should compute the area of the domain appearing in the figure. Let's divide the domain into vertical stripes using the division of $[a, b]$ of the points $x_1, x_2, \ldots, x_{n-1}$. Approximating the area of each strip separately and adding approximations we get an approximate value for the total area and the integral $\int_a^b f(x)dx$. Clearly

$$\int_a^b f(x)dx = \int_a^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{n-1}}^b f(x)dx. \tag{2}$$

If we denote $x_0 = a$ and $x_n = b$, (2) can be written

$$\int_a^b f(x)dx = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)dx. \tag{3}$$

We suppose that the subintervals $[x_k, x_{k+1}]$ have equal length h. We approximate the area of the strip corresponding to $[x_k, x_{k+1}]$ by the area of the trapezoid in the figure. Then we get an approximation

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx h\frac{f(x_k) + f(x_{k+1})}{2}$$

Substituting this to (3) we get

$$\int_a^b f(x)dx \approx \sum_{k=0}^{n-1} h\frac{f(x_k)+f(x_{k+1})}{2}. \tag{4}$$

The graph of function f has been approximated with *straight lines*. In (4) each function value (except $f(x_0)$ and $f(x_n)$) appears twice. Therefore (4) can be written

$$\int_a^b f(x)dx \approx h\left[\frac{f(x_0)}{2}+f(x_1)+f(x_2)+\cdots+f(x_{n-1})+\frac{f(x_n)}{2}\right] \tag{5}$$

This is the so called *trapezoidal rule*.

If f is twice continuously differentiable in $[a,b]$, one can show that the error in the approximation (5) is

$$-\frac{(b-a)}{12}h^2 f''(\xi),$$

where ξ lies between a and b.

Trapezoidal rule

$$\int_a^b f(x)dx = h\left[\frac{f(x_0)}{2}+f(x_1)+f(x_2)+\cdots+f(x_{n-1})+\frac{f(x_n)}{2}\right]+R, \tag{6}$$

where

$$R=-\frac{(b-a)}{12}h^2 f''(\xi), \quad a=x_0\le\xi\le x_n=b.$$

For R we get an upper bound

$$|R|\le\frac{(b-a)}{12}h^2\max_{a\le x\le b}|f''(\xi)| \tag{7}$$

This upper bound will be arbitrarily small if h is taken small enough, because $R\to 0$, when $h\to 0$.

Example 1. We apply (6) to $f(x) = x^2$ on the interval $[0, 1]$. Choose

$$h = \frac{1}{n}, \quad x_0 = 0, \quad x_n = 1, \quad x_k = kh = \frac{k}{n}.$$

Now $f'(x) = 2x$ and $f''(x) = 2$, and thus

$$R = -\frac{1-0}{12}\left(\frac{1}{n}\right)^2 \cdot 2 = -\frac{1}{6n^2}$$

$$\begin{aligned}
\frac{1}{3} = \int_0^1 x^2 dx &= h\left[\frac{x_0^2}{2} + x_1^2 + x_2^2 + \cdots + x_{n-1}^2 + \frac{x_n^2}{2}\right] + R \\
&= h\left[\frac{0}{2} + \left(\frac{1}{n}\right)^2 + \left(\frac{2}{n}\right)^2 + \cdots + \left(\frac{n-1}{n}\right)^2 + \frac{1}{2}\right] - \frac{1}{6n^2} \\
&= \frac{h}{n^2}\left[1^2 + 2^2 + \cdots + (n-1)^2\right] + \frac{h}{2} - \frac{1}{6n^2}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow 1^2 + 2^2 + \cdots + (n-1)^2 &= \frac{n^2}{h}\left[\frac{1}{3} - \frac{h}{2} + \frac{1}{6n^2}\right] \\
&= \frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6} \\
&= \frac{2n^3 - 3n^2 + n}{6}
\end{aligned}$$

$$\begin{aligned}
1^2 + 2^2 + \cdots + (n-1)^2 + n^2 &= \frac{2n^3 - 3n^2 + n}{6} + n^2 \\
&= \frac{2n^3 + 3n^2 + n}{6}
\end{aligned}$$

If f'' is approximately constant, then the error in (6) can be estimated by using *Richardsson's extrapolation* as follows:

First we use the trapezoidal rule with n subintervals ($h = \frac{b-a}{n}$):

$$\int_a^b f(x)dx = T_1 + R_1 \quad \text{where } R_1 = -\frac{(b-a)}{12}\left(\frac{b-a}{n}\right)^2 f''(\xi_1). \qquad (8)$$

Then we use the rule with $2n$ subintervals ($h = \frac{b-a}{2n}$):

$$\int_a^b f(x)dx = T_2 + R_2 \quad \text{where } R_2 = -\frac{(b-a)}{12}\left(\frac{b-a}{2n}\right)^2 f''(\xi_2). \tag{9}$$

Since f'' is approximately constant on $[a, b]$, we have $f(\xi_1) \approx f(\xi_2)$. Then $R_1 \approx 4R_2$, and (8) and (9) imply

$$\begin{cases} \displaystyle\int_a^b f(x)dx \approx T_1 + 4R_2 \\ \displaystyle\int_a^b f(x)dx = T_2 + R_2 \end{cases}$$

Substraction yields the $\frac{1}{3}$-rule

$$R_2 = \frac{T_2 - T_1}{3}. \tag{10}$$

This is an approximation for the error term in the trapezoidal rule with $2n$ subintervals.

Since $\int_a^b f(x)dx = T_2 + R_2$, we can *improve* the approximation T_2 to the integral $\int_a^b f(x)dx$ by using (10). We obtain

$$\int_a^b f(x)dx = T_2 + \frac{T_2 - T_1}{3}. \tag{11}$$

Remark. (11) is actually *Simpson's rule* in a non-standard form. Usually Simpson's rule is derived by dividing $[a, b]$ to an even number ($2n$) of subintervals and writing

$$\int_a^b f(x)dx = \sum_{k=0}^{n-1} \int_{x_{2k}}^{x_{2k+2}} f(x)dx. \tag{12}$$

Here each integral $\int_{x_{2k}}^{x_{2k+2}} f(x)dx$ is approximated by integrating a quadratic polynomial having same values as f at x_{2k}, x_{2k+1} and x_{2k+2}.

Simpson's rule

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(x_0) + 4U + 2J + f(x_{2n})\right] + R, \tag{13}$$

where

$$U = f(x_1) + f(x_3) + \cdots + f(x_{2n-1})$$
$$J = f(x_2) + f(x_4) + \cdots + f(x_{2n-2})$$
$$R = -\frac{(b-a)}{180}h^4 f^{(4)}(\xi) \quad a \le \xi \le b$$

We show next that (11) and (13) agree so that

$$T_2 + \frac{T_2 - T_1}{3} = \frac{h}{3}\left[f(x_0) + 4U + 2J + f(x_{2n})\right],$$

where T_1 and T_2 are as before trapezoidal rule approximations with n and $2n$ subintervals, respectively.

$$\begin{aligned}
T_1 &= 2h\left[\frac{f(x_0)}{2} + \quad + f(x_2) + \quad + f(x_4) + \cdots\right.\\
&\quad \left. + f(x_{2n-2}) + \quad + \frac{f(x_{2n})}{2}\right]\\
4T_2 &= 4h\left[\frac{f(x_0)}{2} + f(x_1) + f(x_2) + f(x_3) + f(x_4) + \cdots\right.\\
&\quad \left. + f(x_{2n-2}) + f(x_{2n-1}) + \frac{f(x_{2n})}{2}\right]\\
4T_2 - T_1 &= h\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots\right.\\
&\quad \left. + 2f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n})\right]\\
&= h\left[f(x_0) + 4U + 2J + f(x_2 n)\right]
\end{aligned}$$

Remark 1. For $n = 2$ we have $J = 0$ and Simpson's rule reads

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

Remark 2. If f is a polynomial of degree ≤ 3, then the approximation given by Simpson's rule yields the exact value of integral ($f^{(4)}(x) \equiv 0$ for such polynomial).

Example 2. Apply Simpson's rule ($h = 0.1$) to the integral

$$\int_0^1 (x^3 + \sin(p\sqrt{3}x))dx.$$

For which values of p the approximation given by Simpson's rule has error $\leq \frac{1}{2} \cdot 10^{-3}$?

$$\begin{aligned} f(x) &= x^3 + \sin(p\sqrt{3}x) \\ f'(x) &= 3x^2 + p\sqrt{3}\cos(p\sqrt{3}x) \\ f''(x) &= 6x - 3p^2 \sin(p\sqrt{3}x) \\ f'''(x) &= 6 - 3\sqrt{3}p^3 \cos(p\sqrt{3}x) \\ f^{(4)}(x) &= 9p^4 \sin(p\sqrt{3}x) \end{aligned}$$

$$|f^{(4)}(\xi)| = |9p^4 \sin(p\sqrt{3}x)| \leq 9p^4$$

$$\Rightarrow \ |R| \leq \frac{b-a}{180}h^4 \max_{a\leq x\leq b} |f^{(4)}(x)| = \frac{1-0}{180}\left(\frac{1}{10}\right)^4 9p^4 = \frac{1}{20} \cdot 10^{-4}p^4.$$

This upper bound is $\leq \frac{1}{2} \cdot 10^{-3}$ if

$$\begin{aligned} \frac{1}{20} \cdot 10^{-4}p^4 &\leq \frac{1}{2} \cdot 10^{-3} \\ \Rightarrow \qquad p^4 &\leq 100 \\ \Rightarrow \qquad p^2 &\leq 10 \\ \Rightarrow \qquad |p| &\leq \sqrt{10}. \end{aligned}$$

The error is at most $\frac{1}{2} \cdot 10^{-3}$ if $|p| \leq \sqrt{10}$ provided that errors in computing function values are neglected.

The error term R can be estimated by using Richardsson's extrapolation just as in the case of the trapezoidal rule, if $f^{(4)}(x)$ is approximately constant.

$$\int_a^b f(x)dx = S_1 + R_1; \quad R_1 = -\frac{(b-a)}{180}h^4 f^{(4)}(\xi_1)$$

Halving each subinterval we apply Simpson's rule with $\frac{h}{2}$:

$$\int_a^b f(x)dx = S_2 + R_2; \quad R_2 = -\frac{(b-a)}{180} \cdot \left(\frac{h}{2}\right)^4 f^{(4)}(\xi_2). \tag{14}$$

If $f^{(4)}(\xi_1) \approx f^{(4)}(\xi_2)$, then $R_1 \approx 16R_2$ and

$$\begin{cases} \int_a^b f(x)dx \approx S_1 + 16R_2 \\ \int_a^b f(x)dx = S_2 + R_2 \end{cases}$$

Substraction yields the $\frac{1}{15}$-rule

$$R_2 \approx \frac{S_2 - S_1}{15}. \tag{15}$$

In order to reach a desired accuracy one can in practice apply Simpson's rule with 2, 4, 8, 16, etc. subintervals and stop the computation when the absolute value on the difference of two consecutive results divided by 15 is less than the desired error bound. Finally the correction

$$\frac{S_2 - S_1}{15}$$

should be added to the last approximation S_2 (see (14)).

2.2 Example

We shall apply the trapezoidal rule and Simpson's rule to the integral

$$\int_0^1 x^4 dx = \Big/_0^1 \frac{x^5}{5} = \frac{1}{5} = 0.2$$

Trapezoidal rule with one subinterval (length $h = 1$):

$$T_1 = 1 \cdot \left[\frac{0^4}{2} + \frac{1^4}{2} \right] = 0.5$$

With two subintervals (length $h = \frac{1}{2}$), we get

$$T_2 = \frac{1}{2} \left[\frac{0^4}{2} + \left(\frac{1}{2} \right)^4 + \frac{1^4}{2} \right] = 0.28125$$

Here the correct error is 0.08125. The $\frac{1}{3}$-rule gives the error estimate

$$T_2 - \int_0^1 x^4 dx \approx -\frac{T_2 - T_1}{3} = -\frac{0.28125 - 0.50000}{3} = 0.07292$$

Formula (7) gives the error bound

$$|R| \leq \frac{b-a}{12} h^2 \max_{a \leq x \leq b} |f''(x)| = \frac{1-0}{12} \left(\frac{1}{2} \right)^2 \max_{0 \leq x \leq 1} |12x^2| = 0.25000$$

The error estimate given by the $\frac{1}{3}$-rule (0.07292) is pretty good while the error bound given by formula (7) (0.25000) is too pessimistic. The $\frac{1}{15}$-rule as well as the formula

$$R = -\frac{b-a}{180} h^4 f^{(4)}(\xi)$$

both give an error estimate with 5 correct decimals. The second estimate is actually exact, because $f^{(4)}(\xi) = 4 \cdot 3 \cdot 2 \cdot 1$ does not depend on ξ.

2.3 Evaluation error

The total error arising from the inaccuracy of the values $f(x_i)$ of the integrand is called the *evaluation error*.

Theorem. *Let $f(x_i)$ and $\tilde{f}(x_i)$ be the exact and computed function values at x_i, respectively, and let ε_i be the error so that*

$$\tilde{f}(x_i) = f(x_i) + \varepsilon_i.$$

Suppose that $|\varepsilon_i| < \varepsilon$ for each i. Then the absolute value of the evaluation error in the trapezoidal rule and Simpson's rule is at most

$$\varepsilon(b-a).$$

Proof. (for the trapezoidal rule)
Denote $f_k = f(x_k)$, $\tilde{f}_k = \tilde{f}(x_k)$. Evaluation error for the trapezoidal rule:

$$\begin{aligned} L &= h\left(\frac{\tilde{f}_0}{2} + \tilde{f}_1 + \cdots + \tilde{f}_{n-1} + \frac{\tilde{f}_n}{2}\right) - h\left(\frac{f_0}{2} + f_1 + \cdots + f_{n-1} + \frac{f_n}{2}\right) \\ &= h\left(\frac{\tilde{f}_0 - f_0}{2} + (\tilde{f}_1 - f_1) + \cdots + (\tilde{f}_{n-1} - f_{n-1}) + \frac{\tilde{f}_n - f_n}{2}\right) \\ &= h\left(\frac{\varepsilon_0}{2} + \varepsilon_1 + \cdots + \varepsilon_{n-1} + \frac{\varepsilon_n}{2}\right) \end{aligned}$$

$$\begin{aligned} |L| &\le h\left(\frac{|\varepsilon_0|}{2} + |\varepsilon_1| + \cdots + |\varepsilon_{n-1}| + \frac{|\varepsilon_n|}{2}\right) \\ &\le h\left(\frac{\varepsilon}{2} + \varepsilon + \cdots + \varepsilon + \frac{\varepsilon}{2}\right) \\ &\le h\varepsilon\left(\frac{1}{2} + 1 + \cdots + 1 + \frac{1}{2}\right) \end{aligned} \tag{16}$$

Here the expression $h\left\{\frac{1}{2} + 1 + \cdots + 1 + \frac{1}{2}\right\}$ also results from the trapezoidal rule applied to the integral $\int_a^b 1dx$. For this integral the trapezoidal rule gives an exact value, because the second derivative appearing in the error formula (for the constant function $\equiv 1$) is zero. Therefore

$$\int_a^b 1dx = b - a = h\left(\frac{1}{2} + 1 + \cdots + 1 + \frac{1}{2}\right).$$

From (16) we thus get

$$|L| \le \varepsilon(b - a).$$

The proof for Simpson's rule is analogous. □

2.4 Difficulties in numerical integration

The above formulas for numerical integration can only be applied when the next two conditions are satisfied:

1. The interval $a \le x \le b$ is finite

2. The integrand $f(x)$ is bounded

Problem 1. In applications we often find integrals with an interval of infinite length, e.g.

$$\int_a^\infty f(x)dx.$$

By definition,

$$\int_a^\infty f(x)dx = \lim_{A\to\infty} \int_a^A f(x)dx;$$

if this limit exists, then the "tail integral" $\int_A^\infty f(x)dx$ approaches zero as $A \to \infty$.

Problem 1 is solved by performing a numerical integration on the finite subinterval $[a, A]$ and by estimating the tail integral from A to ∞.

For example, if we wish to compute $\int_a^\infty f(x)dx$ with an error $\leq \varepsilon$, we could write

$$\int_a^\infty f(x)dx = \int_a^A f(x)dx + \int_A^\infty f(x)dx$$

and choose A so that the tail integral $\int_A^\infty f(x)dx$ has absolute value $\leq \frac{\varepsilon}{2}$. Then we could apply e.g. Simpson's rule to $\int_a^A f(x)dx$ and double the number of subintervals as many times until the error is $\frac{\varepsilon}{2}$. Then the best estimate for $\int_a^A f(x)dx$ should differ from $\int_a^\infty f(x)dx$ by at most ε.

Similarly we could handle integrals such as

$$\int_{-\infty}^a f(x)dx \quad \text{or} \quad \int_{-\infty}^\infty f(x)dx$$

Example of a tail estimate.

$$\int_0^\infty \frac{(\sin x)^2}{x^5+1}dx = \int_0^A \frac{(\sin x)^2}{x^5+1}dx + \int_A^\infty \frac{(\sin x)^2}{x^5+1}dx$$

We wish to choose A so that the tail integral

$$\int_A^\infty \frac{(\sin x)^2}{x^5+1}dx \le 0.001$$

We replace the integrand with a majorant which is easily integrable. Since

$$(\sin x)^2 \le 1 \quad \text{and} \quad \frac{1}{x^5+1} \le \frac{1}{x^5}$$

we have

$$\int_A^\infty \frac{(\sin x)^2}{x^5+1}dx \le \int_A^\infty \frac{1}{x^5}dx = \Big/_A^\infty -\frac{1}{4x^4} = \frac{1}{4A^4}.$$

If $A = 4$, then

$$\frac{1}{4A^4} = \frac{1}{4^5} = \frac{1}{1024} \le 0.001,$$

so that

$$\int_A^\infty \frac{(\sin x)^2}{x^5+1}dx \le 0.001.$$

Remark. In tail estimates usually quite crude estimates of the integrand suffice.

Problem 2. If the integrand is unbounded, we could try to reduce the integral by using a change of variable to another integral with a bounded integrand. Then the interval of integration usually becomes infinite, so that we should use the strategy described in the solution of Problem 1.

Example.

$$\int_0^1 \frac{1+\sin x}{\sqrt{x}}dx$$

Substitute $x = \frac{1}{y}$, $dx = -\frac{1}{y^2}dy$ to get

$$\int_\infty^1 \frac{1+\sin\frac{1}{y}}{\sqrt{\frac{1}{y}}}\left(-\frac{dy}{y^2}\right) = \int_1^\infty \frac{1+\sin\frac{1}{y}}{y^{\frac{3}{2}}}dy$$

Here the integral

$$\int_1^\infty \frac{1+\sin\frac{1}{y}}{y^{\frac{3}{2}}}dy$$

is bounded for $1 \leq y < \infty$. Another way to compute the integral is to divide the integral

$$\int_0^1 \frac{1+\sin x}{\sqrt{x}}dx = \int_0^1 \frac{dx}{\sqrt{x}}dx + \int_0^1 \frac{\sin x}{\sqrt{x}}dx$$

and then substitute $\sin x$ for it's Taylor series

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots .$$

2.5 Example

The primitive of $f(x) = e^{-x^2}$ cannot be expressed in closed form. However

$$\int_0^\infty e^{-x^2}dx = \frac{\sqrt{\pi}}{2} = 0.88623$$

Let's compute this numerically with four correct decimals. Write

$$\int_0^\infty e^{-x^2}dx = \int_0^A e^{-x^2}dx + \int_A^\infty e^{-x^2}dx$$

and determine A so that

$$\int_A^\infty e^{-x^2}dx \leq \frac{1}{4}\cdot 10^{-4} = \frac{1}{2}\cdot\frac{1}{2}\cdot 10^{-4}.$$

For example, we could use the fact that $x^2e^{-x^2}$ is decreasing for $1 \leq x < \infty$ so that

$$e^{-x^2} = \left(x^2e^{-x^2}\right)\cdot\frac{1}{x^2} \leq A^2e^{-A^2}\frac{1}{x^2} \quad \text{if } A \geq 1.$$

Now we get a tail estimate

$$\int_A^\infty e^{-x^2}dx = \int_A^\infty \left(x^2 e^{-x^2}\right)\frac{dx}{x^2} \le A^2 e^{-A^2}\int_A^\infty \frac{dx}{x^2}$$
$$= A^2 e^{-A^2}\cdot\frac{1}{A} = Ae^{-A^2} \qquad (A \ge 1)$$

By choosing $A = 4$ we get

$$\int_4^\infty e^{-x^2}dx \le 4\cdot e^{-16} \approx 4\cdot 10^{-7} \le \frac{1}{4}\cdot 10^{-4}.$$

We estimate $\int_0^4 e^{-x^2}dx$ by using the trapezoidal formula.

2.6 Choice of step length h

If the integrand varies heavily in a subinterval of the interval of integration, it may be a good idea to write

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

and compute separately the integral over the interval ($[a, c]$ in figure) where f has large oscillation. Then we can substantially reduce the computations because in computing the integral $\int_c^b f(x)dx$ the required step length is much larger than in $[a, c]$.

2.7 Remark

Even when $\int_a^b f(x)dx$ can be computed analytically by using a known primitive function F such that $F' = f$, the computations can be so awkward that a numerical solution is to be referred.

3 Nonlinear equations

In this chapter we try to determine approximate values of real roots of a nonlinear equation

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

and estimate the error in these approximations.

3.1 Graphical analysis

A sketch of the graph of $y = f(x)$ can give a preliminary idea of the location of the roots of (1) on the x-axis. The intersection points of the graph with the x-axis then give approximations of these roots.

Sometimes (1) could be replaced with $f_1(x) = f_2(x)$ (if e.g. $f = f_1 - f_2$). Then the x-coordinates of the intersection points of the graphs of $y = f_1(x)$ and $y = f_2(x)$ provide approximations for the roots.

Example 1. If $f(x) = x^2 - \cos x$, then (1) can be written $x^2 = \cos x$. The positive root is then seen to be approximately $x_0 = 0.8$.

3.2 Taylor's formula

Let us recall Taylor's formula in the simplest cases.

The mean value theorem.
If f is differentiable on $[a, b]$, then $\exists\, \xi \in (a, b)$ such that

$$f(b) - f(a) = (b - a)f'(\xi). \tag{2}$$

Graphically this means that a tangent of the graph of $y = f(x)$ is parallel to the line segment L joining the points $\big(a, f(a)\big)$ and $\big(b, f(b)\big)$. The slope of L is

$$\frac{f(b) - f(a)}{b - a}$$

and therefore it agrees with the slope of the tangent at $\big(\xi, f(\xi)\big)$.

Another example of Taylor's formula is

$$f(b) - f(a) = (b - a)f'(a) + \frac{(b - a)^2}{2} f''(\xi), \tag{3}$$

where $a < \xi < b$. If $h = b - a$, then (3) can be written

$$f(a+h) - f(a) = hf'(a) + \frac{h^2}{2} f''(\xi). \tag{4}$$

3.3 General error estimate

Suppose that a is an approximate value of a root $\overline{x}$ of $f(x) = 0$. If a is a good approximation, then $|f(a)|$ should in general be small.

Conversely we have the following estimate.

Theorem 1. *If $\overline{x}$ and a are as above, then*

$$|\overline{x} - a| \leq \frac{|f(a)|}{\min_{x \in I} |f'(x)|}$$

where I is the interval with end points a and $\overline{x}$.

Proof. Mean value theorem (2) yields

$$\begin{aligned} & -f(a) = f(\overline{x}) - f(a) = f'(\xi)(\overline{x} - a) \\ \Rightarrow \quad & |f(a)| = |f'(\xi)||\overline{x} - a| \geq \min_{x \in I} |f'(x)||\overline{x} - a| \\ \Rightarrow \quad & |\overline{x} - a| \leq \frac{|f(a)|}{\min_{x \in I} |f'(x)|} \end{aligned}$$

□

3.4 Fixed point iteration

Let us write $f(x) = 0$ in an equivalent form $x = F(x)$ (two equations are called equivalent if they have exactly the same roots). F can be chosen in different ways, for example

$$x = x - f(x) = F_1(x)$$

$$x = x - cf(x) = F_2(x) \qquad (c \neq 0 \text{ constant})$$

$$x = x - g(x)f(x) = F_3(x) \qquad (0 < |g(x)| < \infty)$$

The equation $f(x) = 0$ has the same roots with $x = F_i(x)$ $(i = 1, 2, 3)$.

In iteration methods we start with an initial value x_0 and define $x_1, x_2, \ldots$ by the *iteration formula*

$$x_{n+1} = F(x_n) \qquad (n = 0, 1, 2, \ldots). \tag{5}$$

Problem. When does the sequence $\{x_n\}$ converge to the desired root $\overline{x}$ of $x = F(x)$?

We shall see that the method works under some assumptions on the slope of F (Theorem 3).

Theorem 2. *Suppose that $\{x_n\}$ converges to a limit α and that F is continuous. Then α is the root of the equation $x = F(x)$.*

Proof. Let $n \to \infty$ in $x_{n+1} = F(x_n)$.

Left-hand side: $x_{n+1} \to \alpha$

Right-hand side: Since F is continuous, then $F(x_n) \to F(\alpha)$.

Then

$$\alpha = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} F(x_n) = F(\alpha).$$

□

The main theorem concerning the convergence of the iteration is:

Theorem 3. *Suppose that the inequality*

$$|F'(x)| \leq m < 1$$

holds on an interval containing the root $\overline{x}$ and each x_n. Then

$$\lim_{n\to\infty} x_n = \overline{x}.$$

Proof.

$$\left.\begin{array}{rcl} x_1 & = & F(x_0) \\ \overline{x} & = & F(\overline{x}) \end{array}\right\} \Rightarrow \quad x_1 - \overline{x} = F(x_0) - F(\overline{x})$$

Mean value theorem (2)

$$\Rightarrow \quad F(x_0) - F(\overline{x}) = F'(\xi)(x_0 - \overline{x})$$

Since $|F'(\xi)| \leq m$, we get

$$|x_1 - \overline{x}| = |F(x_0) - F(\overline{x})| = |F'(\xi)||x_0 - \overline{x}| \leq m|x_0 - \overline{x}|.$$

Hence $|x_1 - \overline{x}| \leq m|x_0 - \overline{x}|$. Similarly

$$|x_2 - \overline{x}| = |F(x_1) - F(\overline{x})| = |F'(\xi_1)||x_1 - \overline{x}| \leq m|x_1 - \overline{x}|.$$

Since $|x_1 - \overline{x}| \leq m|x_0 - \overline{x}|$, it follows that

$$|x_2 - \overline{x}| \leq m|x_1 - \overline{x}| \leq m^2|x_0 - \overline{x}|.$$

Repeating the argument we find that

$$|x_n - \overline{x}| \leq m|x_{n-1} - \overline{x}| \leq m^2|x_{n-2} - \overline{x}| \leq \cdots \leq m^n|x_0 - \overline{x}|.$$

Here $m^n \to 0$ as $n \to \infty$ (because $m < 1$). Therefore the right-hand side $\to 0$. Hence

$$|x_n - \overline{x}| \to 0$$

and the theorem is proved. □

Example 3. Determine the real root of the equation

$$x^3 - x - 1 = 0.$$

From the graph we see that the only real root $\overline{x} \approx 1.3$. We write the equation in an equivalent form

$$x = x^3 - 1 = F(x).$$
$$\Rightarrow F'(x) = 3x^2$$

Therefore $F'(x) > 1$ close to $x = 1.3$. Thus we cannot expect that the sequence $x_n = F(x_{n-1})$ converges. For example, if $x_0 = 1.3$, then $x_1 = 1.197$ and $x_2 \approx 0.71506$, so that the distance from the root $\overline{x}$ is increasing.

Example 4. We write the equation of Example 3 in the form

$$x = F(x) = \frac{1}{x^2 - 1}.$$

Then

$$F'(x) = -\frac{2x}{(x^2 - 1)^2},$$

so that $|F'(x)| > 1$ close to $x = 1.3$. If $x_0 = 1.3$, then $x_1 = 1.4493$ and $x_2 = 0.9087$. The iteration diverges.

Example 5. Write the above equation in the form

$$x = F(x) = (x+1)^{\frac{1}{3}}.$$

Then

$$F'(x) = \frac{1}{3}(x+1)^{-\frac{2}{3}}$$

and

$$F'(1.3) = \frac{1}{3}(2.3)^{-\frac{2}{3}} < 1.$$

Actually

$$|F'(x)| \leq \frac{1}{3} \quad \text{when } x > 0.$$

By induction one can show that $x_n > 0$ if $x_0 > 0$. Thus by Theorem 3 the iteration converges whenever $x_0 > 0$. If $x_0 = 1.3$, we get $x_1 = 1.3200$, $x_2 = 1.3238$.

In the proof of Theorem 3 we derived the inequality

$$|x_n - \overline{x}| \leq m^n |x_0 - \overline{x}|. \tag{6}$$

This can be used to estimate the accuracy of x_n if we know an upper bound for m and $|x_0 - \overline{x}|$.

Problem. How can we decide in general whether the iteration converges for some given initial values x_0?

In Theorem 3 we assumed that $|F'(x)| \leq m < 1$ on an interval I which contains $\overline{x}$ and each x_n. How can we find such an I in practice?

Answer:

- First we find an interval $a \leq x \leq b$ containing the root $\overline{x}$ and the initial point x_0
- I shall be the interval which has the same midpoint as $a \leq x \leq b$ but is three times as long as $a \leq x \leq b$
- m is chosen so that $|F'(x)| \leq m$ on I

- if then $m < 1$, then the hypotheses of Theorem 3 hold on I, inequality (6) holds and the iteration converges
- if $m \geq 1$, start again with a different interval $a \leq x \leq b$.

Example 6. How many iterations should be performed by using the formula

$$x_{n+1} = \frac{-1}{x_n^2 + 2}$$

and the initial point $x_0 = -0.5$ to determine the real root of

$$x^3 + 2x + 1 = 0$$

with four correct decimals?

Since the polynomial $x^3 + 2x + 1$ changes signs on the interval $-0.5 \leq x \leq -0.4$, then $|x_0 - \overline{x}| \leq 0.1$ and I will be the interval $-0.6 \leq x \leq -0.3$.

$$F(x) = \frac{-1}{x^2 + 2}; \quad F'(x) = \frac{2x}{(x^2 + 2)^2}$$

For each $x \in I$ we have

$$|F'(x)| \leq \frac{2 \cdot 0.6}{((-0.3)^2 + 2)^2} = \frac{2 \cdot 0.6}{2.09^2} \leq 0.3$$

We can choose $m = 0.3$. From (6) we get

$$|x_n - \overline{x}| \leq 0.3^n \cdot 0.1$$

For $n = 7$ the right-hand side is $\leq \frac{1}{2} \cdot 10^{-4}$. Thus 7 iterations will suffice.

In error estimates we should not forget the evaluation error.

Theorem 4. *Let $\overline{x}$ be a root of $x = F(x)$ and let ε_n be the evaluation error in $F(x_n)$ so that*

$$x_{n+1} = F(x_n) + \varepsilon_n. \tag{7}$$

Suppose that $|\varepsilon_n| \leq \varepsilon$ and that

$$|F'(x)| \leq m < 1 \tag{8}$$

holds for each x between $\overline{x}$ and x. Then

$$|x_{n+1} - \overline{x}| \leq \frac{m}{1 - m}|x_{n+1} - x_n| + \frac{\varepsilon}{1 - m}. \tag{9}$$

Proof.

$$x_{n+1} - \overline{x} \overset{(7)}{=} F(x_n) + \varepsilon_n - \overline{x} = F(x_n) + \varepsilon_n - F(\overline{x})$$
$$\Rightarrow \quad |x_{n+1} - \overline{x}| \leq |F(x_n) - F(\overline{x})| + |\varepsilon_n|$$

Mean value theorem & (8)

$$\Rightarrow \quad |F(x_n) - F(\overline{x})| \leq m|x_n - \overline{x}|.$$

Hence

$$\begin{aligned} |x_{n+1} - \overline{x}| &\leq m|x_n - \overline{x}| + |\varepsilon_n| \\ &= m|x_n - x_{n+1} + x_{n+1} - \overline{x}| + |\varepsilon_n| \\ &\leq m|x_n - x_{n+1}| + m|x_{n+1} - \overline{x}| + \varepsilon \end{aligned}$$

$$\Rightarrow \quad (1-m)|x_{n+1} - \overline{x}| \leq m|x_n - x_{n+1}| + \varepsilon$$

Since $m < 1$, (9) follows. □

Example 7. We apply Theorem 4 to Example 5.

$$f(x) = x^3 - x - 1; \qquad x = F(x) = (x+1)^{\frac{1}{3}}$$

$$x_0 = 1.3, \quad x_1 = 1.3200, \quad x_2 = 1.3238 \qquad n = 1$$

Now $f(1.3) = -0.103 < 0$ and $f(1.4) = 0.344 > 0$ so $1.3 \leq \overline{x} \leq 1.4$.

$$|F'(x)| = \frac{1}{3(x+1)^{\frac{2}{3}}} \leq \frac{1}{3 \cdot 2.3^{\frac{2}{3}}} \leq 0.2 = m$$

Since $x_1 = 1.3200$ lies between 1.3 and 1.4, we can apply Theorem 4 with $n = 1$. If we assume that the evaluation error in x_2 is $\leq \frac{1}{2} \cdot 10^{-3}$, we obtain

$$\begin{aligned} |x_2 - \overline{x}| &\leq \frac{0.2}{0.8} |1.3238 - 1.3200| + \frac{1}{0.8} \cdot \frac{1}{2} \cdot 10^{-3} \\ &\leq \frac{1}{4} \cdot 4 \cdot 10^{-3} + 0.625 \cdot 10^{-3} \leq 2 \cdot 10^{-3} \end{aligned}$$

Thus the error in x_2 is at most 0.002.

Example 8. Suppose that $\varepsilon = 10^{-12}$ and $m = \frac{1}{2}$ in Theorem 4. If $\overline{x}$ should be determined with an error $\leq 10^{-10}$, we continue the iteration until

$$\frac{\frac{1}{2}}{1-\frac{1}{2}}\,|x_{n+1} - x_n| + \frac{10^{-12}}{1-\frac{1}{2}} \leq 10^{-10},$$

that is, until

$$|x_{n+1} - x_n| \leq 10^{-10} - 2 \cdot 10^{-12} = 0.98 \cdot 10^{-10}.$$

Remark. The first term on the right-hand side of (9) reflects the truncation error and the second term reflects the evaluation error.

We usually apply Theorem 4 by choosing an interval I containing $\overline{x}$ such that (8) holds for some $m < 1$ and each $x \in I$. If $x_n \in I$, we compute x_{n+1} and estimate the error $|x_{n+1} - \overline{x}|$ by using (9). If $x_{n+1} \in I$, we can repeat the procedure. The iteration is stopped when the error is small enough.

The error bound for x_{n+1} given by (9) is always at least equal to the evaluation error in $F(x_n)$. The actual error in x_{n+1} can of course be smaller.

3.5 Newton's method

We write $f(x) = 0$ in the form

$$x = x - \frac{f(x)}{f'(x)} = F(x).$$

Iteration formula

$$\begin{cases} x_0 = \text{ initial value} \\ x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}\ ; \quad n = 0, 1, 2, \ldots \end{cases}$$

The iteration can be performed graphically by drawing a tangent to $y = f(x)$ at $\big(x_n, f(x_n)\big)$ and determine the intersection point of this tangent and x-axis. The equation of this tangent is

$$y - f(x_n) = f'(x_n)(x - x_n)$$

and the intersection point has x-coordinate satisfying

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

so that

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

that is, $x = x_{n+1}$.

Example 9. $f(x) = x^3 - x - 1$

$$\begin{cases} x_0 = 1.3 \\ x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)} = x_n - \dfrac{x_n^3 - x_n - 1}{3x_n^2 - 1} = \dfrac{2x_n^3 + 1}{3x_n^2 - 1} \end{cases}$$

The computations are easier than in Example 5. If $x_0 = 1.3$, we get

$$x_1 = 1.3253 \quad \text{and} \quad x_2 = 1.32472.$$

Example 10. Computing of square roots.

$\sqrt{a}$ is a root of the equation

$$f(x) = x^2 - a = 0.$$

Here $f'(x) = 2x$, and the initial point can be guessed. Then

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

This algorithm is very practical; each step requires just one division and computation of one arithmetic mean.
If $a = 5$ and $x_0 = 2$, we get

$$x_1 = 2.25, \quad x_2 = 2.235, \quad x_3 = 2.2361.$$

Compare with the exact value $\sqrt{5} = 2.23607$.

Example 11. Reciprocals.

$\frac{1}{a}$ is the root of

$$f(x) = \frac{1}{x} - a = 0.$$

Now $f'(x) = -\frac{1}{x^2}$ and thus

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

If $a = 13$ and $x_0 = 0.1$ we get

$$x_1 = 0.07, \quad x_2 = 0.0763, \quad x_3 = 0.076918.$$

The exact value is $\frac{1}{13} = 0.076923$.

The convergence of Newton's method can be studied by using Theorem 3.

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

$$F'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}.$$

If f is two times differentiable, the iteration converges therefore is

$$|F'(x)| = \left| \frac{f(x)f''(x)}{f'(x)^2} \right| \leq m < 1$$

in a neighborhood of the root $\overline{x}$ and $|x_0 - \overline{x}|$ is sufficiently close to $\overline{x}$.

Theorem 4 can be applied to find error estimates. Then ε is an upperbound for the evaluation error in

$$F(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Since x is fixed, this evaluation error is equal to the evaluation error in $\frac{f(x_n)}{f'(x_n)}$.

Example 12. We apply Theorem 4 to Example 9.

$$\begin{aligned} f(x) &= x^3 - x - 1 \\ 1.3 \le \overline{x} &\le 1.4 \\ f(1.3) &= -0.103 < 0 \\ f(1.4) &= 0.344 > 0 \\ f'(x) = 3x^2 - 1; \quad f''(x) &= 6x \end{aligned}$$

As in Example 7 we estimate the derivatives between $x = 1.3$ and $x = 1.4$.

$$|F'(x)| = \left|\frac{f(x)f''(x)}{f'(x)^2}\right| \le \left|\frac{0.344 \cdot 6 \cdot 1.4}{4.1^2}\right| \le 0.2$$

If we assume that $\varepsilon = \frac{1}{2} \cdot 10^{-5}$, we obtain

$$\begin{aligned} |x_2 - \overline{x}| &\le \frac{0.2}{1 - 0.2} |1.32472 - 1.3253| + \frac{\frac{1}{2} \cdot 10^{-5}}{1 - 0.2} \\ &= \frac{1}{4} \cdot 0.00058 + \frac{1}{1.6} \cdot 10^{-5} \\ &= 0.000145 + 0.625 \cdot 10^{-5} \\ &\le 0.000152. \end{aligned}$$

Theorem 5. *Suppose that f is two times differentiable, $f(\overline{x}) = 0$ and $f'(\overline{x}) \neq 0$. If x_n is sufficiently close to $\overline{x}$, then there exists a number ξ between x_n and $\overline{x}$ such that*

$$|x_{n+1} - \overline{x}| = \left|\frac{f''(\xi)}{2f'(x_n)}\right| \cdot |x_n - \overline{x}|^2 \tag{10}$$

Proof. Let $\Delta x = \overline{x} - x_n$ (correction) so that $\overline{x} = x_n + \Delta x$. Taylor's formula $\Rightarrow$

$$0 = f(\overline{x}) = f(x_n + \Delta x) = f(x_n) + \Delta x f'(x_n) + \frac{\Delta x^2}{2} f''(\xi). \tag{11}$$

By hypothesis, f' is continuous and $f'(\overline{x}) \neq 0$. If x_n is sufficiently close to $\overline{x}$, then also $f'(x_n) \neq 0$. Then (11) $\Rightarrow$

$$\begin{aligned} 0 &= \frac{f(x_n)}{f'(x_n)} + \Delta x + \frac{\Delta x^2}{2} \frac{f''(\xi)}{f'(\xi)} \\ &= x_n - x_{n+1} + \overline{x} - x_n + \frac{(\overline{x} - x_n)^2}{2} \frac{f''(\xi)}{f'(\xi)} \\ \Rightarrow \quad x_{n+1} - \overline{x} &= \frac{(\overline{x} - x_n)^2 f''(\xi)}{2f'(x_n)} \quad \Rightarrow \ (10). \end{aligned}$$

□

Formula (10) shows that the convergence pf Newton's method is *quadratic*: the error in x_{n+1} is proportional to the square of the error in x_n. If the proportionality coefficient

$$\left| \frac{f''(\xi)}{2f'(x_n)} \right|$$

is of order 1, then the number of correct decimals will be doubled in each iteration step.

In general fixed point iteration convergence is *linear*; then (10) is replaced with

$$|x_{n+1} - \overline{x}| = |F'(\xi)||x_n - \overline{x}|$$

Order of convergence is defined as the largest value of p such that

$$\lim_{n\to\infty} \frac{|x_{n+1} - \overline{x}|}{|x_n - \overline{x}|^p} = c < \infty.$$

Order of linear convergence is 1 and order of quadratic convergence is 2.

Newton introduced his method in 1687 and it was later applied by Keppler to the equation $x - e\sin x = a$.

3.6 Secant method

If $f'(x_n)$ is replaced by the difference quotient

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

in Newton's method, we get the iteration formula

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} = \frac{x_n f(x_{n-1}) - x_{n-1}f(x_n)}{f(x_{n-1}) - f(x_n)} \qquad (12)$$

In contrast to the previous methods the iteration formula of this *secant method* is no longer of the form

$$x_{n+1} = F(x_n),$$

and we also need two initial values x_0 and x_1. The method can be used for example when the derivative $f'(x_n)$ is difficult to evaluate ($f(x_n)$ could e.g. be defined implicitly etc.).

The converge of the secant method is "superlinear", that is, faster than linear covergenge but slower than quadratic convergence. The order of convergence for the secant method is in fact the golden mean $\gamma = 1.618\ldots$.

If in (12) $\big(x_{n-1}, f(x_{n-1})\big)$ is replaced with $\big(x_0, f(x_0)\big)$, we obtain "classical Regula Falsi". Then all secants pass through a single point $\big(x_0, f(x_0)\big)$. Convergence is slower than in (12).

In another variation of (12) successive approximations x_n and x_{n-1} always lie on different sides of the root. If x_n and x_{n-1} are such approximations, then

$$f(x_n)f(x_{n-1}) < 0.$$

We compute x_{n+1} from (12) and find out whether $f(x_n)f(x_{n+1}) < 0$. If this is the case, we continue iteration. If $f(x_n)f(x_{n+1}) > 0$, then we replace x_n with x_{n-1} in the next iteration step. The advantage of this method is that at each step we get an error estimate, because the root always lies between x_n and x_{n-1} and also between x_{n+1} and x_n.

In the computation of the values of f we must, however, take into account the evaluation error. If the evaluation error in $f(x)$ is at most ε, then $f(x_n)$ is positive if the computed approximate value $\tilde{f}(x_n)$ satisfies $\tilde{f}(x_n) - \varepsilon > 0$.

3.7 Horner's scheme

Horner's scheme is a method to compute the value of a polynomial $p(x)$ and its derivative $p'(x)$ at a given point x_0. For example, the polynomial

$$p(x) = 2x^3 + 5x^2 - 4x + 3$$

can be written in the form

$$p(x) = \big((2x + 5)x - 4\big)x + 3. \tag{13}$$

The value of p at x_0 can be computed in three steps:

$$\begin{aligned} & & c_1 &= 2x_0 + 5 \\ & & c_2 &= c_1 x_0 - 4 \\ & & c_3 &= c_2 x_0 + 3 \\ (13) & \Rightarrow & p(x_0) &= c_3 \end{aligned}$$

General case: Suppose that

$$p(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n.$$

The value $p(x_0)$ can be computed recursively:

$$\begin{cases} c_0 = a_0 \\ c_k = c_{k-1} x_0 + a_k \quad (1 \leq k \leq n) \\ p(x_0) = c_n \end{cases} \tag{14}$$

The computation of $p(x_0)$ then requires n additions and n multiplications.

Claim. $p(x) = (x - x_0)q(x) + c_n$, where $q(x) = c_0 x^{n-1} + c_1 x^{n-2} + \cdots + c_{n-1}$.

Proof.

$$\begin{aligned} (x - x_0)q(x) + c_n &= xq(x) - x_0 q(x) + c_n \\ &= c_0 x^n + c_1 x^{n-1} + \cdots + c_{n-1} x \\ &\quad - x_0(c_0 x^{n-1} + c_1 x^{n-2} + \cdots + c_{n-1}) + c_n \\ &= c_0 x^n + (c_1 - c_0 x_0) x^{n-1} + (c_2 - c_1 x_0) x^{n-2} + \cdots \\ &\quad + (c_{n-1} - c_{n-2} x_0) x + (c_n - c_{n-1} x_0) \\ &= a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n \\ &= p(x). \end{aligned}$$

□

Differentiation:

$$\begin{aligned} & p'(x) &= q(x) + (x - x_0)q'(x) \\ \Rightarrow\ & p'(x_0) &= q(x_0) \end{aligned}$$

$p'(x_0)$ can therefore be computed by Horner's rule by replacing the coefficients a_k with the numbers $c_0, c_1, \ldots, c_{n-1}$:

$$\begin{cases} d_0 = c_0 \\ d_k = d_{k-1}x_0 + c_k \quad (1 \le k \le n-1) \\ p'(x_0) = d_{n-1} \end{cases} \tag{15}$$

Example 13. $p(x) = 2x^3 + 5x^2 - 4x + 3; x_0 = 1$. To compute $p(x_0)$ and $p'(x_0)$ we build Horner's scheme:

		2	5	−4	3
$c_{k-1}x_0 \to$	$x_0 = 1$		2	7	3
$c_k \to$		2	7	3	$6 = p(1)$
$d_{k-1}x_0 \to$	$x_0 = 1$		2	9	
$d_k \to$		2	9	12	$= p'(1)$

Example 14. Determine all real roots of the equation

$$p(x) = x^3 - 2x - 5 = 0.$$

By drawing the graph we see that there is only one real root $\overline{x}$ which lies between $x = 2$ and $x = 3$. Newton's method with $x_0 = 2$:

$$\begin{cases} x_0 = 2 \\ x_{n+1} = x_n - \dfrac{p(x_n)}{p'(x_n)} = x_n - \dfrac{x_n^3 - 2x_n - 5}{3x_n^2 - 2} \end{cases}$$

We use Horner's scheme for the computation of $p(2)$ and $p'(2)$:

	1	0	−2	−5
$x_0 = 2$		$+1 \cdot 2$	$+2 \cdot 2$	$+2 \cdot 2$
	1	2	2	$-1 = p(2)$
$x_0 = 2$		$+1 \cdot 2$	$+4 \cdot 2$	
	1	4	10	$= p'(2)$

Therefore

$$x_1 = x_0 - \frac{p(x_0)}{p'(x_0)} = 2 - \frac{p(2)}{p'(2)} = 2 - \frac{-1}{10} = 2.1$$

The next step yields the scheme

	1	0	-2	-5
$x = 2.1$		$+1 \cdot 2.1$	$+2.1 \cdot 2.1$	$+2.41 \cdot 2.1$
	1	2.1	2.41	$0.061 = p(2.1)$
$x = 2.1$		$+1 \cdot 2.1$	$+4.2 \cdot 2.1$	
	1	4.2	11.23	$= p'(2.1)$

$$x_2 = 2.1 - \frac{p(2.1)}{p'(2.1)} = 2.1 - \frac{0.061}{11.23} = 2.094568$$

Each iteration step after this yields tha same result $x_n = 2.094551$.

Remark. The example shows how Horner's scheme could be used in in connection with Newton's method. In this example the number of computations could be reduced, however, by writing the iteration formula in the form

$$x_{n+1} = x_n - \frac{x_n^3 - 2x_n - 5}{3x_n^2 - 2} = \frac{2x^3 + 5}{3x_n^2 - 2}.$$

3.8 Systems of nonlinear equations

We present an analytic derivation for Newton's method, which can be generalized for several variables.

Let $\overline{x}$ be a root of $f(x) = 0$. If x_n is a good approximation of $\overline{x}$, then $\Delta x = \overline{x} - x_n$ is small. Taylor's formula (4) $\Rightarrow$

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

If $f''(\xi)$ is not very large, then the last term is small, so that

$$\begin{aligned} 0 = f(\overline{x}) &\approx f(x_n) + \Delta x f'(x_n) \\ \Rightarrow \quad \Delta x &\approx -\frac{f(x_n)}{f'(x_n)} \end{aligned}$$

Hence

$$x_n + \Delta x \approx x_n - \frac{f(x_n)}{f'(x_n)} = x_{n+1},$$

Thus x_{n+1} should be a better approximation of $\overline{x}$ than x_n. This was the analytic derivation of Newton's method.

The problem for two variables is to look for approximate solutions of a system

$$\begin{cases} f(x,y) = 0 \\ g(x,y) = 0 \end{cases} \tag{19}$$

The iteration method will be found as before by using Taylor's formula.

Let $(\overline{x}, \overline{y})$ be an exact solution of (19) and let (x_n, y_n) be an approximation of $(\overline{x}, \overline{y})$. Then $\Delta x = \overline{x} - x_n$ and $\Delta y = \overline{y} - y_n$ are small and $\overline{x} = x_n + \Delta x$, $\overline{y} = y_n + \Delta y$. Taylor's formula (Analysis)

$$\begin{aligned} 0 = f(\overline{x}, \overline{y}) &= f(x_n + \Delta x, y_n + \Delta y) \\ &= f(x_n, y_n) + \Delta x D_1 f(x_n, y_n) + \Delta y D_2 f(x_n, y_n) + T_1 \\ &= f(x_n, y_n) + \Delta x \frac{\partial f}{\partial x}(x_n, y_n) + \Delta y \frac{\partial f}{\partial y}(x_n, y_n) + T_1 \end{aligned}$$

Similarly

$$0 = g(\overline{x}, \overline{y}) = g(x_n, y_n) + \Delta x \frac{\partial q}{\partial x}(x_n, y_n) + \Delta y \frac{\partial q}{\partial y}(x_n, y_n) + T_2$$

The remainder T_1 has the form

$$\begin{aligned} T_1 = \frac{1}{2} \Bigg(& \Delta x^2 \frac{\partial^2}{\partial x^2} f(x_n + \xi \Delta x, y_n + \xi \Delta y) \\ & + 2 \Delta x \Delta y \frac{\partial^2}{\partial x \partial y} f(x_n + \xi \Delta x, y_n + \xi \Delta y) \\ & + \Delta y^2 \frac{\partial^2}{\partial y^2} f(x_n + \xi \Delta x, y_n + \xi \Delta y) \Bigg) \quad (0 < \xi < 1) \end{aligned}$$

If these second partial derivatives are not very large, then T_1 (and similarly T_2) is small, so that approximately

$$\begin{cases} 0 = f(x_n, y_n) + \Delta x \dfrac{\partial f}{\partial x}(x_n, y_n) + \Delta y \dfrac{\partial f}{\partial y}(x_n, y_n) \\ 0 = g(x_n, y_n) + \Delta x \dfrac{\partial g}{\partial x}(x_n, y_n) + \Delta y \dfrac{\partial g}{\partial y}(x_n, y_n) \end{cases} \tag{20}$$

We solve Δx and Δy from (20) and set

$$x_{n+1} = x_n + \Delta x, \qquad y_{n+1} = y_n + \Delta y$$

This is Newton's method for the solution of (19).

Example 15.

$$\begin{cases} x^2 - y^2 = 1 \\ x^2 + y^2 = 4 \end{cases}$$

$$f(x,y) = x^2 - y^2 - 1, \quad \frac{\partial f}{\partial x} = 2x, \quad \frac{\partial f}{\partial y} = -2y$$

$$q(x,y) = x^2 + y^2 - 4, \quad \frac{\partial q}{\partial x} = 2x, \quad \frac{\partial q}{\partial y} = 2y$$

This system has a solution $\overline{x} = \sqrt{2.5} = 1.5811$ and $\overline{y} = \sqrt{1.5} = 1.2247$. The system (20) is now

$$\begin{cases} 0 = x_n^2 - y_n^2 - 1 + \Delta x \cdot 2x_n + \Delta y \cdot (-2y_n) \\ 0 = x_n^2 + y_n^2 - 4 + \Delta x \cdot 2x_n + \Delta y \cdot 2y_n \end{cases}$$

Solution:

$$2x_n^2 - 5 + 4x_n \cdot \Delta x = 0 \quad \Rightarrow \quad \Delta x = \frac{5 - 2x_n^2}{4x_n}$$

$$2y_n^2 - 3 + 4y_n \cdot \Delta y = 0 \quad \Rightarrow \quad \Delta y = \frac{3 - 2y_n^2}{4y_n}$$

Iteration formula:

$$\begin{cases} x_{n+1} = x_n + \Delta x = \dfrac{2x_n^2 + 5}{4x_n} = \dfrac{1}{2}x_n + \dfrac{5}{4} \cdot \dfrac{1}{x_n} \\ y_{n+1} = y_n + \Delta y = \dfrac{2y_n^2 + 3}{4y_n} = \dfrac{1}{2}y_n + \dfrac{3}{4} \cdot \dfrac{1}{y_n} \end{cases}$$

Starting from $x_0 = y_0 = 1.4$ we obtain

$$x_1 = \frac{1}{2} \cdot 1.4 + \frac{5}{4} \cdot \frac{1}{1.4} = 1.593$$
$$y_1 = \frac{1}{2} \cdot 1.4 + \frac{3}{4} \cdot \frac{1}{1.4} = 1.236$$
$$x_2 = 1.5812$$
$$y_2 = 1.2248$$

3.9 Ill-conditioned problems

Consider an equation $p(x) = C$, where p is a polynomial. Sometimes a small perturbation of the coefficients of p can result in a dramatic change in the roots of $p(x) = 0$.

For example, if

$$p(x) = (x-1)(x-2)\cdots(x-20) = x^{20} - 210x^{19} + \cdots + 20!,$$

then the roots of $p(x) = 0$ are $x = 1, x = 2, \ldots, x = 20$.

We change the coefficient -210 of x^{19} by replacing it with $-210 - 2^{-23}$. To find the roots of this kind of equation we need a much higher accuracy in computations than 2^{-23}.

4 Approximation

4.1 Introduction

There are a few reasons why approximation is to be referred.

1. Approximation of a given function by a simpler function makes some tasks easier, e.g. evaluation of the function values, numerical integration and differentiation

2. If the values of the function are originally known only at finitely many points (e.g. as measured quantities), then a suitably chosen continuous function could approximate the given function at points where the values are not known.

The approximating function could be e.g.

- a polynomial
- a rational function $\frac{P(x)}{Q(x)}$
- an exponential or a trigonometric function

Consider an approximation of a given function f by a rational function R. In order to get n correct decimals for the approximation on an interval $a \leq x \leq b$, we must have

$$\max_{a \leq x \leq b} |f(x) - R(x)| \leq \frac{1}{2} \cdot 10^{-n}.$$

We say that this number $\max_{a \leq x \leq b} |f(x) - R(x)|$ is the *distance* between f and R on $[a, b]$ or the *norm* of $f - R$ on $[a, b]$, denoted $\|f - R\|$.

Often in practice the values of f are known only at finitely many points $x_1, x_2, \ldots, x_n$ (e.g. as measured quantities). Then we could use the norm

$$\|f - R\| = \max_{1 \leq i \leq n} |f(x_i) - R(x_i)|$$

and try to determine a rational function R whose values *at the points* x_i are sufficiently close to $f(x_i)$.

Problem. Given f find a rational function R such that $\deg R \leq m$ and

$$\|f - R\| \qquad (*)$$

is as small as possible.

The solution depends on the definition of $\|f-R\|$. For example, if we use the above maximum norms, then accidental errors in individual measurements can have a large impact to the approximation R satisfying $(*)$. Therefore it is often better to measure the distance between f and R by using the so called *"least squares norm"*

$$\|f - R\| = \left[\sum_{i=1}^{n} \big(f(x_i) - R(x_i)\big)^2\right]^{\frac{1}{2}} \qquad (1)$$

Remark. The expression in (1) contains the square root in order that the axioms of the inner product space were satisfied. For example the norm (1) satisfies the triangle inequality

$$\|f + g\| \leq \|f\| + \|g\|,$$

where $\|f\| = \|f - 0\|$. Compare with the norm in $\mathbb{R}^2$: if (x_1, y_1) and (x_2, y_2) are points in $\mathbb{R}^2$, then

$$\|(x_1, y_1) - (x_2, y_2)\| = \left[(x_1 - x_2)^2 + (y_1 - y_2)^2\right]^{\frac{1}{2}}.$$

4.2 Polynomial approximation

Approximation by rational functions is in general a difficult task. In the sequel we mainly concentrate on approximation by polynomials. The following fundamental result states that a continuous function can be approximated by polynomials on an interval $[a, b]$ with an arbitrary high precision:

Weierstrass approximation theorem. If f is continuous on $[a, b]$ and $\varepsilon > 0$ then there exists a polynomial $p(x)$ such that

$$\max_{a \leq x \leq b} |f(x) - p(x)| \leq \varepsilon.$$

We will next consider four ways of approximating a given function:

- Taylor polynomial
- interpolation
- least squares
- splines

4.3 Taylor's expansions

If f is n times differentiable in a neighborhood of α, then f has a Taylor's expansion

$$f(x) = f\big(\alpha + (x-\alpha)\big) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(x-\alpha)^k + \frac{f^{(n)}(\xi)}{n!}(x-\alpha)^n,$$

where ξ is a point of the interval I whose end points are x and α. Here the *Taylor polynomial*

$$p(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(x-\alpha)^k$$

is a polynomial approximation to f of degree at most $n-1$. The error is

$$p(x) - f(x) = -\frac{f^{(n)}(\xi)}{n!}(x-\alpha)^n$$

and it has on an interval $a \le x \le b$ the upper bound

$$\|f-p\| = \max_{a\le x\le b} |f(x)-p(x)| \le \frac{\max|f^{(n)}(\xi)|}{n!} \max|x-\alpha|^n$$

In order to minimize this upper bound we should choose $\alpha = \frac{a+b}{2}$, so that α is the midpoint of $[a,b]$.

Example 1. We wish to approximate e^{-x} between $0 \le x \le 10$ to get three correct decimals.

Taylor's expansion at $\alpha = 0$:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$
$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots$$
$$e^{-x} = \sum_{k=0}^{n-1} \frac{(-1)^k}{k!} x^k + (-1)^n \frac{e^{-\xi}}{n!} x^n$$

We should choose n so that

$$\max_{0 \le x \le 10} \left| e^{-x} - \sum_{k=0}^{n-1} \frac{(-1)^k}{k!} x^k \right| \le \max_{0 \le x \le 10} \left| (-1)^n e^{-x} \right| \max_{0 \le x \le 10} \frac{|x|^n}{n!} = \frac{10^n}{n!} \le 0.5 \cdot 10^{-3}$$

The smallest n for which the last inequality is satisfied is $n = 32$.

Taylor's expansion at $\alpha = 5$ is

$$e^{-x} = e^{-[5+(x-5)]} = e^{-5} e^{-(x-5)} = \sum_{k=0}^{n-1} \frac{(-1)^k e^{-5}}{k!} (x-5)^k + (-1)^n \frac{e^{-\xi}}{n!} (x-5)^n$$

The error has an upper bound

$$\max_{0 \le x \le 10} \left| e^{-x} - \sum_{k=0}^{n-1} \frac{(-1)^k e^{-5}}{k!} (x-5)^k \right| \le \frac{5^n}{n!}$$

Here we should have $n \ge 19$ to get $\frac{5^n}{n!} \le \frac{1}{2} \cdot 10^{-3}$.

method	order of polynomial
Taylor: $\alpha = 0$	31
Taylor: $\alpha = 5$	18
other methods	5
rational approximation	5

Remark. Tayrlor's expansion does not in general yield an adequate polynomial approximation with respect to the maximum norm. This is quite natural because the Taylor polynomial depends only on local behaviour of f in a neighborhood of α. On the other hand, in such a small neighborhood of α the expansion can be a very good approximation to f.

Example 2. We wish to approximate $\arctan x$ on the interval $-1 \le x \le 1$ to get 6 correct decimals. Taylor's expansion at $\alpha = 0$:

$$\frac{d}{dx}\arctan x = \frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots$$

$$\Rightarrow \quad \arctan x = \int_0^x \frac{d}{dx}\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

The theorem of Leibnitz guarantees convergence and the remainder term has the estimate

$$\left|\arctan x - \sum_{k=0}^{n-1} \frac{(-1)^k}{2k+1} x^{2k+1}\right| \le \left|\frac{(1)^n}{2n+1} x^{2n+1}\right| \le \frac{1}{2n+1}$$

To get the desired accuracy $\frac{1}{2n+1} \le \frac{1}{2} \cdot 10^{-6}$ we should use a polynomial of degree 1 999 999.

By other methods it is possible to find a polynomial of degree 15 which gives the desired accuracy. A rational approximation of the form

$$R(x) = \frac{a_0 x + a_1 x^3 + a_2 x^5}{1 + a_3 x^2 + a_4 x^4}$$

can also give the desired accuracy, so that

$$\max_{|x| \le 1} |\arctan x - R(x)| \le \frac{1}{2} \cdot 10^{-6}.$$

4.4 Interpolation

Assume that function values $f_i = f(x_i)$ are known for different points $x_0, x_1, \ldots, x_n$. We want to determine a polynomial p_n of degree at most n such that

$$p_n(x_0) = f_0, \ p_n(x_1) = f_1, \ldots, p_n(x_n) = f_n.$$

Such a polynomial is called *interpolation polynomial* and points x_j are called *nodes*. The numbers f_j relating to the nodes can be the values of a known function f (e.g. $\ln x, \sin x$, etc.) such that $f(x_j) = f_j$; then $p_n(x)$ is the *polynomial approximation of* f which has the same values as f in the given points.

The numbers f_j can also be measurements or observed values. In order to get approximate values for f at point x outside the measured points, we use the polynomial $p_n(x)$. If x lies on the interval formed by $x_0, \ldots, x_n$, this is called *interpolation*, otherwise *extrapolation*. In extrapolation, however, the approximation does not usually have good accuracy.

We will later see, that there exists such an interpolation polynomial of degree at most n and it is unique. The uniqueness is follows from the fact that the difference of two possible interpolation polynomials $d_n = p_n - q_n$ is a polynomial of degree at most n which has at least $n + 1$ roots [at nodes $x_0, x_1, \ldots, x_n$ we have by assumption $p_n(x) = q_n(x)$]; therefore d_n vanishes identically and hence $p_n(x) \equiv q_n(x)$.

We will next study different methods for finding $p_n(x)$. All of these methods give a unique polynomial but the forms of the polynomial and the number of required computations vary in different methods.

Special cases: Linear and quadratic interpolation

In linear interpolation we use a straight line passing through (x_0, f_0) and (x_1, f_1) and the assiciated interpolation polynomial is

$$p_1(x) = f_0 + (x - x_0) f[x_0, x_1] \tag{1}$$

where

$$f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} \tag{2}$$

is the *first divided difference* (ensimmäinen jaettu erotus).

From (1) we see that $p_1(x_0) = f_0$, and by (1) and (2) we see that $p_1(x_1) = f_1$.

Example 1. Estimate the finnish population in 1978 by using the following data:

Year	1970	1982
Population (1000)	4598	4842

Solution:

$$p_1(1978) = 4598 + (1978 - 1979)\frac{4842 - 4598}{1982 - 1970} = 4761 \quad \text{(correct 4758)}$$

In *quadratic interpolation* the approximating polynomial $p_2(x)$ is of degree ≤ 2 and its graph passes through the points (x_0, f_0), (x_1, f_1) and (x_2, f_2):

$$p_2(x) = f_0 + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \tag{3}$$

where

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \tag{4}$$

is the *second divided difference*.

From (3) we see that $p_2(x_0) = f_0$ and $p_2(x_1) = f_0 + (x_1 - x_0)\frac{f_1 - f_0}{x_1 - x_0} = f_1$; in addition $p_2(x_2) = f_2$.

Example 2. We know $\ln 8.0 = 2.0794$, $\ln 9.0 = 2.1972$ and $\ln 9.5 = 2.2513$ but what is $\ln 9.2$? We compute divided differences from (2) and (4):

$$\begin{array}{ll} x_0 = 8.0,\ f_0 = 2.0794 & \\ x_1 = 9.0,\ f_1 = 2.1972 & f[x_0, x_1] = 0.1178 \\ x_2 = 9.5,\ f_2 = 2.2513 & f[x_1, x_2] = 0.1082 \end{array} \qquad f[x_0, x_1, x_2] = -0.0064$$

$$\begin{aligned} (3) \Rightarrow\ p_2(x) &= 2.0794 + (x - 8.0) \cdot 0.1178 + (x - 8.0)(x - 9.0) \cdot (-0.0064) \\ &= 0.6762 + 0.2266x - 0.0064x^2 \end{aligned}$$

$p_2(9.2) = 2.2192$ has four correct decimals.

Newton's interpolation and divided differences

Formulas (1) and (3) are special cases of a more general interpolation formula. Note that p_2 is obtained from p_1 by adding the last term of (3) to p_1. We try to accomplish a similar situation in the general case:

$$p_n(x) = p_{n-1}(x) + g_n(x) \tag{5}$$

where $p_{n-1}(x_0) = p_n(x_0) = f_0, \ldots, p_{n-1}(x_{n-1}) = p_n(x_{n-1}) = f_{n-1}$ and $p_n(x_n) = f_n$. In order to determine

$$g_n(x) = p_n(x) - p_{n-1}(x) \tag{5'}$$

note that in view of the above conditions g_n vanishes at $x_0, \ldots, x_{n-1}$. Since in addition $\deg g_n \leq n$ (because $\deg p_n \leq n$, $\deg p_{n-1} \leq n-1$), g_n should be of the form

$$g_n(x) = a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}). \tag{5"}$$

To determine a_n we substitute $x = x_n$ and solve (5") with respect to a_n. Since by (5') $g_n(x_n) = p_n(x_n) - p_{n-1}(x_n) = f_n - p_{n-1}(x_n)$, the result is

$$a_n = \frac{f_n - p_{n-1}(x_n)}{(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1})} \tag{6}$$

For $n = 1$ we have $p_{n-1}(x_1) = p_0(x_1) = f_0$, so that by (6)

$$a_1 = \frac{f_1 - p_0(x_1)}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0} = f[x_0, x_1]$$

and (5) yields formula (1). Similarly, for $n = 2$ we obtain (3) because in (6)

$$f_2 - p_1(x_2) \overset{(1)}{=} f_2 - f_0 - (x_2 - x_0)f[x_0, x_1]$$

so that

$$a_2 = \frac{f_2 - f_0 - (x_2 - x_0)f[x_0, x_1]}{(x_2 - x_0)(x_2 - x_1)} = f[x_0, x_1, x_2]$$

Thus (3) follows from (5) for $n = 2$, when we substitute $g_2(x)$ according to (5"). Similarly

$$a_3 = f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$$

and in general

$$a_k = f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{(x_k - x_0)} \tag{7}$$

Here $f[x_0, \ldots, x_k]$ is *the k:th divided difference.*

The sketch of the proof is as follows. Let $p_{1,n}$ be the interpolation polynomial for the nodes $x_1, \ldots, x_n$ and let $p_{0,n-1}$ be the interpolation polynomial for the nodes $x_0, \ldots, x_{n-1}$. Then

$$p_n(x) = \frac{(x - x_0)p_{1,n}(x) - (x - x_n)p_{0,n-1}(x)}{x_n - x_0}.$$

For $n = k$ we obtain from formula (5)

$$p_k(x) = p_{k-1}(x) + (x - x_0)(x - x_1) \cdots (x - x_{k-1}) f[x_0, \ldots, x_k] \tag{8}$$

For $k = 1, \ldots, n$ we first obtain (1) and (3) (for $k = 1$ and $k = 2$), and finally for $k = n$ we obtain *Newton's interpolation polynomial*

$$\begin{aligned} p_n(x) = f_0 + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ + \cdots + (x - x_0) \cdots (x - x_{n-1}) f[x_0, \ldots, x_n] \end{aligned} \tag{9}$$

In practice one often uses the following *divided difference scheme.*

Example 3. Approximate $\ln 9.2$ by using the given function values.

x_j	$f_j = f(x_j)$	$f[x_j, x_{j+1}]$	$f[x_j, x_{j+1}, x_{j+2}]$	$f[x_j, \ldots, x_{j+3}]$
8.0	2.079442			
		0.117783		
9.0	2.197225		−0.006433	
		0.108134		0.000411
9.5	2.251292		−0.005200	
		0.097735		
11.0	2.397895			

For example

$$-0.005200 = \frac{0.097735 - 0.108134}{11 - 9}.$$

Now formula (9) yields

$$\begin{aligned} p_3(x) = 2.079442 + 0.117783(x - 8.0) - 0.006433(x - 8.0)(x - 9.0) \\ + 0.000411(x - 8.0)(x - 9.0)(x - 9.5) \end{aligned}$$

For $x = 9.2$ we get

$$f(9.2) = \ln 9.2 \approx 2.079442 + 0.141340 - 0.001544 - 0.000030 = 2.219208$$

The exact value $\ln 9.2 = 2.219203$. The accuracy is increasing with k because

$$p_1(9.2) = 2.220782; \quad p_2(9.2) = 2.219238; \quad p_3(9.2) = 2.219208.$$

Newton's forward difference formula

In (9) the nodes are arbitrary. In many applications the distance between consecutive nodes is however a constant h, so that

$$x_1 = x_0 + h, \quad x_2 = x_0 + 2h, \ \ldots \ , x_n = x_0 + nh \tag{10}$$

Then (7) and (9) can be written by using *forward differences* as follows.

The first forward difference of f at x_j is defined as

$$\Delta f_j = f_{j+1} - f_j;$$

the second forward difference of f at x_j is

$$\Delta^2 f_j = \Delta f_{j+1} - \Delta f_j;$$

etc.; in general the k:th forward difference at x_j is

$$\Delta^k f_j = \Delta^{k-1} f_{j+1} - \Delta^{k-1} f_j \qquad (k = 1, 2, \ldots) \tag{11}$$

If (10) holds, we can prove the following relation between divided differences and forward differences:

$$f[x_0, \ldots, x_k] = \frac{1}{k!h^k} \Delta^k f_0. \tag{12}$$

Proof is by induction. For $k = 1$ we have

$$f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{1}{h} [f_1 - f_0] = \frac{1}{1!h} \Delta f_0$$

If (12) holds for k, then

$$f[x_1, \ldots, x_{k+1}] = \frac{1}{k!h^k} \Delta^k f_1,$$
$$f[x_0, \ldots, x_k] = \frac{1}{k!h^k} \Delta^k f_0$$

For $k + 1$ we obtain from (7)

$$\begin{aligned} f[x_0, \ldots, x_{k+1}] &= \frac{f[x_1, \ldots, x_{k+1}] - f[x_0, \ldots, x_k]}{x_{k+1} - x_0} \\ &= \frac{1}{x_{k+1} - x_0} \left[\frac{1}{k!h^k} \Delta^k f_1 - \frac{1}{k!h^k} \Delta^k f_0 \right] \end{aligned}$$

Here in view of (10) $x_{k+1} - x_0 = (k+1)h$, so that

$$f[x_0, \ldots, x_{k+1}] = \frac{1}{(k+1)h} \cdot \frac{1}{k!h^k} \left(\Delta^k f_1 - \Delta^k f_0\right) = \frac{1}{(k+1)!h^{k+1}} \left(\Delta^{k+1} f_0\right)$$

Hence (12) holds even for $k+1$. QED.

Next we put in (9) $x = x_0 + rh$ where r is a real number. Then $x - x_0 = rh$, $x - x_1 = (r-1)h$ (since $x_1 - x_0 = h$) etc. By using (12) we then obtain Newton's forward formula:

$$\begin{aligned} p_n(x) &= f_0 + rh\frac{1}{1!h^1}\Delta f_0 + rh(r-1)h\frac{1}{2!h^2}\Delta^2 f_0 + \cdots \\ &\quad + rh(r-1)h\cdots(r-n+1)h\frac{1}{n!h^n}\Delta^n f_0 \\ &= f_0 + \binom{r}{1}\Delta f_0 + \binom{r}{2}\Delta^2 f_0 + \cdots + \binom{r}{n}\Delta^n f_0 \end{aligned}$$

where the *generalized binomial coefficients* are defined

$$\binom{r}{0} = 1, \quad \binom{r}{s} = \frac{r(r-1)\cdots(r-s+1)}{s!} \qquad (s > 0 \text{ integer}) \tag{13}$$

$$\begin{aligned} p_n(x) &= \sum_{s=0}^{n} \binom{r}{s}\Delta^s f_0 \quad \left(x = x_0 + rh,\ r = \frac{x - x_0}{h}\right) \\ &= f_0 + r\Delta f_0 + \frac{r(r-1)}{2!}\Delta^2 f_0 + \cdots + \frac{r(r-1)\cdots(r-n+1)}{n!}\Delta^n f_0 \end{aligned} \tag{14}$$

Suppose that f is $n+1$ times differentiable. If f is approximated by p_n at x, then one can show that the error in the approximation

$$\begin{aligned} \varepsilon(x) = p_n(x) - f(x) &= -\frac{1}{(n+1)!}(x - x_0)\cdots(x - x_n)f^{(n+1)}(t) \\ &= -\frac{h^{n+1}}{(n+1)!}r(r-1)\cdots(r-n)f^{(n+1)}(t) \end{aligned} \tag{15}$$

where t is in each interval containing x and each x_j.

The idea of the proof relies on Rolle's theorem. Define an auxiliary function

$$Q_n(x) = f(x) - p_n(x) - \gamma(x - x_0)\cdots(x - x_n)$$

so that $Q_n(\overline{x}) = 0$. Then

$$\begin{aligned} &Q_n' \text{ has } n+1 \text{ zeros} \\ &Q_n'' \text{ has } n \text{ zeros} \\ &\vdots \\ &Q_n^{(n+1)} \text{ has 1 zero} \end{aligned}$$

Therefore

$$Q_n^{(n+1)}(t) = f^{(n+1)}(t) - \gamma(n+1)! = 0 \quad \Rightarrow \quad \gamma = \frac{f^{(n+1)}(t)}{(n+1)!}$$

Example 4. Compute $\cosh 0.56$ by using (14) and the following table.

j	x_j	$f_j = \cosh x_j$	Δf_j	$\Delta^2 f_j$	$\Delta^3 f_j$
0	0.5	1.127626			
			0.057839		
1	0.6	1.185465		0.011865	
			0.069704		0.000697
2	0.7	1.255169		0.012562	
			0.082266		
3	0.8	1.337435			

Formula (14) gives

$$\begin{aligned} \cosh 0.56 &\approx 1.127626 + 0.6 \cdot 0.057839 + \frac{0.6(-0.4)}{2} \cdot 0.011865 \\ &\quad + \frac{0.6(-0.4)(-1.4)}{6} \cdot 0.000697 \\ &= 1.127626 + 0.034703 - 0.001424 + 0.000039 \\ &= 1.160944 \end{aligned}$$

Error estimate: Since $\frac{d^4}{dt^4} \cosh t = \cosh t$, from (15) we obtain

$$\varepsilon_3(0.56) = -\frac{0.1^4}{4!} \cdot 0.6(-0.4)(-1.4)(-2.4) \cosh t = A \cosh t,$$

where $A = 0.00000336$ and $0.5 \le t \le 0.8$. Since $\cosh t$ is increasing on the interval $[0.5, 0.8]$ it follows that

$$A \cosh 0.5 \le \varepsilon_3(0.56) \le A \cosh 0.8$$

Therefore

$$p_3(0.56) - A\cosh 0.8 \leq f(x) = \cosh 0.56 \leq p_3(0.56) - A\cosh 0.5$$

Computing the values of the upper and lower bounds we obtain

$$1.160939 \leq \cosh 0.56 \leq 1.160941$$

The exact value (6 decimals) is $\cosh 0.56 = 1.160941$.

Lagrange's interpolating polynomial

Consider the general interpolation problem where the nodes x_j need not be equally spaced. Define polynomials $l_0(x), l_1(x), \ldots, l_n(x)$ such that

$$\begin{aligned} l_0(x) &= (x - x_1)(x - x_2)\cdots(x - x_n) \\ l_k(x) &= (x - x_0)\cdots(x - x_{k-1})(x - x_{k+1})\cdots(x - x_n) \\ l_n(x) &= (x - x_0)(x - x_1)\cdots(x - x_{n-1}) \end{aligned}$$

Lagrange's interpolating polynomial is

$$L_n(x) = \sum_{k=0}^{n} \frac{l_k(x)}{l_k(x_k)} f_k \tag{19}$$

Each term in the sum is a polynomial of degree $\leq n$ taking at x_k the value f_k and vanishing at all other nodes. Therefore $L_n(x_k) = f_k$ for each k, so that L_n is the interpolating polynomial associated with the given data (x_k, f_k).

Example 7. Find $\ln 9.2$ by applying the Lagrange interpolating polynomial and the values of the following table:

x	9.0	9.5	10.0	11.0
$\ln x$	2.19722	2.25129	2.30259	2.39790

$$(19) \Rightarrow \quad L_3(x) = \frac{l_0(x)}{l_0(x_0)} f_0 + \frac{l_1(x)}{l_1(x_1)} f_1 + \frac{l_2(x)}{l_2(x_2)} f_2 + \frac{l_3(x)}{l_3(x_3)} f_3,$$

where

$$\begin{aligned} l_0(x) &= (x - 9.5)(x - 10)(x - 11) \\ l_1(x) &= (x - 9)(x - 10)(x - 11) \text{ etc.} \end{aligned}$$

Therefore

$$\begin{aligned}\ln 9.2 &\approx \frac{-0.43200}{-1.00000}\cdot 2.19722 + \frac{0.28800}{0.37500}\cdot 2.25129 \\ &\quad + \frac{0.10800}{-0.50000}\cdot 2.30259 + \frac{0.04800}{3.00000}\cdot 2.39790 \\ &= 2.21920 \quad \text{(5 correct decimals)}\end{aligned}$$

The use of Lagrange polynomials in numerical work is not recommended, because the computations are laborious and previous work is wasted in the transition to a polynomial of higher degree. However, Lagrange polynomials have considerable theoretical interest.

4.5 Least squares

Example 1. Given 5 points (x_i, y_i) in the plane listed below.

i	x_i	y_i
1	1	1.0
2	1.5	1.7
3	2.0	2.2
4	2.5	2.5
5	3	2.5

We look for a quadratic polynomial $p_2(x) = a + bx + cx^2$ such that

$$\sum_{i=1}^{5} \big(y_i - p_2(x_i)\big)^2$$

is as small as possible. In other words, we wish to find constants a, b and c such that

$$F(a,b,c) = \sum_{i=1}^{5} \big(y_i - a - bx_i - cx_i^2\big)^2$$

is as small as possible. Necessary condition for a minimum

$$\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = \frac{\partial F}{\partial c} = 0$$

leads to the so called *normal equations*

$$\begin{cases} -2\sum_{i=1}^{5}(y_i - a - bx_i - cx_i^2) = 0 \\ -2\sum_{i=1}^{5} x_i(y_i - a - bx_i - cx_i^2) = 0 \\ -2\sum_{i=1}^{5} x_i^2(y_i - a - bx_i - cx_i^2) = 0 \end{cases}$$

Matrix form:

$$\begin{pmatrix} \sum 1 & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{pmatrix}$$

All indices range from 1 to 5. The coefficient matrix is symmetric and one can show that the system has a unique solution which minimizes $F(a, b, c)$.

The coefficient matrix and the vector on the right-hand side can be computed by using the table:

x_i	y_i	x_i^2
1	1.0	1
1.5	1.7	2.25
2.0	2.2	4
2.5	2.5	6.25
3	2.5	9

We first compute the sums

$$\sum_{i=1}^{5} x_i = 10, \quad \sum_{i=1}^{5} y_i = 9.9, \quad \sum_{i=1}^{5} x_i^2 = 22.5$$

Then we compute inner products of the columns:

$$\sum_{i=1}^{5} x_i^3 = 55, \quad \sum_{i=1}^{5} x_i^4 = 142.125, \quad \sum_{i=1}^{5} x_i y_i = 21.7, \quad \sum_{i=1}^{5} x_i^2 y_i = 51.75$$

Since $\sum_{i=1}^{5} 1 = 5$, the normal equations are

$$\begin{pmatrix} 5 & 10 & 22.5 \\ 10 & 22.5 & 55 \\ 22.5 & 55 & 142.125 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 9.9 \\ 21.7 \\ 51.75 \end{pmatrix}$$

The system is solved e.g. by Gauss elimination method. The result is

$$a = -1.1400, \quad b = 2.5886, \quad c = -0.4571.$$

Example 2. The same problem as in Example 1 but we write $p_2(x)$ in the form

$$p_2(x) = a' + b'(x-2) + c'(x-2)^2.$$

Then we obtain the system

$$\begin{pmatrix} \sum 1 & 0 & \sum (x_i-2)^2 \\ 0 & \sum (x_i-2)^2 & 0 \\ \sum (x_i-2)^2 & 0 & \sum (x_i-2)^4 \end{pmatrix} \begin{pmatrix} a' \\ b' \\ c' \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum (x_i-2)y_i \\ \sum (x_i-2)^2 y_i \end{pmatrix}$$

Some of these coefficients vanish by symmetry. From the second equation we get b' while the first and the third equations form a 2×2 system for a' and c'.

The Examples 1 and 2 deal with *least squares approximation*. The function $y = f(x)$ is approximated by a polynomial $p_2(x)$ such that the sum of the squares of the errors $-(f(x) - p_2(x))$ at $x_1, x_2, \ldots, x_5$ is as small as possible.

Instead of $F(a, b, c)$ we could have tried to minimize

$$F_1(a, b, c) = \sum_{i=1}^{5} |y_i - a - bx_i - cx_i^2|$$

or the "maximum norm"

$$F_2(a, b, c) = \max_{1 \le i \le 5} |y_i - a - bx_i - cx_i^2|.$$

Such approximation problems are far more difficult than least squares approximation and will not be considered.

Remark. There is no such a polynomial of degree 2 which attains the values y_i at x_i for all i in Example 1 . Hence we are not looking for an interpolation polynomial but rather an approximation by polynomials.

If we wish to approximate a given function f with a polynomial of degree n, $n \neq 2$, the method is analogous to those in Examples 1 and 2. The method yields $n+1$ normal equations since a polynomial of degree n has $n+1$ coefficients. In the case $n = 1$ the method is called *linear least squares*, as the graph of the approximating polynomial is a straight line.

If the function f is known on the whole interval $[a, b]$ and we wish to perform a least squares approximation of f with a polynomial of degree n, we can try to minimize the integral

$$F(c_0, c_1, \ldots, c_n) = \int_a^b \Big(f(x) - \sum_{k=0}^{n} c_k x^k \Big)^2 dx.$$

Here the "error" is measured with an integral instead of a sum.

Solving the normal equations numerically can be difficult since small perturbations in coefficients can greatly change the solutions of the equations. Therefore it is usually preferred to use other methods for finding an approximation if n is large.

One such method is to replace the polynomial $\sum_{k=0}^{n} c_k x^k$ with a linear combination

$$p(x) = \sum_{k=0}^{n} a_k L_k(x),$$

where the polynomials $L_k(x)$ of degree k have been chosen so that the coefficient matrix in the normal equations in diagonalized. Such polynomials are called *orthogonal polynomials* relating to the given approximation task.

We will illustrate the above method with an example. For simplicity, we use the "integral norm"

$$\|f - p\| = \Big(\int_{-1}^{1} \big(f(x) - p(x) \big)^2 dx \Big)^{\frac{1}{2}},$$

but similarly we could use the "sum norm"

$$\|f-p\| = \Big(\sum_{i=1}^{m} \big(f(x_i)-p(x_i)\big)^2\Big)^{\frac{1}{2}}.$$

Example 3. We wish to approximate $\sin \pi x$ by a cubic polynomial on the interval $-1 \le x \le 1$ so that the integral

$$\int_{-1}^{1} \left(\sin \pi x - \sum_{k=0}^{3} c_k x^k \right)^2 dx$$

is as small as possible. We write

$$p(x) = \sum_{k=0}^{3} a_k L_k(x),$$

where

$$L_0(x) = 1, \quad L_1(x) = x, \quad L_2(x) = \frac{3}{2}x^2 - \frac{1}{2} \quad \text{and} \quad L_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x.$$

In order to minimize

$$F(a_0, a_1, a_2, a_3) = \int_{-1}^{1} \left(\sin \pi x - \sum_{k=0}^{3} a_k L_k(x) \right)^2 dx$$

we set

$$\frac{\partial F}{\partial a_k} = 0 \qquad (k = 0, 1, 2, 3).$$

Then we obtain the following normal equations:

$$\begin{cases} 2\int_{-1}^{1} L_0\Big(\sin\pi x - \sum_{k=0}^{3} a_k L_k(x)\Big)dx = 0 \\ 2\int_{-1}^{1} L_1\Big(\sin\pi x - \sum_{k=0}^{3} a_k L_k(x)\Big)dx = 0 \\ 2\int_{-1}^{1} L_2\Big(\sin\pi x - \sum_{k=0}^{3} a_k L_k(x)\Big)dx = 0 \\ 2\int_{-1}^{1} L_3\Big(\sin\pi x - \sum_{k=0}^{3} a_k L_k(x)\Big)dx = 0 \end{cases} \tag{1}$$

(1) can also be written as

$$\begin{cases} a_0\int_{-1}^{1} L_0^2dx + a_1\int_{-1}^{1} L_0L_1dx + a_2\int_{-1}^{1} L_0L_2dx + a_3\int_{-1}^{1} L_0L_3dx = \int_{-1}^{1} L_0\sin\pi xdx \\ a_0\int_{-1}^{1} L_1L_0dx + a_1\int_{-1}^{1} L_1^2dx + a_2\int_{-1}^{1} L_1L_2dx + a_3\int_{-1}^{1} L_1L_3dx = \int_{-1}^{1} L_1\sin\pi xdx \\ a_0\int_{-1}^{1} L_2L_0dx + a_1\int_{-1}^{1} L_2L_1dx + a_2\int_{-1}^{1} L_2^2dx + a_3\int_{-1}^{1} L_2L_3dx = \int_{-1}^{1} L_2\sin\pi xdx \\ a_0\int_{-1}^{1} L_3L_0dx + a_1\int_{-1}^{1} L_3L_1dx + a_2\int_{-1}^{1} L_3L_2dx + a_3\int_{-1}^{1} L_3^2dx = \int_{-1}^{1} L_3^2\sin\pi xdx \end{cases} \tag{2}$$

The coefficient matrix of (2) is a diagonal matrix because

$$\int_{-1}^{1} L_iL_jdx = 0, \quad \text{when } i \neq j$$

and

$$\int_{-1}^{1} L_i^2 dx = \frac{1}{i+\frac{1}{2}} \quad (0 \le i \le 3).$$

The solution is then

$$a_k = (k+\frac{1}{2}) \int_{-1}^{1} L_k(x) \sin \pi x dx \qquad (0 \le k \le 3)$$

For example,

$$a_0 = \frac{1}{2} \int_{-1}^{1} \sin \pi x dx = 0; \quad a_1 = \frac{3}{2} \int_{-1}^{1} x \sin \pi x dx = \frac{3}{\pi} \quad \text{etc.}$$

Remark. The polynomial $L_k(x)$ are *Legendre polynomials*. They can be defined by the formula

$$L_{k+1}(x) = \frac{2k+1}{k+1} x L_k(x) - \frac{k}{k+1} L_{k-1}(x); \quad L_0(x) = 1, \quad L_1(x) = x$$

4.6 Splines

Increasing the degree of an interpolation polynomial does not always increase accuracy of the interpolation. For example, consider function

$$f(x) = \frac{1}{1+25x^2}$$

on the interval $[-1, 1]$. The nodes of the interpolation polynomials of function f can be chosen such that the maximum error in the interpolation tends to infinity as the degree of polynomial increases. This kind of instability can be avoided by using *splines*.

Splines are piecewise defined polynomials, i.e. continuous functions whose restriction to interval of two consecutive knots is a polynomial. In addition, it is required that the function is sufficiently many times differentiable at knots. Interpolating with such functions is usually numerically stable.

The simplest example of approximation with piecewise defined polynomials is to use straight lines on each subinterval (e.g. trapezoidal rule). However, the graph of such function is not smooth since its first derivative is discontinuous at knots between the end points of the interval. By requiring continuous derivatives, we end up using splines.

In practise, the most important splines are *cubic splines* which are defined as follows. Let f be the given function we wish to approximate on the interval $a \leq x \leq b$. Suppose that the interval is divided into subintervals whose end points are so called *knots*

$$a = x_0 < x_1 < \cdots < x_n = b. \tag{1}$$

The cubic spline relating to knots (1) is a continuously differentiable function on the interval $a \leq x \leq b$ whose restriction to each subinterval with two consecutive knots as end points is a polynomial of degree at most 3. The *cubic spline interpolating function* f is gained by requiring that

$$g(x_0) = f(x_0) = f_0,\ g(x_1) = f(x_1) = f_1, \ldots, g(x_n) = f(x_n) = f_n. \tag{2}$$

Next we will show that a cubic spline satisfying these conditions always exists. It is not unique, however, as the derivatives at the end points of the interval $a \leq x \leq b$ can be defined arbitrarily:

Theorem 1. *Let f be a function defined on the interval $a \leq x \leq b$. Suppose that the knots (1) of the interval are given and let k_0 and k_n be arbitrary real numbers. Then there exits a unique cubic spline g relating to knots (1) such that (2) holds and*

$$g'(x_0) = k_0, \quad g'(x_n) = k_n. \tag{3}$$

Proof. It is easy to see that on each subinterval $I_j = [x_j, x_{j+1}]$ there exists a unique polynomial $p_j(x)$ of degree ≤ 3 such that

$$p_j(x_j) = f(x_j), \quad p_j(x_{j+1}) = f(x_{j+1}) \tag{4}$$

and p_j' has prescribed values at the end points:

$$p_j'(x_j) = k_j, \quad p_j'(x_{j+1}) = k_{j+1} \tag{5}$$

The expression of p_j has the following form where $c_j = \frac{1}{x_{j+1}-x_j}$

$$\begin{aligned} p_j(x) &= f(x_j)c_j^2(x-x_{j+1})^2[1+2c_j(x-x_j)] \\ &\quad + f(x_{j+1})c_j^2(x-x_j)^2[1-2c_j(x-x_{j+1})] \\ &\quad + k_jc_j^2(x-x_j)(x-x_{j+1})^2 + k_{j+1}c_j^2(x-x_j)^2(x-x_{j+1}) \end{aligned} \tag{6}$$

Uniqueness will be left as an exercise.

The restriction of the cubic spline $g(x)$ to I_j will be of the form (6). Since $g(x)$ should be twice continuously differentiable even at each node x_j, we must have

$$p''_{j-1}(x_j) = p''_j(x_j) \qquad (1 \le j \le n-1) \tag{7}$$

By differentiating we get

$$\begin{aligned} p''_j(x) &= 2f(x_j)c_j^2[1+4c_j(x-x_{j+1})+2c_j(x-x_j)] \\ &\quad + 2f(x_{j+1})c_j^2[1-4c_j(x-x_j)-2c_j(x-x_{j+1})] \\ &\quad + (4k_j+2k_{j+1})c_j^2(x-x_{j+1}) + (4k_{j+1}+2k_j)c_j^2(x-x_j) \end{aligned} \tag{8}$$

From (7) we then obtain

$$\begin{aligned} &6f(x_{j-1})c_{j-1}^2 - 6f(x_j)c_{j-1}^2 + (4k_j+2k_{j-1})c_{j-1} \\ &= -6f(x_j)c_j^2 + 6f(x_{j+1})c_j^2 - (4k_j+2k_{j+1})c_j \quad (i \le j \le n-1) \end{aligned} \tag{9}$$

If we denote $\Delta f_{j-1} = f(x_j) - f(x_{j-1})$ and $\Delta f_j = f(x_{j+1}) - f(x_j)$ as in § 4.4, (9) can be written

$$c_{j-1}k_{j-1} + 2(c_{j-1}+c_j)k_j + c_jk_{j+1} = 3(c_{j-1}^2\Delta f_{j-1} + c_j^2\Delta f_j) \tag{10}$$

These equations (10) form a system of $n-1$ linear equations in $n-1$ unknowns $k_1, \ldots, k_{n-1} \quad (1 \le j \le n-1)$. The coefficient matrix of this system is non-singular (without proof). Therefore (10) has a unique solution $k_1, \ldots, k_{n-1}$. Substitutions of this solution in (6) yields polynomials $p_0, p_1, \ldots, p_{n-1}$ defining the unique interpolating cubic spline $g(x)$ satisfying (3) and (7). □

The proof of Theorem 1 provides an algorithm for the determination of the spline. For simplicity we consider the case where the distance of consecutive nodes is a constant h, so that

$$x_1 = x_0 + h, \quad x_2 = x_0 + 2h, \ \ldots, \ x_n = nh$$

Then $c_j = \frac{1}{x_{j+1}-x_j} = \frac{1}{h}$, so that multiplying (10) by h and denoting $f(x_j) = f_j$ we obtain

$$k_{j-1} + 4k_j + k_{j+1} = \frac{3}{h}(f_{j+1} - f_{j-1}) \quad (i \le j \le n-1) \tag{11}$$

Here k_0 and k_n are given , e.g. $k_0 = f'(a)$, $k_n = f'(b)$. In the first step of the algorithm we solve $k_1, \ldots, k_{n-1}$ from the linear system (11). In the next step we determine the coefficients of the spline $g(x)$. On each interval $x_j \le x \le x_{j+1} = x_j + h$ the spline $g(x)$ has a Taylor expansion

$$p_j(x) = a_{j0} + a_{j1}(x - x_j) + a_{j2}(x - x_j)^2 + a_{j3}(x - x_j)^3, \tag{12}$$

where

$$\begin{cases} a_{j0} = p_j(x_j) = f_j \\ a_{j1} = p_j'(x_j) = k_j \quad \text{by (5)} \\ a_{j2} = \dfrac{1}{2}p_j''(x_j) = \dfrac{3}{h^2}(f_{j+1} - f_j) - \dfrac{1}{h}(k_{j+1} + 2k_j) \end{cases} \tag{13}$$

(cf. the right-hand side of (9)). In order to determine a_{j3} we observe that according to (8)

$$p_j''(x_{j+1}) = \frac{6}{h^2}(f_j - f_{j+1}) + \frac{2}{h}(2k_{j+1} + k_j)$$

while by (12)

$$p_j''(x_{j+1}) = 2a_{j2} + 6a_{j3}h$$

Equating the right-hand sides we obtain an equation from which a_{j3} can be solved:

$$a_{j3} = \frac{1}{3h}\left\{\frac{3}{h^2}(f_j - f_{j+1}) + \frac{1}{h}(k_j + 2k_{j+1}) - a_{j2}\right\}$$

Substitution of the expression of a_{j2} from (13) finally yields

$$a_{j3} = \frac{2}{h^3}(f_j - f_{j+1}) + \frac{1}{h^2}(k_{j+1} + k_j). \tag{14}$$

Example 1. We interpolate $f(x) = x^4$ on the interval $-1 \le x \le 1$ by the cubic spline $g(x)$ corresponding to the partition $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ and satisfying $g'(-1) = f'(-1)$, $g'(1) = f'(1)$.

Since $n = 2$, the system (11) consists of a single equation

$$k_0 + 4k_1 + k_2 = \frac{3}{h}(f_2 - f_0).$$

Here $k_0 = f'(-1) = -4$, $k_2 = f'(1) = 4$, $h = 1$, $f_2 = f(1) = 1$ and $f_0 = f(-1) = 1$, so that the solution is $k_1 = 0$. From (13) and (14) we obtain for $j = 0$

$$\begin{aligned}
a_{00} &= f_0 = 1 \\
a_{01} &= k_0 = -4 \\
a_{02} &= 3(f_1 - f_0) - (k_1 + 2k_0) = 3(0 - 1) - (0 - 8) = 5 \\
a_{03} &= 2(f_0 - f_1) + (k_1 + k_0) = 2(1 - 0) + (0 - 4) = -2
\end{aligned}$$

Therefore

$$p_0(x) = 1 - 4(x + 1) + 5(x + 1)^2 - 2(x + 1)^3 = -x^2 - 2x^3.$$

Similarly, for $j = 1$ we obtain

$$\begin{aligned}
a_{10} &= f_1 = 0 \\
a_{11} &= k_1 = 0 \\
a_{12} &= 3(1 - 0) - (4 + 2 \cdot 0) = -1 \\
a_{13} &= 2(0 - 1) + 4 + 0 = 2
\end{aligned}$$

Here

$$p_2(x) = -x^2 + 2x^3.$$

The spline $g(x)$ thus satisfies

$$g(x) = \begin{cases} -x^2 - 2x^3 & \text{when } -1 \le x \le 0 \\ -x^2 + 2x^3 & \text{when } \quad 0 \le x \le 1 \end{cases}$$

Example 2. We interpolate

$$f_0 = f(0) = 1, \quad f_1 = f(2) = 9, \quad f_2 = f(4) = 41, \quad f_3 = f(6) = 41$$

by the cubic spline satisfying $k_0 = 0$, $k_3 = -12$.

Since $n = 3$ and $h = 2$, the system (11) is

$$k_0 + 4k_1 + k_2 = \frac{3}{2}(f_2 - f_0) = 60$$
$$k_1 + 4k_2 + k_3 = \frac{3}{2}(f_3 - f_1) = 48$$

With $k_0 = 0$, $k_3 = -12$ the solution is $k_1 = 12$, $k_2 = 12$.

From (13) and (14) we obtain the coefficients of the spline for $j = 0$:

$$a_{00} = f_0 = 1$$
$$a_{01} = k_0 = 0$$
$$a_{02} = \frac{3}{4}(f_1 - f_0) - \frac{1}{2}(k_1 + 2k_0) = \frac{3}{4}(9 - 1) - \frac{1}{2}(12 + 0) = 0$$
$$a_{03} = \frac{2}{8}(f_0 - f_1) + \frac{1}{4}(k_1 + k_0) = \frac{2}{8}(1 - 9) + \frac{1}{4}(12 + 0) = 1$$

For $0 \le x \le 2$ we then have

$$g(x) = p_0(x) = 1 + x^3.$$

Similarly, for $j = 1$ we obtain

$$\begin{aligned} g(x) = p_1(x) &= 9 + 12(x - 2) + 6(x - 2)^2 - 2(x - 2)^3 \\ &= 25 - 36x + 18x^2 - 2x^3 \qquad (2 \le x \le 4) \end{aligned}$$

Finally, for $j = 2$

$$\begin{aligned} g(x) = p_2(x) &= 41 + 12(x - 4) - 6(x - 4)^2 \\ &= -103 + 60x - 6x^2 \qquad (4 \le x \le 6) \end{aligned}$$

Splines have the following property:

Theorem 2. *Let f be two times continuously differentiable on $a \le x \le b$ and let g be the interpolating cubic spline satisfying (2) and*

$$g'(a) = f'(a) \quad \text{and} \quad g'(b) = f'(b) \tag{15}$$

Then

$$\int_a^b f''(x)^2 dx \ge \int_a^b g''(x)^2 dx \tag{16}$$

and equality holds if and only if $f(x) \equiv g(x)$ for $a \le x \le b$.

Proof. By partial integration

$$\int_a^b g''(x)\,[f''(x)-g''(x)]\,dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} p_j''(x)\,[f''(x)-g''(x)]\,dx$$

$$= \sum_{j=0}^{n-1} \left\{ \Big/_{x_j}^{x_{j+1}} p_j''(x)\,[f'(x)-g'(x)] - \int_{x_j}^{x_{j+1}} p_j'''(x)\,[f'(x)-g'(x)]\,dx \right\}$$

Here the last term vanishes, because $p_j'''(x)$ is a constant and

$$\int_{x_j}^{x_{j+1}} [f'(x)-g'(x)]\,dx = \Big/_{x_j}^{x_{j+1}} [f(x)-g(x)] = 0.$$

Also the sum of the first terms is zero because

$$\sum_{j=0}^{n-1} \Big/_{x_j}^{x_{j+1}} p_j'(x)\,[f'(x)-g'(x)]$$
$$= p_0'(x_1)[f'(x_1)-g(x_1)] - p_0'(x_0)[f'(x_0)-g'(x_0)]$$
$$+ p_1'(x_2)[f'(x_2)-g(x_2)] - p_1'(x_1)[f'(x_1)-g'(x_1)] + \cdots = 0$$

This yields

$$\int_a^b [f''(x)-g''(x)]^2 dx = \int_a^b f''(x)^2 dx - 2\int_a^b f''(x)g''(x)dx + \int_a^b g''(x)^2 dx$$

$$= \int f''(x)^2 dx - \int_a^b g''(x)^2 dx$$

Left-side $\geq 0 \Rightarrow$ right side $\geq 0 \Rightarrow$ (16). □

4.7 Possible applications of polynomial approximation

1° Approximation of functions in a computer

For example the function $\sin x$.

2° **Numerical integration**

An approximate value for $\int_a^b f(x)dx$ can be found by replacing $f(x)$ by a polynomial approximation $p(x)$ such that the error

$$\left| \int_a^b \big(f(x) - p(x)\big)dx \right|$$

is small enough. If we use an interpolating polynomial associated to the points $x_k = a + k\frac{b-a}{n}$ $(0 \leq k \leq n)$, we obtain Newton-Cote's quadrative formula of order n. For $n = 2$ we obtain Simpson's rule.

Sometimes a truncated Taylor's formula may be used. For example,

$$\int_0^{\frac{1}{2}} \frac{\sin x}{x} dx$$

can be computed by using the approximation

$$\frac{\sin x}{x} \approx 1 - \frac{x^2}{3!} + \frac{x^4}{5!}$$

with an accuracy $\pm 10^{-6}$.

3° **Numerical differentiation**

If $p(x)$ is a polynomial approximation to $f(x)$, then $p'(x)$ could be used to approximate the derivative $f'(x)$ of $f(x)$.

An approximation on an entire interval $a \leq x \leq b$ can be obtained by using e.g. an interpolation cubic spline or a least squares approximation. By using interpolation polynomials the error $|f'(x) - p'(x)|$ can approach infinity when the degree of $p(x)$ approaches infinity.

A *local* approximation to $f'(x)$ can be found by differentiation of a suitable interpolating polynomial. A linear interpolating polynomial using the points $\big(x-h, f(x-h)\big)$ and $\big(x+h, f(x+h)\big)$ yields the approximation

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

The corresponding error can be computed using Taylor's formula. Subtracting the equations

$$f(x+h) = f(x) + hf'(x) + \frac{h}{2}f''(x) + \frac{h^3}{6}f'''(\xi_1)$$
$$f(x-h) = f(x) - hf'(x) + \frac{h}{2}f''(x) - \frac{h^3}{6}f'''(\xi_2)$$

yields

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

because

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

for some ξ (provided that $f'''(x)$ is continuous).

An approximation to $f''(x)$ can be found by using a quadratic interpolation polynomial determined by the points

$$\big(x-h, f(x-h)\big), \quad \big(x, f(x)\big) \quad \text{and} \quad \big(x+h, f(x+h)\big).$$

The formula corresponding to (1) is then

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi). \tag{2}$$

The remainder terms in (1) and (2) approach 0 when $h \to 0$. Therefore we can find in principle approximations to $f'(x)$ and $f''(x)$ with an arbitrary accuracy by using function values $f(x)$ and $f(x \pm h)$ for a sufficiently small h.

If $f(x+h)$ and $f(x-h)$ can be computed with an evaluation error $\pm\varepsilon$ and $|f'''(x)| \leq M$ on the interval $[x-h, x+h]$, then the total error in the approximation

$$\frac{f(x+h) - f(x-h)}{2h}$$

is by (1) at most

$$\frac{\varepsilon + \varepsilon}{2h} + \frac{h^2}{6}M = \frac{\varepsilon}{h} + \frac{h^2 M}{6}.$$

The last term is due to the truncation error $-\frac{h^2}{6}f'''(\xi)$ and approaches 0 as $h \to 0$. However, the first term $\frac{\varepsilon}{h}$ increases as $h \to 0$. To minimize the error bound h should be chosen so that

$$T(h) = \frac{\varepsilon}{h} + \frac{h^2 M}{6}$$

is as small as possible. If we let $T'(h) = 0$ we obtain

$$h = \left(\frac{3\varepsilon}{M}\right)^{\frac{1}{3}}.$$

One can also set both error terms equal:

$$\frac{\varepsilon}{h} = \frac{h^2 M}{6} \quad \Rightarrow \quad h = \left(\frac{6\varepsilon}{M}\right)^{\frac{1}{3}}$$

5 Differential equations

5.1 Introduction

We present a few methods for the solution of an initial value problem of an ordinary differential equation of order one

$$\begin{cases} y' = f(x, y) \\ y(a) = \eta \end{cases} \tag{1}$$

A solution $y(x)$ should be differentiable on some given interval $a \leq x \leq b$ and assume the value η at $x = a$. Here $y'(x) = f\big(x, y(x)\big)$.

Theorem 1. *(Picard's theorem) Suppose that f is continuous and that*

$$\left| \frac{\partial f(x, y)}{\partial y} \right| \leq K$$

for each $x \in [a, b]$ and each y. Then the initial value problem (1) has a unique solution.

Example 1. Show that the initial value problem

$$\begin{cases} y' = \dfrac{x^2 \sin y}{1 + x^2} \\ y(0) = 1 \end{cases}$$

has a unique solution on the interval $0 \leq x < \infty$.

The function $f(x, y)$ is everywhere differentiable and hence continuous.

$$\left| \frac{\partial f}{\partial y} \right| = \left| \frac{x^2 \cos y}{1 + x^2} \right| \leq \frac{x^2}{1 + x^2} \leq 1.$$

Thus the condition of Theorem 1 is satisfied.

Sometimes the exact solution can be found analytically. However, even then the computations can be so elaborate that a numerical solution is to be preferred.

Example 2. The initial value problem

$$\begin{cases} y' = x + \dfrac{2y}{1-x^4} \\ y(0) = 1 \end{cases}$$

has the solution

$$y(x) = \left(\frac{1+x}{1-x}\right)^{\frac{1}{2}} e^{\arctan x} \left\{ \int_0^x u \left(\frac{1-u}{1+u}\right)^{\frac{1}{2}} e^{-\arctan u} du + 1 \right\}.$$

5.2 Single-step methods

The simplest single-step method is *Euler's method*. Consider the initial value problem

$$\begin{cases} y' = f(x,y) \\ y(a) = \eta \end{cases} \tag{1}$$

on the interval $a \leq x \leq b$. We subdivide this interval into N subintervals of length $h = \frac{b-a}{N}$ and end points

$$x_k = a + kh \qquad (0 \leq k \leq N)$$

Let y_k be an approximate value to the solution $y(x)$ at $x = x_k$. The error at x_k is then $y_k - y(x_k)$. In Euler's method we compute approximations $y_1, y_2, \ldots, y_N$ recursively by using the formula

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{2}$$

Then $x_0 = a$ and $y_0 = y(a) = \eta$, so that

$$\begin{aligned} y_1 &= y + hf(a,\eta) \\ y_2 &= y_1 + hf(x_1, y_1) \\ y_3 &= y_2 + hf(x_2, y_2) \\ &\vdots \\ y_N &= y_{N-1} + hf(x_{N-1}, y_{N-1}) \end{aligned} \tag{3}$$

Example 3. Apply Euler's method to the initial value problem in Example 2.

$$f(x,y) = x + \frac{2y}{1-x^4}; \qquad x_0 = 0; \quad y_0 = y(0) = 1$$

Using step length $h = 0.1$ we obtain

$$\begin{aligned}
y(0) &= y_0 = 1 \\
y(0.1) &\approx y_1 = 1 + 0.1 \cdot 2 = 1.2 \\
y(0.2) &\approx y_2 = 1.2 + 0.1\left(0.1 + \frac{2 \cdot 1.2}{1 - 10^{-4}}\right) = 1.45 \\
y(0.3) &\approx y_3 = 1.45 + 0.1\left(0.2 + \frac{2 \cdot 1.45}{1 - 16 \cdot 10^{-4}}\right) = 1.76 \\
y(0.4) &\approx y_4 = 2.118 \\
y(0.5) &\approx y_5 = 2.593
\end{aligned}$$

Analytical derivation of Euler's method: The value $y_0 = \eta$ is obtained from the initial condition $y(a) = \eta$. Suppose that the solution $y(x)$ is twice continuously differentiable. From Taylor's formula we get

$$y(x_0 + h) = y(x_0) + hy'(x_0) + \frac{h^2}{2}y''(\xi).$$

If we omit the remainder term we obtain an approximate value to $y(x_0+h) = y(x_1)$:

$$y_1 = y(x_0) + hy'(x_0) = \eta + hf(x_0, y_0) = \eta + f(a, \eta).$$

This is the first formula in (3). The error in y_1 is

$$y_1 - y(x_1) = -\frac{h^2}{2}y''(\xi)$$

Since y'' is continuous, this error $\to 0$ as $h \to 0$.

The second formula in (3) is derived in an analogous way, and in general form

$$y(x_n + h) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(\xi_n)$$

we obtain

$$y_{n+1} = y_n + hf(x_n, y_n)$$

by deleting the remainder term $\frac{h^2}{2}y''(\xi_n)$ and replacing $y'(x_n) = f\big(x_n, y(x_n)\big)$ with the approximation $f(x_n, y_n)$. The error arising from this replacement can be estimated by using the mean value theorem if the condition

$$\left|\frac{\partial f(x,y)}{\partial y}\right| \leq K$$

of Theorem 1 is satisfied:

$$\big|f\big(x_n, y(x_n)\big) - f(x_n, y_n)\big| \leq K\,|y(x_n) - y_n|$$

If in addition $|y''(x)| \leq M$ for each $x \in [a, b]$, one can prove that the global truncation error

$$|y(x_n) - y_n| \leq \frac{hM}{2}\frac{e^{K(x_n-a)} - 1}{K}.$$

The error at the end point $x_N = b$ is therefore at most

$$\frac{hM}{2K}\left(e^{K(b-a)} - 1\right)$$

if we forget the evaluation error.

Euler's method

Problem:

$$\begin{cases} \quad y' = f(x,y) \\ y(a) = \eta \end{cases}$$

Algorithm:

$$\begin{aligned} y_{n+1} &= y_n + hf(x_n, y_n) \\ y_0 &= \eta \\ x_n &= a + nh, \quad h = \frac{b-a}{N} \end{aligned} \tag{4}$$

Error estimate:

$$|y(x_n) - y_n| \leq \frac{hM}{2}\frac{e^{K(x_n-a)} - 1}{K}.$$

Geometric derivation of Euler's method: If the conditions of Theorem 1 are fulfilled, then each point (x, y) in the plane is contained in a solution curve.

In Euler's method we draw a tangent to a solution curve at (x_k, y_k), then y_{k+1} is the y-coordinate of the intersection point of this tangent and the line $x = x_{k+1}$:

$$y_{k+1} = y_k + hf(x_n, y_n)$$

Remark. Instead of Euler's method it is customary to use more accurate methods where the upper bound of the global truncation error is proportional h^p where $p > 1$. As a result of better accuracy one can then use a larger step length h, so that the number N of necessary steps is reduced.

Heun's method

Let

$$y_{n+1}^{(e)} = y_n + hf(x_n, y_n)$$

be the approximate value at x_{n+1} given by Euler's method and let

$$k_1 = hf(x_n, y_n)$$

be the corresponding increase of the function value. We obtain another approximation for k_1 by approximating the solution curve by a secant through (x_n, y_n) with slope $f\big(x_n, y_n^{(e)}\big)$:

$$k_2 = hf(x_{n+1}, y_{n+1}^{(e)}) = hf(x_n + h, y_n + k_1)$$

Heun's method:

$$y_{n+1} = y_n + \frac{1}{2}(k_1 + k_2)$$

The global truncation error in Heun's method is $O(h^2)$, i.e. the error $\leq M \cdot h^2$.

Classical Runge-Kutta method:

$$\begin{aligned} k_1 &= hf\big(x_n, y_n\big) \\ k_2 &= hf\big(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\big) \\ k_3 &= hf\big(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\big) \\ k_4 &= hf\big(x_n + h, y_n + k_3\big) \end{aligned}$$

$$y_{n+1} = y_n + \frac{1}{6}\big(k_1 + 2k_2 + 2k_3 + k_4\big)$$

One can show that the global truncation error is $O(h^4)$.

Example. Compute one step using the classical Runge-Kutta method for the initial value problem

$$\begin{cases} y' = xy \\ y(0) = 1 \end{cases}$$

with step length $h = 0.4$.

Now $f(x, y) = xy$, so we get

$$\begin{aligned} k_1 &= hx_0y_0 = 0 \\ k_2 &= h\big(x_0 + \frac{h}{2}\big)\big(y_0 + \frac{k_1}{2}\big) = h \cdot \frac{h}{2} \cdot 1 = 0.08 \\ k_3 &= h\big(x_0 + \frac{h}{2}\big)\big(y_0 + \frac{k_2}{2}\big) = h \cdot \frac{h}{2}(1 + 0.04) = 0.0832 \\ k_4 &= h\big(x_0 + h\big)\big(y_0 + k_3\big) = 0.173312 \end{aligned}$$

$$y(0.4) \approx y_1 = 1 + \frac{1}{6}\big(0 + 2 \cdot 0.08 + 2 \cdot 0.0832 + 0.173312\big) = 1.083285$$

Truncation error $\approx 2 \cdot 10^{-6}$.

In a *multistep method* the computation of y_{n+1} requires the knowledge of more than one previous approximation, e.g. $y_n, y_{n-1}, y_{n-1}, \ldots, y_{n-p}$.

Example. Midpoint method

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n) \tag{5}$$

Derivation: We integrate the equation $y'(x) = f\big(x, y(x)\big)$ over $[x_{n-1}, x_{n+1}]$ and approximate the integral

$$\int_{x_{n-1}}^{x_{n+1}} f\big(t, y(t)\big)dt$$

with the product

$$f\big(x_n, y(x_n)\big)\big(x_{n+1} - x_{n-1}\big) = 2hf\big(x_n, y(x_n)\big)$$

$$\int_{x_{n-1}}^{x_{n+1}} y'(t)dt = \Big/_{x_{n-1}}^{x_{n+1}} y(t) = y(x_{n+1}) - y(x_{n-1})$$

Replacing the exact values $y(x_k)$ with the approximation y_k we obtain (5). Global truncation error is $O(h^2)$.

Remark. Multistep methods require a separate starting procedure because in the beginning we only know one initial condition $y_0 = \eta$. For example, in using (5) we need y_0 and y_1 for the computation of y_2, because

$$y_2 = y_0 + 2hf(x_1, y_1).$$

Such a starting procedure could be some single-step method.

5.3 Implicit methods

We integrate the equation

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

over $[x_n, x_{n+1}]$.

Left-hand side:

$$\int_{x_n}^{x_{n+1}} y'(t)dt = y(x_{n+1}) - y(x_n)$$

Right-hand side becomes

$$\int_{x_n}^{x_{n+1}} f\big(x, y(x)\big)dx$$

Applying the trapezoidal rule we obtain

$$\int_{x_n}^{x_{n+1}} f\big(x, y(x)\big)dx \approx \frac{h}{2}\left[f\big(x_n, y(x_n)\big) + f\big(x_{n+1}, y(x_{n+1})\big)\right]$$

Replacing on both sides the exact values $y(x_n)$ and $y(x_{n+1})$ with the approximate values y_n and y_{n+1} we obtain:

The trapezoidal method

$$y_{n+1} = y_n + \frac{h}{2}\big[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\big] \tag{1}$$

Global truncation error is $O(h^2)$.

The trapezoidal method is an example of *implicit* methods: y_{n+1} does not depend explicitly on a previous y_n as in the explicit methods discussed so far, but y_{n+1} is obtained as a solution of the equation (1) which may be nonlinear.

Example. Consider again the problem $y' = xy$, $y(0) = 1$. Due to the linearity of the differential equation the equation (1) is linear with respect to y_{n+1}:

$$y_{n+1} = y_n + \frac{h}{2}\big[x_n y_n + x_{n+1} y_{n+1}\big]$$

Using the step length $h = 0.2$ we obtain

$$\begin{aligned} & y_1 = y_0 + 0.1\big[x_n y_n + x_{n+1} y_{n+1}\big] = 1 + 0.1\big(0 + 0.2 y_1\big) \\ \Leftrightarrow\ & y_1 = 1 + 0.02 y_1 \\ \Leftrightarrow\ & y_1 \approx 1.0204 \end{aligned}$$

For $n = 1$ we get

$$\begin{aligned} & y_2 = y_1 + 0.1[x_1 y_1 + x_2 y_2] = 1.0204 + 0.1[0.2 \cdot 1.0204 + 0.4 \cdot y_2] \\ \Rightarrow\ & y(0.4) \approx y_2 = \frac{1.0408}{1 - 0.04} \approx 1.0842 \end{aligned}$$

Truncation error $\approx 9 \cdot 10^{-4}$.

Example. Consider the problem

$$\begin{cases} y' = e^{-y} \\ y(0) = 1 \end{cases}$$

The differential equation is now nonlinear. For $n = 0$ the formula (1) is

$$y_1 = y_0 + \frac{h}{2}\big[e^{-y_0} + e^{-y_1}\big] = 1 + \frac{h}{2}\big[e^{-1} + e^{-y_1}\big].$$

If e.g. $h = 0.2$, then y_1 should be solved from the equation

$$y_1 = 1 + 0.1[e^{-1} + e^{-y_1}]$$

Define an auxiliary function $g(y)$ such that

$$g(y_1) = y_1 - 0.1 \cdot e^{-y_1} - (1 + 0.1e^{-1}) = 0.$$

The solution can be found e.g. by Newton's method. The starting value could be the approximation given by Euler's method

$$y_1^{(o)} = y_0 + he^{-y_0} \approx 1.0736.$$

By Newton's method we obtain

$$\begin{aligned} y_1^{(1)} &= 1.071053 \\ y_1^{(2)} &= 1.071053 \end{aligned}$$

Hence $y_1 = 1.071053$, and we can continue with the next step $n = 1$.

In the trapezoidal method y_{n+1} is solved from an equation $g(y) = 0$, where

$$g(y) = y - \frac{h}{2} f(x_{n+1}, y) - \left[y_n + \frac{h}{2} f(x_n, y_n) \right].$$

If the solution is found by Newton's method as in the previous example, in each iteration step we must compute the derivative

$$g'(y) = 1 - \frac{h}{2} \frac{\partial f}{\partial y}(x_{n+1}, y).$$

An alternative approach to solve $g(y) = 0$ is the fixed point iteration

$$\begin{aligned} y_{n+1}^{(0)} &= y_n + hf(x_n, y_n) \\ y_{n+1}^{(k+1)} &= y_n + \frac{h}{2} \left[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(k)}) \right] \quad (k = 0, 1, 2, \ldots) \end{aligned} \tag{2}$$

The starting value has been computed with Euler's method which serves as a *predictor*. This predicted approximation is corrected using the trapezoidal method. The algorithm is an example of a *predictor-corrector* method.

According to §3.4 a sufficient condition for the convergence of the iteration in (2) is the inequality

$$\left| \frac{h}{2} \frac{\partial f}{\partial y} \right| < 1 \tag{3}$$

in a sufficiently large neighborhood of (x_{n+1}, y_{n+1}). Thus the step length h should be small enough.

Example. Consider the solution of the initial value problem

$$\begin{cases} y' = e^{-y} \\ y(0) = 1 \end{cases}$$

using the predictor-corrector method (2). The iteration will converge towards the solution of (1), if

$$\left| \frac{h}{2} \frac{\partial f}{\partial y} \right| = \frac{h}{2} e^{-y} < 1$$

in a sufficiently large neighborhood of (x_{n+1}, y_{n+1}). Since $e^{-y} > 0$ for each y, then all terms in (2) will be nonnegative. Since in addition $y(0) = y_0 = 1$, we see by induction that $y_{n+1}^{(k+1)} \geq 1$ for each k and n. Therefore (3) will hold if e.g. $h = 0.1$.

If $\left| \frac{\partial f}{\partial y} \right|$ is large, we must use a very small step length to guarantee the convergence of iteration (2). Then Newton's method could be a better choice.

To mention a few methods, Adams-Bashfort is an explicit predictor method while Adams-Moulton is an implicit corrector method. For example, the following predictor-corrector method has a global truncation error $O(h^4)$:

$$y_{n+4} = y_{n+3} + \frac{h}{24}(55f_{n+3} - 59f_{n+2} + 37f_{n+1} - 9f_n)$$
$$y_{n+4} = y_{n+3} + \frac{h}{24}(9f_{n+4} + 19f_{n+3} - 5f_{n+2} + f_{n+1})$$

5.4 Boundary value problems

Consider a *boundary value problem* associated with a second order differential equation

$$\begin{cases} y'' = f(x, y, y') \\ y(a) = \alpha \\ y(b) = \beta \end{cases}$$

where f is a given function. We look for a solution of the given equation defined on the interval $[a, b]$ which has the prescribed values at the end points a and b. The condition $y(a) = \alpha$ and $y(b) = \beta$ are called *boundary conditions.*

Example. A stationary solution of the one-dimensional heat equation

$$\frac{d}{dx}\left(k(x)\frac{dy}{dx}\right) = 0$$

with boundary conditions $y(a) = \alpha$, $y(b) = \beta$ describes the temperature of a thin rod with variable heat conduction properties. Assume that the rod is insulated along its length and that the end points are kept at different temperatures. If the heat conduction coefficient is constant, then the solution of the boundary value problem is a first degree polynomial. If $k(x)$ is nonconstant, the problem should be solved numerically.

Shooting method

We consider the boundary value problem

$$\begin{cases} y'' = f(x, y, y') \\ y(a) = \alpha \\ y(b) = \beta \end{cases} \tag{1}$$

together with the initial value problem

$$\begin{cases} y'' = f(x, y, y') \\ y(a) = \alpha \\ y'(a) = \gamma \end{cases} \tag{2}$$

We assume that both problems have a unique solution. In the shooting method we try to find γ so that the solution of the initial value problem (2) satisfies the boundary condition $y(b) = \beta$. Then the boundary value problem (1) is reduced to the initial value problem (2). When γ has been found, (2) can be represented as a system

$$\begin{cases} y' = v \\ v' = f(x, y, v) \\ y(a) = \alpha \\ v(a) = \gamma \end{cases} \tag{3}$$

This can be solved e.g. using a Runge-Kutta method. γ will be found by improving a suitable initial guess by iteration.

Example. The boundary value problem

$$y'' = -y, \qquad y(0) = 0, \quad y\big(\frac{\pi}{2}\big) = 1$$

has an analytical solution $y(x) = \sin x$. If the solution is approximated by a linear interpolation polynomial satisfying the given boundary conditions, we obtain an initial guess for γ: the slope of the graph of the polynomial

$$\gamma_0 = \frac{1-0}{\frac{\pi}{2}-0} = \frac{2}{\pi} = 0.637.$$

The initial value problem (2) is now

$$\begin{cases} y'' = -y \\ y(0) = 0 \\ y'(0) = \gamma_0 = 0.637 \end{cases}$$

and the system (3) is

$$\begin{cases} y' = v \\ v' = -y \\ y(0) = 0 \\ v(0) = 0.637 \end{cases}$$

Matlab's ode23tx gives $y(\frac{\pi}{2}) = 0.637$. Shooting to the direction of γ_0 we therefore hit below the goal ($y(\frac{\pi}{2}) = 1$). We change to a steeper shooting direction and choose $y'(0) = \gamma_1 = 1.2$. Solving (3) as above we obtain $y(\frac{\pi}{2}) = 1.2$.

Instead of trial and error we can look for the right value of γ by iteration. Let $y(x, \gamma)$ be the solution of the initial value problem (2) as a function of γ and denote

$$g(\gamma) = y(b, \gamma).$$

Then obviously $y(x, \gamma)$ is a solution of (1) exactly when $g(\gamma) = \beta$. To determine γ we therefore have to solve the equation

$$g(\gamma) - \beta = 0. \tag{4}$$

Since g does not have an explicit formula, the values of g must be computed numerically. Also we don't know how to differentiate g, so that we could try

to solve (4) by the *secant method*:

$$\gamma_{n+1} = \gamma_n - \frac{g(\gamma_n) - \beta}{\dfrac{g(\gamma_n) - g(\gamma_{n-1})}{\gamma_n - \gamma_{n-1}}} \tag{5}$$

Example. In the previous example we computed

$$g(\gamma_0) = 0.637 \quad \text{and} \quad g(\gamma_1) = 1.2,$$

where $\gamma_0 = 0.637$ and $\gamma_1 = 1.2$. The secant method gives ($n = 1$)

$$\gamma_2 = 1.2 - \frac{1.2 - 1}{\dfrac{1.2 - 0.637}{1.2 - 0.637}} = 1.2 - 0.2 = 1$$

Then $g(\gamma_2) = y(\frac{\pi}{2}, 1) = 1.0000$; hence we hit directly to the goal and the solution of (4) is $\gamma_2 = 1$.

One can show that if f depends linearly on y and y' as in the previous example, $g(\gamma)$ is a polynomial of γ of degree one. Therefore the secant method immediately gives the correct solution. If the differential equation is nonlinear, we usually have to perform several iterations to solve (4).

Example. We solve the boundary value problem

$$y'' = 1 + yy', \qquad y(0) = 1, \quad y(0.6) = 2$$

by the shooting method using Matlab's ode23tx in the initial value problems.

By linear interpolation we obtain the initial guess

$$\gamma_0 = \frac{2 - 1}{0.6 - 0} = 1.67$$

which yields $g(\gamma_0) = 2.8788$. Choosing a new shooting direction ($\gamma_1 = 0.8$) we obtain $g(\gamma_1) = 1.3544$. The secant method gives $\gamma_2 = 0.8429$; then $g(\gamma_2) = 1.9965$. Further $\gamma_3 = 0.8465$ and $g(\gamma_3) = 2.0000$ is quite close to the correct boundary value.