2

Numerical Methods

Before the computer (*ordinateur* in French) changed the world, numerical mathematics — which mockers referred to as phenomenological — could hardly be counted as one of the supreme disciplines of the mathematical sciences. Whether that is the case today is beside the point, but combined with modeling and simulation it has risen in the hierarchy, and the latter even have to suffer to some extent to justify the existence of the other subjects. In the 1960s the integrimeter, integraph and harmonic oscillator were treated as *instrumental mathematics* in lectures. They have been long forgotten, as have all numerical methods for the hand calculator such as, e.g., extracting roots by subtracting odd numbers.

The turbulent evolution of the computer to the *laptop* allowed *numerical mathematics* to successfully keep pace at some distance. Logarithmic Tables, etc., have long been replaced by the pocket calculator, and linear systems — a central problem — are solved today with three inconspicuous glyphs `A\b`, without resulting in any inconsistencies in style. A multitude of monographs displays what has been achieved so far; only (Golub), (Hairer) and (Rheinboldt70) are stated as examples. The curve of the number of publications with purely numerical themes also seems to be getting somewhat flatter, while the number of problem-related applications is on the rise.

If numerical methods shall be described in a single chapter, it is necessary to concentrate on the essential aspects. The author assumes that the reader is interested primarily in applying existing codes, which is not possible without a minimum of understanding and intuition. This is the premise for the introduction to the mindsets of *numerics* provided here and the discussion of challenging developments as the multiple shooting method and differential-algebraic problems. The numerical part of this book is not limited to the topics treated here; for those possessing the necessary background, further issues are dealt with in later chapters.

2.1 Interpolation and Approximation

In many applications, functions are given only by discrete data sets. Or, a function cannot be integrated in closed form and must be replaced by a simpler one to that end. Then it is approximated *piecewise* by polynomials of *moderate* degree because polynomials of higher degree oscillate more or less strongly in larger intervals. But also approximations by rational functions, exponential functions, and, preferably in periodic problems, trigonometric interpolation are in common use. Let

$$\boxed{\Pi_n \text{ the set of real polynomials } p_n \text{ of degree} \le n}\,.$$

With the usual addition and scalar multiplication, Π_n is a *vector space* of dimension $n+1$ whose basis $\{q_0(x), \ldots, q_n(x)\}$ is chosen according to the individual requirements.

(a) The General Interpolation Problem Let there be given

$$\begin{aligned} &\text{a sequence of support abszissas } \{x_i\}_{i=0}^{\infty}\,,\ x_i \in \mathbb{R}\,,\\ &\text{a sequence of support ordinates } \{f_i\}_{i=0}^{\infty}\,,\ f_i \in \mathbb{R}\,,\\ &\text{a sequence of functions} \qquad\quad \{g_i\}_{i=0}^{\infty}\,,\ g_i \in \mathcal{C}[a,b]\,. \end{aligned}$$

The support abszissas x_i shall be *mutually distinct*; for the other case we refer to (Hoellig) Sect. 3.1. Then, a sequence of functions

$$\{h_n\}_{n=1}^{\infty}\,,\quad h_n(x) = \sum_{i=0}^{n-1} \alpha_i g_i(x)\,, \tag{2.1}$$

is to be found with the *interpolation property*

$$h_n(x_j) = f_j\,,\quad j = 0:n-1\,. \tag{2.2}$$

Writing $\underline{a} = [\alpha_0, \ldots, \alpha_{n-1}]^T$ and $\underline{f} = [f_0, \ldots, f_{n-1}]^T$, (2.1) is, for *fixed* n, equivalent to the linear system of equations

$$A\underline{a} = \underline{f},\quad A = \left[g_i(x_j)\right]_{i,j=0}^{n-1}\,, \tag{2.3}$$

and the interpolation problem (2.1), (2.2) has a unique solution for a regular matrix A.

Theorem 2.1. *(Existence,* HAAR *Condition) Let all n support abscissas x_j, $j = 0:n-1$, be mutually distinct and let every not identical disappearing linear combination of n functions g_i, $i = 0:n-1$, have no more than $n-1$ zeros, then the matrix A is regular and the interpolation problem has a unique solution.*

Proof. If the matrix A is singular, there exists a row vector $\underline{c} \in \mathbb{R}_n$ with $\underline{c}\,A = 0 \in \mathbb{R}_n$. Then the linear combination $h(x) := \sum_{i=0:n-1} c_i g_i(x)$ has n different zeros $x_0, \ldots, x_{n-1}$ because

$$h(x_j) = \sum\nolimits_{i=0}^{n-1} c_i g_i(x_j) = 0,\ j = 0 : n-1\,,$$

in contradiction to the assumption. □

In particular, the HAAR condition is fulfilled for a sequence $\{p_n\}_{n=0}^{\infty}$ of polynomials $p_n \in \Pi_n$ because every not identically disappearing linear combination $h(x) := \sum_{i=0:n-1} c_i p_i(x)$ is a polynomial of degree $\le n-1$ having no more than $n-1$ zeros. However, the matrix A in (2.3) is ill-conditioned in general, hence this linear system of equations is not used commonly for numerical computation of the coefficients α_i.

(b) Interpolating Polynomials To find a *linear recursion formula* for *interpolating polynomials*, let $\{j_0, \ldots, j_m\} \subset \{0, \ldots, n\}$ be an index set with different elements. Then the interpolating polynomial $p_{j_0,\ldots,j_m}(x) \in \Pi_m$ is uniquely determined by

$$\begin{aligned} p_j(x) &= f(x_j)\,,\ m = 0\,,\\ p_{j_0,\ldots,j_m}(x_i) &= f(x_i)\,,\ i = j_0, \ldots, j_m\,,\ m = 1 : n\,, \end{aligned} \tag{2.4}$$

following the Existence Theorem; thus, in particular, $p_{j_0,\ldots,j_m}(x)$ does not depend on a permutation of indices.

Lemma 2.1. *(*AITKEN*) For* $j = 0 : n-m\,,\ m = 1 : n$

$$p_{j,\ldots,j+m}(x) = \frac{1}{x_{j+m} - x_j}\left[(x - x_j)p_{j+1,\ldots,j+m}(x) - (x - x_{j+m})p_{j,\ldots,j+m-1}(x)\right].$$

This formula is used in various applications for computation of interpolating polynomials at a given point x.

Theorem 2.2. *(*CAUCHY*'s Error Representation) Let the function* f *be* $(n+1)$*-times differentiable in* $[a, b]$ *and let* $[u, v, \ldots, w]$ *be the smallest interval* $\mathcal{I} \subset \mathbb{R}$ *containing all* $u, v, \ldots, w \in \mathcal{I}$*. Then* $\forall\, x \in [a,\, b]\ \ \exists\, \xi_x \in [x_0, \ldots, x_n, x]$*:*

$$f(x) - p_n(x; f) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\,\omega(x)\,,\ \ \omega(x) = (x - x_0)\cdots(x - x_n)\,. \tag{2.5}$$

Proof in `SUPPLEMENT\chap02`.

Note that the intermediate values ξ_x change with the value x.

(c) Interpolation after Lagrange We consider the approximation of a function f by an interpolating polynomial in *separated* form

$$\boxed{f(x) \approx p_n(x; f) = \sum_{i=0}^{n} f(x_i) q_i(x)}\,, \tag{2.6}$$

with the basis $\{q_0, \ldots, q_n\}$ of Π_n consisting of LAGRANGE polynomials

$$q_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}, \quad i = 0 : n\,. \tag{2.7}$$

The interpolation property $p_n(x_i; f) = f(x_i)$ is guaranteed here by the specific property $q_i(x_j) = \delta^i{}_j$ (KRONECKER symbol).

Properties: (1°) Interpolation of LAGRANGE is of high theoretical but less practical use because all polynomials $q_i(x)$ have to be computed anew if the set of support nodes is changed or augmented.
(2°) By unique existence, the formula (2.6) is exact for all *monomials* $f(x) = x^k,\ k = 0 : n\,,\ \sum_{i=0:n}(x_i)^k q_i(x) = x^k\,,\ k = 0 : n$; in particular, we obtain a *partition of unity* $\sum_{i=0:n} q_i(x) = 1$ for $k = 0\,.$
(3°) For *equidistant abszissas*, $h = 1/n\,,\ x_i = x_0 + ih\,,\ x = x_0 + sh$, $s \in [0,\ n]\,$, we obtain

$$\frac{x - x_j}{x_i - x_j} = \frac{(x_0 + sh) - (x_0 + jh)}{(x_0 + ih) - (x_0 + jh)} = \frac{s - j}{i - j}\,;$$

hence formula (2.6) is simplified considerably by this translation and scaling of the independent variable:

$$p_n(x(s); f) = \sum_{i=0}^{n} f(x_i) q_i(x(s))\,,\quad q_i(x(s)) = \prod_{j=0, j\neq i}^{n} \frac{s - j}{i - j}\,. \tag{2.8}$$

This representation applies in constructing numerical quadrature formulas as well as in constructing numerical devices for approximations of ordinary differential equations and systems.

(d) Interpolation after Newton Let $f[x_{j_0}, \ldots, x_{j_m}]$ be the highest term of $p_{j_0,\ldots,j_m}(x)$ then, by Lemma 2.1,

$$f[x_j, \ldots, x_{j+m}] = \frac{f[x_{j+1}, \ldots, x_{j+m}] - f[x_j, \ldots, x_{j+m-1}]}{x_{j+m} - x_j}\,. \tag{2.9}$$

These *divided differences* do not depend on the succession of indices because the associated polynomials have this property. On choosing the NEWTON basis for $\Pi_n\,,\ n_0(x) \equiv 1\,,\ n_j(x) = (x - x_0)\cdots(x - x_{j-1})\,,\ j = 1 : n\,,$ we obtain

$$\begin{aligned} p_{0,\ldots,n}(x; f) &= \sum_{j=0}^{n} a_j n_j(x)\,,\quad a_j = f[x_0, \ldots, x_j]\,, \\ &= (\cdots(a_n(x - x_{n-1}) + a_{n-1})(x - x_{n-2}) + a_{n-2})\cdots)(x - x_0) + a_0 \end{aligned} \tag{2.10}$$

because of the interpolation property and the recursion formula

$$p_{0,\ldots,n}(x) = p_{0,\ldots,n-1}(x) + a_n \pi(x)\,,\quad \pi(x) = (x - x_0)\cdots(x - x_{n-1}) \in \Pi_n$$

being typical for this form of the interpolating polynomial.

The well-known TAYLOR polynomial $p_n(x;f) = \sum_{i=0:n} \frac{f^{(i)}(x_0)}{i!}(x-x_0)^i$ may be said to stand on the opposite side of the scale of approximation by polynomials since it uses only one "interpolation point" x_0. By using (2.9), a natural relation may be found between TAYLOR coefficients and divided differences being the coefficients of NEWTON's polynomial:

Lemma 2.2.

$$f[x_i,\dots,x_{i+k}] = \frac{f^{(k)}(\xi)}{k!},\ \ \xi \in [x_i,\dots,x_{i+k}]\,.$$

Proof. Let $p_{i,\dots,i+k}(x)$ be the NEWTON interpolating polynomial of degree $\le k$ with nodes $(x_j,f_j),\ j=i:i+k-1$, and let $(x_{i+k},f_{i+k}) = (\overline{x},f(\overline{x}))$ where all abszissas x_j and $\overline{x}$ shall be mutually distinct. Then we have

$$f(\overline{x}) = p_{i,\dots,i+k}(\overline{x}) = p_{i,\dots,i+k-1}(\overline{x}) + f[x_i,\dots,x_{i+k}](\overline{x}-x_i)\cdots(\overline{x}-x_{i+k-1})$$

at the point $x_{i+k}=\overline{x}$, and, on the other side, the error formula for $p_{i,\dots,i+k}(x)$,

$$f(\overline{x}) = p_{i,\dots,i+k-1}(\overline{x}) + \frac{f^{(k)}(\xi)}{k!}(\overline{x}-x_i)\cdots(\overline{x}-x_{i+k-1})\,.$$

□

(e) By additional **Interpolation of the Derivatives** of f at all abszissas x_i we obtain interpolating polynomials of HERMITE type:

$$f(x) \approx h_{2n+1}(x,f) = \sum_{i=0}^{n} [f(x_i)h_{0,i}(x) + f'(x_i)h_{1,i}(x)] \in \Pi_{2n+1}$$

with HERMITE polynomials

$$\begin{aligned} h_{0,i}(x) &= \big[1-2q_i'(x_i)(x-x_i)\big]q_i(x)^2 &&\Longrightarrow h_{0,i}(x_k) = \delta^i{}_k\,,\ h_{0,i}'(x_k) = 0\,,\\ h_{1,i}(x) &= (x-x_i)q_i(x)^2 &&\Longrightarrow\ h_{1,i}(x_k) = 0\,,\ \ h_{1,i}'(x_k) = \delta^i{}_k\,, \end{aligned}$$

where $q_i(x)$ are the LAGRANGE polynomials again. The error has the same form as in (2.5):

$$f(x) - h_{2n+1}(x,f) = \frac{f^{(2n+1)}(\xi_x)}{(n+1)!}\,\omega(x)\,,\ \ \omega(x) = (x-x_0)^2\cdots(x-x_n)^2.$$

Besides some few exceptions, enhancing the degree n of an interpolating polynomial does *not* improve the approximation, instead a *segmentwise* interpolation with simple polynomials is to be preferred. By requiring some global smoothness, the compound polynomials then lead to the *interpolating spline functions*; cf. **(g)**.

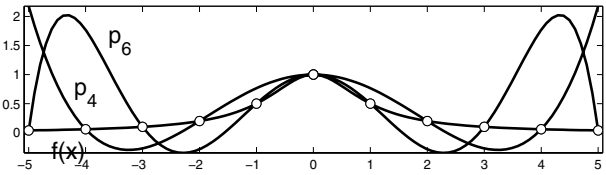

Figure 2.1. Interpolating polynomial of degree $n = 4$, 6 for $f(x) = 1/(1 + x^2)$

(f) Approximation by Beziér Polynomials In a fixed, unpartitioned interval, essential improvement of approximation is attained by abandoning the strong interpolating condition $p_n(x_i; f) = f(x_i)$ in the *interior* of the considered intervall. By the partition of unity,

$$1 = (x + (1 - x))^n = \sum_{i=0}^{n} \binom{n}{i} x^i (1 - x)^{n-i} =: \sum_{i=0}^{n} B_i^n(x)\,,$$

we obtain the basis of *Bernstein* polynomials $B_i^n(x)$ of Π_n and the general representation of $p_n \in \Pi_n$ as BEZIÉR *polynomial* with the BEZIÉR *points* b_i,

$$p_{n,\mathrm{bez}}(x) = b_0 B_0^n(x) + \ldots + b_n B_n^n(x)\,. \tag{2.11}$$

The roots of all $B_i^n(x)$ are placed at the boundary of the interval [0, 1], and precisely one extremal point exists in the interior (maximum point). Therefore BERNSTEIN polynomials possess no turning point in this interval. As a consequence, no *spurious* turning points are dragged in by approximating a function f in this way. However, the approximation is restricted to the unit interval [0, 1] here, otherwise a rescaling becomes necessary. Besides, piecewise interpolation is to be preferred in the present case, too.

Properties: (1°) $\displaystyle\sum_{i=0}^{n} B_i^n(x) = 1\,,\ \sum_{i=0}^{n} \left(\frac{i}{n}\right) B_i^n(x) = x\,.$

(2°) By applying *forward differences* $\Delta b_i = b_{i+1} - b_i$, $\Delta^k b_i = \Delta(\Delta^{k-1} b_i)$, we obtain for the derivatives

$$p_{n,\mathrm{bez}}^{(k)}(x) = \frac{n!}{(n-k)!} \sum_{i=0}^{n-k} (\Delta^k b_i) B_i^{n-k}(x)\,,$$

from which the above mentioned important property follows, namely

$$\boxed{\forall\, i : \Delta^k b_i \geq 0 \implies \forall\, x \in [0,\,1] : p^{(k)}(x) \geq 0}\,.$$

Moreover, the k-th derivatives in $x = 0$ resp. $x = 1$ depend only on the BEZIÉR points $b_0, \ldots, b_k$ resp. $b_{n-k}, \ldots, b_n$.

(3°) Let there be given $s - r + 1$ successive BEZIÉR points $\{b_r, \ldots, b_s\}$ for the abszissas $x = (i - r)/(s - r)$, $i = r : s$, then

$$b_{r,\ldots,s}(x) := \sum_{i=r}^{s} b_i B_{i-r}^{s-r}(x)$$

is the corresponding BEZIÉR polynomial of degree $\leq s-r$ (depending on the succession of points!). By means of the addition theorem for binomial coefficients, the *linear recursion formula* of DE CASTELJAU can be derived,

$$b_{r,\ldots,s}(x) = (1-x)b_{r,\ldots,s-1}(x) + x b_{r+1,\ldots,s}(x)\,,$$

being applied for *pointwise* computation of $p_{n,\mathrm{bez}}(x)$ in place of the algebraic representation (2.11).
(4°) The points $(x_i, b_i) = (i/n, f(i/n)) \in \mathbb{R}^2\,,\ i = 0:n\,,$ are also called BEZIÉR nodes or BEZIÉR points. The BEZIÉR polynomial is attached to the corresponding BEZIÉR polygon like a circus tent to its masts approaching it more and more closely by enhancing the degree resp. the node number. General curves in space are obtained by replacing the BEZIÉR points b_i in (2.11) by vectors,

$$\underline{p}_{n,\mathrm{bez}}(x;\underline{f}) = \sum_{i=0}^{n} \underline{f}\left(\frac{i}{n}\right) B_i^n(x) \in \mathbb{R}^n\,,\ \ \underline{f}(x) \in \mathbb{R}^n\,.$$

(5°) The *approximation by* BEZIÉR *polynomials* provides a basic result of *functional analysis*:

Theorem 2.3. *(*WEIERSTRASS*) Let* $f \in \mathcal{C}[0,1]$ *and*

$$B_n f : x \mapsto \sum_{i=0}^{n} f\left(\frac{i}{n}\right) B_i^n(x)$$

with BERNSTEIN *polynomials* $B_i^n(x)$*. Then*
$\lim_{n\to\infty} \max_{0\leq x\leq 1} |f(x) - B_n f(x)| = 0\,.$

The proof is a simple conclusion of a surprising result of KOROVKIN:

Theorem 2.4. *Let* $L_n : \mathcal{C}[a,b] \mapsto \mathcal{C}[a,b]$ *be a sequence of linear and positive operators* L_n*, i.e.,*

$$\forall\ f,g \in \mathcal{C}[a,b]\ \ \forall\ x \in [a,b] : f(x) \leq g(x) \Longrightarrow L_n(f) \leq L_n(g)\,,$$

and let $f_1(x) = 1\,,\ f_2(x) = x\,,\ f_3(x) = x^2$*. If*

$$\lim_{n\to\infty} \|L_n f_i - f_i\|_\infty = 0\ \ i = 1:3\,,$$

then
$\forall\ f \in \mathcal{C}[a,b] : \lim_{n\to\infty} \|L_n f - f\|_\infty = 0\,.$

Proof (Kosmol) Sect. 4.4.5.

Proof of Theorem 2.3. Apparently, the operators B_n are linear and positive, and we have

$$B_n(f_1, x) = 1\,, \quad B_n(f_2, x) = x\,, \quad B_n(f_3, x) = x^2 + \frac{x - x^2}{n}\,,$$

hence the assumptions of Theorem 2.2 are satisfied. □

A direct, likewise interesting proof of Theorem 2.3 is found in (Yosida) Sect. 0.2.

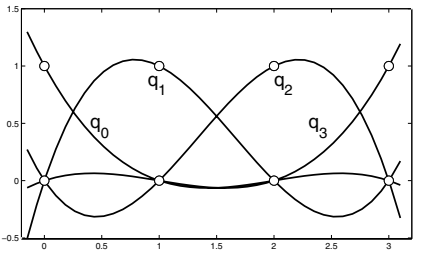

Figure 2.2. LAGRANGE polynomials, $n = 3$

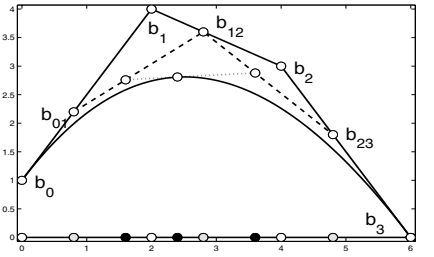

Figure 2.3. BEZIÉR polynomials, $n = 3$, with rescaling

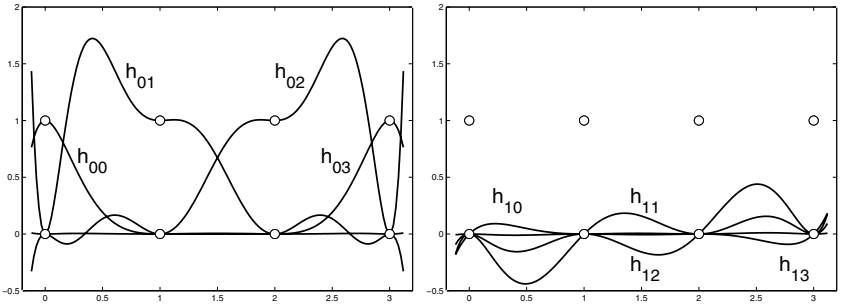

Figure 2.4. HERMITE polynomials, $n = 3$

(g) Interpolating Splines A (continuous) BEZIÉR *curve* consists piecewise of BEZIÉR polynomials having the interpolation property at the ends of their domain, respectively. We consider the special case of a BEZIÉR curve in interval $\mathcal{I} = [0, n \cdot m]$, $m \in \mathbb{N}$. The curve shall consist of BEZIÉR polynomials of degree n with BEZIÉR points $b_{nk}, \ldots, b_{n(k+1)}$ in the subintervals $\mathcal{I}_k = [n(k-1), nk]$, $k = 1 : m$, and it shall attain the values $f(nk)$ at the points nk (n fixed) (Fig. 2.5).

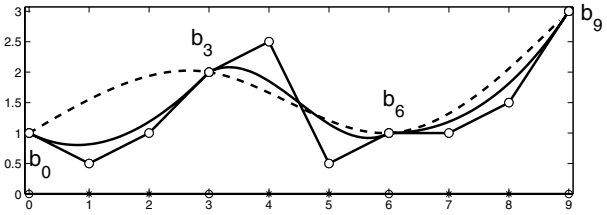

Figure 2.5. BEZIÉR curve and spline, m = n = 3

Definition 2.1. (1°) *A segmented continuous curve of polynomial segments of degree* $\leq n$ *is a* (polynomial) spline *if it is* $(n-1)$*-times continuously differentiable on the entire interval. For* $n = 3$*, the spline is called* cubic spline.
(2°) *Let* $\mathcal{I} = [a, b]$ *and let a partition of* Δ_m *of* $\mathcal{I}$ *be defined by* $a = x_0 < x_1 < \ldots < x_m = b$ *then*

$$S_3(\Delta_m) := \{s \in \mathcal{C}^2(\mathcal{I}),\ \forall\, x \in [x_{i-1}, x_i) : s^{(3)}(x) = \text{const},\ i = 1 : m\}$$

is the vector space of cubic splines.

The dimension of S_3 is $m + 3 = (m + 1) + 2$ hence there are two conditions free for further specification.

For $k \in \mathbb{N}_0$ let

$$\begin{aligned} p_k(x) &:= x^k, \\ q_k(t, x) &:= (t - x)_+^k := \max\{(t - x)^k, 0\} \quad \text{(FÖPPL symbol)}. \end{aligned}$$

The function $q_k(t, x)$ has $k-1$ continuous derivatives in both arguments and the k-th derivative makes a jump of height $k!$ resp. $(-1)^k k!$.

Theorem 2.5. *The set* $S_n(\Delta_m)$ *is a linear space of dimension* $m + n$*. The elements*

$$p_0, \ldots, p_n, q_n(\,\cdot\,, x_1), \ldots, q_n(\,\cdot\,, x_{m-1})$$

constitute a basis of $S_d(\Delta_n)$ *.*

Proof, e.g., (Haemmerlin), p. 246.

Now we consider the case $n = 3$ more exactly. Conceiving $s \in S_3$ as BEZIÉR curve, we have for the BEZIÉR points at distance $x_{i+1} - x_i = 1$:

$$\begin{aligned} s(x_k) = b_{3k} \qquad & \text{because } \ s \in \mathcal{C}[a, b] \\ 2b_{3k} = b_{3k-1} + b_{3k+1} \qquad & \text{because } \ s \in \mathcal{C}^1[a, b] \\ 2b_{3k-1} - b_{3k-2} = d_k = 2b_{3k+1} - b_{3k+2} \quad & \text{because } \ s \in \mathcal{C}^2[a, b]\,. \end{aligned} \tag{2.12}$$

By these relations we obtain

$$\begin{aligned} 4b_{3k-1} - 2b_{3k-2} = 2d_k\,, \qquad & 4b_{3k+1} - 2b_{3k+2} = 2d_k\,, \\ 2b_{3(k-1)+1} - b_{3(k-1)+2} = d_{k-1}\,, \quad & 2b_{3(k+1)-1} - b_{3(k+1)-2} = d_{k+1}\,, \end{aligned}$$

and, by addition of the left and right sides separately,

$$\boxed{3b_{3k-1} = d_{k-1} + 2d_k,\ 3b_{3k+1} = 2d_k + d_{k+1}}\ . \tag{2.13}$$

Accordingly, the numbers b_{3k+1} und $b_{3(k+1)-1} = b_{3k+2}$ divide the line segment between d_k and d_{k+1} into *three* segments.

Furthermore, by (2.12) and (2.13),

$$6b_{3k} = 3b_{3k-1} + 3b_{3k+1} = d_{k-1} + 4d_k + d_{k+1}\,, \tag{2.14}$$

for all *interior* points b_{3k}, $k = 1 : m-1$. Together with (2.13) for b_1 and b_{3m-1}, i.e., for $k = m$ and $k = 0$,

$$3b_{3m-1} = d_{m-1} + 2d_m\,,\ \ 3b_1 = 2d_0 + d_1\,,$$

we obtain the following linear system for the vector $[d_0, \ldots, d_m]^T$ of unknown coefficients (DeBoor points) in case where all data on the right side are given:

$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{m-1} \\ d_m \end{bmatrix} = \begin{bmatrix} 3b_1 \\ 6b_3 \\ \vdots \\ 6b_{3m-3} \\ 3b_{3m-1} \end{bmatrix}\ . \tag{2.15}$$

The matrix is regular and well-conditioned; thus there exists precisely one spline $s \in S_3(\Delta)$ to the data set $\{d_0, \ldots, d_m, b_0, b_{3m}\}$ by the above construction. It is called *cubic interpolating spline* because $s(x_k) = b_{3k} = f_k$ for $k = 0 : m$ (Fig. 2.6).

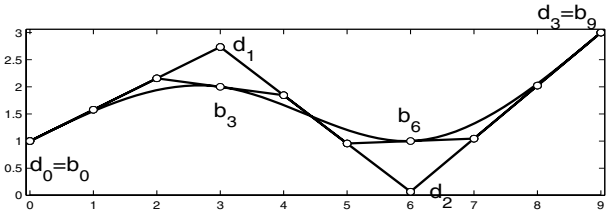

Figure 2.6. Interpolating spline, m = n = 3

Calculation Let $f_k = b_{3k}$ for $k = 0 : m$ be given and let the interpolating spline $s \in S_3(\Delta_m)$ in $[a, b] = [0, m]$, $x_{i+1} - x_i = 1$ to be found.
(1°) Let $f'(a)$, $f'(b)$ be specified. Find b_1, b_{3m-1} by solving

$$f'(a) = s'(0) = 3(b_1 - b_0),\ \ f'(b) = s'(m) = 3(b_{3m} - b_{3m-1}),$$

compute $d_0, \dots, d_m$ by (2.15), compute b_{3k+1}, b_{3k+2} by (2.13),
compute $s(x)$ in $[x_k, x_{k+1}]$ as BEZIÉR polynomial by DE CASTELJAU,

$$s(x_k + \xi) = \sum_{i=0}^{3} b_{3k+i} B_i^3(\xi)\,,$$

using local coordinates $\xi \in [0,\ 1]$.
(2°) Requiring $s''(a) = s''(b) = 0$ we obtain the *natural splines*, $s \in N_3(\Delta_m)$.
For their computation, $d_0 = b_0$ and $d_m = b_{3m}$ are prescribed then, with $n = 3$,

$$\begin{aligned} s''(0) &= n(n-1)(b_2 - 2b_1 + b_0) = 6(-d_0 + b_0) = 0, \\ s''(m) &= n(n-1)(b_{3m} - 2b_{3m-1} + b_{3m-2}) = 6(-d_m + b_{3m}) = 0\,. \end{aligned}$$

Compute $(d_1, \dots, d_{m-1})^T$ by (2.15) without first and last row (because d_0 and d_m fixed)

$$\begin{bmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ \vdots \\ \vdots \\ d_{m-1} \end{bmatrix} = \begin{bmatrix} 6b_3 - b_0 \\ 6b_6 \\ \vdots \\ 6b_{3m-6} \\ 6b_{3m-3} - b_{3m} \end{bmatrix}. \tag{2.16}$$

Because $s''(x_k) = 6(b_{3k} - d_k)/h^2$, $k = 1 : m-1$, $h = x_{i+1} - x_i$ constant, the "moments" $s''(x_k)$ satisfy a similar linear system as the values d_k.

If the exact curvature $\kappa(x) = f''(x)/(1 + f'(x)^2)^{3/2}$ of $f : x \mapsto f(x)$ is replaced by $f''(x)$ approximatively, then the natural splines reveal to be *bending lines*:

Theorem 2.6. *Let Δ_m be an arbitrary partition and let $s \in N_3(\Delta_m)$, i.e., a natural spline, with $s(x_i) = f_i$, $i = 0 : m$. Then*

$$|s|_2^2 := \int_a^b (s''(x))^2 dx = \min\left\{|g|_2^2,\ g \in \mathcal{C}^2[a,b],\ g(x_i) = f_i\right\}.$$

Proof. Let g have the mentioned properties then, using $g''(x)^2 = (s''(x) + g''(x) - s''(x))^2$,

$$\int_a^b g''(x)^2 dx = \int_a^b (s'')^2 dx + 2\int_a^b s''(g'' - s'')\,dx + \int_a^b (g'' - s'')^2 dx\,. \tag{2.17}$$

By partial integration we obtain for the mixed term

$$\begin{aligned} \int_a^b s''(g'' - s'')\,dx &= s''(g' - s')\big|_a^b - \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} s'''(g' - s')\,dx, \\ \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} s'''(g' - s')\,dx &= \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} c_i(g' - s')\,dx = \sum_{i=1}^{m} c_i(g - s)\Big|_{x_{i-1}}^{x_i} = 0\,, \end{aligned}$$

because $g(x_i) = s(x_i) = f_i$ by assumption. Accordingly, the assertion follows if the boundary terms disappear, i.e., if

$$\boxed{s''(g' - s')\Big|_a^b = 0}\,.$$

This condition is fulfilled for instance in the following commonly used cases:

(1°) if $s''(a) = 0 = s''(b)$ (natural spline),
(2°) if $g'(a) = s'(a) = f'(a)$ fixed, $g'(b) = s'(b) = f'(b)$ fixed,
(3°) if $s\,,\ g$ periodic with period $b - a\,.$

□

If calculation shall be performed in an interval $[a, b]$ instead $[0, m]$ then re-scaling becomes necessary; likewise, a change is necessary if the nodes are no longer equidistant. But we do not pursue this matter here; cf. however `SUPPLEMENT\chap02`.

2.2 Orthogonal Polynomials

Let $\overline{\Pi}_n$ be the set of polynomial $p_n \in \Pi_n$ of *exact* degree n and with highest term *one*.

(a) Construction Let $-\infty \le a < b \le \infty$ and let $\omega : [a, b] \to \mathbb{R}_+$ be a non-negative *weight function* with the following properties:

Assumption 2.1. *The* moments $m_k := \displaystyle\int_a^b \omega(x)x^k\,dx\,,\ k \in \mathbb{N}_0\,,$ *exist finitely (possibly being improper integrals), and* $m_0 > 0\,.$

Then two polynomials $p\,,\ q \in \Pi_n$ are called *orthogonal* (w.r.t. the considered interval of integration and the weight function ω) if

$$(p, q) := \int_a^b \omega(x)p(x)q(x)\,dx = 0\,.$$

Theorem 2.7. *(Existence and Construction) Adopt Assumption 2.1.*
(1°) $\forall\, i \in \mathbb{N}_0\ \ \exists!\, p_i \in \overline{\Pi}_i : i \ne k \implies (p_i, p_k) = 0\,.$
(2°) *The orthogonal polynomials are uniquely determined by the three-term recurrence relation (with* $xp : x \mapsto xp(x)$*)*

$$\begin{aligned} &p_{-1}(x) = 0\,,\ \ p_0(x) = 1\,,\ \ p_{i+1}(x) = (x - \delta_{i+1})p_i(x) - \gamma_{i+1}^2 p_{i-1}(x)\,,\ i \ge 0,\\ &\delta_{i+1} = (xp_i, p_i)/(p_i, p_i)\,,\ i \ge 0\,,\ \ \gamma_{i+1}^2 = \begin{cases} 0\,, & i = 0\,,\\ (p_i, p_i)/(p_{i-1}, p_{i-1})\,, & i \ge 1\,. \end{cases} \end{aligned} \tag{2.18}$$

Proof by GRAM-SCHMIDT *orthogonalization* (Stoer), see also `SUPPLEMENT\chap02`.

Obviously, for $p_n \in \overline{\Pi}_n$, it follows that $(p, p_n) = 0$ for all $p \in \Pi_{n-1}$ because orthogonal polynomials are linearly independent and thus form a basis of Π_n . In the remaining part of this section we consider othogonal polynomials p_n as introduced by Theorem 2.7.

Theorem 2.8. *The roots x_i of p_n are real and simple. They all lie in the open interval (a, b) .*

Proof. Let $x_1, \ldots, x_k$ be all roots of p_n of odd multiplicity contained in (a, b) then p_n changes sign precisely at these points. Let

$$a < x_1 < \ldots < x_k < b, \quad q(x) := \prod_{i=1}^{k} (x - x_i),\ k \le n\,,$$

then $p_n(x)q(x)$ does *not* change sign in (a, b) hence $(p_n, q) \neq 0$. Therefore the degree of q must be $k = n$ otherwise we have a contradiction to the above inference to Theorem 2.7. □

(b) The Formulas of Rodriguez To compute orthogonal polynomials $p_n \in \Pi_n$ *explicitely*, we observe the general condition of orthogonality

$$\forall\, q_{n-1} \in \Pi_{n-1} : \int_a^b \omega(x) p_n(x) q_{n-1}(x)\, dx = 0\,, \quad n = 0, 1, \ldots \tag{2.19}$$

and choose for approach in skilful way

$$\omega(x)p_n(x) = \frac{d^n}{dx^n} u_n(x) \implies p_n(x) = \frac{1}{\omega(x)} \frac{d^n}{dx^n} u_n(x) \in \Pi_n\,.$$

Since p_n shall be a polynomial of degree not greater n, obviously

$$\frac{d^{n+1}}{dx^{n+1}} \left[\frac{1}{\omega(x)} \frac{d^n u_n(x)}{dx^n} \right] = \left[\frac{u_n^{(n)}(x)}{\omega(x)} \right]^{(n+1)} = 0\,. \tag{2.20}$$

On the other side, a n-fold partial integration of $\displaystyle\int_a^b u_n^{(n)}(x) q_{n-1}(x)\, dx = 0$ yields

$$\left[u_n^{(n-1)} q_{n-1} - u_n^{(n-2)} q'_{n-1} + - \ldots + (-1)^{n-1} u_n q^{(n-1)} \right] \Big|_a^b = 0\,.$$

This relation is certainly fulfilled for the boundary conditions

$$u_n^{(i)}(a) = 0, \quad u_n^{(i)}(b) = 0, \quad i = 0 : n - 1\,. \tag{2.21}$$

The converse result does also hold and has been proved by (Szegoe):

Theorem 2.9. *Let Assumption 2.1 be fulfilled. Then the boundary value problem (2.20), (2.21) has always a solution u_n and $p_n := u_n/\omega$ is a polynomial of degree n .*

The above boundary value problem has $2n$ boundary conditions for a differential equation of order $2n+1$ hence one condition stands at disposition for normalization.

Example 2.1. LEGENDRE *polynomials:* Interval $(a,b) = (-1,1)$, weight function $\omega(x) \equiv 1\,, u_n^{(2n+1)} = 0\,,\ u_n^{(i)}(\pm 1) = 0\,,\ i = 0 : n-1$ (Fig. reffig0202.1).

$$p_n(x) = \gamma_n \frac{d^n}{dx^n}(x^2-1)^n.$$

The constants γ_n are specified in different ways.

Example 2.2. JACOBI *polynomials:* Interval (a,b) finite, weight function $\omega(x) = (x-a)^\alpha(b-x)^\beta\,, \alpha > -1\,,\ \beta > -1\,.$

$$p_n(x) = \gamma_n \frac{1}{(x-a)^\alpha(b-x)^\beta}\frac{d^n}{dx^n}[(x-a)^{n+\alpha}(b-x)^{n+\beta}]\,.$$

In particular, *shifted* LEGENDRE *polynomials* are obtained for $(a,b) = (0,1)$ and $(\alpha,\beta) = (0,0),\ (1,0),\ (0,1),\ (1,1)$ being applied later on in numerical integration:

$$p_{1,n}(x) = \frac{d^n}{dx^n}\left(x^n(1-x)^n\right)\,, \qquad p_{2,n}(x) = \frac{1}{x}\frac{d^n}{dx^n}\left(x^{n+1}(1-x)^n\right)$$
$$p_{3,n}(x) = \frac{1}{1-x}\frac{d^n}{dx^n}\left(x^n(1-x)^{n+1}\right)\,, p_{4,n}(x) = \frac{1}{x(1-x)}\frac{d^n}{dx^n}x^{n+1}(1-x)^{n+1}\,. \tag{2.22}$$

Example 2.3. CHEBYSHEV *polynomials* $T_n(x)$ with $(a,b) = (-1,1)\,,\ \omega(x) = (1-x^2)^{-1/2}$ are special JACOBI polynomials as well as the above LEGENDRE polynomials. In expanding a function by these polynomials, the values at the boundaries of the interval are more strongly regarded because of the special weight function (Fig. 2.8). By the original condition of orthogonality,

$$\forall\, q_{n-1}(x) \in \Pi_{n-1} : \int_{-1}^{1} \frac{T_n(x)q_{n-1}(x)}{(1-x^2)^{1/2}}\,dx = 0\,, \tag{2.23}$$

a substitution of $x = \cos\varphi$ yields the condition

$$\int_0^\pi T_n(\cos\varphi)q_{n-1}(\cos\varphi)\,d\varphi = 0\,.$$

Because

$$\cos(n+1)\varphi + \cos(n-1)\varphi = 2\cos\varphi\cos n\varphi \tag{2.24}$$

for $n \in \mathbb{N}$, the function $\cos n\varphi$ is a polynomial in $\cos\varphi$, and $(\cos\varphi)^k$ is a linear combination of $\cos j\varphi,\ j = 0:k$, hence

$$q_{n-1}(\cos\varphi) = \sum_{j=0}^{n-1} \gamma_j (\cos\varphi)^j = \sum_{k=0}^{n-1} \delta_k \cos(k\varphi)\,.$$

Thus (2.23) holds if and only if

$$\int_0^\pi T_n(\cos\varphi)\cos(j\varphi)\,d\varphi = 0,\ j = 0:n-1,$$
$$\Longrightarrow\ T_n(\cos\varphi) = \cos(n\varphi) \Longrightarrow\ \boxed{T_n(x) = \cos(n\arccos x)}\,,\ n = 0,1,\ldots.$$

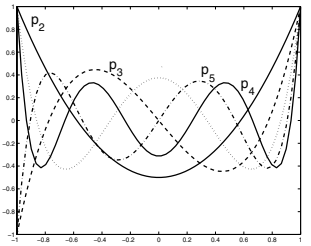

Figure 2.7. LEGENDRE polynomials, $n = 2:6$

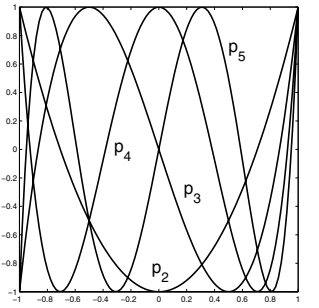

Figure 2.8. CHEBYSHEV polynomials, $n = 2:5$

(c) Minimum Property of Chebyshev Polynomials The recurrence formula (2.24) shows that $T_n(x) = \cos(n\arccos(x))$ has the highest term 2^{n-1}.

Theorem 2.10. *Let $p_n(x)$ be any polynomial of degree $\le n$ with highest term 2^{n-1}. Then there exists at least one $x \in [-1,1]$ such that $|p_n(x)| \ge 1$.*

Proof. Suppose that $|p_n(x)| < 1$ for all $x \in [-1,1]$. $T_n(x)$ takes alternating the values ± 1 at its $n+1$ extremal points $x_i = \cos(i\pi/n),\ i = 0:n$, in $[-1,1]$. Therefore, $T_n(x) - p_n(x)$ is alternating positive or negative at these extremal points and thus $T_n(x) - p_n(x)$ has at least n zero points in $(-1,1)$. However, because of identical highest terms, $T_n(x) - p_n(x)$ is a polynomial of degree $\le n-1$. Accordingly, $T_n(x) - p_n(x) \equiv 0$ in contradiction to the assumption. □

Corollary 2.1. *Let q_n be a polynomial of degree n with highest term a_n then there exists a value $x \in [-1,1]$ such that $|q_n(x)| \ge a_n/2^{n-1}$.*

Proof. Let $a_n \neq 0$ and $q_n^*(x) = q_n(x)2^{n-1}/a_n$. The polynomial q_n^* has the highest term 2^{n-1} hence $|q_n^*(x)| \geq 1$ by Theorem 2.10 for at least one $x_0 \in [-1,1]$. Then $|q_n(x_0)| = |q_n^*(x_0)a_n/2^{n-1}| \geq a_n/2^{n-1}$. □

For an arbitrary polynomial $q_n(x)$ — especially also for a TAYLOR polynomial — there exists a unique expansion by CHEBYSHEV polynomials,

$$q_n(x) = \sum_{i=0}^{n} c_i T_i(x)\,, \quad x \in [-1,1]\,, \quad T_i(x) = \cos(i \arccos(x))\,,$$

because these polynomials are linearly independent by orthogonality.

Theorem 2.11. *If $S_n(x) = \sum_{i=0}^{n} c_i T_i(x)$ are the partial sums of an expansion by* CHEBYSHEV *polynomials then*

$$\max_{-1\leq x\leq 1} |S_{n+1}(x) - S_n(x)| = \inf_{p_n} \max_{-1\leq x\leq 1} |S_{n+1}(x) - p_n(x)|$$

where p_n is an arbitrary polynomial of degree $\leq n$.

Proof. We have $S_n(x) - S_{n-1}(x) = c_n T_n(x)$ hence $|S_n(x) - S_{n-1}(x)| \leq |c_n|\,,\ -1 \leq x \leq 1$. For any arbitrary polynomial p_{n-1} of degree $n-1$, the difference $S_n - p_{n-1}$ has the highest term $c_n 2^{n-1}$ hence

$$|S_n(x) - p_{n-1}(x)| \geq c_n 2^{n-1}/2^{n-1} = c_n\,.$$

at least for one $x \in [-1,1]$ by Corollary 2.1. □

Roughly spoken, the components of an expansion by CHEBYSHEV polynomials decrease in absolute value in the fastest way.

2.3 Numerical Integration

Integrating is an art and differentiating is a handicraft as everybody knows, but from numerical point of view fortunately the situation behaves conversely in some sense. Integration is a smoothing process which has advantageous consequences in numerical approximation whereas a differential quotient has to be replaced always by a difference quotient numerically. Then, in numerator and denominator, subtraction of nearly equal numbers does occur entailing the befeared extinction of leading numbers. However, it should be mentioned at this place that an asymptotic expansion in the sense of Sect. 2.4**(c)** may produce surprisingly exact results; cf. (Rutishauser).

As MATLAB does not know the index zero and also for applications later on, we work in this section throughout with n nodes instead of the usual $n+1$ nodes in interpolating problems.

(a) Integration Rules of Lagrange The computational effort of a numerical integration formula depends on the number of function evaluations. Note once more that we work with n nodes here to compare the individual

rules with each other. Accordingly, we proceed from an interpolating polynomial $p_{n-1}(x)$ of degree $n-1$ of LAGRANGE type, i.e., by (2.6) in slightly modified form,

$$f(x) \approx p_{n-1}(x;f) = \sum_{i=1}^{n} f(x_i) q_i(x)\,, \quad q_i(x) = \prod_{j=1, j\neq i}^{n} \frac{x-x_j}{x_i-x_j} \in \Pi_{n-1}\,, \tag{2.25}$$

and obtain by integration over an interval $(a\,,\,b)$

$$I(f) := \int_a^b f(x)\,dx \approx \sum_{i=1}^{n} f(x_i) \int_a^b q_i(x)\,dx =: \sum_{i=1}^{n} f(x_i)\alpha_i =: I_n(f)\,. \tag{2.26}$$

The n support abszissas x_i shall be *mutually distinct* again, otherwise they can be chosen arbitrarily, in particular, they may lie also in the exterior of the interval of integration. But in this section we suppose always that

$$a \le x_1 < \ldots < x_{n-1} < x_n \le b\,.$$

An *integration* rule has *degree* N if (at least) all polynomials of degree $\le N$ are integrated exactly. Apparently, a LAGRANGE formula (2.26) with n abszissas has degree $N = n-1$; the *maximum degree* N is not greater than $2n-1$, because, inserting the polynomial $f(x) = \Pi_{i=1}^{n}(x-x_i)^2 \in \Pi_{2n}$, we have $I_n(f) = 0$ for the rule and $I(f) > 0$ for the exact integral.

The integration formulas of NEWTON and COTES are of separated type (2.25), too, but with a uniform partition $x_i = a + (i-1)h$ and step length $h = (b-a)/(n-1)\,,\ n \ge 2$. Inserting the translation $x = a + sh$ we obtain by Sect. 2.1

$$\begin{aligned} q_i(x) \;&=\; q_i(a+sh) \;=: \varphi_i(s) = \prod_{j=1,\,j\neq i}^{n} \frac{s-j}{i-j} \in \Pi_{n-1}(s)\,, \quad s \in [0,\,n-1]\,, \\ \alpha_i \;&:=\; \int_a^b q_i(x)dx \;=\; \int_0^{n-1} q_i(a+sh)\frac{dx}{ds}ds = h\int_0^{n-1} \varphi_i(s)ds = h\,\beta_i\,, \\ I_n(f) \;&=\; \sum_{i=1}^{n} f(x_i)\alpha_i \;=\; h\sum_{i=1}^{n} f(x_i)\beta_i\,. \end{aligned}$$

The new weights β_i are now *rational* numbers which depend only on the number n and no longer on the boundaries $a\,,\ b$ of the integral, therefore they can be calculated once for all in tabular form; a substitution of $f(x) \equiv 1$ shows that $\sum_{i=1}^{n} \beta_i = n-1\,,\ n \ge 2$.

Example 2.4. Midpoint rule (1 node):

$$I(f) = (b-a) f\left(\frac{a+b}{2}\right) + \frac{1}{24}(b-a)^3 f''(\xi)\,,$$

Trapezoidal rule ($n = 2$ nodes, $h = b - a$):

$$I(f) = \frac{b-a}{2}[f(a) + f(b)] - \frac{1}{12}(b-a)^3 f''(\xi)\,,$$

SIMPSON*'s rule* ($n = 3$ nodes, $h = (b-a)/2$):

$$I(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] - \frac{(b-a)^5}{2^5 \cdot 90} f^{(4)}(\xi)\,.$$

Note that the intermediate values $\xi \in (a,\, b)$ differ in the individual integration rules.

One observes that the midpoint rule and SIMPSON'S also called KEPLER'S rule have degree n instead of the expected lower degree $n-1$ of the underlying interpolating polynomial. It is however a general property of NEWTON-COTES rules that the degree is n instead $n-1$ for n odd by reason of symmetry.

The general error term $R_n(f)$ in

$$I(f) = I_n(f) + R_n(f)\,, \tag{2.27}$$

is a linear operator satisfying $R_n(\alpha f + \beta g) = \alpha R_n(f) + \beta R_n(g)$. Introducing the FÖPPL-Symbol $(x-t)_+^N := \max\{(x-t)^N,\, 0\}$ again, the following classical result of PEANO displays $R_n(f)$ in elegant integral form; cf. e.g. (Stoer).

Theorem 2.12. *Let an integration rule (2.26) with n nodes have degree N then, for all $f \in \mathcal{C}^{N+1}[a,\, b]$,*

$$R_n(f) = \int_a^b f^{(N+1)}(t) K_n(t)\, dt\,, \quad K_n(t) = \frac{1}{N!} R_n(h_t), \quad h_t : x \mapsto (x-t)_+^N\,.$$

Proof see (Stoer) and `SUPPLEMENT\chap02a`.

$R_n(h_t)$ denotes here the error of the integration rule w.r.t the function $h_t : x \mapsto (x-t)_+^N$ instead of f. Frequently the *kernel* $K_n(t)$ does not change sign in $(a\,,\, b)$ as for instance in NEWTON-COTES rules. Then the mean value theorem of integration yields

$$R_n(f) = f^{(N+1)}(\xi) \int_a^b K_n(t) dt\,, \quad \xi \in (a, b)\,. \tag{2.28}$$

Inserting here the special function $\varphi : x \mapsto x^{N+1}$ for f we obtain

$$R_n(\varphi) = (N+1)! \int_a^b K_n(t) dt \Longrightarrow \int_a^b K_n(t) dt = R_n(\varphi)/(N+1)!\,. \tag{2.29}$$

Fazit: If the integration rule (2.26) has degree N and $K_n(t)$ does not change sign in interval of integration then (2.28) and (2.29) yields the error representation

$$\boxed{R_n(f) = \frac{f^{(N+1)}(\xi)}{(N+1)!}\, R_n(\varphi)\,, \quad \varphi : x \mapsto x^{N+1},\ \xi \in (a,b)} \tag{2.30}$$

(PEANO's error representation); but $R_n(\varphi)$ can be always calculated exactly!

As already noted above, it can be shown for NEWTON-COTES rules with n mutually distinct support abszissas that

$$R_n(f) = \begin{cases} \dfrac{f^{(n)}(\xi)}{(n)!}\, R_n(x^n) & n \text{ even} \\ \dfrac{f^{(n+1)}(\xi)}{(n+1)!}\, R_n(x^{n+1}) & n \text{ odd} \end{cases}$$

where $\xi \in (a,b)$.

(b) Composite Integration Rules As already mentioned in Sect. 2.1, the approximation of f by an interpolating polynomial $p_n \in \Pi_n$ is not improved by enhancing the degree n. Therefore one uses locally polynomials of low degree on a collection of subintervals. The full integral is then approximated by the sum of the approximations of the subintegrals. The resulting *composite rules* are arbitrarily exact in dependence of the node number even for continuous integrand f. For instance, writing $x_i = a + ih,\ i = 0 : m,\ h = (b-a)/m,\ m \in \mathbb{N}$, ($m+1$ support abszissas) we obtain the important *composite trapezoidal rule* from the simple trapezoidal rule

$$\boxed{T(h;f) = \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{m-1} f(x_i) + f(x_m)\right] = I(f) - h^2(b-a)\frac{1}{12}f''(\xi)}\,, \tag{2.31}$$

and the simple SIMPSON rule leads to the corresponding composite rule

$$I(f) = \frac{h}{6}\left[f(x_0) + 2\sum_{i=1}^{m-1} f(x_i) + 4\sum_{i=0}^{m-1} f\left(\frac{x_i + x_{i+1}}{2}\right) + f(x_m)\right] + h^4(b-a)\frac{1}{2880}f^{(4)}(\xi)\,.$$

By applying EULER-MCLAURIN's summation formula (Stoer) it can be shown that formula (2.31) has the following surprising property:

Lemma 2.3. *Let $f \in \mathcal{C}^\infty(\mathbb{R})$ be an $(b-a)$-periodic function then*

$$T(h;f) = \int_a^b f(x)dx + \mathcal{O}(h^p) \quad \forall\, p \in \mathbb{N}\,.$$

In other words, the composite trapezoidal rule is faster convergent than every power of step length h for smooth periodic functions!

(c) Gauß Integration Using an interpolating polynomial of HERMITE form instead of LAGRANGE form for integration, cf. Sect. 2.1**(e)**, we obtain a further type of integration rules, namely,

$$I_n(f) := \sum_{i=1}^{n} \left[f(x_i) \int_a^b h_{0,i}(x)dx + f'(x_i) \int_a^b h_{1,i}(x)dx \right] \tag{2.32}$$

where

$$h_{0,i}(x) = \left[1 - 2q_i'(x_i)(x - x_i)\right] q_i(x)^2\,, \quad h_{1,i}(x) = (x - x_i)q_i(x)^2\,, \tag{2.33}$$

and $q_i(x) \in \Pi_{n-1}$ are the LAGRANGE polynomials. The formula has degree $N = 2n - 1$ for n evaluations of f and n evaluations of the derivative of f.

Choosing now for abszissas the roots $x_i\,,\ i = 1 : n$ of orthogonal polynomials $p_n(x) \in \Pi_n$ w.r.t. the *scalar product* $(f, g) = \int_a^b f(x)g(x)\,dx$ we obtain by Sect. 2.2

$$\int_a^b h_{1,i}(x)\,dx = \int_a^b (x - x_i)q_i(x)^2\,dx = 0$$

because $(x - x_i)q_i(x) = p_n(x)$ and $q_i(x) \in \Pi_{n-1}$. As an inference we have also $\int_a^b h_{0,i}(x)\,dx = \int_a^b q_i(x)^2\,dx$ therefore, by (2.32) and (2.33), we obtain the following integration rules

$$I_n(f) := \sum_{i=1}^{n} f(x_i) \int_a^b q_i(x)^2\,dx \tag{2.34}$$

which have *maximum* degree $N = 2n - 1$ for n nodes.

We summarize the result for a general weight function $\omega(x)$ with the properties of Sect. 2.2 in the following theorem:

Theorem 2.13. *(*GAUSS *Integration) Let $p_n \in \overline{\Pi}_n$ be orthogonal polynomials w.r.t. the scalar product*

$$(f, g) := \int_a^b \omega(x)f(x)g(x)\,dx\,,$$

Let $x_1, \ldots, x_n$ be the roots of p_n, and let

$$A = [p_i(x_j)]_{i=0\ j=1}^{n-1\ n}, \quad \underline{c} = [(p_0, p_0), 0, \ldots, 0]^T\,.$$

(1°) *The matrix A is regular.*
(2°) *Let $\underline{b} = A^{-1}\underline{c}$ and $\underline{b} = [\beta_1, \ldots, \beta_n]^T$ then*

$$\forall\ p \in \Pi_{2n-1} : \int_a^b \omega(x)p(x)\,dx = \sum_{i=1}^{n} \beta_i p(x_i)\,, \tag{2.35}$$

i.e., the integration rule

$$\int_a^b \omega(x)f(x)\,dx = \sum_{i=1}^{n} \beta_i f(x_i) + R_{n,\omega}(x; f) \tag{2.36}$$

has maximum degree $N = 2n - 1$.
(3°) *For error representation in (2.36) we have*

$$\forall\, f \in \mathcal{C}^{2n}[a,\, b]\ \ \exists\, \xi \in (a,\, b) : R_{n,\omega}(x; f) = \frac{f^{(2n)}(\xi)}{(2n)!}(p_n,\, p_n)\,.$$

(4°) *Conversely, if (2.35) holds then the abszissas* x_i *are the roots of the orthogonal polynomials* $p_n(x)$ *and* $A\underline{b} = \underline{c}$ *is fulfilled with* $\underline{b} = [\beta_1, \ldots, \beta_n]^T$.
(5°) *If a rule (2.36) has degree* $N \geq n - 2$ *then the weights* β_i *are positive.*

Proof see (Stoer).

(d) Suboptimal Integration Rules are an important tool in constructing implicit RUNGE-KUTTA methods of maximum order in the next section. Let

$$\underline{b} = [\beta_1, \ldots, \beta_n]^T\,,\ \ \underline{x} = [x_1, \ldots, x_n]^T\,,\ \ F(\underline{x}) = [f(x_1), \ldots, f(x_n)]^T\,.$$

Theorem 2.14. *For* $\delta, \varepsilon \in \{0,\, 1\}$ *, there exists a unique integration rule*

$$\int_0^1 f(x)\, dx \approx \delta\beta_0\, f(0) + \underline{b}^T F(\underline{x}) + \varepsilon\, \beta_{n+1} f(1) \tag{2.37}$$

of maximum degree $\widetilde{N} = 2n + \delta + \varepsilon - 1$.

Choose GAUSS weights and GAUSS nodes $\widetilde{\underline{b}}$, $\underline{x}$ by Theorem 2.13 w.r.t. the weight function $\omega^*(t) = t^\delta(1-t)^\varepsilon$ in [0, 1] and insert $\underline{b} = [\widetilde{\beta}_i/\omega^*(x_i)]_{i=1}^n$. Then the rule is optimal for $(\delta, \varepsilon) = (0, 0)$. For $(\delta, \varepsilon) = (1, 0)$ or $(\delta, \varepsilon) = (0, 1)$, the remaining weight is found by $1 = \delta\beta_0 f(0) + \underline{b}^T \underline{e} + \varepsilon\beta_{n+1} f(1)$. For $(\delta, \varepsilon) = (1, 1)$, the remaining both weights are found by solving

$$\frac{1}{2} = 0 + \underline{b}^T \underline{x} + \beta_{n+1}\,,\ \ 1 = \beta_0 + \underline{b}^T \underline{e} + \beta_{n+1}\,.$$

A comparison with the shifted LEGENDRE polynomials in (2.22) shows that the node abszissas $x_1, \ldots, x_n$ of a rule (2.37) with together n nodes are the roots of the following polynomials:

$$\begin{aligned} &(\delta, \varepsilon) = (0, 0) : p_{1,n}(x)\,, && (\delta, \varepsilon) = (1, 0) : x p_{2,n-1}(x) \\ &(\delta, \varepsilon) = (0, 1) : (1 - x) p_{3,n-1}(x)\,, && (\delta, \varepsilon) = (1, 1) : x(1 - x) p_{4,n-2}(x)\,. \end{aligned} \tag{2.38}$$

Proof of Theorem 2.14 see `SUPPLEMENT\chap02a`. A program for the computation of nodes and weights in all four cases is found in
`KAPITEL02\SECTION_1_2_3`. For integration over an interval (a, b) rescale nodes and weights by $\widehat{x}_i = a + (b - a)x_i$, $\widehat{b}_i = (b - a)b_i$, $i = 1 : n$.

Example 2.5. GAUSS integration with LEGENDRE polynomials:

$$\int_{-1}^1 f(x)\, dx \approx \sum_{i=1}^n \beta_i f(x_i)$$

Table 2.1. GAUSS-LEGENDRE formulas with n nodes

n	x_i	β_i
2	$\pm\sqrt{3}/3$	1
3	0	8/9
	$\pm\sqrt{15}/5$	5/9
4	$\pm\left[525-70\sqrt{30}\right]^{1/2}/35$	$(18+\sqrt{30})/36$
	$\pm\left[525+70\sqrt{3}\right]^{1/2}/35$	$(18-\sqrt{30})/36$
5	0	128/225
	$\pm\left[245-14\sqrt{70}\right]^{1/2}/25$	$(322+13\sqrt{70})/900$
	$\pm\left[245+14\sqrt{70}\right]^{1/2}/25$	$(322-13\sqrt{70})900$

These rules for integration over the interval $[a,\, b]=[-1,\, 1]$ are exact for polynomials of degree n. In a transformation to the interval $[a',\, b']$, the weights and nodes have to be transformed:

$$w_i'=\frac{b'-a'}{b-a}w_i\,,\quad x_i'=a'+\frac{b'-a'}{b-a}(x_i-a)\,.$$

For instance, in transformation to the unit interval $[0,\, 1]$, the weights must be divided by two and the nodes $x_i'=(1+x_i)/2$ are to be used.

(e) Barycentric Coordinates serve mainly to a lucid and concise representation of interpolating polynomials which *live* on arbitrary triangles T in the plane or, more general, on n-simplices in $\mathbb{R}^n$. Also integration rules for general polynomials on these domains can be simplified by this way. We restrict ourselves to the plane and consider an arbitrary triangle T in cartesian (x,y)-coordinates with vertices $P_i(x_i,y_i)$, $i=1,2,3$, being numerated counterclockwise. Then the *double surface area*

$$2|T|=(x_2-x_1)(y_3-y_1)-(x_3-x_1)(y_2-y_1)=x_{21}y_{31}-x_{31}y_{21}\,,\quad(2.39)$$

($x_{21}=x_2-x_1$ etc.) is positive as long as T is non-degenerated. Using the notations of Figure 2.9, the (dimensionless) *barycentric* or *area* coordinates are defined for $0\le\zeta_i\le 1$ by

$$\boxed{\zeta_i=\frac{\text{area of } T_i}{\text{area of } T}\,,\quad i=1,2,3}\,.$$

Accordingly, we have $P_1\simeq(1,0,0)$, $P_2\simeq(0,1,0)$, $P_3\simeq(0,0,1)$ and $\zeta_1+\zeta_2+\zeta_3=1$ whence the barycentric coordinates are *not* linearly independent.

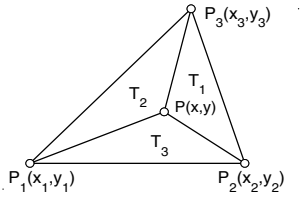

Figure 2.9. Barycentric coordinates

The connection of cartesian and barycentric coordinates is provided by the area rule:

$$2|T_1| = \begin{vmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}, \quad 2|T_2| = \begin{vmatrix} 1 & x & y \\ 1 & x_3 & y_3 \\ 1 & x_1 & y_1 \end{vmatrix}, \quad 2|T_3| = \begin{vmatrix} 1 & x & y \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix}.$$

Expanding w.r.t. the first row and dividing by $2|T|$, we get the affin-linear relations

$$\begin{aligned} \zeta_1 &= \left[(x_2y_3 - x_3y_2) + y_{23}x + x_{32}y\right]/(2|T|) \\ \zeta_2 &= \left[(x_3y_1 - x_1y_3) + y_{31}x + x_{13}y\right]/(2|T|) \\ \zeta_3 &= \left[(x_1y_2 - x_2y_1) + y_{12}x + x_{21}y\right]/(2|T|) \end{aligned} \tag{2.40}$$

(note the cyclic permutation of indices modulo 3). These relations are valid for an arbitrary cartesian coordinate system not necessarily having origin in the center of the triangle. They are used in different applications, in particular for the calculation of partial derivatives, e.g. $\partial\zeta_1/\partial x = y_{23}/(2|T|)$ etc.. Resolution of two equations in (2.40) w.r.t. x and y yields the relation

$$\boxed{1 = \zeta_1 + \zeta_2 + \zeta_3\,,\ x = x_1\zeta_1 + x_2\zeta_2 + x_3\zeta_3\,,\ y = y_1\zeta_1 + y_2\zeta_2 + y_3\zeta_3} \tag{2.41}$$

between arbitrary cartesian and barycentric coordinates. In *unit triangle* $S(\xi,\eta)$ with vertices $Q_1(0,0)\,,\ Q_2(1,0)\,,\ Q_3(0,1)$, we have the relation

$$\boxed{\zeta_1 = 1 - \xi - \eta\,,\ \zeta_2 = \xi\,,\ \zeta_3 = \eta}\,, \tag{2.42}$$

and

$$\int_S \xi^p\eta^q\,d\xi d\eta = \int_0^1 \int_0^{1-\eta} \xi^p\eta^q\,d\xi d\eta = \frac{p!q!}{(p+q+2)!}\,. \tag{2.43}$$

The formula of Holand and Bell (1969) for general triangles T,

$$\boxed{\int_T \zeta_1^m\zeta_2^n\zeta_3^p\,dxdy = 2|T|\,\frac{m!n!p!}{(m+n+p+2)!}} \tag{2.44}$$

then follows by substitution (Bell). Its straightforward generalization to tetrahedrons $T \subset \mathbb{R}^3$ with volume $|T|$ reads

$$\int_T \zeta_1^m\zeta_2^n\zeta_3^p\zeta_4^q\,dxdydz = 6|T|\,\frac{m!n!p!q!}{(m+n+p+q+3)!}\,. \tag{2.45}$$

Example 2.6. (Ciarlet79) Let $\underline{x}_i \in \mathbb{R}^n$, $i = 1 : n+1$, be the vertices of an n-simplex in $\mathbb{R}^n$, e.g., a triangle T in $\mathbb{R}^2$ or a tetrahedron in $\mathbb{R}^3$. Let $\underline{x}_{ij} = (\underline{x}_i + \underline{x}_j)/2$ for $i < j$ the midpoints of the edges, $\underline{x}_{ijk} = (\underline{x}_i + \underline{x}_j + \underline{x}_k)/3$ for $i < j < k$, and $\underline{x}_{iij} = (2\underline{x}_i + \underline{x}_j)/3$ for $i \neq j$. Denote again by Π_m the vector space of polynomials up to degree m with n variables in $\mathbb{R}^n$. Then the following *identities* are valid in $\mathbb{R}^n$:

$$
\begin{aligned}
\forall\, p \in \Pi_1 : p &= \textstyle\sum_{i=1:n+1} p(\underline{x}_i)\zeta_i \\
\forall\, p \in \Pi_2 : p &= \textstyle\sum_{i=1:n+1} p(\underline{x}_i)\zeta_i(2\zeta_i - 1) + \sum_{i<j} p(\underline{x}_{ij})4\zeta_i\zeta_j \\
\forall\, p \in \Pi_3 : p &= 2^{-1} \textstyle\sum_{i=1:n+1} p(\underline{x}_i)\zeta_i(3\zeta_i - 1)(3\zeta_i - 2) \\
&\quad + 2^{-1} \textstyle\sum_{i<j} p(\underline{x}_{ij})9\zeta_i\zeta_j(3\zeta_i - 1) \\
&\quad + \textstyle\sum_{i<j<k} p(\underline{x}_{ijk})27\zeta_i\zeta_j\zeta_k \\
\forall\, p \in \Pi_3 : p &= \textstyle\sum_{i=1:n+1} p(\underline{x}_i)\left(-2\zeta_i^3 + 3\zeta_i^2 - 7\zeta_i \sum_{j<k, j\neq i, k\neq i} \zeta_j\zeta_k\right) \\
&\quad + 27 \textstyle\sum_{i<j<k} p(\underline{x}_{ijk})\zeta_i\zeta_j\zeta_k \\
&\quad + \textstyle\sum_{i\neq j} \nabla p(\underline{x}_i)(\underline{x}_j - \underline{x}_i)\zeta_i\zeta_j(2\zeta_i + \zeta_j - 1)\,.
\end{aligned}
$$

Up to the first both, these identities are not trivial and they are not unique w.r.t. the barycentric coordinates ζ_i because of the linear interdependence. A corresponding formula for MORLEY's second order polynomial and for ARGYRIS' fifth order polynomial is given in Sect. 12.5 (both in $\mathbb{R}^2$ and using normal derivatives in $\underline{x}_{ij}$). See also `SUPPLEMENT\chap09e\chap09f`.

Integration of interpolating polynomials over triangles and more general geometric configurations is a basic tool in the construction of finite elements; see Chap. 9. For triangles we may use (2.42) or, in case of (x, y)-coordinates, the affin-linear transformation

$$
x = x_1 + x_{21}\xi + x_{31}\eta\,, \quad y = y_1 + y_{21}\xi + y_{31}\eta\,, \tag{2.46}
$$

and then apply (2.43) to integrate over the unit triangle S. Or we integrate directly over area coordinates and use BELL's formula (2.44); for instance

$$
\begin{aligned}
\forall\, p \in \Pi_1 : \int_T p(x,y)\,dxdy &= \frac{|T|}{3} \textstyle\sum_{i=1:3} p(\underline{x}_i) \\
\forall\, p \in \Pi_2 : \int_T p(x,y)\,dxdy &= \frac{|T|}{3} \textstyle\sum_{1\le i<j\le 3} p(\underline{x}_{ij})\,.
\end{aligned}
$$

The use of area coordinates together with BELL's formula is the natural choice for triangular elements. It allows to obtain convenient expressions for various integrals in finite element approach without time consuming numerical procedures.

(f) Domain Integrals

(f1) GAUSSian rules apply also to integration over the unit square,

$$
\int_{-1}^{1}\int_{-1}^{1} f(x,y)dx \approx \sum_{i=1}^{n}\sum_{j=1}^{n} \beta_i\,\beta_j\, f(x_i, x_j)\,;
$$

other squares have to be rescaled properly. With the data of Table 2.1 these rules are exact for polynomials

$$p(x,y) = \sum_{i=0}^{N} \sum_{k=0}^{N} a_{ik}\, x^i y^k \text{ with } N \le 2n-1\,,\ n = 2:5\,.$$

(f2) Abszissas and weights of two commonly used GAUSSian rules in unit triangle $S(\xi,\, \eta)$ with vertices $Q(0,0)\,,\ Q(1,0)\,,\ Q(0,1)\,,$

$$\int_S f(\xi,\eta)\, d\xi d\eta \approx |S| \sum_{i=1}^{m} \gamma_i f(\xi_i,\eta_i)\,,\ \ |S| = \frac{1}{2}\,,$$

are given in Table 2.2. These rules are exact for polynomials $p(\xi,\eta) = \sum_{0\le i+k\le n} a_{ik}\, \xi^i \eta^k$ of total degree $n = 2,5$. Integration rules for polynomials on an arbitrary triangle T,

$$\int_T f(x,y) dx dy \approx |T| \sum_{i=1}^{m} \gamma_i\, f(\widetilde{x}_i, \widetilde{y}_i)\,,$$

follow then easily by substitution with the mapping (2.46) (Fig. 2.10).

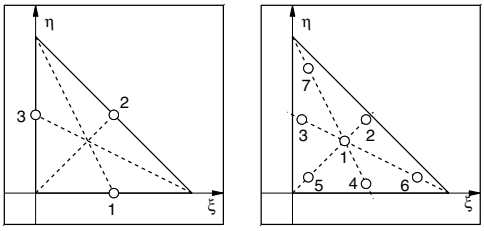

Figure 2.10. GAUSS abszissas in unit triangle, n = 2, 5

Table 2.2.

n	i	ξ_i	η_i	γ_i
2	1	1/2	0	1/3
	2	1/2	1/2	1/3
	3	0	1/2	1/3
5	1	1/3	1/3	0.225
	2	a	a	$(155+\sqrt{15})/1200$
	3	b	a	$(155+\sqrt{15})/1200$
	4	a	b	$(155+\sqrt{15}/1200$
	5	c	c	$(155-\sqrt{15})/1200$
	6	d	c	$(155-\sqrt{15})/1200$
	7	c	d	$(155-\sqrt{15})/1200$

a	$(6+\sqrt{15})/21$
b	$(9-2\sqrt{15})/21$
c	$(6-\sqrt{15})/21$
d	$(9+2\sqrt{15})/21$

(A rule for $n = 3$ using only four nodes is not recommended because of a negative weight γ and that with positive weights has seven nodes as the rule of order 5 .)

(f3) In direct integration over triangle T w.r.t. global (x, y)-coordinates we lastly have to find integrals of monomials

$$\begin{aligned} P_{rs} &= \int_T x^r y^s \, dxdy \\ &= 2|T| \int_S (x_1\zeta_1 + x_2\zeta_2 + x_3\zeta_3)^r (y_1\zeta_1 + y_2\zeta_2 + y_3\zeta_3)^s \, d\zeta_2 d\zeta_3 \\ &= 2|T| \int_S (x_1 + x_{21}\xi + x_{31}\eta)^r (x_1 + y_{21}\xi + y_{31}\eta)^s \, d\xi \, d\eta \end{aligned} \quad . \qquad (2.47)$$

By this way, the integrals P_{rs} are reduced to sums of integrals of the form (2.44) resp. (2.43). The last formula uses again the substitution rule (2.46) for the mapping $g : S \to T$ of (9.20).

Some results are assembled in Table 2.3 for polynomials up to degree $n = 5$ (Bell) where the origin of the cartesian KOS is the center of the triangle for simplicity. The concise representation in this table is however lost beyound degree 5 and for a coordinate system with different position but nowadays a *program* replaces large tables. `KAPITEL02\SECTION_1_2_3\bell1.m` supplies values of the integral (2.47) for arbitrary r , $s \in \mathbb{N}$ in an KOS with arbitrary origin by using Symbolic Mathematics.

Table 2.3.

Order $n = r + s$	$P_{rs}(x,y) = \int_T x^r y^s \, dxdy$
1	$P_{rs}(x,y) = 0$
2	$P_{rs}(x,y) = \lvert T\rvert \, (x_1^r y_1^s + x_2^r y_2^s + x_3^r y_3^s) \, /12$
3	$P_{rs}(x,y) = \lvert T\rvert \, (x_1^r y_1^s + x_2^r y_2^s + x_3^r y_3^s) \, /30$
4	$P_{rs}(x,y) = \lvert T\rvert \, (x_1^r y_1^s + x_2^r y_2^s + x_3^r y_3^s) \, /30$
5	$P_{rs}(x,y) = 2\lvert T\rvert \, (x_1^r y_1^s + x_2^r y_2^s + x_3^r y_3^s) \, /105$

References: (Kardestuncer), (Stoer).

2.4 Initial Value Problems

In this section, vectors are *not* underlined for simple representation. The letter x denotes always the exact solution and y the numerical approximation.

(a) Euler's Method We seek a solution $x : [0, T] \to \mathbb{R}^n$ of the initial value or CAUCHY problem

$$x'(t) = f(t, x(t)), \; 0 \le t \le T, \; x(0) = x_0 \, . \tag{2.48}$$

The problem is said to be *autonomous* if f does not depend explicitely on the independent variable t, i.e. $x'(t) = f(x(t))$.

For solving (2.48) numerically either we can replace $x'(t)$ by a numerical differentiation formula or we can transform the differential equation into an integral equation

$$x(t+\tau) = x(t) + \int_t^{t+\tau} f(s, x(s)) ds \, , \; \tau \text{ step length},$$

and then replace the integral by a numerical integration rule. The most simple case $\int_t^{t+\tau} f(s, x(s)\, ds \simeq \tau f(t, x(t))$ leads immediately to the *explicit* EULER *method*,

$$y(t+\tau) = y(t) + \tau f(t, y(t)), \; t = j\tau \, , \; j = 0, 1, \ldots \, , \; y(0) = x(0) = x_0 \, . \tag{2.49}$$

A substitution of the (unknown) *exact* solution x into the approximation formula (2.49) yields the *defect* or, after dividing by the step length τ, the *discretization error* of this method:

$$d(t, x, \tau) = \frac{x(t+\tau) - x(t)}{\tau} - f(t, x(t)) \, .$$

It measures the exactness with which the exact solution satisfies the *approximation formula* (2.49) and represents also the *local error* in explicit methods as in the present case. If namely $y(t) = x(t)$ is *exact* then we obtain for a single step

$$x(t+\tau) - y(t+\tau) = x(t+\tau) - x(t) + \tau f(t, x(t)) = \tau d(t, x, \tau)\,.$$

The discretization error is always calculated by using a TAYLOR expansion of the solution, e.g., with integral error term,

$$x(t+\tau) = x(t) + \tau\, f(t, x(t)) + \tau^2 \int_0^1 (1-\sigma) x''(t+\sigma\,\tau)\, d\sigma\,,$$

and thus, for the method (2.49),

$$\|d(t, x, \tau)\| \le \tau^p \int_0^1 \|x''(t+\sigma\,\tau)\|\, d\sigma\,, \quad p = 1\,.$$

Accordingly we say that the method (2.49) has order $p = 1$.

Furthermore, we obtain for the *global* error $e(t) = x(t) - y(t)$ by subtraction and application of LIPSCHITZ boundedness, i.e., $\|f(t,u) - f(t,v)\| \le L\,\|u - v\|$ in a suitable domain:

$$\begin{aligned} e(t+\tau) &= e(t) + \tau\,[f(t, x(t)) - f(t, y(t))] + \tau\, d(t, x, \tau)\,, \\ \|e(t+\tau)\| &\le (1 + L\tau)\|e(t)\| + \tau\,\|d(t, x, \tau)\|\,. \end{aligned}$$

An induction then yields an estimation of the global error where the inequality $(1+x)^n \le e^{nx}\,,\ x \ge -1$, is applied for optical reasons:

Lemma 2.4. *(Error Estimation, Convergence) Let $x \in \mathcal{C}^2[0,T]$ be a solution of (2.48) and let f be* LIPSCHITZ*-bounded then*

$$\|e(t)\| \le e^{Lt}\|e(0)\| + \frac{e^{Lt} - 1}{\tau\, L}\, \max_{0 \le s \le t} \tau\, \|d(s, x, \tau)\|\,,\ t = n\tau\,,\ n = 1, 2, \ldots\,.$$

The step length τ is canceled out once. Basically, one power of the step length τ is lost by passing from the local to the global error. This *a-priori error bound* contains the unknown solution x on the right side of the inequality. Therefore it makes only sense in theoretical studies or in comparing different methods with each other but not in practical applications. *A-posteriori error bounds* estimating the error by calculated data are much more difficult to find, therefore one contents himself here usually with *assessed valuation*. The above error bound is sharp for $x' = Lx$ with $L > 0$, and the problem is badly conditioned for large $L \cdot T$. For $L < 0$ this estimation does not make any sense and thus further criteria are necessary to qualify numerical approximations.

(b) General One-step Methods

Example 2.7. The computational device

$$\boxed{y(t+\tau) = y(t) + \tau\big[\omega f(t+\tau, y(t+\tau)) + (1-\omega) f(t, y(t))\big]}\,, \tag{2.50}$$

$0 \le \omega \le 1$, yields the explicit EULER method for $\omega = 0$, the *implicit* EULER *method* for $\omega = 1$, and the *trapezoidal rule* for $\omega = 1/2$. The method has order $p = 2$ for $\omega = 1/2$ and order $p = 1$ else.

A general one-step method can be written in the form

$$y(t+\tau) = y(t) + \tau\, \Phi(t, y(t), \tau)\,,\ t = j\tau\,,\ j = 0, 1, \dots\,,\ y(0) = x_0\,, \tag{2.51}$$

or, if the step length τ is constant throughout iteration, as

$$y_{j+1} = y_j + \tau\, \Phi_j(y_j, \tau)\,,\ j = 0, 1, \dots\,,$$

where $y_j := y(j\tau)$. The device function Φ must satisfy some obvious conditions which are not enumerated here and are fulfilled in normal case; cf. (Hairer). The method is then said to be *explicit* if a *finite* number of evaluations of the right side f (and derivatives of f) suffices for an *exact* computation of Φ, otherwise the method is called *implicit*.

The discretization error is defined in the same way as in **(b)**, and the method is *consistent* (with the differential equation) if

$$\Gamma(x) := \sup_{0 \le \tau \le \tau^*} \sup_{0 \le t \le T-\tau} \frac{1}{\tau^p} \|d(t, x, \tau)\| < \infty \tag{2.52}$$

for some $p \ge 1$ and for *all* solutions $x \in \mathcal{C}^{p+1}[0, T]$. The maximum possible number $p \in \mathbb{N}$ in (2.52) is the order of the method for the considered differential equation and called *order* in general if the method has order p for all sufficiently smooth right sides f of (2.48). The content of Lemma 2.4 remains unchanged by this convention.

(c) Asymptotic Expansion, Extrapolation

Lemma 2.5. *Let the method (2.51) have order $p \ge 1$, let*

$$\Phi(t, x(t), 0) = f(t, x(t))\,,\ \ \operatorname{grad}_x \Phi(t, x(t), \tau) = \operatorname{grad}_x f(t, x(t)) + \mathcal{O}(\tau)\,,$$

and let $\partial\Phi/\partial\tau$ be continuous in τ near $\tau = 0$. Then there exists an error function r *being* independent *of the step length τ such that*

$$\boxed{y(t) = x(t) + r(t)\tau^p + \mathcal{O}(\tau^{p+1})\,,\ \tau \to 0}\,.$$

Proof see (Hairer), vol. I, Sect. 2.8.

This asymptotic representation of the numerical approximation $y(t)$ has two important consequences whenever we apply the method (2.51) once with step length τ and then once more with the reduced step length $q\tau$, $0 < q < 1$,

$$\begin{aligned} y(t, \tau) &= \ \ x(t) + r(t)\tau^p \ \ + \mathcal{O}(\tau^{p+1})\,, \\ y(t, q\tau) &= x(t) + r(t)(q\tau)^p + \mathcal{O}(\tau^{p+1})\,. \end{aligned}$$

(1°) The *weighted* difference

$$z(t) := \frac{q^{-p}y(t,q\tau) - y(t,\tau)}{q^{-p}-1} = x(t) + \mathcal{O}(\tau^{p+1}) \tag{2.53}$$

supplies an improved method of order $p+1$ instead of p with comparable few computational effort.

Example 2.8. The *model problem* $x' = \lambda\, x$ has the solution $x(t) = \kappa\, e^{\lambda\, t}$. We choose $x(0) = 1\,,\ \lambda = 1\,,$ and apply the trapezoidal rule, once with step length $\tau = 1$ and, for comparison, twice with step length $\tau = 1/2$:

$$\begin{aligned} h = 1: \quad y(1) &= \frac{1+0.5}{1-0.5} &&= 3 \\ h = 0.5: \widetilde{y}(1) &= \frac{1+0.25}{1-0.25}\cdot\frac{1+0.25}{1-0.25} = \frac{25}{9} &&= 2.\overline{7}\,. \end{aligned}$$

An application of the *averaging* (2.53) with $p = 2$ and $q = 1/2$ yields the improvement

$$z(1) = \frac{1}{3}\left(4\cdot\frac{25}{9} - 3\right) = \frac{100-27}{27} = 2.703703\ldots$$

with error $\varepsilon = 0.0145\ldots$; the additional amount of work is a neglecting quantity in larger problems.

(2°) The *simple* difference

$$\begin{aligned} y(t,\tau) - y(t,q\tau) &= r(t)(q\tau)^p(q^{-p}-1) + \mathcal{O}(\tau^{p+1})\,, \\ r(t)(q\tau)^p &\simeq \frac{y(t,\tau) - y(t,q\tau)}{q^{-p}-1}\,, \end{aligned}$$

supplies a good estimation of the error function $r(t)$. If, further, D is a diagonal matrix containing suitable weights and C contains tolerances and suitable security factors then, with additional safety bounds,

$$\tau_{\text{new}} = \tau_{\text{old}}\cdot C\cdot \|D[y(q\tau) - y(t)]\|^{-1/p}$$

provides an excellent step length control. Particular advantages are obtained in the case of *imbedded* methods of order p supplying an approximation $z(t)$ of order $p-1$ at the same time,

$$\begin{aligned} y(t) &= x(t) + r(t)\tau^p + \mathcal{O}(\tau^{p+1})\,, \\ z(t) &= x(t) + \widetilde{r}(t)\tau^{p-1} + \mathcal{O}(\tau^p)\,. \end{aligned}$$

The difference yields here *directly* an estimation of the error of $z(t)$:

$$z(t) - y(t) = \widetilde{r}(t)\tau^{p-1} + \mathcal{O}(\tau^p) \simeq z(t) - x(t)\,.$$

(d) Runge-Kutta Methods

Example 2.9. The explicit method of HEUN is obtained from the (implicit) trapezoidal rule (2.50) by inserting

$$f_{j+1}(y_{j+1}) \simeq f_{j+1}\left(y_j + \tau f_j(y_j)\right) .$$

$$\boxed{\begin{array}{l} \text{Trapezoidal rule (p = 2)}: \; y_{j+1} = y_j + \dfrac{\tau}{2}\Big(f_j(y_j) + f_{j+1}(y_{j+1})\Big) \\ \text{Method of HEUN (p = 2)}: y_{j+1} = y_j + \dfrac{\tau}{2}\Big(f_j(y_j) + f_j(y_j + \tau f_j(y_j))\Big) \end{array}} ;$$

the numerical computation is carried out by the scheme

$$k_1 = f_j(y_j)\,, \;\; k_2 = f_j(y_j + \tau\, k_1)\,, \;\; y_{j+1} = y_j + \frac{\tau}{2}\,(k_1 + k_2) \;.$$

A generalization of this concept leads to *multistage* methods or RUNGE-KUTTA methods.

Example 2.10. The classical RUNGE-KUTTA method is a four-stage method of order $p = 4$ in which the function f is four-times evaluated at intermediate steps. Ensuing, a linear combination of these terms forms the forward step; this last step originates mostly from a numerical integration rule, in the present case being SIMPSON's rule:

$$\begin{aligned} k_1 = f(y_j)\,, \;\; k_2 = f_j\left(y_j + \frac{\tau}{2}\,k_1\right)\,, \;\; k_3 = f_j\left(y_j + \frac{\tau}{2}\,k_2\right)\,, \\ k_4 = f_j(y_j + \tau\, k_3)\,, \;\; y_{j+1} = y_j + \tau\,\frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4) \;. \end{aligned}$$

The example shows that the order of a one-step method can be enhanced in a skilful way if several intermediate steps are properly introduced. Even methods of arbitrary high order can be constructed by this way (methods of GRAGG-BULIRSCH-STOER).

A general r-stage one-step method for $x' = f(t,x) \in \mathbb{R}^n$ is a computational device of the form

$$\boxed{\begin{array}{l} k_i(t) = f\left(t + \gamma_i\tau\,,\; y(t) + \tau \displaystyle\sum_{j=1}^{r} \alpha_{ij} k_j(t)\right)\,,\; i = 1:r \\ y(t+\tau) = y(t) + \tau \displaystyle\sum_{i=1}^{r} \beta_i k_i(t) \end{array}} \;. \qquad (2.54)$$

with the *function values* $k_i(t) := f(t+\gamma_i\tau, u_i(t))$ being the unknown quantities. However the subsequent representation is to be preferred in the studies of this method. Let I be the unit matrix, $A = [\alpha^i{}_j] \in \mathbb{R}^r{}_r$, and $b = [\beta_i]$, $c = [\gamma_j]$, $e = [1]$ all together in $\mathbb{R}^r$, and, moreover,

$$\begin{aligned} A \times B &= [\alpha_{ij} B]_{i,j=1}^{r} && \text{KRONECKER product}, \\ U(t) &= [u_i(t)]_{i=1}^{r} && \text{auxiliary vectors}, \ u_i(t) \in \mathbb{R}^n, \\ F(t, U(t)) &= [f(t + \gamma_i \tau, u_i(t))]_{i=1}^{r} \in \mathbb{R}^{r \cdot n}. \end{aligned}$$

The computational device (2.54) is then equivalent to the form with *intermediate values* u_i of the approximation

$$\begin{aligned} U(t) &= e \times y(t) + \tau (A \times I) F(t, U(t)) \in \mathbb{R}^{r \cdot n} \\ y(t + \tau) &= y(t) + \tau (b \times I)^T F(t, U(t)) \quad \in \mathbb{R}^n . \end{aligned} \tag{2.55}$$

For instance, the method of HEUN may now be written as

$$u_1 = y_j , \ \ u_2 = y_j + \tau f_j(u_1) , \ \ y_{j+1} = y_j + \tau \left(f_j(u_1) + f_j(u_2) \right) .$$

The auxiliary quantities $u_i(t)$ may be interpreted as approximations of the exact values $x(t + \gamma_i \tau)$ which however is only of interest in derivation of order conditions.

Properties and Further Notations:

(1°) The method (2.55) is called *explicit* resp. *semi-implicit* if (possibly after renumeration) the matrix A is a strongly lower resp. a lower triangular matrix.

(2°) In normal case, the points $\gamma^i \tau$ are contained in interval $[0, \tau]$, but they are not always mutually distinct; cf. Example 2.10. The intermediate values $u_i(t)$ are approximations of the exact solution at the intermediate points $t + \gamma^i \tau$ as already mentioned. The system of intermediate stages is uniquely solvable if f is LIPSCHITZ bounded and if, but only in implicit methods, the step length τ is sufficiently small.

(3°) The method (2.55) is frequently described by using the BUTCHER matrix $[A|b|c]$ or a similar form. For instance, Example 2.10 can be displayed as

$$\left[\begin{array}{c|c} A & c \\ \hline b & \end{array}\right] = \left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 1 \\ \hline 1/6 & 1/3 & 1/3 & 1/6 & \end{array}\right] .$$

(4°) If $W(t)$ is the solution of

$$W(t) = e \times x(t) + \tau (A \times I) F(t, W(t)) ,$$

then

$$d(t, x, \tau) = \frac{x(t + \tau) - x(t)}{\tau} - (b \times I)^T F(t, W(t))$$

is the *discretization error*.

(5°) The method (2.54) has at least order $p = 1$ if and only if $b^T e = \sum_{i=1}^{r} \beta_i = 1$.

(6°) Table of attainable order p^* of explicit RUNGE-KUTTA methods in dependence of the stage number r (Butcher):

Table 2.4.

r	1	2	3	4	5	6	7	8	9	$r \geq 10$
p^*	1	2	3	**4**	**4**	5	6	6	7	$\leq r-2$

therefore the RUNGE-KUTTA method of order $p = 4$ plays a particular role among all explicit methods.

(e) Multistep Methods A multiple evaluation of the right side f of the underlying differential system can be rather cumbersome. But the order of a method can be enhanced also if the formerly obtained values y_j , y_{j-1} , ... are regarded in the sense of an extrapolation beyound the interval known at the present state. For instance, the device

$$3y_{j+1} - 4y_j + y_{j-1} = 2\tau\, f_j(y_{j+1})\,,\ j = 1, 2, \dots\,,$$

is a well-known implicit method of order $p = 2$ with extraordinary stability properties.

General multistep methods have the form

$$\sum_{i=0}^{k} \alpha_i y_{j+i} = \tau \sum_{i=0}^{k} \beta_i f_{j+i}(y_{j+i})\,,\ j = 0, 1, \dots\,, \tag{2.56}$$

where $\alpha_k \neq 0$ and $|\alpha_0| + |\beta_0| \neq 0$; the method is *explicit* for $\beta_k = 0$ and *implicit* else. The function f has to be evaluated here only once in every t-step. But, on the other side, the starting values $y_1, \dots, y_{k-1}$ must be supplied by some other method.

Properties and Notations:

(1°) Using the polynomials

$$\varrho(\zeta) = \sum_{i=0}^{k} \alpha_i \zeta^i\,,\quad \sigma(\zeta) = \sum_{i=0}^{k} \beta_i \zeta^i$$

and the translation operator $E : y(t) \mapsto Ey(t) := y(t+\tau)$, the device (2.56) can be written more simply as

$$\boxed{\varrho(E)y_j = \tau\sigma(E)f_j\,,\ j = 0, 1, \dots} \,. \tag{2.57}$$

(2°) Application to the model equation $x' = \lambda x$ yields the very simple device

$$\pi(E, \eta)y_j := \varrho(E)y_j - \eta\sigma(E)y_j = 0\,,\ j = 0, 1, \dots$$

where $\eta = \tau\lambda$ with step length τ, and $\pi(E, \eta)$ is the *characteristic polynomial* of the multistep methods describing it completely as well. Furthermore, the discretization error d of the method has the very simple form

$$\tau d(t, x, \tau) = \varrho(E)x(t) - \tau\sigma(E)x'(t)\,.$$

(3°) Order and consistence are defined in the same way as in **(b)**. Whereas the derivation of order conditions and the construction of individual RUNGE-KUTTA-methods and their relatives is a difficult matter and must be leaved to experts, cf. (Hairer), it is much simpler in multistep methods. By reasons of linearity the coefficients are to be adjusted here in a way that, for a prescribed order p, all "differential equations" $x'(t) = t^k$, $k = 0 : p$, are solved exactly. Further possibilities for constructing very special methods arise by using other, especially chosen elementary functions as e^{it}, $\sin(jt)$, $\cos(kt)$ for the adjustment above.

Lemma 2.6. *A multistep method* (ϱ, σ) *has order* $p \geq 1$ *if and only if*

$$\varrho(1) = 0 \;\; \textit{and} \;\; \sum_{i=0}^{k} \left(\alpha_i \frac{i^m}{m} - \beta_i\, i^{m-1} \right) = 0\,,\; m = 1 : p\,.$$

In particular, the method has at least order $p = 1$ if $\varrho(1) = 0$ and $\varrho'(1) - \sigma(1) = 0$.

(4°) The discretization error satisfies in simple way

$$\|\tau d(t, x, \tau)\| \leq \text{const}\, \tau^p \int_t^{t+k\tau} \|x^{(p+1)}(s)\|\, ds\,,$$

and, by induction, an error estimation is deduced in a similar way as in Lemma 2.4 with the same qualitative properties. But, in the present case, the polynomial ϱ must satisfy the following *root condition* or *stability criterium*:

> Every root λ of ϱ satisfies $|\lambda| \leq 1$ and every root λ with $|\lambda| = 1$ is a simple root.

(5°) In order to prevent "spurious solutions", a multistep method should also be *strongly stable* which means that the polynomial ϱ has precisely one root of modulus one, namely the always appearing root $\zeta = 1$, and this root must be a simple root according to the root condition.

(f) Summary

multistage methods	multistep methods
self-starting	not self-starting
high computational effort	low computational effort
simple step length control	difficult step length control

.

(g) Stability

(g1) A differential equation $x'(t) = f(t, x(t))$ is *stable* if the difference of any two solutions remains bounded in absolute value for all $t > 0$, it is *asymptotically stable* if the difference tends to zero in absolute value for $t \to \infty$; cf. Sect. 1.5**(c)**; analogeous notations hold for differential systems w.r.t. an arbitrary submultiplicative norm. The instability of a differential equation is inherited to the numerical approximation in any case, nevertheless, a good step control may supply very acceptable results. If a differential equation is stable then the exact solutions decrease in absolute value during a long t-interval or they remain bounded at least. Of course, this property should be inherited to the numerical approximations, too, but this is not always the case:

Example 2.11. $x' = Ax,\ x_0 = [1, 0, -1]^T,\ x(t) = [x(t), y(t), z(t)]^T.$

$$A = \begin{bmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{bmatrix},\ \text{eigenvalues}: \lambda_1 = -2,\ \lambda_{2,3} = -40 \pm 40\,i.$$

The solution

$$x(t) = \frac{1}{2}e^{-2t} + \frac{1}{2}e^{-40t}(\cos 40t + \sin 40t),$$
$$y(t) = \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-40t}(\cos 40t + \sin 40t),$$
$$z(t) = -e^{-40t}(\cos 40t - \sin 40t).$$

behaves like the solution of $x' = Bx$ for $t > 0.1$ where

$$B = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

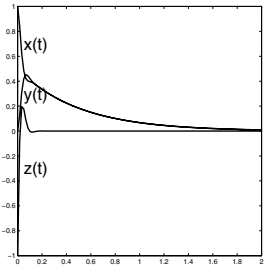

Figure 2.11. Ex. 2.11, solution

The trapezoidal rule provides good approximations whereas the explicit EULER method supplies entirely unacceptable results with the applied step length. In the explicit method, a small step length τ (with high computational effort) is unnecessary for large t whereas, on the other side, a large step length magnifies the high-frequent but fast decreasing parts of the solution in an explosive way.

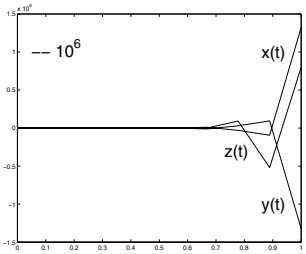

Figure 2.12. Ex. 2.11, EULER explicit, $\tau = 0.1$

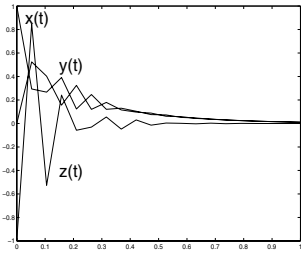

Figure 2.13. Ex. 2.11, Trapezoidal rule, $\tau = 0.1$

(g2) For a more thorough investigation of this phenomenon, we apply the multistage method (2.55) to the *model equation* $x'(t) = \lambda x(t)$ again and obtain with step length τ and $\eta = \tau\lambda$

$$\begin{array}{r} U(t) = ey(t) + \tau\lambda AU(t) \\ y(t+\tau) = y(t) + \tau\lambda b^T U(t) \end{array} \implies \begin{array}{c} U(t) = (I - \eta A)^{-1} ey(t) \\ y(t+\tau) = \left[1 + \eta b^T (I - \eta A)^{-1} e\right] y(t)\,. \end{array}$$

Thereby the computational device

$$y_{j+1} = R(\eta) y_j\,, \quad R(\eta) = 1 + \eta b^T (I - \eta A)^{-1} e\,, \quad R(\infty) := 1 - b^T A^{-1} e \tag{2.58}$$

is derived for the model equation. By using CRAMER's rule, the *stability function* R can be written after brief computation as

$$R(\eta) = \frac{\det\left(I - \eta A + \eta e b^T\right)}{\det(I - \eta A)} =: \frac{P(\eta)}{Q(\eta)}\,, \tag{2.59}$$

with polynomials P and Q, and $Q(\eta) = 1$ in explicit methods. Now the closed set in complex η-plane

$$\boxed{\mathcal{S} := \{\eta \in \mathbb{C} \cup \{\infty\},\ |R(\eta)| \le 1\}}$$

is called *stability region* of the specific one-step method. If we consider more generally the system $x'(t) = Ax(t)$ with *diagonalizable* Matrix A, $A = U\Lambda U^{-1}$, (Λ diagonal matrix of eigenvalues λ_i of A), then the one-step methods obtain the form

$$y_{j+1} = UR(\tau\Lambda)U^{-1} y_j = UR(\tau\Lambda)^j U^{-1} y_0$$

with diagonal matrix

$$R(\tau\Lambda) = \mathrm{diag}(R(\tau\lambda_1), \ldots, R(\tau\lambda_n))\,.$$

As a consequence, all $\eta_i := \tau\lambda_i$ must be contained in the stability region $\mathcal{S}$ if every solution shall be at least bounded for all $t > 0$. This is the well-known COURANT-FRIEDRICHS-LEVY condition for the step length τ, and this condition must always be regarded in stable problems but also in various other cases; see e.g. Sect. 9.7 **(d)**.

Example 2.12. Consider the model equation $x' = \lambda\, x$ with $\eta = \tau\,\lambda$. Then

(A) EULER method explicit $(p = 1)$:	$y_{j+1} = (1+\eta)y_j\,,$
(B) EULER method implizit $(p = 1)$:	$y_{j+1} = (1-\eta)^{-1}y_j\,,$
(C) Trapezoidal rule $(p = 2)$:	$y_{j+1} = \dfrac{2+\eta}{2-\eta}y_j\,,$
(D) Method of HEUN $(p = 2)$:	$y_{j+1} = \left(1+\eta+\dfrac{1}{2}\eta^2\right)y_j\,.$

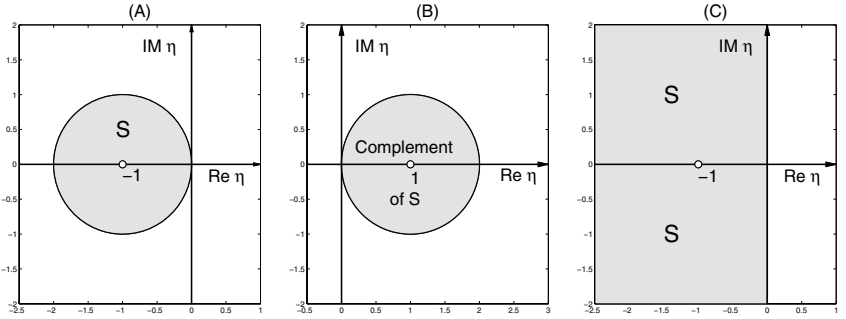

Figure 2.14. Stability regions for example 2.12 without (D)

The discretization error d of a method of type (2.55) of order $p \geq 1$ satisfies $\tau d(t,x,\tau) = \mathcal{O}(\tau^{p+1})$ and, on the other side, a substitution of the model equation with $\lambda = 1$ und $x(0) = 1$ yields the device

$$y(\tau) = R(\tau)y(0) = R(\tau)\,,\quad x(\tau) = e^{\tau}\,.$$

Subtraction yields

$$\tau d(\tau,x,\tau) = x(\tau) - y(\tau) = e^{\tau} - R(\tau) = \mathcal{O}(\tau^{p+1})\,.$$

Accordingly, we have

$$R(\eta) = 1 + \eta + \frac{\eta^2}{2} + \ldots + \frac{\eta^p}{p!} + \mathcal{O}(\eta^{p+1})\,,$$

in every RUNGE-KUTTA method of order p. On the other side $R(\eta) = 1 + \eta + \sum_{i=2}^{r}\kappa_i\eta^i$ in an explicit r-stage method by (2.59). As a consequence, all *explicit* RUNGE-KUTTA methods with same order $\boxed{p = r}$ have the same stability function $R(\eta) = \displaystyle\sum_{i=0}^{p}\eta^i/i!$ (where however $p^* = r \leq 4$ for the *optimum*

order p^* by Table 2.4). By symmetry to the real axis, only the upper half of the implicit curve $|\sum_{i=0}^{p} \eta^i/i!| = 1\,,\ p = 1:6\,,$ is plotted in Figure 2.15.

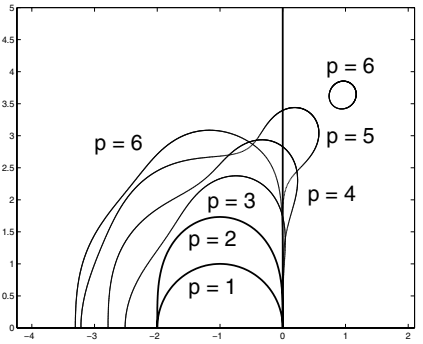

Figure 2.15. Stability regions of explicit RKM with $p = r = 1:4$

(g3) Let us apply a multistep method (2.56) to the model equation $x' = \lambda\, x$ then the result is

$$\pi(E,\eta)y_j = \sum_{i=0}^{k} \gamma_i(\eta)E^i y_j = \sum_{i=0}^{k} \gamma_i(\eta) y_{j+i} = 0\,,\ j = 0,1,\dots\,,$$

by (2.57) or, writing the device as one-step method by introducing the vector $Y_j = [y_j, y_{j+1}, \dots, y_{j+k-1}]^T \in \mathbb{R}^k\,,$

$$Y_{j+1} = F_\pi(\eta)Y_j\,,\quad F_\pi(\eta) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ 0 & & \ddots & \ddots & 0 \\ 0 & & \ddots & 0 & 1 \\ -\gamma_0(\eta)0/\gamma_k(\eta) & \dots & \dots & \dots & -\gamma_{k-1}(\eta)/\gamma_k(\eta) \end{bmatrix}. \tag{2.60}$$

The FROBENIUS matrix $F_\pi(\eta)$ has the characteristic polynomial $\det(\lambda I - F_\pi(\eta)) = \pi(\lambda,\eta)$, hence it is called sometimes *accompanying matrix to the polynomial* π. Every eigenvalue of this matrix possesses precisely one eigenvector. On the other hand the matrix must be a M-matrix by Theorem 1.1; cf. Sect. 1.1 **(c4)**, if all iterations (2.60) shall remain bounded in absolute value (resp. in some norm). Therefore the concept of *stability regions* must be adapted to multistep methods as follows:

Definition 2.2. *Let* $(\varrho\,,\,\sigma)$ *be a multistep method with characteristic polynomial* $\pi(\zeta,\eta) = \varrho(\zeta) - \eta\sigma(\zeta)$*, and let* $\pi(\zeta,\infty) = \sigma(\zeta)$*. Then the stability region* $\mathcal{S} \in \mathbb{C} \cup \{\infty\}$ *is the set of all values* $\eta \in \mathbb{C}$ *with the following two properties:*

(1°) *All roots* $\zeta_i(\eta)$ *of* $\pi(\zeta,\eta)$ *satisfy* $|\zeta_i(\eta)| \le 1$.
(2°) *All roots* $\zeta_i(\eta)$ *of* $\pi(\zeta,\eta)$ *satisfying* $|\zeta_i(\eta)| = 1$ *(unimodular roots) are simple roots of* $\pi(\zeta,\eta)$.

(h) Stiff Differential Systems A system $x'(t) = Ax(t)$ is said to be *stiff* if the eigenvalues λ_i of A have the property

$$\operatorname{Re}\lambda_i \le 0 \text{ and } \max_i |\operatorname{Re}\lambda_i| \gg \min_i |\operatorname{Re}\lambda_i| .$$

Such systems occur e.g. in the motion of mass points being connected with each other by weak and stiff springs at the same time. Furthermore, they appear necessarily in discretization of differential equations as the following simple example shows impressively.

Example 2.13. The eigenvalue problem

$$y''(x) = \lambda^2\, y\,,\ y(0) = y(1) = 0\,, \tag{2.61}$$

has the characteristic pairs

$$(\lambda_j^2\,,\, y_j(x)) = (-j^2\pi^2\,,\, \sin(j\pi x))\,,\ j \in \mathbb{N}\,.$$

If the second derivative is approximated by the central divided difference

$$y''(jh) = h^{-2}\,[y((j+1)h) - 2y(jh) + y(t,(j-1)h)] + \mathcal{O}(h^2)\,,\quad h = 1/(n+1)\,,$$

then we obtain the discrete eigenvalue problem $AY = \widetilde{\lambda}^2 Y \in \mathbb{R}^n$,

$$AY = h^{-2}\begin{bmatrix} -2 & 1 & 0 & \dots & 0 & 0 \\ 1 & -2 & 1 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 1 & -2 \\ 0 & 0 & \dots & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} y(h) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y(nh) \end{bmatrix} = \widetilde{\lambda}^2 \begin{bmatrix} y(h) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y(nh) \end{bmatrix}, \tag{2.62}$$

with characteristic pairs

$$(\widetilde{\lambda}_j^2, Y_j) = \left(-h^{-2}4\sin^2\left(\frac{jh\pi}{2}\right), \left[\sin\left(\frac{jk\pi}{n+1}\right)\right]_{k=1}^{n}\right),\ j = 1:n\,.$$

Note here the very exceptional fact that the eigenvectors of the discretized problem have the same values as the eigenfunctions of the analytic problem (2.61) at corresponding points. The eigenvalues satisfy

$$\widetilde{\lambda}_j^2 = -h^{-2}4\sin^2\left(\frac{jh\pi}{2}\right) = -j^2\pi^2 + \mathcal{O}(j^4h^2) = \lambda_j^2 + \mathcal{O}(j^4h^2)\,,\ j = 1:n\,.$$

Accordingly, $\widetilde{\lambda}_j^2$ is a second-order approximation of the eigenvalue $-j^2\pi^2$ of the differential equation for every *fixed* j and, in particular, the eigenvalues of (2.62) increase beyond every bound in absolute value if the step length h tends to zero.

Let us now consider the *parabolic* initial boundary value problem

$$\begin{aligned} &u_t(t,x) = u_{xx}(t,x)\,,\ 0 \le x \le 1\,,\ 0 \le t\,,\\ &u(t,0) \ = a(t)\,,\ u(t,1) = b(t)\,, u(0,x) = u_0(x)\,,\ u_0(0) = a(0)\,,\ u_0(1) = b(0)\,, \end{aligned} \tag{2.63}$$

which can be solved exactly at least for $a = 0$ and $b = 0$. A discretization in the space variable x in the same way as in (2.61) leads to the initial value problem

$$\begin{aligned} &U'(t) = AU(t) + B(t)\,,\\ &U(t) \ = [u(t,h),\ldots,u(t,nh)]^T\,,\ B(t) = h^{-2}[a(t),\,0,\ldots,0,\,b(t)]^T\,, \end{aligned}$$

with the same matrix A as in (2.62). If this ordinary differential system shall be solved by one of the above considered methods then, by the COURANT-FRIEDRICHS-LEVY condition, the step length $\tau > 0$ must be chosen so small that the value $\widetilde{\lambda}_n^2 \simeq -4\tau/h^2$ still lies in the stability region $\mathcal{S}$. But this restriction of step length can be dropped in the special *implicit* methods (B) and (C) because here the entire negative semi-line belongs to $\mathcal{S}$. By this reason, further criteria on the *shape* of the stability region are introduced: A multistage/multistep method is called

$$\begin{array}{lll} A\text{-}stable & \Longleftrightarrow & \{\eta \in \mathbb{C} \cup \{\infty\},\ \mathrm{Re}\ \eta < 0\} \subset \mathrm{int}\,\mathcal{S} \\ A(\alpha)\text{-}stable & \Longleftrightarrow & \{\eta \in \mathbb{C} \cup \{\infty\},\ \eta \neq 0\,,\ |\pi - \arg\,\eta| < \alpha\} =: \mathcal{S}_\alpha \subset \mathrm{int}\,\mathcal{S} \\ A(0)\text{-}stable & \Longleftrightarrow & \exists\ \alpha > 0 \text{ such that the method is } A(\alpha)\text{-stable} \\ A_0\text{-}stable & \Longleftrightarrow & (-\infty\,,\,0) \subset \mathcal{S} \\ L\text{-}stable & \Longleftrightarrow & \text{method } A\text{-stable and } \infty \subset \mathrm{int}\,\mathcal{S} \end{array} \tag{2.64}$$

The stability function (2.59) shows that the stability region of an *explicit* RUNGE-KUTTA method can never have one of these properties; the same can be verified easily for *explicit* multistep methods. One now recognizes the dilemma in the model Example 2.13: In explicit methods, the step length τ in t-direction must be chosen proportionally to the quadrat of the step length h in x-direction whereas in implicit methods the computational amount of work is considerably magnified.

By the way, not all implicit methods have one of the properties (2.64) which is shown best by plotting the individual stability regions.

Rule for application:

> Use only explicit methods with step control for solving *unstable* differential systems.

(i) Further Examples We consider briefly some methods of the MATLAB ODE suite.

(1°) MATLAB `ode45.m` Imbedded explicit RUNGE-KUTTA method due to DORMAND & PRINCE; cf. (Dormand):

$$\left[\begin{array}{c} A \\ b \end{array}\right] =$$

$$\left[\begin{array}{ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/5 & 0 & 0 & 0 & 0 & 0 & 0 \\
3/40 & 9/40 & 0 & 0 & 0 & 0 & 0 \\
44/45 & -56/15 & 32/9 & 0 & 0 & 0 & 0 \\
19372/6561 & -25360/2187 & 64448/6561 & -212/729 & 0 & 0 & 0 \\
9017/3168 & -355/33 & 46732/5247 & 49/176 & -5103/18656 & 0 & 0 \\
35/384 & 0 & 500/1113 & 125/129 & -2187/6784 & 11/84 & 0 \\
\hline
35/384 & 0 & 500/1113 & 125/129 & -2187/6784 & 11/84 & 0 \\
\hline
5179/57600 & 0 & 7571/16695 & 393/640 & -92097/339200 & 187/2100 & 1/40
\end{array}\right]$$

$$c = [0,\ 1/5,\ 3/10,\ 4/5,\ 8/9,\ 1,\ 1]$$

If one takes the pen-ultimate row for weight vector b in the forward step, then the result is a method of order $p = 5$ and with the last row instead a method of order $p = 4$. The six-stage method has order $p = 5$, the seventh stage being used only for error estimation in step length control. The stability region is given by the curve for $p = 6$ in Figure 2.15.

(2°) ROSENBROCK methods are perhaps not the *ultima ratio* but the result of a long investigation on the *efficiency* of methods for stiff systems. The contradicting requirements on high order, low computational effort and best stability properties as L-stability, cf. (2.64), have finally led to a compromise. Starting point of the deliberations is a RUNGE-KUTTA method of which the matrix A is a (weakly) lower triangular matrix (diagonal implicit methods). Confining ourselves to an autonomeous system $x'(t) = f(x(t))$, the method has the form

$$k_i(t) = f\left(y(t) + \tau \sum_{j=1}^{i-1} a_{ij} k_j(t) + \tau a_{ii} k_i(t)\right)\ ,\ i = 1 : r\,,$$
$$y(t+\tau) = y(t) + \tau \sum_{i=1}^{r} \beta_i k_i(t)\,.$$

These equations are now linearized, for instance the values $k_i(t)$ are replaced by

$$k_i(t) = f(g_i(t)) + \operatorname{grad} f(g_i(t)) a_{ii} k_i(t)\,,\quad g_i(t) = y(t) + \tau \sum_{j=1}^{i-1} a_{ij} k_j(t)\,.$$

Furthermore, the r matrices $\operatorname{grad} f(g_i(t))$ are replaced by a single matrix $J := \operatorname{grad} f(y(t))$ following a proposition of (Calahan) which once more reduces the computational effort considerably. Then the combination

$$k_i(t) = f\left(y(t) + \sum_{j=1}^{i-1} a_{ij}k_j(t)\right) + J\sum_{j=1}^{i} d_{ij}k_j(t)$$
$$y(t+\tau) = y(t) + \tau\sum_{i=1}^{r}\beta_i k_i(t)$$

is chosen in ROSENBROCK methods to attain a higher degree of freedom in the choice of suitable coefficients. The MATLAB program `ode23s.m` of (Shampine82) is of this type where the last evaluation of f in the preceding step is used for first evaluation in the new step:

$$\begin{aligned}
f_0 &= f(t, y(t)) \\
Wk_1 &= f_0 + \tau dT \\
f_1 &= f\left(t + \frac{1}{2}\tau, y(t) + \frac{1}{2}\tau k_1\right) \\
Wk_2 &= f_1 - k_1 + Wk_1 \\
y(t+\tau) &= y(t) + \tau k_2 \\
f_2 &= f(t+\tau, y(t+\tau)) \\
Wk_3 &= f_2 - e\,(k_2 - f_1) - 2\,(k_1 - f_0) \\
\widetilde{y}(t+\tau) &= y(t+\tau) + \frac{\tau}{6}\,(k_1 - 2k_2 + k_3)
\end{aligned}$$

$$d = 1/(2+\sqrt{2})\,, \quad e = 6+\sqrt{2}\,,$$

$$T = \frac{\partial}{\partial t} f(t, y(t))\,, \quad J = \operatorname{grad} f(t, y(t))\,, \quad W = I - hdJ\,.$$

This two-stage method has order $p = 2$, the value $\widetilde{y}(t+\tau)$ is only used for error estimation. Substitution of the model equation $x'(t) = \lambda x(t)$ with $\eta = \tau\lambda$ again yields the device

$$y_{n+1} = R(\eta)y_n\,, \quad R(\eta) = \frac{1 + (1-2d)\eta + (d^2 - 2d + 1/2)\eta^2}{1 - 2d\eta + d^2\eta^2}\,.$$

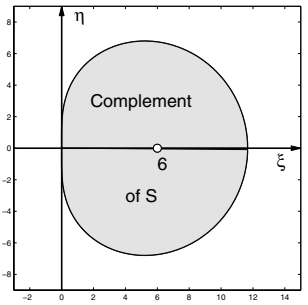

Figure 2.16. Stability region of the Rosenbrock method

(3°) MATLAB `ode113.m` predictor-corrector method
Let the *backward differences* be defined by

$$\nabla^0 f(t) = f(t)\,,\ \nabla f(t) = f(t) - f(t-\tau)\,,$$
$$\nabla^2 f(t) = \nabla(\nabla f(t)) = f(t) - 2f(t-\tau) + f(t-2\tau)\,,\ \text{etc.},$$

and let $p_k(x; f)$ be an interpolating polynomial of degree k with interpolation property

$$p_k((j+i)\tau) = f((j+i)\tau)\,,\ i = 0:k\,,$$

where $j \in \mathbb{N}_0$ is fixed for the present. Then the polynomial may be written in NEWTON-GREGORY form:

$$p_k((j+k+s)\tau; f) = \sum_{i=0}^{k} \binom{s+i-1}{i} \nabla^i f((j+k)\tau)\,,\quad \binom{s-1}{0} = 1\,. \quad (2.65)$$

(3.1°) By integration of (2.65) over the s-interval $(-1, 0)$,

$$x_{j+k} - x_{j+k-1} = \tau \int_{-1}^{0} f((j+k+s)\tau, x(t))\, ds \simeq \tau \int_{-1}^{0} p_k((j+k+s)\tau; f)\, ds\,,$$

the implicit ADAMS methods are generated with k steps and the order $k+1$:

$$y_{j+k} - y_{j+k-1} = \tau \sum_{i=0}^{k} \gamma_i \nabla^i f_{j+k}(y_{j+k})\,,\ j = 0, 1, \dots\,,$$

$$\gamma_i = \int_{-1}^{0} \binom{\xi+i-1}{i}\, d\xi\,.$$

(3.2°) By integration over the s-Intervall $(0, 1)$,

$$x_{j+k+1} - x_{j+k} = \tau \int_{0}^{1} f((j+k+s)\tau, x(t))\, ds \simeq \tau \int_{0}^{1} p_{k-1}((j+k+s)\tau; f)\, ds\,,$$

the explicit ADAMS methods are generated with k steps and order k:

$$y_{j+k} - y_{j+k-1} = \tau \sum_{i=0}^{k-1} \gamma_i^* \nabla^i f_{j+k-1}(y_{j+k-1}) \,, \; j = 0, 1, \ldots \,,$$

$$\gamma_i^* = \int_0^1 \binom{\xi + i - 1}{i} d\xi \,.$$

The coefficients γ_i and γ_i^* do not depend on the step number k and can be computed in advance by recurrence.

(3.3°) Both methods together supply a *predictor-corrector method* for non-stiff differential equations:

In the predictor step, the explicit method (3.2°) of order k is applied once.

In the corrector step, the implicit method (3.1°) of order $k+1$ is repeatedly applied or only once since a single application suffices for order k in the combined method (Fig. 2.17).

Table 2.5. Implicit Adams methods

k	β_0	β_1	β_2	β_3	β_4	β_5
1	$\frac{1}{2}$	$\frac{1}{2}$				
2	$-\frac{1}{12}$	$\frac{8}{12}$	$\frac{5}{12}$			
3	$\frac{1}{24}$	$-\frac{5}{24}$	$\frac{19}{24}$	$\frac{9}{24}$		
4	$-\frac{19}{720}$	$\frac{106}{720}$	$-\frac{264}{720}$	$\frac{646}{720}$	$\frac{251}{720}$	
5	$\frac{27}{1440}$	$-\frac{173}{1440}$	$\frac{482}{1440}$	$-\frac{798}{1440}$	$\frac{1427}{1440}$	$\frac{475}{1440}$

Table 2.6. Explicit Adams methods

k	β_0	β_1	β_2	β_3	β_4	β_5
1	1					
2	$-\frac{1}{2}$	$\frac{3}{2}$				
3	$\frac{5}{12}$	$-\frac{16}{12}$	$\frac{23}{12}$			
4	$-\frac{9}{24}$	$\frac{37}{24}$	$-\frac{59}{24}$	$\frac{55}{24}$		
5	$\frac{251}{720}$	$-\frac{1274}{720}$	$\frac{2616}{720}$	$-\frac{2774}{720}$	$\frac{1901}{720}$	
6	$-\frac{475}{1440}$	$\frac{2877}{1440}$	$-\frac{7298}{1440}$	$\frac{9982}{1440}$	$-\frac{7923}{1440}$	$\frac{4277}{1440}$

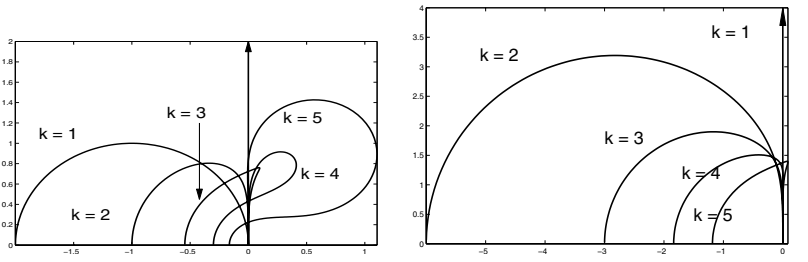

Figure 2.17. Stability regions of explicit and implicit ADAMS methods

(4°) Backward Differentiation Methods (similar methods are applied in `ode15s.m`). Because

$$\frac{d}{ds}\binom{s+i-1}{i}\bigg|_{s=0} = \begin{cases} 0 & \text{for } i = 0 \\ \dfrac{1}{i} & \text{for } i \in \mathbb{N} \end{cases}$$

we obtain from (2.65) by differentiating w.r.t. the variable s

$$f'((j+k)\tau) \simeq \frac{d}{ds}p((j+k)\tau; f) = \frac{1}{\tau}\sum_{i=1}^{k}\frac{1}{i}\nabla^i f((j+k)\tau)\,.$$

In this equation the right side is known and the left side is unknown. We write $y(t) = F(t)$ instead $f(t)$ and $f(t) = F'(t)$ instead $f'(t)$ for the inversion and obtain

$$\tau f_{j+k}(y_{j+k}) = \sum_{i=1}^{k}\frac{1}{i}\nabla^i y_{j+k}\,.$$

Thereby the implicit backward differentiation methods are generated with k steps and order k:

$$\sum_{i=0}^{k}\alpha_i y_{j+i} = \tau\beta_k f_{j+k}(y_{j+k})\,,\ j = 0, 1, \ldots\,,$$

In Figure 2.18 the upper half of the stability regions consists of the *exterior domain* of the plotted curves and of the curves themselves, therefore the "point" ∞ is advantageously contained in the interior of $\mathcal{S}$. For $k > 6$ these methods do no longer fulfill the root criterium.

Table 2.7. Backward differentiation methods

k	β_k	α_0	α_1	α_2	α_3	α_4	α_5	α_6
1	1	- 1	1					
2	2	1	-4	3				
3	6	-2	9	-18	11			
4	12	3	-16	36	-48	25		
5	60	-12	75	-200	300	-300	137	
6	60	10	-72	225	-400	450	-360	147

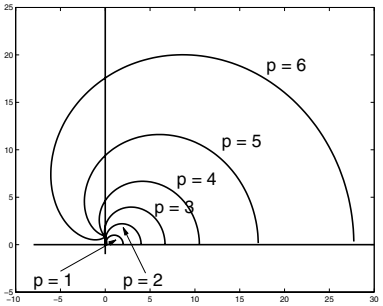

Figure 2.18. Stability regions of backward differentiation methods

(j) Full Implicit Runge-Kutta Methods have recently enjoyed new interest in connection with solving differential-algebraic equations which appear in many technical applications. The results of this subsection are already known since the pioneering work of BUTCHER, but the subsequent algebraized form of the order conditions is presumably due to (Crouzeix75) and (Crouzeix80). For detailled proofs see `SUPPLEMENT\chap02b`. Only for adaption to the notations in this subsection we make the following stipulation:

> A numerical integration rule has *order* p if it has *degree* $p-1$; cf. Sect. 2.3**(a)**, i.e., if it is exact for polynomials $p \in \Pi_{p-1}$.

Let a r-stage RUNGE-KUTTA method (RKM) of order ϱ with BUTCHER matrix $(A,\, b,\, c)$ apply to the trivial differential equation $x'(t) = f(t)$ then the exterior equation or forward step, namely

$$x(t+\tau) - x(t) = \int_t^{t+\tau} f(t)\,dt = \tau \sum_{i=1}^{r} \beta_i f(t+\gamma_i\tau) + \mathcal{O}(\tau^{\varrho+1})\,,$$

corresponds for $t = 0\,,\ \tau = 1$ to a *numerical integration rule*

$$\int_0^1 f(t)\,dt \sim \sum_{i=1}^{r} \beta_i f(\gamma_i)\,.$$

Insertion of the *monomials* $f(t) = t^{k-1}$ shows (by reasons of linearity) that it has order ϱ if and only if

$$\sum_{i=1}^{r} \beta_i \gamma_i^{k-1} = \frac{1}{k}\,,\ k = 1 : \varrho\,. \tag{2.66}$$

In the same way, the interior equations may be considered as integration rules,

$$\int_0^{\gamma_i} f(t)\,dt \sim \sum_{i=1}^{r} \alpha_{ik} f(\gamma_k)\,,\ i = 1 : r\,,$$

and they have order ϱ if and only if

$$\sum_{j=1}^{r} \alpha_{ij} \gamma_j^{k-1} = \frac{1}{k}\,\gamma_i^k\,,\ i = 1 : r\,,\ k = 1 : \varrho\,. \tag{2.67}$$

The exterior equation (integration rule) of a RKM of order ϱ has necessarily order ϱ. The *maximum common* order of the formulas (2.67) is called *interior order* of the underlying RKM. By this way, an explicit RKM has the interior order $\varrho = 1$ because the first equation has degree zero.

Following (Crouzeix75), the order conditions of implicit r-stage RKM may be described in a surprisingly simple algebraic form, but to this end we have to introduce a further condition: By partial integration we obtain the equation

$$\int_0^1 x^{k-1} \int_0^x f(s)\,ds\,dx = \frac{1}{k} \int_0^1 (1 - x^k) f(x)\,dx\,. \tag{2.68}$$

Approximating the exterior integral on *left side* by the exterior equation and the interior integrals by the interior equations of a r-stage RKM yields

$$\int_0^1 x^{k-1} \int_0^x f(s)\,ds\,dx \approx \sum_{i=1}^{r} \beta_i \gamma_i^{k-1} \int_0^{\gamma_k} f(s)\,ds \approx \sum_{i=1}^{r} \beta_i \gamma_i^{k-1} \sum_{j=1}^{r} \alpha_{kj} f(\gamma_j)\,.$$

On the other side, approximating the *right side* by the exterior equation yields

$$\frac{1}{k} \int_0^1 (1 - x^k) f(x)\,dx \approx \frac{1}{k} \sum_{i=1}^{r} \beta_i (1 - \gamma_i^k) f(\gamma_i)\,.$$

Equalizing both sides and substituting for f successively the LAGRANGE polynomials $q_i \in \Pi_{r-1}$ with $q_i(\gamma_j) = \delta^i{}_j\,,\ i = 1, \ldots, r\,,$ yields finally the desired additional condition

$$\sum_{i=1}^{r} \beta_i \gamma_i^{k-1} a_{ij} = \frac{1}{k}\,\beta_j(1-\gamma_j^k)\,,\ j=1:r\,,\ k=1:\varrho\,. \tag{2.69}$$

For simplicity we now introduce the following (nearly historical) abbreviations

$$\begin{array}{|l|}\hline \mathcal{A}(\varrho) :\Longleftrightarrow \text{ the RKM has (at least) order } \varrho \\ \mathcal{B}(\varrho) :\Longleftrightarrow \text{ the exterior equation has (at least) order } \varrho \\ \mathcal{C}(\varrho) :\Longleftrightarrow \text{ the RKM has (at least) interior order } \varrho \\ \mathcal{D}(\varrho) :\Longleftrightarrow \text{ (2.69) holds for } k=1,\dots,\varrho \\ \hline \end{array} \tag{2.70}$$

and the further notations

$$\begin{aligned} &b=[\beta_1,\,\beta_2,\,\dots,\,\beta_r]^T\in\mathbb{R}^r\,,\ c=[\gamma_1,\,\gamma_2,\,\dots,\,\gamma_r]^T\in\mathbb{R}^r\,,\ C=\operatorname{diag}(c)\,,\\ &e=[1,\dots,1]^T\in\mathbb{R}^r\,,\qquad\qquad z_\varrho=[1,\,1/2,\dots,\,1/\varrho]^T\in\mathbb{R}^\varrho\,. \end{aligned} \tag{2.71}$$

Then, by (2.66), (2.67) and (2.68), the stipulations (2.70) are equivalent to

$$\begin{array}{|lll|}\hline \mathcal{A}(\varrho) &\Longleftrightarrow \text{ RKM has order } \varrho & \\ \mathcal{B}(\varrho) &\Longleftrightarrow b^TC^{k-1}e=\dfrac{1}{k}\,, & k=1:\varrho \\ \mathcal{C}(\varrho) &\Longleftrightarrow AC^{k-1}e=\dfrac{1}{k}C^ke\,, & k=1:\varrho \\ \mathcal{D}(\varrho) &\Longleftrightarrow b^TC^{k-1}A=\dfrac{1}{k}(b^T-b^TC^k)\,, & k=1:\varrho \\ \hline \end{array}\,. \tag{2.72}$$

Theorem 2.15. *(*Butcher, Crouzeix, Ehle*) Let a r-stage RKM be given and let all abszissas γ_i be mutually distinct. Then*
(1°) $\mathcal{B}(\varrho)\wedge\mathcal{C}(\xi)\wedge\mathcal{D}(\eta)\Longrightarrow\mathcal{A}(\min\{\varrho,\,2\xi-2,\,\xi+\eta+1\})\,.$
(2°) $\mathcal{B}(\varrho)\wedge\mathcal{C}(r)\Longrightarrow\mathcal{D}(\varrho-r)\,.$
(3°) $\mathcal{B}(\varrho)\wedge\mathcal{D}(r)\Longrightarrow\mathcal{C}(\varrho-r)$ *if all weights* $\beta_i\neq 0\,.$

Consequently, if all abszissas γ_i are mutually distinct and the RKM has property $\mathcal{B}(\varrho)$, then $\mathcal{C}(r)$ or $\mathcal{D}(r)$ determine the crucial property $\mathcal{A}(p)\,.$

For a further *algebraization* of the order conditions let

$$V_\varrho=[\gamma_i^{j-1}]_{i=1\,j=1}^{r\quad\varrho}=\begin{bmatrix}1&\gamma_1&\gamma_1^2&\dots&\gamma_1^{\varrho-1}\\ \vdots&\vdots&\vdots&&\vdots\\ 1&\gamma_r&\gamma_r^2&\dots&\gamma_r^{\varrho-1}\end{bmatrix}\in\mathbb{R}^{r\ \varrho}$$

be the VANDERMONDE matrix. Then, using (2.71), we can write instead of (2.72) in matrix form

$$\boxed{\begin{aligned} \mathcal{B}(\varrho) &\iff V_\varrho^T b = z_\varrho \\ \mathcal{C}(\varrho) &\iff AV_\varrho = \operatorname{diag}(c) V_\varrho \operatorname{diag}(z_\varrho) =: W_\varrho \in \mathbb{R}^r{}_\varrho \\ \mathcal{D}(\varrho) &\iff V_\varrho^T \operatorname{diag}(b) A = (z_\varrho e^T - W_\varrho^T) \operatorname{diag}(b) \end{aligned}} \,. \tag{2.73}$$

Therefore the matrix A of the RKM is uniquely determined by $\mathcal{C}(r)$ if all γ_i are different, and is determined uniquely by $\mathcal{D}(r)$ if in addition all weigths β_i are non-zero.

Corollary 2.2. *(*GAUSS *Methods) The maximum order of the exterior equation, i.e., of the integration rule (2.66), is $p = 2r$ for r stages by Sect. 2.3***(c)***. It is attained if one chooses the roots γ_i , $i = 1 : n$ of $p_{1,n}(x)$ in (2.38). However, $\mathcal{C}(r)$ and $\mathcal{D}(r)$ are equivalent for $\varrho = 2r$ by Theorem 2.15*(2°) *and* (3°) *. Under the above assumptions then, by* (1°)*,*

$$\mathcal{B}(2r) \wedge \Big(\mathcal{C}(r) \vee \mathcal{D}(r)\Big) \Longrightarrow \mathcal{A}(2r).$$

Corollary 2.3. *(*BUTCHER *Methods) Let $\varrho \geq r$ and let all γ_i be different then, by Theorem 2.15*(1°) *and* (2°) *,*

$$\mathcal{B}(\varrho) \wedge \mathcal{C}(r) \Longrightarrow \mathcal{A}(p)\,,\ p = \min\{\varrho, 2r+2, r+\varrho-r+1\} = \varrho\,.$$

For a fixed vector c of abszissas one obtains the BUTCHER *methods of order $\varrho \geq r$ by (2.71) choosing*

$$\big(A, b, c\big) = \big(W_r V_r^{-1},\, V_r^{-1} z_r\,,\, c\big)\,.$$

Corollary 2.4. *(*EHLE *Methods) Let $\varrho \geq 2r-2$, let all γ_i be different and all β_i non-zero, then by Theorem 2.15* (1°) *and* (3°)

$$\mathcal{B}(\varrho) \wedge \mathcal{D}(r) \Longrightarrow \mathcal{A}(p)\,,$$
$$p = \min\{\varrho, 2(\varrho-r)+2, \varrho-r+r+1\} = \min\{\varrho, 2(\varrho-r)+2\} = \varrho\,.$$

For a fixed vector c of abszissas one obtains the EHLE *methods of order $\varrho \geq 2r-2$ by (2.71) choosing*

$$\big(A, b, c\big) = \big(\operatorname{diag}(b)^{-1} V_r^{-T} (z_r e^T - W_r^T) \operatorname{diag}(b)\,,\, V_r^{-T} z_r\,,\, c\big)\,.$$

The roots of the polynomials $p \in \Pi_r$ of (2.38) are chosen for components of the vector c:

$$\begin{aligned}
&\text{GAUSS methods:} && \varrho = 2r, && p_{1,r}(x) && = \left[x^r(1-x)^r\right]^{(r)}\\
&\text{methods of type I:} && \varrho = 2r-1, && xp_{2,r-1}(x) && = \left[x^r(1-x)^{r-1}\right]^{(r-1)}\\
&\text{methods of type II:} && \varrho = 2r-1, && (1-x)p_{3,r-1} && = \left[x^{r-1}(1-x)^r\right]^{(r-1)}\\
&\text{methods of type III:} && \varrho = 2r-2, && x(1-x)p_{4,r-2} && = \left[x^{r-1}(1-x)^{r-1}\right]^{(r-2)}.
\end{aligned} \tag{2.74}$$

The results are summarized in the following table:

Table 2.8.

Type	cond. for γ_i	order	BUTCHER	EHLE	CHIPMAN
GAUSS	(2.72)(1°)	$2r$	$\otimes^A$ =	$\otimes^A$	—
RADAU I B/A	(2.72)(2°)	$2r-1$	$\otimes$	$\otimes^{A,L}$	—
RADAU II A/B	(2.72)(3°)	$2r-1$	$\otimes^{A,L}$	$\otimes$	—
LOBATTO III A/B/C	(2.72)(4°)	$2r-2$	$\otimes^A$	$\otimes^A$	$\otimes^{A,L}$

The CHIPMAN methods have the properties $\mathcal{C}(r)$ and $Ae_1 = \beta_1 e_1\,,\ e_1 = [\delta^1{}_k]_{k=1}^r$ which determine uniquely the matrix A. A-stable methods are marked with $\otimes^A$, L-stable methods with an additional index L.

Example 2.14. The methods of type **Radau II A** have $\gamma_n = 1$ hence the exterior equation and the last row of A are identical. This property has advantages in application to *differential-algebraic problems*. The 3-stage method of order $p = 5$ is A-stable and L-stable, its data are given in the following BUTCHER matrix:

$$\left[\begin{array}{c|c} A & c \\ \hline b & \end{array}\right] = \left[\begin{array}{ccc|c} \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} & \frac{4-\sqrt{6}}{10} \\ \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} & \frac{4+\sqrt{6}}{10} \\ \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} & 1 \\ \hline \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} & \end{array}\right].$$

References: (Hairer), (Shampine97).

2.5 Boundary Value Problems

We look for a solution $x : [0,\,1] \to \mathbb{R}^n$ of the boundary value problem

$$x'(t) = f(t, x(t))\,,\ 0 \le t \le 1\,,\ g(x(0)\,,\, x(1)) = 0 \in \mathbb{R}^n\,, \tag{2.75}$$

confining ourselves to the unit interval by optical reasons and for simple implementation later on. In the other case as, e.g., in periodic problems, a rescaling becomes necessary again: For a problem

$$u'(s) = h(s, u(s))\,,\ 0 \le s \le T\,,\ g(u(0)\,,\, u(T)) = 0 \in \mathbb{R}^n\,,$$

a substitution of $s = T\,t$ yields

$$x'(t) = T\,h(T\,t, x(t))\,,\ 0 \le t \le 1\,,\ g(x(0)\,,\, x(1)) = 0 \in \mathbb{R}^n$$

and the additional factor T has always to be regarded.

(a) The Linear Problem reads:

$$x'(t) = A(t)x(t) + c(t),\ 0 \le t \le 1\,,\ R_0x(0) + R_1x(1) = d \in \mathbb{R}^n\,. \tag{2.76}$$

We choose a uniform partition of the basic t-interval for simplicity,

$$0 = t_1 < t_2 < \ldots < t_m < t_{m+1} = 1\,,\ t_j = (j-1)\tau\,,\ \tau = 1/m\,,$$

beginning with index $j = 1$ w.r.t. the compatibility with MATLAB numeration. Suppose that a *numerical approach* to the differential system on a individual t-interval $[t_j,\, t_{j+1}]$ has the form

$$\boxed{P_jy_j + Q_jy_{j+1} = r_j\,,}$$

then we obtain altogether a large linear system of equations

$$\mathbf{L}(\tau)Y := \begin{bmatrix} P_1 & Q_1 & 0 & \ldots & 0 & 0 \\ 0 & P_2 & Q_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & P_m & Q_m \\ R_0 & 0 & \cdots & \cdots & 0 & R_1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y_{m+1} \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ \vdots \\ \vdots \\ r_m \\ d \end{bmatrix} =: R \tag{2.77}$$

for the unknown values y_j ; and the matrix $\mathbf{L}(\tau)$ must be regular.

Example 2.15. (1°) *Trapezoidal rule*

$$y_{j+1} - y_j - \frac{\tau}{2}\,[A_{j+1}y_{j+1} + A_jy_j] = \frac{\tau}{2}\,(c_{j+1} + c_j)\ ,$$

$$P_j = -I - \frac{\tau}{2}A_j\,,\ Q_j = I - \frac{\tau}{2}A_{j+1}\,,\ r_j = \frac{\tau}{2}\,(c_j + c_{j+1})\ .$$

(2°) *Box scheme*

$$y_{j+1} - y_j - \frac{\tau}{2}A_{j+1/2}\,(y_{j+1} + y_j) = \tau\,c_{j+1/2}\,,$$

$$P_j = -I - \frac{\tau}{2} A_{j+1/2}\,,\ Q_j = I - \frac{\tau}{2} A_{j+1/2}\,,\ r_j = \tau\, c_{j+1/2}\,, \quad \tau = 1/m\,.$$

(3°) *Multiple shooting method*
(3.1°) Solve, for $j = 1, \ldots, m$, the inhomogenous problem with homogenous initial condition

$$x'(t) = A(t)x(t) + c(t)\,,\ t_j \le t \le t_{j+1}\,,\ y(t_j) = 0\,;$$

and denote the solution at point t_{j+1} by r_j.
(3.2°) Solve, for $j = 1, \ldots, m$, the n homogenous initial value problems with inhomogenous initial condition

$$X'(t) = A(t)X(t)\,,\ t_j \le t \le t_{j+1}\,,\ X(t_j) = I \text{ (unit matrix)}\,;$$

and let the solution at point t_{j+1} be the *matrix* V_j. Then

$$y_{j+1} = r_j + V_j y_j \implies y_{j+1} - V_j y_j = r_j\,, \implies \boxed{P_j = -V_j\,,\ Q_j = I}\,.$$

(b) In **nonlinear case** a nonlinear system of equations is produced in the same way and is solved by NEWTON's method. We confine ourselves to the multiple shooting method and apply the flux integral $\Phi(t; t_0, x_0)$ of Sect. 1.6. The numerical solution is denoted by y again.

Multiple shooting method:

(1°) Choose a moderate number m of shooting points in intervall $[0,\ 1]$,

$$[(t_1, y_1), \ldots, (t_m, y_m)]\,,\ y_j \in \mathbb{R}^n\,,\ t_1 = 0\,,\ t_{m+1} = 1\,.$$

(2°) Compute

$$\Phi(t_{j+1}; t_j, y_j) := y_j + \int_{t_j}^{t_{j+1}} f(t, x(t))\, dt\,,\ j = 1 : m\,,$$

by solving the initial value problems

$$x'(t) = f(t, x(t))\ \ t_j \le t \le t_{j+1}\,,\ x(t_j) = y_j\,. \tag{2.78}$$

(3°) Solve the system

$$y_{j+1} - \Phi(t_{j+1}; t_j, y_j) = 0\,,\ j = 1 : m\,,\ g(y_1, y_{m+1}) = 0\,, \tag{2.79}$$

by NEWTON's method. The nonlinear system of equations (2.79) has the form

$$\mathbf{F}(Y) = 0 \in \mathbb{R}^{n(m+1)}\,, \text{ with node vector } Y = [y_1, \ldots, y_{m+1}]^T\,. \tag{2.80}$$

By solving this system with NEWTON's method, a computation of $\operatorname{grad} \mathbf{F}(Y)$ becomes necessary which needs the gradients of Φ at the points (t_j, y_j),

$$\operatorname{grad}_v \Phi(t_{j+1}; t_j, v) = I + \int_{t_j}^{t_{j+1}} \operatorname{grad} f(t, x(t)) \operatorname{grad}_v \Phi(t; t_j, v)\, dt\,. \tag{2.81}$$

The vector field of this matrix-valued flux integral reads:

$$W'(t) = \operatorname{grad} f(t, x(t)W(t)\,,\ W(t)) \in \mathbb{R}^n{}_n\,,$$

and the initial condition for (2.81) is $W(t_j) = I$. Accordingly, in interval $[t_j, t_{j+1}]$, we have to solve again n initial value problems of the form

$$w_k'(t) = \operatorname{grad}_x f(t, x(t))w_k(t) \in \mathbb{R}^n\,,\ t_j \le t \le t_{j+1}\,,\ w_k(t_j) = e_k\,,\ k = 1:n\,, \tag{2.82}$$

where $x(t)$ plays the role of a parameter and $e_k \in \mathbb{R}^n$ is k-th unit vector.

The *simultaneous* solution of all $n+1$ initial value problems (2.78) and (2.81) in every interval $[t_j\,,\,t_{j+1}]$ is essential for success of the method. Then the matrix $\operatorname{grad}\mathbf{F}(Y)$ has the same form as the matrix $\mathbf{L}(\tau)$ in (2.77) where

$$\boxed{P_j = -\operatorname{grad}_v \Phi(t_{j+1}; t_j, y_j)\,,\ \ Q_j = I}\,.$$

The method develops its full power only if the shooting points are chosen properly adapted to the individual problem. Moreover, the NEWTON method must be globalized by a suitable step control. Also, the starting values cannot be chosen arbitrarily but, in simple cases, a linear function respecting the boundary conditions may be sufficient for convergence.

Example 2.16. (Stoer)

$$x_1' = x_2\,,\ \ x_2' = 5\sinh(5\,x_1)\,,\ \ x_1(0) = 0\,,\ x_1(1) = 1\,.$$

For starting trajectory we choose the straight line connecting the boundary points $(0,\, x_1(0))$ and $(1, x_1(1))$. Observe however that $\lim_{t\to 1.0326\ldots} x_1(t) = \infty$, therefore the initially chosen uniform partition of the interval [0, 1] must be adapted to the problem.

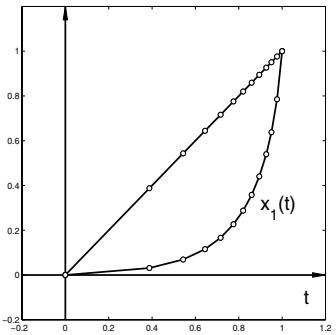

Figure 2.19. Ex. 2.16

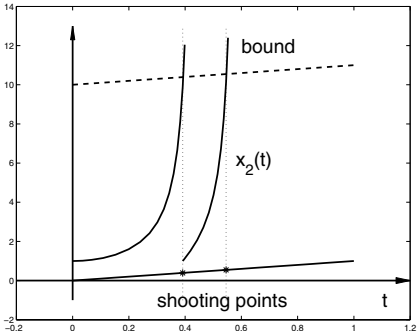

Figure 2.20. Ex. 2.16, adaption

In `KAPITEL04\CONTROL02` some benchmark problems of control theory are solved by NEWTON's method and box scheme.

(c) Boundary Value Problems with Parameter We look for a solution $x(\,\cdot\,,\alpha) : [0\,,\,1] \to \mathbb{R}^n$ of the boundary value problem

$$x'(t;\alpha) = f(t, x(t;\alpha);\alpha)\,,\ 0 \le t \le 1\,,\ g(x(0;\alpha)\,,\,x(1;\alpha);\alpha) = 0 \in \mathbb{R}^n\,, \tag{2.83}$$

where the real parameter α may vary in some interval. Now, the problem has no longer a unique solution but an additional degree of freedom and therefore the numerical solution depends strongly on the chosen initial approximation of x and α which must be given rather accurately.

To apply the multiple shooting method again, let

$$\begin{aligned} \Phi(t;t_0,x_0,\alpha) &= x_0 + \int_{t_0}^{t} f(t, x(t;\alpha);\alpha)\,dt \\ \Phi(t_{j+1};t_j,y_j(\alpha),\alpha) &= y_j(\alpha) + \int_{t_j}^{t_{j+1}} f(t, x(t;\alpha);\alpha)\,dt \end{aligned}$$

be the flux integral belonging to (2.83). Solving the system (2.80) being now of the form

$$\mathbf{F}(V) = 0 \in \mathbb{R}^{n(m+1)+1},\ \ V = [y_1,\ldots,y_{m+1};\alpha]^T \text{ node vector}\,, \tag{2.84}$$

by NEWTON's method, one needs the additional derivative

$$\begin{aligned} H_j := \frac{\partial}{\partial\alpha}\Phi(t_{j+1};t_j,y_j(\alpha),\alpha) &= \frac{\partial}{\partial\alpha}y_j(\alpha) + \int_{t_j}^{t_{j+1}} \frac{\partial}{\partial\alpha} f(t, x(t;\alpha);\alpha)\,ds \\ &+ \int_{t_j}^{t_{j+1}} \operatorname{grad}_x f(t, x(t;\alpha);\alpha)\frac{\partial}{\partial\alpha}x(t;\alpha)\,dt\,. \end{aligned}$$

Therefore the additional initial value problem

$$\begin{aligned} v_j'(t) &= \frac{\partial}{\partial\alpha} f(t, x(t;\alpha);\alpha) + \operatorname{grad}_x f(t, x(t;\alpha);\alpha) v_j(s)\,, \\ v_j(t_j) &= \frac{\partial}{\partial\alpha}(y_j)(\alpha) \in \mathbb{R}^n\,,\ v_1(t_1) = 0\,, \end{aligned} \tag{2.85}$$

has to be solved in every interval $[t_j\,,\,t_{j+1}]$. Note that all $n+2$ initial value problems (2.78), (2.82) and (2.85) must be solved *simultaneously* again. Note also that during the entire iteration not only the numerical approximations V are to be calculated but als the partial derivatives $y_{\alpha,j}\,,\ j = 1 : m+1$ w.r.t. the parameter α else the method may fail.

The JACOBI matrix $\operatorname{grad}\mathbf{F}(V)$ is now a $(m\cdot n, m\cdot n+1)$-matrix in block form

$$\mathbf{L} := \begin{bmatrix} P_1 & Q_1 & 0 & \dots & 0 & 0 & H_1 \\ 0 & P_2 & Q_2 & \ddots & \ddots & 0 & H_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & H_{m-1} \\ 0 & \ddots & \ddots & \ddots & P_m & Q_m & H_m \\ g_{x_0} & 0 & \cdots & \cdots & 0 & g_{x_1} & g_\alpha \end{bmatrix}, \tag{2.86}$$

therefore the MOORE-PENROSE inverse $[\operatorname{grad}\mathbf{F}(V)]^+$ must be used in the GAUSS-NEWTON method, cf. Sect. 1.1**(g)**. An underdetermined linear system of equations

$$[\operatorname{grad}\mathbf{F}(V_j)](V_{j+1}-V_j) = -\mathbf{F}(V_j)$$

is to be solved in every NEWTON step for which the algorithm in Sect. 1.1**(h3)** can be applied. Also, as already mentioned, a good initial approximation V_0 must be known to prevent a convergence to the trivial solution. In rather simple methods also the box scheme may be modified in a suitable way.

2.6 Periodic Problems

(a) Problems with Known Period We seek a T-periodic solution $x : \mathbb{R} \to \mathbb{R}^n$ of the boundary value problem

$$x'(t) = f(t, x(t))\,,\ 0 \le t \le T\,,\ x(0) = x(T) \in \mathbb{R}^n\,. \tag{2.87}$$

If $x(t)$ is a solution, also $x(t+\alpha)$ is a T-periodic solution here for arbitrary $\alpha \in \mathbb{R}$. Therefore an additional *phase condition* must be introduced to ensure uniqueness, but the problem remains nevertheless numerically unstable.

Some possible phase conditions are

(1°) $\quad p(x(0)) := x_k - \eta = 0\,,\ \eta \neq 0;$

(2°) $\quad p(x(0)) := f_k(x(0)) = 0 \implies x_k'(0) = 0.$

But then we have $n+1$ boundary conditions for n unknown functions. If a point on the unknown orbit is known then satisfying results may be obtained also by a good solver for initial value problems.

(b) In **problems with unknown period** T, a transformation to a parameter-dependent problem with known period suggests itself. For instance, the solution $\widetilde{x}$ of

$$\widetilde{x}\,'(s) = T\, f(Ts, \widetilde{x}(s))\,,\ \widetilde{x}(0) = \widetilde{x}(1) \tag{2.88}$$

has period one in s and $x(t) = \widetilde{x}(t/T)$ is a solution of (2.87) with period T. The further treatment of the problem is carried out as in Sect. 2.5**(c)** for parameter-dependent problems. The relatively simple box scheme however

cannot be applied here because of the boundary condition (2.88) and the results of Sect. 2.5**(a)**. In the subsequent examples the *multiple shooting method* is used with *fixed* partition of the underlying t-interval and the solution of an initial value problem with *estimated* initial value for starting trajectory. In particular, this method may serve for final adjustment of periodic solutions.

Example 2.17. Nerve membrane model (Deuflhard84) (Fig. 2.21).

$$\begin{aligned} \dot{u}_1 &= 3(u_2 + u_1 - \frac{1}{3}u_1^3 + \lambda) \\ \dot{u}_2 &= -\frac{1}{3}(u_1 - 0.7 + 0.8u_2)\,. \end{aligned}$$

Transformation to a parameter-dependent problem with period one:

$$\begin{aligned} x_1' &= 3T(x_2 + x_1 - \frac{1}{3}x_1^3 + \lambda) \\ x_2' &= -\frac{T}{3}(x_1 - 0.7 + 0.8x_2)\,. \end{aligned}$$

The initial value problem (2.85) reads:

$$\begin{aligned} v_1' &= 3(x_2 + x_1 - \frac{1}{3}x_1^3 + \lambda) + 3T(1 - x_1^2)v_1 + 3Tv_2 \\ v_2' &= -\frac{1}{3}(x_1 - 0.7 + 0.8x_2) - \frac{T}{3}v_1 - \frac{0.8T}{3}v_2\,. \end{aligned}$$

Test problem : $\lambda = -1$, starting value $(x_1^0, x_2^0, T^0) = (3, 1.5, 12)$.

Example 2.18. Heated flow problem (Deuflhard84) (Fig. 2.22).

$$\dot{u}_1 = -\sigma(u_1 - u_2)\,,\quad \dot{u}_2 = u_1(r - u_3) - u_2\,,\quad \dot{u}_3 = u_1u_2 - bu_3\,.$$

Transformation to a parameter-dependent problem with period one:

$$x_1' = -\sigma T(x_1 - x_2)\,,\quad x_2' = T\,[x_1(r - x_3) - x_2]\,,\quad x_3' = T\,(x_1x_2 - b\,x_3)\,.$$

The initial value problem (2.85) reads:

$$\begin{aligned} v_1' &= -\sigma(x_1 - x_2) && - \sigma T(v_1 + v_2) \\ v_2' &= x_1(r - x_3) - x_2 && + T[(r - x_3)v_1 - v_2 - x_1v_3] \\ v_3' &= x_1x_2 - b\,x_3 + x_2v_1 && + T(x_1v_2 - b\,v_3)\,. \end{aligned}$$

Test problem : $\sigma = 16\,,\ b = 4\,,\ r = 153.083\,$, starting values $(x_1^0, x_2^0, x_3^0, T^0) = (0, -28, 140, 0.95)$.

Example 2.19. ARENSTORF orbits (Arenstorf). In the degenerated three-body problem, three bodies (earth, moon, satellite) are given with masses m_1, m_2 und $m_3 = 0$, and with the following simplifications (Fig. 2.23):

(1°) Earth, moon, satellite move in a plane.
(2°) The distance earth-moon is constant and set to one.
(3°) The influence of the remaining celestial bodies is neglected.

The straight line between earth and moon is chosen for x-axis with common gravity center for origin. Furthermore, $\mu = m_2/(m_1 + m_2) \sim 1/81.45$ is the relative moon mass, $\mu' = 1 - \mu$. Then a system of *two* differential equations is obtained for the motion of the mass-free body in the rotating frame; see Sect. 6.5 **(b)**. For a transformation of the T-periodic problem into a problem with unit period, it is referred to the appertaining MATLAB program.

Example 2.20. Nonlinear oscillators occur in many technical applications. Forced DUFFING equations

$$\ddot{u} + \alpha\dot{u} + \beta u + \gamma u^3 = \delta\cos(\omega t)\,.$$

are a standard model problem in investigation of period doubling, transition to chaos, and bifurcation (in homogenous case); see e.g. (Seydel94). We look here for *harmonic* solutions which have the same period $T = 2\pi/\omega$ as the excitation (else we are led to *strange attractors*). Transformation to a parameter-dependent system with period one by substitution of $t = Ts$, $T = 2\pi/\omega$, yields as above with $y_1(s) = u(t)$

$$\begin{aligned} y_1' &= Ty_2 \\ y_2' &= -T\big(\alpha y_2 + \beta y_1 + \gamma y_1^3 - \delta\cos(2\pi s)\big)\,. \end{aligned}$$

In Fig. 2.24 we have $\alpha = 0.2$, $\beta = 0$, $\gamma = 1$, and at beginning $\delta = 5$. At first an initial value problem is solved to get a start trajectory, then a simple continuation is chosen up to $\delta = 7$. Initial guess of period = 12, final period = 10.2209.

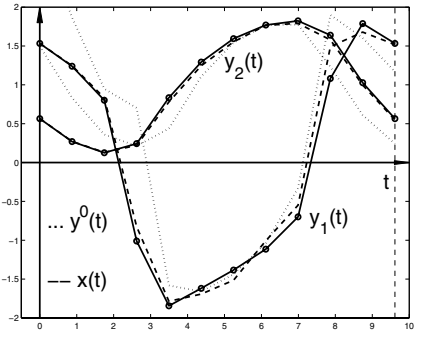

Figure 2.21. Example 2.17

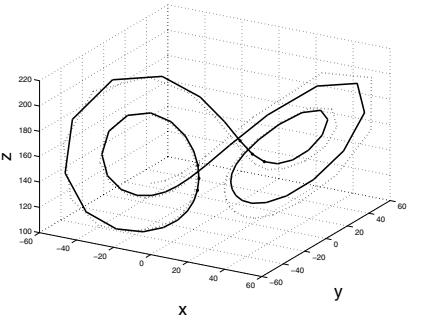

Figure 2.22. Example 2.18

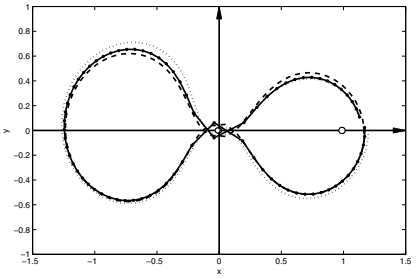

Figure 2.23. Example 2.19

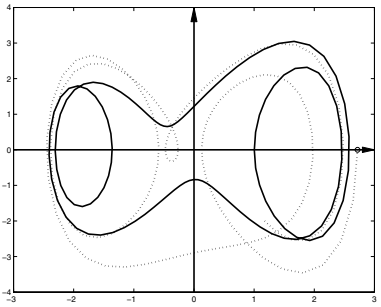

Figure 2.24. Example 2.20

2.7 Differential-Algebraic Problems

Extremal problems are solved in mechanics by their associated variational problem (EULER equations), and frequently a formulation of the extremal function or objective function itself is relinquished at all; cf. Sect. 4.1. Possible equality or inequality restrictions are taken into the objective function via LAGRANGE multiplicators as far as possible; indeed the entire LAGRANGE theory has originated in mechanics of mass points. But the side conditions have nevertheless to be regarded and thus a more or less (rather more) complicated system of differential equations, analytic equations (here apostrophized as "algebraic") and perhaps also inequalities is waiting for numerical approach. Ultimately one is faced with a highly nonlinear *boundary value problem* or a equally nonlinear *initial value problem* of which the solution must satisfy additional restrictions; cf. Chap. 3. Numerical devices for solving such families of problems need a *consistent* start trajectory of which the calculation is often more difficult than the remaining computation. If however the problem is transformed artificially into a control problem, then modern numerical methods can be applied working with (rather) *arbitrary* initial trajectory; cf. Sect. 4.4. If, on the other side, the differential-algebraic problem is a *pure initial value problem* then also special RUNGE-KUTTA methods may be applied. Some of these methods, having been developed in more recent time, shall be considered in the present section. The problem of consistent initial values appears here, too, but is a nonlinear system of equations being solved by the usual methods in the generic case. For some practical applications of these methods we refer to Sect. 11.3 on *multibody problems*.

In this subsection let (x, y) be the theoretical solution and (u, v) its numerical approximation.

(a) Formulation of the Problem At first we consider a *singular* initial value problem in separated form

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)) \in \mathbb{R}^n\,, \quad (x(0), y(0)) = (x_0,\, y_0)\,, \\ \varepsilon\, y'(t) &= g(t, x(t), y(t)) \in \mathbb{R}^m\,, \quad 0 \le \varepsilon \ll 1 \end{aligned} \tag{2.89}$$

depending on the parameter ε. Writing $z(t) = [x(t),\, y(t)]^T$, the system is equivalent to

$$M z'(t) = F(t, z(t)) \in \mathbb{R}^{n+m}\,, \quad z(t) = [x(t),\, y(t)]^T \tag{2.90}$$

where the matrix $M \in \mathbb{R}^{n+m}{}_{n+m}$ is *singular* for $\varepsilon = 0$. Problems of the form $M(x(t))x'(t) = F(t, x(t))$ are transformed by preference in a system

$$x' = y\,, \quad M(x)y - F(x) = 0\,. \tag{2.91}$$

Assumption 2.2. (1°) *The problem (2.89) has a unique solution in* $[0,\, T]$, $0 < T$.

(2°) *The gradient* $\nabla_y g(x, y)$ *is* regular *near the solution* (x, y) *for* $\varepsilon = 0$.

The problem is said to be a *differential-algebraic problem* (DA problem) in the case where $\varepsilon = 0$. In this case the initial values (x_0, y_0) must be *consistent*, i.e., they must satisfy the side condition $g(x_0, y_0) = 0$. Also in the *numerical solution* such initial values must be known or calculated first, at least approximatively. The DA-problem is said to have *index 1* if assumption 2.2 (2°) does hold. Then the function g is invertible w.r.t. y near the solution, $y(t) = G(t, x(t))$, and one obtains by substitution at least *theoretically* an ordinary initial value problem with the differential system $x'(t) = f(t, x(t), G(t, x(t)))$.

Example 2.21. Van der Pol's equation.

The *linear oscillator* $\ddot{x} + \alpha\dot{x} + x = 0$ is damped for $\alpha > 0$ and unstable for $\alpha < 0$. If the parameter α is replaced by $\mu(x^2 - 1)$, $\mu > 0$, then large $|x(t)|$-values lead to a damping and small $|x(t)|$ to an amplification. Transformation in a system of first order yields

$$\dot{x}_1 = x_2\,, \quad \dot{x}_2 = \mu\,(1 - x_1^2)x_2 - x_1\,;$$

If we now write $x_1 = y_1$, $y_2 = \mu\, x_2$, $s = t/\mu$ and ensuing $\mu^2 = 1/\varepsilon$ then we obtain after re-notation

$$\dot{x} = y\,, \; \varepsilon\, \dot{y} = (1 - x^2)y - x\,.$$

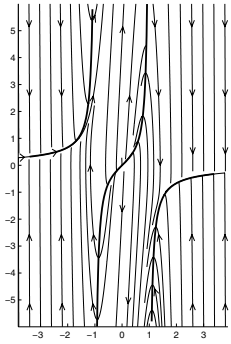

Figure 2.25. Van der Pol's equation

In Figure 2.25 the phase portrait is plotted for $\varepsilon = 0.05$ and the curve $(1 - x^2)y - x = 0$ is marked boldface.

(b) DA-problems are mainly solved by special **Runge-Kutta methods** but also multistep methods may be applied, in particular backward differentiation methods as dealed with in Sect. 2.4**(i)**(4°). If $\varepsilon > 0$ for the present then, with the notations of Sect. 2.4**(d)**, we obtain as *common* method for the *separated* equations

$$\begin{aligned} U(t) &= e \times u(t) + \tau(A \times I)F(t, U(t), V(t)) && \in \mathbb{R}^{r\cdot n} \\ \varepsilon\, V(t) &= \varepsilon\, e \times v(t) + \tau(A \times I)G(t, U(t), V(t)) && \in \mathbb{R}^{r\cdot m} \\ u(t+\tau) &= u(t) + \tau(b \times I)^T F(t, U(t), V(t)) && \in \mathbb{R}^{n} \\ \varepsilon\, v(t+\tau) &= \varepsilon\, v(t) + \tau(b \times I)^T G(t, U(t), V(t)) && \in \mathbb{R}^{m}\,; \end{aligned} \tag{2.92}$$

where $U(t)$ and $V(t)$ are the vectors at the intermediate stages. If now the matrix A is *regular* then the second equation can be written as

$$\tau\, G(t, U(t), V(t)) = \varepsilon\, (A^{-1} \times I)[V(t)) - (e \times v(t))]\,,$$

and, by substitution of the last equation, the parameter ε may be *canceled*. By this way one obtains a *direct approximation* of the DA-problem by a Runge-Kutta method:

$$\begin{aligned} U(t) &= e \times u(t) + \tau(A \times I)F(t, U(t), V(t)) && \in \mathbb{R}^{r\cdot n} \\ 0 &= G(t, U(t), V(t)) && \in \mathbb{R}^{r\cdot m} \\ u(t+\tau) &= u(t) + \tau(b \times I)^T F(t, U(t), V(t)) && \in \mathbb{R}^{n} \\ v(t+\tau) &= (1 - b^T A^{-1} e)v(t) + (b \times I)^T (A^{-1} \times I)V(t) && \in \mathbb{R}^{m}\,, \end{aligned} \tag{2.93}$$

and the stability function satisfies $R(\infty) = 1 - b^T A^{-1} e$; cf. (2.58). However, the algebraic side condition $g(u, v) = 0$ is fulfilled only approximatively in methods of this type (in normal case). This disadvantage is removed if the last equation in (2.93) is replaced by requiring $g(u_{n+1}, v_{n+1}) = 0$. The resulting *indirect* type of methods constitutes an approximation of $x' = f(x, G(x))$ in systems of index 1. If however, besides the regularity of A, also the last row of A is the same as the vector b of weights in the exterior equation (*stiffly accurate methods*) then $g(u_{n+1}, v_{n+1}) = 0$ is fulfilled automatically because of the second equation in (2.93), and the last equation can be dropped. For instance the Rung-Kutta methods of type Radau II A described in Sect. 2.4**(j)** have the just mentioned additional property and thus are suited in a particular way for solving DA-problems.

Let us now apply a Runge-Kutta method to a differential system $M\, x' = f(t, x)$ with *regular* matrix M, then we obtain the computational device

$$\begin{aligned} &(I \times M)\big(U(t) - e \times u(t)\big) = \tau(A \times I)F(t, U(t)) \in \mathbb{R}^{r\cdot n} \\ &u(t+\tau) = (1 - b^T A^{-1} e)u(t) + (b \times I)^T (A^{-1} \times I)U(t) \in \mathbb{R}^{m}\,, \end{aligned} \tag{2.94}$$

in the same way as in the transition of (2.92) to (2.93) and this device works also in a *singular* matrix M. But then the method depends on the condition of the matrix $I \times M - \tau(A \times I)$ hence in particular of the step length τ.

(c) Regular Matrix Pencils Let (λ, u) be a characteristic pair of the *generalized eigenvalue problem*

$$(A + \lambda B)u = 0\,, \quad A, B \in \mathbb{R}^n{}_n\,.$$

Then $x(t) = e^{\lambda t}u$ is a solution of the differential system

$$Bx' + Ax = c(t) \in \mathbb{R}^n \tag{2.95}$$

for $c(t) \equiv 0$. If here, e.g., $A = B$ and $\det(A) = 0$ then $A + \lambda B$ is singular for *all* $\lambda \in \mathbb{R}$, therefore it is tacitly assumed in linear systems (2.95) that the *matrix pencil* $A+\lambda B$ is *regular* such that the associated generalized eigenvalue problem has a *finite* number of nonzero eigenvalues.

Theorem 2.16. *(*Weierstrass, Kronecker*) Let $A+\lambda B$ be a regular matrix pencil then there exist regular matrices P, Q such that*

$$PAQ = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}\,, \quad PBQ = \begin{bmatrix} I & 0 \\ 0 & T \end{bmatrix}\,. \tag{2.96}$$

*The matrix $T = \mathrm{diag}(T_1, \ldots, T_k)$ is a block diagonal matrix with blocks $T_i \in \mathbb{R}^{n_i}{}_{n_i}$ of the form described in Sect. 1.1***(c3)***, and $n_1 + \ldots + n_k = n$.*

Proof see e.g. (Hairer), vol. II, Sect. 6.5.

Let us now multiply (2.95) by P and use the partition

$$\begin{bmatrix} y \\ z \end{bmatrix} = Q^{-1}x\,, \begin{bmatrix} f \\ g \end{bmatrix} = Pc(t)\,,$$

then we obtain two separated systems for for y and z, namely

$$y' = Cy + f(t)\,, \quad Tz' + z = g(t)\,. \tag{2.97}$$

(d) Differential Index The second system in (2.97) has to be solved by recurrence. If for instance $k = 1$ in Theorem 2.16 and $T = T_1 \in \mathbb{R}^m{}_m$ then one starts out from the last row $z_m = g_m(t) \in \mathbb{R}$ and then has to calculate successively the components $z_i(t)\,,\ i = m - 1 : 1\,,$ by $z_i(t) = g_i(t) - z_{i+1}^{(i+1)}(t)$ for which one needs the derivatives $g_m^{(m)}, \ldots, g_2^{(1)}$. Also, with these derivatives, the system $Tz' + z = g(t)$ can be written as *explicit* system,

$$z_{i+1}^{(i+1)} + z_i^{(i)} = g_i^{(i)}(t)\,, \quad i = 1, \ldots, m-1\,, \quad z_m^{(m)} = g_m^{(m)}(t)\,. \tag{2.98}$$

In general, the *differential index* is the number of derivatives being necessary to transform the implicit differential system $F(t, x'(t), x(t)) = 0$ *analytically* into an explicit system. The explicit system $x'(t) = f(t, x(t))$ has index zero by definition, and, e.g., the system (2.98) has index m because m derivatives of g are necessary.

System with index 1. Let the matrix $\nabla_y g(x,y)$ be regular near a solution (x,y) of

$$x'(t) = f(t,x(t),y(t))\,,\quad g(t,x(t),y(t)) \tag{2.99}$$

then we obtain by $0 = \nabla_x g(x,y)x' + \nabla_y g(x,y)y'$ together with (2.99)(1°) the explicit system

$$x' = f(x,y)\,,\quad y' = -\nabla_y g(x,y)^{-1}\nabla_x g(x,y)f(x,y)\,.$$

In this case the system (2.99) has *index 1*.

System with index 2. Let $\nabla_y g(x,y)$ be singular near a solution of (2.99). Then $h(x,y) := \nabla_x g(x,y)f(x,y) = 0$ follows from $g(x,y) = 0$ and

$$\begin{aligned}\nabla_x h(x,y) &= \nabla^2_{xx} g(x,y)f(x,y) + \nabla_x g(x,y)\nabla_x f(x,y)\\ \nabla_y h(x,y) &= \nabla_y \nabla_x g(x,y)f(x,y) + \nabla_x g(x,y)\nabla_y f(x,y)\,.\end{aligned}$$

If $\nabla_y h(x,y)$ is regular then $x' = f(x,y)\,,\ h(x,y) = 0$ is a system with index 1. The system (2.99) has *index 2* in this case. By solving $\nabla_x h(x,y)x' + \nabla_y h(x,y)y' = 0$ w.r.t. y' one obtains again an explicit system of first order,

$$x' = f(x,y)\,,\quad y' = -\nabla_y h(x,y)^{-1}\nabla_x h(x,y)f(x,y)\,.$$

System with index 3. If $\nabla_y g(x,y)$ and $\nabla_y h(x,y)$ are both singular near the solution of (2.99) then $k(x,y) := \nabla_x h(x,y)f(x,y) = 0$ follows from $h(x,y) = 0$. If now $\nabla_y k(x,y)$ is regular then the system $x' = f(x,y)\,,\ k(x,y) = 0$ has index 1. The system (2.99) has *index 3* in this case.

(e) In more recent time also **Semi-Explicit Runge-Kutta Methods**

$$\begin{aligned} U(t) &= e \times u(t) + \tau(A \times I)F(U(t),V(t)) &&\in \mathbb{R}^{r\cdot n}\\ 0 &= G(U(t)) &&\in \mathbb{R}^{r\cdot m}\\ u(t+\tau) &= u(t) + \tau(b \times I)^T F(U(t),V(t)) &&\in \mathbb{R}^{n}\\ 0 &= g(u(t+\tau)) &&\in \mathbb{R}^{m}\end{aligned} \tag{2.100}$$

have been proposed for solving DA-problems of the form

$$x'(t) = f(x(t),y(t))\,,\quad g(x(t)) = 0\,. \tag{2.101}$$

If the matrix A of coefficients is a *triangular matrix* with $\operatorname{diag}(A) = 0$ then we obtain the following device for a single t-step:

Set $u_1 = u$ and compute v_1 with NEWTON's method by
$g\big(u + \tau\alpha_{21} f(u_1,v_1)\big) = 0$ for $i = 2:r$.
Set $u_i = u + \tau\sum_{j=1}^{i-1}\alpha_{ij} f(u_j,v_i)$
and compute v_i with NEWTON's method by
$g(u_i) = 0\,,\ i = 2:r$.
Set $u(t+\tau) = u + \tau\sum_{i=1}^{r}\beta_i f(u_i,v_i)$
and compute $v(t+\tau)$ with NEWTON's method by
$g(u(t+\tau)) = 0$. (2.102)

Theorem 2.17. (1°) *Let the problem (2.101) have a unique solution in* $[0,\, T]\,,\ 0 < T\,.$
(2°) *Let the initial values satisfy* $g(x_0) = 0\,,\ \nabla g(x_0) f(x_0, y_0) = 0\,.$
(3°) *Let* $\nabla g(x) \nabla_y f(x, y)$ *be regular near the solution (system with index 2).*
(4°) *In the matrix A and vector b of the method (2.100), let*

$$\alpha_{i,i-1} \neq 0\,,\ i = 2 : r\,,\ \ \beta_i \neq 0\,,\ i = 1 : r\,.$$

Then the systems in (2.102) have a local unique solution for sufficiently small $\tau\,.$

Proof see (Brasey92), (Brasey93).

Example 2.22. HEM4, 5-stage method of order $p = 4$ by (Brasey92).

$$\left[\begin{array}{c|c} A & c \\ \hline b & \end{array}\right] = \left[\begin{array}{ccccc|c} - & - & - & - & - & - \\ \frac{3}{10} & - & - & - & - & \frac{3}{10} \\ \frac{1+\sqrt{6}}{30} & \frac{11-4\sqrt{6}}{30} & - & - & - & \frac{4-\sqrt{6}}{10} \\ \frac{-79-31\sqrt{6}}{150} & \frac{-1-4\sqrt{6}}{30} & \frac{24+11\sqrt{6}}{25} & - & - & \frac{4+\sqrt{6}}{10} \\ \frac{14+5\sqrt{6}}{6} & \frac{-8+7\sqrt{6}}{6} & \frac{-9-7\sqrt{6}}{4} & \frac{9-\sqrt{6}}{4} & - & 1 \\ \hline 0 & 0 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} & - \end{array}\right]\,.$$

2.8 Hints to the MATLAB programs

```
KAPITEL02/SECTION_1_2_3
Figures of Section 2.1 and 2.2
demo1.m      Test of four Gauss rules in interval
demo2.m      Test of Gauss and Bell rules in arbitrary triangle
bell.m       Exact integration of polynomial
             in arbitrary triangle
gauss_1.m:   Gauss-Legendre integration
gauss_2/3/4.m: Gauss integration, suboptimal, three cases
gauss_t5.m:  Gauss rule of order n = 5 in arbitrary triangle
divdif.m:    Generalized divided differences
KAPITEL02/SECTION_4: Initial Value Problems
Figures of Section 2.4 and stability regions
demo1.m      Arenstorf orbits by using dopri.m,
dopri.m      MATLAB version of FORTRAN version of HAIRER I
dreik_a.m    differential system of restricted
             three-body problem
stab_region.m Program for plots of the stability regions
             of one-step methods
```

```
KAPITEL02/SECTION_5: Boundary Value Problems
adapt01.m    Adaption of shooting points for example
box.m        Box scheme for Newton method
bsp01.m      Example Stoer-Bulirsch, Par. 7.3, Bsp. 1
demo1.m      Masterfile for multiple shooting method
mehrziel.m   Multiple shooting for Newton method
newton.m     Quasi-global Newton method
Kapitel02/SECTION_6: Periodic Problems
bsp01.m      Nerve membran model
bsp02.m      Heat flow problem
bsp03.m      Arenstorf orbit I
bsp04.m      Duffing's equation
demo1.m      Masterfile for multiple shooting method
demo2.m      Periodic solution of Duffing's equation
demo1.m      Some solutions of Duffing's equation
mehrziel_p.m Multiple shooting scheme for Newton's method
             and problems with unknown period
newton_p.m   Quasi-global Newton's method for periodic problems
```

http://www.springer.com/978-3-540-69278-2

Mathematical Methods for Mechanics
A Handbook with MATLAB Experiments
Gekeler, E.W.
2008, XVI, 624 p. 218 illus., Hardcover
ISBN: 978-3-540-69278-2