

## Numerische Methoden

Bevor der *Computer* (franz. *ordinateur*) die Welt veränderte, konnte sich die *Numerische Mathematik* – von Spöttern gern als phänomenologisch bezeichnet – kaum zu den Königsdisziplinen der Mathematischen Wissenschaften rechnen. Ob es heute der Fall ist, sei dahin gestellt, aber im Verbund mit Modellierung und Simulation ist sie in der Hierarchie aufgestiegen und letztere müssen sogar teilweise zur Existenzberechtigung der anderen Fächer herhalten. In den Sechziger Jahren des vorigen Jahrhunderts wurden noch Integrimeter, Integrator und harmonischer Oszillator in Vorlesungen über *Instrumentelle Mathematik* behandelt. Längst sind sie der Vergessenheit anheimgefallen ebenso wie alle numerischen Verfahren für die Handrechenmaschine z.B. das Wurzelziehen durch Subtraktion ungerader Zahlen.

Mit der stürmischen Entwicklung des Computers zum *Laptop* konnte die *Numerische Mathematik* mit einigem Abstand erfolgreich mithalten. Mathematische Tafelwerke sind durch den Taschenrechner ersetzt, und lineare Gleichungssystem – ein zentrales Problem – werden heute mit drei unscheinbaren Glyphen  $\backslash \mathbf{b}$  gelöst, ohne einen Stilbruch zu begehen. Eine Fülle von Monographien dokumentiert das bisher Erreichte, exemplarisch genannt seien nur [Golub], [Hairer] und [Rheinboldt70]. Auch scheint die Kurve der Publikationszahlen mit rein numerischen Themenstellungen etwas flacher zu werden, während auf der anderen Seite problembezogene Anwendungen zunehmen.

Wenn Numerische Methoden in einem einzigen Kapitel beschrieben werden sollen, ist Mut zur Lücke angesagt. Der Verfasser geht davon aus, dass der Leser in erster Linie vorhandene Codes anwenden will, was aber nicht ohne ein Mindestmaß an Verständnis und Einfühlungsvermögen geht. Unter dieser Prämisse soll hier in die Denkweise der *Numerik* eingeführt aber auch anspruchsvollere Entwicklungen wie Mehrzielverfahren und differential-algebraische Probleme angesprochen werden. Der Numerische Teil des Buches erschöpft sich nicht in den hier behandelten Themen, weitere Problemkreise werden später behandelt, wenn die notwendigen theoretischen Grundlagen vorhanden sind.

## 2.1 Interpolation und Approximation

Vielfach liegt eine Funktion  $f$  nur als Datensatz vor oder muss durch eine leichter integrierbare Funktion ersetzt werden. Dann wird sie im Regelfall *stückweise* durch ein Polynom *mäßigen Grades* approximiert, weil Polynome hohen Grades in größeren Intervallen stark oszillieren. Es sei

$$\Pi_n \text{ die Menge der reellen Polynome } p_n \text{ vom Grad } \leq n \text{ .}$$

Mit der üblichen Addition und skalaren Multiplikation ist  $\Pi_n$  ein *Vektorraum* der Dimension  $n + 1$ , dessen Basis  $\{q_0(x), \dots, q_n(x)\}$  je nach den gestellten Anforderungen geeignet gewählt wird.

**(a) Das allgemeine Interpolationsproblem** Gegeben sei

$$\begin{aligned} &\text{eine Folge von Stützstellen } \{x_i\}_{i=0}^\infty, x_i \in \mathbb{R}, \\ &\text{eine Folge von Stützwerten } \{f_i\}_{i=0}^\infty, f_i \in \mathbb{R}, \\ &\text{eine Folge von Funktionen } \{g_i\}_{i=0}^\infty, g_i \in \mathcal{C}[a, b]. \end{aligned}$$

Die Stützstellen  $x_i$  sollen *alle verschieden* sein, für den anderen Fall verweisen wir auf [Hoellig] § 3.1. Gesucht ist eine Folge von Funktionen

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

mit der *Interpolationseigenschaft*

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

Schreibt man  $a = [\alpha_0, \dots, \alpha_{n-1}]^T$  und  $f = [f_0, \dots, f_{n-1}]^T$ , dann ist (2.1) für festes  $n$  äquivalent mit dem linearen Gleichungssystem

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

und das Interpolationsproblem (2.1), (2.2) ist eindeutig lösbar, wenn die Matrix  $A$  regulär ist.

**Satz 2.1.** (*EE-Satz, HAAR-Bedingung*) Wenn alle  $n$  Stützstellen  $x_j$ ,  $j = 0 : n - 1$ , verschieden sind und jede nicht identisch verschwindende Linearkombination von  $n$  Funktionen  $g_i$ ,  $i = 0 : n - 1$ , höchstens  $n - 1$  Nullstellen hat, dann ist die Matrix  $A$  regulär und damit das Interpolationsproblem eindeutig lösbar.

*Beweis.* Ist die Matrix  $A$  singulär, dann gibt es einen Vektor  $\underline{c} \in \mathbb{R}_n$  mit  $\underline{c}A = 0 \in \mathbb{R}_n$ . Dann hat aber die Linearkombination  $h(x) := \sum_{i=0:n-1} c_i g_i(x)$  die  $n$  verschiedenen Nullstellen  $x_0, \dots, x_{n-1}$  wegen

$$h(x_j) = \sum_{i=0:n-1} c_i g_i(x_j) = 0, \quad j = 0 : n-1,$$

im Widerspruch zur Voraussetzung.  $\square$

Die HAAR-Bedingung ist insbesondere für eine Folge  $\{p_n\}_{n=0}^\infty$  von Polynomen  $p_n \in \Pi_n$  erfüllt, weil dann jede nicht identisch verschwindende Linearkombination  $h(x) := \sum_{i=0:n-1} c_i p_i(x)$  als Polynom vom Grad  $\leq n-1$  höchstens  $n-1$  Nullstellen haben kann. I.A. hat aber die Matrix  $A$  in (2.3) eine schlechte Kondition, deswegen wird dieses lineare Gleichungssystem nicht zur Berechnung der Gewichte  $\alpha_i$  verwendet.

**(b) Interpolationspolynome** Zur Bestimmung einer *linearen Rekursionsformel* für *Interpolationspolynome* sei  $\{j_0, \dots, j_m\} \subset \{0, \dots, n\}$  eine Indexmenge mit lauter verschiedenen Einträgen, dann ist das Interpolationspolynom  $p_{j_0, \dots, j_m}(x) \in \Pi_m$  nach dem EE-Satz eindeutig bestimmt durch

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

insbesondere ist  $p_{j_0, \dots, j_m}(x)$  unabhängig von der Reihenfolge der Indizes.

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

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

Diese Formel wird in verschiedenen Anwendungen zur schnellen Berechnung von Interpolationspolynomen an einer festen Stelle  $x$  verwendet.

**Satz 2.2.** (CAUCHY-Restglied) Die Funktion  $f$  sei  $(n+1)$ -mal differenzierbar in  $[a, b]$ , und es sei  $[u, v, \dots, w]$  das kleinste Intervall  $\mathcal{I} \subset \mathbb{R}$  mit  $u, v, \dots, w \in \mathcal{I}$ .

$\forall x \in [a, b] \quad \exists \xi_x \in [x_0, \dots, x_n, x]:$

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

Beweis SUPPLEMENT\chap02.

**(c) Interpolation nach Lagrange** Wir betrachten die Approximation einer Funktion  $f$  durch ein Interpolationspolynom in der *separierten* Form

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

mit der Basis  $\{q_0, \dots, q_n\}$  von  $\Pi_n$  der LAGRANGE-Grundpolynome

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

Wegen  $q_i(x_j) = \delta^i_j$  (KRONECKER-Symbol) ist die Interpolationseigenschaft  $p_n(x_i; f) = f(x_i)$  gesichert.

### Eigenschaften:

- (1°) Die Interpolation nach LAGRANGE ist von hohem theoretischen aber geringen praktischen Wert, weil bei Hinzunahme einer weiteren Stützstelle alle Grundpolynome  $q_i(x)$  neu berechnet werden müssen.
- (2°) Nach dem EE-Satz ist die Formel für alle *Monome*  $f(x) = x^k$ ,  $k = 0 : n$ , exakt:  $\sum_{i=0:n} (x_i)^k q_i(x) = x^k$ ,  $k = 0 : n$ , insbesondere gilt die *Zerlegung der Eins*  $\sum_{i=0:n} q_i(x) = 1$ .
- (3°) Bei *äquidistanten Stützstellen*,  $h = 1/n$ ,  $x_i = x_0 + ih$ ,  $x = x_0 + sh$ ,  $s \in [0, n]$ , vereinfacht sich die Formel (2.6) wegen

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

erheblich zu

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

Diese Darstellung wird zur Konstruktion von numerischen Quadraturformeln verwendet wie auch zur Konstruktion von Näherungsverfahren für Anfangswertprobleme gewöhnlicher Differentialsysteme.

**(d) Interpolation nach Newton** Es sei  $f[x_{j_0}, \dots, x_{j_m}]$  der Höchstkoeffizient von  $p_{j_0, \dots, j_m}(x)$ , dann gilt nach Lemma 2.1

$$f[x_j, \dots, x_{j+m}] = \frac{f[x_{j+1}, \dots, x_{j+m}] - f[x_j, \dots, x_{j+m-1}]}{x_{j+m} - x_j}. \quad (2.9)$$

Diese *dividierten Differenzen* hängen nicht von der Reihenfolge der Indizes ab, weil die zugehörigen Polynome diese Eigenschaft haben. Wählt man die NEWTON-Basis für  $\Pi_n$ ,  $n_0(x) \equiv 1$ ,  $n_j(x) = (x - x_0) \cdots (x - x_{j-1})$ ,  $j = 1 : n$ , dann gilt

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

wegen der für dieses Polynom typischen Rekursionsformel

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

und der Interpolationseigenschaft.

Das TAYLOR-Polynom  $p_n(x; f) = \sum_{i=0:n} \frac{f^{(i)}(x_0)}{i!} (x - x_0)^i$  steht sozusagen am anderen Ende der Skala bei der Approximation mit Polynomen. Mit Hilfe von (2.9) lässt sich eine natürliche Beziehung zwischen den TAYLOR-Koeffizienten und den dividierten Differenzen als Koeffizienten des NEWTON-Polynoms herstellen:

**Lemma 2.2.**  $f[x_i, \dots, x_{i+k}] = \frac{f^{(k)}(\xi)}{k!}, \quad \xi \in [x_i, \dots, x_{i+k}].$

*Beweis.* Es sei  $p_{i,\dots,i+k}(x)$  das NEWTONsche Interpolationspolynom vom Grad  $\leq k$  zu den Knoten  $(x_j, f_j)$ ,  $j = i : i + k - 1$ , und  $(x_{i+k}, f_{i+k}) = (\bar{x}, f(\bar{x}))$ , wobei alle Abszissen  $x_j$  und  $\bar{x}$  verschieden sein sollen. Dann gilt einerseits an der Stelle  $x_{i+k} = \bar{x}$

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

und andererseits die Restgliedformel für  $p_{i,\dots,i+k}(x)$

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

(e) Bei **Hinzunahme der Ableitungen** von  $f$  an den Stützstellen  $x_i$  entstehen Interpolationspolynome nach HERMITE:

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

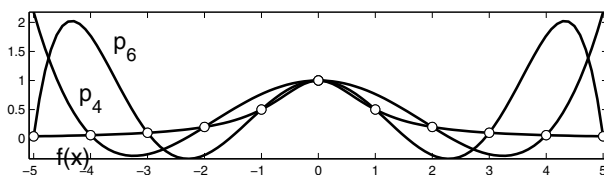
mit den HERMITE-Grundpolynomen

$$\begin{aligned} h_{0,i}(x) &= [1 - 2q'_i(x_i)(x - x_i)]q_i(x)^2 \implies h_{0,i}(x_k) = \delta^i_k, \quad h'_{0,i}(x_k) = 0, \\ h_{1,i}(x) &= (x - x_i)q_i(x)^2 \implies h_{1,i}(x_k) = 0, \quad h'_{1,i}(x_k) = \delta^i_k, \end{aligned}$$

wobei  $q_i(x)$  die LAGRANGE-Grundpolynome sind. Das Restglied hat die gleiche Form wie in (2.5):

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

Die Erhöhung des Grades  $n$  bei einem Interpolationspolynom verbessert i.a. *nicht* die Approximationsgüte, daher ist *segmentweise Interpolation* mit einfachen Polynomen vorzuziehen. Bei globalen Glattheitsforderungen an die zusammengesetzten Polynome ergeben sich dann die *interpolierenden Spline-Funktionen*, vgl. (g).



**Abb. 2.1.** Interpolationspolynom vom Grad  $n = 4, 6$  für  $f(x) = 1/(1+x^2)$

**(f) Approximation mit Beziér-Polynomen** In einem festen – nicht gestückelten – Intervall werden wesentlich bessere Approximationsergebnisse erzielt, wenn die strenge Interpolationsbedingung  $p_n(x_i; f) = f(x_i)$  für Stützpunkte im *Innern* aufgegeben wird. Aus der Zerlegung der Eins

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

ergibt sich die Basis der *Bernstein-Polynome*  $B_i^n(x)$  von  $\Pi_n$  und die allgemeine Darstellung von  $p_n \in \Pi_n$  als *BEZIÉR-Polynom* mit den *BEZIÉR-Punkten*  $b_i$ ,

$$p_{n,\text{bez}}(x) = b_0 B_0^n(x) + \dots + b_n B_n^n(x). \quad (2.11)$$

Die Nullstellen aller  $B_i^n(x)$  liegen auf dem Rand des Einheitsintervalls  $[0, 1]$ , und im Innern existiert genau ein Extremalpunkt (Maximumpunkt). Daher haben die *BERNSTEIN-Polynome* keine Wendepunkte im Einheitsintervall, und durch die Approximation einer Funktion  $f$  mit *BEZIÉR-Polynomen* werden *keine unerwünschten* Wendepunkte eingeschleppt. Allerdings muss man sich auf die *Approximation im Einheitsintervall* beschränken, andernfalls muss umskaliert werden; im Übrigen ist auch hier die *stückweise* Approximation vorzuziehen.

### Eigenschaften:

$$(1^\circ) \quad \sum_{i=0}^n B_i^n(x) = 1, \quad \sum_{i=0}^n \left(\frac{i}{n}\right) B_i^n(x) = x.$$

(2°) Mit den *Vorwärts-Differenzen*  $\Delta b_i = b_{i+1} - b_i$ ,  $\Delta^k b_i = \Delta(\Delta^{k-1} b_i)$  berechnet man die Ableitungen

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

woraus die o.g. wichtige Eigenschaft

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

folgt. Ferner hängen die  $k$ -ten Ableitungen in  $x = 0$  bzw.  $x = 1$  nur von den *BEZIÉR-Punkten*  $b_0, \dots, b_k$  bzw.  $b_{n-k}, \dots, b_n$  ab.

- (3°) Sind  $s - r + 1$  aufeinanderfolgende BEZIÉR-Punkte  $\{b_r, \dots, b_s\}$  für die Abszissen  $x = (i - r)/(s - r)$ ,  $i = r : s$ , gegeben, dann ist

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

das zugehörige BEZIÉR-Polynom vom Grad  $\leq s - r$  (hängt von der Reihenfolge der Punkte ab!). Mit Hilfe des Additionstheorems für Binomialkoeffizienten lässt sich die *lineare Rekursionsformel* von DE CASTELJAU herleiten,

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

die an Stelle der algebraischen Darstellung (2.11) zur *punktweisen* Berechnung von  $p_{n,\text{bez}}(x)$  verwendet wird.

- (4°) Die Punkte  $(x_i, b_i) = (i/n, f(i/n)) \in \mathbb{R}^2$ ,  $i = 0 : n$ , heißen BEZIÉR-Knoten oder ebenfalls BEZIÉR-Punkte. Das BEZIÉR-Polynom hängt an dem zugehörigen BEZIÉR-Polygon wie ein Zirkuszelt an seinen Masten und schmiegt sich immer besser an, wenn der Polynomgrad und entsprechend die Knotenanzahl erhöht wird. Allgemeine Raumkurven ergeben sich, wenn die BEZIÉR-Punkte  $b_i$  in (2.11) durch Vektoren ersetzt werden,

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

- (5°) Die *Approximation mit BEZIÉR-Polynomen* liefert ein fundamentales Ergebnis der *Funktionalanalysis*:

**Satz 2.3.** (WEIERSTRASS) *Es sei  $f \in C[0, 1]$  und*

$$B_n f : x \mapsto \sum_{i=0}^n f \left( \frac{i}{n} \right) B_i^n(x) \text{ mit den BERNSTEIN-Polynomen } B_i^n(x),$$

*dann gilt  $\lim_{n \rightarrow \infty} \max_{0 \leq x \leq 1} |f(x) - B_n f(x)| = 0$ .*

Der Beweis dieses Satzes ist eine einfache Folgerung aus einem überraschenden Ergebnis von KOROVKIN:

**Satz 2.4.** *Es sei  $L_n : C[a, b] \mapsto C[a, b]$  ein Folge von linearen und positiven Operatoren, d.h.*

$$\forall f, g \in C[a, b] \quad \forall t \in [a, b] : f(t) \leq g(t) \implies L_n(f) \leq L_n(g),$$

*und es sei  $f_1(t) = 1$ ,  $f_2(t) = t$ ,  $f_3(t) = t^2$ . Wenn*

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

*gilt, dann folgt  $\forall f \in C[a, b] : \lim_{n \rightarrow \infty} \|L_n f - f\|_\infty = 0$ .*

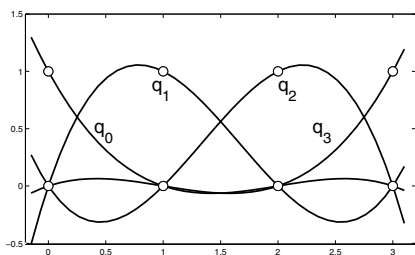
Beweis [Kosmol] § 4.4.5.

*Beweis* von Satz 2.3. Die Operatoren  $B_n$  sind offenbar linear und positiv, und es gilt

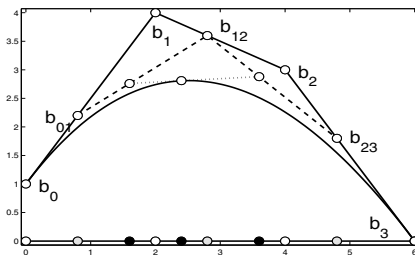
$$B_n(f_1, t) = 1, \quad B_n(f_2, t) = t, \quad B_n(f_3, t) = t^2 + \frac{t - t^2}{n},$$

daher sind die Voraussetzungen von Satz 2.4 erfüllt.  $\square$

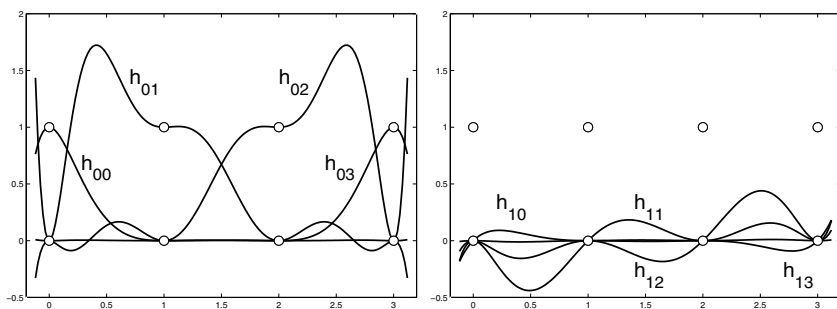
Einen direkten, ebenso interessanten Beweis von Satz 2.3 findet man in [Yosida] § 0.2.



**Abb. 2.2.** LAGRANGE-Grundpolynome,  $n = 3$

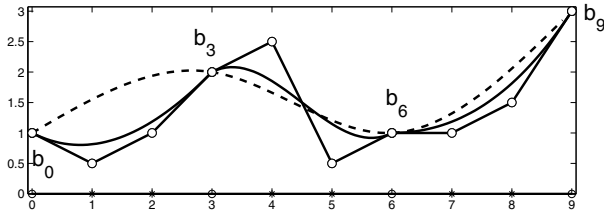


**Abb. 2.3.** BEZIÉR-Polynom,  $n = 3$  mit Umskalierung



**Abb. 2.4.** HERMITE-Grundpolynome,  $n = 3$

**(g) Interpolationssplines** Eine (stetige) BEZIÉR-Kurve besteht stückweise aus BEZIÉR-Polynomen, die jeweils an den Enden ihres Definitionsintervalls die Interpolationseigenschaft haben. Wir betrachten den Spezialfall einer BEZIÉR-Kurve im Intervall  $\mathcal{I} = [0, n \cdot m]$ ,  $m \in \mathbb{N}$ . In den Teilintervallen  $\mathcal{I}_k = [n(k-1), nk]$ ,  $k = 1 : m$ , soll die Kurve aus BEZIÉR-Polynomen vom Grad  $n$  mit den BEZIÉR-Punkten  $b_{nk}, \dots, b_{n(k+1)}$  bestehen, sie soll also an den Stellen  $nk$  vorgegebene Werte  $f(nk)$  annehmen ( $n$  fest) (Abb. 2.5).

Abb. 2.5. BEZIÉR-Kurve und Spline,  $m = n = 3$ 

**Definition 2.1.** (1°) Eine segmentierte stetige Kurve mit Polynomsegmenten vom Grad  $\leq n$  heißt (Polynom-)Spline, wenn sie insgesamt  $(n-1)$ -mal stetig differenzierbar ist. Für  $n = 3$  heißt der Spline kubischer Spline.

(2°) Es sei  $\mathcal{I} = [a, b]$ , und es sei durch  $a = x_0 < x_1 < \dots < x_m = b$  eine Segmentierung (Unterteilung)  $\Delta_m$  von  $\mathcal{I}$  gegeben, dann ist

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

der Vektorraum der kubischen Splines.

Die Dimension von  $S_3$  ist  $m + 3 = (m + 1) + 2$ , es sind also zwei Bedingungen frei.

Für  $k \in \mathbb{N}_0$  sei

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

$q_k(t, x)$  hat  $k - 1$  stetige Ableitungen nach beiden Argumenten und die  $k$ -te Ableitung macht einen Sprung der Höhe  $k!$  bzw.  $(-1)^k k!$ .

**Satz 2.5.** (Basis-Splines, B-Splines) Die Menge  $S_n(\Delta_m)$  ist ein linearer Raum der Dimension  $m + n$ . Die Elemente

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

bilden eine Basis von  $S_d(\Delta_n)$ .

Beweis. [Hämmerlin], S. 246.

Wir betrachten nun den Fall  $n = 3$ . Fasst man  $s \in S_3$  als BEZIÉR-Kurve auf, so gilt für die BEZIÉR-Punkte bei Abstand  $x_{i+1} - x_i = 1$ :

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

Hieraus folgt

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

und dann durch Addition der linken und der rechten Gleichungen

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

Die Zahlen  $b_{3k+1}$  und  $b_{3(k+1)-1} = b_{3k+2}$  dreiteilen also die Strecke zwischen  $d_k$  und  $d_{k+1}$ .

Ferner gilt für alle *inneren* Punkte  $b_{3k}$ ,  $k = 1 : m - 1$ , nach (2.12) und (2.13)

$$6b_{3k} = 3b_{3k-1} + 3b_{3k+1} = d_{k-1} + 4d_k + d_{k+1}. \quad (2.14)$$

Zusammen mit (2.13) für  $b_1$  und  $b_{3m-1}$ , d.h. für  $k = m$  und  $k = 0$ ,

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

ergibt sich das folgende lineare Gleichungssystem für den Vektor mit den unbekannten Koeffizienten  $[d_0, \dots, d_m]^T$  (DEBOOR-Punkte), falls die Daten auf der rechten Seite vorgegeben sind:

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

Zu  $\{d_0, \dots, d_m, b_0, b_{3m}\}$  existiert nach obiger Konstruktion genau ein Spline  $s \in S_3(\Delta)$  (Abb. 2.6). Er heißt *kubischer Interpolationsspline* wegen  $s(x_k) = b_{3k} = f_k$  für  $k = 0 : m$ .

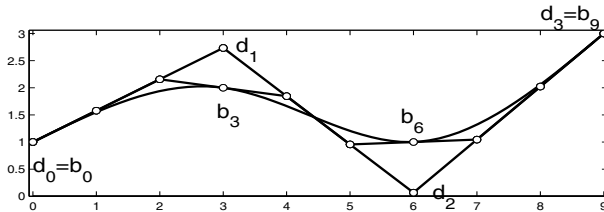


Abb. 2.6. Interpolationsspline  $m = n = 3$

**Berechnung** Es sei  $f_k = b_{3k}$ ,  $k = 0 : m$  gegeben und der interpolierende Spline  $s \in S_3(\Delta_m)$  in  $[a, b] = [0, m]$ ,  $x_{i+1} - x_i = 1$  gesucht.

(1°) Es sei  $f'(a)$ ,  $f'(b)$  gegeben. Berechne  $b_1$ ,  $b_{3m-1}$  aus

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

berechne  $d_0, \dots, d_m$  aus (2.15), berechne  $b_{3k+1}, b_{3k+2}$  aus (2.13),  
 berechne  $s(x)$  in  $[x_k, x_{k+1}]$  als BEZIÉR-Polynom nach DE CASTELJAU,

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

in den lokalen Koordinaten  $\xi \in [0, 1]$ .

(2°) Wenn  $s''(a) = s''(b) = 0$  gesetzt wird, ergeben sich *natürliche Splines*,  $s \in N_3(\Delta_m)$ . Zu ihrer Berechnung wird  $d_0 = b_0$  und  $d_m = b_{3m}$  fest vorgegeben, dann gilt mit  $n = 3$

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

Berechne  $(d_1, \dots, d_{m-1})^T$  aus (2.15) ohne die erste und letzte Zeile (weil  $d_0$  und  $d_m$  fest)

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

Es gilt  $s''(x_k) = 6(b_{3k} - d_k)/h^2$ ,  $k = 1 : m-1$ ,  $h = x_{i+1} - x_i$  konstant, daher erfüllen die „Momente“  $s''(x_k)$  ein ähnliches Gleichungssystem wie die Werte  $d_k$ .

Wenn die exakte Krümmung  $\kappa(x) = f''(x)/(1 + f'(x)^2)^{3/2}$  von  $f : x \mapsto f(x)$  näherungsweise durch  $f''(x)$  ersetzt wird, erweisen sich die natürlichen kubischen Interpolationssplines als Biegelinien („Straklatten“):

**Satz 2.6.** *Es sei  $s \in N_3(\Delta_m)$ , d.h. ein natürlicher Spline und  $\Delta_m$  beliebig, mit  $s(x_i) = f_i$ ,  $i = 0 : m$ . Dann gilt*

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

*Beweis.* Es sei  $g$  eine beliebige Vergleichsfunktion mit der gleichen Interpolationseigenschaft, dann gilt  $g''(x)^2 = (s''(x) + g''(x) - s''(x))^2$ , also

$$\int_a^b g''(x)^2 dx = \int_a^b (s'')^2 dx + 2 \int_a^b s''(g'' - s'') dx + \int_a^b (g'' - s'')^2 dx. \quad (2.17)$$

Mit partieller Integration folgt für den gemischten Term

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

weil nach Voraussetzung  $g(x_i) = s(x_i) = f_i$ . Also folgt die Behauptung, wenn die Randterme verschwinden, d.h., wenn gilt

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

Diese Bedingung ist u.a. in den folgenden drei wichtigen Fällen erfüllt:

- (1°) wenn  $s''(a) = 0 = s''(b)$  (natürlicher Spline),
- (2°) wenn  $g'(a) = s'(a) = f'(a)$  fest,  $g'(b) = s'(b) = f'(b)$  fest,
- (3°) wenn  $s, g$  periodisch sind mit der Periode  $b - a$ . □

Bei einem Intervall  $[a, b]$  anstatt  $[0, m]$  muss umskaliert werden; ebenso ist eine Abänderung notwendig bei nicht äquidistanten Stützstellen. Auf eine weitergehende Behandlung der Spline-Funktionen soll aber an dieser Stelle verzichtet werden; vgl. jedoch SUPPLEMENT\chap02.

## 2.2 Orthogonale Polynome

Es sei  $\overline{\Pi}_n$  die Menge der Polynome  $p_n \in \Pi_n$  vom exakten Grad  $n$  und Höchstkoeffizient Eins.

**(a) Konstruktion** Es sei  $-\infty \leq a < b \leq \infty$ , und es sei  $\omega : [a, b] \rightarrow \mathbb{R}_+$  eine nichtnegative *Gewichtsfunktion* mit den folgenden Eigenschaften:

**Voraussetzung 2.1.** Die Momente  $m_k := \int_a^b \omega(x) x^k dx$ ,  $k \in \mathbb{N}_0$ , existieren (ev. als uneigentliche Integrale), und es ist  $m_0 > 0$ .

Zwei Polynome  $p$  und  $q$  heißen dann *orthogonal* (bezüglich des gewählten Integrationsintervalls und der Gewichtsfunktion  $\omega$ ), wenn

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

**Satz 2.7.** (*Existenz und Konstruktion*) (1°) Unter Voraussetzung 2.1 gilt

$\forall i \in \mathbb{N}_0 \exists! p_i \in \overline{\Pi}_i : i \neq k \implies (p_i, p_k) = 0.$

(2°) Die Orthogonalpolynome sind eindeutig bestimmt durch die dreigliedrige Rekursionsformel (mit  $xp : x \mapsto xp(x)$ )

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

Beweis durch GRAM-SCHMIDT-Orthogonalisierung [Stoer], siehe auch SUPPLEMENT\chap02.

Natürlich gilt für orthogonale Polynome  $p_n \in \overline{\Pi}_n$  stets  $p \in \Pi_{n-1} \implies (p, p_n) = 0$ , weil die orthogonalen Polynome als linear unabhängige Funktionen eine Basis des  $\Pi_{n-1}$  bilden. Im restlichen Teil dieses Abschnitts betrachten wir orthogonale Polynome  $p_n$  gemäß Satz 2.7.

**Satz 2.8.** Die Wurzeln  $x_i$  von  $p_n$  sind reell, einfach und liegen im offenen Intervall  $(a, b)$ .

*Beweis.* Es seien  $x_1, \dots, x_k$  alle Wurzeln von  $p_n$ , die in  $(a, b)$  liegen und ungerade Vielfachheit haben, dann wechselt also  $p_n$  genau an diesen Stellen das Vorzeichen. O.B. sei etwa

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

dann wechselt  $p_n(x)q(x)$  das Vorzeichen in  $(a, b)$  *nicht*, und es gilt also  $(p_n, q) \neq 0$ . Daher muss gelten  $\text{Grad } q = k = n$ , sonst ergibt sich ein Widerspruch zu obigen Folgerung aus Satz 2.7.  $\square$

**(b) Die Formeln von Rodriguez** Zur expliziten Konstruktion von Orthogonalpolynomen  $p_n \in \Pi_n$  beachten wir die allgemeine Orthogonalitätsbedingung

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

und machen den Ansatz

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

Dann ergibt sich aus (2.19) die Differentialgleichung

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

Die  $n$ -malige partielle Integration von  $\int_a^b u_n^{(n)}(x) q_{n-1}(x) dx = 0$  ergibt

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

Diese Gleichung ist sicher erfüllt für die Randbedingungen

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

**Satz 2.9.** *Unter Voraussetzung 2.1 hat das Randwertproblem (2.20), (2.21) stets eine Lösung  $u_n$  und  $p_n := u_n/\omega$  ist ein Polynom vom Grad  $n$ .*

Beweis [Szegoe].

Das obige Randwertproblem hat  $2n$  Randbedingungen für eine Differentialgleichung der Ordnung  $2n+1$ ; es ist also eine Bedingung zur Normierung frei.

**Beispiel 2.1.** LEGENDRE-Polynome:  $(a, b) = (-1, 1)$ ,  
 $\omega(x) \equiv 1$ ,  $u_n^{(2n+1)} = 0$ ,  $u_n^{(i)}(\pm 1) = 0$ ,  $i = 0 : n-1$ .

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

Die Konstanten  $\gamma_n$  werden in unterschiedlicher Weise festgelegt. (Abb. 2.7).

**Beispiel 2.2.** JACOBI-Polynome:  $(a, b)$  endlich,  $\omega(x) = (x-a)^\alpha (b-x)^\beta$ ,  
 $\alpha > -1$ ,  $\beta > -1$ .

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

Insbesondere ergeben sich für  $(a, b) = (0, 1)$  und  $(\alpha, \beta) = (0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ ,  $(1, 1)$  verschobene LEGENDRE-Polynome, die später noch gebraucht werden:

$$\begin{aligned} p_{1,n}(x) &= \frac{d^n}{dx^n} (x^n (1-x)^n), & p_{2,n}(x) &= \frac{1}{x} \frac{d^n}{dx^n} (x^{n+1} (1-x)^n) \\ p_{3,n}(x) &= \frac{1}{1-x} \frac{d^n}{dx^n} (x^n (1-x)^{n+1}), & p_{4,n}(x) &= \frac{1}{x(1-x)} \frac{d^n}{dx^n} x^{n+1} (1-x)^{n+1}. \end{aligned} \quad (2.22)$$

**Beispiel 2.3.** TSCHEBYSCHJEFF-Polynome  $T_n(x)$  mit  $(a, b) = (-1, 1)$ ,  
 $\omega(x) = (1-x^2)^{-1/2}$  sind ebenso wie die LEGENDRE-Polynome spezielle JACOBI-Polynome. Wegen der besonderen Form der Gewichtsfunktion wird bei der Entwicklung einer Funktion nach diesen Polynomen der Fehler an den Intervallenden besonders stark bewichtet. Aus der ursprünglichen Orthogonalitätsbedingung

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

ergibt sich durch Substitution  $x = \cos \varphi$  die Orthogonalitätsbedingung

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

Wegen

$$\cos(n+1)\varphi + \cos(n-1)\varphi = 2 \cos \varphi \cos n\varphi \quad (2.24)$$

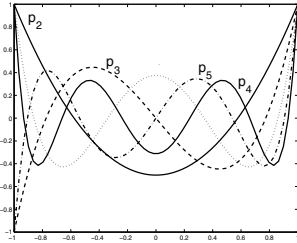
für  $n \in \mathbb{N}$  ist  $\cos n\varphi$  ein Polynom in  $\cos \varphi$ , und  $(\cos \varphi)^k$  ist eine Linearkombination von  $\cos j\varphi$ ,  $j = 0 : k$ , also

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

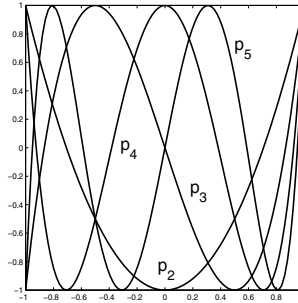
Daher gilt (2.23) genau dann, wenn

$$\int_0^\pi T_n(\cos \varphi) \cos(j\varphi) d\varphi = 0, \quad j = 0 : n-1,$$

$$\Rightarrow T_n(\cos \varphi) = \cos(n\varphi) \Rightarrow \boxed{T_n(x) = \cos(n \arccos x)}, \quad n = 0, 1, \dots$$



**Abb. 2.7.** LEGENDRE-Polynome,  
 $n = 2 : 6$



**Abb. 2.8.** TSCHEBYSCHJEFF-Polynome,  
 $n = 2 : 5$

**(c) Minimaleigenschaft von Tchebyscheff-Polynomen** Die Rekursionsformel (2.24) zeigt, dass  $T_n(x) = \cos(n \arccos(x))$  den Höchstkoeffizient  $2^{n-1}$  hat.

**Satz 2.10.** *Es sei  $p_n(x)$  ein Polynom vom Grad  $\leq n$  mit dem Höchstkoeffizienten  $2^{n-1}$ . Dann gibt es mindestens ein  $x \in [-1, 1]$  mit  $|p_n(x)| \geq 1$ .*

*Beweis.* Es gelte  $|p_n(x)| < 1$  für alle  $x \in [-1, 1]$ .  $T_n(x)$  nimmt an seinen  $n+1$  Extremalstellen  $x_i = \cos(i\pi/n)$ ,  $i = 0 : n$ , im Intervall  $[-1, 1]$  alternierend die Werte  $\pm 1$  an. Daher ist  $T_n(x) - p_n(x)$  an den Extremalstellen abwechselnd

positiv oder negativ. D.h.  $T_n(x) - p_n(x)$  hat in  $(-1, 1)$  mindestens  $n$  Nullstellen. Wegen gleicher Höchstkoeffizienten ist aber  $T_n(x) - p_n(x)$  ein Polynom vom Grad  $\leq n - 1$ . Daraus folgt  $T_n(x) - p_n(x) \equiv 0$  im Widerspruch zur Voraussetzung.  $\square$

**Folgerung 2.1.** *Es sei  $q_n$  ein Polynom vom Grad  $n$  mit Höchstkoeffizient  $a_n$ . Dann gibt es ein  $x \in [-1, 1]$  mit  $|q_n(x)| \geq a_n/2^{n-1}$ .*

*Beweis.* Es sei  $a_n \neq 0$  und  $q_n^*(x) = q_n(x)2^{n-1}/a_n$ . Das Polynom  $q_n^*$  hat den Höchstkoeffizient  $2^{n-1}$ , daher gilt  $|q_n^*(x)| \geq 1$  nach Satz 2.10 für mindestens ein  $x_0 \in [-1, 1]$ . Daraus folgt  $|q_n(x_0)| = |q_n^*(x_0)a_n/2^{n-1}| \geq a_n/2^{n-1}$ .  $\square$

Für ein beliebiges Polynom  $q_n(x)$  – insbesondere auch für ein TAYLOR-Polynom – gibt es eine eindeutige Entwicklung nach TSCHEBYSCHJEFF-Polynomen,

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

weil diese Polynome wegen der Orthogonalität linear unabhängig sind.

**Satz 2.11.** *Sind  $S_n(x) = \sum_{i=0}^n c_i T_i(x)$  die Partialsummen einer Entwicklung nach TSCHEBYSCHJEFF-Polynomen, dann gilt*

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

wobei  $p_n$  ein beliebiges Polynom vom Grad  $\leq n$  ist.

*Beweis.* Es gilt  $S_n(x) - S_{n-1}(x) = c_n T_n(x)$ , also  $|S_n(x) - S_{n-1}(x)| \leq |c_n|$ ,  $-1 \leq x \leq 1$ . Für ein beliebiges Polynom  $p_{n-1}$  vom Grad  $n - 1$  hat  $S_n - p_{n-1}$  den Höchstkoeffizient  $c_n 2^{n-1}$ . Also gilt nach der Folgerung zu Satz 2.10 für mindestens ein  $x \in [-1, 1]$

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

Grob gesagt nehmen also bei einer Entwicklung nach TSCHEBYSCHJEFF-Polynomen die Koeffizienten betragsweise am schnellsten ab.

## 2.3 Numerische Integration

Bekanntlich ist das Integrieren eine Kunst und das Differenzieren Handwerk, vom numerischen Standpunkt gesehen kehrt sich aber die Sachlage in gewisser Weise um. Das Integrieren ist ein glättender Prozess, was sich in der numerischen Approximation vorteilhaft auswirkt, die Ableitung dagegen muss stets durch einen Differenzenquotienten ersetzt werden, bei dem im Zähler und im

Nenner Subtraktion annähernd gleich großer Zahlen auftritt, also die gefürchtete Auslöschung führender Ziffern. Es sei aber an dieser Stelle erwähnt, dass eine asymptotische Entwicklung im Sinne von §2.4(c) zu verblüffend genauen Resultaten führen kann; vgl. [Rutishauser].

Wegen der Indizierung ohne Null bei MATLAB<sup>®</sup> und wegen der GAUSSschen Quadraturformeln ist es hier sinnvoll, mit  $n$  anstatt  $n + 1$  Stützstellen zu arbeiten, damit die einzelnen Regeln besser miteinander verglichen werden können.

**(a) Quadratur nach Lagrange** Der Rechenaufwand einer numerischen Integrationsformel hängt von der Anzahl der Stützstellen ab, an denen der Integrand  $f$  ausgewertet werden muss. Wir gehen von einem Interpolationspolynom  $p_{n-1}(x)$  vom Grad  $n - 1$  und LAGRANGE-Typ aus, d.h. nach (2.6) in leicht modifizierter Form

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

und erhalten nach Integration über das Intervall  $(a, b)$

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

wobei die  $n$  Stützstellen  $x_i$  wieder *alle verschieden* sein sollen. Im Übrigen ist ihre Wahl jedoch frei, insbesondere können sie auch außerhalb des Integrationsintervalls liegen. In diesem Abschnitt soll aber gelten

$$a \leq x_1 < \dots < x_{n-1} < x_n \leq b.$$

Eine Quadraturformel hat den *Genauigkeitsgrad*  $N$ , wenn genau alle Polynome vom Grad  $\leq N$  exakt integriert werden. Der Genauigkeitsgrad einer LAGRANGE-Formel (2.26) mit  $n$  Stützstellen (!) ist mindestens  $N = n - 1$ . Der *maximale Genauigkeitsgrad* einer solchen Formel ist  $N = 2n - 1$ . Setzt man nämlich das Polynom  $f(x) = \prod_{i=1}^n (x - x_i)^2 \in \Pi_{2n}$  ein, so ist  $I_n(f) = 0$  und das exakte Integral  $I(f) > 0$ .

Die NEWTON-COTES-Formeln sind ebenfalls vom LAGRANGE-Typ (2.25), aber die Stützstellen werden äquidistant gewählt,  $x_i = a + (i - 1)h$ ,  $h = (b - a)/(n - 1)$ ,  $n \geq 2$ . Mit der Substitution  $x = a + (s - 1)h$  folgt nach § 2.1

$$\begin{aligned} q_i(x) &= q_i(a + (s - 1)h) =: \varphi_i(s) = \prod_{j=1, j \neq i}^n \frac{s - j}{i - j} \in \Pi_{n-1}(s), \quad s \in [1, n], \\ \alpha_i &:= \int_a^b q_i(x) dx = \int_1^n q_i(a + (s - 1)h) \frac{dx}{ds} ds = h \int_1^n \varphi_i(s) ds = h \beta_i, \\ I_n(f) &= \sum_{i=1}^n f(x_i) \alpha_i = h \sum_{i=1}^n f(x_i) \beta_i. \end{aligned}$$

Die Gewichte  $\beta_i$  sind jetzt *rationale* Zahlen, die nur noch von der Stützstellenzahl  $n$  abhängen und sich damit bei Änderung der Intervallgrenzen nicht mehr ändern; für  $f(x) \equiv 1$  ergibt sich  $\sum_{i=1}^n \beta_i = n - 1, n \geq 2$ .

**Beispiel 2.4.** *Mittelpunktregel* (ein Knoten):

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

(*Sehnen-*)Trapezregel ( $n = 2$  Knoten,  $h = b - a$ ):

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

KEPLERsche Fassregel ( $n = 3$  Knoten,  $h = (b-a)/2$ ):

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

Die KEPLER-Regel wird im angelsächsischen Sprachgebrauch SIMPSON-Regel genannt. Mittelpunktregel und KEPLER-Regel haben den gleichen Genauigkeitsgrad  $n$ . Allgemein ist aus Symmetriegründen der Grad  $n$  statt  $n-1$  für ungerades  $n$  bei NEWTON-COTES-Formeln.

Über die – von Formel zu Formel verschiedenen – Zwischenstellen  $\xi \in (a, b)$  ist i.A. nichts Näheres bekannt, deswegen scheidet i.d.R. eine Abschätzung des Restgliedes  $R_n(f)$  in

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

über das CAUCHY-Restglied – vgl. Satz 2.2 – aus. Mit dem FÖPPL-Symbol  $(x-t)_+^N := \max\{(x-t)^N, 0\}$  gilt aber das folgende Resultat von PEANO, vgl. [Stoer]:

**Satz 2.12.** *Wenn die Quadraturformel (2.26) mit  $n$  Knoten den Genauigkeitsgrad  $N$  hat, dann gilt für alle  $f \in C^{N+1}[a, b]$*

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

Dabei ist  $R_n(h_t)$  der Fehler für die Funktion  $h_t : x \mapsto (x-t)_+^N$  anstatt  $f$ . In vielen Fällen wie z.B. bei den NEWTON-COTES-Formeln, wechselt  $K_n(t)$  im Integrationsintervall  $(a, b)$  nicht das Vorzeichen. Dann ergibt der Mittelwertsatz der Integralrechnung

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

Setzt man hier für  $f$  die spezielle Funktion  $\varphi : x \mapsto x^{N+1}$  ein, dann folgt

$$R_n(\varphi) = (N+1)! \int_a^b K_n(t) dt \implies \int_a^b K_n(t) dt = R_n(\varphi)/(N+1)!. \quad (2.29)$$

**Fazit:** Wenn die Integrationsformel (2.26) den Genauigkeitsgrad  $N$  hat und  $K_n(t)$  im Integrationsintervall nicht das Vorzeichen wechselt, dann ergeben (2.28) und (2.29) für den Fehler

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

(Restglied nach PEANO);  $R_n(\varphi)$  kann aber stets exakt berechnet werden!

**(b) Summierte Quadraturformeln** Wie schon in § 2.1 erwähnt, wird die Approximation von  $f$  durch ein Interpolationspolynom i.d.R. nicht besser, wenn der Polynomgrad  $n$  erhöht wird. Man bleibt daher bei Polynomen niederen Grades und unterteilt stattdessen das Integrationsintervall.

Die resultierenden *summierten Quadraturformeln* sind sogar bei stetigem Integranden  $f$  beliebig genau in Abhängigkeit von der Stützstellenanzahl. Mit  $x_i = a + ih$ ,  $i = 0 : m$ ,  $h = (b-a)/m$ ,  $m \in \mathbb{N}$  erhalten wir z.B. aus der Sehnentrapezregel die wichtige *summierte Sehnentrapezregel*

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

und aus der KEPLER-Regel die SIMPSON-Regel

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

Die Formel (2.31) sticht noch durch eine besondere Eigenschaft hervor:

**Lemma 2.3.** Ist  $f \in C^\infty(\mathbb{R})$   $(b-a)$ -periodisch, dann gilt

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

Bei glatten periodischen Funktionen konvergiert also die summierte Sehnentrapezregel schneller als jede Potenz der Schrittweite  $h$  (!).

**(c) Quadratur nach Gauß** Wird an Stelle des LAGRANGE-Polynoms das Interpolationspolynom in der HERMITESchen Form integriert – vgl. § 2.1(e), so erhält man Quadraturformeln vom Typ

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

mit

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

und den LAGRANGE-Grundpolynomen  $q_i(x) \in \Pi_{n-1}$ . Die Formel hat den Genauigkeitsgrad  $N = 2n - 1$  bei  $n$ -maliger Auswertung von  $f$  und  $n$ -maliger Auswertung der Ableitung von  $f$ .

Werden nun die Stützstellen  $x_i$ ,  $i = 1 : n$  als Nullstellen von Orthogonalpolynomen  $p_n(x) \in \Pi_n$  bezüglich des Skalarprodukts  $(f, g) = \int_a^b f(x)g(x) dx$  gewählt, dann folgt aus § 2.2

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

wegen  $(x - x_i)q_i(x) = p_n(x)$  und  $q_i(x) \in \Pi_{n-1}$ . Außerdem gilt dann  $\int_a^b h_{0,i}(x) dx = \int_a^b q_i(x)^2 dx$ , daher ergeben sich bei dieser Wahl aus (2.32) und (2.33) Quadraturformeln

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

mit maximalem Genauigkeitsgrad  $N = 2n - 1$  bei  $n$  Stützstellen.

Für eine allgemeine Gewichtsfunktion  $\omega(x)$  mit den Eigenschaften aus § 2.2 fassen wir das Ergebnis in dem folgenden Satz zusammen:

**Satz 2.13.** (Quadratur nach GAUSS) *Es seien  $p_n \in \overline{\Pi}_n$  Orthogonalpolynome bezüglich des Skalarprodukts  $(f, g) := \int_a^b \omega(x)f(x)g(x) dx$ , es seien  $x_1, \dots, x_n$  die Nullstellen von  $p_n$ , und es sei*

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

(1°) *Die Matrix  $A$  ist regulär.*

(2°) *Ist  $\underline{b} = A^{-1}\underline{c}$  und  $\underline{b} = [\beta_1, \dots, \beta_n]^T$ , dann gilt*

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

*d.h. die Quadraturformel*

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

*hat den maximalen Genauigkeitsgrad  $N = 2n - 1$ .*

(3°) Für das Restglied in (2.36) gilt

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

(4°) Wenn umgekehrt (2.35) gilt, dann sind die Stützstellen  $x_i$  die Nullstellen der Orthogonalpolynome  $p_n(x)$ , und es gilt  $A\mathbf{b} = \mathbf{c}$  mit  $\mathbf{b} = [\beta_1, \dots, \beta_n]^T$ .

(5°) Wenn eine Formel (2.36) den Genauigkeitsgrad  $N \geq n - 2$  hat, sind die Gewichte  $\beta_i$  positiv.

Beweis [Stoer].

**(d) Suboptimale Quadraturformeln** sind ein wichtiges Hilfsmittel bei der Konstruktion von impliziten RUNGE-KUTTA-Verfahren maximaler Ordnung im nächsten Abschnitt; dazu sei

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

**Satz 2.14.** Für  $\delta, \varepsilon \in \{0, 1\}$  existiert eine eindeutige Quadraturformel

$$\int_0^1 f(t) dt \approx \delta \beta_0 f(0) + \mathbf{b}^T F(\mathbf{x}) + \varepsilon \beta_{n+1} f(1) \quad (2.37)$$

mit Genauigkeitsgrad  $\tilde{N} = 2n + \delta + \varepsilon - 1$ , der für diesen Formeltyp maximal ist.

Wähle die GAUSS-Gewichte und -Knoten nach Satz 2.13 bez. der Gewichtsfunktion  $\omega^*(t) = t^\delta(1-t)^\varepsilon$  in  $[0, 1]$  und setze  $\mathbf{b} = [\tilde{\beta}_i/\omega^*(x_i)]_{i=1}^n$ . Dann ist die Regel optimal für  $(\delta, \varepsilon) = (0, 0)$ . Für  $(\delta, \varepsilon) = (1, 0)$  oder  $(\delta, \varepsilon) = (0, 1)$  ergeben sich die übrigen Gewichte aus  $1 = \delta \beta_0 f(0) + \mathbf{b}^T \mathbf{e} + \varepsilon \beta_{n+1} f(1)$ . Für  $(\delta, \varepsilon) = (1, 1)$  ergeben sich die beiden übrigen Gewichte aus

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

Ein Vergleich mit den verschobenen LEGENDRE-Polyomen in (2.22) zeigt, dass die Stützstellen  $x_1, \dots, x_n$  einer Regel (2.37) mit insgesamt  $n$  Knoten jeweils die Wurzeln der folgenden Polynome sind:

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

Beweis von Satz 2.14 siehe SUPPLEMENT\chap02a. Ein Programm zur Berechnung der Knoten und Gewichte in allen vier Fällen findet man in KAPITEL02\SECTION\_1\_2\_3. Zur Integration über das Intervall  $(a, b)$  müssen Knoten und Gewichte umskaliert werden gemäß  $\hat{x}_i = a + (b-a)x_i$ ,  $\hat{b}_i = (b-a)b_i$ ,  $i = 1 : n$ .

**Beispiel 2.5.** GAUSS-Quadratur mit LEGENDRE-Polynomen:

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

**Tabelle 2.1.** GAUSS-LEGENDRE-Formeln mit  $n$  Stützstellen

| n | $x_i$                                                                                             | $\beta_i$                                                                                                       |
|---|---------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| 2 | $\pm \frac{1}{3}\sqrt{3}$                                                                         | 1                                                                                                               |
| 3 | 0<br>$\pm \frac{1}{5}\sqrt{15}$                                                                   | $\frac{8}{9}$<br>$\frac{5}{9}$                                                                                  |
| 4 | $\pm \frac{1}{35} [525 - 70\sqrt{30}]^{1/2}$<br>$\pm \frac{1}{35} [525 + 70\sqrt{30}]^{1/2}$      | $\frac{1}{2} + \frac{1}{36}\sqrt{30}$<br>$\frac{1}{2} - \frac{1}{36}\sqrt{30}$                                  |
| 5 | 0<br>$\pm \frac{1}{25} [245 - 14\sqrt{70}]^{1/2}$<br>$\pm \frac{1}{25} [245 + 14\sqrt{70}]^{1/2}$ | $\frac{128}{225}$<br>$\frac{161}{450} + \frac{13}{900}\sqrt{70}$<br>$\frac{161}{450} - \frac{13}{900}\sqrt{70}$ |

Diese Formeln gelten für das Integrationsintervall  $[a, b] = [-1, 1]$ . Bei der Transformation auf ein Intervall  $[a', b']$  müssen die Gewichte und die Stützstellen transformiert werden:

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

Zum Beispiel müssen bei der Transformation auf das Einheitsintervall  $[0, 1]$  die Gewichte halbiert und die Stützstellen  $x'_i = (1 + x_i)/2$  verwendet werden.

**(e) Baryzentrische Koordinaten** dienen zur vereinfachten Darstellung von Polynomen und ihrer Quadratur in Dreiecken oder allgemeiner in  $n$ -Simplices im  $\mathbb{R}^n$ . Wir beschränken uns hier auf die Ebene und betrachten ein beliebiges Dreieck  $T$  im kartesischen  $(x, y)$ -Koordinatensystem mit den Ecken  $P_i(x_i, y_i)$ ,  $i = 1, 2, 3$ , die im Gegenuhrzeigersinn numeriert sind. Der doppelte Flächeninhalt

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

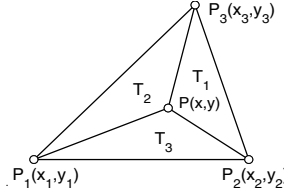
( $x_{21} = x_2 - x_1$  usw.) ist dann positiv, solange  $T$  nicht degeneriert ist. Mit den Bezeichnungen aus Abb. 2.9 sind die (dimensionslosen) *baryzentrischen* oder *Schwerpunktkoordinaten* für  $0 \leq \zeta_i \leq 1$  definiert durch

$$\zeta_i = \frac{\text{Fläche von } T_i}{\text{Fläche von } T}, \quad i = 1, 2, 3.$$

Es gilt also

$$P_1 \simeq (1, 0, 0), \quad P_2 \simeq (0, 1, 0), \quad P_3 \simeq (0, 0, 1)$$

und  $\zeta_1 + \zeta_2 + \zeta_3 = 1$ , damit sind die baryzentrischen Koordinaten *linear abhängig*.



**Abb. 2.9.** Baryzentrische Koordinaten

Der Zusammenhang zwischen kartesischen und baryzentrischen Koordinaten wird für  $(x, y) \in T$  durch die Flächenformel hergestellt:

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

Entwickelt man jeweils nach der ersten Zeile und teilt durch  $2|T|$ , dann ergibt sich für ein beliebiges kartesisches Koordinatensystem

$$\begin{aligned} \zeta_1 &= \frac{1}{2|T|} [(x_2 y_3 - x_3 y_2) - y_{32} x + x_{32} y] \\ \zeta_2 &= \frac{1}{2|T|} [(x_3 y_1 - x_1 y_3) + y_{31} x - x_{31} y] \\ \zeta_3 &= \frac{1}{2|T|} [(x_1 y_2 - x_2 y_1) - y_{21} x + x_{21} y] \end{aligned} \quad (2.40)$$

(man beachte die zyklische Permutation der Indizes modulo 3). Diese Beziehungen gelten allgemein ohne dass der Ursprung des KOS im Zentrum des Dreiecks liegen muss. Sie werden verschiedentlich angewendet etwa bei der Berechnung von Ableitungen, z.B.  $\partial \zeta_1 / \partial x = y_{23} / (2|T|)$  etc.. Auflösen zweier Gleichungen in (2.40) nach  $x$  und  $y$  liefert die Beziehung zwischen kartesischen und baryzentrischen Koordinaten

$$\begin{aligned} 1 &= \zeta_1 + \zeta_2 + \zeta_3 \\ x &= x_1 \zeta_1 + x_2 \zeta_2 + x_3 \zeta_3 \\ y &= y_1 \zeta_1 + y_2 \zeta_2 + y_3 \zeta_3 \end{aligned} \quad (2.41)$$

Im *Einheitsdreieck*  $S(\xi, \eta)$  mit den Ecken  $Q_1(0, 0)$ ,  $Q_2(1, 0)$ ,  $Q_3(0, 1)$  gilt die Beziehung

$$\zeta_1 = 1 - \xi - \eta, \quad \zeta_2 = \xi, \quad \zeta_3 = \eta, \quad (2.42)$$

und

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

Daraus folgt die für allgemeine Dreiecke  $T \subset \mathbb{R}^2$  gültige Formel von HOLLAND und Bell (1969)

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

durch Substitution [Bell]. Ihre direkte Verallgemeinerung auf Tetraeder  $T \subset \mathbb{R}^3$  mit Volumen  $|T|$  lautet

$$\int_T \zeta_1^m \zeta_2^n \zeta_3^p \zeta_4^q dx dy dz = 6|T| \frac{m!n!p!q!}{(m+n+p+q+3)!}. \quad (2.45)$$

**Beispiel 2.6.** [Ciarlet79] Es seien  $\underline{x}_i \in \mathbb{R}^n$ ,  $i = 1 : n+1$ , die Ecken eines  $n$ -Simplex im  $\mathbb{R}^n$ , z.B. eines Dreieckes  $T$  im  $\mathbb{R}^2$  oder eines Tetraeders im  $\mathbb{R}^3$ . Es seien  $\underline{x}_{ij} = (\underline{x}_i + \underline{x}_j)/2$  für  $i < j$  die Mittelpunkte der Kanten  $\underline{x}_{ijk} = (\underline{x}_i + \underline{x}_j + \underline{x}_k)/3$  für  $i < j < k$ , und  $\underline{x}_{iij} = (2\underline{x}_i + \underline{x}_j)/3$  für  $i \neq j$ . Mit  $\Pi_m$  bezeichnen wir wieder den Vektorraum der Polynome vom Grad  $\leq m$  mit  $n$  Variablen im  $\mathbb{R}^n$ . Dann gelten die folgenden Identitäten im  $\mathbb{R}^n$ :

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

Bis auf die ersten beiden sind diese Darstellungen nichttrivial, und sie sind nicht eindeutig wegen der linearen Abhängigkeit der baryzentrischen Koordinaten  $\zeta_i$ . Eine entsprechende Darstellung für MORLEYS Polynom zweiten Grades und das Polynom fünften Grades von ARGYRIS' in  $\mathbb{R}^2$  findet man in § 12.5 bzw. in [Gekeler08], §12.2 (c). (Beide benützen die Normalableitungen in  $\underline{x}_{ij}$ ). Siehe auch SUPPLEMENT\chap09e\chap09f.

Die Integration von Interpolationspolynomen über Dreiecke und allgemeinere geometrische Gebilde ist eine grundlegende Aufgabe bei der Konstruktion von finiten Elementen; siehe Kap. 9. Für Dreiecke kann man die Formel für die affin lineare Transformation

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

anwenden und anschließend mit der Formel (2.43) über das Einheitsdreieck  $S$  integrieren oder man integriert direkt mit Hilfe von baryzentrischen Koordinaten und der Formel (2.44) von BELL, z.B. ergibt sich

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

**(f) Gebietsintegrale (f1)** Die Quadratur nach GAUSS wird auch zur Integration über das Einheitsquadrat verwendet:

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

Mit den Daten aus Tab. 2.1 ist diese Formel exakt für Polynome

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

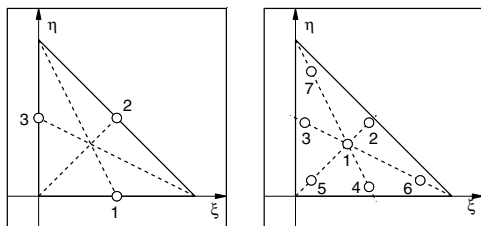
**(f2)** Stützstellen und Gewichte zweier beliebiger GAUSS-Formeln im Einheitsdreieck  $S(\xi, \eta)$  mit den Ecken  $Q(0,0)$ ,  $Q(1,0)$ ,  $Q(0,1)$ ,

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

sind in der folgenden Tabelle angegeben:

**Tabelle 2.2.**

| $n$ | $i$ | $\xi_i$ | $\eta_i$ | $\gamma_i$               |                         |
|-----|-----|---------|----------|--------------------------|-------------------------|
| 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                    | a $(6 + \sqrt{15})/21$  |
|     | 2   | a       | a        | $(155 + \sqrt{15})/1200$ | b $(9 - 2\sqrt{15})/21$ |
|     | 3   | b       | a        | $(155 + \sqrt{15})/1200$ | c $(6 - \sqrt{15})/21$  |
|     | 4   | a       | b        | $(155 + \sqrt{15})/1200$ | d $(9 + 2\sqrt{15})/21$ |
|     | 5   | c       | c        | $(155 - \sqrt{15})/1200$ |                         |
|     | 6   | d       | c        | $(155 - \sqrt{15})/1200$ |                         |
|     | 7   | c       | d        | $(155 - \sqrt{15})/1200$ |                         |



**Abb. 2.10.** Stützstellen nach GAUSS im Einheitsdreieck

Die Formeln sind exakt für Polynome  $p(\xi, \eta) = \sum_{0 \leq i+k \leq n} a_{ik} \xi^i \eta^k$  vom Grad  $n \leq 2$  bzw.  $n \leq 5$  (im Einheitsdreieck). Eine Formel für  $n = 3$  mit vier Knoten hat eine negatives Gewicht  $\gamma$  und ist daher nicht empfehlenswert, eine weitere Formel für  $n = 3$  mit positiven Gewichten hat sieben Knoten wie die Formel für  $n = 5$ .

Integrationsregeln für Polynome

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

über ein beliebiges Dreieck  $T$  ergeben sich in einfacher Weise durch Substitution mit (2.46).

**(f3)** Bei der direkten Integration von Polynomen über ein Dreieck  $T$  in globalen  $(x, y)$ -Koordinaten sind letztlich Integrale von Monomen zu berechnen. Durch Substitution von (2.41) erhält man

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

Damit werden die Integrale  $P_{rs}$  in Summen von Integralen der Form (2.44) bzw. (2.43) zerlegt. Die zweite Formel benützt wieder die Substitutionsregel (2.46) für die Abbildung  $g : S \rightarrow T$  von (9.21).

Für Polynome vom Grad  $n \leq 5$  sind die Ergebnisse in Tab. 2.3 angegeben [Bell], wobei aber aus Gründen der einfachen Darstellung der Ursprung des KOS im Zentrum des Dreiecks liegt. Die einfache Darstellung in dieser Tabelle gilt nicht mehr für Polynome höheren Grades und für KOS mit anderem Ursprung, aber heutzutage ersetzt ein *Programm* große Tabellen. KAPITEL02\SECTION\_1\_2\_3\bell11.m liefert die Werte des Integrals (2.47) für beliebige  $r, s \in \mathbb{N}$  in einem KOS mit beliebigem Ursprung (mit MATLAB<sup>®</sup> SYMBOLIC MATH. TOOLBOX).

Tabelle 2.3.

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

References: [Kardestuncer], [Stoer].

## 2.4 Anfangswertprobleme

In diesem Abschnitt werden Vektoren *nicht* unterstrichen!

**(a) Das Euler-Verfahren** Gesucht ist eine Lösung  $x : [0, T] \rightarrow \mathbb{R}^n$  des Anfangswert- oder CAUCHY-Problems

$$x'(t) = f(t, x(t)), \quad 0 \leq t \leq T, \quad x(0) = x_0. \quad (2.48)$$

Das Problem heißt *autonom*, wenn  $f$  nicht explizit von  $t$  abhängt.

Zur numerischen Lösung von (2.48) kann entweder  $x'(t)$  durch eine Differenzenformel ersetzt werden, oder aber man wandelt die Differentialgleichung in eine Intergleichung um,

$$x(t + \tau) = x(t) + \int_t^{t+\tau} f(s, x(s)) ds, \quad \tau \text{ Schrittweite,}$$

und ersetzt das Integral durch eine numerische Integrationsformel. Der einfachste Fall  $\int_t^{t+\tau} f(s, x(s)) ds \simeq \tau f(t, x(t))$  führt zum *expliziten EULER-Verfahren*,

$$y(t + \tau) = y(t) + \tau f(t, y(t)), \quad t = j\tau, \quad j = 0, 1, \dots, \quad y(0) = x(0) = x_0. \quad (2.49)$$

Das Einsetzen der exakten Lösung in die Näherungsformel (2.49) ergibt den *Defekt* oder nach Division durch  $\tau$  den *Diskretisierungsfehler*

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

Er misst die Genauigkeit, mit der die exakte Lösung der *Näherungsformel* (2.49) genügt und stellt bei einem expliziten Verfahren wie im vorliegenden

Fall auch den *lokalen Fehler* dar. Ist nämlich  $y(t) = x(t)$  *exakt*, dann folgt für einen Schritt

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

Der Diskretisierungsfehler wird stets mit einer TAYLOR-Entwicklung berechnet, z.B. gilt mit dem Integralrestglied

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

also für das Verfahren (2.49)

$$\|d(t, x, \tau)\| \leq \tau^1 \int_0^1 \|x''(t + \sigma \tau)\| d\sigma.$$

Man sagt daher: Das Verfahren (2.49) hat die Ordnung  $p = 1$ .

Für den *globalen Fehler*  $e(t) = x(t) - y(t)$  folgt durch Subtraktion unter Verwendung der LIPSCHITZ-Beschränkung

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

Eine Induktion ergibt dann die Abschätzung des globalen Fehlers, wobei noch aus optischen Gründen die Ungleichung  $(1 + x)^n \leq e^{nx}$ ,  $|x| \leq 1$ , verwendet wird:

**Lemma 2.4.** (*Fehlerabschätzung, Konvergenz*) Ist  $x \in C^2[0, T]$  Lösung von (2.48) und ist  $f$  LIPSCHITZ-beschränkt, dann gilt

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

Die Schrittweite  $\tau$  hebt sich also einmal heraus. Grundsätzlich unterscheiden sich die Ordnung des lokalen und des globalen Fehlers um den Faktor Eins.

Bei dieser *A-priori-Fehlerabschätzung* geht die unbekannte Lösung  $x$  in die Fehlerschranke ein. *A-posteriori-Fehlerabschätzungen*, die den Fehler mit Hilfe der berechneten Daten einschränken, sind schwierig herzuleiten, daher begnügt man sich i.d.R. mit *Fehlerschätzungen*. Für  $x' = Lx$  mit  $L > 0$  ist die obige Abschätzung scharf, und das Problem ist für großes  $L \cdot T$  schlecht konditioniert. Für  $L < 0$  ist die Abschätzung nicht sinnvoll, woraus sich die Notwendigkeit ergibt, neben der Diskretisierungsordnung weitere Gütekriterien für numerische Verfahren einzuführen.

**(b) Allgemeine Einschrittverfahren****Beispiel 2.7.** Die Iterationsvorschrift

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

$0 \leq \omega \leq 1$ , ergibt für  $\omega = 0$  das explizite EULER-Verfahren, für  $\omega = 1$  das *implizite* EULER-Verfahren und für  $\omega = 1/2$  die Trapezregel. Für  $\omega = 1/2$  hat das Verfahren die Ordnung  $p = 2$  und sonst die Ordnung  $p = 1$ .

Ein allgemeines Einschrittverfahren lässt sich schreiben als

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

oder als

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

mit  $y_j := y(j\tau)$ , wenn die Schrittweite konstant ist. Die *Verfahrensfunktion*  $\Phi$  muss einigen naheliegenden Bedingungen genügen, die aber im Regelfall erfüllt sind; vgl. [Hairer]. Das Verfahren heißt *explizit*, wenn zur *exakten* Berechnung die rechte Seite  $f$  der Differentialgleichung nur an endlich vielen Stellen ausgewertet werden muss, im andern Fall heißt das Verfahren *implizit*.

Der Diskretisierungsfehler ist wie in **(b)** definiert, und das Verfahren heißt *konsistent* (mit der Differentialgleichung), wenn für ein  $p \geq 1$  und für *alle* Lösungen  $x \in \mathcal{C}^{p+1}[0, T]$  gilt

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

Die maximal mögliche Zahl  $p$  in (2.52) heißt Ordnung des Verfahrens für die gegebene Differentialgleichung und *Ordnung* allgemein, wenn das Verfahren die Ordnung  $p$  für alle hinreichend glatten rechten Seiten  $f$  von (2.48) hat. An der Aussage von Lemma 2.4 ändert sich nichts.

**(c) Asymptotische Entwicklung, Extrapolation**

**Lemma 2.5.** Wenn das Verfahren (2.51) die Ordnung  $p \geq 1$  hat sowie

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

gilt, und  $\partial\Phi/\partial\tau$  in einer Nullumgebung von  $\tau$  stetig ist in  $\tau$ , dann gibt es eine von  $\tau$  unabhängige Fehlerfunktion  $r$  mit

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

Beweis [Hairer], Bd. I, § 2.8.

Diese asymptotische Aussage hat zwei wichtige Konsequenzen, wenn wir das Verfahren (2.51) einmal mit der Schrittweite  $\tau$  und dann noch einmal mit der reduzierten Schrittweite  $q\tau$ ,  $0 < q < 1$ , anwenden,

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

(1°) Die *bewichtete* Differenz

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

liefert ein verbessertes Verfahren mit der Ordnung  $p+1$  statt  $p$  mit vergleichsweise wenig Rechenaufwand.

**Beispiel 2.8.** Das *Testproblem*  $x' = \lambda x$  hat die Lösung  $x(t) = \kappa e^{\lambda t}$ . Wir wählen  $x(0) = 1$ ,  $\lambda = 1$ , und wenden die Trapezregel an, einmal mit der Schrittweite  $\tau = 1$  und zum Vergleich zweimal mit der Schrittweite  $\tau = 0.5$ :

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

Eine Anwendung der *Mittelung* (2.53) mit  $p = 2$  und  $q = 1/2$  ergibt praktisch ohne zusätzlichen Rechenaufwand die Verbesserung

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

mit dem Fehler  $\varepsilon = 0.0145 \dots$ .

(2°) Die *einfache* Differenz

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

liefert eine gute Schätzung für den globalen Diskretisierungsfehler. Mit einer Diagonalmatrix  $D$  von Gewichten, mit Toleranzen und Sicherheitsfaktoren, die in einer Konstante  $C$  zusammengefasst werden können, sowie weiteren Sicherheitsschranken ergibt

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

eine hervorragende Schrittweitensteuerung. Besonders günstig ist die asymptotische Situation bei *eingebetteten* Verfahren der Ordnung  $p$ , die gleichzeitig eine Näherung  $z(t)$  der Ordnung  $p - 1$  liefern,

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

Die Differenz ergibt dann direkt eine Schätzung für den Fehler von  $z(t)$ :

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

#### (d) Runge-Kutta-Verfahren

**Beispiel 2.9.** Das explizite Verfahren von HEUN entsteht aus der impliziten Trapezregel (2.50) durch die Substitution

$$f_{j+1}(y_{j+1}) \simeq f_{j+1}(y_j + \tau f_j(y_j)), \quad f_j(y_j) := f(t_0 + j\tau, y_j).$$

|                                                                                                                                                                                                                             |  |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|
| Trapezregel ( $p = 2$ ): $y_{j+1} = y_j + \frac{\tau}{2} \left( f_j(y_j) + f_{j+1}(y_{j+1}) \right)$<br>Verf. von HEUN ( $p = 2$ ): $y_{j+1} = y_j + \frac{\tau}{2} \left( f_j(y_j) + f_{j+1}(y_j + \tau f_j(y_j)) \right)$ |  |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|

Die Berechnung erfolgt in jedem Schritt nach dem Schema ( $t = t_0 + j\tau$ )

$$k_1(t) = f(t, y_j), \quad k_2(t) = f\left(t + \tau, y_j + \tau k_1\right), \quad y_{j+1} = y_j + \frac{\tau}{2} (k_1(t) + k_2(t)).$$

Die Verallgemeinerung dieser Idee führt zu den *mehrstufigen* Verfahren oder RUNGE-KUTTA-Verfahren.

**Beispiel 2.10.** Das klassische RUNGE-KUTTA-Verfahren ist ein vierstufiges Verfahren mit der Ordnung  $p = 4$ , bei dem viermal die Funktion  $f$  in Zwischenstufen ausgewertet und dann in einem Vorwärtsschritt eine Linearkombination der Ergebnisse gebildet wird; dieser letzte Schritt entsteht meistens aus einer numerischen Integrationsformel (hier der KEPLERSchen Fassregel):

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

Das Beispiel zeigt, wie die Ordnung von Einschrittverfahren auf kunstvolle Weise erhöht werden kann, wenn mehrere Zwischenstufen eingeführt werden, sogar Verfahren beliebig hoher Ordnung können auf diese Weise konstruiert werden (Verfahren von GRAGG-BULIRSCH-STOER).

Ein allgemeines  $r$ -stufiges Einschrittverfahren für  $x' = f(t, x) \in \mathbb{R}^n$  ist eine Rechenvorschrift der Form

$$\begin{aligned}
 k_i(t) &= f \left( t + \gamma_i \tau, y(t) + \tau \sum_{j=1}^r \alpha_{ij} k_j(t) \right), \quad i = 1 : r \\
 y(t + \tau) &= y(t) + \tau \sum_{i=1}^r \beta_i k_i(t)
 \end{aligned}
 \tag{2.54}$$

mit den *Funktionswerten*  $k_i(t) := f(t + \gamma_i \tau, u_i(t))$  als unbekannten Größen. Für das Studium der Eigenschaften ist aber die nachfolgende Darstellung günstiger, dazu sei  $I$  die Einheitsmatrix,

$$\begin{aligned}
 A &= [\alpha_{ij}] \in \mathbb{R}^r_r, \quad b = [\beta_i], \quad c = [\gamma_j], \quad e = [1] \text{ alle in } \mathbb{R}^r, \text{ sowie} \\
 A \times B &= [\alpha_{ij} B]_{i,j=1}^r \quad \text{Kronecker-Produkt,} \\
 U(t) &= [u_i(t)]_{i=1}^r \quad \text{Hilfsvektoren, } u_i(t) \in \mathbb{R}^n, \\
 F(t, U(t)) &= [f(t + \gamma_i \tau, u_i(t))]_{i=1}^r \in \mathbb{R}^{r \cdot n}.
 \end{aligned}$$

Die Rechenvorschrift (2.54) ist dann äquivalent zu der Form mit *Zwischenwerten*

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

Zum Beispiel lässt sich das Verfahren von HEUN auch schreiben als

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

Die Hilfsgrößen  $u_i(t)$  können als Näherungen von  $x(t + \gamma_i \tau)$  aufgefasst werden, was aber nur für die Herleitung von Ordnungsbedingungen bedeutsam ist.

### Eigenschaften und weitere Bezeichnungen:

- (1°) Das Verfahren (2.55) heißt *explizit* bzw. *semi-implizit*, wenn (nach einer ev. Umnummerierung) die Matrix  $A$  eine streng untere bzw. eine untere Dreiecksmatrix ist; in den anderen Fällen heißt das Verfahren *implizit*.
- (2°) Im Regelfall liegen die Werte  $\gamma^i \tau$  im Intervall  $[0, \tau]$ , sind aber nicht immer alle verschieden; vgl. Beispiel 2.10. Die Zwischenwerte  $u_i(t)$  sind wie gesagt Näherungen an den Zwischenstellen  $t + \gamma^i \tau$ . Das System der Zwischenstufen ist eindeutig lösbar, wenn  $f$  LIPSCHITZ-beschränkt und die Schrittweite  $\tau$  hinreichend klein ist, bei expliziten Verfahren entfällt die zweite Bedingung.
- (3°) Mit Hilfe der BUTCHER-Matrix  $[A|b|c]$  o.ä. lässt sich ein Verfahren (2.55) in Kurzform beschreiben, z.B. gilt für das Beispiel 2.10 in optisch angepasster Schreibweise

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

(4°) Ist  $W(t)$  die Lösung von  $W(t) = e \times x(t) + \tau(A \times I)F(t, W(t))$ , dann ist

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

der *Diskretisierungsfehler*.

(5°) Ein Verfahren (2.54) hat die Ordnung  $p \geq 1$ , g.d.w.  $b^T e = \sum_{i=1}^r \beta_i = 1$ .

(6°) Tabelle der erreichbaren Ordnung  $p^*$  von expliziten RUNGE-KUTTA-Verfahren bei vorgegebener Stufenzahl  $r$  nach [Butcher]:

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

Daher nimmt das RUNGE-KUTTA-Verfahren der Ordnung  $p = 4$  eine besondere Stellung ein.

**(e) Mehrstellenverfahren** Die mehrmalige Auswertung der rechten Seite  $f$  der Differentialgleichung bei mehrstufigen Verfahren kann sehr zeitraubend sein. Die Ordnung des Verfahrens lässt sich aber auch erhöhen, wenn im Sinne einer Extrapolation die bisher berechneten Werte  $y_j, y_{j-1}, \dots$  berücksichtigt werden. Zum Beispiel ist ein bekanntes implizites Verfahren der Ordnung  $p = 2$  mit sehr guten Stabilitätseigenschaften durch die Vorschrift

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

gegeben. Allgemein haben Mehrstellenverfahren oder Mehrschrittverfahren die Form

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

mit  $\alpha_k \neq 0$  und  $|\alpha_0| + |\beta_0| \neq 0$ . Ist  $\beta_k = 0$ , so ist das Verfahren *explizit* im andern Fall *implizit*. Die Funktion  $f$  muss hier in jedem Schritt nur einmal ausgewertet werden, dagegen müssen die Startwerte  $y_1, \dots, y_{k-1}$  durch ein anderes Verfahren vorgegeben werden.

**Eigenschaften und Bezeichnungen:**

(1°) Mit den Polynomen  $\varrho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i$ ,  $\sigma(\zeta) = \sum_{i=0}^k \beta_i \zeta^i$  und dem Translationsoperator  $E : y(t) \mapsto Ey(t) := y(t + \tau)$ , lässt sich (2.56) vereinfacht schreiben als

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

(2°) Einsetzen der Testgleichung  $x' = \lambda x$  ergibt mit  $\eta = \tau\lambda$

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

Mit Hilfe des *charakteristischen Polynoms*  $\pi(E, \eta)$  ergibt sich sofort der Diskretisierungsfehler des Verfahrens zu

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

- (3°) Ordnung und Konsistenz sind wie in **(b)** definiert. Im Gegensatz zu den RUNGE-KUTTA-Verfahren lässt sich aber bei Mehrstellenverfahren eine vorgegebene Ordnung  $p$  leicht erreichen, wenn die Koeffizienten so abgeglichen werden, dass die Approximation exakt ist für alle „Differentialgleichungen“  $x'(t) = t^k$ ,  $k = 0 : p$ . Weitere Möglichkeiten zur Konstruktion von Mehrschrittverfahren ergeben sich, wenn der Abgleich über Funktionen  $e^{it}$ ,  $\sin(jt)$ ,  $\cos(kt)$  oder eine gewisse Kombination aller dieser Grundfunktionen erfolgt.

**Lemma 2.6.** *Ein Mehrschrittverfahren  $(\varrho, \sigma)$  hat die Ordnung  $p \geq 1$  g.d.w.*

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

Insbesondere hat das Verfahren die Ordnung  $p \geq 1$ , wenn  $\varrho(1) = 0$  und  $\varrho'(1) - \sigma(1) = 0$ .

- (4°) Für den Diskretisierungsfehler ergibt sich in einfacher Weise

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

und dann durch Induktion eine Fehlerabschätzung mit der gleichen qualitativen Aussage wie in Lemma 2.4. Dazu muss aber das Polynom  $\varrho$  die *Wurzelbedingung* erfüllen:

Die Wurzeln von  $\varrho$  sind betragsweise  $\leq 1$ ,  
und die Wurzeln vom Betrag Eins sind einfache Wurzeln.

- (5°) Damit bei der Anwendung keine „Geisterlösungen“ auftreten, sollte das Mehrschrittverfahren *stark stabil* sein, d.h. das Polynom  $\varrho$  hat genau eine Wurzel  $\zeta$  vom Betrag Eins, nämlich  $\zeta = 1$ , und diese ist einfache Wurzel von  $\varrho$ .

## (f) Zusammenfassung

| mehrstufige Verfahren          | Mehrstellenverfahren             |
|--------------------------------|----------------------------------|
| starten selbst                 | starten nicht selbst             |
| Rechenaufwand hoch             | Rechenaufwand nieder             |
| Schrittweitensteuerung einfach | Schrittweitensteuerung schwierig |

**(g) Stabilität**

**(g1)** Die Differentialgleichung  $x'(t) = f(t, x(t))$  heißt *stabil*, wenn die Differenz je zweier Lösungen im ganzen  $t$ -Intervall beschränkt bleibt und *asymptotisch stabil*, wenn die Differenz zusätzlich für  $t \rightarrow \infty$  gegen Null geht, vgl. §1.5(c). Die Instabilität der Differentialgleichung überträgt sich in jedem Fall auf die numerische Approximation, trotzdem kann eine gute Schrittweitensteuerung hervorragende Ergebnisse liefern. Bei einer stabilen Differentialgleichung fallen über einen längeren Zeitraum gesehen die Lösungen betragsweise exponentiell ab oder sie bleiben wenigstens beschränkt. Diese Eigenschaft muss sich natürlich auf die numerische Approximation vererben, was aber nicht immer gewährleistet ist:

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

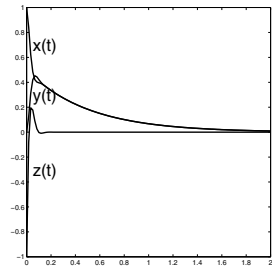
$$A = \begin{bmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{bmatrix}, \text{ Eigenwerte : } \lambda_1 = -2, \lambda_{2,3} = -40 \pm 40i.$$

Die Lösung

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

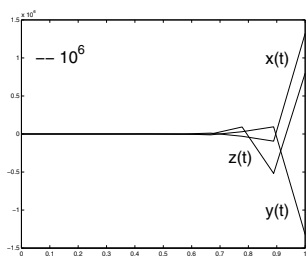
verhält sich ab  $t = 0.1$  wie die Lösung von  $x' = Bx$  mit

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

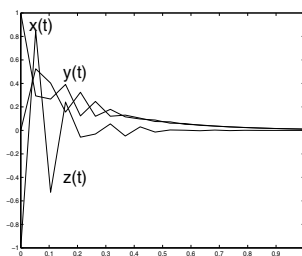


**Abb. 2.11.** Bsp. 2.11, Lösung

Während das explizite EULER-Verfahren mit dieser Schrittweite völlig inakzeptabel ist, liefert die Trapezregel brauchbare Ergebnisse. Bei dem expliziten Verfahren ist eine kleine Schrittweite unnötig bei großem  $t$ , während eine große Schrittweite die hochfrequenten (aber schnell abklingenden) Anteile der Lösung explosionsartig verstärkt.



**Abb. 2.12.** Bsp. 2.11, EULER explizit,  $\tau = 0.1$



**Abb. 2.13.** Bsp. 2.11, Trapezregel,  $\tau = 0.1$

**(g2)** Zum Studium dieses Phänomens wenden wir das mehrstufige Verfahren (2.55) auf die *Testgleichung*  $x'(t) = \lambda x(t)$  an und erhalten mit  $\eta = \tau\lambda$

$$\begin{aligned} U(t) &= ey(t) + \tau\lambda AU(t) & \implies & U(t) = (I - \eta A)^{-1}ey(t) \\ y(t + \tau) &= y(t) + \tau\lambda b^T U(t) & \implies & y(t + \tau) = [1 + \eta b^T (I - \eta A)^{-1}e] y(t). \end{aligned}$$

Daraus ergibt sich für die Testgleichung die Iterationsvorschrift

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

(wobei  $R(\infty)$  nur für reguläre Koeffizientenmatrix  $A$  definiert ist). Aus der CRAMERSchen Regel folgt nach kurzer Rechnung die *Stabilitätsfunktion*

$$R(\eta) = \frac{\det(I - \eta A + \eta eb^T)}{\det(I - \eta A)} =: \frac{P(\eta)}{Q(\eta)}, \quad (2.59)$$

mit Polynomen  $P$  und  $Q$ , und bei expliziten Verfahren ist  $Q(\eta) = 1$ . Die Menge

$$\mathcal{S} := \{\eta \in \mathbb{C}, |R(\eta)| \leq 1\}$$

in der komplexen  $\eta$ -Ebene heißt *Stabilitätsbereich* des Einschrittverfahrens. Betrachten wir nun das System  $x'(t) = Ax(t)$  mit einer *diagonalisierbaren* Matrix  $A$ ,  $A = U\Lambda U^{-1}$ ,

( $\Lambda$  Diagonalmatrix der Eigenwerte  $\lambda_i$  von  $A$ ), dann folgt für das Einschrittverfahren

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

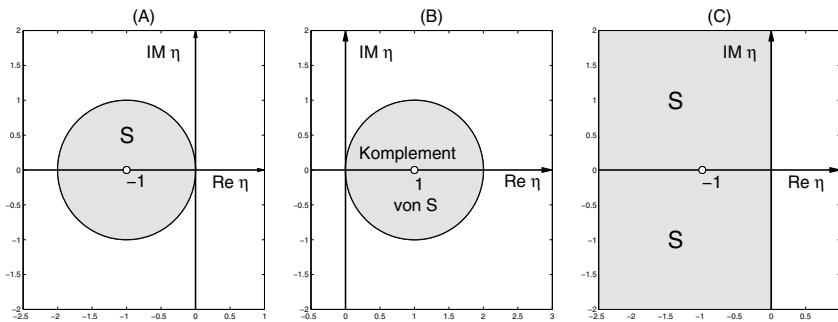
mit der Diagonalmatrix

$$R(\tau\Lambda) = \text{diag}(R(\tau\lambda_1), \dots, R(\tau\lambda_n)).$$

Es müssen also alle  $\eta_i := \tau\lambda_i$  im Stabilitätsbereich  $\mathcal{S}$  liegen, wenn jede Lösung wenigstens beschränkt bleiben soll. Dies ist die COURANT-FRIEDRICHS-LEVY-Bedingung für die Schrittweite  $\tau$ . Sie muss bei allen stabilen Problemen beachtet werden aber auch in vielen anderen Fällen; siehe z.B. § 9.7 (d).

**Beispiel 2.12.** Für die Testgleichung  $x' = \lambda x$  gilt mit  $\eta = h \lambda$

- (A) EULER-Verfahren explizit ( $p = 1$ ):  $y_{j+1} = (1 + \eta)y_j$ ,
- (B) EULER-Verfahren implizit ( $p = 1$ ):  $y_{j+1} = (1 - \eta)^{-1}y_j$ ,
- (C) Trapezregel ( $p = 2$ ):  $y_{j+1} = \frac{2 + \eta}{2 - \eta}y_j$ ,
- (D) Verfahren von HEUN ( $p = 2$ ):  $y_{j+1} = \left(1 + \eta + \frac{1}{2}\eta^2\right)y_j$ .



**Abb. 2.14.** Stabilitätsbereiche für Beispiel 2.12 ohne (D)

Hat ein Verfahren vom Typ (2.55) die Ordnung  $p \geq 1$ , dann gilt einerseits für den Diskretisierungsfehler  $\tau d(t, x, \tau) = \mathcal{O}(\tau^{p+1})$  und andererseits folgt für die Testgleichung mit  $\lambda = 1$  und  $x(0) = 1$

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

Subtraktion ergibt

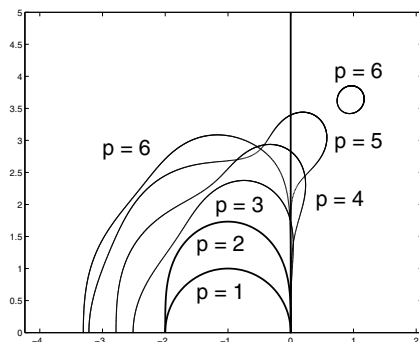
$$\tau d(\tau, x, \tau) = x(\tau) - y(\tau) = e^\tau - R(\tau) = \mathcal{O}(\tau^{p+1}).$$

Folglich gilt für jedes RUNGE-KUTTA-Verfahren der Ordnung  $p$

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

und alle *expliziten* RUNGE-KUTTA-Verfahren mit  $\boxed{p = r}$  haben die gleiche

Stabilitätsfunktion  $R(\eta) = \sum_{i=0}^p \eta^i / i!$ , weil dann  $P(\eta)$  ein Polynom vom Grad  $p$  ist nach (2.59). Wegen der Symmetrie zur reellen Achse ist in Abb. 2.15 jeweils nur die obere Hälfte des Stabilitätsbereichs eingezeichnet.



**Abb. 2.15.** Stabilitätsbereiche für explizite RKV mit  $p = r = 1 : 6$

**(g3)** Wenden wir die Testgleichung auf ein Mehrschrittverfahren (2.56) an, so ergibt sich nach (2.57)

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

oder, mit  $Y_j = [y_j, y_{j+1}, \dots, y_{j+k-1}]^T \in \mathbb{R}^k$  als Einschrittverfahren geschrieben,

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

Die FROBENIUS-Matrix  $F_\pi(\eta)$  hat das charakteristische Polynom

$\det(\lambda I - F_\pi(\eta)) = \pi(\lambda, \eta)$ , daher heißt sie auch Begleitmatrix zum Polynom  $\pi$ . Außerdem besitzt sie zu jedem Eigenwert genau einen Eigenvektor. Andererseits muss sie nach Satz 1.1 eine M-Matrix sein – vgl. § 1.1 **(c4)**, wenn alle Folgen (2.60) beschränkt bleiben sollen, daher ist der Begriff des *Stabilitätsbereichs* bei Mehrstellenverfahren etwas abzuändern:

**Definition 2.2.** Es sei  $(\varrho, \sigma)$  ein Mehrschrittverfahren mit dem charakteristischen Polynom  $\pi(\zeta, \eta) = \varrho(\zeta) - \eta\sigma(\zeta)$ , und es sei  $\pi(\zeta, \infty) = \sigma(\zeta)$ . Dann besteht der Stabilitätsbereich  $\mathcal{S} \in \mathbb{C} \cup \{\infty\}$  aus der Menge aller Punkte  $\eta$ , für die gilt:

(1°) Alle Wurzeln  $\zeta_i(\eta)$  von  $\pi(\zeta, \eta)$  erfüllen  $|\zeta_i(\eta)| \leq 1$ .

(2°) Alle Wurzeln  $\zeta_i(\eta)$  von  $\pi(\zeta, \eta)$  mit  $|\zeta_i(\eta)| = 1$  (unimodulare Wurzeln) sind einfache Wurzeln von  $\pi(\zeta, \eta)$ .

**(h) Steife Differentialsysteme** Ein System  $x'(t) = Ax(t)$  heißt *steif*, wenn für alle Eigenwerte  $\lambda_i$  von  $A$  gilt

$$\operatorname{Re} \lambda_i \leq 0 \text{ und } \operatorname{Max}_i |\operatorname{Re} \lambda_i| \gg \operatorname{Min}_i |\operatorname{Re} \lambda_i|.$$

Ein solches System entsteht z.B., wenn Massepunkte untereinander gleichzeitig durch sehr weiche und sehr harte Federn verbunden sind. Zwangsläufig treten steife System bei der Diskretisierung von Differentialgleichungen auf, wie das folgende Beispiel zeigt.

**Beispiel 2.13.** Das Eigenwertproblem

$$y''(x) = \lambda^2 y, \quad y(0) = y(1) = 0, \quad (2.61)$$

hat die charakteristischen Paare

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

Wird die zweite Ableitung diskretisiert vermöge

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

dann entsteht das Eigenwertproblem  $AY = \tilde{\lambda}^2 Y \in \mathbb{R}^n$ ,

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

mit den charakteristischen Paaren

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

Ausnahmsweise stimmen hier die Eigenvektoren des diskretisierten Problems mit den Werten der Eigenfunktionen des analytischen Problems (2.61) an den entsprechenden Stellen überein. Für die Eigenwerte gilt

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

Für jedes feste  $j$  ist  $\tilde{\lambda}_j^2$  eine Approximation zweiter Ordnung des Eigenwerts  $-j^2\pi^2$  der Differentialgleichung, insbesondere wachsen die Eigenwerte von (2.62) betragsmäßig über jede Schranke, wenn die Schrittweite  $h$  gegen Null geht.

Diskretisieren wir nun das *parabolische* Anfangsrandwertproblem

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

ebenso wie (2.61) in der Raumveränderlichen  $x$ , so ergibt sich das Anfangswertproblem

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

mit der Matrix  $A$  aus (2.62). Das Modellproblem (2.63) kann wenigstens für  $a(t) = b(t) = 0$  exakt gelöst werden. Soll es mit einem numerischen Verfahren gelöst werden, so muss nach der COURANT-FRIEDRICHS-LEVY-Bedingung die Schrittweite  $\tau$  so klein gewählt werden, dass der Punkt  $\tau \tilde{\lambda}_n^2 \simeq -4\tau/h^2$  noch im Stabilitätsbereich  $\mathcal{S}$  liegt. Bei den *impliziten* Verfahren (B) und (C) entfällt dagegen diese Schrittweitenbeschränkung, weil die ganze negative Halbgerade zu  $\mathcal{S}$  gehört. Aus diesem Grund werden *Formkriterien* für den Stabilitätsbereich eingeführt: Ein Verfahren heißt

|                                                                                                                                                                                                    |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| $A\text{-stabil} \iff \{\eta \in \mathbb{C}, \operatorname{Re} \eta \leq 0\} \subset \mathcal{S}$                                                                                                  | . (2.64) |
| $A(\alpha)\text{-stabil} \iff \{\eta \in \mathbb{C}, \eta \neq 0,  \pi - \operatorname{Arg} \eta  \leq \alpha\} =: \mathcal{S}_\alpha \subset \mathcal{S}$                                         |          |
| $A(0)\text{-stabil} \iff \exists \alpha > 0 : \text{Verfahren } A(\alpha)\text{-stabil}$                                                                                                           |          |
| $A_0\text{-stabil} \iff (-\infty, 0] \subset \mathcal{S}$                                                                                                                                          |          |
| $L\text{-stabil} \iff \text{Verfahren } A\text{-stabil und } \lim_{\operatorname{Re} \eta \rightarrow -\infty} R(\eta) = 0 \text{ oder } A\text{-stabil und } R(\infty) = 0 \text{ bei Existenz.}$ |          |

( $R$  Stabilitätsfunktion); s. auch [Hairer] II, p. 45, Prop. 3.8 und p. 269.

Wegen der Darstellung (2.59) für die Stabilitätsfunktion kann ein explizites RUNGE-KUTTA-Verfahren niemals eine der in (2.64) genannten Eigenschaften haben, das gleiche lässt sich auch für explizite Mehrschrittverfahren einfach nachprüfen. Man erkennt das Dilemma: Bei expliziten Verfahren und muss im Beispiel 2.13 die Schrittweite  $\tau$  in  $t$ -Richtung proportional zum Quadrat der Schrittweite  $h$  in  $x$ -Richtung gewählt werden. Bei impliziten Verfahren vergrößert sich dagegen der Rechenaufwand erheblich. Im Übrigen haben nicht alle impliziten Verfahren automatisch eine der Eigenschaften (2.64). Im Einzelfall lässt sich dies am besten durch Plotten des Stabilitätsbereichs nachprüfen.

Regel:

Für Anfangswertprobleme mit instabiler Differentialgleichung nur explizite Verfahren mit Schrittweitensteuerung verwenden!

(i) **Weitere Beispiele** Wir beschreiben kurz einige Verfahren aus der MATLAB® ODE Suite.

(1°) MATLAB® `ode45.m` Eingebettetes explizites RUNGE-KUTTA-Verfahren nach DORMAND & PRINCE; vgl. [Dormand]:

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

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

Verwendet man die zweitletzte Zeile für  $b$ , ergibt sich ein Verfahren der Ordnung  $p = 5$  und mit der letzten Zeile ein Verfahren der Ordnung  $p = 4$ . Das sechsstufige Verfahren hat die Ordnung  $p = 5$ , die siebte Stufe wird nur zur Fehlerschätzung verwendet. Der Stabilitätsbereich ist der Gleiche wie bei einem expliziten sechsstufigen RUNGE-KUTTA-Verfahren der Ordnung  $p = 6$ ; vgl. Abb. 2.15.

(2°) ROSENBROCK-Verfahren sind vielleicht nicht die *ultima ratio* aber doch das Ergebnis einer langen Reihe von Überlegungen zur *Effizienz* von Verfahren für steife Systeme. Die sich widersprechenden Forderungen an eine hohe Ordnung, geringen Rechenaufwand und beste Stabilitätseigenschaften wie L-Stabilität, vgl. (2.64), haben schließlich zu einem Kompromiss geführt. Ausgangspunkt ist ein RUNGE-KUTTA-Verfahren, bei dem die Koeffizientenmatrix eine untere Dreiecksmatrix ist (diagonal implizite Verfahren). Wenn wir uns zunächst auf ein autonomes System  $x'(t) = f(x(t))$  beschränken, hat es nach (2.54) die Form

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

$$y(t + \tau) = y(t) + \tau \sum_{i=1}^r \beta_i k_i(t).$$

Die Gleichungen werden nun linearisiert, z.B. können die Funktionswerte  $k_i(t)$  durch

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

ersetzt werden. Nach einem Vorschlag von [Calahan] werden zusätzlich die  $r$  Matrizen  $\text{grad } f(g_i(t))$  durch eine einzige Matrix  $J := \text{grad } f(y(t))$  ersetzt, was den Rechenaufwand pro Iterationsschritt noch einmal erheblich reduziert. Bei den ROSENBROCK-Verfahren wird dann die Kombination

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

$$y(t + \tau) = y(t) + \tau \sum_{i=1}^r \beta_i k_i(t)$$

gewählt, um eine größere Freiheit in der Wahl der Koeffizienten zu erreichen. Das MATLAB®-Programm `ode23s.m` von [Shampine82] ist von diesem Typ, dabei wird die letzte Auswertung von  $f$  im vorhergehenden Schritt als erste Auswertung im neuen Schritt verwendet:

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

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

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

Das zweistufige Verfahren hat die Ordnung  $p = 2$ , der Wert  $\tilde{y}(t + \tau)$  wird nur zur Fehlerschätzung verwendet. Einsetzen der Testgleichung  $x'(t) = \lambda x(t)$  ergibt mit  $\eta = \tau \lambda$  wieder die Iterationsvorschrift

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

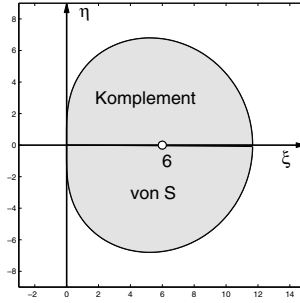


Abb. 2.16. Stabilitätsbereich des Rosenbrock-Verfahrens

(3°) MATLAB® `ode113.m` Prädiktor-Korrektor-Verfahren

Es seien die *Rückwärtsdifferenzen* definiert durch

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

und es sei  $p_k(x; f)$  ein Interpolationspolynom vom Grad  $k$  mit der Interpolationseigenschaft

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

wobei zunächst  $j \in \mathbb{N}_0$  fest gegeben ist. Dann lässt sich dieses Polynom in der NEWTON-GREGORY-Form schreiben als

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

(3.1°) Durch Integration von (2.65) über das  $s$ -Intervall  $(-1, 0)$ ,

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

entstehen die impliziten ADAMS-Verfahren mit  $k$  Stellen und der Ordnung  $k+1$ :

$$\begin{aligned}y_{j+k} - y_{j+k-1} &= \tau \sum_{i=0}^k \gamma_i \nabla^i f_{j+k}(y_{j+k}), \quad j = 0, 1, \dots, \\ \gamma_i &= \int_{-1}^0 \binom{\xi + i - 1}{i} d\xi.\end{aligned}$$

(3.2°) Durch Integration über das  $s$ -Intervall  $(0, 1)$  entstehen aus

$$x_{j+k+1} - x_{j+k} = \tau \int_0^1 f((j + k + s)\tau, x(t)) ds \simeq \tau \int_0^1 p_{k-1}((j + k + s)\tau; f) ds,$$

die expliziten ADAMS-Verfahren mit  $k$  Stellen und der Ordnung  $k$ :

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

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

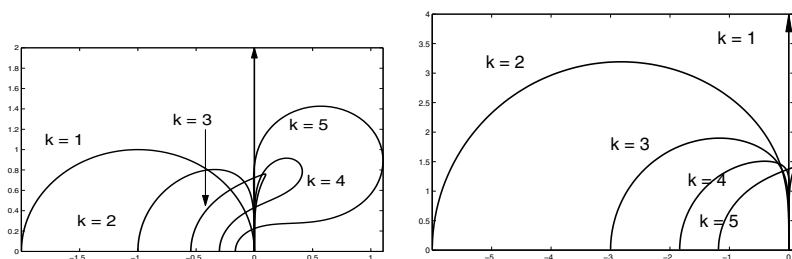
Die Koeffizienten  $\gamma_i$  und  $\gamma_i^*$  hängen nicht von der Stellenzahl  $k$  ab und lassen sich rekursiv berechnen.

(3.3°) Zusammen ergeben beide Verfahren ein *Prädiktor-Korrektor-Verfahren* für nichtsteife Differentialgleichungen:

Im Prädiktorschritt wird einmal ein explizites Verfahren (3.2°) der Ordnung  $k$  angewendet. Im Korrektorschritt wird mit dem impliziten Verfahren (3.1°) der Ordnung  $k + 1$  iteriert, wobei aber ein Iterationsschritt für die Ordnung  $k$  des kombinierten Verfahrens genügt.

**Tabelle 2.4.** Explizite Adams-Verfahren und implizite Adams-Verfahren

| $k$ | $\beta_0$           | $\beta_1$           | $\beta_2$            | $\beta_3$           | $\beta_4$            | $\beta_5$           | $\beta_0$         | $\beta_1$           | $\beta_2$          | $\beta_3$           | $\beta_4$           | $\beta_5$          |
|-----|---------------------|---------------------|----------------------|---------------------|----------------------|---------------------|-------------------|---------------------|--------------------|---------------------|---------------------|--------------------|
| 1   | 1                   |                     |                      |                     |                      |                     | $\frac{1}{2}$     | $\frac{1}{2}$       |                    |                     |                     |                    |
| 2   | $-\frac{1}{2}$      | $\frac{3}{2}$       |                      |                     |                      |                     | $-\frac{1}{12}$   | $\frac{8}{12}$      | $\frac{5}{12}$     |                     |                     |                    |
| 3   | $\frac{5}{12}$      | $-\frac{16}{12}$    | $\frac{23}{12}$      |                     |                      |                     | $\frac{1}{24}$    | $-\frac{5}{24}$     | $\frac{19}{24}$    | $\frac{9}{24}$      |                     |                    |
| 4   | $-\frac{9}{24}$     | $\frac{37}{24}$     | $-\frac{59}{24}$     | $\frac{55}{24}$     |                      |                     | $-\frac{19}{720}$ | $\frac{106}{720}$   | $-\frac{264}{720}$ | $\frac{646}{720}$   | $\frac{251}{720}$   |                    |
| 5   | $\frac{251}{720}$   | $-\frac{1274}{720}$ | $\frac{2616}{720}$   | $-\frac{2774}{720}$ | $\frac{1901}{720}$   |                     | $\frac{27}{1440}$ | $-\frac{173}{1440}$ | $\frac{482}{1440}$ | $-\frac{798}{1440}$ | $\frac{1427}{1440}$ | $\frac{475}{1440}$ |
| 6   | $-\frac{475}{1440}$ | $\frac{2877}{1440}$ | $-\frac{7298}{1440}$ | $\frac{9982}{1440}$ | $-\frac{7923}{1440}$ | $\frac{4277}{1440}$ |                   |                     |                    |                     |                     |                    |



**Abb. 2.17.** Stabilitätsbereiche expliziter und impliziter ADAMS-Verfahren

(4°) Rückwärtsdifferenzenverfahren (ähnliche Verfahren werden in `odes15s.m` verwendet). Wegen

$$\left. \frac{d}{ds} \binom{s+i-1}{i} \right|_{s=0} = \begin{cases} 0 & \text{für } i = 0 \\ \frac{1}{i} & \text{für } i \in \mathbb{N} \end{cases}$$

folgt aus (2.65) durch Ableiten nach  $s$

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

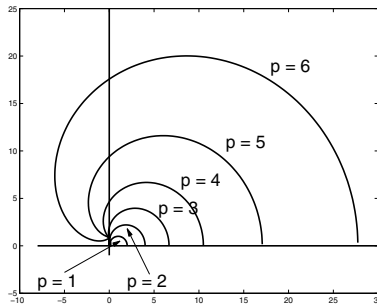
In dieser Gleichung ist die rechte Seite gegeben und die linke gesucht. Für die Umkehrung schreiben wir  $y(t) = F(t)$  statt  $f(t)$  sowie  $f(t) = F'(t)$  statt  $f'(t)$  und erhalten

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

Hieraus entstehen die impliziten Rückwärtsdifferenzenverfahren

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

mit  $k$  Stellen und der Ordnung  $k$  (Tab. 2.5).



**Abb. 2.18.** Stabilitätsbereiche der Rückwärtsdifferenzenverfahren

In Abb. 2.18 besteht die obere Hälfte des Stabilitätsbereichs jeweils aus dem *Außengebiet* der eingezeichneten Kurven und den Kurven selbst, es gilt also stets  $\infty \in \text{int}(\mathcal{S})$ . Für  $k > 6$  erfüllen diese Verfahren nicht mehr das Wurzelkriterium.

**Tabelle 2.5.** Rückwärtsdifferenzenverfahren

| $k$ | $\beta_k$ | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ |
|-----|-----------|------------|------------|------------|------------|------------|------------|------------|
| 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        |

**(j) Voll implizite Runge-Kutta-Verfahren** haben in jüngerer Zeit wieder größeres Interesse gefunden im Zusammenhang mit der Lösung *differential-algebraischer Gleichungen*, die bei vielen technischen und mechanischen Problemen auftreten. Die Ergebnisse dieses Abschnitts sind mehrheitlich schon seit den Arbeiten von BUTCHER bekannt, die nachfolgende algebraisierte Form ist aber vermutlich [Crouzeix75] und [Crouzeix80] zu verdanken. Für ausführliche Beweise siehe SUPPLEMENT\chap02b.

Aus Darstellungsgründen führen wir für diesen Abschnitt die folgende Bezeichnung ein:

Eine numerische Quadraturformel hat die *Ordnung*  $p$ , wenn sie den Genauigkeitsgrad  $p - 1$  hat, d.h. wenn sie exakt ist für alle Polynome mit  $p$  Koeffizienten; vgl. § 2.3(a).

Wird ein  $r$ -stufiges RUNGE-KUTTA-Verfahren (RKV) der Ordnung  $p$  mit der BUTCHER-Matrix  $(A, b, c)$  auf die triviale Differentialgleichung  $x'(t) = f(t)$  angewendet, dann entspricht die äußere Gleichung – auch *Kellerzeile* genannt, nämlich die Gleichung

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

für  $t = 0$ ,  $\tau = 1$  einer *numerischen Quadraturformel*

$$\int_0^1 f(t) dt \sim \sum_{i=1}^r \beta_i f(\gamma_i). \quad (2.66)$$

Werden hier die *Monome*  $f(t) = t^{k-1}$  eingesetzt, so zeigt sich auf Grund der Linearität, dass diese Formel die Ordnung  $\varrho$  hat, g.d.w.

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

Ebenso lassen sich den *inneren* Gleichungen eines RKV Quadraturformeln zuordnen gemäß

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

und sie haben die Ordnung  $\varrho$ , wenn

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

Bei einem RKV der Ordnung  $\varrho$  muss die äußere Gleichung notwendigerweise die Ordnung  $\varrho$  haben. Die *größte gemeinsame* Ordnung der Formeln (2.67)

heißt *innere Ordnung* des RKV. Damit hat ein *explizites* RKV die innere Ordnung  $\varrho = 1$ , weil die erste Gleichung den Grad Null hat.

Die Bedingungen für die Koeffizienten voll impliziter  $r$ -stufiger RKV der Ordnung  $p$  lassen sich in erstaunlich einfacher Weise beschreiben. Dazu bedarf es aber einer Zusatzbedingung, die für beliebige stetige Funktionen  $f$  durch partielle Integration gewonnen wird:

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

Wird das äußere Integral der linken Seite durch die äußere Gleichung und die inneren Integrale durch die entsprechenden inneren Gleichungen eines RKV approximiert, so ergibt sich insgesamt für die linke Seite

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

Approximieren wir dagegen die rechte Seite von (2.68) durch die äußere Gleichung des RKV, so ergibt sich

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

Wir fordern nun, dass beide Approximationen gleich sind, wenn wir für  $f$  nacheinander die LAGRANGE Polynome  $q_i \in \Pi_{r-1}$  mit  $q_i(\gamma_j) = \delta^i_j$ ,  $i = 1, \dots, r$ , einsetzen. Dann erhalten wir als Ergebnis die gewünschte zusätzliche Bedingung

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

Zur Formalisierung der Ordnungsbedingungen definieren wir nun die folgenden vier Eigenschaften eines RKV:

|                                                                                                                                                                                                                                                                                                                                                                      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| $\mathcal{A}(\varrho) : \iff$ das Verfahren hat <i>mindestens</i> die Ordnung $\varrho$<br>$\mathcal{B}(\varrho) : \iff$ die äußere Gleichung hat <i>mindestens</i> die Ordnung $\varrho$<br>$\mathcal{C}(\varrho) : \iff$ das Verfahren hat <i>mindestens</i> die innere Ordnung $\varrho$<br>$\mathcal{D}(\varrho) : \iff$ (2.69) gilt für $k = 1, \dots, \varrho$ | . (2.70) |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|

Ferner seien

$$b = [\beta_1, \beta_2, \dots, \beta_r]^T \in \mathbb{R}^r, \quad c = [\gamma_1, \gamma_2, \dots, \gamma_r]^T \in \mathbb{R}^r, \quad C = \text{diag}(c), \\ e = [1, \dots, 1]^T \in \mathbb{R}^r, \quad z_\varrho = [1, 1/2, \dots, 1/\varrho]^T \in \mathbb{R}^\varrho.$$

Dann sind die Eigenschaften (2.70) nach (2.66), (2.67) und (2.69) gleichbedeutend mit

$$\begin{array}{lcl}
 \mathcal{A}(\varrho) & \Longleftrightarrow & \text{RKV hat Ordnung } \varrho \\
 \mathcal{B}(\varrho) & \Longleftrightarrow & b^T C^{k-1} e = \frac{1}{k}, \quad k = 1 : \varrho \\
 \mathcal{C}(\varrho) & \Longleftrightarrow & A C^{k-1} e = \frac{1}{k} C^k e, \quad k = 1 : \varrho \\
 \mathcal{D}(\varrho) & \Longleftrightarrow & b^T C^{k-1} A = \frac{1}{k} (b^T - b^T C^k), \quad k = 1 : \varrho
 \end{array} \quad (2.71)$$

**Satz 2.15.** (BUTCHER, CROUZEIX, EHLE) *Wenn alle Stützabszissen  $\gamma_i$  verschieden sind, gilt*

$$(1^\circ) \quad \mathcal{B}(\varrho) \wedge \mathcal{C}(\xi) \wedge \mathcal{D}(\eta) \Longrightarrow \mathcal{A}(\text{Min}\{\varrho, 2\xi - 2, \xi + \eta + 1\}),$$

$$(2^\circ) \quad \mathcal{B}(\varrho) \wedge \mathcal{C}(r) \Longrightarrow \mathcal{D}(\varrho - r),$$

$$(3^\circ) \quad \mathcal{B}(\varrho) \wedge \mathcal{D}(r) \Longrightarrow \mathcal{C}(\varrho - r), \text{ wenn alle Gewichte } \beta_i \text{ ungleich Null sind.}$$

Wenn also alle Abszissen  $\gamma_i$  verschieden sind und das RKV die Eigenschaft  $\mathcal{B}(\varrho)$  hat, dann bestimmt  $\mathcal{C}(r)$  oder  $\mathcal{D}(r)$  eindeutig die entscheidende Eigenschaft  $\mathcal{A}(p)$ . Zur weiteren *Algebraisierung*  $V = [\gamma_i^{j-1}]_{i,j=1}^r$  die VANDERMONDESche Matrix), . Dann können wir an Stelle von (2.71) in Matrix-Vektor-Form schreiben

$$\begin{array}{lcl}
 \mathcal{B}(\varrho) & \Longleftrightarrow & V_\varrho^T b = z_\varrho \\
 \mathcal{C}(\varrho) & \Longleftrightarrow & A V_\varrho = \text{diag}(c) V_\varrho \text{diag}(z_\varrho) =: W_\varrho \in \mathbb{R}^{r_\varrho} \\
 \mathcal{D}(\varrho) & \Longleftrightarrow & V_\varrho^T \text{diag}(b) A = (z_\varrho e^T - W_\varrho^T) \text{diag}(b)
 \end{array} \quad (2.72)$$

Die Koeffizientenmatrix  $A$  eines RKV ist daher eindeutig bestimmt durch  $\mathcal{C}(r)$ , wenn alle  $\gamma_i$  verschieden sind, und sie ist eindeutig bestimmt durch  $\mathcal{D}(r)$ , wenn alle Gewichte  $\beta_i$  ungleich Null sind.

**Folgerung 2.2.** GAUSS-Verfahren. *Die maximale Ordnung der äußeren Gleichung, also der Quadraturformel (2.66) ist  $p = 2r$  bei  $r$  Stützstellen nach § 2.3(c). Sie wird erreicht, wenn man die Nullstellen  $\gamma_i$ ,  $i = 1 : n$ , von  $p_{1,n}(x)$  in (2.38) wählt. Für  $\varrho = 2r$  sind aber  $\mathcal{C}(r)$  und  $\mathcal{D}(r)$  nach Satz 2.15(2°) und (3°) äquivalent. Unter der angegebenen Voraussetzung folgt dann aus (1°)*

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

**Folgerung 2.3.** BUTCHER-Verfahren. Wenn  $\varrho \geq r$  ist und alle  $\gamma_i$  verschieden sind, dann folgt Satz 2.15 (1°) und (2°)

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

Bei festem Abszissenvektor  $c$  erhält man die BUTCHER-Verfahren der Ordnung  $\varrho$  nach (2.72) aus

$$(A, b, c) = (WV^{-T}, V^{-1}z, c).$$

**Folgerung 2.4.** EHLE-Verfahren. Wenn  $\varrho \geq 2r - 2$  ist, alle  $\gamma_i$  verschieden und alle  $\beta_i$  ungleich Null sind, folgt aus Satz 2.15 (1°) und (3°)

$$\begin{aligned} \mathcal{B}(\varrho) \wedge \mathcal{D}(r) &\implies \mathcal{A}(p), \\ p &= \min\{\varrho, 2(\varrho - r) + 2, \varrho - r + r + 1\} = \min\{\varrho, 2(\varrho - r) + 2\} = \varrho. \end{aligned}$$

Bei festem Abszissenvektor  $c$  erhält man die EHLE-Verfahren der Ordnung  $\varrho \geq 2r - 2$  nach (2.72) aus

$$(A, b, c) = (B^{-1}V^{-1}(ze^T - W^T)B, V^{-1}Z, c).$$

Für die Komponenten des Abszissenvektors  $c$  werden die Wurzeln der Polynome in (2.38) gewählt:

$$\begin{aligned} \text{GAUSS-Verfahren:} \quad \varrho &= 2r, \quad p_{1,r}(x) &= [x^r(1-x)^r]^{(r)} \\ \text{Verfahren vom Typ I:} \quad \varrho &= 2r - 1, \quad xp_{2,r-1}(x) &= [x^r(1-x)^{r-1}]^{(r-1)} \\ \text{Verfahren vom Typ II:} \quad \varrho &= 2r - 1, \quad (1-x)p_{3,r-1} &= [x^{r-1}(1-x)^r]^{(r-1)} \\ \text{Verfahren vom Typ III:} \quad \varrho &= 2r - 2, \quad x(1-x)p_{4,r-2} &= [x^{r-1}(1-x)^{r-1}]^{(r-2)}. \end{aligned} \tag{2.73}$$

Das Ergebnis ist in der folgenden Tabelle zusammengefasst:

**Tabelle 2.6.**

| Typ               | Bedg. an $\gamma_i$ | Ordnung  | BUTCHER         | EHLE            | CHIPMAN         |
|-------------------|---------------------|----------|-----------------|-----------------|-----------------|
| GAUSS             | (2.73)(1°)          | $2r$     | $\otimes^A =$   | $\otimes^A$     | —               |
| RADAU I B/A       | (2.73)(2°)          | $2r - 1$ | $\otimes$       | $\otimes^{A,L}$ | —               |
| RADAU II A/B      | (2.73)(3°)          | $2r - 1$ | $\otimes^{A,L}$ | $\otimes$       | —               |
| LOBATTO III A/B/C | (2.73)(4°)          | $2r - 2$ | $\otimes^A$     | $\otimes^A$     | $\otimes^{A,L}$ |

Die CHIPMAN-Verfahren haben die Eigenschaft  $\mathcal{C}(r)$  und  $Ae_1 = \beta_1 e_1$ ,  $e_1 = [\delta^1_k]_{k=1}^r$ , woraus die Koeffizientenmatrix  $A$  vollständig berechnet werden kann. A-stabile Verfahren sind mit  $\otimes^A$  bezeichnet, L-stabile Verfahren mit dem zusätzlichen Index L.

**Beispiel 2.14.** Bei den Verfahren **Radau II A** ist  $\gamma_n = 1$ , daher stimmen die Kellerzeile und die letzte Zeile von  $A$  miteinander überein, was bei der Anwendung auf *differential-algebraische Probleme* vorteilhaft ist. Das 3-stufige Verfahren der Ordnung  $p = 5$  ist A-stabil und L-stabil, seine Daten sind in der folgenden BUTCHER-Matrix angegeben:

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

Literatur: [Hairer], [Shampine97].

## 2.5 Randwertprobleme

Gesucht ist eine Lösung  $x : [0, 1] \rightarrow \mathbb{R}^n$  des Randwertproblems

$$x'(t) = f(t, x(t)), \quad 0 \leq t \leq 1, \quad g(x(0), x(1)) = 0 \in \mathbb{R}^n, \quad (2.74)$$

wobei wir uns aus optischen Gründen und auch wegen der Implementierung auf das Einheitsintervall beschränken. Bei einem anderen Definitionsintervall muss umskaliert werden: Für ein Problem

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

ergibt die Substitution  $s = Tt$

$$x'(t) = Th(Tt, x(t)), \quad 0 \leq t \leq 1, \quad g(x(0), x(1)) = 0 \in \mathbb{R}^n.$$

(a) **Das lineare Problem** hat die Form

$$x'(t) = A(t)x(t) + c(t), \quad 0 \leq t \leq 1, \quad R_0x(0) + R_1x(1) = d \in \mathbb{R}^n \quad (2.75)$$

Der Einfachheit halber wählen wir eine *äquidistante* Unterteilung des Definitionsintervalls,

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

beginnend mit dem Index  $j = 1$ , um die Kompatibilität mit der MATLAB<sup>®</sup>-Indizierung zu wahren. Hat die *numerische Approximation* der Differentialgleichung auf dem einzelnen  $t$ -Intervall  $[t_j, t_{j+1}]$  die Form

$$\boxed{P_j y_j + Q_j y_{j+1} = r_j},$$

dann ergibt sich ein lineares Gleichungssystem

$$\mathbf{L}(\tau)Y := \begin{bmatrix} P_1 & Q_1 & 0 & \dots & 0 & 0 \\ 0 & P_2 & Q_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & P_m & Q_m \\ R_0 & 0 & \dots & \dots & 0 & R_1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y_{m+1} \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ r_m \\ d \end{bmatrix} =: R \quad (2.76)$$

für die gesuchten Werte  $y_j$ , dessen Matrix regulär sein muss.

**Beispiel 2.15.** (1°) *Trapezregel*

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

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

(2°) *Boxschema*

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

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

(3°) *Mehrzielverfahren*

(3.1°) Löse für  $j = 1, \dots, m$  das inhomogene Problem mit homogener Anfangsbedingung

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

die Lösung an der Stelle  $t_{j+1}$  sei  $r_j$ .

(3.2°) Löse für  $j = 1, \dots, m$  die  $n$  homogenen Anfangswertprobleme mit inhomogenen Anfangsbedingungen

$$X'(t) = A(t)X(t), \quad t_j \leq t \leq t_{j+1}, \quad X(t_j) = I \text{ (Einheitsmatrix)};$$

die Lösung an der Stelle  $t_{j+1}$  sei die Matrix  $V_j$ . Dann gilt

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

**(b) Im nichtlinearen Fall** wird auf die gleiche Weise ein nichtlineares Gleichungssystem aufgestellt und mit dem NEWTON-Verfahren gelöst. Wir beschränken uns auf das Mehrzielverfahren und verwenden das Flussintegral  $\Phi(t; t_0, x_0)$  aus Abschnitt 1.6. Die numerische Lösung sei wieder mit  $y$  bezeichnet.

Mehrzielverfahren:

(1°) Wähle eine moderate Anzahl  $m$  von Schießpunkten im Intervall  $[0, 1]$ ,

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

(2°) Berechne

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

durch Lösung der Anfangswertprobleme

$$x'(t) = f(t, x(t)) \quad t_j \leq t \leq t_{j+1}, \quad x(t_j) = y_j. \quad (2.77)$$

(3°) Löse das Gleichungssystem

$$y_{j+1} - \Phi(t_{j+1}; t_j, y_j) = 0, \quad j = 1 : m, \quad g(y_1, y_{m+1}) = 0, \quad (2.78)$$

mit dem NEWTON-Verfahren. Das nichtlineare Gleichungssystem (2.78) hat die Form

$$\mathbf{F}(Y) = 0 \in \mathbb{R}^{n(m+1)}, \quad \text{mit dem Knotenvektor } Y = [y_1, \dots, y_{m+1}]^T. \quad (2.79)$$

Soll dieses System mit dem NEWTON-Verfahren gelöst werden, so muss  $\text{grad } \mathbf{F}(Y)$  berechnet werden, wozu der Gradienten von  $\Phi$  an den Stellen  $(t_j, y_j)$  benötigt wird,

$$\text{grad}_v \Phi(t_{j+1}; t_j, v) = I + \int_{t_j}^{t_{j+1}} \text{grad } f(t, x(t)) \text{grad}_v \Phi(t; t_j, v) dt. \quad (2.80)$$

Das Vektorfeld zu diesem matrixwertigen Flussintegral ist

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

und die Anfangsbedingung für (2.80) ist  $W(t_j) = I$ . Es sind also im Intervall  $[t_j, t_{j+1}]$  die  $n$  Anfangswertprobleme

$$w'_k(t) = \text{grad}_x f(t, x(t))w_k(t) \in \mathbb{R}^n, \quad t_j \leq t \leq t_{j+1}, \quad w_k(t_j) = e_k, \quad k = 1 : n, \quad (2.81)$$

zu lösen, wobei  $x(t)$  die Rolle eines Parameters spielt und  $e_k \in \mathbb{R}^n$  der  $k$ -te Einheitsvektor ist.

Entscheidend für das Mehrzielverfahren ist die *simultane* Lösung der  $n + 1$  Anfangswertprobleme (2.77) und (2.80) in jedem Intervall  $[t_j, t_{j+1}]$ . Die Matrix  $\text{grad } \mathbf{F}(Y)$  hat dann die gleiche Form wie die Matrix  $\mathbf{L}(\tau)$  in (2.76) mit

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

Das Verfahren entfaltet seine volle Kraft erst, wenn die Schießpunktabszissen  $t_j$  geeignet gewählt werden, was man aber dem Computer überlassen kann. Außerdem muss das NEWTON-Verfahren durch eine geeignete Schrittweitensteuerung globalisiert werden. Die Startwerte können nicht beliebig sein, in einfachen Fällen genügt aber eine lineare Funktion, die die Randbedingungen erfüllt.

**(c) Randwertprobleme mit Parameter** Gesucht ist ein Element aus der Lösungsschar  $x(\cdot, a) : [0, 1] \rightarrow \mathbb{R}^n$  des Randwertproblems

$$x'(t; a) = f(t, x(t; a); a), \quad 0 \leq t \leq 1, \quad g(x(0; a), x(1; a); a) = 0 \in \mathbb{R}^n, \quad (2.82)$$

wobei der reelle Parameter  $a$  in einem gewissen Intervall frei variieren darf. Weil das Problem nun einen zusätzliche Freiheitsgrad hat, hängt die numerische Lösung von der gewählten Anfangsnäherung für  $x$  und  $a$  ab, von der man eine ziemlich genaue Vorstellung haben muss.

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

das zu (2.82) gehörende Flussintegral. Zur Lösung des Gleichungssystems (2.79), also jetzt

$$\mathbf{F}(V) = 0 \in \mathbb{R}^{n(m+1)+1}, \quad V = [y_1, \dots, y_{m+1}; a]^T \text{ Knotenvektor}, \quad (2.83)$$

mit dem NEWTON-Verfahren benötigt man zusätzlich die Ableitung

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

Es ist also im Intervall  $[t_j, t_{j+1}]$  zusätzlich das Anfangswertproblem

$$\begin{aligned} v_j'(t) &= \frac{\partial}{\partial a} f(t, x(t; a); a) + \text{grad}_x f(t, x(t; a); a) v_j(s), \\ v_j(t_j) &= \frac{\partial}{\partial a} (y_j)(a) \in \mathbb{R}^n, \quad v_1(t_1) = 0, \end{aligned} \quad (2.84)$$

zu lösen. Wichtig ist auch hier, dass im Mehrzielverfahren alle  $n + 2$  Anfangswertprobleme (2.77), (2.81) und (2.84) *simultan* gelöst werden. Man beachte, dass während der ganzen Iteration neben der numerischen Approximation  $V$  auch die Ableitungen  $y_{a,j}$ ,  $j = 1 : m + 1$ , berechnet werden, nur dann sind befriedigende Resultate zu verzeichnen.

Die JACOBI-Matrix  $\text{grad } \mathbf{F}(V)$  ist nun eine  $(m \cdot n, m \cdot n + 1)$ -Matrix der Blockform

$$\mathbf{L} := \begin{bmatrix} P_1 & Q_1 & 0 & \dots & 0 & 0 & H_1 \\ 0 & P_2 & Q_2 & \ddots & \ddots & 0 & H_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & H_{m-1} \\ 0 & \ddots & \ddots & \ddots & P_m & Q_m & H_m \\ g_{x_0} & 0 & \dots & \dots & 0 & g_{x_1} & g_a \end{bmatrix}, \quad (2.85)$$

deswegen wird im GAUSS-NEWTON-Verfahren die MOORE-PENROSE-Inverse  $[\text{grad } \mathbf{F}(V)]^+$  verwendet, vgl. Abschnitt 1.1(g). In jedem GAUSS-NEWTON-Schritt ist ein unterbestimmtes lineares Gleichungssystem

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

zu lösen, wozu der Algorithmus aus § 1.1(h3) verwendet werden kann. Damit das Verfahren nicht gegen die triviale Lösung konvergiert, muss eine gute Anfangsnäherung  $V_0$  vorliegen. Bei einfacheren Problemen kann auch das Box-Schema entsprechend modifiziert werden.

**Beispiel 2.16.** [Stoer]

$$\begin{aligned} x'_1 &= x_2 \\ x'_2 &= 5 \sinh(5x_1) \end{aligned}, \quad x_1(0) = 0, \quad x_1(1) = 1.$$

Als Starttrajektorie wird die gerade Verbindung der Punkte  $(0, x_1(0))$  und  $(1, x_1(1))$  gewählt. Es gilt aber  $\lim_{t \rightarrow 1.0326\dots} x_1(t) = \infty$ , daher muss die zunächst äquidistante Anfangsunterteilung des Intervalls  $[0, 1]$  dem Problem angepasst werden.

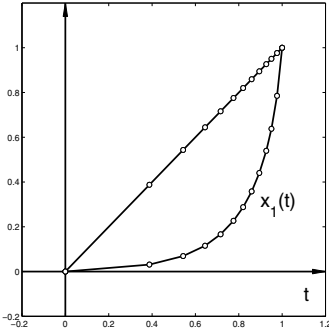


Abb. 2.19. Beispiel 2.16

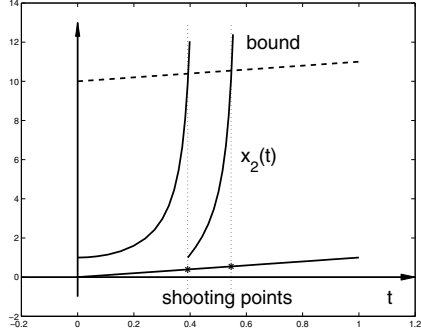


Abb. 2.20. Beispiel 2.16, Anpassung

## 2.6 Periodische Probleme

**(a) Probleme mit bekannter Periode** Gesucht ist eine  $T$ -periodische Lösung  $x : \mathbb{R} \rightarrow \mathbb{R}^n$  des Randwertproblems

$$x'(t) = f(t, x(t)), \quad 0 \leq t \leq T, \quad x(0) = x(T) \in \mathbb{R}^n. \quad (2.86)$$

Hier ist mit  $x(t)$  immer auch  $x(t + \alpha)$ ,  $\alpha \in \mathbb{R}$ , eine  $T$ -periodische Lösung, für die Eindeutigkeit ist daher eine *Phasenbedingung* notwendig, aber trotzdem bleibt das Problem numerisch instabil. Mögliche Phasenbedingungen sind etwa

- (1°)  $p(x(0)) := x_k - \eta = 0, \quad \eta \neq 0;$
- (2°)  $p(x(0)) := f_k(x(0)) = 0 \implies x'_k(0) = 0.$

Damit ergeben sich aber  $n + 1$  Randbedingungen für  $n$  unbekannte Funktionen. Wenn ein Punkt auf dem Orbit bekannt ist, lassen sich befriedigende Ergebnisse auch mit einem guten Löser für Anfangswertprobleme erzielen.

**(b) Probleme mit unbekannter Periode** Wenn die Periode  $T$  unbekannt ist, dann empfiehlt sich eine Transformation auf ein parameterabhängiges Problem mit bekannter Periode, z.B. hat die Lösung  $\tilde{x}$  von

$$\tilde{x}'(s) = T f(Ts, \tilde{x}(s)), \quad \tilde{x}(0) = \tilde{x}(1) \quad (2.87)$$

die Periode Eins in  $s$  und  $x(t) = \tilde{x}(t/T)$  ist eine Lösung von (2.86) mit der Periode  $T$ . Die weitere Behandlung des Problems verläuft wie in Abschnitt 2.5(c) für parameterabhängige Probleme. Wegen der Randbedingung in (2.87) und dem in § 2.5(a) beschriebenen Sachverhalt lässt sich das relativ einfache Boxschema hier nicht anwenden. In den nachfolgenden Beispielen wird das Mehrzielverfahren mit *fester* Intervallunterteilung verwendet. Als Starttrajektorie dient die Lösung eines *Anfangswertproblems* mit *geschätztem* Startwert.

**Beispiel 2.17.** Nervenmembranmodell [Deuffhard84]

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

Umwandlung in ein parameterabhängiges Problem mit Periode  $T = 1$ :

$$\begin{aligned}x'_1 &= 3T(x_2 + x_1 - \frac{1}{3}x_1^3 + \lambda) \\ x'_2 &= -\frac{T}{3}(x_1 - 0.7 + 0.8x_2).\end{aligned}$$

Das Anfangswertproblem (2.84) hat die Form

$$\begin{aligned}v'_1 &= 3(x_2 + x_1 - \frac{1}{3}x_1^3 + \lambda) + 3T(1 - x_1^2)v_1 + 3Tv_2 \\ v'_2 &= -\frac{1}{3}(x_1 - 0.7 + 0.8x_2) - \frac{T}{3}v_1 - \frac{0.8T}{3}v_2.\end{aligned}$$

Testproblem :  $\lambda = -1$ , Startwerte  $(x_1^0, x_2^0, T^0) = (3, 1.5, 12)$ .

**Beispiel 2.18.** Wärmeleitproblem [Deuffhard84]

$$\begin{aligned}\dot{u}_1 &= -\sigma(u_1 - u_2) \\ \dot{u}_2 &= u_1(r - u_3) - u_2 \\ \dot{u}_3 &= u_1u_2 - bu_3.\end{aligned}$$

Umwandlung in ein parameterabhängiges Problem mit Periode  $T = 1$ :

$$\begin{aligned}x'_1 &= -\sigma T(x_1 - x_2) \\ x'_2 &= T[x_1(r - x_3) - x_2] \\ x'_3 &= T(x_1x_2 - bx_3).\end{aligned}$$

Das Anfangswertproblem (2.84) hat die Form

$$\begin{aligned}v'_1 &= -\sigma(x_1 - x_2) & -\sigma T(v_1 + v_2) \\ v'_2 &= x_1(r - x_3) - x_2 & + T[(r - x_3)v_1 - v_2 - x_1v_3] \\ v'_3 &= x_1x_2 - bx_3 + x_2v_1 & + T(x_1v_2 - bv_3).\end{aligned}$$

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

**Beispiel 2.19.** ARENSTORF-Orbits [Arenstorf] Beim entarteten Dreikörperproblem sind drei Körper (Erde, Mond, Satellit) mit den Massen  $m_1$ ,  $m_2$  und  $m_3 = 0$  gegeben, sowie die folgenden Vereinfachungen:

(1°) Erde, Mond und Satellit bewegen sich in einer Ebene.

- (2°) Der Abstand Erde–Mond ist konstant gleich Eins.  
 (3°) Der Einfluss der übrigen Himmelskörper wird vernachlässigt.

Wählt man die Achse Erde–Mond als x-Achse mit dem gemeinsamen Schwerpunkt der beiden Himmelskörper als Ursprung, einen Umlauf des Mondes als Zeiteinheit und setzt  $\mu = m_2/(m_1 + m_2) \sim 1/81,45$  (relative Mondmasse),  $\mu' = 1 - \mu$ , so ergibt sich ein System von *zwei* Differentialgleichungen, vgl. § 6.2(e). Für die Umwandlung in ein 1-periodisches Problem mit der ursprünglichen Periode  $T$  wird auf das entsprechende MATLAB®-Programm verwiesen.

**Beispiel 2.20.** Die inhomogene DUFFINGSche Gleichung

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

ist ein Modellproblem zum Studium von Periodenverdopplung, Übergang zum Chaos und (im homogenen Fall) für Bifurkation periodischer Lösungen [Seydel94]. *Harmonische* Lösungen haben hier die gleiche Periode  $T = 2\pi/\omega$  wie die rechte Seite, im andern Fall entstehen „seltsame Attraktoren“ (strange attractors). Transformation auf ein parameterabhängiges System mit Periode Eins durch Substitution von  $t = Ts$ ,  $T = 2\pi/\omega$ , ergibt wie oben mit  $y_1(s) = u(t)$

$$y_1' = Ty_2, \quad y_2' = -T(\alpha y_2 + \beta y_1 + \gamma y_1^3 - \delta \cos(2\pi s)).$$

In Abb. 2.24 ist  $\alpha = 0.2$ ,  $\beta = 0$ ,  $\gamma = 1$ , und am Anfang  $\delta = 5$ . Für eine Starttrajektorie wird zuerst ein Anfangswertproblem gelöst, anschließend einfache Fortsetzung bis  $\delta = 7$ . Geschätzte Anfangsperiode ist  $T = 12$ , finale Periode  $T = 10.2209$ .

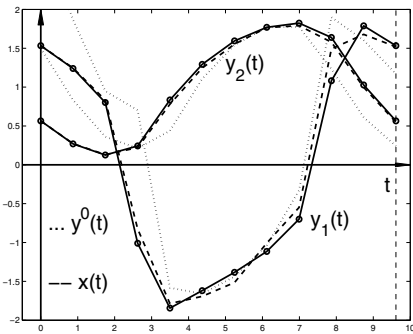


Abb. 2.21. Beispiel 2.17

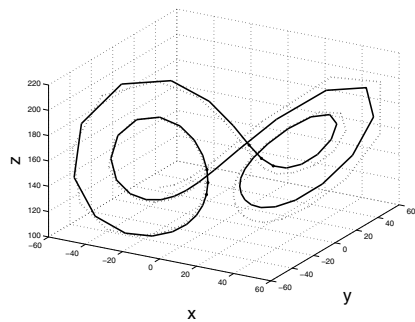


Abb. 2.22. Beispiel 2.18

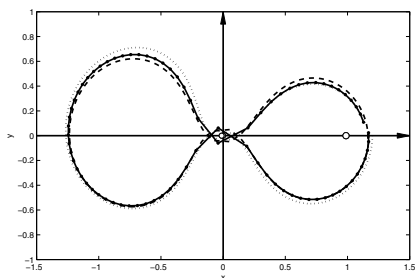


Abb. 2.23. Beispiel 2.19

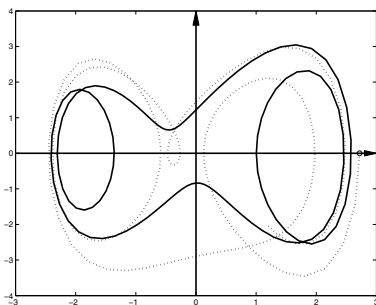


Abb. 2.24. Beispiel 2.20

## 2.7 Differential-algebraische Probleme

In der Mechanik wird ein Extremalproblem im Regelfall über die Variationsgleichungen gelöst – vgl. § 4.1. Wenn zusätzliche Zwangsbedingungen vorliegen und diese Bedingungen Gleichungsrestriktionen sind, dann entsteht ein *differential-algebraisches* Problem, das nach einer ev. Diskretisierung der Raumveränderlichen (*Linienmethode*) zum einen Teil aus einem System von gewöhnlichen Differentialgleichungen und zum andern aus einem System von algebraischen Gleichungen besteht, wobei die letzteren in vielen Fällen Flächen im Raum beschreiben (holonome Zwangsbedingungen), auf denen die Lösung des Differentialsystems lebt. Letztlich ist man dann mit einem *Randwertproblem* oder einem *Anfangswertproblem* konfrontiert, das bei anspruchsvolleren Aufgaben hochgradig nichtlinear ist. Numerische Verfahren zur Lösung nichtlinearer Randwertprobleme benötigen eine *konsistente Starttrajektorie*, die oft schwierig zu finden ist. Wenn ein solches Problem aber künstlich in ein *Kontrollproblem* umgewandelt wird, können moderne numerische Methoden der Optimierung eingesetzt werden, die ohne Starttrajektorie auskommen – vgl. §4.4. Für differential-algebraische Probleme, die reine Anfangswertprobleme sind, wurden spezielle RUNGE-KUTTA-Verfahren entwickelt, von denen in diesem Abschnitt einige behandelt werden. Das Problem der konsistenten Anfangswerte tritt hier ebenfalls auf, es ist aber im Regelfall nur ein nichtlineares Gleichungssystem und mit den üblichen Methoden zu lösen. Für praktische Beispiele zu diesen Verfahren verweisen wir auf den Abschnitt 11.3 über *Mehrkörperprobleme*.

Im Folgenden sei  $(x, y)$  die theoretische Lösung und  $(u, v)$  ihre numerische Approximation.

**(a) Problemstellung** Wir betrachten zunächst ein *singuläres* Anfangswertproblem in der separierten Form

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)) \in \mathbb{R}^n, \quad (x(0), y(0)) = (x_0, y_0), \\ \varepsilon y'(t) &= g(t, x(t), y(t)) \in \mathbb{R}^m, \quad 0 \leq \varepsilon \ll 1 \end{aligned} \quad (2.88)$$

mit hinreichend glatten Daten. Fasst man die abhängigen Veränderlichen in  $z(t) = [x(t), y(t)]^T$  zusammen, so kann das Differentialsystem auch in der Form

$$Mz'(t) = F(t, z(t)) \in \mathbb{R}^{n+m}, \quad z(t) = [x(t), y(t)]^T \quad (2.89)$$

geschrieben werden mit einer Matrix  $M \in \mathbb{R}^{n+m}_{n+m}$ , die für  $\varepsilon = 0$  *singulär* wird. Probleme von der Form  $M(x(t))x'(t) = F(t, x(t))$  werden am Besten umgewandelt in ein System

$$x' = y, \quad M(x)y - F(x) = 0. \quad (2.90)$$

**Voraussetzung 2.2.** (1°) Das Problem (2.88) sei für  $t \in [0, T]$ ,  $0 < T$ , *eindeutig lösbar*.

(2°) Für  $\varepsilon = 0$  sei  $\nabla_y g(x, y)$  in der Nähe der Lösung  $(x, y)$  *regulär*.

Ist  $\varepsilon = 0$ , dann spricht man von einem *differential-algebraischen* Problem (DA-Problem), und die Anfangswerte müssen dann *konsistent* sein, d.h. es muss  $g(x_0, y_0) = 0$  gelten. Bei der *numerischen Lösung* müssen solche Startwerte zunächst berechnet werden oder wenigstens näherungsweise bekannt sein. Unter Vor. 2.2(2°) heißt das DA-Problem *vom Index 1*,  $g$  ist in der Nähe der Lösung auflösbar nach  $y$ ,  $y(t) = G(t, x(t))$ , und man erhält durch Einsetzen *theoretisch* ein gewöhnliches Anfangswertproblem mit dem Differentialsystem  $x'(t) = f(t, x(t), G(t, x(t)))$ .

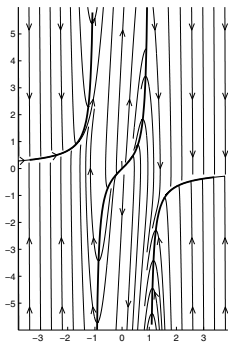
**Beispiel 2.21.** VAN DER POL-Gleichung.

Der *lineare Oszillator*  $\ddot{x} + \alpha\dot{x} + x = 0$  ist gedämpft für  $\alpha > 0$  und instabil für  $\alpha < 0$ . Wird der Parameter  $\alpha$  durch  $\mu(x^2 - 1)$ ,  $\mu > 0$ , ersetzt, so erhält man für große  $|x(t)|$ -Werte eine Dämpfung und für kleine  $|x(t)|$  eine Verstärkung. Umwandlung in ein System erster Ordnung ergibt

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

setzt man  $x_1 = y_1$ ,  $y_2 = \mu x_2$ ,  $s = t/\mu$  und schreibt anschließend  $\mu^2 = 1/\varepsilon$ , so ergibt sich nach einer Neubenennung

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



**Abb. 2.25.** VAN DER POL-Gleichung

In Abb. 2.25 ist das Phasenporträt für  $\varepsilon = 0.05$  und die Kurve  $(1 - x^2)y - x = 0$  eingezeichnet.

(b) DA-Probleme werden vorwiegend mit speziellen RUNGE-KUTTA-Verfahren gelöst, obwohl auch Mehrstellenverfahren angewendet werden können insbesondere die Rückwärtsdifferenzenverfahren aus § 2.4(i)(4°). Ist zunächst  $\varepsilon > 0$ , dann folgt mit den Bezeichnungen aus § 2.4(d) bei einem *gemeinsamen* Verfahren für die *getrennten Gleichungen*

$$\begin{aligned}
U(t) &= e \times u(t) + \tau(A \times I)F(t, U(t), V(t)) && \in \mathbb{R}^{r \cdot n} \\
\varepsilon V(t) &= \varepsilon e \times v(t) + \tau(A \times I)G(t, U(t), V(t)) && \in \mathbb{R}^{r \cdot m} \\
u(t + \tau) &= u(t) + \tau(b \times I)^T F(t, U(t), V(t)) && \in \mathbb{R}^n \\
\varepsilon v(t + \tau) &= \varepsilon v(t) + \tau(b \times I)^T G(t, U(t), V(t)) && \in \mathbb{R}^m;
\end{aligned} \tag{2.91}$$

dabei sind  $U(t)$  und  $V(t)$  die Vektoren der Zwischenstufen. Ist nun die Matrix  $A$  des Verfahrens *regulär*, dann folgt aus der zweiten Gleichung

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

und bei Einsetzen in die letzte Gleichung lässt sich der Parameter  $\varepsilon$  *streichen*. Insgesamt erhält man auf diese Weise eine *direkte Approximation* des DA-Problems mit RUNGE-KUTTA-Verfahren,

$$\begin{aligned}
U(t) &= e \times u(t) + \tau(A \times I)F(t, U(t), V(t)) && \in \mathbb{R}^{r \cdot n} \\
0 &= G(t, U(t), V(t)) && \in \mathbb{R}^{r \cdot m} \\
u(t + \tau) &= u(t) + \tau(b \times I)^T F(t, U(t), V(t)) && \in \mathbb{R}^n \\
v(t + \tau) &= (1 - b^T A^{-1} e)v(t) + (b \times I)^T (A^{-1} \times I)V(t) && \in \mathbb{R}^m,
\end{aligned} \tag{2.92}$$

und es gilt  $R(\infty) = 1 - b^T A^{-1} e$  für die Stabilitätsfunktion; vgl. (2.58). Allerdings wird bei diesem Typ von Verfahren die algebraische Nebenbedingung  $g(u, v) = 0$  i.d.R. nur näherungsweise erfüllt. Dieser Nachteil wird aber beseitigt, wenn die letzte Gleichung in (2.92) durch die Forderung  $g(u_{n+1}, v_{n+1}) = 0$  ersetzt wird. Dieser *indirekte* Verfahrenstyp stellt eine konsistente Approximation von  $x'(t) = f(x, G(x))$  bei Systemen vom Index 1 dar. Wenn *zusätzlich* zur Regularität von  $A$  noch die letzte Zeile von  $A$  mit dem Vektor  $b$  der Gewichte in der Kellerzeile übereinstimmt (*steif genaue Verfahren*), dann ist  $g(u_{n+1}, v_{n+1}) = 0$  wegen der zweiten Gleichung in (2.92) automatisch erfüllt, und die letzte Gleichung in (2.92) entfällt. Zum Beispiel haben die in § 2.4(j) beschriebenen RUNGE-KUTTA-Verfahren vom Typ RA-DAU II A die genannten Eigenschaften und eignen sich daher in besonderer Weise zur Lösung von DA-Problemen.

Wenden wir ein RUNGE-KUTTA-Verfahren auf das Differentialsystem  $M x' = f(t, x)$  mit *regulärer* Matrix  $M$  an, so erhalten wir in der gleichen Weise wie beim Übergang von (2.91) zu (2.92) die Vorschrift

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

und diese Vorschrift bleibt auch bei *singulärer* Matrix  $M$  sinnvoll, aber dann hängt das Verfahren von der Kondition der Matrix  $I \times M - \tau(A \times I)$  also insbesondere von der Schrittweite  $\tau$  ab.

**(c) Reguläre Matrizenpaare** Ist  $(\lambda, u)$  charakteristisches Paar des *verallgemeinerten Eigenwertproblems*

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

dann ist  $x(t) = e^{\lambda t}u$  Lösung des Differentialsystems

$$Bx' + Ax = c(t) \in \mathbb{R}^n \quad (2.94)$$

mit  $c(t) \equiv 0$ . Ist hier z.B.  $A = B$ ,  $\det(A) = 0$ , dann ist  $A + \lambda B$  singulär für alle  $\lambda \in \mathbb{R}$ , daher wird bei linearen Systemen (2.94) stillschweigend vorausgesetzt, dass das Matrizenpaar „matrix pencil“  $(A, B)$  *regulär* ist mit  $\det(A + \lambda B) \neq 0$ .

**Satz 2.16.** (WEIERSTRASS, KRONECKER) Für reguläre Paare  $(A, B)$  gibt es reguläre Matrizen  $P, Q$ , so dass gilt

$$PAQ = \begin{bmatrix} C & O \\ O & I \end{bmatrix}, \quad PBQ = \begin{bmatrix} I & O \\ O & T \end{bmatrix}; \quad (2.95)$$

dabei ist  $T = \text{diag}(T_1, \dots, T_k)$  eine Blockdiagonalmatrix mit Blöcken  $T_i \in \mathbb{R}^{n_i}_{n_i}$  der in § 1.1(c3) beschriebenen Form und  $n_1 + \dots + n_k = n$ .

Beweis [Hairer], Bd. II, § 6.5.

Multipliziert man (2.94) mit  $P$  und verwendet die Zerlegung

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

dann ergeben sich zwei getrennte System für  $y$  und  $z$ ,

$$y' = Cy + f(t), \quad Tz' + z = g(t). \quad (2.96)$$

**(d) Differentialindex** Das zweite System in (2.96) muss rekursiv gelöst werden. Ist z.B.  $k = 1$  in Satz 2.16 und  $T = T_1 \in \mathbb{R}^{m_m}_m$ , dann geht man von der letzten Zeile  $z_m = g_m(t) \in \mathbb{R}$  aus und hat dann sukzessive die Komponenten  $z_i(t)$ ,  $i = m - 1 : 1$ , aus  $z_i(t) = g_i(t) - z_{i+1}^{(i+1)}(t)$  zu berechnen, wozu die Ableitungen  $g_m^{(m)}, \dots, g_2^{(1)}$  benötigt werden. Mit diesen Ableitungen lässt sich  $Tz' + z = g(t)$  auch als *explizites* System schreiben,

$$z_{i+1}^{(i+1)} + z_i^{(i)} = g_i^{(i)}(t), \quad i = 1, \dots, m - 1, \quad z_m^{(m)} = g_m^{(m)}(t). \quad (2.97)$$

Allgemein ist der *Differentialindex* die Anzahl der notwendigen Ableitungen, um ein implizites Differentialsystem  $F(t, x'(t), x(t)) = 0$  *analytisch* in ein explizites System umzuformen. Das explizite System  $x'(t) = f(t, x(t))$  hat definitionsgemäß den Index Null, und z.B. das System (2.97) hat den Index  $m$ , weil  $m$ -mal abgeleitet werden muss.

*System mit Index 1.* Wenn die Matrix  $\nabla_y g(x, y)$  in der Nähe der Lösung  $(x, y)$  von

$$x'(t) = f(t, x(t), y(t)), \quad g(t, x(t), y(t)) \quad (2.98)$$

regulär ist, dann folgt aus  $0 = \nabla_x g(x, y)x' + \nabla_y g(x, y)y'$  zusammen mit (2.98)(1°) das explizite System

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

Das System (2.98) hat in diesem Fall den Index 1.

*System mit Index 2.* Es sei  $\nabla_y g(x, y)$  in der Nähe der Lösung von (2.98) singulär. Aus  $g(x, y) = 0$  folgt  $h(x, y) := \nabla_x g(x, y) f(x, y) = 0$  und

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

Wenn  $\nabla_y h(x, y)$  regulär ist, dann ist  $x' = f(x, y)$ ,  $h(x, y) = 0$  ein System mit Index 1. Das System (2.98) hat in diesem Fall den Index 2. Nach Auflösung von  $\nabla_x h(x, y)x' + \nabla_y h(x, y)y' = 0$  erhält man wieder ein explizites System erster Ordnung,

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

*System mit Index 3.* Ist sowohl  $\nabla_y g(x, y)$  als auch  $\nabla_y h(x, y)$  in der Nähe der Lösung von (2.98) singulär, dann folgt  $k(x, y) := \nabla_x h(x, y) f(x, y) = 0$  aus  $h(x, y) = 0$ . Ist nun  $\nabla_y k(x, y)$  regulär, dann hat das System  $x' = f(x, y)$ ,  $k(x, y) = 0$  den Index 1. Das System (2.98) hat in diesem Fall den Index 3.

(e) In jüngerer Zeit wurden auch **halb-explizite** RUNGE-KUTTA-Verfahren zur Lösung von DA-Problemen

$$x'(t) = f(x(t), y(t)), \quad g(x(t)) = 0 \quad (2.99)$$

vorgeschlagen,

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

Wenn die Koeffizientenmatrix  $A$  eine *untere Dreiecksmatrix* mit  $\text{diag}(A) = 0$  und  $\beta_r \neq 0$  ist, dann ergibt sich für einen Zeitschritt die folgende Rechenvorschrift:

|                                                                                                                                                                                                                                                                                                                                  |         |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|
| Setze $u_1 = u$<br>Für $i = 2 : r$ setze<br>$u_i = u + \tau \sum_{j=1}^{i-1} \alpha_{ij} f(u_j, v_j)$<br>berechne $v_{i-1}$ mit dem NEWTON-Verfahren aus<br>$g(u_i) = 0$<br>Setze $u(t + \tau) = u + \tau \sum_{i=1}^r \beta_i f(u_i, v_i)$<br>Berechne $v_r = v(t + \tau)$ mit dem NEWTON-Verfahren aus<br>$g(u(t + \tau)) = 0$ | (2.101) |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|

**Satz 2.17.** (1°) Das Problem (2.99) sei für  $t \in [0, T]$ ,  $0 < T$ , eindeutig lösbar.

(2°) Für die Startwerte gelte  $g(x_0) = 0$ ,  $\nabla g(x_0) f(x_0, y_0) = 0$ .

(3°) In der Nähe der Lösung sei  $\nabla g(x) \nabla_y f(x, y)$  regulär (System mit Index 2).

(4°) In der Matrix  $A$  und dem Vektor  $b$  in (2.100) sei

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

Dann haben die Systeme in (2.101) für hinreichend kleine  $\tau$  eine lokal eindeutige Lösung.

Beweis [Brasey92], [Brasey93].

**Beispiel 2.22.** HEM4, 5-stufiges Verfahren der Ordnung  $p = 4$  nach [Brasey92].

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

Nichtstationäre NAVIER-STOKES-Gleichungen in Geschwindigkeits-Druck-Form ( $(\underline{u}, p)$ -Form) erweisen sich als DAE-Problem (2.99) nach einer Diskretisierung in den Raumveränderlichen mit einer Finite-Element-Methode; s. § 9.6. Die Geschwindigkeit  $\underline{u}$  spielt die Rolle von  $x$  und der skalare Druck  $p$  die Rolle von  $y$ . Die Inkompressibilitätsbedingung  $\text{div } \underline{u} = 0$  entspricht der „algebraischen“ Bedingung  $g(x) = 0$ . Deswegen bieten sich hier halb-implizite RKV

an. Die numerischen Ergebnisse sind recht akzeptabel, insbesondere weil die Nebenbedingung ein *lineares* Gleichungssystem mit *konstanter* Matrix für die Komponenten des Drucks ergibt.

## 2.8 Hinweise zu den MATLAB®-Programmen

### KAPITEL02/SECTION\_1\_2\_3

Abbildungen zu den Abschnitten 2.1 und 2.2

```
demo1.m      Vier Gauss-Formeln im Intervall
demo2.m      Gauss- und Bell-Formeln im allgem. Dreieck
bell.m       Exakte Integration eines Polynoms
              in einem allgemeinen Dreieck
gauss_1.m:    Gauss-Legendre-Integration
gauss_2/3/4.m: Gauss-Integration, suboptimal, drei Faelle
gauss_t5.m:   Gauss-Integration, Ordnung n = 5, allgem. Dreieck
divdif.m:     Verallgemeinerte dividierte Differenzen
```

### KAPITEL02/SECTION\_4: Anfangswertprobleme

Abbildungen zu Abschnitt 2.4 und Stab.-bereiche

```
demo1.m      Arenstorf-Orbits mit dopri.m,
dopri.m      MATLAB-Version der FORTRAN-Version von HAIRER
dreik_a.m    Differentialsystem des eingeschaenkten
              Dreikoerperproblems
stab_region.m Stabilitaetsbereiche von Einschnitt-
              verfahren
```

### KAPITEL02/SECTION\_5: Randwertprobleme

```
adapt01.m    Anpassung der Schiesspunkte in Bsp. 2.16
box.m        Box-Schema fuer Newton-Verfahren
bsp01.m      Beispiel Stoer-Bulirsch, Par. 7.3, Bsp. 1
demo1.m      Masterfile fuer Mehrzielverfahren
mehrziel.m   Mehrzielverfahren fuer Newton-Verfahren
newton.m     Quasi-globales Newton-Verfahren
```

### Kapitel02/SECTION\_6: Periodische Probleme

```
bsp01.m      Nervenmembranmodell
bsp02.m      Waermeleitproblem
bsp03.m      Arenstorf-Orbit I
bsp04.m      Duffingsche Gleichung
demo1.m      Masterfile fuer Mehrzielverfahren
demo2.m      Periodische Loesung d. Duffingschen Gleichung
demo3.m      Weitere Loesungen d. Duffingschen Gleichung
mehrziel_p.m Mehrzielverfahren fuer Newtonverfahren
              und Probleme mit unbekannter Periode
newton_p.m   Quasiglobales Newtonverf. fuer periodische Probleme
```

Mathematische Methoden zur Mechanik  
Ein Handbuch mit MATLAB®-Experimenten  
Gekeler, E.W.  
2010, XVIII, 645 S. 217 Abb., Softcover  
ISBN: 978-3-642-14252-9