W. Oevel

Numerik Dynamischer Systeme

Skript zur Vorlesung, JWG-Universität Frankfurt, Wintersemester 96/97

Inhalt

Literatur:

J. Stoer und R. Bulirsch: *Numerische Mathematik 2*, Springer 1990.

H.R. Schwarz: *Numerische Mathematik*, Teubner 1993.

P. Deuflhard und F. Bornemann: *Numerische Mathematik II: Integration gewöhnlicher Differentialgleichungen*, de Gruyter 1994.

J.C. Butcher: *The Numerical Analysis of Ordinary Differential Equations*, Wiley 1987.

E. Hairer, S.P. Nørsett und G. Wanner: *Solving Ordinary Differential Equations I: Nonstiff Problems*, Springer 1993.

A.M. Stuart und A.R. Humphries: *Dynamical Systems and Numerical Analysis*, Cambridge University Press 1996.

Kapitel 0

Vorbemerkungen

Ein (autonomes) dynamisches System ist ein gekoppeltes System gewöhnlicher Differentialgleichungen

$$\frac{d}{dt}\begin{pmatrix} y_1(t) \\ \vdots \\ y_N(t) \end{pmatrix} = \begin{pmatrix} f_1(y_1,\ldots,y_N) \\ \vdots \\ f_N(y_1,\ldots,y_N) \end{pmatrix} , \quad \text{kurz} \quad \frac{dy}{dt} = f(y)$$

mit $y \in \mathbb{R}^N$ und einem "Vektorfeld" $f : \mathbb{R}^N \to \mathbb{R}^N$. Es geht um das Anfangswertproblem

$$\frac{dy}{dt} = f(y) \quad \text{mit vorgegebenem Startwert} \quad y(t_0) = y_0 \ .$$

Ein mögliches Lösungsverfahren (das klassische Runge-Kutta-Schema ≈ 1900) besteht aus einem "Zeitschritt": $y(t_0) \to y(t_0 + h)$ mit der "Schrittweite" h: mit den "Zwischenstufen"

$$k_1 = f(y_0)\ , \quad k_2 = f(y_0 + \tfrac{h}{2}\, k_1)\ , \quad k_3 = f(y_0 + \tfrac{h}{2}\, k_2)\ , \quad k_4 = f(y_0 + hk_3)$$

berechnet man

$$y_1 \ := \ y_0 \ + \ \frac{h}{6}\ (k_1 + 2\,k_2 + 2\,k_3 + k_4)$$

als Approximation von $y(t_0 + h)$, wobei $\|y_1 - y(t_0 + h)\| = O(h^5)$ ("Konvergenzordnung 4") gilt. Der nächste Zeitschritt ist analog mit y_1 statt y_0. Man erhält so $y_2 \approx y(t_0 + h + h)$ usw.

Ziel der Vorlesung: Theorie für eine große Familie solcher Verfahren (Runge-Kutta-Theorie).

Speziell soll die Anwendung auf “Hamilton-Systeme”

$$\frac{d}{dt}\begin{pmatrix} q_1 \\ \vdots \\ q_n \\ p_1 \\ \vdots \\ p_n \end{pmatrix} = \begin{pmatrix} \partial H/\partial p_1 \\ \vdots \\ \partial H/\partial p_n \\ -\partial H/\partial q_1 \\ \vdots \\ -\partial H/\partial q_n \end{pmatrix}$$

betrachtet werden, welche durch eine “Hamiltonfunktion” $H : \mathbb{R}^{2n} \to \mathbb{R}$ erzeugt werden. Der Spezialfall $H(q,p) = \frac{1}{2}\langle p,p\rangle + V(q)$ mit $q,p \in \mathbb{R}^n$ liefert die Newtonschen Bewegungsgleichungen

$$\underbrace{\frac{d^2q}{dt^2}}_{\text{Beschleunigung}} = \underbrace{-\nabla_q V(q)}_{\text{Kraftfeld}} .$$

Hamilton-Systeme zeigen ein spezielles Verhalten, dies sollte von der Numerik berücksichtigt werden (“symplektische Integration”, seit $\approx$ 1983).

Bemerkung: Die Differentialgleichung $dy/dt = f(y)$ ist äquivalent zur Integralgleichung

$$y(t_0+h) = y(t_0) + \int_{t_0}^{t_0+h} f(y(t))\, dt ,$$

d.h., die Lösung entspricht einer Integration (allerdings mit unbekanntem Integranden). Daher gibt es bei den Verfahren starke Anleihen bei der numerischen Quadratur (Integration).

Kapitel 1

Polynominterpolation

Interpolationsaufgabe: zur Wertetabelle $(x_0, y_0), \ldots, (x_n, y_n) \in \mathbb{R}^2$ mit paarweise verschiedenen x_i finde ein Polynom $P_n(x)$ vom Grad $\leq n$, das

$$P_n(x_i) \;=\; y_i \;, \quad i = 0, \ldots, n$$

erfüllt.

Satz 1.1:

Es existiert ein eindeutiges Interpolationspolynom $P_n(x)$. Eine mögliche Darstellung ist

$$P_n(x) \;=\; \sum_{j=0}^{n} y_j \, L_j(x) \quad \textit{(Lagrange-Darstellung)}$$

mit den ***Lagrange-Polynomen***

$$L_j(x) = \frac{(x-x_0)\;\cdots\;(x-x_{j-1})(x-x_{j+1})\;\cdots\;(x-x_n)}{(x_j-x_0)\cdots(x_j-x_{j-1})(x_j-x_{j+1})\cdots(x_j-x_n)} \;,\; j = 0, \ldots, n \;.$$

Beweis: Mit dem Ansatz $P_n(x) = a_0 + a_1 x + \cdots + a_n x^n$ stellen die Interpolationsbedingungen das Gleichungssystem

$$\begin{pmatrix} 1 & x_0 & x_0^2 \ldots & x_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 \ldots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix}$$

dar. Die Vandermonde-Matrix ist für paarweise verschiedene x_i invertierbar, womit Existenz und Eindeutigkeit folgt. Die Lagrange-Polynome erfüllen offensichtlich

$$L_j(x_i) \;=\; \begin{cases} 1 & \text{für} \quad i = j \;, \\ 0 & \text{für} \quad i \neq j \;, \end{cases}$$

so daß $P_n(x) := \sum_{j=0}^{n} y_j \, L_j(x)$ die Interpolationsbedingungen $P_n(x_i) = y_i$ erfüllt.

Q.E.D.

Angenommen, (x_i, y_i) ist Wertetabelle einer glatten Funktion $f : \mathbb{R} \to \mathbb{R}$, d.h., $y_i = f(x_i)$, $i = 0, \ldots, n$. Wie gut approximiert das Interpolationspolynom P_n die Funktion f zwischen den Stützstellen?

Satz 1.2: (Interpolationsfehler)

Sei $f \in C^{n+1}(\mathbb{R}, \mathbb{R})$ ($n+1$-fach stetig differenzierbar von $\mathbb{R}$ nach $\mathbb{R}$). Für das Interpolationspolynom $P_n(x)$ zu $(x_0, f(x_0)), \ldots, (x_n, f(x_n))$ gilt

$$f(x) - P_n(x) \;=\; \frac{f^{(n+1)}(\eta)}{(n+1)!} \, (x - x_0)(x - x_1) \cdots (x - x_n)$$

mit einem Zwischenwert

$$\eta = \eta(x_0, \ldots, x_n, x) \;\in\; \Big(\min(x_0, \ldots, x_n, x) \, , \, \max(x_0, \ldots, x_n, x) \Big) \, .$$

Beweis: siehe jedes beliebige Buch zur “Einführung in die Numerik”.

Kapitel 2

Einige Quadraturverfahren

Gegeben sei ein Integrationsintervall $[x, x+h]$. Ziel: finde **Knoten** (**Stützstellen**) $x_0, \ldots, x_n$ und **Gewichte** $b_0, \ldots, b_n$, so daß der **Quadraturfehler**

$$\int_x^{x+h} f(\xi)\, d\xi \; - \; h \sum_{j=0}^{n} b_j\, f(x_j)$$

"möglichst klein" ist. Die Werte $x_0, \ldots, x_n, b_0, \ldots, b_n$ werden als die "Daten der Quadraturformel"

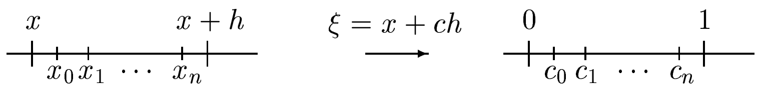
$$Q_n[f] \;=\; h \sum_{j=0}^{n} b_j\, f(x_j)$$

bezeichnet. Idee zur Bestimmung der Daten: fordere, daß die Quadraturformel exakt ist für alle Polynome bis zu möglichst hohem Grad.

Bemerkung 2.1: *Es ist sinnvoll, das Integrationsintervall $\xi \in [x, x+h]$ mittels $\xi(c) = x + ch$ auf das Standardintervall $c \in [0,1]$ abzubilden:*

$$\int_x^{x+h} f(\xi)\, d\xi \;=\; h \int_0^1 f(x+ch)\, dc\ .$$

Den Stützstellen $x_i \in [x, x+h]$ ensprechen dann $c_i \in [0,1]$ mit $x_i = x + c_i h$:



Die c_i beschreiben die "Verteilung der Stützstellen" unabhängig von der Lage x und der Länge h des Integrationsintervalls. Also: alternativ zu $x_0, \ldots, x_n, b_0, \ldots, b_n$ suche nach $c_0, \ldots, c_n, b_0, \ldots, b_n$, so daß

$$\int_x^{x+h} f(\xi)\, d\xi \;=\; \underbrace{h \sum_{j=0}^{n} b_j\, f(x + c_j h)}_{Q_n[f]} \;+\; \textit{Fehler}[f]\ .$$

2.1 Newton-Cotes-Formeln

Die Stützstellen $x_0, \dots, x_n$ bzw. $c_0, \dots, c_n$ (mit $x_i = x + c_i h$) sind vorgegeben. Aufgabe: finde $b_0, \dots, b_n$.

Satz 2.2: (Newton-Cotes-Quadratur)

Mit den Lagrange-Polynomen

$$L_j(\xi) \;=\; \prod_{\substack{k=0\\k\neq j}}^{n} \frac{\xi - x_k}{x_j - x_k} \quad \text{bzw.} \quad L_j^*(c) \;=\; \prod_{\substack{k=0\\k\neq j}}^{n} \frac{c - c_k}{c_j - c_k}$$

zu $x_0, \dots, x_n$ bzw. $c_0, \dots, c_n$ wähle

$$b_j \;=\; \frac{1}{h}\int_x^{x+h} L_j(\xi)\, d\xi \;\overset{(\xi = x + ch)}{=}\; \int_0^1 L_j^*(c)\, dc\ , \quad j = 0, \dots, n\ .$$

Dann ist die Quadraturformel

$$Q_n[f] \;:=\; h\sum_{j=0}^{n} b_j\, f(x_j) \;=\; h\;\sum_{j=0}^{n} b_j\, f(x + c_j h)$$

für alle Polynome P bis zum Grad n exakt: $\displaystyle\int_x^{x+h} P(\xi)\, d\xi = Q_n[P]$.

Beweis: Für Polynome P vom Grad $\leq n$ gilt $P(\xi) \equiv \displaystyle\sum_{j=0}^{n} P(x_j)\, L_j(\xi)$, denn P ist seine eigene Polynominterpolierende. Es folgt

$$\int_x^{x+h} P(\xi)\, d\xi \;=\; \sum_{j=0}^{n} P(x_j)\, \underbrace{\int_x^{x+h} L_j(\xi)\, d\xi}_{h\, b_j} \;=\; Q_n[P]\ .$$

Q.E.D.

Bemerkung 2.3: *Die so gewählten Gewichte hängen damit nicht vom Intervall $[x, x+h]$ ab, sondern nur von der Knotenverteilung $c_0, \dots, c_n$.*

Bemerkung 2.4: *Die Gewichte b_j sind die Lösung eines linearen Vandermonde-Systems*

$$\sum_{j=0}^{n} b_j\, c_j^i \;=\; \frac{1}{i+1}\ , \quad i = 0, \dots, n\ .$$

Beweis: Für $i = 0, \ldots, n$ gilt

$$
\begin{array}{ccc}
\displaystyle\int_x^{x+h} (\xi - x)^i \, d\xi & \stackrel{\text{(exakt)}}{=} & \displaystyle h \sum_{j=0}^{n} b_j \, (x_j - x)^i \\
\Big\| \; {\scriptstyle \xi = x + ch} & & \Big\| \; {\scriptstyle x_j - x = c_j h} \\
\displaystyle h^{i+1} \int_0^1 c^i \, dc = \frac{h^{i+1}}{i+1} & = & \displaystyle h^{i+1} \sum_{j=0}^{n} b_j \, c_j^i \; .
\end{array}
$$

Q.E.D.

Interpretation 2.5: *Zum Integrand f sei $P(\xi) = \sum_{j=0}^{n} f(x_j) \, L_j(\xi)$ das Interpolationspolynom zu den Stützstellen x_j. Es folgt*

$$
\int_x^{x+h} P(\xi) \, d\xi \;=\; \sum_{j=0}^{n} f(x_j) \int_x^{x+h} L_j(\xi) \, d\xi \;=\; h \sum_{j=0}^{n} b_j \, f(x_j) \;=\; Q_n[f] \; ,
$$

also

$$
\boxed{\text{Quadraturformel = exaktes Integral über das Interpolationspolynom.}}
$$

Bezeichnung 2.6: *Die Quadraturformel $Q_n[f]$ in Satz 2.2 heißt* ***Newton-Cotes-Formel*** *zum Interpolationsgrad n bzgl. der Stützstellen $x_0, \ldots, x_n$.*

Beispiel 2.7: *Wähle äquidistante Stützstellen $x_i = x + \frac{i}{n} h$, $i = 0, \ldots, n$.*

$n = 0$ (Riemann-Approximation):

$$
\int_x^{x+h} f(\xi) \, d\xi \;=\; h \, f(x) \;+\; \textit{Fehler}_0[f]
$$

$n = 1$ (Trapezformel):

$$
\int_x^{x+h} f(\xi) \, d\xi \;=\; h \, \frac{f(x) + f(x+h)}{2} \;+\; \textit{Fehler}_1[f]
$$

$n = 2$ (Simpsonformel):

$$
\int_x^{x+h} f(\xi) \, d\xi \;=\; \frac{h}{6} \, \Big(f(x) + 4 f(x + \tfrac{h}{2}) + f(x+h) \Big) \;+\; \textit{Fehler}_2[f]
$$

$n = 3$ (Newton's 3/8-Regel):

$$
\int_x^{x+h} f(\xi) \, d\xi = \frac{h}{8} \, \Big(f(x) + 3 f(x + \tfrac{h}{3}) + 3 f(x + \tfrac{2h}{3}) + f(x+h) \Big) \;+\; \textit{Fehler}_3 \, [f]
$$

<u>$n = 4$</u> *(Milne-Formel):*

$$\int_x^{x+h} f(\xi)\,d\xi \;=\; \frac{h}{90}\,\Big(7f(x) + 32\,f(x+\tfrac{h}{4}) + 12\,f(x+\tfrac{h}{2})$$

$$+32\,f(x+\tfrac{3h}{4}) + 7\,f(x+h)\Big) \;+\; \mathit{Fehler}_4\,[f]$$

Mit der Interpretation 2.5, also

$$\text{Quadraturfehler} \;=\; \int \Big(f(\xi) \;-\; \text{Interpolationspolynom}(\xi)\Big)\,d\xi$$

und Satz 1.2 folgen Fehlerabschätzungen:

Satz 2.8:

Sei $Q_n[f]$ die Quadratur-Formel aus Satz 2.2 zu den Stützstellen $x_j = x + c_j h \in [x, x+h]$ mit der Verteilung $c_0 < c_1 < \ldots < c_n$ in $[0,1]$. Für $f \in C^{n+1}([x,x+h],\mathbb{R})$ gilt

$$\left|\int_x^{x+h} f(\xi)\,d\xi \;-\; Q_n[f]\right| \;\le\; d_n\,h^{n+2} \max_{\xi\in[x,x+h]} |f^{(n+1)}(\xi)|$$

mit

$$d_n \;:=\; \frac{1}{(n+1)!}\int_0^1 |c-c_0|\,\cdots\,|c-c_n|\,dc\;.$$

Für gerades n und $f \in C^{n+2}([x,x+h],\mathbb{R})$ gilt sogar

$$\left|\int_x^{x+h} f(\xi)\,d\xi \;-\; Q_n[f]\right| \;\le\; e_n\,h^{n+3} \max_{\xi\in[x,x+h]} |f^{(n+2)}(\xi)|$$

mit

$$e_n := \frac{1}{(n+2)!}\int_0^1 |c-\tfrac{1}{2}|\,|c-c_0|\,\cdots\,|c-c_n|\,dc\;,$$

falls die Stützstellen symmetrisch im Intervall liegen:

$$x_{n-i} = 2\,x + h - x_i\;,\quad \text{d.h.,}\quad c_{n-i} = 1 - c_i\;,\quad i = 0,\ldots,n\;.$$

Beweis: Nach Satz 1.2 gilt für das Interpolationspolynom $P_n(\xi)$ durch $(x_0, f(x_0)), \ldots, (x_n, f(x_n))$:

$$
\begin{aligned}
f(\xi) - P_n(\xi) &= \frac{f^{(n+1)}(\eta)}{(n+1)!}\,(\xi - x_0)\cdots(\xi - x_n) \quad \text{mit} \quad \eta \in [x, x+h] \\
\Longrightarrow \quad \left|\int_x^{x+h} f(\xi)\,d\xi - Q_n[f]\right| &= \left|\int_x^{x+h} \frac{f^{(n+1)}(\eta)}{(n+1)!}\,(\xi - x_n)\cdots(\xi - x_n)\,d\xi\right| \\
&\le \frac{1}{(n+1)!}\int_x^{x+h} |\xi - x_0|\cdots|\xi - x_n|\,d\xi \;\left(\max_{\eta\in[x,x+h]} |f^{(n+1)}(\eta)|\right) \\
&\overset{(*)}{=} h^{n+2}\underbrace{\frac{1}{(n+1)!}\int_0^1 |c - c_0|\cdots|c - c_n|\,dc}_{d_n} \;\left(\max_{\eta\in[x,x+h]} |f^{(n+1)}(\eta)|\right)
\end{aligned}
$$

mit der Substitution $\xi = x + ch$ in $(*)$. Für gerades n nehme einen beliebigen weiteren Punkt $x_{n+1} \in [x, x+h]$ mit $x_{n+1} \notin \{x_0, \ldots, x_n\}$ hinzu. Mit dem Interpolationspolynom P_{n+1} durch $(x_0, f(x_0)), \ldots, (x_{n+1}, f(x_{n+1}))$ folgt wie oben mit $n \mapsto n+1$:

$$
\left|\int_x^{x+h} f(\xi)\,d\xi \;-\; \int_x^{x+h} P_{n+1}(\xi)\,d\xi\right| \;\le
$$
$$
\frac{h^{n+3}}{(n+2)!}\int_0^1 |c - c_0|\cdots|c - c_n|\,|c - c_{n+1}|\,dc \;\left(\max_{\eta\in[x,x+h]} |f^{(n+2)}(\eta)|\right).
$$

Es wird gezeigt, daß bei symmetrischen Stützstellen

$$
\int_x^{x+h} P_{n+1}(\xi)\,d\xi \;=\; \int_x^{x+h} P_n(\xi)\,d\xi \;\equiv\; Q_n[f]
$$

gilt. Beachte dazu $P_{n+1}(\xi) = P_n(\xi) + \alpha\,(\xi - x_0)\cdots(\xi - x_n)$ mit einem gewissem $\alpha \in \mathbb{R}$ (offensichtlich gilt $P_{n+1}(x_i) = P_n(x_i) = f(x_i)$ für $i = 0, \ldots, n$. Mit $P_{n+1}(x_{n+1}) = f(x_{n+1})$ wird α eindeutig festgelegt). Es folgt

$$
\int_x^{x+h} P_{n+1}(\xi)\,d\xi \;=\; Q_n[f] \;+\; \alpha\int_x^{x+h} (\xi - x_0)\cdots(\xi - x_n)\,d\xi\;.
$$

Für eine symmetrische Verteilung $x_{n-i} = 2x + h - x_i$ folgt mit $\xi = 2x + h - \eta$:

$$
\begin{aligned}
I \;&:=\; \int_x^{x+h} (\xi - x_0)\cdots(\xi - x_n)\,d\xi \\
&=\; -\int_{x+h}^{x} (2x + h - \eta - x_0)\cdots(2x + h - \eta - x_n)\,d\eta \\
&=\; \int_x^{x+h} (x_n - \eta)\cdots(x_0 - \eta)\,d\eta \;=\; (-1)^{n+1}\,I\;.
\end{aligned}
$$

Für gerades n folgt $I = 0$, also

$$\left| \int_x^{x+h} f(\xi)\, d\xi \; - \; Q_n[f] \right| \;\leq$$

$$\frac{h^{n+3}}{(n+2)!} \int_0^1 |c - c_0| \cdots |c - c_n|\, |c - c_{n+1}|\, dc \quad \Big(\max_{\eta \in [x, x+h]} |f^{(n+2)}(\eta)| \Big)$$

mit beliebigem $c_{n+1} = \frac{x_{n+1} - x}{h} \in [0,1]$, z.B. $c_{n+1} = 1/2$ (aus Stetigkeitsgründen darf nun c_{n+1} auch mit einem der $c_0, \ldots, c_n$ übereinstimmen).

Q.E.D.

Bemerkung 2.9: *Für äquidistante Stützstellen $x_i = x + \frac{i}{n}\, h$, $i = 0, \ldots, n$, kann man (mit wesentlich größerem Aufwand) sogar*

$$\int_x^{x+h} f(\xi)\, d\xi \; - \; Q_n[f] \; =$$

$$\begin{cases} -h^{n+2}\, \dfrac{f^{(n+1)}(\eta)}{(n+1)!} \displaystyle\int_0^1 c\,(c - \tfrac{1}{n})(c - \tfrac{2}{n}) \cdots (c-1)\, dc\,, & n \text{ ungerade} \\[2ex] -h^{n+3}\, \dfrac{f^{(n+2)}(\eta)}{(n+2)!} \displaystyle\int_0^1 c\,(c - \tfrac{1}{n})(c - \tfrac{2}{n}) \cdots (c - \tfrac{1}{2})^2 \cdots (c-1)\, dc\,, & n \text{ gerade} \end{cases}$$

mit geeignetem Zwischenwert $\eta \in (x, x+h)$ beweisen (keine Betragszeichen!).

2.2 Gauß-Quadratur

Idee: die Knoten x_i (bzw. ihre Verteilung c_i) sind nicht vorgegeben, sondern sollen "optimal" gewählt werden. Schreibtechnisch ist es hier schöner, n Knoten/Gewichte $c_1, \ldots, c_n / b_1, \ldots, b_n$ (statt $c_0, \ldots, c_n / b_0, \ldots, b_n$) zu betrachten.

Bemerkung 2.10: *Der maximal erreichbare Exaktheitsgrad einer Quadraturformel*

$$G_n[f] \; = \; h \sum_{j=1}^{n} b_j\, f(x_j) \; = \; h \sum_{j=1}^{n} b_j\, f(x + c_j h)$$

mit n Knoten ist $2n - 1$.

Beweis: Die Konstruktion für $2n-1$ folgt (Gauß-Quadratur). Für das Polynom

$$P_{2n}(\xi) \; = \; (\xi - x_1)^2 \cdots (\xi - x_n)^2$$

vom Grad $2n$ gilt $G_n[P_{2n}] = 0 \neq \displaystyle\int_x^{x+h} P_{2n}(\xi)\, d\xi > 0$.

Q.E.D.

Satz 2.11: (Gauß-Quadratur)
Die Quadraturformel

$$\int_x^{x+h} f(\xi)\,d\xi \;=\; \underbrace{h\sum_{j=1}^{n} b_j\, f(x+c_j h)}_{n.te\ \text{Gaußformel}\ G_n[f]} \;+\; \textit{Fehler}_n\,[f]$$

ist genau dann exakt für alle Polynome vom Grad $\leq 2n-1$, *wenn* $c_1,\ldots,c_n,b_1,\ldots,b_n$ *das Gleichungssystem*

$$\sum_{j=1}^{n} b_j\, c_j^{i-1} \;=\; \frac{1}{i}\,, \quad i=1,\ldots,2n \qquad (\#)$$

lösen ($2n$ *Gleichungen für* $2n$ *Unbekannte).*

Beweis: Setze die Monome $f_i(\xi)=(\xi-x)^{i-1}$ $(i=1,\ldots,2n)$ ein:

$$\int_x^{x+h}(\xi-x)^{i-1}\,d\xi \;\stackrel{(\xi=x+ch)}{=}\; h\int_0^1 (ch)^{i-1}dc \;=\; \frac{h^i}{i}\,,$$

$$G_n[f_i] \;=\; h\sum_{j=1}^{n} b_j\,(c_j h)^{i-1} \;=\; h^i\sum_{j=1}^{n} b_j\, c_j^{i-1}\,.$$

Q.E.D.

Fakten:

+ das System (#) hat eine bis auf Permutation eindeutige reelle Lösung,
+ es gilt $c_1,\ldots,c_n\in(0,1)$, d.h., $x_j=x+c_j h\in(x,x+h)$,
+ es gilt $b_j>0$ (gut für numerische Stabilität),
− für $n>5$ haben die Daten $c_1,\ldots,c_n,b_1,\ldots,b_n$ keine geschlossene Darstellung,
+ man kann $c_1,\ldots,b_n$ aber numerisch schnell und stabil berechnen.

Zugang über Orthogonalpolynome:

Definition 2.12: $f,g:[x,x+h]\to\mathbb{R}$ *seien (quadratisch) integrierbar.*

a) $\ll f,g\gg\;:=\int_x^{x+h} f(\xi)\,g(\xi)\,d\xi$ *heißt* **Skalarprodukt**,

b) f **orthogonal** g *bedeutet* $\ll f,g\gg\;=0$.

Definition und Satz 2.13: Definiere rekursiv

$$P_k(\xi) \;=\; \xi^k \;+\; \sum_{j=0}^{k-1} \frac{\ll \xi^k, P_j \gg}{\ll P_j, P_j \gg} \, P_j(\xi) \;, \quad k = 1, 2, \dots$$

mit dem Start $P_0(\xi) \equiv 1$.

a) Es gilt $\ll P_n, P \gg = 0$ *für alle Polynome* P *vom Grad* $< n$.

b) P_n *hat genau* n *einfache Nullstellen* $x_j = x + c_j h \in (x, x+h)$, $j = 1, \dots, n$. *Achtung:* x_j, c_j *hängen von* n *ab!*

c) Wählt man der Newton-Cotes-Quadratur (Satz 2.2) entsprechend $b_j = \frac{1}{h}\int_x^{x+h} L_j(\xi)\, d\xi = \int_0^1 L_j^*(c)\, dc$ *mit den Lagrangre-Polynomen* $L_j(\xi)$ *bzw.* $L_j^*(c)$ *zu* (x_i) *bzw.* (c_i), *so sind* $c_1, \dots, c_n, b_1, \dots, b_n$ *die Daten der Gauß-Formel* $G_n[f]$. *Es gilt* $b_j > 0$.

Beweis: a) Induktion nach n mit der Induktionsbehauptung

$$\ll P_k, P_j \gg = 0 \quad \forall\, k, j \in \{0, \dots, n\} \;,\; k \neq j \;.$$

<u>Schritt $n \to n+1$:</u> zu zeigen ist nur $\ll P_{n+1}, P_j \gg = 0 \;\forall j = 0, \dots, n$.

$$\begin{aligned}
\ll P_{n+1}, P_j \gg \;&=\; \ll \xi^{n+1} - \sum_{k=0}^{n} \frac{\ll \xi^{n+1}, P_k \gg}{\ll P_k, P_k \gg} P_k \;,\; P_j \gg \\
&=\; \ll \xi^{n+1}, P_j \gg \;-\; \sum_{k=0}^{n} \frac{\ll \xi^{n+1}, P_k \gg}{\ll P_k, P_k \gg} \underbrace{\ll P_k, P_j \gg}_{0 \text{ für } k \neq j} \\
&=\; \ll \xi^{n+1}, P_j \gg \;-\; \ll \xi^{n+1}, P_j \gg \;=\; 0 \;.
\end{aligned}$$

Jedes P vom Grad $< n$ läßt sich als $P(\xi) = \sum_{j=0}^{n-1} \alpha_j \, P_j(\xi)$ schreiben, es folgt

$$\ll P_n, P \gg \;=\; \sum_{j=0}^{n-1} \alpha_j \; \ll P_n, P_j \gg \;=\; 0 \;.$$

b) Seien $x_1, \dots, x_j$ die paarweise verschiedenen reellen Nullstellen von P_n in $(x, x+h)$ mit Vielfachheiten $n_1, \dots, n_j$. Betrachte

$$P(\xi) = (\xi - x_1)^{m_1} \dots (\xi - x_j)^{m_j} \quad \text{mit} \quad m_i = \begin{cases} 1 & \text{für ungerades } n_i \\ 0 & \text{für gerades } n_i \end{cases}$$

mit denselben Vorzeichenwechseln wie P_n auf $(x, x+h)$. Damit folgt $\ll P_n, P \gg \neq 0$. Falls $j < n$ gilt, so ist $\mathrm{grad}(P) = m_1 + \dots + m_j < n$, und mit a) folgt der Widerspruch $\ll P_n, P \gg = 0$.

c) Mit $b_j = \dfrac{1}{h}\displaystyle\int_x^{x+h} L_j(\xi)\,d\xi$ gilt nach Satz 2.2 (mit $n \to n-1$), daß $G_n[f]$ bis zum Polynomgrad $n-1$ exakt ist. Für jedes Polynom P vom Grad $\leq 2n-1$ existiert eine Zerlegung

$$P(\xi) = \alpha(\xi) P_n(\xi) + \beta(\xi)$$

mit Polynomen $\alpha(\xi), \beta(\xi)$ vom Grad $\leq n-1$ (Polynomdivision mit Rest):

$$\int_x^{x+h} P(\xi)\,d\xi \; - \; G_n[P]$$

$$= \underbrace{\int_x^{x+h} \alpha(\xi)\,P_n(\xi)\,d\xi}_{= \ll \alpha, P_n \gg = 0} + \int_x^{x+h} \beta(\xi)\,d\xi - h \sum_{j=1}^{n} b_j \Big(\alpha(x_j)\;\underbrace{P_n(x_j)}_{0} + \beta(x_j)\Big)$$

$$= \int_x^{x+h} \beta(\xi)\,d\xi \; - \; G_n[\beta] \; = \; 0 \; ,$$

da die Quadratur bis zum Grad $n-1$ exakt ist. Damit ist $G_n[.]$ auch bis zum Grad $2n-1$ exakt. Es gilt $b_j = \dfrac{1}{h}\displaystyle\int_x^{x+h} L_j(\xi)\,d\xi$, aber auch

$$b_j = \sum_{k=1}^{n} b_k \underbrace{L_j^2(x_k)}_{\delta_{kj}} \; = \; \frac{1}{h}\int_x^{x+h} L_j^2(\xi)\,d\xi \; > \; 0 \; ,$$

da mit $\mathrm{grad}(L_j^2) = 2n-2$ die Quadratur exakt ist.

Q.E.D.

Bezeichnung 2.14: *Die Orthogonalpolynome P_n heißen* **"Legendre-Polynome"** *über dem Intervall $[x, x+h]$ (die Literatur benutzt das Standardintervall $[-1,1]$, also $x=-1, h=2$). Sei $P_n^*(c)$ das n.te Legendre-Polynom über unserem Standardintervall $[0,1]$. Es gilt $P_n(x+ch) = h^n\,P_n^*(c)$, was unmittelbar aus der Rodriguez-Darstellung in Satz 2.16 folgt.*

Hilfssatz 2.15:

Das Polynom $\widetilde{P}_n$ vom Grad n habe die Eigenschaft $\ll \widetilde{P}_n, P \gg = 0$ für alle Polynome P vom Grad $< n$. Dann folgt $\widetilde{P}_n(\xi) = const_n\,P_n(\xi)$.

Beweis: Mit $\widetilde{P}_n = \alpha_n P_n + \alpha_{n-1} P_{n-1} + \cdots + \alpha_0 P_0$ folgt für $j = 0, \ldots, n-1$:

$$0 \; = \; \ll \widetilde{P}_n, P_j \gg \; = \; \alpha_j \ll P_j, P_j \gg \quad ,$$

d.h., $\alpha_1 = \ldots = \alpha_{n-1} = 0$.

Q.E.D.

Satz 2.16: (Rodriguez-Formel)

Es gilt die Darstellung

$$P_n(\xi) = \frac{n!}{(2n)!} \, \frac{d^n}{d\xi^n} \, (\xi - x)^n (\xi - (x+h))^n \ .$$

Beweis: Setze $g_{2n}(\xi) := (\xi - x)^n (\xi - (x+h))^n$, sei P ein beliebiges Polynom vom Grad $< n$. Durch partielle Integration folgt

$$\begin{aligned} \ll g_{2n}^{(n)}, P \gg \ &= \ \underbrace{[g_{2n}^{(n-1)}(\xi) P(\xi)]_{\xi=x}^{\xi=x+h}}_{=0} - \ll g_{2n}^{(n-1)}, P' \gg \\ \stackrel{\text{(analog)}}{=} \ll g_{2n}^{(n-2)}, P'' \gg \ &= \ \ldots \ = \ (-1)^n \ \ll g_{2n}, \underbrace{P^{(n)}}_{\equiv 0} \gg \ . \end{aligned}$$

Damit ist $g_{2n}^{(n)}$ orthogonal auf allen Polynomen von Grad $< n$, so daß mit 2.15 $g_{2n}^{(n)}(\xi) = \text{const}_n \, P_n(\xi)$ folgt. Der Faktor folgt aus der Normierung

$$g_{2n}^{(n)}(\xi) \ = \ \frac{d^n}{d\xi^n} \, (\xi^{2n} + \cdots) = \frac{(2n)!}{n!} \, (\xi^n + \cdots) \ .$$

Q.E.D.

Folgerung 2.17: *Offensichtlich ist g_{2n} eine gerade Funktion bezüglich Spiegelung an der Intervallmitte. Da jede Ableitung die Parität ändert, folgt*

$$P_n \ \textit{ist eine} \ \left\{ \begin{array}{c} \textit{gerade} \\ \textit{ungerade} \end{array} \right\} \ \textit{Funktion für} \ \left\{ \begin{array}{c} \textit{gerades } n \\ \textit{ungerades } n \end{array} \right\},$$

d.h., $P_n(x + ch) = (-1)^n P_n(\,x + (1-c)h\,)$. Über dem Standardintervall $[0,1]$ gilt speziell $P_n^*(c) = (-1)^n P_n^*(1-c)$. Die Nullstellen sind damit symmetrisch um die Intervallmitte verteilt. Für ungerades n ist die Intervallmitte eine der Nullstellen.

Gauß-Daten 2.18:

$$
\begin{aligned}
n=1: \quad & P_1^*(c) = c - \tfrac{1}{2} \\
& c_1 = \tfrac{1}{2}\,, \quad b_1 = 1 \\
n=2: \quad & P_2^*(c) = c^2 - c + \tfrac{1}{2} \\
& c_1 = \tfrac{1}{2} - \tfrac{1}{2\,\sqrt{3}}\,, \quad b_1 = \tfrac{1}{2} \\
& c_2 = \tfrac{1}{2} + \tfrac{1}{2\,\sqrt{3}}\,, \quad b_2 = \tfrac{1}{2} \\
n=3: \quad & P_3^*(c) = (c - \tfrac{1}{2})\,(c^2 - c + \tfrac{1}{10}) \\
& c_1 = \tfrac{1}{2} - \tfrac{1}{2}\,\sqrt{\tfrac{3}{5}}\,, \quad b_1 = \tfrac{5}{18} \\
& c_2 = \tfrac{1}{2}\,, \quad b_2 = \tfrac{4}{9} \\
& c_3 = \tfrac{1}{2} + \tfrac{1}{2}\,\sqrt{\tfrac{3}{5}}\,, \quad b_3 = \tfrac{5}{18} \\
n=4: \quad & c_1 = \tfrac{1}{2} - \tfrac{1}{2}\,\sqrt{15 + 2\,\sqrt{30}}, \quad b_1 = \tfrac{1}{4} - \tfrac{\sqrt{30}}{72} \\
& c_2 = \tfrac{1}{2} - \tfrac{1}{2}\,\sqrt{15 - 2\,\sqrt{30}}, \quad b_2 = \tfrac{1}{4} + \tfrac{\sqrt{30}}{72} \\
& c_3 = \tfrac{1}{2} + \tfrac{1}{2}\,\sqrt{15 - 2\,\sqrt{30}}, \quad b_3 = b_2 \\
& c_4 = \tfrac{1}{2} + \tfrac{1}{2}\,\sqrt{15 + 2\,\sqrt{30}}, \quad b_4 = b_1 \\
n=5: \quad & c_1 = \tfrac{1}{2} - \tfrac{1}{6}\,\sqrt{5 + 2\,\sqrt{\tfrac{10}{7}}}\,, \quad b_1 = \tfrac{161}{900} - \tfrac{13\,\sqrt{70}}{1800} \\
& c_2 = \tfrac{1}{2} - \tfrac{1}{6}\,\sqrt{5 - 2\,\sqrt{\tfrac{10}{7}}}\,, \quad b_2 = \tfrac{161}{900} + \tfrac{13\,\sqrt{70}}{1800} \\
& c_3 = \tfrac{1}{2}\,, \quad b_3 = \tfrac{64}{225} \\
& c_4 = \tfrac{1}{2} + \tfrac{1}{6}\,\sqrt{5 - 2\,\sqrt{\tfrac{10}{7}}}\,, \quad b_4 = b_2 \\
& c_5 = \tfrac{1}{2} + \tfrac{1}{6}\,\sqrt{5 + 2\,\sqrt{\tfrac{10}{7}}}\,, \quad b_5 = b_1
\end{aligned}
$$

Bemerkung 2.19: *Sei $m = x + \frac{h}{2}$ die Intervallmitte. Die Legendre-Polynome über $[x, x+h]$ erfüllen die Rekursionen (z.B. [Abramowitz, Stegun: Handbook of Math. Functions, Dover 1970])*

$$
\begin{aligned}
P_{n+1}(\xi) &= (\xi - m)\,P_n(\xi) \;-\; \frac{h^2}{4}\,\frac{n^2}{4n^2-1}\,P_{n-1}(\xi)\,, \\
P'_{n+1}(\xi) &= (\xi - m)P'_n(\xi) \;-\; \frac{h^2}{4}\,\frac{n^2}{4n^2-1}\,P'_{n-1}(\xi) \;+\; P_n(\xi)\,.
\end{aligned}
$$

Mit $P_0(\xi) = 1$, $P'_0(\xi) = 0$, $P_1(\xi) = \xi - m$, $P'_1(\xi) = 1$ können so P_n und P'_n rekursiv an jeder Stelle numerisch stabil ausgewertet werden. Die Nullstellensu-

che über das Newton-Verfahren ist damit problemlos. Gute Startwerte für die Nullstellen $(x <)\ x_1 < \ldots < x_n\ (< x + h)$ *von* P_n *sind durch*

$$x_j^{(0)} \;=\; m \;-\; \frac{h}{4}\,\Big[\;\cos\Big(\frac{j\,\pi}{n+1}\Big) \;+\; \cos\Big(\frac{(2j-1)\,\pi}{2n}\Big)\Big]\,, \quad j = 1, \ldots, n$$

gegeben.

Satz 2.20: (Fehler der Gauss-Quadratur)

Es gilt

$$\int_x^{x+h} f(\xi)\,d\xi \;-\; G_n[f] \;=\; \frac{\ll P_n, P_n \gg}{(2n)!}\, f^{(2n)}(\eta)$$

mit einem Zwischenwert $\eta \in (x, x+h)$ *und*

$$\ll P_n, P_n \gg \;=\; h^{2n+1} \int_0^1 (P_n^*(c))^2 dc \;=\; h^{2n+1}\,\frac{(n!)^4}{(2n)!(2n+1)!}\,.$$

Beweis: siehe z.B. [Stoer, Numerische Mathematik 1].

Kapitel 3

Dynamische Systeme

Definition 3.1: *Ein Differentialgleichungssystem 1. Ordnung*

$$\frac{dy}{dt} \;=\; f(t,y)\ ; \quad y \in \mathbb{R}^N\ ; \quad f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$$

heißt ***dynamisches System*** *auf dem* ***Phasenraum*** $\mathbb{R}^N$*. Der Parameter t wird die* ***Zeit*** *genannt. Das* ***Vektorfeld*** *$f(t,y)$ heißt* ***autonom****, wenn es nicht explizit von der Zeit abhängt: $dy/dt = f(y)$.*

Bemerkung 3.2: *Mit $z := (t,y)^T \in \mathbb{R} \times \mathbb{R}^N$, $dz/dt = g(z) := (1, f(t,y))^T$ läßt sich jedes System durch Vergrößerung des Phasenraums autonom machen:*

$$\frac{dy}{dt} \;=\; f(t,y) \quad \Longleftrightarrow \quad \frac{d}{dt}\begin{pmatrix} t \\ y \end{pmatrix} \;=\; \begin{pmatrix} 1 \\ f(t,y) \end{pmatrix}\ .$$

Bemerkung 3.3: *Eine Differentialgleichung höherer Ordnung*

$$\frac{d^k y}{dt^k} \;=\; f\Big(t, y, \frac{dy}{dt}\ , \ldots,\ \frac{d^{k-1}y}{dt^{k-1}}\Big)\ ; \quad y \in \mathbb{R}^n \qquad (\#)$$

läßt sich als dynamisches System auf einem größeren Phasenraum interpretieren:

$$\frac{d}{dt}\begin{pmatrix} t \\ z_0 \\ z_1 \\ \vdots \\ z_{k-2} \\ z_{k-1} \end{pmatrix} \;=\; \begin{pmatrix} 1 \\ z_1 \\ z_2 \\ \vdots \\ z_{k-1} \\ f(t, z_0, \ldots, z_{k-1}) \end{pmatrix} \;\in\; \underbrace{\mathbb{R} \times \mathbb{R}^n \times \ldots \times \mathbb{R}^n}_{\mathbb{R}^N \;=\; \mathbb{R}^{1+nk}}\ .$$

Setzt man $z_0 = y$, so folgt $z_i = d^i y/dt^i \in \mathbb{R}^n$, $i = 0, \ldots, k-1$. Der letzte Block $dz_{k-1}/dt = dy^k/dt^k = f(t, z_0, \ldots, z_{k-1})$ repräsentiert dann (#).

Damit reicht es, numerische Verfahren zu Lösung von $dy/dt = f(y)$ zu entwickeln, die für beliebiges (glattes) f auf beliebigen Phasenräumen funktionieren. Es geht hier um das **Anfangswertproblem** (AWP)

$$\frac{dy}{dt} = f(y) \; ; \quad y(t_0) = y_0 \; . \tag{\#}$$

Randwertaufgaben (Vorgabe von Daten zu verschiedenen Zeiten) erfordern andere Methoden.

Existenz und Eindeutigkeit für $dy/dt = f(t, y)$ auf dem $\mathbb{R}^N$ sind unter minimalen Voraussetzungen an das Vektorfeld garantiert:

Satz 3.4: (Existenzsatz von Peano)
Sei $f(t, y)$ stetig auf einem offenem Gebiet $\Omega \subset \mathbb{R} \times \mathbb{R}^N$, sei $(t_0, y_0) \in \Omega$. Dann hat das AWP (#) mindestens eine Lösung, die sich bis zum Rand von Ω fortsetzen läßt.

Beweis: siehe z.B. [W. Walter, Gewöhnliche Differentialgleichungen, Springer, Kap. II.10].

Bemerkung 3.5: *Es können die Fälle auftreten:*

a) die Lösungen existieren für alle $t \in [t_0, \infty)$,

b) $\lim_{t \to T-0} \|y(t)\| = \infty$ für $t_0 \le t < T < \infty$,

c) $\lim_{t \to T-0} dist\Big((t, y(t)) \, , \, \partial\Omega \Big) = 0$ für $t_0 \le t < T < \infty$
(mit dist = Abstand zum Rand $\partial\Omega$).

Beispiele für diese Fälle sind z.B. in [Deuflhard & Bornemann] zu finden. Analoges gilt für Zeiten $t \le t_0$.

Satz 3.6: (Eindeutigkeit der Lösung)
Sei $f(t, y)$ stetig auf einem offenem Gebiet $\Omega \subset \mathbb{R} \times \mathbb{R}^N$ und Lipschitz-stetig bzgl. y, d.h., es existiert L mit

$$|f(t, y_1) - f(t, y_2)| \;\le\; L\, |y_1 - y_2| \quad \forall\, (t, y_1), (t, y_2) \in \Omega.$$

Dann ist die Lösung aus Satz 3.4 eindeutig.

Beweis: siehe z.B. [W. Walter].

Bemerkung 3.7: *<u>Lokale</u> Lipschitz-Stetigkeit reicht für Eindeutigkeit: zu jedem $(t,y) \in \Omega \subset \mathbb{R} \times \mathbb{R}^N$ existiere eine Umgebung $U(t,y)$ und eine Lipschitz-Konstante $L = L(t,y)$, so daß*

$$|f(t,y_1) - f(x,y_2)| \;\leq\; L\,|y_1 - y_2| \quad \forall\ (t,y_1),(t,y_2) \in U(t,y)\ .$$

Zusammenfassung: Sei $f(t,y)$ stetig und stetig differenzierbar bezüglich y (und damit lokal Lipschitz-stetig in y). Dann existiert lokal, d.h., für hinreichend kleine h, eine eindeutige Lösung $y(t_0+h)$ des AWP (#).

Bemerkung 3.8: *Für spezielle Gebiete gelten weitergehende Aussagen, z.B.: Sei $\Omega = [t_0,t_1] \times \mathbb{R}^N$, $f : \Omega \to \mathbb{R}^N$ stetig und global Lipschitz-stetig bzgl. y (d.h., die L-Konstante ist global gültig auf Ω). Dann existiert die Lösung $y(t)\ \forall\ t \in [t_0,t_1]$. Vergleiche z.B. [Walter, II.10, Satz VII].*

Im folgenden wird $f(t,y)$ als hinreichend glatt (mindestens stetig differenzierbar) vorausgesetzt.

Definition 3.9:

Sei $y(t;t_0,y_0)$ die Lösung des AWP $dy/dt = f(t,y)\ ;\ y(t_0) = y_0$. Die Abbildung

$$F_{t,t_0} : \begin{array}{l} \mathbb{R}^N \to \mathbb{R}^N \\ y_0 \to y(t;t_0,y_0) \end{array}$$

*heißt **Fluß** des dynamischen Systems (auch "**Zeitschritt**" $t_0 \to t$ genannt).*

Satz 3.10: (Gruppenstruktur des Flusses)

Es gilt $F_{t,t_1} \circ F_{t_1 t_0} = F_{t,t_0}$ für alle Zeiten $t_0, t_1, t \in \mathbb{R}$, für welche die Flüsse definiert sind.

Beweis: Sei $u(t) = (F_{t,t_1} \circ F_{t_1,t_0})(y_0) = y(t;t_1,y(t_1;t_0,y_0))$. Diese Funktion erfüllt

$$\frac{du}{dt} = f(t,u) \quad \text{ und } \quad u(t_1) = y(t_1;t_0,y_0)\ .$$

Sei $v(t) = F_{t,t_0}(y_0) = y(t;t_0,y_0)$. Diese Funktion erfüllt

$$\frac{dv}{dt} = f(t,v) \quad \text{ und } \quad v(t_1) = y(t_1;t_0,y_0)\ .$$

Also: u und v sind beide Lösungen des AWP $dz/dt = f(t,z); z(t_1) = y(t_1;t_0,y_0)$. Aus der Eindeutigkeit von Lösungen folgt $u = v$.

Q.E.D.

Satz 3.11: (Gruppenstruktur autonomer Systeme)
Für autonome Differentialgleichungen $dy/dt = f(y)$ hängt F_{t,t_0} nur von der Differenz $t - t_0$ ab:

$$F_{t_0+h,t_0} = F_{t_1+h,t_1} \quad \forall\, t_0, t_1, h \in \mathbb{R} .$$

Mit der Notation $F_h \equiv F_{t_0+h,t_0}$ (unabhängig von t_0) gilt

$$F_0 = id ,$$
$$F_{h_2} \circ F_{h_1} = F_{h_1} \circ F_{h_2} = F_{h_1+h_2} , \qquad \text{(“Gruppenstruktur des Flusses”)}$$
$$F_h \circ F_{-h} = F_{-h} \circ F_h = id .$$

Beweis: Seien

$$u(h) = F_{t_0+h,t_0}(y_0) = y(t_0 + h; t_0, y_0) \ , \ v(h) = F_{t_1+h,t_1}(y_0) = y(t_1 + h; t_1, y_0) \ .$$

Beide Funktionen erfüllen dieselbe Differentialgleichung:

$$\frac{du}{dh} = f(u) \ , \quad \frac{dv}{dh} = f(v) \ .$$

Außerdem gilt $u(0) = y(t_0; t_0, y_0) = y_0 = y(t_1; t_1, y_0) = v(0)$. Mit der Eindeutigkeit von Anfangswertproblemen folgt $u = v$. Zur Gruppenstruktur:
- $F_0 = F_{t_0,t_0} = id$ ist klar ,
- $F_{h_2} \circ F_{h_1} \equiv F_{t_0+h_1+h_2,t_0+h_1} \circ F_{t_0+h_1,t_0} \overset{(3.10)}{=} F_{t_0+h_1+h_2,t_0} \equiv F_{h_1+h_2}$,
- $F_h \circ F_{-h} = F_{h-h} = F_0 = id.$

Q.E.D.

Bemerkung 3.12: *Sei $y(t) = y(t; t_0, y_0)$ die Lösung des AWP $dy/dt = f(y)$, $y(t_0) = y_0$. Mit*

$$F_h(y(t)) = (F_{t+h,t} \circ F_{t,t_0})(y_0) = F_{t+h,t_0}(y_0) = y(t + h)$$

wirkt F_h <u>auf Lösungskurven</u> als Zeitverschiebung $F_h : y(t) \to y(t + h)$.

Bemerkung 3.13: *Sei F_h der Fluß von $dy/dt = f(y)$ und $\widetilde{F}_h$ sei der Fluß von $dy/dt = -f(y)$. Dann gilt $F_h \circ \widetilde{F}_h = \widetilde{F}_h \circ F_h = id$, also $\widetilde{F}_h = (F_h)^{-1} = F_{-h}$.*

Beweis: Sei

$$y(t; t_0, y_0) \quad \text{die Lösung des AWP} \quad \frac{dy}{dt} = f(y) \ ; \ y(t_0) = y_0$$

und

$$\widetilde{y}(t; t_0, y_0) \quad \text{die Lösung des AWP} \quad \frac{dy}{dt} = -f(y); \ y(t_0) = y_0 \ .$$

Es gilt $\widetilde{y}(t; t_0, y_0) = y(2\,t_0 - t; t_0, y_0)$, denn diese Funktion löst das zweite AWP, Eindeutigkeit. Damit folgt $F_{t_0-t} \equiv F_{2\,t_0-t,t_0} = \widetilde{F}_{t,t_0} \equiv \widetilde{F}_{t-t_0}$.

Q.E.D.

Beispiele 3.14:

a) *Eine lineare Differentialgleichung* $dy/dt = Ay$ *mit einer konstanten Matrix* A *hat die Lösung* $y(t) = e^{(t-t_0)A}y_0$, *also* $F_{t,t_0} = F_{t-t_0} = e^{(t-t_0)A}$.

b) *Das skalare System* $dy/dt = y^2$, $y(t) \in \mathbb{R}$, *hat die Lösung*

$$y(t) = \frac{y(t_0)}{1-(t-t_0)\,y(t_0)} \ .$$

Damit ist der Fluß die Abbildung

$$F_h : y \to \frac{y}{1-hy} \ , \qquad \text{definiert für} \quad \begin{cases} h \in (-\infty, \frac{1}{y}) & \text{für } \ y > 0 \ , \\ h \in (\frac{1}{y}, \infty) & \text{für } \ y < 0 \ , \\ h \in (-\infty, \infty) & \text{für } \ y = 0 \ . \end{cases}$$

Zusammenfassung: "Lösen" eines autonomen Systems $dy/dt = f(y)$ auf $\mathbb{R}^N$ heißt: finde die 1-parametrige Abbildungsschar $F_h : \mathbb{R}^N \to \mathbb{R}^N$, welche

$$F_0 = id \ , \quad F_{h_2} \circ F_{h_1} = F_{h_2+h_1} \ , \quad F_{-h} = F_h^{-1}$$

erfüllt. Numerisch: finde Approximationen von F_h. Ein numerischer Integrator $I_h : \mathbb{R}^N \to \mathbb{R}^N$ ist eine 1-parametrige Schar von Abbildungen mit möglichst kleinem "Verfahrensfehler" $F_h - I_h$.

Numerisches Lösungsverfahren 3.15: *Zur Lösung des AWP*

$$\frac{dy}{dt} = f(y), \ y(t_0) = y_0$$

wähle Stützwerte $t_0, \ t_1 = t_0 + h_0, \ t_2 = t_1 + h_1, \ \ldots$ *mit* ***"Schrittweiten"*** h_i *und berechne iterativ*

$$y_{k+1} = I_{h_k}(y_k) \ \ \text{mit dem Start} \ \ y_0 \ ,$$

also

$$\begin{aligned} y_{exakt}(t_0 + h_0 + h_1 + \cdots + h_n) &= F_{h_n} \circ F_{h_{n-1}} \circ \ldots \circ F_{h_0}(y_0) \\ &\approx I_{h_n} \circ I_{h_{n-1}} \circ \ldots \circ I_{h_0}(y_0) \end{aligned}$$

<u>Ziel:</u> finde I_h, so daß $I_h - F_h$ klein ist für <u>kleines</u> h (systematische Konstruktionen von I_h folgen später). Problem: schon ein einzelner Schritt $y_0 \to y_1 = I_{h_0}(y_0)$ führt auf eine Nachbartrajektorie. Damit ist zu klären: wie weit können benachbarte Trajektorien im Laufe der Zeit auseinander laufen?

Bemerkung 3.16: *Das AWP $dy/dt = f(y)$, $y(t_0) = y_0$ ist äquivalent zur Integralgleichung*

$$y(t) \;=\; y_0 \;+\; \int_{t_0}^{t} f(y(t))\,dt \;.$$

Lemma 3.17: (Gronwall)

Sei $g : [0,\infty) \to \mathbb{R}$ stetig und es gelte

$$g(t) \le g(0) + L \int_0^t g(h)\,dh \quad \forall\, t \in [0,\infty)$$

mit einer Konstanten $L \ge 0$. Dann folgt $g(t) \le g(0)\,e^{tL}$ $\forall\, t \in [0,\infty]$.

Beweis: Sei $\epsilon > 0$ beliebig. Zeige $g(t) \le (g(0)+\epsilon)\,e^{tL}$ (womit die behauptete Abschätzung folgt, da für jedes t ein beliebig kleines ϵ gewählt werden kann). Angenommen,

$$T := \inf\{\; h \in [0,\infty) \;;\; g(h) \ge (g(0)+\epsilon)\,e^{hL} \;\} \;,$$

existiert. Es gilt

$$g(h) \;<\; (g(0)+\epsilon)\;e^{hL} \quad \forall\; h \in [0,T)$$

und damit

$$g(T) \;=\; \lim_{h\to T-0} g(h) \;\le\; (g(0)+\epsilon)\,e^{TL} \;.$$

Wäre $g(T) \;<\; (g(0)+\epsilon)\,e^{TL}$, so gälte auch $g(h) \;<\; (g(0)+\epsilon)\,e^{hL}$ auf einer Umgebung von T im Widerspruch zu $T = \inf\{...\}$. Also gilt

$$g(T) \;=\; (g(0)+\epsilon)\,e^{TL} \;.$$

Es folgt der Widerpruch

$$\begin{aligned} g(T) \;&\le\; g(0) \;+\; L \int_0^T g(h)\,dh \\ &<\; g(0)+\epsilon \;+\; L \int_0^T (g(0)+\epsilon)\,e^{h\,L}\,dh \;=\; (g(0)+\epsilon)\,e^{T\,L} \,. \end{aligned}$$

Q.E.D.

Hiermit ergibt sich der folgende

Satz 3.18: (Abhängigkeit der Lösung von den Anfangsdaten)

Der Fluß F_h des Systems $dy/dt = f(y)$ mit Lipschitz-stetigem Vektorfeld $\|f(y_0) - f(z_0)\| \le L\,\|y_0 - z_0\|$ ist wieder Lipschitz-stetig mit

$$\|F_h(y_0) - F_h(z_0)\| \;\le\; e^{hL}\,\|y_0 - z_0\| \;, \quad h \ge 0 \;.$$

Beweis: Es seien $y(t_0+h) = F_h(y_0)$ und $z(t_0+h) = F_h(z_0)$ die Lösungen zu den Anfangswerten y_0 bzw. z_0 zum Zeitpunkt t_0. Mit Bemerkung 3.16 gilt

$$\begin{aligned} F_h(y_0) &= y_0 + \int_{t_0}^{t_0+h} \frac{dy(t)}{dt}\,dt &&= y_0 + \int_0^h f(y(t_0+\tau))\,d\tau \\ F_h(z_0) &= z_0 + \int_{t_0}^{t_0+h} \frac{dz(t)}{dt}\,dt &&= z_0 + \int_0^h f(z(t_0+\tau))\,d\tau \end{aligned}$$

und folglich

$$\begin{aligned} \|F_h(y_0) - F_h(z_0)\| &\le \|y_0 - z_0\| + \int_0^h \|f(y(t_0+\tau)) - f(z(t_0+\tau))\|\,d\tau \\ &\le \|y_0 - z_0\| + L\int_0^h \|y(t_0+\tau)) - z(t_0+\tau))\|\,d\tau\ . \end{aligned}$$

Das Gronwall-Lemma mit $g(h) = \|F_h(y_0) - F_h(z_0)\|$, $g(0) = \|y_0 - z_0\|$ liefert sofort die Behauptung.

Q.E.D.

Interpretation:

+ Der Fluß $F_h(y)$ ist Lipschitz-stetig in y (auch glatter, wenn $f(y)$ glatter ist).

– Die Lipschitz-Konstante e^{hL} wächst exponentiell mit der Zeit h:

Langzeitintegration ist numerisch prinzipiell ein Problem für solche Systeme, für welche die Abschätzung aus Satz 3.18 realistisch ist („instabile Differentialgleichungen").

Kapitel 4

Einschrittverfahren: Runge-Kutta(RK)-Theorie

Ziel: zum Fluß F_h von $dy/dt = f(y)$ auf dem $\mathbb{R}^N$ finde *systematisch* approximierende Abbildungen I_h, so daß $\|F_h(y) - I_h(y)\|$ klein ist *für kleine Werte* von h (d.h., die Taylor-Entwicklungen von F_h und I_h sollen in möglichst hoher Ordnung in h übereinstimmen).

4.1 Notation und Definitionen, Bäume, etc.

Definition 4.1:

Gegeben glattes $f : \mathbb{R}^N \to \mathbb{R}^M$, $f(y) = \begin{pmatrix} f_1(y_1, \ldots, y_N) \\ \vdots \\ f_M(y_1, \ldots, y_N) \end{pmatrix}$.

Die n.te Ableitung $f^{(n)}$ *bei* $y \in \mathbb{R}^N$ *ist die* n*-lineare Abbildung*

$$\begin{array}{rccl} f^{(n)}(y): & \mathbb{R}^N \times \cdots \times \mathbb{R}^N & \to & \mathbb{R}^M \\ & (v_1, \ldots, v_n) & \to & f^{(n)}(y)[v_1, \ldots, v_n] \end{array}$$

mit $\Big(f^{(n)}(y)[v_1, \ldots, v_n]\Big)_i = \sum_{j_1=1}^{N} \cdots \sum_{j_n=1}^{N} \frac{\partial^n f_i(y_1, ..., y_n)}{\partial y_{j_1} \ldots \partial y_{j_n}} (v_1)_{j_1} \cdots (v_n)_{j_n}$,

$i = 1, \ldots, M$.

Notation:

$f^{(1)}(y) \equiv f'(y)$ *(mit* $f'(y)[v]$ = *Jacobi-Matrix* $f'(y)$ *wirkend auf* v*),*

$f^{(2)}(y) \equiv f''(y)$

usw.

Bemerkung 4.2:
Die Taylor-Reihe für $f : \mathbb{R}^N \to \mathbb{R}^M$ um $y \in \mathbb{R}^N$ ist in dieser Notation:

$$f(y+v) = f(y) + \frac{1}{1!} f'(y)[v] + \frac{1}{2!} f''(y)[v,v] + \cdots \ .$$

Eigenschaften 4.3:

a) *Symmetrie:* $f^{(n)}(y)[v_1,..,v_i,..,v_j,..,v_n] = f^{(n)}(y)[v_1,..,v_j,..,v_i,..,v_n]$.

b) *"Produktregel": für beliebige $g_i : \mathbb{R}^N \to \mathbb{R}^N$ gilt*

$$\Big(f^{(n)}(y)[g_1(y),\ldots,g_n(y)] \Big)'[v] \ = \ f^{(n+1)}(y)[g_1(y),\ldots,g_n(y),v]$$
$$+ \ f^{(n)}(y)[g_1'(y)[v],\ldots,g_n(y)] \ + \ \cdots \ + \ f^{(n)}(y)[g_1(y),\ldots,g_n'(y)[v]] \ .$$

Beweis: Einsetzen der Definition von $f^{(n)}$.

Beobachtung 4.4: *Die Taylor-Reihe der Lösung $y(t_0+h) = F_h(y_0)$ des AWP $dy/dt = f(y)$, $y(t_0) = y_0$ ist konstruierbar:*

$$y(t_0+h) = y(t_0) + h \frac{dy}{dt}_{|t_0} + \frac{h^2}{2!}\frac{d^2y}{dt^2}_{|t_0} + \cdots \ ,$$

wobei $y(t_0) = y_0$, $\frac{dy}{dt}_{|t_0} = f(y_0)$ und $\frac{d}{dt} g(y(t)) = g'(y)\left[\frac{dy}{dt}\right] = g'(y)[f(y)]$, also

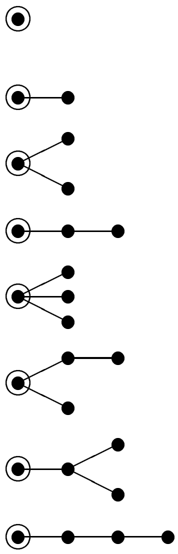
$$\frac{d^k y}{dt^k} = \underbrace{\Big(\ldots \Big((f'(y)[f(y)])'[f(y)] \Big)' \ldots \Big)' [f(y)]}_{k-1 \text{ Richtungsableitungen}} \ .$$

Damit:

$$\begin{aligned}
y(t_0) &= y_0 \\
\frac{dy}{dt}(t_0) &= f(y_0) \\
\frac{d^2y}{dt^2}(t_0) &= f'[f] \qquad (\equiv f'(y_0)[f(y_0)]) \\
\frac{d^3y}{dt^3}(t_0) &= f''[f,f] + f'[f'[f]] \\
\frac{d^4y}{dt^4}(t_0) &= f'''[f,f,f] + 3\, f''[f'[f],f] + f'[f''[f,f]] + f'[f'[f'[f]]] \\
&\ \vdots
\end{aligned}$$

Wir brauchen systematische *Beschreibung all dieser*

Terme	durch	"gewurzelte Bäume"
f	$\longleftrightarrow$	
$f'[f]$	$\longleftrightarrow$	
$f''[f,f]$	$\longleftrightarrow$	
$f'[f'[f]]$	$\longleftrightarrow$	
$f'''[f,f,f]$	$\longleftrightarrow$	
$f''[f'[f],f]$	$\longleftrightarrow$	
$f'[f''[f,f]]$	$\longleftrightarrow$	
$f'[f'[f'[f]]]$	$\longleftrightarrow$	



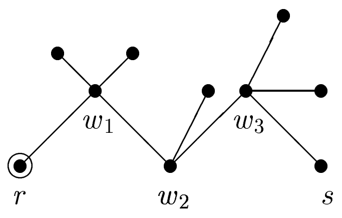
Definition 4.5: *Ein **numerierter gewurzelter Baum** (labelled rooted tree)* $\lambda\rho\tau = (V,E,r)$ *ist ein Tripel aus*

einer ***Knotenmenge***	$V = \{v_1, \dots, v_n\}$	*(vertices)* ,
einer ***Kantenmenge***	$E \subset V \times V$	*(edges)* ,
einer ***Wurzel***	$r \in V$	*(root)* ,

so daß zu jedem $s \in V \backslash \{r\}$ *genau eine Kantenfolge* $e_1, e_2, \dots, e_k \in E$ *der Form*

$$e_1 = (r, w_1), \dots,\ e_i = (w_{i-1}, w_i), \dots,\ e_k = (w_{k-1}, s)$$

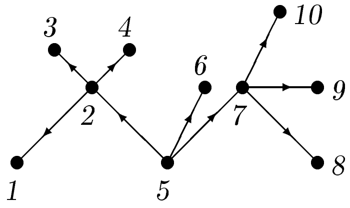
*existiert (ein **Weg** der **Länge** k von der Wurzel nach s):*



Bezeichnung: die Anzahl der Knoten $n = |V|$ *heißt **Ordnung*** $|\lambda\rho\tau|$ *des Baums.*

Bemerkung 4.6: *Für jedes* $s \in V\backslash\{r\}$ *gibt es genau eine Kante der Form* $(v,s) \in E$ *(Aufgabe 4). Daraus folgt* $|V| = |E| + 1$.
Bezeichnung: $(v,s) =$ (**Vater, Sohn**). *Ein Knoten kann mehrere Söhne haben oder auch keine (**Endknoten**, **Blatt**).*

Bemerkung 4.7: *Gewurzelte Bäume sind gerichtete (Kanten sind geordnete Paare) zusammenhängende (... es existiert ein Weg ...) Graphen ohne geschlossene Wege (... genau ein Weg ...). Die Angabe der Wurzel $r \in V$ ist eigentlich überflüssig, sie ist aus den Richtungen $(v,s) \in E$ ("vom Vater zum Sohn") rekonstruierbar: starte irgendwo, gehe zum Vater, zu dessen Vater, ... $\to$ Wurzel.*

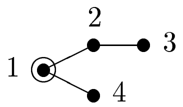


$$\Longrightarrow \qquad r = 5$$

Definition 4.8:

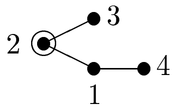
*Zwei numerierte gewurzelte Bäume $\lambda\rho\tau = (V, E, r)$ und $\lambda\rho\tau' = (V', E', r')$ heißen **isomorph**, wenn eine Bijektion $I : V \to V'$ existiert mit $I(r) = r'$ und $(I(v), I(s)) \in E' \;\forall\; (v,s) \in E$.*

Beispiel 4.9:



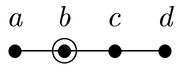
$V = \{\, 1\,,\, 2\,,\, 3\,,\, 4 \,\}$, $r = 1$

isomorph zu



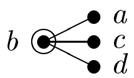
$V' = \{\, 1\,,\, 2\,,\, 3\,,\, 4 \,\}$, $r' = 1$

isomorph zu



$V'' = \{\, a\,,\, b\,,\, c\,,\, d \,\}$, $r'' = b$

nicht isomorph zu



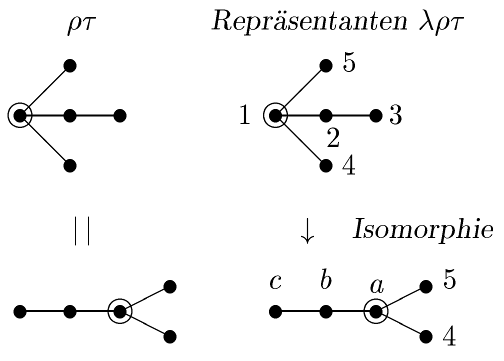
Nach Aufabe 4d) ist Isomorphie eine Äquivalenzrelation auf der Menge aller gewurzelten Bäume.

Definition 4.10:

*Ein **gewurzelter Baum** $\rho\tau$ ist eine Äquivalenzklasse numerierter gewurzelter Bäume unter Isomorphie. Schreibweise:*

$$\rho\tau = [\lambda\rho\tau] \quad \textit{mit Repräsentant} \quad \lambda\rho\tau \in \rho\tau\ .$$

Beispiel 4.11:



$V = \{ 1 , 2 , 3 , 4 , 5 \}$

$\downarrow \downarrow \downarrow \downarrow \downarrow$

$V' = \{ a , b , c , 4 , 5 \}$

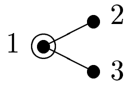
Definition 4.12:

Sei $\lambda\rho\tau = (V, E, r)$ *und* $S_V = \{\pi : V \to V; \pi\,\text{bijektiv}\}$ *die Permutationsgruppe über* V*. Die Untergruppe*

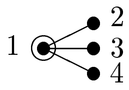
$$G(\lambda\rho\tau) := \{\pi \in S_V; \pi(r) = r; (\pi(v), \pi(s)) \in E \ \forall\ (v, s) \in E\}$$

von S_V *heißt* *Symmetriegruppe* *von* $\lambda\rho\tau$ *(d.h.,* $G(\lambda\rho\tau)$ *ist die Menge aller Isomorphien* $\lambda\rho\tau \to \lambda\rho\tau$*).*

Beispiel 4.13:



$$G(\lambda\rho\tau) \quad = \quad \left\{ \frac{1\,2\,3}{1\,2\,3} \,,\, \frac{1\,2\,3}{1\,3\,2} \right\}$$



$$G(\lambda\rho\tau) \quad = \quad \{\text{alle Permutationen von 2,3,4}\}$$

Bemerkung 4.14: *Sei* $I: \ \lambda\rho\tau = (V, E, r) \ \to \ \lambda\rho\tau' = (V', E', r')$ *eine Isomorphie, dann sind* $G(\lambda\rho\tau)$ *und* $G(\lambda\rho\tau')$ *als Gruppen isomorph:*

$$G(\lambda\rho\tau') = \{ \ I \circ \pi \circ I^{-1} \ ; \ \pi \in G(\lambda\rho\tau) \ \} \ .$$

Damit ist die Symmetriegruppe unabhängig vom Repräsentanten $\lambda\rho\tau$ *eines Baumes* $\rho\tau$*.*

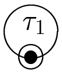
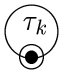
Definition 4.15:

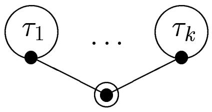
Die **Symmetrie** $\sigma(\rho\tau)$ *eines Baumes ist die Anzahl der Symmetrien*

$$\sigma(\rho\tau) = |G(\lambda\rho\tau)|$$

eines beliebigen Repräsentanten $\lambda\rho\tau \in \rho\tau$*.*

Notation 4.16: ("Produkt von Bäumen")

Seien $\rho\tau_1 =$  $, \ldots, \rho\tau_k =$  . *Setze*

$$[\![\ \rho\tau_1 \cdots \rho\tau_k\ ]\!] :=$$ 

und $\quad [\![\ \rho\tau_1^{m_1} \cdots \rho\tau_k^{m_k}\ ]\!] := [\![\ \underbrace{\rho\tau_1 \cdots \rho\tau_1}_{m_1} \ \cdots\ \underbrace{\rho\tau_k \cdots \rho\tau_k}_{m_k}\ ]\!]\ .$

Satz 4.17: (rekursive Berechnung von Symmetrien)
Seien $\rho\tau_1, \ldots, \rho\tau_k$ *paarweise verschieden. Dann gilt*

$$\sigma([\![\ \rho\tau_1^{m_1} \cdots \rho\tau_k^{m_k}\ ]\!]) \ = \prod_{i=1}^{k} m_i!\ (\sigma(\rho\tau_i))^{m_i}\ .$$

Beweis: vergleiche [Butcher, Satz 144A].

Sei $\lambda\rho\tau = (V, E, r) \in \rho\tau := [\![\ \rho\tau_1^{m_1} \cdots \rho\tau_k^{m_k}\ ]\!]$. Es seien $e \subset E$ die von r ausgehenden Kanten. Betrachte die disjunkten Zerlegungen

$$V = \{r\} \cup (V_{11} \cup \ldots \cup V_{1m_1}) \cup \ldots \cup (V_{k1} \cup \ldots \cup V_{km_k})$$

und

$$E = e \cup (E_{11} \cup \ldots \cup E_{1m_1}) \cup \ldots \cup (E_{k1} \cup \ldots \cup E_{km_k})\ ,$$

so daß

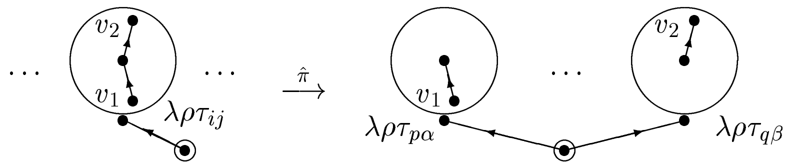
$$\lambda\rho\tau_{ij} = (V_{ij}, E_{ij}, r_{ij}) \in \rho\tau_i \quad \text{mit} \quad i = 1, \ldots, k\ , \quad j = 1, \ldots, m_i,$$

numerierter Repräsentant der j.ten Kopie von $\rho\tau_i$ ist.

Sei $\pi_{ij} : V_{ij} \to V_{ij}$ eine Symmetrie von $\lambda\rho\tau_{ij}$. Für jedes i sind die Kopien $\lambda\rho\tau_{i\alpha}$ und $\lambda\rho\tau_{i\beta}$ mit beliebigem $\alpha, \beta \in \{1, \ldots, m_i\}$ austauschbar, so daß $\lambda\rho\tau_{i1}, \ldots, \lambda\rho\tau_{im_i}$ beliebig permutiert werden können. Formal: mit Isomorphien $I_{i,\alpha,\beta} : V_{i\alpha} \to V_{i\beta}$ liefern Permutationen $p_i : \{1, \ldots, m_i\} \to \{1, \ldots, m_i\}$ und Symmetrien $\pi_{ij} : V_{ij} \to V_{ij}$ der Teilbäume die Symmetrien

$$\begin{array}{rcl} \hat{\pi}(p_1, \ldots, p_k; \pi_{11}, \ldots, \pi_{1m_1}, \ldots, \pi_{k1}, \ldots, \pi_{km_k}) : & V \to V & \\ v \in V_{ij} \quad \to & \pi_{ip_i(j)}(\ I_{i,j,p_i(j)}(v)\ ) \in V_{ip_i(j)} & \end{array} \tag{\#}$$

von $\lambda\rho\tau$. Andere Symmetrien existieren nicht. Angenommen, es existiert eine Symmetrie $\hat{\pi} : V \to V$ von $\lambda\rho\tau$, welche die Indizes V_{ij} von $\lambda\rho\tau_{ij}$ nicht vollständig auf die Indizes $V_{i\alpha}$ einer anderen Kopie $\lambda\rho\tau_{i\alpha}$ abbildet: angenommen



$$v_1, v_2 \in V_{ij} \quad \xrightarrow{\hat{\pi}} \quad v_1 \in V_{p\alpha}\ , v_2 \in V_{q\beta} \quad \text{mit} \quad (p,\alpha) \neq (q,\beta)\ .$$

Wäre $\hat{\pi}$ eine Symmetrie, so müßte es immer noch einen (gerichteten) Weg von v_1 nach v_2 geben. Widerspruch!

Damit ist die Anzahl der Symmetrien (#) von $\lambda\rho\tau$ gegeben durch die Anzahl der Permutationen $p_1, \ldots, p_k$ mal der Anzahl der Symmetrien

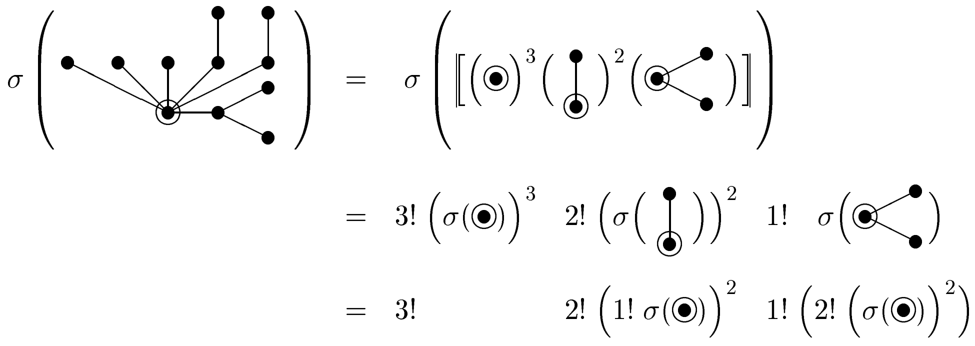
$$\pi_{11}, \ldots, \pi_{1m_1}, \ldots, \pi_{k1}, \ldots, \pi_{km_k}\ ,$$

also $\quad \sigma(\rho\tau) \ = \ m_1! \cdots m_k!\ (\sigma(\rho\tau_1))^{m_1} \cdots (\sigma(\rho\tau_k))^{m_k}\ .$

Q.E.D.

Mit $\sigma(\circledcirc) = 1$ erlaubt Satz 4.17 die rekursive Berechnung der Symmetrie komplizierter Bäume aus den Symmetrien einfacher Bäume:

Beispiel 4.18:



$= \ 24\ .$

Definition 4.19:

Ein numerierter Baum $\lambda\rho\tau = (V, E, r)$ mit $V \subset \mathbb{N}$ heißt **monoton (numeriert)**, *wenn $v < s\ \forall\ (v,s) \in E$.*

Notwendigerweise muß $r = \min V$ gelten.

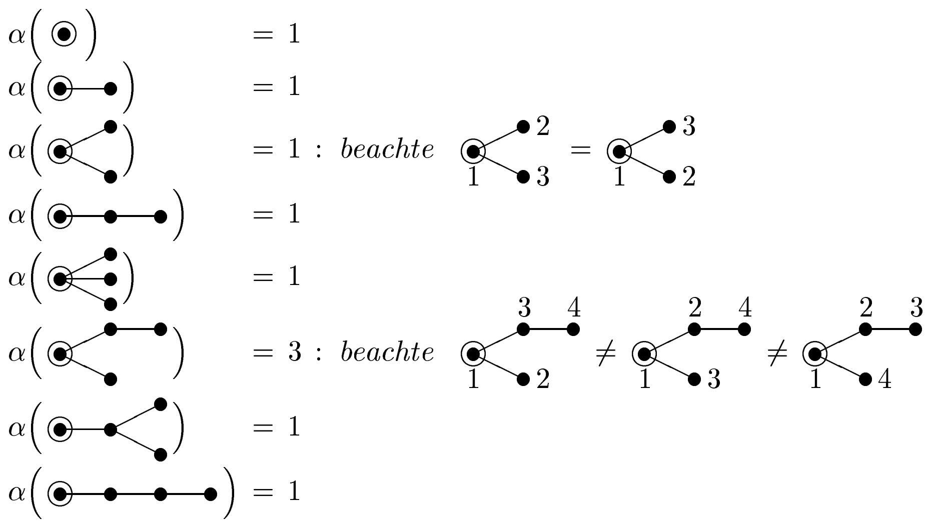
Definition 4.20:

Sei $V \subset \mathbb{N}$ mit $|V| = |\rho\tau|$ fixiert. Die **Anzahl der monotonen Numerierungen** *von $\rho\tau$ ist*

$$\alpha(\rho\tau) \;:=\; |\,\{\ \lambda\rho\tau = (V,E,r) \in \rho\tau\ ;\ \lambda\rho\tau \text{ ist monoton}\ \}\,|\ .$$

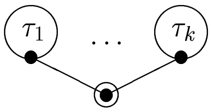
Offensichtlich ist α unabhängig von der gewählten Indexmenge V.

Beispiel 4.21:



Definition 4.22:

Die **Dichte** *γ eines Baums $\rho\tau = [\![\ \rho\tau_1 \cdots \rho\tau_k\ ]\!] =$*

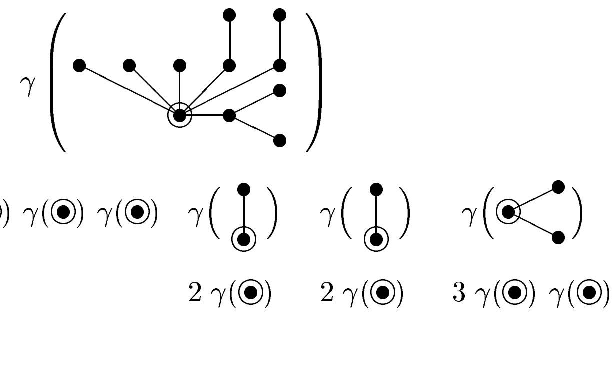


(die $\rho\tau_i$ dürfen übereinstimmen) ist rekursiv durch

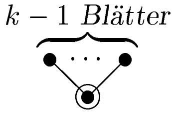
$$\gamma(\rho\tau) \;=\; |\rho\tau|\ \gamma(\rho\tau_1)\ \cdots\ \gamma(\rho\tau_k)$$

mit $\gamma(\odot) = 1$ definiert.

Beispiel 4.23: *a)*



b) $\gamma(\underbrace{\odot\!-\!\bullet\!-\!\bullet \cdots \bullet\!-\!\bullet}_{k \text{ Knoten}}) \;=\; k\,\gamma(\underbrace{\odot\!-\!\bullet \cdots \bullet\!-\!\bullet}_{k-1 \text{ Knoten}}) \;=\; \ldots \;=\; k!$

c) $\gamma\Bigg($ 

$\Bigg) \;=\; k\; \underbrace{\gamma(\odot) \cdots \gamma(\odot)}_{k-1} \;=\; k\,.$

Satz 4.24:

Für jeden gewurzelten Baum $\rho\tau$ gilt: $\quad \alpha(\rho\tau) \;=\; \dfrac{|\rho\tau|\,!}{\gamma(\rho\tau)\,\sigma(\rho\tau)}\,.$

Beweis: vergleiche [Butcher, Satz 145E].

Induktion nach $|\rho\tau|$. Für $\rho\tau = \odot$ ist die Behauptung richtig. Betrachte nun $\rho\tau = [\![\, (\rho\tau_1)^{m_1} \cdots (\rho\tau_k)^{m_k} \,]\!]$ mit paarweise verschiedenen $\rho\tau_1, \ldots, \rho\tau_k$.
Konstruiere monotone $\lambda\rho\tau = (V, E, r) \in \rho\tau$ mit $V = \{1, \ldots, n\}$, $n = |\rho\tau|$. Betrachte dazu die disjunkten Zerlegungen

$$\{2, \ldots, n\} = (V_{11} \cup \ldots \cup V_{1m_1}) \cup \ldots \cup (V_{k1} \cup \ldots \cup V_{km_k})$$

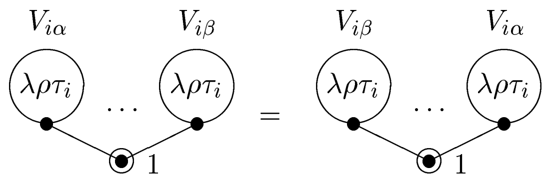
mit $|V_{ij}| = |\rho\tau_i|$, $j = 1, \ldots m_i$ (beachte $r = 1$, so daß die Unterbäume mit $2, \ldots, n$ numeriert werden). Es gibt hierfür

$$\frac{(n-1)\,!}{(|\rho\tau_1|\,!)^{m_1} \cdots (|\rho\tau_k|\,!)^{m_k}}$$

Möglichkeiten. Benutze V_{ij} zur Numerierung der j.ten Kopie von $\rho\tau_i$. Sie kann in $\alpha(\rho\tau_i)$-facher Weise monoton numeriert werden. Dies ergibt

$$\frac{(n-1)\,!\;\alpha(\rho\tau_1)^{m_1} \cdots \alpha(\rho\tau_k)^{m_k}}{(|\rho\tau_1|\,!)^{m_1} \cdots (|\rho\tau_k|\,!)^{m_k}}$$

monotone $\lambda\rho\tau \in \rho\tau$. Davon sind diejenigen identisch, die für fixiertes i durch Permutation der Indexmengen V_{ij}, $j = 1, \ldots, m_i$, entstehen:



Ergebnis: $\quad \boxed{\alpha(\rho\tau) = \dfrac{(n-1)!}{m_1! \cdots m_k!} \left(\dfrac{\alpha(\rho\tau_1)}{|\rho\tau_1|\,!}\right)^{m_1} \cdots \left(\dfrac{\alpha(\rho\tau_k)}{|\rho\tau_k|\,!}\right)^{m_k}}\,.$

Induktiv gelte $\dfrac{\alpha(\rho\tau_i)}{|\rho\tau_i|\,!} = \dfrac{1}{\gamma(\rho\tau_i)\sigma(\rho\tau_i)}$ (beachte $|\rho\tau_i| < |\rho\tau|$). Es folgt

$$\begin{aligned}\alpha(\rho\tau) &= \frac{n\,(n-1)\,!}{n\,(\gamma(\rho\tau_1))^{m_1}\cdots(\gamma(\rho\tau_k))^{m_k}\; m_1!\cdots m_k!\,(\sigma(\rho\tau_1))^{m_1}\cdots(\sigma(\rho\tau_k))^{m_k}}\\ &= \frac{n!}{\gamma(\rho\tau)\,\sigma(\rho\tau)}\end{aligned}$$

mit Satz 4.17 und Definition 4.22.

Q.E.D.

4.2 Die Taylor-Entwicklung der exakten Lösung

Ziel: formalisiere Beobachtung 4.4.

Definition 4.25:

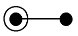
Sei $y \in \mathbb{R}^N$ und (glattes) $f : \mathbb{R}^N \to \mathbb{R}^N$ vorgegeben. Einem gewurzelten Baum $\rho\tau$ wird das **elementare Differential** *$D_{f,y}(\rho\tau) \in \mathbb{R}^N$ (an der Stelle y) durch die rekursive Definition*

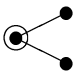
$$D_{f,y}([\![\,\rho\tau_1\cdots\rho\tau_k\,]\!]) = f^{(k)}(y)[\,D_{f,y}(\rho\tau_1)\,,\,\ldots\,,\,D_{f,y}(\rho\tau_k)\,]$$

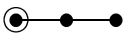
mit $D_{f,y}(\odot) = f(y)$ zugeordnet.

Beispiel 4.26:

$D_{f,\cdot}\Big($  $\Big) = f$

$D_{f,\cdot}\Big($  $\Big) = f'[f]$

$D_{f,\cdot}\Big($  $\Big) = f''[f,f]$

$D_{f,\cdot}\Big($  $\Big) = f'[f'[f]]$

$\vdots$

Dies sind die "Terme" der Taylor-Entwicklung aus 4.4.

Satz 4.27:

Für eine Lösung y von $\dfrac{dy}{dt} = f(y)$ gilt mit α aus Definition 4.20:

$$\frac{d^n y}{dt^n} = \sum_{\substack{\rho\tau\\ |\rho\tau|=n}} \alpha(\rho\tau) D_{f,y}(\rho\tau)\ , \quad n = 1, 2, \ldots\ .$$

Beweis: Zu zeigen: $$\frac{d^n y}{dt^n} \;=\; \sum_{\substack{\lambda\rho\tau \\ \text{monoton} \\ |\lambda\rho\tau| = n}} \mathrm{D}_{f,y}([\lambda\rho\tau]) \;.$$

Induktion nach n:

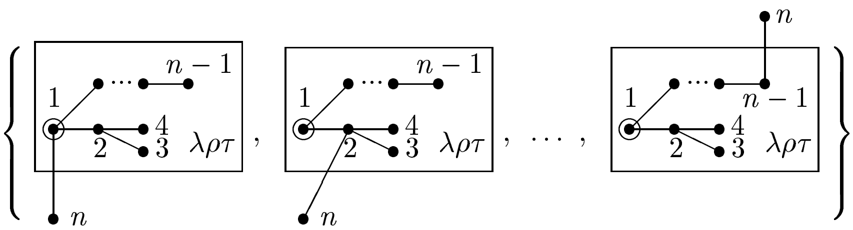
<u>$n = 1$:</u> $$\frac{dy}{dt} = \mathrm{D}_{f,y}(\odot) = f(y) \quad \text{ist erfüllt.}$$

<u>$n-1 \to n$:</u> $$\frac{d^n y}{dt^n} = \frac{d}{dt} \sum_{\substack{\lambda\rho\tau \\ \text{monoton} \\ |\lambda\rho\tau| = n-1}} \mathrm{D}_{f,y}([\lambda\rho\tau]) \;=\; \sum_{\substack{\lambda\rho\tau \\ \text{monoton} \\ |\lambda\rho\tau| = n-1}} \Big(\mathrm{D}_{f,y}([\lambda\rho\tau])\Big)'[f(y)] \;.$$

Mit der "Produktregel" 4.3.b) gilt offensichtlich

$$\Big(\mathrm{D}_{f,y}([\lambda\rho\tau])\Big)'[f(y)] \;=\; \sum_{\widetilde{\lambda\rho\tau} \in M_{\lambda\rho\tau}} \mathrm{D}_{f,y}([\widetilde{\lambda\rho\tau}])$$

mit

$M_{\lambda\rho\tau} =$ 

(hierbei sei o.B.d.A. $\lambda\rho\tau$ monoton mit $V = \{1, \ldots, n-1\}$ numeriert). Dies sind wieder monoton numerierte Bäume. Jeder monoton mit $\tilde{V} = \{1, \ldots, n\}$ numerierte Baum $\widetilde{\lambda\rho\tau}$ entsteht auf diese Weise. Es folgt

$$\sum_{\substack{\lambda\rho\tau \\ \text{monoton} \\ |\lambda\rho\tau| = n-1}} \sum_{\widetilde{\lambda\rho\tau} \in M_{\lambda\rho\tau}} \mathrm{D}_{f,y}([\widetilde{\lambda\rho\tau}]) \;=\; \sum_{\substack{\widetilde{\lambda\rho\tau} \\ \text{monoton} \\ |\widetilde{\lambda\rho\tau}| = n}} \mathrm{D}_{f,y}([\widetilde{\lambda\rho\tau}]) \;.$$

Q.E.D.

Bezeichnung 4.28:

Die Taylor-Entwicklung

$$F_h(y) \;\simeq\; y + \sum_{n=1}^{\infty} \frac{h^n}{n!} \sum_{\substack{\rho\tau \\ |\rho\tau| = n}} \alpha(\rho\tau)\, D_{f,y}(\rho\tau)$$

des Flusses F_h mittels elementarer Differentiale heißt **Butcher-Reihe**. *Jeder Term ist mittels $f, f', f'', \ldots$ an jeder Stelle y berechenbar.*

4.3 Verfahrensfehler

Definition 4.29:

Ein **Einschritt-Verfahren** *(Integrator) der (lokalen)* **Konsistenzordnung** p *zur Lösung von* $\dfrac{dy}{dt} = f(y)$ *auf dem* $\mathbb{R}^N$ *ist eine 1-parametrige Schar von Abbildungen* $I_h : \mathbb{R}^N \to \mathbb{R}^N$ *mit dem* **lokalen Verfahrensfehler**

$$F_h(y) - I_h(y) \; = \; O(h^{p+1}) \; .$$

Der Term $e(y)$ *in* $O(h^{p+1}) = e(y)\, h^{p+1} + O(h^{p+2})$ *heißt* **führender Fehlerkoeffizient**.

Beispiel 4.30: *Durch Abschneiden der Taylor-Entwicklung 4.28 nach der* p*.ten Ordnung in* h *ergeben sich die* **Taylor-Verfahren:**

Ordnung 1: $I_h(y) = y + hf(y)$ *(Euler-(Polygonzug-)Verfahren)*

Ordnung 2: $I_h(y) = y + hf(y) + \dfrac{h^2}{2}\, f'(y)[f(y)]$

Ordnung p: $I_h(y) = y + \displaystyle\sum_{n=1}^{p} \frac{h^n}{n!} \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \alpha(\rho\tau) D_{f,y}(\rho\tau)$

Problem: alle partiellen y*-Ableitungen von* f *bis zur Ordnung* $p-1$ *sind zu implementieren und in jedem Zeitschritt*

$$y(t) \to I_h(y(t)) \approx y(t+h)$$

auszuwerten. Damit sind Taylor-Verfahren in der Praxis i.a. kaum einsetzbar.

Bemerkung 4.31: *Die Frage, ob die Taylor-Reihe 4.28 den exakten Fluß darstellt, stellt sich nicht:*

$$F_h(y) = y + \sum_{n=1}^{p} \frac{h^n}{n!} \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \alpha(\rho\tau) D_p(\rho\tau) \; + \; \text{Fehler}(h,p)$$

Analytizität heißt: $\lim\limits_{p\to\infty} \text{Fehler}(h,p) = 0$ *bei festem* h*.*

Numerischer Fehler der Taylor-Verfahren:

$$F_h - I_h = \text{Fehler}(h,p) \quad \underline{\text{bei festem } p} \text{ und wählbarem (kleinen) } h.$$

Definition 4.32:

Zu festem $T \in \mathbb{R}$ betrachte $h(n) = T/n$, $n \in \mathbb{N}$. Ein Verfahren I_h heißt (global) **konvergent** *mit der* **Konvergenzordnung** *p, wenn*

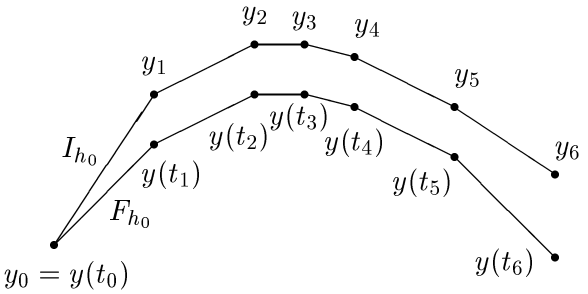
$$F_T(y) - (\underbrace{I_{h(n)} \circ \cdots \circ I_{h(n)}}_{n})(y) \;=\; O(h(n)^p) \;=\; O\Big(\frac{1}{n^p}\Big)$$

("globaler Verfahrensfehler").

Bemerkung 4.33: *Hierbei werden zur Integration über ein längeres Zeitintervall T n äquidistante Zeitschritte mit $h = T/n$ betrachtet. In der Praxis wird man mit variablen Schrittweiten $h_0, h_1, \ldots, h_{n-1}$ (**"adaptiv"**) arbeiten:*

$$t_{i+1} = t_i + h_i \ , \quad y_{i+1} = I_{h_i}(y_i) \ .$$

Dann wird der Zeitschritt $F_{t_n,t_0} \equiv F_{h_{n-1}} \circ \ldots \circ F_{h_0} \approx I_{h_{n-1}} \circ \ldots \circ I_{h_0}$ durch $y(t_n) = F_{t_n - t_0}(y_0) \approx y_n$ approximiert:



Satz 4.34: (Globale Fehler aus lokalen Fehlern)

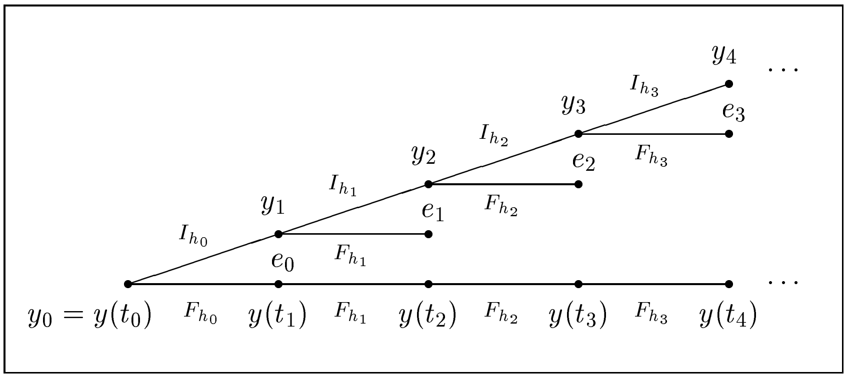
Sei $y(t) = F_{t-t_0}(y_0)$ die exakte Lösung des AWP $\dfrac{dy}{dt} = f(y)$, $y(t_0) = y_0$ mit Lipschitz-stetigem $f : \|f(y) - f(z)\| \le L\,\|y - z\|$. Sei I_h ein numerisches Verfahren, mit dem mittels positiver Schrittweiten $h_0, h_1, \ldots$ numerische Stützwerte $y_{i+1} = I_{h_i}(y_i)$ zur Approximation von $y(t)$ zu den Zeiten $t_{i+1} = t_i + h_i$ berechnet werden. Es seien

$$e_i := F_{h_i}(y_i) - I_{h_i}(y_i)$$

die lokalen Fehler und

$$E_i := y(t_i) - y_i = (F_{h_{i-1}} \circ \ldots \circ F_{h_0})(y_0) - (I_{h_{i-1}} \circ \ldots \circ I_{h_0})(y_0)$$

die globalen Fehler:



Es gilt

$$\begin{aligned} \|E_n\| &\leq \sum_{i=0}^{n-1} \|e_i\| \, e^{(t_n - t_{i+1})L} \leq \left(\sum_{i=0}^{n-1} \|e_i\| \right) e^{(t_n - t_0)L} \\ &\leq n \left(\max_{i=0 \dots n-1} \|e_i\| \right) e^{(t_n - t_0)L} . \end{aligned}$$

Für konstante Schrittweiten $h := h_0 = h_1 = \ldots = T/n > 0$ gilt auch

$$\|E_n\| \leq \left(\max_{i=0 \dots n-1} \|e_i\| \right) \frac{e^{TL} - 1}{e^{hL} - 1} < \left(\max_{i=0 \dots n-1} \|e_i\| \right) \frac{e^{TL} - 1}{hL} .$$

Beweis: Anwendung von Satz 3.18 auf

$$\begin{aligned} E_{i+1} &= y(t_{i+1}) - y_{i+1} \\ &= \underbrace{F_{h_i}(y(t_i)) - F_{h_i}(y_i)}_{\text{Satz 3.18}} + \underbrace{F_{h_i}(y_i) - I_{h_i}(y_i)}_{e_i} \end{aligned}$$

liefert die Rekursion

$$\|E_{i+1}\| \leq e^{h_i L} \|E_i\| + \|e_i\| .$$

Mit $E_0 = 0$ folgt

$$\begin{aligned} \|E_1\| &\leq \|e_0\| \\ \|E_2\| &\leq e^{h_1 L} \|E_1\| + \|e_1\| \leq e^{h_1 L} \|e_0\| + \|e_1\| \\ \|E_3\| &\leq e^{h_2 L} \|E_2\| + \|e_2\| \leq e^{(h_1 + h_2)L} \|e_0\| + e^{h_2 L} \|e_1\| + \|e_2\| \\ &\vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \end{aligned}$$

und damit

$$\|E_n\| \overset{(*)}{\leq} \sum_{i=0}^{n-1} e^{(h_{i+1}+\cdots+h_{n-1})L}\,\|e_i\| \;\leq\; \sum_{i=0}^{n-1} e^{(t_n-t_{i+1})L}\,\|e_i\| \;\leq\; e^{(t_n-t_0)L}\sum_{i=0}^{n-1}\|e_i\|\,.$$

Im äquidistanten Fall folgt aus $(*)$:

$$\|E_n\| \;\leq\; \sum_{i=0}^{n-1} e^{(n-1-i)\,hL}\,\|e_i\| \;\leq\; \Big(\sum_{j=0}^{n-1} e^{jhL}\Big)\Big(\max_{i=0..n-1}\|e_i\|\Big)$$

mit

$$\sum_{j=0}^{n-1}(e^{hL})^j \;=\; \frac{e^{nhL}-1}{e^{hL}-1} \;\leq\; \frac{e^{nhL}-1}{hL}\,.$$

Q.E.D.

Bemerkung 4.35:

Für ein festes Integrationsintervall $t_0 \to t_0+T$ gilt bei einem Verfahren I_h der Konsistenzordnung p

$$\|e_i\| = O(h^{p+1}) = O\Big(\frac{1}{n^{p+1}}\Big) \quad \textit{mit } h = T/n,$$

also

$$\|E_n\| \leq n\,\Big(\max_{i=0..n-1}\|e_i\|\Big)\,e^{TL} = O\Big(\frac{1}{n^p}\Big)\,.$$

Merke: lokale Konsistenzordnung = globale Konvergenzordnung.

Bezeichnung 4.36:

*Ein Verfahren mit Ordnung $p \geq 1$ heißt **konsistent/konvergent**. Man kann über ein gegebenes Zeitintervall $[t_0, t_0+T]$ beliebig genau integrieren, wenn man nur $h = T/n$ klein genug wählt:*

$$F_T(y_0) - (I_{T/n}\circ\ldots\circ I_{T/n})(y_0) = O\Big(\Big(\frac{T}{n}\Big)^p\Big) \overset{n\to\infty}{\longrightarrow} 0\,.$$

Satz 4.37: (Globale Fehler aus lokalen Verfahrens- und Rundungsfehlern)

In Satz 4.34 seien

$$y_{i+1} = I_{h_i}(y_i) + \varepsilon_i$$

die mit den absoluten Auswertungs-(Rundungs-)Fehlern ε_i von $I_h(y_i)$ behafteten tatsächlich berechneten Stützwerte. Es gelten die Abschätzungen aus Satz 4.34 mit $\|e_i\|$ ersetzt durch $\|e_i\| + \|\varepsilon_i\|$, speziell

$$\|E_n\| \leq \Big(\max_{i=0\ldots n-1}(\|e_i\| + \|\varepsilon_i\|)\Big)\,\frac{e^{TL}-1}{hL}\,.$$

Beweis: Vergleiche mit dem Beweis von Satz 4.34:

$$E_{i+1} \;=\; y(t_{i+1}) - y_{i+1} \;=\; \underbrace{F_{h_i}(y(t_i)) - F_{h_i}(y_i)}_{\text{Satz 3.18}} \;+\; \underbrace{F_{h_i}(y_i) - I_{h_i}(y_i)}_{e_i} \;-\; \varepsilon_i$$

$$\Longrightarrow \quad \|E_{i+1}\| \;\le\; e^{h_i L}\,\|E_i\| + \|e_i\| + \|\varepsilon_i\| \;.$$

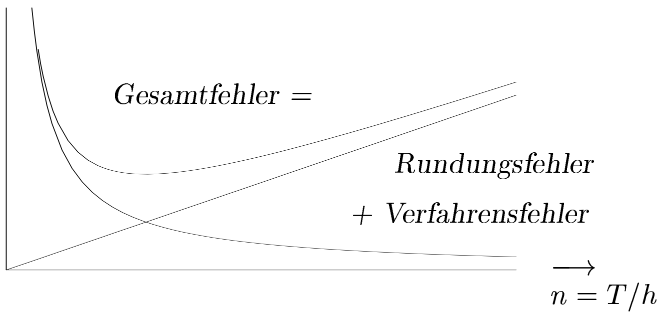
Alle Abschätzungen folgen hieraus.

Q.E.D.

Bemerkung 4.38:

Im äquidistanten Fall $h = T/n$ folgt für ein Verfahren der Ordnung p mit $\max\limits_{i=0..n-1} \|e_i\| = e\,h^{p+1} + O(h^{p+2})$ und $\varepsilon := \max\limits_{i=0..n-1} \|\varepsilon_i\|$:

$$\begin{aligned} \|E_n\| \;&\le\; \Big(e\,h^{p+1} + O(h^{p+2}) + \varepsilon\Big)\,\frac{e^{TL}-1}{hL} \\ &=\; \Big(\,e\,h^p + \frac{\varepsilon}{h} \;+\; O(h^{p+1})\,\Big)\,\frac{e^{TL}-1}{L} \;. \end{aligned}$$



Dabei wird $e\,h^p + \varepsilon/h$ minimal, wenn $e\,h^{p+1} = \varepsilon/p$ gilt, also

$$\text{lokaler Verfahrensfehler} \;=\; \frac{\text{Rundungsfehler}}{\text{Ordnung}} \;.$$

Es folgt die **Faustregel***:*

> *die Schrittweite eines Verfahrens darf höchstens so klein gewählt werden, daß der lokale Verfahrensfehler von der Größenordnung der absoluten Rundungsfehler ist (intuitiv klar):*

$$h_{min} \;\approx\; \left(\frac{\varepsilon}{e\,p}\right)^{\frac{1}{p+1}} \;.$$

4.4 Schrittweitensteuerung

Wichtig in der Praxis! Man wird versuchen, einen Zeitschritt mit möglichst großer Schrittweite durchzuführen unter der Nebenbedingung, den lokalen Verfahrensfehler unter einer gegebenen Schranke ε zu halten, also

$$\|F_h(y) - I_h(y)\| \;=\; \|e(y)\, h^{p+1}\| \;+\; O(h^{p+2}) \underset{(\approx)}{\leq} \varepsilon \;.$$

Der benötigte führende Fehlerkoeffizient $e(y)$ ist selten analytisch abzuschätzen und ist daher i.a. numerisch zu approximieren. Probiere dazu das Verfahren mit verschiedenen Schrittweiten aus, speziell: vergleiche $I_h(y)$ mit $I_{h/2}(I_{h/2}(y))$.

Lemma 4.39:

Für ein Verfahren $I_h(y) = F_h(y) - e(y)\, h^{p+1} + O(h^{p+2})$ *der Ordnung* p *gilt*

$$F_{h/2}(I_{h/2}(y)) \;=\; F_h(y) \;-\; e(y) \left(\frac{h}{2}\right)^{p+1} \;+\; O(h^{p+2}) \;.$$

Beweis: Für den Lipschitz-stetigen Fluß gilt $F_h(y) = y + h\, G_h(y)$ mit Lipschitz-stetigem $G_h(y)$, also

$$F_h(\tilde{y}) - F_h(\hat{y}) \;=\; \tilde{y} - \hat{y} \;+\; h\, O(\tilde{y} - \hat{y}) \;.$$

Mit $\tilde{y} = I_h(y)$, $\hat{y} = F_h(y)$ ergibt sich

$$\begin{aligned} F_h(I_h(y)) - F_{2h}(y) \;&=\; \underbrace{I_h(y) - F_h(y)}_{-e(y)\, h^{p+1} + O(h^{p+2})} \;+\; \underbrace{h\, O\Big(I_h(y) - F_h(y)\Big)}_{O(h^{p+2})} \\ &=\; -\, e(y)\, h^{p+1} \;+\; O(h^{p+2}) \;. \end{aligned}$$

Die Behauptung folgt mit $h \to h/2$.

Q.E.D.

Einerseits gilt

$$I_h(y) \;=\; F_h(y) \;-\; e(y)\, h^{p+1} \;+\; O(h^{p+2})$$

andererseits bei Lipschitz-stetigem $e(y)$:

$$\begin{aligned} I_{h/2}(I_{h/2}(y)) \;&=\; F_{h/2}(I_{h/2}(y)) \;-\; \underbrace{e(I_{h/2}(y))}_{e(y)+O(h)} \left(\frac{h}{2}\right)^{p+1} \;+\; O(h^{p+2}) \\ &\overset{(4.39)}{=}\; F_h(y) \;-\; 2\, e(y) \left(\frac{h}{2}\right)^{p+1} \;+\; O(h^{p+2}) \;. \end{aligned}$$

Hiermit folgt

$$I_{h/2}(I_{h/2})(y) \; - \; I_h(y) \; = \; e(y)\, h^{p+1} \left(1 - \frac{1}{2^p}\right) \; + \; O(h^{p+2}).$$

Dies liefert einen Schätzwert des führenden lokalen Fehlerterms:

$$\boxed{e(y)\, h^{p+1} \; = \; \frac{2^p}{2^p - 1} \Big(I_{h/2}(I_{h/2}(y)) \; - \; I_h(y)\Big) \; + \; O(h^{p+2}) \; .}$$

Bemerkung 4.40: *Ein Zeitschritt mit dieser Abschätzung des lokalen Verfahrensfehlers führt zum dreifachen Aufwand: I_h und zweimal $I_{h/2}$ sind auszuwerten. Bei speziellen Verfahren (siehe Sektion 4.5.4) können solche Abschätzungen günstiger berechnet werden.*

Automatische Schrittweitensteuerung 4.41:

Finde h so, daß für das Verfahren I_h der Ordnung p am Punkt y gilt:

$$\|F_h(y) - I_h(y)\| \; \approx \; \varepsilon \; = \; \textit{vorgegebene Genauigkeit}.$$

Wähle dazu ein h und berechne

$$E(h, y) \; := \; \frac{2^p}{2^p - 1} \Big(I_{h/2}(I_{h/2}(y)) \; - \; I_h(y)\Big) \; .$$

Gilt $\|E(h,y)\| \approx \varepsilon$, dann akzeptiere h als Schrittweite und $I_h(y)$ (oder besser $I_h(y) + E(h,y)$) als Approximation von $F_h(y)$.

Wenn nicht, so ist

$$\tilde{h} \; := \; h \left(\frac{\varepsilon}{\|E(h,y)\|}\right)^{\frac{1}{p+1}}$$

Kandidat für eine Schrittweite mit lokalem Verfahrensfehler $\approx \varepsilon$:

$$\begin{aligned} \|T_{\tilde{h}}(y) - I_{\tilde{h}}(y)\| &= \|e(y)\, \tilde{h}^{p+1}\| \; + \; O(\tilde{h}^{p+2}) \\ &= \|e(y)\, h^{p+1}\| \left(\frac{\tilde{h}}{h}\right)^{p+1} \; + \; O(\tilde{h}^{p+2}) \\ &= \|E(h,y)\| \left(\frac{\tilde{h}}{h}\right)^{p+1} \; + \; O(\tilde{h}^{p+2}) \; = \; \varepsilon \; + \; O(\tilde{h}^{p+2}) \; . \end{aligned}$$

Berechne $E(\tilde{h}, y)$ und teste erneut $\|E(\tilde{h}, y)\| \approx \varepsilon$ usw. Hierdurch wird die Schrittweite bei Bedarf verkleinert oder vergrößert.

4.5 Runge-Kutta-Verfahren

Approximiere die Taylor-Reihe

$$F_h(y) \;=\; y \;+\; h\,f(y) \;+\; \frac{h^2}{2!}\,f'(y)[f(y)] \;+\; \frac{h^3}{3!}\Big(f''[f,f]+f'[f'[f]]\Big) \;+\; O(h^4)$$

auf möglichst hohe Ordnung bei geringem Aufwand (möglichst unter Vermeidung von Ableitungen $f', f'', \ldots$).

Idee: werte f an verschiedenen Stellen aus und bilde Linearkombinationen, z.B.

$$\begin{aligned} I_h(y) \;&=\; y \;+\; h\,f\Big(y+\frac{h}{2}\,f(y)\Big) \qquad \text{(Runge)} \\ &\overset{(4.2)}{=} \underbrace{y \;+\; h\,f(y) \;+\; \frac{h^2}{2}\,f'(y)[f(y)]}_{\text{ok}} \;+\; \underbrace{\frac{h^3}{8}\,f''(y)[f(y),f(y)] \;+\; O(h^4)}_{\text{Fehler}} \;. \end{aligned}$$

4.5.1 Die RK-Familie

Definition 4.42:

*Ein s-**stufiges RK-Verfahren** zur Lösung von $dy/dt = f(y)$ ist eine Abbildung der Form*

$$I_h(y) \;=\; y \;+\; h\,\sum_{j=1}^{s} b_j\,f(y_j) \;,$$

*wobei die **Zwischenstufen** y_j die Lösung eines Gleichungssystems der Form*

$$y_i \;=\; y \;+\; h\,\sum_{j=1}^{s} a_{ij}\,f(y_j) \;, \quad i=1,\ldots,s$$

sind.

Bemerkung 4.43: *Eine Implementierung geschieht meist mit $k_j = hf(y_j)$ in der äquivalenten Form*

$$I_h(y) \;=\; y \;+\; \sum_{j=1}^{s} b_j\,k_j \;,$$

wobei $$k_i \;=\; h\,f\Big(y+\sum_{j=1}^{s} a_{ij}\,k_j\Big) \;, \quad i=1,\ldots,s \;.$$

Bemerkung 4.44: *Für nichtautonome Systeme* $\dfrac{dy}{dt} = f(t,y)$ *in der Form* $\dfrac{d}{dt}\begin{pmatrix} t \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ f(t,y) \end{pmatrix}$ *ergibt sich* $I_h\left(\begin{pmatrix} t \\ y \end{pmatrix}\right) = \begin{pmatrix} t \\ y \end{pmatrix} + h \sum_{j=1}^{s} b_j \begin{pmatrix} 1 \\ f(t_j, y_j) \end{pmatrix}$ *mit*

$$\begin{pmatrix} t_i \\ y_i \end{pmatrix} = \begin{pmatrix} t \\ y \end{pmatrix} + h \sum_{j=1}^{s} a_{ij} \begin{pmatrix} 1 \\ f(t_j, y_j) \end{pmatrix} , \quad i = 1, \ldots, s ,$$

also

$$t_i = t + c_i\, h \qquad \text{mit} \quad c_i = \sum_{j=1}^{s} a_{ij} .$$

Der numerische Zeitschritt $t \to t + h$ *ist damit gegeben durch*

$$I_{t+h,t}(y) = y + h \sum_{j=1}^{s} b_j\, f(t + c_j\, h, y_j)$$

mit
$$y_i = y + h \sum_{j=1}^{s} a_{ij}\, f(t + c_j h, y_j) , \quad i = 1, \ldots, s .$$

Bezeichnung 4.45: *Die Parameter des Verfahrens werden als* ***Butcher-Schema***

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \quad = \quad \begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array} \qquad \text{mit} \qquad c_i = \sum_{j=1}^{s} a_{ij}$$

angegeben.

Bemerkung 4.46: *Sei* $\pi : \{1, \ldots, s\} \to \{1, \ldots, s\}$ *eine Permutation, sei*

$$\tilde{a}_{ij} = a_{\pi(i),\pi(j)} , \quad \tilde{c}_i = c_{\pi(i)} , \quad \tilde{b}_j = b_{\pi(j)} .$$

Dann erzeugen $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ *und* $\begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline & \tilde{b}^T \end{array}$ *dieselbe Abbildung* I_h *(mit vertauschten Zwischenstufen* $\tilde{y}_i = y_{\pi(i)}$*).*

Bemerkung 4.47: (Reduktion der Stufenzahl)

a) *Eine Null<u>spalte</u>* $a_{1j} = a_{2j} = \ldots = b_j = 0$ *kann zusammen mit der entsprechenden Zeile herausgestrichen werden* (y_j *wird definiert, aber nirgends verwendet).*

b) *Eine Null<u>zeile</u>* $c_i = a_{i1} = \ldots = a_{is} = 0$ *ist nicht trivial:* $y_i = y$*.*

c) *Zwei identische Zeilen* $a_{i_1 j} = a_{i_2 j}$, $j = 1, \dots, s$, *können zusammengefaßt werden, da* $y_{i_1} = y_{i_2}$ *folgt. Definiere die neue Spalte* i_1

$$\begin{aligned} a_{ii_1}^{(neu)} &= a_{ii_1}^{(alt)} + a_{ii_2}^{(alt)} \ , \quad i \ = \ 1, \dots, s \ , \\ b_{i_1}^{(neu)} &= b_{i_1}^{(alt)} + b_{i_2}^{(alt)} \end{aligned}$$

als Summe der alten Spalten und streiche die Zeile und Spalte i_2.

Bemerkung 4.48: *Die Lösung des speziellen AWP* $\dfrac{dy}{dt} = f(t)$, $y(t_0) = 0$, *mit* $f : \mathbb{R} \to \mathbb{R}$ *ist* $y(t_0 + h) = \displaystyle\int_{t_0}^{t_0+h} f(t)\, dt$. *Die RK-Abbildung*

$$I_h(y(t_0)) \ = \ h \sum_{j=1}^{s} b_j \, f(t_0 + c_j \, h)$$

wird damit zu einer Quadraturformel mit s *Knoten* c_j *und Gewichten* b_j. *Für* $f(t) = (t - t_0)^{k-1}$ *mit exakter polynomialer Lösung* $y(t_0 + h) = h^k/k$ *folgt*

$$I_h(y(t_0)) \ - \ y(t_0 + h) \ = \ h^k \left(\sum_{j=1}^{s} b_j \, c_j^{k-1} \ - \ \frac{1}{k} \right) \ = \ O(h^{p+1}) \ ,$$

wo p *die Ordnung des RK-Verfahrens ist. Es folgen die* ***Quadraturbedingungen***

$$\sum_{j=1}^{s} b_j \, c_j^{k-1} \ = \ \frac{1}{k} \ , \quad k = 1, \dots, p$$

als notwendige Bedingungen an die Parameter, um Ordnung p *zu erreichen. Die Ordnung des RK-Verfahrens entspricht damit dem polynomialen Exaktheitsgrad* $p - 1$ *als Quadraturformel.*

Folgerung: ***die maximal mögliche Ordnung eines*** s***-stufigen RK-Verfahrens ist*** $2s$.

4.5.2 Ordnungstheorie

Ziel: identifiziere die Butcher-Reihe von I_h.

Hilfssatz 4.49: (Die Butcher-Reihe des impliziten Euler-Verfahrens)

Sei $Y = I_h(y)$ *die Lösung von* $Y = y + h\, f(Y)$ *("implizites Euler-Verfahren"). Dann gilt*

$$\frac{1}{n!} \, \frac{d^n}{dh^n} \, Y_{\big|h=0} \ = \ \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \frac{1}{\sigma(\rho\tau)} \, D_{f,y}(\rho\tau) \ ,$$

d.h., $$I_h(y) \simeq y + \sum_{n=1}^{\infty} h^n \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \frac{1}{\sigma(\rho\tau)} D_{f,y}(\rho\tau) \ .$$

Beweis: [Butcher, Satz 303 C], mehr Kombinatorik.

Definition 4.50:

Das Butcher-Schema $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ *sei gegeben. Zum Baum* $\rho\tau$ *wähle* $\lambda\rho\tau = (V, E, r) \in \rho\tau$ *mit* $V = \{1, \ldots, n\}$. *Definiere die* **RK-Gewichte**

$$\Phi(\rho\tau) = \sum_{j_1=1}^{s} \cdots \sum_{j_n=1}^{s} b_{j_r} \prod_{(\alpha,\beta)\in E} a_{j_\alpha j_\beta}$$

und

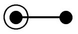
$$\Phi_i(\rho\tau) = \sum_{j_1=1}^{s} \cdots \sum_{j_n=1}^{s} a_{i j_r} \prod_{(\alpha,\beta)\in E} a_{j_\alpha j_\beta}$$

für $i = 1, \ldots, s$. *Diese Definitionen sind unabhängig vom gewählten Repräsentanten* $\lambda\rho\tau \in \rho\tau$.

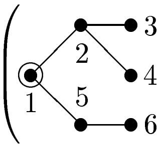
Beispiel 4.51:

$\Phi($ 

$) = \sum_{j=1}^{s} b_j \ ,$

$\Phi($ 

$) = \sum_{j=1}^{s}\sum_{k=1}^{s} b_j\, a_{jk} = \sum_{j=1}^{s} b_j\, c_j \ ,$

$\Phi\Bigg($ 

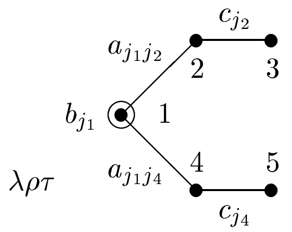
$\Bigg)$

$$= \sum_{j_1,j_2,j_5} b_{j_1}\, a_{j_1 j_2} \underbrace{\Big(\sum_{j_3} a_{j_2 j_3}\Big)}_{c_{j_2}} \underbrace{\Big(\sum_{j_4} a_{j_2 j_4}\Big)}_{c_{j_2}} a_{j_1 j_5} \underbrace{\Big(\sum_{j_6} a_{j_5 j_6}\Big)}_{c_{j_5}}$$

$$= \sum_{j_1,j_2,j_5} b_{j_1}\, a_{j_1 j_2}\, c_{j_2}^2\, a_{j_1 j_5}\, c_{j_5} \ .$$

Bemerkung 4.52: *Für eine Kante* (α, β) *mit einem* *<u>Endknoten</u>* β *kann eine Summation ausgeführt werden und liefert* $\sum_{j_\beta=1}^{s} a_{j_\alpha j_\beta} = c_{j_\alpha}$. *Es folgt die anschauliche Konstruktion des Gewichtes* Φ: *hefte an die Wurzel eine Kopie von* b, *an jede Kante eine Kopie von* A, *die für "Endkanten" zu einer Kopie von* c

vereinfacht werden kann. Dann multipliziere alles und addiere:



$$\longrightarrow \quad \Phi(\rho\tau) = \sum_{j_1=1}^{s} \sum_{j_2=1}^{s} \sum_{j_4=1}^{s} b_{j_1}\, a_{j_1 j_2}\, c_{j_2}\; a_{j_1 j_4}\, c_{j_4}\,.$$

Bemerkung 4.53: *Es gelten die rekursiven Darstellungen*

$$\begin{aligned}
\Phi(\llbracket\, \rho\tau_1 \cdots \rho\tau_k \,\rrbracket) &= \sum_{j=1}^{s} b_j\, \Phi_j(\rho\tau_1)\, \cdots\, \Phi_j(\rho\tau_k)\ , \\
\Phi_i(\llbracket\, \rho\tau_1 \cdots \rho\tau_k \,\rrbracket) &= \sum_{j=1}^{s} a_{ij}\, \Phi_j(\rho\tau_1)\, \cdots\, \Phi_j(\rho\tau_k)\ , \quad i = 1, \ldots, s
\end{aligned}$$

mit $\Phi(\odot) = \sum_{j=1}^{s} b_j$ *und* $\Phi_i(\odot) = c_i\,.$

Satz 4.54: (Die Butcher-Reihe eines RK-Verfahrens)
Für die RK-Abbildung 4.42 gilt

$$y_i \;\simeq\; y \;+\; \sum_{n=1}^{\infty} h^n \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \frac{\Phi_i(\rho\tau)}{\sigma(\rho\tau)}\; D_{f,y}(\rho\tau)$$

mit $i = 1, \ldots, s$ *und*

$$I_h(y) \;\simeq\; y \;+\; \sum_{n=1}^{\infty} h^n \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \frac{\Phi(\rho\tau)}{\sigma(\rho\tau)}\; D_{f,y}(\rho\tau)\ .$$

Beweis: Fasse die Zwischenstufen $y_1, \ldots, y_s$ und $y_{s+1} := I_h(y) \in \mathbb{R}^N$ zum Vektor $\hat{Y} = (y_1, \ldots, y_{s+1}) \in \mathbb{R}^{N\times(s+1)}$ zusammen, setze $a_{s+1,j} := b_j$, $j = 1, \ldots, s$, und $\Phi_{s+1}(\rho\tau) := \Phi(\rho\tau)$. Mit $\hat{f} : \mathbb{R}^{N\times(s+1)} \to \mathbb{R}^{N\times(s+1)}$:

$$\hat{f}(\hat{Y}) \;:=\; \left(\sum_{j=1}^{s} a_{1j}\, f(y_j)\ ,\ \ldots\ ,\ \sum_{j=1}^{s} a_{s+1,j}\, f(y_j) \right)$$

ist die RK-Abbildung 4.42 definiert durch das implizite Euler-Verfahren

$$\hat{Y} = \hat{y} + h\, \hat{f}(\hat{Y})$$

auf $\mathbb{R}^{N\times(s+1)}$ mit $\hat{y} = (y,\dots,y)$. Hilfssatz 4.49 liefert die Butcher-Reihe

$$\hat{Y} \;\simeq\; \hat{y} \;+\; \sum_{n=1}^{\infty} h^n \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \frac{1}{\sigma(\rho\tau)}\, \mathrm{D}_{\hat{f},\hat{y}}(\rho\tau)\ .$$

Die Behauptung folgt dann mit

$$\mathrm{D}_{\hat{f},\hat{y}}(\rho\tau) \;=\; \Big(\Phi_1(\rho\tau)\ \mathrm{D}_{f,y}(\rho\tau)\ ,\ \dots\ ,\ \Phi_{s+1}(\rho\tau)\ \mathrm{D}_{f,y}(\rho\tau)\Big)\ \in \mathbb{R}^{N\times(s+1)}\ .$$

Dies ergibt sich per Induktion nach $|\rho\tau|$. Start:

$$\begin{aligned}
\mathrm{D}_{\hat{f},\hat{y}}(\odot) \;=\; \hat{f}(\hat{y}) \;&=\; \Big(\ \sum_{j=1}^{s} a_{1j}\, f(y)\ \ ,\ \dots\ ,\ \sum_{j=1}^{s} a_{s+1,j}\, f(y)\ \Big) \\
&=\; \Big(\Phi_1(\odot)\ \mathrm{D}_{f,y}(\odot)\ ,\ \dots\ ,\ \Phi_{s+1}(\odot)\ \mathrm{D}_{f,y}(\odot)\Big).
\end{aligned}$$

Induktionsschritt: sei $\rho\tau = [\![\ \rho\tau_1 \cdots \rho\tau_k\ ]\!]$. Mit

$$\Big(v_1^{(\alpha)}\ ,\ \dots\ ,\ v_{s+1}^{(\alpha)}\Big) \;:=\; \mathrm{D}_{\hat{f},\hat{y}}(\rho\tau_\alpha)$$

$$=\; \Big(\Phi_1(\rho\tau_\alpha)\ \mathrm{D}_{f,y}(\rho\tau_\alpha)\ ,\ \dots\ ,\ \Phi_{s+1}(\rho\tau_\alpha)\ \mathrm{D}_{f,y}(\rho\tau_\alpha)\Big)\ ,\quad \alpha = 1,\dots,k$$

gilt mit den rekursiven Darstellungen 4.25 und 4.53:

$$\mathrm{D}_{\hat{f},\hat{y}}(\rho\tau) \;=\; \hat{f}^{(n)}(\hat{y})\Big[\mathrm{D}_{\hat{f},\hat{y}}(\rho\tau_1),\dots,\mathrm{D}_{\hat{f},\hat{y}}(\rho\tau_k)\Big]$$

$$\begin{aligned}
&=\; \hat{f}^{(n)}(\hat{y})\Big[\Big(v_1^{(1)},\dots,v_{s+1}^{(1)}\Big)\ ,\ \dots\ ,\ \Big(v_1^{(k)},\dots,v_{s+1}^{(k)}\Big)\Big] \\
&=\; \Big(\sum_{j=1}^{s} a_{1j}\, f^{(n)}(y)[v_j^{(1)},\dots,v_j^{(k)}]\ ,\ \dots\ ,\ \sum_{j=1}^{s} a_{s+1,j}\, f^{(n)}(y)[v_j^{(1)},\dots,v_j^{(k)}]\Big) \\
&=\; \Big(\sum_{j=1}^{s} a_{1j}\Phi_j(\rho\tau_1)\cdots\Phi_j(\rho\tau_k)\ f^{(n)}(y)[\mathrm{D}_{f,y}(\rho\tau_1),\dots,\mathrm{D}_{f,y}(\rho\tau_k)]\ ,\ \dots\ , \\
&\qquad\qquad \sum_{j=1}^{s} a_{s+1,j}\Phi_j(\rho\tau_1)\cdots\Phi_j(\rho\tau_k)\ f^{(n)}(y)[\mathrm{D}_{f,y}(\rho\tau_1),\dots,\mathrm{D}_{f,y}(\rho\tau_k)]\Big) \\
&=\; \Big(\sum_{j=1}^{s} \Phi_1(\rho\tau)\,\mathrm{D}_{f,y}(\rho\tau)\ ,\ \dots\ ,\ \sum_{j=1}^{s} \Phi_{s+1}(\rho\tau)\,\mathrm{D}_{f,y}(\rho\tau)\Big).
\end{aligned}$$

Q.E.D.

Korollar 4.55: (Die Ordnungsgleichungen)

Mit 4.24 ergibt der Vergleich der Butcher-Reihen 4.28 und 4.54

$$F_h(y) - I_h(y) \;\simeq\; \sum_{n=1}^{\infty} h^n \sum_{\substack{\rho\tau \\ |\rho\tau|=n}} \frac{1}{\sigma(\rho\tau)} \left(\frac{1}{\gamma(\rho\tau)} \;-\; \Phi(\rho\tau) \right) D_{f,y}(\rho\tau) \;.$$

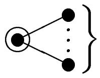
Damit ist ein RK-Verfahren genau dann von der Ordnung p, wenn für alle Bäume $\rho\tau$ mit $|\rho\tau| \le p$ die **Ordnungsgleichungen**

$$\Phi(\rho\tau) \;=\; \frac{1}{\gamma(\rho\tau)}$$

erfüllt sind. Der führende Fehlerterm hat die Darstellung

$$h^{p+1} \sum_{\substack{\rho\tau \\ |\rho\tau|=p+1}} \frac{1}{\sigma(\rho\tau)} \left(\frac{1}{\gamma(\rho\tau)} - \Phi(\rho\tau) \right) D_{f,y}(\rho\tau) \;.$$

Bemerkung 4.56: *Dies ist ein System polynomialer Gleichungen für die Butcher-Parameter (Tafel 4.1). Die* **Konsistenzbedingung** $\sum_{j=1}^{s} b_j \;=\; 1$ *garantiert Konsistenz/Konvergenz. Die "Büschel"*



$k - 1$ *Blätter*

liefern die **Quadraturbedingungen**

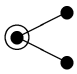
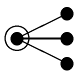
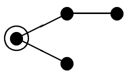
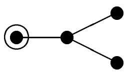
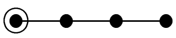
$$\sum_{j=1}^{s} b_j\, c_j^{k-1} \;=\; \frac{1}{k} \;, \qquad k = 1, \ldots, p$$

aus Bemerkung 4.48.

Bemerkung 4.57: *Sei a_p = Anzahl aller Bäume $\rho\tau$ mit genau p Knoten*[1] *Die Anzahl der Gleichungen $\sum_{k=1}^{p} a_k$ steigt schnell mit der gewünschten Ordnung p:*

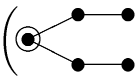
p	1	2	3	4	5	6	7	8	9	10	...	20
a_p	1	1	2	4	9	20	48	115	286	719	...	12 826 228
$\sum_{k=1}^{p} a_k$	1	2	4	8	17	37	85	200	486	1205	...	20 247 374

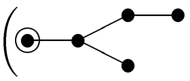
[1] Es gilt die formale Potenzreihenidentiät (siehe z.B. [Butcher]) $\sum_{p=0}^{\infty} a_{p+1}\, x^p \;=\; \prod_{p=1}^{\infty} \frac{1}{(1-x^p)^{a_p}}$, aus der diese Zahlen durch Koeffizientenvergleich rekursiv bestimmt werden können.

Ordnung	$\rho\tau$	$\Phi(\rho\tau)$	$=$	$\dfrac{1}{\gamma(\rho\tau)}$
1		$\sum_{j=1}^{s} b_j$	$=$	1
2		$\sum_{j=1}^{s} b_j\, c_j$	$=$	$\frac{1}{2}$
3		$\sum_{j=1}^{s} b_j\, c_j^2$	$=$	$\frac{1}{3}$
3		$\sum_{j=1}^{s}\sum_{k=1}^{s} b_j\, a_{jk}\, c_k$	$=$	$\frac{1}{6}$
4		$\sum_{j=1}^{s} b_j\, c_j^3$	$=$	$\frac{1}{4}$
4		$\sum_{j=1}^{s}\sum_{k=1}^{s} b_j\, c_j\, a_{jk}\, c_k$	$=$	$\frac{1}{8}$
4		$\sum_{j=1}^{s}\sum_{k=1}^{s} b_j\, a_{jk}\, c_k^2$	$=$	$\frac{1}{12}$
4		$\sum_{j=1}^{s}\sum_{k=1}^{s}\sum_{l=1}^{s} b_j\, a_{jk}\, a_{kl}\, c_l$	$=$	$\frac{1}{24}$

Tafel 4.1: Die ersten Ordnungsgleichungen.

Bemerkung 4.58: *In der Anwendung von RK-Verfahren auf* <u>*skalare*</u> *Gleichungen $dy/dt = f(y)$, $y \in \mathbb{R}$, fallen einige elementare Differentiale zusammen, so daß im Vergleich der Butcher-Reihen von F_h und I_h nicht für jeden Baum getrennt $\Phi(\rho\tau) = 1/\gamma(\rho\tau)$ gefordert zu werden braucht, z.B.*

$D_{f,y}$ () $= f''(y)\, f'(y)\, f(y)\, f'(y)\, f(y)$

$=$ $D_{f,y}$ () $= f'(y)\, f''(y)\, f'(y)\, f(y)\, f(y)\ .$

4.5.3 Explizite RK-Verfahren

Definition 4.59:

Ein RK-Verfahren $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ *mit streng unterer Dreiecksmatrix* A *heißt* **explizit**. *Dann ist ein Zeitschritt* $I_h(y) = y + \sum_{j=1}^{s} b_j\, k_j$ *nach Bemerkung 4.43 in der Form*

$$k_1 := h\, f(y)\ , \quad k_i := h\, f\Big(y + \sum_{j=1}^{i-1} a_{ij}\, k_j\Big)\ , \quad i = 2, \ldots, s$$

mit s *Auswertungen von* f *ausführbar.*

Bemerkung 4.60: *Mit Bemerkung 4.46 reicht es, wenn* A *durch Permutation auf strenge Dreiecksform gebracht werden kann.*

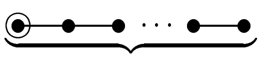
Bemerkung 4.61:

a) *Es existieren explizite RK-Verfahren beliebig hoher Ordnung (bei hinreichend hoher Stufenzahl), siehe Bemerkung 4.85.*

b) *Ein explizites* s*-stufiges Verfahren hat höchstens die Ordnung* s*. Es gilt ([Butcher]):*

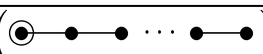
Stufenzahl	1	2	3	4	5	6	7	8	9
erreichbare Ordnung	1	2	3	4	4	5	6	6	7

Beweis von Ordnung $\leq s$: Mit $e = (1, \ldots, 1)^T$ gilt

$\Phi\Big($  $\Big) = \langle b, A^s e \rangle = 0$,

$s + 1$ Knoten

da $A^s = 0$ (Explizitheit $\Rightarrow A$ ist nilpotent). Aber

$1 / \gamma\big($  $\big) = \dfrac{1}{(s+1)\,!}$.

Q.E.D.

Einige explizite RK-Verfahren 4.62: *Die Ordnungsgleichungen werden durch die folgenden Butcher-Parameter erfüllt (Einsetzen und Verifizieren). Mit Stufenzahl* s *und Ordnung* p*:*

$s=1$, $p=1$

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

“Euler-” oder **“Polygonzug-Verfahren”**

$$I_h(y) = y + h\,f(y)$$

$s=2$, $p=2$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

“Runge 2.ter Ordnung”

$$I_h(y) = y + h\,f(\,y + \tfrac{h}{2}\,f(y)\,)$$

$s=2$, $p=2$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

“Heun 2.ter Ordnung”

$$I_h(y) = y + \tfrac{h}{2}\,f(y) + \tfrac{h}{2}\,f(\,y + h\,f(y)\,)$$

$s=3$, $p=3$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

“Heun 3. Ordnung”

$s=3$, $p=3$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

“Kutta 3.ter Ordnung”

$s=3$, $p=3$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}$$

“RK 3.ter Ordnung”

$s=4$, $p=4$

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{2}{3} & -\frac{1}{3} & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}$$

“3/8-Verfahren”

$s=4$, $p=4$

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Das **“klassische” RK-Verfahren 4. Ordnung**. Dies ist <u>das</u> RK-Verfahren <u>schlechthin</u>!

Das "klassische" Verfahren 4. Ordnung bietet einen akzeptablen Kompromiß zwischen Aufwand (Stufenzahl) und Genauigkeit (Ordnung).

Bemerkung 4.63: *Nach Bemerkung 4.48 werden die Verfahren in Anwendung auf $dy/dt = f(t)$ zu Quadraturformeln*

$$\int_{t_0}^{t_0+h} f(t)\,dt \;=\; h \sum_{j=1}^{s} b_j\, f(t_0 + c_j\, h) \;+\; O(h^{p+1}) \;.$$

Hierbei gilt:

Euler-Verfahren	$\longrightarrow$	*einfache Riemann-Summe*
Runge 2.ter Ordnung	$\longrightarrow$	*einfache Riemann-Summe*
Heun 2.ter Ordnung	$\longrightarrow$	*Trapez-Regel*
Heun 3.ter Ordnung	$\longrightarrow$	
Kutta 3.ter Ordnung	$\longrightarrow$	*Simpson-Regel*
RK 3.ter Ordnung	$\longrightarrow$	*Simpson-Regel*
3/8-Verfahren	$\longrightarrow$	*3/8-Regel*
klassisches RK-Verfahren	$\longrightarrow$	*Simpson-Regel.*

4.5.4 Eingebettete Verfahren, Schrittweitensteuerung

Versuche, mittels zweier unterschiedlicher Verfahren eine Abschätzung des lokalen Verfahrensfehlers zu berechnen und zur Schrittweitensteuerung einzusetzen. Idee (Fehlberg): suche explizite Verfahren I_h bzw. $\hat{I}_h$ der Ordnung p bzw. $\hat{p} = p + 1$, die gemeinsame Zwischenstufen haben, so daß der Aufwand zur simultanen Ausführung beider Verfahren nicht wesentlich größer ist als der Aufwand jedes einzelnen Verfahrens. Damit sollte die Butcher-Matrix von I_h eine Teilmatrix (**"Einbettung"**) der Butcher-Matrix von $\hat{I}_h$ sein.

Nimmt man an, daß

$$\|F_h(y) - \hat{I}_h(y)\| \;=\; O(h^{p+2}) \;\ll\; \|F_h(y) - I_h(y)\| \;=\; \|e(y)\, h^{p+1}\| + O(h^{p+2})$$

gilt, so liefert

$$E(h, y) \;=\; \hat{I}_h(y) - I_h(y) \quad (\; = e(y)\, h^{p+1} + O(h^{p+2})\;)$$

eine Fehlerschätzung für das Verfahren niederer Ordung:

$$\|F_h(y) - \hat{I}_h(y)\| \;\ll\; \|F_h(y) - I_h(y)\| \;\approx\; \|E(h, y)\| \;.$$

Wie in der Schrittweitensteuerung 4.41 geht man zu

$$h \quad\to\quad h \left(\frac{\varepsilon}{\|E(h, y)\|}\right)^{\frac{1}{p+1}}$$

über, um $\|F_h(y) - \hat{I}_h(y)\| \ll \|F_h(y) - I_h(y)\| \approx \epsilon$ zu erreichen.

Variante 1: Akzeptiere $I_h(y)$ als Approximation von $F_h(y)$. Vorteil: die gefundene Schrittweite ist optimal groß. Ärgerlich: die bessere Approximation $\hat{I}_h(y)$ wird nur zur Schrittweitensteuerung verwendet.

Variante 2: Akzeptiere $\hat{I}_h(y)$ als Approximation von $F_h(y)$. Ärgerlich: die benutzten Schrittweiten sind tendenziell zu klein, es gilt $\|F_h(y) - \hat{I}_h(y)\| \ll \epsilon$.

Beispiel 4.64: *Das Verfahren "Runge 2.ter Ordnung"* $\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$ *ist eingebettet in "Kutta 3.ter Ordnung":*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Mit

$$k_1 = h\, f(y)\ , \qquad I_h(y) = y + k_2\ ,$$
$$k_2 = h\, f\Big(y + \frac{1}{2}\, k_1\Big)\ , \qquad \hat{I}_h(y) = y + \frac{1}{6}\ (k_1 + 4\, k_2 + k_3)$$
$$k_3 = h\, f(y - k_1 + 2\, k_2)\ ,$$

ergibt

$$E(h,y) = \frac{1}{6}\ (k_1 - 2\, k_2 + k_3) \qquad (\ = \hat{I}_h(y) - I_h(y)\ )$$

eine heuristische Abschätzung des lokalen Fehlers:

$$\|F_h(y) - \hat{I}_h(y)\| \ \ll\ \|F_h(y) - I_h(y)\| \ \approx\ \|E(h,y)\|\ .$$

Beispiel 4.65: *Das Fehlberg4(5)-Verfahren:*

$$\begin{array}{c|cccccc} 0 & 0 & & & & & \\ \frac{2}{9} & \frac{2}{9} & 0 & & & & \\ \frac{1}{3} & \frac{1}{12} & \frac{1}{4} & 0 & & & \\ \frac{3}{4} & \frac{69}{128} & -\frac{243}{128} & \frac{135}{64} & 0 & & \\ 1 & -\frac{17}{12} & \frac{27}{4} & -\frac{27}{5} & \frac{16}{15} & 0 & \\ \frac{5}{6} & \frac{65}{432} & -\frac{5}{16} & \frac{13}{16} & \frac{4}{27} & \frac{5}{144} & 0 \\ \hline & \frac{1}{9} & 0 & \frac{9}{20} & \frac{16}{45} & \frac{1}{12} & 0 \\ & \frac{47}{450} & 0 & \frac{12}{25} & \frac{32}{225} & \frac{1}{30} & \frac{6}{25} \\ & -\frac{1}{150} & 0 & \frac{3}{100} & -\frac{16}{75} & -\frac{1}{20} & \frac{6}{25} \end{array}
\quad
\begin{array}{l} \leftarrow\ k_1 = h\, f(y) \\ \leftarrow\ k_2 = h\, f(y + \frac{2}{9}\, k_1) \\ \qquad \vdots \\ \qquad \vdots \\ \qquad \vdots \\ \leftarrow\ k_6 = h\, f(y + \frac{65}{432}\, k_1 + \cdots) \\ \leftarrow\ b^T \quad (\textit{Ordnung 4, 5 Stufen}) \\ \leftarrow\ \hat{b}^T \quad (\textit{Ordnung 5, 6 Stufen}) \\ \leftarrow\ \hat{b}^T - b^T \quad (\textit{Fehlerschätzer}) \end{array}$$

Also:

$$E(h,y) = -\frac{1}{150}\,k_1 + \frac{3}{100}\,k_3 - \frac{16}{75}\,k_4 - \frac{1}{20}\,k_5 + \frac{6}{25}\,k_6$$

ist eine Abschätzung des führenden Fehlerterms des Verfahrens 4.ter Ordnung

$$I_h(y) = y + \frac{1}{9}\,k_1 + \frac{9}{20}\,k_3 + \frac{16}{45}\,k_4 + \frac{1}{12}\,k_5\ .$$

Bemerkung 4.66: (FSAL-Prinzip, First Same As Last) *Es gibt einige Verfahren, für die sich Variante 1 anbietet. Wenn nämlich der Vektor* $(b_1,\ldots,b_s)$ *des niederen (s-stufigen) Verfahrens* I_h *mit der letzten Stufenzeile* $a_{s+1,1},\ldots,a_{s+1,s}$ *des höheren* ($s+1$*-stufigen) Verfahrens* $\hat{I}_h$ *übereinstimmt, so gilt* $y_{s+1} = I_h(y)$. *Akzeptiert man* $\tilde{y} = I_h(y)$ *als Approximation von* $F_h(y)$, *so liegt die erste Stufe* $\tilde{k}_1 = h\,f(\tilde{y})$ *des nächsten Schrittes schon als letzte Stufe* $k_{s+1} = h\,f(y_{s+1})$ *des vorherigen Schrittes vor. Die Fehlerabschätzung ist damit umsonst.*

Beispiel 4.67: *Das Fehlberg 3(4)-Verfahren mit FSAL:*

$$\begin{array}{c|ccccc}
0 & 0 & & & & \\
\frac{1}{4} & \frac{1}{4} & 0 & & & \\
\frac{4}{9} & \frac{4}{81} & \frac{32}{81} & 0 & & \\
\frac{6}{7} & \frac{57}{98} & -\frac{432}{343} & \frac{1053}{686} & 0 & \\
1 & \frac{1}{6} & 0 & \frac{27}{52} & \frac{49}{156} & 0 \\
\hline
 & \frac{1}{6} & 0 & \frac{27}{52} & \frac{49}{156} & 0 \\
 & \frac{43}{288} & 0 & \frac{243}{416} & \frac{343}{1872} & \frac{1}{12} \\
 & -\frac{5}{288} & 0 & \frac{27}{416} & -\frac{245}{1872} & \frac{1}{12}
\end{array}
\quad
\begin{array}{l}
\\ \\ \\ \\ \\
\longleftarrow\ b^T \quad (\textit{Ordnung 3, 4 Stufen}) \\
\longleftarrow\ \hat{b}^T \quad (\textit{Ordnung 4, 5 Stufen}) \\
\longleftarrow\ \hat{b}^T - b^T \quad (\textit{Fehlerschätzer})
\end{array}$$

Hierbei ist

$$y_5 \ = \ y + \frac{1}{6}\,k_1 + \frac{27}{52}\,k_3 + \frac{49}{156}\,k_4 \ = \ I_h(y)\ ,$$

d.h., $k_5 = h\,f(y_5) = h\,f(I_h(y))$ *ist die erste Stufe des nächsten Zeitschritts (bei Schrittweitenwechsel* $h \to \tilde{h}$ *ist sie* $\tilde{h}\,f(y_5)$*).*

Bemerkung 4.68: *Die eingebetteten RK-Fehlberg-Verfahren (weitere siehe z.B. [Butcher] oder [Hairer, Nørsett & Wanner]) sind in der Praxis ausgezeichnete allround-Methoden für "nichtsteife" Systeme (steife Systeme: siehe Sektion 4.7).*

4.5.5 Implizite RK-Verfahren

Problem: Bei nicht expliziten RK-Verfahren auf dem $\mathbb{R}^N$ ist in jedem Zeitschritt $I_h(y)$ zunächst ein nichtlineares Gleichungssystem

$$y_i \ = \ y \ + \ h\sum_{j=1}^{s} a_{ij}\,f(y_j)\ , \quad i = 1,\ldots,s$$

für die Zwischenstufen numerisch zu lösen. Praktische Durchführung:

Newton-Verfahren $(y_1^{(k)}, \ldots, y_s^{(k)}) \to (y_1^{(k+1)}, \ldots, y_s^{(k+1)})$ auf $\mathbb{R}^{N\times s}$. Startwerte sind z.B. durch $y_i^{(0)} = y$ gegeben oder (genauer) über beliebige explizite Verfahren konstruierbar, wobei $y_i \approx F_{c_i h}(y)$ gilt: vergleiche die Bemerkungen 4.76 und 4.86. Problem: zur Ausführung eines Newton-Schrittes wird f' benötigt (Aufwand!).

Alternativ:

Fixpunkt-Iteration 4.69:

$$y_i^{(k+1)} = y + h \sum_{j=1}^{s} a_{ij}\, f(y_j^{(k)}) \;, \qquad \begin{array}{l} i = 1, \ldots, s \;, \\ k = 0, 1, 2, \ldots \end{array}$$

mit Start $y_1^{(0)} = \ldots = y_s^{(0)} = y$.

Satz 4.70: (Approximation impliziter Verfahren durch explizite)

Es sei $\|A\|_\infty$ *die Zeilensummennorm der Butcher-Matrix* $A = (a_{ij})$. *Bei Lipschitz-stetigem Vektorfeld* $\|f(y) - f(z)\| \le L\,\|y - z\|$ *konvergiert die Fixpunktiteration 4.69 für* $|h|\,\|A\|_\infty L < 1$ *gegen die eindeutige Lösung* y_i^*. *Nach* $q-1$ *Schritten gilt*

$$y_i^{(q-1)} \;=\; y_i^* + O(h^q) \;,$$

so daß

$$I_h^{(q-1)}(y) := y + h \sum_{j=1}^{s} b_j\, f(y_j^{(q-1)}) \;=\; \underbrace{y + h \sum_{j=1}^{s} b_j\, f(y_j^*)}_{I_h^{(exakt)}(y)} \;+\; O(h^{q+1}) \;.$$

Beweis: Sei analog zum Beweis von Satz 4.54

$$\hat{Y} = (y_1, \ldots, y_s) \in \mathbb{R}^{N\times s} \;, \quad \hat{y} = (y, \ldots, y) \in \mathbb{R}^{N\times s}$$

und

$$\hat{f}(\hat{Y}) \;:=\; \left(\sum_{j=1}^{s} a_{1j}\, f(y_j) \;,\; \ldots \;,\; \sum_{j=1}^{s} a_{sj}\, f(y_j) \right) \;,$$

womit die exakten Zwischenstufen $\hat{Y}^* = (y_1^*, \ldots, y_s^*)$ als Lösung des Fixpunktproblems

$$\hat{Y} = \hat{y} + h\, \hat{f}(\hat{Y}) =: \Psi_h(\hat{Y})$$

definiert sind. Eine Kontraktionskonstante von $\Psi_h : \mathbb{R}^{N\times s} \to \mathbb{R}^{N\times s}$ bezüglich der Norm $\|\hat{Y}\|_\infty = \|(y_1,\ldots,y_s)\|_\infty := \max\limits_{i=1..s} \|y_i\|$ mit beliebiger Norm $\|\,.\,\|$ auf $\mathbb{R}^N$ ist durch $|h|\,\|A\|_\infty L$ gegeben:

$$\begin{aligned}
\|\Psi_h(\hat{Y}) - \Psi_h(\hat{Z})\|_\infty &= |h| \max_{i=1..s} \Big\| \sum_{j=1}^{s} a_{ij} \Big(f(y_j) - f(z_j) \Big) \Big\| \\
&\le |h| \max_{i=1..s} \sum_{j=1}^{s} |a_{ij}|\, \|f(y_j) - f(z_j)\| \\
&\le |h| \Big(\max_{i=1..s} \sum_{j=1}^{s} |a_{ij}| \Big) \Big(\max_{j=1..s} \|f(y_j) - f(z_j)\| \Big) \\
&\le |h|\, \|A\|_\infty\, L \max_{j=1..s} \|y_j - z_j\| \;=\; |h|\, \|A\|_\infty\, L\, \|\hat{Y} - \hat{Z}\|_\infty \,,
\end{aligned}$$

wobei $\hat{Z} = (z_1,\ldots,z_s)$. Konvergenz und Eindeutigkeit der Lösung folgt damit aus dem Banachschen Fixpunktsatz. In der Iteration $\hat{Y}^{(q)} = \Psi_h(\hat{Y}^{(q-1)})$ mit dem Start $\hat{Y}^{(0)} = \hat{y}$ gilt

$$\|\hat{Y}^{(q)} - \hat{Y}^*\|_\infty \;\le\; \Big(h\, \|A\|_\infty\, L \Big)^q \|\hat{Y}^{(0)} - \hat{Y}^*\|_\infty \;=\; O(h^{q+1}) \,,$$

da

$$\begin{aligned}
\|\hat{Y}^{(0)} - \hat{Y}^*\|_\infty &= \|(y,\ldots,y) - (y_1^*,\ldots,y_s^*)\|_\infty \\
&= |h| \max_{i=1..s} \Big\| \sum_{j=1}^{s} a_{ij}\, f(y_j^*) \Big\| \;=\; O(h) \,.
\end{aligned}$$

Q.E.D.

Bemerkung 4.71: *Mit $p-1$ Schritten der Fixpunktiteration kann ein implizites RK-Verfahren p.ter Ordnung approximativ durchgeführt werden, ohne daß die Ordnung verlorengeht. Die resultierenden* <u>*expliziten*</u> *Verfahren (Tafel 4.2) heißen* **Prädiktor-Korrektor-Verfahren**. *Ihre effektive Stufenzahl (die Anzahl der benötigten f-Auswertungen) ist gemäß Bemerkung 4.47 nach Herausstreichen redundanter Stufen $1+(p-1)s$ (bzw. $1+(p-1)(s-1)$, falls (a_{ij}) eine Nullzeile enthält).*

Beispiel 4.72: *Das* **Trapezverfahren**

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{l} y_1 = y \\ y_2 = y + \frac{h}{2} f(y) + \frac{h}{2} f(y_2) \\ I_h(y) = y_2 \end{array}$$

der Ordnung 2 ist durch

$$Y = I_h(y) = y + \frac{h}{2} f(y) + \frac{h}{2} f(Y)$$

$$
\begin{array}{c|c||c|c|c|c|c}
\begin{matrix} y_1^{(0)} \\ \vdots \\ y_s^{(0)} \end{matrix} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & \begin{matrix} 0 & & \\ \vdots & \ddots & \\ 0 & \dots & 0 \end{matrix} & & & & \\
\hline
\begin{matrix} y_1^{(1)} \\ \vdots \\ y_s^{(1)} \end{matrix} & \begin{matrix} c_1 \\ \vdots \\ c_s \end{matrix} & \begin{matrix} a_{11} & \dots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \dots & a_{ss} \end{matrix} & \begin{matrix} 0 & & \\ \vdots & \ddots & \\ 0 & \dots & 0 \end{matrix} & & & \\
\hline
\begin{matrix} y_1^{(2)} \\ \vdots \\ y_s^{(2)} \end{matrix} & \begin{matrix} c_1 \\ \vdots \\ c_s \end{matrix} & \begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix} & \begin{matrix} a_{11} & \dots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \dots & a_{ss} \end{matrix} & \begin{matrix} 0 & & \\ \vdots & \ddots & \\ 0 & \dots & 0 \end{matrix} & & \\
\hline
\vdots & & \vdots & \ddots & \ddots & \ddots & \\
\hline
\begin{matrix} y_1^{(p-1)} \\ \vdots \\ y_s^{(p-1)} \end{matrix} & \begin{matrix} c_1 \\ \vdots \\ c_s \end{matrix} & \begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \vdots \end{matrix} & \dots & \begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix} & \begin{matrix} a_{11} & \dots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \dots & a_{ss} \end{matrix} & \begin{matrix} 0 & & \\ \vdots & \ddots & \\ 0 & \dots & 0 \end{matrix} \\
\hline\hline
 & & 0 \ \dots \ 0 & \dots & 0 \ \dots \ 0 & 0 \ \dots \ 0 & b_1 \ \dots \ b_s
\end{array}
$$

Tafel 4.2: Explizites Butcher-Schema zur Approximation eines impliziten Schemas (a_{ij}) der Ordnung p.

definiert. Ein Schritt der Fixpunktiteration mit $Y^{(0)} = y$ reicht, um die Ordnung zu erhalten:

$$Y^{(1)} = y + \frac{h}{2} f(y) + \frac{h}{2} f(Y^{(0)}) = y + h\, f(y)\ .$$

Das resultierende approximative Trapezverfahren

$$
\begin{aligned}
Y^{(1)} &= y + h\, f(y)\ , && \textit{(Prädiktor)} \\
I_h^{(1)}(y) &= y + \frac{h}{2} f(y) + \frac{h}{2} f(Y^{(1)}) && \textit{(Korrektor)}
\end{aligned}
$$

ist identisch mit Heun 2.ter Ordnung:
$$
\begin{array}{c|cc}
0 & 0 & 0 \\
1 & 1 & 0 \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
$$

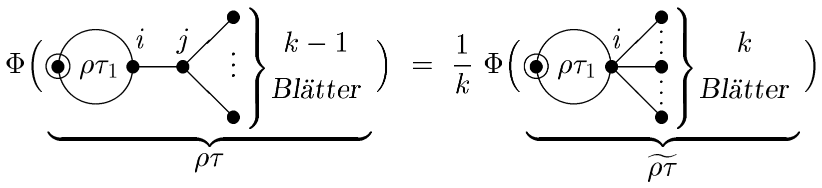
Bemerkung 4.73: *Für steife Systeme (siehe Sektion 4.7) mit großen Lipschitz-Konstanten L müssen unrealistisch kleine Schrittweiten gewählt werden, um die Konvergenz der Fixpunktiteration zu garantieren. Man sollte dann das Newton-Verfahren benutzen.*

4.5.6 Die Gauß-Legendre-Verfahren

Hilfssatz 4.74: (Abhängigkeit von Ordnungsgleichungen)

Für ein $k \in \mathbb{N}$ gelte $$\sum_{j=1}^{s} a_{ij}\, c_j^{k-1} = \frac{c_i^k}{k}\ , \quad i = 1, \ldots, s\ .$$

Dann folgt



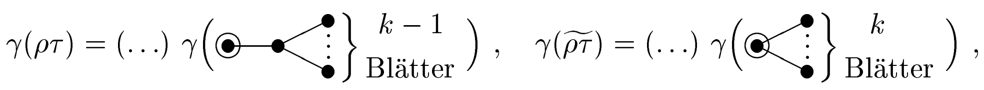
für alle Bäume $\rho\tau$, $\widetilde{\rho\tau}$ der angegebenen Form mit beliebigem Teilbaum $\rho\tau_1$. Mit $\gamma(\rho\tau) = k\,\gamma(\widetilde{\rho\tau})$ sind die Ordnungsgleichungen für $\rho\tau$ und $\widetilde{\rho\tau}$ äquivalent:

$$\Phi(\rho\tau) \;-\; \frac{1}{\gamma(\rho\tau)} \;=\; \frac{1}{k}\,\Big(\;\Phi(\widetilde{\rho\tau}) \;-\; \frac{1}{\gamma(\widetilde{\rho\tau})}\;\Big)\ .$$

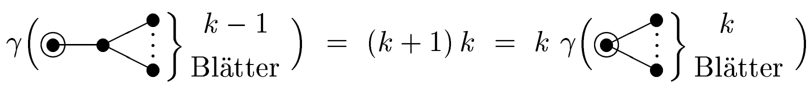
Beweis: Sei (i,j) die Kante, die $\lambda\rho\tau_1 \in \rho\tau_1$ mit dem restlichen "Büschel" in $\rho\tau$ verbindet. Aus der Definition 4.50 folgt

$$\Phi(\rho\tau) = \sum_{i=1}^{s} (\ldots)_i \sum_{j=1}^{s} a_{ij}\, c_j^{k-1}\ , \quad \Phi(\widetilde{\rho\tau}) = \sum_{i=1}^{s} (\ldots)_i\, c_i^k\ ,$$

wobei $(\ldots)_i$ die beiden Bäumen gemeinsamen Beiträge der Kanten in $\rho\tau_1$ darstellt. Unmittelbar ergibt sich $\Phi(\rho\tau) = \Phi(\widetilde{\rho\tau})/k$. Mit der rekursiven Definition 4.22 der Dichte gilt



wobei $(\ldots)$ die gemeinsamen Beiträge aus dem Teilbaum $\rho\tau_1$ darstellt. Mit



folgt $\gamma(\rho\tau) = k\,\gamma(\widetilde{\rho\tau})$.

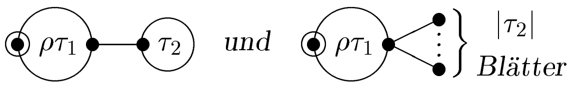
Q.E.D.

Satz 4.75: ("Büschelreduktion")

*Unter der Voraussetzung (**"simplifying assumptions"**)*

$$Simp^{(q)} \; : \qquad \sum_{j=1}^{s} a_{ij}\, c_j^{k-1} = \frac{c_i^k}{k} \;, \quad i = 1, \ldots, s \;, \quad k = 1, \ldots, q$$

sind die Ordnungsgleichungen der Bäume



mit beliebigen Teilbäumen $\rho\tau_1$ und τ_2 äquivalent, wenn $|\tau_2| \leq q$ gilt.

Beweis: Wende rekursiv Hilfssatz 4.74 auf die Blätter in τ_2 an, die Schritt für Schritt an den Baum $\rho\tau_1$ "herangeschoben" werden können, bis τ_2 zum "Büschel" wird.

Q.E.D.

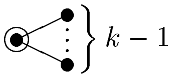
Bemerkung 4.76: *Zur Interpretation der simplifying assumptions: die Zwischenstufe y_i eines RK-Schemas ist eine Approximation des exakten Zeitschrittes $F_{c_i h}$. Vergleich der Butcher-Reihen 4.28 und 4.54 liefert nämlich*

$$F_{c_i h}(y) - y_i \;\simeq\; \sum_{n=1}^{\infty} h^n \sum_{\substack{\rho\tau \\ |\rho\tau| = n}} \frac{1}{\sigma(\rho\tau)} \left(\frac{c_i^n}{\gamma(\rho\tau)} \;-\; \Phi_i(\rho\tau) \right) D_{f,y}(\rho\tau) \;.$$

Damit approximiert y_i den exakten Fluß $F_{c_i h}$ bis auf $O(h^{p_i+1})$, wenn für alle Bäume $\rho\tau$ mit $|\rho\tau| \leq p_i$ die **Stufenordnungsgleichungen**

$$\Phi_i(\rho\tau) \;=\; \frac{c_i^{|\rho\tau|}}{\gamma(\rho\tau)}$$

erfüllt sind (p_i heißt dann i.te **Stufenordnung***). Mit $c_i = \sum_j a_{ij}$ gilt $\Phi_i(\odot) = c_i$, so daß für jedes RK-Verfahren die Stufenordnungen mindestens 1 sind. Die simplifying assumptions $Simp^{(q)}$ sind die Stufenordnungsgleichungen $i = 1, \ldots, s$ zu den "Büscheln"*

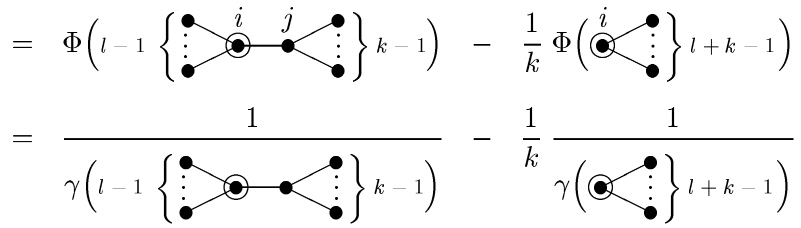


$$, \quad k = 1, \ldots, q \;.$$

Bemerkung 4.77: *Bei Verfahren hoher Ordnung sind die simplifying assumptions notwendigerweise erfüllt. Es gelte $b_j \neq 0$, die c_i seien paarweise verschieden. Für ein (nach Bemerkung 4.61.b implizites) s-stufiges Verfahren der Ordnung $p > s$ muß die Butcher-Matrix $Simp^{(p-s)}$ erfüllen.*

Beweis: Sei $k \in \{1, \ldots, p-s\}$. Für $l = 1, \ldots, p-k$ gilt

$$\sum_{i=1}^{s}\sum_{j=1}^{s} b_i\, c_i^{l-1}\, a_{ij}\, c_j^{k-1} \quad - \quad \frac{1}{k}\sum_{i=1}^{s} b_i\, c_i^{l+k-1}$$



$$= \frac{1}{(k+l)\,k} \quad - \quad \frac{1}{k}\,\frac{1}{k+l} \quad = \quad 0\ ,$$

da die Bäume $k + l \leq p$ Knoten haben. Speziell sind mit $k \leq p - s$ alle Werte $l = 1, \ldots, s$ zulässig, so daß

$$\sum_{i=1}^{s} b_i\, c_i^{l-1}\left(\sum_{j=1}^{s} a_{ij}\, c_j^{k-1} - \frac{c_i^k}{k}\right) = 0\ , \quad l = 1, \ldots, s$$

folgt. Dies kann als homogenes lineares System von s Gleichungen ($l = 1, \ldots, s$) für die simplying assumptions aufgefaßt werden, dessen Koeffizientenmatrix durch $\mathrm{diag}(b_1, \ldots, b_s)$ und die Vandermonde-Matrix (c_i^{l-1}) gegeben ist.

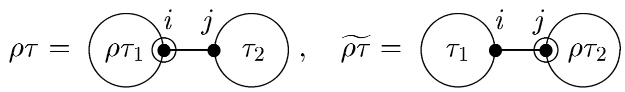
Q.E.D.

Hilfssatz 4.78:

Es gelte die **Symplektizitätsbedingung**

$$Symp: \quad b_i\, a_{ij} + b_j\, a_{ji} \;=\; b_i\, b_j\ , \quad i, j = 1, \ldots, s\ .$$

Dann folgt für "wurzelverschobene" Baumpaare



mit beliebigen Teilbäumen τ_1, τ_2:

$$\Phi(\rho\tau) \;+\; \Phi(\widetilde{\rho\tau}) \;=\; \Phi(\rho\tau_1)\,\Phi(\rho\tau_2)\ .$$

Für die Ordnungsgleichungen folgt

$$\Phi(\rho\tau) - \frac{1}{\gamma(\rho\tau)} \;=\; -\left(\Phi(\widetilde{\rho\tau}) - \frac{1}{\gamma(\widetilde{\rho\tau})}\right) + \;\Phi(\rho\tau_1)\,\Phi(\rho\tau_2) \;-\; \frac{1}{\gamma(\rho\tau_1)}\,\frac{1}{\gamma(\rho\tau_2)}\;.$$

Beweis: Sei (i,j) bzw. (j,i) die Kante, die numerierte Repräsentanten von τ_1 und τ_2 in $\rho\tau$ bzw. $\widetilde{\rho\tau}$ verbindet (die Wurzelverschiebung ändert lediglich die Orientierung dieser Kante). Mit der Definition 4.50 gilt

$$\Phi(\rho\tau) \;=\; \sum_{i=1}^{s}\sum_{j=1}^{s} b_i\,a_{ij}\,(\stackrel{\tau_1}{\cdots})_i\,(\stackrel{\tau_2}{\cdots})_j\;, \quad \Phi(\widetilde{\rho\tau}) \;=\; \sum_{i=1}^{s}\sum_{j=1}^{s} b_j\,a_{ji}\,(\stackrel{\tau_1}{\cdots})_i\,(\stackrel{\tau_2}{\cdots})_j\;,$$

wobei $(\stackrel{\tau_1}{\cdots})_i$ bzw. $(\stackrel{\tau_2}{\cdots})_j$ die Beiträge aus den Kanten in τ_1 bzw. τ_2 sind. Aus der Symplektizitätsbedingung folgt

$$\begin{aligned}\Phi(\rho\tau) + \Phi(\widetilde{\rho\tau}) \;&=\; \sum_{i=1}^{s}\sum_{j=1}^{s} b_i\,b_j\,(\stackrel{\tau_1}{\cdots})_i\,(\stackrel{\tau_2}{\cdots})_j \\ &=\; \Big(\sum_{i=1}^{s} b_i\,(\stackrel{\tau_1}{\cdots})_i\Big)\,\Big(\sum_{j=1}^{s} b_j\,(\stackrel{\tau_2}{\cdots})_j\Big) \;=\; \Phi(\rho\tau_1)\;\Phi(\rho\tau_2)\;.\end{aligned}$$

Für die Dichte gilt mit der rekursiven Definition 4.22

$$\frac{\gamma(\rho\tau)}{|\rho\tau|} \;=\; \frac{\gamma(\rho\tau_1)}{|\rho\tau_1|}\,\gamma(\rho\tau_2)\;, \quad \frac{\gamma(\widetilde{\rho\tau})}{|\widetilde{\rho\tau}|} \;=\; \frac{\gamma(\rho\tau_2)}{|\rho\tau_2|}\,\gamma(\rho\tau_1)\;,$$

also mit $|\rho\tau| = |\widetilde{\rho\tau}| = |\rho\tau_1| + |\rho\tau_2|$: $\quad \dfrac{1}{\gamma(\rho\tau)} + \dfrac{1}{\gamma(\widetilde{\rho\tau})} = \dfrac{1}{\gamma(\rho\tau_1)}\,\dfrac{1}{\gamma(\rho\tau_2)}\;.$

Q.E.D.

Sind die Ordnungsgleichungen für die Teilbäume $\rho\tau_1$, $\rho\tau_2$ erfüllt, so ist die Ordnungsgleichung unabhängig von der Position der Wurzel:

$$\Phi(\rho\tau) - \frac{1}{\gamma(\rho\tau)} \;=\; -\left(\Phi(\widetilde{\rho\tau}) - \frac{1}{\gamma(\widetilde{\rho\tau})}\right).$$

Folgerung 4.79: (Invarianz unter Wurzelverschiebung)

Unter der Symplektizitätsbedingung 4.78 sind die Ordnungsgleichungen aller durch Wurzelverschiebung entstehenden Bäume äquivalent, wenn die Ordnungsgleichungen für alle Bäume niedrigerer Knotenzahl erfüllt sind.

Bemerkung 4.80: *Eine skalare Funktion $E : \mathbb{R}^N \to \mathbb{R}$ heißt* **Erhaltungssatz** *des dynamischen Systems $dy/dt = f(y)$ auf dem $\mathbb{R}^N$, wenn überall $E'(y)[f(y)] = \langle \nabla_y E(y), f(y) \rangle = 0$ gilt. Auf den Lösungskurven $y(t)$ folgt dann*

$$\frac{d}{dt}\, E(y(t)) \;=\; E'(y(t))[f(y(t))] \;=\; 0 \;,$$

d.h., E bleibt im Lauf der Zeit konstant. Lineare Erhaltungssätze der Form $E(y) = \langle C, y \rangle$ sind durch einen konstanten Vektor $C \in \mathbb{R}^N$ mit $\langle C, f(y) \rangle = 0 \ \forall\, y \in \mathbb{R}^N$ gegeben. Die exakte Lösung ist dann auf eine durch den Normalenvektor C gegebene Hyperfläche im $\mathbb{R}^N$ eingeschränkt. Dies gilt auch für die numerische Lösung: für jedes RK-Verfahren folgt

$$E(I_h(y)) - E(y) \;=\; \langle C, I_h(y) - y \rangle \;=\; h \sum_{j=1}^{s} b_j \,\langle C, f(y_j) \rangle \;=\; 0 \;.$$

Verfahren mit der Symplektizitätsbedingung 4.78 erhalten auch (in der Praxis sehr häufig auftretende) quadratische Erhaltungssätze der Form $E(y) = \langle y, By \rangle$, wo B eine symmetrische $N \times N$-Matrix ist:

$$\begin{aligned}
E(I_h(y)) - E(y) &= \langle I_h(y), B I_h(y) \rangle - \langle y, By \rangle \\
&= \langle I_h(y) - y, B\,(I_h(y) - y)\,\rangle + 2\,\langle I_h(y) - y, By \rangle \\
&= \langle I_h(y) - y, B\,(I_h(y) - y)\,\rangle + 2\,h \textstyle\sum_i b_i\, \langle f(y_i), By \rangle \\
&\overset{(*)}{=} \langle I_h(y) - y, B\,(I_h(y) - y)\,\rangle - 2\,h \textstyle\sum_i b_i\, \langle f(y_i), B(y_i - y) \rangle \\
&= \langle h \Big(\textstyle\sum_i b_i f(y_i) \Big), B\; h \Big(\textstyle\sum_j b_j f(y_j) \Big) \rangle \\
&\qquad\qquad - \; 2\,h \textstyle\sum_i b_i\, \langle f(y_i), B\,h \Big(\textstyle\sum_j a_{ij} f(y_j) \Big) \rangle \\
&= h^2 \textstyle\sum_i \sum_j (b_i\, b_j - b_i\, a_{ij} - b_j\, a_{ji})\, \langle f(y_i), B f(y_j) \rangle \\
&= 0 \;,
\end{aligned}$$

wobei in $(*)$

$$0 \;=\; E'(y)[f(y)] \;=\; 2\,\langle f(y), By \rangle \quad \forall\, y \in \mathbb{R}^N$$

benutzt wurde. Die numerische Invarianz quadratischer Erhaltungssätze ist eine für die Praxis sehr attraktive Eigenschaft eines Integrators.

Satz 4.81: (Die Gauß-Legendre-Verfahren, Butcher 1963)

Seien $c_1, \ldots, c_s$ die Nullstellen des Legendre-Polynoms

$$P_s^*(c) \;=\; \frac{d^s}{dc^s}\, c^s\, (1 - c)^s \;.$$

Das s-stufige RK-Verfahren ("Gauß-Legendre-Verfahren") mit dem durch

$$
\begin{aligned}
Quad^{(s)}: \quad \sum_{j=1}^{s} b_j\, c_j^{k-1} &= \frac{1}{k}\,, \quad k=1,\ldots,s \\
Simp^{(s)}: \quad \sum_{j=1}^{s} a_{ij}\, c_j^{k-1} &= \frac{c_i^k}{k}\,, \quad i,k=1,\ldots,s
\end{aligned}
$$

eindeutig festgelegten Butcher-Schema (s lineare Gleichungen für (b_j), s^2 lineare Gleichungen für (a_{ij})) hat die Ordnung $2s$. Das Schema erfüllt die Symplektizitätsbedingung $Symp$ aus 4.78.

Beweis: Seien

$$
\begin{aligned}
C: &\quad c_1,\ldots,c_s \text{ sind Legendre-Nullstellen,} \\
Quad^{(2s)}: &\quad \sum_{j=1}^{s} b_j\, c_j^{k-1} = \frac{1}{k}\,, \quad k=1,\ldots,2\,s\,.
\end{aligned}
$$

Es wird gezeigt:

$$
\begin{tikzcd}
 & Simp^{(s)} \arrow[dr] \arrow[drr] & & \\
 & & Symp \arrow[r] & \text{Ordnung } 2s \\
C \arrow[dr] & & & \\
 & Quad^{(2s)} \arrow[uur] \arrow[uurr] & & \\
Quad^{(s)} \arrow[ur] & & &
\end{tikzcd}
$$

$\underline{C, Quad^{(s)} \Rightarrow Quad^{(2s)}}$: Die Bedingungen C und $Quad^{(s)}$ besagen, daß (c_j), (b_j) die Daten der Gauß-Quadratur sind: die Quadraturformel

$$
\int_{t_0}^{t_0+h} f(t)\,dt \;=\; h \sum_{j=1}^{s} b_j\, f(t_0+c_j h) \;+\; \text{Fehler}
$$

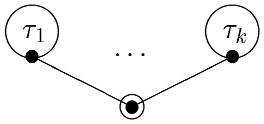
ist für alle Polynome f bis zum Grad $2s-1$ exakt. Die Monome $f(t)=(t-t_0)^{k-1}$ mit $k=s+1,\ldots,2s$ liefern die zusätzlichen Quadraturbedingungen.

$\underline{Simp^{(s)}, Quad^{(2s)} \Rightarrow Symp}$: Für beliebige $k,l\in\{1,\ldots,s\}$ gilt

$$
\begin{aligned}
&\sum_{i=1}^{s}\sum_{j=1}^{s} c_i^{l-1}\,(b_i\,a_{ij}+b_j\,a_{ji}-b_i\,b_j)\,c_j^{k-1} \\
=\;& \frac{1}{k}\sum_{i=1}^{s} b_i\,c_i^{k+l-1} + \frac{1}{l}\sum_{j=1}^{s} b_j\,c_j^{k+l-1} - \Big(\sum_{i=1}^{s} b_i\,c_i^{l-1}\Big)\Big(\sum_{j=1}^{s} b_j\,c_j^{k-1}\Big) \\
=\;& \frac{1}{k}\,\frac{1}{k+l} + \frac{1}{l}\,\frac{1}{k+l} - \frac{1}{l}\,\frac{1}{k} \;=\; 0\,.
\end{aligned}
$$

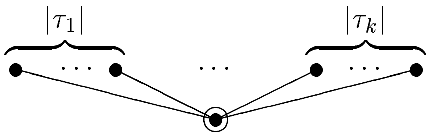
Dies sind die Komponenten (l, k) der Matrixgleichung $V^T MV = 0$ mit der Vandermonde-Matrix $(V_{jk}) = (c_j^{k-1})$ und $(M_{ij}) = (b_i a_{ij} + b_j a_{ji} - b_i b_j)$. Da die Legendre-Nullstellen c_i paarweise verschieden sind, folgt $M_{ij} = 0$.

$\underline{Simp^{(s)}, Symp, Quad^{(2s)} \Rightarrow \text{Ordnung } 2s}$: Induktiv wird gezeigt: hat das Verfahren die Ordnung $p - 1 < 2s$, dann hat es auch die Ordnung p. Betrachte dazu einen beliebigen Baum $\rho\tau$ mit p Knoten. Mittels Satz 4.79 kann die Wurzel in das "Zentrum" des Baums verschoben werden, bis alle von der Wurzel ausgehenden Teilbäume höchstens $p/2 \leq s$ Knoten haben:



$$|\tau_1| \leq s, \ldots, |\tau_k| \leq s \ .$$

Mittels Satz 4.75 können alle Teilbäume reduziert werden, so daß ein "Büschel" mit $p \leq s$ Knoten entsteht:



Mit $Quad^{(2s)}$ sind die Ordnungsgleichungen der "Büschel" bis zu $2s$ Knoten erfüllt.

Q.E.D.

Bemerkung 4.82: *Die Lösung der linearen Gleichungen $Quad^{(s)}$ und $Simp^{(s)}$ für (b_j) und (a_{ij}) ist mit den Lagrange-Polynomen*

$$L_j(c) \ = \ \prod_{\substack{k=1 \\ k \neq j}}^{s} \frac{c - c_k}{c_j - c_k}$$

zu $c_1, \ldots, c_s$ durch

$$a_{ij} \ = \ \int_0^{c_i} L_j(c)\, dc \ , \quad b_j \ = \ \int_0^1 L_j(c)\, dc \ , \quad i, j = 1, \ldots, s$$

darstellbar.

Bemerkung 4.83: *Die s-stufigen RK-Verfahren der Ordnung $2s$ sind bis auf Permutation der Butcher-Parameter gemäß Bemerkung 4.46 eindeutig: $(c_i), (b_j)$ sind als Daten der Gauß-Quadratur festgelegt. Mit Bemerkung 4.77 folgt notwendigerweise $Simp^{(s)}$, womit auch die Matrix (a_{ij}) festgelegt ist.*

Die ersten Gauß-Legendre-Verfahren 4.84: *Für das 1-stufige Verfahren der Ordnung 2*

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

ist der Zeitschritt $y \to Y = I_h(y)$ *mit*

$$y_1 = y + \frac{h}{2}\, f(y_1)\ , \quad Y = y + h\, f(y_1) \quad \Rightarrow \quad y_1 = \frac{1}{2}\,(y+Y)$$

als Lösung der Gleichung

$$Y \;=\; y \;+\; h\, f\Big(\, \frac{1}{2}\,(y+Y)\,\Big)$$

*definiert (***"implizite Mittelpunktsregel"***).*

Das 2-stufige Verfahren 4.ter Ordnung ist

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Das 3-stufige Verfahren 6.ter Ordnung ist

$$\begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

Bemerkung 4.85: *Nach Bemerkung 4.71 können diese Verfahren mit* $2s-1$ *Schritten der Fixpunktiteration 4.69 explizit gemacht werden, ohne daß die Ordnung verlorengeht. Damit existieren explizite Verfahren der Ordnung* $2s$ *mit* $2s^2 - s + 1$ *Stufen.*

Bemerkung 4.86: *Satz 4.81 läßt sich verallgemeinern: seien* $c_1, \ldots, c_s$ *paarweise verschieden, aber sonst beliebig. Legt man* (b_j), (a_{ij}) *durch* $Quad^{(s)}$, $Simp^{(s)}$ *fest (vergleiche Bemerkung 4.82), so hat das resultierende RK-Verfahren genau die Ordnung* p*, welche die durch* $(c_j), (b_j)$ *gegebene Quadraturformel*

$$\int_{t_0}^{t_0+h} f(t)\, dt \;=\; h \sum_{j=1}^{s} b_j\, f(t_0 + c_j\, h) \;+\; O(h^{p+1})$$

hat, d.h., $p-1$ ist der polynomiale Exaktheitsgrad. Es gilt hierbei stets $s \leq p \leq 2s$, denn mit $Quad^{(s)}$ sind die b_j als die Gewichte der Newton-Cotes-Quadratur zu den Knoten c_j gewählt, so daß alle Polynome bis zum Grad $s-1$ exakt integriert werden. Durch bestimmte Wahl der c_i kann höherer Exaktheitsgrad bis hin zur Ordnung $2s$ (Gauß-Legendre-Quadratur) erreicht werden. Die so konstruierten Verfahren heißen **"vom Kollokationstyp"**, *ihre Stufenordnungen 4.76 sind stets s. Der Beweis kann graphentheoretisch analog zu Satz 4.81 geführt werden, wobei allerdings i.a. nicht die Symplektizitätsbedingung Symp aus 4.78 gilt, sondern durch "row simplifying assumptions" [Butcher, Formel (342c)] ersetzt wird, mit denen Bäume ähnlich wie in Hilfssatz 4.74/Satz 4.75 "von der Wurzel" her vereinfacht werden können ([Butcher, Theorem 342C]). Ein alternativer Beweis ist z.B. in [Deuflhard&Bornemann] zu finden.*

4.6 Zeitumkehr: adjungierte Verfahren

Die Inverse des exakten Flußes ist wieder der Fluß: $(F_h)^{-1} = F_{-h}$. Für den numerischen Fluß I_h wird i.a. ein anderes Verfahren zur Invertierung benötigt:

Definition 4.87:

Das einem Verfahren I_h **adjungierte** *Verfahren ist $I_h^* = (I_{-h})^{-1}$. Ein Verfahren mit $I_h = I_h^*$ heißt* **symmetrisch** *(***selbstadjungiert, reflexiv, reversibel***).*

Satz 4.88:

a) *Das s-stufige RK-Verfahren mit den Butcher-Parametern*

$$c_i^* = 1 - c_i \ , \quad a_{ij}^* = b_i - a_{ij} \ , \quad b_j^* = b_j \ , \quad i,j = 1,\ldots,s$$

liefert die Adjungierte des s-stufigen RK-Verfahrens mit den Parametern $(c_i), (a_{ij}), (b_j)$.

b) *Ein konsistentes s-stufiges RK-Verfahren, dessen Butcher-Parameter die Symmetrie*

$$c_{\pi(i)} = 1 - c_i \ , \quad a_{\pi(i),\pi(j)} = b_i - a_{ij} \ , \quad b_{\pi(j)} = b_j \ , \quad i,j = 1,\ldots,s$$

mit einer beliebigen Permutation $\pi : \{1,\ldots,s\} \to \{1,\ldots,s\}$ aufweisen, ist symmetrisch. Es ist notwendigerweise implizit.

Beweis: a) Sei I_h^* das Verfahren mit $(c_i^*), (a_{ij}^*), (b_j^*)$. Es wird gezeigt, daß $I_{-h}^* \circ I_h$ die identische Abbildung ist, womit I_h^* als das adjungierte Verfahren identifiziert ist. Sei $\tilde{y} = I_h(y) = y + h \sum_j b_j f(y_j)$. Die Zwischenstufen $\tilde{y}_i^*$ in der Auswertung von $I^*_{-h}(\tilde{y})$ sind durch

$$\begin{aligned}\tilde{y}_i^* &= \tilde{y} - h\sum_{j=1}^s a_{ij}^*\, f(\tilde{y}_i^*) \;=\; y + h\sum_{j=1}^s (b_i - a_{ij}^*)\, f(\tilde{y}_i^*) \\ &= y + h\sum_{j=1}^s a_{ij}\, f(\tilde{y}_i^*)\ , \quad i = 1,\ldots,s\end{aligned}$$

definiert. Die durch $y_i = y + h\sum_j a_{ij} f(y_j)$ definierten Stufen des Schrittes $I_h(y)$ bilden die (für hinreichend kleines h) eindeutige Lösung dieser Gleichungen. Mit $\tilde{y}_i^* = y_i$ folgt

$$I^*_{-h}(I_h(y)) = \tilde{y} - h\sum_{j=1}^s b_j^*\, f(\tilde{y}_j^*) = y + h\sum_{j=1}^s \Big(b_j\, f(y_j) - b_j^*\, f(\tilde{y}_j^*)\Big) = y\ .$$

b) folgt unmittelbar aus a) und Bemerkung 4.46. Die Diagonale der Butcher-Matrix kann nicht identisch verschwinden: für ein explizites Verfahren mit $a_{11} = \ldots = a_{ss} = 0$ würde im Widerspruch zur Konsistenz $b_i = a_{\pi(i),\pi(i)} + a_{ii} = 0$ für alle $i = 1,\ldots,s$ folgen.

Q.E.D.

Die "Spiegelung" $I_h \to I_h^*$ erhält die Ordnung:

Satz 4.89:

Für die Adjungierte I_h^* eines RK-Verfahrens I_h der Ordnung p mit

$$F_h(y) - I_h(y) \;=\; e(y)\, h^{p+1} + O(h^{p+2})$$

gilt

$$F_h(y) - I_h^*(y) \;=\; (-1)^p\, e(y)\, h^{p+1} + O(h^{p+2})\ .$$

Beweis: Es gelte $F_h(y) = I_h^*(y) + e^*(y)\, h^{p^*+1} + O(h^{p^*+2})$. Mit Lipschitz-stetigem führenden Fehlerkoeffizienten $e(y)$ von I_h und

$$I_{-h}(\tilde{y} + \Delta y) \;=\; I_{-h}(\tilde{y}) \;+\; \Delta y \;+\; h\, O(\Delta y)$$

folgt

$$\begin{aligned} y \;=\; F_{-h}(F_h(y)) \;&=\; I_{-h}(F_h(y)) \;+\; e(F_h(y))\,(-h)^{p+1} \;+\; O(h^{p+2}) \\ =\; I_{-h}\Big(\, I_h^*(y)) &+ e^*(y)\, h^{p^*+1} + O(h^{p^*+2})\,\Big) \;+\; e(F_h(y))\,(-h)^{p+1} \;+\; O(h^{p+2}) \\ =\; \underbrace{I_{-h}(I_h^*(y))}_{y} &\;+\; e^*(y)\, h^{p^*+1} \;+\; e(y)\,(-h)^{p+1} \;+\; O(h^{\min(p,p^*)+2})\ . \end{aligned}$$

Es folgt $p^* = p$ und $e^*(y) = (-1)^p e(y)$.

Q.E.D.

Satz 4.90:

Ein symmetrisches RK-Verfahren hat die Ordnung p, wenn die Ordnungsgleichungen $\Phi(\rho\tau) = 1/\gamma(\rho\tau)$ für alle Bäume mit ungerader Knotenzahl $|\rho\tau| \leq p$ erfüllt sind. Die Ordnung ist stets gerade.

Beweis: Die Ordnungsgleichungen bis zur ungeraden Ordnung q seien erfüllt. Mit $I_h = I_h^*$ gilt nach Satz 4.89 für den führenden Fehlerkoeffizienten $e(y) = (-1)^q\, e(y) = -e(y)$, d.h., $e(y) = 0$. Die Ordnungsgleichungen der geraden Ordnung $q+1$ sind damit automatisch erfüllt. Für die Ordnungsgleichungen bis zur Ordnung p sind damit nur die ungeraden Ordnungsgleichungen zu fordern.

Q.E.D.

Beispiel 4.91: *Für ein symmetrisches RK-Verfahren 4.88.b) folgt*

$$\sum_{j=1}^{s} b_j\, c_j = \sum_{j=1}^{s} b_{\pi(j)}\, c_{\pi(j)} = \sum_{j=1}^{s} b_j\,(1 - c_j) = \sum_{j=1}^{s} b_j \;-\; \sum_{j=1}^{s} b_j\, c_j \ ,$$

d.h.,

$$\Phi(\odot\!\!-\!\!\bullet) \;=\; \sum_{j=1}^{s} b_j\, c_j \;=\; \frac{1}{2}\sum_{j=1}^{s} b_j \;=\; \frac{1}{2}\,\Phi(\odot)\ .$$

Mit $\gamma(\odot\!\!-\!\!\bullet) = 2\,\gamma(\odot)$ ist jedes konsistente symmetrische RK-Verfahren bereits von zweiter Ordnung.

Bemerkung 4.92: *Die Symmetrieforderung ist damit hilfreich in der Konstruktion impliziter Verfahren, da die Ordnungsgleichungen von Bäumen mit gerader Knotenzahl nicht betrachtet zu werden brauchen.*

Bemerkung 4.93: *Für den globalen Fehler eines RK-Verfahrens der Ordnung p mit konstanter Schrittweite $h = h(n) = T/n$ gelte eine asymptotische Entwicklung der Form*

$$F_T(y_0) - (\underbrace{I_h \circ \ldots \circ I_h}_{n})(y_0) \;=\; e_p(T, y_0)\, h^p \;+\; e_{p+1}(T, y_0)\, h^{p+1} + \cdots$$

Für symmetrische Verfahren enthält dies Reihe nur gerade Potenzen in h [Hairer, Nørsett & Wanner, Theorem 8.9]. Symmetrische Verfahren bieten sich damit an, durch Extrapolation (simultane Auswertung mit mehreren Schrittweiten, daraus resultierende Fehlerabschätzungen und Korrekturen wie in Abschnitt 4.4) verbessert zu werden.

Satz 4.94:

Die Gauß-Legendre-Verfahren 4.81 sind symmetrisch.

Beweis: Mit der Anordnung $0 < c_1 < \ldots < c_s < 1$ der Legendre-Wurzeln gilt $c_{\pi(i)} = 1 - c_i$ mit $\pi(i) = s + 1 - i$. Für die zugeordneten Lagrange-Polynome folgt

$$L_{\pi(j)}(c) = L_j(1-c) \ ,$$

und daraus mit Bemerkung 4.82

$$\begin{aligned} a_{\pi(i),\pi(j)} &= \int_0^{c_{\pi(i)}} L_{\pi(j)}(c)\,dc \;=\; \int_0^{1-c_i} L_j(1-c)\,dc \\ &= \int_{c_i}^1 L_j(c)\,dc \;=\; \int_0^1 L_j(c)\,dc - \int_0^{c_i} L_j(c)\,dc \;=\; b_i - a_{ij} \ , \end{aligned}$$

und

$$b_{\pi(j)} \;=\; \int_0^1 L_{\pi(j)}(c)\,dc \;=\; \int_0^1 L_j(1-c)\,dc \;=\; \int_0^1 L_j(c)\,dc \;=\; b_j \ .$$

Q.E.D.

4.7 A-Stabilität, steife Systeme

Idee: versuche gewisse qualitative Eigenschaften spezieller Systeme bei numerischer Approximation zu erhalten. Bei sogenannten asymptotisch stabilen dynamischen Systemen laufen alle Lösungskurven für große Zeiten gegen einen Grenzpunkt (Attraktor). Der Prototyp eines solchen Systems ist das **skalare Testproblem**

$$\frac{dy}{dt} \;=\; \lambda\, y \ , \quad y(t), \lambda \in \mathbb{C}$$

dessen Lösungen $y(t) = y(t_0)\, e^{\lambda(t-t_0)}$ gegen 0 konvergieren, falls der Realteil $\Re e(\lambda)$ negativ ist. Numerische Verfahren, die dieses Verhalten <u>mit beliebigen Schrittweiten</u> erhalten, heißen A-stabil.

Satz 4.95: (Die Stabilitätsfunktion eines RK-Verfahrens)

Der Zeitschritt $I_h(y) = p(\lambda\, h)\, y$ eines s-stufigen RK-Verfahren $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ *angewendet auf das Testproblem $dy/dt = \lambda y$ in $\mathbb{C}$ ist die Multiplikation mit einem skalaren Faktor $p(\lambda h)$, der* **Stabilitätsfunktion** *des Verfahrens. Mit dem euklidischen Skalarprodukt $\langle .,. \rangle$ auf $\mathbb{R}^s$ und $e = (1, \ldots, 1)^T \in \mathbb{R}^s$ ist*

$$p(z) \;=\; 1 \;+\; z\,\langle b, (\mathbb{1} - z\,A)^{-1} e\rangle \;\overset{(*)}{=}\; \frac{\det(\,\mathbb{1} - z\,(A - e\,b^T)\,)}{\det(\mathbb{1} - zA)}$$

eine rationale Funktion des Parameters $z = \lambda h$. Für explizite Verfahren ist die Stabilitätsfunktion ein Polynom.

Beweis: Für das Testproblem sind die Zwischenstufen in $I_h(y)$ gegeben durch

$$y_i - \lambda\, h \sum_{j=1}^{s} a_{ij}\, y_j = y \ , \quad \text{d.h.,} \quad \begin{pmatrix} y_1 \\ \vdots \\ y_s \end{pmatrix} = (\mathbb{1} - z\, A)^{-1} \begin{pmatrix} y \\ \vdots \\ y \end{pmatrix} .$$

Für explizite Verfahren gilt dabei

$$(\mathbb{1} - zA)^{-1} = 1 + z\, A + z^2\, A^2 + \cdots + z^{s-1}\, A^{s-1} \ , \quad A^s = 0 \ .$$

Die Identität $(*)$ folgt aus

$$(\mathbb{1} - zA)^{-1} (\, \mathbb{1} - z\,(A - e\, b^T)\,) \ = \ \mathbb{1} + z\,(1 - zA)^{-1} e\, b^T$$

und $\det(\mathbb{1} + x\, b^T) = 1 + \langle x, b\rangle$ mit $x = z\,(1 - zA)^{-1} e$.

Q.E.D.

Bemerkung 4.96: *Die Taylor-Entwicklung von $p(z)$ um $z = 0$ ist*

$$p(z) \ = \ 1 + z\,\langle b, e\rangle + z^2\,\langle b, Ae\rangle + z^3\,\langle b, A^2 e\rangle + \cdots \ .$$

Mit

$$\langle b, A^{k-1} e\rangle \ = \ \Phi(\rho\tau_k) \ , \quad \rho\tau_k \ = \ \underbrace{\circledcirc\!-\!\bullet \ \cdots \ \bullet\!-\!\bullet}_{k \text{ Knoten}}$$

und $\Phi(\rho\tau_k) = 1/\gamma(\rho\tau_k) = 1/k!$ folgt

$$p(z) \ = \ 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^q}{q!} \ + \ O(z^{q+1}) \ = \ e^z + O(z^{q+1}) \ ,$$

wenn das Verfahren die Ordnung q hat. Die Stabilitätsfunktion kodiert damit die Ordnungsgleichungen der "gestreckten" Bäume und liefert eine Approximation der e-Funktion (der exakte Fluß $F_h = e^{\lambda h}$ von $dy/dt = \lambda y$). Mit hinreichend kleinen Schrittweiten, genauer, für $|z| = |\lambda h| \ll 1$, approximiert damit das Verfahren die exakte Lösung.

Bemerkung 4.97: *Für ein lineares System $dy/dt = By$ auf dem $\mathbb{R}^N$ liefert das RK-Verfahren den Zeitschritt $I_h(y) = p(hB)y$ durch Multiplikation mit der Matrix $p(hB)$, die für diagonalisierbares $B = T\,\mathrm{diag}(\lambda_1, \ldots, \lambda_N)\, T^{-1}$ durch*

$$p(hB) \ = \ T\,\mathrm{diag}\Big(p(\lambda_1 h), \ldots, p(\lambda_N h)\Big)\, T^{-1}$$

gegeben ist. Damit beschreibt die Wirkung des Integrators auf das Testproblem $dy/dt = \lambda y$ (mit komplexen λ) vollständig die Wirkung auf beliebige lineare Systeme. Der exakte Fluß ist durch

$$F_h \ = \ e^{hB} \ = \ T\,\mathrm{diag}\Big(e^{\lambda_1 h}, \ldots, e^{\lambda_N h}\Big)\, T^{-1}$$

gegeben.

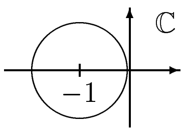
Definition 4.98:

Der **Stabilitätsbereich** *eines Integrators mit Stabilitätsfunktion* $p : \mathbb{C} \rightarrow \mathbb{C}$ *ist*

$$S \;=\; \{\, z \in \mathbb{C}\,;\, |p(z)| < 1 \,\} \;.$$

Der Integrator heißt **A-stabil**, *wenn der Stabilitätsbereich die linke Halbebene umfaßt:* $|p(z)| < 1 \;\forall\; z \in \mathbb{C}$ *mit* $\Re e(z) < 0$.

Beispiel 4.99: *Für das skalare Testproblem* $dy/dt \,=\, f(y) \,=\, \lambda\, y$ *liefert das Euler-Verfahren* $I_h(y) \,=\, y + hf(y) \,=\, (1 + \lambda\, h)y$, *also* $p(z) \,=\, 1 + z$. *Der Stabilitätsbereich ist der Einheitskreis in der komplexen Ebene mit dem Zentrum* $(-1, 0)$:



Für die durch $Y \;=\; y \,+\, hf((y \,+\, Y)/2) \;=\; y \,+\, \lambda h(y \,+\, Y)/2$ *definierte implizite Mittelpunktsregel 4.84 gilt*

$$I_h(y) \;=\; Y \;=\; \frac{1 + \lambda\, h/2}{1 - \lambda\, h/2}\; y \;=\; p(\lambda\, h)\, y \;, \quad p(z) \;=\; \frac{1 + z/2}{1 - z/2} \;.$$

Die Forderung

$$|p(z)|^2 \;=\; p(z)p(\bar z) \;=\; \frac{1 + (z + \bar z)/2 + z\bar z/4}{1 - (z + \bar z)/2 + z\bar z/4} \;<\; 1$$

ist äquivalent zu $\Re e(z) \,=\, (z \,+\, \bar z)/2 \,<\, 0$. *Der Stabilitätsbereich ist die offene linke komplexe Halbebene, d.h., die implizite Mittelpunktsregel ist A-stabil.*

Bemerkung 4.100: *Für explizite Verfahren ist die Stabilitätsfunktion 4.95 ein Polynom und damit in der linken komplexen Halbebene unbeschränkt: explizite Verfahren können nicht A-stabil sein.*

Bemerkung 4.101: *(Zur Interpretation des Stabilitätsbereichs)*
Ein lineares AWP $dy/dt = By$, $y(t_0) = y_0$ *auf dem* $\mathbb{R}^N$ *mit diagonalisierbarem* $B = T\,\mathrm{diag}(\lambda_1, \ldots, \lambda_N)\, T^{-1}$ *hat die Lösungen*

$$y(t) = T\,\mathrm{diag}\Big(e^{\lambda_1 (t - t_0)}, \ldots, e^{\lambda_N (t - t_0)}\Big)\, T^{-1}\, y_0 \;.$$

Gilt für alle Eigenwerte $\Re e(\lambda_i) \,<\, 0$, *so ist das System "asymptotisch stabil":* $\lim\limits_{t \to \infty} y(t) = 0$ *für alle Startwerte. Mit dem iterierten Zeitschritt*

$$(\underbrace{I_h \circ \ldots \circ I_h}_{n})\,(y_0) \;=\; p(hB)^n y_0 \;=\; T\,\mathrm{diag}\Big(p(\lambda_1 h)^n, \ldots, p(\lambda_N h)^n\Big)\, T^{-1}\, y_0$$

eines Integrators mit konstanter Schrittweite wird das Abklingen der exakten Lösung numerisch genau dann beschrieben, wenn für alle Eigenwerte $|p(\lambda_i h)| < 1$ *gilt, d.h.,*

$$\lambda_i h \in S \ . \qquad (\#)$$

Bei A-stabilen Verfahren ist die Bedingung (#) *für jedes* $h > 0$ *automatisch erfüllt, es folgt*

$$\lim_{n\to\infty} \underbrace{(I_h \circ I_h \circ \cdots \circ I_h)}_{n} (y_0) = 0$$

für alle Startpunkte y_0 *und* <u>*für alle Schrittweiten* $h > 0$</u>.

Ist das Verfahren nicht A-stabil, so kann die Forderung (#) zu dramatischen Einschränkungen an die Wahl der Schrittweite führen, wie folgendes Beispiel zeigt:

Beispiel 4.102: *Das AWP*

$$\frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -1000 & 1 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} , \quad \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

hat die exakte Lösung

$$y_1(t) = \frac{1}{999} \left(e^{-t} - e^{-1000\,t} \right) , \quad y_2(t) = e^{-t} \ .$$

Für $t > 1/1000$ *trägt der vom Eigenwert* $\lambda_1 = -1000$ *stammende Anteil der Lösung praktisch nichts mehr bei. Bei Verwendung der Euler-Methode* $I_h(y) = (1 + hB)\, y$ *folgt mit der Diagonalisierung* $B = T\,\mathrm{diag}(-1000, -1)\, T^{-1}$:

$$\underbrace{(I_h \circ \ldots \circ I_h)}_{n} (y_0) = T \begin{pmatrix} (1-1000\,h)^n & 0 \\ 0 & (1-h)^n \end{pmatrix} T^{-1}\ y_0 \ .$$

Der durch e^{-t} *gegebene führende Term der Lösung wird für einige Zeitschritte numerisch approximiert, wenn* $h \ll 1$ *gilt. Wählt man jedoch Schrittweiten* $2/1000 < h \ll 1$, *so erzeugt der Eigenwert* $\lambda_1 = -1000$ *eine numerische Katastrophe: es gilt* $1 - 1000\,h < -1$, *so daß* $(1 - 1000\,h)^n$ *mit wachsendem* n *unter wechselnden Vorzeichen explodiert, die numerische Lösung dominiert und völlig unbrauchbar macht. Man muß Schrittweiten* $h < 2/1000$ *wählen (dies ist die Bedingung* $\lambda_1 h \in S$, *d.h.,* $-1000h \in (-2, 0)$*), damit der von diesem Eigenwert stammende Anteil der numerischen Lösung nicht explodiert.*

Mit diesem Beispiel wird folgendes heuristisches Prinzip plausibel:

> *Zur Integration linearer Systeme muß die Schrittweite* h *so klein gewählt werden, daß für alle Eigenwerte mit negativem Realteil* $\lambda_i h \in S$ *gilt.*

Ist dies nicht der Fall, so wird in der numerischen Lösung $y_n = I_h \circ \cdots \circ I_h(y_0)$ ein mit n anwachsender Term erzeugt, obwohl in der exakten Lösung der entsprechende Term exponentiell abklingt. Damit bestimmt der negativste Realteil der Eigenwerte die Schrittweite (und damit den Rechenaufwand), obwohl dieser Eigenwert zur exakten Lösung praktisch nichts beiträgt. Es soll ein Maß eingeführt werden, das das Auftreten solcher numerischer Probleme anzeigt.[2]

Definition 4.103:

Ein lineares System $dy/dt = By$ heißt **steif**, *wenn mit den Eigenwerten λ_i von B gilt*

$$\sigma(B) := \frac{\max\limits_i |\Re e(\lambda_i)|}{|\max\limits_i \Re e(\lambda_i)|} \gg 1 \ .$$

Für nichtlineare Systeme $dy/dt = f(y)$ auf dem $\mathbb{R}^N$ betrachte als lokales Steifheitsmaß $\sigma(f'(y))$ mit der Jacobi-Matrix erster partieller Ableitungen $f'(y) = (\partial f_i/\partial y_j)$.

Erklärung: Die Eigenwerte seien in der Form $\Re e(\lambda_1) \geq \ldots \geq \Re e(\lambda_N)$ geordnet. Dann tritt $\sigma(B) \gg 1$ in der Situation $\Re e(\lambda_N) \ll -|\Re e(\lambda_1)|$ ein. Um den durch λ_1 gegebenen führenden Term der exakten Lösung in seiner Größenordnung numerisch richtig beschreiben zu können, muß $h\,|\Re e(\lambda_1)| \ll 1$ gelten. Damit der von λ_N erzeugte Anteil der numerischen Lösung abfällt und nicht –analog zu Beispiel 4.102– eine numerische Katastrophe erzeugt, wird die Schrittweite durch die weitere Forderung $h\,\lambda_N \in S$ eingeschränkt. Bei kleinen Stabilitätsbereichen kann dies zu wesentlich kleineren Schrittweiten zwingen als die Forderung $h\,|\Re e(\lambda_1)| \ll 1$, die für eine qualitativ richtige numerische Beschreibung des führenden Exponentialterms benötigt wird.

Für nichtlineare Systeme $dy/dt = f(y)$ gilt in der Umgebung eines Punktes y_0 die Taylor-Entwicklung

$$f(y) \approx f(y_0) + B\,(y - y_0) \ , \quad B = f'(y_0) \ .$$

Man rechnet leicht nach, daß ein Zeitschritt des RK-Verfahrens I_h mit der Stabilitätsfunktion p für $f(y) = f(y_0) + B(y - y_0)$ zu

$$I_h(y) - I_h(y_0) = p(hB)\,(y - y_0)$$

führt, d.h., lokal beschreibt $p(hB)$ das Auseinanderlaufen numerischer Trajektorien. Für Eigenwerte von B mit sehr negativem $\Re e(\lambda_i)$ gibt es Nachbartrajektorien, die sehr schnell auf die Trajektorie $F_h(y_0)$ zulaufen. Numerisch werden jedoch stattdessen stark divergierende Punkte erzeugt, wenn die Schrittweite so

[2]Es gibt in der Literatur kein allgemein akzeptiertes Steifheitsmaß.

groß ist, daß nicht alle Eigenwerte λ_i von B mittels $z = \lambda_i h$ in den Stabilitätsbereich hineingezogen werden. Damit dies nicht zu extrem kleinen Schrittweiten zwingt, sollte man einen A-stabilen Integrator nehmen, für den die Forderung $|p(h\lambda_i)| < 1$ automatisch erfüllt ist.

Faustregel 4.104: *Steife Systeme sollten mit A-stabilen Verfahren integriert werden, um nicht zu kleine Schrittweiten wählen zu müssen. In der Lösung steifer Problem werden somit implizite Verfahren wichtig. Obwohl in einem einzelnen Schritt gegenüber expliziten Verfahren höherer Rechenaufwand auftritt, kann bei A-stabilen impliziten Integratoren durch größere Schrittweiten ein geringerer Gesamtaufwand erreicht werden.*

Satz 4.105:

Ein RK-Verfahren $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ *mit der Eigenschaft Symp 4.78 und* $b_i > 0$ *ist A-stabil.*

Beweis: Sei $B = \mathrm{diag}(b_1, \ldots, b_s)$ und $e = (1, \ldots, 1)^T$.

a) Mit $Ax = \lambda x \in \mathbb{C}^s$ und dem üblichen komplexen Skalarprodukt $\langle x, y \rangle = \sum_i \bar{x}_i y_i$ folgt aus *Symp*, d.h., $BA + A^T B = b\, b^T$:

$$(\lambda + \bar{\lambda})\, \langle x, Bx \rangle \;=\; \langle x, (BA + A^T B)x \rangle \;=\; \langle x, b \rangle\, \langle b, x \rangle \;=\; |\langle b, x \rangle|^2 \;\geq\; 0\;.$$

Damit gilt $\Re e(\lambda) \geq 0$ für alle Eigenwerte λ von A. Für mindestens einen Eigenwert muß dabei $\Re e(\lambda) > 0$ erfüllt sein, da für die Diagonalelemente $a_{ii} = b_i/2$ gilt. Mit $\sum_i \lambda_i = \mathrm{tr}(A) = \sum_i a_{ii} = \sum_i b_i/2 > 0$ können nicht alle Eigenwerte rein imaginär sein.

b) Aus *Symp* folgt $A - e\, b^T = -B^{-1} A^T B$ und somit

$$\det(\mathbb{1} - z(A - e\, b^T)) = \det(B^{-1}(\mathbb{1} + zA^T)B) = \det(\mathbb{1} + zA^T) = \det(\mathbb{1} + zA).$$

Für die Stabilitätsfunktion 4.95 ergibt sich mit $\mu = 1/z$

$$p(z) \;=\; \frac{\det(\mathbb{1} + zA)}{\det(\mathbb{1} - zA)} \;=\; \frac{\det(\mu\mathbb{1} + A)}{\det(\mu\mathbb{1} - A)} \;=\; \prod_{i=1}^{s} \frac{\mu + \lambda_i}{\mu - \lambda_i} \qquad (\#)$$

mit den Eigenwerten λ_i von A. Für $\Re e(z) < 0$, d.h., $\Re e(\mu) < 0$, ist $|p(z)| < 1$ zu zeigen. Für beliebige komplexe Zahlen μ, λ gilt

$$\Big(|\mu - \lambda|\, |\mu - \bar{\lambda}|\Big)^2 - \Big(|\mu + \lambda|\, |\mu + \bar{\lambda}|\Big)^2 = -8\, |\mu|^2 |\lambda|^2\, \Re e(\mu)\, \Re e(\lambda)\;.$$

Mit $\Re e(\mu) < 0$ und $\Re e(\lambda) \geq 0$ folgt

$$\frac{|\mu + \lambda|\, |\mu + \bar{\lambda}|}{|\mu - \lambda|\, |\mu - \bar{\lambda}|} \;\leq\; 1 \quad \text{bzw.} \quad \frac{|\mu + \lambda|}{|\mu - \lambda|} \;\leq\; 1$$

für jedes konjugierte Eigenwertpaar $\lambda, \bar{\lambda}$ von A bzw. für jeden rellen Eigenwert λ von A. Für mindestens einen Eigenwert bzw. ein Eigenwertpaar gilt dabei mit $\Re e(\lambda) > 0$ die strenge Ungleichung < 1 statt ≤ 1. Für die aus diesen Faktoren aufgebaute Stabilitätsfunktion (#) gilt damit in der linken Halbebene $|p(z)| < 1$.

Q.E.D.

Folgerung 4.106: *Die Gauß-Legendre Verfahren 4.81 sind A-stabil, da die Gewichte b_i der Gauß-Quadratur positiv sind und nach dem Beweis von Satz 4.81 die Symplektizitätsbedingung gilt.*

Kapitel 5

Symplektische Integration

Definition 5.1:

Mittels einer skalaren Funktion $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *definiert man ein* **Hamilton-System**

$$\frac{d}{dt}\begin{pmatrix} q_1 \\ \vdots \\ q_n \\ p_1 \\ \vdots \\ p_n \end{pmatrix} = \begin{pmatrix} \partial H/\partial p_1 \\ \vdots \\ \partial H/\partial p_n \\ -\partial H/\partial q_1 \\ \vdots \\ -\partial H/\partial q_n \end{pmatrix}.$$

Bezeichung: $H(q_1, \ldots, q_n, p_1, \ldots, p_n)$ *heißt* **Hamilton-Funktion** *("Energie" des Systems). Kompakte Schreibweise als dynamisches System:*

$$\frac{dy}{dt} = f(y) = \begin{pmatrix} 0 & \mathbb{1} \\ -\mathbb{1} & 0 \end{pmatrix}\begin{pmatrix} \nabla_q H \\ \nabla_p H \end{pmatrix} = \mathcal{P}\,\nabla_y H$$

mit $y = (q,p)^T \in \mathbb{R}^n \times \mathbb{R}^n$ *und* $\mathcal{P} = \begin{pmatrix} 0 & \mathbb{1} \\ -\mathbb{1} & 0 \end{pmatrix}$.

Beispiel 5.2: *Der Spezialfall* $H(q,p) = \frac{1}{2}\langle p,p\rangle + V(q)$ *("kinetische Energie" + "potentielle Energie") liefert die Newtonschen Bewegungsgleichungen:*

$$\left.\begin{aligned} \frac{dq}{dt} &= \nabla_p H = p \\ \frac{dp}{dt} &= -\nabla_q H = -\nabla_q V(q) \end{aligned}\right\} \iff \underbrace{\frac{d^2q}{dt^2}}_{\text{Beschleunigung}} = \underbrace{-\nabla_q V(q)}_{\text{Kraftfeld}}.$$

Definition 5.3:
Eine invertierbare Abbildung $F : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ heißt **symplektisch (kanonisch)**, *wenn mit der Jacobi-Matrix $F'(y)$ die Matrixgleichung*

$$F'(y)\,\mathcal{P}\,(F'(y))^T \;=\; \mathcal{P} \quad \forall\, y \in \mathbb{R}^{2n}$$

gilt.

Bemerkung 5.4: *Diese Bedingung läßt sich bequemer als "Invarianz der symplektischen 2-Form $dq \wedge dp$" formulieren: mit*

$$y \;=\; (q,p)^T \;=\; (q_1,\ldots,q_n,p_1,\ldots,p_n)^T \in \mathbb{R}^n \times \mathbb{R}^n$$

und

$$F(y) \;=\; (Q,P)^T \;=\; (Q_1,\ldots,Q_n,P_1,\ldots,P_n)^T \in \mathbb{R}^n \times \mathbb{R}^n$$

ist die Symplektizität von F äquivalent zu

$$dQ \wedge dP \;:=\; \sum_{\alpha=1}^{n} dQ_\alpha \wedge dP_\alpha \;=\; \sum_{\alpha=1}^{n} dq_\alpha \wedge dp_\alpha \;=:\; dq \wedge dp\ .$$

Satz 5.5:
Für jedes Hamilton-System $dy/dt = \mathcal{P}\nabla H$ gilt:

a) H ist ein Erhaltungssatz: $dH(y(t))/dt = 0$,
b) der Fluß F_h ist eine symplektische Abbildung: $F_h'(y)\mathcal{P}(F_h'(y))^T = \mathcal{P}$.

Beweis: a) Mit der Schiefsymmetrie von $\mathcal{P}$ folgt

$$\frac{dH}{dt} \;=\; \langle \nabla H, \frac{dy}{dt}\rangle \;=\; \langle \nabla H, \mathcal{P}\nabla H\rangle \;=\; 0\ .$$

b) Hamiltonische Vektorfelder $f = \mathcal{P}\,\nabla H$ sind durch die Matrixgleichung

$$f'(y)\,\mathcal{P} \;+\; \mathcal{P}\,(f'(y))^T \;=\; 0 \qquad (\#)$$

charakterisiert: der Ausdruck

$$\langle a, f'(y)\mathcal{P}b\rangle = \langle a, \mathcal{P}(\nabla H)'(y)\mathcal{P}b\rangle = -\langle \mathcal{P}a, (\nabla H)'(y)\mathcal{P}b\rangle = -H''(y)[\mathcal{P}a, \mathcal{P}b]$$

ist symmetrisch in den beliebigen Vektoren $a, b \in \mathbb{R}^{2n}$. Damit folgt

$$0 = \langle a, f'(y)\mathcal{P}b\rangle - \langle b, f'(y)\mathcal{P}a\rangle = \langle a, \Big(f'(y)\mathcal{P} + \mathcal{P}(f'(y))^T\Big)b\rangle \quad \forall\, a,b \in \mathbb{R}^{2n}\ .$$

Sei $Y = F_h(y)$. Durch Vertauschung der (partiellen) Ableitungen nach h bzw. nach y (symbolisiert durch $'$) folgt

$$\frac{d}{dh}F_h'(y) \;=\; \Big(\frac{d}{dh}F_h(y)\Big)' \;=\; \Big(f(F_h(y))\Big)' \;=\; f'(Y)Y' \;.$$

Damit folgt für $\Delta(h;y) := F_h'(y)\mathcal{P}(F_h'(y))^T - \mathcal{P} = Y'\mathcal{P}Y'^T - \mathcal{P}$:

$$\frac{d\Delta}{dh} \;=\; f'(Y)Y'\mathcal{P}Y'^T + Y'\mathcal{P}Y'^T(f'(Y))^T \;.$$

Subtraktion von (#) an der Stelle Y liefert die lineare Differentialgleichung

$$\frac{d\Delta}{dh} \;=\; f'(Y)\,\Delta + \Delta\,(f'(Y))^T$$

für Δ, wobei mit $F_0(y) = y$ offensichtlich $\Delta(0;y) = 0$ gilt.Die Lösung dieses AWP für Δ ist $\Delta(h;y) = 0$ für alle h.

Q.E.D.

Bemerkung 5.6: *Es gilt auch folgende Umkehrung von Satz 5.5.b). Ein dynamisches System $dy/dt = f(y)$ ist genau dann (lokal) von der Form $f(y) = \mathcal{P}\,\nabla H$, wenn die Flußabbildungen F_h für alle h symplektisch sind. Verfolgt man den Beweis von Satz 5.5 rückwärt, so folgt aus der Symplektizität des Flusses die Bedingung (#) für das Vektorfeld. Diese besagt aber, daß die Rotation von $\mathcal{P}^{-1}f$ verschwindet, so daß $\mathcal{P}^{-1}f$ ein Gradientfeld mit einem Potential H ist. Also:*

Die Symplektizität des Flusses ist <u>die</u> Eigenschaft von Hamilton-Systemen.

Bemerkung 5.7: *Die Symplektizität des Flusses F_h impliziert gewisse qualitative Eigenschaften. Die offensichtlichste ist die Erhaltung des Phasenraumvolumens (Satz von Liouville): aus $F_h'\mathcal{P}(F_h')^T = \mathcal{P}$ folgt $\det(F_h') = 1$ ($\det(F_h') = -1$ kann nicht auftreten, da diese Größe stetig von h abhängt und $F_0' = \mathbb{1}$ gilt). Mit der Substitutionsregel für Volumintegrale folgt*

$$\int_{F_h(\Omega)} d^{2n}Y \;\stackrel{(Y=F_h(y))}{=}\; \int_\Omega \det(F_h')\;d^{2n}y = \int_\Omega d^{2n}y \;.$$

Damit ist z.B. die Existenz von Attraktoren (Punkte, die eine ganze Umgebung von Startpunkten anziehen) in der Hamilton-Dynamik ausgeschlossen.

Alle qualitativen Eigenschaften von Hamilton-Systemen, die aus der Symplektizität des Flußes folgen, lassen sich leicht auf den numerischen Fluß übertragen:

Definition 5.8:

Ein Integrator $I_h : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ heißt **symplektisch**, *wenn I_h eine symplektische Abbildung ist.*

Man verzichtet hierbei gänzlich auf die Forderung, den skalaren Erhaltungssatz H des Hamilton-Systems auch numerisch zu erhalten.

Satz 5.9: (Sanz-Serna, Lasagni, Suris 1988)

Die von $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ erzeugte s-stufige RK-Abbildung I_h ist symplektisch für alle Hamilton-Systeme $dy/dt = f(y) = \mathcal{P}\nabla H$, wenn die Symplektizitätsbedingung

$$Symp: \quad b_i\, a_{ij} + b_j\, a_{ji} \;=\; b_i\, b_j \;, \quad i,j = 1,\ldots,s$$

aus Hilfssatz 4.78 erfüllt ist.

Beweis: Sei $y = (q,p)^T \in \mathbb{R}^{2n}$ und $f(y) = (Q(q,p), P(q,p))^T$ mit $Q = \nabla_p H$, $P = -\nabla_q H$. Mit den Zwischenstufen

$$q_i = q + h \sum_{j=1}^{s} a_{ij}\, Q_j \;, \quad Q_j := Q(q_j, p_j)$$

$$p_i = p + h \sum_{j=1}^{s} a_{ij}\, P_j \;, \quad P_j := P(q_j, p_j)$$

ist der Zeitschritt $I_h(q,p) = (\tilde{q}, \tilde{p})^T$ durch

$$\tilde{q} \;=\; q \;+\; h \sum_{j=1}^{s} b_j\, Q_j \;, \quad \tilde{p} \;=\; p \;+\; h \sum_{j=1}^{s} b_j\, P_j$$

definiert. Damit gilt

$$dq_j \;=\; dq \;+\; h \sum_{i=1}^{s} a_{ji}\, dQ_i \;, \quad d\tilde{q} \;=\; dq \;+\; h \sum_{i=1}^{s} b_i\, dQ_i$$

$$dp_i \;=\; dp \;+\; h \sum_{j=1}^{s} a_{ij}\, dP_j \;, \quad d\tilde{p} \;=\; dq \;+\; h \sum_{j=1}^{s} b_j\, dP_j$$

und folglich

$$
\begin{aligned}
& d\tilde{q} \wedge d\tilde{p} - dq \wedge dp \\
= \;& \Big(dq + h \sum_{i=1}^{s} b_i\, dQ_i\Big) \wedge \Big(dp + h \sum_{j=1}^{s} b_j\, dP_j\Big) - dq \wedge dp \\
= \;& h \sum_{i=1}^{s} b_i\, dQ_i \wedge dp \;+\; h \sum_{j=1}^{s} b_j\, dq \wedge dP_j \;+\; h^2 \sum_{i,j=1}^{s} b_i\, b_j\, dQ_i \wedge dP_j \\
= \;& h \sum_{i=1}^{s} b_i\, dQ_i \wedge \Big(dp + h \sum_{j=1}^{s} a_{ij}\, dP_j\Big) \;+\; h \sum_{j=1}^{s} b_j \Big(dq + h \sum_{i=1}^{s} a_{ji}\, dQ_i\Big) \wedge dP_j \\
& \qquad + \; h^2 \sum_{i,j=1}^{s} (b_i\, b_j - b_i\, a_{ij} - b_j\, a_{ji}) \; dQ_i \wedge dP_j \\
= \;& h \sum_{i=1}^{s} b_i\, dQ_i \wedge dp_i \;+\; h \sum_{j=1}^{s} b_j\, dq_j \wedge dP_j \\
& \qquad + \; h^2 \sum_{i,j=1}^{s} (b_i\, b_j - b_i\, a_{ij} - b_j\, a_{ji}) \; dQ_i \wedge dP_j \\
= \;& h \sum_{i=1}^{s} b_i \Big(\underbrace{dQ_i \wedge dp_i + dq_i \wedge dP_i}_{\Delta_i}\Big) + h^2 \sum_{i,j=1}^{s} (b_i\, b_j - b_i\, a_{ij} - b_j\, a_{ji}) \; dQ_i \wedge dP_j \; .
\end{aligned}
\tag{\#}
$$

Für Hamilton-Systeme gilt $\Delta_i = 0$: mit

$$
dQ_i = \frac{\partial Q}{\partial q}\Big|_{q_i,p_i} dq_i \;+\; \frac{\partial Q}{\partial p}\Big|_{q_i,p_i} dp_i \;, \quad dP_i = \frac{\partial P}{\partial q}\Big|_{q_i,p_i} dq_i \;+\; \frac{\partial P}{\partial p}\Big|_{q_i,p_i} dp_i
$$

folgt

$$
\Delta_i \;=\; dQ_i \wedge dp_i + dq_i \wedge dP_i \;=
$$

$$
\frac{\partial Q}{\partial q}\Big|_{q_i,p_i} dq_i \wedge dp_i \;+\; \frac{\partial Q}{\partial p}\Big|_{q_i,p_i} \underbrace{dp_i \wedge dp_i}_{0} \;+\; \frac{\partial P}{\partial q}\Big|_{q_i,p_i} \underbrace{dq_i \wedge dq_i}_{0} \;+\; \frac{\partial P}{\partial p}\Big|_{q_i,p_i} dq_i \wedge dp_i
$$

$$
= \;\Big(\frac{\partial Q}{\partial q}\Big|_{q_i,p_i} + \;\frac{\partial P}{\partial p}\Big|_{q_i,p_i}\Big)\; dq_i \wedge dp_i \;.
$$

Für Hamilton-Systeme mit $Q = \nabla_p H \equiv \partial H/\partial p$, $P = -\nabla_q H \equiv -\partial H/\partial q$ ergibt sich

$$
\frac{\partial Q}{\partial q} \;+\; \frac{\partial P}{\partial p} \;=\; \frac{\partial^2 H}{\partial p \partial q} \;-\; \frac{\partial H}{\partial q \partial p} \;=\; 0 \;.
$$

Damit folgt aus (#) die Transformation

$$
d\tilde{q} \wedge d\tilde{p} - dq \wedge dp \;=\; h^2 \sum_{i,j=1}^{s} (b_i\, b_j - b_i\, a_{ij} - b_j\, a_{ji}) \; dQ_i \wedge dP_j
$$

der symplektischen 2-Form unter eine beliebigen RK-Abbildung.

Q.E.D.

Bemerkung 5.10: *Gilt $b_{j_0} = 0$ für ein symplektisches RK-Verfahren, so folgt aus Symp für jedes i mit $b_i \neq 0$:*

$$b_i\, a_{ij_0} \;+\; b_{j_0}\, a_{j_0 i} \;=\; b_i\, b_{j_0} \quad\Longrightarrow\quad a_{ij_0} \;=\; 0\;.$$

Damit taucht die Zwischenstufe j_0 gar nicht in den definierenden Gleichungen der Zwischenstufen i mit $b_i \neq 0$ auf. Da die Zwischenstufe j_0 somit nicht zur Abbildung I_h beiträgt, ist sie redundant und kann aus dem Butcher-Schema gestrichen werden: die Stufenzahl reduziert sich. Man kann also o.B.d.A. bei symplektischen Verfahren annehmen, daß alle $b_i \neq 0$ sind.

Bemerkung 5.11: *Symplektische Verfahren sind notwendigerweise implizit: aus Symp folgt auf der Diagonalen $a_{ii} = b_i/2$. Allerdings kann man fordern, daß der streng obere Dreiecksanteil von (a_{ij}) verschwindet. Aus Symp folgt dann $a_{ij} = b_j$ für $i < j$. Dies führt zu der Klasse der* **diagonalimpliziten symplektischen** *RK-Verfahren der Form*

$$\begin{array}{c|ccccc}
c_1 & b_1/2 & 0 & \dots & \dots & 0 \\
c_2 & b_1 & b_2/2 & \ddots & \ddots & \vdots \\
c_3 & b_1 & b_2 & b_3/2 & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & 0 \\
c_s & b_1 & b_2 & b_3 & \dots & b_s/2 \\
\hline
 & b_1 & b_2 & b_3 & \dots & b_s
\end{array}$$

Ein Zeitschritt I_h läßt sich als Komposition

$$I_h \;=\; \hat{I}_{b_s h} \circ \dots \circ \hat{I}_{b_1 h}\;,$$

von mehreren Schritten der impliziten Mittelpunktsregel $\hat{I}_h$ interpretieren (das 1-stufige Gauß-Legendre-Verfahren 4.84 der Ordnung 2). Eine Implementierung ist damit recht einfach, die Kosten sind (für implizite Verfahren) minimal, da nur Zeitschritte eines 1-stufigen Verfahrens durchzuführen sind. Es existieren diagonalimplizite symplektische Verfahren beliebig hoher Ordnung, allerdings mit sehr hohen Stufenzahlen.

Das 1-stufige Verfahren 2.ter Ordnung ist die implizite Mittelpunktsregel:

$$\begin{array}{c|c}
\frac{1}{2} & \frac{1}{2} \\
\hline
 & 1
\end{array}$$

Die 2-stufigen Verfahren haben auch nur maximal die Ordnung 2. Für 3-stufige Verfahren lassen sich die Ordnungsgleichungen bis zur Ordnung 4 erfüllen:

$$\begin{array}{c|ccc} c_1 & b_1/2 & 0 & 0 \\ c_2 & b_1 & b_2/2 & 0 \\ c_3 & b_1 & b_2 & b_3/2 \\ \hline & b_1 & b_2 & b_3 \end{array} \qquad \begin{aligned} b_1 &= \frac{1}{3}\left(2 + 2^{1/3} + 2^{-1/3}\right) , \\ b_2 &= 1 - 2\,b_1 , \\ b_3 &= b_1 . \end{aligned}$$

Dieses Verfahren erfüllt das Symmetriekriterium in Satz 4.88.b mit der Permutation $\pi(i) = 4-i$ und ist damit symmetrisch. Die diagonalimpliziten symplektischen Verfahren mit den Stufenzahlen $s = 4, 5, 6$ führen auch nur zur maximalen Ordnung 4. Erst mit $s = 7$ Stufen läßt sich Ordnung 6 erreichen. Die Lösungen $b_1, \ldots, b_7$ der Ordnungsgleichungen lassen sich aber nicht mehr in einfacher Form darstellen und sind numerisch zu bestimmen.

Bemerkung 5.12: *Die s-stufigen Gauß-Legendre-Verfahren aus Satz 4.81 haben zusammengefaßt folgende bemerkenswerten qualitativen Eigenschaften:*

+ *sie sind symplektisch (Beweis von Satz 4.81),*
+ *sie erhalten quadratische Erhaltungssätze (Bemerkung 4.80),*
+ *sie sind symmetrisch (Satz 4.94),*
+ *sie sind A-stabil (Folgerung 4.106).*

Weiterhin gilt:

+ *sie haben die maximale Ordnung $2s$,*
− *ein Zeitschritt ist recht teuer, da die Butcher-Matrix vollbesetzt ist.*

Bemerkung 5.13: *Eine vollständige Klassifizierung aller symplektischen symmetrischen RK-Verfahren mit Stufen $s \leq 6$ ist in*

W. Oevel and M. Sofroniou: *Symplectic Runge-Kutta-Schemes II: Classification of Symmetric Methods, preprint 1996*

zu finden.

Kapitel 6

Mehrschrittverfahren

Idee: seien $y_0, \ldots, y_n$ Approximationen der Lösung des AWP $dy/dt = f(y)$, $y(t_0) = y_0$, zu äquidistanten Zeiten $t_i = t_0 + ih$. Bestimme als Zeitschritt die nächste Approximation $y_{n+1} \approx y(t_n + h)$ nicht nur aus y_n, sondern aus mehreren der vorliegenden Stützwerte:

$$y_{n+1} \; = \; I_h(y_n, y_{n-1}, \ldots, y_{n-s+1}) \; .$$

Speziell wird die Klasse der expliziten s-Schrittverfahren

$$\begin{aligned} y_{n+1} \; = \; & - a_{s-1}\, y_n - a_{s-2}\, y_{n-1} - \; \cdots \; - a_0\, y_{n-s+1} \\ & + h \Big(b_{s-1}\, f(y_n) + b_{s-2}\, f(y_{n-1}) + \; \cdots \; + b_0\, f(y_{n-s+1}) \Big) \end{aligned}$$

betrachtet. Bei impliziten s-Schrittverfahren wird y_{n+1} als Lösung von

$$\begin{aligned} y_{n+1} - h\, b_s\, f(y_{n+1}) \; = \; & -a_{s-1}\, y_n - a_{s-2}\, y_{n-1} - \; \cdots \; - a_0\, y_{n-s+1} \\ & + h \Big(b_{s-1}\, f(y_n) + b_{s-2}\, f(y_{n-1}) + \; \cdots \; + b_0\, f(y_{n-s+1}) \Big) \end{aligned}$$

bestimmt. Offensichtlich gilt:

+ man braucht im expliziten Fall pro Zeitschritt $t_n \rightarrow t_{n+1}$ <u>nur eine Funktionsauswertung</u> $f(y_n)$, die anderen benötigten Werte liegen schon aus früheren Zeitschritten gespeichert vor,
+ Fehlerschätzungen sind umsonst, da mehrere simultan ausgeführte Verfahren mit unterschiedlicher Ordnung sich auf dieselben Werte $f(y_n), \ldots, f(y_{n-s+1})$ stützen,
− man braucht s Startwerte $y_0, \ldots, y_{s-1}$ (z.B. durch ein Einschrittverfahren zu erzeugen), um die folgenden Zeitschritte durchführen zu können,

– ein Schrittweitenwechsel ist aufwendig, da die Koeffizienten aus der Annahme $y_n \approx F_h(y_{n-1}) \approx F_{2h}(y_{n-2}) \approx \ldots$ bestimmt werden. Man kann einen Wechsel durch Interpolation durchführen, oder man erzeugt mittels eines Einzelschrittverfahrens mit der neuen Schrittweite die benötigten Stützpunkte.

Definition 6.1:

Ein **lineares s-Schrittverfahren** *zur Lösung von $dy/dt = f(y)$ auf dem $\mathbb{R}^N$ ist eine Abbildung $y_s = I_h(y_{s-1}, \ldots, y_0)$, definiert als Lösung von*

$$\sum_{j=0}^{s} a_j\, y_j \;=\; h \sum_{j=0}^{s} b_j\, f(y_j) \;, \quad a_s = 1 \;.$$

Sie ist explizit für $b_s = 0$.

6.1 Ordnungstheorie

Beispiel 6.2: *Betrachte die* **explizite Mittelpunktsregel (das "leap-frog"-Verfahren)**

$$y_{n+1} \;=\; y_{n-1} \;+\; 2\,h\,f(y_n) \;.$$

Unter der Annahme

$$y_{n-1} \;=\; F_{-h}(y_n) \;=\; y_n - h\,f(y_n) + \frac{h^2}{2}\, f'(y_n)[f(y_n)] \;+\; O(h^3)$$

folgt

$$y_{n+1} \;=\; y_n + h\,f(y_n) + \frac{h^2}{2}\, f'(y_n)[f(y_n)] \;+\; O(h^3) \;=\; F_h(y_n) + O(h^3) \;,$$

d.h., der Zeitschritt $y_n \to y_{n+1}$ ist von 2.ter Ordnung.

Definition 6.3:

Mit dem exakten Fluß F_h von $dy/dt = f(y)$ ist der **lokale Verfahrensfehler** *des Verfahrens 6.1*

$$e(h,y) \;:=\; F_h(y) \;-I_h\Big(y, F_{-h}(y), F_{-2h}(y), \ldots, F_{(s-1)h}(y)\Big)$$

Das Verfahren 6.1 hat die **lokale Konsistenzordnung** *p, wenn $e(h,y) = O(h^{p+1})$ gilt.*

Hilfssatz 6.4:

Das Verfahren 6.1 ist genau dann von der Ordnung p, wenn

$$d(h,y) \;:=\; \sum_{j=0}^{s} a_j\, F_{jh}(y) \;-\; h \sum_{j=0}^{s} b_j\, f(F_{jh}(y)) \;=\; O(h^{p+1})$$

gilt.

Beweis: Sei $y_i := F_{ih}(y)$, $i = 0, \ldots, s$, und $\hat{y} := I_h(y_{s-1}, y_{s-2}, \ldots, y_0)$. Aus

$$\begin{aligned} y_s \;+\; \sum_{j=0}^{s-1} a_j\, y_j \;&=\; h\, b_s\, f(y_s) \;+\; h \sum_{j=0}^{s-1} b_j\, f(y_j) \;+\; d(h,y)\;, \\ \hat{y} \;+\; \sum_{j=0}^{s-1} a_j\, y_j \;&=\; h\, b_s\, f(\hat{y}) \;+\; h \sum_{j=0}^{s-1} b_j\, f(y_j) \end{aligned}$$

folgt mit $\tilde{y} = F_{(s-1)h}(y)$:

$$\begin{aligned} e(h,\tilde{y}) \;=\; y_s - \hat{y} \;&=\; h\, b_s \Big(f(y_s) - f(\hat{y}) \Big) \;+\; d(h,y) \\ &=\; h\, b_s \Big(f(y_s) - f(y_s - e(h,\tilde{y})) \Big) \;+\; d(h,y)\;. \end{aligned}$$

Für $b_s = 0$ folgt die Behauptung mit $e(h,\tilde{y}) = d(h,y)$. Für implizite Verfahren folgt mit der Lipschitz-Konstante L von f

$$\|e(h,\tilde{y})\| \;\le\; |h|\,|b_s|\,L\,\|e(h,\tilde{y})\| \;+\; \|d(h,y)\|\;, \quad \text{d.h.} \quad \|e(h,\tilde{y})\| \;\le\; \frac{\|d(h,y)\|}{1 - |h|\,|b_s|\,L}$$

für hinreichend kleines $|h|$. Analog gilt

$$\|d(h,y)\| = \Big\| e(h,\tilde{y}) - h\, b_s \Big(f(y_s) - f(y_s - e(h,\tilde{y})) \Big) \Big\| \le (1 + |h|\,|b_s|\,L)\, \|e(h,\tilde{y})\|\,,$$

so daß allgemein $O(e(h,\tilde{y})) = O(d(h,y))$ im Limes $h \to 0$ folgt.

Q.E.D.

Satz 6.5: (Ordnungsgleichungen)

Das Verfahren 6.1 ist genau dann von der Ordnung p, wenn

$$\left.\begin{aligned} \sum_{j=0}^{s} a_j \;&=\; 0\;, \\ \sum_{j=1}^{s} j\, a_j \;&=\; \sum_{j=0}^{s} b_j\;, \end{aligned}\right\} \quad \textit{(Konsistenzbedingung)}$$

$$\sum_{j=1}^{s} j^k\, a_j \;=\; k \sum_{j=1}^{s} j^{k-1}\, b_j\;, \quad k = 2, \ldots, p$$

gilt (ein <u>*lineares*</u> *Gleichungssystem für die Parameter* $a_0, \ldots, a_{s-1}$, $b_1, \ldots, b_s$*). Beachte* $a_s = 1$*. Der führende lokale Fehlerterm ist von der Form*

$$F_h(y) - I_h\Big(y, F_{-h}(y), \ldots, F_{(1-s)h}(y)\Big) \;=\; C_{p+1}\, h^{p+1}\, \frac{d^{p+1}y}{dt^{p+1}} \;,$$

wobei $d^{p+1}y/dt^{p+1} \equiv d^{p+1}F_h(y)/dh^{p+1}$ *die Zeitableitung der durch* y *laufenden exakten Lösungskurve ist und*

$$C_{p+1} \;=\; \frac{1}{(p+1)!} \left(\sum_{j=0}^{s} j^{p+1}\, a_j \;-\; (p+1) \sum_{j=0}^{s} j^p\, b_j \right) .$$

Beweis: Mit der Lösung $y(t)$ von $dy/dt = f(y)$ und $y^{(k)} := d^k y/dt^k$ gilt

$$y(t + jh) \;=\; F_{jh}(y(t)) \;=\; \sum_{k=0}^{\infty} \frac{(jh)^k}{k!}\, y^{(k)}(t)$$

sowie

$$f(y(t + jh)) \;=\; \frac{d}{dt}\, y(t + jh) \;=\; \sum_{k=0}^{\infty} \frac{(jh)^k}{k!}\, y^{(k+1)}(t) \;.$$

Mit Hilfssatz 6.4 folgt die Behauptung aus

$$d(h, y(t)) \;=\; \sum_{j=0}^{s} a_j\, F_{jh}(y(t)) \;-\; h \sum_{j=0}^{s} b_j\, f(F_{jh}(y(t)))$$

$$= \Big(\sum_{j=0}^{s} a_j\Big)\, y(t) \;+\; \sum_{k=1}^{\infty} \frac{h^k}{k!} \left(\sum_{j=0}^{s} \Big(a_j\, j^k \;-\; k\, b_j\, j^{k-1}\Big) \right) y^{(k)}(t) \;.$$

Q.E.D.

6.2 Stabilität

Lokale Konsistenz führt bei Mehrschrittverfahren nicht automatisch zur lokalen Konvergenz. Dies wird verständlich, wenn man den trivialen Spezialfall $dy/dt = 0$ betrachtet. Die Stützpunkte sind dann durch

$$\sum_{j=0}^{s} a_j\, y_{n+j} \;=\; 0 \qquad (\#)$$

definiert.

Bemerkung 6.6: *Sei λ eine n-fache Nullstelle des Polynoms*

$$\rho(z) \;=\; a_s\, z^s + a_{s-1}\, z^{s-1} + \cdots + a_0 \;.$$

Man rechnet leicht nach, daß dann $y_k = \lambda^k$, $y_k = k\lambda^k$, …, $y_k = k^{n-1}\lambda^k$ Lösungen der Differenzengleichung (#) sind. Die allgemeine Lösung von (#) ist die Linearkombination

$$y_k \;=\; P_1(k)\,\lambda_1^k \;+\; P_2(k)\,\lambda_2^k \;+\; \cdots \;,$$

wo $\lambda_1, \lambda_2, \ldots$ die paarweise verschiedenen Wurzeln von ρ mit den Vielfachheiten $n_1, n_2, \ldots$ und $P_1, P_2, \ldots$ beliebige Polynome des Grades $n_1 - 1, n_2 - 1, \ldots$ sind. Gibt es eine Wurzel $|\lambda| > 1$, so explodiert die Lösung exponentiell für $k \to \infty$. Eine mehrfache Nullstelle mit $|\lambda| = 1$ führt zur polynomialen Explosion.

Definition 6.7:

Das **charakteristische Polynom** *des Verfahrens 6.1 ist*

$$\rho(z) \;=\; \sum_{j=0}^{s} a_j\, z^j \;.$$

Das Verfahren heißt **stabil**, *wenn die* **Dahlquistsche Wurzelbedingung** *gilt:*

a) $|\lambda| \leq 1$ *für alle Wurzeln von* ρ,

b) $|\lambda| < 1$ *für alle mehrfachen Wurzeln von* ρ.

Bemerkung 6.8: *Bei Konsistenz ist $\lambda = 1$ wegen $\rho(1) = \sum_{j=0}^{s} a_j = 0$ stets eine der Wurzeln des charakteristischen Polynoms.*

Bemerkung 6.9: *Die Stabilitätsforderung ist eine massive Einschränkung an die Koeffizienten des Verfahrens. Es gelten die* **Dahlquist-Schranken** *(siehe z.B. [Hairer, Nørsett & Wanner]): die Konsistenzordnung p eines stabilen linearen s-Schrittverfahrens 6.1 erfüllt*

a) $p \leq s + 2$ *für gerades* s,

b) $p \leq s + 1$ *für ungerades* s,

c) $p \leq s$ *für* $b_s \leq 0$ *(d.h., speziell für explizite Verfahren mit* $b_s = 0$*).*

Eine spezielle Familie stabiler Verfahren ist charakterisiert durch

Satz 6.10:

Alle konsistenten Verfahren mit $a_0 \leq 0, \ldots, a_{s-1} \leq 0, a_s = 1$ sind stabil.

Beweis: Konsistenz impliziert $\sum_{j=0}^{s} a_j = 0$. Für $|z| > 1$ folgt mit der umgekehrten Dreiecksungleichung

$$|\rho(z)| \;=\; |z|^s \left| 1 + \frac{a_{s-1}}{z} + \cdots + \frac{a_0}{z^s} \right| \;\geq\; |z|^s \left(1 - \frac{|a_{s-1}|}{|z|} - \cdots - \frac{|a_0|}{|z^s|} \right)$$

$$> |z|^s \Big(1 - |a_{s-1}| - \cdots - |a_0| \Big) \;=\; |z|^s \Big(1 + a_{s-1} + \cdots + a_0 \Big) \;=\; 0 \;,$$

so daß z nicht Wurzel sein kann. Für $|z| \geq 1$ gilt

$$\begin{aligned} |\rho'(z)| \;&=\; s\,|z|^{s-1} \left| 1 + \frac{s-1}{s}\,\frac{a_{s-1}}{z} + \cdots + \frac{1}{s}\,\frac{a_1}{z^{s-1}} \right| \\ &\geq\; s\,|z|^{s-1} \left(1 - \frac{s-1}{s}\,\frac{|a_{s-1}|}{|z|} - \cdots - \frac{1}{s}\,\frac{|a_1|}{|z^{s-1}|} \right) \\ &>\; s\,|z|^{s-1} \Big(1 - |a_{s-1}| - \cdots - |a_1| \Big) \;=\; s\,|z|^{s-1}\,(-a_0) \;\geq\; 0 \;, \end{aligned}$$

so daß z nicht mehrfache Wurzel sein kann.

Q.E.D.

6.3 Konvergenz

Definition 6.11:

Das Verfahren 6.1

$$y_k \;=\; I_h(y_{k-1}, \ldots, y_{k-s}) \;, \quad k = s, \ldots, n$$

heißt **global konvergent von der Ordnung** p*, wenn mit konstanter Schrittweite* $h = T/n$ *und exakten Startwerten*

$$y_k \;=\; F_{kh}(y_0) \;, \quad k = 0, \ldots, s-1$$

gilt

$$y_n \;-\; F_T(y_0) \;=\; O\Big(\frac{1}{n^p}\Big) \;.$$

Es gilt

$$\boxed{\text{Lokale Konsistenz + Stabilität = Globale Konvergenz.}}$$

Ein allgemeiner Beweis ist technisch aufwendig (siehe z.B. [Deuflhard & Bornemann, Kapitel 7.1.3]. Wir folgen hier [Schwarz] und betrachten nur die Verfahren aus Satz 6.10:

Satz 6.12: (Globale Fehler aus lokalen Fehlern)
Für ein explizites Verfahren 6.1

$$y_k \;=\; I_h(y_{k-1},\ldots,y_{k-s})\;, \quad k=s,\ldots,n$$

der stabilen Klasse 6.10 mit konstanter Schrittweite $h = T/n$ und den Startwerten $y_0,\ldots,y_{s-1}$ gilt für den globalen Fehler zur Zeit T

$$\|y_n - F_T(y_0)\| \;\le\; \left(\underbrace{\max_{j=1..s-1} \|y_j - F_{jh}(y_0)\|}_{Startfehler} \;+\; \frac{\max\limits_{j=s..n}\|e_j\|}{|h|BL} \right) e^{n|h|BL}$$

mit der Lipschitz-Konstanten L des Vektorfeldes f, $B = \sum_{j=0}^{s-1} |b_j|$ und den lokalen Fehlern $e_s,\ldots,e_n$ des Mehrschrittverfahrens.

Beweis: Mit $t_k = t_0 + kh$ und $y(t_k) = F_{kh}(y_0)$ gilt

$$y_k \;=\; \sum_{j=0}^{s-1} \Big(-a_j\, y_{k-s+j} + h\, b_j\, f(y_{k-s+j}) \Big)\;, \quad k=s,s+1,\ldots$$

und

$$y(t_k) \;=\; \sum_{j=0}^{s-1} \Big(-a_j\, y(t_{k-s+j}) + h\, b_j\, f(y(t_{k-s+j})) \Big) \;+\; e_k\;,$$

wobei $e_k = e(h, y(t_{k-1}))$ der lokale Verfahrensfehler aus Definition 6.3 bzw. Hilfssatz 6.4 ist. Für den globalen Fehler $E_k := y(t_k) - y_k$ am Stützpunkt y_k folgt

$$E_k \;=\; \sum_{j=0}^{s-1} \Big(-a_j\, E_{k-s+j} + h\, b_j \Big(f(y(t_{k-s+j})) - f(y_{k-s+j}) \Big) \Big) \;+\; e_k$$

und damit

$$\|E_k\| \;\le\; \sum_{j=0}^{s-1} (-a_j + |h||b_j|L)\, \|E_{k-s+j}\| \;+\; \|e_k\|\;.$$

Mit $c_j := -a_j + |h||b_j|L$ erhält man die rekursive Fehlerfortpflanzung

$$\|E_k\| \;\le\; \sum_{j=0}^{s-1} c_j\, \|E_{k-s+j}\| \;+\; \|e_k\|\;, \quad k=s,s+1,\ldots\;.$$

Mit $c := \sum_{j=0}^{s-1} c_j = \sum_{j=0}^{s-1} (-a_j + |h|\,|b_j|\,L) = 1 + |h|BL$ und

$$\tilde{E}_k \;:=\; \max(\|E_1\|, \|E_2\| \ldots, \|E_k\|)$$

folgt die vereinfachte Rekursion

$$\tilde{E}_k \;\leq\; c\,\tilde{E}_{k-1} + \|e_k\|\;, \quad k = s, s+1, \ldots\;.$$

Hieraus ergibt sich die Fortpflanzung des Startfehlers $\tilde{E}_{s-1}$ und der lokalen Fehler $e_s, \ldots, e_n$ des Mehrschrittverfahrens:

$$\begin{aligned}\tilde{E}_n \;&\leq\; c^{n-s+1}\,\tilde{E}_{s-1} \;+\; c^{n-s}\,\|e_s\| \;+\; c^{n-s-1}\,\|e_{s+1}\| \;+\; \cdots \;+\; \|e_n\|\\ &\leq\; c^{n-s+1}\,\tilde{E}_{s-1} \;+\; (c^{n-s} + c^{n-s-1} + \cdots + 1)\,\max_{j=s..n}\|e_j\|\;.\end{aligned}$$

Mit

$$c^{n-s+1} \;\leq\; c^n \;\leq\; (1 + |h|BL)^n \;\leq\; e^{n|h|BL}$$

und

$$\begin{aligned}1 \;+\; c \;+\; \cdots \;+\; c^{n-s} \;&=\; \frac{c^{n-s+1} - 1}{c - 1} \;\leq\; \frac{c^n - 1}{c - 1}\\ &=\; \frac{(1 + |h|BL)^n - 1}{|h|BL} \;\leq\; \frac{e^{n|h|BL} - 1}{|h|BL} \;\leq\; \frac{e^{n|h|BL}}{|h|BL}\end{aligned}$$

folgt die Behauptung

$$\tilde{E}_n \;\leq\; \left(\tilde{E}_{s-1} \;+\; \frac{\max_{j=s..n}\|e_j\|}{|h|BL}\right) e^{n|h|BL}\;.$$

Q.E.D.

Folgerung 6.13: *Berechnet man aus y_0 die Startwerte $y_1, \ldots, y_{s-1}$ mit einem Einschrittverfahren der Ordnung $p-1$ und hat das s-Schrittverfahren die lokale Konsistenzordnung p, so folgt die globale Konvergenzordnung p. Also: man muß auch die Startwerte in hinreichend hoher Ordnung approximieren.*

6.4 Konkrete Verfahren

6.4.1 Adams-Bashforth-Methoden

Die expliziten s-Schrittverfahren vom **Adams-Bashforth-Typ** sind von der Form

$$AB_s: \qquad y_{n+1} \;=\; y_n \;+\; h\Big(\,b_{s-1}\,f(y_n) + \cdots + b_0\,f(y_{n-s+1})\,\Big)\;.$$

$$\begin{aligned}
AB_1: \quad y_{n+1} &= y_n + h\,f(y_n) \quad \text{(Euler-Verfahren)}\,,\\
AB_2: \quad y_{n+1} &= y_n + \frac{h}{2}\Big(\,3\,f(y_n) - f(y_{n-1})\,\Big)\,,\\
AB_3: \quad y_{n+1} &= y_n + \frac{h}{12}\Big(\,23\,f(y_n) - 16\,f(y_{n-1}) + 5\,f(y_{n-2})\,\Big)\,,\\
AB_4: \quad y_{n+1} &= y_n + \frac{h}{24}\Big(\,55\,f(y_n) - 59\,f(y_{n-1}) + 37\,f(y_{n-2}) - 9\,f(y_{n-3})\,\Big)\,,\\
AB_5: \quad y_{n+1} &= y_n + \frac{h}{720}\Big(\,1901\,f(y_n) - 2774\,f(y_{n-1}) + 2616\,f(y_{n-2})\\
&\qquad -1274\,f(y_{n-3}) + 251\,f(y_{n-4})\,\Big)\,,\\
AB_6: \quad y_{n+1} &= y_n + \frac{h}{1440}\Big(\,4277\,f(y_n) - 7923\,f(y_{n-1}) + 9982\,f(y_{n-2})\\
&\qquad -7298\,f(y_{n-3}) + 2877\,f(y_{n-4}) - 475\,f(y_{n-5})\,\Big)\,.
\end{aligned}$$

Tafel 6.1: Adams-Bashforth-Verfahren

Das charakteristische Polynom $\rho(z) = z^{s-1}(z-1)$ erfüllt die Wurzelbedingung 6.7, so daß diese Verfahren stabil sind. Mit den s Parametern $b_0, \ldots, b_{s-1}$ lassen sich die Ordnungsgleichungen aus Satz 6.5 bis $p = s$ erfüllen (dies ist nach Bemerkung 6.9.c die maximal erreichbare Ordnung). Man erhält die in Tafel 6.1 angegebenen Verfahren. Durch simultane Ausführung der Verfahren AB_s und AB_{s+1} ergibt sich eine Schätzung des Verfahrensfehlers für AB_s. Die führenden Fehlerkoeffizienten gemäß Satz 6.5 sind

	AB_1	AB_2	AB_3	AB_4	AB_5	AB_6
Ordnung $p = s$	1	2	3	4	5	6
C_{p+1}	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$
$\approx$	0.5	0.417	0.375	0.349	0.330	0.316

6.4.2 Nyström-Methoden

Die expliziten s-Schrittverfahren vom **Nyström-Typ** sind von der Form

$$N_s: \qquad y_{n+1} = y_{n-1} + h\Big(\,b_{s-1}\,f(y_n) + \cdots + b_0\,f(y_{n-s+1})\,\Big)\,.$$

Das charakteristische Polynom $\rho(z) = z^{s-2}(z^2-1)$ erfüllt die Wurzelbedingung 6.7, so daß diese Verfahren stabil sind. Mit den s Parametern $b_0, \ldots, b_{s-1}$ lassen sich die Ordnungsgleichungen aus Satz 6.5 bis $p = s$ erfüllen, man erhält

$$
\begin{aligned}
N_2: \quad y_{n+1} &= y_{n-1} + 2\,h\,f(y_n) \quad \text{(leap-frog)}\,, \\
N_3: \quad y_{n+1} &= y_{n-1} + \frac{h}{3}\Big(\,7\,f(y_n) - 2\,f(y_{n-1}) + f(y_{n-2})\,\Big)\,, \\
N_4: \quad y_{n+1} &= y_{n-1} + \frac{h}{3}\Big(\,8\,f(y_n) - 5\,f(y_{n-1}) + 4\,f(y_{n-2}) - f(y_{n-3})\,\Big)\,, \\
N_5: \quad y_{n+1} &= y_{n-1} + \frac{h}{90}\Big(\,269\,f(y_n) - 266\,f(y_{n-1}) + 294\,f(y_{n-2}) \\
&\qquad\qquad -146\,f(y_{n-3}) + 29\,f(y_{n-4})\,\Big)\,, \\
N_6: \quad y_{n+1} &= y_{n-1} + \frac{h}{90}\Big(\,297\,f(y_n) - 406\,f(y_{n-1}) + 574\,f(y_{n-2}) \\
&\qquad\qquad -426\,f(y_{n-3}) + 169\,f(y_{n-4}) - 28\,f(y_{n-5})\,\Big)\,.
\end{aligned}
$$

Tafel 6.2: Nyström-Verfahren

die in Tafel 6.2 aufgelisteten Verfahren. Die führenden Fehlerkoeffizienten gemäß Satz 6.5 sind

	N_2	N_3	N_4	N_5	N_6
Ordnung $p = s$	2	3	4	5	6
C_{p+1}	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{29}{90}$	$\frac{14}{45}$	$\frac{1139}{3780}$
$\approx$	0.333	0.333	0.322	0.311	0.301

6.4.3 Adams-Moulton-Methoden

Die impliziten s-Schrittverfahren vom **Adams-Moulton-Typ** sind von der Form

$$AM_s: \qquad y_{n+1} \;=\; y_n \;+\; h\Big(\,b_s\,f(y_{n+1}) + b_{s-1}\,f(y_n) + \cdots + b_0\,f(y_{n-s+1})\,\Big)\,.$$

Das charakteristische Polynom $\rho(z) = z^{s-1}(z-1)$ erfüllt die Wurzelbedingung 6.7, so daß diese Verfahren stabil sind. Mit den s Parametern $b_0, \ldots, b_s$ lassen sich die Ordnungsgleichungen aus Satz 6.5 bis $p = s + 1$ erfüllen, man erhält die in Tafel 6.3 angegebenen Verfahren. Die führenden Fehlerkoeffizienten

$$
\begin{aligned}
AM_1: \quad y_{n+1} &= y_n + \frac{h}{2}\Big(\, f(y_{n+1}) + f(y_n) \Big) \quad \text{(implizite Trapezregel)}\ ,\\
AM_2: \quad y_{n+1} &= y_n + \frac{h}{12}\Big(\, 5\, f(y_{n+1}) + 8\, f(y_n) - f(y_{n-1}) \Big)\ ,\\
AM_3: \quad y_{n+1} &= y_n + \frac{h}{24}\Big(\, 9\, f(y_{n+1}) + 19\, f(y_n) - 5\, f(y_{n-1}) + f(y_{n-2}) \Big)\ ,\\
AM_4: \quad y_{n+1} &= y_n + \frac{h}{720}\Big(\, 251\, f(y_{n+1}) + 646\, f(y_n) - 264\, f(y_{n-1})\\
&\qquad +106\, f(y_{n-2}) - 19\, f(y_{n-3}) \Big)\ ,\\
AM_5: \quad y_{n+1} &= y_n + \frac{h}{1440}\Big(\, 475\, f(y_{n+1}) + 1427\, f(y_n) - 798\, f(y_{n-1})\\
&\qquad +482\, f(y_{n-2}) - 173\, f(y_{n-3}) + 27\, f(y_{n-4}) \Big)\ ,\\
AM_6: \quad y_{n+1} &= y_n + \frac{h}{60480}\Big(\, 19087\, f(y_{n+1}) + 65112\, f(y_n) - 46461\, f(y_{n-1})\\
&\qquad +37504\, f(y_{n-2}) - 20211\, f(y_{n-3}) + 6312\, f(y_{n-4})\\
&\qquad -863\, f(y_{n-5}) \Big)\ .
\end{aligned}
$$

Tafel 6.3: Adams-Moulton-Verfahren

gemäß Satz 6.5 sind

	AM_1	AM_2	AM_3	AM_4	AM_5	AM_6
Ordnung $p = s + 1$	2	3	4	5	6	7
C_{p+1}	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$	$-\frac{275}{24192}$
$\approx$	-0.0833	-0.0417	-0.0263	-0.0188	-0.0143	-0.0114

Der wesentliche Unterschied zu den AB-Methoden sind die erheblich kleineren Fehlerkoeffizienten C_{p+1}. Allerdings muß für den implizit definierten Zeitschritt eine Gleichung gelöst werden, was i.a. durch ein Newton-Verfahren oder eine Fixpunktiteration geschehen kann. Ein geringer Rechenaufwand ist dabei erforderlich, wenn man eine Approximation $y_{n+1}^{(p)}$ ("Prädiktor") durch ein explizites Verfahren erzeugt, so daß mit diesem Startwert ein einziger Schritt der Fixpunktiteration reicht, um die Ordnung zu erhalten. Wählt man ein AB-Verfahren zur Berechnung der Approximation $y_{n+1}^{(p)}$, so erhält man die expliziten Adams-Bashforth-Moulton-Methoden.

6.4.4 Adams-Bashforth-Moulton-Methoden

In den expliziten Verfahren vom **Adams-Bashforth-Moulton-Typ** wird zunächst das Adams-Bashforth-Verfahren AM_{s+1} verwendet, um eine Approximation $y_{n+1}^{(p)}$ (Prädiktor) des Zeitschritts zu finden. Dieser wird in die rechte Seite des Adams-Moulton-Verfahrens AM_s eingesetzt (Korrektorschritt), was einem Schritt der Fixpunktiteration zur Lösung des AM_s-Schritts mit dem Startwert $y_{n+1}^{(p)}$ entspricht:

$$
\begin{aligned}
&AB_{s+1}M_s: \\
&\quad AB_{s+1}: \; y_{n+1}^{(p)} = y_n + h\left(b_s\, f(y_n) + \cdots + b_0\, f(y_{n-s})\right), \\
&\quad AM_s: \; y_{n+1} = y_n + h\,\tilde{b}_s\, f(y_{n+1}^{(p)}) + h\left(\tilde{b}_{s-1}\, f(y_n) + \cdots + \tilde{b}_0\, f(y_{n-s+1})\right).
\end{aligned}
$$

Die Ordnung und der führende Fehlerkoeffizient dieser Kombination sollen bestimmt werden. Für das Ergebnis $\tilde{y}_{n+1}$ eines impliziten AM_s-Schrittes der Ordnung $s+1$, definiert durch

$$
\tilde{y}_{n+1} = y_n + h\,\tilde{b}_s\, f(\tilde{y}_{n+1}) + h\left(\tilde{b}_{s-1}\, f(y_n) + \cdots + \tilde{b}_0\, f(y_{n-s+1})\right), \qquad (\#)
$$

gilt

$$
F_h(y_n) - \tilde{y}_{n+1} \;=\; C_{s+2}^{AM}\, h^{s+2}\, \frac{d^{s+2}y}{dt^{s+2}} \;+\; O(h^{s+3}) \qquad (\#\#)
$$

mit dem Fehlerkoeffizienten C_{s+2}^{AM} von AM_s. Für den Prädiktorschritt AB_{s+1} mit der Ordnung $s+1$ gilt $F_h(y_n) - y_{n+1}^{(p)} = O(h^{s+2})$. Es folgt

$$
\tilde{y}_{n+1} - y_{n+1}^{(p)} \;=\; O(h^{s+2}) \,.
$$

Für den durch

$$
y_{n+1} \;=\; y_n + h\,\tilde{b}_s\, f(y_{n+1}^{(p)}) \;+\; h\Big(\tilde{b}_{s-1}\, f(y_n) + \cdots + \tilde{b}_0\, f(y_{n-s+1})\Big)
$$

definierten Korrektorschritt folgt durch Differenzbildung mit (#) bei Lipschitzstetigem f:

$$
\begin{aligned}
\tilde{y}_{n+1} - y_{n+1} \;&=\; h\,\tilde{b}_s\left(f(\tilde{y}_{n+1}) - f(y_{n+1}^{(p)})\right) \\
&=\; h\,\tilde{b}_s\, O\Big(\tilde{y}_{n+1} - y_{n+1}^{(p)}\Big) \;=\; O(h^{s+3}) \,.
\end{aligned}
$$

Für den Schritt $AB_{s+1}M_s$ ergibt sich hieraus zusammen mit (##):

$$
\begin{aligned}
F_h(y_n) - y_{n+1} \;&=\; F_h(y_n) - \tilde{y}_{n+1} \quad + \quad \tilde{y}_{n+1} - y_{n+1} \\
&=\; C_{s+2}^{AM}\, h^{s+2}\, \frac{d^{s+2}y}{dt^{s+2}} \;+\; O(h^{s+3}) \,.
\end{aligned}
$$

Das $AB_{s+1}M_s$-Verfahren erbt damit die Ordnung $s+1$ und den führenden Fehlerkoeffizienten des impliziten Korrektorschritts AM_s. Dieselbe Ordnung kann man auch erhalten, wenn man den Prädiktor mit AB_s statt mit AB_{s+1} berechnet. Der resultierende Fehlerkoeffizient ist dann aber eine Kombination von C^{AM}_{s+2} mit dem wesentlich größeren Fehlerkoeffizienten des AB_s-Prädiktorschrittes, verstärkt durch die Lipschitz-Konstante von f.
Die angegebene Version $AB_{s+1}M_s$ der Ordnung $s+1$ verbindet die kleinen Fehlerkoeffizienten der impliziten AM_s-Verfahren mit der kostengünstigen Durchführung eines expliziten Zeitschritts (eine neue f-Auswertung bei s gespeicherten früheren f-Werten für AB_{s+1}, eine zusätzliche Auswertung $f(y^{(p)}_{n+1})$ im Korrektorschritt AM_s).
In der Praxis sind die ABM-Verfahren wie z.B.

$$AB_3M_2 : \ \text{(Ordnung 3)}$$

$$\begin{aligned}
AB_3 :\ & y^{(p)}_{n+1} = y_n + \frac{h}{12}\Big(23\,f(y_n) - 16\,f(y_{n-1}) + 5\,f(y_{n-2})\Big)\,,\\
AM_2 :\ & y_{n+1} = y_n + \frac{h}{12}\Big(5\,f(y^{(p)}_{n+1}) + 8\,f(y_n) - f(y_{n-1})\Big)
\end{aligned}$$

oder

$$AB_4M_3 : \ \text{(Ordnung 4)}$$

$$\begin{aligned}
AB_4 :\ & y^{(p)}_{n+1} = y_n + \frac{h}{24}\Big(55\,f(y_n) - 59\,f(y_{n-1}) + 37\,f(y_{n-2}) - 9\,f(y_{n-3})\Big)\,,\\
AM_3 :\ & y_{n+1} = y_n + \frac{h}{24}\Big(9\,f(y^{(p)}_{n+1}) + 19\,f(y_n) - 5\,f(y_{n-1}) + f(y_{n-2})\Big)
\end{aligned}$$

oder

$$AB_5M_4 : \ \text{(Ordnung 5)}$$

$$\begin{aligned}
AB_5 :\ & y^{(p)}_{n+1} = y_n + \frac{h}{720}\Big(1901\,f(y_n) - 2774\,f(y_{n-1}) + 2616\,f(y_{n-2})\\
& \qquad -1274\,f(y_{n-3}) + 251\,f(y_{n-4})\Big)\,,\\
AM_4 :\ & y_{n+1} = y_n + \frac{h}{720}\Big(251\,f(y^{(p)}_{n+1}) + 646\,f(y_n) - 264\,f(y_{n-1})\\
& \qquad +106\,f(y_{n-2}) - 19\,f(y_{n-3})\Big)
\end{aligned}$$

häufig eingesetzte und für nichtsteife Systeme recht effiziente Verfahren.

6.4.5 Steife Systeme, BDF-Methoden

Betrachte das skalare lineare Testproblem $dy/dt = \lambda\, y$, $y(t), \lambda \in \mathbb{C}$. Bei negativem Realteil $\Re e(\lambda) < 0$ gilt **asymptotische Stabilität**:

$$\lim_{t\to\infty} y(t) \;=\; \lim_{t\to\infty} y(t_0)\, e^{\lambda\,(t-t_0)} \;=\; 0$$

für alle Startwerte $y(t_0)$. Ein lineares s-Schrittverfahren 6.1 heißt in Analogie zu den Einschrittverfahren (Kapitel 4.7) A-stabil, wenn mit (fehlerbehafteten, also beliebigen) Startwerten $y_0, \ldots, y_{s-1}$ und beliebigen konstanten Schrittweiten h auch für die numerische Lösung $\lim\limits_{n\to\infty} y_n = 0$ gilt. Die Stützwerte sind durch

$$\sum_{j=0}^{s} a_j\, y_{n+j} \;=\; h \sum_{j=0}^{s} b_j\, f(y_{n+j}) \;=\; h\,\lambda \sum_{j=0}^{s} b_j\, y_{n+j} \;,$$

also als Lösung der Differenzengleichung

$$\sum_{j=0}^{s} (a_j - \mu\, b_j)\, y_{n+j} \;=\; 0 \;, \quad \mu := h\,\lambda$$

definiert. Mit Bemerkung 6.6 gilt $y_n \to 0$ für beliebiges Startwerte genau dann, wenn $|z| < 1$ gilt für alle Wurzeln z des Polynoms

$$\rho_\mu(z) \;:=\; \sum_{j=0}^{s} (a_j - \mu\, b_j)\, z^j \;.$$

Der **Bereich absoluter Stabilität** wird damit definiert als

$$S \;:=\; \{\; \mu \in \mathbb{C} \;;\; |z| < 1 \text{ für alle } z \in \mathbb{C} \text{ mit } \rho_\mu(z) = 0 \;\} \;.$$

Da die Wurzeln von ρ_μ stetig von μ abhängen, impliziert die Dahlquistsche Wurzelbedingung 6.7.a, daß $\mu = 0$ auf dem Rand von S liegen muß.

Das Verfahren heißt **A-stabil**, wenn S die gesamte linke Halbebene

$$C_- := \{\mu \in \mathbb{C}\,;\, \Re e(\mu) < 0\}$$

enthält. Bei asymptotisch stabilem Testproblem mit $\mu = h\lambda \in \mathbb{C}_- \subset S$ folgt dann $\lim\limits_{n\to\infty} y_n = 0$ für beliebige Schrittweiten h.

Leider existieren keine A-stabilen linearen s-Schrittverfahren hoher Ordnung. Es gilt die **2.te Dahlquist-Schranke** (siehe z.B. [Deuflhard & Bornemann, Satz 7.36]):

Für ein A-stabiles lineares s-Schrittverfahren 6.1 der Ordnung p gilt $p \le 2$.

Dies ist kein Widerspruch zur Existenz A-stabiler RK-Verfahren höherer Ordnung aus Kapitel 4.7: diese Verfahren sind nicht von dem einfachen Typ 6.1. Für steife Systeme geeignete lineare Mehrschrittverfahren höherer Ordnung sollten, wenn sie schon nicht A-stabil sein können, zumindestens große Stabilitätsbereiche haben. Die impliziten AM-Verfahren haben zwar im Vergleich zu den expliziten AB-Methoden größere Stabilitätsbereiche, diese sind aber immer noch recht klein. Die folgende Klasse von impliziten s-Schrittverfahren vom *BDF*-Typ (backward differentiation formula)

$$a_s\, y_{n+1} \ + \ a_{s-1}\, y_n \ + \ \cdots \ + \ a_0\, y_{n-s+1} \ = \ h\, f(y_{n+1})$$

sind für $s = 1, \ldots, 6$ wesentlich besser für steife Systeme geeignet als die AM-Methoden. Die Ordnungsgleichungen aus Satz 6.5 (diesmal mit der Normierung $b_s = 1$ statt $a_s = 1$) sind bis zur Ordnung $p = s$ lösbar, man findet:

$$\begin{aligned}
&\mathit{BDF}_1: \quad y_{n+1} - y_n \ = \ h\ f(y_{n+1}) \quad \text{(implizites Euler-Verfahren, } A\text{-stabil)}\ , \\
&\mathit{BDF}_2: \quad \frac{3}{2}\, y_{n+1} - 2\, y_n + \frac{1}{2}\, y_{n-1} \ = \ h\ f(y_{n+1})\ , \\
&\mathit{BDF}_3: \quad \frac{11}{6}\, y_{n+1} - 3\, y_n + \frac{3}{2}\, y_{n-1} - \frac{1}{3}\, y_{n-2} \ = \ h\ f(y_{n+1})\ , \\
&\mathit{BDF}_4: \quad \frac{25}{12}\, y_{n+1} - 4\, y_n + 3\, y_{n-1} - \frac{4}{3}\, y_{n-2} + \frac{1}{4}\, y_{n-3} \ = \ h\ f(y_{n+1})\ , \\
&\mathit{BDF}_5: \quad \frac{137}{60}\, y_{n+1} - 5\, y_n + 5\, y_{n-1} - \frac{10}{3}\, y_{n-2} + \frac{5}{4}\, y_{n-3} - \frac{1}{5}\, y_{n-3} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \ h\ f(y_{n+1})\ , \\
&\mathit{BDF}_6: \quad \frac{49}{20}\, y_{n+1} - 6\, y_n + \frac{15}{2}\, y_{n-1} - \frac{20}{3}\, y_{n-2} + \frac{15}{4}\, y_{n-3} - \frac{6}{5}\, y_{n-4} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \frac{1}{6}\, y_{n-5} \ = \ h\ f(y_{n+1})\ .
\end{aligned}$$

Ihr Stabilitätsbereich S umfaßt einen großen Teil der linken Halbebene C_- (siehe z.B. [Deuflhard & Bornemann], [Schwarz, Figur 9.9]). Allerdings wird $C_- \setminus S$ mit zunehmendem s (d.h., mit zunehmender Ordnung $p = s$) größer. Für *BDF*-Verfahren mit $s > 6$ liegt nicht einmal mehr der Nullpunkt von $\mathbb{C}$ auf dem Rand von S, womit die Dahlquistsche Wurzelbedingung 6.7 verletzt ist und die entsprechenden *BDF*-Verfahren unbrauchbar sind.