

# MOSIM

## Modellierung und Simulation

2. Juni 2016



---

## Vorwort

Die vorliegende Ausarbeitung ist der Versuch einer Ausformulierung der Vorlesungen MOSIM/I und MOSIM/II, gehalten von Prof. Dr. Wensch im Wintersemester 2015/2016 und Sommersemester 2016 an der Technischen Universität Dresden. Die vorliegende Ausarbeitung wurde erweitert durch Beiträge aus [Strehmel et al. \(2012\)](#) und [Hairer et al. \(1993\)](#) und enthält noch orthografische und grammatikalische Fehler, ferner sind Abbildungen noch nachzutragen.

Dresden, 2. Juni 2016.



---

# Inhaltsverzeichnis

---

## Teil I Nichtsteife Differentialgleichungen

---

<b>1</b>	<b>Theoretische Grundlagen</b>	3
1.1	Problemstellung	3
1.2	Existenz, Eindeutigkeit und Sensitivität	5
1.3	Notwendige Bezeichnungen und Konzepte	7
<b>2</b>	<b>Modellierung mittels gewöhnlicher Differentialgleichungen</b>	11
2.1	Grundlagen	11
2.2	Reaktionskinetik	11
2.3	Populationsgenetik	12
2.4	Mechanik	13
2.4.1	Mehrkörpersysteme in Zustandsform	13
2.4.2	Mehrkörpersysteme in Deskriptorform	15
2.4.3	Hamilton-Systeme	16
<b>3</b>	<b>Einschrittverfahren</b>	19
3.1	Grundlagen	19
3.2	Explizite Runge/Kutta-Verfahren	20
3.3	Ordnungsaussagen und B-Reihen für Runge/Kutta-Verfahren	23
3.3.1	Taylor-Entwicklung der exakten Lösungen	24
3.3.2	Taylor-Entwicklung numerischer Lösungen	27
3.3.3	Ordnungsbedingungen für Runge/Kutta-Verfahren	31
3.4	Konstruktion expliziter Runge/Kutta-Verfahren	32
3.4.1	Verfahren bis zur Ordnung $p = 3$	32
3.4.2	Verfahren höherer Ordnung	34

3.5	Konvergenz	38
3.6	Schrittweitensteuerung	41
3.6.1	Grundprinzip	41
3.6.2	Fehlerschätzung mittels Richardson-Extrapolation	44
3.6.3	Fehlerschätzung mittels Einbettung	46
3.7	Stetige explizit Runge/Kutta-Verfahren	48
<b>4</b>	<b>Extrapolationsverfahren</b>	<b>51</b>
4.1	Asymptotische Entwicklung des globalen Fehlers	51
4.2	Extrapolationsschritte	54
4.3	Implementierung	56
4.4	Extrapolation mit symmetrischen Verfahren	56
4.4.1	Gespiegelte Verfahren	57
4.4.2	Symmetrische Verfahren	59
4.4.3	Gragg/Bulirsch/Stoer-Verfahren	60
<b>5</b>	<b>Qualitative Analyse von Differentialgleichungsmodellen</b>	<b>63</b>
5.1	Stabilität von Fixpunkten	63
5.2	Phasenlinien	65
5.2.1	Stabilität von Fixpunkten	65
5.3	Stabilität im $\mathbb{R}^n$	66
5.4	1-dimensionale Populationsmodelle	67
5.5	Interagierende Populationsmodelle	69
5.6	Bifurkation	74

---

## Teil II Steife Differentialgleichungen

---

<b>6</b>	<b>Theoretische Grundlagen steifer Probleme</b>	<b>81</b>
6.1	Stabilität von Differentialgleichungen	81
6.2	Einseitige Lipschitz-Bedingung	84
6.3	Steife Differentialgleichungen	85
6.4	Auftreten steifer Systeme	86
6.5	Numerisches Fehlverhalten	92

<b>7</b>	<b>Implizite Runge/Kutta-Verfahren</b>	1
7.1	Verfahrensvorschrift und Ordnung	1
7.2	Vereinfachende Bedingungen	3
7.3	Implizite Runge/Kutta-Verfahren höherer Ordnung	9
7.3.1	Verfahren maximaler Ordnung, Gauß-Verfahren	14
7.3.2	Radau-Verfahren	16
7.3.3	Lobatto-Verfahren	19
	<b>Sachverzeichnis</b>	21
	<b>Literaturverzeichnis</b>	23





## Nichtsteife Differentialgleichungen



# Theoretische Grundlagen

## 1.1 Problemstellung

Ausgangspunkt der Untersuchungen sind Anfangswertprobleme für gewöhnliche Differentialgleichung erster Ordnung der Form

$$y' = \mathbf{f}(t, y(t)), \quad y(t_0) = y_0, \quad (1.1)$$

dabei sind  $y'(t)$  die Ableitung der Funktion  $y : I \rightarrow \mathbb{R}^n$  nach  $t$  und  $f : I \times \Omega \rightarrow \mathbb{R}^n$  wird als *rechte Seite* bezeichnet, wobei  $I \subset \mathbb{R}$  als (Zeit-) *Intervall* und  $\Omega \subset \mathbb{R}^n$  als *Zustands-* oder *Phasenraum* heißen; das Gebiet  $\Omega_0 := I \times \Omega$  heißt *erweiterter Phasenraum*.

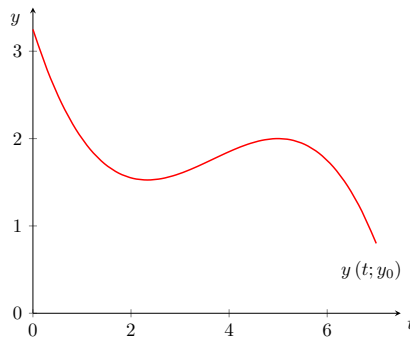


Abb. 1.1: Erweiterter Phasenraum

Die vektorwertige Differentialgleichung  $y(t)$  bildet das *nichtautonom* System

$$\begin{aligned} y'_1(t) &= f_1(t, y_1(t), \dots, y_n(t)) & y_1(t_0) &= y_{0,1}, \\ y'_2(t) &= f_2(t, y_1(t), \dots, y_n(t)) & y_2(t_0) &= y_{0,2}, \\ &\vdots & &\vdots \\ y'_n(t) &= f_n(t, y_1(t), \dots, y_n(t)) & y_n(t_0) &= y_{0,n} \end{aligned} \quad (1.2)$$

*Beispiel 1.1.1 (Räuber/Beute-Modell).* Für die Beutepopulation steht in unbegrenzten Umfang Nahrung zur Verfügung. Bei Abwesenheit von Räubern hängt die Größe  $y_2(t)$  der Beutepopulation lediglich von der Geburtenüberschussrate  $\alpha > 0$  ab, so dass

$$y_1' = \alpha y_1. \quad (1.3)$$

Bei der Anwesenheit von Räubern verringert sich die Beutepopulation in Abhängigkeit von der Kontaktrate  $\beta > 0$  zwischen den Begegnungen  $y_1(t) y_2(t)$  von Beute- und Räubern, so dass (1.3) erweitert wird zu

$$y_1' = \alpha y_1 - \beta y_1 y_2. \quad (1.4)$$

Die Räuberpopulation ernährt sich nun ausschließlich von Beutetieren. Bei Abwesenheit von Beute übertrifft die Todesrate die Geburtenrate, so dass die Größe  $y_2(t)$  der Räuberpopulation mit der Rate  $\gamma > 0$  abnimmt, d.h. es gilt

$$y_2' = -\gamma y_2. \quad (1.5)$$

Bei Anwesenheit von Beute wächst die Räuberpopulation proportional  $\delta > 0$  zu den Begegnung  $y_1(t) y_2(t)$  zwischen Beute und Räubern, so dass 1.5 erweitert wird zu

$$y_2' = -\gamma y_2 + \delta y_1 y_2. \quad (1.6)$$

Fasst man (1.4) und (1.6) zusammen, dann erhält man das System

$$y_1' = y_1 (\alpha - \beta y_2(t)), \quad y_2' = y_2 (\delta y_1 - \gamma) \quad (1.7)$$

In manchen Fällen werden an Stelle des Standardproblems (1.1) zwei Sonderformen gewöhnlicher Differentialgleichungen betrachtet: Bei *autonomen* Anfangswertproblemen

$$y' = f(y), \quad y(t_0) = y_0 \quad (1.8)$$

ist die rechte Seite nicht von der Variable  $t$  abhängig. Ferner kann jedes nichtautonome Anfangswertproblem (1.1) mittels der Transformation

$$z(t) = \begin{pmatrix} y(t) \\ t \end{pmatrix}$$

in das autonome System

$$z' = \begin{pmatrix} f(t, y(t)) \\ 1 \end{pmatrix} = g(z), \quad z(t_0) = \begin{pmatrix} y_0 \\ t_0 \end{pmatrix} \quad (1.9)$$

überführt werden.

In manchen Anwendungen treten aber auch Anfangswertprobleme *gewöhnlicher Differentialgleichungen  $m$ -ter Ordnung*

$$y^{(m)} = f\left(t, y(t), y'(t), \dots, y^{(m-1)}(t)\right), \quad y^{(i)}(t_0) = y_{0,i} = y_{0,i}, \quad i = 0, \dots, m-1 \quad (1.10)$$

mit  $y = (y_1, \dots, y_n)^\top$ ,  $f = (f_1, \dots, f_n)^\top$  auf. Diese können auf ein System 1. Ordnung (1.1) zurückgeführt werden. Setzt man

$$z_i(t) = y^{(i-1)}(t) \quad \text{für } i = 0, \dots, m-1,$$

mit den Anfangsbedingungen  $z_i(t_0) = y_{0,i}$ , dann erhält man das System

$$\begin{aligned} z_i' &= z_{i+1}(t) \\ z_{m-1}' &= f(t, z_0, z_1, \dots, z_{m-1}) \end{aligned}, \quad i = 1, \dots, m-1 \quad (1.11)$$

von  $m \cdot n$  Differentialgleichungen 1. Ordnung mit den Anfangsbedingungen

$$z_i(t_0) = y_{0,i-1}, \quad i = 1, \dots, m.$$

Das heißt, mit Ausnahme von speziellen Verfahren für Gleichungen 2. Ordnung benötigt man numerische Verfahren nur für autonome Systeme 1. Ordnung.

*Beispiel 1.1.2 (Pendel; harmonischer Oszillator).* Unter Vernachlässigung von Reibung lässt sich die Bewegung eines Massepunktes  $M$  mit Masse  $m$  durch eine Differentialgleichung 2. Ordnung beschreiben. Gemäß dem Zweiten Newtonschen Gesetz ist die Kraft  $F$  gleich der Masse  $m$  mal Beschleunigung  $a$ . Im Pendelfall gilt

$$F = m \cdot l \cdot \varphi''(t),$$

wobei  $l$  die Länge des Pendels ist und  $\varphi''$  die Winkelbeschleunigung. Die Kraft  $F$ , die zum Zeitpunkt  $t$  auf  $M$  wirkt ist die tangentielle Komponente der nach unten gerichtete Schwerkraft  $m \cdot g$ ,

$$-m \cdot g \cdot \sin \varphi(t),$$

damit ergibt sich die vereinfachte Differentialgleichung

$$\varphi''(t) = -\frac{g}{l} \sin \varphi(t) \quad (1.12)$$

Die Winkelbeschleunigung ist die Ableitung der Winkelgeschwindigkeit  $\omega$  nach der Zeit, d.h. es gilt

$$\omega'(t) = -\frac{g}{l} \sin \varphi(t),$$

so dass sich (1.12) auf eine Differentialgleichung erster Ordnung zurückführen lässt mit

$$\varphi'(t) = \omega(t).$$

## 1.2 Existenz, Eindeutigkeit und Sensitivität

Bei vielen Anfangswertproblemen kann es vorkommen, dass keine geschlossene Formeln für eine Lösung gefunden werden kann. Für den Definitionsbereich  $D$  der Funktion  $D \subset I \times \mathbb{R}^n$  lässt sich folgende einschränkende Zylindermenge  $Z_{a,b}$  für positive  $a, b > 0$  definieren:

$$Z_{a,b} := \{(t, y) : |t - t_0| \leq a, \|y - y_0\| \leq b\}, \quad a, b > 0 \quad (1.13)$$

Der folgende *Satz von Peano* 1.2.1 liefert die Bedingung für die Existenz mindestens einer Lösung:

**Theorem 1.2.1 (Peano).** *Gegeben sei eine stetige Funktion  $f : Z_{a,b} \rightarrow \mathbb{R}^n$  auf der offenen Teilmenge  $Z_{a,b} \subset \mathbb{R}^{1+n}$ . Dann besitzt jedes Anfangswertproblem*

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0, \quad (t_0, y_0) \in \Omega_0 \quad (1.14)$$

*eine lokale Lösung, d.h. es gibt ein  $\varepsilon = \varepsilon(t_0, y_0) > 0$  derart, so dass das Anfangswertproblem auf dem Intervall  $[t_0, t_0 + \varepsilon]$  mindestens eine Lösung besitzt.*

*Beweis.* (Aulbach, 2010) □

**Definition 1.2.2 (Lipschitz-stetig).** *Die Funktion  $f(t, y)$  genügt Zylinder  $Z_{a,b}$  der Lipschitz-Bedingung, wenn es ein  $L > 0$  gibt, so dass*

$$\|f(t, x) - f(t, y)\| \leq L \|x - y\|, \quad \text{für alle } (t, x), (t, y) \in Z_{a,b}. \quad (1.15)$$

*Zum Nachweis der Lipschitz-Eigenschaft von  $f(t, y)$  ist es hinreichend zu zeigen, dass  $f$  auf der offenen Menge  $D \supset Z_{a,b}$  stetig differenzierbar bezüglich  $y$  ist. Das folgende Lemma 1.2.3 liefert ein Kriterium für die Lipschitz-Eigenschaft.* △

**Lemma 1.2.3.** *Ist die Jacobi-Matrix  $\mathcal{J}(f) := Df(t, y) = (\partial_{y_j} f_i(t, y))_{i,j}$  beschränkt, d.h.*

$$\|\mathcal{J}(f)\| \leq L, (t, y) \in Z_{a,b},$$

*dann genügt  $f$  der Lipschitz-Bedingung bezüglich  $L > 0$ .*

*Beweis.* Der Mittelwertsatz für Vektorfunktionen

$$f(t, u) - f(t, v) = \int_0^1 f_y(t, u + \theta(u - v))(u - v) d\theta, (t, u), (t, v) \in Z_{a,b} \quad (1.16)$$

liefert die Abschätzung

$$\begin{aligned} \|f(t, u) - f(t, v)\| &= \left\| \int_0^1 f_y(t, u + \theta(u - v))(u - v) d\theta \right\| \\ &\leq \int_0^1 \|f_y(t, u + \theta(u - v))\| d\theta \|u - v\| \\ &\leq L \|u - v\| \end{aligned}$$

mit der Lipschitz-Konstanten  $L = \max_{(t,\xi) \in Z_{a,b}} \|f_y(t, \xi)\|$ , womit das Lemma bewiesen wurde.  $\square$

Damit lässt sich der folgende Existenz- und Eindeigkeitssatz von Picard-Lindelöf 1.2.4 bewiesen werden.

**Theorem 1.2.4 (Picard-Lindelöf).** *Ist  $D$  eine offene Teilmenge des  $\mathbb{R}^{1+n}$ , und sei  $f : D \rightarrow \mathbb{R}^n$  stetig und bezüglich  $y$  Lipschitz-stetig, so besitzt jedes Anfangswertproblem der Form*

$$y' = f(t, y), y(t_0) = y_0 \quad (1.17)$$

*eine eindeutig bestimmte, lokale Lösung, d.h. es existiert ein  $\varepsilon = \varepsilon(t_0, \mathbf{y}_0) > 0$  derart, so dass das Anfangswertproblem auf dem Intervall  $[t_0, t_0 + \varepsilon]$  genau eine Lösung besitzt.*

*Beweis.* Aulbach (2010)  $\square$

Es bleibt die Frage was passiert, wenn die Funktion  $f$  lediglich stetig ist. In diesem Fall lassen sich zwei Situationen unterscheiden. So kann es für große Werte Intervalle zu Polstellen kommen, wie bei  $y'(t) = y(t)^2$  oder für kleine Intervalle zum Verlust der Eindeutigkeit kommen, wie zum Beispiel bei  $y'(t) = \sqrt{|y|}$ . Dies führt zum Begriff der *Sensitivität*.

*Beispiel 1.2.5.* Betrachtet wird das lineare Anfangswertproblem

$$y' = \lambda y, y(0) = y_0$$

mit der Lösung

$$y(t) = y_0 e^{\lambda t}.$$

Für  $\lambda < 0$  konvergieren die Lösungen für  $t \rightarrow \infty$  gegen 0 für alle Startwerte  $y_0$ , wohin gegen die Lösungen für  $\lambda > 0$  und alle Startwerte  $y_0$  explodieren, wie in den Abbildungen zu sehen ist.

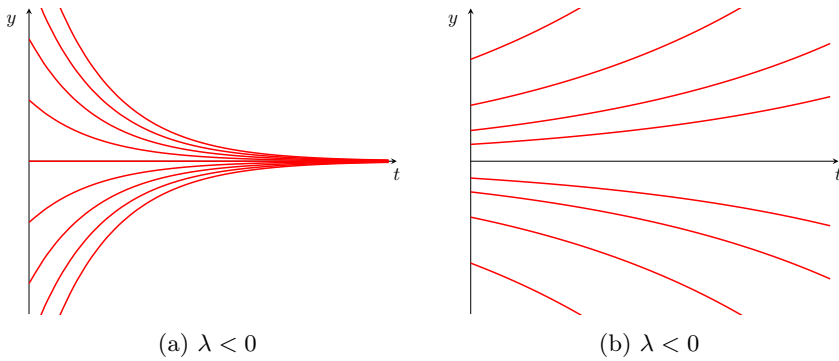


Abb. 1.2: Lösungen des Anfangswertproblems  $y' = \lambda y$  abhängig von Parameter  $\lambda$  und unterschiedlichen Startwerten  $y_0$ .

Werden zu einem Anfangswertproblem zwei Lösungen gefunden, dann können diese wie folgt abgeschätzt werden:

**Lemma 1.2.6.** *Genügt  $f$  eine Lipschitz-Bedingung auf einem Streifen  $S \subset \Omega_0$ , so gilt für zwei Lösungen  $y, (t), w(t)$  die Abschätzung*

$$\|y(t) - w(t)\| \leq \exp[L(t - t_0)] \|y_0 - w_0\| \quad (1.18)$$

für alle  $t \in I$ , d.h. die Lösung hängt stetig vom Anfangswert  $y_0$  ab.

Mit Hilfe der Abschätzung (1.18) lassen sich drei Klassen von Anfangswertproblemen identifizieren: △

- (1) Für moderate  $L(t - t_0)$  spricht man von *nichtsteifen* Lösungen.
- (2) Für große  $L(t - t_0)$ , aber gut konditionierte Systeme spricht man von *steifen* Lösungen.
- (3) Für große  $L(t - t_0)$  und schlecht konditionierte Systeme spricht man von *instabilen* Lösungen.

## 1.3 Notwendige Bezeichnungen und Konzepte

Zur Untersuchung des Lösungsverhaltens werden oft *Landau-Symbolen*  $\mathcal{O}(\cdot), \mathcal{O}(\cdot)$  verwendet. Für zwei Funktionen  $g, h$  sind sie gegeben durch

- $g(x)$  groß  $\mathcal{O}$  von  $h(x)$ ,

$$g(x) = \mathcal{O}(h(x)) \text{ für } x \rightarrow x_0, \text{ falls } \lim_{x \rightarrow x_0} \frac{\|g(x)\|}{\|h(x)\|} < K < \infty.$$

- $g(x)$  klein  $\mathcal{o}$  von  $h(x)$ ,

$$g(x) = \mathcal{o}(h(x)) \text{ für } x \rightarrow x_0, \text{ falls } \lim_{x \rightarrow x_0} \frac{\|g(x)\|}{\|h(x)\|} = 0.$$

*Beispiel 1.3.1.* sEs gilt:

- (1)  $x^3 + x^5 = \mathcal{O}(x^5)$  für  $x \rightarrow \infty$ .
- (2)  $x^3 + x^5 = \mathcal{O}(x^3)$  für  $x \rightarrow 0$ .
- (3)  $\sin x = \mathcal{O}(x)$  für  $x \rightarrow 0$ .
- (4)  $1 - \cos x = \mathcal{O}(x^2)$  für  $x \rightarrow 0$ , denn mittels Taylor-Entwicklung gilt

$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{24} \pm \cdots = 1 - \frac{x^2}{2} \cos \xi + \mathcal{O}(x^4).$$

Um beispielsweise Konvergenzaussagen für den Vektorraum  $\mathbb{R}^n$  treffen zu können werden Vektornormen und ihre zugeordneten Matrixnormen benötigt. Die Abbildung

$$\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$$

heißt *Vektornorm*, wenn die drei Vektornormaxiome

- (i) Für alle  $\mathbf{x} \in \mathbb{R}^n$  gilt  $\|\mathbf{x}\| \geq 0$  und  $\|\mathbf{x}\| = 0$  gilt genau dann, wenn  $\mathbf{x} = \mathbf{0}$ .
- (ii) Für alle  $\mathbf{x} \in \mathbb{R}^n$  und  $\lambda \in \mathbb{R}$  gilt  $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$ .
- (iii) Für alle  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  gilt  $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ .

erfüllt sind.

Für Normen werden meist die so genannte  $p$ -Normen

$$\|\mathbf{x}\|_p := \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad \mathbf{x} \in \mathbb{R}^n, p \in \bar{\mathbb{N}}$$

verwendet. Speziell gilt:

- $\|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i|$  heißt *Betragssummennorm*.
- $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^n |x_i|^2}$  heißt *Euklidische Norm*.
- $\|\mathbf{x}\|_\infty := \max_i |x_i|$  heißt *Maximumsnorm*.

△

Die Betragssummennorm wird auch als Manhattan-Norm bezeichnet, da damit auf einem Gitter der kürzeste Abstand zwischen zwei Punkten bestimmt werden kann.

**Definition 1.3.2.** Zu jeder  $p$ -Vektornorm  $\|\cdot\|_p$  gibt es eine zugeordnete Matrixnorm  $\|\cdot\|_p$

$$\|\mathbf{A}\| := \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|, \quad \mathbf{A} \in \mathbb{R}^{m \times n} \quad (1.19)$$

Mit Definition 1.3.2 gilt für die Einheitsmatrix  $\|\mathbf{I}\|_p = 1$  und folgende Ungleichung

$$\|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{x}\|_p.$$

**Definition 1.3.3.** Für die speziellen  $p$ -Vektornormen  $\|\cdot\|_1$ ,  $\|\cdot\|_2$  und  $\|\cdot\|_\infty$  heißen die durch sie induzierten  $p$ -Matrixnormen



- $\|\mathbf{A}\|_1 := \max_j \sum_{i=1}^n |a_{ij}|$  Spaltensummennorm.
- $\|\mathbf{A}\|_2 := \sqrt{\rho(\mathbf{A}^\top \mathbf{A})}$  Spektralnrm.
- $\|\mathbf{x}\|_\infty := \max_i \sum_{j=1}^n |a_{ij}|$  Zeilensummennorm.



Die euklidische Spektralnrm ist durch ein Skalarprodukt erzeugt, differenzierbar in  $\mathbb{R}^n \setminus \{\mathbf{0}\}$  nicht exakt berechenbar. Allerdings liefert der Spektralradius  $\rho(\mathbf{M})$  einer Matrix  $\mathbf{M}$ , also der betragsmäßig größte Eigenwert der der Matrix  $\mathbf{M}$  eine unter Abschätzung der Matrixnorm

$$\rho(\mathbf{M}) \leq \|\mathbf{M}\|, \mathbf{M} \in \mathbb{R}^{n \times n}.$$

Die in den folgenden Kapitel behandelten Verfahren nutzen ferner ein wichtiges Konzept, die *Taylor-Entwicklung um  $x_0$*  in  $\mathbb{R}$

$$f(x) = \sum_{k=0}^n \frac{1}{k!} f^{(k)}(x_0) (x - x_0)^k + \begin{cases} \mathcal{O}\left((x - x_0)^{n+1}\right), \\ \frac{1}{(n+1)!} f^{(n+1)}(\xi) (x - x_0)^n, \end{cases} \quad \xi \in [x, x_0] \quad (1.20)$$

erklärt. Durch den Übergang von  $\mathbb{R} \rightarrow \mathbb{R}^n$  mit  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$  und  $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^n$  erhält man

$$f(\mathbf{x}) = \sum_{|\alpha|=0}^n \frac{1}{\alpha!} (\mathbf{x} - \mathbf{x}_0)^\alpha D^\alpha f(\mathbf{x}) + \mathcal{O}\left(\|\mathbf{x} - \mathbf{x}_0\|^{n+1}\right).$$

*Beispiel 1.3.4.* Für  $\alpha = 2$  gilt

$$h \mapsto \sum_{i,j} \frac{\partial^2 f_h}{\partial x_i \partial x_j} (x_i - x_{0,i}) (x_j - x_{0,j}),$$

wobei allgemein für den Tensor der  $(n+1)$ -ten Stufe

$$\left[ \frac{\partial^n}{\partial \mathbf{x}^n} \left( \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(k)} \right) \right]_k = \sum_{i_1, \dots, i_n} \frac{\partial^n f_h}{\partial x_{i_1} \dots \partial x_{i_n}} \left[ v_{i_1}^{(1)} v_{i_2}^{(2)} \dots v_{i_n}^{(n)} \right]$$

gilt.

Ein weiteres, wichtiges Hilfsmittel ist der *Mittelwertsatz für für vektorielle Funktionen*  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$  Wegen  $F(\theta) = f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x}))$  gilt  $f(\mathbf{x}) = F(0)$  und  $f(\mathbf{y}) = F(1)$  gilt

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= F(1) - F(0) = \int_0^1 F'(\theta) d\theta \\ &= \int_0^1 f_{\mathbf{x}}(\mathbf{x} + \theta(\mathbf{x} - \mathbf{y})) (\mathbf{y} - \mathbf{x}) d\theta \\ &= (\mathbf{y} - \mathbf{x}) \int_0^1 f_{\mathbf{x}}(\mathbf{x} + \theta(\mathbf{x} - \mathbf{y})) d\theta = (\mathbf{y} - \mathbf{x}) f_{\mathbf{x}}^*, \end{aligned}$$

und demnach

$$\begin{aligned} \|f(\mathbf{x}) - f(\mathbf{y})\| &\leq \left\| (\mathbf{y} - \mathbf{x}) \int_0^1 f_{\mathbf{x}}(\mathbf{x} + \theta(\mathbf{x} - \mathbf{y})) d\theta \right\| \\ &\leq \|\mathbf{y} - \mathbf{x}\| \left\| \int_0^1 f_{\mathbf{x}}(\mathbf{x} + \theta(\mathbf{x} - \mathbf{y})) d\theta \right\| \\ &\leq \|\mathbf{y} - \mathbf{x}\| \int_0^1 \|f_{\mathbf{x}}(\mathbf{x} + \theta(\mathbf{x} - \mathbf{y}))\| d\theta. \end{aligned}$$



# Modellierung mittels gewöhnlicher Differentialgleichungen

## 2.1 Grundlagen

Bei der Beschreibung eines Modells müssen zunächst die beteiligten Größen identifiziert werden. Dabei werden jene Größen,

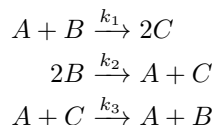
- die bestimmt werden sollen als *Zustandsgrößen* bezeichnet.
- die unveränderlich sind als *Parameter* bezeichnet.
- die einstellbar sind als *Steuergrößen* bezeichnet.

In diesem Sinne werden Anfangswertprobleme, deren Parameter und Steuergrößen bekannt sind, als *direktes Problem* bezeichnet. Sollen diese mittels gewöhnlicher Differentialgleichungen gelöst werden, so wird häufig die Zeit  $t$  als unabhängige Größe betrachtet, und man spricht dann von *Prozessen*.

Werden hingegen Mechanismen im Raum mit diskreten Massepunkten oder homogene Probleme betrachtet, so lassen sich diese ebenfalls mittels gewöhnlicher Differentialgleichungen lösen; für kontinuierliche und inhomogene hingegen mittels partieller Differentialgleichungen.

## 2.2 Reaktionskinetik

Bei der Umsetzung der gelösten Substanzen  $A, B, \dots$  in einem homogenen Medium sind die zu bestimmenden Zustandsgrößen die Konzentrationen  $y_A, y_B, \dots$  der Substanzen  $A, B, \dots$ . Die Kinetik der autokatalytischen Reaktionen



wird durch das System der Differentialgleichungen

$$\begin{aligned}\dot{y}_A &= -k_1 y_A y_B + k_2 y_B^2, \\ \dot{y}_B &= -k_1 y_A y_B - k_2 y_B^2 + k_3 y_A y_C, \\ \dot{y}_C &= -2k_1 y_A y_B + k_2 y_B^2 - k_3 y_A y_C,\end{aligned}$$

beschrieben. Wegen  $\dot{y}_A + \dot{y}_B + \dot{y}_C = 0$ , heißt das System *invariant*.

*Beispiel 2.2.1 (Robertson-Problem).* Das Robertson-Problem beschreibt die Kinetik einer autokatalytischen Reaktion mit den Reaktionskonstanten  $k_1 = .04$ ,  $k_2 = 3 \cdot 10^7$  und  $k_3 = 10^4$ . Die Reaktionen laufen mit unterschiedlich starken Geschwindigkeiten ab. Dieses Problem wird in Kapitel 6 unter dem Begriff *Steifheit* näher untersucht.

## 2.3 Populationsgenetik

Ist eine Population ausreichend groß, dann lässt sich für deren Wachstum ein kontinuierliches System zugrunde legen. Angenommen das biologische System besteht aus  $N$  verschiedenen Spezies. Bezeichnet  $p_i(t)$  die Populationsgröße der  $i$ -ten Spezies, dann lässt sich das Modell wie folgt

$$p'_i = \alpha_i p_i - \beta_i p_i + \gamma_{ij} p_i p_j - \delta_{ik} p_i p_k \quad (2.1)$$

beschreiben. Dabei bezeichnen die Parameter

- $\alpha_i$  die *Reproduktionsrate* der Spezies  $i$ ,
- $\beta_i$  die *Sterberate* der Spezies  $i$ ,
- $\gamma_{ij}$  die *nahrungsabhängige Vermehrungsrate*, und
- $\delta_{ik}$  die *Entnahmerate*.

Dabei sei angemerkt, dass die Parameter  $\alpha_i, \dots, \delta_{ik}$  keineswegs fest sein müssen, sondern auch von  $t$  abhängen können.

*Beispiel 2.3.1 (Lotka/Volterra-Modell).* Im einfachen Lotka-Volterra-Modella mit zwei Spezies lauten die zwei Differentialgleichungen

$$\dot{p}_1 = \lambda_1 p_1 - \delta_{12} p_1 p_2, \quad \dot{p}_2 = -\lambda_2 p_2 + \gamma_{21} p_1 p_2, \quad (2.2)$$

wobei  $\lambda_i$  die *Nettoreproduktionsrate* der Spezies  $i$  ist.

Das so beschriebene Modell ist periodisch und da die speziellen Lösungen eindeutig sind, können sich die Niveaulinien nicht schneiden.

Das System ist im Gleichgewicht, wenn immer  $\dot{p}_1 = \dot{p}_2 = 0$  ist. In diesem Fall gilt

$$0 = p_1 (\lambda_1 - \delta_{12} p_2), \quad 0 = p_2 (\gamma_{21} p_1 - \lambda_2)$$

und das System ist konstant für

$$\begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} = \begin{pmatrix} \lambda_2 / \gamma_{21} \\ \lambda_1 / \delta_{12} \end{pmatrix}.$$

In allen anderen Fällen ist das System nichtkonstant, sondern periodisch. Um dies zu zeigen muss die Invariante des Systems bestimmt werden, d.h.  $F(x, y)$  mit  $\dot{F}(p_1, p_2) = 0$ , d.h. für  $p = (p_1, p_2)^\top$  dann

$$\partial_1 F(p_1, p_2) \cdot \dot{p}_1 + \partial_2 F(p_1, p_2) \cdot \dot{p}_2 = \partial_1 F(p_1, p_2) (\lambda_1 - \delta_{12} p_2) + \partial_2 F(p_1, p_2) (\gamma_{21} p_1 - \lambda_2).$$

Setzt man

$$\partial_1 F(p_1, p_2) = G(p_1, p_2) (\lambda_1 - \delta_{12} p_2), \quad \partial_2 F(p) = G(p_1, p_2) (\gamma_{21} p_1 - \lambda_2),$$

dann gilt  $\partial_2 \partial_1 F(p) = \partial_1 \partial_2 F(p)$  sowie

$$\partial_1 F(p_1, p_2) = \frac{\lambda_2 - \delta_{21} p_2}{p_1}, \quad \partial_2 F(p_1, p_2) = \frac{\lambda_1 - \delta_{21} p_2}{p_2}, \quad (2.3)$$

für  $G(p_1, p_2) = \frac{1}{p_1 p_2}$ .

Lösen dieser Differentialgleichung liefert

$$F(p_1, p_2) = \lambda_2 \ln p_1 - \delta_{21} p_1 + \lambda_1 \ln p_2 - \delta_{12} p_2, \quad (2.4)$$

deren Niveaulinien geschlossene Kurven sind. Das Ergebnis findet sich in [2.1](#).

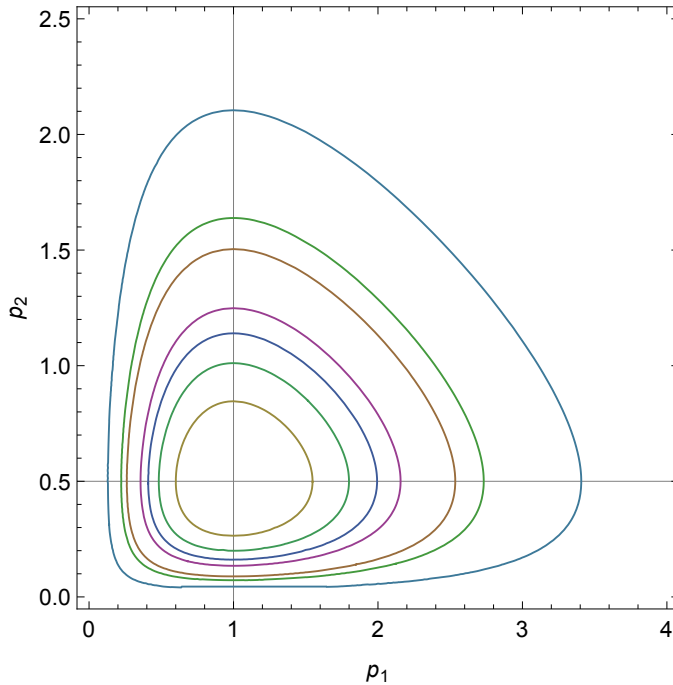


Abb. 2.1: Phasenlinien des Räuber/Beute-Modells mit  $\lambda_1 = 2/3$ ,  $\lambda_2 = 4/3$ ,  $\delta_{12} = 1 = \delta_{21}$  zu verschiedenen Anfangswerten. Der Fixpunkt liegt in  $(1, 1/2)$ .

## 2.4 Mechanik

### 2.4.1 Mehrkörpersysteme in Zustandsform

Ein Mehrkörpersystem ist ein mechanisches System von Einzelkörpern, Massepunkte oder Starrkörper, die über starre Verbindungen, Federn oder Dämpfer miteinander verbunden sind.

*Beispiel 2.4.1 (Doppelpendel).* Ein Doppelpendel, vgl. [2.2](#), besitzt zwei Massepunkte  $m_1, m_2$ . Die Pendellänge der Masse  $m_1$  sei  $l_1$  und der Winkel Öffnungswinkel sei  $\varphi_1$ . Entsprechend bezeichnet  $l_2$  und  $\varphi_2$  die Pendellänge und den Winkel zum Massepunkt  $m_2$ .

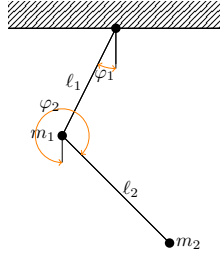


Abb. 2.2: Mathematisches Doppelpendel

Die Lage der zwei Massepunkte kann durch als Zustandsform  $(\varphi_1(t), \varphi_2(t))$  oder als Deskriptorform  $((x_1(t), y_1(t)), (x_2(t), y_2(t)))$  ausgedrückt werden. Zur Beschreibung des Doppelpendels in der Zustandsform bietet sich das *Hamilton-Prinzip* an:

Schritt 1: Wähle einen minimalen Satz von Zustandsgrößen  $q(t) = (q_1(t), \dots, q_{n_q}(t))^T$ .

Schritt 2: Erstelle ein Gleichungssystem der kinetischen Energie  $E_{\text{kin}} \equiv T(q, \dot{q})$  aller vorkommender Körper und der potentiellen Energie  $E_{\text{pot}} \equiv V(q)$  des Systems.

Schritt 3: Formuliere die Lagrange-Funktion  $L(q, \dot{q}) = T(q, \dot{q}) - V(\dot{q})$  für die kinetischen und potentiellen Energien.

Schritt 4: Löse das *Hamiltonische Wirkungsfunktional*  $\int_{t_0}^T L(q, \dot{q}) dt$  für stationäre Werte  $q(t_0) = \alpha$  und  $q(\tau) = \beta$ ,

$$\delta \int_{t_0}^T L(q, \dot{q}) dt = 0. \quad (2.5)$$

Zur Bestimmung der Bewegungsgleichung des Doppelpendels in der Zustandsform sind  $q(t) \equiv (q_1, q_2)^T$  die verallgemeinerten Koordinaten des Doppelpendels.

Ist  $\mathbf{M} : \mathbb{R}^2 \rightarrow \mathbb{R}^{n \times n}$  die symmetrische, positiv definite Massematrix des Doppelpendels, so bestimmt sich die kinetische Energie des Doppelpendels aus der homogenen, quadratischen Form

$$E_{\text{kin}} \equiv T(q, \dot{q}) = \frac{1}{2} \sum_{i,j} m_{ij}(q) q_i q_j = \frac{1}{2} \dot{q}^T \mathbf{M}(q) \dot{q},$$

womit die Lagrange-Funktion

$$L(q, \dot{q}) = \frac{1}{2} \dot{q}^T \mathbf{M}(q) \dot{q} - V(q).$$

Wird  $q(t)$  in eine Kurvenschar eingebettet  $q(t) + \delta q(t)$  für ein kleines  $|\delta|$ , so erhält man für das Hamiltonische Wirkungsfunktional

$$\begin{aligned}
\delta \int_{t_0}^{\tau} L(q, \dot{q}) dt &= \int_{t_0}^{\tau} L(q + \delta q, \dot{q} + \delta \dot{q}) dt - \int_{t_0}^{\tau} L(q, \dot{q}) dt \\
&= \int_{t_0}^{\tau} \left( \frac{\partial L}{\partial q} \delta q + \frac{\partial L}{\partial \dot{q}} \delta \dot{q} \right) dt \\
&= \int_{t_0}^{\tau} \frac{\partial L}{\partial q} \delta q dt - \int_{t_0}^{\tau} \frac{D}{Dt} \left( \frac{\partial L}{\partial \dot{q}} \right) \delta q dt + \left. \frac{\partial L}{\partial \dot{q}} \delta q \right|_{t_0}^{\tau} \\
&= \int_{t_0}^{\tau} \left( \frac{\partial L}{\partial q} \delta q - \frac{D}{Dt} \left( \frac{\partial L}{\partial \dot{q}} \right) \right) \delta q dt = 0,
\end{aligned} \tag{2.6}$$

da der Klammerausdruck für jedes  $\delta q$  verschwinden muss.

Setzt man nun für ein konkretes Doppelpendel

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} -l_1 \cos \varphi_1 \\ -l_1 \sin \varphi_1 \end{pmatrix}, \quad \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} -l_1 \cos \varphi_1 - l_2 \cos \varphi_2 \\ -l_1 \sin \varphi_1 - l_2 \sin \varphi_2 \end{pmatrix}$$

mit

$$\begin{aligned}
T &= \frac{1}{2} (m_1 (\dot{x}_1^2 + \dot{y}_1^2) + m_2 (\dot{x}_2^2 + \dot{y}_2^2)), \\
V &= m_1 g y_1 + m_2 g y_2,
\end{aligned}$$

dann ergibt sich aus (2.6) aufgrund der Euler-Lagrange-Gleichungen

$$\frac{D}{Dt} \left( \frac{\partial L}{\partial \dot{q}} \right) = \frac{\partial L}{\partial q} \tag{2.7}$$

also

$$\frac{D}{Dt} \left( \frac{\partial T}{\partial \dot{q}} \right) = \frac{\partial T}{\partial q} - \frac{\partial V}{\partial q}.$$

Damit lautet die **Lagrange-Bewegungsgleichung 2. Art** aus

$$\frac{D}{Dt} \left( \frac{\partial T}{\partial \dot{q}} \right) = \mathbf{M}(q) \ddot{q} = \frac{\partial T}{\partial q} - \frac{\partial V}{\partial q}$$

Im Fall des Doppelpendels erhält man die Lösung durch die Abbildung  $(q_1, q_2) \mapsto (\varphi_1, \varphi_2)$ .

### 2.4.2 Mehrkörpersysteme in Deskriptorform

Soll zum Beispiel das Doppelpendel aus Beispiel 2.4 als System mit „einfachen“ Koordinaten beschrieben werden, so kann dies als Minimierungsproblem

$$f(\mathbf{z}) \rightarrow \min_{g(\mathbf{z})=0} \tag{2.8}$$

unter Nebenbedingungen  $g(\mathbf{z})$  begriffen werden. Existieren *keine aktiven* Nebenbedingungen, dann ist die notwendige Bedingung

$$\nabla f(\mathbf{x}) \equiv \mathbf{0}. \tag{2.9}$$

Existieren hingegen  $s \in \mathbb{N}$  *aktive* Nebenbedingungen, dann lässt sich (2.8) mittels Lagrange-Ansatz

$$\Lambda(\mathbf{z}; \boldsymbol{\lambda}) := f + \sum_{k=1}^s \lambda_k g_k(\mathbf{z}), \quad \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_s) \quad (2.10)$$

lösen. In diesem Fall ist die notwendige Bedingung

$$\nabla \Lambda(\mathbf{x}; \boldsymbol{\lambda}) \equiv \mathbf{0}$$

*Beispiel 2.4.2.* Betrachte wird ein Doppelpendel. Zum Zeitpunkt  $t$  seien die Position  $\mathbf{Y}$  der Massepunkte  $\mathbf{m}_1, \mathbf{m}_2$  gegeben durch  $\mathbf{z}^\top = (x_1, y_1, x_2, y_2)$ . Gesucht ist dann der Vektor  $\mathbf{q}$  derart, für den

$$\delta \int_{t_0}^{t_1} L(\mathbf{z}, \dot{\mathbf{z}}) dt \equiv 0$$

gilt. Das zugehörige System mit Nebenbedingungen ist gegeben durch

$$\begin{aligned} E_{\text{kin}} &= T = \frac{1}{2} m_1 (\dot{x}_1^2 + \dot{y}_1^2) + \frac{1}{2} m_2 (\dot{x}_2^2 + \dot{y}_2^2), \\ E_{\text{kin}} &= V = \frac{1}{2} m_1 (\dot{x}_1^2 + \dot{y}_1^2) + \frac{1}{2} m_2 (\dot{x}_2^2 + \dot{y}_2^2), \\ 0 &\equiv g(x_1, y_1, x_2, y_2) = g(\mathbf{z}) = \left( \begin{array}{c} x_1^2 + y_1^2 - l_1^2 \\ (x_1 - x_2)^2 + (y_1 - y_2)^2 - l_2^2 \end{array} \right). \end{aligned}$$

Wird  $g$  mittels Lagrange-Multiplikator an  $L(\mathbf{q}, \dot{\mathbf{q}})$  gekoppelt, dann liefert

$$\delta \int_{t_0}^{t_1} L(\mathbf{z}, \dot{\mathbf{z}}) - \boldsymbol{\lambda}^\top g(\mathbf{z}) dt \equiv 0$$

mittels den Euler-Lagrange-Gleichungen

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{z}_i} = \frac{\partial L}{\partial z_i} + \boldsymbol{\lambda}^\top \frac{\partial g}{\partial z_i}$$

das notwendige das Gleichungssystem

$$\begin{aligned} m_1 \ddot{x}_1 &= +2\lambda_1 x_1 + 2\lambda_2 (x_1 - x_2) \\ m_1 \ddot{y}_1 &= -m_1 g + 2\lambda_1 y_1 + 2\lambda_2 (y_1 - y_2) \\ m_2 \ddot{x}_2 &= -2\lambda_2 (x_1 - x_2) \\ m_2 \ddot{y}_2 &= -m_2 g - 2\lambda_2 (y_1 - y_2) \end{aligned} \quad (2.11)$$

ist die gesuchte *Deskriptorform*. Für die Nebenbedingungen hat man die notwendigen Bedingungen

$$\begin{aligned} 0 &\equiv x_1^2 + y_1^2 - l_1^2, \\ 0 &\equiv (x_1 - x_2)^2 + (y_1 - y_2)^2 - l_2^2, \end{aligned} \quad (2.12)$$

und diese heißen *Algebradifferentialgleichungen*. Lösen von (2.11) und (2.12) liefert dann die gewünschte Bewegungsgleichung.

### 2.4.3 Hamilton-Systeme

Hamilton-Systeme sind dynamische Systeme, die durch die Hamiltonsche Mechanik beschrieben wird und dient der Betrachtung der Entwicklung von Bewegungen im Phasenraum und lassen sich aus den Euler-Lagrange-Gleichungen entwickeln. Betrachtet wird der *Impuls*



$$p_i := \frac{\partial L}{\partial \dot{q}_i}. \quad (2.13)$$

So ist für  $L = \sum_i \frac{1}{2} m_i \dot{q}_i^2$  der Impuls dann  $p_i = m_i \dot{q}_i$ . Setzt man

$$H := \sum_i p_i \dot{q}_i - L, \quad (2.14)$$

so erhält man für das Differential

$$\begin{aligned} dH &= \sum_i (\dot{q}_i \cdot dp_i + p_i dq_i) \\ &= \sum_i \dot{q}_i dp_i - \sum_i \frac{\partial L}{\partial q_i} dq_i, \end{aligned}$$

wobei das Differential einer Funktion  $f(x, y)$  gegeben ist durch

$$df(x, y) = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy.$$

Stellt man  $H$  als  $H(\mathbf{p}, \mathbf{q})$  dar, dann gilt

$$\frac{\partial H}{\partial q_i} = \frac{\partial L}{\partial q_i} = \frac{D}{Dt} p_i - \dot{p}_i,$$

und dies liefert die Hamiltonschen Gleichungen

$$\begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i}, \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i}, \end{aligned} \quad (2.15)$$

für die Gesamtenergie  $H$  des System.



## Einschrittverfahren

Zur numerischen Lösung von Anfangswertproblemen

$$y' = f(t, y), \quad y(t_0) = y_0, \quad (3.1)$$

die auf dem Gebiet  $S := \{(t, y) | t_0 \leq t \leq T, y \in \mathbb{R}^n\}$  stetig sind und dort der Lipschitz-Bedingung genügen, können auf einem Gitter Einschrittverfahren herangezogen werden. Ihre Behandlung wird der Gegenstand der Betrachtungen in diesem Kapitel sein. Zur Unterscheidung von exakter und numerischer Lösung wird mit  $y$  die exakte Lösung, und mit  $u$  die numerische Lösung bezeichnet.

### 3.1 Grundlagen

Grundlage der numerischen Behandlung ist die *Diskretisierung* oder *Zerlegung* des Integrationsintervalls  $[t_0, T]$  in ein *Punktgitter*  $I_h = \{t_0, t_1, \dots, t_N\}$  mit den *Gitterpunkten*  $t_m$ ,

$$t_0 < t_1 < t_2 < \dots < t_N \leq T,$$

und den Schrittweiten  $h_m := t_{m+1} - t_m$ . Dies dient zur Bestimmung von Näherungen  $u_m \approx y(t_m)$  als Funktionswerte der *Gitterfunktion*  $u_h : I_h \rightarrow \mathbb{R}^n$  sind.

**Definition 3.1.1 (Einschrittverfahren).** Ein Einschrittverfahren zur Bestimmung einer Gitterfunktion  $u_h$  ist gegeben durch

$$u_{m+1} = u_m + h_m \varphi(t_m, u_m, h_m; f), \quad u_0 = y_0. \quad (3.2)$$

Die Funktion  $\varphi$  heißt Verfahrens- oder Inkrementfunktion.

*Beispiel 3.1.2 (Euler-Verfahren).* Die Lösung des Anfangswertproblems (3.1) kann auf ein *Quadraturproblem*

$$y(t_m + h) = y(t_m) + \int_{t_m}^{t_{m+1}} f(\tau, y(\tau)) d\tau \quad (3.3)$$

zurückgeführt werden. Man erhält eine Näherungslösung  $u_{m+1}$  mittels

- *explizitem Euler-Verfahren*: Ersetzt man die Lösung  $y(t_m)$  durch  $u_m$  und die Ableitung  $f(t_m, y(t_m))$  durch den Vorwärts-Differenzenquotient  $(u_{m+1} - u_m)/h_m$ , so erhält man das explizite Euler-Verfahren

$$u_{m+1} = u_m + h_m f(t_m, u_m), \quad u_0 = y_0, \quad m = 0, \dots, N-1. \quad (3.4)$$

- *impliziten Euler-Verfahren*: Ersetzt man die Lösung  $y(t_m)$  durch  $u_m$  und die Ableitung  $f(t_m, y(t_m))$  durch den Rückwärts-Differenzenquotient  $(u_m - u_{m-1})/h_{m-1}$ , so erhält man nach Indexverschiebung das implizite Euler-Verfahren

$$u_{m+1} = u_m + f h_m(t_{m+1}, u_{m+1}), \quad m = 0, \dots, N-1 \quad (3.5)$$

Für das Quadraturproblem (3.3) heißt dies, dass im Fall des expliziten Euler-Verfahren das Integral durch das Produkt aus Integranden am *linken Rand* des Intervalls mit der Intervallbreite betrachtet wird, während im impliziten Euler-Verfahren das Integral durch das Produkt aus Integranden am *rechten Rand* des Intervalls mit der Intervallbreite betrachtet wird. Die Abbildungen (3.1a) und (3.1b) verdeutlichen die Unterschiedlichen Ansätze.

	[Abb:2.1.2]	[Abb:2.1.3]
(a)	Ex-(b)	Im-
plizites	plizites	
Euler-	Euler-	
Verfahren	Verfahren	

Abb. 3.1: Vorwärts- und Rückwärts-Differenzenquotient

**Definition 3.1.3 (Konsistenz).** Sei  $K > 0$ . Ein Einschrittverfahren  $\Phi$  ist konsistent von der Ordnung  $p \in \mathbb{N}$ , wenn für ein hinreichend glattes Anfangswertproblem

$$y' = f(t, y), \quad y(t_0) = y_0 \quad (3.6)$$

wenn

$$\|y(t_m + h) - u_{m+1}\| \leq K h^{p+1}, \quad (3.7)$$

d.h. wenn die Taylor-Entwicklung der exakten Lösung  $y(t_m + h)$  und der numerischen Lösung bis einschließlich des Terms  $h^p$  übereinstimmen.

### 3.2 Explizite Runge/Kutta-Verfahren

Die wichtigste Klasse von Einschrittverfahren zur Lösung numerischer Probleme sind die *Runge/Kutta-Verfahren*. Betrachtet man das explizite Euler-Verfahren als abgebrochenes Taylor-Entwicklung der Funktion  $y(t_m + h)$  nach dem zweiten Glied. Bricht man die Taylor-Entwicklung von  $y(t_m; h)$  nach dem dritten Glied ab, so erhält man

$$u_{m+1} = u_m + h f(t_m, u_m) + \frac{h_m^2}{2} (f_t(t_m, u_m) + f_y(t_m, u_m) f(t_m, u_m)). \quad (3.8)$$

Setzt man im Quadraturproblem (3.3)

$$u_{m+1}^{(1)} := u_m, \quad u_{m+1}^{(2)} := u_m + hf(t_{m+1}, u_m),$$

so erhält man hieraus mit (3.8)

$$u_{m+1} = u_m + \frac{h}{2} \left( f(t_m, u_{m+1}^{(1)}) + f(t_{m+1}, u_{m+1}^{(2)}) \right). \quad (3.9)$$

Das Verfahren (3.8) ist von der Konsistenz  $p = 2$ , das Verfahren (3.9) ist von der Konsistenz  $p = 1$ . Offensichtlich müssen bei Verfahren höherer Konsistenz zusätzliche Differentiale bestimmt werden, die solche Verfahren unter Umständen unnötig verlangsamen. Dem gegenüber hat das im Folgenden zu definierende Runge/Kutta-Verfahren den Vorteil, dass auf die Ableitung von  $f(t, y(t))$  vollständig verzichtet werden kann: Bei Verfahren dieser Art spricht man von Runge/Kutta-Verfahren:

**Definition 3.2.1 (Runge/Kutta-Verfahren).** Sei  $s \in \mathbb{N}$ . Ein Einschrittverfahren der Form

$$\begin{aligned} u_{m+1} &= u_m + h \sum_{i=1}^s b_i f(t_m + c_i h, u_{m+1}^{(i)}) \\ u_{m+1}^{(i)} &= u_m + h \sum_{j=1}^s a_{ij} f(t_m + c_j h, u_{m+1}^{(j)}), \quad i = 1, \dots, s \end{aligned} \quad (3.10)$$

heißt  $s$ -stufiges Runge/Kutta-Verfahren. Dabei bezeichnet

- $\mathbf{c} = (c_1, \dots, c_s)^\top$  den Knotenvektor,
- $\mathbf{A} = (a_{ij})$  die Verfahrensmatrix,
- $\mathbf{b} = (b_1, \dots, b_s)$  den Gewichtsvektor, und
- $s$  die Stufenzahl des Runge/Kutta-Verfahrens.

Ist die Verfahrensmatrix  $\mathbf{A}$  eine strikte, untere Dreiecksmatrix, d.h.  $a_{ij} = 0$  für  $j \geq i$ , dann lassen sich die Hilfsgrößen  $u_{m+1}^{(i)}$  explizit aus (3.10) berechnen. In diesem Fall spricht man von einem *expliziten Runge/Kutta-Verfahren*, sonst von einem *impliziten Runge/Kutta-Verfahren*. In der Regel lassen sich implizite Verfahren nicht ohne weiteres lösen, sondern bedürfen nicht-lineare Lösungsstrategien. △

Eine zu (3.10) äquivalente Formulierung des  $s$ -stufigen Runge/Kutta-Verfahrens ist gegeben durch △

$$\begin{aligned} u_{m+1} &= u_m + h \sum_{i=1}^s b_i k_i(t_m, u_{m+1}; h) \\ k_i(t_m, u_{m+1}; h) &= f\left(t_m + c_i h, u_m + h \sum_{j=1}^s a_{ij} k_j(t_m, u_m; h)\right), \quad i = 1, \dots, s. \end{aligned} \quad (3.11)$$

Die Formulierung (3.11) ist für die Implementierung in der Regel besser geeignet, denn  $f$  wird an jeder Stelle  $u_{m+1}^{(i)}$  nur einmal berechnet und basiert auf der Steigungswerten  $k_i(t_m, u_m; h)$ , die Formulierung (3.10) auf den Stufenwerten  $u_{m+1}^{(i)}$ .

**Lemma 3.2.2.** Sei  $u_m$  die numerische Lösung eines konsistenten Runge/Kutta-Verfahrens zur Lösung des Anfangswertproblems

$$y' = f(t, y), \quad y(t_0) = y_0.$$

Ist  $z_m$  die numerische Lösung des zugehörigen autonomen Systems und ist die Knotenbedingung erfüllt, dann gilt

$$z_m = \begin{pmatrix} u_m \\ t_m \end{pmatrix}.$$

*Beweis.* [Strehmel et al. \(2012, Satz 2.4.1\)](#)

□

Zur Darstellung der Koeffizienten  $s$ -stufiger Runge/Kutta-Verfahren bedient man sich häufig der Tableau-Schreibweise

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^\top \end{array}.$$

Da für explizite Runge/Kutta-Verfahren  $a_{ij} = 0$  für  $j \geq i$  wird dieses Tableau üblicherweise als *Butcher-Schema*

$$\begin{array}{c|ccc} c_1 = 0 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{21} & a_{32} & \\ \vdots & \vdots & & \ddots \\ c_s & a_{s1} & \dots & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array}$$

angegeben.

*Beispiel 3.2.3.* Das spezielle Anfangswertproblem

$$y' = f(t), \quad y(t_m) = y_m$$

ist äquivalent zu dem *Quadraturproblem*

$$y(t_m + h) = y(t_m) + \int_{t_m}^{t_{m+1}} f(\tau) d\tau.$$

Damit ist ein  $s$ -stufiges Runge/Kutta-Verfahren

$$u_{m+1} = u_m + h \sum_{i=1}^s b_i f(t_m + c_i + h)$$

offensichtlich ein Verfahren zur numerischen Integration

**Lemma 3.2.4.** Es gilt äquivalent:

(a) Ein  $s$ -stufiges Runge/Kutta-Verfahren ist konsistent.

(b) Es gilt

$$\sum_{i=1}^s b_i = 1.$$

*Beweis.* Sei

$$\phi(t, y(t); h) = \sum_{i=1}^s b_i f\left(t_m + c_i h, \tilde{u}_{m+1}^{(i)}\right)$$

die Verfahrensfunktion für das betrachtete,  $s$ -stufige Runge/Kutta-Verfahren, wobei  $\tilde{u}_{m+1}^{(i)}$  den Näherungswert zum Stufenwert  $\tilde{u}_m = y(t_m)$  bezeichnet. Erfüllt  $f$  die Lipschitz-Bedingung, dann gilt für  $h = 0$

$$\phi(t, y(t); 0) = f(t, y(t))$$

genau dann, wenn  $\sum_{i=1}^s b_i = 1$ . Dann gilt

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\|e(t+h)\|}{h} &= \lim_{h \rightarrow 0} \frac{\|y(t+h) - \tilde{u}(t+h)\|}{h} \\ &= \lim_{h \rightarrow 0} \frac{\|y(t+h) - y(t) - h\phi(t, y(t); h)\|}{h} \\ &= \|y'(t) - \phi(t, y(t); 0)\| = \|f(t, y(t)) - f(t, y(t))\| \\ &= 0, \end{aligned}$$

womit die Behauptung gezeigt wurde. □

Ist die *Knotenbedingung*

$$c_i = \sum_{j=1}^s a_{ij}, \quad i = 1, \dots, s \quad (3.12)$$

erfüllt, dann können Bedingungen für die Konsistenzordnung von Runge/Kutta-Verfahren anhand der Betrachtung autonomer Systeme formuliert werden.

### 3.3 Ordnungsaussagen und B-Reihen für Runge/Kutta-Verfahren

Nach Lemma 3.2.3 kann jedes nichtautonome Problem auf ein autonomes Problem zurückgeführt werden kann. Ist ferner  $f$  auf  $S$  stetig differenzierbar, dann bietet es sich an die Taylor-Entwicklung der exakten Lösung  $y(t_m + h)$  und der numerischen Lösung  $\tilde{u}_{m+1}$  an der Stelle  $t_m + h$  zu betrachten

$$\begin{aligned} y(t_m + h) &= y(t_m) + \sum_{k=1}^p \frac{h^k}{k!} y^{(k)}(t_m) + \mathcal{O}(h^{p+1}), \\ u_h(t_m + h) &= y(t_m) + \sum_{k=1}^p \frac{h^k}{k!} \frac{d^k}{dh^k} u_h^{(k)}(t_m) + \mathcal{O}(h^{p+1}), \end{aligned}$$

Eine naive Herangehensweise für die Differentialgleichung  $y' = f(y)$  und  $f'(y) = f_y(y)$  liefert für die ersten drei Ableitungen

$$y' = f(y) \quad (3.13)$$

$$y'' = \frac{d}{dt} y' = \frac{d}{dt} f(y) = f_y(y) y'(t) = f_y f(y) \quad (3.14)$$

$$y''' = \frac{d}{dt} y'' = \frac{d}{dt} (f'(y) f(y)) = f_{yy}(f, f) + f_y f_y f. \quad (3.15)$$

Mit  $u_{m+1}^{(2)} = u_m + hf(u_m)$  und  $u_{m+1} = y_m + h/2 \left( f(u_{m+1}^{(1)}) + f(u_{m+1}^{(2)}) \right)$  erhält man

$$\begin{aligned}
 u_{m+1} &= u_m + \frac{h}{2} \left( f(u_{m+1}^{(1)}) + f(u_m + hf(u_m^{(1)})) \right) \\
 &= u_m + f(u_{m+1}^{(1)}) + \frac{h^2}{2} f'(u_m) f(u_m) + \frac{1}{4} h^3 f''(u_m) f(u_m) f(u_m) + \mathcal{O}(h^4) \\
 &= u_m + f(u_{m+1}^{(1)}) + \frac{h^2}{2} f'(u_m) f(u_m) + \frac{1}{6} h^3 \frac{6}{4} f''(u_m) f(u_m) f(u_m) + \mathcal{O}(h^4) \\
 &= u_m + u_{m+1}^{(1)} + \frac{h^2}{2} u_{m+1}^{(2)} + u_{m+1}^{(3)} + \mathcal{O}(h^4)
 \end{aligned}$$

### 3.3.1 Taylor-Entwicklung der exakten Lösungen

Folgenden werden autonome Systeme  $y' = f(y)$  mit  $y \in \mathbb{R}$  betrachtet. Die rechte Seite sei ferner auf einem Gebiet  $S$  hinreichend oft stetig differenzierbar. Ist  $V$  ein Vektorraum über  $\mathbb{R}$ , und  $V^d$  das  $n$ -fache, kartesische Produkt  $V \times \cdots \times V$ , heißt eine Funktion  $f : V^d \rightarrow \mathbb{R}$  mit den Eigenschaften

- (i)  $f(\mathbf{v}_1, \dots, \mathbf{v}_i + \tilde{\mathbf{v}}_i, \dots, \mathbf{v}_d) = f(\mathbf{v}_1, \dots, \mathbf{v}_i, \dots, \mathbf{v}_d) + f(\mathbf{v}_1, \dots, \mathbf{v}_i \tilde{\mathbf{v}}_i, \dots, \mathbf{v}_d)$
- (ii)  $f(\mathbf{v}_1, \dots, \alpha \mathbf{v}_i, \dots, \mathbf{v}_d) = \alpha f(\mathbf{v}_1, \dots, \mathbf{v}_i \tilde{\mathbf{v}}_i, \dots, \mathbf{v}_d)$

für alle  $i = 1, \dots, d$  und  $\alpha$  Tensor der Ordnung  $d$  auf  $V$ .

In Komponentenschreibweise

$$y'_i = f_i(y_1, \dots, y_n)$$

ergibt sich das Differential  $f_y f$  zu

$$(f_y f)_i = \frac{d}{dt} y'_i = \sum_k \frac{\partial f_i}{\partial y_k} f_k.$$

Für die  $i$ -te Komponente der Differentiale  $f_{yy}(f, f)$  und  $f_y f_y f$  erhält man die Darstellung

$$(f_{yy}(f, f))_i = \sum_{k,l} \frac{\partial^2 f_i}{\partial y_k \partial y_l} f_k f_l, \quad (f_y f_y f)_i = \sum_{k,l} \frac{\partial f_i}{\partial y_k} \frac{\partial f_k}{\partial y_l} f_l,$$

wobei  $f_{yy}$  ein Tensor der dritten Stufe ist. Offensichtlich durch wiederholte Anwendung von Produkt und Kettenregel die Darstellung höherer Ableitungen schnell unüberschaubar, da die elementaren Differentiale einzeln mitgeführt werden müssen. Eine von [Butcher \(1963\)](#) eingeführte Idee besteht in der Darstellung von der Ableitungen durch Wurzelbäume. Zunächst wird  $f(y)$  ein Baum zugeordnet, der nur aus der Wurzel  $\bullet$  besteht. Differenziert man weiter, dann wird an diese Wurzel ein neues Element angehängen. Das elementare Differential  $f_y f$  wird dann durch den Baum

$$y'' = f_y f, \quad \begin{array}{c} \bullet \\ | \\ \bullet \end{array} \begin{array}{c} f \\ | \\ f' \end{array} \quad (3.16)$$

dargestellt. Die Struktur des elementaren Differential (3.15) wird mit den Bäumen



$$y''' = f_{yy}(f, f) + f_y f_y f, \quad \begin{array}{c} f \quad f \\ \diagdown \quad \diagup \\ f'' \end{array} + \begin{array}{c} f \\ | \\ f' \\ | \\ f' \end{array} \quad (3.17)$$

Eine Ableitung von  $f$  entspricht also einem Knoten mit mit einem Ast, eine zweite Ableitung von  $f$  ein Knoten mit zwei Ästen. Übersichtshalber wird im Folgenden die Beschriftung der Knoten weggelassen. Die Ableitung des Differentials  $(f_{yy}(f, f))$  ist

$$\begin{aligned} (f_{yy}(f, f))' &= f_{yyy}(f, f, f) + f_{yy}(f_y f, f) + f_{yy}(f, f_y f), & \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} \\ &= f_{yyy}(f, f, f) + 2f_{yy}(f_y f, f), & \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + 2 \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} \end{aligned}$$

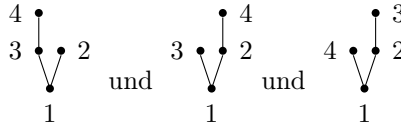
und die Ableitung von  $f_y f_y f$  ist

$$(f_y f_y f)' = f_{yy}(f, f_y, f) + f_y f_{yy}(f, f) + f_y f_y f_y, \quad \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array}$$

Damit ergibt sich  $y^{(4)}$  zu

$$y^{(4)} = f_{yyy}(f, f, f) + 3f_{yy}(f_y f, f) + f_y f_{yy}(f, f) + f_y f_y f_y, \quad \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + 3 \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} + \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array} \quad (3.18)$$

wobei die Bäume von  $f_{yy}$  zusammengefasst werden können, denn dies Operatoren sind symmetrisch im Sinne von  $f''(x_1, x_2) = f''(x_2, x_1)$ . Dies ist aber nur dann möglich, wenn die Knoten der Bäume nicht ausgezeichnet sind. Zeichnet man die Knoten Man betrachte die Bäume von  $f_{yy}(f_y f, f)$ ,  $f_{yy}(f, f_y f)$  und  $f_{yy}(f, f, f)$



Bäume dieser Art werden auch *ausgezeichnete* oder *numerierte* Bäume  $\mathcal{LT}_n$  mit genau  $n$  Knoten bezeichnet, wobei jeder Sohnknoten an den Vaterknoten angehängt wird. Durch das anhängen eines Baumes an eine neue Wurzel können neue Bäume erzeugt werden. Seien  $\tau_1, \dots, \tau_k$  beliebige Bäume, dann erzeugt die Operation  $[\cdot]$  einen neuen Baum mit

$$[\tau_1, \tau_2, \dots, \tau_k] := \begin{array}{c} \tau_1 \quad \tau_2 \quad \dots \quad \tau_k \\ \diagdown \quad \diagup \quad \dots \quad \diagup \\ \bullet \end{array},$$

und durch die mehrfache Anwendung von  $[\cdot]$  kann jeder Baum erzeugt werden. Der Baum zu  $\triangle$

dem Differential  $f_{yy}(f_{yy}(f_{yy}(f, f), f_y f_y f))$  besitzt die Struktur  $\begin{array}{c} \vdots \\ \vee \\ \vdots \end{array}$  und kann erzeugt werden durch  $[\begin{array}{c} \vdots \\ \vee \\ \vdots \end{array}] = [[\cdot, \cdot], [\cdot]]$ , d.h. die Bäume können rekursiv erzeugt werden. Ist  $\mathcal{T}$  die Menge aller Bäume, dann

$$\begin{aligned} \mathcal{T} &= \left\{ \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array}, \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array}, \begin{array}{c} \vdots \\ \vee \\ \vdots \end{array}, \dots \right\} \\ &:= \{\cdot\} \cup \{[\tau_1, \tau_2, \dots, \tau_k] \mid \tau_i \in \mathcal{T}\} \end{aligned} \quad (3.19)$$

Da die Bäume Darstellung elementarer Differentiale sind, können diese auch rekursiv definiert werden

**Definition 3.3.1 (elementares Differential).** Ein Wurzelbaum  $\tau$  wird rekursiv einem elementaren Differential

$$F(\tau)(y) = \begin{cases} f(y) & \tau = \bullet \\ f^{(k)}(y)[F(\tau_1)(y), \dots, F(\tau_k)(y)] & \tau = [\tau_1, \tau_2, \dots, \tau_k] \end{cases}$$

zugeordnet.

Die Taylor-Entwicklung bis zur vierten Stufe ist

$$y(t+h) = y(t) + y'(t)h + \frac{1}{2!}y''(t)h^2 + \frac{1}{3!}y'''(t)h^3 + \frac{1}{4!}y^{(4)}(t)h^4 + \mathcal{O}(h^5).$$

Mit der Definition 3.3.1 erhält man dann

$$\begin{aligned} y(t+h) &= y(t) + F(\bullet)h + \frac{1}{2!}F(\textcircled{\bullet})h^2 + \frac{1}{3!}\left(F(\textcircled{\textcircled{\bullet}}) + F(\textcircled{\textcircled{\bullet}})\right) \\ &\quad + \frac{1}{4!}\left(F(\textcircled{\textcircled{\textcircled{\bullet}}}) + 3F(\textcircled{\textcircled{\textcircled{\bullet}}}) + F(\textcircled{\textcircled{\textcircled{\bullet}}}) + F(\textcircled{\textcircled{\textcircled{\bullet}}})\right)h^4 + \mathcal{O}(h^5). \end{aligned}$$

Dies lässt sich verallgemeinern

**Theorem 3.3.2.** Die Taylor-Entwicklung der exakten Lösung ist rekursiv

$$y(t+h) = y(t) + \sum_{k=1}^p \frac{h^k}{k!} \sum_{\tau \in \mathcal{L}\mathcal{T}} F(\tau)(y(t)) + \mathcal{O}(h^{p+1}). \quad (3.20)$$

Die Darstellung (3.20) nicht unbedingt praktikabel, daher

**Definition 3.3.3.** Die Anzahl der Knoten eines Baumes  $\tau \in \mathcal{T}$  heißt Ordnung  $\rho = \rho(\tau)$ , so dass  $\mathcal{T}_\rho$  die Menge aller Bäume mit  $\rho$  Knoten ist. Für  $\tau \in \mathcal{T}$  heißt dann  $\alpha(\tau)$  die Anzahl der möglichen monotonen Nummerierungen.

Damit kann ein allgemeines Ergebnis für die  $\rho$ -fache Ableitung der exakten Lösung gilt:

**Theorem 3.3.4.** Die exakte Lösung von  $y_i^{(\rho)} = f_i(y_1, \dots, y_n)$  genügt

$$y^{(\rho)}(t_m+h) = \sum_{\tau \in \mathcal{L}\mathcal{T}_\rho} F(\tau)(y(t)) \sum_{\tau \in \mathcal{T}_\rho} \alpha(\tau) F(\tau)(y(t)). \quad (3.21)$$

*Beweis.* Die Fälle der Ordnung  $\rho = 1, 2, 3$  wurden bereits behandelt. Für die vierte Ableitung erhält man die Struktur (3.18) und man erhält diese dadurch, dass an jeden Baum (3.17) mit drei Knoten wieder neue Bäume angehängt werden und die Knoten wie in Abbildung (3.2) neu ausgezeichnet werden.

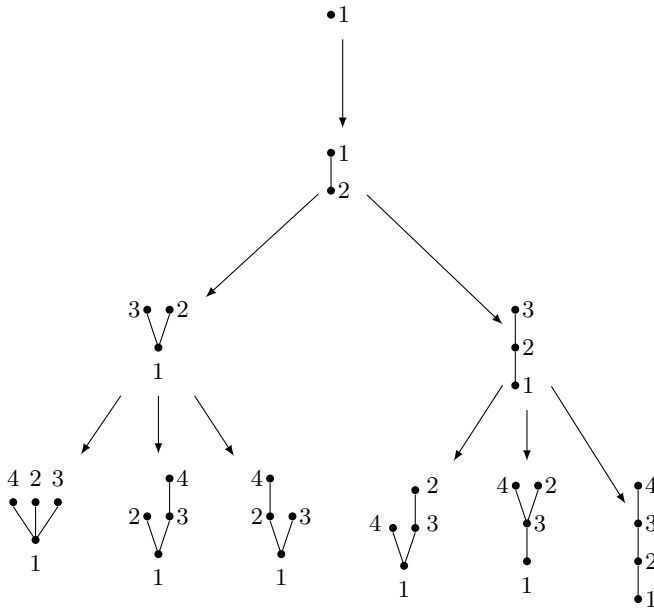


Abb. 3.2: Entwicklung der Ableitungen der exakten Lösung.

Folglich besitzt die  $\rho$ -fache Ableitung ausgezeichnete Bäume der Ordnung  $\rho$ . Werden Bäume mit identischen elementaren Differentialen zusammengefasst, dann erhält man die Bedingung (3.21), was zu zeigen war.  $\square$

### 3.3.2 Taylor-Entwicklung numerischer Lösungen

Gesucht sind nun die Taylor-Entwicklungen für Stufen der Lösung  $\tilde{u}_{m+1}$  des Runge/Kutta-Verfahrens

$$u_{m+1}^{(i)}(h) = y_n + h \sum_j a_{ij} f\left(\tilde{u}_{m+1}^{(j)}\right)$$

an der Stelle um  $h$ . Übersichtshalber bezeichne  $\partial^k \tilde{u}_{m+1}$  die  $k$ -te Ableitung von  $\tilde{u}_{m+1}^{(k)}$ , dann lassen sich die ersten Ableitungen als

$$\begin{aligned} \partial^0 \tilde{u}_{m+1}^{(i)} &= y(t_m), \\ \partial \tilde{u}_{m+1}^{(i)} &= \sum_{j \leq s} a_{ij} f\left(\tilde{u}_{m+1}^{(j)}\right), \\ \partial^2 \tilde{u}_{m+1}^{(i)} &= 2 \sum_{j,k} a_{ij} a_{jk} f'\left(\tilde{u}_{m+1}^{(j)}\right) f\left(\tilde{u}_{m+1}^{(k)}\right). \end{aligned}$$

schreiben. Für höhere Ableitungen gilt das folgenden

**Lemma 3.3.5.** *Es gilt*

$$\left. \frac{d^k}{dh^k} \right|_{h=0} h f(f(y(h))) = k \left. \frac{d^{k-1}}{dh^{k-1}} \right|_{h=0} f(f(y(h))). \quad (3.22)$$

*Beweis.* Durch die  $k$ -fache Anwendung der Produktregel ist das Ergebnis immer Null, es sei denn  $h$  wird genau einmal, und  $f(y)$  genau  $k-1$  mal differenziert. Also gibt es  $k$  Möglichkeiten.  $\square$

**Definition 3.3.6 (Dichte; Symmetrie).** Bezeichne  $\overline{\mathcal{T}}_p := \{\tau \in \mathcal{T} \mid \rho(\tau) \leq p\}$  die Menge aller Bäume bis zur Ordnung  $p$ .

(1) Die Symmetrie  $\sigma(\tau)$  eines Baumes ist rekursiv definiert durch

$$\begin{aligned} \sigma(\cdot) &= 1 \\ \sigma\left(\left[\tau_1^{l_1}, \tau_2^{l_2}, \dots, \tau_k^{l_k}\right]\right) &= l_1! \sigma(\tau_1)^{l_1} l_2! \sigma(\tau_2)^{l_2} \cdots l_k! \sigma(\tau_k)^{l_k}, \end{aligned} \quad (3.23)$$

wobei die Exponenten  $l_i$  die Anzahl gleicher Teilbäume angeben,

$$\left[\tau_1^{l_1}, \tau_2^{l_2}, \dots, \tau_k^{l_k}\right] := \left[\underbrace{\tau_1, \dots, \tau_1}_{l_1}, \underbrace{\tau_2, \dots, \tau_2}_{l_2}, \dots, \underbrace{\tau_k, \dots, \tau_k}_{l_k}\right].$$

(2) Die Dichte  $\gamma(\tau)$  eines Baumes  $\tau = [\tau_1, \tau_2, \dots, \tau_k]$  ist rekursiv definiert durch

$$\begin{aligned} \gamma(\cdot) &= 1 \\ \gamma(\tau) &= \rho(\tau) \gamma(\tau_1) \gamma(\tau_2) \cdots \gamma(\tau_k). \end{aligned} \quad (3.24)$$

$\triangle$

Der Zugang zu dieser Definition erscheint zunächst schwer. Die elementaren Gewichte erhält als das Produkt der Ordnung  $\rho(\tau)$  und all der Ordnung  $\rho(\tau_i)$  die dadurch entstehen, dass man nacheinander die Väter entfernt.

Die Symmetrie eines Baumes muss letztendlich kombinatorisch bestimmt werden. Das folgende Beispiel zeigt, wie die Werte berechnet werden können, wobei ausgenutzt wird, dass mittels der Operation  $[\cdot]$  komplexe Bäume mit (3.19) rekursiv dargestellt werden können.

*Beispiel 3.3.7.* Der Baum  $\tau = \mathfrak{V} \updownarrow \mathfrak{I} = [\mathfrak{V}, \mathfrak{I}] = [[\cdot^2], [[\cdot]]]$  besitzt sieben Knoten, und damit die Ordnung  $\rho\left(\mathfrak{V} \updownarrow \mathfrak{I}\right) = 7$ . Für die Dichte gilt

$$\begin{aligned} \gamma\left(\mathfrak{V} \updownarrow \mathfrak{I}\right) &= 7\gamma(\mathfrak{V})\gamma(\mathfrak{I}) \\ &= 7 \cdot (3 \cdot 1 \cdot 1) \cdot (3 \cdot 2 \cdot 1) \\ &= 126. \end{aligned}$$

Der Teilbaum  $\mathfrak{V}$  besitzt eine Spiegelsymmetrie, womit

$$\begin{aligned} \sigma\left(\mathfrak{V} \updownarrow \mathfrak{I}\right) &= \sigma\left(\mathfrak{V} \updownarrow \mathfrak{I}\right) \sigma(\mathfrak{I}) \\ &= \sigma\left([\cdot^2]\right) \sigma([\cdot]) \\ &= 2 \cdot 1 = 2. \end{aligned}$$

Die Menge  $\mathcal{ST}$  ist die Menge der Wurzelbäume, die höchstens in der Wurzel eine Verzweigung finden. Insbesondere werden mit  $\mathcal{LST}$  die Menge der ausgezeichneten Einfachbäume bezeichnet. Offensichtlich können auch einfache Bäume ausgezeichnet werden, in diesem Fall wird die Menge mit  $\mathcal{LST}$  bezeichnet, womit

$\triangle$

**Definition 3.3.8.** Sei  $\tau = [\tau_1, \dots, \tau_l]$  ein einfacher ausgezeichneter Baum.. Dann heißt

$$\widehat{F}(\tau) = f_{y \dots y} \left( y^{\rho(\tau_1)}, \dots, y^{\rho(\tau_l)} \right)$$

elementares Differential des einfachen Baumes  $\tau$ .

**Lemma 3.3.9 (Faà di Bruno).** Es gilt

$$f(y)^{(k)} = \sum_{\tau \in \mathcal{LST}_k} \widehat{F}(\tau). \quad (3.25)$$


Mit dem Lemma von Faà die Bruno können nun die Taylor-Entwicklungen der numerischen Lösungen angegeben werden. Zu diesem Zweck wird Folgendes benötigt

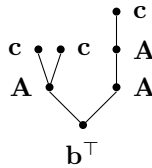
**Definition 3.3.10 (elementares Gewicht).** Das elementare Gewicht  $\Phi(\tau)$  eines  $s$ -stufigen RK-Verfahrens ist definiert durch

$$\begin{aligned} \Phi(\bullet) &= \sum_i b_i \\ \Phi([\tau_1, \tau_2, \dots, \tau_k]) &= \sum_i b_i \left( \tilde{\Phi}_i(t_1) \tilde{\Phi}_i(t_2) \cdots \tilde{\Phi}_i(t_k) \right) \end{aligned} \quad (3.26)$$

mit den Vektoren

$$\begin{aligned} \tilde{\Phi}(\bullet) &= \mathbf{c} \\ \tilde{\Phi}([\tau_1, \tau_2, \dots, \tau_k]) &= \left( \sum_j a_{ij} \left( \tilde{\Phi}_j(t_1) \tilde{\Phi}_j(t_2) \cdots \tilde{\Phi}_j(t_k) \right) \right)_{i=1}^s \end{aligned} \quad (3.27)$$

*Beispiel 3.3.11.* Um die elementaren Gewichte eines Baumes zu bestimmen geht man am besten grafisch vor. Dabei wird die Wurzel mit  $\mathbf{b}^\top$  ausgezeichnet, die inneren Knoten mit  $\mathbf{A}$  und die Blätter mit  $\mathbf{c}$ . Betrachtet man den Baum , dann



Zur Berechnung und übersichtlichen Darstellung bezeichne der Operator  $*$  das gewöhnliche Matrix-Vektor-Produkt entlang der Kanten und  $\otimes$  die komponentenweise Multiplikation auf gleicher Ebene. Dann erhält man für das elementare Gewicht

$$\begin{aligned} \Phi(\text{tree}) &= \mathbf{b}^\top * (\mathbf{A} * (\mathbf{c} \otimes \mathbf{c})) \otimes (\mathbf{A} * \mathbf{A} * \mathbf{c}) \\ &= \sum_i b_i \tilde{\Phi}_i(\mathbf{v}) \tilde{\Phi}_i(\mathbf{!}) \\ &= \sum_{i,j,k,l} b_i (a_{ij} c_j^2) (a_{ik} a_{kl} c_l) \end{aligned} \quad (3.28)$$

Mit der letzten Definition kann nun die die Entwicklung der numerischen Lösung angegeben werden

**Theorem 3.3.12.** Die Entwicklung der numerische Lösung  $\tilde{u}_{m+1}$  eines  $s$ -stufigen Runge/Kutta-Verfahrens ist

$$\tilde{u}_{m+1}^{(i)} = y(t_m) + \sum_{\tau \in \bar{\mathcal{T}}_p} \tilde{\Phi}_i(\tau) \frac{h^{\rho(\tau)}}{\sigma(\tau)} F(\tau)(y(t_m)) + \mathcal{O}(h^{p+1}), \quad i = 1, \dots, s, \quad (3.29)$$

$$\tilde{u}_{m+1} = y(t_m) + \sum_{\tau \in \bar{\mathcal{T}}_p} \Phi(\tau) \frac{h^{\rho(\tau)}}{\sigma(\tau)} F(\tau)(y(t_m)) + \mathcal{O}(h^{p+1}). \quad (3.30)$$

*Beweis.* [Strehmel et al. \(2012\)](#) □

Werden anstelle allgemeiner Bäume ausgezeichnete Bäume betrachtet, dann erhält man für die asymptotische  $k$ -te Abbildung der numerischen Lösung

**Korollar 3.3.13.** Die  $k$ -te Ableitung des Stufenwertes  $\tilde{u}_{m+1}^{(i)}$  genügt

$$\begin{aligned} \left. \frac{d^k}{dh^k} \tilde{u}_{m+1}^{(i)} \right|_{h=0} &= \sum_{\tau \in \mathcal{LT}_k} \gamma(\tau) \tilde{\Phi}_i(\tau) F(\tau)(y(t_m)) \\ &= \sum_{\tau \in \mathcal{LT}_k} \gamma(\tau) \sum_j a_{ij} \tilde{\Phi}_j(\tau) F(\tau)(y(t_m)). \end{aligned} \quad (3.31)$$

Die  $k$ -te Ableitung der numerischen Lösung  $\tilde{u}_{m+1}$  genügt

$$\begin{aligned} \left. \frac{d^k}{dh^k} \tilde{u}_{m+1} \right|_{h=0} &= \sum_{\tau \in \mathcal{LT}_k} \gamma(\tau) \sum_i b_i \tilde{\Phi}_i(\tau) F(\tau)(y(t_m)) \\ &= \sum_{\tau \in \mathcal{T}_k} \alpha(\tau) \gamma(\tau) \sum_i b_i \tilde{\Phi}_i(\tau) F(\tau)(y(t_m)). \end{aligned} \quad (3.32)$$

*Beweis.* Der Beweis der asymptotische Taylor-Entwicklung für die numerische Lösung wird mittels vollständiger Induktion gezeigt für (3.31) gezeigt, da hier aus dann auch (3.32) folgt.

- $n = 1$ : Im Fall  $n = 1$  gilt

$$\begin{aligned} \left. \frac{d}{dh} \tilde{u}_{m+1}^{(i)} \right|_{h=0} &= \left. \frac{d}{dh} \left( y(t_m) + h \sum_j a_{ij} f(\tilde{u}_{m+1}^{(j)}) \right) \right|_{h=0} \\ &= 1 \left( \sum_j a_{ij} f(\tilde{u}_{m+1}^{(j)}) \right) \\ &= \sum_{j \leq i-1} a_{ij} f(\tilde{u}_m) = \gamma(\cdot) \Phi_i(\cdot) F(\cdot), \end{aligned}$$

womit die Behauptung für  $n = 1$  erfüllt ist.

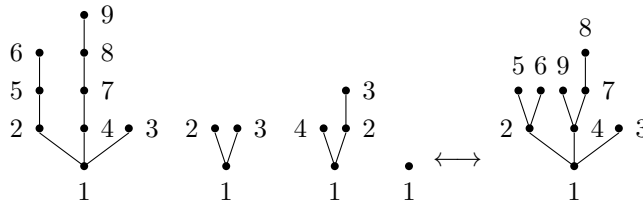
- $n - 1 \rightarrow n$ : Sei angenommen, die Behauptung wurde bereits für ein  $q \in \{1, \dots, n - 1\}$  gezeigt. Sei  $\tau = [\tau_1, \dots, \tau_r]$  und  $\rho(\tau_i) = k_i$

$$\begin{aligned}
\left. \frac{d^n}{dh^n} \tilde{u}_{m+1}^{(i)} \right|_{h=0} &= n \sum_j a_{ij} f \left( \tilde{u}_{m+1}^{(j)} \right) \\
&= n \sum_j a_{ij} \sum_{\tau \in \mathcal{LST}_n} \hat{F}(\tau) \\
&= n \sum_j a_{ij} \sum_{\tau \in \mathcal{LST}_n} \underbrace{f y \dots y}_{\tau\text{-fach}} \left( \partial^{k_1} \tilde{u}_{m+1}^{(j)}, \dots, \partial^{k_r} \tilde{u}_{m+1}^{(j)} \right) \\
&= n \sum_j a_{ij} \sum_{\tau \in \mathcal{LST}_n} f_{y \dots y} \left( \sum_{t_1 \in \mathcal{LT}_{k_1}} \gamma(t_1) \tilde{\Phi}_j(t_1) F(t_1), \dots, \sum_{t_r \in \mathcal{LT}_{k_r}} \gamma(t_r) \tilde{\Phi}_j(t_r) F(t_r) \right) \\
&= \sum_{\tau \in \mathcal{LST}_n} \sum_{t_1 \in \mathcal{LT}_{k_1}} \dots \sum_{t_r \in \mathcal{LT}_{k_r}} n \gamma(t_1) \dots n \gamma(t_r) \\
&\quad \cdot \sum_j a_{ij} \tilde{\Phi}_j(t_1) \dots \tilde{\Phi}_j(t_r) f_{y \dots y}(F(t_1), \dots, F(t_r)) . \\
&= \sum_{\substack{t \in \mathcal{LT}_n \\ t=[t_1, \dots, t_n]}} n \gamma(t_1) \dots n \gamma(t_r) \sum_j a_{ij} \tilde{\Phi}_j(t_1) \dots \tilde{\Phi}_j(t_r) f_{y \dots y}(F(t_1), \dots, F(t_r)) \\
&= \sum_{\substack{t \in \mathcal{LT}_n \\ t=[t_1, \dots, t_n]}} \gamma(t) \tilde{\Phi}_i(t) F(t)
\end{aligned}$$

womit die Behauptung auf für beliebige  $n$  gezeigt wurde.

□

Im Beweis wurde ausgenutzt dass mehrere Teilbäume zu einem Baum zusammengesetzt werden. Das bedeutet, dass es eine 1:1-Korrespondenz gibt, z.B. △



### 3.3.3 Ordnungsbedingungen für Runge/Kutta-Verfahren

Betrachtet man die Bedingungen (3.30) dann gilt

**Theorem 3.3.14.** *Für ein Runge/Kutta-Verfahren sind folgende Aussagen äquivalent.*

- (a) *Das Runge/Kutta-Verfahren hat die Ordnung  $p$ .*
- (b) *Für alle Bäume  $\tau \in T$  mit  $\rho(\tau) \leq p$  gilt die Bedingung*

$$\frac{1}{\gamma(\tau)} = \Phi(\tau) = \sum_i b_i \left( \tilde{\Phi}_i(t_1) \tilde{\Phi}_i(t_2) \dots \tilde{\Phi}_i(t_k) \right). \quad (3.33)$$

*Beispiel 3.3.15.* Die Ordnungsbedingungen für den Baum aus Beispiel 3.3.11 erhält man mit Satz 3.3.14

$$\frac{1}{126} = \sum_{i,j,k,l} b_i (a_{ij} c_j^2) (a_{ik} a_{kl} c_l).$$

Setzt man  $\mathbf{1} = (1, \dots, 1)^\top$ ,  $\mathbf{C} = \text{diag}(c_i)$  und  $\mathbf{c}^k = \mathbf{C}^k \mathbf{1}$ , dann können die Ordnungsbedingungen für Runge/Kutta-Verfahren bis  $p = 4$  kompakt in Tabelle 3.1 zusammenfassen

$p$	$\tau$	$F(\tau)$	$\sigma(\tau)$	$\gamma(\tau)$	Bedingung	$\Phi(\tau)$
1	•	$f$	1	1	$1 = \sum b_i = \mathbf{b}^\top \mathbf{1}$	
2	‡	$f'f$	1	2	$\frac{1}{2} = \sum b_i c_i = \mathbf{b}^\top \mathbf{c}$	
3	✱	$f''(f, f)$	2	3	$\frac{1}{3} = \sum b_i c_i^2 = \mathbf{b}^\top \mathbf{c}^2$	
	‡	$f'f'f$	1	6	$\frac{1}{6} = \sum b_i a_{ij} c_i = \mathbf{b}^\top \mathbf{A} \mathbf{c}$	
4	✱✱	$f'''(f, f, f)$	6	4	$\frac{1}{4} = \sum b_i c_i^3 = \mathbf{b}^\top \mathbf{c}^3$	
	‡	$f''(f, f', f)$	1	8	$\frac{1}{8} = \sum b_i c_i a_{ij} c_j = \mathbf{b}^\top \mathbf{c} \mathbf{A} \mathbf{c}$	
	✱	$f'f''(f, f)$	2	12	$\frac{1}{12} = \sum b_i a_{ij} c_j^2 = \mathbf{b}^\top \mathbf{A} \mathbf{c}^2$	
	‡	$f'f'f'f$	1	24	$\frac{1}{24} = \sum b_i a_{ij} a_{jk} c_k = \mathbf{b}^\top \mathbf{A} \mathbf{A} \mathbf{c}$	

Tabelle 3.1: Ordnungsbedingungen für implizite Runge/Kutta-Verfahren bis  $p = 4$ .

## 3.4 Konstruktion expliziter Runge/Kutta-Verfahren

### 3.4.1 Verfahren bis zur Ordnung $p = 3$

Ausgehend von der Tabelle 3.1 können nun explizite Runge/Kutta-Verfahren konstruiert werden. Dabei wird vorausgesetzt, dass die Knotenbedingung

$$c_i = \sum_{j \leq i-1} a_{ij}, \quad i = 1, \dots, s \quad (3.34)$$

und insbesondere  $c_1 = 0$  erfüllt ist. Das einzige einstufige explizite Runge/Kutta-Verfahren ist das bereits genannte Euler-Verfahren

*Beispiel 3.4.1 (Euler).* Für  $p = s = 1$  gilt  $b_1 = 1$  und dies liefert

$$\begin{aligned} u_{m+1}^{(1)} &= u_m, \\ u_{m+1} &= u_m + hf \left( u_{m+1}^{(1)} \right) = u_m + hf(u_m). \end{aligned}$$

Setzt man  $b_1 + b_2 = 1$ , dann besitzen zweistufige Runge/Kutta-Verfahren mit (3.34) das Parameterschema

$$\begin{array}{c|c} 0 & c_2 \\ \hline c_2 & 1 - \frac{1}{2c_2} \quad \frac{1}{2c_2} \end{array}, \quad c_2 \neq 0.$$



*Beispiel 3.4.2 (Runge; Heun).* Sei  $p = s = 2$ . Für die Knoten  $c_2 = 1/2$  erhält man das Verfahren von *Runge*, dies entspricht der *Mittelpunktregel* eines Quadraturproblems. In kompakter Schreibweise heißt dies

$$u_{m+1} = u_m + hf \left( u_m + \frac{h}{2} f(u_m) \right). \quad (3.35)$$

Für  $c_2 = 1$  erhält man das Verfahren von *Heun*, dies entspricht der *Trapezregel* eines Quadraturproblems. In kompakter Schreibweise heißt dies

$$u_{m+1} = u_m + \frac{h}{2} (f(u_m) + f(u_m + hf(u_m))) \quad (3.36)$$

Für ein Verfahren der Ordnung  $p = 3$  sind mindestens  $s = 3$  Stufen notwendig und jedes Verfahren muss das Parameterschema

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \end{array}.$$

Zusammen mit Tabelle 3.1 müssen dann die vier Bedingungen

$$b_1 + b_2 + b_3 = 1 \quad (3.37a)$$

$$b_2 c_2 + b_3 c_3 = \frac{1}{2} \quad (3.37b)$$

$$b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3} \quad (3.37c)$$

$$b_3 a_{32} c_2 = \frac{1}{6} \quad (3.37d)$$

Werden die Knoten  $c_2$  und  $c_3$  als freie Parameter aufgefasst, mit  $c_2 \neq 0$ , dann können folgende Fälle unterschieden werden.

Fall A: Für  $c_2 \neq c_3 \neq 0$  erhält man durch Lösen von (3.37b) und (3.37c) die Gewichte

$$b_2 = \frac{3c_3 - 2}{6c_2(c_3 - c_2)}, \quad b_3 = \frac{2 - 3c_3}{6c_3(c_3 - c_2)}.$$

Ist der Parameter  $b_3 \neq 0$  frei, dann folgt aus (3.37d), dass  $c_2 \neq 2/3$ , womit aus (3.37a) und (3.37d) die Parameter

$$b_1 = 1 - b_2 - b_3, \quad a_{32} = \frac{1}{6b_3 c_2}.$$

Das Butcher-Schema für diesen Fall ist

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & & c_2 & \\ c_3 & \frac{c_3(3c_2 - 3c_2^2 - c_3)}{c_2(2 - 3c_2)} & \frac{c_3(c_3 - c_2)}{c_2(2 - 3c_2)} & \\ \hline & \frac{-3c_3 + 6c_2 c_3 + 2 - 3c_3}{6c_2 c_3} & \frac{c_3 - 2}{6c_2(c_3 - c_2)} & \frac{2 - 3c_3}{6c_3(c_3 - c_2)} \end{array}$$

Fall B: Sei  $c_3 = 0$ . Aus (3.37b) und (3.37c) folgt  $c_2 = 2/3$  und  $b_2 = 3/4$ . Setzt man  $b_3 \neq 0$ , dann erhält man die Parameter

$$b_1 = \frac{1}{4} - b_3, \quad a_{32} = \frac{1}{4b_3}.$$

Das Butcher-Schema für diesen Fall ist

$$\begin{array}{c|ccc} 0 & & & \\ 2 & 2 & & \\ \frac{2}{3} & \frac{2}{3} & & \\ 3 & 1 & 1 & \\ 0 & -\frac{1}{4b_3} & \frac{1}{4b_3} & \\ \hline & \frac{1}{4} - b_3 & \frac{3}{4} & b_3 \end{array}.$$

Fall C: Sei  $c_2 = c_3$ . Aus (3.37b) und (3.37c) ergibt sich  $c_2 = c_3 = 2/3$  und  $b_2 + b_3 = \frac{3}{4}$ . Ist der Parameter  $b_3 \neq 0$  frei, dann folgt aus (3.37d), dann erhält man die Parameter

$$b_1 = \frac{1}{4}, \quad b_2 = \frac{3}{4} - b_3, \quad a_{32} = \frac{1}{4b_3}$$

Das Butcher-Schema für diesen Fall ist

$$\begin{array}{c|ccc} 0 & & & \\ 2 & 2 & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & 2 & 1 & 1 \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{4b_3} & \frac{1}{4b_3} \\ \hline & \frac{1}{4} & \frac{3}{4} - b_3 & b_3 \end{array}.$$

*Beispiel 3.4.3 (Heun; Kutta).* Setzt man  $c_2 = 1/3$  und  $c_3 = 2/3$ , dann liefert der Fall **Fall A** des Verfahren von *Heun*

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

. Setzt man  $c_2 = 1/2$  und  $c_3 = 1$ , dann erhält man das Verfahren von *Kutta*

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

dies liefert die *Simpson-Regel* im Fall eines Quadraturproblems.

### 3.4.2 Verfahren höherer Ordnung

Für Runge/Kutta-Verfahren erhält man aus der Tabelle 3.1 acht Ordnungsbedingungen

$$\begin{array}{c} \cdot, \mathbf{I}, \mathbf{V}, \mathbf{I}, \mathbf{V}, \mathbf{Y}, \mathbf{I} \\ \cdot, \mathbf{I}, \mathbf{V}, \mathbf{I}, \mathbf{V}, \mathbf{Y}, \mathbf{I} \end{array}$$

hinzu kommen drei weitere Knotenbedingungen bei 13 Variablen: drei  $c_i$ , vier  $b_i$  und sechs  $a_{ij}$ . Dies liefert das unterbestimmte Gleichungssystem

$$b_1 + b_2 + b_3 + b_4 = 1 \quad (3.38a)$$

$$b_2 c_2 + b_3 c_3 + b_4 c_4 = \frac{1}{2} \quad (3.38b)$$

$$b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = \frac{1}{3} \quad (3.38c)$$

$$b_2 a_{32} c_2 + b_3 a_{42} c_3 + b_4 a_{43} c_4 = \frac{1}{6} \quad (3.38d)$$

$$b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = \frac{1}{4} \quad (3.38e)$$

$$b_3 c_3 a_{32} c_2 + b_4 c_4 a_{42} c_2 + b_4 c_4 a_{43} c_3 = \frac{1}{8} \quad (3.38f)$$

$$b_3 a_{32} c_2^2 + b_4 a_{42} c_2^2 + b_4 a_{43} c_3^2 = \frac{1}{12} \quad (3.38g)$$

$$b_4 a_{43} a_{32} c_2 = \frac{1}{24} \quad (3.38h)$$

Beim Lösen des Gleichungssystems zeigt sich, dass stets  $c_4 = 1$  gilt.

**Lemma 3.4.4.** *Jedes explizite 4-stufige Runge/Kutta-Verfahren der Ordnung  $p = 4$  erfüllt die vereinfachende Bedingung*

$$D(1) : \quad \sum_{i \leq s} b_i a_{ij} = b_j (1 - c_j), \quad j = 1, \dots, s \quad (3.39)$$

*Beweis.* Das Produkt  $\mathbf{D} \cdot \mathbf{E}$  der Matrizen

$$\mathbf{D} = \begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ b_1 c_1 & b_2 c_2 & b_3 c_3 & b_4 c_4 \\ \sum_i b_i a_{i1} & \sum_i b_i a_{i2} & \sum_i b_i a_{i3} & \sum_i b_i a_{i4} \end{pmatrix} = \begin{pmatrix} \mathbf{b}^\top \\ \mathbf{b}^\top \mathbf{C} \\ \mathbf{b}^\top \mathbf{A} \end{pmatrix},$$

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & c_2 & c_2^2 & 0 \\ 1 & c_3 & c_3^2 & \sum_j a_{3j} c_j \\ 1 & c_4 & c_4^2 & \sum_j a_{4j} c_j \end{pmatrix} = (\mathbf{1}, \mathbf{C}\mathbf{1}, \mathbf{C}^2\mathbf{1}, \mathbf{A}\mathbf{C}\mathbf{1})$$

berechnet sich zu

$$\mathbf{DE} = \begin{pmatrix} \mathbf{b}^\top \mathbf{1} & \mathbf{b}^\top \mathbf{C}\mathbf{1} & \mathbf{b}^\top \mathbf{C}^2\mathbf{1} & \mathbf{b}^\top \mathbf{A}\mathbf{C}\mathbf{1} \\ \mathbf{b}^\top \mathbf{C}\mathbf{1} & \mathbf{b}^\top \mathbf{C}^2\mathbf{1} & \mathbf{b}^\top \mathbf{C}^3\mathbf{1} & \mathbf{b}^\top \mathbf{C}\mathbf{A}\mathbf{C}\mathbf{1} \\ \mathbf{b}^\top \mathbf{A}\mathbf{1} & \mathbf{b}^\top \mathbf{A}\mathbf{C}\mathbf{1} & \mathbf{b}^\top \mathbf{A}\mathbf{C}^2\mathbf{1} & \mathbf{b}^\top \mathbf{A}\mathbf{A}\mathbf{C}\mathbf{1} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{8} \end{pmatrix} = \begin{pmatrix} \cdot & \mathbf{I} & \mathbf{V} & \mathbf{I} \\ \mathbf{I} & \mathbf{V} & \mathbf{V} & \mathbf{V} \\ \mathbf{I} & \mathbf{I} & \mathbf{Y} & \mathbf{I} \\ \mathbf{I} & \mathbf{I} & \mathbf{Y} & \mathbf{I} \end{pmatrix}$$

Offensichtlich ist die erste Zeile die Summe der zweiten und dritten Zeile, und es gilt  $(1, -1, -1)\mathbf{DE} = \mathbf{0}$ , so dass zwei Fälle betrachtet werden können:

Fall 1: Im Fall  $(1, -1, -1)$   $\mathbf{D} \neq \mathbf{0}$  ist  $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$  singular. Offensichtlich kann die erste Spalte  $\mathbf{e}_1$  von  $\mathbf{E}$  keine Linearkombination der anderen sein und es gibt einen Vektor  $\alpha = (\alpha_2, \alpha_3, \alpha_4) \neq \mathbf{0}$  mit  $\alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3 + \alpha_4 \mathbf{e}_4 = \mathbf{0}$ , womit

$$\begin{aligned} \mathbf{0} &= \mathbf{D} (\alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3 + \alpha_4 \mathbf{e}_4) \\ &= \alpha_2 \begin{pmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{6} \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ \frac{1}{3} \\ \frac{1}{4} \\ \frac{1}{12} \end{pmatrix} + \alpha_4 \begin{pmatrix} 1 \\ \frac{1}{6} \\ \frac{1}{8} \\ \frac{1}{24} \end{pmatrix} \end{aligned}$$

Offensichtlich sind die beiden letzten Vektoren von  $\mathbf{DE}$  linear abhängig und es gilt  $\alpha_2 = 0$  und ferner  $\alpha_3 = 0$ , da der Eintrag  $E_{23} = c_2^2 \neq 0$  ist, womit aber auch  $\alpha_4 = 0$  gelten müsste. Dann ist aber  $\mathbf{E}$  regulär, was zum Widerspruch führt.

Fall 2: Im Fall  $(1, -1, 1)$   $\mathbf{D} = \mathbf{0}$  berechnet sich die erste Zeile von  $\mathbf{D}$  als Summe der zweiten und dritten Zeile von  $\mathbf{D}$ , d.h. es gilt  $b_j = b_j c_j + \sum_i b_i a_{ij}$ , und damit die vereinfachende Bedingung (3.39) und für  $j = 4$  liefert die vereinfachende Bedingung (3.39)

$$0 = \sum_i b_i a_{ij} = b_4 (1 - c_4),$$

da für explizite Verfahren  $a_{ij} = 0$  für alle  $j \in \{1, \dots, 4\}$ , und hieraus folgt., dass  $c_4 = 1$ .

Damit wurde das Lemma bewiesen.  $\square$

Für den Abschluss der Konstruktion müssen  $c_2, c_3$  so gewählt werden, so dass die Knoten  $\mathbf{c} = (0, c_2, c_3, 1)$  paarweise verschieden sind, damit die *Vandermon-Matrix*

$$\mathbf{V} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ c_1 & c_2 & c_3 & c_4 \\ c_1^2 & c_2^2 & c_3^2 & c_4^2 \\ c_1^3 & c_2^3 & c_3^3 & c_4^3 \end{pmatrix} \quad (3.40)$$

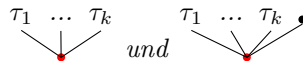
regulär. Dann gilt mit den Ordnungsbedingungen, dass

$$\begin{aligned} \mathbf{b}^\top \mathbf{V} &= \left( \sum_i b_i c_i, \sum_i b_i c_i, \sum_i b_i c_i^2, \sum_i b_i c_i^3 \right) \\ &= \left( 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4} \right). \end{aligned}$$

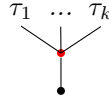
Fast man dies als Quadraturformeln auf, dann besagt dies, dass alle Polynome vom Höchstgrad drei exakt integriert werden. In diesem Fall können auch andere Knoten benutzt werden.

Beachtet man bei der Konstruktion die vereinfachende Bedingung  $D(1)$ , dann gilt der

**Theorem 3.4.5.** *Erfüllt ein Runge/Kutta-Verfahren die vereinfachende Bedingung  $D(1)$ , und die Ordnungsbedingungen für die Bäume*



so sind auch die Ordnungsbedingungen für den Baum



automatisch erfüllt.

*Beweis.* [Strehmel et al. \(2012\)](#)

□

Mit dem Satz 3.4.5 können nun Bäume auch dann reduziert werden.

*Beispiel 3.4.6.* Sei angenommen, die Ordnungsbedingungen für die Bäume auf der linken Seite sind erfüllt, und die zugehörigen Runge/Kutta-Verfahren erfüllen die Bedingung  $D(1)$ . Dann erfüllen auch die Bäume auf der rechten Seite die Ordnungsbedingungen,

$$\begin{array}{l} \vdots \text{ und } \cdot \Rightarrow \vdots \\ \vee \text{ und } \vdots \Rightarrow \vdots \\ \vee\vee \text{ und } \vee \Rightarrow \vee \\ \vee \text{ und } \vdots \Rightarrow \vdots \end{array}$$

Mit diesen Erkenntnissen verbleiben nun drei  $D(1)$ -Bedingungen für  $j = 1, 2, 3$ , drei Knotenbedingungen und eine Bedingung für den Baum  $\vee$ , der nicht weiter reduziert werden kann, da von der Wurzel zwei Äste ablaufen. Mit vorgegebenen Knoten  $\mathbf{c}$  sind diese sieben Gleichungen lösen sechs Variablen  $a_{ij}$ , wobei die Gleichungen linear abhängig sind. Mit  $D(1)$  erhält man

$$\begin{aligned} \sum_j \left( \sum_i b_{ij} a_{ij} \right) &= \sum_j (b_j (1 - c_j)) = \sum_j b_j - \sum_j b_j c_j = \frac{1}{2}, \\ \sum_i b_i \sum_j a_{ij} &= \sum_i b_i c_i = \frac{1}{2}, \end{aligned}$$

dies liefert das lineare Gleichungssystem  $\mathbf{K}\mathbf{a} = \mathbf{z}$  mit

$$\underbrace{\begin{pmatrix} 1 & & & & \\ & 1 & 1 & & \\ & & & 1 & 1 & 1 \\ \hline b_2 & b_3 & & b_4 & & \\ & & b_3 & & b_4 & \\ & & & & & b_4 \\ \hline & & b_3 c_3 c_2 & b_4 c_4 c_2 & b_4 c_4 c_3 & \end{pmatrix}}_{=\mathbf{K}} \underbrace{\begin{pmatrix} a_{21} \\ a_{31} \\ a_{32} \\ a_{41} \\ a_{42} \\ a_{43} \end{pmatrix}}_{=\mathbf{a}} = \underbrace{\begin{pmatrix} c_2 \\ c_3 \\ b_1 (1 - c_1) \\ b_2 (1 - c_2) \\ b_3 (1 - c_3) \\ 1/8 \end{pmatrix}}_{=\mathbf{z}}$$

Die erste Zeile liefert  $a_{21}$ , die zweite Zeile  $a_{31}$  und die dritte Zeile  $a_{41}$ . Die Zeile sechs liefert  $a_{43}$  und die Zeilen fünf und sieben die Koeffizienten  $a_{32}, a_{42}$ .

*Beispiel 3.4.7 (Kuttasche 3/8-Regel).* Führt man die Konstruktion auf die *Newtonsche 3/8-Regel* zurück, dann sind die Gewichte und die Knoten gegeben durch

$$\mathbf{b} = (1/8, 3/8, 3/8, 1/8)^\top, \quad \mathbf{c} = (0, 1/3, 2/3, 1)^\top$$

und das Gleichungssystem (3.38a) bis (3.38h) liefert das Butcher-Schema

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

Beispiel 3.4.8 (Klassisches Runge/Kutta-Verfahren). Führt man die Konstruktion auf die Simpsonregel zurück, dann sind die Gewichte und die Knoten gegeben durch

$$\mathbf{b} = (1/6, 1/3, 1/3, 1/6)^\top, \qquad \mathbf{c} = (0, 1/2, 1/2, 1)^\top$$

und das Gleichungssystem (3.38a) bis (3.38h) liefert das Butcher-Schema

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

Es lassen sich auch Verfahren noch höherer Ordnung konstruieren die aber oft nicht mehr praktikabel sind, es gelten aber folgende Schranken:

**Theorem 3.4.9 (Butcher-Schranken).** Für die maximale erreichbare Ordnung  $p$  eines expliziten Runge/Kutta-Verfahrens mit  $s$  Stufen gilt  $p \leq s$ . Ferner gelten folgende verschärften Abschätzungen für die minimale Stufenzahl  $s_{\min}$

$$\begin{array}{c|cccccccc} p & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ s_{\min} & 1 & 2 & 3 & 4 & 6 & 7 & 9 & 11 \end{array} \begin{array}{l} p \geq 9 \\ s_{\min} \geq p + 3 \end{array}.$$

Beweis. [Strehmel et al. \(2012\)](#) □

### 3.5 Konvergenz

Der Einfachheit halber wird ein gleichabständiges Gitter zu einem Einschrittverfahren betrachtet. Für die Untersuchung von Konsistenz und Konvergenz eines Runge/Kutta-Verfahrens sind zwei Arten von Fehler entscheidend

**Definition 3.5.1 (lokaler Diskretisierungsfehler; Konsistenzordnung).** Sei  $\tilde{u}_{m+1}$  die numerische Lösung eines Einschrittverfahrens mit Startwert  $u_m = y(t_m)$  nach einem Schritt,

$$\tilde{u}_{m+1} = y(t_m) + h\varphi(t_m, u_m; h). \tag{3.41}$$

Dann heißt

$$\eta_{m+1} = \eta(t_m, h) = y(t_{m+1}) - \tilde{u}_{m+1}, \quad m = 0, \dots, N - 1 \tag{3.42}$$

lokaler Diskretisierungsfehler. Ein Einschrittverfahren heißt

(i) konsistent, wenn für alle Anfangswertaufgaben

$$\lim_{h \rightarrow 0} \left\| \frac{\eta(t+h)}{h} \right\| = 0, \quad t \in [t_0, T] \quad (3.43)$$

(i) konsistent von der Ordnung  $p \in \mathbb{N}$ , wenn für eine hinreichend oft stetig partiell differenzierbare rechte Seite  $f$  die Abschätzung

$$\|\eta(t+h)\| \leq Ch^{p+1}, \quad h \in (0, h], \quad t \in [t_0, T - t_h] \quad (3.44)$$

für ein von  $h$  unabhängiges  $C$  gilt.

Anstelle der lokalen Fehlerbetrachtung, kann auch die globale Fehlerentwicklung untersucht werden:

**Definition 3.5.2 (globaler Diskretisierungsfehler; Konvergenzordnung).** Ein Einschrittverfahren heißt

(i) konvergent, wenn für alle Anfangswertprobleme für den globalen Diskretisierungsfehler

$$E_m = y(t_m) - u_h(t_m) \quad (3.45)$$

die Beziehung

$$\max_m \|E_m\| \rightarrow 0, \quad h_{\max} \rightarrow 0$$

gilt.

(i) konvergent von der Ordnung  $p^*$ , wenn die Abschätzung

$$\max_m \|E_m\| \leq Ch_{\max}^{p^*}, \quad h_{\max} \in (0, H], \quad t_m \in [t_0, T], \quad (3.46)$$

mit einer von  $h$  unabhängigen Konstante  $C$  gilt.

Für eine  $p$ -mal stetig differenzierbare Verfahrensfunktion  $\varphi$  folgt aus der Definition der Konvergenzordnung, dass  $\varphi$  genau dann von der Konvergenzordnung  $p^*$  ist, wenn für den lokalen Fehler  $E_m = \mathcal{O}(h^{p^*})$  gilt. △

Der folgende Satz gibt den Zusammenhang zwischen Konsistenzordnung und Konvergenzordnung wider.

**Theorem 3.5.3.** Ein Einschrittverfahren zur Lösung eines Anfangswertproblems sei konsistent mit der Ordnung  $p$ . Ist die Verfahrensfunktion  $\varphi$  Lipschitz-stetig mit Lipschitz-Konstante  $L_\varphi > 0$ , dann ist das Verfahren konvergent mit der Ordnung  $p$ .

*Beweis.* Die Idee des Beweises ist eine Zerlegung des globalen Diskretisierungsfehlers, so dass sich dieser über die lokalen Diskretisierungsfehler abgeschätzt werden kann.

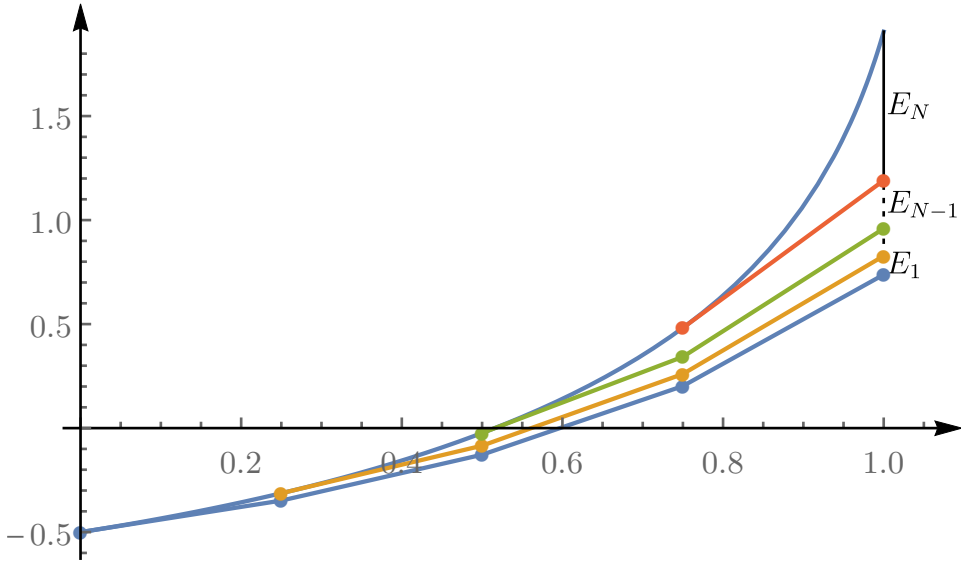


Abb. 3.3: Zerlegung des globalen Diskretisierungsfehler am Beispiel des Euler-Verfahrens zum Anfangsproblem  $y'(t) = e^{y(t)}(1+t)$ ,  $y(0) = -1/2$ .

Nach Definition des globalen Diskretisierungsfehler (3.45) am Gitterpunkt  $t_m$  gilt

$$\begin{aligned}
 E_m &= y(t_m) - u_m \\
 &= y(t_m) - \tilde{u}_m + \tilde{u}_m - u_m \\
 &= y(t_m) - \tilde{u}_m + y(t_{m-1}) + h_{m-1}\varphi(t_{m-1}, y(t_{m-1}); h_{m-1}) \\
 &\quad - u_{m-1} - h_{m-1}\varphi(t_{m-1}, u_{m-1}; h_{m-1}).
 \end{aligned} \tag{3.47}$$

Bezeichne nun  $\tilde{u}_{m+1,n}$  das Ergebnis nach einem Schritt mit exaktem Startvektor  $u_{m,n} = y(t_n)$ , d.h.

$$\tilde{u}_{m+1,n} = y(t_n) + h_m\varphi(t_m, y(t_n); h), \tag{3.48}$$

dann lässt sich der globale Diskretisierungsfehler  $E := E_N = y(t_N) - u_N$  mit (3.47) und (3.48) berechnen als die Summe der globalen Diskretisierungsfehler  $E_m$  mit  $m = 0, \dots, N$ , d.h.

$$E = \sum_{n=0}^{N-1} (\tilde{u}_{N,n} - \tilde{u}_{N,n+1}), \tag{3.49}$$

so dass sich (3.49) auch abschätzen lässt mit

$$\|E\| \leq \sum_{n=0}^{N-1} \|\tilde{u}_{N,n} - \tilde{u}_{N,n+1}\|.$$

Nach Voraussetzung ist die Verfahrensfunktion  $\varphi$  Lipschitz-stetig, so dass

$$\begin{aligned}
 \|\tilde{u}_{m+1,n} - \tilde{u}_{m+1,n+1}\| &\leq \|y(t_n) - y(t_{n+1})\| + h_m L \|y(t_n) - y(t_{n+1})\| \\
 &= (1 + h_m L) \|y(t_n) - y(t_{n+1})\|,
 \end{aligned}$$

so dass



$$\begin{aligned}\|\tilde{u}_{N,n} - \tilde{u}_{N,n+1}\| &\leq e^{L \sum_{i=1}^{m-1} h_i} \|\tilde{u}_{m+1,n} - \tilde{u}_{m+1,n+1}\| \\ &= e^{L \sum_{i=1}^{m-1} h_i} \eta(t_m, h_h)\end{aligned}$$

$$\|E\| \leq e^{L(T-t_0)} -$$

Betrachtet werden weitere numerische Lösungen  $\tilde{u}_{m+1,n}$  mit Startwerten in  $\tilde{u}_{m,n} = y(t_n)$  nach einem Schritt

mit  $\tilde{u}_m = y(t_m)$  die in  $t_m$  auf der exakten Lösung starten. Dann gilt

$$\tilde{u}_{m+1} = \tilde{u}_m + h_m \varphi(t_m, \tilde{u}_m; h),$$

so dass

$$\hat{\eta}_N = \sum_{n=0}^{N-1} \tilde{u}_N - \tilde{u}_N^{n+1}.$$

Dann ist

$$\|e_N\| \leq \sum_{n=0}^{N-1} (u_N^n - u_N^{n+1}) \leq \sum_{n=0}^{N-1} \|u_N^n - u_N^{n+1}\|.$$

Nutzt man die Lipschitz-Stetigkeit aus, dann gilt

$$\begin{aligned}\|u_{m+1}^n - u_{m+1}^{n+1}\| &\leq \|u_{m+1}^n - u_{m+1}^{n+1}\| + h_m L_\phi \|u_m^n - u_m^{n+1}\| \\ &\leq (1 + h_m) L_\phi \|u_m^n - u_m^{n+1}\| \\ &\leq e^{h_m L_\phi} \|u_m^n - u_m^{n+1}\| \\ &\leq e^{L(T-t_0)} \mathcal{O}(h_n^{p+1}),\end{aligned}$$

und dies liefert die Abschätzung

$$\|E\| \leq e^{L(T-t_0)} \sum_{n=0}^{N-1} \mathcal{O}(h_n^{p+1}) = \mathcal{O}(h^p),$$

falls  $h_n = h$  konstant ist, beziehungsweise bei nicht konstanter Schrittweite

$$\begin{aligned}\|E\| &\leq e^{L(T-t_0)} \sum_{n=0}^{N-1} h_n \mathcal{O}(h_{\max}^p) \\ &\leq (T - t_0) e^{L_\phi(T-t_0)} \mathcal{O}(h_{\max}^p),\end{aligned}$$

was zu zeigen war. □

## 3.6 Schrittweitensteuerung

### 3.6.1 Grundprinzip

Häufig werden konstante Schrittweiten betrachtet. Der Satz 3.5.3 besagt, dass die Konvergenz eines Einschrittverfahrens durch die Konsistenz des Verfahrens abgeschätzt werden kann. Dabei ist die Effizienz eines konstruierten Verfahrens von der Schrittweite  $h$  abhängig. Ist ein

vorgegebener globaler Fehler groß, so kann die Schrittweite sehr grob gewählt werden. Soll hingegen der Fehler minimiert werden, so müssen hingegen die Schrittweiten entsprechend klein gewählt werden.

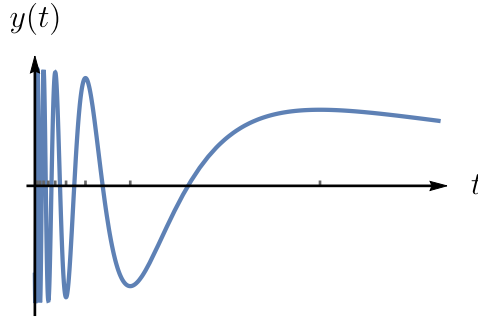


Abb. 3.4: Das Verfahren oszilliert bei kleinen Schrittweiten stark, für große Schrittweiten ist das Verfahren glatt.

Ein Einschrittverfahren, dass die Schrittweite selbst reguliert, sollte also effizienter sein, als eine Rechnung mit konstanten Schrittweiten.

Dabei geht man in der Regel wie folgt vor:

Schritt 1: Schätzung des lokalen Fehlers.

Schritt 2: Modell für lokalen Fehler mit  $\eta_m \approx C(t_m) h_m^{p+1}$ .

Schritt 3: Prognose für die nächste Schrittweite  $h_{m+1}$  mit  $\eta_{m+1} \approx C(t_{m+1}) h_{m+1}^{p+1}$ .

Für die Berechnung zweier Gleitpunktzahlen  $z, w$  gilt nun

$$fl(z \circ w) = (z \circ w)(1 + \varepsilon), \quad \varepsilon = \varepsilon(z, w), \quad |\varepsilon| \leq \text{eps},$$

dabei bezeichnet  $fl$  die in der Gleitpunktrechnung ausgeführte Operation und  $\text{eps} \approx 10^{-16}$  die relative Computergenauigkeit. Für die Verfahrensfunktion  $\varphi(t_m, u_m; h_m)$  eines Einschrittverfahrens gilt damit für die Gleitpunktrechnung

$$fl(\varphi) = \varphi(1 + \varepsilon_1), \quad |\varepsilon_1| \leq C \text{eps},$$

wobei  $C$  eine für  $h_m \rightarrow 0$  beschränkte Konstante ist. Damit ist aber auch

$$\begin{aligned} fl(h_m fl(\varphi)) &= h_m fl(\varphi)(1 + \varepsilon_2) = h_m \varphi(1 + \varepsilon_1)(1 + \varepsilon_2) \\ fl(u_m + fl(h_m fl(\varphi))) &= (u_m + h_m \varphi(1 + \varepsilon_1)(1 + \varepsilon_2))(1 + \varepsilon_3) \\ &= (u_m + h_m) \varphi(1 + \varepsilon_3) + h_m \varphi(\varepsilon_1 + \varepsilon_2) + \mathcal{O}(\text{eps}^2), \end{aligned}$$

d.h. die computeralgebraische Auswertung berechnet nicht die Näherungslösung  $u_{m+1}$ , sondern eine verfälschte Näherungslösung  $v_{m+1}$ , und bei jedem Schritt addieren sich die Abweichungen zwischen den Näherungslösungen  $u_{m+1}, v_{m+1}$ , so dass man bestrebt ist die Schrittweiten so zu wählen, dass der addierte Rundungsfehler klein bleibt.

*Beispiel 3.6.1.* Betrachtet wird ein Euler-Verfahren und die Taylor-Entwicklung der exakten Lösung für  $t_{m+1} = t + h = t_m + h$ , d.h.

$$\begin{aligned} u_{m+1} &= u_m + hf(u_m) \\ y(t+h) &= u_m + hf(u_m) + \frac{h^2}{2} f_y f(u_m) + \dots \end{aligned}$$

Dann gilt für den lokalen Diskretisierungsfehler

$$\begin{aligned} \eta(t_m, h) &= y(t+h) - \tilde{u}_{m+1} \\ &= \frac{1}{2} f_y f(u_m) h^2 + \mathcal{O}(h^3) \\ &= C(t) h^2 + \mathcal{O}(h^3) \end{aligned} \tag{3.50}$$

Für die Prognose der neuen Schrittweite wird davon ausgegangen, dass

$$C(t_m) \approx C(t_{m+1})$$

gilt. Das heißt, dass die Schrittweite so gewählt wird, dass der lokale Fehler um einen vorgegebenen Toleranzwert  $\kappa_m \approx C(t_m) h_{m+1}^{p+1}$  liegt, das heißt

$$\frac{\kappa_m}{\eta_m} \approx \left( \frac{h_{m+1}}{h_m} \right)^{p+1},$$

also

$$h_{m+1} = h_m \left( \frac{\kappa_m}{\eta_m} \right)^{\frac{1}{p+1}}.$$

In der Praxis zerlegt man den Toleranzvektor  $\kappa$  mit den Komponenten  $\kappa_i$  in einen absolut und einem relativ Teil,  $\kappa_{a,i}$ ,  $\kappa_{r,i}$ . Da bei der numerischen Lösung Genauigkeits- und Rundungsfehler auftreten können, kann der lokale Fehler  $\eta_m$  Null annehmen, so dass oft der *relativen Fehler*

$$\rho_i := \frac{\eta_i}{\kappa_{a,i} + \kappa_{r,i} |u_i|} \tag{3.51}$$

der  $i$ -ten Komponente betrachtet wird. In diesem Fall gilt für die neue Schrittweite

$$h_{m+1} = h_m \frac{1}{\|\rho_i\|^{1/p+1}}, \tag{3.52}$$

wobei häufig die Toleranz für alle Komponenten gleich gewählt wird. Dabei wird in der Praxis häufig noch ein Sicherheitsfaktor  $0.8 \lesssim \beta \lesssim 0.9$  hinzumultipliziert, so dass sich die neue Schrittweite zu

$$h_{neu} = \frac{\beta}{\|\rho\|^{1/p+1}} h =: \alpha \cdot h$$

Dabei darf  $h$  weder zu schnell wachsen, noch zu schnell fallen. In diesem Fall als Vorschlag für die neue Schrittweite

$$h_{neu} = h \cdot \tilde{\alpha}, \quad \tilde{\alpha} := \min \left\{ \alpha_{\max}, \max \{ \alpha_{\min} \}, \alpha (1/\rho)^{1/(p+1)} \right\}$$

nehmen, wobei üblicherweise  $2 \lesssim \alpha_{\max} \lesssim 4$  sowie  $\alpha_{\min} \approx 0.25$  gewählt wird. Ist  $\tilde{\alpha}$  sehr groß oder sehr klein, dann ändert sich das Lösung sehr stark und das bisher betrachtete Fehlermodell ist nur bedingt gültig.

Das Grundsalgorithmus eines ODE-Lösers ist

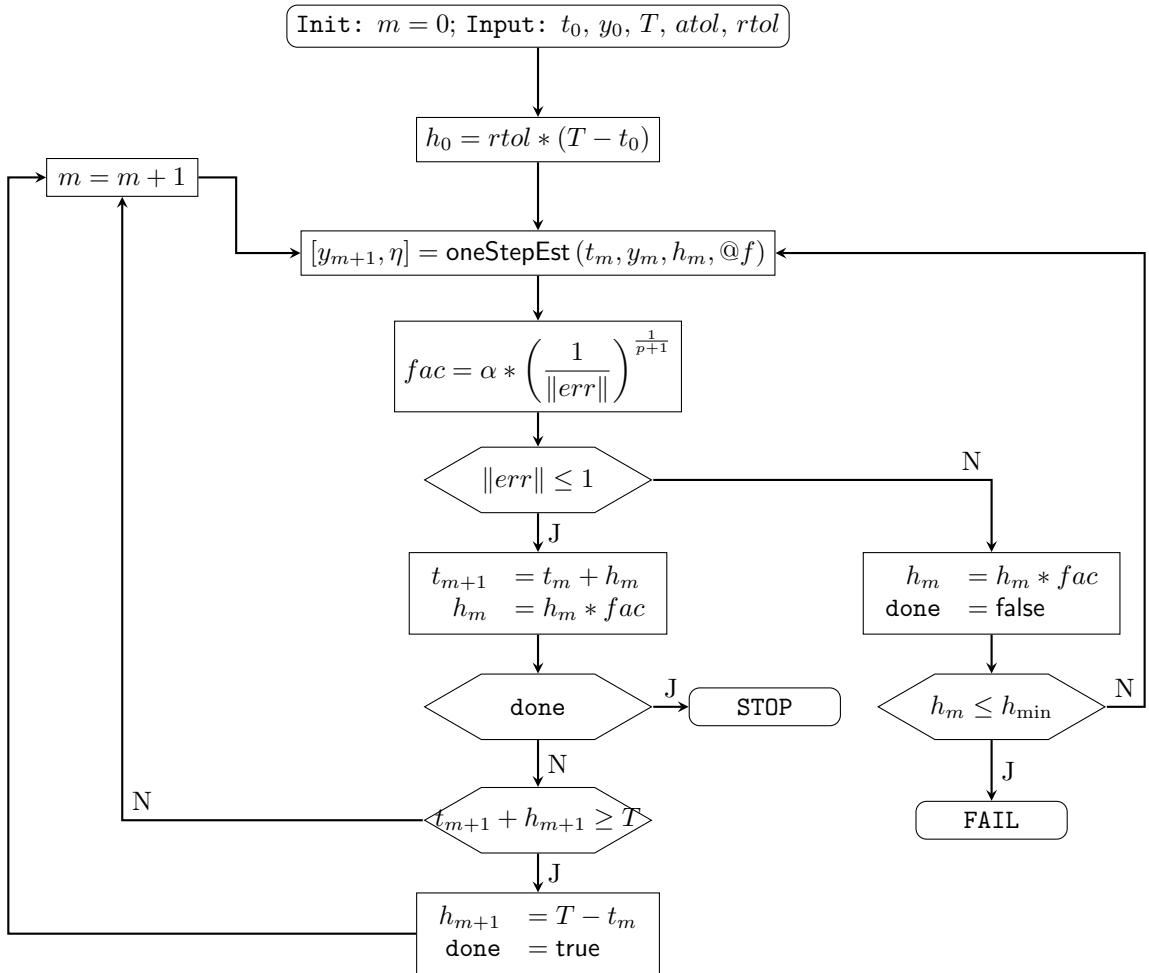


Abb. 3.5

### 3.6.2 Fehlerschätzung mittels Richardson-Extrapolation

Betrachtet wird ein  $s$ -stufiges, explizites Runge/Kutta-Verfahren mit Konsistenzordnung  $p$ . Bezeichne  $\tilde{y}(t)$  die exakte Lösung zum Startwert  $\tilde{y}(t_m) = u_m$ . Zur Berechnung des lokalen Fehlers werden zwei Näherungslösungen an der Stelle  $t_m + h$  mit

$$u_h = u_h(t_m + h), \quad u_{2 \times h/2} = u_{h/2}(t_m + h)$$

mit  $u_{h/2} = u_{h/2}(t_m + h/2)$  als Näherungslösung nach einem Schritt (vgl. Abbildung 4.1).

Die beiden Näherungslösungen unterscheiden sich dadurch, dass die linke mittels einer ganzen Schrittweite  $h$  berechnet wird, die rechte mittels zweier Schritte und halber Schrittweite  $h/2$ . Für den lokalen Fehler gilt dann

$$\tilde{y}(t_m + h) - u_h = C(t_m) h^{p+1} + \mathcal{O}(h^{p+2}), \quad (3.53)$$

$$\tilde{y}(t_m + h) - u_{2 \times h/2} = 2C(t_m) \left(\frac{h}{2}\right)^{p+1} + \mathcal{O}(h^{p+2}). \quad (3.54)$$

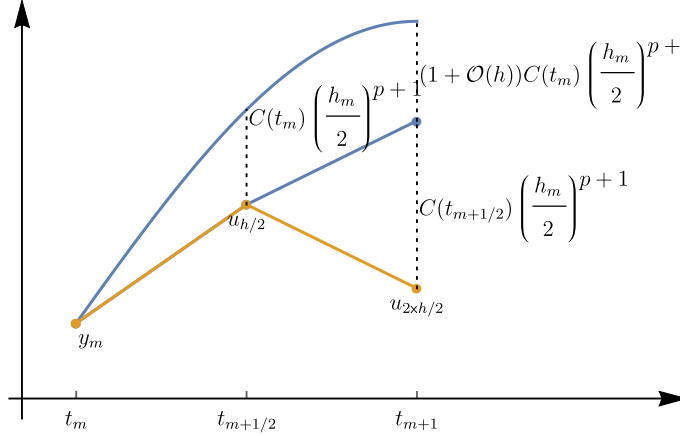


Abb. 3.6

Subtrahiert man (3.54) von (3.53), dann erhält man

$$\begin{aligned} u_h(t_m + h) - u_{2 \times h/2}(t_m + h) &= C(t_m) h^{p+1} - \frac{1}{2^p} C(t_m) h^{p+1} + \mathcal{O}(h^{p+2}) \\ &= \frac{2^{p-1} - 1}{2^p} C(t_m) h^{p+1} + \mathcal{O}(h^{p+2}) \end{aligned}$$

und nach Umstellen

$$2C(t_m) = \frac{u_{2 \times h/2} - u_h}{2^{p-1}} \left(\frac{h}{2}\right)^{-(p+1)} + \mathcal{O}(h),$$

und dies die Fehlerschätzung mittels

$$\tilde{y}(t_m + h) - u_{2 \times h/2} = \frac{u_{2 \times h/2} - u_h}{2^{p-1}} + \mathcal{O}(h^{p+2}). \quad (3.55)$$

Führt man die Extrapolation durch den Übergang von  $u_h$  und  $u_{2 \times h/2}$  zu

$$w_h = u_{2 \times h/2} + \frac{u_{2 \times h/2} - u_h}{2^{p-1}}$$

aus, dann erhält man eine Näherungslösung mit Konsistenz  $p + 1$  für die Lösung  $\tilde{y}(t)$ . Da sich das Grundverfahren geändert hat, ändert sich auch der Fehlerschätzer. Hat man den Fehlerschätzer für  $u_{2 \times h/2}$ , und nutzt den genaueren Wert  $w_h$ , dann hat man keinen Fehlerschätzer.

In der Praxis werden häufig absolute  $\kappa_a := atol$  und relative Toleranzen  $\kappa_r := rtol$  vorgegeben, so dass für den Fehler dann lokalen Fehler dann

$$\left\| \frac{1}{sk} (\tilde{y}(t_m + h) - u_{2 \times h/2}) \right\| \approx \left\| \frac{1}{2^{p-1} - 1} \frac{1}{sk} (u_{2 \times h/2} - u_h) \right\| =: err \leq \kappa, \quad sk = \kappa_a + \max \{ |u_m|, |u_{2 \times h/2}| \}$$

gilt. Die neue Schrittweite  $h_{neu}$  wird dabei so gewählt, so dass der Fehler (3.54) einer vorgegebenen Genauigkeit  $\kappa$  entspricht. Ist  $err > 1$ , dann wird die Schrittweite verkleinert. Ist  $err < 1$ , dann wird der Schritt  $t_m \rightarrow t_m + h$  mit  $u_{2 \times h/2}$  und  $w_h$  durchgeführt.

Da bei der Extrapolation auch ein Zusammenhang zwischen der Toleranz und dem globalen Fehler existiert mit  $\eta_m \approx \kappa h$  gilt, lässt sich der globale Fehler  $E$  auch Abschätzen mit

$$E \approx \sum_{m=0}^{N-1} \eta_m \approx \kappa (T - t_0) =: Err. \quad (3.57)$$

### 3.6.3 Fehlerschätzung mittels Einbettung

Die Idee der Fehlerschätzung durch Einbettung ist ein Runge/Kutta-Verfahren durch die Einführung zusätzlicher Stufen eines zweiten Runge/Kutta-Verfahren verschiedener Ordnung bei gleichem Knotenvektor  $\mathbf{c}$  und gleicher Verfahrensmatrix aber unterschiedlichen Gewichten  $b_i$  und  $\hat{b}_i$ . Das zugehörige Parameterschema ist dann

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & \\ c_3 & a_{31} & a_{32} & & \\ \vdots & \vdots & & \ddots & \\ c_s & a_{s1} & a_{s2} & \dots & a_{s,s-1} \\ \hline & b_1 & b_2 & \dots & b_s \\ \hline & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_s \end{array}$$

Vereinfachend können autonome Differentialgleichungen betrachtet werden. Dann

**Definition 3.6.2.** Sei  $s \in \mathbb{N}$ . Ein eingebettetes Runge/Kutta-Verfahren ist ein Einschrittverfahren der Gestalt

$$\begin{aligned} u_{m+1} &= u_m + h \sum_{i \leq s} b_i f(u_{m+1}^{(i)}), \quad u_{m+1}^{(i)} = u_m + h \sum_{j \leq s} a_{ij} f(u_{m+1}^{(j)}) \\ \hat{u}_{m+1} &= u_m + h \sum_{i \leq s} \hat{b}_i f(u_{m+1}^{(i)}) \end{aligned} \quad (3.58)$$

wobei beide Verfahren Näherungslösungen  $u_{m+1}$  und  $\hat{u}_{m+1}$  von verschiedener Konsistenzordnung  $p$  und  $q$  sind.

Für den lokalen Fehler des eingebetteten Verfahrens sind zwei Fälle zu unterscheiden.

- $q > p$ : Sei  $q > p$ , dann gilt für den lokalen Fehler des Verfahrens  $u_{m+1}$  der Ordnung  $p$

$$\begin{aligned} \tilde{y}(t_m + h) - u_{m+1} &= C(t_m) h^{p+1} + \mathcal{O}(h^{p+2}) \\ &= \tilde{y}(t_m + h) - \hat{u}_{m+1} + \hat{u}_{m+1} - u_{m+1} \\ &= \hat{u}_{m+1} - u_{m+1} + \mathcal{O}(h^{q+1}), \end{aligned}$$

womit

$$\hat{u}_{m+1} - u_{m+1} = C(t_m) h^{p+1} + \mathcal{O}(h^{p+2}), \quad q > p. \quad (3.59)$$

- $q < p$ : Der Fall  $q < p$  ist analog.

Die Differenz  $\hat{u}_{m+1} - u_{m+1}$  ist damit ein Schätzer für den lokalen Fehler  $\eta_{m+1}$  des Runge/Kutta-Verfahrens der Konsistenzordnung  $q^* = \min\{p, q\}$  und es gilt analog zu (3.56), dass  $err = \left\| \frac{1}{sk} (u_{m+1} - \hat{u}_{m+1}) \right\|$ , woraus sich die Schrittweite aus

$$h_{neu} = h \cdot \tilde{\alpha}, \quad \tilde{\alpha} := \min \left\{ \alpha_{\max}, \max \{ \alpha_{\min} \}, \alpha (1/err)^{1/(q^*+1)} \right\}.$$

Dieser Fehleransatz ist in der Regel mit geringeren Rechenaufwand als bei der Fehlerabschätzung mittels Richardson-Extrapolation durchführbar. △

*Beispiel 3.6.3 (Fehlberg-Konstruktion; Runge/Kutta/Fehlberg-Verfahren).* Ein eingebettetes Runge/Kutta-Verfahren mit  $p(q)$  oder RKFP( $q$ ) bedeutet, dass das Hauptverfahren von der Konsistenzordnung  $p$  ist, der Fehlerschätzer  $\hat{u}_{m+1}$  von der Konsistenzordnung  $q$ . Da nun  $f(u_{m+1})$  im nächsten Schritt benötigt wird, kann es vorher ausgerechnet werden und für die Gewichte  $\hat{b}_i$  verwendet werden. Dies wird auch als *Fehlberg-Konstruktion* bezeichnet, wobei die von Fehlberg konstruierten Verfahren mit  $q > p$  entwickelt wurden.

- RKF1 (2): Ein 2-stufiges Runge/Kutta/Fehlberg-Verfahren mit 1 (2) besitzt das Parameterschema

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

- RKF1 (2): Ein 2-stufiges Runge/Kutta/Fehlberg-Verfahren mit 1 (2) besitzt das Parameterschema

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

Für das Verfahren RKF4 (5) hat man vereinfachend  $u_{m+1} = y(t_{m+1}) + \mathcal{O}(h^5)$  und  $\hat{u}_{m+1} = y(t_{m+1}) + \mathcal{O}(h^6)$ , und der Schätzer  $\hat{u}_{m+1} - u_{m+1}$  für den Hauptteil des lokalen Diskretisierungsfehler ist dann sehr klein. Dieses Verfahren hat jedoch lediglich historische Bedeutung, da man zwar den Fehler für das Verfahren geringere Ordnung schätzt, aber man eigentlich an Verfahren mit höherer Konsistenz interessiert ist, wie zum Beispiel △

*Beispiel 3.6.4 (Verfahren von Dormand/Prince mit 5 (4); DOPRI5 (4)).* Das von Dormand und Prince entwickelte Verfahren ist ein 7-stufiges verfahren. Wendet man die Fehlberg-Konstruktion an, dann sind aber je Schritt nur sechs Funktionsaufrufe benötigt. Denn das Verfahren ist so konstruiert, so dass  $a_{7i} = b_i$  für  $i = 1, \dots, 6$  gilt, und  $b_7 = 0$ . Damit gilt  $u_{m+1} = u_{m+1}^{(7)}$  und der letzte Funktionsaufruf  $f(u_{m+1}^{(7)})$  ist identisch mit dem ersten Funktionsaufruf  $f(u_{m+1})$  im nächsten Schritt.

### 3.7 Stetige explizit Runge/Kutta-Verfahren

Damit ein Verfahren möglichst effizient ist, ist man bestrebt bei vorgegebener Genauigkeit  $\kappa = \text{tol}$  mit möglichst großen Schrittweiten durchzuführen. Allerdings liegt die Lösung  $u_h$  häufig auf einem inhomogenen Gitter außerhalb des durch die Schrittweitensteuerung erzeugten Punktgitters  $I_h$ . Dementsprechend ist man daran interessiert, dass ein Runge/Kutta-Verfahren auch eine Näherungslösung für den Zeitpunkt

$$t = t_m + \theta h, \quad 0 < \theta \leq 1 \quad (3.60)$$

liefert.

**Definition 3.7.1 (Runge/Kutta-Verfahren; stetig-explizites).** Eine Interpolationsvorschrift –

$$\begin{aligned} v(t_m + \theta h) &= u_m + h \sum_{i \leq s} \widehat{b}_i(\theta) f\left(t_m + c_i h, u_{m+1}^{(i)}\right) \\ u_{m+1}^{(i)} &= u_m + h \sum_{j=1}^{i-1} a_{ij} f\left(t_m + c_j h, u_{m+1}^{(j)}\right), \quad i = 1, \dots, s \end{aligned} \quad (3.61)$$

heißt stetiges explizites Runge/Kutta-Verfahren, wenn die Gewicht Polynome  $\widehat{b}_i(\theta)$ ,  $i = 1, \dots, s$  mit

$$\widehat{b}_i(0) = 0, \quad \widehat{b}_i(1) = b_i$$

sind.

*Beispiel 3.7.2.* Betrachtet wird eine lineare Interpolation mit den Stützstellen  $(t_m, t_{m+1})$  und den Stützwerten  $(u_m, u_{m+1})$ . Ein stetiges explizites  $s$ -stufige Runge/Kutta-Verfahren mit linearem Interpolationspolynom

$$v(t_m + \theta h) = u_m + \theta(u_{m+1} - u_m) \quad (3.62)$$

ist dann gegeben durch

$$\begin{aligned} v(t_m + \theta h) &= u_m + \theta h \sum_{i \leq s} \widehat{b}_i f\left(t_m + c_i h, u_{m+1}^{(i)}\right), \quad 0 \leq \theta \leq 1 \\ u_{m+1}^{(i)} &= u_m + h \sum_{i \leq s} \widehat{b}_i f\left(t_m + c_i h, u_{m+1}^{(i)}\right) \end{aligned} \quad (3.63)$$

**Definition 3.7.3.** Für eins-stufiges Runge/Kutta-Verfahren mit der Ordnung  $p \geq 1$  ist das stetige  $s$ -stufige Runge/Kutta-Verfahren von der gleichmäßigen Ordnung  $q$ , wenn für den lokalen Fehler

$$y(t_m + \theta h) - v(t_m + \theta h), \quad v(t_m) = y(t_m)$$

die Abschätzung

$$\|y(t_m + \theta h) - v(t_m + \theta h)\| = \mathcal{O}(h^{q+1}), \quad \forall \theta \in [0, 1]$$

gilt.



Die Ordnung bestimmt lässt sich nun aus der Taylor-Entwicklungen über elementare Differentiale zu

$$y(t_m + \theta h) = u_m + \sum_{k \leq q} \frac{h^k}{k!} \theta^k \sum_{\tau \in \mathcal{LT}_k} F(\tau) + \mathcal{O}(h^{q+1})$$

$$v(t_m + \theta h) = u_m + \sum_{k \leq q} \frac{h^k}{k!} \sum_{\tau \in \mathcal{LT}_k} \gamma(\theta) \widehat{\Phi}_{s+1}(\theta)(\tau) F(\tau) + \mathcal{O}(h^{q+1})$$

**Theorem 3.7.4.** *Ein  $s$ -stufiges stetiges Runge/Kutta-Verfahren hat die gleichmäßige Ordnung  $q$ , wenn*

$$\widehat{\Phi}(\theta)(\tau) = \frac{\theta^k}{\gamma(\tau)}, \forall \tau \in \mathcal{LT}_k, 1 \leq k \leq q.$$

*Beispiel 3.7.5.* Die Bedingungen für ein stetiges Runge/Kutta-Verfahren der gleichmäßigen Höchstordnung  $q = 3$  lauten (vgl. Tabelle 3.1)

$p$	$\tau$	Bedingung	$\widehat{\Phi}(\tau)$
1	$\bullet$	$\theta = \widehat{b}_1(\theta) + \widehat{b}_2(\theta) + \widehat{b}_3(\theta)$	
2	$\dagger$	$\frac{\theta^2}{2} = \widehat{b}_2(\theta) c_2 + \widehat{b}_3(\theta) c_3$	
3	$\forall$	$\frac{\theta^3}{3} = \widehat{b}_2(\theta) c_2^2 + \widehat{b}_3(\theta) c_3^2$	
	$\vdots$	$\frac{\theta^3}{6} = \widehat{b}_3(\theta) a_{32} c_3$	



## Extrapolationsverfahren

Extrapolationsverfahren sind Verfahren die der Erhöhung der Konvergenzordnung dienen. Die Grundidee der Richardson-Extrapolation ist mehrere Lösungen in  $t_{m+1} = t_m + H$  eines Problems für unterschiedliche Schrittweiten zu berechnen.

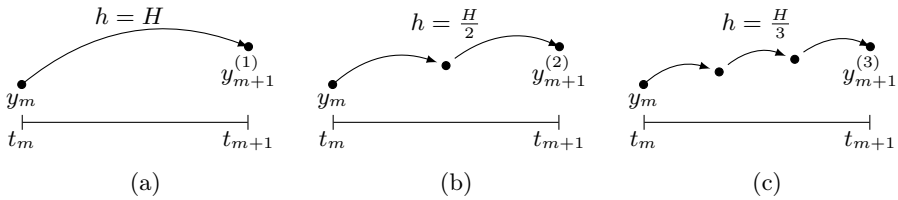


Abb. 4.1: Stufenwerte in Abhängigkeit von  $h$

### 4.1 Asymptotische Entwicklung des globalen Fehlers

Betrachtet wird dazu ein  $s$ -stufiges Runge/Kutta-Verfahren der Ordnung  $p$  mit der Verfahrensfunktion  $\Phi$ . Gesucht ist eine numerische Lösung  $u_h(t) = u_h(m \cdot h)$  ausgehend vom Startpunkt  $(0, y_0)$  mit Schrittweiten  $h$ . Es bezeichnet wie im

**Definition 4.1.1 (Diskretisierungsfehler; globaler).** Sei  $u_h(t) = y_m$  mit  $t = m \cdot h$  die numerische Lösung in  $t$ . Dann heißt

$$\eta_h^*(t) = y(t) - u_h(t) \quad (4.1)$$

globaler Diskretisierungsfehler.

Angenommen auf einem gleichabständigen Gitter besitze ein Einschrittverfahren der Ordnung  $p$  zur Lösung des Anfangswertproblems

$$\dot{y} = f(t, y(t)), \quad y(t_0) = 0$$

die Gestalt

$$u_h(t+h) = u_h + h\varphi(t, u_h(t); h), \quad u_h(t_0) \equiv y_0, \quad t \in [t_0, T] \quad (4.2)$$

Dann gilt nach Definition des lokalen Diskretisierungsfehler 3.1.3 und der Konsistenzordnung ??

$$\eta(t+h) = y(t+h) - y(t) - h\varphi(t, y(t); h) = \mathcal{O}(h^{p+1}), \quad h \rightarrow 0,$$

falls  $\Phi$  auf dem Intervall  $I_h$  hinreichend glatt ist. Dies ist für Runge/Kutta-Verfahren dann erfüllt, wenn  $f(t, y)$  hinreichend glatt ist. Daraus folgt, dass eine asymptotische  $h$ -Entwicklung des lokalen Diskretisierungsfehlers

$$\eta(t+h) = d_{p+1}(t)h^{p+1} + d_{p+2}(t)h^{p+2} + \dots + d_{N+1}(t)h^{N+1} + \mathcal{O}(h^{N+2}) \quad (4.3)$$

Für die asymptotische  $h$ -Entwicklung des globalen Fehlers  $\eta_h^*(t)$  gilt dann der folgende Satz:

**Theorem 4.1.2.** *Sei die rechte Seite  $f(t, y)$  eines Anfangswertproblems, und die zugehörige Verfahrensfunktion  $\Phi$  eines Einschrittverfahrens hinreichend glatt. Dann hat der globale Fehler  $\eta_h^*(t)$  nach  $n$  Schritten in  $t^* = t_0 + nh$  eine asymptotische  $h$ -Entwicklung der Form*

$$\eta_h^*(t^*) = e_p(t^*)h^p + e_{p+1}(t^*)h^{p+1} + \dots + e_N(t^*)h^N + E_{n+1}(t^*, h), \quad t^* = t_0 + nhm \quad (4.4)$$

wobei das Restglied  $E_{n+1}(t^*, h)$  für  $0 < h \leq h_0$  beschränkt ist.

*Beweis.* Zunächst ist zu zeigen, dass

$$y(t) - u_h(t) = e_p(t_m)h^p + \mathcal{O}(h^{p+1}) \quad (4.5)$$

gilt. Sei

$$\widehat{u}_h(t) := u_h(t) + e_p(t)h^p$$

ein neues Einschrittverfahren mit Verfahrensfunktion  $\widehat{\varphi}$ . Für dieses neue Verfahren soll dann die Beziehung

$$\begin{array}{ccc} u_h(t) & \xrightarrow{\Phi} & u_h(t+h) \\ \uparrow e_p(t)h^p & & \uparrow e_p(t+h)h^p \\ \widehat{u}_h(t) & \xrightarrow{\widehat{\varphi}} & \widehat{u}_h(t+h) \end{array}$$

Interpretiert man dies als neues Einschrittverfahren mit Verfahrensvorschrift  $\widehat{\varphi}$ , dann gilt

$$\widehat{u}_h(t+h) = \widehat{u}_h(t) + h\widehat{\varphi}(t, \widehat{u}_h(t); h), \quad \widehat{u}_h(t_0) = y_0 \quad (4.6)$$

Der Zusammenhang zwischen (4.2) und (4.6) ist in Abbildung 4.2 dargestellt, und folglich besitzt die Verfahrensfunktion  $\widehat{\varphi}$  die Form

$$\begin{aligned} \widehat{\varphi}(t, \widehat{u}_h(t); h) &= \varphi(t, u_h(t); h) + (e_p(t+h) - e_p(t))h^{p-1} \\ &= \varphi(t, \widehat{u}_h(t) - e_p(t)h^p; h) + (e_p(t+h) - e_p(t))h^{p-1}. \end{aligned}$$

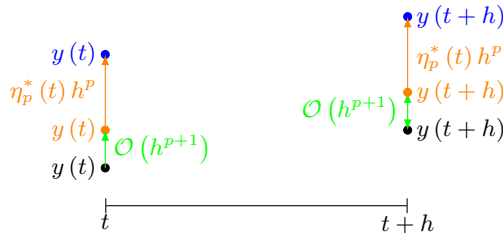


Abb. 4.2

Angenommen, die Verfahrensfunktion  $\widehat{\varphi}$  soll ein Verfahren der Ordnung  $p+1$  liefern. Dann gilt für den lokalen Diskretisierungsfehler von (4.6)

$$\begin{aligned}\widehat{\eta}(t+h) &= y(t+h) - y(t) - h\widehat{\varphi}(t, y(t); h) \\ &= y(t+h) - y(t) - h\varphi(t, y(t) - e_p(t)h^p; h) - (e_p(t+h) - e_p(t))h^p \\ &= y(t+h) - y(t) - h\varphi(t, y(t); h) - (e_p(t+h) - e_p(t))h^p \\ &\quad + h(\varphi(t, y(t); h) - \varphi(t, y(t) - e_p(t)h^p; h)).\end{aligned}$$

Mit den Entwicklungen

$$\begin{aligned}\varphi(t, y(t); h) - \varphi(t, y(t) - e_p(t)h^p; h) &= \frac{d}{dy}\varphi(t, y(t) - e_p(t)h^p; h)e_p(t)h^p + \mathcal{O}(h^{2p}) \\ e_p(t+h) - e_p(t) &= e'_p(t)h + \mathcal{O}(h^2)\end{aligned}$$

erhält man für den lokalen Diskretisierungsfehler

$$\widehat{\eta}(t+h) = \left(d_{p+1}(t) + \frac{d\varphi}{dy}(t, y(t); h)e_p(t) - e'_p(t)\right)h^{p+1} + \mathcal{O}(h^{p+2}). \quad (4.7)$$

Da das Verfahren konsistent sein soll, muss die Bedingung  $\varphi(t, y(t); 0) = f(t, y(t))$  gelten, und dies liefert

$$\frac{d\varphi}{dy}(t, y(t); h) = f_y(t, y(t)),$$

eingesetzt in (4.7) erhält man dann

$$\widehat{\eta}(t+h) = (d_{p+1}(t) + f_y(t, y(t))e_p(t) - e'_p(t))h^{p+1} + \mathcal{O}(h^{p+2}). \quad (4.8)$$

Fasst man  $e'_p(t)$  als Anfangswertproblem auf,

$$e'_p(t) = f_y(t, y(t))e_p(t) + d_{p+1}(t), \quad e_p(t_0) = 0 \quad (4.9)$$

auf, dann ist (4.6) ein Verfahren der Ordnung  $p+1$  und für den globalen Diskretisierungsfehler  $\widehat{\eta}_h^*(t^*) = y(t^*) - \widehat{u}_h(t^*)$  gilt dann  $\widehat{\eta}_h^*(t^*) = \mathcal{O}(h^{p+1})$  und damit

$$\eta_h(t^*) - e_p(t^*)h^p = \mathcal{O}(h^{p+1}),$$

womit die Behauptung gezeigt wurde. □

*Anmerkung 4.1.3.* Ohne Beschränkung kann für das Restglied  $E_{N+1}(t^*, h) = \mathcal{O}(h^{N+1})$  angenommen werden kann. Ferner

ist  $u_h(t)$  und damit  $\eta_h(t)$  nur für  $\frac{t}{h} = M \in \mathbb{N}$  definiert.

Die Eigenschaft (4.9) gilt wegen:

**Theorem 4.1.4.** *Gilt für die asymptotische  $h$ -Entwicklung des lokalen Fehlers*

$$\eta(t+h) = d_{p+1}(t)h^{p+1} + \mathcal{O}(h^{p+2}),$$

so genügt der führende Term in der  $h$ -Entwicklung des globalen Fehlers dem Anfangswertproblem

$$e'_p(t) = f_y(t, y(t))e_p(t) + d_{p+1}(t), \quad e_p(t_0) = y(t_0)$$

*Anmerkung 4.1.5.* Vereinfachend werde die autonome Differentialgleichung  $y = f(y)$  betrachtet. Für verschwindende Störungen gilt

$$\delta'_y = f_y \delta_y,$$

da

$$\begin{aligned} y'_1 &= f(y_1) \rightarrow (y_1 - y_2)' = f_y(y_1 - y_2) + \mathcal{O}(\|y_1 - y_2\|^2) \\ y'_2 &= f(y_2) \end{aligned}$$

## 4.2 Extrapolationsschritte

Zu einem Einschrittverfahren  $\Phi$  der Ordnung  $p$  sei eine Makroschrittweite  $H > 0$ , und eine monoton fallende Folge von Mikroschritten

$$\{h_1, h_2, \dots\}, \quad h_i = \frac{H}{n_i}, \quad n_i \in \mathbb{N}, \quad n_i < n_{i+1}. \quad (4.10)$$

wie in den Abbildungen 4.14.1a, 4.1b und 4.1c. Zu diesem Einschrittverfahren  $\Phi$  werden nun die Gitterpunkte von  $[t_m, t_m + H]$  für alle  $i = 1, \dots, M$  berechnet mit

$$T_{i,m+1} := u_{h_i}(t_m + H) \quad (4.11)$$

als numerische Lösung für  $y(t_0 + H)$ . Für  $n_1 = 1$ ,  $n_2 = 2$  und  $n_3 = 3$  mit  $p = 1$  erhält man beispielsweise mittels Reihenentwicklung

$$\begin{aligned} T_{1,m+1} &= y(t_m + H) + e_1(t_m + H)H + e_2(t_m + H)H^2 + \mathcal{O}(H^4) \\ T_{2,m+1} &= y(t_m + H) + e_1(t_m + H)\frac{H}{2} + e_2(t_m + H)\frac{H^2}{4} + \mathcal{O}(H^4) \\ T_{3,m+1} &= y(t_m + H) + e_1(t_m + H)\frac{H}{3} + e_2(t_m + H)\frac{H^2}{9} + \mathcal{O}(H^4) \end{aligned}$$

mit  $e_3(t_m + H)H^3 + \mathcal{O}(H^4) = \mathcal{O}(H^4)$ . Dies lässt sich kompakt schreiben als

$$\begin{pmatrix} T_{1,m+1} \\ T_{2,m+1} \\ T_{3,m+1} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \frac{1}{2} & \frac{1}{4} \\ 1 & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \begin{pmatrix} y(t_m + H) \\ e_1(t_m + H)H \\ e_2(t_m + H)H^2 \end{pmatrix} + \mathcal{O}(H^{p+M-1})$$

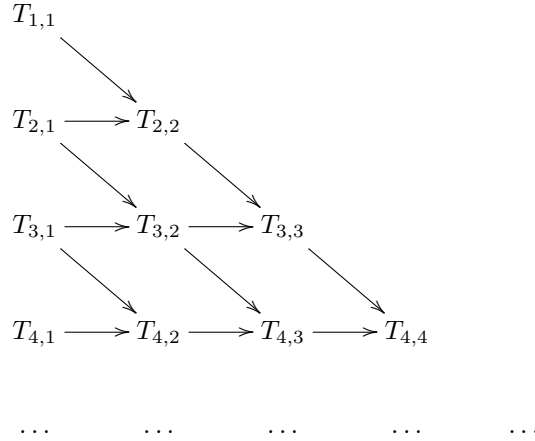
Mit (4.10) führt dies zu einem allgemeinen, linearen Gleichungssystem

$$\begin{pmatrix} T_{1,m+1} \\ T_{2,m+1} \\ \vdots \\ T_{M,m+1} \end{pmatrix} = \begin{bmatrix} 1 & \frac{1}{n_1^p} & \frac{1}{n_1^{p+1}} & \cdots & \frac{1}{n_1^{p+M-2}} \\ 1 & \frac{1}{n_2^p} & \frac{1}{n_2^{p+1}} & \cdots & \frac{1}{n_2^{p+M-2}} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \frac{1}{n_M^p} & \frac{1}{n_M^{p+1}} & \cdots & \frac{1}{n_M^{p+M-2}} \end{bmatrix} \begin{pmatrix} y(t_m + H) \\ e_p(t_m + H)H^p \\ e_{p+1}(t_m + H)H^{p+1} \\ \vdots \\ e_{p+M-2}(t_m + H)H^{p+M-2} \end{pmatrix} + \mathcal{O}(H^{p+M-1}),$$

wobei die Koeffizientenmatrix regulär ist. Dieses lineare Gleichungssystem lässt sich nun mittels den üblichen Mitteln lösen. Eleganter hingegen ist die Anwendung des *Aitken/Neville-Algorithmus*

$$T_{l,k+1} = T_{l,k} + \frac{T_{l,k} - T_{l-1,k}}{(n_l/n_{l-k}) - 1}. \quad (4.12)$$

Hierzu können die  $T_{i,k}$  in einem *Extrapolationstableau*, dem *Neville-Tableau*, angeordnet werden:



Zur Berechnung des Extrapolationstableau können verschiedene Verfeinerungsfolgen  $F = \{n_i\}_{i \leq M}$  gewählt werden:

- *harmonische Folge*

$$F_H = \{1, 2, 3, 4, 5 \dots\}, n_i = i$$

- *Romberg-Folge*

$$F_R = \{1, 2, 4, 6, 8 \dots\}, n_i = 2^{i-1}$$

- *Bulirsch-Folge*

$$F_B = \{1, 2, 3, 4, 6, 12, \dots\}, n_i = 2n_{i-2}, i \geq 4$$

Mittels Extrapolationsverfahren können nun aus einfachen Verfahren mit niedriger Ordnung Verfahren mit höheren Ordnungen konstruiert werden. Wählt man für ein Extrapolationsverfahren ein explizites Runge/Kutta-Verfahren, dann entspricht jedes  $T_{l,k}$  wieder einem expliziten Runge/Kutta-Verfahren. Legt man beispielsweise für die Extrapolation ein explizites Euler-Verfahren mit harmonische Verfeinerungsfolge  $F_H$  zugrunde, dann wird  $f$

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

mal oft aufgerufen. Für  $T_{3,3}$  erhält man beispielsweise ein Runge/Kutta-Verfahren der Ordnung 3. Allgemein gilt

**Theorem 4.2.1.** *Sei  $f(t, y)$  hinreichend glatt. Liegt der Extrapolation ein explizites Euler-Verfahren mit harmonischer Verfeinerungsfolge zugrunde, dann ist  $T_{M,M}$  ein explizites Runge/Kutta-Verfahren der Ordnung  $p = M$  mit der Stufenzahl*

$$s = \frac{p(p-1)}{2} + 1.$$

Der Satz 4.2.1 besagt also, dass der Rechenaufwand eines extrapolierten Verfahrens zwar mit dem Grad der Verfeinerung steigt, gleichzeitig die Ordnung steuerbar ist. Allerdings sind die Stufenzahlen nicht optimal. Zum Beispiel ergibt der Satz die folgende Übersicht der Stufenzahlen ⚠

$p$	1	2	3	4
$s$ , klassisch	1	2	3	4
$s$ , extrapoliert	1	2	4	7

### 4.3 Implementierung

Um die Extrapolation zu implementieren wird versucht, in der asymptotischen  $h$ -Entwicklung des globalen Fehlers (4.4) möglichst viele Terme zu eliminieren. Zu diesem Zweck wird ein Interpolationspolynom

$$P(h) = e_0 - e_p h^p - \dots - e_{p+M-2} h^{p+M-2}$$

gesucht, welches für  $M$  die Interpolationsbedingungen

$$P(h_i) = T_{i,m+1}, \quad i = l, l-1, \dots, l-M+1,$$

erfüllt. Dies ergibt sich aus der Betrachtung einer Interpolationsaufgabe: Zu gegebenen Stützstellen  $x_0, \dots, x_N$  und Stützwerten  $f_0, \dots, f_N$  wird ein Polynom  $P$  gesucht, dass die Bedingungen  $P(x_i) = f_i$  für  $i = 0, \dots, N$  erfüllt.

Setzt man nun  $P_{[i,l]}(x)$  für das Interpolationspolynom durch  $x_i, \dots, x_{i+l-1}$ , dann erhält man die rekursive Darstellung

$$P_{[i,l]}(x) = (ax + b) P_{[i,l-1]}(x) + (1 - ax - b) P_{[i+1,l-1]}(x)$$

mit

$$ax + b = \begin{cases} 1 & \text{falls } x = x_i \\ 0 & \text{falls } x = x_{i+l-1} \end{cases},$$

also

$$ax + b = \frac{x - x_{i+l-1}}{x_i - x_{i+l-1}}.$$

Die Idee der Extrapolation ist es, von  $h = \frac{H}{n_1}, \frac{H}{n_2}, \dots$  auf  $h \rightarrow 0$  zu schließen. Für  $x = 0$  erhält man dann die rekursive Vorschrift

$$\begin{aligned} P_{[i,1]}(0) &= f_i \\ P_{[i,l]}(0) &= \frac{-x_{i+l-1}}{x_i - x_{i+l-1}} P_{[i,l-1]}(0) + \frac{x_i}{x_i - x_{i+l-1}} P_{[i+1,l-1]}(0). \end{aligned}$$

### 4.4 Extrapolation mit symmetrischen Verfahren

Die bisherigen Erkenntnisse bilden die Basis für hespiegelte und symetrische Einschrittverfahren. Dieser Abschnitt entwickelt zunächst aus einem Einschrittverfahren  $\Phi$  ein gespiegeltes



Einschrittverfahren  $\widehat{\Phi}$ . Anschließend werden symmetrische Einschrittverfahren untersucht, für die eine asymptotische  $h$ -Entwicklung des globalen Fehlers der Form

$$\eta_h^*(t) = e_2(t) h^2 + e_4^2(t) h^4 + \dots$$

gilt. Üblicherweise wird man hier  $p = 2$  setzen. Anschließend wird sich zeigen, dass es keine expliziten, symmetrischen Runge/Kutta-Verfahren gibt. Dies führt dann zu Mehrschrittverfahren.

#### 4.4.1 Gespiegelte Verfahren

Zur Spiegelung eines Verfahrens geht man wie folgt vor:

- (1) Starte in  $u_h(t+h)$ .
- (2) Ersetze in der Verfahrensvorschrift die Schrittweite  $h$  durch  $-h$ ,

$$u_{-h}(t-h) = u_{-h}(t) - h\varphi(t, u_{-h}(t); -h).$$

- (3) Ersetze  $t$  durch  $t+h$ ,

$$u_{-h}(t) = u_{-h}(t+h) - h\varphi(t+h, u_{-h}(t+h); -h).$$

- (4) Löse nach  $u_{-h}(t+h)$  auf,

$$u_{-h}(t+h) = u_{-h}(t+h) + h\varphi(t+h, u_{-h}(t+h); -h). \quad (4.13)$$

Das Vorgehen zur Spiegelung eines Einschrittverfahrens ist in 4.3 dargestellt.

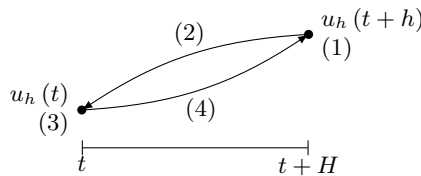


Abb. 4.3: Spiegelung eines Einschrittverfahrens

Schreibt man für (4.13) verkürzend

$$u_{-h}(t+h) = u_{-h}(t) + h\widehat{\varphi}(t, u_{-h}(t); h), \quad (4.14)$$

dann lässt sich folgende Definition angeben.

**Definition 4.4.1** (gespiegeltes Einschrittverfahren (Strehmel et al., 2012)). Das Einschrittverfahren

$$u_{-h}(t+h) = u_{-h}(t) + h\widehat{\varphi}(t, u_{-h}(t); h),$$

heißt zum Einschrittverfahren  $\Phi$ ,

$$u_h(t+h) = u_h(t) + h\varphi(t, u_h(t); h)$$

gespiegeltes Einschrittverfahren  $\widehat{\Phi}$ , und  $\widehat{\varphi}$  heißt gespiegelte Vefahrensfunktion.

**Beispiel 4.4.2.** (1) Die Spiegelung des expliziten Euler-Verfahrens liefert das implizite Euler-Verfahren

$$u_{-h}(t+h) = u_{-h}(t) + hf(t+h, u_{-h}(t+h)).$$

△

Durch die Spiegelung des impliziten Euler-Verfahrens erhält man dann wieder das explizite.

(2) Die Spiegelung der expliziten Trapez-Regel liefert

$$u_{-h}(t+h) = u_{-h}(t) + \frac{h}{2} (f(t+h, u_{-h}(t+h)) + f(t, u_{-h}(t)))$$

(3) Die Spiegelung eines  $s$ -stufigen expliziten Runge/Kutta-Verfahrens liefert

$$u_{-h}(t+h) = u_{-h}(t) + h \sum_{i \leq s} b_i k_i$$

$$k_i = f \left( t + (1 - c_i)h, u_{-h}(t) + h \sum_{j \leq s} (b_j - a_{ij}) k_j \right), i = 1, \dots, s$$

Die Spiegelung der expliziten Verfahren lässt sich verallgemeinern.

**Theorem 4.4.3.** Die Spiegelung eines  $s$ -stufigen Runge/Kutta-Verfahrens ist wider ein  $s$ -stufiges Runge/Kutta-Verfahren mit den Koeffizienten

$$\begin{array}{c|cccc} (1-c_1) & (b_1-a_{11}) & (b_2-a_{12}) & \dots & (b_s-a_{1s}) \\ \vdots & \vdots & \vdots & & \vdots \\ (1-c_s) & (b_1-a_{s1}) & (b_2-a_{s2}) & \dots & (b_s-a_{ss}) \\ \hline & b_1 & b_2 & \dots & b_s \end{array} \quad (4.15)$$

*Beweis.* Die Behauptung folgt aus der alternativen Darstellung eines Runge/Kutta-Verfahrens und den Schritten zur Bestimmung eines gespiegelten Verfahrens.  $\square$

**Korollar 4.4.4.** Die Spiegelung eines expliziten  $s$ -stufigen Runge/Kutta-Verfahrens ist nicht explizit.

*Beweis.* Da für ein explizites Verfahren die Einträge der Hauptdiagonalen  $a_{ii}$  stets alle Null sind, sind die zumindest die Hauptdiagonaleinträge des gespiegelten Verfahrens ungleich Null.

□

**Theorem 4.4.5.** Die Spiegelung  $\widehat{\Phi}$  eines gespiegelten  $\widehat{\Phi}$  Verfahrens ergibt das Ausgangsverfahren  $\Phi$ .

Für die Konsistenzordnung eines gespiegelten Verfahrens gilt

**Theorem 4.4.6.** *Das gespiegelte Einschrittverfahren  $\widehat{\Phi}$  hat die gleiche Konsistenzordnung des Ausgangsverfahrens.*

*Beweis.* Die Diskretisierung des lokalen Fehlers liefert

$$\eta(t+h) = y(t+h) - y(t) - h\varphi(t, y(t); h) = d_{p+1}(t) h^{p+1} + \mathcal{O}(h^{p+2}).$$

Wendet man die Schritte zur Spiegelung eines Runge/Kutta-Verfahren an,

$$y(t) - y(t+h) + h\varphi(t+h), y(t+h; -h) = d_{p+1}(t+h) (-h)^{p+1} + \mathcal{O}(h^{p+2}), \quad (4.16)$$

dann ist der lokale Fehler des gespiegelten Verfahrens gegeben durch

$$\begin{aligned} \widehat{\eta}(t+h) &= y(t+h) - y(t) - h\widehat{\varphi}(t, y(t); h) \\ &= y(t+h) - y(t) - h\varphi(t+h; y(t+h); -h), \end{aligned}$$

wobei bei der Umformung  $d_{p+1}(t+h) = d_{p+1}(t) + \mathcal{O}(h)$  ausgenutzt wurde. Hieraus folgt nun, dass das gespiegelte Verfahren die Ordnung  $p$  hat, was die Behauptung war.  $\square$

Von Beachtung für Lipschitz-stetige  $f$  und Differentialgleichungen  $v' = f(v)$  und  $u' = f(u)$   $\triangle$  mit  $v(t) - u(t) = \mathcal{O}(h^p)$ , dass

$$\begin{aligned} u(t+h) - v(t+h) &= \left( u(t) + hf(u(t)) + \frac{h^2}{2} f_u f(u(t)) \right) - \left( v(t) + hf(v(t)) + \frac{h^2}{2} f_v f(v(t)) \right) \\ &= u(t) - v(t) + \mathcal{O}(h \|u(t) - v(t)\|). \end{aligned}$$

**Theorem 4.4.7.** *Bei genügender Glattheit von  $f(t, y)$  besitzt das gespiegelte Verfahren  $\widehat{\Phi}$  eine asymptotische  $h$ -Entwicklung des Diskretisierungsfehler  $\eta_{-h}(t)$  der Form*

$$\eta_{-h}(t^*) = e_p(t^*) (-h)^p + e_{p+1}(t^*) (-h)^{p+1} + \dots + e_N(t^*) (-h)^N + E_{N+1}(t^*, -h) (-h)^{N+1},$$

wobei das  $E_{N+1}(t^*, -h)$  für hinreichend kleine  $h$  beschränkt ist.

*Beweis.* Der Beweis ist analog zum Satz 4.1.2.  $\square$

#### 4.4.2 Symmetrische Verfahren

Zur Konstruktion möglichst effizienter Extrapolationsverfahren sind symmetrische Einschrittverfahren hilfreich.

**Definition 4.4.8 (Einschrittverfahren; symmetrisch).** *Ein Einschrittverfahren heißt symmetrisch, wenn für die Verfahrensfunktionen  $\varphi$  und  $\widehat{\varphi}$  die Beziehung  $\varphi = \widehat{\varphi}$  gilt.*

Aus der Definition symmetrischer Einschrittverfahren folgt

**Theorem 4.4.9.** *Ein symmetrisches Einschrittverfahren besitzt in  $t^* = t_0 + nh$  stets eine asymptotische  $h^2$ -Entwicklung der Form*

$$\eta_h(t^*) = e_{2\gamma}(t^*) h^{2\gamma} + e_{2\gamma+2}(t^*) h^{2\gamma+2} + \dots, \quad e_{2j}(t_0) = 0. \quad (4.17)$$

*Beweis.* Wegen  $\varphi = \widehat{\varphi}$  folgt aus (4.14) die Gleichheit  $u_{-h}(t+h) = u_h(t+h)$ . Mit den Sätzen 4.1.2 und 4.4.7 folgt, dass die ungeraden Glieder der asymptotischen Entwicklung verschieden, d.h.  $e_{2j+1}(t^*) = 0$  für  $j = \gamma, \gamma+1, \dots$ . Damit wurde die Behauptung gezeigt.  $\square$

### 4.4.3 Gragg/Bulirsch/Stoer-Verfahren

Der letzte Satz 4.4.9 besagt, dass symmetrische Einschrittverfahren besonders effizient sind. Dafür sind jedoch die üblichen Runge/Kutta-Verfahren ungeeignet. Nutzt man Mehrschrittverfahren auf Basis der expliziten Mittelpunktsregel, dann erhält man den Algorithmus

$$u_1 = u_0 + hf(t_0, u_0) \quad (4.18a)$$

$$u_{m+1} = u_{m-1} + 2hf(t_m, u_m), \quad m = 1, 2, \dots, 2N \quad (4.18b)$$

Das Problem ist, dass für diesen Ansatz Einschrittverfahren benötigt werden. Schreibt man die Mittelpunktsregel als Einschrittverfahren für das verdoppelte System

$$v' = f(t, v), \quad v(t_0) = v_0$$

$$w' = f(t, w), \quad w(t_0) = w_0$$

Für gegebene  $u_{2m}, v_{2m}$  gilt dann

$$w_{2m+1} = w_{2m} + h(t_{2m}, v_{2m}) \quad (4.19a)$$

$$v_{2m+1} = v_{2m} + h(t_{2m+1}, w_{2m+1}) \quad (4.19b)$$

$$v_{2m+2} = v_{2m+1} + h(t_{2m+1}, w_{2m+1}) \quad (4.19c)$$

$$w_{2m+2} = w_{2m+1} + h(t_{2m+2}, v_{2m+2}) \quad (4.19d)$$

Die Gleichungen (4.19a) und (4.19c) sind offensichtlich explizit Euler-Verfahren, während (4.19b) und (4.19d) implizite Euler-Verfahren sind. Ferner erhält man aus (4.19b) und (4.19c)

$$v_{2m+2} = v_{2m} + 2hf(t_{2m+1}, w_{2m+1}) \quad (4.20)$$

Systeme der Form  $v' = f(t, w)$  und  $w' = g(t, v)$  heißen (4.20) *Strömer/Verlet-Verfahren*. Nun sei  $H = 2h$  mit

$$\bar{v}_m := v_{2m},$$

$$\bar{w}_m = w_{2m}.$$

Durch Anwendung der Spiegelung (4.19a) bis (4.19d) erhält man das *symplektische Strömer-Verlet-Verfahren*

$$\bar{w}_{m+1/2} = \bar{w}_m + h(t_m, \bar{v}_m) \quad (4.21a)$$

$$\bar{v}_{m+1} = \bar{v}_m + h\bar{w}_{m+1/2} \quad (4.21b)$$

$$\bar{w}_{m+1} = \bar{w}_{m+1/2} + h(t_{2m+2}, w_{2m+2}) \quad (4.21c)$$

Das Verfahren (4.21a), (4.21b) und (4.21c) ist von der Ordnung  $p = 2$  und hier explizit. Die Werte  $u_{2m}$  und  $v_m$  besitzen eine asymptotische  $h^2$ -Entwicklung. Wird dies als Mittelpunktsregel implementiert, dann erhält man

$$u_1 = y_0 + h(t_0, u_0)$$

$$u_{m+1} = u_{m+1} + 2hf(t_m, u_m), \quad u_{2m} = \bar{v}_{2m}, \quad u_{2m+1} = \bar{w}_{2m+1},$$

und zur Extrapolation wird  $u_{2m} = v_{2m}$  verwendet. Zur Durchführung des Verfahrens wählt man eine monoton fallende Schrittweitenfolge  $\{h_i\} = \{H/n_i\}$  mit einer geraden Unterteilungsfolge  $\{n_i\}$ , z.B. die

- doppelte Romberg-Folge

$$F_{2R} = \{2, 4, 8, 16, 32, 64, 128, 256, \dots\}$$

- doppelte Bulirsch-Folge

$$F_{2R} = \{2, 4, 6, 8, 12, 16, 24, 32, \dots\}$$

- doppelte harmonische Folge

$$F_{2H} = \{2, 4, 6, 8, 10, 12, 14, 16, \dots\}$$

und setzt

$$T_{i,1} = S_{h_i}(t_0 + H), \quad S_{h_i}(t_0 + H) = \frac{1}{4}(u_{2N-1} + 2u_{2N} + u_{2N+1})$$

Dieses Verfahren ist symmetrisch und es existiert eine asymptotische  $h^2$ -Entwicklung des globalen Fehlers.



## Qualitative Analyse von Differentialgleichungsmodellen

Die bisherige Analyse von Differentialgleichungen bezog sich meist auf einen festen Endzeitpunkt  $T < \infty$ . Dies lässt aber die Untersuchung des Langzeitverhaltens einer Differentialgleichung außer Acht. Der Einfachheit halber werden in diesem Kapitel autonome Systeme  $\dot{y} = f(y(t))$  betrachtet.

### 5.1 Stabilität von Fixpunkten

Betrachtet wird im Folgenden das autonome Anfangswertproblem

$$\dot{y} = f(y(t)), \quad y(t_0) = 0. \quad (5.1)$$

Für den einfachen Fall  $\dot{y} = \lambda y(t)$  mit Anfangswert  $y(0) = y_0$  hängt das Langzeitverhalten der Lösung

$$y(t) = y_0 e^{\lambda t}$$

vom Vorzeichen von  $\lambda$  ab. Grundsätzlich können drei Fälle unterschieden werden:

- $\lambda > 0$ : Im Fall  $\lambda > 0$  wächst die Lösung  $y(t)$  exponentiell für  $t \rightarrow \infty$ , d.h. für große Werte von  $t$  ist die Berechnung von  $\lambda$  wenig sinnvoll, da auch Störungen verstärkt werden.
- $\lambda = 0$ : Im Fall  $\lambda = 0$  ist die Lösung  $y(t)$  konstant für alle  $t \geq 0$ , d.h. für große Werte von  $t$  ist die Berechnung von  $\lambda$  unproblematisch, wobei Störungen erhalten bleiben.
- $\lambda < 0$ : Im Fall  $\lambda < 0$  fällt die Lösung  $y(t)$  exponentiell für  $t \rightarrow \infty$ , d.h. für große Werte von  $t$  ist die Berechnung von  $\lambda$  unproblematisch, da auch Störungen gedämpft werden.

Dieses bereits bekannte Lösungsverhalten kann auch auf nichtlineare Lösungen verallgemeinert werden.

**Definition 5.1.1 (Lösungsverhalten; Stabilität).** Sei  $\psi(t) := y(t; y_0) \in C^1[0, \infty)$  die Lösung eines Anfangswertproblems. Dann heißt  $\psi$  in Vorwärtsrichtung  $t \rightarrow \infty$

- (1) Ljapunov-stabil, wenn für ein  $\eta > 0$  die Differentialgleichung für alle Anfangswerte  $y_0$  mit  $\|y_0 - \psi(0)\| \leq \eta$  Lösungen auf  $[0, \infty)$  haben und zu jedem  $\varepsilon > 0$  ein  $\delta(\varepsilon) > 0$  existiert, so dass für alle  $t \geq 0$  aus

$$\|y_0 - \psi(0)\| \leq \delta(\varepsilon) \quad \text{stets} \quad \|y(t; y_0) - \psi(t)\| \leq \varepsilon \quad (5.2)$$

folgt.

- (2) asymptotisch stabil, wenn  $\psi$  Ljapunov-stabil ist und zusätzlich ein  $\bar{\alpha} > 0$  existiert, so dass aus

$$\|y_0 - \psi(0)\| \leq \bar{\alpha} \quad \text{stets} \quad \|y(t; \psi) - \psi(t)\| \xrightarrow{t \rightarrow \infty} 0 \quad (5.3)$$

folgt.

- (3) exponentiell stabil, wenn  $\psi$  Ljapunov-stabil ist und zusätzlich Konstanten  $C$  und  $\sigma$  existieren, so dass

$$\|y(t; y_0) - \psi(t)\| \leq Ce^{-\sigma t} \|y_0 - \psi(0)\| \quad (5.4)$$

gilt.

- (4) instabil, wenn sie nicht stabil ist.

△

Synonym mit Ljapunov-stabile Lösung wird eine solche Lösung auch *strikt stabil* genannt. Für die Umkehrung der Richtung,  $t \rightarrow \infty$  wird die Definition umgekehrt, d.h. eine Lösung, die in Vorwärtsrichtung asymptotisch stabil ist, ist in Rückwärtsrichtung instabil.

Aus der Definition 5.1.1 und der zugehörigen Bemerkung folgt, dass für die Charakterisierung des Lösungsverhaltens von  $\psi$  die Richtung der Zeitachse entscheidend ist, da bei Zeitumkehr sich die stabilen und instabilen Lösungsräume vertauschen.

**Definition 5.1.2 (Fixpunkt).** Ein Vektor  $\mathbf{y}^* \in \mathbb{R}^n$  heißt Fixpunkt oder Gleichgewichtslage, falls  $f(t, \mathbf{y}^*) = 0$  für alle  $t \geq 0$  gilt.

Von Interesse ist nun die Stabilität von Fixpunkten. Unter Verwendung der Definition 5.1.1 kann ein Fixpunkt charakterisiert werden, als

- (strikt) stabil, falls es eine Umgebung  $\Omega_0$  von  $\mathbf{y}^*$  gibt, so dass für jede Lösung  $\psi(t)$  mit  $y_0 \in \Omega_0$  die Bedingung

$$\lim_{t \rightarrow \infty} \psi(t) = \mathbf{y}^*$$

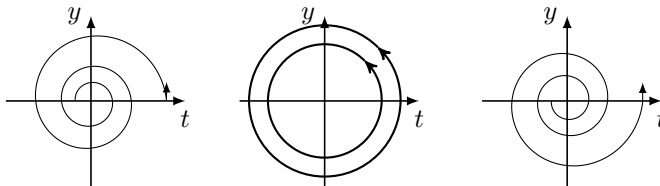
erfüllt ist.

- neutral stabil, falls für eine hinreichend kleine Umgebung  $\Omega_0$  von  $\mathbf{y}^*$  eine Umgebung  $V$  von  $\mathbf{y}^*$  existiert, so dass aus

$$\psi(0) \in V \quad \text{stets} \quad \psi(t) \in \Omega_0 \quad \text{für alle } t \geq 0$$

folgt.

- instabil, falls er weder strikt stabil noch neutral stabil ist.



(a) stabiler Fixpunkt (b) neutral stabiler Fixpunkt (c) instabiler Fixpunkt

Abb. 5.1: Arten von Fixpunkten



*Beispiel 5.1.3.*

Ohne Beschränkung kann man sich im Folgenden auf die Stabilität von *Fixpunkten autonomer Differentialgleichungen* konzentrieren. Dies erlaubt die Charakterisierung von Fixpunkten mittels Analyse von Phasenlinien. △

## 5.2 Phasenlinien

Bevor die Problematik der Stabilität weiter verfolgt wird, bietet es sich zunächst ein simples Beispiel an.

*Beispiel 5.2.1.* Betrachtet wird das Anfangswertproblem

$$\dot{y} = y(1 - y), \quad y(0) = y_0. \quad (5.5)$$

Die rechte Seite von (5.5) besitzt zwei Fixpunkte  $y_1^* = 0$  und  $y_2^* = 1$ .

In dem Phasenportrait des Beispiels erreichen die Trajektorien den Fixpunkte nie, dies ist aber durch aus möglich. △

*Beispiel 5.2.2.* Die Differentialgleichungen

$$y' = \sqrt{|y|}, \quad y' = \begin{cases} \sqrt{|y|} & y \geq 0 \\ -\sqrt{|y|} & y < 0 \end{cases}$$

besitzt die exakten Lösungen

### 5.2.1 Stabilität von Fixpunkten

Um Fixpunkte für allgemeine, nichtlineare Differentialgleichungen zu bestimmen muss man in der Regel auf die Taylor-Entwicklung zurückgreifen. Für die rechte Seite von (5.1) gilt dann

$$y' = f(y) = f(y^*) + f'(y^*)(y - y^*) + \mathcal{O}(\|y - y^*\|^2). \quad (5.6)$$

Da  $f(y^*) = 0$  ist, wird durch  $f'(y^*)$  die Stabilität bestimmt. Zur Vorbereitung auf den nächsten Satz bezeichnet

- $\sigma := \{\lambda \in \mathbb{C} | \det(\mathbf{A} - \lambda \mathbf{I}) = 0\}$  das Spektrum einer Matrix  $\mathbf{A} \in \mathbb{C}^{n \times n}$ .
- $\nu(\mathbf{A}) = \max_{\lambda \in \sigma(\mathbf{A})} \Re(\lambda)$  die Spektralabzisse.

**Theorem 5.2.3.** *Sei  $\mathbf{y}^* \in \Omega_0$  ein Fixpunkt der autonomen Differentialgleichung  $y' = f(\mathbf{y}(t))$  mit hinreichend oft stetig differenzierbaren rechten Seite. Dann sind folgende Aussagen äquivalent:*

- Der Fixpunkt  $\mathbf{y}^*$  ist asymptotisch stabil.
- Die Spektralabzisse ist negativ  $\nu(Df(\mathbf{y}^*)) < 0$ .

*Beweis.* Deuffhard and Bornemann (2013)

□

### 5.3 Stabilität im $\mathbb{R}^n$

Sei  $\mathbf{z} := (\mathbf{y} - \mathbf{y}^*)$  und  $\mathbf{A} := f'(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*)$ . Dann kann die Entwicklung (5.6) der rechte Seite als ein lineares System mit konstanten Koeffizienten dargestellt werden

$$\mathbf{z}' = \mathbf{A}\mathbf{z}(t) + \mathcal{O}(\|\mathbf{z}\|^2). \quad (5.7)$$

Dann sind im Sinne von Satz 5.2.3 die Eigenwerte  $\lambda$  von  $\mathbf{A}$  entscheidend. Ist  $\mathbf{z}^*$  ein Fixpunkt, dann gibt es folgende Lösungsdarstellungen

- **einfache Eigenwerte:** Ist  $\mathbf{v}$  der Eigenvektor zum Eigenwert  $\lambda$ , dann ist  $\mathbf{z}(t) = e^{\lambda t}\mathbf{v}$  eine Lösung.
- **komplexe Eigenwerte:** Ist  $\mathbf{v}$  der Eigenvektor zum komplexen Eigenwert  $\lambda = \alpha + i\beta$ , dann ist  $\mathbf{z}(t) = \langle \Re e^{\lambda t}\mathbf{v}, \Im e^{\lambda t}\mathbf{v} \rangle$  eine Lösung. Mit  $\lambda = \alpha + i\beta$  gilt ferner  $e^{\lambda t} = e^{\alpha t}(\cos \beta t + i \sin \beta t)$ .
- **mehrfache Eigenwerte:** Sind  $\mathbf{v}_0, \mathbf{v}_1, \dots$  Eigenvektoren zum Eigenwert, dann ist  $\mathbf{z}(t) = e^{\lambda t}\mathbf{v}_0 + te^{\lambda t}\mathbf{v}_1, \dots$  eine Lösung.

Die Stabilität wird nun gesichert, wenn

**Theorem 5.3.1.** *Sei  $\mathbf{y}^* \in \Omega$  ein Fixpunkt eines linearen Systems mit konstanten Koeffizienten. Dann gilt:*

- (1) *Der Fixpunkt  $\mathbf{y}^*$  ist genau dann asymptotisch stabil, wenn für jeden Eigenwert  $\lambda_i$  von  $\mathbf{A}$  gilt*

$$\Re \lambda_i < 0.$$

- (2) *Der Fixpunkt  $\mathbf{y}^*$  ist genau dann neutral stabil, wenn für jeden Eigenwert  $\lambda_i$  von  $\mathbf{A}$  gilt*

$$\Re \lambda_i \leq 0$$

*und zu jeder  $k$ -fachen Nullstelle  $\lambda_i$  des charakteristischen Polynom mit  $\Re \lambda_i = 0$  gehören genau  $k$  linear unabhängige Eigenvektoren.*

- (3) *Der Fixpunkt  $\mathbf{y}^*$  ist genau dann instabil, wenn für einen Eigenwert  $\lambda_i$  von  $\mathbf{A}$  gilt*

$$\Re \lambda_i > 0.$$

*Beispiel 5.3.2.* Sei  $\lambda = 0$  ein zweifacher Erwartungswert für die Matrizen

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Die Eigenvektoren  $\mathbf{v}_{1,2}$  zur Matrix  $\mathbf{A}$  und  $\mathbf{w}$  zur Matrix  $\mathbf{B}$  sind

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

und damit der Fixpunkt  $\mathbf{z}^* = 0$  von  $\mathbf{A}$  neutral stabil, während der von  $\mathbf{B}$  instabil ist.

## 5.4 1-dimensionale Populationsmodelle

Ein *einfaches Populationsmodell* zur Beschreibung der Populationsgröße  $p(t)$  ist gegeben durch

$$p' = (\lambda - \mu) p(t) = \alpha p(t), \quad \lambda, \mu > 0$$

mit Geburtsrate  $\lambda$  und Sterberate  $\mu$  bzw. Nettozuwachsrate  $\alpha$ . Die rechte Seite hat ihren Fixpunkt bei  $p^* = 0$ . Offensichtlich ist der Fixpunkt für  $\alpha < 0$  stabil, und für  $\alpha > 0$  instabil, denn für  $\alpha > 0$  würde dies unbegrenztes Populationswachstum implizieren. Dies ist allerdings ein eher unrealistisches Modell. Daher erscheint die Annahme einer Sättigungsgrenze  $K$  sinnvoll. Dies führt zu einem Populationsmodell der Form

$$p' = \alpha p(t) \left(1 - \frac{p(t)}{K}\right) \quad (5.8)$$

Dieses *logarithmische Populationsmodell mit beschränkten Ressourcen* besitzt offensichtlich die beiden Fixpunkte  $p_1^* = 0$  und  $p_2^* = K$ . Die Gleichung (5.8) lässt sich analytisch mittels Trennung der Variablen lösen. Aus

$$\alpha = \frac{p'}{p(1 - p/K)}$$

folgt mittels Partialbruchzerlegung

$$\frac{1}{p(1 - p/K)} = \frac{-K}{p^2 - Kp} = \frac{1}{p} - \frac{1}{p - K}$$

und Integration die Gleichheit

$$\ln \left( \frac{p(t)}{p(t) - K} \right) = \alpha t + c_0,$$

ergibt. Dies liefert

$$p(t) = \frac{c_1 K e^{\alpha t}}{c_1 e^{\alpha t} - 1} = \frac{K}{1 - c_2 (e^{-\alpha t})}$$

und mit der Anfangsbedingung  $p(t_0) = p_0$  folgt dann

$$p(t) = \frac{K p_0}{p_0 + (K - p_0) e^{-\alpha(t-t_0)}}. \quad (5.9)$$

Diese exakte Lösung die folgenden grundsätzlichen Eigenschaften für  $\alpha > 0$ :

- (1) Die Lösung  $p(t)$  ist stets nichtnegativ für  $p_0 \geq 0$ .
- (2) Für  $t \rightarrow \infty$  konvergiert die Lösung, wenn  $p_0 > 0$  gegen den Gleichgewichtspunkt  $p^* = K$ .

Durch die Einführung eines Rivalitätsterm kann das einfache Modell erweitert werden:

- *konstante Entnahme*  $c$ :
- *dynamische Entnahme*  $\gamma y$ :

Wird aus der Population regelmäßig oder konitnuierlich eine konstante Menge  $\bar{c}$  entnommen, dann erweitert sich die Differentialgleichung (5.8) zu

$$p' = \alpha p(t) \left(1 - \frac{p(t)}{K}\right) - c, \quad (5.10)$$

was zu von  $p$  abhängigen Fixpunkten  $p^* = c^*(p)$  führt:

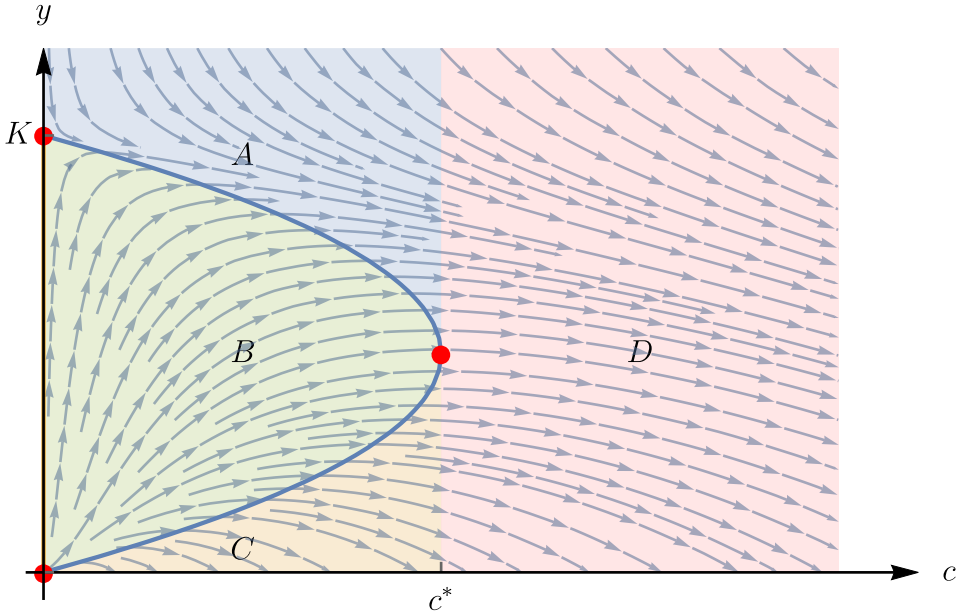


Abb. 5.2: Für  $c < c^*$  ist die Bejagung möglich, falls  $(c, y) \in A$  und die Population bricht zusammen, falls  $(c, y) \in C$ . Für  $c > c^*$  ist  $(c, y) \in D$  und die Population bricht zusammen.

Wird aus der Population regelmäßig oder konitnuierlich eine von der Population abhängigen Menge  $\gamma$  entnommen, dann erweitert sich die Differentialgleichung (5.8) zu

$$\begin{aligned} p' &= \alpha p(t) \left(1 - \frac{p(t)}{K}\right) - \gamma p \\ &= \alpha p(t) \left(1 - \frac{\gamma}{\alpha} - \frac{p(t)}{K}\right) \\ &= (\alpha - \gamma) p(t) \left(1 - \frac{p(t)}{K(1 - \gamma/\alpha)}\right) \end{aligned} \quad (5.11)$$

mit den Fixpunkten  $p_1^* = 0$  und  $p_2^* = K \left(1 - \frac{\gamma}{\alpha}\right)$ . Damit die Population stabil ist, muss die Entnahme  $(c, p)$  in den Bereichen A und B liegen. Dies ist dann erfüllt, wenn  $\gamma < \alpha$  ist, ansonsten ist die Population instabil, und folglich lässt sich auch eine maximale Entnahme bestimmen. Bezeichne

$$R = \gamma p_2^* = \gamma K \left(1 - \frac{\gamma}{\alpha}\right)$$

die Entnahmerate, bei der die entnommene Menge einerseits maximal, andererseits die Population stabil ist. In diesem Fall muss  $\frac{dR}{d\gamma} = 0$  gelten, dies liefert

$$\gamma = \frac{\alpha}{2}.$$

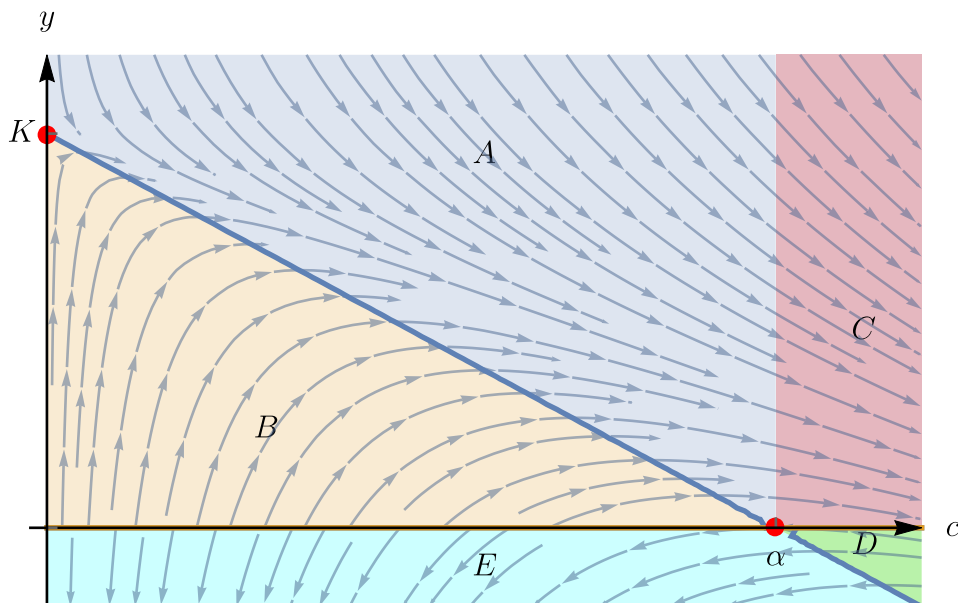


Abb. 5.3: Für  $c < \alpha$  ist die Bejagung möglich, falls  $(c, y) \in A, B$  und Population ist stabil. Für  $c > \alpha$  ist die Population instabil.

## 5.5 Interagierende Populationsmodelle

In den einführenden Bemerkungen in Kapitel ?? wurde bereits ein Modell mit zwei Populationen eingeführt. Wird ein Modell mit zwei Populationen betrachtet, dann können diese in besonderer Beziehung zueinander stehen:

- *Räuber/Beute-Verhältnis*
- *Konkurrenz*
- *Symbiose*

*Beispiel 5.5.1* (Räuber/Beute-Modell). Das *Räuber/Beute-Modell* wurde bereits eingeführt. In diesem Modell werden zwei Populationen, Beutetiere  $x(t)$  und Raubtiere  $y(t)$  unterschieden. Für  $t > t_0$  ist dies ein Anfangswertproblem mit zwei gewöhnlichen Differentialgleichungen, die die *Lotka/Volterra-Gleichungen*

$$x' = \alpha x - \beta xy = (\alpha - \beta y) x, \quad x(t_0) = x_0 \quad (5.12a)$$

$$y' = \delta xy - \gamma y = (\delta x - \gamma) y, \quad y(t_0) = y_0 \quad (5.12b)$$

mit  $\alpha, \beta, \gamma, \delta > 0$  erfüllen. Zur qualitativen Untersuchung des Modells müssen die Phasenebene, die Fixpunkte und die Zeroklinien

$$\mathcal{Z}_x : 1 - \beta y = 0, \quad \mathcal{Z}_y : 1 - \delta x = 0 \quad (5.13)$$

Die Darstellung der zugehörigen Lösung kann sowohl in der  $xy$ -Ebene geschehen oder komponentenweise über die Zeitachse. Zur beispielhaften Untersuchung der Stabilität sei das Modell speziell gegeben durch  $f(x, y)$  mit

$$\begin{aligned} x' &= (1 - y)x \\ y' &= (-2 + x)y \end{aligned}$$

Die Fixpunkte sind dann

$$\mathbf{z}_1^* = (0, 0), \quad \mathbf{z}_2^* = (2, 1)$$

mit den zugehörigen Spektralabzissen  $\nu(Df(\mathbf{z}_i^*))$  der Jacobi-Matrien

$$\nu(Df(\mathbf{z}_1^*)) = 1, \quad \nu(Df(\mathbf{z}_2^*)) = 0$$

d.h. der Fixpunkte  $\mathbf{z}_1^*$  ist instabil, der Fixpunkte  $\mathbf{z}_2^*$  ist neutral stabil.

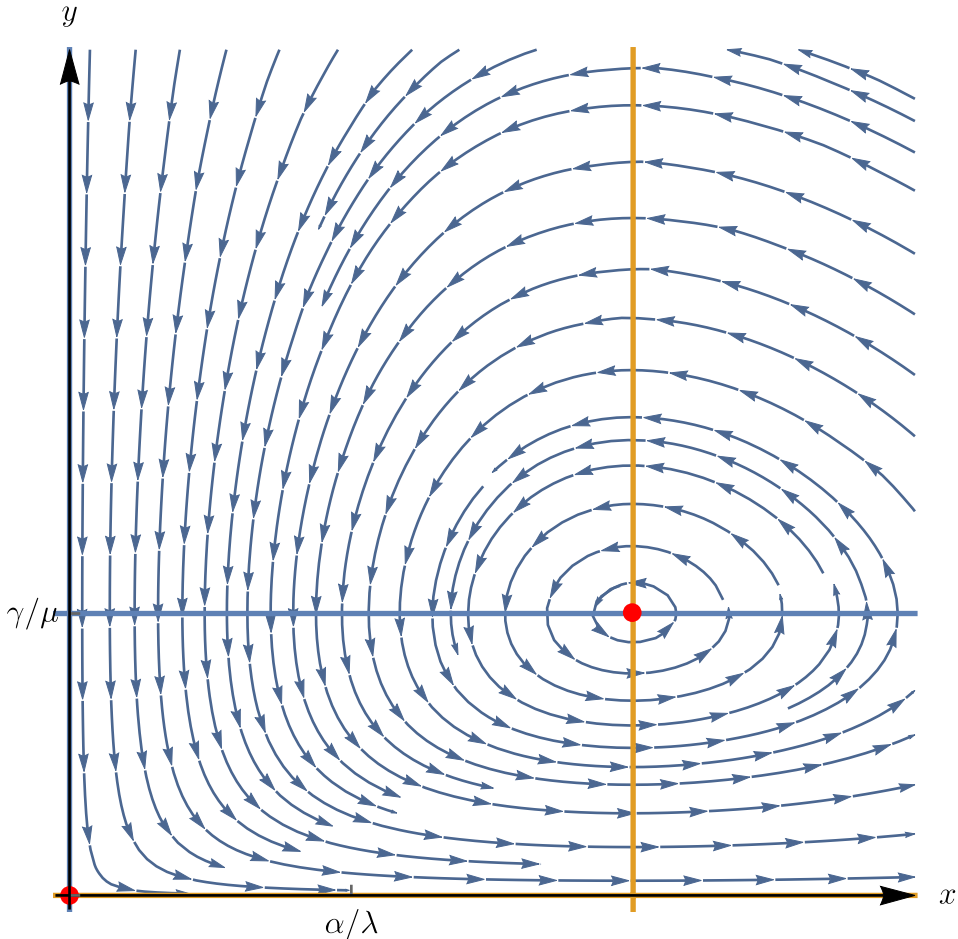


Abb. 5.4: Richtungsfeld im Räuber/Beute-Modell

*Beispiel 5.5.2 (Konkurrenzmodell).* Im *Konkurrenzmodell* kommen zwei Populationen vor die um beschränkte Ressourcen konkurrieren. Dies ist einerseits eine Erweiterung des logarithmischen Populationsmodell dar, andererseits durch Einführung von *Reibungstermen* eine Modifikation des Räuber/Beute-Modells,

$$x' = (\alpha - \beta y) x - \lambda x^2 = (\alpha - \beta y - \lambda x) x, \quad x(t_0) = x_0 \quad (5.14a)$$

$$y' = (\gamma - \delta x) y - \mu y^2 = (\gamma - \delta x - \mu y) y, \quad y(t_0) = y_0 \quad (5.14b)$$

mit  $\alpha, \beta, \gamma, \delta, \lambda, \mu > 0$ . Zur qualitativen Untersuchung des Modells müssen die Phasenebene, die Fixpunkte und die Zeroklinien

$$\mathcal{Z}_x : \alpha - \beta y - \lambda x = 0, \quad \mathcal{Z}_y : \gamma - \delta x - \mu y = 0 \quad (5.15)$$

untersucht werden. Werden beide Geraden in die gewöhnlichen Geradengleichungen in Abhängigkeit von  $x$  geschrieben, dann besitzt  $\mathcal{Z}_x$  eine negative Steigung, während  $\mathcal{Z}_y$  positive Steigung besitzt. Haben beide Geraden im positiven Quadranten keinen gemeinsamen Punkt, dann gibt drei Fixpunkte,  $\mathbf{z}_1^* = (0, 0)$ ,  $\mathbf{z}_2^* = (\alpha/\lambda, 0)$ , und  $\mathbf{z}_3^* = (0, \gamma/\mu)$ . Haben beide Geraden hingegen einen gemeinsamen Punkt im positiven Quadranten, dann ist

$$\mathbf{z}_4^* = \left( \frac{\beta\gamma - \alpha\mu}{\beta\delta - \lambda\mu}, \frac{\alpha\delta - \gamma\lambda}{\beta\delta - \lambda\mu} \right)$$

ein Fixpunkt.

Zur beispielhaften Untersuchung der Stabilität sei das Modell speziell gegeben durch  $f(x, y)$  mit

$$\begin{aligned} x' &= 2x \left( 1 - \frac{x}{2} \right) - xy = (2 - x - y) x \\ y' &= 3y \left( 1 - \frac{y}{3} \right) - 2xy = (3 - y - 2x) y \end{aligned}$$

Die Fixpunkte sind dann

$$\mathbf{z}_1^* = (0, 0), \quad \mathbf{z}_2^* = (2, 0), \quad \mathbf{z}_3^* = (0, 3), \quad \mathbf{z}_4^* = (1, 1),$$

mit den zugehörigen Spektralabzissen  $\nu(Df(\mathbf{z}_i^*))$  der Jacobi-Matrizen

$$\nu(Df(\mathbf{z}_1^*)) = 3, \quad \nu(Df(\mathbf{z}_2^*)) = -1, \quad \nu(Df(\mathbf{z}_3^*)) = -1, \quad \nu(Df(\mathbf{z}_4^*)) = \sqrt{2} - 1,$$

d.h. die die Fixpunkte  $\mathbf{z}_1^*, \mathbf{z}_4^*$  sind instabil, und die Fixpunkte  $\mathbf{z}_2^*, \mathbf{z}_3^*$  sind asymptotisch stabil.

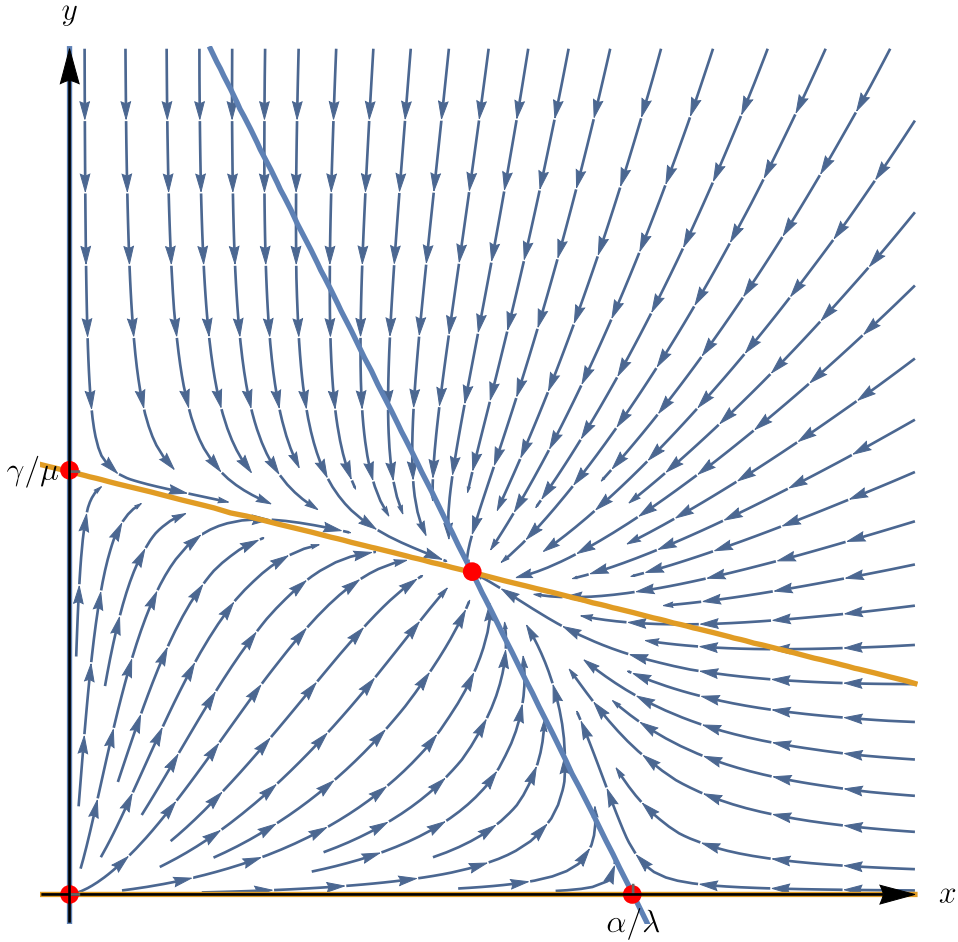


Abb. 5.5: Richtungsfeld im Konkurrenz-Modell

*Beispiel 5.5.3 (Symbiosemodell).* Im *Symbiosemodell* kommen gibt es zwei Populationen vor, deren Kooperation um beschränkte Ressourcen Vorteile für die gemeinsame Population birgt. Dies ist wieder ein logarithmisches Populationsmodell,

$$x' = (\alpha + \beta y)x - \lambda x^2 = (\alpha + \beta y - \lambda x) \quad , \quad x(t_0) = x_0 \quad (5.16a)$$

$$y' = (\gamma + \delta x)y - \mu y^2 = (\gamma + \delta x - \mu y)y \quad , \quad y(t_0) = y_0 \quad (5.16b)$$

mit  $\alpha, \beta, \gamma, \delta, \kappa > 0$ . Zur qualitativen Untersuchung des Modells müssen die Phasenebene, die Fixpunkte und die Zeroklinen

$$Z_x : \alpha + \beta y - \lambda x = 0, \quad Z_y : \gamma + \delta x - \mu y = 0 \quad (5.17)$$

untersucht werden. Werden beide Geraden in die gewöhnlichen Geradengleichungen in Abhängigkeit von  $x$  geschrieben, dann besitzen  $Z_x$  und  $Z_y$  positive Steigung. Haben beide Geraden im positiven Quadranten keinen gemeinsamen Punkt, dann gibt drei Fixpunkte,  $\mathbf{z}_1^* = (0, 0)$ ,  $\mathbf{z}_2^* = (\alpha/\lambda, 0)$ , und  $\mathbf{z}_3^* = (0, \gamma/\mu)$ . Haben beide Geraden hingegen einen gemeinsamen Punkt in positiven Quadranten, dann ist



$$\mathbf{z}_4^* = \left( \frac{\alpha\mu + \beta\gamma}{\lambda\mu - \beta\delta}, \frac{\alpha\delta + \gamma\lambda}{\lambda\mu - \beta\delta} \right)$$

ein Fixpunkt.

Zur beispielhaften Untersuchung der Stabilität sei das Modell speziell gegeben durch  $f(x, y)$  mit

$$x' = \left(1 + \frac{y}{2} - x\right)x, \quad y' = \left(1 + \frac{x}{2} - y\right)y.$$

Die Fixpunkte sind dann

$$\mathbf{z}_1^* = (0, 0), \quad \mathbf{z}_2^* = (1, 0), \quad \mathbf{z}_3^* = (0, 1), \quad \mathbf{z}_4^* = (2, 2),$$

mit den zugehörigen Spektralabzissen  $\nu(Df(\mathbf{z}_i^*))$  der Jacobi-Matrizen

$$\nu(Df(\mathbf{z}_1^*)) = 1, \quad \nu(Df(\mathbf{z}_2^*)) = \frac{3}{2}, \quad \nu(Df(\mathbf{z}_3^*)) = \frac{3}{2}, \quad \nu(Df(\mathbf{z}_4^*)) = -1,$$

d.h. die die Fixpunkte  $\mathbf{z}_1^*, \mathbf{z}_2^*, \mathbf{z}_3^*$  sind instabil, und die Fixpunkte  $\mathbf{z}_2^*, \mathbf{z}_3^*$  sind asymptotisch stabil.

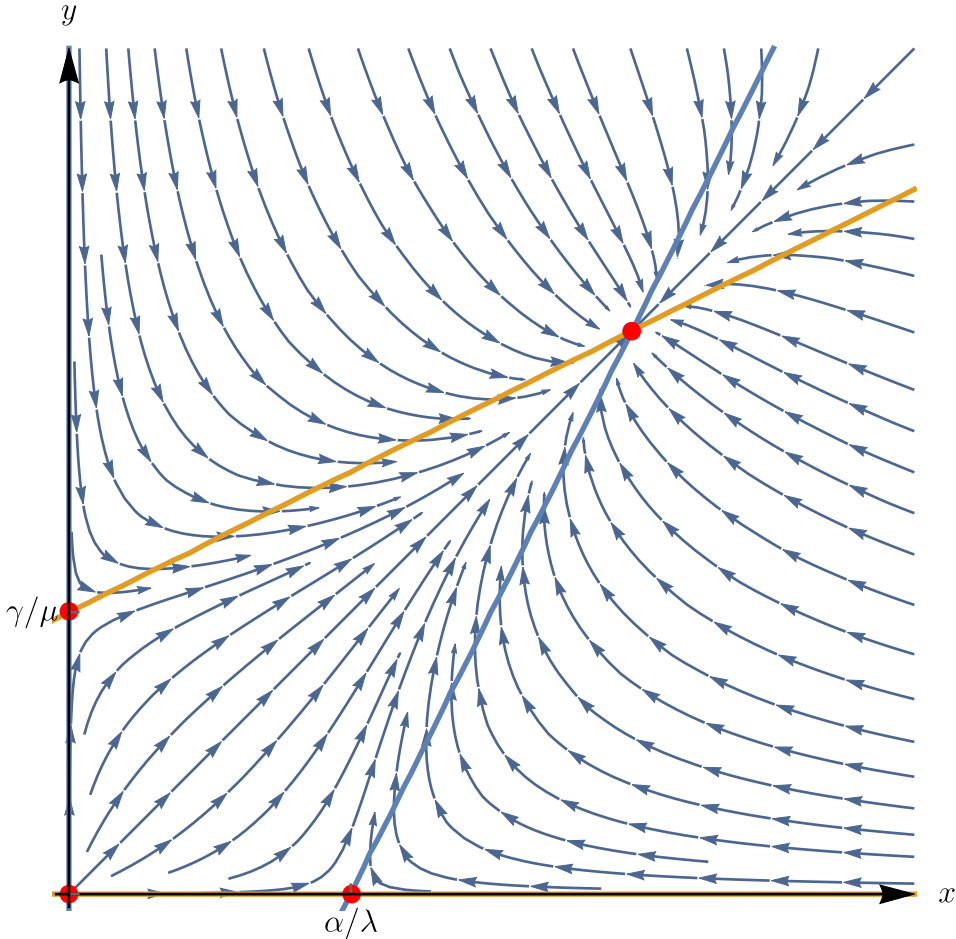


Abb. 5.6: Richtungsfeld im Symbiose-Modell

## 5.6 Bifurkation

Betrachtet wird eine parameterabhängige Differentialgleichung

$$y' = f(y; \mu), \quad (5.18)$$

wobei der Parameter  $\mu$  einen Einfluss auf das Lösungsverhalten hat, d.h. die für  $\mu < \mu_0$  hat die Lösung von (5.18) ein anderes Lösungsverhalten als für  $\mu > \mu^*$ . *Bifurkation* in diesem Zusammenhang bedeutet, dass die Lösung am *Bifurkationspunkt*  $\mu^*$  ihr Verhalten ändert, und damit auch die Fixpunkte und deren Typ.

*Beispiel 5.6.1.* Die skalare Differentialgleichung

$$y' = \mu y, \quad y(t_0) = y_0$$

die Lösung  $y(t) = y_0 e^{\mu t}$  besitzt den Fixpunkt  $y^* = 0$ . nebst Bifurkationspunkt  $\mu^* = 0$ . Für  $\mu < 0$  ist der Fixpunkt asymptotisch stabil, während er für  $\mu > 0$  instabil ist, d.h. der die Stabilität des Fixpunktes ändert sich am Bifurkationspunkt  $\mu^* = 0$ .

Im Folgenden werden vier verschiedene Arten von Bifurkationspunkten untersucht werden.

*Beispiel 5.6.2.* Die skalare Differentialgleichung

$$y' = \mu - y^2, \quad \mu \in \mathbb{R}_{\geq 0}$$

besitzt die Fixpunkte  $y_1^* = \sqrt{\mu}$  und  $y_2^* = -\sqrt{\mu}$  nebst Bifurkationspunkt  $\mu^* = 0$ . Für  $\sqrt{\mu^*} = y_1^*$  ist der Fixpunkt  $y_1^*$  stabil, und für  $y_2^* = -\sqrt{\mu^*}$  instabil. Ferner existiert für  $\mu^* < 0$  kein Fixpunkt. Bifurkationspunkte dieser Art heißen *Sattelnoden*.

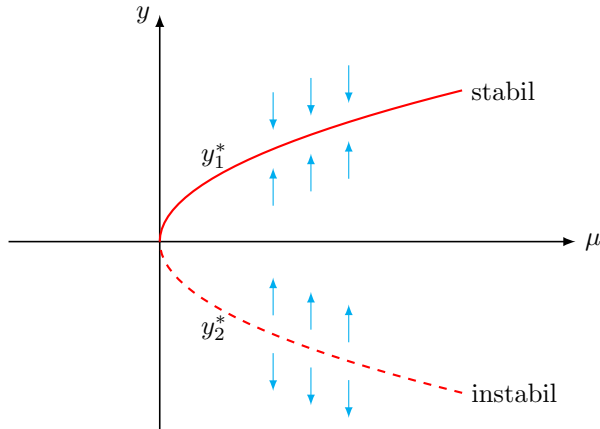


Abb. 5.7: Die Fixpunkte  $y_i^*$ ,  $i = 1, 2$  aufgetragen über dem Bifurkationsparameter  $\mu$ .

*Beispiel 5.6.3.* Die skalare Differentialgleichung

$$y' = \mu y - y^3 = y(\mu - y^2), \quad \mu \in \mathbb{R}_{\geq 0}$$

besitzt die Fixpunkte  $y_1^* = 0$ ,  $y_2^* = \sqrt{\mu}$  und  $y_3^* = -\sqrt{\mu}$  nebst Bifurkationspunkt  $\mu^* = 0$ . Für  $\mu^* > 0$  ist  $y_1^*$  instabil, und für  $\mu < 0$  stabil. Ferner sind für  $\mu > 0$  die Fixpunkte  $y_2^*$

und  $y_3^*$  stabil, während sie für  $\mu < 0$  nicht existieren. Bifurkationspunkte dieser Art heißen *Heugabelpunkte*.

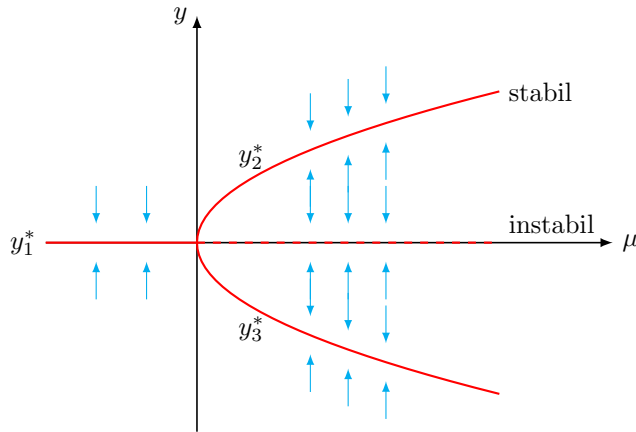


Abb. 5.8: Am Bifurkationspunkt ändert der Fixpunkt  $y_1^*$  seine Stabilität. Es entstehen zwei neue stabile Fixpunkt  $y_2^*$  und  $y_3^*$ .

*Beispiel 5.6.4.* Die skalare Differentialgleichung

$$y' = -y^3 + y - \mu$$

besitzt die Fixpunkte gegeben durch  $\mu = -y^3 + y$ , wobei die Bifurkationspunkte gegeben sind durch  $\mu_1^* = \frac{1}{3}$  und  $\mu_2^* = \frac{-1}{3}$ . In Abbildung 5.9 springen die Fixpunkte für  $\mu > \mu_1^*$  auf den unteren Fixpunktast, für  $\mu < \mu_1^*$  ändern die Fixpunkte ihr Verhalten nicht.

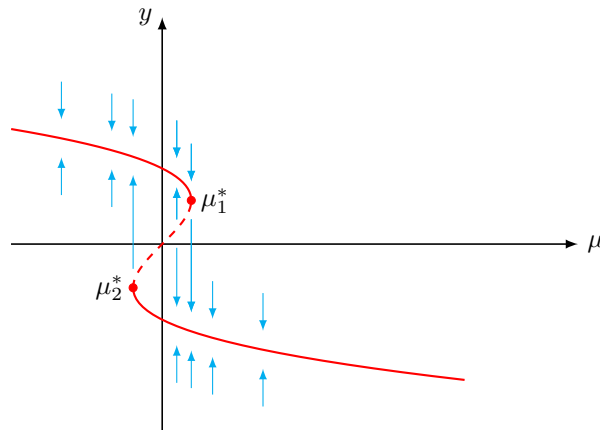


Abb. 5.9: Sprung der Fixpunkt an den Bifurkationspunkten.

Ein häufig untersuchtes biologisches Phänomen ist die Populationsentwicklung des Fichtenspinners *Choristoneura occidentalis*

*Beispiel 5.6.5 (Spruce Budworm).* Die Population des kanadischen Fichtenspinners vermehrt sich alles ca. 15 Jahre explosionsartig. Das Populationsmodell basiert auf folgenden Annahmen:

- (1) Es gibt zwei Populationen, eine Fichtenspinner- und Vogelpopulation.
- (2) Die Fichtenspinner haben begrenzte Kapazitäten zur Verfügung.
- (3) Die Vögel ernähren sich nicht regelmäßig von den Fichtenspinnern, sondern in Abhängigkeit der Populationsentwicklung. Sind viele Fichtenspinner vorhanden, dann werden diese bis zu einer Obergrenze gefressen, sonst wird auf alternative Nahrungsquellen zurückgegriffen.

Dies kann als eindimensionales Populationsmodell beschrieben werden

$$y' = \alpha y \left(1 - \frac{y}{K}\right) - \frac{By^2}{y^2 + A^2}. \quad (5.19)$$

Mit der Fixpunkteigenschaft  $y' = 0$  gilt

$$\frac{\alpha}{B} \left(1 - \frac{y}{K}\right) = \frac{y}{y^2 + A^2}$$

Setzt man  $\alpha' = \alpha/B$  und  $g(y) := y/(y^2 + A^2)$ , dann gilt die Beziehung

$$\alpha' \left(1 - \frac{y}{K}\right) = g(y),$$

die linke Seite beschreibt eine Gerade  $h(y)$ , die rechte Seite eine stetige Funktion. Abhängig von den Schnittpunkten von  $h$  und  $g$  haben die Fixpunkte verschiedene Stabilitätseigenschaften, d.h. der stabile Fixpunkt springt.

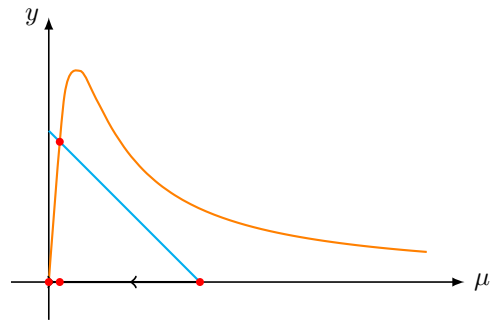
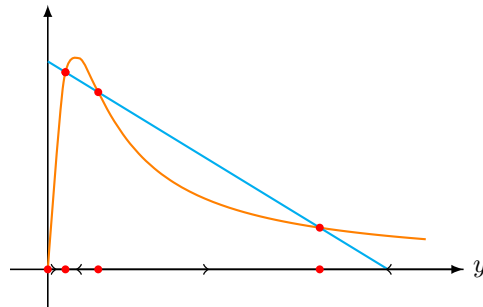
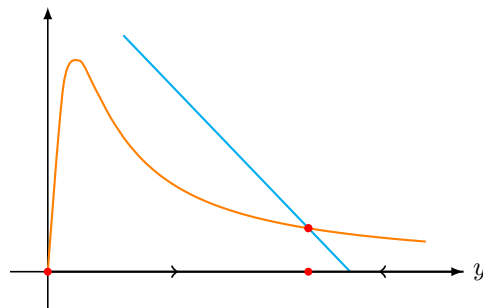
(a) Ein Schnittpunkt von  $h$  und  $g$ (b) Drei Schnittpunkte von  $h$  und  $g$ (c) Ein Schnittpunkt von  $h$  und  $g$ 

Abb. 5.10: Sprung der Fixpunkt an den Bifurkationspunkten.



## Steife Differentialgleichungen





## Theoretische Grundlagen steifer Probleme

### 6.1 Stabilität von Differentialgleichungen

Bei der numerischen Behandlung eines mathematischen Problems  $f$

In der Numerik können zwei Größen zur Bewertung eines Verfahrens herangezogen werden:

- **Kondition**
- **Stabilität**

Ein numerisches Verfahren  $\tilde{y} = \tilde{f}(x)$  zu einem mathematischen Problem  $y = f(x)$  ist *stabil*, wenn der Einfluss von Störungen auf die Lösung des Problems gering ist. Demgegenüber ist ein mathematisches Problem  $y = f(x)$  *gut konditioniert*, wenn die numerische Lösung eines mathematischen Problems unabhängig von der Störung der Eingangsdaten  $x, \tilde{x}$  ist.

*Beispiel 6.1.1 (Hoher Rundungsfehlereinfluss).* Bestimmt werden soll die Ableitung  $f'(x)$  mittels Differenzenquotient. Dann gilt

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

Berücksichtigt man bei der computeralgebraischen Umsetzung die *Floating Point Arithmetik* mit  $|\varepsilon_1|, |\varepsilon_2| < \text{eps} := 10^{-16}$ , dann gilt

$$\begin{aligned} \frac{f(x+h)(1+\varepsilon_1) - f(x)(1+\varepsilon_2)}{h} &\approx \frac{f(x+h) - f(x)}{h} + \frac{2\text{eps}}{h} \mathcal{O}(f) \\ &\approx \frac{f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3) - f(x)}{h} + \frac{2\text{eps}}{h} \mathcal{O}(f) \\ &\approx f'(x) + \frac{h}{2}f'' + \frac{2\text{eps}}{h}f + \mathcal{O}(h^2) \\ &\approx f'(x) + h + \frac{\text{eps}}{h}, \end{aligned}$$

wobei die  $h + \text{eps}/h$  für  $h = \sqrt{\text{eps}}$  minimal ist. Betrachtet man anstelle einer Funktion  $f$  die numerische Lösung einer Differentialgleichung, dann ist ein Problem gut konditioniert, wenn die Fehlerfortpflanzung in etwa der analytischen oder exakten Lösung entspricht. △

Für die Charakterisierung der Stabilität eines Fixpunktes  $y^*$  mit

$$\dot{y} = f(y), f(y^*) = 0$$

können unterschiedliche Stabilitätsbegriffe angeführt werden:

**Definition 6.1.2.** Sei  $y^* \in \mathcal{C}^1[0, \infty)$  ein Gleichgewicht einer Differentialgleichung  $\dot{y} = f(t, y)$ . Dann heißt  $y^*$

- (1) Ljapunov-stabil, wenn für ein  $\eta > 0$  die Differentialgleichung für alle Anfangswerte  $y_0$  mit  $\|y_0 - y^*(0)\| \leq \eta$  Lösungen auf  $[0, \infty)$  haben und zu jedem  $\varepsilon > 0$  ein  $\delta(\varepsilon) > 0$  existiert, so dass für alle  $t \geq 0$  aus

$$\|y_0 - y^*(0)\| \leq \delta(\varepsilon) \quad \text{stets} \quad \|y(t; y_0) - y^*(t)\| \leq \varepsilon \quad (6.1)$$

folgt.

- (2) asymptotisch stabil, wenn  $y^*$  Ljapunov-stabil ist und zusätzlich ein  $\bar{\alpha} > 0$  existiert, so dass aus

$$\|y_0 - y^*(0)\| \leq \bar{\alpha} \quad \text{stets} \quad \|y(t; y_0) - y^*(t)\| \xrightarrow{t \rightarrow \infty} 0 \quad (6.2)$$

folgt.

- (3) exponentiell stabil, wenn  $y^*$  Ljapunov-stabil ist und zusätzlich Konstanten  $C$  und  $\sigma$  existieren, so dass

$$\|y(t; y_0) - y^*(t)\| \leq C e^{-\sigma t} \|y_0 - y^*(0)\| \quad (6.3)$$

gilt.

△

Aus den Definitionen der unterschiedlichen Stabilitätsbegriffe folgt:

**Korollar 6.1.3.** Sei  $y^* \in \mathcal{C}^1[0, \infty)$  ein Gleichgewicht einer Differentialgleichung  $\dot{y} = f(t, y)$ .

- (1) Ist  $y^*$  exponentiell stabil, dann ist es auch asymptotisch stabil.  
 (2) Ist  $y^*$  asymptotisch stabil, dann ist es auch Ljapunov-stabil.

Für den einfachen, eindimensionalen Fall mit  $\dot{y} = \lambda y$  erhält man die spezielle Lösung  $y(t) = e^{\lambda t} y_0$ . Abhängig von  $\lambda$  gilt

- $\lambda < 0$ : Im Fall  $\lambda < 0$  ist die Lösung exponentiell stabil.
- $\lambda = 0$ : Im Fall  $\lambda = 0$  ist die Lösung *neutral stabil*, d.h. sie ist zwar Ljapunov-stabil, aber nicht asymptotisch stabil.
- $\lambda > 0$ : Im Fall  $\lambda > 0$  ist die Lösung unbeschränkt.

*Beispiel 6.1.4.* Die allgemeine Lösung des linearen Systems

$$\dot{\mathbf{y}} = \mathbf{A} \mathbf{y}, \mathbf{y}(0) = \mathbf{y}_0$$

mit  $\mathbf{A} \in \mathbb{R}^{n \times n}$  ist

$$\mathbf{y}(t) = \exp(t\mathbf{A}) \mathbf{y}_0. \quad (6.4)$$

Zur Analyse der Stabilität der Nulllösung des lineares Systems kann das Eigensystem von  $\mathbf{A}$  mit Eigenwerte  $\lambda_1, \dots, \lambda_n$  und Eigenvektoren  $\mathbf{v}_1, \dots, \mathbf{v}_n$  betrachtet werden. Sei hierzu  $\mathbf{T}$  eine Transformationsmatrix mit den Eigenvektoren von  $\mathbf{A}$  als Spalten und  $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Dann gilt  $\mathbf{AT} = \mathbf{T}\mathbf{\Lambda}$ . Setzt man nun  $\mathbf{y} = \mathbf{T}\mathbf{x}$ , dann gilt  $\dot{\mathbf{y}} = \mathbf{T}\dot{\mathbf{x}}$  und  $\mathbf{A}\mathbf{y} = \mathbf{AT}\mathbf{x}$ , so dass  $\mathbf{T}\dot{\mathbf{x}} = \mathbf{T}\mathbf{\Lambda}\mathbf{x}$ , also  $\dot{\mathbf{x}} = \mathbf{\Lambda}\mathbf{x}$ , dies ist das System

$$\begin{aligned}\dot{x}_1 &= \lambda_1 x_1, \\ \dot{x}_2 &= \lambda_2 x_2, \\ &\vdots \\ \dot{x}_n &= \lambda_n x_n.\end{aligned}$$

Allgemein gilt für das Diagonalisieren mittels einer analytische Funktion  $g$ , dass

△

$$g(\mathbf{A}) = \mathbf{T} \begin{bmatrix} g(\lambda_1) & & \\ & \ddots & \\ & & g(\lambda_n) \end{bmatrix} \mathbf{T}^{-1}, \quad \mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1}$$

Dann lässt sich (6.4) als Potenzreihe schreiben,

$$\begin{aligned}\exp(t\mathbf{A}) &= \exp(t\mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1}) = \mathbf{I} + t\mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1} + \frac{t^2}{2}\mathbf{T}\mathbf{\Lambda}^2\mathbf{T}^{-1} + \dots \\ &= \mathbf{I} + t\mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1} + \frac{t^2}{2}\mathbf{T}\mathbf{\Lambda}^2\mathbf{T}^{-1} + \frac{t^3}{6}\mathbf{T}\mathbf{\Lambda}^3\mathbf{T}^{-1} + \dots \\ &= \mathbf{T} \exp(t\mathbf{\Lambda}) \mathbf{T}^{-1} = \mathbf{T} \begin{pmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{pmatrix} \mathbf{T}^{-1}\end{aligned}$$

Zusammenfassend gilt folgende Stabilitätsaussage für lineare Systeme

**Theorem 6.1.5.** *Sei  $\mathbf{y}^*$  die Nulllösung eines linearen Anfangswertproblems*

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \quad \mathbf{y}(0) = \mathbf{y}_0.$$

Dann gilt:

- (1) *Die Nulllösung  $\mathbf{y}^*$  ist genau dann asymptotisch stabil, wenn für die Eigenwerte von  $\mathbf{A}$*

$$\Re \lambda_i < 0, \quad \forall i$$

*gilt.*

- (2) *Die Nulllösung  $\mathbf{y}^*$  ist genau dann Ljapunov-stabil, wenn für die Eigenwerte von  $\mathbf{A}$*

$$\Re \lambda_i \leq 0, \quad \forall i$$

*gilt, und zu jeder  $k$ -fachen Nullstelle  $\lambda_i$  des charakteristischen Polynoms mit  $\Re \lambda_i = 0$  gehören  $k$  linear unabhängige Eigenvektoren.*

Die Aussage 6.1.5(2) besagt, dass  $\mathbf{y}^*$  genau dann Ljapunov-stabil ist, wenn  $\Re \lambda_i \leq 0$  für alle  $i$  einer diagonalisierbaren Matrix  $\mathbf{A}$  gilt, oder dass im Fall, dass  $\mathbf{A}$  nicht diagonalisierbar ist, im Fall  $\Re \lambda_i = 0$  kein Jordan-Block vorliegt, d.h. ist  $\mathbf{A}$  nicht diagonalisierbar, dann müssen die Jordan-Blöcke betrachtet werden.

△

## 6.2 Einseitige Lipschitz-Bedingung

Die Aussage aus dem Satz 6.1.5 lässt sich nicht ohne weiteres auf nichtlineare Systeme übertragen. Für diese ist der folgende Begriff notwendig.

**Definition 6.2.1 (Dissipativität).** Eine Differentialgleichung  $\dot{y} = f(t, y)$  heißt dissipativ oder kontraktiv, wenn

$$\|y(t_2) - v(t_2)\| \leq \|y(t_1) - v(t_1)\|$$

für  $t_2 > t_1$  bei beliebigen Lösungen  $y(t)$  und  $v(t)$ .

Wünschenswert wäre es leicht erkennen zu können, ob ein System dissipativ ist. Sei  $\langle \cdot, \cdot \rangle$  ein Skalarprodukt im  $\mathbb{R}^n$  mit zugehörigem 2-Normquadrat  $\|y\|_2^2 = \langle y, y \rangle$ . Differenziert man die 2-Norm, dann gilt

$$\begin{aligned} \frac{d}{dt} \|y(t) - v(t)\|_2^2 &= \frac{d}{dt} \langle y(t) - v(t), y(t) - v(t) \rangle \\ &= 2 \langle \dot{y}(t) - \dot{v}(t), y(t) - v(t) \rangle \\ &= 2 \langle f(t, y) - f(t, v), y(t) - v(t) \rangle, \end{aligned}$$

und liefert folgende Definition:

**Definition 6.2.2 (einseitige Lipschitz-Bedingung).** Eine Funktion  $f$  genügt einer einseitigen Lipschitz-Bedingung, wenn

$$\langle f(t, y) - f(t, v), y(t) - v(t) \rangle \leq \ell(t) \|y - v\|_2^2, \quad t \geq 0, \quad y, v \in \mathbb{R}^n. \quad (6.5)$$

Dabei heißt  $\ell \in \mathbb{R}$  einseitige Lipschitz-Konstante.

**Theorem 6.2.3.** Genügt die rechte Seite einer Differentialgleichung einer einseitigen Lipschitz-Bedingung, dann gilt

$$\|y(t_2) - v(t_2)\|_2 \leq \exp \left[ \int_{t_1}^{t_2} \ell(\tau) d\tau \right] \|y(t_1) - v(t_1)\|_2$$

Ist  $\ell(t) \leq 0$ , so ist das System dissipativ.

*Beweis.* Seien  $y, v$  zwei Lösungen der rechten Seite einer Differentialgleichungen.

$$\psi(t) := \|y(t) - v(t)\|^2 = \langle y(t) - v(t), y(t) - v(t) \rangle.$$

Wie oben behandelt ist  $\psi(t)$  stetig differenzierbar mit

$$\dot{\psi} = 2 \langle f(t, y) - f(t, v), y(t) - v(t) \rangle.$$

Da die rechte Seite der Differentialgleichung  $f$  die einseitige Lipschitz-Bedingung erfüllt, gilt

$$\dot{\psi} \leq 2\ell(t) \|y(t) - v(t)\|^2 = 2\ell(t) \psi(t), \quad t \geq 0.$$

Für zwei beliebige  $t_1, t_2$  mit  $t_2 \geq t_1$  gilt mittels Integration

$$\psi(t_2) \leq \psi(t_1) + 2 \int_{t_1}^{t_2} \ell(\tau) \psi(\tau) d\tau, \quad (6.6)$$

und die Behauptung folgt dann aus dem Gronwall-Lemma ?? angewandt auf (6.6),

$$\psi(t_2) \leq \exp \left[ 2 \int_{t_1}^{t_2} \ell(\tau) dt \right] \psi(t_1),$$

und nach Ziehen der Wurzel auf beiden Seiten. □

△ Im einfachen Fall  $\dot{y} = f(t, y) = \lambda y$  bedeutet

- Lipschitz-Bedingung, dass  $|\lambda(y - v)| \leq L|y - v|$ , und die Konstante  $L \geq |\lambda|$  ist nach *zwei Seiten* beschränkt.
- einseitige Lipschitz-Bedingung, dass  $\langle \lambda(y - v), y - v \rangle \leq \ell|y - v|^2$ , und  $\lambda \leq \ell$  ist nach einer Seite beschränkt.

Erfüllt eine Funktion  $f$  die einseitige Lipschitz-Bedingung, dann sind die Eigenwerte nach oben beschränkt und es können beliebig kleine Eigenwerte zu gelassen werden. Gilt ferner  $\ell(t) \leq 0$ , genau dann ist  $f$  dissipativ.

## 6.3 Steife Differentialgleichungen

Curtiss and Hirschfelder (1952) und Dahlquist (1963) beobachteten, dass explizite Runge/Kutta-Verfahren, aufgrund ihres beschränkten Stabilitätsgebietes, erhebliche Schwierigkeiten bei der numerischen Behandlung bestimmter Differentialgleichungen haben. Der Grund liegt im Wesentlichen darin, dass die Lipschitz-Bedingung nur für große Werte  $L \gg 1$  erfüllt ist, was dazu führt, dass die Schrittweiten für eine hohe Genauigkeit sehr klein gewählt werden müssen, bei großen Schrittweiten das Problem jedoch nicht gut genug approximiert werden kann.

Steife Differentialgleichungen können durch zwei Phasen charakterisiert werden:

- (1) *transiente Phase*: In der transienten Phase besitzt das System eine eine Dynamik, bei der nach einer Auslenkung die Gleichgewichtslösung erreicht wird, d.h. man befindet sich auf der noch nicht glatten Lösung.
- (2) *steife Phase*: In der steifen Phase ist das Gleichgewicht nazu erreicht und Abweichungen werden stark gedämpft und die Gleichgewichtslösung wird nach einer kleinen Auslenkung wieder schnell erreicht.

Die Bezeichnung „steif“ für Differentialgleichung wurde von Curtiss and Hirschfelder (1952) für solche Reaktanden eingeführt, dereb einem Gleichgewichtszustand sich schnell einstellte, während langsame Reakanden eingefroren wurden, d. h. also *steif* wurden. Allerdings gibt es bis dato keine exakte Definition für Steifheit, so dass lediglich eine ungenügende Charakterisierung angegeben werden kann:

**Definition 6.3.1 (Steifheit).** *Ein System heißt steif, wenn die rechte Seite einer Differentialgleichung die Lipschitz-Bedingung  $L$  und die einseitige Lipschitz-Bedingung  $\ell$  erfüllt mit*

$$L \gg 1, \ell \lesssim 0 \quad \text{bzw.} \quad hL \gg 1, h\ell \lesssim 0,$$

wobei  $h$  eine typische, den Toleranzanforderungen angepasste, Schrittweite ist.

Im einfachen Fall von

$$\dot{y} = \lambda y$$

erhält man für die Lipschitz-Bedingung bzw. die einseitige Lipschitz-Bedingung

$$L = |\lambda|, \ell = \lambda$$

denn

$$|\lambda(y - v)| = L|y - v|, \quad \lambda(y - v)^2 = \langle f(y) - f(v), y - v \rangle \leq \ell \|y - v\|^2.$$

*Beispiel 6.3.2 (Testproblem; Prothero/Robinson-Gleichung).* Zur Veranschaulichung des Begriffs der Steifheit wird folgendes skalare, nichtautonome Testproblem betrachtet:

$$\dot{y} = \lambda (y(t) - \varphi(t)) + \dot{\varphi}(t) \quad (6.7)$$

Die allgemeine, exakte Lösung lautet

$$y(t; y_0) = \varphi(t) + y_0 e^{\lambda t} \quad (6.8)$$

Für den speziellen Fall

$$\dot{y} = 0.5 (y(t) - \cos t) - \sin t, \quad y_0 = 1.5$$

erhält man die exakte Lösung

$$y(t) = 1.5e^{\lambda t} + \cos t.$$

Für  $\lambda < 0$  ist die Lösung  $y^*(t) = \cos t$  asymptotisch stabil, für  $\lambda > 0$  instabil, vgl.

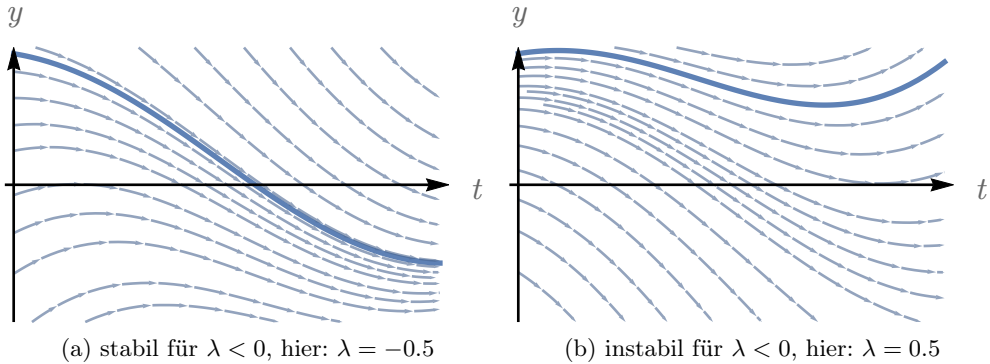
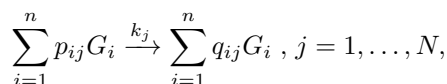


Abb. 6.1: Stabiles und instabiles Lösungsverhalten der speziellen Testproblems

## 6.4 Auftreten steifer Systeme

Ein typisches Problem steifer Systeme tritt, wie von [Curtiss and Hirschfelder \(1952\)](#) notiert, bei manchen Reaktionsgleichungen auf. Die analytische Behandlung von Reaktionsgleichungen wird durch die *Reaktionskinetik* behandelt.

Betrachtet wird ein abgeschlossenes, homogenes und konstantes System mit  $N$  chemischen Reaktionen zwischen  $n$  chemischen Substanzen  $G_i$ ,  $i = 1, \dots, n$ . Die  $j$ -te chemische Reaktion wird beschrieben durch den Umsatz von  $p_{ij}$  Einheiten  $G_i$  in  $q_{ij}$  Einheiten von  $G_i$ ,



abhängig von der Reaktionsgeschwindigkeit  $k_j$ . Da das System konstant ist, können anstelle der Substanzen  $G_i$  auch deren Konzentrationen  $y_i(t)$ , betrachtet werden. In diesem Fall verhält sich die Reaktionsgeschwindigkeit der  $j$ -ten Reaktion proportional zu

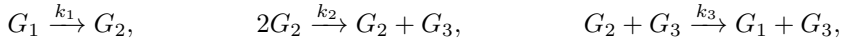
$$k_j \prod_{l=1}^n y_l^{p_{lj}}(t).$$

Fasst man den Verbrauch an  $G_i$  und die Produktion von  $G_i$  zusammen, dann erhält man ein System von Differentialgleichungen

$$\dot{y}_i = \sum_{j=1}^N (q_{ij} - p_{ij}) k_j \prod_{l=1}^n y_l^{p_{lj}}(t), \quad y_i(t_0) = y_{i,0} \quad (6.9)$$

für alle  $i = 1, \dots, n$ .

*Beispiel 6.4.1 (Reaktionskinetik).* Untersucht werden sollen die chemische Reaktionen (6.9)



mit

$$k_1 = 0.4, \quad k_2 = 3 \times 10^7, \quad k_3 = 3 \times 10^4.$$

Wendet man (6.9) auf die drei Reaktionen an, dann erhält man das System

$$\dot{y}_1 = -k_1 y_1 + k_3 y_2 y_3, \quad (6.10a)$$

$$\dot{y}_2 = k_1 y_1 - k_2 y_2^2 - k_3 y_2 y_3, \quad (6.10b)$$

$$\dot{y}_3 = k_2 y_2^2. \quad (6.10c)$$

. Offensichtlich laufen die Reaktionen sehr unterschiedlich ab. Die erste Reaktion (6.10a) läuft im Vergleich zu den anderen sehr langsam ab, während die zweite Reaktion (6.10b) sehr schnell abläuft. Da das System konservativ ist, gilt  $\dot{y}_1 + \dot{y}_2 + \dot{y}_3 = 0$  und damit  $y_1(t) + y_2(t) + y_3(t) = 1$ .

Setzt man voraus, dass zu Beginn der Reaktion lediglich die Substanz  $G_1$  vorhanden ist,  $\mathbf{y}_0 \equiv (1, 0, 0)$ , dann stellt sich ein Gleichgewicht bei  $\mathbf{y}^* = (0, 0, 1)$  ein. Die hohen Reaktionskonstanten führen zu einer großen Norm der Jacobi-Matrix und damit zu einer großen Lipschitz-Konstanten  $L \gg 1$ , womit das betrachtete Beispiel steif ist.

Ein Methode zur Lösung parabolischer Anfangs-Randwert-Aufgaben liegt in der Verwendung der *Linienmethode* oder *finite Differenzenmethode*, indem der betrachtete Raum durch ein Gitter diskretisiert wird. Betrachtet man das Intervall  $[0, 1]$ , dann kann dieses in  $N - 1$  Teilintervalle  $x_i, x_{j+1}$  mit  $0 = x_0 < x_1 < \dots < x_{N-1} < x_N = 1$  unterteilt werden. Dann kann Anstelle von  $u(t, x_i)$  die Einschränkung  $u_i(t)$  betrachtet werden, d.h.  $u_i(t)$  „lebt“ auf den *Linien*.

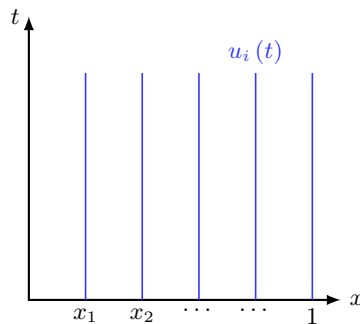


Abb. 6.2: Prinzip der Linienmethode.

Im Ergebnis erhält man dadurch ein Anfangswertproblem gewöhnlicher Differentialgleichungen, welches im Allgemeinen steif ist, wobei allerdings die Steifheit durch eine höhere Feinheit des Gitters erhöht wird. Ein Anwendung der Linienmethode ist die Wärmeleitungsgleichung:

*Beispiel 6.4.2 (Wärmeleitungsgleichung).* Betrachtet wird der Wärmefluss

$$\mathbf{q} = -\lambda \nabla T, \quad (6.11)$$

der in der Zeitspanne  $dt$  durch die Fläche  $dA$  fließt. Dabei bezeichnet  $T$  die Temperatur und  $\lambda$  die Wärmeleitfähigkeit.

In einem homogenen Medium beträgt die Energie eines Volumen  $V$  ist

$$U = c_p \rho T, \quad (6.12)$$

dabei bezeichnet  $c_p$  die spezifische Wärmekapazität und  $\rho$  die Massendichte. Ist der Wärmefluss nicht konstant, dann gilt für die Wärmeverteilung

$$Q = (\mathbf{q} \times \mathbf{n}) dA dt.$$

Für ein Testvolumen  $V$  ergibt sich für ein homogenes die Wärmeleitung

$$\frac{d}{dt} \int_V U(t, \mathbf{x}) dV = \int_{\partial V} \frac{Q}{dt} dA = - \int_{\partial V} (\mathbf{q} \times \mathbf{n}) dA = - \int_V \nabla \mathbf{q} dV = \int_V \frac{\partial U}{\partial t} dV, \quad (6.13)$$

wobei im  $\mathbb{R}^3$

$$\nabla = \frac{\partial}{\partial x_1} \mathbf{e}_1 + \frac{\partial}{\partial x_2} \mathbf{e}_2 + \frac{\partial}{\partial x_3} \mathbf{e}_3$$

gilt. Dies Gleichungskette (6.13) liefert

$$\frac{\partial U}{\partial t} = -\nabla \mathbf{q} = c_p \rho \frac{\partial T}{\partial t} = -\nabla (-\lambda \nabla T) = \lambda \nabla^2 T. \quad (6.14)$$

Sind  $c_p \rho$  und  $\lambda$  konstant, dann lautet die Wärmeleitungsgleichung

$$\frac{\partial T}{\partial t} = \frac{\lambda}{c_p \rho} \nabla^2 T = \delta \nabla^2 T, \quad (6.15)$$

wobei  $\delta$  den Temperaturleitkoeffizienten bezeichnet. Im inhomogenen Medien mit Wärmequellen lautet die Wärmeleitungsgleichung

$$\frac{\partial T}{\partial t} = \frac{\lambda}{c_p \rho} \nabla^2 T = \delta \nabla^2 T + q(t, \mathbf{x}).$$

Im 1-dimensionalen Fall lautet das Anfangswertproblem für ein homogenes Medium ohne Wärmequelle mit den Dirichlet-Randbedingungen

$$\begin{aligned} \dot{u} &= \delta \frac{\partial^2 u}{\partial x^2}, \quad u(t, 0) = u(t, 1) = 0 \\ u(0, x) &= g(x) \end{aligned}$$

Die Schrittweite im gleichabständigen Gitter lautet  $\Delta x = \frac{1}{N}$  mit Punkten  $x_i = i \Delta x$ . Ersetzt man die Ortsableitungen durch zentrale Differenzenquotienten mit der Approximation

$$u_i(t) \approx u(t, x_i)$$



der Linienmethode und vernachlässigt den auftretenden Fehler, so erhält man

$$\begin{aligned}\frac{\partial^2}{\partial x^2} u(t, x_i) &= \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} \\ u_i(0) &= g(x_i) \\ \dot{u}_i &= \frac{\delta}{\Delta x^2} (u_{i+1} - 2u_i + u_{i-1})\end{aligned}\tag{6.16}$$

für alle  $i = 1, \dots, N-1$ , bzw. in kompakter Form

$$\dot{\mathbf{U}} = \begin{pmatrix} \dot{u}_1 \\ \vdots \\ \dot{u}_{N-1} \end{pmatrix} = \mathbf{A} \mathbf{U}(t) + \mathbf{q}(t), \quad \mathbf{U}(0) = \mathbf{g}_0$$

mit

$$\mathbf{A} = \frac{\delta}{\Delta x^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 1 & -2 \end{pmatrix}, \quad \mathbf{g}_0 = (g(x_1), \dots, g(x_{N-1}))^\top,$$

wobei  $q(t)$  eine Wärmequelle ist. Es stellt sich nun die Frage, was die Eigenwerte von  $\mathbf{T}$  sind. Es gilt

$$\mathbf{T}\mathbf{x} = \lambda\mathbf{x}$$

mit  $x_0 = x_N = 0$ . Also gilt die Differenzengleichung

$$x_{k+1} - 2x_k + x_{k-1} = \lambda x_k, \quad k = 1, \dots, N-1.$$

Um  $x_k$  zu berechnen benötigt man die beiden Vorgänger von  $x_k$ , wobei  $x_1$  frei gewählt werden kann. Die Lösung der Differenzengleichung berechnet sich nun ähnlich der einer Differentialgleichung. Sei

$$y^{(n)} + a_{n-1}y^{(n-1)} + \dots + a_1y' + a_0y = 0,$$

dann erhält man mittels Operatorschreibweise  $D := \frac{d}{dt}$ , die Operatorgleichung

$$p(D)y = 0, \quad p(x) = x^n + \sum_{\nu=0}^{n-1} a_\nu x^\nu.\tag{6.17}$$

Setzt man  $y = e^{\lambda t}$  als Eigenfunktion von  $D$ , dann gilt

$$p(\lambda)y = 0$$

$$p(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0.$$

Der Hintergrund hierfür ist folgender: Im endlich dimensionalen Fall gilt für eine lineare Abbildung  $\mathbf{A} : \mathbf{x} \mapsto \mathbf{A}\mathbf{x}$  das Eigenwertproblem  $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ . Im unendlich dimensionalen Fall erhält man mittels des Operators  $D : y \mapsto Dy$  die Lösung  $y = e^{\lambda t}$  von  $Dy = \lambda y$ . Ferner gilt

$$\begin{aligned}DDy &= D(\lambda y) = \lambda y = \lambda^2 y, \\ P(D)y &= p(\lambda)y.\end{aligned}$$

Sei nun  $y = \{y_i\}_{i \in \mathbb{N}_0}$  eine unendliche Zahlenfolge und bezeichne  $S$  einen *Shift-* bzw. *Verschiebungsoperator*,

$$S((y_0, y_1, y_2, \dots)) = (y_1, y_2, y_3, \dots)$$

dann erhält man als Lösung von  $Sy = \lambda y$  mit Folgenglieder  $(Sy)_k = \lambda y_k$

$$y_{k+1} = \lambda y_k.$$

Mit  $y = 1$  erhält man dann  $y_1 = \lambda$ ,  $y_2 = \lambda^2$  usw. Also ist  $y_k = \lambda^k$ . Interpretiert  $D$  nun als Shiftoperator, dann erhält man die Gleichung

$$y_{k+n} + a_{n-1}y_{k+n-1} + a_1y_{k+1} + a_0y_k = 0. \quad (6.18)$$

Setzt man nun  $y_k = \mu^k$  als Eigenfunktion, dann liefert jede Lösung  $\mu$  von

$$p(\mu) = 0$$

eine Lösung der Differenzengleichung.

Greift diesen Gedanken für das Problem der Wärmeleitung auf, dann erhält man mit (6.16)

$$y_{k+1} - (2 + \lambda)y_k + y_{k-1} = 0$$

also

$$y_k = \alpha_1 \mu_1^k + \alpha_2 \mu_2^k, \mu_1 + \mu_2 = 2 + \lambda, \mu_1 \mu_2 = 1.$$

Dies liefert das Gleichungssystem

$$\begin{aligned} 0 &\equiv y_0 = \alpha_1 \mu_1^0 + \alpha_2 \mu_2^0 = \alpha_1 + \alpha_2, \\ y_1 &= \alpha_1 \mu_1^1 + \alpha_2 \mu_2^1, \\ &\vdots \\ y_{N-1} &= \alpha_1 \mu_1^{N-1} + \alpha_2 \mu_2^{N-1}, \\ 0 &\equiv y_N = \alpha_1 \mu_1^N + \alpha_2 \mu_2^N. \end{aligned}$$

Nach Voraussetzung von  $\mu_1, \mu_2$  liefert die letzte Bedingung

$$\mu_1^N - \mu_2^N = 0$$

und damit  $\mu_1^{2N} = 1$ . Die  $n$ -te Ableitung von  $\mu_1$  ist dann

$$\mu_1^{(n)} = \exp \left[ \frac{\pi i}{2N} n \right], n = 0, 1, \dots, N-1, N, N+1, \dots, 2N-1. \quad (6.19)$$

Folgende Fälle können nun unterschieden werden:

- $n = 0$ : Der Fall  $n = 0$  führt (6.19) auf  $y_k = 0$ .
- $1 \leq n \leq N-1$ : Der Fall  $1 \leq n \leq N-1$  ist unproblematisch
- $n = N$ : Der Fall  $n = N$  führt die Gleichung auf  $\mu_1 = -1$  und  $y_k = 0$ .
- $n \geq N+1$ : Der Fall  $n \geq N+1$  werden die Rollen von  $\mu_1$  und  $\mu_2$  getauscht.

Also erhält man

$$\begin{aligned}
 y_k^{(n)} &= \frac{1}{2} \left( \exp \left[ \pi i \frac{nk}{N} \right] - \exp \left[ -\pi i \frac{nk}{N} \right] \right) \\
 &= \frac{1}{2i} 2i \sin \pi \frac{nk}{N} \\
 &= \sin \frac{\pi nk}{N}, \quad n = 1, \dots, N-1.
 \end{aligned}$$

Für die Eigenwerte gilt dann

$$\begin{aligned}
 y_{k+1}^{(n)} - 2y_k^{(n)} + y_{k-1}^{(n)} &= \sin \frac{\pi n(k+1)}{N} - 2 \sin \frac{\pi nk}{N} + \sin \frac{\pi n(k-1)}{N} \\
 &= \sin \frac{\pi kn}{N} \cos \frac{\pi n}{N} - 2 \sin \frac{\pi nk}{N} \\
 &= 2 \left( \cos \frac{\pi n}{N} - 1 \right) \sin \frac{\pi kn}{N},
 \end{aligned}$$

dies entspricht wieder einer Eigenfunktion.

Für  $x_k = \frac{k}{N}$  erhält man

$$y(x_k) = \sin(\pi n x_k)$$

also im kontinuierlichen

$$y(x) = \sin(\pi n x),$$

dies wiederum liefert die Eigenfunktionen von  $\partial^2/\partial \mathbf{x}^2$  zu  $y(0) = y(1) = 0$ . Die zugehörigen Eigenwerte sind dann

$$\lambda_n = -\pi^2 n^2, \quad \tilde{\lambda}_n = 2N^2 \left( \cos \frac{\pi n}{N} - 1 \right),$$

im kontinuierlichen und diskreten Fall sind, wobei letzteres sich aus  $1/\Delta x^2 = N^2$  ergibt. Der Vergleich zeigt, dass je mehr Diskretisierungspunkte Gewählt werden,  $\Delta x \rightarrow 0$ , dass

$$\lambda_1 \approx -\pi^2, \quad \lambda_{N-1} \rightarrow -\infty.$$

Da nun die Eigenwerte durchweg einerseits negative Realteile, andererseits von stark unterschiedlicher Größenordnungen sind, liegt bei feinerem Ortsgitter ein steifes Anfangswertproblem vor. und es gilt △

$$u(t, x_k) = \sum_{n=1}^{N-1} \sin \left( \frac{\pi nk}{N} \right) \exp \left[ \hat{\lambda}_n t \right] y_{n,0},$$

wobei  $y_{n,0}$  sich aus dem Anfangswert ergibt mit  $y(0, x) = y_0(x)$ .

*Beispiel 6.4.3 (Fibonacci).* Betrachtet wird die Folge  $\{y_n\}_{n \in \mathbb{N}}$  mit

$$y_{n+2} = y_{n+1} + y_n.$$

Setzt man  $y_n = \lambda^n$ , dann erhält man die *Fibonacci-Differenzgleichung*

$$\lambda^{n+2} - \lambda^{n+1} - \lambda^n = 0, \quad \lambda_0 = 0, \quad \lambda_1 = 1. \quad (6.20)$$

Im einfachen Fall erhält man die charakteristische Funktion

$$\psi(\xi) = \xi^2 - \xi - 1 = 0$$

mit den Nullstellen

$$\xi_1 = \frac{1 + \sqrt{5}}{2}, \quad \xi_2 = \frac{1 - \sqrt{5}}{2}.$$

Die allgemeine Lösung von (6.20) ist dann

$$y_n = c_0 \left( \frac{1 + \sqrt{5}}{2} \right)^n + c_1 \left( \frac{1 - \sqrt{5}}{2} \right)^n$$

Einsetzen der Anfangswerte liefert dann  $c_0 = -c_1 = \sqrt{\frac{1}{5}}$ , womit

$$y_n = \left( \frac{1}{\sqrt{5}} \left( \frac{1 + \sqrt{5}}{2} \right)^n - \left( \frac{1 - \sqrt{5}}{2} \right)^n \right). \quad (6.21)$$

Dies liefert die Fibonacci-Zahlen  $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \dots$ .

## 6.5 Numerisches Fehlerverhalten

Betrachtet werden nun die expliziten und impliziten Euler-Verfahren für das Prothero/Robinson-Testproblem

$$\dot{y} = \lambda(y(t) - \varphi(t)) + \dot{\varphi}(t).$$

Um bei der numerischen Behandlung des Problems die Fehlerfortpflanzung zu untersuchen werden nun zwei numerische Lösungen  $u_m, \tilde{u}_m$  mit Fehler

$$\Delta u_m = \tilde{u}_m - u_m.$$

Im expliziten Euler-Verfahren erhält man als Vorschriften

$$\begin{aligned} u_{m+1} &= u_m + h(\lambda(u_m - \varphi(t_m))) + \dot{\varphi}(t_m), \\ \tilde{u}_{m+1} &= \tilde{u}_m + h(\lambda(\tilde{u}_m - \varphi(t_m))) + \dot{\varphi}(t_m), \end{aligned}$$

woraus sich für den Fehler

$$\Delta u_{m+1} = \Delta u_m + h\lambda\Delta u_m = (1 + h\lambda)\Delta u_m \quad (6.22)$$

ergibt.

Im impliziten Euler-Verfahren erhält man als Vorschriften

$$\begin{aligned} u_{m+1} &= u_m + h(\lambda(u_{m+1} - \varphi(t_{m+1}))) + \dot{\varphi}(t_{m+1}), \\ \tilde{u}_{m+1} &= \tilde{u}_m + h(\lambda(\tilde{u}_{m+1} - \varphi(t_{m+1}))) + \dot{\varphi}(t_{m+1}), \end{aligned}$$

woraus sich für den Fehler

$$\begin{aligned} \Delta u_{m+1} &= \Delta u_m + h\lambda\Delta u_{m+1} \frac{1}{(1 - h\lambda)} \Delta u_m \\ &= \frac{1}{1 - h\lambda} \Delta u_m \end{aligned} \quad (6.23)$$

ergibt.

Es stellt sich nun dabei die Frage, wann Fehler gedämpft werden. Sei  $z = h\lambda \in \mathbb{C}$ . Betrachtet man die qualitativen Funktionen

$$f(z) = 1 + z,$$

$$g(z) = \frac{1}{1 - z},$$

dann erhält man zusammen mit dem exakten Fall folgende Stabilitätsgebiete (vgl. Kapitel ??)

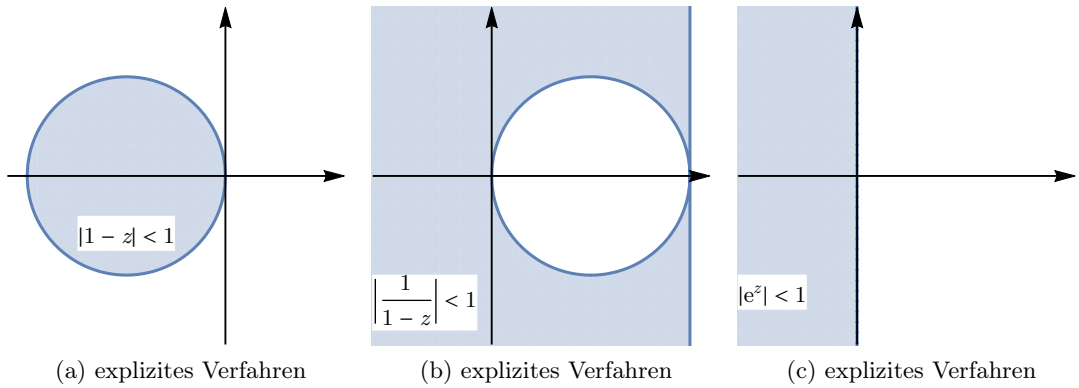


Abb. 6.3: Stabilitätsgebiet



---

## Inhaltsverzeichnis





## Implizite Runge/Kutta-Verfahren

### 7.1 Verfahrensvorschrift und Ordnung

Wie im Fall eines expliziten Verfahrens wird mit  $u_m$  die Näherungslösung zur Lösung der linken Seite einer Anfangswertaufgabe bezeichnet. Die Definition aus 3 aufgreifend, lässt sich ein implizites Runge/Kutta-Verfahren wie folgt definieren:

**Definition 7.1.1 (Runge/Kutta-Verfahren; implizit).** Ein implizites,  $s$ -stufiges Runge/Kutta-Verfahren für das Anfangswertproblem

$$\dot{y} = f(t, y), \quad y(t_0) = y_0 \quad (7.1)$$

ist gegeben durch

$$u_{m+1} = u_m + h \sum_{i \leq s} b_i f\left(t_m + c_i h, u_{m+1}^{(i)}\right), \quad (7.2a)$$

$$u_{m+1}^{(i)} = u_m + h \sum_{j \leq s} a_{ij} f\left(t_m + c_j h, u_{m+1}^{(j)}\right), \quad i = 1, \dots, s. \quad (7.2b)$$

Wie im Fall expliziter Runge/Kutta-Verfahren beschreibt (7.2b) wieder die *Aufdatierung*. △ Ferner lassen sich auch für implizite Runge/Kutta-Verfahren Butcher-Tableaus aufstellen. Sei  $\mathbf{A}$  die Verfahrensmatrix,  $\mathbf{c}$  der Knotenvektor und  $\mathbf{b}$  der Gewichtungsfaktor, dann

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array}$$

Eine alternative Formulierung von (7.2a) und (7.2b) wäre einerseits

$$u_{m+1} = u_m + h \sum_{i \leq s} b_i k_i(t_m, u_m; h), \quad (7.3a)$$

$$k_i(t_m, u_m; h) = f\left(t_m + c_i h, u_m + h \sum_{j \leq s} a_{ij} k_j(t_m, u_m; h)\right), \quad i = 1, \dots, s. \quad (7.3b)$$

, andererseits

$$u_{m+1} = u_m + h \sum_{i \leq s} b_i U_{m+1}^{(i)} \quad , \quad (7.4a)$$

$$u_m^{(i)} = u_{m+1} + \sum_{j \leq s} a_{ij} U_{m+1}^{(j)} \quad , i = 1, \dots, s. \quad (7.4b)$$

$$U_m^{(i)} = f \left( t_m + c_i h, u_{m+1}^{(i)} \right) \quad , i = 1, \dots, s. \quad (7.4c)$$

*Anmerkung 7.1.2.* (1) Das Verfahren lässt sich kompakt ähnlich der expliziten Behandlung auch in kompakter Schreibweise formulieren,

$$\frac{\mathbf{c} | \mathbf{A}}{\mathbf{b}^\top}$$

- (2) In der praktischen Anwendung kann es sinnvoll sein, lediglich die  $k_i$  und nicht die  $u_m^{(i)}$  zu speichern.
- (3) Die Darstellung (7.4a), (7.4b) und (7.4c) kann auch für implizite Differentialgleichungen verwendet werden. Denn gilt  $0 = F(t, y, \dot{y})$ , dann auch  $0 = F(t_m + c_i h, u_{m+1}^{(i)}, U_{m+1}^{(i)})$ .

### Ordnung

Wieder wird mit  $\eta(t_m, h) = y(t_{m+1}) - \tilde{u}_{m+1}$  bei  $u_m = y(t_m)$  der lokale Fehler an der Stelle  $t_{m+1}$  bezeichnet und mit  $\eta^* = y(t_m) - u_h(t_m)$  der globale Fehler bezeichnet. Es ist bekannt, dass für ein konsistentes Verfahren der Ordnung  $p$  mit  $\eta = \mathcal{O}(h^{p+1})$ , das Verfahren konvergent von der Ordnung  $p$  ist mit  $\eta^* = \mathcal{O}(h^p)$ .

### Ordnungsbedingung

Um die Ordnungsbedingungen zu bestimmen, wird die Taylor-Entwicklung der exakten und numerischen Lösung betrachtet. Im exakten Fall gilt

$$y(t_m + h) = y(t_m) + \sum_{k=1}^p \frac{h^k}{k!} \sum_{\tau \in \mathcal{LT}_k} F(\tau)$$

$$u_{m+1} = y(t_m) + \sum_{k=1}^p \frac{h^k}{k!} \sum_{\tau \in \mathcal{LT}_k} \gamma(\tau) \Phi_{s+1}(\tau) F(\tau)$$

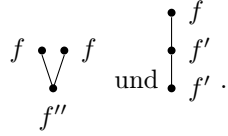
mit Bäumen aus  $\mathcal{T}$  bzw. ausgezeichneten Bäumen  $\mathcal{LT}$ , wobei mit  $\square \in \mathcal{T}$ , und  $\tau_1, \dots, \tau_k \in \mathcal{T}$  auch  $[\tau_1, \dots, \tau_k] \in \mathcal{T}$  gilt. Das elementare Differential  $y'' = f_y f$  wird dargestellt durch den Baum

$$\begin{array}{c} \bullet \quad f \\ | \\ \bullet \quad f' \end{array} .$$

Leitet man weiter ab,

$$y''' = f_{yy}(f, f) + f_y f_y f,$$

dann erhält man die Baumstruktur



Zusammenfassend erhält man :

$\tau = []$		$\tau = [\tau_1, \dots, \tau_k]$	
$\gamma$	1	$\rho(\tau)$	$\gamma(\tau_1) \cdots \gamma(\tau_k)$
$\Phi_i$	$\sum_j a_{ij}$	$\sum_j a_{ij} \Phi_j$	$(\tau_1) \cdots \Phi_j(\tau_k)$
$F$	$f$	$f_{y \dots y}$	$(F(\tau_1), \dots, F(\tau_k))$
$\rho$	1	$1 + \rho(\tau_1) + \cdots + \rho(\tau_k)$	

Tabelle 7.1: Ordnungsbedingungen für B-Bäume

Ferner gilt für die Ordnungsbedingungen, dass ein Runge/Kutta-Verfahren genau dann die Ordnung  $p$  hat, wenn die Bedingung

$$\Phi_{s+1}(\tau) = \frac{1}{\gamma(\tau)}, \quad \rho(\tau) \leq p$$

erfüllt ist. An dieser Stelle sei der Übersicht halber an die Ordnungsbedingungen für Runge/Kutta-Verfahren bis  $p = 4$  erinnert (vgl. Tabelle 3.1).

## 7.2 Vereinfachende Bedingungen

Für die Konstruktion impliziter Runge/Kutta-Verfahren können ähnlich den expliziten Runge/Kutta-Verfahren für die Koeffizienten  $\mathbf{A}$ ,  $\mathbf{b}$  und  $\mathbf{c}$  aus den Butcher-Tableaus hilfreiche Bedingungen für eine vereinfachte Konstruktion formuliert werden.

**Definition 7.2.1 (vereinfachende Bedingungen).** *Die Bedingungen*

$$B(p) : \quad \sum_{i \leq s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \dots, p \quad (7.5)$$

$$C(q) : \quad \sum_{j \leq s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad k = 1, \dots, q, \quad i = 1, \dots, s \quad (7.6)$$

$$D(r) : \quad \sum_{i \leq s} b_i c_i^{k-1} a_{ij} = \frac{1}{k} b_j (1 - c_j^k), \quad k = 1, \dots, r, \quad j = 1, \dots, s \quad (7.7)$$

heißen vereinfachende Bedingungen.

*Anmerkung 7.2.2.* Die vereinfachenden Bedingungen  $B(p)$  und  $C(q)$  lassen eine einfache Interpretation zu. △

(1)  $B(p)$  ist äquivalent zu den Ordnungsbedingungen für alle *buschartigen Bäume*, also Bäume mit Höchsttiefe 1,



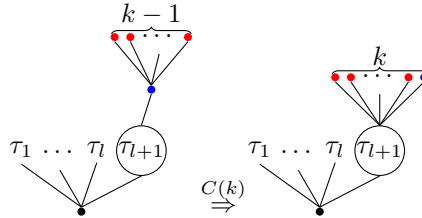
(2)  $C(q)$  ist gerade die Stufenordnung  $q$ . Sei  $\tau$  ein Baum, der an einer beliebigen Stelle einen Teilbaum der Form

$$\tau_1 = \begin{array}{c} \bullet \quad \bullet \quad \cdots \quad \bullet \\ \diagup \quad \diagdown \quad \diagup \quad \diagdown \\ \bullet \end{array} = [[^{k-1}]]$$

mit  $k + 1$  Knoten besitzt. Das elementare Gewicht  $\Phi(\tau)$  enthält den Faktor  $\sum_j a_{ij} c_j^{k-1}$ . Reduziert man den Baum  $\tau_1$  auf einen buschartigen Baum  $t_1 = \begin{array}{c} \bullet \quad \bullet \quad \cdots \quad \bullet \\ \diagup \quad \diagdown \quad \diagup \quad \diagdown \\ \bullet \end{array}$  mit  $k + 1$  Knoten. Das elementare Gewicht des zugehörigen neuen Baum  $t$  besitzt dann Faktor  $c_i^k$ , anstelle von  $\sum_j a_{ij} c_j^{k-1}$ . Für die Dichte von  $\tau_1$  und  $t_1$  gilt

$$\gamma(\tau_1) = (k + 1)k, \quad \gamma(t_1) = k + 1.$$

Damit folgt aus der Ordnungsbedingung  $\Phi(\tau) = \frac{1}{\gamma(\tau)}$  für den Baum  $\tau$  sofort die Ordnungsbedingung  $\Phi(t) = \frac{1}{\gamma(t)}$  für den Baum  $t$ , falls  $C(k)$  erfüllt ist.



Verwendet man für die beiden Bäume die Operation  $[\cdot]$ , dann lässt sich die die Reduktion durch

$$[\tau_1, \dots, \tau_l, [\dots [[[\cdot], \dots, [\cdot]]]]] \xRightarrow{C(k)} [\tau_1, \dots, \tau_l, [\dots [[[\cdot], \dots, [\cdot]]]]]$$

darstellen.

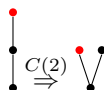
Durch die wiederholte Anwendung der Reduktion erhält man:



**Lemma 7.2.3 (Strehmel et al., 2012).** *Gilt die vereinfachende Bedingung  $C(q)$ , und lässt sich ein Baum  $\tau$  auf einen Baum  $t$  dadurch reduzieren, dass alle Teilbäume  $\tau_i$  mit  $q+1$  Knoten durch buschartige Teilbäume  $t_i$  mit  $q+1$  Knoten ersetzt werden, dann sind folgende folgende Aussagen äquivalent:*

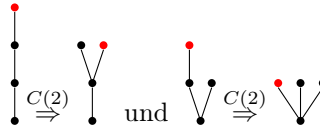
- (a)  $\Phi(\tau) = \frac{1}{\gamma(\tau)}$ .
- (b)  $\Phi(t) = \frac{1}{\gamma(t)}$ .

Beispiel 7.2.4 ( $C(2)$ ).

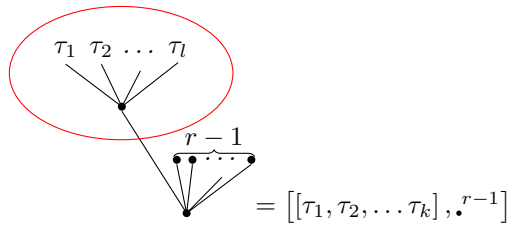
(1)  $p = 3$  : Der Baum  $\begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \bullet \end{array}$  kann mit  $C(2)$  wie folgt reduziert werden



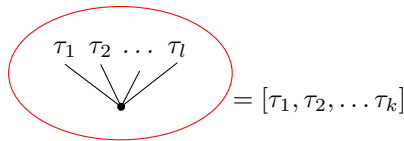
(2)  $p = 4$ : Die Bäume  und  können mit  $C(2)$  wie folgt reduziert werden



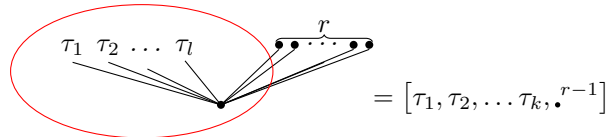
**Lemma 7.2.5 (Strehmel et al. 2012).** *Gilt die vereinfachende Bedingung  $D(r)$ , so ist die Ordnungsbedingung für einen Baum der Form*



*automatisch erfüllt, falls die Ordnungsbedingungen für*

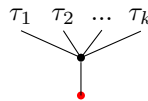


*und*

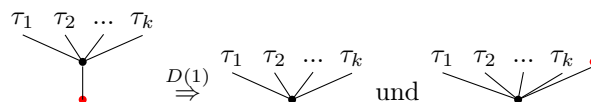


*erfüllt sind.*

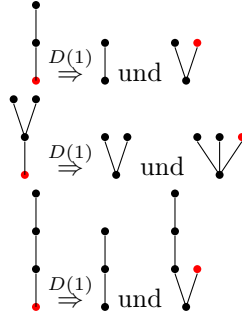
**Beispiel 7.2.6 ( $D(1)$ ).** Bäume der Form



werden mit Lemma 7.2.5 und  $D(1)$  wie folgt reduziert



Für  $p \leq 4$  ergibt sich



**Theorem 7.2.7.** *Ein Verfahren mit den Vereinfachenden Bedingungen  $B(p)$ ,  $C(q)$  und  $D(r)$  hat die Ordnung*

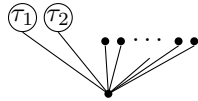
$$p' = \min \{p, 2q + 2, q + r + 1\}. \quad (7.8)$$

*Beweis.* Es ist zu zeigen, dass die Ordnungsbedingungen für alle Bäume mit bis zu  $p$  Knoten mithilfe der vereinfachenden Bedingungen  $C(q)$  und  $D(r)$  auf die Ordnungsbedingungen für buschartige Bäume gleicher Knotenzahl reduziert werden können. Sei  $\tau = [\tau_1, \tau_2, \dots, \tau_k]$  ein beliebiger Baum mit  $\rho^* := \rho(\tau)$   $\rho^* \leq 2q + 2$  und  $\rho^* \leq q + r + 1$ .

Besitzt jeder Teilbaum  $[\tau_\nu]$  höchstens  $q + 1$  Knoten, dann kann nach Bemerkung 7.2.2 der Baum  $\tau$  auf einen buschartigen Baum der Tiefe 1 mit  $\rho^*$  Knoten reduzieren werden. Die Bedingung  $C(q)$  besagt nun,

Die vereinfachenden Bedingungen  $C(q)$  und  $D(r)$  können iterativ angewandt werden, bis eine Reduktion nicht mehr möglich ist. Dann können zwei Alternativen getrennt untersucht werden:

Sei zunächst angenommen, an der Wurzel des Baumes  $\tau$  hängen mindestens zwei Bäume mit einer Mindesttiefe von zwei

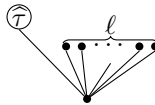


Da eine Reduktion mittels  $C(q)$  nicht anwendbar ist, gilt Dann  $\rho(\tau_1), \rho(\tau_2) \geq q + 1$ , so dass

$$\rho(\tau) \geq 2(q + 1) + 1 = 2q + 3,$$

was zum Widerspruch führt.

Sei nun angenommen, an der Wurzel des Baum  $\tau$  hängt genau ein Baum  $\hat{\tau}$  mit Mindesttiefe von eins,



Lässt sich der Baum  $\tau$  einerseits nicht weiter über  $C(q)$  weiter reduzieren, dann gilt  $\rho(\hat{\tau}) \geq q + 1$ , und andererseits nicht weiter über  $D(r)$  reduzieren, dann muss  $\ell \geq r$  gelten. In diesem Fall gilt

$$\rho(\tau) = q + 1 + r + 1 = q + r + 2,$$

was ebenfalls zum Widerspruch führt. Damit wurde die Behauptung gezeigt.  $\square$

△ Explizite Verfahren erfüllen höchstens die vereinfachende Bedingung  $C(1)$ , d.h.

$$c_i = \sum_{j \leq s} a_{ij}, \quad i = 1, \dots, s,$$

denn für  $C(2)$  gilt

$$\frac{1}{2}c_i^2 = \sum_{j \leq s} a_{ij}c_j, \quad i = 1, \dots, s. \quad (7.9)$$

Insbesondere für  $i = 2$  gilt

$$\frac{1}{2}c_2^2 = a_{21} = 0,$$

was offensichtlich keine sinnvolle zweite Stufe sein kann.

**Lemma 7.2.8.** *Für ein  $s$ -stufiges Runge/Kutta-Verfahren seien die Knoten  $c_1, \dots, c_s$  paarweise verschieden. Für ein  $m \in \mathbb{N}$  mit  $1 \leq m \leq s$  gilt:*

- (1) *Aus den vereinfachenden Bedingungen  $B(s+m)$  und  $C(s)$  folgt stets die vereinfachende Bedingung  $D(m)$ .*
- (2) *Aus den vereinfachenden Bedingungen  $B(s+m)$  und  $D(s)$  und den Gewichten  $b_i \neq 0$  für  $i = 1, \dots, s$  folgt stets die vereinfachende Bedingung  $C(m)$ .*

*Beweis.* Zunächst sei

$$d_j^{(k)} := \sum_{i \leq s} b_i c_i^{k-1} a_{ij} - \frac{b_j}{k} (1 - c_j^k), \quad j = 1, \dots, s, \quad k = 1, \dots, m. \quad (7.10)$$

Für dieses  $d_j^{(k)}$  ist zu zeigen, dass  $d_j^{(k)} \equiv 0$  für alle  $j = 1, \dots, s$  gilt. Nach Voraussetzung einerseits

$$\frac{1}{\nu} = \sum_{i \leq s} b_i c_i^{\nu-1}, \quad \nu = 1, \dots, s+m,$$

andererseits

$$\frac{1}{\nu} c_i^\nu = \sum_{j \leq s} a_{ij} c_j^{\nu-1}, \quad i = 1, \dots, s, \quad \nu = 1, \dots, s.$$

Eingesetzt in (7.10) liefert dies

$$\begin{aligned} \sum_{j \leq s} d_j^{(k)} c_j^{\nu-1} &= \sum_{i, j \leq s} b_i a_{ij} c_i^{k-1} c_j^{\nu-1} - \sum_{j \leq s} \frac{b_j}{k} (1 - c_j^k) c_j^{\nu-1} \\ &= \sum_{i \leq s} b_i c_i^{k-1} \sum_{j \leq s} a_{ij} c_j^{\nu-1} - \frac{1}{k} \sum_{j=1}^s b_j c_j^{\nu-1} + \frac{1}{k} \sum_{j \leq s} b_j c_j^{k+\nu-1} \\ &= \frac{1}{\nu} \sum_{i \leq s} b_i c_i^{k+\nu-1} - \frac{1}{k\nu} + \frac{1}{k(k+\nu)} \\ &= \frac{1}{\nu(k+\nu)} - \frac{1}{k\nu} + \frac{1}{k(k+\nu)} \\ &= 0. \end{aligned} \quad (7.11)$$

Bezeichne nun

$$\mathbf{d}^{(k)} := (d_1^{(k)}, \dots, d_s^{(k)}) \quad \text{und} \quad \mathbf{V} := \begin{bmatrix} 1 & c_1 & \dots & c_1^{s-1} \\ \vdots & & & \vdots \\ 1 & c_s & \dots & c_s^{s-1} \end{bmatrix}.$$

Dann lässt sich (7.11) als die Beziehung

$$0 = \sum_{j \leq s} d_j^{(k)} c_j^{\nu-1} = \mathbf{d}^{(k)} \mathbf{V}$$

schreiben, und da die  $c_1, \dots, c_s$  paarweise verschieden sind, ist  $\mathbf{V}$  regulär, womit  $\mathbf{d}^{(k)} = \mathbf{0}$  gelten muss. Damit wurde (1) gezeigt.

Um die Behauptung (2) zu zeigen setzt man

$$c_i^{(k)} = \sum_{i \leq s} a_{ij} c_j^{k-1} - \frac{c_i^k}{k}, \quad (7.12)$$

und es gilt dann analog

$$\begin{aligned} \sum_{i \leq s} b_i c_i^{\nu-1} c_i^{(k)} &= \sum_{i \leq s} \sum_{j \leq s} b_i c_i^{\nu-1} a_{ij} c_j^{k-1} - \frac{1}{k} \sum_{i \leq s} b_i c_i^{k+\nu+1} \\ &= \sum_{j \leq s} \frac{1}{\nu} b_j (1 - c_j^\nu) c_j^{k-1} - \frac{1}{k(k+\nu)} \\ &= \frac{1}{\nu k} - \frac{1}{\nu(k+\nu)} - \frac{1}{k(k+\nu)} \\ &= 0. \end{aligned} \quad (7.13)$$

Bezeichne nun

$$\mathbf{c}^{(k)} := (c_1^{(k)}, \dots, c_s^{(k)}) \quad \text{und} \quad \mathbf{B} := \text{diag}(b_1, \dots, b_s),$$

dann lässt sich (7.13) als die Beziehung

$$0 = \sum_{j \leq s} b_j c_j^{(k)} c_j^{\nu-1} = \mathbf{c}^{(k)} \mathbf{B} \mathbf{V}$$

schreiben. Nach Voraussetzung sind die Gewichte  $b_i \neq 0$  für  $i = 1, \dots, s$ , womit  $\mathbf{B} \mathbf{V}$  regulär ist, und folglich  $\mathbf{c}^{(k)} = \mathbf{0}$  gelten muss, womit auch (2) gezeigt wurde.  $\square$

Aus dem letzten Lemma 7.2.8 erhält man das folgende Ergebnis:

**Korollar 7.2.9.** *Ein Runge/Kutta-Verfahren mit den vereinfachenden Bedingungen  $B(s+m)$  und  $C(s)$  hat die Ordnung  $p = s + m$ .*

△

Dieses Ergebnis erinnert an das Koallakationsverfahren. Erfüllt ein  $s$ -stufiges Runge/Kutta-Verfahren mit paarweise verschiedenen  $c_i$  und Ordnung  $p = s + m$  sowie die vereinfachende Bedingung  $C(s)$ , dann können durch die geeignete Vorgabe des Knotenvektors  $\mathbf{c}$  die Gewichte  $b_i$  eindeutig berechnet werden.

Aus dem Lemma 7.2.8 erhält man aber auch folgende Ergebnis:

**Korollar 7.2.10.** *Ein Runge/Kutta-Verfahren mit den vereinfachenden Bedingungen  $B(s+m)$  und  $D(s)$  hat die Ordnung  $p = s + m$ , falls  $b_i \neq 0$ ,  $i = 1, \dots, s$ , und  $m \geq s - 2$  gilt.*



### 7.3 Implizite Runge/Kutta-Verfahren höherer Ordnung

Um im Folgenden Verfahren höherer Ordnung  $p \geq s - 2$  konstruieren zu können, sei zunächst an die numerische Integration erinnert. Dieser Ansatz der *Quadraturformel* unterstellt, dass für gegebene  $n + 1$  Stützstellen  $a \leq x_0 < x_1 < \dots < x_n \leq b$  das Integral von  $f$  auf dem Intervall  $[a, b]$  „exakt“ durch die Summe der gewichteten Funktionswerte

$$Q_{n+1}(f) := \sum_{\nu=0}^n \omega_{\nu} f(x_{\nu}), \quad \omega_{\nu} \in \mathbb{R}, \quad \nu = 0, \dots, n \quad (7.14)$$

dargestellt werden kann, also

$$Q_{n+1}(f) \approx \int_a^b f(x) dx \quad (7.15)$$

gilt. Man ist also danach bestrebt eine Quadraturformel zu bestimmen, so dass die Konvergenz

$$\lim_{n \rightarrow \infty} Q_{n+1}(f) = \int_a^b f(x) dx \quad \text{bzw.} \quad \lim_{n \rightarrow \infty} R_{n+1}(f) = 0 \quad (7.16)$$

für die Quadraturformel bzw. für den *Quadraturfehler*  $R_{n+1}(f) := \int_a^b f(x) dx - Q_{n+1}(f)$  gilt.

Eine wichtige Klasse von Quadraturformeln basieren auf der Interpolation der Funktion  $f$  durch ein Polynom  $p \in \mathbb{P}_n$  vom Grad  $n$ . Eine Quadraturformel  $Q_{n+1}$  heißt *Interpolationsquadratur*, falls das Polynom von Grad  $n$  exakt integriert wird

$$R_{n+1}(p_n) = 0.$$

Werden für die Interpolation von  $f$  die Lagrange-Interpolation

$$p(x) = \sum_{\nu=0}^n f_{\nu} \ell_{\nu}(x), \quad \ell_{\nu}(x) := \prod_{\substack{i=0 \\ i \neq \nu}}^n \frac{x - x_{\nu}}{x_i - x_{\nu}}$$

verwendet, dann sind die Gewichte  $\omega_{\nu}$  offensichtlich gerade die Lagrange-Polynome. In diesem Fall gilt

$$\int_a^b p(x) dx = \sum_{\nu=0}^n f_{\nu} \int_a^b \ell_{\nu}(x) dx.$$

Aus der numerischen Mathematik ist bekannt, dass *Newton/Cotes-Formeln* für  $n + 1$  Stützstellen bei ungeradem  $n$ , Polynome vom Grad höchstens  $n$  exakt, bei geradem  $n$  Polynome vom Grad höchstens  $n + 1$  integriert werden können. Durch die Wahl geeigneter Stützstellen und geeigneter Polynome können aber auch Quadraturformeln mit höherer Genauigkeit gewonnen werden.

**Theorem 7.3.1 (Quadraturformel; maximale Ordnung).** *Können die Stützstellen  $x_0, \dots, x_n$  einer Quadraturformel*

$$Q_{n+1}(f) = \sum_{\nu=0}^n \omega_{\nu} f_{\nu}, \quad \omega_{\nu} \in \mathbb{R}$$

*paarweise verschieden und frei gewählt werden, dann integriert die Quadraturformel bestensfalls Polynome vom Höchstgrad  $2n + 1$ .*

*Beweis.* Sei  $m = n + 1$  und seien  $x_0, \dots, x_{m-1}$  paarweise verschiedene Stützstellen. Ferner sei

$$M(x) := \prod_{\nu \leq m} (x - x_\nu)^2$$

ein Polynom vom Grad  $2m$ . Ohne Beschränkung kann für das Integrationsintervall  $[a, b] = [-1, 1]$  gewählt werden. Dann gilt einerseits

$$\int_{-1}^1 M(x) dx > 0,$$

andererseits

$$Q_{n+1}(f) = \sum_{\nu=0}^n \omega_\nu M(x_\nu) = 0.$$

Es werden also bestenfalls höchstens Polynome  $p$  vom Grad  $\deg p \leq 2m - 1$  integriert, womit die Behauptung gezeigt wurde.  $\square$

Eine Klasse von Quadraturformeln, die Polynome vom Höchstgrad  $2n + 1$  integriert sind die *Gauß-Quadraturformeln*. Der Ansatz geht hier davon aus, dass die zu integrierende Funktion  $f$  in eine mit  $\omega$  gewichtete Funktion  $\Phi$  zerlegt werden kann,  $f(x) = \omega(x) \Phi(x)$ , wobei sich die Funktion  $\Phi(x)$  durch ein Polynom  $M$  approximiert werden kann. In diesem Fall gilt dann die Approximation

$$\int_a^b f(x) dx = \int_a^b \omega(x) \Phi(x) dx \approx \int_a^b \omega(x) M(x) dx = \sum_{\nu=0}^n \omega_\nu \Phi_\nu, \quad \Phi_\nu \equiv f(x_\nu).$$

$\triangle$

Dies erlaubt einen weiteren Zugang für die vereinfachende Bedingungen  $B(p)$  und  $C(q)$ . Betrachtet man das Anfangswertproblem

$$\dot{y} = f(t), \quad y(t_m) = 0.$$

Die exakte Lösung ist dann

$$y(t_m + h) = \int_{t_m}^{t_m+h} f(t) dt.$$

Bezeichnet  $\tilde{u}_{m+1}$  die Näherungslösung des exakten Problems, das bei  $t_m$  startet, dann liefert ein  $s$ -stufige Runge/Kutta-Verfahren

$$\tilde{u}_{m+1} = h \sum_{i \leq s} b_i f(t_m + c_i h).$$

Wird nun  $f(t)$  durch ein Polynom  $M(t) = (t - t_m)^{k-1}$  approximiert, und setzt man  $\omega(t) \equiv 1$ , dann gilt

$$y(t_m + h) = \int_{t_m}^{t_m+h} (t - t_m)^{k-1} dt = \frac{1}{k} h^k, \quad \tilde{u}_{m+1} = h^k \sum_{i \leq s} b_i c_i^{k-1}.$$

Damit wird der lokale Diskretisierungsfehler

$$\eta(t_m + h) = \left( \frac{1}{k} - \sum_{i \leq s} b_i c_i^{k-1} \right) h^k$$

minimal, wenn  $B(p)$  erfüllt ist.

**Korollar 7.3.2.** *Sei angenommen, ein Runge/Kutta-Verfahren erfüllt die Bedingung  $B(p)$ . Dann gilt.*

- (1) *Polynome vom Höchstgrad  $p - 1$  werden exakt integriert.*
- (2) *Für Polynome vom Höchstgrad  $p - 1$  besitzt die Quadraturformel den Genauigkeitsgrad  $p - 1$ .*

Betrachtet man für ein  $s$ -stufiges Runge/Kutta-Verfahren anstelle Näherungslösungen  $\tilde{u}_{m+1}$  am Ende eines Schrittes die Näherungslösungen  $\tilde{u}_{m+1}^{(i)}$  der Zwischenschritte, dann liefert der gleiche Ansatz mittels Quadratur über das Polynom  $M(t)$  und Gewichtsfunktion  $\omega(t)$

$$y(t_m + c_i h) \int_{t_m}^{t_m + c_i h} M(t) dt = \frac{1}{k} c_i^k h^k, \quad \tilde{u}_{m+1}^{(i)} = h \sum_{j \leq s} a_{ij} c_j^{k-1} h^{k-1}.$$

Dann wird der lokale Diskretisierungsfehler der  $i$ -ten Stufe

$$\eta(t_m + c_i h) = \left( \frac{1}{k} c_i^k - \sum_{j \leq s} a_{ij} c_j^{k-1} \right) h^k$$

minimal, wenn  $C(q)$  erfüllt ist.

Bei der weiteren Betrachtung kann man sich auf ein Intervall  $[-1, 1]$  beschränken, da mittels  $\triangle$  Variablentransformation jedes Intervall  $[a, b]$  auf ein Intervall  $[-1, 1]$  zurückgeführt werden kann.

Damit ein möglichst hoher Exaktheitsgrad erreicht werden kann, müssen die Stützstellen  $x_0, \dots, x_n$  paarweise voneinander verschieden sein. Dies ist dann der Fall, wenn sie die Nullstellen eines  $n$ -ten orthogonalen Polynoms  $M_n$  sind. Eine Klasse von Polynomen  $[-1, 1]$  für die diese Eigenschaft gilt sind die Legendre-Polynome:

**Definition 7.3.3 (Legendre-Polynom).** *Ein Polynom  $L_n$  vom Grad  $n$  der Form*

$$L_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n} \left( (x^2 - 1)^n \right), \quad n \in \mathbb{N}_0 \quad (7.17)$$

heißt Legendre-Polynom.

Speziell sind

$$\begin{aligned} L_0(x) &= 1, & L_1(x) &= x, \\ L_2(x) &= \frac{1}{2} (3x^2 - 1), & L_3(x) &= \frac{1}{2} (5x^3 - 3x). \end{aligned}$$

Eine wichtige Eigenschaft von Legendre-Polynomen ist ihre Orthogonalität:

**Lemma 7.3.4.** *Für  $m, n \in \mathbb{N}_0$  gilt*

$$\langle L_m, L_n \rangle := \int_{-1}^1 L_m(x) L_n(x) dx = \frac{1}{2n+1} \delta_{mn}. \quad (7.18)$$

Eine weitere Eigenschaft von Legendre-Polynomen ist die Tatsache, dass ein Legendre-Polynom  $L_n$  vom Grad  $n$  genau  $n$  verschiedene Nullstellen besitzt:

**Lemma 7.3.5.** Für alle  $m \in \mathbb{N}_0$  besitzt ein Legendre-Polynom  $L_m(x)$  genau  $n$  einfache Nullstellen.

Mit diesen Eigenschaften von Legendre-Polynomen gilt, dass Gauß-Quadraturformeln mit  $m$  Knoten Polynome vom Höchstgrad  $2m - 1$  exakt integrieren.

**Theorem 7.3.6 (Gauß-Quadraturformel).** Sei  $m = n + 1$ . Sind  $x_0, \dots, x_n$  die Nullstellen des  $m$ -ten Legendre-Polynoms  $L_m$ , dann gibt es genau eine Quadraturformel, die Gauß-Quadraturformel,

$$Q_{n+1}^G(f) = \sum_{\nu=0}^n \omega_\nu f(x_\nu), \quad x_\nu \in [-1, 1], \quad (7.19)$$

die Polynome vom Höchstgrad  $2m - 1$  exakt integriert, wobei die Gewichte

$$\omega_\nu = \int_{-1}^1 \prod_{\substack{j=0 \\ j \neq \nu}}^n \left( \frac{x - x_i}{x_\nu - x_j} \right)^2 dx, \quad \nu = 0, \dots, n \quad (7.20)$$

gerade die Lagrange-Polynome sind.

*Beweis.* Der Beweis kann in drei Schritten gezeigt werden ?.

**Schritt 1:** Zunächst ist zu zeigen, dass eine solche Quadratformel existiert. Wegen Lemma 7.3.5 besitzt  $L_m(x)$   $m = n + 1$  verschiedene Nullstellen. Zu jedem dieser Nullstellen existiert nun eine offene Newton/Cotes-Formel, die Polynome vom Mindestgrad  $n = m - 1$  exakt integriert. Sei nun  $q(x)$  ein Polynom vom Höchstgrad  $\deg q \leq m - 1$  und  $r(x)$  ein Polynom vom Höchstgrad  $\deg(r) \leq m - 1$ , so dass ein Polynom  $p(x)$  der Form

$$p(x) = q(x) L_m(x) + r(x)$$

existiert, und für dessen Integral

$$\int_{-1}^1 p(x) dx = \int_{-1}^1 q(x) L_m(x) dx + \int_{-1}^1 r(x) dx$$

gilt. Da nach Lemma 7.3.4 paarweise verschiedene Legendre-Polynome zueinander orthogonal sind, lässt sich das Polynom  $q$  als Linearkombination

$$q(x) = \sum_{\nu=0}^{m-1} \lambda_\nu L_\nu(x)$$

dargestellt werden, so dass dann aber

$$\int_{-1}^1 q(x) L_m(x) dx = \sum_{\nu=0}^{m-1} \lambda_\nu \int_{-1}^1 L_\nu(x) L_m(x) dx = 0,$$

d.h.

$$\int_{-1}^1 p(x) dx = \int_{-1}^1 r(x) dx.$$

Damit gilt aber auch für die Newton/Cotes-Formel bezüglich der Nullstellen, dass

$$\sum_{\nu=0}^n \omega_{\nu} p(x_{\nu}) = \sum_{\nu=0}^n \omega_{\nu} r(x_{\nu}).$$

Da Newton/Cotes-Formeln aber Polynome vom Höchstgrad  $m - 1$  bei  $m$  Stützstellen aber exakt integrieren, gilt dann die Gleichheit

$$\sum_{\nu=0}^n \omega_{\nu} r(x_{\nu}) dx = \int_{-1}^1 r(x) dx \int_{-1}^1 p(x) dx.$$

Damit integriert die Gauß-Quadraturformel  $Q_{n+1}^G(f)$  jedes Polynom vom Höchstgrad  $2m - 1$  exakt.

Schritt 2: Als nächstes ist zu zeigen, dass die Gewichte der Gauß-Quadraturformel tatsächlich die Lagrange-Polynome sind. Nach Definition des Lagrange-Polynoms ist  $\ell_{\nu}(x)$  ein Polynom vom Grad  $n = m - 1$ . Wie bereits gezeigt wurde, integriert die Newton/Cotes-Formel an den Nullstellen des Legendre-Polynom jedes Polynom vom Höchstgrad  $2m - 1$  exakt. Da  $\ell_{\nu}^2(x)$  ein Polynom vom Grad  $2m - 2$  ist, wird dieses nun auch exakt integriert. Damit gilt

$$\int_{-1}^1 \ell_{\nu}^2(x) dx = \int_{-1}^1 \prod_{\substack{i=0 \\ i \neq \nu}}^n \left( \frac{x - x_i}{x_{\nu} - x_i} \right)^2 = \sum_{k=0}^n \omega_k \ell_{\nu}(x_k) = \omega_k \delta_{k\nu} = \omega_{\nu} \quad (7.21)$$

Damit wurden die Form der Gewichte gezeigt.

Schritt 3: Es bleibt zu zeigen, dass die Gauß-Quadraturformel eindeutig ist. Sei angenommen, es gebe eine weitere Gauß-Quadraturformel der Form

$$V_{n+1}^G(f) := \sum_{\nu=0}^n \rho_{\nu} f(x_{\nu})$$

mit paarweise verschiedenen Stützstellen  $y_0, \dots, y_n$  und Gewichten  $\rho_{\nu}$ . Nach den bisherigen Ergebnissen gibt es dann Polynome vom Höchstgrad  $2m - 1$ , die exakt integriert werden. Sei  $\widehat{\ell}_{\nu}(y)$  ein zweites Lagrange-Polynom bezüglich  $y$ , und  $v(y)$  ein Polynom vom Grad  $2m - 1 = 2m + 1$  der Form  $g(y) := \widehat{\ell}_{\nu}(y) L_m(y)$ , denn  $L_m(y)$  ist vom Grad  $m$  und  $\widehat{\ell}_{\nu}(y)$  ein Polynom vom Grad  $n$ . Da die Gauß-Quadraturformel das Polynom  $v(y)$  dann exakt integriert gilt

$$\begin{aligned} \int_{-1}^1 v(y) dy &= \int_{-1}^1 \widehat{\ell}_{\nu}(y) L_m(y) dy \\ &= \sum_{k=0}^n \rho_k \widehat{\ell}_{\nu}(y_k) L_m(y_k) \\ &= \sum_{k=0}^n \rho_k \delta_{k\nu} L_m(y_i) = \rho_{\nu} L_m(y_{\nu}). \end{aligned}$$

Nutzt man wieder die Orthogonalität der Legendre-Polynom, dann lässt sich  $\widehat{\ell}_{\nu}(x)$  wieder als Linearkombination  $\sum_{k=0}^n \lambda_{\nu} L_k(x)$  darstellen. Damit gilt aber auch, dass

$$\int_{-1}^1 \widehat{\ell}_{\nu}(y) L_m(y) dy = \rho_{\nu} L_m(y_{\nu}) = 0$$

gilt. Da die Gewichte  $\rho_{\nu}$  echt positiv sein müssen, muss  $L_m(y_{\nu}) = 0$  gelten. Damit sind aber die Stützstellen  $y_0, \dots, y_n$  gerade wieder die Nullstellen des Legendre-Polynoms, und damit gilt  $y_{\nu} = x_{\nu}$  für alle  $\nu = 0, \dots, n$ . Damit gilt aber auch nach Definition der Gewichte, dass  $\omega_{\nu} = \rho_{\nu}$  für alle  $\nu = 1, \dots, n$  gilt.

Damit wurde der Satz gezeigt.  $\square$

$\triangle$  Da offensichtlich  $\int_{-1}^1 \ell_\nu^2(x) dx > 0$  ist, besagt das Ergebnis (7.21), dass die Gewichte positiv sind.

Mit diesen Überlegungen können nun Verfahren der Ordnung  $p \geq s-2$  gesucht werden. Führt man für die vereinfachenden Bedingungen 7.5, (7.5) und (7.7) folgende Matrixschreibweisen

$$\begin{aligned} \mathbf{V} &:= \left( c_i^{j-1} \right)_{ij} = \begin{bmatrix} 1 & c_1 & \dots & c_1^{s-1} \\ \vdots & & & \vdots \\ 1 & c_s & \dots & c_s^{s-1} \end{bmatrix}, \quad \mathbf{N} := \text{diag} \left( \frac{1}{j} \right) = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & \frac{1}{s} \end{bmatrix} \\ \mathbf{B} &:= \text{diag}(b_i) = \begin{bmatrix} b_1 & & \\ & \ddots & \\ & & b_s \end{bmatrix}, \quad \mathbf{C} := \text{diag}(c_i) = \begin{bmatrix} c_1 & & \\ & \ddots & \\ & & c_s \end{bmatrix} \end{aligned} \quad (7.22)$$

ein, dann lassen sich diese in der Form

$$\begin{aligned} B(s): \quad & \mathbf{b}^\top \mathbf{V} = \mathbf{1}^\top \mathbf{N} \\ C(s): \quad & \mathbf{A} \mathbf{V} = \mathbf{C} \mathbf{V} \mathbf{N} \\ D(s): \quad & \mathbf{V}^\top \mathbf{B} \mathbf{A} = \mathbf{N} \mathbf{1} \mathbf{b}^\top - \mathbf{N} \mathbf{V}^\top \mathbf{C} \mathbf{B} \\ & = \mathbf{N} (\mathbf{1} \mathbf{1}^\top - \mathbf{V}^\top \mathbf{C}) \mathbf{B} \end{aligned} \quad (7.23)$$

$\triangle$  schreiben, wobei  $\mathbf{A}$  wieder die Verfahrensmatrix und  $\mathbf{b}$  der Gewichtsvektor des Runge/Kutta-Verfahrens ist. Für vorgegebene  $c_i$  werden also durch  $B(s)$  die Gewichte  $b_i$ , und durch  $C(s)$  die Verfahrensmatrix  $\mathbf{A}$  festgelegt. Sind die  $c_i$  paarweise verschieden, und gilt  $b_i \neq 0$ ,  $i = 1, \dots, s$ , dann wird durch  $D(s)$  ebenfalls die Verfahrensmatrix  $\mathbf{A}$  festgelegt. Um eine höhere Ordnung zu erreichen, sind nun solche  $c_i$  gesucht, so dass von  $B(s)$  auf  $B(s+m)$  geschlossen werden kann.

### 7.3.1 Verfahren maximaler Ordnung. Gauß-Verfahren

Gesucht sind nun paarweise verschiedene Knoten  $c_i$  eines Runge/Kutta-Verfahrens mit der vereinfachenden Bedingung  $B(s)$ , so dass auf ein Verfahren der Ordnung  $p = 2s$  bestimmt werden kann. Wird ein  $s$ -stufiges Runge/Kutta-Verfahren auf ein Quadraturproblem angewandt, dann gilt für ein beliebiges  $f(x)$  vom Grad  $r < s$  die Bedingung

$$\int_0^1 f(x) dx = \sum_{i \leq s} b_i f(c_i).$$

Für  $f(x) = x^{k-1}$  gilt im exakten Fall die Bedingung

$$\frac{1}{k} x^k \Big|_{x=0}^1 = \frac{1}{k} = \sum_{i \leq s} b_i c_i^{k-1}, \quad (7.24)$$

d.h. die Gewichte  $b_i$  sind so zu bestimmen, so dass (7.24) exakt ist.

Sei  $Q(x) = f(x) M(x)$  mit  $f(x) = x^{\nu-1}$ ,  $\nu = 1, \dots, n$  und Polynom

$$M(x) := \prod_{i \leq s} (x - c_i).$$

Dann gilt im exakten Fall die Bedingung

$$0 = \langle M, v \rangle = \int_0^1 M(x) v(x) dx = \sum_{i \leq s} b_i M(c_i) v(c_i), \quad i = 1, \dots, s$$

d.h.  $M(x)$  auf dem Intervall  $[0, 1]$  orthogonal zu  $1, x, x^2, \dots, x^{s-1}$  bezüglich des Skalarprodukts  $\langle \cdot, \cdot \rangle$ . Bezeichne  $X$  den von  $\{1, x, x^s, \dots, x^s\}$  aufgespannten,  $(s+1)$ -dimensionalen Vektorraum. Dann gibt es genau einen 1-dimensionalen Untervektorraum, der zu  $\{1, x, x^s, \dots, x^{s-1}\}$  orthogonal ist. Damit ist  $M(x)$  eindeutig festgelegt als das Polynom  $P_s(x)$  aus  $\mathbb{P}_s = \text{span}\{1, x, x^2, \dots, x^s\}$ , das orthogonal auf  $\mathbb{P}_{s-1}$  steht mit  $P_s(x) = \sum_{\nu=0}^s a_\nu x^\nu$  und Hauptkoeffizient  $a_s = 1$ .

Das Polynom  $P_s(x)$  lässt sich auf mehreren Wegen bestimmen, zum Beispiel mittels Gram/Schmidt-Orthogonalisierungsverfahren. Betrachtet man zum Beispiel  $\{1, x, \dots\}$ , dann muss mit dem Gram/Schmidt-Orthogonalisierungsverfahren

$$\int_0^1 (x + \alpha \cdot 1) \cdot 1 \equiv 0$$

gelten. Lösen des Integrals liefert  $\frac{1}{2} + \alpha \equiv 0$ , und damit  $\alpha = -\frac{1}{2}$ , also sind  $1, \frac{1}{2}x - \frac{1}{2}$  orthogonal.

Man kann aber auch zeigen, dass dies verschobene *Legendre-Polynome*

$$\widehat{L}_s(x) = \frac{1}{s!} \frac{d^s}{dx^s} [x^s (x-1)^s], \quad (7.25)$$

sind.

Mittels der Gauß/Lagrange-Polynome kann ein Bezug zum CG-Verfahren hergestellt werden. △  
Orthogonalisiert man die Vektoren  $\mathbf{v}, \mathbf{A}\mathbf{v}, \mathbf{A}^2\mathbf{v}, \dots$ , so erhält man die *3-Term-Rekursion*

$$\mathbf{v}_{k+1} = \mathbf{A}\mathbf{v}_k + \alpha \mathbf{v}_k + \beta \mathbf{v}_{k-1},$$

falls  $\mathbf{A}$  symmetrisch ist. Dabei bestimmt sich der neue Vektor  $\mathbf{v}_{k+1}$  aus der Orthogonalisierung von  $\mathbf{A}\mathbf{v}_k$  mit all den Vektoren, die bereits bestimmt wurden sind. Das elegante bei der 3-Term-Rekursion ist nun, dass nur die letzten beiden Vektor  $\mathbf{v}_k, \mathbf{v}_{k-1}$  benutzt werden müssen, da diese automatisch orthogonal zu allen anderen Vektoren sind. Betrachtet man  $\mathbf{A}$  als linearen Operator in den Polynomvektorraum ist,

$$\mathbf{A} \rightarrow p(x) \mapsto xp(x) \quad \text{und} \quad \mathbf{v} \rightarrow p(x)$$

wobei  $x$  symmetrisch ist, mit

$$\int (xp(x)) q(x) dx = \int (p(x)) (x \cdot q(x)) dx$$

Wendet man dies auf die Legendre-Polynome an, dann gilt

$$\widehat{L}_{s+1}(x) = x\widehat{L}_s(x) + \alpha L_s(x) + \beta L_{s-1}(x) \quad (7.26)$$

Um eins-stufiges Gauß-Verfahren nun zu konstruieren wählt man für die Knoten  $c_i$  die paarweise verschiedenen Nullstellen des verschobenen Legendre-Polynoms und bestimmt den Gerichtsvektor  $\mathbf{b}$  aus der vereinfachenden Bedingung  $B(s)$ , die Verfahrensmatrix auf der vereinfachenden Bedingung  $C(s)$

*Beispiel 7.3.7.* Die Gauß-Verfahren mit  $s = 1, 2$  und  $p = 2s$  sind

- $s = 1, p = 2$ :

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

- $s = 2, p = 4$ :

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \hline \frac{3-\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

### 7.3.2 Radau-Verfahren

Die verschiedenen Gauß-Verfahren besitzen für die Stufenzahl  $s$  die maximale Ordnung eines Runge/Kutta-Verfahrens, allerdings sind die Stabilitätseigenschaften nicht optimal. Es lassen sich aber Verfahren der Ordnung  $p = 2s - 1$  konstruieren, die bessere Stabilitätseigenschaften besitzen, diese heißen Radau-Verfahren.

Gibt man die Knoten  $c_1, \dots, c_s$  vor, dann ist das Polynom  $M(x) = (x - c_1) \cdots (x - c_s)$  orthogonal zu  $1, x, \dots, x^{s-2}$ , dann lässt sich mittels der 3-Term-Rekursion das Polynom  $M_s(x)$  darstellen als

$$M_{\alpha, \beta}(x) = \alpha \hat{L}_s(x) + \beta \hat{L}_{s-1}(x), \quad \alpha, \beta \in \mathbb{R}$$

**Theorem 7.3.8.** *Ein Runge/Kutta-Verfahren besitze die Ordnung  $p = 2s - 1$ . Dann sind die Knoten  $c_i$  die Nullstellen eines Polynoms der Form*

$$M_{s, \xi}(x) = \hat{L}_s(x) + \xi \hat{L}_{s-1}(x), \quad \xi \in \mathbb{R} \quad (7.27)$$

ist.

*Beweis.* Ein Runge/Kutta-Verfahren der Ordnung  $p = 2s - 1$  erfüllt die Bedingung  $B(2s - 1)$ . Sei  $Q(x)$  ein beliebiges Polynom vom Grad  $\deg Q < 2s - 1$ . Dann gilt im exakten Fall

$$\int_0^1 Q(x) dx = \sum_{i \leq s} b_i Q(c_i).$$

Existiert ein Polynom  $q(x)$  der Form

$$q(x) = \prod_{i \leq s} (x - c_i),$$

nun ein beliebiges Polynom  $v(x)$  mit  $\deg v < s - 1$ , so dass  $Q(x) = q(x)v(x)$ , dann gilt

$$\langle q, v \rangle = \int_0^1 q(x)v(x) dx = \sum_{i \leq s} b_i q(c_i)v(c_i) = 0,$$

d.h.  $q(x)$  ist im Intervall  $[0, 1]$  orthogonal zu allen Polynomen  $v(x)$  mit  $\deg v < s - 1$ . Da für ein beliebiges  $\sigma \in \mathbb{N}_0$  die verschobene Legendre-Polynome  $\hat{L}_\sigma(x)$  im Intervall  $[0, 1]$  orthogonal sind, ist  $q(x)$  eine Linearkombination von  $\hat{L}_s(x)$  und  $\hat{L}_{s-1}(x)$ , also

$$q(x) = \lambda M_{s, \xi}(x), \quad \lambda \in \mathbb{R},$$

und da die Knoten  $c_i$  eine Runge/Kutta-Verfahrens die Nullstellen von  $q(x)$  sind, sind diese auch die Nullstellen von  $M_{s, \xi}(x)$  sind.  $\square$



△ Setzt man  $\xi = 1$  oder  $\xi = -1$ , dann liefern die Quadraturmethoden die Knoten

- $c_1 = 0$ : Für  $\xi = 1$  gilt  $c_1 = 0$  und das Runge/Kutta-Verfahren heißt *Radau-I-Verfahren*.
- $c_s = 1$ : Für  $\xi = -1$  gilt  $c_s = 1$  und das Runge/Kutta-Verfahren heißt *Radau-II-Verfahren*.

Mittels den Bedingung lassen sich die Knoten der Radau-Verfahren bestimmen:

**Theorem 7.3.9.** *Es gilt:*

- (1) Die Knoten  $\{c_1, \dots, c_s\} \subset [0, 1)$  eines Radau-I-Verfahrens sind  $s$  paarweise verschiedenen Nullstellen des Polynoms

$$M_{s,1}(x) = \frac{d^{s-1}}{dx^{s-1}} \left( x^s (x-)^{s-1} \right)$$

mit  $c_1 = 0$ .

- (2) Die Knoten  $\{c_1, \dots, c_s\} \subset (0, 1]$  eines Radau-II-Verfahrens sind  $s$  paarweise verschiedenen Nullstellen des Polynoms

$$M_{s,-1}(x) = \frac{d^{s-1}}{dx^{s-1}} \left( x^{s-1} (x-)^s \right)$$

mit  $c_s = 1$ .

*Beweis.* Es genügt den Fall  $\xi = 1$  zu betrachten. Dann gilt mit (7.25)

$$\begin{aligned} M_{s,1}(x) &= \frac{1}{s!} \frac{d^s}{dx^s} (x^s (x-1)^s) + \frac{1}{(s-1)!} \frac{d^{s-1}}{dx^{s-1}} \left( x^{s-1} (x-1)^{s-1} \right) \\ &= \frac{1}{(s-1)!} \frac{d^{s-1}}{dx^{s-1}} \left( \frac{1}{s} \frac{d}{dx} (x^s (x-1)^s) + x^{s-1} (x-1)^{s-1} \right) \\ &= \frac{2}{(s-1)!} \frac{d^{s-1}}{dx^{s-1}} \left( x^s (x-1)^{s-1} \right). \end{aligned}$$

Das Polynom  $f(x) = x^s (x-1)^{s-1}$  besitzt eine  $s$ -fache Nullstelle in  $x = 0$  und eine  $(s-1)$ -fache in  $x = 1$ . Die  $(s-1)$ -fache Anwendung des Satzes von Rolle liefert, dass  $f^{(s-1)}$  eine einfache Nullstelle in  $x = 0$  und  $(s-1)$  einfache Nullstellen auf  $(0, 1)$ , und da ein Polynom vom Grad  $s$  genau  $s$  Nullstellen besitzt folgt hieraus die Behauptung.  $\square$

Da die sich Knoten  $c_i, i = 1, \dots, s$ , eines Rand-Verfahrens aus den Nullstellen des entsprechenden Polynoms  $M_{s,1}(x)$  und  $M_{s,-1}$  sind, werden hierdurch die Gewichte  $\mathbf{b}$  aus der Bedingung  $B(s)$  bestimmt werden. Für die Bestimmung der Verfahrensmatrix  $\mathbf{A}$  können zwei Fälle unterschieden werden: △

- (i) **Radau-IA-Verfahren:** Im Radau-IA-Verfahren ist  $A$  durch  $D(s)$  bestimmt.
- (ii) **Radau-IIA-Verfahren:** Im Radau-IIA-Verfahren ist  $A$  durch  $C(s)$  bestimmt.

**Theorem 7.3.10.** *Die  $s$ -stufigen Radau-IA-Verfahren und Radau-IIA-Verfahren habe die Konsistenzordnung  $p = 2s - 1$ .*

*Beweis.* Nach Konstruktion der Gauß/Radau-Quadraturformel mit Polynom

$$M(x) = \frac{d^{s-1}}{dx^{s-1}} x^s (x-1)^{s-1}$$

ist dies ein Verfahren der Genauigkeit  $2s-2$ , und erfüllt damit  $B(2s-1)$ . Einerseits erfüllt ein Radau-IA-Verfahren nach Konstruktion die Bedingung  $D(s)$  und damit nach Lemma 7.2.8(2) auch die Bedingung  $C(s-1)$ , wobei dann  $a_{i1} = b_1$ . Andererseits erfüllt ein Radau-IIA-Verfahren nach Konstruktion die Bedingung  $C(S)$  und damit nach Lemma 7.2.8(1) auch  $D(s-1)$ . Dies entspricht einem Kollokationsverfahren mit  $u_{m+1}^{(s)} = u_{m+1}$  und  $a_{si} = b_i$ . Zusammen mit Satz 7.2.7 folgt daraus die Behauptung.  $\square$

*Beispiel 7.3.11 (Radau-IA-Verfahren).*

(1)  $s = 1$ : Für  $s = 1$  erhält man das Tableau

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

⚠ wobei dabei zu beachten ist, dass die Knotenbedingung nicht erfüllt ist, wobei  $C(1)$  aber nur für  $s > 1$  erfüllt sein muss.

(2)  $s = 2$ : Für  $s = 2$  erhält man das Tableau

$$\begin{array}{c|cc} 0 & \frac{1}{4} & \frac{-1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline \frac{3}{4} & \frac{1}{4} & \frac{3}{4} \\ & \frac{1}{4} & \frac{1}{4} \end{array}.$$

(3) Für  $s = 3$  erhält man das Tableau

$$\begin{array}{c|ccc} 0 & \frac{1}{9} & \frac{-1-\sqrt{6}}{18} & \frac{-1+\sqrt{6}}{18} \\ \frac{6-\sqrt{6}}{10} & \frac{1}{9} & \frac{88+7\sqrt{6}}{360} & \frac{88-43\sqrt{6}}{360} \\ \frac{6+\sqrt{6}}{10} & \frac{1}{9} & \frac{88+43\sqrt{6}}{360} & \frac{88+7\sqrt{6}}{360} \\ \hline & \frac{1}{9} & \frac{16+\sqrt{6}}{36} & \frac{16-\sqrt{6}}{36} \end{array}.$$

*Beispiel 7.3.12 (Radau-IIA-Verfahren).*

(1)  $s = 1$ : Für  $s = 1$  erhält man das Tableau

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

⚠ wobei die gerade das implizite Euler-Verfahren ist.

(2)  $s = 2$ : Für  $s = 2$  erhält man das Tableau

$$\begin{array}{c|cc} 1 & 5 & -1 \\ \frac{3}{4} & \frac{12}{3} & \frac{12}{1} \\ 1 & \frac{4}{3} & \frac{4}{1} \\ \hline & \frac{4}{3} & \frac{1}{4} \\ & \frac{4}{4} & \frac{1}{4} \end{array}.$$

(3) Für  $s = 3$  erhält man das Tableau

$4 - \sqrt{6}$	$88 - 7\sqrt{6}$	$88 + 7\sqrt{6}$	$-2 + 3\sqrt{6}$
$\frac{10}{4 + \sqrt{6}}$	$\frac{360}{269 + 169\sqrt{6}}$	$\frac{360}{269 + 169\sqrt{6}}$	$\frac{225}{2 - 3\sqrt{6}}$
$\frac{10}{1}$	$\frac{1800}{16 - \sqrt{6}}$	$\frac{1800}{16 + \sqrt{6}}$	$\frac{225}{1}$
	$\frac{36}{16 - \sqrt{6}}$	$\frac{36}{16 + \sqrt{6}}$	$\frac{9}{1}$
	$\frac{36}{36}$	$\frac{36}{36}$	$\frac{9}{9}$

### 7.3.3 Lobatto-Verfahren

Dieser letzte Methode ist ein Runge/Kutta-Verfahren der Ordnung  $2s - 2$ . Analog zum Beweis von Satz 7.3.9 lässt sich folgendes zeigen:

**Theorem 7.3.13.** Die Knoten  $c_i$  eines  $s$ -stufiges Runge/Kutta-Verfahrens der Ordnung  $p = 2s - 2$  sind die Nullstellen des Polynoms

$$\hat{P}_{s,\xi\mu} = \hat{P}_s(x) + \xi \hat{P}_{s-1} + \mu \hat{P}_{s-1}(x), \quad \xi, \mu \in \mathbb{R}.$$

Setzt man  $\xi = 0$  und  $\mu = -1$ , dann lässt sich die Konstruktion eines solchen Runge/Kutta-Verfahrens auf die Lobatto-Formeln des Quadraturproblems zurückführen, die mit  $c_0 = 0$  und  $c_1 = 1$  Polynome bis Höchstgrad  $2s - 3$  exakt integrieren. Dies führt zu den *Lobatto-III-Verfahren*: △

**Theorem 7.3.14.** Die Knoten  $c_i$  eines  $s$ -stufigen Lobatto-III-Verfahrens sind die  $s$  paarweise verschieden Nullstellen im Intervall  $[0, 1]$  des Polynoms

$$M(x) = \frac{d^{s-2}}{dx^{s-2}} x^{s-1} (x-1)^{s-1} \quad (7.28)$$

mit  $c_0 = 0$  und  $c_s = 1$ .

*Beweis.* Analog zu Satz 7.3.10.

Die Gewichte  $b_i$  werden durch die Quadraturmethode vorgegeben und die Verfahrensmatrix  $\mathbf{A}$  werden durch die vereinfachenden Bedingungen bestimmt

Lobatto-IIIA-Verfahren:  $\mathbf{A}$  wird durch  $C(s)$  bestimmt.

Lobatto-IIIB-Verfahren:  $\mathbf{A}$  wird durch  $D(s)$  bestimmt.

Lobatto-IIIC-Verfahren:  $\mathbf{A}$  wird durch  $C(s-1)$  und den zusätzlichen Bedingung  $a_{i1} = b_1$  für alle  $i = 1, \dots, s$  bestimmt.

*Beispiel 7.3.15 (Lobatto-IIIA-Verfahren).* Für  $s = 2, 3$  lauten die Butcher-Schemata des Lobatto-IIIA-Verfahrens

$0 \mid 0 \ 0$	$0 \mid 0 \ 0 \ 0$
$1 \mid \frac{1}{2} \ \frac{1}{2}$	$\frac{1}{2} \mid \frac{5}{24} \ \frac{1}{3} \ \frac{-1}{24}$
$\hline \frac{1}{2} \ \frac{1}{2}$	$\hline \frac{1}{6} \ \frac{3}{6} \ \frac{1}{6}$
	$\hline \frac{1}{6} \ \frac{3}{6} \ \frac{1}{6}$

Für den Fall  $s = 2$  geht das Lobatto-IIIA-Verfahren in die Trapezregel über.

*Beispiel 7.3.16 (Lobatto-IIIB-Verfahren).* Für  $s = 2, 3$  lauten die Butcher-Schemata des Lobatto-IIIB-Verfahrens

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

Für  $s = 2$  verstößt das Verfahren gegen die Knotenbedingung. Der Vorteil des Verfahrens ist, dass durch die Nullen in den letzten Spalten der Stufenwert  $u_{m+1}^{(s)}$  explizit berechnet werden.

*Beispiel 7.3.17 (Lobatto-IIIC-Verfahren).* Für  $s = 2, 3$  lauten die Butcher-Schemata des Lobatto-IIIA-Verfahrens

$$\begin{array}{c|c}
 0 & \frac{1}{2} & -\frac{1}{2} \\
 & \frac{1}{2} & \frac{1}{2} \\
 1 & \frac{1}{2} & -\frac{1}{2} \\
 \hline
 & \frac{1}{2} & \frac{1}{2} \\
 & \frac{1}{2} & \frac{1}{2}
 \end{array}
 \qquad
 \begin{array}{c|c}
 0 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \\
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\
 \frac{1}{2} & \frac{1}{6} & \frac{12}{12} & -\frac{1}{12} \\
 & \frac{1}{6} & \frac{1}{2} & \frac{1}{12} \\
 1 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \\
 \hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{6}
 \end{array}.$$

---

# Sachverzeichnis

## B

Butcher-Schema, 22

## D

Dichte, 28

Diskretisierungsfehler

globaler -, 51

Dissipativität, 84

## E

Einschrittverfahren, 19

gespiegeltes -, 57

symmetrisch, 59

einseitige Lipschitz-Bedingung, 84

elementares Differential, 26

elementares Gewicht, 29

### Euler-Verfahren

explizites -, 20

implizites -, 20

## F

Faà die Bruno, 29

Fehler

relativer-, 43

Fibonacci, 91

Fixpunkt, 64

## G

Gauß-Quadraturformel, 12

Gitterfunktion, 19

## I

Interpolationsquadratur, 9

## K

Knotenbedingung, 23

Konsistenz, 20

## L

Legendre-Polynom, 15

Legendre-Polynom, 11

Lipschitz-stetig, 5

## M

Modell

Lotka-Volterra, 12

## O

Ordnung, 26

## P

Populationsmodelle, 67

Prothero/Robinson-Gleichung, 86

## Q

Quadraturformel, 9

Quadraturproblem, 19

## R

Räuber/Beute-Modell, 3

Reaktionskinetik, 86

Robertson-Problem, 12

Runge/Kutta-Verfahren, 21

implizites -, 1

## S

Satz

Eindeutigkeitsss. von Picard-Lindelöf, 6

Existenzs. von Peano, 5

Stabilität, 63

steif, 85

Strömer/Verlet-Verfahren, 60

Symmetrie, 28

## T

Testproblem, 86

## V

Verfahren von

-Heun, 33

-Runge, 33

Verfahrensfunktion-, 19

## Z

Zylindermenge, 5



- Aulbach, B., 2010. Gewöhnliche Differenzialgleichungen, 2nd Edition. Spektrum Akad. Verl., München.
- Brokate, M., Henze, N., Hettlich, F., Meister, A., Schranz-Kirlinger, G., Sonar, T., 2016. Grundwissen Mathematikstudium. Berlin, Heidelberg.
- Butcher, J. C., 1963. Coefficients for the study of Runge-Kutta integration processes. Journal of the Australian Mathematical Society 3 (02), 185.
- Curtiss, C. F., Hirschfelder, J. O., 1952. Integration of Stiff Equations. Proceedings of the National Academy of Sciences 38 (3), 235–243.
- Dahlquist, G. G., 1963. A Special Stability Problem for Linear Multistep Methods. BIT: Numerical Mathematics 3, 27–43.
- Deuffhard, P., Bornemann, F. A., 2013. Numerische Mathematik 2, 4th Edition. Gewöhnliche Differentialgleichungen. de Gruyter, Berlin.  
URL <http://www.worldcat.org/title/numerische-mathematik-2-integration-gewoehnlich-oclc/311882110>
- Hairer, E., Nørsett, S. P., Wanner, G., 1993. Nonstiff problems, 2nd Edition. Vol. 8 of Springer Series in Computational Mathematics. Springer, Berlin.
- Hairer, E., Wanner, G., 2010. Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, 2nd Edition. Vol. 14 of Springer Series in Computational Mathematics. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg.  
URL <http://dx.doi.org/10.1007/978-3-642-05221-7>
- Strehmel, K., Weiner, R., Podhaisky, H., 2012. Numerik gewöhnlicher Differentialgleichungen, 2nd Edition. Vieweg+Teubner Verlag, Wiesbaden.