Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Theorie und Numerik gewöhnlicher Differentialgleichungen

by

Wolfgang Hackbusch

Lecture note no.: 22 2004

Theorie und Numerik gewöhnlicher Differentialgleichungen*

Wolfgang Hackbusch
Max-Planck-Institut *Mathematik in den Naturwissenschaften*
Inselstr. 22-26, D-04103 Leipzig, Germany
email: wh@mis.mpg.de

Zusammenfassung

Der Hauptteil dieser Vorlesung befasst sich mit der numerischen Lösung der Anfangswertaufgaben bei gewöhnlichen Differentialgleichungen. Speziellere Kapitel behandeln die asymptotischen Entwicklungen und die Schrittweitensteuerung bei Einschrittverfahren. Den steifen Differentialgleichungen ist ein eigenes Kapitel gewidmet. Die Mehrschrittverfahren erlauben bei gleichen Kosten, höhere Konsistenzordnungen zu erzielen; allerdings hat man Stabilitätsbedingungen zu erfüllen. Die Vorlesung schließt mit der Behandlung von Randwertaufgaben.

Inhaltsverzeichnis

*Als zweistündige Vorlesung gehalten an der Christian-Albrechts-Universität zu Kiel im Wintersemester 2003/04

1 Anfangswertaufgaben für gewöhnliche Differentialgleichungen

1.1 Aufgabenstellung

Situation: Gegeben sei ein endliches oder halbunendliches Intervall

$$I = [x_0, x_E] \text{ oder } I = [x_0, \infty)$$

und eine stetige Funktion

$$f = f(x, y), \qquad f : I \times \mathbb{R}^n \to \mathbb{R}^n.$$

Gesucht wird eine stetig differenzierbare Funktion $y = y(x) : I \to \mathbb{R}^n$, sodass die folgende Differentialgleichung erfüllt ist:

$$y'(x) = f(x, y(x)) \qquad \text{für alle } x \in I\,. \tag{1.1.1}$$

Notation 1.1.1 *y heißt* Lösung *der Differentialgleichung. Wir sprechen auch dann von einer Lösung y, wenn y nur auf einem Teilbereich $I' \subset I$ definiert ist (vgl. Satz 1.3.1).*

(1.1.1) heißt genauer Differentialgleichung erster Ordnung.

*Ist $n > 1$, so heißt (1.1.1) genauer Differentialgleichungs**system**. Im Falle von $n = 1$ spricht man von einer* skalaren *Differentialgleichung.*

Der Graph $\{(x, y(x)) : x \in [x_0, x_E]\}$ heißt auch Trajektorie.

Im allgemeinen gibt es unendlich viele Lösungen, für die Eindeutigkeit benötigen wir *Anfangswerte*. Das folgende Problem bezeichnet man als *Anfangswertaufgabe*:

$$y(x_0) = y_0 \in \mathbb{R}^n, \qquad y' = f(x, y) \text{ in } I. \tag{1.1.2}$$

Dabei ist $y' = f(x, y)$ die verkürzte Schreibweise für (1.1.1).

Bemerkung 1.1.2 *a) Der Definitionsbereich von f kann eingeschränkt werden, z.B.*

$$f \in C(U), \quad U \supset \{(x, y(x)) \in I \times \mathbb{R}^n : y \text{ Lösung der Differentialgleichung}\},$$

sodass $(x, y(x))$ für alle $x \in I$ bezüglich des zweiten Argumentes im Inneren von $U_x := \{(x, z) \in U\}$ liegt.

b) Falls höhere Ableitungen in der Differentialgleichung auftreten, spricht man von einer Differentialgleichung höherer Ordnung. *Eine* explizite *Differentialgleichung m-ter Ordnung für $z \in \mathbb{R}^k$ lautet*

$$z^{(m)} = F(x, z, z', \ldots, z^{(m-1)}) \tag{1.1.3a}$$

und benötigt m Anfangswertbedingungen

$$z(x_0) = z_0, \quad z'(x_0) = z_1, \ldots, \quad z^{(m-1)}(x_0) = z_{m-1}. \tag{1.1.3b}$$

c) Die Anfangswertaufgabe (1.1.3a,b) kann in eine Differentialgleichung erster Ordnung mit $n = k \cdot m$ umgeformt werden. Man setze dazu $y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$ mit $y_i := z^{(i-1)}$ für $1 \le i \le m$. Dann gilt

$$y_1' = y_2, \quad y_2' = y_3, \ldots, \quad y_m' = z^{(m)} = F(x, z, z', \ldots, z^{(m-1)}) = F(x, y_1, y_2, \ldots, y_m),$$

also $y' = f(x, y)$ mit $f(x, y) = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ y_m \\ F(x, y_1, \ldots, y_m) \end{pmatrix}$. Der Anfangswert ist $y(x_0) = y_0$ mit $y_0 = \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{m-1} \end{pmatrix}$.

Teil c) der Bemerkung zeigt, dass man o.B.d.A. die Anfangswertaufgabe (1.1.2) für Differentialgleichungen *erster* Ordnung betrachten kann. Trotzdem kann es sinnvoll sein, spezielle numerische Verfahren für (1.1.3a,b) zu entwerfen (dies wird aber nicht in der Vorlesung behandelt).

Bemerkung 1.1.3 *Sei $n = 1$. Die Funktion $f : I \times \mathbb{R} \to \mathbb{R}$ kann als "Richtungsfeld" interpretiert werden: Jedem Paar $(x, y) \in I \times \mathbb{R}$ wird eine Richtung $f(x, y)$ zugeordnet. Jede Lösung $y(x)$ ist dann eine Funktion, deren Tangente in jedem Punkt $(x, y(x))$ mit der Richtung des Feldes übereinstimmt.*

Bei dieser Interpretation können auch Werte $f(x, y) = \infty$ ohne Schwierigkeiten als senkrechte Richtungen zugelassen werden. Konsequenterweise sind dann Funktionen y durch Kurven zu ersetzen.

1.2 Verallgemeinerungen

1.2.1 Allgemeinere Randwertvorgaben

Übungsaufgabe 1.2.1 *Im Falle von (1.1.2) mit $I = [x_0, x_E]$ und Anfangswert bei x_0 wird in positiver x-Richtung gelöst. Ebenso kann der Anfangswert (eigentlich "Endwert") bei x_E vorgegeben werden ($y(x_E) = y_E$) und die Lösung in negativer x-Richtung bestimmt werden. Entsprechend kann (1.1.1) in $I = (-\infty, x_E]$ mit dem "Endwert" bei x_E formuliert werden. Man zeige, dass man die Formulierung o.B.d.A. auf die Situation von (1.1.2) beschränken kann.*

Im Falle von $n > 1$ können Randbedingungen gemischt bei x_0 und x_E formuliert werden. Das *allgemeine Randbedingung* hat die Form $R(y(x_0), y(x_E)) = 0$ mit n Gleichungen, d.h. $R : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ (vgl. §5).

Beispiel für $n = 2$: Die Differentialgleichungen $y_1' = y_1 + (x - 1)\, y_2$, $y_2' = 0$ mit der periodischen Randbedingung $y_1(0) = y_1(1) = 1$ (zwei Randbedingungen!) für y_1 und keiner Bedingung für y_2 haben die eindeutige Lösung $y_1 = e^x - (e - 1)\, x$, $y_2 = e - 1$.

1.2.2 Implizite Differentialgleichungen

Mit $F : I \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ kann die implizite Differentialgleichung

$$F(x, y, y') = 0 \tag{1.2.1}$$

formuliert werden. Wenn (1.2.1) zumindest lokal nach dem letzten Argument auflösbar ist, erhält man (1.1.1).

1.2.3 Algebrodifferentialgleichungen

Die Differentialgleichung $y' = f(x, y)$ kann gemischt mit zusätzlichen algebraischen Gleichungen auftreten.

Ein Beispiel wäre die Kombination von n Differentialgleichungen $y' = f(x, y, z)$ ($y \in \mathbb{R}^n, z \in \mathbb{R}^m$) mit m algebraischen Gleichungen $z = g(x, y)$, wobei n Anfangsbedingungen zu erfüllen sind. Die beste Umformung wäre in diesem Falle die Elimination der Unbekannten z: Die neue Differentialgleichung lautet

$$y' = F(x, y) := f(x, y, g(x, y)).$$

Eine andere mögliche Umformung sieht wie folgt aus: Ableitung von $z = g(x, y)$ nach x liefert

$$z'(x) = g_x(x, y(x)) + g_y(x, y(x))\, y'(x) = g_x(x, y(x)) + g_y(x, y(x))\, f(x, y(x), z(x)) =: G(x, y, z).$$

Man setze $Y = \binom{y}{z}$ und $F(x, Y) := \binom{f(x,y,z)}{G(x,y,z)}$. Dann stellt $Y' = F(x, Y)$ eine $(n + m)$-dimensionale Differentialgleichung dar. Die zusätzlichen m Anfangswerte ergeben sich aus $z(x_0) = g(x_0, y_0)$ und sind deshalb nicht frei wählbar.

Als weiteres Beispiel betrachten wie einen Spezialfall, in dem (1.2.1) nicht nach y' auflösbar ist. Seien stattdessen die p Gleichungen $F_i(x, y, y') = 0$ für $1 \le i \le p$ nach $(y_1, \ldots, y_p)$ auflösbar, während die Gleichungen $F_i(x, y, y') = F_i(x, y) = 0$ für $p < i \le n$ nicht von y' abhängen. Dann ist (1.2.1) in p Differentialgleichungen und $n - p$ algebraische Gleichungen getrennt worden.

1.2.4 Schwächere Glattheitsbedingungen

Bisher wurde y als stetig differenzierbar und f als stetig vorausgesetzt. Falls f z.B. bezüglich x unstetig ist, kann eine Lösung $y' = f(x, y)$ nicht stetig differenzierbar sein. Alternativen wären stückweise stetige f und stetige und stückweise stetig differenzierbare y.

Beispiel: Die stetige und stückweise stetig differenzierbare Funktion $y(x) = |x|$ erfüllt (stückweise) die Differentialgleichung $y' = f(x, y)$ mit $f(x, y) := \operatorname{sign}(x)$.

Auf die Differenzierbarkeit kann man verzichten, wenn man zur integrierten Formulierung übergeht:

$$y(x) = y_0 + \int_{x_0}^{x} f(\xi, y(\xi))\, d\xi \qquad (x_0, y_0 \text{ aus (1.1.2)}) \tag{1.2.2}$$

Lemma 1.2.2 *a) Wenn die Integralgleichung (1.2.2) eine stetige Lösung y (in der Umgebung von x_0) besitzt und f stetig ist, ist y die stetig differenzierbare Lösung der Anfangswertaufgabe (1.1.2).*

b) Umgekehrt ist jede Lösung von (1.1.2) eine Lösung der Integralgleichung (1.2.2).

c) Damit sind (1.2.2) und (1.1.2) unter der Voraussetzung, dass f stetig ist, äquivalent.

1.3 Erinnerung an die Theorie

1.3.1 Existenz

Satz 1.3.1 (Peano)[1] *Sei $f \in C(I \times Y)$ mit $I = [x_0, x_E]$, Y kompakt und y_0 liege im Inneren von Y. Dann gibt es mindestens eine Lösung der Anfangswertaufgabe (1.1.2), die in $I = [x_0, \tilde{x}_E]$, $x_0 < \tilde{x}_E \le x_E$, definiert ist. Der Graph $\{(x, y(x) : x_0 \le x \le \tilde{x}_E\}$ startet bei (x_0, y_0) und endet auf dem Rand von $(x_0, x_E] \times Y$, d.h. entweder bei $x = x_E$ oder bei $x = \tilde{x}_E < x_E$ und $y(\tilde{x}_E) \in \partial Y$.*

Bemerkung 1.3.2 *$\tilde{x}_E = x_E$ braucht nicht zu gelten, im Gegenteil: $\tilde{x}_E - x_0$ kann beliebig klein sein. Ein Beispiel hierfür ist*

$$n = 1, \quad f(x, y) = y^2/c, \quad c \in (0, 1), \quad x_0 = 0, \quad y_0 = 1.$$

Dann ist die Lösung $y = \dfrac{c}{c - x}$ nur auf $[0, c)$ definiert. Wird ein kompaktes $Y = [-R, R]$ $(R > 1)$ als Wertemenge gewählt, wird der Rand von $[0, 1] \times Y$ bei $\tilde{x}_E = c\dfrac{R-1}{R}$ erreicht.

1.3.2 Eindeutigkeit

Im allgemeinen ist die Lösung der Anfangswertaufgabe nicht eindeutig. Ein Beispiel[2] hierfür ist

$$n = 1,\ f(x, y) = \operatorname{sign}(y) \sqrt[+]{|y|}, \qquad x_0 = 0,\ y_0 = 0,$$

denn sämtliche Funktionen

$$y(x) = \left\{ \begin{array}{ll} 0 & \text{für } x \le \hat{x} \\ \pm\dfrac{(x - \hat{x})^2}{4} & \text{für } x > \hat{x} \end{array} \right\}$$

für beliebiges $\hat{x} > 0$ sind Lösungen der Anfangswertaufgabe.

Definition 1.3.3 *Eine Funktion $\varphi \in C(U)$ heißt (global) Lipschitz-stetig*[3]*, falls es eine sogenannte Lipschitz-Konstante $C \in \mathbb{R}$ gibt, sodass*

$$\|\varphi(x) - \varphi(x')\| \le C \|x - x'\| \qquad \textit{für alle } x, x' \in U.$$

Satz 1.3.4 (Picard-Lindelöf)[4] *Sei $y \in C^1(I)$ Lösung der Anfangswertaufgabe (1.1.2), $U \subset I \times \mathbb{R}^n$ mit U_x wie in Bemerkung 1.1.2a, f Lipschitz-stetig in U bezüglich des zweiten Argumentes y:*

$$\|f(x, y_1) - f(x, y_2)\| \le L \|y_1 - y_2\| \qquad \textit{für alle } (x, y_1), (x, y_2) \in U\,. \tag{1.3.1}$$

Dann ist y die einzige Lösung der Anfangswertaufgabe.

Beweis. Einsetzen von $y_0 = \tilde{y}_0$ in (1.3.2) aus dem späteren Satz 1.3.6 liefert $y(x; y_0) = y(x; \tilde{y}_0)$, damit folgt die Eindeutigkeit. ∎

Korollar 1.3.5 *Ist f stetig und Lipschitz-stetig bezüglich y und $y(\cdot)$ Lösung, dann folgt Existenz und Eindeutigkeit.*

---

[1] Giuseppe Peano, geb. 27. Aug. 1858 in Cuneo (Italien), gest. 20. April 1932 in Turin

[2] Man beachte, dass die Nichteindeutigkeit nur in einer Richtung gilt. Wird ein Wert y_E bei x_E vorgegeben und nach links gelöst, ist die Lösung eindeutig.

[3] Rudolf Otto Sigismund Lipschitz, geb. 14. Mai 1832 in Königsberg, gest. 7. Okt. 1903 in Bonn

[4] Charles Émile Picard, Schwiegersohn von Hermite, geb. 24. Juli 1856 in Paris, gest. 11. Dez. 1941 in Paris
Ernst Leonard Lindelöf, geb. 7. März 1870 in Helsingfors, gest. 4. Juni 1946 in Helsinki

1.3.3 Abhängigkeit der Lösung von Anfangswert und rechter Seite

1.3.3.1 Abhängigkeit von Anfangswert

Satz 1.3.6 *Seien $y_0, \tilde{y}_0 \in \mathbb{R}^n$ zwei Anfangswerte, $I = [x_0, x_E]$, $f \in C(U)$ mit $(x_0, y_0), (x_0, \tilde{y}_0) \in U$. Es mögen die Lösungen zu y_0 und $\tilde{y}_0$ in U existieren (Notation: $y(x; y_0)$, $y(x; \tilde{y}_0)$). In U gelte (1.3.1) mit der Konstanten L. Dann hängen die Lösungen Lipschitz-stetig vom Anfangswert ab:*[5]

$$\|y(x; y_0) - y(x; \tilde{y}_0)\| \leq e^{L|x-x_0|} \|y_0 - \tilde{y}_0\| \qquad \text{für alle } x \in I\,. \tag{1.3.2}$$

Übungsaufgabe 1.3.7 *Zusätzlich zu den Voraussetzungen des Satzes 1.3.6 gelte für $1 \leq k \leq \ell$, dass $\dfrac{\partial^k f}{\partial y^k}$ stetig ist. Dann hängt $y(x; y_0)$ ℓ-fach stetig differenzierbar vom Anfangswert y_0 ab.*

1.3.3.2 Hilfsmittel und Beweis Für den Beweis benötigen wir noch ein Resultat über Integralungleichungen.

Lemma 1.3.8 *Eine Funktion $\varphi \in C(I, \mathbb{R})$ erfülle mit*

$$\varphi_0 \geq 0, \quad L \geq 0, \quad E \geq 0 \tag{1.3.3a}$$

die Bedingung

$$\varphi(x) \leq \varphi_0 + L \int_{x_0}^{x} \varphi(\xi)\, d\xi + E(x - x_0) \qquad \text{für alle } x \in I\,. \tag{1.3.3b}$$

Dann gilt $\varphi(x) \leq \begin{cases} e^{L(x-x_0)}\varphi_0 + \frac{E}{L}\left(e^{L(x-x_0)} - 1\right) & \text{für } L > 0, \\ \varphi_0 + E(x - x_0) & \text{für } L = 0. \end{cases}$

Beweis. Zunächst zeigen wir folgende **Zwischenbehauptung:** Für alle $n \in \mathbb{N}_0$ gilt:

$$\varphi(x) \leq e^{L(x-x_0)}\varphi_0 + \frac{c}{n!}L^n(x - x_0)^n + E\sum_{k=1}^{n} \frac{L^{k-1}(x - x_0)^k}{k!} \quad \text{mit } c := \|\varphi\|_\infty\,.$$

Induktion: Für $n = 0$ gilt die Behauptung, da $\varphi(x) \leq \|\varphi\|_\infty \leq e^{L(x-x_0)}\varphi_0 + c$. Die Induktionsbehauptung für $n - 1 \geq 0$ wird auf den Integranden angewandt:

$$\begin{aligned}
\varphi(x) &\leq \varphi_0 + L\int_{x_0}^{x} \varphi(\xi)\, d\xi + E(x - x_0) \leq \varphi_0 + (x - x_0)E + \\
&\quad + L\int_{x_0}^{x} \left[e^{L(\xi-x_0)}\varphi_0 + \frac{c}{(n-1)!}L^{n-1}(\xi - x_0)^{n-1} + E\sum_{k=1}^{n-1} \frac{L^{k-1}(\xi - x_0)^k}{k!} \right] d\xi \\
&= \varphi_0 + (x - x_0)E + L\left[\frac{\varphi_0}{L}(e^{L(x-x_0)} - 1) + \frac{c}{n!}L^{n-1}(x - x_0)^n + E\sum_{k=2}^{n} \frac{L^{k-2}(x - x_0)^k}{k!} \right] \\
&\leq e^{L(x-x_0)}\varphi_0 + \frac{c}{n!}L^n(x - x_0)^n + E\sum_{k=1}^{n} \frac{L^{k-1}(x - x_0)^k}{k!},
\end{aligned}$$

wobei $L > 0$ angenommen wurde. Aber für $L = 0$ ist die Zwischenbehauptung offensichtlich.

Die Behauptung des Lemmas folgt nun für $n \to \infty$. ∎

[5] Da $x \geq x_0$ für $x \in I$, stimmen $e^{L|x-x_0|}$ und $e^{L(x-x_0)}$ überein. Die Schreibweise $|x - x_0|$ wurde gewählt um anzudeuten, dass auch bei einer Umkehr der Berechnungsrichtung ($x < x_0$, vgl. Übungsaufgabe 1.2.1) der Exponent nicht negativ werden kann.

Beweis zu Satz 1.3.6. Für $s := y_0$ oder $s := \tilde{y}_0$ mit $y' = f(x,y)$, $y(x_0) = s$ gilt

$$y(x;s) = y(x_0;s) + \int_{x_0}^{x} y'(\xi,s)\,d\xi = s + \int_{x_0}^{x} f(\xi, y(\xi;s))\,d\xi\,.$$

Man definiere $\delta(x) := y(x;y_0) - y(x;\tilde{y}_0)$ und $\varphi(x) := \|\delta(x)\|$.
Dann schätzt man

$$\delta(x) = y_0 - \tilde{y}_0 + \int_{x_0}^{x} \left[f(\xi, y(\xi;y_0)) - f(\xi, y(\xi;\tilde{y}_0))\right]\,d\xi$$

ab durch

$$\begin{aligned}
\varphi(x) &\le \|y_0 - \tilde{y}_0\| + \Big\| \int_{x_0}^{x} \left[f(\xi, y(\xi;y_0)) - f(\xi, y(\xi;\tilde{y}_0))\right]\,d\xi \Big\| \\
&\le \|y_0 - \tilde{y}_0\| + \int_{x_0}^{x} \|f(\xi, y(\xi;y_0)) - f(\xi, y(\xi;\tilde{y}_0))\|\,d\xi \\
&\le \|y_0 - \tilde{y}_0\| + \int_{x_0}^{x} L\,\|y(\xi;y_0)) - y(\xi;\tilde{y}_0))\|\,d\xi \\
&= \varphi_0 + L\int_{x_0}^{x} \varphi(\xi)\,d\xi\,,
\end{aligned}$$

wobei $\varphi_0 := \|y_0 - \tilde{y}_0\|$. Die Behauptung folgt mit Lemma 1.3.8 für $E = 0$. ∎

1.3.3.3 Numerisches Beispiel Die Stetigkeit darf nicht darüber hinwegtäuschen, dass der exponentielle Faktor $e^{L|x-x_0|}$ dafür verantwortlich sein kann, dass kleine Änderungen des Anfangswertes große Änderungen der Lösung hervorrufen können. Hierzu das folgende Beispiel: Die Lösung von

$$y' = 10\left(y - \frac{x^2}{1+x^2}\right) + \frac{2x}{(1+x^2)^2}, \qquad y(0) = y_0,$$

lautet $y(x) = y_0 e^{10x} + \frac{x^2}{1+x^2}$. Abbildung 1.3.1 zeigt die Lösungen zu $y_0 = 0$ und $y_0 = 0.0001$.

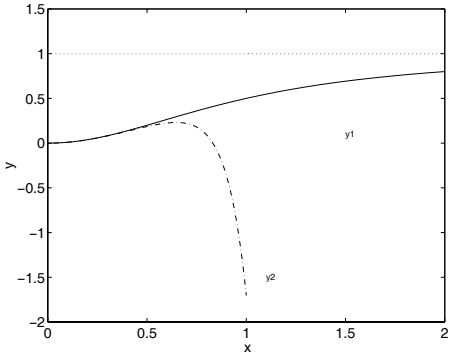


Abbildung 1.3.1: Oben: $y_1(x) = \frac{x^2}{1+x^2}$ zu $y_0 = 0$; unten: $y_2(x) = -10^{-4}e^{10x} + \frac{x^2}{1+x^2}$ zu $y_0 = -0.0001$

1.3.3.4 Abhängigkeit von f Falls die Funktion $f(\cdot,\cdot)$ nur approximativ als $\tilde{f}(\cdot,\cdot)$ berechnet werden kann oder falls aus anderen Gründen nur eine Näherung $\tilde{f}$ von f bekannt ist, stellt sich die Frage, wie die Lösung der Differentialgleichung auf eine Änderung von f reagiert. Man beachte, dass von $\tilde{f}$ die Bedingung (1.3.1) nicht verlangt wird.

Satz 1.3.9 *Seien $y, \tilde{y}$ Lösungen zu $y' = f(x,y)$ bzw. $\tilde{y}' = \tilde{f}(x,\tilde{y})$ in U mit $y(x_0) = \tilde{y}(x_0) = y_0$. f erfülle (1.3.1) und es gebe ein $\varepsilon > 0$, sodass für alle $(x,y) \in U$ die Ungleichung*

$$\|f(x,y) - \tilde{f}(x,y)\| \le \varepsilon$$

gilt. Dann existiert auch die Lösung $\tilde{y}$, und es gilt[5]

$$\|y(x) - \tilde{y}(x)\| \le \begin{cases} \frac{\varepsilon}{L}\left(e^{L|x-x_0|} - 1\right) & \text{für } L > 0, \\ \varepsilon\,|x - x_0| & \text{für } L = 0. \end{cases}$$

Beweis. Es ist

$$\begin{aligned} \varphi(x) := \|y(x) - \tilde{y}(x)\| &= \Big\| \int_{x_0}^{x} \Big[f(\xi, y(\xi)) - \tilde{f}(\xi, \tilde{y}(\xi)) \Big]\, d\xi \Big\| \\ &\le \int_{x_0}^{x} \Big[\|f(\xi, y(\xi)) - f(\xi, \tilde{y}(\xi))\| + \|f(\xi, \tilde{y}(\xi)) - \tilde{f}(\xi, \tilde{y}(\xi))\| \Big]\, d\xi \\ &\le L \int_{x_0}^{x} [\varphi(\xi) + \varepsilon]\, d\xi \le L \int_{x_0}^{x} \varphi(\xi)\, d\xi + \varepsilon(x - x_0)\,. \end{aligned}$$

Die Behauptung folgt mit $\varphi_0 = 0, E = \varepsilon$ aus Lemma 1.3.8. ∎

1.4 Lineare Differentialgleichungssysteme

Für verschiedene Klassen von Differentialgleichungen gibt es Lösungsdarstellungen, bei der aber noch bestimmte Integrale auftreten können. Eine der wichtigsten Klassen sind die linearen Differentialgleichungen $y' = A(x)y$ (d.h. $f(x,y) = A(x)y$ mit einer $n \times n$-Matrixfunktion $A(x)$).

1.4.1 Lineare Differentialgleichungssysteme mit konstanten Koeffizienten

Für lineare Differentialgleichungen mit konstanten Koeffizienten A ergeben sich vollkommen explizite Lösungsdarstellungen.

Satz 1.4.1 *Die Anfangswertaufgabe $y' = Ay$, $y(x_0) = y_0$ hat eine eindeutige Lösung $y(x) = e^{A(x-x_0)}y_0$, wobei für eine Matrix X die Exponentialfunktion durch*

$$e^X := \sum_{\nu=0}^{\infty} \frac{1}{\nu!} X^\nu$$

definiert ist.

Falls A diagonalisierbar ist, d.h. $A = TDT^{-1}$, $D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$, so kann die Lösung in der Form

$$y(x) = \sum_{k=1}^{n} y_0^{(k)} e^{\lambda_k (x-x_0)} x^{(k)}$$

dargestellt werden, wobei $x^{(k)}$ die Eigenvektoren (d.h. $Ax^{(k)} = \lambda_k x^{(k)}$) und $y_0^{(k)}$ die Koeffizienten der Basisdarstellung $y(x_0) = \sum_{k=1}^{n} y_0^{(k)} x^{(k)}$ sind.

Die Aussage des Satzes 1.4.1 ist ein Spezialfall des nachfolgenden Satzes 1.4.3.

Übungsaufgabe 1.4.2 *Für eine beliebige $n \times n$-Matrix X gilt: a) Die e^X definierende Reihe konvergiert. b) e^X ist regulär.*

1.4.2 Lineare Differentialgleichungssysteme mit nichtkonstanten Koeffizienten

Satz 1.4.3 *Die Anfangswertaufgabe* $y' = A(x)y$, $y(x_0) = y_0$ *hat die Lösung* $y(x) = T(x, x_0)y_0$, *wobei* $T(x, x_0)$ *die sogenannte* Fundamentalmatrix *ist. Sie erfüllt das matrixwertige Anfangswertproblem*

$$\frac{d}{dx}T(x, x_0) = A(x)T(x, x_0), \quad T(x_0, x_0) = I\,.$$

Für alle x *ist* $T(x, x_0)$ *regulär. Falls die Matrizen* $A(x), A(x')$ *für alle* $x, x' \in I$ *vertauschbar sind, gilt*

$$T(x, x_0) := \exp(\int_{x_0}^{x} A(\xi)\,d\xi)$$

(dieser Fall liegt insbesondere für $n = 1$ *vor).*

Beweis. a) Für $y(x) = T(x, x_0)y_0$ rechnet man nach, dass

$$y'(x) = \frac{d}{dx}T(x, x_0)y_0 = A(x)T(x, x_0)y_0 = A(x)y(x)$$

und $y(x_0) = T(x_0, x_0)y_0 = y_0$.

b) Die matrixwerte Funktion $B(x)$ sei mit ihrer Ableitung $B'(x)$ vertauschbar. Dann gilt

$$\frac{d}{dx}\left(B(x)^\nu\right) = \nu B'(x)B(x)^{\nu-1}.$$

(Was ist ohne Vertauschbarkeit das Resultat?)

c) An Hand der Potenzreihe $\exp(B(x)) = \sum_{\nu=0}^{\infty} \frac{1}{\nu!}B(x)^\nu$ erhält man unter der Voraussetzung von b), dass $\frac{d}{dx}\exp(B(x)) = B'(x)\exp(B(x))$.

d) Für $B(x) = \int_{x_0}^{x} A(\xi)\,d\xi$ gilt $B'(x) = A(x)$. Falls die Matrizen $A(x), A(x')$ für alle $x, x' \in I$ vertauschbar sind, ist die Voraussetzung von b) erfüllt. Nach Teil c) folgt

$$\frac{d}{dx}T(x, x_0) = \frac{d}{dx}\exp(\int_{x_0}^{x} A(\xi)\,d\xi) = A(x)\exp(\int_{x_0}^{x} A(\xi)\,d\xi) = A(x)T(x, x_0).$$

e) Schließlich ist $T(x_0, x_0) = \exp(\int_{x_0}^{x_0} A(\xi)\,d\xi) = \exp(O) = I$. ∎

2 Einschrittverfahren

Bevor wir allgemeine Einschrittverfahren beschreiben, wird das Euler-Verfahren[6] als Prototyp eines Einschrittverfahrens behandelt.

2.1 Euler-Verfahren

2.1.1 Notationen, etc.

Gegeben sei die Anfangswertaufgabe

$$y' = f(x, y), \quad y(x_0) = y_0.$$

Dann erhält man für eine *Schrittweite* $h > 0$ die *Stützstellen* $x_\nu = x_0 + h\nu$ für $\nu \in \mathbb{N}$, wobei für endliches I die Indizes ν nach oben beschränkt sind. (Später werden wir die Schrittweite in jedem Schritt variabel halten. Die konstante Wahl h dient nur der Vereinfachung).

Aus der Integralgleichungsformulierung (1.2.2) erhält man die Näherung

$$y(x_1) = y(x_0 + h) = y_0 + \int_{x_0}^{x_1} y'(\xi)\, d\xi = y_0 + \int_{x_0}^{x_1} f(\xi, y(\xi))\, d\xi \approx y_0 + hf(x_0, y_0),$$

indem man das Integral durch eine Rechtecksformel (konstante Interpolation des Integranden) mit der Stützstelle x_0 approximiert. Wenn man $x_0, x_1 = x_0 + h$ durch $x_\nu, x_{\nu+1} = x_\nu + h$ ersetzt, lautet das Resultat

$$y_{\nu+1} = y_\nu + hf(x_\nu, y_\nu), \tag{2.1.1}$$

das zusammen mit dem Anfangswert y_0 eine Rekursionsformel für die y_ν darstellt. Der Rechenaufwand beträgt etwa

$$\text{Häufigkeit der Auswertung von } f \times \text{Rechenaufwand der Auswertung von } f.$$

Für das Euler-Verfahren auf $I = [x_0, x_E]$ ist dies $\frac{x_E - x_0}{h} \times$Rechenaufwand der Auswertung von f.

Eine andere (ältere) Bezeichnung für das Euler-Verfahren lautet *Polygonzugverfahren*. Dabei liegt die Vorstellung zugrunde, dass man die Punkte (x_ν, y_ν) (y_ν: Lösung von (2.1.1)) durch Streckenzüge verbindet und so den Graph einer stückweise linearen Funktion erhält.

Notation 2.1.1 *Wir definieren* $\eta_\nu := \eta(x_\nu, h) := y_\nu$. ***Achtung:*** $\eta(x, h)$ *ist nur für die Stützstellen* $x \in \{x_0 + \nu h \in I : \nu \in \mathbb{N}_0\}$ *definiert. Für* η *gilt die analoge Rekursionsformel*

$$\eta(x_0, h) = y_0, \quad \eta(x_{\nu+1}, h) = \eta(x_\nu, h) + hf(x_\nu, \eta(x_\nu, h)), \tag{2.1.1'}$$

oder in kürzerer Schreibweise:

$$\eta_0 = y_0, \quad \eta_{\nu+1} = \eta_\nu + hf(x_\nu, \eta_\nu). \tag{2.1.1"}$$

Zur Formel (2.1.1") seien zwei Bemerkungen gemacht:
a) $\eta_{\nu+1}$ hängt nur von nur von η_ν (d.h. nicht von der "älteren Vergangenheit" $\eta_{\nu-1}, \eta_{\nu-2}, \ldots$). Dies ist das Kennzeichen der *Einschrittverfahren*.
b) Der neue (unbekannte) Wert $\eta_{\nu+1}$ kann direkt aus (2.1.1") berechnet werden. Benötigt wird nur die Funktionsauswertung von f. Deshalb nennt man Gleichung (2.1.1") ein *explizites* Verfahren. (Im impliziten Fall wäre noch eine nichtlineare Gleichung für $\eta_{\nu+1}$ zu lösen).

2.1.2 Fehleranalyse

Im folgenden wird die y_0-Abhängigkeit von $\eta(x, h)$ untersucht.

Übungsaufgabe 2.1.2 *Die Ungleichung* $1 + x \le e^x$ *gilt für alle* $x \in \mathbb{R}$.

[6] Leonhard Euler, geb. 15. April 1707 in Basel, gest. 18. Sept. 1783 in St. Petersburg

Satz 2.1.3 *Seien $\eta(x; y_0; h)$ und $\eta(x; \tilde{y}_0; h)$ die Euler-Resultate zu den beiden Anfangswerten y_0 und $\tilde{y}_0$, f sei Lipschitz-stetig bezüglich y mit Lipschitz-Konstante L. Dann gilt*[5]

$$\|\eta(x; y_0; h) - \eta(x; \tilde{y}_0; h)\| \leq e^{L|x-x_0|} \|y_0 - \tilde{y}_0\| . \tag{2.1.2}$$

Beweis. Definiere $\eta_\nu := \eta(x_\nu; y_0; h)$, $\tilde{\eta}_\nu := \eta(x_\nu; \tilde{y}_0; h)$. Dann gilt

$$\begin{aligned}
\|\eta_{\nu+1} - \tilde{\eta}_{\nu+1}\| &= \|(\eta_\nu + hf(x_\nu, \eta_\nu)) - (\tilde{\eta}_\nu + hf(x_\nu, \tilde{\eta}_\nu))\| \\
&\leq \|\eta_\nu - \tilde{\eta}_\nu\| + h \|f(x_\nu, \eta_\nu) - f(x_\nu, \tilde{\eta}_\nu)\| \\
&\leq \|\eta_\nu - \tilde{\eta}_\nu\| + hL \|\eta_\nu - \tilde{\eta}_\nu\| = (1 + hL) \|\eta_\nu - \tilde{\eta}_\nu\| \\
&\overset{\text{Übungsaufgabe 2.1.2}}{\leq} e^{hL} \|\eta_\nu - \tilde{\eta}_\nu\| .
\end{aligned}$$

Die rekursive Anwendung dieser Ungleichung liefert $\|\eta_\nu - \tilde{\eta}_\nu\| \leq e^{\nu hL} \|\eta_0 - \tilde{\eta}_0\|$. Die Behauptung folgt mit $\nu h = x_\nu - x_0$, $\eta_0 = y_0$ und $\tilde{\eta}_0 = \tilde{y}_0$. ∎

Lemma 2.1.4 *Sei f Lipschitz-stetig in x und y, y sei (exakte) Lösung der Differentialgleichung. Dann gilt die Taylor-Entwicklung*[7]

$$\begin{aligned}
y(x_{\nu+1}) = y(x_\nu + h) &= y(x_\nu) + hy'(x_\nu) + h^2 R(x_\nu, x_{\nu+1}) \\
&= y(x_\nu) + hf(x_\nu, y(x_\nu)) + h^2 R(x_\nu, x_{\nu+1})
\end{aligned} \tag{2.1.3}$$

mit einer Restgliedabschätzung $\|R\| \leq const$ (an dieser Stelle geht die Lipschitz-Stetigkeit von f bezüglich beider Argumente ein).

Beweis. Die Taylor-Entwicklung lautet $y(x_{\nu+1}) = y(x_\nu) + hy'(x_\nu) + \frac{1}{2}h^2 y''(x_\nu + \vartheta h)$, falls y'' existiert, und allgemeiner $y(x_{\nu+1}) = y(x_\nu) + hy'(x_\nu) + \frac{1}{2}h^2 L_{y'}$ mit der Lipschitz-Konstanten von y', falls y' Lipschitz-stetig ist. Da $y' = f(x, y)$, ist letztere Bedingung erfüllt. Also ist R beschränkt durch die Lipschitz-Konstante $L_{y'} = L_f$ (Lipschitz-Konstante von f) in $[x_\nu, x_{\nu+1}]$. ∎

Satz 2.1.5 *Sei f Lipschitz-stetig in x und y. Dann konvergiert das Euler-Verfahren für $h \to 0$ gegen die exakte Lösung*[5]*:*

$$\|\eta(x, h) - y(x)\| \leq hK \frac{1}{L} (e^{L|x-x_0|} - 1). \tag{2.1.4}$$

Dabei ist L die Lipschitz-Konstante aus (1.3.1) und $K = L_f$ die Lipschitz-Konstante von f bezüglich beider Argumente. Da auf der rechten Seite die erste Potenz $h = h^1$ steht, sagt man, dass das Verfahren von erster Ordnung *konvergiert.*

Beweis. Definiere $\delta(x) := \eta(x, h) - y(x)$. Dann lässt sich

$$\begin{aligned}
\delta(x_{\nu+1}) &= \eta(x_{\nu+1}, h) - y(x_{\nu+1}) \\
&\overset{(2.1.3)}{=} [\eta(x_\nu) + hf(x_\nu, \eta(x_\nu, h))] - \big[y(x_\nu) + hf(x_\nu, y(x_\nu)) + h^2 R(x_\nu, x_{\nu+1})\big] \\
&= \eta(x_\nu, h) - y(x_\nu) + h\,(f(x_\nu, \eta(x_\nu, h)) - f(x_\nu, y(x_\nu))) + h^2 R(x_\nu, x_{\nu+1})
\end{aligned}$$

mittels

$$\begin{aligned}
\|\delta(x_{\nu+1})\| &\leq \|\eta(x_\nu, h) - y(x_\nu)\| + h \|f(x_\nu, \eta(x_\nu, h)) - f(x_\nu, y(x_\nu))\| + h^2 \|R\| \\
&\leq \|\delta(x_\nu)\| + hL \|\delta(x_\nu)\| + h^2 \|R\| \\
&= (1 + hL) \|\delta(x_\nu)\| + h^2 L_f
\end{aligned}$$

abschätzen. Die Behauptung folgt aus Lemma 2.1.6 mit $a_\nu := \|\delta(x_\nu)\|$ mit $a_0 = 0$, $B := L_f$, $k = 2$. ∎

[7] Brook Taylor, geb. 18. Aug. 1685 in Edmonton, Middlesex, England, gest. 29. Dez. 1731 in Somerset House, London

Lemma 2.1.6 *Seien Konstanten $L, B, h, k, a_0 \geq 0$ gegeben. Wenn die Größen a_ν $(\nu \geq 1)$ die Ungleichung*

$$a_{\nu+1} \leq (1+hL)a_\nu + h^k B \qquad \text{für alle } \nu \geq 0 \tag{2.1.5}$$

erfüllen, so gilt

$$a_\nu \leq e^{\nu hL} a_0 + h^{k-1} B \left\{ \begin{array}{ll} \nu h & \text{für } L = 0, \\ \frac{e^{\nu hL}-1}{L} & \text{für } L > 0. \end{array} \right\}.$$

Beweis. a) Per Induktion zeigen wir die folgende Zwischenbehauptung:

$$a_\nu \leq A_\nu := \sum_{\mu=0}^{\nu-1} (1+hL)^\mu h^k B + (1+hL)^\nu a_0. \tag{2.1.6}$$

Für $\nu = 0$ lautet die rechtet Seite $A_0 := a_0$, sodass $a_0 \leq a_0$ zutrifft. Die Behauptung gelte für ν. Dann findet man für $\nu + 1$, dass

$$\begin{aligned} a_{\nu+1} &\underset{(2.1.5)}{\leq} (1+hL)a_\nu + h^k B \underset{(2.1.6)}{\leq} (1+hL)A_\nu \\ &\leq \underbrace{\sum_{\mu=0}^{\nu-1} (1+hL)^{\mu+1} h^k B + h^k B}_{=\sum_{\mu=0}^{\nu}(1+hL)^\mu h^k B} + (1+hL)^{\nu+1} a_0 = A_{\nu+1}. \end{aligned}$$

b) Mit Übungsaufgabe 2.1.2 gilt $(1+hL)^\nu a_0 \leq e^{\nu hL} a_0$. Für die Summe $h^k B \sum_{\mu=0}^{\nu-1} (1+hL)^\mu$ erhält man $h^k B\nu$ im trivialen Fall $L = 0$. Sonst folgt für die geometrische Summe

$$h^k B \sum_{\mu=0}^{\nu-1} (1+hL)^\mu = h^k B \frac{(1+hL)^\nu - 1}{(1+hL) - 1} = h^k B \frac{(1+hL)^\nu - 1}{hL} \underset{\text{Übung 2.1.2}}{\leq} h^{k-1} B \frac{e^{\nu hL} - 1}{L}.$$

■

2.1.3 Numerisches Beispiel

Für $n = 1$ hat das Anfangswertproblem

$$y'(x) = f(x,y) := y(x) \text{ für } x \in I := [0,2], \qquad y(0) = y_0 := 1 \tag{2.1.7}$$

die Lösung $y(x) = e^x$. Die folgende Tabelle zeigt die Lösung $\eta(2,h)$ des Euler-Verfahrens im Endpunkt $x = 2$ für verschiedene Schrittweiten h.

Schrittweite h	Lösung $\eta(2,h)$	absoluter Fehler $\lvert\eta(2,h) - e^2\rvert$	relativer Fehler $\frac{\lvert\eta(2,h)-e^2\rvert}{e^2}$
$2/5 = 0.4000$	5.37824_{10}+00	2.01082_{10}+00	3.73880_{10}-01
$2/10 = 0.2000$	6.19174_{10}+00	1.19732_{10}+00	1.93374_{10}-01
$2/20 = 0.1000$	6.72750_{10}+00	6.61556_{10}-01	9.83361_{10}-02
$2/40 = 0.0500$	7.03999_{10}+00	3.49067_{10}-01	4.95835_{10}-02
$2/80 = 0.0250$	7.20957_{10}+00	1.79488_{10}-01	2.48958_{10}-02
$2/160 = 0.0125$	7.29802_{10}+00	9.10352_{10}-02	1.24740_{10}-02

Die Fehler bestätigen das Resultat (2.1.4): Bei Halbierung der Schrittweite halbiert sich in etwa auch der Fehler, d.h. das Euler-Verfahren ist von erster Ordnung genau. Umgekehrt zeigt sich der Nachteil eines Verfahrens erster Ordnung: Eine hohe Genauigkeit erreicht man nur mit sehr kleiner Schrittweite und entsprechend großem Aufwand.

2.2 Allgemeine Einschrittverfahren, Konsistenz

2.2.1 Notationen, etc.

Die allgemeine Form *expliziter Einschrittverfahren* ist

$$\eta_{i+1} = \eta_i + h\phi(x_i, \eta_i; h, f)\,, \tag{2.2.1a}$$

wobei $x_i = x_0 + ih$, $\eta_i = \eta(x_i, h)$. ϕ wird die *Inkrementfunktion* genannt. Das Argument f von ϕ bedeutet, dass die Auswertung von ϕ die Auswertung von f an einer geeigneten Stelle verwenden darf (bei der Implementierung von ϕ ist hier f als Funktion zu übergeben). Für das Euler-Verfahren ist beispielsweise $\phi(x_i, \eta_i; h, f) := f(x_i, \eta_i)$.

Implizite Einschrittverfahren lauten

$$\eta_{i+1} = \eta_i + h\phi(x_i, \eta_i, \eta_{i+1}; h, f)\,. \tag{2.2.1b}$$

(2.2.1b) ist im allgemeinen ein nichtlineares Gleichungssystem für $\eta_{i+1} \in \mathbb{R}^n$. Natürlich kann (2.2.1a) als Spezialfall von (2.2.1b) aufgefasst werden. Die naheliegende Frage, ob das nichtlineares Gleichungssystem (2.2.1b) überhaupt nach η_{i+1} auflösbar ist, kann zumindest für $h \to 0$ positiv beantwortet werden.

Satz 2.2.1 *Die Lipschitz-Stetigkeit* $\|\phi(x, y, \eta; h, f) - \phi(x, y, \tilde\eta; h, f)\| \le L\,\|\eta - \tilde\eta\|$ *wird vorausgesetzt*[8]*. Dann ist (2.2.1b) für hinreichend kleine h (hinreichend ist* $h < \frac{1}{L}$*) eindeutig lösbar.*

Beweis. Für feste η_i, x_i, h definiere $\varphi(\eta) := \eta_i + h\phi(x_i, \eta_i, \eta; h, f)$. η_{i+1} ist genau dann eine Lösung von (2.2.1b), wenn es die Fixpunktgleichung

$$\eta_{i+1} = \varphi(\eta_{i+1})$$

erfüllt. φ ist Lipschitz-stetig mit der Lipschitz-Konstanten $L_\varphi = hL$:

$$\begin{aligned}\|\varphi(\tilde\eta) - \varphi(\eta)\| &= \|[\eta_i + h\phi(x_i, \eta_i, \tilde\eta; h, f)] - [\eta_i + h\phi(x_i, \eta_i, \eta; h, f)]\| \\ &= h\,\|\phi(x_i, \eta_i, \tilde\eta; h, f) - \phi(x_i, \eta_i, \eta; h, f)\| \le hL\,\|\tilde\eta - \eta\|\,.\end{aligned}$$

Für $h < \frac{1}{L}$ ist $L_\varphi < 1$. Die Behauptung folgt damit aus dem Banachschen Fixpunktsatz. ∎

Der Banachsche[9] Fixpunktsatz (vgl. [12, Satz 4.2.1]) zeigt darüberhinaus, dass die Fixpunktiteration zum Ziel führt.

Bemerkung 2.2.2 *a) (2.2.1b) ist durch folgende Fixpunktiteration lösbar:*

$$\eta_{i+1}^{(j+1)} := \eta_i + h\phi(x_i, \eta_i, \eta_{i+1}^{(j)}; h, f)\,. \tag{2.2.2}$$

Hierbei kann $\eta_{i+1}^{(0)}$ *beliebig gewählt werden. Die natürliche Wahl ist* $\eta_{i+1}^{(0)} = \eta_i$*.*

b) Da der Fehler von $\eta_{i+1}^{(j+1)}$ *der Abschätzung* $\|\eta_{i+1}^{(j+1)} - \eta_{i+1}\| \le hL\|\eta_{i+1}^{(j)} - \eta_{i+1}\|$ *genügt (d.h.* $\eta_{i+1}^{(j)} = \eta_{i+1} + \mathcal{O}(h^j\|\eta_{i+1}^{(0)} - \eta_{i+1}\|)$*) und für den Fehler von* $\eta(\cdot, h)$ *aus (2.2.1b) die Größenordnung* $\mathcal{O}(h^k)$ *für ein k erwartet werden muss (z.B.* $k = 1$ *im Falle von (2.1.4)), reicht eine feste Anzahl von Iterationsschritten (2.2.2).*

Die Kombination von Startwert und Iteration liefert unter Berücksichtigung der Bemerkung 2.2.2b die sogenannten *Prädiktor-Korrektor-Formeln:*

1. Mit einem *expliziten Prädiktor-Verfahren* $\phi_{\text{Präd}}$ berechne man den Startwert

$$\eta_{i+1}^{(0)} := \eta_i + h\phi_{\text{Präd}}(x_i, y_i; h, f).$$

2. Man führe einen oder mehrere Iterationsschritte (2.2.2) mit dem impliziten *Korrektor-Verfahren* $\phi_{\text{Korr}} := \phi$ durch.

[8] Im Allgemeinen wird die Lipschitz-Stetigkeit von ϕ nur erreichbar sein, wenn auch die Lipschitz-Stetigkeit (1.3.1) von f gilt.

[9] Stefan Banach, geb. 30. März 1892 in Krakau, gest. 31. Aug. 1945 in Lemberg (Lwow)

Zur Illustration geben wir die folgenden Beispiele für Einschrittverfahren an:

$$\phi(x,y;h,f) = f(x,y) \qquad \text{(Euler-Verfahren)} \tag{2.2.3a}$$

$$\phi(x,y_1,y_2;h,f) = f(x+h,y_2) \qquad \text{(implizites Euler-Verfahren)} \tag{2.2.3b}$$

$$\phi(x,y_1,y_2;h,f) = \frac{1}{2}(f(x,y_1)+f(x+h,y_2)) \qquad \text{(Trapezformel)} \tag{2.2.3c}$$

$$\phi(x,y;h,f) = f(x+\frac{h}{2},y+\frac{h}{2}f(x,y)) \qquad \text{(modifiziertes Euler-Verfahren)} \tag{2.2.3d}$$

$$\phi(x,y;h,f) = \frac{1}{2}\Big(f(x,y)+f(x+h,y+hf(x,y))\Big) \qquad \text{(Verfahren von Heun}^{10}\text{)} \tag{2.2.3e}$$

Bemerkung 2.2.3 *a) (2.2.3e) ist interpretierbar als Prädiktor-Korrektor-Formel, wobei (2.2.3a) als Prädiktor und (2.2.3c) als Korrektor benutzt wird.*

b) Jede Prädiktor-Korrektor-Kombination mit fester Iterationszahl lässt sich als neues explizites *Einschrittverfahren auffassen.*

2.2.2 Konsistenz

Definition 2.2.4 *Sei $(x,y) \in \mathbb{R}^{n+1}$ fest und $z(\cdot)$ die Lösung von $z(x)=y$, $z'(t)=f(t,z(t))$ für $t \geq x$. Dann heißt*

$$\tau(x,y,h) = \begin{cases} \frac{z(x+h)-y}{h} - \phi(x,y;h,f) & \textit{(expliziter Fall)} \\ \frac{z(x+h)-y}{h} - \phi(x,y,z(x+h);h,f) & \textit{(impliziter Fall)} \end{cases} \tag{2.2.4}$$

der lokale Diskretisierungsfehler *bei (x,y).*

Im expliziten Fall ist $z(x+h) = y+h\phi(x,y;h,f)+h\tau(x,y,h)$ die *exakte* Lösung der Anfangswertaufgabe in $x+h$ (mit Anfangswert in x), während $\eta(x+h,h) := y+h\phi(x,y;h,f)$ der Wert des Einschrittverfahrens ist. Der in *einem* Lösungsschritt gemachte Fehler beträgt daher gerade $h\tau(x,y,h)$. Analoges gilt im impliziten Fall.

Offensichtlich ist es sinnvoll zu fordern, dass τ möglichst klein ist. Mindestens muss $\lim_{h\to 0}\tau(x,y,h) \to 0$ gelten. Wegen

$$\lim_{h\to 0}\frac{z(x+h)-y}{h} = \lim_{h\to 0}\frac{z(x+h)-z(x)}{h} = z'(x) = f(x,y)$$

impliziert $\lim_{h\to 0}\tau(x,y,h)\to 0$, dass

$$\lim_{h\to 0}\phi(x,y;h,f) = f(x,y). \tag{2.2.5}$$

Definition 2.2.5 *ϕ ist konsistent zur Differentialgleichung $y'=f(x,y)$, falls (2.2.5) für $f \in C^0(U)$ gilt.*

Definition 2.2.6 *Sei $f \in C^p(U)$ (es reicht auch $f \in C^{p-1}(U)$ und f^{p-1} Lipschitz-stetig). Dann beschreibt ϕ ein Verfahren der Konsistenzordnung p, falls $\tau(x,y,h) = \mathcal{O}(h^p)$ ist (für $(x,y) \in U$).*

Wie man vermuten kann und später auch gezeigt wird, wird das Verfahren umso genauere Resultate liefern, je höher die Konsistenzordnung ist. Es ist daher Ziel, eine möglichst hohe Konsistenzordnung zu erreichen (für vernünftige "Kosten"). Die Ordnungen der Beispiele (2.2.3a-e) sind, wie man mit Taylor-Entwicklung nachrechnen kann:

(2.2.3a)	(2.2.3b)	(2.2.3c)	(2.2.3d)	(2.2.3e)
1	1	2	2	2

Wie kann man also ein Verfahren höherer Ordnung konstruieren?

1) Die Taylor-Entwicklung $z(x+h) = z(x)+h\sum_{k=1}^{p}\frac{1}{k!}z^{(k)}(x)h^{k-1}+\mathcal{O}(h^{p+1})$ legt nahe,

$$\phi := \sum_{k=1}^{p}\frac{1}{k!}z^{(k)}(x)h^{k-1}$$

zu verwenden. Hierzu müssen die Ableitungen $z^{(k)}$ geeignet dargestellt werden. Für $k = 1$ ist

$$z'(x) = f(x, z(x)) \tag{2.2.6a}$$

und daher einfach mit f ausdrückbar. Für $k = 2$ findet man

$$z''(x) = \frac{d}{dx} f(x, z(x)) = f_x(x, z(x)) + f_z(x, z(x))z' = f_x + f_z f, \tag{2.2.6b}$$

d.h. für die Implementierung dieses Termes benötigt man die ersten partiellen Ableitungen von f. Entsprechend findet man (komplizierter werdende) Formeln für $z^{(k)}$, die nur f und seine Ableitungen bis zur Ordnung $k-1$ erfordern. Wenn man $\phi(x, y; h, f)$ so interpretiert, dass mit f auch seine Ableitungen verfügbar sind, ergibt sich ein Einschrittverfahren der Ordnung p, da $z(x + h) = y + h\phi(x, y; h, f) + h\tau(x, y, h)$ mit $\tau(x, y, h) = \mathcal{O}(h^p)$.

2) Einen anderen Zugang bietet die Integralformulierung (1.2.2). Sei I eine Quadraturformel der Ordnung $p-1$ (d.h. I ist exakt für alle Polynome vom Grad $\le p-1$). Die Darstellung von I sei $I(\varphi) := h \sum_\nu \gamma_\nu \varphi(x^{(\nu)})$ mit $x \le x^{(0)} < x^{(1)} < \ldots \le x + h$. Dann ist

$$z(x + h) = z(x) + \int_x^{x+h} f(t, z(t))\, dt \approx z(x) + I(f(\cdot, z(\cdot))) = y + h \sum_\nu \gamma_\nu f(x^{(\nu)}, z(x^{(\nu)})) + \mathcal{O}(h^{p+1}).$$

Nun bleibt das nicht-triviale Problem, $z(x^{(\nu)})$ bis auf $\mathcal{O}(h^p)$ durch Näherungen zu ersetzen, die neben f (und seinen Ableitungen) nur $z(x)$ (und impliziten Fall zusätzlich $z(x + h)$) enthalten.

3) Einfacher ist ein Ansatz von ϕ mit freien Parametern, zum Beispiel

$$\phi(x, y; h, f) = \alpha f(x, y) + \beta f(x + \gamma h, y + \delta h f(x, y))$$

mit vier freien Parametern und nur zwei f-Auswertungen. Man entwickelt ϕ in eine Taylor-Reihe:

$$\phi(x, y; h, f) = [\alpha + \beta]\, f(x, y) + h\beta\, [\gamma f_x(x, y) + \delta f_y(x, y) f(x, y)] + \mathcal{O}(h^2).$$

Es soll $\tau(x, y, h) = \mathcal{O}(h^p)$ (in diesem Fall mit $p = 2$) gelten, was

$$\begin{aligned} \phi(x, y; h, f) = \frac{z(x + h) - y}{h} + \mathcal{O}(h^2) &= \sum_{k=1}^{2} \frac{1}{k!} z^{(k)}(x) h^{k-1} + \mathcal{O}(h^2) \\ &\underset{(2.2.6a,b)}{=} f(x, z(x)) + \frac{h}{2} [f_x + f_z f] + \mathcal{O}(h^2) \end{aligned}$$

impliziert. Koeffizientenvergleich liefert drei nichtlineare Gleichungen für die vier Parameter $\alpha, \beta, \gamma, \delta$:

$$\alpha + \beta = 1, \quad \beta\gamma = \frac{1}{2}, \quad \beta\delta = \frac{1}{2}.$$

Spezielle Lösungen sind

1. $\alpha = \beta = \frac{1}{2}$, $\gamma = \delta = 1$ (dies reproduziert das Einschrittverfahren (2.2.3e))
2. $\alpha = 0$, $\beta = 1$, $\gamma = \delta = \frac{1}{2}$ (ergibt (2.2.3d))

4) Ein weiterer Zugang mittels Runge-Kutta-artigen Verfahren wird in §2.3 beschrieben.

2.2.3 Numerisches Beispiel

Für das Anfangswertproblem (2.1.7) liefert die Trapezregel (2.2.3c) die folgenden Resultate:

Schrittweite h	Lösung $\eta(2,h)$	absoluter Fehler $\lvert\eta(2,h)-e^2\rvert$	relativer Fehler $\frac{\lvert\eta(2,h)-e^2\rvert}{e^2}$
$2/5 = 0.4000$	7.59375_{10}+00	2.04694_{10}-01	2.69556_{10}-02
$2/10 = 0.2000$	7.43878_{10}+00	4.97246_{10}-02	6.68451_{10}-03
$2/20 = 0.1000$	7.40140_{10}+00	1.23439_{10}-02	1.66778_{10}-03
$2/40 = 0.0500$	7.39214_{10}+00	3.08057_{10}-03	4.16736_{10}-04
$2/80 = 0.0250$	7.38983_{10}+00	7.69806_{10}-04	1.04171_{10}-04

Der Abfall der Fehler entspricht dem Faktor 4 und zeigt damit an, dass das Verfahren von zweiter Ordnung genau ist. Aufgrund dieser Tatsache erhält man Fehler, die deutlich kleiner als für das Euler-Verfahren in §2.1.3 sind.

2.3 Runge-Kutta-Verfahren

Das klassische Verfahren von Runge[11] und Kutta[12] (1895) lautet:

$$k_1 := f(x,y), \tag{2.3.1a}$$

$$k_2 := f(x+\frac{h}{2}, y+\frac{h}{2}k_1), \tag{2.3.1b}$$

$$k_3 := f(x+\frac{h}{2}, y+\frac{h}{2}k_2), \tag{2.3.1c}$$

$$k_4 := f(x+h, y+hk_3), \tag{2.3.1d}$$

$$\phi(x,y;h,f) := \frac{1}{6}(k_1+2k_2+2k_3+k_4) \tag{2.3.1e}$$

Innerhalb der Klasse der *RK-Methoden* (Runge-Kutta-artigen Methoden) ist (2.3.1a-e) ein sogenanntes *vierstufiges* Verfahren. Offenbar benötigt es pro Schritt einen Aufwand von vier f-Auswertungen. Durch die Einführung der Zwischengrößen $k_1,\ldots,k_4$ erhält das Runge-Kutta-Verfahren seine einfache Gestalt. Elimination dieser Zwischengrößen ergäbe die explizite, aber sehr umständliche Formel $\phi(x,y;h,f) = \frac{1}{6}\left[f(x,y) + 2f(x+\frac{h}{2}, y+\frac{h}{2}f(x,y)) + \ldots\right]$.

Das Euler-Verfahren ordnet sich als triviale, einstufige RK-Methode ein: $k_1 := f(x,y)$, $\phi(x,y;h,f) := k_1$. Aber auch das Heun-Verfahren (2.2.3e) lässt sich als zweistufiges RK-Verfahren interpretieren: $k_1 := f(x,y)$, $k_2 := f(x+h, y+hk_1)$, $\phi(x,y;h,f) := \frac{1}{2}(k_1+k_2)$.

Satz 2.3.1 *Das Runge-Kutta-Verfahren (2.3.1a-e) ist ein explizites Einschrittverfahren der Konsistenzordnung $p = 4$.*

Beweis. a) Ungeachtet der Vierstufigkeit in (2.3.1a-d) ist ϕ aus (2.3.1e) direkt aus dem Anfangswert y berechenbar, ist also explizit.

b) Die Konsistenzordnung ergibt sich aus einer (länglichen) Taylor-Entwicklung (vgl. [3]). ∎

Die Koeffizienten der k_i in der letzten Gleichung (2.3.1e) haben einen direkten Bezug zu Quadraturformeln. Dazu nehme man den Spezialfall an, dass f nicht von y abhängt. Für $y'(x) = f(x)$, also $y(x) = y(x_0) + \int_{x_0}^{x} f(t)\,dt$, erhält man

$$k_1 = f(x),\ k_2 = k_3 = f(x+\frac{h}{2}),\ k_4 = f(x+h),$$
$$\phi(x,y;h,f) = \frac{1}{6}\left[f(x) + 4f(x+\frac{h}{2}) + f(x+h)\right].$$

Damit ist $h\phi$ die Simpson-Quadraturformel für $\int_x^{x+h} f(t)\,dt$, deren Restglied $\mathcal{O}(h^5) = \mathcal{O}(h^{p+1})$ beträgt. Somit ist die Konsistenzordnung $p = 4$ wenigstens für diesen Spezialfall bestätigt.

Es sei angemerkt, dass (2.3.1a-e) ist nicht die einzige vierstufige RK-Methode der Konsistenzordnung $p = 4$.

[11] Carle David Tolmé Runge, geb. 30. Aug. 1856 in Bremen, gest. 3. Jan. 1927 in Göttingen

[12] Martin Wilhelm Kutta, geb. 3. Nov. 1867 in Pitschen (Oberschlesien), gest. 25. Dez. 1944 in Fürstenfeldbruck

2.3.1 Numerisches Beispiel

Für das Anfangswertproblem (2.1.7) liefert das Runge-Kutta-Verfahren (2.3.1a-e) die folgenden Resultate:

Schrittweite h	Lösung $\eta(2,h)$	absoluter Fehler $\lvert\eta(2,h)-e^2\rvert$	relativer Fehler $\frac{\lvert\eta(2,h)-e^2\rvert}{e^2}$
$2/5 = 0.4000$	7.38679_{10}+00	2.26237_{10}-03	3.06273_{10}-04
$2/10 = 0.2000$	7.38889_{10}+00	1.66857_{10}-04	2.25822_{10}-05 (Faktor: 13.56)
$2/20 = 0.1000$	7.38904_{10}+00	1.13316_{10}-05	1.53356_{10}-06 (Faktor: 14.73)
$2/40 = 0.0500$	7.38906_{10}+00	7.38300_{10}-07	9.99181_{10}-08 (Faktor: 15.35)
$2/80 = 0.0250$	7.38906_{10}+00	4.71143_{10}-08	6.37622_{10}-09 (Faktor: 15.67)

Die Fehlerverbesserungsfaktoren entsprechen $16 = 2^4$ und zeigen somit die vierte Ordnung an.

Man beachte, dass das Runge-Kutta-Verfahren für $h = 0.0500$ ebenso viele f-Auswertungen benötigt wie das Euler-Verfahren für $h = 0.0125$. Die erzielten absoluten Fehler sind aber 7.38300_{10}-07 (Runge-Kutta) und 9.10352_{10}-02 (Euler).

In der Definition 2.2.6 der Konsistenz ist eine hinreichende Glattheit von f gefordert. Dass dies notwendig ist, zeigt das folgende Beispiel mit

$$y'(x) = f(x,y) := 1.1 \cdot |x|^{0.1}, \qquad y(0) = 0.$$

Die in Abbildung 2.3.1 illustrierte Konvergenz des Runge-Kutta-Verfahrens (RK4) ist von erster statt vierter Ordnung (ebenso wie für das Euler-Verfahren).

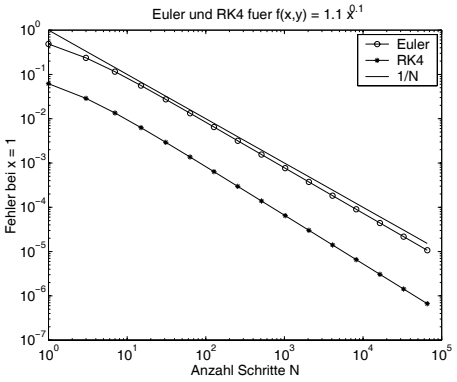


Abbildung 2.3.1: Runge-Kutta-Verfahren für nichtglattes f

2.3.2 Allgemeine Runge-Kutta-Verfahren

Definition 2.3.2 *ϕ heißt (explizites) m-stufiges Runge-Kutta-Verfahren, falls es die folgende Form hat:*

$$\begin{aligned} k_1 &= f(x,y), \\ k_2 &= f(x+\alpha_2 h, y+h\beta_{21}k_1), \\ &\vdots \\ k_m &= f(x+\alpha_m, y+h\textstyle\sum_{i=1}^{m-1}\beta_{mi}k_i), \qquad \phi(x,y;h,f) = \sum_{i=1}^{m}\gamma_i k_i\,. \end{aligned} \tag{2.3.2}$$

Zur kürzeren Notation wird das sogenannte Runge-Kutta-Schema *verwendet:*

$$\begin{array}{c|cccc} 0 & & & & \\ \alpha_2 & \beta_{21} & & & \\ \vdots & \vdots & \ddots & & \\ \alpha_m & \beta_{m1} & \cdots & \beta_{mm-1} & \\ \hline & \gamma_1 & \cdots & \gamma_{m-1} & \gamma_m \end{array} \tag{2.3.2'}$$

wobei hierin $\alpha_1 := 0$ festgelegt ist.

Das klassische Verfahren (2.3.1a-e) sieht in dieser Notation wie folgt aus:

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

Bemerkung 2.3.3 *Das Runge-Kutta-Verfahren (2.3.2) ist genau dann konsistent, wenn* $\sum_{i=1}^{m} \gamma_i = 1$ *ist.*

Beweis. Für $h \to 0$ folgt $k_i \to f(x,y)$, also $\phi(x,y;h,f) = \lim_{h\to 0} \sum_{i=1}^{m} \gamma_i k_i = \left(\sum_{i=1}^{m} \gamma_i\right) f(x,y)$. ∎

Übungsaufgabe 2.3.4 *Welches Quadraturverfahren entsteht aus (2.3.2) für* $f(x,y) = f(x)$*?*

Bemerkung 2.3.5 *Ein m-stufiges Runge-Kutta-Verfahren benötigt m f-Auswertungen pro Schritt. Sei* $p_{\text{opt}}(m)$ *die optimale Konsistenzordnung, die mit einem m-stufigen Runge-Kutta-Verfahren erreichbar ist. Die tabellierten Werte von* $p_{\text{opt}}(m)$ *lauten:*

m	**1**	**2**	**3**	**4**	5	**6**	**7**	8	**9**	10	**11**
p_{opt}	**1**	**2**	**3**	**4**	4	**5**	**6**	6	**7**	7	**8**

Damit ist das klassische Runge-Kutta-Verfahren (2.3.1a-e) das beste, das mit m f-Auswertungen die Ordnung $p_{\text{opt}}(m) = m$ *erreicht. Die zugehörigen Verfahren findet man in Grigorieff [1].*

2.3.3 Implizite Runge-Kutta-Verfahren

Die β-Koeffizienten in (2.3.2') bilden eine strikte, untere Dreiecksmatrix. Damit ist gesichert, dass k_{i+1} explizit aus $k_1, \ldots, k_i$ berechnet werden kann. Implizite Runge-Kutta-Verfahren haben dagegen die Form

$$\begin{aligned} &k_j = f(x + \alpha_j h, y + h\sum_{\ell=1}^{m} \beta_{j\ell} k_\ell) \quad \text{für } j \in \{1, \ldots, m\}, \\ &\phi = \textstyle\sum_{\ell=1}^{m} \gamma_\ell k_\ell \,. \end{aligned} \tag{2.3.3}$$

mit dem dazugehörigen Schema

$$\begin{array}{c|ccc} \alpha_1 & \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \vdots & & \vdots \\ \alpha_m & \beta_{m1} & \cdots & \beta_{mm} \\ \hline & \gamma_1 & \cdots & \gamma_m \end{array}$$

Bemerkung 2.3.6 *Gleichung (2.3.3) ist ein System von m (nichtlinearen) Gleichungen für* $k_1, \ldots, k_m$*.*

Ein Spezialfall sind diagonal-implizite Runge-Kutta-Verfahren. Hier gilt $\beta_{ij} = 0$ für $j > i$. Man erhält ein gestaffeltes System: Für jedes k_i ist eine nichtlineare Gleichung zu lösen:

$$\begin{array}{c|ccc} \alpha_1 & \beta_{11} & & \\ \vdots & & \ddots & \\ \alpha_m & \beta_{m1} & \cdots & \beta_{mm} \\ \hline & \gamma_1 & \cdots & \gamma_m \end{array}$$

Übungsaufgabe 2.3.7 *Man übertrage die Resultate zur Auflösbarkeit der impliziten Gleichungen aus Satz 2.2.1 auf die impliziten Runge-Kutta-Verfahren.*

2.3.4 Runge-Kutta-Gauß-Formeln

Falls in (2.3.3) alle Koeffizienten α_i, $\beta_{i\ell}$, γ_i frei wählbar sind, kann die optimale Lösung angegeben werden. Es ergeben sich die Runge-Kutta-Gauß-Formeln, die charakterisiert sind durch:

- α_k sind die Gauß-Stützstellen[13] im Intervall $[0,1]$ (zur Gewichtsfunktion $\omega = 1$), also die auf $[0,1]$ transformierten Stützstellen des Legendre-Polynoms[14] L_m.
- γ_k sind die zugehörigen Gewichte.
- $\beta_{k\ell}$ sind eindeutig bestimmt.

Satz 2.3.8 *Für Runge-Kutta-Gauß-Verfahren gilt die Konsistenzordnung $p = 2m$.*

Bemerkung 2.3.9 *Ist f nur von x abhängig, dann ist $h\phi(x_i, y_i; h, f)$ die Gauß-Quadratur von $\int_x^{x+h} f(\xi)\, d\xi$.*

Die Runge-Kutta-Gauß-Formeln der Ordnungen $m = 1, 2$ lauten wie folgt:

1. $m = 1$, $p = 2$. Dann ist $\eta_{i+1} = \eta_i + hk_1$ mit $k_1 = f(x_i + \frac{h}{2}, \eta_i + \frac{h}{2}k_1)$, d.h.

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \tag{2.3.4a}$$

2. $m = 2$, $p = 4$.

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{2.3.4b}$$

Beispiele höherer Ordnung findet man in Grigorieff [1].

2.3.5 Eingebettete Runge-Kutta-Verfahren

In dem späteren Kapitel 2.7.4 über Schrittweitensteuerung werden sogenannte "eingebettete Runge-Kutta-Verfahren" eine Rolle spielen. Hierbei handelt es sich um ein Paar von Runge-Kutta-Formeln mit übereinstimmenden Koeffizienten α_i und β_{ij}:

$$\begin{array}{c|ccccc} 0 & & & & & \\ \alpha_2 & \beta_{21} & & & & \\ \vdots & \vdots & \ddots & & & \\ \alpha_m & \beta_{m1} & \cdots & \beta_{mm-1} & & \\ \hline & \gamma_{1,1} & \cdots & \gamma_{1,m-1} & \gamma_{1,m} \\ & \gamma_{2,1} & \cdots & \gamma_{2,m-1} & \gamma_{2,m} \end{array}$$

Die $\gamma_{1,i}$ in der vorletzten Zeile sind hierbei die Koeffizienten zum ersten Verfahren ϕ_1, die $\gamma_{2,i}$ die zum zweiten Verfahren ϕ_2. Beide Verfahren sollen sich in der Ordnung unterscheiden. Da die α_i und β_{ij} für ϕ_1 und ϕ_2 gleich sind, besteht der Hauptrechenaufwand in der Berechnung der k_i (d.h. man bekommt praktisch zwei Runge-Kutta-Resultate für den Preis eines einzigen).

Beispiel 2.3.10 ([4, 5]) *In Tabelle 2.3.1 ist das erste Verfahren von fünfter und das zweite Verfahren von vierter Ordnung. Die Dormand-Princeschen Koeffizienten für ein Verfahren der Ordnungen 8 und 7 sind z.B. in [3, Seite 225] angegeben.*

[13] Johann Carl Friedrich Gauß, geb. 30. April 1777 in Braunschweig, gest. 23. Feb. 1855 in Göttingen
[14] Adrien-Marie Legendre, geb. 18. Sept. 1752 in Paris, gest. 10. Jan. 1833 in Paris

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Tabelle 2.3.1: Eingebettete Runge-Kutta-Verfahren der Ordnungen 5 und 4 nach Dormand-Prince [4]

2.4 Konvergenz von Einschrittverfahren

Zunächst wird gezeigt, dass es reicht, nur explizite Verfahren $\eta_{i+1} := \eta_i + h\phi(x, \eta_i; h, f)$ zu diskutieren. Man beachte, dass das ϕ aus dem nächsten Lemma nicht praktisch verfügbar ist, sondern nur für die theoretische Aussage verwendet wird, dass zu einem impliziten Verfahren ein äquivalentes explizites existiert.

Lemma 2.4.1 *Sei ein implizites Verfahren gegeben mit $\eta_{i+1} := \eta_i + h\varphi(x_i, \eta_i, \eta_{i+1}; h, f)$, wobei φ Lipschitz-stetig bezüglich η_i und η_{i+1} ist und $h \in (0, h_0]$ für ein hinreichend kleines $h_0 > 0$. Dann gibt es ein $\phi(x, y; h, f)$ mit $\phi(x_i, \eta_i; h, f) = \varphi(x_i, \eta_i, \eta_{i+1}; h, f)$, das Lipschitz-stetig bezüglich η_i ist.*

Beweis. Nach Satz 2.2.1 ist $\eta_{i+1} = \eta_i + h\varphi$ lösbar. Damit existiert die Umkehrfunktion $\psi(x, \eta_i; h, f)$ mit $\eta_{i+1} = \psi(x_i, \eta_i; h, f)$, wobei ψ Lipschitz-stetig bezüglich η_i ist (Satz über implizite Funktionen). Setze

$$\phi(x, y; h, f) := \varphi(x, y, \psi(x, y; h, f); h, f).$$

Für die Lipschitz-Konstante gilt dann $L_{\phi_{\text{bzgl. } y}} \le L_{\varphi_{\text{bzgl. } y}} + L_{\varphi_{\text{bzgl. } \eta_{i+1}}} \cdot L_{\psi_{\text{bzgl. } y}}$. ∎

Definition 2.4.2 *Sei f Lipschitz-stetig bezüglich y (Bedingung (1.3.1)). Ein Verfahren ϕ heißt konvergent, falls*

$$\lim_{h \to 0} \max_{\substack{x \in I \\ x = x_i \text{ für ein } i}} \|\eta(x, h) - y(x)\| = 0\,.$$

Für den später diskutierten Fall $x_{i+1} = x_i + ih_i$ variierender h_i ist $\lim_{h\to 0}$ durch $\lim_{\max(h_i)\to 0}$ zu ersetzen.

Satz 2.4.3 *Voraussetzungen:*
1) $I = [x_0, x_E]$, $x_E > x_0$, $h_0 > 0$, $p \in \mathbb{N}$, $\gamma \in (0, \infty]$.
2) $y \in C_L^p(I)$, d.h. die Lösung von $y(x_0) = y_0$, $y' = f(x, y)$ habe eine Lipschitz-stetige p-te Ableitung $y^{(p)}$.
3) $G := \{(x, y, h) : x \in I,\ \|y - y(x)\| \le \gamma,\ 0 < h \le h_0\}$.
4) Für die Inkrementfunktion des Einschrittverfahrens gelte $\phi \in C(G)$ und es existiere $M > 0$, sodass

$$\|\phi(x, y_1; h, f) - \phi(x, y_2; h, f)\| \le M \|y_1 - y_2\| \qquad \text{für alle } (x, y_1, h),\ (x, y_2, h) \in G.$$

5) p sei die Konsistenzordnung, d.h. es gibt ein $N \in \mathbb{R}$, sodass für alle $x \in I$, $0 < h \le h_0$ gilt

$$\|\tau(x, y(x), h)\| \le N h^p\,.$$

Dann gibt es ein $h_1 \in (0, h_0]$, sodass für alle Schrittweiten h mit $0 < h \le h_1$ die Näherung $\eta(x, h)$ wohldefiniert[15] ist und[5]

$$\|\eta(x, h) - y(x)\| \le N |h|^p \frac{e^{M|x - x_0|} - 1}{M} \qquad \text{für alle } x = x_0 + \nu h,\ \nu \in \mathbb{Z} \tag{2.4.1}$$

erfüllt. Im Falle $\gamma = \infty$ ist $h_1 = h_0$.

[15] Die Näherungen $\eta(x, h)$ sind wohldefiniert, wenn stets $(x, \eta(x, h), h) \in G$ gilt. Andernfalls ist die Inkrementfunktion nicht mehr definiert und das Einschrittverfahren nicht durchführbar.

Definition 2.4.4 *Falls (2.4.1) gilt, heißt p die (Konvergenz-)Ordnung von ϕ.*

Die wesentliche Aussage von Satz 2.4.3 lautet: $\left.\begin{matrix}\text{Konsistenz impliziert Konvergenz}\\ \text{Konsistenzordnung = Konvergenzordnung}\end{matrix}\right\}$.

Beweis von Satz 2.4.3. Wähle $h_1 \in (0, h_0]$ sodass $N\, h_1^p\, \dfrac{e^{M|x_E - x_0|} - 1}{M} \le \gamma$ (man beachte, dass $\gamma = \infty$ die Wahl $h_1 = h_0$ erlaubt). Setze $\delta_\nu := \|\eta(x_\nu, h) - y(x_\nu)\|$.

Wir wollen beweisen, dass

$$\delta_0 = 0, \quad \delta_\nu \le (1 + hM)\delta_{\nu-1} + Nh^{p+1} \qquad \text{für alle } \nu \in \mathbb{N} \text{ mit } x_\nu \in I. \tag{2.4.2}$$

Angenommen, die Größen δ_ν erfüllen (2.4.2), so folgt aus Lemma 2.1.6 (mit $a_\nu := \delta_\nu$, $L := M$, $B := N$, $k := p + 1$), dass

$$\delta_\nu \le Nh^p \frac{e^{\nu h M} - 1}{M} \underset{h \le h_1}{\le} N\, h_1^p\, \frac{e^{|x_E - x_0| M} - 1}{M} \underset{\text{Wahl von } h_1}{\le} \gamma.$$

Dies bedeutet, dass - solange (2.4.2) gilt - die Näherungen $\eta(x_\nu, h)$ (genauer: die Tripel $(x_\nu, \eta(x_\nu, h), h)$) in dem Definitionsbereich G liegen und das Einschrittverfahren somit wohldefiniert ist. Man beachte, dass definitionsgemäß $(x_\nu, y(x_\nu), h) \in G$ gilt.

(2.4.2) wird per Induktion bewiesen. Für $\nu = 0$ ist $\eta(x_0, h) = y_0$, sodass sich $\delta_0 = 0$ ergibt. Sei (2.4.2) für $\nu - 1 \in \mathbb{N}_0$ angenommen. Dann liefert

$$\begin{aligned}
&\delta_\nu = \|\eta(x_\nu, h) - y(x_\nu)\| = \|\eta(x_{\nu-1}, h) + h\phi(x_{\nu-1}, \eta(x_{\nu-1}, h), h) - y(x_\nu)\| \\
&= \|\eta(x_{\nu-1}, h) - y(x_{\nu-1}) - h\left(\tfrac{y(x_\nu) - y(x_{\nu-1})}{h} - \phi(x_{\nu-1}, y(x_{\nu-1}), h)\right) \\
&\qquad + h\left(\phi(x_{\nu-1}, \eta(x_{\nu-1}, h); h, f) - \phi(x_{\nu-1}, y(x_{\nu-1}); h, f)\right)\| \\
&\le \|\eta(x_{\nu-1}, h) - y(x_{\nu-1})\| + h\,\|\tau(x_{\nu-1}, y(x_{\nu-1}), h)\| + h\,\|\phi(x_{\nu-1}, \eta(x_{\nu-1}, h); h, f) - \phi(x_{\nu-1}, y(x_{\nu-1}); h, f)\| \\
&\le \delta_{\nu-1} + hNh^p + hM\delta_{\nu-1} = (1 + hM)\delta_{\nu-1} + Nh^{p+1}
\end{aligned}$$

die Ungleichung der Induktionsbehauptung für ν. ∎

2.5 Rundungsfehlereinfluss

Bisher sind wir davon ausgegangen, dass das Einschrittverfahren ohne Rundungsfehler durchgeführt wird. In der Praxis sind Rundungsfehler nicht nur unvermeidbar, sondern ihre Anzahl steigt auch proportional zur Anzahl der Schritte, d.h. proportional zu h^{-1}. Das ein Problem besteht, sieht man für sehr kleine Schrittweiten: Sei $\eta_0 = 1$. Sobald $|h\phi| < eps$, liefert die Gleitkomma-Arithmetik $\eta_1 = \eta_0 +_{gl} h\phi = \eta_0$ und folglich $\eta_i = \eta_0$ für alle i.

Wir machen die folgenden **Annahmen:**

1. Für $(x, y, h) \in G$ (mit G wie in Satz 2.4.3) gelte

$$\|\tilde{\phi}(x, y; h, f) - \phi(x, y; h, f)\| \le \varepsilon\,, \tag{2.5.1a}$$

   wobei $\tilde{\phi}$ die mit Rundungsfehlern behaftete Inkrementfunktion ϕ bezeichnet.

2. Der Multiplikationsfehler in $h *_{gl} \phi$ wird vernachlässigt, da $h\phi$ eine Größenordnung kleiner als η_i angenommen werden kann.

3. Der Additionsfehler in $\tilde{\eta}_{i-1} + (h\tilde{\phi})$ erfüllt

$$\left\|(\tilde{\eta}_{i-1} +_{gl} h\tilde{\phi}) - (\tilde{\eta}_{i-1} + h\tilde{\phi})\right\| \le eps\left\|\tilde{\eta}_{i-1} +_{gl} h\tilde{\phi}\right\| = eps\,\|\tilde{\eta}_i\|\,. \tag{2.5.1b}$$

Wir bezeichnen mit η_i das rundungsfehlerfreie Resultat zu ϕ, mit $\tilde{\eta}_i$ das rundungsfehlerbehaftete Resultat zu $\tilde{\phi}$ und $+_{gl}$. Die Differenz sei $\delta_i := \|\eta_i - \tilde{\eta}_i\|$, wobei $\delta_0 = 0$ gilt. M sei die Lipschitz-Konstante von ϕ. Damit erhalten wir

$$\begin{aligned}\delta_{i+1} = \|\eta_{i+1} - \tilde{\eta}_{i+1}\| &= \left\|\eta_i + h\phi(x_i, \eta_i; h, f) - (\tilde{\eta}_i +_{gl} h\tilde{\phi}(x_i, \tilde{\eta}_i; h, f))\right\| \\ &\leq eps\,\|\tilde{\eta}_{i+1}\| + \delta_i + h\,\|\phi(x_i, \eta_i; h, f) - \phi(x_i, \tilde{\eta}_i; h, f)\| + h\|\phi(x_i, \tilde{\eta}_i; h, f) - \tilde{\phi}(x_i, \tilde{\eta}_i; h, f)\| \\ &\leq eps(\|\eta_{i+1}\| + \delta_{i+1}) + (1 + hM)\delta_i + h\varepsilon\,.\end{aligned}$$

Da $\eta_i(x, h) \to y(x)$ und $\|y(x)\| \leq \|y\|_\infty$ in I gilt, gibt es eine Konstante E mit $\|\eta_i\| \leq E$ für alle $x_i \in I$. Aus $\delta_{i+1} \leq (1 + hM)\delta_i + h\varepsilon + epsE + eps\,\delta_{i+1}$ schließt man wegen $eps < 1$ auf

$$\delta_{i+1} \leq \frac{(1 + hM)\delta_i + h\varepsilon + epsE}{1 - eps}.$$

Lemma 2.1.6 mit $L = \dfrac{M + \dfrac{eps}{h}}{1 - eps}$, $h = a_0 = 0$, $B = \dfrac{h\varepsilon + epsE}{1 - eps}$ liefert $\delta_i \leq \dfrac{h\varepsilon + epsE}{h} \dfrac{e^{(M + \frac{eps}{h})\frac{|x_i - x_0|}{1 - eps}} - 1}{M + \frac{eps}{h}}$.
Da wir realistischerweise $Mh \gg eps$ und $1 \gg eps$ annehmen dürfen, ist $M + \frac{eps}{h} \approx M$ und $1 - eps \approx 1$ und erlaubt die Vereinfachung

$$\delta_i \lesssim \frac{h\varepsilon + epsE}{h} \frac{e^{M|x_i - x_0|} - 1}{M}.$$

Damit gilt der folgende Satz:

Satz 2.5.1 *Sei ϕ Lipschitz-stetig in y mit Konstante M, wobei $Mh \gg eps$ angenommen sei. Es gelte (2.5.1a) mit $\varepsilon = c_\varphi \cdot eps$ mit eps wie in (2.5.1b). Sei $E := \|y\|_\infty + 1$ (siehe oben). Dann gilt in erster Näherung*

$$\|\eta(x, h) - \tilde{\eta}(x, h)\| \leq eps \left(c_\varphi + \frac{E}{h}\right) \frac{e^{M|x_i - x_0|} - 1}{M}\,.$$

Korollar 2.5.2 *Es gelten die Voraussetzungen von Satz 2.5.1, und (2.4.1) sei erfüllt. Dann gilt*

$$\|\tilde{\eta}(x, h) - y(x)\| \leq \left(Nh^p + eps\Big(c_\varphi + \frac{E}{h}\Big)\right) \frac{e^{M|x_i - x_0|} - 1}{M}\,.$$

Es liegt also keine "numerische Konvergenz" für $h \to 0$ vor; im Gegenteil: die rechte Seite strebt gegen unendlich. Stattdessen stellt sich die Frage, wann die rechte Seite minimal wird. Man definiere hierzu $g(h) := Nh^p + eps\left(c_\varphi + \frac{E}{h}\right)$. Die Nullstelle der Ableitung $g'(h) = pNh^{p-1} - \frac{epsE}{h^2}$ liefert die "optimale" Schrittweite $h_{opt} = \left(\frac{epsE}{pN}\right)^{\frac{1}{p+1}}$ und den Wert $g(h_{opt}) = c_\varphi eps + (p + 1)N\left(\frac{epsE}{pN}\right)^{\frac{p}{p+1}}$.

Korollar 2.5.3 *Sei eps die relative Maschinengenauigkeit und p die Ordnung des Verfahrens. Dann ist $h_{opt} = \mathcal{O}(eps^{\frac{1}{p+1}})$ und für den Fehler gilt $\|\tilde{\eta}(x, h) - y(x)\| \leq \mathcal{O}(eps^{\frac{p}{p+1}})$.*

Folgerung 2.5.4 *Eine hohe Ordnung p bringt Vorteile:*

Euler:	$p = 1$	$h_{opt} = \mathcal{O}(eps^{1/2})$	*Fehler* $\approx \mathcal{O}(eps^{1/2})$	*Kosten* $\approx \mathcal{O}(eps^{-1/2})$,
Runge-Kutta:	$p = 4$	$h_{opt} = \mathcal{O}(eps^{1/5})$	*Fehler* $\approx \mathcal{O}(eps^{4/5})$	*Kosten* $\approx \mathcal{O}(eps^{-1/5})$,
	$p \to \infty$	$h_{opt} \to \mathcal{O}(1)$	*Fehler* $\approx \mathcal{O}(eps)$	*Kosten* $\to \mathcal{O}(1)$.

Die letzte Spalte weist auf einen weiteren Vorteil einer hohen Ordnung hin: Der Rechenaufwand ist proportional[16] *zu $1/h$ und damit geringer, wenn h_{opt} größer wird. Zusammen gibt es einen doppelten Effekt: Ein Verfahren der hohen Ordnung p erzielt den kleineren Fehler $\mathcal{O}(eps^{\frac{p}{p+1}})$ zu geringeren Kosten $\mathcal{O}(eps^{\frac{-1}{p+1}})$.*

[16] Die Darstellung ist etwas vereinfacht. Runge-Kutta-Verfahren der Ordnung p benötigen pro Schritt die wachsende Anzahl von $\beta(p)$ f-Auswertungen, wobei $\beta(\cdot)$ die Umkehrfunktion von $p_{\text{opt}}(\cdot)$ aus Bemerkung 2.3.5 ist. In der Aussage ist eine konstante Anzahl von f-Auswertungen unterstellt.

2.6 Asymptotische Entwicklung von $\eta(x,h)$

2.6.1 Existenz einer asymptotische Entwicklung

Mit dem Konvergenzresultat gilt $\eta(x,h) = y(x) + \mathcal{O}(h^p)$, wobei p die Ordnung des Verfahrens bezeichnet. Das Ziel ist nun eine *asymptotische Entwicklung* der Form

$$\eta(x,h) = y(x) + e_p(x)h^p + e_{p+1}(x)h^{p+1} + \ldots + e_{q-1}h^{q-1} + e_q(x,h)h^q \tag{2.6.1a}$$

für ein $q \geq p$ mit

$$\|e_q(x,h)\| \leq const \text{ für } x \in I,\ h \to 0, \qquad e_p, \ldots, e_{q-1} \text{ unabhängig von } h. \tag{2.6.1b}$$

Satz 2.6.1 *Seien $f(x,y) \in C^{q+1}(U)$ und ϕ ein Einschrittverfahren der Ordnung p mit $\phi \in C^{q+1}(U \times [0,h_0])$. Dann sind die Gleichungen (2.6.1a,b) erfüllt.*

Beweisskizze. 1) Wir machen den Ansatz (2.6.1a): $\eta(x,h) = \sum_{\nu=0}^{q} e_\nu h^\nu$, $e_0 = y$, $e_1 = \ldots = e_{p-1} = 0$, wobei e_ν bis auf e_q nur von x abhängt (zur Vereinfachung der Schreibweise wird das Argument h in e_q unterdrückt). Einsetzen in $0 = \eta_{i+1} - \eta_i - h\phi(x_i, \eta_i; h, f)$ liefert

$$0 = \sum_{\nu=0}^{q} e_\nu(x_i + h)h^\nu - \sum_{\nu=0}^{q} e_\nu(x_i)h^\nu - h\phi\left(x_i, \sum_{\nu=0}^{q} e_\nu(x_i)h^\nu; h, f\right)$$

Taylor-Entwicklung um $x = x_i$ liefert für $h^\nu e_\nu(x_i + h)$

$$\begin{array}{rcccccccc} e_0(x+h) & = & e_0(x) & + & he_0'(x) & + & \frac{h^2}{2}e_0''(x) & + & \frac{h^3}{6}e_0'''(x) \quad + \ldots \\ he_1(x+h) & = & & & he_1(x) & + & h^2e_1'(x) & + & \frac{h^3}{2}e_1''(x) \quad + \ldots \\ h^2e_2(x+h) & = & & & & & h^2e_2(x) & + & h^3e_2'(x) \quad + \ldots \end{array}$$

Subtrahiert man von dieser Summe $\sum_{\nu=0}^{q} e_\nu(x)h^\nu$, verbleibt

$$\sum_{\nu=0}^{q} e_\nu(x+h)h^\nu - \sum_{\nu=0}^{q} e_\nu(x)h^\nu = he_0'(x) + h^2\left[\frac{1}{2}e_0''(x) + e_1'(x)\right] + h^3\left[\frac{1}{6}e_0'''(x) + \frac{1}{2}e_1''(x) + e_2'(x)\right] + \ldots$$

Verwickelter ist die Entwicklung von $h\phi(x, \sum_{\nu=0}^{q} e_\nu h^\nu; h, f)$:

$$h\phi(x, \sum_{\nu=0}^{q} e_\nu h^\nu; h, f) = h\phi(x, e_0; 0, f) + h^2\left[\phi_y(x, e_0; 0, f)e_1 + \phi_h(x, e_0; 0, f)\right] + \ldots$$

Koeffizientenvergleich der h-Potenzen liefert:

$$\begin{array}{ll} h^0: & 0 = e_0(x) - e_0(x) \\ h^1: & 0 = e_0'(x) - \phi(x, e_0(x); 0, f) = y'(x) - f(x,y) \qquad \text{(Konsistenz: } e_0 = y\text{)} \\ h^2: & 0 = \frac{1}{2}e_0''(x) + e_1'(x) - \phi_y(x, e_0(x); 0, f)e_1(x) - \phi_h(x, e_0(x); 0, f) \end{array}$$

Die erste Gleichung ist trivialerweise erfüllt. Die zweite Zeile liefert $e_0 = y$, da $\phi(x, e_0(x); 0, f) = \lim_{h\to 0} \phi(x, e_0(x); h, f) = f(x, e_0(x))$ wegen der Konsistenzbedingung (2.2.5). Die Gleichung zur h^2-Potenz ist eine Differentialgleichung für e_1, da $e_0 = y$ und seine Ableitungen bekannt sind. Es stellt sich heraus, dass auch die weiteren Gleichungen zu $h^{\nu+1}$ $(2 \leq \nu + 1 \leq q)$ lineare Differentialgleichungen für e_ν ergeben, die von der Form

$$e_\nu(x_0) = 0, \quad e_\nu'(x) = \phi_y(x, y(x); 0, f)e_\nu + g_\nu(x, y, e_1, e_2, \ldots, e_{\nu-1}) \tag{2.6.2}$$

sind (g_ν enthält auch die Ableitungen seiner Argumente e_μ, $\mu < \nu$). Der Anfangswert $e_\nu(x_0) = 0$ $(\nu > 0)$ ergibt sich aus $\eta(x_0, h) = y_0$. Die obigen Gleichungen bilden ein *gestaffeltes System* von linearen Differentialgleichungen für e_1, $e_2, \ldots, e_{q-1}$.

2) Man setze $y(x,h) = \sum_{\nu=0}^{q-1} e_\nu(x)h^\nu$ mit den oben berechneten Funktionen e_ν. Nun beweise man, dass $y(x,h) - \eta(x,h) = \mathcal{O}(h^q)$ gilt. Damit ist $e_q(x,h)h^q = \mathcal{O}(h^q)$ gezeigt, also $e_q(x,h) = \mathcal{O}(1)$ wie gefordert.

3) Aus $\tau = \mathcal{O}(h^p)$ folgt $\eta(x,h) = y(x) + \mathcal{O}(h^p)$. Damit sind $e_1 = e_2 = \ldots = e_{p-1} = 0$ (offenbar muss der Ausdruck $g_\nu(x, y, e_1, e_2, \ldots, e_{\nu-1}) = g_\nu(x, y, 0, 0, \ldots, 0)$ für $\nu < p$ verschwinden). ∎

Bemerkung 2.6.2 *Die Aussage von Satz 2.6.1 lässt sich auch für implizite Einschrittverfahren nachweisen.*

Eine spezielle implizite Methode ist die Trapezformel (2.2.3c). Ihre Anwendung auf $f = f(x)$ liefert die asymptotische Entwicklung der summierte Trapezquadraturformel. Bekanntlich besitzt aber die summierte Trapezformel eine asymptotische Entwicklung in h^2, d.h. $e_\nu = 0$ für ungerade ν. Ob diese vorteilhafte Eigenschaft auch für das Einschrittverfahren vorliegt, lässt sich anhand der Inkrementfunktion einfach überprüfen.

Bemerkung 2.6.3 *Die Inkrementfunktion ϕ des (notwendigerweise impliziten) Einschrittverfahrens sei für $h \in [-h_0, h_0]$ definiert und erfülle die Symmetriebedingung*

$$\eta_{i+1} = \eta_i + h\phi(x_i, \eta_i, \eta_{i+1}; h, f) \quad \Longleftrightarrow \quad \eta_i = \eta_{i+1} - h\phi(x_i, \eta_{i+1}, \eta_i; -h, f).$$

Dann gilt $e_\nu = 0$ für ungerade ν, d.h. in der asymptotischen Entwicklung treten nur Potenzen von h^2 auf.

Die Trapezformel (2.2.3c) erfüllt offenbar diese Symmetriebedingung, ebenso die "mid-Euler"-Formel

$$\phi(x, y_0, y_1; h, f) := f(x + \frac{h}{2}, \frac{y_0 + y_1}{2}).$$

2.6.2 Extrapolationsverfahren

Die Existenz einer asymptotische Entwicklung ist die Voraussetzung, um das folgende *Extrapolationsverfahren* anwenden zu können. Die Entwicklung wird hier etwas anders notiert, da wir nur die nichtverschwindenden h-Potenzen aufschreiben wollen:

$$\eta(x,h) = y(x) + e_1(x)h^{p_1} + e_2(x)h^{p_2} + \ldots e_m(x)h^{p_m} + E_{m+1}(x,h)h^{p_{m+1}} \quad \text{mit } \|E_{m+1}(x,h)\| = \mathcal{O}(1)$$

und $0 < p_1 < p_2 < \ldots < p_{m+1}$.

$x \in I$ sei fest. Für mehrere Schrittweiten h_i $(0 \le i \le m)$ mit $\dfrac{x}{h_i} \in \mathbb{Z}$ seien $\eta(x; h_i)$ gegeben. Dann ist

$$\eta(x,h_i) = y(x) + \sum_{\nu=1}^{m} h_i^{p_\nu} e_\nu(x) + E_{m+1}(x,h_i)h_i^{p_{m+1}} \qquad \text{für } i = 0, \ldots, m$$

Zunächst stelle man sich vor, dass der Restterm nicht aufträte: $\eta(x,h_i) = y(x) + \sum_{\nu=1}^{m} h_i^{p_\nu} e_\nu(x)$. Dies beschriebe ein System mit $m+1$ Gleichungen und den $m+1$ Unbekannten $y(x), e_1(x), \ldots, e_m(x)$. Offenbar könnten wir jetzt die *exakte* Lösung $y(x)$ erhalten.

Allgemein bestimme man die Lösung $(\gamma_0, \ldots, \gamma_m)$ der Gleichungen

$$\sum_{\nu=0}^{m} \gamma_\nu = 1, \quad \sum_{\nu=0}^{m} h_i^{p_1}\gamma_\nu = \sum_{\nu=0}^{m} h_i^{p_2}\gamma_\nu = \ldots = \sum_{\nu=0}^{m} h_i^{p_m}\gamma_\nu = 0.$$

Dann ist

$$\eta_{ex}(x, h_0, \ldots, h_m) := \sum_{\nu=0}^{m} \gamma_\nu \eta(x, h_i) = y(x) + \sum_{\nu=0}^{m} \gamma_\nu E_{m+1}(x, h_i) h_i^{p_{m+1}},$$

d.h. die *extrapolierte* Lösung η_{ex} stimmt bis auf $\mathcal{O}\left(\max_i(h_i^{p_{m+1}})\right)$ mit der exakten Lösung $y(x)$ überein.

Die naheliegende Anwendung der Extrapolation sähe so aus, dass man beginnend bei $x = x_0$ mit einem Einschrittverfahren ϕ bis $x = x_E$ rechnet, um dann den extrapolierten Wert $\eta_{ex}(x_E, h_0, \ldots, h_m)$ als Endresultat zu nehmen.

Günstiger ist ein *wiederholter Neustart:*
1) Man wähle eine grobe Schrittweite H mit $\frac{x_E - x_0}{H} \in \mathbb{Z}$, die die Teilintervalle $I_\mu = [x_0 + (\mu - 1)\, H, x_0 + \mu H]$ definiert.
2) Im Falle von $\mu = 0$ ist der Anfangwert y_0 am linken Randpunkt von I_0 gegeben.
3) Sei der Anfangwert y_μ am linken Randpunkt $x_0 + (\mu - 1)\, H$ von I_μ gegeben. Man berechne $\eta(x_0 + \mu H, h_i)$ für verschiedene Schrittweiten h_i mit $H/h_i \in \mathbb{N}$ und nehme den extrapolierten Wert $\eta_{ex}(x_0 + \mu H, h_0, \ldots, h_m)$ als neuen Anfangswert $y_{\mu+1}$ für das nächste Intervall $I_{\mu+1}$.

Bemerkung 2.6.4 *Für die Wahl der Schrittweiten $h_0, \ldots, h_m$ gibt es verschiedene Möglichkeiten:*
1) Die klassische Wahl ist die Romberg-Folge[17] *$h_i = 2^{-i}h_0$. Der Aufwand wird im wesentlich durch die kleinste Schrittweite $h_m = 2^{-m}h_0$ bestimmt und ist für größere m unerfreulich hoch.*
2) Die Bulirsch-Folge *ist durch h_0, $h_1 := \frac{h_0}{2}$ und $h_i := \frac{h_{i-2}}{3}$ für $i > 1$ definiert und führt zu $h_m \approx \sqrt{3}^{-m} h_0$.*
3) Eine langsame Nullfolge wie $h_i = h_0/i$ ist billiger, aber instabil für $m \to \infty$. Die genaue Stabilitätsbedingung lautet: $h_{i+1} \leq \alpha h_i$ für ein $0 < \alpha < 1$ und alle $i \geq 0$ (vergleiche [12, Kapitel 2.4.4: Romberg-Quadratur]).

2.7 Schrittweitensteuerung

Bisher war h fest, die Stützstellen haben ein äquidistantes Gitter gebildet. Jetzt soll $x_{\nu+1} = x_\nu + h_\nu$ mit möglicherweisen unterschiedlichen Schrittweiten h_ν gelten. Wir erhalten folgenden algemeinen Algorithmus:

$$\text{Start } \eta_0 = y(x_0);\ h_0 \text{ vorgegeben} \tag{2.7.1a}$$
$$\text{Iteration } \eta_{i+1} := \eta_i + h_i \phi(x_i, \eta_i, [\eta_{i+1},]\, h_i) \tag{2.7.1b}$$
$$h_{i+1} \text{ bestimmen (Schrittweitensteuerung)} \tag{2.7.1c}$$

Eventuell wird in (2.7.1b) die Schrittweite h_i verworfen und geeignet verbessert.

Für die Schrittsteuerung gibt es zwei Ziele:

- *Effizienz:* Das Resultat (z.B. $y(x_E)$) soll mit möglichst wenig Aufwand (im Allgemeinen also mit möglichst wenigen f-Auswertungen) erhalten werden, wobei die berechnete Näherung (mit einer gewissen Sicherheit) eine vorgegebene Genauigkeit erfüllt.
- *Robustheit, Zuverlässigkeit:* Die Schrittweitensteuerung soll nicht zur für eine spezielle Differentialgleichung funktionieren, sondern für eine möglichst[18] große Klasse von Problemen.

2.7.1 Notwendigkeit der Schrittweitensteuerung

Wir geben zwei Beispiele. Abbildung 2.3.1 zeigt, dass das klassische Runge-Kutta-Verfahren für die Differentialgleichung mit nichtglattem $f(x, y) := 1.1 \cdot |x|^{0.1}$ nicht die erhoffte vierte Ordnung ergibt. Da f mit wachsendem x glatter wird (d.h. die Ableitungen werden kleiner), liegt es nahe, die Schrittweiten mit der Entfernung von $x = 0$ größer werden zu lassen. Ein geeignetes "graduiertes Gitter" ist durch $x_i = (i/N)^{4/1.1}$, $i = 0, \ldots, N$, gegeben. Die Schrittweite ergibt sich hier als $h_i := x_{i+1} - x_i$. Das klassische Runge-Kutta-Verfahren für diese Schrittweiten liefert, wie man in Abbildung 2.7.1 sieht, wieder Konvergenz vierter Ordnung.

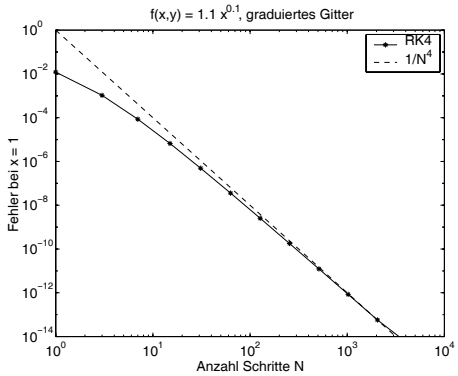


Abbildung 2.7.1: Runge-Kutta-Verfahren für nichtglattes f und graduiertes Gitter $x_i = (i/N)^{4/1.1}$

Das zweite Beispiel ist die Kepler-Bahn[19] eines Körpers um einen festen Massepunkt bei $\mathbf{x} = 0$ (z.B. Kometenbahn um die Sonne). Die Bahn $\mathbf{x}(t)$ im $\mathbb{R}^3$ lautet $\ddot{\mathbf{x}}(t) = \mathbf{x}(t)/r^3$ mit der Euklidischen Norm

[17] Werner Romberg, geb. 16. Mai 1909 in Berlin, gest. 20. Febr. 2003 in Heidelberg
[18] Eine absolute Sicherheit kann nicht erwartet werden (vgl. Bemerkung 2.7.1)
[19] Johannes Kepler, geb. 27. Dez. 1571 in Weil, Württemberg, gest. 15 Nov. 1630 in Regensburg

$r := |\mathbf{x}(t)|$. Gemäß Bemerkung 1.1.2c kann dieses Differentialgleichungssystem zweiter Ordnung in ein System erster Ordnung umgeschrieben werden. Der Graph $(t, \mathbf{x}(t))$ beschreibt bekanntlich eine Ellipse mit $\mathbf{x} = 0$ als einem Brennpunkt. Verwendet man Polarkoordinaten (r, φ), so lautet der zweite Keplersche Satz $r\dot{\varphi} = const.$ Im Falle einer langgestreckten Ellipse (wie bei Kometenbahnen üblich) wird der stark gekrümmte Ellipsenteil bei kleinem r besonders schnell durchlaufen, d.h. $\mathbf{x}(t)$ hat große Ableitungen, wenn $\mathbf{x}(t)$ dem Zentrum $\mathbf{x} = 0$ nahekommt, während der überwiegende Teil der Ellipse fast linear verläuft. Entsprechend wird man vermuten, dass man Schrittweiten verwenden sollte, die sich mit kleiner werdendem r verfeinern.

Das erste der Beispiele ist eines der seltenen, in dem die geeignete Schrittweite explizit vorhergesagt werden kann. Im Allgemeinen muss die Schrittweite numerisch bestimmt werden. Die hierfür möglichen Techniken werden als nächstes diskutiert.

2.7.2 Genauigkeit pro Schritt

Die nachfolgend beschriebenen Techniken versuchen, den lokalen Konsistenzfehler $\tau(x_i, \eta_i, h_i)$ zu kontrollieren. Pro Schritt werden 1) die bisherigen Fehler fortgepflanzt und 2) der neue Fehlerterm $h_i\tau(x_i, \eta_i, h_i)$ addiert. Im Grund kann man nur versuchen, eine Ungleichung

$$\|\tau(x_i, \eta_i, h_i)\| \leq \tau_0$$

einzuhalten. Der an sich viel interessante globale Fehler $\|y(x_i) - \eta_i\|$ ist nicht direkt zugänglich. Der Grund ist, dass der Effekt der Fehlerfortpflanzung während der Rechnung (ohne weiteres Wissen über die Lösung der Differentialgleichung) schwer einzuschätzen ist[20]. Im besten Fall wird der Fehler gedämpft, sodass man einen globalen Fehler proportional zu $\max(h_i)\tau_0$ erwartet. Werden die Fehler unverändert weitergetragen (oder - genauer gesagt - ebenso verstärkt wie die Lösung), wäre der globale Fehler in der Größenordnung von $\sum_i h_i \|\tau(x_i, \eta_i, h_i)\| \leq (x_E - x_0)\,\tau_0$ zu erwarten. Diese Abschätzung ist der Grund, warum man $\|\tau(x_i, \eta_i, h_i)\|$ durch τ_0 beschränkt und nicht den mit h_i multiplizierten Fehler $h_i \|\tau(x_i, \eta_i, h_i)\|$. Wenn dagegen die lokalen Fehler exponentiell ansteigen (ohne dass die Lösung das gleiche Wachstum hat - im diesem Fall liegt ein schlecht konditioniertes Problem vor!), müssten die ersten lokalen Fehler wesentlich kleiner sein als die späteren.

2.7.3 Steuerung durch Halbierung

Gegeben seien $y_0 = \eta_0$ und eine Genauigkeitsschranke τ_0 (hier wird $\tau_0 \gg eps$ angenommen, damit der Rundungsfehlereinfluss vernachlässigt werden kann). Wir definieren $e(x; h) := \eta(x; h) - y(x)$, wobei die verwendete (konstante) Schrittweite h als Parameter auftritt. Nach einem Schritt stimmt $-e(x_0 + h; h)$ mit $h\tau(x_0, y_0; h), h)$ überein. Gesucht wird daher ein h mit

$$\|e(x_0 + h; h)\| \approx \tau_0 h. \tag{2.7.2}$$

Wir nehmen an, dass sich die Lösung in der Form

$$\eta(x; h) = y(x) + h^p e_p(x) + \ldots \qquad (p\text{: Konsistenzordnung})$$

entwickeln lässt, womit $e(x_0 + h; h) = h^p e_p(x_0 + h) + \ldots$ gilt. (2.6.2) entnimmt man $e_p(x_0) = 0$. Taylor-Entwicklung liefert daher

$$e_p(x_0 + h) = e_p(x_0) + he'_p(x_0) + \ldots = he'_p(x_0) + \ldots,$$

sodass $e(x_0 + h; h) = h^{p+1}e'_p(x_0) + \ldots$ Man beachte, dass nach Definition des lokalen Diskretisierungsfehlers $e(x_0 + h; h) = -h\tau(x_0, \eta_0; h)$ gilt, sodass das obige p wirklich mit der Konsistenzordnung übereinstimmt.

Kombiniert man diese Aussage mit (2.7.2), so erhält man

$$h \approx \sqrt[p]{\frac{\tau_0}{\left\|e'_p(x_0)\right\|}}. \tag{2.7.3}$$

[20] Will man wirklich den globalen Fehler unter Kontrolle halten, muss man zusätzlich eine sogenannte adjungierte Differentialgleichung mit Anfangswerten bei x_E lösen. In diesem Falle müsste man iterativ vorgehen: Zunächst sind rohe Näherungen auf dem gesamten Intervall I zu ermitteln (und abzuspeichern), dann kann man in einem zweiten Durchgang Schätzungen für die geeigneten Schrittweiten h_i erhalten.

Zur numerischen Approximation des unbekannten Wertes $e'_p(x_0)$ ermittelt man wie folgt zwei Bestimmungsgleichungen. Im Folgenden ist H eine "Versuchsschrittweite"; naheliegend ist die vorhergehende Schrittweite: $H = h_i$. Ein einziger Schritt mit einer Schrittweite H liefert

$$\eta(x_0 + H; H) = y_0 + H^p e_p(x_0 + H) + \dots,$$

während zwei Schritte mit der halben Schrittweite

$$\eta(x_0 + H; \frac{H}{2}) = y_0 + \left(\frac{H}{2}\right)^p e_p(x_0 + H) + \dots$$

ergeben. Subtraktion beider Gleichung führt auf

$$\begin{aligned}\eta(x_0 + H; H) - \eta(x_0 + H; \frac{H}{2}) &= H^p e_p(x_0 + H) - 2^{-p} H^p e_p(x_0 + H) + \dots \\ &= (1 - 2^{-p}) H^p e_p(x_0 + H) + \dots = H^{p+1}(1 - 2^{-p}) e'_p(x_0) + \dots,\end{aligned}$$

wobei "..." für höhere H-Terme steht. Nach Vernachlässigung von "..." erhält man

$$e'_p(x_0) \approx \frac{1}{H^{p+1}(1 - 2^{-p})} \left[\eta(x_0 + H; H) - \eta(x_0 + H; \frac{H}{2})\right]. \tag{2.7.4}$$

Zusammen mit (2.7.3) ergibt dies

$$h \quad \approx \quad H \cdot \sqrt[p]{\frac{\tau_0(1 - 2^{-p})H}{\left\|\eta(x_0 + H; H) - \eta(x_0 + H; \frac{H}{2})\right\|}} \tag{2.7.5}$$

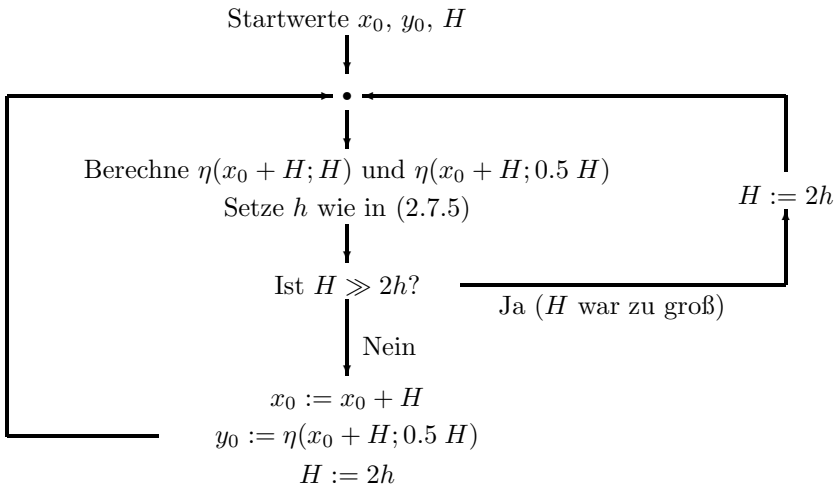
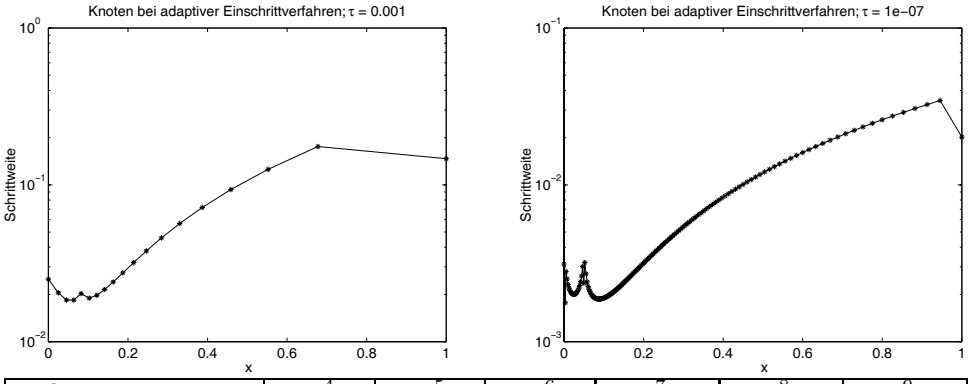


Abbildung 2.7.2: Algorithmus der Schrittweitensteuerung mittels Halbierung

Die Schrittweitensteuerung kann wie in Abbildung 2.7.2 aussehen. Allerdings hat man in der praktischen Implementation noch weitere Sicherheitsmaßnahmen zu treffen: 1) Sobald $h \approx \frac{H}{2}$ oder $h \approx H$ (z.B. in der Form $h/\frac{H}{2} \in [\rho, 1/\rho]$ für ein $\rho \in (0, 1)$) gilt, sollte die neue Schrittweiten $h := \frac{H}{2}$ bzw. $h := H$ akzeptiert werden, da dann die bereits vorhandenen Resultate $\eta(x_0 + \frac{H}{2}; \frac{H}{2})$ bzw. $\eta(x_0 + H; H)$ verwendet werden können. 2) Im Falle einer Vergröberung ($h_{i+1} > h_i$) ist sicherzustellen, dass der Quotient h_{i+1}/h_i nicht zu groß wird. 3) Ebenso sollten Maßnahmen getroffen werden, dass die Schrittweite nicht zu klein wird ($h \geq h_{\min}$), wenn die berechnete Lösung keine Singularität zeigt. 4) Schließlich ist noch auszuschließen, dass die Schrittweiten-Schleife nicht zwischen kleinen und großen Schrittweiten ohne Ende iteriert.

Als Beispiel wird die Anfangswertaufgabe $y'(x) = f(x,y) := -200x\,(y(x))^2$, $y(0) = 1$ verwendet. Die Lösung $y(x) = 1/\left[1+100x^2\right]$ variiert stark in $[0,1]$. Als Einschrittverfahren wird das klassische Runge-Kutta-Verfahren verwendet. Abbildung 2.7.3 zeigt die Wahl der Knotenpunkte x_i. Maximale und minimale Schrittweite unterscheiden sich um einen Faktor 10. Die angegebene Zahl der benötigten Schritte ist entscheidend für den Aufwand.



Toleranz τ_0	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
Anzahl der Schritte	33	55	93	163	286	506
Fehler	2.0_{10}-7	2.7_{10}-8	1.0_{10}-8	4.5_{10}-10	8.5_{10}-11	4.8_{10}-12

Abbildung 2.7.3: Adaptives Runge-Kutta-Verfahren bei der Lösung $y(x) = 1/\left[1+100x^2\right]$

2.7.4 Steuerung durch zwei Verfahren verschiedener Ordnung

Seien ϕ_1 und ϕ_2 zwei Einschrittverfahren der Ordnungen p und $p+1$, das heißt für die lokalen Diskretisierungsfehler gilt $\tau_1(x;h) = \mathcal{O}(h^p)$ und $\tau_2 = \mathcal{O}(h^{p+1})$. Man berechne mit einer Versuchsschrittweite H

$$\eta_1(x_1;H) := y_0 + H\phi_1(x_0,y_0;H) \quad \text{und} \quad \eta_2(x_1;H) := y_0 + H\phi_2(x_0,y_0;H).$$

Wir setzen voraus, dass sich der Konsistenzfehler in H entwickeln lässt: $\tau_1(x_0,H) = c_\tau H^p + \mathcal{O}(H^{p+1})$. Dann ist auch

$$\eta_1(x_1;H) - \eta_2(x_1;H) = [y(x_1) - H\tau_1] - [y(x_1) - H\tau_2] = H\tau_2 - H\tau_1 = -c_\tau H^{p+1} + \mathcal{O}(H^{p+2}). \tag{2.7.6}$$

d.h. $\|c_\tau\| \approx \|\eta_1(x_1;H) - \eta_2(x_1;H)\|\,/H^{p+1}$ ist näherungsweise bekannt.

Das Ziel ist es nun, h so zu bestimmen, dass gilt

$$\|\tau_1(x_0,h)\| = \|\eta_1(x_0+h;h) - y(x_0+h)\|\,/h \approx \tau_0 \tag{2.7.7}$$

(vgl. (2.7.2)), wobei y Lösung zum Anfangswert $y(x_0) = y_0$ ist. Unter Vernachlässigung höherer Terme gilt

$$y(x_0+h) - \eta_1(x_0+h;h) = h\tau_1 \approx c_\tau h^{p+1}.$$

Aus $\|\eta_1(x_0+h;h) - y(x_0+h)\| \approx \|c_\tau\|\,h^{p+1} \underset{(2.7.6)}{\approx} \|\eta_1(x_0;H) - \eta_2(x_0;H)\|\,h^{p+1}/H^{p+1} \underset{(2.7.7)}{\approx} \tau_0 h$ folgt

$$h \approx H \sqrt[p]{\frac{\tau_0 H}{\|\eta_1(x_1;H) - \eta_2(x_1;H)\|}}\,.$$

Zur praktische Durchführung wähle man ϕ_1 und ϕ_2 so, dass beide simultan berechnet werden können, ohne doppelten Aufwand zu benötigen. Typisch sind *eingebettete Runge-Kutta-Verfahren*, wie sie in §2.3.5 eingeführt wurden.

Bemerkung 2.7.1 *Eine "hundertprozentig sichere" Steuerung gibt es nicht.*

Beweis. Eine deterministische Steuerung sei vorgegeben. Die Steuerung möge im Falle der Differentialgleichung $y = f := 0$ die Stützstellen $(x_\nu, 0) \in \mathbb{R}^{n+1}$ benutzen. Man wähle ein $f \in C^\infty$, dass $\not\equiv 0$ ist, aber für alle x_ν den Wert $f(x_\nu, 0) = 0$ besitzt. Dann hat $y' = f(x, y)$ nicht die Lösung $y = 0$, das Näherungsverfahren produziert aber $\eta_i = 0$. ∎

Übungsaufgabe 2.7.2 *Die Funktion $y(x) = \tanh(ax)$ ist Lösung von $y' = a\left(1 - y^2\right)$ und nähert sich für $a \to \infty$ punktweise der Signumfunktion $\operatorname{sign}(x)$. Man wähle den Anfangswert $y(-1) = \tanh(-a)$ und löse das Anfangswertproblem in $[-1, 1]$ für verschiedene $a \geq 1$ unter Verwendung der Schrittweitensteuerung.*

3 Steife Differentialgleichungen

Insbesondere aus der Chemie stammen Differentialgleichungen, für die die Standardverfahren keine vernünftige Resultate liefern, obwohl das Konvergenzresultat aus §2.4 gültig bleibt. Zwar sind die Gleichungen der chemischen Reaktionskinetik nichtlinear. Der Effekt lässt sich aber am übersichtlichsten an linearen Differentialgleichungen erklären.

3.1 Begriff der Steifheit

Im folgenden sei

$$y' = Ay, \qquad y(0) = y_0, \tag{3.1.1}$$

ein Anfangswertproblem mit $n \geq 2$ Differentialgleichungen angenommen, wobei A eine konstante und diagonalisierbare $n \times n$-Matrix sei. Die Eigenwerte von A seien

$$\{\lambda_1, \ldots, \lambda_n\} =: \sigma(A) \tag{3.1.2}$$

($\sigma(A)$ wird das *Spektrum* von A genannt). Nach Satz 1.4.1 ist die Lösung eine Linearkombination von $y_\nu e^{\lambda_\nu x}$ (y_ν Eigenvektor zum Eigenwert λ_ν)

$$y = \sum_{\nu=1}^{n} \alpha_\nu y_\nu e^{\lambda_\nu x}.$$

Das Wachstum der Komponenten wird durch den Realteil $\Re e(\lambda_\nu)$ beschrieben:

$\Re e(\lambda_\nu) \gg 1$	stark wachsend
$\lvert\Re e(\lambda_\nu)\rvert \lesssim 1$	stationär
$\Re e(\lambda_\nu) \ll -1$	stark fallend

Die Differentialgleichung heißt *steif* (englisch: stiff), wenn die Differenz $\max_{\nu,\mu} |\Re e(\lambda_\nu) - \Re e(\lambda_\mu)|$ groß wird. Im folgenden gehen wir von der realistischen Annahme aus, dass die Differentialgleichung so skaliert ist, dass $\Re e(\lambda_\nu) \sim 1$ für den größten Realteil gilt.

Definition 3.1.1 *Die Eigenwerte (3.1.2) seien nach der Größe des Realteils sortiert: $\Re e(\lambda_1) \geq \ldots \geq \Re e(\lambda_n)$. Die Differentialgleichung (3.1.1) heißt steif, falls gilt*

$$|\Re e(\lambda_1)| \sim 1, \tag{3.1.3a}$$

$$\Re e(\lambda_n) \ll -1\,. \tag{3.1.3b}$$

Beispiel 3.1.2 *Gegeben sei die Differentialgleichung*

$$y' = \begin{bmatrix} -50 & -51 \\ -51 & -50 \end{bmatrix} y\,.$$

Die Eigenwerte von A sind dann $\lambda_1 = 1$, $\lambda_2 = -101$. Die allgemeine Lösung ist dann

$$y(x) = \alpha e^{x} \begin{bmatrix} +1 \\ -1 \end{bmatrix} + \beta e^{-101x} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{mit } \alpha, \beta \in \mathbb{R}.$$

Wir stellen nun zwei Anfangswertaufgaben: Zunächst sei $y(0) = \begin{bmatrix} 10 \\ 9 \end{bmatrix}$, also $\alpha = 0.5$, $\beta = 9.5$ und $y_1(x) = \frac{1}{2}e^x + \frac{19}{2}e^{-101x}$ sowie $y_2(x) = -\frac{1}{2}e^x + \frac{19}{2}e^{-101x}$. In $[0, 0.0045]$ fällt der e^{-101x}-Term von 1 auf 0.01, in $[0.045, \infty)$ verhält sich y_1 somit ähnlich wie der glatte Anteil $0.5e^x$. Die numerische Lösung sollte daher wie folgt ablaufen:

- in $[0, 0.0045]$ kleine Schrittweiten (bis e^{-101x} abgeklungen ist),
- für $x \geq 0.045$ große Schrittweiten, da e^x leicht zu approximieren ist.

Bei der zweiten Anfangswertaufgabe sei der Anfangswert $y(0) = \begin{bmatrix} +1 \\ -1 \end{bmatrix}$, also $y(x) = e^x \begin{bmatrix} +1 \\ -1 \end{bmatrix}$. Die Abklingphase entfällt jetzt, und es sind sofort größere Schrittweiten möglich. Ein numerischer Versuch mit dem Euler- und dem klassischen Runge-Kutta-Verfahren und der Schrittweite $h = 1/10$ liefert aber für die erste Komponente y_1 bzw. η_1:

x	y_1	$\eta_{1,\text{Euler}}$	$\eta_{1,\text{RK4}}$
0	1	1	1
0.1	1.105	1.100	1.105
⋮	⋮	⋮	⋮
1	2.718	2.594	$3.1_{10}6$

(3.1.4a)

Für die verfälschten Anfangswerte

$$\tilde{y}_0 = \begin{bmatrix} 1.01 \\ -0.99 \end{bmatrix}, \quad \text{also } \tilde{y}_1(x) = e^x + 0.01e^{-101x} \approx e^x \tag{3.1.4b}$$

erhält man sogar

x	$\tilde{y}_1$	$\tilde{\eta}_{1,\text{Euler}}$	$\tilde{\eta}_{1,\text{RK4}}$
0	1.01	1.01	1.01
0.1	1.105	1.009	1.143
0.2	1.221	2.038	$9.24_{10}2$
0.3	1.350	-6.205	$2.8_{10}5$
⋮			
1	2.718	$3.89_{10}7$	Überlauf

(3.1.4c)

3.2 Ursache der numerischen Schwierigkeiten

Das *Euler-Verfahren* hat die Form $\eta_{i+1} = \eta_i + hA\eta_i = (I + hA)\eta_i$. Damit ist

$$\eta(x_k; h) = (I + hA)^k y_0$$

und für $y_0 = \sum_{\nu=1}^{n} \alpha_\nu y_\nu$ (y_ν Eigenvektoren von A) gilt

$$\eta(x_k; h) = \sum_{\nu=1}^{n} \alpha_\nu (1 + h\lambda_\nu)^k y_\nu \,.$$

Sei $x_0 := 0$. Für festes $x = x_k = kh$ und $h = \frac{x}{k} \to 0$ (d.h. $k = \frac{x}{h} \to \infty$) gilt

$$\lim_{k\to\infty} \left(1 + \frac{x}{k}\lambda_\nu\right)^k = e^{\lambda_\nu x} \,,$$

sodass $\eta(x_k; h)$ tatsächlich gegen die exakte Lösung konvergiert. Aber für

$$h > -\frac{2}{\Re e(\lambda_\nu)} \text{ und } \Re e(\lambda_\nu) < 0 \tag{3.2.1a}$$

folgt

$$\Re e\,(1 + h\lambda_\nu) < -1 \,, \tag{3.2.1b}$$

sodass $(1 + h\lambda_\nu)^k$ oszilliert und $|1 + h\lambda_\nu|^k$ exponentiell explodiert. Um dieses unerwünschte Verhalten auszuschließen, ist h klein genug zu wählen:

Folgerung 3.2.1 *Für die Schrittweite muss mindestens gelten, dass*

$$h < \frac{2}{|\Re e(\lambda_n)|}\,. \tag{3.2.1c}$$

Für das erste Anfangswertproblem in Beispiel 3.1.2 gilt $\alpha_1 = 1$ für den harmlosen Eigenwert $\lambda_1 = 1$ und $\alpha_2 = 0$ für $\lambda_2 = -101$. Wegen $\alpha_2 = 0$ sollte der Effekt des Faktors $|1 + h\lambda_\nu|^k = |1 - 10.1|^k = 9.1^k$ nicht sichtbar sein, aber aufgrund von Rundungsfehlern bleibt $\alpha_2 = 0$ nicht exakt erhalten. In (3.1.4c) ist $\alpha_2 = 0.01$. Der Wert $0.01 * 9.1^{10} = 3.894\,2_{10}7$ entspricht genau dem beobachteten Euler-Resultat.

Als nächstes untersuchen wir das *Runge-Kutta-Verfahren*: Für $y' = Ay$ (d.h. $f(x, y) = Ay$) ist

$$\begin{aligned}
&k_1 = Ay\,, && k_3 = A(y + \tfrac{h}{2}k_2) = Ay + \tfrac{h}{2}A^2y + \tfrac{h^4}{4}A^3y\,,\\
&k_2 = A(y + \tfrac{h}{2}k_1) = Ay + \tfrac{h}{2}A^2y\,, && k_4 = A(y + hk_3) = Ay + hA^2y + \tfrac{h^2}{2}A^3y + \tfrac{h^3}{4}A^4y,\\
&\phi = \frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4) = Ay + \frac{h}{2}A^2y + \frac{h^2}{6}A^3y + \frac{h^3}{24}A^4y,\\
&\eta_{\mu+1} = \eta_\mu + h\phi(x_\mu, \eta_\mu) = \left(I + hA + \frac{h^2}{2}A^2 + \frac{h^3}{6}A^3 + \frac{h^4}{24}A^4\right)\eta_\mu =: P_4(hA)\eta_\mu\,.
\end{aligned}$$

Zusammenfassend stellen wir fest, dass die Rekursion $\eta_{\mu+1} = P_4(hA)\eta_\mu$ gilt. Damit lautet die Runge-Kutta-Lösung

$$\eta_\mu = (P_4(hA))^\mu\, y_0 \qquad \text{mit dem Polynom } P_4(\xi) = \sum_{\nu=0}^{4} \frac{\xi^\nu}{\nu!}.$$

Sei nun A diagonalisierbar: $Y^{-1}AY = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. Dann wird auch $P_4(hA)$ durch Y diagonalisiert:

$$Y^{-1}P_4(hA)Y = \mathrm{diag}\{P_4(h\lambda_1), \ldots, P_4(h\lambda_n)\}\,.$$

Wie für jedes nichtkonstante Polynom gilt $\lim_{\xi\to-\infty} |P_4(\xi)| = \infty$. Also existiert $\xi_0 := \min\{\xi \in \mathbb{R} : |P_4(x)| \le 1$ für alle $x \in [\xi, 0]\}$. Für die Schrittweite h muss dann gefordert werden:

$$-h\,|\Re e(\lambda_n)| = h\Re e(\lambda_n) \ge \xi_0, \text{ also } h \le \left|\frac{\xi_0}{\Re e(\lambda_n)}\right|\,.$$

Präziser kann man nach allen komplexen $\zeta \in \mathbb{C}$ fragen, für die $|P_4(\zeta)| \le 1$ gilt. Dies definiert das in Abbildung 3.2.1 wiedergegebene *Stabilitätsgebiet*. Die Schrittweite h muss so gewählt werden, dass $h\lambda_\nu$ für alle Eigenwerte im Stabilitätsgebiet liegt.

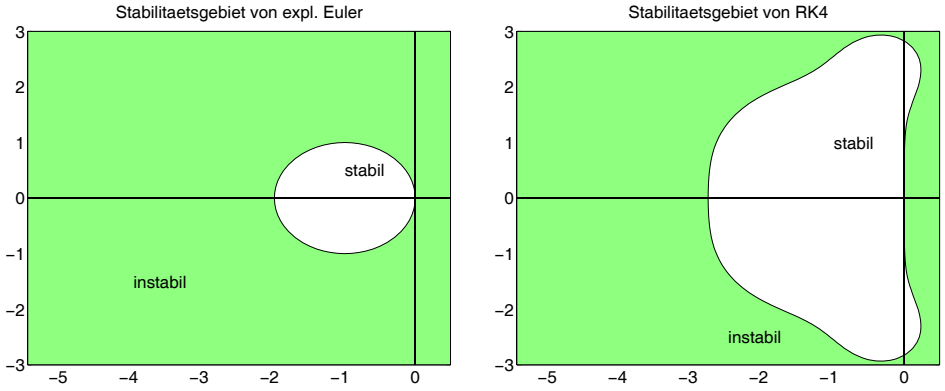


Abbildung 3.2.1: Stabilitätsgebiete

Das *implizite Euler-Verfahren* zeigt ein anderes Verhalten. Hier ist

$$\eta(x+h;h) = \eta(x;h) + hA\eta(x+h;h), \quad \text{d.h. } \eta(x+h;h) = (I-hA)^{-1}\eta(x;h)\,.$$

Für die Anfangswertaufgabe (3.1.4b) und $h = 1/10$ erhält man

x	$\tilde{y}_1(x)$	$\tilde{\eta}_{1,\text{impl.Euler}}$
0	1.01	1.01
⋮	⋮	⋮
1	2.718	2.868

. Der Grund für dieses Verhalten folgt aus $\tilde{\eta}_{\text{impl.Euler}}(x_k,h) = (I-hA)^{-k}\,\tilde{y}_0$. Die erste Komponente $\tilde{\eta}_{1,\text{impl.Euler}}(x_k,h)$ aus der obigen Tabelle hat damit die folgende Darstellung:

$$\tilde{\eta}_{1,\text{impl.Euler}}(x_k,h) = \frac{1}{(1-h)^k} + \frac{1}{100}\frac{1}{(1+101h)^k} \underset{h=\frac{1}{10}}{=} \left(\frac{10}{9}\right)^k + \frac{1}{100}\left(\frac{10}{111}\right)^k.$$

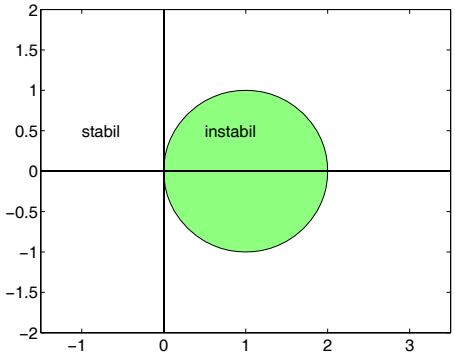


Abbildung 3.2.2: Stabilitätsgebiet für das implizite Euler-Verfahren

3.3 Stabilitätsbedingungen

Wir betrachten das Modellproblem (Testgleichung)

$$y' = Ay, \qquad \text{wobei } A \text{ eine konstante } n \times n\text{-Matrix ist.}$$

Wir setzen voraus, dass das (implizite) Einschrittverfahren $\eta(x+h;h) = \eta(x;h) + h\phi(\eta(x;h),\eta(x+h;h);h,f)$ bei Anwendung auf die Testgleichung die Form

$$\eta(x+h;h) = R(hA)\eta(x;h) \tag{3.3.1}$$

annimmt, wobei $R(\xi)$ eine rationale Funktion ist (für alle bisher beschriebenen Einschrittverfahren ist dies der Fall).

Beispiel 3.3.1 *Die rationale Funktion $R(\xi)$ lautet für spezielle Verfahren wie folgt:*

Euler-Verfahren (2.1.1'): $$R(\xi) = 1+\xi, \tag{3.3.2a}$$

Implizites Euler-Verfahren (2.2.3b): $$R(\xi) = \frac{1}{1-\xi}, \tag{3.3.2b}$$

Trapezformel (2.2.3c): $$R(\xi) = \frac{1+\xi/2}{1-\xi/2}, \tag{3.3.2c}$$

Klassisches Runge-Kutta-Verfahren (2.3.1a-e): $$R(\xi) = 1+\xi+\frac{\xi^2}{2!}+\frac{\xi^3}{3!}+\frac{\xi^4}{4!} \tag{3.3.2d}$$

Übungsaufgabe 3.3.2 *Für das implizite Runge-Kutta-Gauß-Verfahren ($m = 2$, vgl. (2.3.4b)) ermittle man $R(\xi)$.*

Lemma 3.3.3 *Das Einschrittverfahren habe die Konsistenzordnung p und erfülle (3.3.1) für $f(y) = Ay$. Dann gilt*

$$R(\xi) = e^{\xi} + \mathcal{O}(\xi^{p+1}) \quad \textit{bzgl. } \xi \to 0. \tag{3.3.3}$$

Beweis. Es sei $y' = Ay$ und x fest mit $y(x) = \eta(x;h)$. Nach Satz 1.4.1 ist $y(x+h) = e^{hA}y(x)$. Damit gilt für den lokalen Diskretisierungsfehler

$$\mathcal{O}(h^{p+1}) = \eta(x+h;h) - y(x+h) = R(hA)\eta(x) - y(x+h) = \left[R(hA) - e^{hA}\right] y(x)\,.$$

Man wähle den Anfangswert $y(x)$ als einen Eigenvektor von A zum Eigenwert λ_ν:

$$(R(h\lambda_\nu) - e^{h\lambda_\nu})y(x) = \mathcal{O}(h^{p+1}).$$

Ersetzen von $h\lambda_\nu$ durch ξ liefert die Aussage. ∎

Im Beweis wurde die folgende Eigenschaft verwendet:

Lemma 3.3.4 *a) Ist λ Eigenwert von A, so ist $R(h\lambda)$ Eigenwert von $R(hA)$.*
b) Ist e Eigenvektor von A zum Eigenwert λ, so ist e auch Eigenvektor von $R(hA)$ zum Eigenwert $R(h\lambda)$.

Aus (3.3.1) folgt das

Lemma 3.3.5 *Die Lösung des Einschrittverfahrens auf $y' = Ay$ mit konstanter $n \times n$-Matrix A lautet*

$$\eta(x_k;h) = R(hA)^k \eta(x_0;h)\,. \tag{3.3.4}$$

Es sei $\eta(x_0) = \sum_{\nu=1}^n \alpha_\nu e_\nu$ mit $Ae_\nu = \lambda_\nu e_\nu$ (hier wurde die Diagonalisierbarkeit von A angenommen). Die Näherung

$$\eta(x_k;h) \underset{(3.3.4)}{=} \sum_{\nu=1}^n \alpha_\nu (R(h\lambda_\nu))^k e_\nu$$

soll

$$y(x_k) = \sum_{\nu=1}^n \alpha_\nu e^{h\lambda_\nu k} e_\nu$$

approximieren. Falls $\Re e(\lambda_\nu) \ll -1$, ist $e^{\lambda_\nu hk}$ sehr klein für $x = hk \geq \mathcal{O}(1)$. Hieraus schließen wir: Auch $(R(h\lambda_\nu))^k$ muss (sehr) klein sein, zumindest aber beschränkt. Diese Überlegung führt zu folgender Definition.

Definition 3.3.6 *R gehöre zum Verfahren ϕ. Dann heißt ϕ* absolut stabil *genau dann, wenn*

$$|R(z)| < 1 \qquad \textit{für } z \in \mathbb{C} \textit{ mit } \Re e(z) < 0. \tag{3.3.5}$$

Eine einfache Folgerung ist, dass $\Re e(\lambda_\nu) < 0$ (für alle ν) für beliebiges $h > 0$ die Ungleichung $|R(h\lambda_\nu)| < 1$ impliziert.

Explizite Verfahren liefern im allgemeinen ein Polynom $R(z)$, insbesondere ergeben explizite Runge-Kutta-Verfahren stets Polynome. Hierzu ist ein negatives Resultat festzuhalten:

Lemma 3.3.7 *Falls R ein Polynom ist, ist ϕ nicht absolut stabil.*

Beweis. 1) Falls $\operatorname{grad} R \leq 0$, d.h. $R(z) = a_0$, folgt aus (3.3.3), dass $R(z) = 1$, sodass (3.3.5) nicht zutrifft.
2) Falls $\operatorname{grad} R \geq 1$, folgt $|R(z)| \to \infty$ für $z \to -\infty$ im Widerspruch zu (3.3.5). ∎

Damit bleiben die impliziten Verfahren als Kandidaten für absolut stabile Verfahren übrig. Das folgende Lemma gibt hinreichende und notwendige Bedingungen an.

Lemma 3.3.8 *Es sei $R(z)$ rational und erfülle die Konsistenzbedingung $R(0) = 1$ (vgl. (3.3.3)). Dann ist ϕ genau dann absolut stabil, wenn gilt*

$$R(z) \neq \textit{const} \tag{3.3.6a}$$

$$R(z) \textit{ ist holomorph für } z \textit{ mit } \Re e(z) < 0 \tag{3.3.6b}$$

$$|R(z)| \leq 1 \qquad \textit{für } \Re e(z) = 0 \tag{3.3.6c}$$

Beweis. 1) "absolut stabil $\Rightarrow$ (3.3.6a-c)":
ad (3.3.6a): Wäre $R(z) = const$, müsste nach Voraussetzung $R(0) = 1$ auch $R(z) = 1$ im Widerspruch zu (3.3.5) gelten.
ad (3.3.6b): Wegen (3.3.5) ist R in $\{\zeta \in \mathbb{C} : \Re e(\zeta) < 0\}$ beschränkt, kann also dort keine Polstellen besitzen, was (3.3.6b) beweist.
ad (3.3.6c): Aus (3.3.5) folgt (3.3.6c) durch Grenzübergang.

2) "absolut stabil $\Leftarrow$ (3.3.6a-c)": Die Transformation $\zeta = \frac{1+z}{1-z}$ bildet die linke Halbebene $z \in \{z' \in \mathbb{C} : \Re e(z') < 0\}$ auf den Kreis $\zeta \in \{\zeta' \in \mathbb{C} : |\zeta'| < 1\}$ ab, wobei

$$\begin{array}{lcl} \Re e(z) = 0 & \Rightarrow & |\zeta| = 1, \\ z = 0 & \Rightarrow & \zeta = 1, \\ z = -\infty & \Rightarrow & \zeta = -1. \end{array}$$

Die Transformation definiert die neue Funktion $\hat{R}(\zeta) = R\left(\frac{1-\zeta}{1+\zeta}\right) = R(z)$, die offenbar wieder rational ist. Außerdem hat $\hat{R}$ die Eigenschaften

$$\hat{R} \text{ ist holomorph für } |\zeta| < 1, \qquad |\hat{R}| \le 1 \quad \text{für } |\zeta| = 1.$$

Nach dem Maximumprinzip für holomorphe Funktionen gilt

$$|\hat{R}| \le 1 \qquad \text{für } |\zeta| \le 1;$$

genauer sind zwei Alternativen möglich: Entweder ist $\hat{R}$ konstant ($|\hat{R}| = 1$) oder $|\hat{R}| < 1$ für alle $|\zeta| < 1$. Der erste Fall impliziert $R = const$ im Widerspruch zu (3.3.6a). Also gilt der zweite Fall und beweist (3.3.5). ∎

Man möchte nun Verfahren finden mit den folgenden Eigenschaften:

$$\begin{array}{ll} R(z) = e^z + \mathcal{O}(z^{p+1}) & \text{(lokale Eigenschaft)} \\ |R(z)| < 1 \quad \text{für alle } z \in \mathbb{C} \text{ mit } \Re e(z) < 0 & \text{(globale Eigenschaft)} \end{array} \tag{3.3.7}$$

Welche rationalen Funktionen sind hierfür geeignet?

Definition 3.3.9 *Sei $g(z)$ in $z = 0$ holomorph. P_{jk} und Q_{jk} seien Polynome vom Grad $\le k$ beziehungsweise $\le j$. Wenn $g(z)Q_{jk}(z) - P_{jk}(z) = \mathcal{O}(|z|^{j+k+1})$ für $z \to 0$ gilt, heißt $R_{jk} := \frac{P_{jk}}{Q_{jk}}$ die Padé-Approximation[21] von g zum Index (j, k).*

Die Padé-Approximation zum Index $(0, k)$ ist das Taylor-Polynom vom Grad k.

Wegen der ersten Eigenschaft in (3.3.7) interessiert insbesondere die Approximation von $g(z) = e^z$. Das folgende Resultat findet man bei Gautschi [6, Theorem 5.5.1]:

Satz 3.3.10 *Die Padé-Approximation von e^x zum Index (j, k) lautet $R_{jk} = \dfrac{P_{jk}}{Q_{jk}}$ mit*

$$P_{jk}(z) = \sum_{\ell=0}^{k} \frac{k!}{(k-\ell)!} \frac{(k+j-\ell)!}{(k+j)!} \frac{z^\ell}{\ell!}, \qquad Q_{jk}(z) = P_{kj}(-z). \tag{3.3.8}$$

Padé-Tabelle der Approximation R_{jk} von e^z :

$j \diagdown k$	0	1	2
0	1	$1 + x$	$1 + x + \frac{x^2}{2}$
1	$\dfrac{1}{1-x}$	$\dfrac{1+\frac{x}{2}}{1-\frac{x}{2}}$	$\dfrac{1+\frac{2}{3}x+\frac{x^2}{6}}{1-\frac{x}{3}}$
2	$\dfrac{1}{1-x+\frac{x^2}{2}}$	$\dfrac{1+\frac{x}{3}}{1-\frac{2}{3}x+\frac{x^2}{6}}$	$\dfrac{1+\frac{x}{2}+\frac{x^2}{12}}{1-\frac{x}{2}+\frac{x^2}{12}}$

[21] Henri Eugène Padé, geb. 17. Dez. 1863 in Abbeville, Picardy, Frankreich, gest. 9. Juli 1953 in Aix-en-Provence

Lemma 3.3.11 *$Q_{jj}(z)$ aus (3.3.8) hat nur Nullstellen in $\{z \in \mathbb{C} : \Re e(z) > 0\}$.*

Beweis. Anwendung des Routh-Hurwitz-Kriteriums (vgl. [6, p. 306]). ∎

Satz 3.3.12 *Verfahren mit $R(z) = R_{jj}$ $(j \geq 1)$ (diagonale Padé-Gruppe) und $R(z) = R_{j+1,j}$ $(j \geq 0)$ (subdiagonale Padé-Gruppe) sind absolut stabil.*

Beweis. a) Wegen Lemma 3.3.11 liegen sämtliche Polstellen von R in der rechten Halbebene, d.h. R ist in der linken Halbebene holomorph. Da R nicht konstant ist, bleibt nur noch (3.3.6c) nachzuweisen. Für rein imaginäres z gilt $\bar{z} = -z$. Mit (3.3.8) folgt $|Q_{jk}(z)| = |P_{kj}(-z)| = |P_{kj}(\bar{z})| = \left|\overline{P_{kj}(z)}\right| = |P_{kj}(z)|$, also $|R(z)| = |P_{jj}(z)| \,/\, |Q_{jj}(z)| = 1$. Somit folgt die absolute Stabilität aus Lemma 3.3.8.

b) Für den Fall $R_{j+1,j}$ vergleiche man [6, pp. 306-308]. ∎

Satz 3.3.13 *Implizite Runge-Kutta-Verfahren vom Gauß-Typ (m-stufig) führen zu $R(z) = \frac{P_{mm}(z)}{Q_{mm}(z)}$ und sind daher absolut stabil.*

Der Stabilitätsbegriff kann noch weiter verfeinert werden. Bisher ist $|R(z)| \approx 1$ für $\Re e(z) < 0$ zugelassen, insbesondere darf der Limes $|z| \to \infty$ für z mit $\Re e(z) < 0$ zu $\lim |R(z)| = 1$ führen. $|R(h\lambda_\nu)| \approx 1$ führt dazu, dass die entsprechende Komponente in der Größenordnung unverändert bleibt, obwohl $\exp(\lambda_\nu x)$ gegen null strebt. Daher muss man wie folgt vorgehen:

1. Kleine Schrittweiten wählen, bis die Komponenten, die zu λ_ν mit $\Re e(\lambda_\nu) \ll \Re e(\lambda_1)$ gehören, abgeklungen sind.
2. Danach dürfen wieder größere Schrittweiten gewählt werden, da die in 1) erwähnten Komponenten nicht mehr verstärkt werden.

Eine günstigere Situation (Vermeidung der Phase 1) entsteht für Einschrittverfahren der nachfolgenden Klasse.

Definition 3.3.14 *Ein Einschrittverfahren heißt* stark absolut stabil*, wenn es absolut stabil ist und außerdem $R(z) \to 0$ für $\Re e(z) \to -\infty$ gilt.*

Offenbar erfüllt eine rationale Funktion $R = \frac{P}{Q}$ die Bedingung $\lim_{z\to-\infty} R(z) = 0$ genau dann, wenn $grad(Q) > grad(P)$ ist. Daher gilt:

Folgerung 3.3.15 *a) Die diagonalen Padé-Approximationen sind nicht stark absolut stabil.*
b) Padé-Approximationen zum Index (j,k), $j < k$, erfüllen die Bedingung $\lim_{z\to-\infty} R(z) = 0$.

Satz 3.3.16 *Verfahren mit der subdiagonalen Padé-Approximation $R = R_{j+1,j}$ sind stark absolut stabil. Speziell gilt für $j = 0$: Die implizite Euler-Formel ist stark absolut stabil.*

Bisher wurde nur das Modellproblem (3.1.1) untersucht. Im Falle einer allgemeinen (nichtlinearen) Differentialgleichung $y' = f(x, y)$ hat man die Ableitung ($n \times n$-Matrix)

$$A(x, y) := \frac{\partial f(x, y)}{\partial y}$$

anstelle der in (3.1.1) als konstant angenommenen Matrix zu untersuchen. Wenn die Eigenwerte von $A(x, y)$ die Eigenschaft (3.1.3a,b) besitzen, hat man absolut stabile Verfahren anzuwenden. Wegen der Abhängigkeit von x, y kann es durchaus passieren, dass sich eine nichtlineare Differentialgleichung nur für einen Teilbereich $I_0 \subset I$ steif verhält.

Bei der Lösung eines Differentialgleichungssystems muss man entweder *a-priori*-Information über ihre eventuelle Steifheit haben und entsprechende absolut stabile Verfahren einsetzen. Oder man zusätzlich zur Schrittweitensteuerung (zumindest ab und zu) an $A(x, y)$ testen, ob das System steif ist. Im positiven Fall wähle man ein (stark) absolut stabiles Verfahren. Andernfalls wähle man ein "billigeres" explizites Verfahren.

4 Mehrschrittverfahren

4.1 Allgemeine Mehrschrittverfahren

Hier sei h wieder eine *konstante* Schrittweite. Näherungen $\eta_j = \eta(x_0 + jh; h)$ seien bekannt (das heißt für die Implementierung: abgespeichert) für $0 \le j \le r-1$, wobei $r \ge 1$ die Schrittzahl ist. Dabei ist $r = 1$ uninteressant, da dann wieder Einschrittverfahren entstehen.

Zur Berechnung von η_r kann man die r vorhergehenden η-Werte benutzen:

$$\eta_r := -\sum_{j=0}^{r-1} a_j \eta_j + h\phi(x_0, \eta_r, \eta_{r-1}, \ldots, \eta_0; h, f)$$

und allgemein das Mehrschrittverfahren

$$\eta_{j+r} := -\sum_{\nu=0}^{r-1} a_\nu \eta_{j+\nu} + h\phi(x_j, \eta_{j+r}, \eta_{j+r-1}, \ldots, \eta_j; h, f) \tag{4.1.1}$$

aufstellen. Falls ϕ das Argument η_{j+r} besitzt, ist das Verfahren implizit. Das Mehrschrittverfahren (4.1.1) wird genauer r-Schrittverfahren genannt.

Definition 4.1.1 *Ein* lineares *r-Schrittverfahren liegt vor, falls gilt*

$$\phi(x_j, \eta_{j+r}, \eta_{j+r-1}, \ldots, \eta_j; h; f) = \sum_{\mu=0}^{r} b_\mu \underbrace{f(x_{j+\mu}, \eta_{j+\mu})}_{=:f_{j+\mu}}. \tag{4.1.2}$$

Definition 4.1.2 *Für $f \in C^1$ sei z die Lösung der Anfangswertaufgabe $z'(t) = f(t, z(t))$, $z(x) = y$. Der* lokale Diskretisierungsfehler *ist dann*

$$\tau(x, y; h) := \frac{1}{h}\left[\sum_{\nu=0}^{r} a_\nu z(x + \nu h) - h\phi(x, z(x + rh), \ldots, z(x); h, f)\right],$$

wobei $a_r := 1$ ist, während die anderen a_ν aus (4.1.1) stammen.

Definition 4.1.3 *Seien $\phi \in C^0$, $f \in C^1$. Ein Mehrschrittverfahren heißt* konsistent*, wenn*

$$\lim_{h \to 0} \tau(x, y; h) = 0.$$

Das Verfahren hat die Konsistenzordnung *$p > 0$, falls $\tau(x, y; h) = \mathcal{O}(h^p)$ für alle $f \in C^p$ und $h \to 0$ gilt.*

Start eines Mehrschrittverfahrens: Neben $y_0 = y(x_0) = \eta_0$ werden zusätzlich $\eta_1, \ldots, \eta_{r-1}$ benötigt.

Es gilt zu beachten, dass $\eta_1, \ldots, \eta_{r-1}$ im Allgemeinen nicht fehlerfrei sind. Wir machen daher den folgenden Ansatz: Seien

$$\eta_j = y(x_j) + \varepsilon_j \qquad \text{für } 0 \le j \le r-1 \tag{4.1.3}$$

Startwerte mit Fehlern ε_j. Für $j \ge 0$ wende man das mit $h\varepsilon_{j+r}$ gestörte Mehrschrittverfahren an:

$$\sum_{\nu=0}^{r} a_\nu \eta_{j+\nu} = h\phi(x_j, \eta_{j+r}, \eta_{j+r-1}, \ldots, \eta_j; h; f) + h\varepsilon_{j+r}\,, \tag{4.1.4}$$

wobei $a_r = 1$ wie in Definition 4.1.2. Die Lösung bezeichne man mit $\eta(x; \vec{\varepsilon}, h)$, wobei $\vec{\varepsilon} = (\varepsilon_i)_{i=0}^{j+r}$.

Für die Konvergenzdefinition sind zwei Versionen möglich. Die schwächere Form ist wiedergegeben in

Definition 4.1.4 *Sei $\vec{\varepsilon} = (\varepsilon_i)_{i=0}^{r-1}$ (d.h. $\varepsilon_i = 0$ für $i \ge r$). Ein Mehrschrittverfahren (4.1.1) heißt konvergent, falls für alle $f \in C^1$ gilt, dass $\eta(x, \vec{\varepsilon}, h) \to y(x)$ für $h \to 0$, $\sup_j \|\varepsilon_j\| \to 0$.*

Die stärke Version der Konvergenz verlangt $\eta(x, \vec{\varepsilon}, h) \to y(x)$ für Störungen $\vec{\varepsilon} = (\varepsilon_i)_{i=0}^{j+r}$ auch in (4.1.4).

4.2 Beispiele

4.2.1 Adams-Bashforth-Verfahren

In

$$y(x_{j+1}) = y(x_j) + \int_{x_j}^{x_{j+1}} f(\xi, y(\xi))\, d\xi\ .$$

ist $\Psi(\xi) := f(\xi, y(\xi))$ der Integrand. Man benutze eine Quadratur für $\int_{x_j}^{x_{j+1}} \Psi(\xi)\, d\xi$ mit den Stützstellen x_j, $x_{j-1}, \ldots, x_{j-q}$ ($q \geq 1$ vorgegeben), wobei $x_{j-1}, \ldots, x_{j-q}$ außerhalb des Integrationsbereichs liegen. Für das Interpolationspolynom[22] p_q gilt $grad(p_q) \leq q$ und $p_q(x_\nu) = \phi(x_\nu)$ für $j - q \leq \nu \leq j$. Damit erhält man folgende Quadraturformel:

$$I_{x_j}^{x_{j+1}}(\phi) := \int_{x_j}^{x_{j+1}} p_q(\xi)\, d\xi = h \sum_{\mu=0}^{q} \beta_{q\mu} \Psi(x_{j-\mu})\,, \quad \text{wobei } \beta_{q\mu} = \int_0^1 \prod_{\ell \in \{0,\ldots,q\}\setminus\{\mu\}} \frac{\ell - \xi}{\ell - \mu}\, d\xi$$

(der letzte Integrand ist das μ-te Lagrange-Polynom[23] zu den skalierten Stützstellen $0, -1, \ldots, -q$; vgl. [15, Abschnitt 2.1.1]). Setzt man nun die Quadratur in die obige Gleichung für $y(x_{j+1})$ ein, so erhält man mit $f_j := f(x_j, \eta_j)$ das Adams-Bashforth-Verfahren:[24]

$$\eta_{j+1} = \eta_j + h \sum_{\mu=0}^{q} \beta_{q\mu} f_{j-\mu}\,. \tag{4.2.1a}$$

Bemerkung 4.2.1 *(4.2.1a) ist ein lineares, explizites $(q+1)$-Schrittverfahren mit*

$$a_r = 1,\ a_{r-1} = -1,\ a_{r-2} = \ldots = a_0 = 0,\ \ r = q+1, \tag{4.2.1b}$$

$$b_r = 0,\ b_\mu = \beta_{q,r-1-\mu}\ \ (0 \leq \mu \leq q)\,. \tag{4.2.1c}$$

Der Quadraturfehler ist

$$I_{x_j}^{x_{j+1}}(\Psi) - \int_{x_j}^{x_{j+1}} \Psi(\xi)\, d\xi = \mathcal{O}(h^{q+2}),$$

also ist $\tau = \mathcal{O}(h^{q+1})$. Die Konsistenzordnung des Verfahrens ist demnach $p = q + 1 = r$.

Beispiel 4.2.2 *Für $q = 0$, also $r = 1$, ist $\beta_{00} = 1$ und es liegt das Euler-Verfahren vor.*
Für $q = 1$, also $r = 2$, lauten die Koeffizienten $\beta_{10} = 3/2$, $\beta_{11} = -1/2$ und definieren das 2-Schrittverfahren

$$\eta_{j+2} = \eta_{j+1} + \frac{h}{2}(3 f_{j+1} - f_j).$$

Der entscheidende Vorteil der Mehrschrittverfahren (4.2.1a) liegt in der Tatsache, dass die $f_{j-\mu}$ für $\mu > 0$ schon aus früheren Rechnungen bekannt sind und nicht neu ausgewertet werden müssen:

Bemerkung 4.2.3 *Der Aufwand von (4.2.1a) ist im wesentlichen eine f-Auswertung pro Schritt.*

4.2.2 Adams-Moulton-Verfahren

Die Voraussetzungen sind wie in Abschnitt 4.2.1, die Stützstellen der Quadratur sind aber nun x_{j+1}, $x_j, \ldots$, x_{j-q+1}:

$$\eta_{j+1} = \eta_j + h \sum_{\mu=0}^{q} \beta_{q\mu}\, f_{j+1-\mu} \qquad \text{mit } \beta_{q\mu} = \int_0^1 \prod_{\ell \in \{0,\ldots,q\}\setminus\{\mu\}} \frac{\ell - 1 + \xi}{\ell - \mu}\, d\xi. \tag{4.2.2}$$

(4.2.2) ist ein implizites lineares r-Schrittverfahren der Konsistenzordnung $q + 1$, wobei $r = \max(1, q)$.

[22] treffender wäre der Name "Extrapolationspolynom"
[23] Joseph-Louis Lagrange, geb. 25. Jan. 1736 in Turin, gest. 10. April 1813 in Paris
[24] John Couch Adams, geb. 5. Juni 1819 in Laneast, gest. 21. Jan. 1892 in Cambridge
Francis Bashforth, geb. 1819 in Thurnscoe (nahe Doncaster), gest. 1912 in Woolhall Spa, Lincolnshire

Beispiel 4.2.4 *a) Für $q = 0$ ist $\beta_{00} = 1$ und man erhält das implizite Euler-Verfahren.*
b) Für $q = 1$ sind $\beta_{10} = \beta_{11} = 0,5$ und man erhält das Trapezverfahren mit $\eta_{j+1} = \eta_j + \frac{h}{2}(f_{j+1} + f_j)$.
c) Ein echtes Mehrschrittverfahren liegt erst für $q > 1$ vor. Häufig wird ein Schritt von Adams-Moulton als Korrektor verwendet, wobei Adams-Bashforth als Prädiktor dient.

4.2.3 Mittelpunktsregel

Hier betrachte man

$$y(x_{j+r}) = y(x_j) + \int_{x_j}^{x_{j+r}} f(\xi, y(\xi))\, d\xi \ .$$

Mögliche Stützstellen der Quadratur sind x_j, $x_{j+1}, \ldots, x_{j+r}$ und man erhält ein Verfahren der Form

$$\eta_{j+r} = \eta_j + h \sum_{\mu=0}^{r} \beta_{r\mu} f_{j+\mu} \ ,$$

wobei der μ-te Summand entfällt, wenn $x_{j+\mu}$ nicht als Quadraturstützstelle verwendet wird. Das Verfahren ist explizit (implizit), falls $\beta_{rr} = 0$ ist ($\beta_{rr} \neq 0$).

Ein Spezialfall ist die Mittelpunktsregel (midpoint rule) mit $r = 2$ und x_{j+1} als einziger Stützstelle. Dies liefert die Koeffizienten $\beta_{20} = \beta_{22} = 0$, $\beta_{21} = 2$:

$$\eta_{j+2} = \eta_j + 2hf(x_{j+1}, \eta_{j+1})\,. \tag{4.2.3}$$

Dies ist ein explizites Zweischrittverfahren der Konsistenzordnung 2. Zum Aufwand gilt wieder die Aussage der Bemerkung 4.2.3.

4.2.4 BDF-Verfahren

Das Kürzel BDF steht für "backward differentiation formulae". Bei der Herleitung geht man nicht von der Integralformulierung (1.2.2) aus, sondern wendet die Idee der "Kollokation" auf die Differentialgleichung an. Aus der Quadratur wird eine Interpolation: Das Interpolationspolynom durch $(x_{j+\nu}, \eta_{j+\nu})\,,\ 0 \le \nu \le r$, lautet in der Lagrange-Formulierung

$$P(x; \eta_{j+r}) = \sum_{\nu=0}^{r} \eta_{j+\nu} L_{j+\nu}(x).$$

Man beachte, dass die Werte $\eta_{j+\nu}$ für $\nu < r$ fixiert sind, während η_{j+r} eine freie Variable ist, die noch zu bestimmen ist. Da die Interpolierende die Funktion $y(x)$ annähern soll, liegt es nahe, die Differentialgleichung im Punkt x_{j+r} zur Bestimmung von η_{j+r} zu verwenden:[25]

$$P'(x_{j+r}; \eta_{j+r}) = f(x_{j+r}, \eta_{j+r}). \tag{4.2.4}$$

Mit den Größen $\alpha_\nu := hL'_{j+\nu}(x_{j+r})$ erhält man hieraus das implizite r-Schrittverfahren

$$\sum_{\nu=0}^{r} \alpha_\nu \eta_{j+\nu} = hf(x_{j+r}, \eta_{j+r}). \tag{4.2.5}$$

Man beachte, dass hier die Normierung $\alpha_r = 1$ nicht zutrifft. Hierzu könnte man (4.2.5) durch α_r teilen. Wie die Notation $\alpha_\nu = hL'_{j+\nu}(x_{j+r})$ suggeriert, hängt α_ν nicht von j ab, wenn - wie üblich - äquidistante Gitterpunkte $x_j = x_0 + jh$ vorliegen. Aber im Prinzip sind auch nichtäquidistante x_j möglich, wenn man in (4.2.5) $\alpha_\nu = \alpha_{\nu,j} = hL'_{j+\nu}(x_{j+r})$ für jedes j neu auswertet.

Obwohl $P'(x_{j+r}; \eta_{j+r})$ nach einer exakten Ableitung aussieht, stellt $P'(x_{j+r}; \eta_{j+r}) = \frac{1}{h}\sum_{\nu=0}^{r} \alpha_\nu \eta_{j+\nu}$ nur eine numerische Differenzenformel mit Hilfe der $\{(x_{j+\nu}, \eta_{j+\nu}) : 0 \le \nu \le r\}$ dar.

[25] Geht man von $P'(x; \eta_{j+r}) \approx f(x, P(x; \eta_{j+r}))$ aus, gibt es mehrere Möglichkeiten, zu einer Gleichung für η_{j+r} zu gelangen. Erzwingt man $P'(\hat{x}; \eta_{j+r}) = f(\hat{x}, P(\hat{x}; \eta_{j+r}))$ in einem Punkt $\hat{x}$, spricht man von "Kollokation im Punkte $\hat{x}$". Für die Wahl $\hat{x} = x_{j+r}$ gilt $P(x_{j+r}; \eta_{j+r}) = \eta_{j+r}$ und ergibt (4.2.4). Eine Alternative ist eine gewichtete Gleichung der Form $\int \chi(x)\,(P'(x; \eta_{j+r}) - f(x, P(x; \eta_{j+r})))\, dx = 0$ bzw. $\sum_\mu \chi_\mu \left(P'(x^{(\mu)}; \eta_{j+r}) - f(x^{(\mu)}, P(x^{(\mu)}; \eta_{j+r}))\right)$.

Übungsaufgabe 4.2.5 *Man prüfe nach, dass die Koeffizienten α_ν für $r \in \{1, \ldots, 6\}$ wie folgt lauten:*

	α_0	α_1	α_2	α_3	α_4	α_5	α_6
$r=1$	-1	1		*implizites Euler-Verfahren*			
$r=2$	$\frac{1}{2}$	-2	$\frac{3}{2}$				
$r=3$	$-\frac{1}{3}$	$\frac{3}{2}$	-3	$\frac{11}{6}$			
$r=4$	$\frac{1}{4}$	$-\frac{4}{3}$	3	-4	$\frac{25}{12}$		
$r=5$	$-\frac{1}{5}$	$\frac{4}{5}$	$-\frac{10}{3}$	5	-5	$\frac{137}{60}$	
$r=6$	$\frac{1}{6}$	$-\frac{6}{5}$	$\frac{15}{4}$	$-\frac{20}{3}$	$\frac{15}{2}$	-6	$\frac{49}{20}$

4.2.5 Starten eines Mehrschrittverfahrens

Zum Start eines r-Schrittverfahrens braucht man neben $\eta_0 = y_0$ auch die weiteren Werte $\eta_1, \ldots, \eta_{r-1}$. Eine mögliche **Konstruktion** sieht so aus:

- Mit dem Einschrittverfahren ϕ_1 erhält man η_1 aus $\eta_0 = y_0$.
- Mit einem Ein - oder Zweischrittverfahren erhält man η_2 aus η_1 bzw. η_0 und η_1.
- $\vdots$
- Mit einem Ein-,$\ldots$,$(r-2)$- oder $(r-1)$-Schrittverfahren erhält man η_{r-1}.

Die Ordnung dieser Verfahren darf $p-1$ sein, wenn insgesamt die Ordnung p erreicht werden soll. Ein Beispiel dafür ist die Mittelpunktsregel (4.2.3) mit Konsistenzordnung 2, bei dem der Start mit dem Euler-Verfahren (Konsistenzordnung 1) gemacht wird.

Mehrschrittverfahren sind im Allgemeinen an eine konstante Schrittweite gebunden. Damit wird ein Neustart auch dann notwendig, wenn man die Schrittweite wechseln will. Eine Ausnahme stellen die BDF-Verfahren dar (vgl. §4.2.4 und [3, Abschnitt 7.4.2]).

4.2.6 Gegenbeispiele zur Konvergenz

Die Vermutung, dass aus Konsistenz Konvergenz folgt, ist falsch!

Beispiel 4.2.6 *Ein allgemeines explizites Zweischrittverfahren ist*

$$\eta_{j+2} + a_1\eta_{j+1} + a_0\eta_j = h(b_1 f_{j+1} + b_0 f_j) \; .$$

Die vier freien Parameter können so gewählt werden, dass die Konsistenzordnung maximal wird. Dazu führt man eine Taylor-Entwicklung für den Diskretisierungsfehler durch:

$$\tau = \alpha_{-1}h^{-1} + \alpha_0 h^0 + \alpha_1 h^1 + \alpha_2 h^2 + \mathcal{O}(h^3) \; .$$

Die Bedingungen $\alpha_{-1} = \alpha_0 = \alpha_1 = \alpha_2 = 0$ sind erfüllt für $a_1 = 4$, $a_0 = -5$, $b_1 = 4$, $b_0 = 2$. Damit erhält man das Verfahren:

$$\eta_{j+2} + 4\eta_{j+1} - 5\eta_j = h(4f_{j+1} + 2f_j) \; . \tag{4.2.6}$$

Die Konsistenzordnung von (4.2.6) ist drei.

Nun wende man das Verfahren auf die Anfangswertaufgabe $y' = -y$, $y(0) = 1$ an mit $\eta_1 := e^{-h}$, $h = 0.01$. Da die exakte Lösung $y(x) = e^{-x}$ ist, ist η_1 fehlerfrei. Das Verfahren liefert:

j	x_j	$\eta_j - y(x_j)$
0	0.00	0
1	0.01	0
2	0.02	-0.16_{10}-8
3	0.03	$+0.50_{10}$-8
4	0.04	-0.30_{10}-7
5	0.05	$+0.14_{10}$-6
$\vdots$	$\vdots$	$\vdots$
99	0.99	$+0.13_{10}60$
100	1.00	$-0.65_{10}60$

Es ist also $\eta(1,h) = -0.65_{10}60$ statt $y(1) = \dfrac{1}{e} \approx 0.37$.

Beispiel 4.2.7 *Seien $\eta_j, \eta_{j+1}, \ldots, \eta_{j+r-1}$ gegeben und P_{r-1} das Interpolationspolynom durch (x_ν, η_ν) für $j \le \nu \le j+r-1$. Setze nun*

$$\eta_{j+r} = P_{r-1}(x_{j+r})\,. \tag{4.2.7a}$$

Wegen $\eta_{j+\mu} = y(x_{j+\mu})$ für $0 \le \mu \le r-1$ gilt für den lokalen Diskretisierungsfehler $\tau = \mathcal{O}(h^r)$. (4.2.7a) ist also ein lineares r-Schrittverfahren der Ordnung $r-1$.

Im Spezialfall $r = 2$ gilt

$$\eta_{j+2} = 2\eta_{j+1} - \eta_j\,. \tag{4.2.7b}$$

Das Verfahren hat zwar Konsistenzordnung 1, kann aber ebensowenig wie (4.2.7a) gegen die Lösung konvergieren, da es f nicht berücksichtigt.

Um von Konsistenz auf Konvergenz zu schließen, benötigt man als zusätzliche Eigenschaft die "Stabilität" des Verfahrens.

4.3 Lösung linearer Differenzengleichungen

4.3.1 Lösungsraum $\mathcal{F}_0$

Definition 4.3.1 *Seien $a_0, a_1, \ldots, a_{r-1} \in \mathbb{C},\ a_0 \neq 0,\ \beta_j \in \mathbb{C}$ für $j \in \mathbb{N}_0$. Dann heißt*

$$u_{j+r} + \sum_{\nu=0}^{r-1} a_\nu u_{j+\nu} = \beta_j \qquad \text{für alle } j \in \mathbb{N}_0 \tag{4.3.1}$$

lineare Differenzengleichung (für die Werte $u_j \in \mathbb{C}$, $j \in \mathbb{N}_0$).

Bemerkung 4.3.2 *a) Sind $u_0, u_1, \ldots, u_{r-1}$ gegeben, so definiert (4.3.1) die nachfolgenden u_j $(j \ge r)$ eindeutig.*
b) Mit $a_r := 1$ nimmt (4.3.1) die Form $\sum_{\nu=0}^{r} a_\nu u_{j+\nu} = \beta_j$ für alle $j \in \mathbb{N}_0$ an.
c) Falls $\beta_j = 0$ für alle $j \ge 0$, heißt die Differenzengleichung homogen*:*

$$\sum_{\nu=0}^{r} a_\nu u_{j+\nu} = 0 \quad \text{für alle } j \in \mathbb{N}_0\,. \tag{4.3.2}$$

d) Zur Lösung von (4.3.2) macht man den Ansatz $u_j = z^j$ mit $z \in \mathbb{C} \setminus \{0\}$. Einsetzen liefert:

$$\sum_{\nu=0}^{r} a_\nu u_{j+\nu} = \sum_{\nu=0}^{r} a_\nu z^{j+\nu} = z^j \sum_{\nu=0}^{r} a_\nu z^\nu = 0\,.$$

Diese Bedingung ist wegen $z^j \neq 0$ genau dann erfüllt, wenn $\sum_{\nu=0}^{r} a_\nu z^\nu = 0$ gilt.

Definition 4.3.3 *Das charakteristische Polynom der Differenzengleichungen (4.3.1) und (4.3.2) lautet*

$$\psi(\xi) := \sum_{\nu=0}^{r} a_\nu \xi^\nu = 0\,. \tag{4.3.3}$$

Bemerkung 4.3.4 *Die Menge $\mathcal{F}_0$ der Lösungen von (4.3.2) bildet einen r-dimensionalen Vektorraum.*

Beweis. Offensichtlich ist $\mathcal{F}_0$ ein Vektorraum. Es bleibt also zu zeigen, dass $\mathcal{F}_0$ die Dimension r besitzt.

Die Abbildung $\phi : (u_j)_{j\in\mathbb{N}_0} \in \mathcal{F}_0 \mapsto \begin{pmatrix} u_0 \\ \vdots \\ u_{r-1} \end{pmatrix} \in \mathbb{C}^r$ ist surjektiv und nach Bemerkung 4.3.2a auch injektiv, also bijektiv. Da $\dim \mathbb{C}^r = r$, ist auch $\dim \mathcal{F}_0 = r$. ∎

4.3.2 Darstellung der Lösungen

Im Folgenden soll eine Basis von $\mathcal{F}_0$ bestimmt werden. Gemäß Bemerkung 4.3.2d definiert man $z_1, \ldots, z_m$ ($m \le r$) als die verschiedenen Nullstellen $z_i \neq 0$ des charakteristischen Polynoms ψ. Dann sind mit $(z_i^j)_{j\in\mathbb{N}_0}$ für $1 \le j \le m$ bereits m linear unabhängige Lösungen von (4.3.2) gefunden. Im Falle $m = r$ ist damit eine Basis gefunden.

Die Bedingung $z_i \neq 0$ wird diskutiert in

Bemerkung 4.3.5 *$z = 0$ ist genau dann Nullstelle von ψ, wenn $a_0 = 0$, was in Definition 4.3.1 ausgeschlossen wurde, da dann eine Differenzengleichung mit kleinerem r vorliegt. Trotzdem lässt sich eine zugehörige Lösung der homogenen Differenzengleichung angeben: $u_j = \delta_{j0}$ ($j \ge 0$; Kronecker-Symbol δ).*

Ist $z = 0$ eine k-fache Nullstelle von ψ, so gilt $a_0 = a_1 = \ldots = a_{k-1}$. Die zugehörigen Lösungen der homogenen Differenzengleichung sind $(u_j^{(\ell)})_{j\in\mathbb{N}_0}$ mit $u_j^{(\ell)} = \delta_{j\ell}$ für $\ell = 0, \ldots, k-1$.

Es sei daran erinnert, dass (ein Polynom oder eine holomorphe Funktion) ψ genau dann eine (mindestens) k-fache Nullstelle z besitzt, wenn

$$\psi(z) = \psi'(z) = \ldots = \psi^{(k-1)}(z) = 0. \tag{4.3.4}$$

Sei nun $z \neq 0$ eine k-fache Nullstelle von ψ. Wir wollen zeigen, dass für $0 \le \ell \le k-1$ die Folge $u_j := j^\ell z^j$ eine Lösung der homogenen Differenzengleichung ist.

Der Vektorraum der Polynome hat neben der üblichen Monombasis $\{x^0, x^1, \ldots, x^n\}$ auch die Basis $\left\{\prod_{k=0}^{m-1}(x-k) : m = 0, \ldots, n\right\}$. Die $u_j = j^\ell z^j$ bilden also genau dann eine Basis, wenn auch die Folgen $v_j = j\,(j-1) \times \ldots \times (j-\ell+1)\, z^j$ eine Basis bilden.

Mit der Leibniz-Regel erhält man aus (4.3.4) $\left(\frac{d}{d\zeta}\right)^\ell \left(\xi^j \psi(\xi)\right) = 0$ für $\xi = z$ und alle $0 \le \ell \le k-1$. Die explizite Darstellung von $(\frac{d}{d\zeta})^\ell \left(\xi^j\psi(\xi)\right) = (\frac{d}{d\xi})^\ell \sum_{\nu=0}^r a_\nu \xi^{j+\nu}$ lautet

$$\left(\xi^j \psi(\xi)\right)^{(\ell)}\Big|_{\xi=z} = \sum_{\nu=0}^{r} a_\nu z^{j+\nu-\ell}\,(j+\nu)\,(j+\nu-1) \times \ldots \times (j+\nu-\ell+1)\,. \tag{4.3.5}$$

Einsetzen von $v_j = j\,(j-1) \times \ldots \times (j-\ell+1)\, z^j$ in die Differenzengleichung (4.3.2) liefert

$$\sum_{\nu=0}^{r} a_\nu v_{j+\nu} = \sum_{\nu=0}^{r} a_\nu z^{j+\nu}\,(j+\nu)\,(j+\nu-1) \times \ldots \times (j+\nu-\ell+1)\,.$$

Dieser Ausdruck ist das Produkt von (4.3.5) mit z^ℓ, also beweist $\sum_{\nu=0}^r \alpha_\nu v_{j+\nu} = 0$ die Behauptung.

Da $z = 0$ nach Bemerkung 4.3.5 ausgeschlossen werden kann, erhalten wir den

Satz 4.3.6 *$z_1, \ldots, z_m \in \mathbb{C} \setminus \{0\}$ seien die verschiedenen Nullstellen von ψ mit Vielfachheiten $\nu_1, \ldots, \nu_m$, wobei $\sum_{i=1}^m \nu_i = r = \operatorname{grad}(\psi)$. Dann bilden die r Folgen $(u_j)_{j\in\mathbb{N}_0} = (j^\nu z_i^j)_{j\in\mathbb{N}_0}$ mit $0 \le \nu \le \nu_i$, $1 \le i \le m$ eine Basis des linearen Vektorraumes $\mathcal{F}_0$ aller Lösungen der homogenen Differenzengleichung (4.3.2).*

Korollar 4.3.7 *Alle Lösungen der inhomogenen Differenzengleichungen (4.3.1) lassen sich schreiben als*

spezielle Lösung + Lösung der homogenen Differenzengleichung (4.3.2).

4.3.3 Stabilität

Eine informelle Definition der Stabilität ist, dass alle homogenen Lösungen beschränkt bleiben.

Definition 4.3.8 *Die Differenzengleichung (4.3.1) heißt stabil, wenn alle Nullstellen z von ψ folgende Bedingung erfüllen:*

$$|z| < 1 \text{ oder } (|z| = 1 \text{ und } z \text{ ist einfache Nullstelle}). \tag{4.3.6}$$

Satz 4.3.9 *Die Lösungen der homogenen Differenzengleichung (4.3.2) bleiben für beliebige Anfangswerte $u_0, \ldots, u_{r-1}$ genau dann beschränkt, wenn Stabilität wie in (4.3.6) definiert gilt. Dabei bedeutet Beschränktheit $\sup_{j\ge0} |u_j| \le C \max_{0\le j\le r-1} |u_j|$ (Diese Definition ist äquivalent zu $\sup_{j\ge0} |u_j| < \infty$.).*

Beweis. a) Zuerst sei die Beschränktheit $\sup_{j\geq 0}|u_j|<\infty$ vorausgesetzt. Sei z eine Nullstelle des charakteristischen Polynoms. Wir wählen die spezielle Lösung $u_j=z^j$. Wegen der Beschränktheit von $|u_j|$ ist $|z|\leq 1$. Für einen indirekten Beweis von (4.3.6) nehmen wir an: Es gelte $|z|=1$ und z sei mindestens eine zweifache Nullstelle. Dann ist auch $u_j=jz^j$ eine Lösung und es gilt

$$\lim_{j\to 0}|u_j|=\lim_{j\to 0}j|z|^j=\lim_{j\to 0}j=\infty\ .$$

Das ist ein Widerspruch zur Voraussetzung, dass die u_j beschränkt sind. Also erfüllt jede Nullstelle z die Bedingung (4.3.6).

b) Zu beliebigen Anfangswerten $u_0,\ldots,u_{r-1}$ existiert nach Satz 4.3.6 eine Lösung der Form

$$u_j=\sum_{\nu\mu}a_{\nu\mu}\,j^\nu z_\mu^j,$$

wobei jetzt vorausgesetzt sei, dass alle Nullstellen z_μ die Bedingung (4.3.6) erfüllen. Für z_μ mit $|z_\mu|<1$ gilt $\lim_{j\to 0}j^\nu z_\mu^j=0$ und damit $\sup_{j\geq 0}\left|j^\nu z_\mu^j\right|<\infty$, während für z_μ mit $|z_\mu|=1$ nur der Exponent $\nu=0$ auftritt, sodass $\left|j^\nu z_\mu^j\right|=\left|z_\mu^j\right|=1$. Damit ist auch die Linearkombination $u_j=\sum_{\nu\mu}a_{\nu\mu}j^\nu z_\mu^j$ beschränkt. ∎

Lemma 4.3.10 *Die Differenzengleichung ist genau dann stabil, wenn es eine zugeordnete Matrixnorm $\|\cdot\|$ gibt, sodass gilt*

$$\|A\|\leq 1\ , \tag{4.3.7a}$$

wobei die Begleitmatrix $A\in\mathbb{C}^{r\times r}$ des charakteristischen Polynoms aus (4.3.3) folgende Gestalt hat:

$$A=\begin{pmatrix}0 & 1 & & 0\\ & \ddots & \ddots & \\ 0 & & 0 & 1\\ -a_0 & -a_1 & \ldots & -a_r\end{pmatrix}\ . \tag{4.3.7b}$$

Beweis. 1) Es gelte $\|A\|\leq 1$. Jede Lösung von (4.3.2) erfüllt

$$U_{j+1}:=\begin{pmatrix}u_{j+1}\\ \vdots\\ u_{j+r}\end{pmatrix}=A\begin{pmatrix}u_j\\ \vdots\\ u_{j+r-1}\end{pmatrix}=AU_j\ . \tag{4.3.7c}$$

Damit gilt

$$|u_j|\leq C\,\|U_j\|=C\left\|A^jU_0\right\|\leq C\left\|A^j\right\|\|U_0\|\leq C\,\|U_0\|\leq C'\max_{0\leq j\leq r-1}|u_j| \tag{4.3.7d}$$

und nach Satz 4.3.9 ist die Differenzengleichung stabil. In (4.3.7d) wurde ausgenutzt, dass die Maximumnorm $\|U_j\|_\infty:=\max_{0\leq k\leq r-1}|u_{j+k}|$ und $\|U_j\|$ äquivalent sind.

2) Nun sei die Differenzengleichung als stabil vorausgesetzt. Wegen

$$\det(\xi I-A)=\psi(\xi)$$

sind die Eigenwerte von A gerade die Nullstellen von ψ.

2a) Die Nullstellen z_i $(i=1,\ldots,r)$ seien zunächst alle durch $|z_i|<1$ beschränkt. Man definiere T als die Matrix, die A in Jordan-Normalform[26] bringt, $D:=\mathrm{diag}(1,\varepsilon,\ldots,\varepsilon^{r-1})$ mit $\varepsilon>0$ und eine Vektornorm durch

$$\|x\|_{\text{neu}}:=\|D^{-1}T^{-1}x\|_2.$$

Dann ist die zugeordnete Matrixnorm

$$\|X\|_{\text{neu}}=\left\|D^{-1}T^{-1}XTD\right\|_2$$

(Beweis als Übung).

[26] Marie Ennemond Camille Jordan, geb. 5. Jan. 1838 in La Croix-Rousse, Lyon, gest. 22. Jan. 1922 in Paris

Die neue Matrixnorm auf A angewandt liefert

$$\|A\|_{\text{neu}} = \left\|D^{-1}T^{-1}ATD\right\|_2 = \left\| D^{-1} \begin{pmatrix} z_1 & \mathcal{O}(1) & 0 \\ & \ddots & \mathcal{O}(1) \\ 0 & & z_r \end{pmatrix} D \right\|_2$$

In der Nebendiagonalen steht entweder 0 oder 1. In beiden Fällen: $\mathcal{O}(1)$

$$= \left\| \begin{pmatrix} z_1 & \mathcal{O}(\varepsilon) & 0 \\ & \ddots & \mathcal{O}(\varepsilon) \\ 0 & & z_r \end{pmatrix} \right\|_2$$

Die Transformation mit D multipliziert die obere Nebendiagonale mit ε.

$$\leq \left\| \begin{pmatrix} z_1 & & 0 \\ & \ddots & \\ 0 & & z_r \end{pmatrix} \right\|_2 + \left\| \begin{pmatrix} 0 & \mathcal{O}(\varepsilon) & 0 \\ & \ddots & \mathcal{O}(\varepsilon) \\ 0 & & 0 \end{pmatrix} \right\|_2$$

$$\leq \max_{i=1}^{r} |z_i| + \varepsilon \leq 1$$

für ein hinreichend kleines ε.

2b) Sei nun $|z_r| \leq \ldots \leq |z_{m+1}| < |z_m| = \ldots = |z_1| = 1$, und $z_1, \ldots, z_m$ seien einfache Nullstellen. Die Jordan-Normalform hat dann die Gestalt

$$A' = T^{-1}AT = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \text{ mit } A_1 = \operatorname{diag}(z_1, \ldots, z_m),\ A_2 = \begin{pmatrix} z_{m+1} & \mathcal{O}(1) & \\ & \ddots & \mathcal{O}(1) \\ 0 & & z_r \end{pmatrix}. \tag{4.3.7e}$$

Dank $\|A'\|_2 = \max\{\|A_1\|_2, \|A_2\|_2\}$ und Teil 2a mit $D := \operatorname{diag}(1, \ldots, 1, 1, \varepsilon, \ldots, \varepsilon^{r-m-1})$ erhält man wieder $\|A\|_{\text{neu}} \leq 1$. ∎

Satz 4.3.11 *Die Differenzengleichung sei stabil, die Anfangswerte seien durch α beschränkt:*

$$|u_0|, \ldots, |u_{r-1}| \leq \alpha.$$

Für ein hinreichend kleines $\gamma \geq 0$ gelte

$$\sum_{\mu=0}^{r} a_\mu u_{j+\mu} = \beta_j \qquad \textit{mit } |\beta_{j+r}| \leq \beta + \gamma \max_{0 \leq \mu \leq j+r} |u_\mu| \,. \tag{4.3.8}$$

Dann gibt es k, $k' > 0$ mit

$$|u_j| \leq kk'\alpha\, e^{\frac{j\gamma k}{1-\gamma k}} + \begin{cases} jk\beta/(1-\gamma k) & \textit{für } \gamma = 0, \\ \frac{\beta}{\gamma}(e^{\frac{j\gamma k}{1-\gamma k}} - 1) & \textit{für } \gamma > 0 \textit{ und } \gamma k < 1. \end{cases}$$

Beweis. 1) Seien A und $\|\cdot\|_{\text{neu}}$ wie in Lemma 4.3.10. Seien k, k' die Konstanten in $\|U\|_\infty \leq k\|U\|_{\text{neu}}$ und $\|U\|_{\text{neu}} \leq k'\|U\|_\infty$, die die Äquivalenz der beiden Normen auf dem $\mathbb{C}^r$ beschreiben.

2) Man setze

$$U_j := \begin{pmatrix} u_j \\ \vdots \\ u_{j+r-1} \end{pmatrix}, \qquad e := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^r \,.$$

Dann ist die Differenzengleichung (4.3.8) äquivalent zu $U_{j+1} = AU_j + \beta_{j+r}e$. Nach Lemma 4.3.10 gilt $\|A\|_{\text{neu}} \leq 1$. Eine Skalierung der Vektornorm $\|\cdot\|_{\text{neu}}$ ändert die Matrixnorm nicht. Daher sei o.B.d.A.[27] $\|e\|_{\text{neu}} = 1$. Im Weiteren schreiben wir $\|\cdot\|$ für $\|\cdot\|_{\text{neu}}$.

3) Mit $\xi_j := \max_{0 \leq \mu \leq j} |u_\mu|$ folgt

$$\|U_{j+1}\| = \|AU_j + \beta_{j+r}e\| \leq \underbrace{\|A\|}_{\leq 1}\|U_j\| + |\beta_{j+r}|\underbrace{\|e\|}_{=1} \leq \|U_j\| + |\beta_{j+r}| \leq \|U_j\| + \beta + \gamma\xi_{j+r} \,.$$

[27] Die Skalierung wirkt sich allerdings auf k und k' aus.

Man definiere $\eta_0 := \|U_0\|$ und $\eta_{j+1} = \eta_j + \beta + \gamma\xi_{j+r}$. Offenbar gilt $\|U_j\| \le \eta_j$ und $\eta_{j+1} \ge \eta_j$. Es folgt

$$\xi_{j+r} = \max_{0\le\mu\le j+r} |u_\mu| = \max_{0\le\mu\le j+1} \|U_\mu\|_\infty \le \max_{0\le\mu\le j+1} k\,\|U_\mu\| = k \max_{0\le\mu\le j+1} \eta_\mu = k\eta_{j+1}.$$

Um die Ungleichung $\eta_{j+1} = \eta_j + \beta + \gamma\xi_{j+r} \le \eta_j + \beta + \gamma k\eta_{j+1}$ nach η_{j+1} auflösen zu können, muss $\gamma k < 1$ gelten: $\eta_{j+1} \le \frac{1}{1-\gamma k}\eta_j + \frac{\beta}{1-\gamma k}$. Nach Lemma 2.1.6 (mit $h = 1$, $L = \gamma k/(1-\gamma k)$, $B = \beta/(1-\gamma k)$) ist die Abschätzung

$$\eta_j \le \eta_0\, e^{\frac{j\gamma k}{1-\gamma k}} + \begin{cases} j\beta/(1-\gamma k) & \text{für } \gamma = 0, \\ \frac{\beta}{\gamma k}(e^{\frac{j\gamma k}{1-\gamma k}} - 1) & \text{für } \gamma > 0 \text{ und } \gamma k < 1 \end{cases}$$

gültig. Zusammen mit $\eta_0 = \|U_0\| \le k' \|U_0\|_\infty \le k'\alpha$ und $|u_j| \le \|U_j\|_\infty \le k\,\|U_j\|$ folgt die Behauptung. ∎

4.4 Konvergenz von Mehrschrittverfahren

Definition 4.4.1 *Ein r-Schrittverfahren der Form (4.1.1) (d.h. $\sum_{\nu=0}^r a_\nu\eta_{j+\nu} = h\phi(x_j, \eta_{j+r}, \ldots, \eta_j, h, f)$ oder in der gestörten Form (4.1.4)) heißt* stabil, *falls für alle Nullstellen ξ des charakteristischen Polynoms ψ gilt:*

$$|\xi| < 1 \quad \textit{oder} \quad (|\xi| \le 1 \textit{ und } \xi \textit{ ist einfache Nullstelle}).$$

Satz 4.4.2 *Das Verfahren (4.1.1) sei für $\phi \equiv 0$ gegeben. Dann folgt aus der Konvergenz des Verfahrens die Stabilität.*

Beweis. Wegen $\phi \equiv 0$ liegt die homogene Aufgabe (4.3.2) vor: $\sum_{\nu=0}^r a_\nu\eta_{j+\nu} = 0$. Die exakte Lösung ist $y \equiv 0$. Man definiere $\eta(x, \vec{\varepsilon}, h)$ durch $\eta(jh, \vec{\varepsilon}, h) = \eta_j$, wobei η_j (4.3.2) erfüllt und die Anfangswerte $\eta_j = y(x_j) + \varepsilon_j$ für $0 \le j \le r-1$ besitzt. Wegen der Konvergenz des Verfahrens gilt

$$\lim_{h\to 0, \|\vec{\varepsilon}\|\to 0} \eta(x, \vec{\varepsilon}, h) = y(x)\,.$$

Hier ist $y \equiv 0$ und damit $\eta_j = \varepsilon_j$. Für hinreichend kleine h und $\|\vec{\varepsilon}\|$ ist somit $\|\eta(x, \vec{\varepsilon}, h)\|_\infty \le 1$. Satz 4.3.9 ist anwendbar und beweist die Stabilität. ∎

Für die Inkrementfunktion ϕ in (4.1.1) wird die folgende Eigenschaft vorausgesetzt.

Bedingung 4.4.3 *Für jede Lipschitz-stetige Funktion f existiere M_f mit*

$$|\phi(x, u_r, \ldots, u_0; h, f) - \phi(x, v_r, \ldots, v_0; h, f)| \le M_f \max_{0\le j\le r} |u_j - v_j|\ . \tag{4.4.1}$$

Bemerkung 4.4.4 *Forderung (4.4.1) ist für lineare Mehrschrittverfahren (4.1.2) erfüllt.*

Satz 4.4.5 *Ein Mehrschrittverfahren der Form (4.1.1) sei konsistent und stabil. f sei Lipschitz-stetig, und ϕ erfülle (4.4.1). Dann konvergiert das Verfahren. Diese Aussage gilt auch, wenn die stärkeren Störungen aus (4.1.4) vorliegen.*

Beweis. η_j sei Lösung von (4.1.4) mit Störungen ε_j ($|\varepsilon_j| \le \varepsilon$) in den Anfangswerten $\eta_j = y(x_j) + \varepsilon_j$ und in der Differenzengleichung $\sum_{\nu=0}^r a_\nu\eta_{j+\nu} = h\phi(x_j, \eta_{j+r}, \ldots, \eta_j; h, f) + h\varepsilon_{j+r}$. Die Fehler seien $e_j := \eta_j - y(x_j)$. Die Anfangswerte sind $e_j = \varepsilon_j$ ($0 \le j \le r-1$). Der Konsistenzfehler lautet

$$h\tau_{j+r} = \sum_{\nu=0}^r a_\nu y((j+\nu)h) - h\phi(x_j, \eta_{j+r}, \ldots, \eta_j; h, f).$$

Damit erfüllt e_j die Rekursion $\sum_{\nu=0}^r a_\nu e_{j+\nu} = c_{j+r}$ mit

$$c_{j+r} = h(\phi(x_j, \eta_{j+r}, \ldots, \eta_j; h, f) - \phi(x_j, y((j+r)h, \ldots, y(jh); h, f)) + h(\varepsilon_{j+r} - \tau_{j+r}).$$

Die rechte Seite wird abgeschätzt durch

$$|c_{j+r}| \le hM_f \max_{0\le\mu\le r} \left|\eta_{j+\mu-y((j+\mu)h)}\right| + h(\varepsilon + \sup_{0\le\mu\le r} \tau_{j+\mu}) = hM_f \max_{0\le\mu\le r} |e_{j+\mu}| + h(\varepsilon + \sup_{0\le\mu\le r} \tau_{j+\mu})\,.$$

Damit gelten die Voraussetzungen von Satz 4.3.11 mit $\alpha = \varepsilon$, $\gamma = hM_f$, $\beta = h(\varepsilon + \sup \tau_{j+\mu})$. Wegen $h \to 0$ wird γ hinreichend klein, und es ist

$$\begin{aligned} |e_j| &\le kk'\varepsilon \exp(\frac{jhM_f k}{1 - hM_f k}) + \frac{h(\varepsilon + \sup \tau_{j+\mu})}{hM_f} \left(\exp(\frac{jhM_f k}{1 - hM_f k}) - 1 \right) \\ &\le kk'\varepsilon e^{(x-x_0)2M_f k} + \frac{\varepsilon + \sup \tau_{j+\mu}}{M_f} \left(\left(\exp \frac{(x - x_0)M_f k}{1 - hM_f k} \right) - 1 \right) \qquad \text{für } x_j = x_0 + jh \le x \end{aligned}$$

Wegen der Konsistenz des Verfahrens geht $\sup |\tau_\mu|$ gegen 0. Zusammen mit $\varepsilon \to 0$ folgt auch $\lim_{h \to 0} |e_j| \to 0$, d.h. die Konvergenz des Verfahrens. ∎

Korollar 4.4.6 *Die Voraussetzungen seien wie in Satz 4.4.5, p sei die Konsistenzordnung. Dann ist p auch die Konvergenzordnung.*

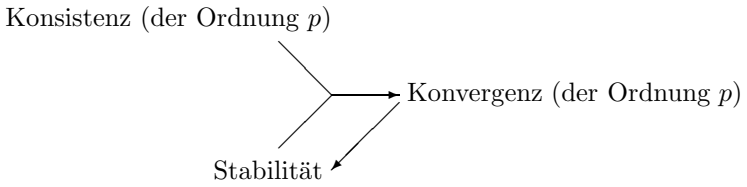


Abbildung 4.4.1: Der Zusammenhang von Konsistenz, Konvergenz und Stabilität

Beispiel 4.4.7 *a)* ***Adams-Bashforth:*** *Es sind $a_r = 1$, $a_{r-1} = -1$, $a_{r-2} = \ldots = a_0 = 0$. Also ist $\psi(\xi) = \xi^r - \xi^{r-1} = \xi^{r-1}(\xi - 1)$. Die Nullstellen sind also $\xi_1 = 1$, $\xi_2 = \ldots = \xi_r = 0$. Das Verfahren ist also stabil.*

b) ***Adams-Moulton:*** *Da sich das Verfahren von Adams-Bashforth nur in den β unterscheidet, folgt die Stabilität analog.*

c) ***Mittelpunktsregel:*** *Hier ist $r = 2$ und $a_2 = 1$, $a_1 = 0$, $a_0 = -1$. Damit ist $\psi(\xi) = \xi^2 - 1$ mit den Nullstellen $\xi_1 = 1$ und $\xi_2 = -1$.*

d) ***(4.2.6):*** *Sei $\psi(\xi) = \xi^2 + 4\xi - 5$. Dann sind die Nullstellen $\xi_1 = 1$ und $\xi_2 = -5$. Wegen $|\xi_2| > 1$ ist das Verfahren nicht stabil.*

e) ***(4.2.7b):*** *Sei $\psi(\xi) = \xi^2 - 2\xi + 1$. Dann ist $\xi = 1$ zweifache Nullstelle und das Verfahren nicht stabil.*

f) ***BDF-Verfahren:*** *Stabilität liegt genau dann vor, wenn $r \le 6$ (vgl. Übungsaufgabe 4.2.5).*

4.5 Konsistenz linearer Mehrschrittverfahren

Bei linearen Mehrschrittverfahren sollten die Koeffizienten b_ν so gewählt werden, dass der lokale Diskretisierungsfehler

$$\tau(x, y; h) = \frac{1}{h} \left[\sum_{\nu=0}^{r} a_\nu z(x + \nu h) - h \sum_{\nu=0}^{r} b_\nu f(x + \nu h, z(x + \nu h)) \right]$$

klein wird (z Lösung von $z' = f(t, z)$, $z(t) = y$).

Sei $f \equiv 0$ und $y = 1$, sodass $z \equiv 1$ die exakte Lösung ist. Für diesen Spezialfall ist $\tau(x, y; h) = \frac{1}{h} \sum_{\nu=0}^{r} a_\nu$. Wegen $\tau(x, y; h) \to 0$ für $h \to 0$ muss $\sum_{\nu=0}^{r} a_\nu = 0$ gelten. Mit Hilfe des charakteristischen Polynoms $\psi(\zeta) = \sum_{\nu=0}^{r} a_\nu \zeta^\nu$ schreiben wir diese Bedingung als

$$\psi(1) = 0 \tag{4.5.1a}$$

Neben ψ definieren wir das Polynom

$$\sigma(\zeta) := \sum_{\nu=0}^{r} b_\nu \zeta^\nu, \tag{4.5.1b}$$

das zu den Koeffizienten b_ν gehört. Zu ψ, σ sei die Funktion φ definiert:

$$\varphi(\zeta) := \frac{\psi(\zeta)}{\log \zeta} - \sigma(\zeta). \tag{4.5.1c}$$

Lemma 4.5.1 *a) φ ist beschränkt für $\zeta = 1$ genau dann, wenn (4.5.1a).*
b) Unter der Voraussetzung (4.5.1a) ist φ holomorph in $\zeta = 1$.

Beweis. Falls (4.5.1a) nicht gilt, ist φ unbeschränkt bei $\zeta = 1$. Unter der Voraussetzung (4.5.1a) nimmt der Ausdruck $\frac{\psi(\zeta)}{\log \zeta}$ bei $\zeta = 1$ die Form $\frac{0}{0}$ an. Nach der Regel von l'Hospital gilt

$$\lim_{\zeta \to 1} \frac{\psi(\zeta)}{\log \zeta} = \frac{\frac{d}{d\zeta}\psi(\zeta)}{\frac{d}{d\zeta}\log \zeta}\Big|_{\zeta=1} = \frac{\psi'(1)}{1}.$$

Dies beweist die Beschränktheit und darüber hinaus $\varphi(1) = \psi'(1) - \sigma(1)$. Da nun bei $\zeta = 1$ eine hebbare Singularität vorliegt, ist φ auch in $\zeta = 1$ holomorph. ∎

Satz 4.5.2 *Sei $\psi(1) = 0$. Dann hat das lineare Mehrschrittverfahren (4.1.2) genau dann die (lokale) Konsistenzordnung p, wenn $\zeta = 1$ eine p-fache Wurzel von φ ist.*

Beweis. a) Das Mehrschrittverfahren habe die Konsistenzordnung p. Wir wählen die Differentialgleichung $y' = y$. Dann ist (bis auf einen Faktor) $z(t) = e^t$ und

$$\tau(x,y,h) = \frac{\sum_{\nu=0}^r a_\nu e^{x+\nu h} - h\sum_{\nu=0}^r b_\nu e^{x+\nu h}}{h} = e^x \frac{\sum_{\nu=0}^r a_\nu (e^h)^\nu - h\sum_{\nu=0}^r b_\nu (e^h)^\nu}{h}$$
$$= e^x \left(\frac{\psi(e^h)}{h} - \sigma(e^h) \right) = e^x \varphi(e^h).$$

Wir setzen $\delta := e^h - 1 = h \cdot e^{\theta h}$ $(0 < \theta < h$; Zwischenwertsatz der Differentialrechnung). Damit lässt sich δ nach oben und unten durch $const \cdot h$ abschätzen. Die Taylor-Entwicklung von $\varphi(e^h)$ um $h = 0$ existiert wegen Lemma 4.5.1b und wird als Taylor-Entwicklung von $\varphi(1+\delta)$ geschrieben:

$$\varphi(e^h) = \varphi(1) + \varphi'(1)\delta + \ldots + \frac{\varphi^{(p-1)}(1)}{(p-1)!}\delta^{p-1} + \mathcal{O}(\delta^p).$$

Da $\delta^k \sim h^k$, folgt aus $\tau(x,y,h) = \mathcal{O}(h^p) = \mathcal{O}(\delta^p)$, dass die δ^k-Terme mit $k < p$ nicht auftreten dürfen, d.h. $\varphi(1) = \varphi'(1) = \ldots = \varphi^{(p-1)}(1) = 0$. Also ist 1 eine p-fache Nullstelle von φ. Ist umgekehrt 1 eine p-fache Nullstelle, zeigt die Taylor-Entwicklung, dass $\tau(x,y,h) = e^x\varphi(e^h) = \mathcal{O}(\delta^p) = \mathcal{O}(h^p)$ und somit Konsistenz der Ordnung p vorliegt. ∎

4.6 Optimale Ordnung linearer Mehrschrittverfahren

Da $a_r = 1$ festgelegt ist (vgl. Definition 4.1.2) und die Bedingung (4.5.1a) vorliegt, verbleiben von den Koeffizienten $a_0, \ldots, a_r$ noch $r-1$ freie Parameter.

Die Zahl der Koeffizienten b_ν ist r für explizite Verfahren ($b_r = 0$) und $r+1$ für implizite Verfahren.

Zusammen ergeben sich $2r-1$ $[2r]$ freie Parameter. Diese können verwendet werden, um $\varphi(1) = \varphi'(1) = \ldots = \varphi^{(p-1)}(1) = 0$ ("p-fache Nullstelle bei 1") mit $p = 2r-1$ $[p = 2r]$ zu erfüllen. Dies beweist die

Folgerung 4.6.1 *Es gibt ein explizites [implizites] lineares r-Schrittverfahren der Konsistenzordnung $2r-1$ $[2r]$. Für $r = 1, 2$ lauten diese Verfahren und ihre Ordnungen p wie folgt:*

$r = 1$	*explizit*	*Euler-Verfahren*	$p = 1 = 2r-1$
$r = 1$	*implizit*	*Trapezformel*	$p = 2 = 2r$
$r = 2$	*explizit*	*(4.2.6)*	$p = 3 = 2r-1$
$r = 2$	*implizit*	*(4.6.1)*	$p = 4 = 2r$

wobei im letzten Fall die Koeffizienten gegeben sind durch

$$a_0 = -1,\ a_1 = 0,\ a_2 = 1,\ b_0 = b_2 = \tfrac{1}{3},\ b_1 = \tfrac{4}{3}. \tag{4.6.1}$$

Bemerkung 4.6.2 *Das Verfahren (4.6.1) ist stabil, da $\psi(\zeta) = \zeta^2 - 1$ die Nullstellen $\zeta_1 = 1$ und $\zeta_2 = -1$ hat. Das Verfahren (4.2.6) wurde schon in Beispiel 4.4.7d als instabil erkannt (vgl. Beispiel 4.2.6). Für Einschrittverfahren ($r = 1$) liegt prinzipiell Stabilität vor.*

Für alle $r > 2$ sind Verfahren der Ordnung $2r$ instabil, wie der folgende Satz von Dahlquist[28] besagt.

[28] Germund Dahlquist, geb. 16. Jan. 1925, Emeritus an der Universität Stockholm

Satz 4.6.3 (Dahlquist) *Die höchste erreichbare Konsistenzordnung p eines stabilen r-Schrittverfahren ist*

$$p = r + 1 \quad \textit{für } r \textit{ ungerade} \tag{4.6.2a}$$
$$p = r + 2 \quad \textit{für } r \textit{ gerade}\,. \tag{4.6.2b}$$

Eine hinreichende und notwendige Voraussetzung für (4.6.2b) ist, dass alle Wurzeln von ψ den Betrag 1 haben und σ speziell bestimmt wird (vgl. Beweis).

Beweis. 1) Wir zeigen zunächst, dass stabile r-Schrittverfahren der Ordnungen $p = r + 1$ existieren. Dazu wird die Variable ζ im Argument der Polynome $\psi(\zeta)$ und $\sigma(\zeta)$ durch z substituiert:

$$\zeta = \frac{1+z}{1-z}, \qquad z = \frac{\zeta - 1}{\zeta + 1}. \tag{4.6.3a}$$

Dies liefert die Funktionen

$$p(z) := \left(\frac{1-z}{2}\right)^r \psi\left(\frac{1+z}{1-z}\right), \qquad s(z) = \left(\frac{1-z}{2}\right)^r \sigma\left(\frac{1+z}{1-z}\right). \tag{4.6.3b}$$

Bemerkung 4.6.4 *a) $p(z)$ und $s(z)$ sind wieder Polynome vom Grad $\le r$.*

b) Ist $\zeta_0 \neq -1$ eine Wurzel von $\psi(\zeta)$ der Vielfachheit k, so hat $p(z)$ bei $z_0 = \frac{\zeta_0 - 1}{\zeta_0 + 1}$ eine Nullstelle gleicher Vielfachheit.

c) Mit (4.6.3a) wird der komplexe Einheitskreis $|\zeta| < 1$ auf die linke Halbebene $\Re\mathrm{e}\, z < 0$ abgebildet; insbesondere werden $\zeta = 1$ auf $z = 0$ und $\zeta = -1$ auf $z = \infty$ abgebildet.

Beweis. a) $p(z)$ ist eine Linearkombination von $\left(\frac{1+z}{1-z}\right)^\nu (1-z)^r = (1-z)^{r-\nu}(1+z)^\nu$ für $0 \le \nu \le r$.

b) Ist k die exakte Vielfachheit von ζ_0, so gilt $\psi(\zeta) = (\zeta - \zeta_0)^k \psi_0(\zeta)$ mit $\psi_0(\zeta_0) \neq 0$. Folglich ist

$$\begin{aligned} p(z) &= \left(\frac{1-z}{2}\right)^r \left(\frac{1+z}{1-z} - \zeta_0\right)^k \psi_0\left(\frac{1+z}{1-z}\right) = (1 + z - (1-z)\,\zeta_0)^k \left[\frac{1}{2^r}(1-z)^{r-k}\,\psi_0\left(\frac{1+z}{1-z}\right)\right] \\ &= (z - z_0)^k \left[\frac{(1-z)^{r-k}}{2^r\,(\zeta_0+1)^k}\psi_0\left(\frac{1+z}{1-z}\right)\right] \end{aligned}$$

Da $\zeta_0 = \infty$ ausgeschlossen ist, tritt $z_0 = 1$ nicht auf, und die eckige Klammer ist bei $z = z_0$ ungleich null. ∎

Wegen $\psi(1) = 0$ (vgl. (4.5.1a)) und der Stabilität ist $\zeta = 1$ eine Wurzel der Ordnung 1. Gemäß Bemerkung 4.6.4c hat $p(z)$ eine einfache Nullstelle bei $z = 0$. Damit hat p die Gestalt

$$p(z) = \alpha_1 z + \alpha_2 z^2 + \ldots + \alpha_\ell z^\ell \qquad \text{mit } \alpha_1 \neq 0,\ \ell = \operatorname{grad} p \le k. \tag{4.6.4a}$$

Ohne Beschränkung der Allgemeinheit darf

$$\alpha_1 > 0 \tag{4.6.4b}$$

angenommen werden. (Andernfalls wird die Gleichung des Mehrschrittverfahrens mit -1 skaliert, was $\psi(\zeta)$ in $-\psi(\zeta)$ ändert.) Wir wollen nun zeigen, dass

$$\alpha_\mu \ge 0 \qquad \text{für alle } 1 \le \mu \le r. \tag{4.6.4c}$$

Dazu schreiben wir die Nullstellen von p als $z_\nu = x_\nu + i\,y_\nu$ $(x_\nu, y_\nu \in \mathbb{R})$. Dann muss

$$p(z) = \alpha_\ell\, z \prod_\nu (z - z_\nu) = \alpha_\ell\, z \prod_{\nu \text{ mit } y_\nu = 0} (z - x_\nu) \prod_{\nu \text{ mit } y_\nu \neq 0} \left((z - x_\nu)^2 + y_\nu^2\right) \tag{4.6.4d}$$

gelten, wobei das letzte Produkt über alle Paare von konjugiert komplexen Nullstellen geführt wird. Die Stabilität impliziert $|\zeta_\nu| \le 1$, was nach Bemerkung 4.6.4c zu $\Re\mathrm{e}\, z_\nu \le 0$, d.h. $x_\nu \le 0$ führt. Aus $z - x_\nu = z + |x_\nu|$ ersieht man, dass alle Polynomkoeffizienten von $p(z)$ das gleiche Vorzeichen (oder Vorzeichen null) haben müssen. Mit (4.6.4b) folgt (4.6.4c).

Mit φ aus (4.5.1c) setzen wir nun

$$g(z) := \left(\frac{1-z}{2}\right)^r \varphi\left(\frac{1+z}{1-z}\right) = \frac{p(z)}{\log\frac{1+z}{1-z}} - s(z)$$

und beachten die folgenden Implikationen:

$$\begin{aligned} & g(z) \text{ hat bei } z=0 \text{ eine } p\text{-fache Nullstelle} \\ \Leftrightarrow\quad & \varphi(\zeta) \text{ hat bei } \zeta = 1 \text{ eine } p\text{-fache Nullstelle} \\ \Leftrightarrow\quad & p = \text{Konsistenzordnung} \end{aligned}$$

(vgl. Satz 4.5.2). Da $p(z)/\log\frac{1+z}{1-z}$ holomorph bei $z=0$ ist (vgl. Lemma 4.5.1), gibt es eine Potenzreihenentwicklung

$$\frac{z}{\log\frac{1+z}{1-z}}\frac{p(z)}{z} = \beta_0 + \beta_1 z + \beta_2 z^2 + \dots . \tag{4.6.5a}$$

Damit g eine p-fache Nullstelle $z=0$ besitzt, muss $s(z) = \sum_{\mu=0}^{r} \beta'_\mu z^\mu$ mit $\beta'_\mu = \beta_\mu$ für $0 \le \mu \le p-1$ gelten. Da nach (4.6.2a) $p = r+1 > r = \operatorname{grad} s$ gilt, ist das Polynom $s(z)$ bereits eindeutig festgelegt:

$$s(z) = \beta_0 + \beta_1 z + \ldots + \beta_r z^r, \tag{4.6.5b}$$

wobei $r = p-1$. Aus $s(z)$ gemäß (4.6.5b) erhält man $\sigma(\zeta)$ (vgl. (4.6.3b)). Damit hat man ein Verfahren der Ordnung $p = r+1$ konstruiert.

2) Als Nächstes wollen wir die Annahme, die Ordnung $p > r+1$ sei für ungerades r erreichbar, zum Widerspruch führen. Der Vergleich von (4.6.5a) und (4.6.5b) zeigt, dass in (4.6.5a)

$$\beta_\mu = 0 \qquad \text{für } r+1 \le \mu \le p-1 \tag{4.6.5c}$$

gelten müsste (die ersten β_μ, $0 \le \mu \le r$, würden durch die Wahl (4.6.5b) neutralisiert). Die Funktion $z/\log\frac{1+z}{1-z}$ ist eine gerade Funktion in z, sodass die Potenzreihenentwicklung

$$\frac{z}{\log\frac{1+z}{1-z}} = c_0 + c_2 z^2 + c_4 z^4 + \dots$$

lautet. Seien α_μ die Koeffizienten aus (4.6.4a), wobei $\alpha_\mu := 0$ für $\mu > \ell = \operatorname{grad} p$ gesetzt sei. Koeffizientenvergleich beider Seiten aus (4.6.5a) zeigt

$$\begin{aligned} \beta_0 &= c_0\alpha_1, & \dots \qquad & \beta_{2\nu} = c_0\alpha_{2\nu+1} + c_2\alpha_{2\nu-1} + \ldots + c_{2\nu}\alpha_1, \\ \beta_1 &= c_0\alpha_2, & \dots \qquad & \beta_{2\nu+1} = c_0\alpha_{2\nu+2} + c_2\alpha_{2\nu} \quad + \ldots + c_{2\nu}\alpha_2. \end{aligned}$$

Wir werden noch zeigen, dass $c_{2\nu} < 0$ für alle $\nu \ge 1$. Für ungerades r folgt somit

$$\beta_{r+1} = c_0\underbrace{\alpha_{r+2}}_{=0} + \underbrace{c_2}_{<0}\underbrace{\alpha_r}_{\ge 0} + \underbrace{c_4}_{<0}\underbrace{\alpha_{r-2}}_{\ge 0} + \ldots + \underbrace{c_{r+1}}_{<0}\underbrace{\alpha_1}_{>0} < 0 \tag{4.6.5d}$$

im Widerspruch zu $\beta_{r+1} = 0$ (vgl. (4.6.5c)).

Zum Beweis von $c_{2\nu} < 0$ benutzen wir das

Lemma 4.6.5 ([9]) *Für $f(t) = \sum_{\nu=0}^{\infty} A_\nu t^\nu$ und $g(t) = \sum_{\nu=0}^{\infty} B_\nu t^\nu$ gelte*

$$f(t)g(t) \equiv 1, \qquad A_\nu > 0 \ (\nu \ge 0), \qquad A_{\nu+1}A_{\nu-1} > A_\nu^2 \ (\nu \ge 1).$$

Dann ist $B_\nu < 0$ für alle $\nu \ge 1$.

Beweis. O.B.d.A. sei $A_0 = 1$, was $B_0 = 1$ impliziert. Koeffizientenvergleich in $f(t)g(t) \equiv 1$ zeigt

$$0 = A_n + \sum_{\nu=1}^{n} B_\nu A_{n-\nu} \quad (n \ge 1) \quad \text{und} \quad -B_{n+1} = A_{n+1} + \sum_{\nu=1}^{n} B_\nu A_{n-\nu+1} \quad (n \ge 0).$$

Die letzte Identität für $n = 0$ zeigt $B_1 < 0$. Man multipliziert die erste Gleichung mit A_{n+1}, die zweite mit $-A_n$ und addiert: $A_n B_{n+1} = \sum_{\nu=1}^{n} B_\nu \left(A_{n+1}A_{n-\nu} - A_n A_{n-\nu+1}\right)$. Hieraus schließt man per Induktion, dass $B_{n+1} < 0$, falls $A_{n+1}A_{n-\nu} - A_n A_{n-\nu+1} > 0$. Diese Ungleichung ist aber mit $\frac{A_{n+1}}{A_n} > \frac{A_{n-\nu+1}}{A_{n-\nu}}$ identisch und folgt aus der Voraussetzung $A_{\nu+1}A_{\nu-1} > A_\nu^2$, die als $\frac{A_{\nu+1}}{A_\nu} > \frac{A_\nu}{A_{\nu-1}}$ geschrieben werden kann. ∎

Wir wenden dies Lemma auf $f(z^2) = \frac{1}{z}\log\frac{1+z}{1-z} = 2 + \frac{2}{3}z^2 + \frac{2}{5}z^4 + \ldots$ an. Die Koeffizienten $A_\nu = \frac{2}{2\nu+1} > 0$ erfüllen wegen

$$A_{\nu+1}A_{\nu-1} = \frac{2}{2\nu+3}\cdot\frac{2}{2\nu-1} = \frac{4}{(2\nu+1)^2-4} > \frac{4}{(2\nu+1)^2} = A_\nu^2$$

die Voraussetzung des Lemmas. Da $B_\nu = c_{2\nu}$, ist der Beweis des Teils 2) vollständig.

3) Für gerades r lautet die (4.6.5d) entsprechende Summe

$$\beta_{r+1} = c_0\underbrace{\alpha_{r+2}}_{=0} + \underbrace{c_2}_{<0}\underbrace{\alpha_r}_{\geq 0} + \underbrace{c_4}_{<0}\underbrace{\alpha_{r-2}}_{\geq 0} + \ldots + \underbrace{c_r}_{<0}\underbrace{\alpha_2}_{\geq 0} \leq 0,$$

wobei das Gleichheitszeichen genau dann gilt, wenn $\alpha_2 = \alpha_4 = \ldots = \alpha_r = 0$. Letzteres ist äquivalent dazu, dass $p(z)$ ungerade ist:

$$p(z) = -p(-z) \qquad \text{für alle } z.$$

Damit muss mit jeder Wurzel z_ν von p auch $-z_\nu$ eine Nullstelle sein. Da aber $\Re e\, z_\nu \leq 0$ wegen der Stabilität gilt, folgt $\Re e\, z_\nu = 0$ für alle z_ν, was der Bedingung $|\zeta_\nu| = 1$ entspricht. Umgekehrt gilt: Wenn $|\zeta_\nu| = 1$ für alle Wurzeln von ψ zutrifft, folgen $\Re e\, z_\nu = 0$ und

$$p(z) = const \cdot z \prod_\nu \left(z^2 - (\Im m\, z_\nu)^2\right)$$

und damit $p(z) = -p(-z)$. Somit ist die Ordnung $p = r + 2$ erreichbar.

4) Dass die Ordnung $p = r + 3$ (r gerade) nicht erreichbar ist, folgt nach Teil 1): $\beta_{r+2} < 0$. ∎

4.7 Asymptotische Entwicklung für die Mittelpunktsregel

In §4.2.3 wurde die Mittelpunktsregel

$$\eta_{j+2} = \eta_j + 2hf_{j+1} \qquad (f_{j+1} := f(x_{j+1}, \eta_{j+1}) \tag{4.7.1a}$$

eingeführt. Die Startwerte des Verfahrens werden mit dem Euler-Verfahren bestimmt:

$$\eta_0 = y_0, \quad \eta_1 = \eta_0 + hf_0. \tag{4.7.1b}$$

In §2.6 (vgl. (2.6.1a)) wurde eine asymptotische Entwicklung $\eta(x,h) = y(x) + e_p(x)h^p + e_{p+1}(x)h^{p+1} + \ldots + e_{q-1}h^{q-1} + e_q(x,h)h^q$ für Einschrittverfahren bewiesen. In Bemerkung 2.6.3 wurde eine Symmetriebedingung formuliert, eine Entwicklung $\eta(x,h) = y(x) + e_1(x)h^2 + e_2(x)h^4 + \ldots$ in h^2 nach sich zieht. Auch die Mittelpunktsregel ist symmetrisch, sodass man eine h^2-Entwicklung erwarten kann. Die Tatsache, dass die Mittelpunktsregel ein Zweischrittverfahren ist, spiegelt sich aber ein einem anderen Verhalten der asymptotische Entwicklung wieder. Die Wurzeln des charakteristischen Polynoms ψ sind ± 1 und die zugehörige Lösung der homogenen Differenzengleichungen demnach $\eta_j = a1^j + b\,(-1)^j$ $(a, b \in \mathbb{R})$. Entsprechend lautet der Ansatz für die asymptotische Entwicklung

$$\eta(x,h) = y(x) + \sum_{k=1}^{\ell-1} h^{2k}\left[u_k(x) + (-1)^{\frac{x-x_0}{h}}v_k(x)\right] + \mathcal{O}(h^{2\ell})\,. \tag{4.7.1c}$$

Man beachte, dass der Exponent $\frac{x-x_0}{h}$ ganzzahlig ist, da nur Argumente $x = x_j = x_0 + jh$ auftreten.

Satz 4.7.1 *Sei $f \in C^{2\ell}$ für ein $\ell \in \mathbb{N}$. Dann erfüllt die Lösung von (4.7.1a,b) die Darstellung (4.7.1c).*

Der Beweis findet sich bei Gragg [7].

Folgerung 4.7.2 *Die Extrapolation (in h^2) ist anwendbar für $x = x_0 + 2jh$. Also ist zum Beispiel eine Extrapolation aus $\eta(x_0 + 2h_0; h_0), \eta(x_0 + 2h_0; \frac{h_0}{2})$ und $\eta(x_0 + 2h_0; \frac{h_0}{3})$ möglich ($j = 2$, $j = 4$ und $j = 6$).*

5 Randwertaufgaben für gewöhnliche Differentialgleichungen

5.1 Aufgabenstellung, Theorie

Gegeben sei die Differentialgleichungen

$$y'(x) = f(x, y(x)) \qquad \text{in } a \le x \le b,\ y \in \mathbb{R}^n \tag{5.1.1}$$

Anstelle von n Anfangsbedingungen $y_i(x_0) = y_{i,0}$ $(1 \le i \le n)$ im Falle des Anfangswertproblems (1.1.2) werden α Randbedingungen am linken Randpunkt a und $\beta = n - \alpha$ Randbedingungen am rechten Randpunkt b vorgeschrieben:

$$\varphi_1(y(a)) = \varphi_2(y(a)) = \ldots = \varphi_\alpha(y(a)) = 0, \tag{5.1.2a}$$
$$\psi_1(y(b)) = \psi_2(y(b)) = \ldots = \psi_\beta(y(b)) = 0, \tag{5.1.2b}$$
$$a + \beta = n. \tag{5.1.2c}$$

Dabei sind $\varphi_i(y)$ und $\psi_j(y)$ nichtlineare Funktionen von $\mathbb{R}^n$ in $\mathbb{R}$.

Für $a = n$ und $\beta = 0$ erhält man das Anfangswertproblem (1.1.2) zurück (mit nichtlinearer impliziter Gleichung für y_0). Umgekehrt entspricht $a = 0$ und $\beta = n$ dem Anfangswertproblem in umgekehrter Richtung.

Die Randbedingungen (5.1.2a-c) können verallgemeinert werden:

$$r(y(a), y(b) = 0 \qquad \text{mit } r : \mathbb{R}^{2n} \to \mathbb{R}^n. \tag{5.1.3}$$

Die Formulierung (5.1.3) erlaubt zum Beispiel periodische Randbedingungen:

$$y_i(a) = y_i(b) \qquad \text{für } 1 \le i \le n.$$

Beispiel 5.1.1 *Die Differentialgleichung $u''(x) + c^2 u(x) = 0$ in $[a, b] = [0, 1]$ mit $c \neq 0$ und den Randbedingungen $u(0) = 0$ und $u(1) = 1$ wird als Differentialgleichungssystem erster Ordnung umgeschrieben: $y_1 = u$, $y_2 = u'$, $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ (vgl. Bemerkung 1.1.2c). Es entsteht*

$$y' = \begin{bmatrix} 0 & 1 \\ -c^2 & 0 \end{bmatrix} y, \qquad y_1(0) = 0,\ y_1(1) = 1.$$

Die allgemeine Lösung der Differentialgleichung für $c \neq 0$ ist

$$y(x) = \alpha \begin{pmatrix} \sin cx \\ c\cos cx \end{pmatrix} + \beta \begin{pmatrix} \cos cx \\ -c\sin cx \end{pmatrix} \qquad \text{mit } \alpha, \beta \in \mathbb{R}.$$

Die Bedingung $y_1(0) = 0$ erzwingt $\beta = 0$, während $y_1(1) = 1$ auf $\alpha = 1/\sin c$ führt, falls $\sin c \neq 0$. Man kann drei Fälle unterscheiden:

- $\sin c \neq 0$ *und* $c \neq 0$*: eine eindeutige Lösung* $y = \frac{1}{\sin c} \begin{pmatrix} \sin cx \\ c\cos cx \end{pmatrix}$ *existiert;*
- $\sin c = 0$ *und* $c \neq 0$*: es existiert keine Lösung;*
- $c = 0$*:* $y = \begin{pmatrix} x \\ 1 \end{pmatrix}$ *ist eindeutige Lösung.*

Im Falle von "$\sin c = 0,\ c \neq 0$*" hat das homogene Problem*

$$y' = \begin{bmatrix} 0 & 1 \\ -c^2 & 0 \end{bmatrix} y, \qquad y_1(0) = 0,\ y_1(1) = 0,$$

eine nichttriviale Lösung $y = \alpha \begin{pmatrix} \sin cx \\ c\cos cx \end{pmatrix}$ *($\alpha \in \mathbb{R}$ beliebig).*

Auch für das allgemeine Randwertproblem (5.1.1-3) gilt:

- Eine Lösung braucht nicht zu existieren.
- Bei nichtlinearen Randwertaufgaben können mehrere isolierte Lösungen existieren.

Ein Beispiel zum zweiten Punkt ist $y' = y^2 - 1$ mit der periodischen Randbedingung $y(0) = y(1)$. Es gibt genau die zwei Lösungen $y(x) = +1$ und $y(x) = -1$.

Eine hinreichende Bedingung für Existenz und Eindeutigkeit wird der Satz 5.1.3 enthalten. Wir stellen dazu die Randwertaufgabe[29]

$$-u'' + q(x,u) = 0, \qquad u(a) = u_a,\ u(b) = u_b \tag{5.1.4a}$$

auf. Dabei ist $q(x,u)$ als differenzierbar in u angenommen und

$$q_u(x,u) \geq 0 \qquad \text{für alle } u \in \mathbb{R} \tag{5.1.4b}$$

vorausgesetzt. Wir werden das Randwertproblem formal auf die Anfangswertaufgabe

$$-u'' + q(x,u) = 0, \qquad u(a) = u_a,\ u'(a) = \gamma \tag{5.1.4c}$$

zurückführen. Die Lösung von (5.1.4c) wird mit $u(x;\gamma)$ bezeichnet. Da $u(x;\gamma)$ nicht auf dem gesamten Intervall $I = [a,b]$ definiert zu sein braucht (vgl. Bemerkung 1.3.2), sei zunächst eine hinreichende Bedingung angegeben.

Übungsaufgabe 5.1.2 *Die Bedingung (5.1.4b) sei zu $0 \leq q_u(x,u) \leq q_{\max}$ für alle $x \in [a,b]$, $u \in \mathbb{R}$ verstärkt. Man zeige, dass die Lösung $u(x;\gamma)$ von (5.1.4c) für alle $\gamma \in \mathbb{R}$ in $[a,b]$ existiert.* Hinweis: *Man beachte $q(x,u(x)) = q(x,0) + \int_0^{u(x)} q_u(x,v)dv$ und leite $|q(x,u(x))| \leq A + B\,|u(x)|$ für geeignete Konstanten A, B ab. Anschließend zeige man, dass $|u(x;\gamma)| \leq U(x,\gamma)$, wobei man die Lösung $U(x,\gamma)$ der linearen Differentialgleichung $U'' = A + BU$, $U(a) = |u_a|$, $U'(a) = |\gamma|$ exakt löse.*

Satz 5.1.3 *Es existiere eine Lösung der Differentialgleichung aus (5.1.4c) auf $[a,b]$ mit $u(a) = u_a$ (keine Vorgabe bezüglich $u'(a)$ oder $u(b)$). Ferner sei q nach u differenzierbar und erfülle (5.1.4b). Dann ist die Randwertaufgabe (5.1.4a) eindeutig lösbar.*

Beweis. 1) Wir definieren $u(x;\gamma)$ als die Lösung des Anfangswertproblems und notieren den rechten Randwert als

$$\Phi(\gamma) := u(b,\gamma). \tag{5.1.5}$$

Nach Voraussetzung ist Φ für ein $\gamma = \gamma_0$ und aus Stetigkeitsgründen (vgl. Satz 1.3.6) in einer Umgebung von γ_0 definiert.

2) Wir wollen zeigen, dass $\Phi(\gamma)$ monoton wachsend ist und als Bildbereich $(-\infty,\infty)$ besitzt (d.h. $\Phi(\gamma)$ ist in einer Umgebung $(\gamma',\gamma'') \subset (-\infty,\infty)$ von γ_0 definiert und $\lim_{\gamma \searrow \gamma'} \Phi(\gamma) = -\infty$, $\lim_{\gamma \nearrow \gamma''} \Phi(\gamma) = \infty$). Gemäß Übungsaufgabe 1.3.7 ist u nach γ differenzierbar. Auf Grund der Gleichung (5.1.4a) gilt dies auch für u''. Differentiation von $u''(x,\gamma) = q(x,u(x,\gamma))$ nach γ liefert

$$u''_\gamma(x,\gamma) = q_u(x,u(x,\gamma))\,u_\gamma(x,\gamma), \qquad u_\gamma(a,\gamma) = 0,\ u'_\gamma(a,\gamma) = 1. \tag{5.1.6}$$

Für festes γ erfüllt $v(x) := u_\gamma(x,\gamma)$ die Differentialgleichung

$$v'' = Q(x,\gamma)v, \qquad v(a) = 0,\ v'(a) = 1,$$

wobei $Q(x,\gamma) := q_u(x,u(x,\gamma)) \geq 0$. Die lineare Differentialgleichung $v'' = Qv$ hat eine Lösung für alle $x \in [a,b]$. Die Lösung erfüllt für alle $x > a$ die Ungleichungen $v > 0$, $v' \geq 1$, $v'' \geq 0$. Damit ist $v(x) \geq x - a$ und speziell $v(b) \geq b - a$. Letzteres zeigt

$$\Phi_\gamma(\gamma) = u_\gamma(b,\gamma) = v(b) \geq b - a > 0.$$

[29] Eine allgemeinere rechte Seite in $-u'' + q(x,u) = f(x)$ kann o.B.d.A durch 0 ersetzt werden, da q in $q - f$ abgeändert werden kann, ohne dass sich q_u ändert.

Hieraus schließt man $\Phi(\gamma) \geq \Phi(0) + \gamma\,(b-a)$ und $\Phi(-\gamma) \leq \Phi(0) - \gamma\,(b-a)$. Damit enthält der Bildbereich von Φ den gesamten Bereich $(-\infty, \infty)$. Wegen der Monotonie gibt es genau ein γ^* mit $\Phi(\gamma^*) = u_b$. Also ist $u(b, \gamma^*)$ die einzige Lösung des Randwertproblems. ∎

Entsprechende Aussagen kann man auch für allgemeinere Randbedingungen und für Differentialgleichungen der Form

$$-\left(p(x)u'\right)' + q(x, u, u') = 0$$

mit $p > 0$ unter geeigneten Voraussetzungen beweisen (vgl. [6, Kap. 7]).

5.2 Diskretisierung durch Differenzenverfahren

Wir wählen ein äquidistantes Gitter zur Schrittweite $h = (b-a)/n$, $n \in \mathbb{N}$:

$$x_0 = a \leq x_1 = a + h \leq \ldots \leq x_n = b, \qquad x_\nu = a + \nu h.$$

Die Diskretisierung von $-u'' + q(x,u) = 0$ ersetzt in allen inneren Gitterpunkten $x_\nu \in (a,b)$ die zweite Ableitung durch die dividierte zweite Differenz. Dabei bezeichnet u_ν wie üblich die Näherung zu $u(x_\nu)$:

$$\frac{1}{h^2}\left\{-u_{\nu-1} + 2u_\nu - u_{\nu+1}\right\} + q(x_\nu, u_\nu) = 0 \quad \text{für } 1 \leq \nu \leq n-1. \tag{5.2.1}$$

Bemerkung 5.2.1 *a) Falls $u \in C^4([a,b])$, beträgt der Konsistenzfehler $\mathcal{O}(h^2)$, d.h. die exakten Werte $u_\nu := u(x_\nu)$ erfüllen die Gleichungen*

$$\frac{1}{h^2}\left\{-u_{\nu-1} + 2u_\nu - u_{\nu+1}\right\} + q(x_\nu, u_\nu) = \varepsilon_\nu \quad \textit{für } 1 \leq \nu \leq n-1$$

mit $|\varepsilon_\nu| \leq const * h^2$.

b) Hinreichend für die Glattheit $u \in C^4([a,b])$ sind: (i) es gibt überhaupt eine Lösung in $[a,b]$ und (ii) $q(x, u(x)) \in C^2([a,b])$.

Der Beweis benutzt die Taylor-Entwicklung von $u_{\nu\pm1} = u(x_\nu \pm 1)$ mit Restglied $h^4 u''''(x_\nu \pm \vartheta_\pm h)$. Verwendet man dagegen die Integralformulierung des Restgliedes, sieht man, dass schon die Lipschitz-Stetigkeit der dritten Ableitungen von u ausreicht.

Ergänzt man (5.2.1) um die Randwertvorgaben, so erhält man das diskrete nichtlineare Randwertproblem

$$u_0 = u_a, \qquad u_n = u_b, \tag{5.2.2a}$$

$$\frac{-u_{\nu-1} + 2u_\nu - u_{\nu+1}}{h^2} + q(x_\nu, u_\nu) = 0 \quad \text{für } 1 \leq \nu \leq n-1. \tag{5.2.2b}$$

Der nachfolgende Satz zeigt, dass sich die Konsistenzordnung $\mathcal{O}(h^2)$ auch auf die Konvergenzordnung überträgt.

Satz 5.2.2 *$u \in C^4([a,b])$ sei Lösung der Randwertaufgabe (5.1.4a) und q erfülle (5.1.4b) und $\frac{\partial}{\partial u} q \in C^1([a,b] \times \mathbb{R})$. Dann existiert die diskrete Lösung $(u_\nu)_{\nu=0}^n$ eindeutig. Ferner gibt es ein C, sodass*

$$|u(x_\nu) - u_\nu| \leq Ch^2 \qquad \textit{für alle } 0 \leq \nu \leq n. \tag{5.2.3}$$

Beweis. a) Der Beweis zur eindeutigen Lösbarkeit wird auf Lemma 5.3.2i-iii verschoben.

b) Wir setzen $u_h := (u_\nu)_{\nu=0}^n \in \mathbb{R}^{n+1}$ und schreiben (5.2.2a,b) als

$$F_h(u_h) = 0 \qquad \text{mit } F_h : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}. \tag{5.2.4}$$

$u_h^{\text{exakt}} := (u(x_\nu))_{\nu=0}^n$ sei die Beschränkung der exakten Lösung auf die Gitterpunkte. Gemäß Bemerkung 5.2.1 können wir den Konsistenzfehler in der folgenden Form schreiben:

$$F_h(u_h^{\text{exakt}}) = h^2 c_h \qquad \text{mit } c_h = (c_\nu)_{\nu=0}^n \text{ und } |c_\nu| \leq K. \tag{5.2.5}$$

Die Gleichungen (5.2.4) und (5.2.5) differieren um $\mathcal{O}(h^2)$. Es bleibt zu zeigen, dass sich auch die zugehörigen Lösungen um $\mathcal{O}(h^2)$ unterscheiden. Nach Diskussion der Stabilität von F_h im nachfolgenden Unterkapitel wird dies Resultat von Satz 5.3.3 geliefert. ∎

5.3 Stabilitätsanalysis des Differenzenverfahrens

Im Folgenden setzen wir $N := n+1$ ($h = 1/n$) und untersuchen das Verhalten der Familie $\{F_h : N \in \mathbb{N}\}$ von Abbildungen.

Definition 5.3.1 *Eine Familie von (nichtlinearen) Abbildungen $F_h : \mathbb{R}^N \to \mathbb{R}^N$ heißt* stabil *(bezüglich der Norm $\|\cdot\|$), falls es ein $C > 0$ gibt, sodass*

$$\|F_h(u) - F_h(v)\| \geq \frac{1}{C}\|u - v\| \qquad \text{für alle } u, v \in \mathbb{R}^N. \tag{5.3.1}$$

Lemma 5.3.2 *Es sei die Stabilität (5.3.1) vorausgesetzt. Dann gelten die Aussagen:*
(i) $F_h(u_h) = y$ ($y \in \mathbb{R}^N$ beliebig) hat höchstens eine Lösung. (Damit existiert die Umkehrabbildung F_h^{-1} auf $Y_h := Bild(F_h)$.)
(ii) Die Umkehrabbildung $F_h^{-1} : Y_h \to \mathbb{R}^N$ ist Lipschitz-stetig mit der Lipschitz-Konstanten C aus (5.3.1).
(iii) Wenn $F_h \in C^1(\mathbb{R}^N)$, existiert die Umkehrabbildung F_h^{-1} auf $Y_h = \mathbb{R}^N$.

Beweis. a) Seien u, v zwei Lösungen: $F_h(u) = F_h(v) = y$. Aus (5.3.1) schließt man $0 = \|F_h(u) - F_h(v)\| \geq \frac{1}{C}\|u - v\|$, also $u = v$.

b) Zu beliebigen $y, z \in Y_h \subset \mathbb{R}^N$ seien $u := F_h^{-1}(y)$ und $v := F_h^{-1}(z)$ die nach (i) eindeutigen Lösungen. Dann ist

$$\left\|F_h^{-1}(y) - F_h^{-1}(z)\right\| = \|u - v\| \underset{(5.3.1)}{\leq} C\|F_h(u) - F_h(v)\| = \|y - z\|.$$

c) $y \in \mathbb{R}^N$ sei beliebig. Ferner sei $y_0 := F_h(0)$ das Bild des Null-Vektors. Wir untersuchen nun die Differentialgleichung

$$\frac{d}{dt}u(t) = F_u(u(t))^{-1}(y - y_0) \qquad \text{mit dem Anfangswert } u(0) = 0 \in \mathbb{R}^N. \tag{5.3.2}$$

Dabei ist $F_u = \frac{d}{du}F_h$ die Jacobi-Matrix. Dass die Inverse von F_u existiert, folgt aus der Ungleichung

$$\begin{aligned}\|F_u(u)w\| &= \left\|\lim_{\varepsilon\to 0}\frac{F(u+\varepsilon w) - F(u)}{\varepsilon}\right\| \\ &= \lim_{\varepsilon\to 0} \underbrace{\frac{1}{\varepsilon}\|F(u+\varepsilon w) - F(u)\|}_{\frac{1}{\varepsilon}\|F(u+\varepsilon w)-F(u)\| \underset{(5.3.1)}{\geq} \frac{1}{C\varepsilon}\|(u+\varepsilon w)-u\| = \frac{1}{C}\|w\|} \geq \frac{1}{C}\|w\|,\end{aligned}$$

die die Matrixnorm-Abschätzung $\left\|F_u^{-1}(u)\right\| \leq C$ gleichmäßig in u beweist. Damit existiert die Lösung des Anfangswertproblems (5.3.2) für alle $t \geq 0$. Die Kettenregel liefert

$$\frac{d}{dt}F_h(u(t)) = F_u(u(t))\frac{du}{dt} \underset{(5.3.2)}{=} F_u(u(t))F_u(u(t))^{-1}(y - y_0) = y - y_0.$$

Hieraus folgt mit $F_h(u(0)) \underset{(5.3.2)}{=} F_h(0) = y_0$, dass

$$F_h(u(t)) = F_h(u(0)) + \int_0^t \frac{d}{dt}F_h(u(s))ds = y_0 + t(y - y_0).$$

Damit hat $F_h(v) = y$ die Lösung $v = u(1)$ von. Da $y \in \mathbb{R}^N$ beliebig war, existiert die Umkehrabbildung F_h^{-1} auf $Y_h = \mathbb{R}^N$. ∎

Satz 5.3.3 *Wir setzen voraus: (i) F_h sei stabil bezüglich $\|\cdot\|$, (ii) u_h sei Lösung von $F_h(u_h) = 0$, (iii) F_h sei konsistent von der Ordnung k, d.h.*

$$F_h(u_h^{\text{exakt}}) = h^k c_h \qquad \textit{mit } \|c_h\| \leq \textit{const}. \tag{5.3.3}$$

Dann gilt

$$\left\|u_h^{\text{exakt}} - u_h\right\| = \mathcal{O}(h^k).$$

Beweis. $\left\|u_h^{\text{exakt}} - u_h\right\| = \left\|F_h^{-1}(h^k c_h) - F_h^{-1}(0)\right\| \le C \left\|h^k c_h - 0\right\| = \mathcal{O}(h^k)$. ∎

Vor dem Beweis des nächsten Lemmas sei an den Zwischenwertsatz der Differentialrechnung im vektorwertigen Fall erinnert.

Bemerkung 5.3.4 *Sei $G : \mathbb{R}^N \to \mathbb{R}^N$ differenzierbar und $g_i : \mathbb{R}^N \to \mathbb{R}^1$ bezeichnet die i-te Komponente.*
a) Für alle $u, v \in \mathbb{R}^N$ gilt $G(u) - G(v) = \tilde{G}' \cdot (u - v)$, wobei die $N \times N$-Matrix $\tilde{G}'$ die Zeilen $\tilde{g}_i' = \operatorname{grad} g_i(u + \vartheta_i (v - u)$ mit geeigneten Zwischenwert $\vartheta_i \in (0, 1)$ besitzt.
b) Es gilt $\|G(u) - G(v)\|_\infty \le \max\{\|G_u(u + \vartheta (v - u))\|_\infty : 0 \le \vartheta \le 1\} \|u - v\|_\infty$.

Beweis. a) $G_i(u) - G_i(v) = \langle \tilde{g}_i', u - v\rangle$ gilt nach dem üblichen Zwischenwertsatz der Differentialrechnung. Hieraus folgt $G(u) - G(v) = \tilde{G}' \cdot (u - v)$.

b) Nach Definition der Zeilensummennorm gilt

$$\|\tilde{g}_i'\|_1 := \sum_{j=1}^{N} |\tilde{g}_{i,j}'| \le \|G_u(u + \vartheta_i (v - u))\|_\infty \le \max\{\|G_u(u + \vartheta (v - u))\|_\infty =: \gamma.$$

Damit folgt

$$|G_i(u) - G_i(v)| = |\langle \tilde{g}_i', u - v\rangle| \le \|\tilde{g}_i'\|_1 \|u - v\|_\infty \le \gamma \|u - v\|_\infty$$

für alle i, also $\|G(u) - G(v)\|_\infty \le \gamma \|u - v\|_\infty$. ∎

Lemma 5.3.5 *F_h sei gemäß (5.2.2a,b) definiert. Es gelte (5.1.4b): $q_u \ge 0$. Dann ist F_h stabil bezüglich der Maximumnorm $\|\cdot\|_\infty$.*

Beweis. Die Dimension ist $N = n + 1$. Die Ableitung von F_h die folgende, in Blockgestalt wiedergegebene Jacobi-Matrix

$$\frac{dF_h(u)}{du} = \begin{bmatrix} 1 & & \\ & A & \\ & & 1 \end{bmatrix}$$

$$\text{mit } A = B + C,\ B = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix},\ C = \operatorname{diag}\{q_u(x_\nu, u_\nu)_{\nu=1}^{n-1}\}$$

(nicht angeschriebene Matrixeinträge sind Nullen). Im Sinne einer komponentenweisen Ungleichung gilt $\frac{dF_h}{du} \ge \hat{B} := \begin{bmatrix} 1 & & \\ & B & \\ & & 1 \end{bmatrix}$ wegen $q_u \ge 0$. Die Untermatrix B ist invertierbar, damit auch $\hat{B}$. Standardüberlegungen führen auf $\left(\frac{dF_h}{du}\right)^{-1} \le \hat{B}^{-1}$ und die entsprechende Ungleichung $\left\|(\frac{dF_h}{du})^{-1}\right\|_\infty \le \|\hat{B}^{-1}\|_\infty$ für die Zeilensummennorm (vgl. [10, Satz 4.3.17]). Es ist $\|\hat{B}^{-1}\|_\infty = \max\{1, \|B^{-1}\|_\infty\}$. Zum Beweis von $\|B^{-1}\|_\infty \le \frac{1}{8}(b - a)^2$ sei auf [10, Satz 4.3.16] verwiesen (der dortige Vektor $\mathbf{w}$ hat die Koeffizienten $w_\nu = \frac{1}{8}(b - a)^2 - \frac{1}{2}\left(\frac{a+b}{2} - x_\nu\right)^2$).

Nun verwenden wir Bemerkung 5.3.4b mit $G := F_h^{-1}$ angewandt auf $F_h(u), F_h(v)$. Dabei nutzen wir $DF_h^{-1} = (\frac{dF_h}{du})^{-1}$ und $\left\|(\frac{dF_h}{du})^{-1}\right\|_\infty \le \|\hat{B}^{-1}\|_\infty$ für beliebige Argumente:

$$\|u - v\|_\infty \le \|\hat{B}^{-1}\|_\infty \|F_h(u) - F_h(v)\|_\infty \le C \|F_h(u) - F_h(v)\|_\infty$$

mit $C := \max\{1, \frac{1}{8}(b - a)^2\}$. ∎

5.4 Iterative Lösung der Differenzengleichungen

Die Gleichungen (5.2.2a,b) stellen ein nichtlineares System dar. Eine mögliche iterative Lösung kann wie folgt aussehen. Als Startvektor u_h^0 wird ein beliebiger Vektor mit $u_0^0 = u_a$, $u_n^0 = u_b$ gewählt. Gegeben die m-te Iterierte u_h^m, wird die $m+1$-te Iterierte u_h^{m+1} definiert durch $u_h^{m+1} = \left(u_\nu^{m+1}\right)_{\nu=0}^n$ mit

$$u_0^{m+1} := u_a, \quad u_n^{m+1} := u_b, \quad u_\nu^{m+1} + h^2 q(x_\nu, u_\nu^{m+1}) = \frac{1}{2}\left[u_{\nu-1}^m + u_{\nu+1}^m\right] \quad \text{für } 1 \le \nu \le n-1. \tag{5.4.1}$$

Jede der $n-1$ letzten Gleichungen ist eine skalare nichtlineare Gleichung vom Typ $w + h^2 q(x_\nu, w) = r$ ($w, r \in \mathbb{R}$). Wieder sichert $q_u \ge 0$ die Lösbarkeit. Im Prinzip lässt sich diese Gleichung mittels des Newton-Verfahrens[30] lösen.

Lemma 5.4.1 *Die in (5.4.1) definierte Iteration konvergiert gegen die exakte Lösung von (5.2.2a,b).*

Beweis. Seien $d_\nu^m := u_\nu^m - u_\nu$ (u_ν Lösung von (5.2.2a,b)) die Iterationsfehler. Subtraktion der Gleichung $u_\nu + h^2 q(x_\nu, u_\nu) = \frac{1}{2}[u_{\nu-1} + u_{\nu+1}]$ von (5.4.1) liefert

$$\left(1 + h^2 q_u(x_\nu, \tilde{u}_\nu^{m+1})\right) d_\nu^{m+1} = \frac{1}{2}\left[d_{\nu-1}^m + d_{\nu+1}^m\right] \quad \text{für } 1 \le \nu \le n-1,$$

wobei $\tilde{u}_\nu^{m+1}$ aus dem Zwischenwertsatz der Differentialgleichung resultiert. Wegen $q_u \ge 0$ folgt

$$\left|d_\nu^{m+1}\right| \le \frac{1}{2}\left|d_{\nu-1}^m + d_{\nu+1}^m\right|.$$

Wir definieren die gewichtete Maximumsnorm

$$\|u\| := \max\{\omega_\nu |u_\nu| : 1 \le \nu \le n-1\} \quad \text{mit } \omega_\nu := 1/\sin\frac{\nu\pi}{n} > 0.$$

Da der Sinus in $[0,\pi]$ konvex ist, gilt $\frac{1}{2}\left[\sin\frac{(\nu+1)\pi}{n} + \sin\frac{(\nu-1)\pi}{n}\right] < \sin\left(\frac{\nu\pi}{n}\right)$, sodass $\frac{1}{2}\left(\frac{1}{\omega_{\nu+1}} + \frac{1}{\omega_{\nu-1}}\right) < \frac{1}{\omega_\nu}$ (diese Ungleichung gilt auch für $\nu = 1$ oder $\nu = n-1$, wenn $\frac{1}{\omega_0} = \frac{1}{\omega_n}$ als 0 interpretiert wird). Wir setzen

$$\rho := \max\left\{\frac{\omega_\nu}{2}\left(\frac{1}{\omega_{\nu-1}} + \frac{1}{\omega_{\nu+1}}\right) : 1 \le \nu \le n-1\right\} < 1$$

und schließen, dass

$$\omega_\nu \left|d_\nu^{m+1}\right| \le \frac{\omega_\nu}{2}\left|d_{\nu-1}^m + d_{\nu+1}^m\right| \le \frac{\omega_\nu}{2}\left(\frac{\omega_{\nu-1}\left|d_{\nu-1}^m\right|}{\omega_{\nu-1}} + \frac{\omega_{\nu+1}\left|d_{\nu+1}^m\right|}{\omega_{\nu+1}}\right) \le \frac{\omega_\nu}{2}\left(\frac{\|d^m\|}{\omega_{\nu-1}} + \frac{\|d^m\|}{\omega_{\nu+1}}\right) \le \rho\,\|d^m\|.$$

Maximierung über alle ν liefert die Konvergenz: $\|d^{m+1}\| \le \rho\|d^m\|$ (d.h. $\|d^m\| \le \rho^m \left\|d^0\right\|$). ∎

In der Praxis sind zwei Verbesserungen nötig: Das Newton-Verfahren zur Lösung von $w + h^2 q(x_\nu, w) = r$ wird durch einfachere genäherte Methoden ersetzt, und die Jacobi-Iteration $u^{m+1} \mapsto u^m$ sollte durch schneller konvergente Iterationen ersetzt werden (vgl. [11]).

5.5 Mehrzielmethode

In §5.2 war das Randwertproblem zuerst diskretisiert worden und dann z.B. mit iterativen Verfahren gelöst worden. Man kann versuchen, auch hier die Verfahren zur Lösung der *Anfangswert*aufgaben mit ihrer adaptiven Schrittweitenwahl einzusetzen. Dann würde die Diskretisierung als Teil des Lösungsverfahrens erfolgen.

[30] Sir Isaac Newton, geb. 4. Jan. 1643 in Woolsthorpe, gest. 31. März 1727 in London

5.5.1 Einfaches Schießverfahren

Beim Beweis der Existenz einer Lösung der Randwertaufgabe wurde in (5.1.5) die Funktion $\Phi(\gamma) = u(b, \gamma)$ verwendet. Hier war u die Lösung zu den Anfangswerten $u(a) = u_a,\ u'(a) = \gamma$. Es stellte sich heraus, dass γ so gewählt werden kann, dass die gewünschte Bedingung $u(b) = u_b$ getroffen wird. Dieses Vorgehen soll jetzt algorithmisch genutzt werden.

Wenn die Differentialgleichung linear ist (zum Beispiel $q(x, u) = \alpha(x)u + \beta(x)$ in (5.1.4c)), braucht man nur zwei Lösungen u_I und u_{II} von (5.1.4c) zu unterschiedlichen Anfangwerten $\gamma_I := u_I'(a) \neq \gamma_{II} := u_{II}'(a)$ zu ermitteln. Wenn das Randwertproblem überhaupt lösbar ist, unterscheiden sich die Endwerte $u_I(b), u_{II}(b)$. Ist die Bedingung $u(b) = u_b$ gestellt, so prüft man nach, dass die Linearkombination

$$u(x) := \frac{u_{II}(b) - u_b}{u_{II}(b) - u_I(b)} u_I(x) + \frac{u_b - u_I(b)}{u_{II}(b) - u_I(b)} u_{II}(x)$$

die Randwerte $u(a) = u_a$, $u(b) = u_b$ besitzt und wieder die Differentialgleichung erfüllt. Indem man die Ein- oder Mehrschrittverfahren aus §§2-4 verwendet, erhält man Approximationen von $u_I(x), u_{II}(x)$, die entsprechend linear kombiniert werden können.

Im nichtlinearen Fall wendet man das Newton-Verfahren auf die nichtlineare Gleichung

$$F(\gamma) := u(b, \gamma) - u_b = 0$$

an. Die Ableitung F' lässt sich auf verschiedene Weisen annähern:

- "numerische Differentiation": Nachdem man $u(b, \gamma)$ zu γ berechnet hat, wird das Anfangswertproblem (5.1.4c) mit $\gamma + \varepsilon$ für ein kleines $\varepsilon \neq 0$ berechnet und $F'(\gamma)$ durch $(u(b, \gamma + \varepsilon) - u(b, \gamma))/\varepsilon$ genähert.
- "analytische Differentiation": Es ist möglich, neben der Funktionsauswertung auch die Ableitung mit ähnlichem Aufwand exakt zu berechnen (vgl. [8]).
- "Linearisiertes Problem": Die Lösung $u_\gamma(\cdot, \gamma)$ von (5.1.6) liefert $F'(\gamma) = u_\gamma(b, \gamma)$.

Die Problematik des Schießverfahrens kann bereits am linearen Randwertproblem

$$u'' - 100\, u = 0 \text{ in } [0, 100], \qquad u(0) = 1, \quad u(100) = 2$$

studiert werden. Das Anfangswertproblem $u'' - 100\, u = 0$ mit $u(0) = 1$ und $u'(0) = \gamma$ hat die Lösung $\frac{10+\gamma}{20} e^{10x} + \frac{10-\gamma}{20} e^{-10x}$. Für $\gamma = -10$ ergibt sich $u(100, \gamma) = e^{-1000} = 5.076 \times 10^{-435}$, sodass mit Unterlauf bestensfalls der Wert 0 zu erwarten ist. Für einen anderen Wert von γ ist wegen $e^{1000} = 1.9701 \times 10^{434}$ ein Überlauf zu erwarten. Damit ist die Auswertung von $F(\gamma) = u(b, \gamma) - u_b$ praktisch unmöglich, während die Lösung

$$\frac{2 - \exp(-1000)}{1 - \exp(-2000)} e^{10(x-100)} + \frac{1 - 2\exp(-1000)}{1 - \exp(-2000)} e^{-10x} \approx 2e^{10(x-100)} + e^{-10x}$$

der Randwertaufgabe in dem überschaubaren Wertebereich $[0, 2]$ verläuft. Während also die Randwertaufgabe wohlkonditioniert ist, ist das zugehörige Anfangswertproblem schlecht gestellt.

Das Beispiel wäre weniger dramatisch ausgefallen, wenn das Intervall $[0, 100]$ durch $[0, 1]$ ersetzt worden wäre. Entsprechend wird die Abhilfe gewählt. Im anschließenden Mehrzielverfahren wird das Gesamtintervall in kleinere unterteilt und das Schießverfahren auf den kleineren Intervallen durchgeführt.

5.5.2 Mehrzielmethode

Wir kehren wieder zu dem System (5.1.1) von n Gleichungen erster Ordnung in $[a, b]$ zurück und stellen die allgemeine Randbedingung (5.1.3):

$$\begin{aligned} y'(x) &= f(x, y(x)) && \text{in } a \le x \le b,\ y \in \mathbb{R}^n, \\ r(y(a), y(b)) &= 0 && \text{mit } r : \mathbb{R}^{2n} \to \mathbb{R}^n. \end{aligned}$$

Bei der Mehrzielmethode wird eine Intervallzerlegung

$$a = x_1 < x_2 < \ldots < x_m = b$$

vorgegeben (vgl. [16, Kap. 7.3.5]). Zusammen mit (noch zu bestimmenden) Anfangswerten $s_k \in \mathbb{R}^n$ erhält man die m Anfangwertprobleme

$$y'(x) = f(x, y(x)) \quad \text{in } x_k \le x \le x_{k+1} \qquad \text{mit } y(x_k) = s_k \qquad (\text{für alle } 1 \le k \le m-1). \tag{5.5.1}$$

Hier sind die Intervalle $[x_k, x_{k+1}]$ als hinreichend klein angenommen, sodass die oben diskutierten Konditionsprobleme nicht auftreten. Die Lösungen seien als

$$y(x; x_k, s_k) \qquad \text{für } x \in [x_k, x_{k+1}] \text{ Lösung von (5.5.1)}$$

notiert. Bei allgemeiner Wahl der s_k ist

$$y(x_{k+1}; x_k, s_k) \neq y(x_{k+1}; x_{k+1}, s_{k+1}) = s_{k+1}$$

zu erwarten. Um zu einer global glatten Funktion zu kommen, sind die Gleichungen

$$y(x_{k+1}; x_k, s_k) = s_{k+1}$$

als neue Bedingungen zu stellen. Da wir diese Gleichung auch für $k = m-1$ aufschreiben wollen, wird s_m als eine eigene Variable eingeführt.

Die Freiheitsgrade sind die $m \cdot n$ Größen des Vektors

$$s := (s_1, s_2, \ldots, s_m) \in \mathbb{R}^{m \cdot n}.$$

Die Bedingungen setzen sich aus der eigentlichen Randbedingung (5.1.3) in der Form $r(s_1, s_m) = 0$ und den $m-1$ Übergangsbedingungen $y(x_{k+1}; x_k, s_k) = s_{k+1}$ $(1 \le k \le m-1)$ zusammen. Die Funktion $F(s) : \mathbb{R}^{m \cdot n} \to \mathbb{R}^{m \cdot n}$ fasst alle Bedingungen zusammen:

$$F(s) := \begin{bmatrix} y(x_2; x_1, s_1) - s_2 \\ \vdots \\ y(x_m; x_{m-1}, s_{m-1}) - s_m \\ r(s_1, s_m) \end{bmatrix} = 0$$

Die Mehrzielmethode besteht aus der Lösung von $F(s) = 0$ mit Hilfe des Newton-Verfahrens. Die Auswertung von F erfordert die Berechnung der Anfangswertaufgaben $y(x_{k+1}; x_k, s_k)$ mit Methoden aus §§2-4. Zur Berechnung der Jacobi-Matrix F' beachte man dessen Blockstruktur

$$F'(s) = \begin{bmatrix} G_1 & -I & & & & \\ & G_2 & -I & & & \\ & & \ddots & \ddots & & \\ & & & G_{m-2} & -I & \\ & & & & G_{m-1} & -I \\ A & & & & & B \end{bmatrix} \qquad \text{mit} \quad \begin{cases} G_k := \dfrac{\partial}{\partial s_k} y(x_{k+1}; x_k, s_k), \\ A := \dfrac{\partial}{\partial s_1} r(s_1, s_m), \\ B := \dfrac{\partial}{\partial s_m} r(s_1, s_m). \end{cases}$$

Die Berechnung der G_k kann einer der drei in §5.5.1 genannten Möglichkeiten folgen.

Danksagung. Die Abbildungen entstammen einem Vorlesungsskript von Herrn PD Dr. M. Melenk, dem hiermit herzlich für die Unterstützung gedankt sei.

Literatur

[1] R. D. Grigorieff: *Numerik gewöhnlicher Differentialgleichungen, Band 1.* Teubner-Verlag, Stuttgart, 1972.

[2] R. D. Grigorieff: *Numerik gewöhnlicher Differentialgleichungen, Band 2: Mehrschrittverfahren.* Teubner-Verlag, Stuttgart. 1977.

[3] P. Deuflhard und F. Bornemann: *Numerische Mathematik II. Integration gewöhnlicher Differentialgleichungen.* Zweite Auflage, W de Gruyter, Berlin, 2002.

[4] J. R. Dormand und P. J. Prince: A family of embedded Runge-Kutta formulae. *J. Comput. Appl. Math.* **6** (1980) 19-26.

[5] J. R. Dormand und P. J. Prince: Higher order embedded Runge-Kutta formulae. *J. Comput. Appl. Math.* **7** (1981) 67-75.

[6] W. Gautschi: *Numerical Analysis. An Introduction.* Birkhäuser, Boston, 1997 (darin: §§5-6).

[7] W. Gragg: On extrapolation algorithms for ordinary initial value problems. *J. SIAM Numer. Anal. Ser. B* **2** (1965) 384-403.

[8] A. Griewank: *Evaluating Derivatives. Principles and Techniques for Algorithmic Differentiation.* SIAM, Philadelphia, 2000.

[9] Th. Kaluza: Über Koeffizienten reziproker Potenzreihen. *Math. Zeitschrift* **28** (1928) 161-170.

[10] W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen.* Zweite Auflage, Teubner-Verlag, Stuttgart, 1996.

[11] W. Hackbusch: *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* Zweite Auflage, Teubner-Verlag, Stuttgart, 1993.

[12] W. Hackbusch: *Der Stabilitätsbegriff in der Numerik.* Lecture Notes 20, Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig, 2003.

[13] E. Hairer, S. P. Nørsett und G. Wanner: *Solving Ordinary Differential Equations. Part 1.* Zweite Auflage, Springer-Verlag, Berlin, 1993.

[14] W. Luther, K. Niederdrenk, F. Reutter und H. Yserentant: *Gewöhnliche Differentialgleichungen.* Vieweg, Braunschweig, 1987.

[15] J. Stoer: *Einführung in die Numerische Mathematik I.* Achte Auflage, Springer-Verlag, Berlin, 1999.

[16] J. Stoer und R. Bulirsch: *Einführung in die Numerische Mathematik II.* Vierte Auflage, Springer-Verlag, Berlin, 2000.

Index