Numerik gewöhnlicher Differentialgleichungen

Steffen Börm

Stand 16. April 2010

Inhaltsverzeichnis

1 Einleitung

Bevor wir uns der Analyse und numerische Behandlung von gewöhnlichen Differentialgleichungen, insbesondere von Anfangswertproblemen, zuwenden, sollen zunächst einige mehr oder weniger einfache Probleme vorgestellt werden, die sich mit Hilfe derartiger Gleichungen beschreiben lassen.

1.1 Federpendel

Ein sehr einfaches Beispiel für ein Anfangswertproblem ist das abstrakte Federpendel: Es besteht aus einer Masse m, die mittels einer Feder mit dem Nullpunkt verbunden ist und nach oben oder unten ausgelenkt werden kann. Die Auslenkung zu einem bestimmten Zeitpunkt t bezeichnen wir mit $u(t)$.

Falls die Masse sich nicht im Nullpunkt befindet, ist die Feder angespannt und übt eine Kraft aus, die die Masse in den Nullpunkt zurückzieht. Im abstrakten Fall ist diese Kraft $F(t)$ durch das Gesetz

$$F(t) = -cu(t)$$

gegeben, wobei c die *Federkonstante* ist.

Gemäß den Newtonschen Axiomen bewirkt die Kraft F eine Beschleunigung $a(t)$ der Masse, die proportional zum Kehrwert von m ist, es gilt also

$$F(t) = ma(t), \qquad a(t) = \frac{1}{m}F(t).$$

Die Beschleunigung ist die Ableitung der Geschwindigkeit $v(t)$ der Masse, und die Geschwindigkeit ist die Ableitung der Auslenkung $u(t)$, so dass wir die Gleichungen

$$u'(t) = v(t), \qquad v'(t) = a(t) = \frac{1}{m}F(t) = -\frac{c}{m}u(t)$$

erhalten. Indem wir $u(t)$ und $v(t)$ zu einem Vektor

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

zusammenfassen, erhalten wir die kompakte Schreibweise

$$y'(t) = \begin{pmatrix} 0 & 1 \\ -c/m & 0 \end{pmatrix} y(t),$$

mit der die Bewegungen des abstrakten Pendels vollständig beschrieben werden können.

Wenn die Auslenkung $u(t_0)$ und die Geschwindigkeit $v(t_0)$ zu einem bestimmten Zeitpunkt t_0 bekannt sind, können wir Auslenkung und Geschwindigkeit zu jedem späteren Zeitpunkt $t \geq t_0$ als Lösung des Systems

$$y(t_0) = \begin{pmatrix} u(t_0) \\ v(t_0) \end{pmatrix}, \qquad y'(t) = \begin{pmatrix} 0 & 1 \\ -c/m & 0 \end{pmatrix} y(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq t_0}$$

bestimmen. Ein derartiges Gleichungssystem, bei dem die Lösung zu einem Anfangszeitpunkt t_0 und die Ableitung der Lösung zu jedem Zeitpunkt t bekannt sind, nennt man *Anfangswertproblem*.

In unserem Fall haben wir es mit einem besonders einfachen System zu tun, dass sich analytisch lösen lässt. Da $y'(t)$ und $y(t)$ über die von t unabhängige Matrix

$$A := \begin{pmatrix} 0 & 1 \\ -c/m & 0 \end{pmatrix}$$

verbunden sind, ist

$$y(t) := e^{tA} y(t_0) \qquad \text{für alle } t \in \mathbb{R}_{\geq t_0}$$

eine Lösung des Systems, denn für diese Funktion gilt

$$y(t_0) = e^{0A} y(t_0) = y(t_0), \qquad y'(t) = Ae^{tA} y(t_0) = Ay(t), \qquad \text{für alle } t \in \mathbb{R}_{\geq t_0}.$$

Die Exponentialfunktion der Matrix At lässt sich beispielsweise über die Exponentialreihe approximieren. Wesentlich eleganter ist es, die Matrix A mit Hilfe einer Ähnlichkeitstransformation

$$T^{-1} A T = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}$$

zu diagonalisieren und die Exponentialfunktion durch

$$e^{tA} = e^{tTDT^{-1}} = T e^{tD} T^{-1} = T \begin{pmatrix} e^{t\lambda_1} & 0 \\ 0 & e^{t\lambda_2} \end{pmatrix} T^{-1}$$

zu berechnen. Im Falle des Federpendels stellt sich heraus, dass beide Eigenwerte λ_1, λ_2 rein imaginär und zueinander konjugiert sind, so dass die Matrix e^{At} und damit auch die Lösung y periodisch ist.

Die Exponentialfunktion eines rein imaginären Wertes steht in enger Beziehung zu Sinus- und Cosinus-Funktionen, deshalb überrascht es nicht, dass wir auch direkt den Ansatz

$$\begin{aligned} u(t) &= \alpha \sin(\omega t) + \beta \cos(\omega t), \\ v(t) &= \alpha\omega \cos(\omega t) - \beta\omega \sin(\omega t) \qquad \text{für alle } t \in \mathbb{R}_{\geq t_0} \end{aligned}$$

verwenden können. Der Parameter ω hängt von c und m ab, während die Parameter α und β verwendet werden können, um sicherzustellen, dass die Anfangsbedingungen erfüllt sind.

1.2 Mehrkörperprobleme

Das abstrakte Federpendel ist ein relativ einfaches Beispiel, weil die Ableitung y' und die Funktion y lediglich durch eine Matrix, also eine lineare Abbildung, gekoppelt sind und sich deshalb die Lösung analytisch angeben lässt.

Die in der Praxis auftretenden Probleme sind in der Regel nicht so einfach zu behandeln. Ein Beispiel ist das *Mehrkörperproblem*, bei dem n Massen $m_1, \dots, m_n$ zu einem Zeitpunkt t an n verschiedenen Positionen $x_1(t), \dots, x_n(t)$ im zwei- oder höherdimensionalen Raum liegen und mittels der Gravitation aufeinander einwirken.

In diesem Fall übt die Masse m_i auf die Masse m_j eine Kraft von

$$F_{ij}(t) = \varrho m_i m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3}$$

aus, wobei ϱ die Gravitationskonstante ist. Insgesamt wirkt also eine Kraft von

$$F_i(t) = \varrho m_i \sum_{\substack{j=1 \\ j \neq i}}^{n} m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3}$$

auf die Masse m_i, und entsprechend der Newton-Axiome entsteht dadurch eine Beschleunigung von

$$a_i(t) = \varrho \sum_{\substack{j=1 \\ j \neq i}}^{n} m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3}. \tag{1.1}$$

Wie im Falle des Federpendels benutzen wir die Newtonschen Axiome um festzustellen, dass a_i die Ableitung der Geschwindigkeit v_i und v_i die Ableitung des Ortes x_i ist, also

$$x_i'(t) = v_i(t), \qquad v_i'(t) = a_i(t) = \varrho \sum_{\substack{j=1 \\ j \neq i}}^{n} m_j \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|^3} \qquad \text{für alle } t \in \mathbb{R}$$

gilt. Wir fassen die Orte x_i, die Geschwindigkeiten v_i und die Beschleuigungen a_i zu Vektoren

$$x(t) := \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \qquad v(t) := \begin{pmatrix} v_1(t) \\ \vdots \\ v_n(t) \end{pmatrix}, \qquad a(t) := \begin{pmatrix} a_1(t) \\ \vdots \\ a_n(t) \end{pmatrix}$$

zusammen und schreiben die Differentialgleichung in der Form

$$\begin{pmatrix} x'(t) \\ v'(t) \end{pmatrix} = \begin{pmatrix} v(t) \\ a(t) \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}.$$

Gemäß (1.1) können wir $a(t)$ als Funktion A von x schreiben, erhalten also

$$a(t) = A(x(t)) \qquad \text{für alle } t \in \mathbb{R}.$$

Zur weiteren Vereinfachung fassen wir $x(t)$ und $y(t)$ zu einem Vektor

$$y(t) := \begin{pmatrix} x(t) \\ v(t) \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}$$

zusammen und führen die Funktion

$$f(y) := \begin{pmatrix} y_2 \\ A(y_1) \end{pmatrix}$$

ein, um die kompakte Darstellung

$$y'(t) = f(y(t)) \qquad \text{für alle } t \in \mathbb{R}$$

zu erhalten.

Im Falle des Federpendels war f lediglich eine lineare Abbildung, im Falle des Mehrkörperproblems ist f nicht linear, und die einfachen analytischen Lösungsansätze für lineare Probleme lassen sich nicht mehr verwenden.

Für $n \leq 3$ ist es noch möglich, die Lösung y wenigstens formal (etwa durch spezielle Reihenentwicklungen) darzustellen, für $n > 3$ dagegen sind numerische Approximationsverfahren das Mittel der Wahl.

Da diese Verfahren in der Regel eine große Anzahl ähnlicher Rechenoperationen erfordern, waren sie eine der wichtigsten praktischen Anwendungen früher Rechenmaschinen.

1.3 Wärmeleitung

Als Beispiel für ein zwar lineares, aber trotzdem nicht triviales, Anfangswertproblem untersuchen wir die Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t}(x,t) = \kappa \frac{\partial^2 u}{\partial x^2}(x,t) \qquad \text{für alle } t \in \mathbb{R}_{>0}, x \in [0,1]. \tag{1.2}$$

Sie beschreibt die Erwärmung oder Abkühlung eines Drahtes der Länge 1: $x \in [0,1]$ gibt die Position auf dem Draht an, t den Zeitpunkt, und $u(x,t)$ ist die Temperatur im Punkt x zum Zeitpunkt t. Wir nehmen zur Vereinfachung an, dass die Randbedingungen

$$u(0,t) = u(1,t) = 0 \qquad \text{für alle } t \in \mathbb{R}_{>0}$$

gelten, dass also die Temperatur an den beiden Endpunkten des Drahts fixiert ist.

Wir approximieren zunächst die Ortsableitung durch einen Differenzenquotienten: Die Taylor-Entwicklung von u um einen Punkt x ergibt

$$u(x+h,t) = u(x,t) + hu'(x,t) + \frac{h^2}{2}u''(x,t) + \frac{h^3}{6}u^{(3)}(x,t) + \frac{h^4}{24}u^{(4)}(x_+,t),$$
$$u(x-h,t) = u(x,t) - hu'(x,t) + \frac{h^2}{2}u''(x,t) - \frac{h^3}{6}u^{(3)}(x,t) + \frac{h^4}{24}u^{(4)}(x_-,t)$$

für Punkte $x_+ \in [x, x+h]$ und $x_- \in [x-h, x]$. Addition der beiden Gleichungen und Division durch h^2 ergibt

$$\frac{u(x-h,t) - 2u(x,t) + u(x+h,t)}{h^2} = u''(x) + \frac{h^2}{24}(u^{(4)}(x_+,t) + u^{(4)}(x_-,t)),$$

also können wir die zweite Ableitung durch den Differenzenquotienten auf der rechten Seite approximieren.

Wir ersetzen nun das kontinuierliche Intervall $[0,1]$ durch $n \in \mathbb{N}$ diskrete Punkte $0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1$, die durch

$$x_i := \frac{i}{n+1} \qquad \text{für alle } i \in \{0, \ldots, n+1\}$$

gegeben sind. Wir ersetzen $u(x,t)$ durch den Vektor $y(t) = (y_i(t))_{i=1}^n$ mit

$$y_i(t) = u(x_i, t) \qquad \text{für alle } t \in \mathbb{R}_{>0}, i \in \{1, \ldots, n\},$$

und stellen fest, dass die Wärmeleitungsgleichung (1.2) durch die Gleichungen

$$y_i'(t) = \begin{cases} \frac{\kappa}{h^2}(y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)) & \text{falls } 1 < i < n, \\ \frac{\kappa}{h^2}(-2y_i(t) + y_{i+1}(t)) & \text{falls } i = 1, \\ \frac{\kappa}{h^2}(y_{i-1}(t) - 2y_i(t)) & \text{falls } 1 = n, \end{cases} \qquad \text{für alle } i \in \{1, \ldots, n\}$$

und alle $t \in \mathbb{R}_{>0}$ mit $h = 1/(n+1)$ approximiert wird.

Kompakt lässt sich dieses System als

$$y'(t) = Ay(t) \qquad \text{für alle } t \in \mathbb{R}_{>0}$$

schreiben, wobei die Matrix $A \in \mathbb{R}^{n \times n}$ durch

$$A := -\frac{\kappa}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$

gegeben ist.

2 Theoretische Grundlagen

Bevor wir numerische Lösungsverfahren für gewöhnliche Differentialgleichungen untersuchen können, müssen wir zunächst klären, unter welchen Bedingungen diese Gleichungen überhaupt eine Lösung besitzen. In Hinblick auf die numerische Behandlung ist ebenfalls wichtig, wie empfindlich die Lösung auf Störungen der Parameter, insbesondere des Startwertes, reagiert.

2.1 Allgemeine Problemstellung

Wir konzentrieren uns auf die Analyse des Anfangswertproblems

$$y(a) = y_0, \qquad y'(t) = f(t, y(t)) \qquad \text{für alle } t \in [a, b] \tag{2.1}$$

auf einem Intervall $[a, b]$ mit einem Startwert y_0 in einem Banachraum V und einer Funktion $f : [a, b] \times V \to V$. Gesucht ist eine mindestens einmal stetig differenzierbare Funktion $y : [a, b] \to V$.

Das allgemeinere Problem

$$\begin{aligned} y(a) = y_0, \quad y'(a) = y_1, \quad \ldots, \quad y^{(m-1)}(a) = y_{m-1}, \\ y^{(m)}(t) = f(t, y(t), y'(t), \ldots, y^{(m-1)}(t)) \qquad \text{für alle } t \in [a, b] \end{aligned}$$

lässt sich auf die Form (2.1) zurückführen, indem wir den Hilfsvektor

$$w(t) := \begin{pmatrix} y(t) \\ y'(t) \\ \vdots \\ y^{(m-1)}(t) \end{pmatrix} \qquad \text{für alle } t \in [a, b]$$

einführen und das erweiterte System

$$w(t) = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{m-1} \end{pmatrix}, \quad w'(t) = \begin{pmatrix} 0 & w_2(t) & & & \\ & 0 & w_3(t) & & \\ & & \ddots & \ddots & \\ & & & 0 & w_m(t) \\ & & & & f(t, w_1(t), w_2(t), \ldots, w_m(t)) \end{pmatrix}$$

für alle $t \in [a, b]$ lösen. Die umgekehrte Vorgehensweise, also die Elimination von Ableitungsvariablen, haben wir bereits im Falle des abstrakten Federpendels kennengelernt.

2.2 Existenz und Eindeutigkeit

Das zentrale Hilfsmittel für den Beweis von Existenz und Eindeutigkeit von Lösungen des Problems (2.1) ist der Fixpunktsatz von Banach:

Satz 2.1 (Banach) *Sei X eine vollständige Teilmenge eines normierten Raumes. Sei $\Phi : X \to X$ eine Abbildung, und sei $L \in [0, 1)$ eine Zahl mit*

$$\|\Phi(x) - \Phi(y)\| \leq L\|x - y\| \qquad \text{für alle } x, y \in X. \tag{2.2}$$

Dann besitzt Φ einen Fixpunkt in X, es existiert also ein $x_* \in X$ mit

$$\Phi(x_*) = x_*.$$

Dieser Fixpunkt ist eindeutig bestimmt.

Beweis. (vgl. [2]) Sei $x_0 \in X$. Wir definieren die Folge $(x_n)_{n\in\mathbb{N}_0}$ durch

$$x_{n+1} = \Phi(x_n) \qquad \text{für alle } n \in \mathbb{N}_0.$$

Unser Ziel ist es, nachzuweisen, dass $(x_n)_{n\in\mathbb{N}_0}$ eine Cauchy-Folge ist.

Zunächst beweisen wir

$$\|x_{n+1} - x_n\| \leq L^n\|x_1 - x_0\| \tag{2.3}$$

durch Induktion für alle $n \in \mathbb{N}_0$. Für $n = 0$ ist (2.3) trivial.

Gelte nun (2.3) für ein $n \in \mathbb{N}_0$. Nach Voraussetzung gilt dann

$$\begin{aligned}\|x_{n+2} - x_{n+1}\| = \|\Phi(x_{n+1}) - \Phi(x_n)\| &\overset{(2.2)}{\leq} L\|x_{n+1} - x_n\| \\ &\leq LL^n\|x_1 - x_0\| = L^{n+1}\|x_1 - x_0\|,\end{aligned}$$

und der Induktionsschritt ist bewiesen.

Seien nun $n \in \mathbb{N}_0$ und $m \in \mathbb{N}_{\geq n}$. Dann gilt

$$\begin{aligned}\|x_m - x_n\| = \left\|\sum_{i=1}^{m-n} x_{n+i} - x_{n+i-1}\right\| &\leq \sum_{i=1}^{m-n} \|x_{n+i} - x_{n+i-1}\| \overset{(2.3)}{\leq} \sum_{i=1}^{m-n} L^{n+i-1}\|x_1 - x_0\| \\ &= \|x_1 - x_0\|L^n \sum_{i=0}^{m-n-1} L^i \leq \|x_1 - x_0\|L^n \sum_{i=0}^{\infty} L^i = \|x_1 - x_0\|\frac{L^n}{1-L}.\end{aligned}$$

Sei $\epsilon \in \mathbb{R}_{>0}$. Wir wählen $n_0 \in \mathbb{N}_0$ so, dass

$$\|x_1 - x_0\|\frac{L^{n_0}}{1-L} \leq \epsilon$$

gilt. Für alle $n, m \in \mathbb{N}_0$ mit $n_0 \leq n \leq m$ gilt dann

$$\|x_m - x_n\| \leq \|x_1 - x_0\|\frac{L^n}{1-L} \leq \|x_1 - x_0\|\frac{L^{n_0}}{1-L} \leq \epsilon,$$

also ist $(x_n)_{n\in\mathbb{N}_0}$ eine Cauchy-Folge.

Da X vollständig ist, muss es ein $x_* \in X$ mit

$$\lim_{n\to\infty} \|x_* - x_n\| = 0$$

geben, und wir müssen nur noch nachprüfen, dass x_* auch ein Fixpunkt von Φ ist.

Sei dazu $\epsilon \in \mathbb{R}_{>0}$. Da $(x_n)_{n\in\mathbb{N}_0}$ gegen x_* konvergiert, gibt es ein $n \in \mathbb{N}_0$ so, dass

$$\|x_* - x_n\| \leq \epsilon/4, \qquad\qquad \|x_* - x_{n+1}\| \leq \epsilon/4$$

gelten, und wir erhalten

$$\begin{aligned}
\|x_* - \Phi(x_*)\| &= \|x_* - x_n + x_n - x_{n+1} + x_{n+1} - \Phi(x_*)\| \\
&\leq \|x_* - x_n\| + \|x_n - x_{n+1}\| + \|\Phi(x_n) - \Phi(x_*)\| \\
&\leq \|x_* - x_n\| + \|x_n - x_{n+1}\| + L\|x_* - x_n\| \leq \epsilon/2 + \|x_n - x_{n+1}\| \\
&= \epsilon/2 + \|x_n - x_* + x_* - x_{n+1}\| \leq \epsilon/2 + \|x_n - x_*\| + \|x_* - x_{n+1}\| \leq \epsilon,
\end{aligned}$$

also folgt $x_* = \Phi(x_*)$, und x_* ist als Fixpunkt identifiziert.

Zum Nachweis der Eindeutigkeit wählen wir einen zweiten Fixpunkt x_{**} und erhalten

$$\|x_* - x_{**}\| = \|\Phi(x_*) - \Phi(x_{**})\| \leq L\|x_* - x_{**}\|,$$

also folgt aus $L < 1$ bereits $x_* = x_{**}$. ∎

Mit Hilfe des Fixpunktsatzes können wir nun die Lösbarkeit einer gewöhnlichen Differentialgleichung untersuchen. Die Grundlage ist die folgende Beobachtung:

Lemma 2.2 (Integralformulierung) *Sei eine stetige Funktion $f \in C([a,b]\times V, V)$ gegeben. Eine Funktion $y \in C^1([a,b], V)$ löst das Anfangswertproblem (2.1) genau dann, wenn*

$$y(t) = y_0 + \int_a^t f(s, y(s))\, ds \qquad\qquad \text{für alle } t \in [a,b] \tag{2.4}$$

gilt. Insbesondere ist eine stetige Funktion, die (2.4) löst, auch bereits einmal stetig differenzierbar.

Beweis. Im ersten Schritt gehen wir davon aus, dass y das Anfangswertproblem löst. Nach Hauptsatz der Differential- und Integralrechnung gilt dann

$$y_0 + \int_a^t f(s, y(s))\, ds = y_0 + \int_a^t y'(s)\, ds = y(a) + y(t) - y(a) = y(t),$$

für alle $t \in [a,b]$, also die Integralgleichung (2.4).

Im zweiten Schritt gehen wir davon aus, dass $y \in C([a,b], V)$ die Gleichung (2.4) erfüllt. Für $t = a$ folgt aus ihr unmittelbar $y(a) = y_0$. Für ein $t \in [a,b)$ und $h \in (0, b-t]$ erhalten wir

$$\frac{y(t+h) - y(t)}{h} = \frac{1}{h}\int_t^{t+h} f(s, y(s))\, ds = f(\eta, y(\eta))$$

mit einem $\eta \in [t, t+h]$ gemäß dem Mittelwertsatz der Integralrechnung. Für $h \to 0$ folgt $\eta \to t$ und somit

$$y'(t) = \lim_{h \to 0} \frac{y(t+h) - y(t)}{h} = f(t, y(t)),$$

also ist y stetig differenzierbar und erfüllt die gewöhnliche Differentialgleichung (2.1) für alle $t \in [a, b)$.

Für $t = b$ folgt die Aussage, indem wir entsprechend den linksseitigen Differenzenquotienten zur Approximation der Ableitung einsetzen. ∎

Mit Hilfe der Integralformulierung können wir eine Aussage über die Lösbarkeit der gewöhnlichen Differentialgleichung treffen.

Satz 2.3 (Picard-Lindelöf) *Die Funktion $f \in C([a, b] \times V, V)$ erfülle die* globale Lipschitz-Bedingung

$$\|f(t, x) - f(t, y)\| \leq L\|x - y\| \qquad \text{für alle } t \in [a, b] \text{ und } x, y \in V. \tag{2.5}$$

Dann besitzt das Anfangswertproblem (2.1) eine eindeutige Lösung $y \in C^1([a, b], V)$.

Beweis. (vgl. [4]) Wir führen den Beweis mit Hilfe des Banachschen Fixpunktsatzes 2.1: Dazu führen wir den Operator Φ durch

$$\Phi[u](x) := y_0 + \int_a^x f(s, u(s))\, ds \qquad \text{für alle } t \in [a, b], u \in C([a, b], V)$$

ein und untersuchen die von ihm induzierte Fixpunktiteration. Nach Lemma 2.2 wissen wir nämlich, dass ein Fixpunkt von Φ eine Lösung von (2.1) ist.

Damit wir Satz 2.1 anwenden können, müssen wir eine geeignete Norm auf dem Raum $C([a, b], V)$ einführen. Wir verwenden die gewichtete Supremumsnorm

$$\|u\|_e := \sup \left\{ e^{-2Lx} \|u(x)\| \;:\; x \in [a, b] \right\}, \qquad \text{für alle } u \in C([a, b], V)$$

und stellen fest, dass bezüglich dieser Norm

$$\begin{aligned}
e^{-2Lx}\|\Phi[u](x) - \Phi[v](x)\| &= e^{-2Lx} \left\| \int_a^x f(t, u(t)) - f(t, v(t))\, dt \right\| \\
&\leq e^{-2Lx} \int_a^x \|f(t, u(t)) - f(t, v(t))\|\, dt \\
&\leq L e^{-2Lx} \int_a^x \|u(t) - v(t)\|\, dt \\
&= L e^{-2Lx} \int_a^x e^{2Lt} e^{-2Lt} \|u(t) - v(t)\|\, dt \\
&\leq L e^{-2Lx} \int_a^x e^{2Lt} \|u - v\|_e\, dt \\
&= L e^{-2Lx} \|u - v\|_e \int_a^x e^{2Lt}\, dt
\end{aligned}$$

$$= Le^{-2Lx}\|u-v\|_e \frac{1}{2L}(e^{2Lx} - e^{2La}) \leq \frac{1}{2}\|u-v\|_e$$

für alle $x \in [a,b]$ und alle $u, v \in C([a,b], V)$ gilt. Also folgt

$$\|\Phi[u] - \Phi[v]\|_e \leq \frac{1}{2}\|u-v\|_e \qquad \text{für alle } u, v \in C([a,b], V),$$

und da die Norm $\|\cdot\|_e$ äquivalent zur Maximumnorm auf dem Raum der stetigen Funktionen ist, können wir Satz 2.1 anwenden, um zu folgern, dass ein eindeutig bestimmter Fixpunkt $y \in C([a,b], V)$ mit $\Phi[y] = y$ existiert.

Nach Lemma 2.2 ist diese Funktion y auch die eindeutig bestimmte Lösung des Anfangswertproblems. ∎

Satz 2.3 ist nicht nur ein Existenz- und Eindeutigkeitsresultat, er bietet uns auch ein Konstruktionsverfahren für die Lösung des Anfangswertproblems:

Bemerkung 2.4 (Picard-Iteration) *Ausgehend von einer beliebigen Funktion u_0 können wir, wie im Satz 2.1, die Folge $u_{n+1} := \Phi(u_n)$ konstruieren, und Satz 2.3 impliziert, dass diese Folge gegen die Lösung des Anfangswertproblems (2.1) konvergieren wird. Diese Konstruktion trägt den Namen* Picard-Iteration.

Für die Praxis ist diese Konstruktion nur dann anwendbar, wenn sich die einzelnen Iterierten u_n geeignet im Rechner darstellen lassen, etwa mit Hilfe einer Diskretisierung.

2.3 Störungen der Daten

Für die numerische Behandlung des Anfangswertproblems (2.1) ist neben der prinzipiellen Lösbarkeit auch der Einfluss von Störungen relevant, schließlich wird im praktischen Algorithmus in der Regel mit Gleitpunktarithmetik beschränkter Genauigkeit gearbeitet.

Ein wichtiges Hilfsmittel für die Analyse ist die *Grönwallsche Ungleichung*, von der wir hier nur die folgende vereinfachte Variante benötigen:

Lemma 2.5 (Grönwall) *Seien $[a,b] \subseteq \mathbb{R}$ ein Intervall, $u \in C([a,b], \mathbb{R})$ eine Funktion und $\alpha \in \mathbb{R}, \beta \in \mathbb{R}_{\geq 0}$ Konstanten, die*

$$u(t) \leq \alpha + \beta \int_a^t u(s)\,ds \qquad \textit{für alle } t \in [a,b] \tag{2.6}$$

erfüllen. Dann gilt

$$u(t) \leq \alpha e^{\beta(t-a)} \qquad \textit{für alle } t \in [a,b].$$

Beweis. (vgl. [3]) Wir führen die Hilfsfunktion $v \in C([a,b))$ mit

$$v(t) := e^{-\beta(t-a)} \int_a^t \beta u(s)\,ds \qquad \text{für alle } t \in [a,b]$$

ein und erhalten mit Produktregel und (2.6) die Abschätzung

$$\begin{aligned} v'(t) &= -\beta e^{-\beta(t-a)} \int_a^t \beta u(s)\,ds + e^{-\beta(t-a)}\beta u(t) \\ &= \beta e^{-\beta(t-a)} \left(u(t) - \int_a^t \beta u(s)\,ds \right) \overset{(2.6)}{\leq} \beta\alpha e^{-\beta(t-a)} \end{aligned}$$

für alle $t \in [a,b]$. Aus $v(a) = 0$ folgt

$$\begin{aligned} e^{-\beta(t-a)} \int_a^t \beta u(s)\,ds = v(t) = v(t) - v(a) = \int_a^t v'(s)\,ds &\leq \beta\alpha \int_a^t e^{-\beta(s-a)}\,ds \\ = \beta\alpha \left(-\frac{1}{\beta} \right) (e^{-\beta(t-a)} - e^{-\beta(a-a)}) &= \alpha - \alpha e^{-\beta(t-a)} \end{aligned}$$

und indem wir mit $e^{\beta(t-a)}$ multiplizieren erhalten wir aus (2.6)

$$u(t) \leq \alpha + \int_a^t \beta u(s)\,ds \leq \alpha + e^{\beta(t-a)}(\alpha - \alpha e^{-\beta(t-a)}) = \alpha + \alpha e^{\beta(t-a)} - \alpha = \alpha e^{\beta(t-a)}.$$

Das ist die zu beweisende Ungleichung. ∎

Mit Hilfe dieses Korollars und des Lemmas 2.2 können wir nun den Einfluss von Störungen der Anfangsdaten untersuchen:

Satz 2.6 (Störungen) *Sei $U \subseteq V$. Die Funktion $f \in C([a,b] \times U, U)$ erfülle die bereits aus Satz 2.3 bekannte globale Lipschitz-Bedingung (2.5). Die Funktion $g \in C([a,b] \times U, U)$ erfülle die Bedingung*

$$\|f(t,x) - g(t,x)\| \leq M \qquad \text{für alle } t \in [a,b], x \in U$$

mit einer von t und x unabhängigen Konstanten $M \in \mathbb{R}_{\geq 0}$.

Seien $y_0, z_0 \in U$, und seien $y, z \in C^1([a,b], U)$ Lösungen der Anfangswertprobleme

$$\begin{aligned} y(a) &= y_0, & y'(t) &= f(t, y(t)), \\ z(a) &= z_0, & z'(t) &= g(t, z(t)) \qquad \text{für alle } t \in [a,b]. \end{aligned}$$

Dann gilt die Abschätzung

$$\|y(t) - z(t)\| \leq \begin{cases} \|y_0 - z_0\| e^{L(t-a)} + \frac{M}{L}\left(e^{L(t-a)} - 1\right) & \text{falls } L > 0, \\ \|y_0 - z_0\| + M(t-a) & \text{falls } L = 0 \end{cases} \quad \text{für alle } t \in [a,b],$$

kleine Störungen der Anfangsdaten und der rechten Seite führen also auch nur zu kleinen Störungen der Lösung.

Beweis. Mit Lemma 2.2 erhalten wir

$$y(t) - z(t) = y_0 - z_0 + \int_a^t f(s, y(s)) - g(s, z(s))\,ds,$$

$$\begin{aligned}\|y(t)-z(t)\| &\le \|y_0-z_0\| + \int_a^t \|f(s,y(s))-g(s,z(s))\|\,ds \\ &\le \|y_0-z_0\| + \int_a^t \|f(s,y(s))-f(s,z(s))\| + \|f(s,z(s))-g(s,z(s))\|\,ds \\ &\overset{(2.5)}{\le} \|y_0-z_0\| + \int_a^t L\|y(s)-z(s)\| + M\,ds. \end{aligned} \tag{2.7}$$

Im Falle $L = 0$ impliziert (2.7) bereits

$$\|y(t)-z(t)\| \le \|y_0-z_0\| + M(t-a) \qquad \text{für alle } t \in [a,b],$$

so dass wir uns auf den Fall $L > 0$ konzentrieren können. Wir setzen

$$u(t) := \|y(t)-z(t)\| + \frac{M}{L} \qquad \text{für alle } t \in [a,b],$$

um die Gleichung

$$\begin{aligned} u(t) &= \|y(t)-z(t)\| + \frac{M}{L} \\ &\le \|y_0-z_0\| + \frac{M}{L} + \int_a^t L\|y(s)-z(s)\| + M\,ds = u(a) + L\int_a^t u(s)\,ds, \end{aligned}$$

zu erhalten. Nun wenden wir Lemma 2.5 auf

$$\alpha := u(a) = \|y_0-z_0\| + \frac{M}{L}, \qquad \beta := L$$

an, um die Abschätzung

$$\begin{aligned}\|y(t)-z(t)\| = u(t) - \frac{M}{L} &\le u(a)e^{L(t-a)} - \frac{M}{L} \\ &= \left(\|y_0-z_0\| + \frac{M}{L}\right)e^{L(t-a)} - \frac{M}{L} \\ &= \|y_0-z_0\|e^{L(t-a)} + \frac{M}{L}\left(e^{L(t-a)}-1\right) \qquad \text{für alle } t \in [a,b] \end{aligned}$$

zu erhalten. ∎

Bemerkung 2.7 *Die kleinste mögliche Konstante für M in Satz 2.6 ist offenbar gerade die Maximumnorm von $f-g$, also*

$$M = \|f-g\|_\infty := \sup\{\|f(t,x)-g(t,x)\| \ : \ (t,x) \in [a,b]\times U\},$$

so dass sich die Aussage des Satzes auch abstrakt als

$$\|y(t)-z(t)\| \le c_1(t)\|y_0-z_0\| + c_2(t)\|f-g\|_\infty \qquad \textit{für alle } t \in [a,b]$$

schreiben lässt. Der Satz beschreibt also die Lipschitz-stetige Abhängigkeit der Lösung zum Zeitpunkt t von den Anfangsdaten und der rechten Seite.

3 Einschrittverfahren

Die exakte Lösung eines Anfangswertproblems der Form (2.1) wird sich im allgemeinen Fall nicht exakt berechnen lassen. Stattdessen müssen wir auf eine Approximation zurückgreifen: Statt nach einer geschlossenen Formel für die Lösung zu suchen, beschränken wir uns darauf, sie nur in einzelnen Punkten $t_0, \dots, t_n \in [a, b]$ näherungsweise zu berechnen.

Wir sind natürlich an Verfahren interessiert, die uns eine möglichst genaue Näherung zur Verfügung stellen, und das bei möglichst geringem Rechen- und Speicheraufwand.

Ein möglicher Zugang wäre etwa die Picard-Iteration (vgl. Bemerkung 2.4): Wir könnten die Iterierten durch ihre Werte in Punkten des Intervalls approximieren und zur Berechnung der Integrale eine Quadraturformel verwenden, die nur diese Punktwerte benötigt. Der Nachteil dieses Zugangs besteht darin, dass *alle* Punktwerte gleichzeitig im Speicher gehalten werden müssen.

Wir suchen stattdessen nach einem Verfahren, bei dem wir die Werte zu den verschiedenen Zeitpunkten der Reihe nach berechnen können. Ein *Einschrittverfahren* versucht, den Wert zu einem Zeitpunkt x_{i+1} nur auf Grundlage des Wertes zum unmittelbar vorangehenden Zeipunkt x_i zu approximieren.

Um die Diskussion der Lösbarkeit zu vermeiden setzen wir, sofern nicht gesondert erwähnt, im folgenden Kapitel voraus, dass die rechte Seite f des Anfangswertproblems (2.1) im zweiten Argument Lipschitz-stetig ist (vgl. Bedingung (2.5)). Satz 2.3 impliziert dann die eindeutige Lösbarkeit für beliebige Startwerte in V und Startpunkte in $[a, b]$.

3.1 Euler-Verfahren

Wir untersuchen zunächst ein besonders einfaches Einschrittverfahren: Das *Euler-Verfahren* basiert auf der Idee, die Ableitung im Anfangswertproblem (2.1) durch einen Differenzenquotienten zu ersetzen:

$$\frac{y(t+h) - y(t)}{h} \approx y'(t) = f(t, y(t)), \qquad y(t+h) \approx y(t) + hf(t, y(t)).$$

Für $n \in \mathbb{N}$ setzen wir

$$t_i := a\frac{n-i}{n} + b\frac{i}{n} = a + hi \qquad \text{für } h := \frac{1}{n}, i \in \{0, \dots, n\}$$

und berechnen induktiv die Näherungen

$$\eta(t_{i+1}) := \eta(t_i) + hf(t_i, \eta(t_i)) \qquad \text{für alle } i \in \{0, \dots, n-1\}.$$

Offenbar ist dieses Verfahren sehr effizient durchführbar: In jedem Schritt muss f einmal ausgewertet und eine Linearkombination berechnet werden, und es brauchen nur jeweils $\eta(t_{i+1})$ und $\eta(t_i)$ gleichzeitig im Speicher gehalten zu werden.

Da der Wert $\eta(t_{i+1})$ jeweils direkt berechnet werden kann, spricht man von einem *expliziten* Verfahren.

Wenn wir statt des „rechtsseitigen“ Differenzenquotienten den „linksseitigen“ Quotienten verwenden, erhalten wir den Ansatz

$$\frac{y(t) - y(t-h)}{h} \approx y'(t) = f(t, y(t)), \qquad y(t) \approx y(t-h) + hf(t, y(t)),$$

der nicht mehr direkt berechnet werden kann: Jetzt ist die neue Approximation $\eta(t_{i+1})$ durch das potentiell nichtlineare Gleichungssystem

$$\eta(t_{i+1}) - hf(t_{i+1}, \eta(t_{i+1})) = \eta(t_i)$$

gegeben, das wir mit einem geeigneten Verfahren auflösen müssen.

Ein brauchbarer Ansatz hierzu ist eine Fixpunkt-Iteration mit dem Operator

$$\Psi(x) := \eta(t_i) + hf(t_{i+1}, x) \qquad \text{für alle } x \in V.$$

Falls f Lipschitz-stetig im zweiten Argument ist, also

$$\|f(t_{i+1}, x) - f(t_{i+1}, y)\| \leq L\|x - y\| \qquad \text{für alle } x, y \in V$$

gilt, erhalten wir

$$\|\Psi(x) - \Psi(y)\| = \|hf(t_{i+1}, x) - hf(t_{i+1}, y)\| \leq hL\|x - y\| \qquad \text{für alle } x, y \in V,$$

und der Satz 2.1 von Banach garantiert Konvergenz gegen einen eindeutig bestimmten Fixpunkt x^*, falls wir $h < 1/L$ sicherstellen können. Dieser Fixpunkt erfüllt

$$x^* = \Psi(x^*) = \eta(t_i) + hf(t_{i+1}, x^*), \qquad x^* - hf(t_{i+1}, x^*) = \eta(t_i),$$

ist also die gesuchte Lösung $\eta(t_{i+1}) = x^*$.

Falls f hinreichend oft differenzierbar ist und gute Startwerte bekannt sind, kann man natürlich statt der Fixpunkt-Iteration auch alternative Ansätze wie beispielsweise das Newton-Verfahren verwenden, um $\eta(t_{i+1})$ zu berechnen.

Da bei dieser Variante $\eta(t_{i+1})$ nur indirekt über eine Gleichung gegeben ist, spricht man von einem *impliziten* Verfahren.

Das Euler-Verfahren kann dank Lemma 2.2 auch als Quadraturverfahren interpretiert werden: Die Lösung y des Anfangswertproblems auf dem Intervall $[t_i, t_{i+1}]$ erfüllt die Gleichung

$$y(t) = y(t_i) + \int_{t_i}^{t} f(s, y(s))\, ds \qquad \text{für alle } t \in [t_i, t_{i+1}],$$

so dass wir insbesondere

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))\, ds$$

erhalten. Wenn wir das Integral per Rechteckregel approximieren, also durch

$$\int_{t_i}^{t_{i+1}} f(s, y(s))\, ds \approx h f(t_i, y(t_i)) \quad \text{oder} \quad \int_{t_i}^{t_{i+1}} f(s, y(s))\, ds \approx h f(t_{i+1}, y(t_{i+1})),$$

erhalten wir wieder das explizite beziehungsweise implizite Euler-Verfahren.

Wenn wir die Integrale mit anderen Quadraturformeln approximieren, erhalten wir weitere, in der Regel implizite, Einschrittverfahren.

3.2 Konvergenz

Natürlich ist das Euler-Verfahren nur dann nützlich, wenn es auch eine hinreichend gute Approximation der tatsächlichen Lösung berechnet. Wir müssen also untersuchen, ob und, falls ja, wie schnell die approximative Lösung gegen die echte Lösung konvergiert.

Die Theorie basiert auf Vergleichen zwischen exakten und approximativen Lösungen zu verschiedenen Startwerten: Unter den Annahmen von Satz 2.3 definieren wir für alle Startwerte $y_* \in V$ und alle Startzeitpunkte $t_* \in [a, b]$ die Funktion $y(\cdot; t_*, y_*) : [t_*, b] \to V$ als Lösung des Anfangswertproblems

$$y(t_*; t_*, y_*) = y_*, \qquad \frac{\partial}{\partial t} y(t; t_*, y_*) = f(t, y(t; t_*, y_*)) \qquad \text{für alle } t \in [t_*, b]. \tag{3.1}$$

Nach Satz 2.3 ist $y(\cdot; t_*, y_*)$ eindeutig definiert.

Infolge der Eindeutigkeit muss für $t_*, s_* \in [a, b]$ mit $t_* \leq s_*$ auch die Gleichung

$$y(t; t_*, y_*) = y(t; s_*, y(s_*; t_*, y_*)) \qquad \text{für alle } t \in [s_*, b] \tag{3.2}$$

gelten: Im Punkt s_* stimmen beide Seiten der Gleichung überein, also müssen sie auch auf dem gesamten Intervall $[s_*, b]$ übereinstimmen.

Nun benötigen wir eine ähnliche Funktion für die approximativen Lösungen. Wir untersuchen ein allgemeines explizites Einschrittverfahren der Form

$$\eta(t_{i+1}) := \eta(t_i) + h\Phi(t_i, \eta(t_i), h) \qquad \text{für alle } i \in \{0, \ldots, n-1\} \tag{3.3}$$

mit der *Inkrementfunktion* Φ. Das explizite Euler-Verfahren beispielsweise können wir per $\Phi(t, x, h) := f(t, x)$ auf diese allgemeine Form zurückführen.

Die Intervalle $[t_*, b]$ werden durch ihre diskreten Gegenstücke

$$\begin{aligned} [t_*, b]_h &:= [t_*, b] \cap \{t_0, \ldots, t_n\} \\ &= \{t_i = a + ih \ : \ i \in \{0, \ldots, n\}, a + ih \geq t_*\} \qquad \text{für alle } t_* \in [a, b] \end{aligned}$$

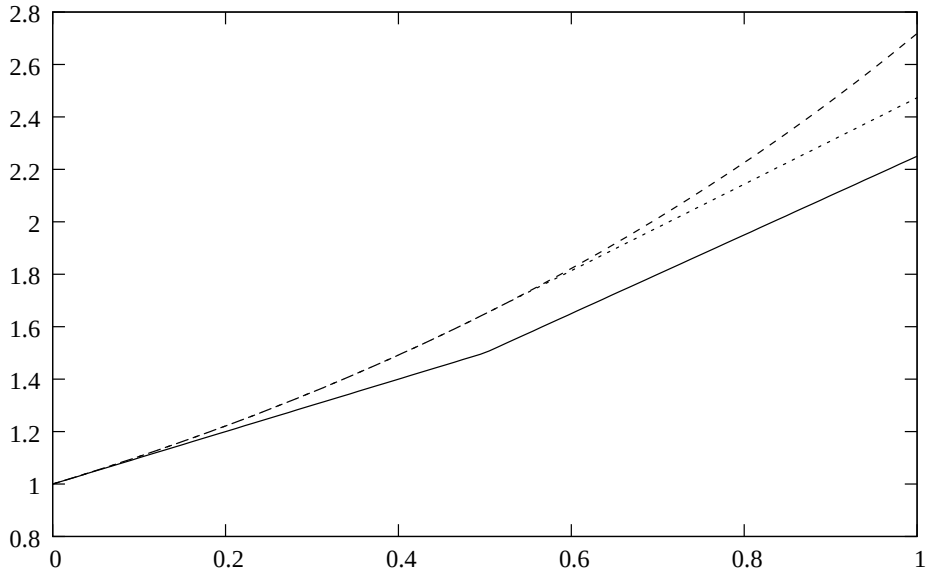

Abbildung 3.1: Ansatz zur Fehlerabschätzung: Zwischen der echten Lösung $y(t; x_0, y_0)$ (oben) und der approximativen Lösung $\eta(t; x_0, y_0)$ (unten) wird die halbapproximative Lösung $\eta(t; x_1, y(t_1; x_0, y_0))$ eingeschoben

ersetzt, und die Funktionen y durch diskrete Funktionen η: Für einen Startwert $y_* \in V$ und einen Startzeitpunkt $t_* \in [a, b]_h$ definieren wir die Funktion $\eta(\cdot; t_*, y_*) : [t_*, b]_h \to V$ induktiv durch

$$\eta(t_{i+1}; t_*, y_*) := \eta(t_i; t_*, y_*) + h\Phi(t_i, \eta(t_i; t_*, y_*), h) \qquad \text{für alle } t_i \in [t_*, b]_h.$$

Aus der induktiven Definition folgt für $t_*, s_* \in [a, b]_h$ mit $t_* \leq s_*$ die der Gleichung (3.2) entsprechende Fortsetzungseigenschaft

$$\eta(t; t_*, y_*) = \eta(t; s_*, \eta(s_*; t_*, y_*)) \qquad \text{für alle } t \in [s_*, b]_h. \tag{3.4}$$

Mit Hilfe der Fortsetzungseigenschaften (3.2) und (3.4) können wir nun eine Darstellung für den Approximationsfehler finden:

Wir wählen $t_*, s_*, t \in [a, b]_h$ mit $t_* \leq s_* \leq t$. Wegen (3.2) wissen wir, dass die Fortsetzung der exakten Lösung y ab s_* mit dem exakten Startwert $y(s_*; t_*, y_*)$ wieder die exakte Lösung y sein wird. Wegen (3.4) wissen wir, dass die Fortsetzung der approximativen Lösung η ab s_* mit dem approximativen Startwert $\eta(s_*; t_*, y_*)$ wieder die approximative Lösung η sein wird. Für unsere Abschätzung schieben wir nun zwischen den beiden Funktionen die *approximative* Fortsetzung mit dem *exakten* Startwert $y(s_*; t_*, y_*)$ ein, erhalten also

$$\begin{aligned} y(t; t_*, y_*) - \eta(t; t_*, y_*) = {} & y(t; s_*, y(s_*; t_*, y_*)) - \eta(t; s_*, y(s_*; t_*, y_*)) \\ & + \eta(t; s_*, y(s_*; t_*, y_*)) - \eta(t; s_*, \eta(s_*; t_*, y_*)). \end{aligned} \tag{3.5}$$

Für $s_* > t_*$ lässt sich die Norm der ersten Differenz auf ein Anfangswertproblem (mit exaktem Startwert) auf dem kleineren Intervall $[s_*, b]_h$ zurückführen, während sich die zweite Differenz mit Hilfe einer diskreten Variante des Satzes 2.6 abschätzen lässt:

Lemma 3.1 (Diskrete Störungen) *Sei die Inkrementfunktion Φ auf $[a,b]_h \times V \times \{h\}$ Lipschitz-stetig im zweiten Argument, es gelte also*

$$\|\Phi(t,x,h) - \Phi(t,z,h)\| \le L\|x - z\| \qquad \text{für alle } x, z \in V, t \in [a,b]_h \tag{3.6}$$

mit einer von t, x, z und h unabhängigen Konstanten $L \in \mathbb{R}_{\ge 0}$. Dann ist die Differenz zweier Näherungslösungen beschränkt durch

$$\|\eta(t;t_0,y_0) - \eta(t;t_0,z_0)\| \le \|y_0 - z_0\| e^{L(t-t_0)} \qquad \text{für alle } y_0, z_0 \in V, t \in [a,b]_h.$$

Beweis. Wir beweisen

$$\|\eta(t_i;t_0,y_0) - \eta(t_i;t_0,z_0)\| \le \|y_0 - z_0\| e^{L(t_i-t_0)} \qquad \text{für alle } i \in \{0,\dots,n\} \tag{3.7}$$

per (endlicher) Induktion. Für $i = 0$ ist die Aussage trivial.

Gelte nun (3.7) für ein $i \in \{0,\dots,n-1\}$. Nach Definition haben wir

$$\begin{aligned}
&\|\eta(t_{i+1};t_0,y_0) - \eta(t_{i+1};t_0,z_0)\| \\
&\quad = \| \left(\eta(t_i;t_0,y_0) + h\Phi(t_i,\eta(t_i;t_0,y_0),h)\right) - \left(\eta(t_i;t_0,z_0) + h\Phi(t_i,\eta(t_i;t_0,z_0),h)\right) \| \\
&\quad \le \|\eta(t_i;t_0,y_0) - \eta(t_i;t_0,z_0)\| + h\|\Phi(t_i,\eta(t_i;t_0,y_0),h) - \Phi(t_i,\eta(t_i;t_0,z_0),h)\| \\
&\quad \le \|\eta(t_i;t_0,y_0) - \eta(t_i;t_0,z_0)\| + hL\|\eta(t_i;t_0,y_0) - \eta(t_i;t_0,y_0)\| \\
&\quad = (1+hL)\|\eta(t_i;t_0,y_0) - \eta(t_i;t_0,y_0)\| \le e^{hL}\|\eta(t_i;t_0,y_0) - \eta(t_i;t_0,y_0)\|
\end{aligned}$$

so dass wir aus der Induktionsvoraussetzung

$$\|\eta(t_{i+1};t_0,y_0) - \eta(t_{i+1};t_0,z_0)\| \le \|y_0 - z_0\| e^{hL} e^{L(t_i-t_0)} = \|y_0 - z_0\| e^{L(t_{i+1}-t_0)}$$

schließen können. Damit ist (3.7) auch für $i+1$ bewiesen. ∎

Anstelle der Lipschitz-Stetigkeit von f setzt diese Abschätzung die der Inkrementfunktion Φ voraus. Im Falle des expliziten Euler-Verfahrens ist beides ohnehin äquivalent. Für allgemeinere Verfahren ist eine entsprechende Aussage in der Regel leicht nachzuweisen.

Dank Lemma 3.1 nimmt (3.5) die Form

$$\begin{aligned}
\|y(t;t_*,y_*) - \eta(t;t_*,y_*)\| \le{}& \|y(t;s_*,y(s_*;t_*,y_*)) - \eta(t;s_*,y(s_*;t_*,y_*))\| \\
&+ \|y(s_*;t_*,y_*) - \eta(s_*;t_*,y_*)\| e^{L(t-s_*)}
\end{aligned} \tag{3.8}$$

an, der zweite Term ist also (bis auf einen Skalierungsfaktor) gerade der Fehler, den das Näherungsverfahren zwischen s_* und t_* verursacht. Indem wir $s_* = t_* + h$ wählen, entspricht dieser Term also dem *lokalen* Fehler, den wir in einem einzelnen Schritt des Verfahrens in Kauf nehmen.

Um den Gesamtfehler zu erhalten genügt es, diese einzelnen lokalen Fehler mit den entsprechenden Faktoren aufzusummieren. Da unser Ansatz jeweils die Werte der *exakten* Lösung als Startwerte für lokale und globale Approximationen verwendet, bietet es sich an, die betreffenden Werte abkürzend mit

$$y_i := y(t_i; t_0, y_0) \qquad \text{für alle } i \in \{0, \dots, n\}$$

zu bezeichnen. Durch Induktion der Abschätzung (3.8) erhalten wir die folgende Abschätzung des Gesamtfehlers:

Satz 3.2 (Konvergenz) *Sei die Inkrementfunktion Lipschitz-stetig im zweiten Argument, es gelte also (3.6). Sei*

$$K_\Phi := \max\{\|y(t_{i+1}; t_i, y_i) - \eta(t_{i+1}; t_i, y_i)\| \ : \ i \in \{0, \dots, n-1\}\} \tag{3.9}$$

der maximale lokale *Fehler. Dann folgt die Abschätzung*

$$\|y(t) - \eta(t)\| \le \begin{cases} \frac{K_\Phi}{Lh}(e^{L(t-a)} - 1) & \textit{falls } L > 0, \\ \frac{K_\Phi}{h}(t-a) & \textit{falls } L = 0 \end{cases} \qquad \textit{für alle } t \in [a,b]_h.$$

Beweis. Sei $t \in [a,b]_h$ gegeben. Der Fall $t = a$ ist trivial, wir beschränken uns also auf $t = t_i$ mit $i \in \{1, \dots, n\}$.

Zunächst beweisen wir die Hilfsbehauptung

$$\|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| \le K_\Phi \sum_{\ell=j+1}^{i} e^{Lh(i-\ell)} \qquad \text{für alle } j \in \{0, \dots, i-1\} \tag{3.10}$$

durch (endliche, absteigende) Induktion über j.

Wir haben

$$\|y(t; t_{i-1}, y_{i-1}) - \eta(t; t_{i-1}, y_{i-1})\| = \|y(t_i; t_{i-1}, y_{i-1}) - \eta(t_i; t_{i-1}, y_{i-1})\| \le K_\Phi$$

nach Voraussetzung, also ist (3.10) für $j = i-1$ bewiesen.

Sei nun $j \in \{1, \dots, i-1\}$ so gegeben, dass (3.10) gilt. Wie bereits gezeigt folgt aus Lemma 3.1 und (3.2) sowie (3.4) die Abschätzung

$$\begin{aligned} \|y(t; t_{j-1}, y_{j-1}) - \eta(t; t_{j-1}, y_{j-1})\| &= \|y(t; t_j, y_j) - \eta(t; t_j, \eta(t_j; t_{j-1}, y_{j-1}))\| \\ &\le \|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| + \|\eta(t; t_j, y_j) - \eta(t; t_j, \eta(t_j; t_{j-1}, y_{j-1}))\| \\ &\le \|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| + \|y_j - \eta(t_j; t_{j-1}, y_{j-1})\| e^{L(t-t_j)} \\ &\le \|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| + K_\Phi e^{Lh(i-j)}. \end{aligned}$$

Wir wenden die Induktionsvoraussetzung an, um

$$\|y(t; t_{j-1}, y_{j-1}) - \eta(t; t_{j-1}, y_{j-1})\| \le K_\Phi \sum_{\ell=j+1}^{i} e^{Lh(i-\ell)} + K_\Phi e^{Lh(i-j)} = K_\Phi \sum_{\ell=j}^{i} e^{Lh(i-\ell)}$$

zu erhalten. Das ist die Abschätzung (3.10) für $j-1$, und die Induktion ist vollständig.

Um die gesuchte Aussage zu beweisen, wenden wir (3.10) auf $j=0$ an und finden

$$\|y(t;a,y_0)-\eta(t;a,y_0)\| \le K_\Phi \sum_{\ell=1}^{i} e^{Lh(i-\ell)}.$$

Im Falle $L=0$ ist die Summe gerade i, also $(t-a)/h$, und wir sind fertig.

Im Falle $L>0$ erhalten wir $e^{Lh}>1$ und beschränken die geometrische Reihe durch

$$\begin{aligned}\sum_{\ell=1}^{i} e^{Lh(i-\ell)} &= e^{Lhi}\sum_{\ell=1}^{i} e^{-Lh\ell} = e^{L(t-a)}\sum_{\ell=1}^{i}(e^{-Lh})^\ell = e^{L(t-a)}\frac{e^{-Lh}-e^{-Lh(i+1)}}{1-e^{-Lh}}\\ &= e^{L(t-a)}\frac{1-e^{-Lhi}}{e^{Lh}-1} \le e^{L(t-a)}\frac{1-e^{-L(t-a)}}{1+Lh-1} = \frac{1}{Lh}(e^{L(t-a)}-1),\end{aligned}$$

so dass insgesamt die Abschätzung

$$\|y(t;a,y_0)-\eta(t;a,y_0)\| \le \frac{K_\Phi}{Lh}(e^{L(t-a)}-1)$$

bewiesen ist. ∎

3.3 Konsistenz

Aus Satz 3.2 folgt, dass für die Konvergenz des Näherungsverfahrens das Verhalten des Faktors K_Φ/h ausschlaggebend ist. Wenn wir in (3.9) die Definition von $\eta(t_{i+1};t_i,y_i)$ einsetzen, erhalten wir

$$\begin{aligned}\frac{K_\Phi}{h} &= \max\left\{\frac{\|y(t_{i+1};t_i,y_i)-y(t_i;t_i,y_i)-h\Phi(t_i,y_i,h)\|}{h} \;:\; i\in\{0,\dots,n-1\}\right\}\\ &= \max\left\{\left\|\frac{y(t_i+h;t_i,y_i)-y(t_i;t_i,y_i)}{h}-\Phi(t_i,y_i,h)\right\| \;:\; i\in\{0,\dots,n-1\}\right\}\end{aligned}$$

Für $h\to 0$ wird der linke Term gegen $y'(t)$ konvergieren, also gegen $f(t,y(t))$. Offenbar kann also das Näherungsverfahren nur dann erfolgreich sein, wenn für $h\to 0$ die Inkrementfunktion Φ gegen f konvergiert.

Definition 3.3 (Diskretisierungsfehler) *Sei Φ eine Inkrementfunktion. Wir definieren den* lokalen Diskretisierungsfehler *zu dem durch Φ gegebenen expliziten Einschrittverfahren durch*

$$\tau(t;x,h) := \frac{y(t+h;t,x)-x}{h}-\Phi(t,x,h) \qquad \textit{für alle } h\in\mathbb{R}_{>0}, t\in[a,b-h], x\in V.$$

Bei dieser Definition ist zu beachten, dass wegen Satz 2.3 und wegen der Lipschitz-Stetigkeit von f die Funktion τ für alle $x \in V$ wohldefiniert ist.

Wie bereits gesehen gilt

$$\frac{K_\Phi}{h} \leq \max\{\|\tau(t_i; y_i, h)\| \ : \ i \in \{0, \dots, n-1\}\},$$

nach Satz 3.2 ist es also für die Konvergenz der Näherungslösung sehr erstrebenswert, dass $\tau(t; y(t), h)$ für $h \to 0$ gleichmäßig gegen Null geht.

Definition 3.4 (Konsistenz) *Sei Φ eine Inkrementfunktion. Das durch sie definierte explizite Einschrittverfahren heißt* konsistent mit dem Anfangswertproblem (2.1)*, falls*

$$\lim_{h \to 0} \sup\{\|\tau(t; y(t), h)\| \ : \ t \in [a, b-h]\} = 0 \tag{3.11}$$

gilt. Das Verfahren heißt von der Ordnung p konsistent mit dem Problem *für ein $p \in \mathbb{N}$, falls es Konstanten $C_y, h_y \in \mathbb{R}_{>0}$ so gibt, dass*

$$\sup\{\|\tau(t; y(t), h)\| \ : \ t \in [a, b-h]\} \leq C_y h^p \qquad \text{für alle } h \in (0, h_y) \tag{3.12}$$

gilt. Offenbar impliziert diese Bedingung bereits, dass Φ auch konsistent ist.

Die Konsistenz eines Verfahrens lässt sich auf ein sehr einfaches Kriterium zurückführen:

Lemma 3.5 (Konsistenz) *Sei f stetig und sei y Lösung des Anfangswertproblems (2.1). Sei Φ stetig auf*

$$\Delta := \{(t, h) \ : \ t \in [a, b], h \in [0, b-t]\}.$$

Das durch Φ definierte explizite Einschrittverfahren ist genau dann konsistent mit dem Problem, wenn

$$f(t, y(t)) = \Phi(t, y(t), 0) \qquad \text{für alle } t \in [a, b] \tag{3.13}$$

erfüllt ist.

Beweis. Wir beweisen zunächst zwei Hilfsbehauptungen.

Vorbetrachtung 1: Für alle $t \in [a, b)$, $h \in (0, t-b]$ gilt

$$\begin{aligned}
\|\tau(t; y(t), h)\| &= \left\| \frac{y(t+h; t, y(t)) - y(t)}{h} - \Phi(t, y(t), h) \right\| \\
&= \left\| \frac{y(t+h) - y(t)}{h} - \Phi(t, y(t), h) \right\| \\
&= \|y'(t_+) - \Phi(t, y(t), h)\| = \|f(t_+, y(t_+)) - \Phi(t, y(t), h)\|
\end{aligned} \tag{3.14}$$

mit einem $t_+ \in [t, t+h]$.

Vorbetrachtung 2: Da f und y stetig sind, ist

$$t \mapsto f(t, y(t))$$

stetig auf der kompakten Menge $[a, b]$, also insbesondere gleichmäßig stetig. Also finden wir für beliebige $\epsilon \in \mathbb{R}_{>0}$ ein $\delta_1 \in \mathbb{R}_{>0}$ so, dass

$$\|f(t, y(t)) - f(s, y(s))\| \leq \epsilon/3 \qquad \text{für alle } t, s \in [a, b] \text{ mit } |t - s| \leq \delta_1 \tag{3.15}$$

gilt. Da Φ stetig auf der kompakten Menge Δ ist, finden wir für beliebige $\epsilon \in \mathbb{R}_{>0}$ ein $\delta_2 \in \mathbb{R}_{>0}$ so, dass

$$\|\Phi(t, y(t), h) - \Phi(s, y(s), h')\| \leq \epsilon/3 \qquad \text{für alle } (t, h), (s, h') \in \Delta \text{ mit } |t - s| + |h - h'| \leq \delta_2 \tag{3.16}$$

gilt. Mit Hilfe dieser Abschätzungen können wir nun den eigentlichen Beweis führen.

Teil 1: Sei nun Φ konsistent und $h \in \mathbb{R}_{>0}$ sowie $t \in [a, b - h]$ gegeben. Sei $\epsilon \in \mathbb{R}_{>0}$. Seien $\delta_1, \delta_2 \in \mathbb{R}_{>0}$ wie in (3.15) und (3.16) gewählt. Aus (3.11) folgt, dass es ein $\delta_3 \in \mathbb{R}_{>0}$ mit

$$\|\tau(t, y(t), h)\| \leq \epsilon/3 \qquad \text{für alle } t \in [a, b - h], h \in (0, \delta_3)$$

geben muss. Wir setzen $\delta := \min\{\delta_1, \delta_2, \delta_3\}$.

Seien $h \in (0, \delta)$ und $t \in [a, b-h]$ gegeben. Wir fixieren $t_+ \in [t, t+h]$ mit der Eigenschaft (3.14) und erhalten

$$\begin{aligned}\|f(t, y(t)) - \Phi(t, y(t), 0)\| &\leq \|f(t, y(t)) - f(t_+, y(t_+))\| + \|f(t_+, y(t_+)) - \Phi(t, y(t), h)\| \\ &\quad + \|\Phi(t, y(t), h) - \Phi(t, y(t), 0)\| \\ &\leq \epsilon/3 + \|\tau(t, y(t), h)\| + \epsilon/3 \leq \epsilon,\end{aligned}$$

und da ϵ beliebig gewählt war folgt (3.13).

Teil 2: Setzen wir nun voraus, dass (3.13) gilt. Sei $\epsilon \in \mathbb{R}_{>0}$, und seien wieder $\delta_1, \delta_2 \in \mathbb{R}_{>0}$ wie in (3.15) und (3.16) gewählt. Wir setzen $\delta := \min\{\delta_1, \delta_2\}$.

Seien $h \in (0, \delta)$ und $t \in [a, b - h]$ gegeben. Wir fixieren wieder $t_+ \in [t, t + h]$ mit der Eigenschaft (3.14) und erhalten

$$\begin{aligned}\|\tau(t, y(t), h)\| &= \|f(t_+, y(t_+)) - \Phi(t, y(t), h)\| \\ &\leq \|f(t_+, y(t_+)) - f(t, y(t))\| + \|f(t, y(t)) - \Phi(t, y(t), 0)\| \\ &\quad + \|\Phi(t, y(t), 0) - \Phi(t, y(t), h)\| \leq \epsilon/3 + \epsilon/3 = 2\epsilon/3,\end{aligned}$$

also folgt die Konsistenz. ∎

Aus diesem Lemma folgt direkt, dass das explizite Euler-Verfahren konsistent ist, schließlich ist es durch $\Phi(t, x, h) = f(t, x)$ definiert. Falls f Lipschitz-stetig *in beiden Argumenten* ist, ist das Euler-Verfahren sogar konsistent von Ordnung 1:

Lemma 3.6 (Konsistenz Euler) *Sei f Lipschitz-stetig im ersten und zweiten Argument, es gelte also*

$$\|f(t,x) - f(s,z)\| \leq L_f(|t-s| + \|x-z\|) \qquad \text{für alle } t,s \in [a,b], x,z \in V \tag{3.17}$$

mit einer von t, s, x und z unabhängigen Konstante $L_f \in \mathbb{R}_{\geq 0}$.

Dann ist das explizite Euler-Verfahren konsistent von Ordnung 1.

Beweis. Wir können $h_y \in \mathbb{R}_{>0}$ beliebig wählen und setzen

$$M_f := \max\{\|f(t,y(t))\| \ : \ t \in [a,b]\}, \qquad C_y := L_f(1+M_f).$$

Da die Funktion

$$t \mapsto \|f(t,y(t))\|$$

stetig und das Intervall $[a,b]$ kompakt ist, sind M_f und C_y wohldefiniert.

Seien nun $h \in (0, h_y)$ und $t \in [a, b-h]$ gegeben. Nach Definition gilt

$$\begin{aligned}\|\tau(t,y(t),h)\| &= \left\| \frac{y(t+h;t,y(t)) - y(t)}{h} - \Phi(t,y(t),h) \right\| \\ &= \|y'(t_+) - f(t,y(t))\| = \|f(t_+,y(t_+)) - f(t,y(t))\|\end{aligned}$$

für ein $t_+ \in [t, t+h]$.

Aus der Lipschitz-Stetigkeit von f folgern wir nun

$$\|\tau(t,y(t),h)\| = \|f(t_+,y(t_+)) - f(t,y(t))\| \leq L_f(|t-t_+| + \|y(t) - y(t_+)\|). \tag{3.18}$$

Auf den zweiten Term wenden wir wieder den Zwischenwertsatz an, um

$$\|y(t) - y(t_+)\| = |t-t_+| \, \|y'(t_{++})\| = |t-t_+| \, \|f(t_{++}, y(t_{++})\| \leq M_f |t-t_+|$$

für ein $t_{++} \in [t, t_+]$ zu erhalten. Insgesamt folgt aus (3.18)

$$\|\tau(t,y(t),h)\| \leq L_f(1+M_f)|t-t_+| = C_y h,$$

also die zu zeigende Behauptung. ∎

Bemerkung 3.7 (Alternativer Beweis) *Falls y' Lipschitz-stetig ist, falls es also eine Konstante $L_y \in \mathbb{R}_{\geq 0}$ mit*

$$\|y'(t) - y'(s)\| \leq L_y |t-s| \qquad \text{für alle } t,s \in [a,b]$$

gibt, erhalten wir

$$\begin{aligned}\|\tau(t,y(t),h)\| &= \left\| \frac{y(t+h;t,y(t)) - y(t)}{h} - \Phi(t,y(t),h) \right\| \\ &= \|y'(t_+) - f(t,y(t))\| = \|y'(t_+) - y'(t)\| \leq L_y \|t_+ - t\| \leq L_y h\end{aligned}$$

für ein $t_+ \in [t, t+h]$ und beliebige $h \in \mathbb{R}_{>0}$, $t \in [a, b-h]$.

Falls y zweimal stetig differenzierbar ist, kann per Mittelwertsatz L_y durch das Maximum von y'' abgeschätzt werden.

Indem wir eine Konsistenzaussage mit dem Konvergenzsatz 3.2 kombinieren, erhalten wir eine Fehlerabschätzung für das Einschrittverfahren:

Satz 3.8 (Konsistenz und Konvergenz) *Sei Φ eine Inkrementfunktion, und sei das durch sie definierte explizite Einschrittverfahren konsistent mit (2.1). Dann gilt*

$$\lim_{\substack{n\to\infty \\ h=(b-a)/n}} \sup\{\|y(t;a,y_0) - \eta(t;a,y_0)\| \ : \ t \in [a,b]_h\} = 0,$$

die diskreten Näherungslösungen konvergieren also gegen die exakte Lösung.

Falls das Verfahren konsistent von Ordnung p ist, gibt es Konstanten $C, h_y \in \mathbb{R}_{>0}$ mit

$$\sup\{\|y(t;a,y_0) - \eta(t;a,y_0)\| \ : \ t \in [a,b]_h\} \le Ch^p \quad \textit{für alle } h = (b-a)/n \in (0,h_y).$$

Beweis. Der Fall $b = a$ ist trivial, deshalb beschränken wir uns auf den Fall $b > a$.

Sei K_Φ die Konstante aus Satz 3.2, und sei

$$C_K := \begin{cases} \frac{1}{L}(e^{L(b-a)} - 1) & \text{falls } L > 0, \\ b - a & \text{falls } L = 0. \end{cases}$$

Aus Satz 3.2 folgt

$$\|y(t;a,y_0) - \eta(t;a,y_0)\| \le C_K \frac{K_\Phi}{h} \qquad \text{für alle } h \in \mathbb{R}_{>0}, t \in [a,b]_h.$$

Sei $\epsilon \in \mathbb{R}_{>0}$. Da das durch Φ definierte Verfahren konsistent mit (2.1) ist, gibt es ein $\delta \in \mathbb{R}_{>0}$ so, dass die Abschätzung

$$\max\{\|\tau(t;y(t),h)\| \ : \ t \in [a,b-h]\} \le \frac{\epsilon}{C_K} \qquad \text{für alle } h \in (0,\delta)$$

gilt. Damit erhalten wir auch

$$\|y(t;a,y_0) - \eta(t;a,y_0)\| \le C_K \frac{K_\Phi}{h} = C_K \frac{\epsilon}{C_K} = \epsilon \qquad \text{für alle } t \in [a,b]_h.$$

und $n \in \mathbb{N}$ mit $h = (b-a)/n < \delta$.

Sei nun das Verfahren konsistent von Ordnung p, und seien $C_y, h_y \in \mathbb{R}_{>0}$ die Konstanten aus (3.12). Dann haben wir

$$\|y(t;a,y_0) - \eta(t;a,y_0)\| \le C_K \frac{K_\Phi}{h} \le C_K C_y h^p \qquad \text{für alle } t \in [a,b]_h$$

und $n \in \mathbb{N}$ mit $h = (b-a)/n < h_y$. Die gewünschte Aussage folgt für $C := C_K C_y$. ∎

Korollar 3.9 (Konvergenz Euler) *Sei f in beiden Argumenten Lipschitz-stetig, es gelte also (3.17).*

Dann gibt es eine Konstante $C_{\text{eu}} \in \mathbb{R}_{>0}$ so, dass für alle $h \in \mathbb{R}_{>0}$ die per explizitem Euler-Verfahren mit Schrittweite h berechnete Näherungslösung die Abschätzung

$$\|y(t;a,y_0) - \eta(t;a,y_0)\| \le C_{\text{eu}} h \qquad \textit{für alle } t \in [a,b]_h$$

erfüllt. Insbesondere konvergiert die Näherung für $h \to 0$ gegen die exakte Lösung.

Beweis. Kombiniere Satz 3.8 mit Lemma 3.6. ∎

3.4 Lokalisierte Konvergenzaussagen

Satz 3.2 erfordert die Lipschitz-Stetigkeit der Inkrement-Funktion Φ auf dem gesamten Raum V und bietet eine Fehlerabschätzung ohne weitere Einschränkungen an K_Φ.

In der Praxis passiert es häufig, dass die Funktion f, und damit in der Regel auch die von ihr abhängende Inkrementfunktion Φ, nur in einer Umgebung der exakten Lösung Lipschitz-stetig sind. In dieser Situation kann es sinnvoll sein, Φ Lipschitz-stetig auf den gesamten Raum V fortzusetzen und dann die bereits bewiesenen Aussagen auf die modifizierte Inkrementfunktion anzuwenden.

Im Interesse der Einfachheit beschränken wir uns in diesem Abschnitt auf den Fall, dass V ein Hilbert-Raum ist. Die Grundlage unseres Fortsetzungsarguments ist die Projektion beliebiger Vektoren auf die Einheitskugel:

Lemma 3.10 (Projektion) *Wir definieren die Projektion*

$$\Pi_1(x) := \begin{cases} x & \text{falls } \|x\| \le 1, \\ \frac{x}{\|x\|} & \text{ansonsten} \end{cases} \qquad \text{für alle } x \in V.$$

Dann gilt

$$\|\Pi_1(x) - \Pi_1(z)\| \le \|x - z\| \qquad \text{für alle } x, z \in V.$$

Beweis. Seien $x, z \in V$, und sei ohne Beschränkung der Allgemeinheit $\|x\| \le \|z\|$ angenommen.

Wir unterscheiden drei Fälle: Falls $\|z\| \le 1$ gilt, folgt auch $\|x\| \le 1$ und damit $\Pi_1(x) = x$, $\Pi_1(y) = y$, so dass die Aussage trivial ist.

Falls $\|x\| \le 1 < \|z\|$ gilt, impliziert die Cauchy-Schwarz-Ungleichung bereits

$$2\langle x, z\rangle \le 2\|x\|\,\|z\| \le 2\|z\| < \|z\| + \|z\|^2$$

und wir erhalten

$$\begin{aligned} \|\Pi_1(x) - \Pi_1(z)\|^2 &= \left\| x - \frac{z}{\|z\|} \right\|^2 = \|x\|^2 - 2\langle x, z\rangle \frac{1}{\|z\|} + 1 \\ &= \|x\|^2 - 2\langle x, z\rangle + 2\langle x, z\rangle \left(1 - \frac{1}{\|z\|}\right) + \|z\|^2 - \left(1 - \frac{1}{\|z\|^2}\right)\|z\|^2 \\ &= \|x - z\|^2 + \left(1 - \frac{1}{\|z\|}\right)\left(2\langle x, z\rangle - \left(1 + \frac{1}{\|z\|}\right)\|z\|^2\right) \\ &\le \|x - z\|^2 + \left(1 - \frac{1}{\|z\|}\right)\left(\|z\| + \|z\|^2 - \|z\|^2 - \|z\|\right) = \|x - z\|^2. \end{aligned}$$

Im Falle $1 < \|x\| \le \|z\|$ schließlich ergibt sich

$$\|\Pi_1(x) - \Pi_1(z)\|^2 = \left\| \frac{x}{\|x\|} - \frac{z}{\|z\|} \right\|^2 = 2 - 2\langle x, z\rangle \frac{1}{\|x\|\,\|z\|}$$

$$= \|x\|^2 - 2\langle x, z\rangle + \|z\|^2 - (\|x\|^2 + \|z\|^2 - 2) + \left(2 - \frac{2}{\|x\|\,\|z\|}\right)\langle x, z\rangle$$
$$\leq \|x - z\|^2 - (2\|x\|\,\|z\| - 2) + \left(2 - \frac{2}{\|x\|\,\|z\|}\right)\|x\|\,\|z\| = \|x - z\|^2,$$

und der Beweis ist abgeschlossen. ∎

Wir sind nicht an Projektionen auf die Einheitskugel interessiert, sondern auf potentiell kleinere Kugeln mit Radius $\gamma \in \mathbb{R}_{>0}$, die sich einfach per

$$\Pi_\gamma(x) := \gamma\Pi_1(x/\gamma) \qquad \text{für alle } x \in V$$

definieren lassen. Offenbar gilt auch hier

$$\begin{aligned}\|\Pi_\gamma(x) - \Pi_\gamma(z)\| &= \gamma\|\Pi_1(x/\gamma) - \Pi_1(z/\gamma)\| \\ &\leq \gamma\|x/\gamma - z/\gamma\| = \|x - z\| \qquad \text{für alle } x, z \in V,\end{aligned}$$

und wir können die Fortsetzung von Φ konstruieren:

Korollar 3.11 (Lokalisierung) *Sei $y_0 \in V$, und sei $y : [a, b] \to V$ eine Lösung des Anfangswertproblems (2.1).*

Sei $\gamma \in \mathbb{R}_{>0}$. Wir definieren die Umgebungen

$$S(t) := \{x \in V \;:\; \|x - y(t)\| \leq \gamma\} \qquad \textit{für alle } t \in [a, b]_h$$

und setzen voraus, dass die Inkrementfunktion auf ihnen im zweiten Argument Lipschitzstetig ist, dass also

$$\|\Phi(t, x) - \Phi(t, z)\| \leq L\|x - z\| \qquad \textit{für alle } t \in [a, b]_h, x, z \in S(t)$$

für ein $L \in \mathbb{R}_{\geq 0}$ gilt. Es sei K_Φ/h klein genug, um

$$\gamma \geq \begin{cases} \frac{K_\Phi}{Lh}\left(e^{L(t-a)} - 1\right) & \textit{falls } L > 0, \\ \frac{K_\Phi}{h}(t - a) & \textit{falls } L = 0 \end{cases} \tag{3.19}$$

für die in (3.9) definierte Konstante sicherzustellen. Dann folgt für alle $t \in [a, b]_h$ die Abschätzung

$$\|y(t) - \eta(t)\| \leq \begin{cases} \frac{K_\Phi}{Lh}(e^{L(t-a)} - 1) & \textit{falls } L > 0, \\ \frac{K_\Phi}{h}(t - a) & \textit{falls } L = 0. \end{cases}$$

Beweis. Da die Inkrementfunktion Φ nur lokal Lipschitz-stetig ist, setzen wir sie zu einer global Lipschitz-stetigen Funktion Φ_* fort: Falls ein Vektor nicht in der durch $S(t)$ definierten Umgebung der Lösung liegt, wird er in diese Menge projiziert, indem wir

$$x_t := y(t) + \Pi_\gamma(x - y(t)) \qquad \text{für alle } t \in [a, b]_h, x \in V$$

setzen, denn nach Definition von Π_γ gilt dann $\|x_t - y(t)\| \leq \gamma$, also $x_t \in S(t)$.

Die Funktion Φ_* definieren wir durch

$$\Phi_*(t,x,h) := \Phi(t,x_t,h) \qquad \text{für alle } t \in [a,b]_h, x \in V,$$

und wir erhalten dank Lemma 3.10 die Abschätzung

$$\begin{aligned}\|\Phi_*(t,x,h) - \Phi_*(t,z,h)\| &= \|\Phi(t,x_t,h) - \Phi(t,z_t,h)\| \\ &\le L\|x_t - z_t\| \le L\|x - z\| \qquad \text{für alle } t \in [a,b]_h, x, z \in V,\end{aligned}$$

die Inkrementfunktion Φ_* ist also *global* Lipschitz-stetig im zweiten Argument.

Analog zu η definieren wir Näherungslösungen η_* für die fortgesetzte Inkrementfunktion Φ_* durch

$$\eta_*(t_{i+1}; t_*, y_*) := \eta_*(t_i; t_*, y_*) + h\Phi_*(t_i, \eta_*(t_i; t_*, y_*), h) \qquad \text{für alle } t_i \in [t_*, b]_h.$$

Wegen $y(t) \in S(t)$ folgt $\Phi_*(t, y(t), h) = \Phi(t, y(t), h)$, also auch $K_\Phi = K_{\Phi_*}$, und aus Satz 3.2 und (3.19) erhalten wir die Abschätzung

$$\|y(t; a, y_0) - \eta_*(t; a, y_0)\| \le \gamma \qquad \text{für alle } t \in [a,b]_h.$$

Daraus folgt $\eta_*(t; a, y_0) \in S(t)$, und nach Definition von Φ_* also auch $\eta_*(t; a, y_0) = \eta(t; a, y_0)$. Damit ist die gesuchte Aussage bewiesen. ∎

4 Verfahren höherer Ordnung

4.1 Konsistenzkriterium

Wie wir in Satz 3.8 gesehen haben, entscheidet die Konsistenzordnung p darüber, wie schnell sich der Fehler der Näherungslösung reduziert. Im Falle des Euler-Verfahrens bewirkt eine Halbierung der Schrittweite lediglich eine Halbierung des Fehlers, während bei einem Verfahren p-ter Ordnung der Fehler bereits um den Faktor 2^{-p} reduziert würde.

Wenn wir eine gewisse Genauigkeit $\epsilon \in \mathbb{R}_{>0}$ erreichen wollen, muss

$$\epsilon \sim h^p \sim n^{-p}$$

gelten, wir benötigen also

$$n \sim \epsilon^{-1/p}$$

Schritte des Einschrittverfahrens. Das bedeutet, dass ein Verfahren erster Ordnung eine Million Schritte zum Erreichen einer Genauigkeit von 10^{-6} benötigt, während ein Verfahren zweiter Ordnung mit nur tausend Schritten auskommen könnte, also im Idealfall *tausendmal* schneller wäre. In der Realität erfordert ein Schritt eines Verfahrens höherer Ordnung einen etwas höheren Rechenaufwand, wird aber trotzdem sehr viel schneller als ein Verfahren niedrigerer Ordnung sein.

Wir sind also daran interessiert, möglichst einfache Verfahren möglichst hoher Ordnung zu konstruieren. Dazu gehen wir ähnlich wie in Bemerkung 3.7 vor: Wegen (3.2) haben wir $y(t+h;t,y(t)) = y(t+h)$, und wenn wir $y \in C^{p+1}[a,b]$ voraussetzen, folgt mit dem Satz von Taylor

$$\begin{aligned}
\tau(t,y(t),h) &= \frac{y(t+h;t,y(t)) - y(t)}{h} - \Phi(t,x,h) \\
&= \frac{y(t+h) - y(t)}{h} - \Phi(t,x,h) \\
&= \frac{1}{h}\left(\sum_{\nu=0}^{p} \frac{h^\nu}{\nu!} y^{(\nu)}(t) + \frac{h^{p+1}}{(p+1)!} y^{(p+1)}(t_+) - y(t)\right) \\
&\quad - \sum_{\nu=0}^{p-1} \frac{h^\nu}{\nu!} \frac{\partial^\nu \Phi}{\partial h^\nu}(t,x,0) - \frac{h^p}{p!} \frac{\partial^p \Phi}{\partial h^p}(t,x,h_+) \\
&= \sum_{\nu=1}^{p} \frac{h^{\nu-1}}{\nu!} y^{(\nu)}(t) + \frac{h^p}{(p+1)!} y^{(p+1)}(t_+)
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{\nu=0}^{p-1} \frac{h^\nu}{\nu!} \frac{\partial^\nu \Phi}{\partial h^\nu}(t,x,0) - \frac{h^p}{p!} \frac{\partial^p \Phi}{\partial h^p}(t,x,h_+) \\
&= \sum_{\nu=0}^{p-1} \frac{h^\nu}{\nu!} \left(\frac{y^{(\nu+1)}(t)}{\nu+1} - \frac{\partial^\nu \Phi}{\partial h^\nu}(t,x,0) \right) + \frac{h^p}{p!} \left(\frac{y^{(p+1)}(t_+)}{p+1} - \frac{\partial^p \Phi}{\partial h^p}(t,x,h_+) \right)
\end{aligned} \tag{4.1}
$$

für Zwischenstellen $t_+ \in [t, t+h]$ und $h_+ \in [0,h]$. Da wir ein Näherungsverfahren konstruieren wollen, steht uns die Lösung y im Allgemeinen nicht zur Verfügung, also drücken wir sie mit Hilfe von f aus: Es gilt

$$
f_y(t) := f(t, y(t)) = y'(t) \qquad \text{für alle } t \in [a,b], \tag{4.2}
$$

da y die Lösung von (2.1) ist, und wir erhalten

$$
\tau(t, y(t), h) = \sum_{\nu=0}^{p-1} \frac{h^\nu}{\nu!} \left(\frac{f_y^{(\nu)}(t)}{\nu+1} - \partial_h^\nu \Phi(t,x,0) \right) + \frac{h^p}{p!} \left(\frac{f_y^{(p)}(t_+)}{p+1} - \partial_h^p \Phi(t,x,h_+) \right). \tag{4.3}
$$

Eine Konsistenzordnung von p können wir nur erwarten, falls die erste Summe verschwindet, falls also die Ableitungen bis zur Ordnung $p-1$ von f_y und Φ übereinstimmen.

Lemma 4.1 (Konsistenzkriterium) *Sei $p \in \mathbb{N}$, und sei $y \in C^{p+1}[a,b]$. Sei Φ eine im dritten Argument p-mal stetig differenzierbare Inkrementfunktion, die*

$$
(\nu+1) \frac{\partial^\nu \Phi}{\partial h^\nu}(t, y(t), 0) = f_y^{(\nu)}(t) \qquad \textit{für alle } t \in [a,b], \nu \in \{0, \dots, p-1\}
$$

erfüllt. Dann ist das durch Φ definierte Verfahren von p-ter Ordnung konsistent.

Beweis. (vgl. [2]) Sei $h_y \in \mathbb{R}_{>0}$. Da $y^{(p+1)}$ stetig ist, ist

$$
C_1 := \max\{\|y^{(p+1)}(t)\| \ : \ t \in [a,b]\} = \max\{\|f_y^{(p)}(t)\| \ : \ t \in [a,b]\} < \infty
$$

als Maximum einer stetigen Funktion auf dem kompakten Intervall $[a,b]$ wohldefiniert. Da Φ im dritten Argument p-mal stetig differenzierbar ist, ist

$$
C_2 := \max\left\{ \left\| \frac{\partial^p \Phi}{\partial h^p}(t, y(t), h) \right\| \ : \ t \in [a,b], h \in [0, b-t] \right\} < \infty
$$

als Maximum einer stetigen Funktion auf der (nach Heine-Borel) kompakten Menge

$$
\Delta := \{(t,h) \ : \ t \in [a,b], h \in [0, b-t]\}
$$

ebenfalls wohldefiniert.

Sei $h \in (0, h_y]$ und $t \in [a, b-h]$. Durch Einsetzen in (4.3) folgt sofort

$$
\tau(t, y(t), h) = \frac{h^p}{p!} \left(\frac{f_y^{(p)}(t_+)}{p+1} - \partial_h^p \Phi(t,x,h_+) \right) = \frac{h^p}{p!} \left(\frac{y^{(p+1)}(t_+)}{p+1} - \partial_h^p \Phi(t,x,h_+) \right)
$$

mit Zwischenpunkten $t_+ \in [t, t+h]$ und $h_+ \in [0, h]$. Nach Definition von C_1 und C_2 erhalten wir also

$$\|\tau(t, y(t), h)\| \leq h^p \left(\left\| \frac{f_y^{(p)}(t_+)}{(p+1)!} \right\| + \left\| \frac{\partial_h^p \Phi(t, x, h_+)}{p!} \right\| \right) \leq h^p \left(\frac{C_1}{(p+1)!} + \frac{C_2}{p!} \right) = C_y h^p$$

für die Konstante

$$C_y := \frac{C_1}{(p+1)!} + \frac{C_2}{p!},$$

und damit ist (3.12) bewiesen. ∎

Dieses Resultat lässt sich als Verallgemeinerung von Lemma 3.5 interpretieren: $\Phi(t, x, 0) = f(t, x)$ impliziert Konsistenz. Falls f differenzierbar (es reicht auch Lipschitz-Stetigkeit) ist, folgt Konsistenz erster Ordnung. Falls zusätzliche Ableitungen existieren und stetig sind, erhalten wir höhere Konsistenzordnungen.

Lemma 4.1 kann verwendet werden, um Einschrittverfahren beliebig hoher Konsistenzordnung zu entwickeln, indem man die Struktur der Funktion f_y ausnutzt: Nach (4.2) gilt

$$\begin{aligned} f_y'(t) &= Df(t, y(t)) \cdot (1, y'(t)) = Df(t, y(t)) \cdot (1, f(t, y(t))) \\ &= \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial x}(t, y(t)) f(t, y(t)) \qquad \text{für alle } t \in [a, b], \end{aligned}$$

wir können also diese Größe berechnen, falls uns die Ableitungen von f zur Verfügung stehen. Wir definieren

$$\Phi(t, x, h) := f(t, x) + h \left(\frac{\partial f}{\partial t}(t, x) + \frac{\partial f}{\partial x}(t, x) x \right) \quad \text{für alle } t \in [a, b], h \in [0, b-t], x \in V,$$

und Lemma 4.1 zeigt, dass das durch diese Inkrementfunktion definierte Verfahren die Konsistenzordnung 2 besitzt.

Wenn uns höhere Ableitungen von f_y, also letztendlich von f, zur Verfügung stehen, können wir in dieser Weise prinzipiell Verfahren beliebig hoher Konsistenzordnung konstruieren.

In der Praxis stehen uns sehr oft die Ableitungen von f nicht zur Verfügung, so dass wir uns für Verfahren interessieren, die eine höhere Konsistenzordnung auch ohne diese zusätzliche Information erzielen (schließlich können wir eine Funktion statt per Taylor-Entwicklung auch durch Lagrange-Interpolation approximieren).

Beispiel 4.2 (Heun) *Es ist möglich, aus einem Quadraturverfahren höherer Ordnung eine Inkrementfunktion höherer Konsistenzordnung zu konstruieren. Als Beispiel verwenden wir die Trapezregel*

$$\int_t^{t+h} g(s)\, ds \approx \frac{h}{2}(g(t) + g(t+h)),$$

die bei einem zweimal differenzierbaren Integranden einen Fehler der Ordnung h^3 aufweist.

Wir wenden diese Regel auf die Integraldarstellung aus Lemma 2.2 an und erhalten

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))\, ds \approx y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y(t+h))),$$

also ein zunächst implizites Einschrittverfahren.

Falls h klein genug ist, dürfen wir erwarten, dass wir mit der Fixpunktiteration

$$y^{(i+1)}(t+h) := y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y^{(i)}(t+h))) \qquad \text{für alle } i \in \mathbb{N}_0$$

schnell eine gute Näherung von $y(t+h)$ erhalten können.

Wenn wir zur Bestimmung des Startwerts das explizite Euler-Verfahren verwenden, erhalten wir $y^{(0)}(t+h) := y(t) + hf(t, y(t))$ und nach dem ersten Iterationsschritt

$$\begin{aligned} y^{(1)}(t+h) &:= y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y^{(0)}(t+h)) \\ &= y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y(t) + hf(t, y(t)))). \end{aligned}$$

Für eine hinreichend kleine Schrittweite h können wir davon ausgehen, dass dieser Wert bereits eine gute Näherung von $y(t+h)$ darstellt. Die entsprechende Inkrementfunktion lautet

$$\Phi(t, x, h) := \frac{1}{2}\left(f(t, x) + f(t+h, x + hf(t, x))\right)$$

und definiert das Heun-Verfahren.

Falls f einmal stetig differenzierbar ist, folgt aus $\Phi(t, x, 0) = f(t, x)$ nach Lemma 4.1 bereits, dass das Verfahren von erster Ordnung konsistent ist. Falls f zweimal stetig differenzierbar ist, erhalten wir per Kettenregel

$$\begin{aligned} \frac{\partial \Phi}{\partial h}(t, x, h) &= \frac{1}{2} Df(t+h, x + hf(t, y)) \cdot (1, f(t, x)), \\ f_y'(t) &= Df(t, y(t)) \cdot (1, y'(t)) = Df(t, y(t)) \cdot (1, f(t, y(t))), \\ 2\frac{\partial \Phi}{\partial h}(t, y(t), 0) &= f_y'(t), \end{aligned}$$

also ist das Verfahren gemäß Lemma 4.1 in diesem Fall von zweiter Ordnung konsistent.

Statt der Trapezregel können wir auch mit der Mittelpunktregel arbeiten, die ebenfalls einen Fehler in der Größenordnung von h^3 erwarten lässt. Auch hier müssen wir den Wert im Mittelpunkt des Intervalls geeignet approximieren, beispielsweise durch das Euler-Verfahren:

Beispiel 4.3 (Runge) *Analog können wir auch die Mittelpunktregel verwenden, um eine Inkrementfunktion zu konstruieren: Wir gehen wieder von Lemma 2.2 aus und setzen*

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))\, ds \approx y(t) + hf\left(t + \frac{h}{2}, y\left(t + \frac{h}{2}\right)\right).$$

Wie schon in Beispiel 4.2 verwenden wir das explizite Euler-Verfahren, um die Approximation

$$y\left(t+\frac{h}{2}\right) \approx y(t) + \frac{h}{2} f(t, y(t))$$

zu gewinnen und erhalten

$$y(t+h) \approx y(t) + hf\left(t+\frac{h}{2}, y(t) + \frac{h}{2} f(t, y(t))\right).$$

Die zugehörige Inkrementfunktion

$$\Phi(t, x, h) := f\left(t+\frac{h}{2}, x + \frac{h}{2} f(t, x)\right)$$

definiert das Runge- *oder auch* Euler-Collatz-Verfahren.

Falls f einmal stetig differenzierbar ist, folgt aus $\Phi(t, x, 0) = f(t, x)$ per Lemma 4.1, dass das Verfahren von erster Ordnung konsistent ist. Falls f zweimal stetig differenzierbar ist, impliziert die Kettenregel

$$\begin{aligned}
\frac{\partial \Phi}{\partial h}(t, x, h) &= Df\left(t+\frac{h}{2}, x+\frac{h}{2} f(t, x)\right) \cdot \left(\frac{1}{2}, \frac{1}{2} f(t, x)\right) \\
&= \frac{1}{2} Df\left(t+\frac{h}{2}, x+\frac{h}{2} f(t, x)\right) \cdot (1, f(t, x)), \\
2\frac{\partial \Phi}{\partial h}(t, y(t), 0) &= Df(t, y(t)) \cdot (1, f(t, y(t))) = f_y'(t),
\end{aligned}$$

und dank Lemma 4.1 können wir schließen, dass das Verfahren auch von zweiter Ordnung konsistent ist.

Es stellt sich die Frage, ob man durch Quadraturformeln höherer Ordnung auch zu Inkrementfunktionen höherer Ordnung gelangen kann. Im Prinzip können wir eine Quadraturformel

$$\int_t^{t+h} f(s, y(s))\, ds \approx \sum_{i=1}^{m} \omega_i f(s_i, y(s_i))$$

verwenden, müssen dann aber brauchbare Näherungswerte für $y(s_i)$ in allen Quadraturpunkten zur Verfügung stellen. Eine Analyse der Fehlerfortpflanzung im Quadraturverfahren zeigt, dass eine Quadraturordnung von $p+1$, also eine Konsistenzordnung von p, nur dann zu erwarten ist, wenn die Näherungswerte für $y(s_i)$ genau bis auf einen Fehler der Ordnung $p-1$ sind. Diese Eigenschaft ist im Allgemeinen nur schwer sicherzustellen.

Im Falle des Heun- und Euler-Collatz-Verfahrens profitieren wir davon, dass das explizite Euler-Verfahren eine Näherung erster Ordnung für $y(t+h)$ bzw. $y(t+h/2)$ zur Verfügung stellt, für höhere Ordnungen ist dieses Ziel schwieriger zu erreichen.

4.2 Runge-Kutta-Verfahren

Sowohl das Heun- als auch das Euler-Collatz-Verfahren basieren darauf, die Funktion f in Punkten auszuwerten, die von vorangehenden Auswertungen der Funktion abhängen können. Allgemein haben wir also die Struktur

$$k_i = f\left(t + c_i h, x + h\sum_{j=1}^{i-1} a_{ij} k_j\right) \qquad \text{für } i \in \{1, \dots, s\}, \tag{4.4a}$$

$$\Phi(t, x, h) = \sum_{i=1}^{s} b_i k_i \tag{4.4b}$$

eines expliziten Verfahrens, das mit Hilfe von s Auswertungen der Funktion f eine Inkrementfunktion definiert. Die entscheidenden Parameter sind die Vektoren $c \in \mathbb{R}^s$, die die Zeitpunkte für die Auswertungen angeben, die Matrix $A = (a_{ij})_{i,j=1}^s$, die angibt, wie die i-te Funktionsauswertung von den vorangehenden Auswertungen beeinflusst wird, und der Vektor $b \in \mathbb{R}^s$, der beschreibt, wie die einzelnen Funktionsauswertungen kombiniert werden müssen, um die Inkrementfunktion zu erhalten.

Die durch (4.4) beschriebenen Verfahren bezeichnen wir als *Runge-Kutta-Verfahren* der Stufe s. Die bisher betrachteten Einschrittverfahren lassen sich als Runge-Kutta-Verfahren interpretieren: Das explizite Euler-Verfahren ist einstufig mit den Parametern

$$A = \begin{pmatrix} 0 \end{pmatrix}, \qquad c = \begin{pmatrix} 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \end{pmatrix},$$

das Heun-Verfahren ist zweistufig mit

$$A = \begin{pmatrix} 0 & \\ 1 & 0 \end{pmatrix}, \qquad c = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix},$$

und das Runge-Verfahren ist ebenfalls zweistufig mit

$$A = \begin{pmatrix} 0 & \\ 1/2 & 0 \end{pmatrix}, \qquad c = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Kompakt lassen sich Runge-Kutta-Verfahren in Form des *Butcher-Schemas*

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

schreiben, die drei oben erwähnten Verfahren nehmen dann die Form

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

an. Anschaulich entsprichen bei diesem Schema die ersten s Zeilen jeweils einer Auswertung von f: die c-Spalte gibt den Zeitpunkt an, die restlichen Spalten beschreiben

den Ort. Die letzten s Spalten des Butcher-Schemas gehören jeweils zu einer der Zwischengrößen k_i und beschreiben, wie diese Größe skaliert werden muss, bevor sie in die Berechnung der nächsten Größe beziehungsweise des Endergebnisses eingeht.

Durch Taylor-Entwicklung lassen sich aus den in Lemma 4.1 gegebenen Bedingungen nichtlineare Gleichungssysteme herleiten, die zur Konstruktion von Runge-Kutta-Verfahren höherer Ordnung verwendet werden können. Dem Gleichungssystem kann man entnehmen, dass Quadraturformeln einen guten Lösungsansatz bieten: Für den Vektor c lassen sich Quadraturpunkte auf dem Intervall $[0, 1]$ verwenden, für den Vektor b die entsprechenden Quadraturgewichte.

Wenn man die Simpson-Quadraturformel

$$\int_0^1 g(s)\,ds \approx \frac{1}{6}(g(0) + 4g(1/2) + g(1))$$

aus Symmetriegründen in der Form

$$\int_0^1 g(s)\,ds \approx \frac{1}{6}g(0) + \frac{1}{3}g(1/2) + \frac{1}{3}g(1/2) + \frac{1}{6}g(1)$$

schreibt und die passenden Koeffizienten a_{ij} berechnet, erhält man das Schema

0	0			
1/2	1/2	0		
1/2	0	1/2	0	
1	0	0	1	0
	1/6	1/3	1/3	1/6

das das *klassischen Runge-Kutta-Verfahren* vierter Stufe beschreibt. Es lässt sich nachweisen, dass dieses Verfahren die Konsistenzordnung vier besitzt.

Entsprechend kann man auch die 3/8-Quadraturformel von Newton verwenden, die durch

$$\int_0^1 g(s)\,ds \approx \frac{1}{8}g(0) + \frac{3}{8}g(1/3) + \frac{3}{8}g(2/3) + \frac{1}{8}g(1)$$

gegeben ist und zu dem Butcher-Schema

0	0			
1/3	1/3	0		
2/3	−1/3	1	0	
1	1	−1	1	0
	1/8	3/8	3/8	1/8

führt. Man kann zeigen, dass auch das zu diesem Schema gehörende Runge-Kutta-Verfahren die Konsistenzordnung vier besitzt.

Im Interesse einer hohen Genauigkeit sind wir natürlich an Verfahren möglichst hoher Konsistenzordnung interessiert. Die Konstruktion derartiger Verfahren ist im Allgemeinen relativ schwierig, aber es ist immerhin möglich, eine obere Schranke für die maximal erreichbare Ordnung anzugeben, indem man ein einfaches Beispielproblem analysiert.

Lemma 4.4 (Exponentialfunktion) *Sei ein explizites Runge-Kutta-Verfahren der Stufe s durch (A, b, c) gegeben.*

Wendet man es auf das für $\lambda \in \mathbb{R}$ gegebene einfache Anfangswertproblem

$$y_\lambda(0) = 1, \qquad y'_\lambda(t) = \lambda y_\lambda(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0} \tag{4.5}$$

an, dessen Lösung offenbar durch $y_\lambda(t) = e^{\lambda t}$ gegeben ist, so gilt für die durch das Verfahren definierte Näherungslösung

$$\eta_\lambda(t+h; t, x) := x + h\Phi(t, x, h) = g(\lambda h)x \quad \text{für alle } h \in \mathbb{R}_{>0}, t \in [a, b-h], x, \lambda \in \mathbb{R}$$

mit einem Polynom $g \in \mathcal{P}_s$, das nur von A, b und c abhängt. Dieses Polynom wird manchmal als Stabilitätsfunktion *bezeichnet.*

Beweis. Im trivialen Fall $\lambda = 0$ setzen wir $g \equiv 1$ und sind fertig.

Sei nun $\lambda \neq 0$ angenommen. Einsetzen der Differentialgleichung in (4.4) führt auf die Gleichung

$$\begin{aligned} k_i &= \lambda \left(x + h \sum_{j=1}^{i-1} a_{ij} k_j \right), \\ \lambda^{-1} k_i &= \left(x + h \sum_{j=1}^{i-1} a_{ij} k_j \right) = x + h\lambda \sum_{j=1}^{i-1} a_{ij} (\lambda^{-1} k_j) \qquad \text{für alle } i \in \{1, \ldots, s\}, \end{aligned}$$

die es nahelegt, einen polynomialen Zusammenhang zwischen den skalierten Hilfsvektoren $\lambda^{-1} k_i$ zu vermuten.

Konkret wollen wir nun die Existenz von Polynomen $q_i \in \mathcal{P}_{i-1}$ beweisen, die

$$\lambda^{-1} k_i = q_i(\lambda h)x \qquad \text{für alle } i \in \{1, \ldots, s\} \tag{4.6}$$

erfüllen. Wir verwenden dazu eine abschnittsweise Induktion, zeigen also

$$\lambda^{-1} k_i = q_i(\lambda h)x \qquad \text{für alle } i \in \{1, \ldots, \ell\} \tag{4.7}$$

für alle $\ell \in \{1, \ldots, s\}$. Der Induktionsanfang $\ell = 1$ ist einfach: Für $i = 1$ haben wir

$$\lambda^{-1} k_i = \lambda^{-1} f(t_i, x) = x,$$

also gilt die Behauptung mit $q_1 = 1 \in \mathcal{P}_{i-1}$.

Gelte nun (4.7) für ein $\ell \in \{1, \ldots, s-1\}$. Dann erhalten wir nach (4.4) und dank der Induktionsvoraussetzung

$$\lambda^{-1} k_{\ell+1} = \lambda^{-1} f \left(t + c_{\ell+1} h, x + h \sum_{j=1}^{\ell} a_{\ell+1,j} k_j \right) = x + \lambda h \sum_{j=1}^{\ell} a_{\ell+1,j} \lambda^{-1} k_j$$

$$= x + \lambda h \sum_{j=1}^{\ell} a_{\ell+1,j} q_j(\lambda h) x = \left(1 + \lambda h \sum_{j=1}^{\ell} a_{\ell+1,j} q_j(\lambda h)\right) x = q_{\ell+1}(\lambda h) x$$

für das Polynom

$$q_{\ell+1}(\zeta) := 1 + \zeta \sum_{j=1}^{\ell} a_{\ell+1,j} q_j(\zeta),$$

das in $\mathcal{P}_\ell$ liegen muss, da die Polynome q_j für $j \leq \ell$ höchstens in $\mathcal{P}_{\ell-1}$ liegen können. Damit ist die Induktion abgeschlossen und (4.7) sowie auch (4.6) bewiesen.

Die Näherungslösung ist nach (4.4) gegeben durch

$$\begin{aligned}\eta(t+h;t,x) &= x + h\Phi(t,x,h) = x + h\sum_{i=1}^{s} b_i k_i = x + h\lambda \sum_{i=1}^{s} b_i \lambda^{-1} k_i \\ &= x + h\lambda \sum_{i=1}^{s} b_i q_i(\lambda h) x = \left(1 + \lambda h \sum_{i=1}^{s} b_i q_i(\lambda h)\right) x = g(\lambda h) x\end{aligned}$$

für das Polynom

$$g(\zeta) := 1 + \zeta \sum_{i=1}^{s} b_i q_i(\zeta),$$

das wegen $q_i \in \mathcal{P}_{i-1}$ in $\mathcal{P}_s$ enthalten sein muss. ∎

Offenbar steht die Stabilitätsfunktion g in enger Beziehung zum Approximationsfehler: Die Differenz zwischen exakter Lösung y_λ und approximativer Lösung η_λ ist gerade durch

$$y_\lambda(t+h) - \eta_\lambda(t+h;t,y_\lambda(t)) = e^{\lambda(t+h)} - g(\lambda h) e^{\lambda t} = e\lambda t(e^{\lambda h} - g(\lambda h))$$

gegeben, für eine hohe Konsistenzordnung muss also g eine möglichst gute Approximation der Exponentialfunktion sein. Aus dieser Beobachtung ergibt sich die folgende Schranke für die von einem s-stufigen expliziten Runge-Kutta-Verfahren erreichbare Konsistenzordnung:

Lemma 4.5 (Maximale Ordnung) *Die Konsistenzordnung p eines s-stufigen expliziten Runge-Kutta-Verfahrens beträgt höchstens s. Im Falle $p = s$ gilt*

$$g(\zeta) = \sum_{i=0}^{m} \frac{\zeta^i}{i!}. \tag{4.8}$$

Beweis. Sei ein s-stufiges explizites Runge-Kutta-Verfahren gegeben, und sei g die entsprechende Stabilitätsfunktion. Wir untersuchen den lokalen Diskretisierungsfehler für das Problem (4.5). Er ist gegeben durch

$$\tau(t,x,h) = \frac{y_\lambda(t+h) - y_\lambda(t)}{h} - \Phi(t, y_\lambda(t), h) = \frac{y_\lambda(t+h) - (y_\lambda(t) + h\Phi(t, y_\lambda(t), h))}{h}$$

$$= \frac{y_\lambda(t+h) - \eta(t+h;t,y_\lambda(t))}{h} = \frac{e^{\lambda(t+h)} - g(\lambda h)e^{\lambda t}}{h} = e^{\lambda t}\frac{e^{\lambda h} - g(\lambda h)}{h}.$$

Durch Einsetzen der Exponentialreihe erhalten wir

$$\tau(t,x,h) = \frac{e^{\lambda t}}{h}\left(\sum_{i=0}^{s}\frac{(\lambda h)^i}{i!} - g(\lambda h)\right) + \frac{e^{\lambda t}}{h}\frac{(\lambda h)^{s+1}}{(s+1)!} + \frac{e^{\lambda t}}{h}\frac{(\lambda h_+)^{s+2}}{(s+2)!}$$

für ein $h_+ \in [0,h]$.

Falls (4.8) gilt, folgt unmittelbar

$$\frac{e^{\lambda t}\lambda^{s+1}}{(s+1)!}h^s \le \|\tau(t,x,h)\| \le \frac{e^{\lambda t}\lambda^{s+1}}{(s+1)!}h^s + \frac{e^{\lambda t}\lambda^{s+2}}{(s+2)!}h^{s+1} \qquad \text{für alle } h \in \mathbb{R}_{>0},$$

also ist das Verfahren konsistent von Ordnung s und eine höhere Ordnung auch nicht möglich.

Falls (4.8) nicht gilt, kann das entscheidende Polynom

$$\zeta \mapsto \sum_{i=0}^{s}\frac{\zeta^i}{i!} - g(\zeta)$$

in Null höchstens eine Nullstelle der Ordnung s besitzen, also gibt es eine Konstante $c \in \mathbb{R}_{>0}$ und ein $h_0 \in \mathbb{R}_{>0}$ so, dass

$$\left|\sum_{i=0}^{s}\frac{\zeta^i}{i!} - g(\zeta)\right| \ge c\zeta^s \qquad \text{für alle } \zeta \in (0,h_0).$$

gilt, also auch

$$\begin{aligned}|\tau(t,x,h)| &\ge \frac{e^{\lambda t}}{h}\left|\sum_{i=0}^{s}\frac{(\lambda h)^i}{i!} - g(\lambda h)\right| - \frac{e^{\lambda t}\lambda^{s+1}}{(s+1)!}h^s \\ &\ge ce^{\lambda t}\frac{(\lambda h)^s}{h} - \frac{e^{\lambda t}\lambda^{s+1}}{(s+1)!}h^s = ce^{\lambda t}\lambda^s h^{s-1} - \frac{e^{\lambda t}\lambda^{s+1}}{(s+1)!}h^s \quad \text{für alle } h \in (0,h_0/\lambda).\end{aligned}$$

Demzufolge kann die Konsistenzordnung in diesem Fall höchstens $s-1$ betragen. ∎

Die Umkehrung dieser Aussage gilt nicht: Es kann vorkommen, dass zu einer gegebenen Konsistenzordnung p kein p-stufiges Runge-Kutta-Verfahren existiert.

Bemerkung 4.6 (Implizite Verfahren) *Für ein explizites Runge-Kutta-Verfahren muss die zugehörige Matrix A eine strikte untere Dreiecksmatrix sein. Wäre sie es nicht, könnten die Größen $k_1,\dots,k_s$ nicht der Reihe nach explizit berechnet werden.*

Wenn wir beliebige Matrizen A zulassen, erhalten wir im Allgemeinen implizite *Runge-Kutta-Verfahren. Diese Verfahren unterliegen nicht der in Lemma 4.5 bewiesenen Schranke für die Konsistenzordnung, sondern können wesentlich höhere Genauigkeiten erzielen.*

Beispielsweise können auch in diesem Fall die Vektoren c und b entsprechend einer Quadraturformel gewählt werden, etwa entsprechend einer Gauß-Formel. Es ist bekannt, dass eine Gauß-Formel mit s Quadraturpunkten exakt bis zur Ordnung $2s-1$ ist, und man kann zeigen, dass das mit Hilfe einer derartigen Formel definierte implizite Runge-Kutta-Verfahren die Konsistenzordnung $2s$ besitzt.

Es ist auch bekannt, dass eine Quadraturformel mit s Quadraturpunkten keine höhere Exaktheitsordnung erreichen kann, dementsprechend kann auch ein s-stufiges Runge-Kutta-Verfahren die Konsistenzordnung von $2s$ nicht überschreiten.

Ein so definiertes allgemeines Runge-Kutta-Verfahren hat den Nachteil, dass in jedem Schritt ein nichtlineares Gleichungssystem mit s unbekannten Vektoren k_i gelöst werden muss, wodurch ein hoher Rechenaufwand zustande kommen kann.

Einen Mittelweg beschreiten semi-implizite *Runge-Kutta-Verfahren, bei denen die Matrix A zwar eine untere Dreiecksmatrix ist, aber Diagonaleinträge ungleich Null zugelassen werden. Dann können die Vektoren $k_1, k_2, \ldots, k_s$ der Reihe nach bestimmt werden, indem eine Folge von s nichtlinearen Gleichungssystemen für jeweils einen einzelnen Vektor gelöst wird.*

4.3 Extrapolationsverfahren

Die Abschätzung von Satz 3.2 in Kombination mit $e^{Lh}/L \approx h$ lässt uns erwarten, dass bei einem Verfahren der Konsistenzordnung p für hinreichend kleine Schrittweiten h eine Abschätzung der Form

$$\|y(t+h;t,y(t)) - \eta(t+h;t,y(t))\| \leq Ch^{p+1} \qquad \text{für alle } h \in (0,h_y) \tag{4.9}$$

gilt. Falls wir annehmen, dass sich der Fehler um den Punkt $h = 0$ in eine Taylor-Reihe der Ordnung $p+2$ entwickeln lässt, müssen wegen (4.9) die ersten $p+1$ Terme verschwinden, und es bleibt eine Abschätzung der Form

$$\|y(t+h;t,y(t)) - \eta(t+h;t,y(t)) - c_\tau h^{p+1}\| \leq C_c h^{p+2} \qquad \text{für alle } h \in (0,h_y) \tag{4.10}$$

mit einer geeigneten Konstanten $C_c \in \mathbb{R}_{\geq 0}$ und dem Vektor c_τ, der mit dem Term der Ordnung $p+1$ der Taylor-Entwicklung korrespondiert.

Mit Hilfe dieser Abschätzung können wir die Konsistenzordnung des Verfahrens verbessern: Wir berechnen eine Lösung $\eta_1(t+h;t,y(t))$ mit dem ursprünglichen Verfahren und eine zweite Lösung $\eta_2(t+h;t,y(t))$ mit einem Verfahren zu der halbierten Schrittweite $h/2$. Die Voraussetzung (4.10) impliziert dann

$$\begin{aligned} y(t+h;t,y(t)) &\approx \eta_1(t+h;t,y(t)) + c_\tau h^{p+1}, \\ y(t+h;t,y(t)) &\approx \eta_2(t+h;t,y(t)) + c_\tau h(h/2)^p, \end{aligned}$$

und Multiplikation der zweiten Gleichung mit 2^p und Subtraktion beider Gleichungen ergibt

$$(2^p-1)y(t+h;t,y(t)) \approx 2^p\eta_2(t+h;t,y(t)) - \eta_1(t+h;t,y(t)),$$

$$y(t+h;t,y(t)) \approx \eta_3(t+h;t,y(t)) := \frac{2^p\eta_2(t+h;t,y(t)) - \eta_1(t+h;t,y(t))}{2^p-1}.$$

Ausgehend von (4.10) erhalten wir

$$\begin{aligned}
&\|y(t+h;t,y(t)) - \eta_3(t+h;t,y(t))\| \\
&\qquad = \left\| \frac{2^p-1}{2^p-1} y(t+h;t,y(t)) - \frac{2^p\eta_2(t+h;t,y(t)) - \eta_1(t+h;t,y(t))}{1-2^p} \right\| \\
&\qquad = \frac{1}{2^p-1} \|2^p(y(t+h;t,y(t)) - \eta_2(t+h;t,y(t))) \\
&\qquad\qquad - (y(t+h;t,y(t)) - \eta_1(t+h;t,y(t)))\| \\
&\qquad = \frac{1}{2^p-1} \|2^p(y(t+h;t,y(t)) - \eta_2(t+h;t,y(t)) - c_\tau h(h/2)^p) \\
&\qquad\qquad - (y(t+h;t,y(t)) - \eta_1(t+h;t,y(t)) - c_\tau h^{p+1}\| \\
&\qquad \le \frac{1}{2^p-1} (2^p \|y(t+h;t,y(t)) - \eta_2(t+h;t,y(t)) - c_\tau h(h/2)^p\| \\
&\qquad\qquad + \|y(t+h;t,y(t)) - \eta_1(t+h;t,y(t)) - c_\tau h^{p+1}\|) \\
&\qquad \le \frac{1}{2^p-1} (2^p C_c h^{p+2} + C_c h^{p+2}) = \frac{C_c(2^p+1)}{2^p-1} h^{p+2},
\end{aligned}$$

also besitzt das durch η_3 gegebene Einschrittverfahren die Konsistenzordnung $p+1$.

Die Konstruktion von η_3 entspricht der bekannten *Grenzwertextrapolation*: Im Idealfall würden wir mit Schrittweite 0 rechnen wollen, aber das ist praktisch nicht durchführbar. Stattdessen rechnen wir mit den Schrittweiten h und $h/2$ und bestimmen das lineare Interpolationspolynom, das die für diese Schrittweiten erhaltenen Werte trifft, und werten es an der Stelle Null aus. Dreh- und Angelpunkt der Extrapolation ist die Kenntnis einer Entwicklung der Form (4.10): Wenn eine solche *asymptotische Entwicklung* vorliegt, können wir Einschrittverfahren hoher Ordnung durch einfache Kombination von Verfahren niedriger Ordnung konstruieren, allerdings wird der Rechenaufwand dabei in der Regel deutlich höher als etwa bei einem Runge-Kutta-Verfahren ausfallen: Schon in unserem Beispiel erfordert der Extrapolationsansatz den *dreifachen* Rechenaufwand des Basisverfahrens.

Da bei einem Extrapolationsansatz die Schrittweite als entscheidender Parameter berücksichtigt werden muss, benötigen wir eine etwas erweiterte Notation. Wenn wir eine verbesserte Approximation mit Hilfe von $n \in \mathbb{N}$ Teilschritten der Schrittweite $h_n := h/n$ berechnen wollen, müssen wir

$$\begin{aligned}
\eta_n(t) &:= \eta(t), \\
\eta_n(t+(i+1)h_n) &:= \eta_n(t+ih_n) + h_n\Phi(t+ih_n, \eta_n(t+ih_n), h_n)
\end{aligned}$$

für alle $i \in \{0,\dots,n-1\}$ bestimmen und erhalten schließlich die Näherung $\eta_n(t+h)$. Ideal wäre $n=\infty$, denn dann wäre $h_n = 0$ und aus unserem Konvergenzsatz 3.2 und der Konsistenz des zugrundeliegenden Näherungsverfahrens würde $\eta_n = y(t+h)$ folgen.

Da die Berechnung unendlich vieler Zwischenschritte zu lange dauern würde, müssen wir uns auf endlich viele von ihnen beschränken und hoffen, dass wir trotzdem eine gute Approximation des Falls $h_n = 0$ erhalten. Abstrakt bedeutet das, dass wir die Funktion

$$\chi_{t,h} : \{h_n = h/n \ : \ n \in \mathbb{N}\} \to V, \qquad h_n \mapsto \eta_n(t+h)$$

untersuchen und sie stetig in Null fortsetzen möchten. Damit das sinnvoll möglich ist, müssen wir etwas über das Verhalten von $\chi_{t,h}$ in der Nähe der Null wissen: Wir brauchen eine *asymptotische Entwicklung*.

Genauer gesagt gehen wir davon aus, dass es Konstanten $C_c \in \mathbb{R}_{>0}$ und $h_y \in \mathbb{R}_{>0}$ und Funktionen $e_1, \ldots, e_k \in C([a,b], V)$ so gibt, dass

$$\|\eta_n(t+h) + e_1(t)h_n^p + e_2(t)h_n^{p+\omega} + \ldots + e_k(t)h_n^{p+\omega(k-1)} - y(t+h)\| \leq C_c h_n^{p+\omega k} \quad (4.11)$$

für alle $h \in (0, h_y)$, alle $t \in [a, b-h]$ und alle $n \in \mathbb{N}$ gilt. Aus dieser Abschätzung folgt, dass

$$\widehat{\chi}_{t,h}(h_n) := y(t+h) - e_1(t)h_n^p - e_2(t)h_n^{p+\omega} - \ldots - e_k(t)h_n^{p+\omega(k-1)} \quad (4.12)$$

eine gute Approximation von $\eta_n(t+h)$ ist. Wenn wir dieses Polynom konkret berechnen könnten, hätten wir die Möglichkeit, den Wert $\widehat{\chi}_{t,h}(0) = y(t+h)$ zu bestimmen, also die exakte Lösung. Leider ist es nicht so einfach: Das Polynom $\widehat{\chi}_{t,h}$ lässt sich nicht direkt bestimmen.

Stattdessen approximieren wir es: Da $\widehat{\chi}_{t,h}(h_n)$ eine gute Approximation von $\eta_n(t+h)$ ist, können wir eine gute Approximation von $\widehat{\chi}_{t,h}$ gewinnen, indem wir $\eta_n(t+h)$ in ausgewählten Punkten h_n interpolieren. Das so konstruierte Interpolationspolynom $\tilde{\chi}_{t,h}$ erfüllt

$$\tilde{\chi}_{t,h}(h_n) = \eta_n(t+h) \approx \widehat{\chi}_{t,h}(h_n) \quad (4.13)$$

in allen Interpolationspunkten h_n, und wir können es in Null auswerten, um

$$\tilde{\chi}_{t,h}(0) \approx \widehat{\chi}_{t,h}(0) = y(t+h)$$

zu bestimmen. Zur Definition der Stützstellen der Interpolation geben wir Schrittzahlen $n_0 < n_1 < \ldots < n_k$ mit korrespondierenden Schrittweiten $h_{n_i} = h/n_i$ vor, berechnen die Näherungslösungen

$$\eta_i := \eta_{n_i}(t+h) \qquad \text{für alle } i \in \{0, \ldots, k\},$$

und suchen ein Interpolationspolynom $\tilde{\chi}_{t,h}$, das

$$\tilde{\chi}_{t,h}(h_{n_i}) = \eta_i \qquad \text{für alle } i \in \{0, \ldots, k\} \quad (4.14)$$

erfüllt. Infolge der speziellen Gestalt von $\widehat{\chi}_{t,h}$ bietet es sich an, das Polynom $\tilde{\chi}_{t,h}$ in dem Polynomraum

$$\mathcal{V}_k := \left\{ x \mapsto \alpha_0 + \sum_{j=1}^{k} \alpha_j x^{p+\omega(j-1)} \ : \ \alpha_0, \ldots, \alpha_k \in V \right\}$$

zu suchen. Zur Analyse empfiehlt es sich, auch das reellwertige Gegenstück

$$\mathcal{W}_k := \left\{ x \mapsto \beta_0 + \sum_{j=1}^{k} \beta_j x^{p+\omega(j-1)} \;:\; \beta_0, \ldots, \beta_k \in \mathbb{R} \right\}$$

zu untersuchen. Aus Existenz- und Eindeutigkeitsaussagen für den reellwertigen Fall lassen sich dann einfach die korrespondierenden Aussagen für den allgemeinen Fall gewinnen. Da $\mathcal{W}_k$ von den üblichen Polynomräumen abweicht (einige Monome fehlen), müssen wir zunächst Existenz und Eindeutigkeit eines Interpolanten $q \in \mathcal{W}_k$ diskutieren.

Lemma 4.7 (Interpolationsaufgabe) *Seien Stützstellen $x_0 > x_1 > \ldots > x_k > 0$ und Stützwerte $q_0, \ldots, q_k \in \mathbb{R}$ gegeben. Es gibt genau ein $q \in \mathcal{W}_k$ mit*

$$q(x_i) = q_i \qquad \text{für alle } i \in \{0, \ldots, k\}.$$

Beweis. (vgl. [1, Lemma 4.33]) Wir betrachten die Abbildung

$$\Sigma : \mathbb{R}^{k+1} \to \mathbb{R}^{k+1}, \qquad (\beta_0, \ldots, \beta_k) \mapsto \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_i^{p+(j-1)\omega} \right)_{i=0}^{k},$$

die offenbar linear ist. Falls wir nachweisen können, dass sie auch injektiv ist, folgt aus Dimensionsgründen, dass sie auch surjektiv ist, also eindeutig invertierbar. Da Σ gerade die Koeffizienten $\beta_0, \ldots, \beta_k \in \mathbb{R}$ auf die Werte des durch sie definierten Polynoms

$$q : \mathbb{R} \to \beta_0 + \sum_{j=1}^{k} \beta_j x^{p+(j-1)\omega} \qquad \text{für alle } x \in \mathbb{R} \tag{4.15}$$

in den Interpolationspunkten $x_1, \ldots, x_{k+1}$ abbildet, würde das die eindeutige Lösbarkeit der Interpolationsaufgabe implizieren.

Zum Nachweis der Injektivität fixieren wir $\beta_0, \ldots, \beta_k \in \mathbb{R}$ so, dass $\Sigma(\beta_0, \ldots, \beta_k) = 0$ gilt. Wir führen das Polynom $q_* \in \mathcal{P}_{k-1}$ durch

$$q_*(x) = \sum_{j=1}^{k} \beta_j x^{j-1} \qquad \text{für alle } x \in \mathbb{R}$$

ein und stellen fest, dass

$$q(x) = \alpha_0 + x^p q_*(x^\omega) \qquad \text{für alle } x \in \mathbb{R}$$

mit dem in (4.15) definierten $q \in \mathcal{W}_k$ gilt. Daraus folgt

$$-\alpha_0 = x_i^p q_*(x_i^\omega) \qquad \text{für alle } i \in \{0, \ldots, k\}. \tag{4.16}$$

Wir definieren die Funktion

$$\varphi : \mathbb{R} \to \mathbb{R}, \qquad \zeta \mapsto -\beta_0 \zeta^{-p/\omega}$$

und erhalten

$$x_i^p \varphi(x_i^\omega) = -\beta_0 x_i^p x_i^{-p} = -\beta_0 = x_i^p q_*(x_i^\omega) \qquad \text{für alle } i \in \{0, \dots, k\},$$

die Funktion q_* interpoliert also φ in allen Punkten $x_0^\omega, \dots, x_k^\omega$.

Insbesondere interpoliert q_* die Funktion φ in den Punkten $x_1^\omega, \dots, x_k^\omega$, so dass wir die bekannte Formel für den Interpolationsfehler anwenden können: Es gibt ein $\zeta \in [x_0^\omega, \dots, x_k^\omega]$ so, dass sich der Interpolationsfehler im Punkt x_0^ω durch

$$q_*(x_0^\omega) - \varphi(x_0^\omega) = \frac{\varphi^{(k)}(\zeta)}{k!}(x_0^\omega - x_1^\omega) \cdots (x_0^\omega - x_k^\omega)$$

darstellen lässt. Wir wissen aber bereits, dass in x_0^ω die Funktion φ und das Polynom q_* ebenfalls übereinstimmen, also folgt

$$0 = \frac{\varphi^{(k)}(\zeta)}{k!}(x_0^\omega - x_1^\omega) \cdots (x_0^\omega - x_k^\omega).$$

Da die x_i paarweise verschieden und positiv sind, sind auch die x_i^ω paarweise verschieden und positiv, also muss

$$0 = \varphi^{(k)}(\zeta)$$

gelten. Wegen $\zeta > 0$ und $p > 0$ können wir aus der Definition von φ bereits $\beta_0 = 0$ folgern.

Durch Einsetzen in (4.16) und Dividieren durch x_i^p erhalten wir

$$q_*(x_i^\omega) = 0 \qquad \text{für alle } i \in \{0, \dots, k\}.$$

Da die x_i^ω paarweise verschieden sind und q_* nur ein Polynom der Ordnung $k-1$ ist, folgt per Identitätssatz $q_* = 0$, also auch $\beta_1 = \ldots = \beta_k = 0$.

Demzufolge enthält der Kern von Σ nur die Null, somit muss Σ injektiv und damit auch bijektiv sein, und die durch

$$(\beta_0, \dots, \beta_k) := \Sigma^{-1}(q_0, \dots, q_k)$$

gegebenen Koeffizienten definieren eindeutig das gesuchte Polynom q. ∎

Indem wir das Lemma auf die kanonischen Einheitsvektoren anwenden, erhalten wir die zu unserer Interpolationsaufgabe gehörenden Lagrange-Polynome: Für jedes $i \in \{0, \dots, k\}$ existiert genau ein $\mathcal{L}_i \in \mathcal{W}_k$ mit

$$\mathcal{L}_i(x_j) = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } j \in \{0, \dots, k\}. \tag{4.17}$$

Der Interpolant zu Werten $q_0, \ldots, q_k \in \mathbb{R}$ lässt sich damit als

$$q := \sum_{i=0}^{k} q_i \mathcal{L}_i \in \mathcal{W}_k$$

definieren. Wir können aber mit Hilfe der Lagrange-Polynome auch die uns eigentlich interessierende Interpolationsaufgabe im vektorwertigen Raum $\mathcal{W}_k$ lösen:

Lemma 4.8 (Vektorwertige Interpolationsaufgabe) *Seien wieder die Stützstellen $x_0 > x_1 > \ldots > x_k > 0$ und Stützwerte $\chi_0, \ldots, \chi_k \in V$ gegeben. Es gibt genau ein $\chi \in \mathcal{V}_k$ mit*

$$\chi(x_i) = \chi_i \qquad \text{für alle } i \in \{0, \ldots, k\}. \tag{4.18}$$

Beweis. Unter Verwendung der in (4.17) eingeführten Lagrange-Polynome definieren wir χ durch

$$\chi := \sum_{i=0}^{k} \chi_i \mathcal{L}_i.$$

Aus $\mathcal{L}_i \in \mathcal{W}_k$ folgt direkt $\chi \in \mathcal{V}_k$, und da es sich um Lagrange-Polynome handelt, erhalten wir auch

$$\chi(x_j) = \sum_{i=0}^{k} \chi_i \mathcal{L}_i(x_j) = \chi_j \qquad \text{für alle } i, j \in \{0, \ldots, k\}.$$

Zum Nachweis der Eindeutigkeit fixieren wir ein zweites Polynom $\chi' \in \mathcal{W}_k$, das ebenfalls die Eigenschaft (4.18) besitzt. Daraus folgt, dass die Differenz $\widehat{\chi} := \chi - \chi' \in \mathcal{V}_k$ in allen Punkten x_i gleich Null sein muss. Nach Definition des Raums $\mathcal{V}_k$ können wir Vektoren $\alpha_0, \ldots, \alpha_k \in V$ so wählen, dass

$$\widehat{\chi}(x) = \alpha_0 + \sum_{j=1}^{k} \alpha_j x^{p+\omega(j-1)} \qquad \text{für alle } x \in \mathbb{R}$$

gilt. Sei $i \in \{0, \ldots, k\}$. Wir betrachten das durch

$$\widehat{\chi}_i(x) := \langle \alpha_i, \widehat{\chi}(x) \rangle = \langle \alpha_i, \alpha_0 \rangle + \sum_{j=1}^{k} \langle \alpha_i, \alpha_j \rangle x^{p+\omega(j-1)} \qquad \text{für alle } x \in \mathbb{R}$$

definierte Polynom. Offenbar ist es ein Element von $\mathcal{W}_k$, das in allen Punkten x_j verschwindet, also impliziert die Eindeutigkeitsaussage von Lemma 4.7 bereits $\widehat{\chi}_i \equiv 0$ und damit insbesondere $0 = \langle \alpha_i, \alpha_i \rangle = \|\alpha_i\|^2$. Da i beliebig gewählt war, folgt $\widehat{\chi} = 0$ und damit auch $\chi = \chi'$. ∎

Zur Definition des Polynoms $\tilde{\chi}_{t,h} \in \mathcal{V}_k$ wenden wir das Lemma 4.8 auf die Punkte $x_0 = h_{n_0}, \ldots, x_k = h_{n_k}$ an und die Werte $\chi_0 = \eta_0, \ldots, \chi_k = \eta_k$ an und erhalten

$$\tilde{\chi}_{t,h} := \sum_{i=0}^{k} \eta_i \mathcal{L}_i \in \mathcal{V}_k.$$

Nun müssen wir überprüfen, dass das Extrapolationsverfahren auch die gewünschte Konsistenzordnung erreicht. Da das Polynom $\tilde{\chi}_{t,h}$ gemäß (4.13) durch Interpolation *gestörter* Werte von $\hat{\chi}_{t,h}$ entsteht, müssen wir analysieren, wie diese Störung sich auf den gesuchten Wert $\tilde{\chi}_{t,h}(0)$ auswirkt.

Wir verwenden zur Analyse die *Lebesgue-Zahl* unserer Interpolationsaufgabe:

Lemma 4.9 (Lebesgue-Zahl) *Es gibt eine Konstante $\Lambda(n_0, \dots, n_k)$, die nur von $n_0, \dots, n_k$, aber nicht von h abhängt, und die*

$$\left\| \sum_{i=0}^{k} \chi_i \mathcal{L}_i(0) \right\| \leq \Lambda(n_0, \dots, n_k) \max\{\|\chi_i\| \ : \ i \in \{0, \dots, k\}\} \quad \textit{für alle } \chi_0, \dots, \chi_k \in V$$

erfüllt.

Beweis. Wir bezeichnen die Lagrange-Polynome zu den Interpolationspunkten $\hat{x}_0 = 1/n_0, \dots, \hat{x}_k = 1/n_k$ mit $\widehat{\mathcal{L}}_0, \dots, \widehat{\mathcal{L}}_k$ und definieren die Lebesgue-Zahl durch

$$\Lambda(n_0, \dots, n_k) := \sum_{i=0}^{k} |\widehat{\mathcal{L}}_i(0)|.$$

gegeben ist. Sei $i \in \{0, \dots, k\}$ fixiert. Es gilt

$$\widehat{\mathcal{L}}_i(\hat{x}_j) = \mathcal{L}_i(x_j) = \mathcal{L}_i(h\hat{x}_j) \qquad \text{für alle } j \in \{0, \dots, k\},$$

also impliziert die Eindeutigkeitsaussage von Lemma 4.7 bereits

$$\widehat{\mathcal{L}}_i(x) = \mathcal{L}_i(hx) \qquad \text{für alle } x \in \mathbb{R},$$

und damit insbesondere $\widehat{\mathcal{L}}_i(0) = \mathcal{L}_i(0)$.

Seien nun $\chi_0, \dots, \chi_k \in V$ gegeben, und sei $C := \max\{\|\chi_i\| \ : \ i \in \{0, \dots, k\}\}$. Dann gilt

$$\left\| \sum_{i=0}^{k} \chi_i \mathcal{L}_i(0) \right\| \leq \sum_{i=0}^{k} \|\chi_i\| \ |\mathcal{L}_i(0)| \leq C \sum_{i=0}^{k} |\mathcal{L}_i(0)| = C \sum_{i=0}^{k} |\widehat{\mathcal{L}}_i(0)| = C\Lambda(n_0, \dots, n_k),$$

und der Beweis ist abgeschlossen. ∎

Wir wenden diese Abschätzung auf die Differenz $\tilde{\chi}_{t,h} - \widehat{\chi}_{t,h} \in \mathcal{V}_k$ an und finden

$$\|\tilde{\chi}_{t,h}(0) - \widehat{\chi}_{t,h}(0)\| \leq \Lambda(n_0, \dots, n_k) \max\{\|\tilde{\chi}_{t,h}(x_i) - \widehat{\chi}_{t,h}(x_i)\| \ : \ i \in \{0, \dots, k\}\},$$

Störungen in den Stützwerten werden also schlimmstenfalls um den Faktor $\Lambda(n_0, \dots, n_k)$ verstärkt.

Aus der Wahl von $\tilde{\chi}_{t,h}$ und $\widehat{\chi}_{t,h}$ (siehe (4.12) und (4.14)) folgt

$$\|\tilde{\chi}_{t,h}(x_i) - \widehat{\chi}_{t,h}(x_i)\| = \|\eta_{n_i}(t+h) + e_1(t)h_{n_i}^p + \ldots + e_k(t)h_{n_i}^{p+\omega(k-1)} - y(t+h)\|$$

$$\leq C_c h_{n_i}^{p+\omega k} \leq C_c h^{p+\omega k} \qquad \text{für alle } i \in \{0, \ldots, k\}.$$

Mit Hilfe der Lebesgue-Zahl erhalten wir

$$\|\tilde{\chi}_{t,h}(0) - y(t+h)\| \leq \Lambda(1/n_0, \ldots, 1/n_k) C_c h^{p+\omega k},$$

die gewünschte Konsistenzbedingung mit einer von h unabhängigen Konstanten.

Um nachzuweisen, dass das Extrapolationsverfahren wie gewünscht funktioniert, müssen wir die Existenz von asymptotischen Entwicklungen der Form (4.11) diskutieren. Da diese Abschätzung im Wesentlichen den Unterschied zwischen der Näherungslösung $\eta_n(t+h) + e_1(t)h_n^p + \ldots + e_k(t)h_n^{p+\omega(k+1)}$ und der Lösung $y(t+h)$ beschreibt, liegt es nahe, die Untersuchung auf Grundlage der Konsistenzordnung zu führen.

Wir beschränken uns auf den Fall $\omega = 1$ und $k = 2$ und müssen zeigen, dass

$$\|\eta_n(t+h) + e(t+h)h_n^p - y(t+h)\| \leq C_c h_n^{p+1} \tag{4.19}$$

für alle $h \in (0, h_y)$, $t \in [a, b-h]$ und $n \in \mathbb{N}$ gilt.

Unser Ziel ist es, die Funktion e zu konstruieren. Dazu verfolgen wir den Ansatz, ein zweites Einschrittverfahren mit der Inkrementfunktion $\widetilde{\Phi}$ so zu konstruieren, dass es die höhere Konsistenzordnung $p+1$ besitzt und

$$\widetilde{\eta}_n(t + ih_n) = \eta_n(t + ih_n) + e(t + ih_n)h_n^p \tag{4.20}$$

für seine Näherungslösung $\widetilde{\eta}_n$ gilt. Dann folgt nämlich aus dem Konvergenzsatz 3.2 bereits (4.19). Die Näherungslösung $\widetilde{\eta}_n(t+h)$ ist gegeben durch

$$\begin{aligned} \widetilde{\eta}_n(t) &:= \eta(t), \\ \widetilde{\eta}_n(t + (i+1)h_n) &:= \widetilde{\eta}_n(t + ih_n) + h_n \widetilde{\Phi}(t + ih_n, \widetilde{\eta}_n(t + ih_n), h_n), \end{aligned}$$

also nimmt die Bedingung (4.20) die Form

$$\begin{aligned} \eta_n(t + (i+1)h_n) + e(t + (i+1)h_n)h_n^p &= \widetilde{\eta}_n(t + ih_h) \\ &= \widetilde{\eta}_n(t + ih_n) + h_n \widetilde{\Phi}(t + ih_n, \widetilde{\eta}_n(t + ih_n), h_n) \\ &= \eta_n(t + ih_n) + e(t + ih_n)h_n^p + h_n \widetilde{\Phi}(t + ih_n, \eta_n(t + ih_n) + e(t + ih_n)h_n^p, h_n) \end{aligned}$$

an und wir gelangen mit

$$\eta_n(t + (i+1)h_n) = \eta_n(t + ih_n) + h_n \Phi(t + ih_n, \eta_n(t + ih_n, h_n)$$

zu dem Ansatz

$$\widetilde{\Phi}(t, x, h_n) := \Phi(t, x - e(t)h_n^p, h_n) + (e(t + (i+1)h_n) - e(t + ih_n))h_n^{p-1}.$$

Nun betrachten wir den lokalen Diskretisierungsfehler des neuen Näherungsverfahrens $\widetilde{\Phi}$, der durch

$$\widetilde{\tau}(t, y(t), h_n) = \frac{y(t + h_n) - y(t)}{h} - \widetilde{\Phi}(t, y(t), h_n)$$

$$= \frac{1}{h_n}\left(y(t+h_n) - y(t) - h_n\Phi(t, y(t) - e(t)h_n^p, h_n)\right.$$
$$\left.-(e(t+(i+1)h_n) - e(t+ih_n))h_n^p\right)$$

gegeben ist. Wir verwenden die Taylor-Entwicklung von Φ im zweiten Argument, um

$$\Phi(t, y(t) - e(t)h_n^p, h_n) = \Phi(t, y(t), h_n) - e(t)h_n^p\frac{\partial\Phi}{\partial x}(t, y(t), h_n)$$
$$+ \frac{e^2(t)h_n^{2p}}{2}\frac{\partial^2\Phi}{\partial x^2}(t, y(t) + \xi e(t)h_n^p, h_n)$$

mit einem Zwischenpunkt $\xi \in [0, 1]$ zu erhalten. Den zweiten Term dieser Entwicklung entwickeln wir im dritten Argument, um

$$\frac{\partial\Phi}{\partial x}(t, y(t), h_n) = \frac{\partial\Phi}{\partial x}(t, y(t), 0) + h_n\frac{\partial^2\Phi}{\partial x\partial h}(t, y(t), h_+)$$

mit einem Zwischenpunkt $h_+ \in [0, h_n]$ zu gewinnen. Da Φ mindestens von erster Ordnung konsistent ist, gilt $\Phi(t, x, 0) = f(t, x)$, also auch

$$\frac{\partial\Phi}{\partial x}(t, y(t), h_n) = \frac{\partial f}{\partial x}(t, y(t)) + h_n\frac{\partial^2\Phi}{\partial x\partial h}(t, y(t), h_+).$$

Schließlich wenden wir die Taylor-Entwicklung auch in der Form

$$e(t+(i+1)h_n) - e(t+ih_n) = h_n e'(t+ih_n) + \frac{h_n^2}{2}e''(t+(i+\zeta)h_n)$$

mit einem Zwischenpunkt $\eta \in [0, 1]$ an, um auch den letzten Term von $\widetilde{\tau}$ zu vereinfachen.

Wir sammeln die Restterme in

$$R(t, h_n, i) := \frac{e^2(t)h_n^{p-1}}{2}\frac{\partial^2\Phi}{\partial x^2}(t, y(t) + \xi e(t)h_n^p, h_n)$$
$$+ \frac{\partial^2\Phi}{\partial x\partial h}(t, y(t), h_+) + \frac{1}{2}e''(t+(i+\zeta)h_n)$$

und erhalten

$$\widetilde{\tau}(t, y(t), h_n) = \frac{1}{h_n}\left(y(t+h_n) - y(t) - h_n\Phi(t, y(t), h_n) + e(t)h_n^{p+1}\frac{\partial f}{\partial x}(t, y(t)) - h_n^{p+1}e'(t)\right)$$
$$+ h_n^{p+1}R(t, h_n, i)$$
$$= \tau(t, y(t), h_n) + h_n^p\left(e(t)\frac{\partial f}{\partial x}(t, y(t)) - e'(t)\right) + h_n^{p+1}R(t, h_n, i)$$

Jetzt müssen wir die Konsistenz des ursprünglichen Einschrittverfahrens einsetzen. Da es eine Konsistenzordnung von p besitzt, folgt aus der Taylor-Entwicklung (4.1) des lokalen Diskretisierungsfehlers bereits die Darstellung

$$\tau(t, y(t), h) = \frac{h^p}{p!}\left(\frac{y^{(p+1)}(t)}{p+1} - \frac{\partial^p\Phi}{\partial h^p}(t, x, 0)\right) + \frac{h^{p+1}}{(p+1)!}\left(\frac{y^{(p+2)}(t_+)}{p+2} - \frac{\partial^{p+1}\Phi}{\partial h^{p+1}}(t, x, h_+)\right)$$

mit $x = y(t)$, $h_+ \in [0, h]$ und $t_+ \in [t, t+h]$. Wir definieren

$$d(t) := \frac{y^{(p+1)}(t)}{(p+1)!} - \frac{\partial^p \Phi}{\partial h^p}(t, y(t), 0) \qquad \text{für alle } t \in [a, b]$$

und erhalten für die Konstante

$$C_1 := \max \left\{ \frac{1}{(p+1)!} \left\| \frac{y^{(p+2)}(t_+)}{p+2} - \frac{\partial^{p+1}\Phi}{\partial h^{p+1}}(t, x, h_+) \right\| \; : \; h \in (0, h_y), \right.$$
$$\left. t \in [a, b-h], h_+ \in [0, h] \right\}$$

die lokale Abschätzung

$$\|\tau(t, y(t), h) - d(t)h^p\| \leq C_1 h^{p+1} \qquad \text{für alle } h \in (0, h_y), t \in [a, b-h].$$

Einsetzen in $\widetilde{\tau}$ führt zu

$$\|\widetilde{\tau}(s, y(s), h_n)\| \leq h_n^p \left\| d(s) - e(s)\frac{\partial f}{\partial x}(s, y(s)) + e'(s) \right\| + h_n^{p+1}(C_1 + \|R(s, h_n, i)\|).$$

Wenn wir also $e(t)$ als Lösung der gewöhnlichen Differentialgleichung

$$e(t) = 0, \qquad e'(s) = e(s)\frac{\partial f}{\partial x}(s, y(s)) - d(s) \qquad s \in [t, t+h]$$

definieren, folgt unmittelbar

$$\|\widetilde{\tau}(s, y(s), h_n)\| \leq h_n^{p+1}(C_1 + \|R(s, h_n, i)\|),$$

und wir können wie üblich eine obere Schranke für $\|R(s, h_n, i)\|$ finden und so beweisen, dass $\widetilde{\Phi}$ die Konsistenzordnung $p+1$ besitzt.

Anwendung des Satzes 3.2 ergibt dann

$$\|\eta_n(t+h) - e(t+h)h_n^p - y(t+h)\| = \|\widetilde{\eta}_n(t+h) - y(t+h)\| \leq C h_n^{p+1}$$

für eine Konstante $C \in \mathbb{R}_{>0}$, und das ist die gesuchte Abschätzung.

Wir können den beschriebenen Prozess induktiv fortführen, um durch Addition weiterer Terme die Konsistenzordnung weiter zu erhöhen und damit Aussagen der Form (4.11) zu gewinnen. Voraussetzung ist immer, dass Φ und f hinreichend oft differenzierbar sind, wir können also auch mit einem Extrapolationsverfahren nur dann höhere Konsistenzordnungen erreichen, wenn die Lösung hinreichend glatt ist.

5 Schrittweitensteuerung

5.1 Einschrittverfahren variabler Schrittweite

Wie wir bereits in Bemerkung 3.7 gesehen haben, hängt die Konsistenzordnung mit dem Verhalten der Lösung zusammen. Falls die Lösung y zweimal stetig differenzierbar ist, folgt beispielsweise für das explizite Euler-Verfahren

$$\|\tau(t, y(t), h)\| = \|y'(t_+) - y'(t)\| = (t_+ - t)\|y''(t_{++})\|$$

mit Zwischenpunkten $t_+ \in [t, t+h]$ und $t_{++} \in [t, t_+]$. Daraus folgt

$$\|\tau(t, y(t), h)\| \leq h \max\{\|y''(s)\| \ : \ s \in [t, t+h]\}, \tag{5.1}$$

also eine lokale Konsistenzabschätzung.

Im Falle von Bemerkung 3.7 haben wir das Maximum auf der rechten Seite dieser Abschätzung durch das Maximum über das gesamte Intervall $[a, b]$ weiter abgeschätzt und dadurch eine global gleichmäßige Abschätzung für den Diskretisierungsfehler erhalten.

In der Praxis kann eine derartige gleichmäßige Abschätzung unattraktiv sein: falls die zweite Ableitung nur auf einem kleinen Teilstück von $[a, b]$ sehr große Werte annimmt, würden Teile der Lösung unnötig genau approximiert, und dadurch wäre der Rechenaufwand zu hoch.

Wesentlich attraktiver wäre es, die Schrittweite *adaptiv* zu wählen, sie also dort nur dort klein zu wählen, wo das Verhalten der Lösung es erfordert.

Zur näheren Untersuchung dieser Vorgehensweise fixieren wir *beliebige* Punkte

$$a = t_0 < t_1 < \ldots < t_n = b,$$

deren Abstände durch

$$h_i := t_{i+1} - t_i \qquad \text{für alle } i \in \{0, \ldots, n-1\}$$

gegeben sind. Insbesondere sind wir an dem Fall interessiert, dass diese Abstände nicht mehr unbedingt gleich groß sind, sondern sich dem Verhalten der Lösung anpassen.

Natürlich müssen wir dabei darauf achten, dass die Genauigkeit der Lösung weiterhin gesteuert werden kann. Zur Analyse des Fehlers gehen wir wie im Beweis des Satzes 3.2 vor, allerdings müssen wir den Beweis etwas verallgemeinern und eine etwas weniger scharfe Abschätzung in Kauf nehmen.

Der Beweis des Satzes basiert auf Lemma 3.1, das unverändert auch für den Fall nichtäquidistanter Punkte t_i gilt. Selbstverständlich könnte man auch hier darüber nachdenken, statt der globalen Lipschitz-Konstanten L lokale Lipschitz-Konstanten L_i zu verwenden, um eine unregelmäßige Fehlerfortpflanzun zu beschreiben, wir beschränken uns im Interesse der Einfachheit auf die globale Lipschitz-Bedingung (3.6).

Bei konstanter Schrittweite erhielten wir im Beweis von Satz 3.2 eine geometrische Summe, die sich direkt ausrechnen lässt. Bei variabler Schrittweite ist das nicht mehr der Fall, aber wir erhalten immerhin noch die folgende Aussage:

Satz 5.1 (Variable Schrittweite) *Sei Φ Lipschitz-stetig im zweiten Argument, es gelte also (3.6). Sei*

$$\tau_0 := \max\{\|\tau(t_i; y(t_i), h_i)\| \ : \ i \in \{0, \dots, n-1\}\} \tag{5.2}$$

der maximale lokale Diskretisierungsfehler. Dann gilt die Abschätzung

$$\|y(t) - \eta(t)\| \le \tau_0(t-a)e^{L(t-a)} \qquad \text{für alle } t \in [a,b]_h.$$

Beweis. Sei $t \in [a,b]_h$. Da der Fall $t = a$ trivial ist, beschränken wir uns auf $t = t_i$ mit $i \in \{1, \dots, n\}$. Wir verwenden zur Abkürzung wieder

$$y_j := y(t_j) \qquad \text{für alle } j \in \{0, \dots, n\}$$

und beweisen zunächst die Hilfsbehauptung

$$\|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| \le \tau_0(t-t_j)e^{L(t-t_{j+1})} \qquad \text{für alle } j \in \{0, \dots, i-1\} \tag{5.3}$$

durch (endliche, absteigende) Induktion über j.

Der Induktionsanfang $j = i-1$ ergibt sich aus

$$\begin{aligned} h_j\tau(t_j, y_j, h_j) &= y(t_j + h_j; t_j, y_j) - y_j - h_j\Phi(t_j, y_j, h_j) = y(t_{j+1}) - (y_j + h_j\Phi(t_j, y_j, h_j)) \\ &= y(t_{j+1}) - \eta(t_{j+1}; t_j, y_j) = y(t) - \eta(t; t_j, y_j) \end{aligned}$$

direkt, denn es gilt

$$\|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| = \|y(t) - \eta(t; t_j, y_j)\| = h_j\|\tau(t_j, y_j, h_j)\| \le (t-t_j)\tau_0.$$

Sei nun $j \in \{1, \dots, i-1\}$ so gewählt, dass (5.3) gilt. Durch Einschieben von $\eta(t; t_j, y_j)$ erhalten wir

$$\begin{aligned} \|y(t; t_{j-1}, y_{j-1}) - \eta(t; t_{j-1}, y_{j-1})\| &= \|y(t; t_j, y_j) - \eta(t; t_j, \eta(t_j; t_{j-1}, y_{j-1}))\| \\ &\le \|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| + \|\eta(t; t_j, y_j) - \eta(t; t_j, \eta(t_j; t_{j-1}, y_{j-1}))\| \\ &\le \|y(t; t_j, y_j) - \eta(t; t_j, y_j)\| + \|y_j - \eta(t_j; t_{j-1}, y_{j-1})\|e^{L(t-t_j)}. \end{aligned}$$

Dank der Definition des lokalen Diskretisierungsfehlers folgt

$$\|y_j - \eta(t_j; t_{j-1}, y_{j-1})\| = \|y(t_j) - (y(t_{j-1}) + h_{j-1}\Phi(t_{j-1}, y_{j-1}, h_{j-1}))\|$$

$$= h_{j-1}\|\tau(t_{j-1}, y_{j-1}, h_{j-1})\| \leq (t_j - t_{j-1})\tau_0,$$

und zusammen mit der Induktionsbehauptung erhalten wir

$$\begin{aligned}\|y(t; t_{j-1}, y_{j-1}) - \eta(t; t_{j-1}, y_{j-1})\| &\leq \tau_0(t - t_j)e^{L(t-t_{j+1})} + \tau_0(t_j - t_{j-1})e^{L(t-t_j)} \\ &\leq \tau_0(t - t_{j-1})e^{L(t-t_j)},\end{aligned}$$

also die erforderliche Abschätzung.

Indem wir (5.3) auf $j = 0$ anwenden, erhalten wir die gewünschte Aussage. ∎

5.2 Kombination zweier Verfahren unterschiedlicher Ordnung

Sei eine Genauigkeit $\epsilon \in \mathbb{R}_{>0}$ gegeben. Satz 5.1 besagt, dass es ausreicht, die lokalen Diskretisierungsfehler $\tau(t_i; y(t_i), y_i)$ unter die Schranke

$$\frac{\epsilon}{(b-a)e^{L(b-a)}}$$

zu drücken, um zu garantieren, dass die Näherungslösung in allen Punkten des diskreten Intervalls $[a, b]_h$ höchstens einen Fehler von ϵ aufweist.

Wie wir in (5.1) bereits gesehen haben, hängt der lokale Diskretisierungsfehler aber nur von dem lokalen Verhalten der Lösung ab und kann deshalb auch lokal gesteuert werden, indem man die Schrittweite h_i geeignet wählt. Das Ziel ist es also, die Schrittweiten so zu wählen, dass einerseits eine gewisse Genauigkeit erzielt wird und andererseits möglichst wenige der potentiell aufwendigen Funktionsauswertungen durchgeführt werden müssen.

Da wir den Fehler nie exakt kennen können, sind wir auf Abschätzungen angewiesen, die sich praktisch berechnen lassen. Eine gute Heuristik besteht darin, zwei Einschrittverfahren Φ_1 und Φ_2 unterschiedlicher Ordnung zu verwenden.

Wir bezeichnen die zu Φ_1 und Φ_2 gehörenden Näherungslösungen mit η_1 und η_2 und die entsprechenden lokalen Diskretisierungsfehler mit τ_1 und τ_2 und fordern, dass die beiden Verfahren von p-ter beziehungsweise $(p + 1)$-ter Ordnung sind, i.e., dass Konstanten $h_y, C_1, C_2 \in \mathbb{R}_{>0}$ mit

$$\|\tau_1(t, y(t), h)\| \leq C_1 h^p, \qquad \|\tau_2(t, y(t), h)\| \leq C_2 h^{p+1} \qquad \text{für alle } h \in (0, h_y)$$

existieren. Wir gehen davon aus, dass sich der Fehler τ_1 in Null in einer Taylor-Reihe der Ordnung $p + 1$ entwickeln lässt, dass also ein Vektor c_τ und eine Konstante $C_c \in \mathbb{R}_{>0}$ mit

$$\|\tau_1(t, y(t), h) - h^p c_\tau\| \leq C_c h^{p+1} \qquad \text{für alle } h \in (0, h_y) \tag{5.4}$$

existieren. Aufgrund dieser Annahme gilt

$$\begin{aligned}&\eta_1(t + h; t, y(t)) - \eta_2(t + h; t, y(t)) \\ &\qquad = \eta_1(t + h; t, y(t)) - y(t + h) + y(t + h) - \eta_2(t + h; t, y(t))\end{aligned}$$

$$
\begin{aligned}
&= -h\tau_1(t, y(t), h) + h\tau_2(t, y(t), h) \\
&= -hh^p c_\tau + hh^p c_\tau - h\tau_1(t, y(t), h) + h\tau_2(t, y(t), h) \\
&= -h^{p+1} c_\tau + h(h^p c_\tau - \tau_1(t, y(t), h)) + h\tau_2(t, y(t), h),
\end{aligned}
$$

also bringen wir $h^{p+1}c_\tau$ auf die linke Seite der Gleichung und erhalten

$$
\begin{aligned}
\left\| c_\tau - \frac{\eta_1(t+h; t, y(t)) - \eta_2(t+h; t, y(t))}{h^{p+1}} \right\| &= h \frac{\|\tau_1(t, y(t), h) - h^p c_\tau\| + \|\tau_2(t, y(t), h)\|}{h^{p+1}} \\
&\leq (C_c + C_2)h,
\end{aligned}
$$

wir können also zumindest den führenden Term von $\tau_1(t, y(t), h)$ approximativ bestimmen. Das bietet uns die Möglichkeit, die Schrittweite zu regeln: Wenn h hinreichend klein ist, ist

$$
c'_\tau := \frac{\eta_1(t+h; t, y(t)) - \eta_2(t+h; t, y(t))}{h^{p+1}}
$$

eine gute Approximation von c_τ, und es gilt

$$
\begin{aligned}
\|\tau_1(t, y(t), \tilde{h})\| &\leq \tilde{h}^p \|c_\tau\| + C_c \tilde{h}^{p+1} \\
&\leq \tilde{h}^p \|c'_\tau\| + \tilde{h}^p \|c_\tau - c'_\tau\| + C_c \tilde{h}^{p+1} \\
&\leq \tilde{h}^p \|c'_\tau\| + (2C_c + C_2)\tilde{h}^{p+1} \qquad \text{für alle } \tilde{h} \in (0, h_y).
\end{aligned}
$$

Um eine gute Heuristik zu erhalten, vernachlässigen wir den zweiten Term und müssen

$$
\tilde{h}^p \|c'_\tau\| \leq \tau_0
$$

sicherstellen. Das ist äquivalent zu

$$
\begin{aligned}
\tilde{h}^p &\leq \frac{\tau_0}{\|c'_\tau\|}, \\
\tilde{h}^p &\leq \frac{\tau_0 h^{p+1}}{\|\eta_1(t+h; t, y(t)) - \eta_2(t+h; t, y(t))\|}, \\
\tilde{h} &\leq h \left(\frac{\tau_0 h}{\|\eta_1(t+h; t, y(t)) - \eta_2(t+h; t, y(t))\|} \right)^{1/p}.
\end{aligned}
$$

Außer für $t = t_0$ ist $y(t)$ nicht bekannt, hier müssen wir also stattdessen unsere Näherungslösung $\eta(t)$ verwenden und erhalten die folgende *heuristische* Schrittweitensteuerung:

- Berechne $\eta_1(t+h; t, \eta(t))$.
- Berechne $\eta_2(t+h; t, \eta(t))$.
- Berechne $\alpha := \tau_0 h / \|\eta_1(t+h; t, \eta(t)) - \eta_2(t+h; t, \eta(t))\|$.
- Setze $\tilde{h} := h \sqrt[p]{\alpha}$.

- Berechne $\eta(t) = \eta_1(t + \tilde{h}; t, \eta(t))$.

Natürlich sind Verfeinerungen dieser Methode denkbar. Falls $\tilde{h}$ sich nicht sehr von h unterscheidet, kann man etwa $\tilde{h} := h$ setzen und den bereits berechneten Wert $\eta_1(t + h; t, \eta(t))$ für $\eta(t)$ verwenden.

Die Schrittweitensteuerung mit Hilfe zweier Verfahren unterschiedlicher Ordnung funktioniert im Allgemeinen nur dann, wenn beide Verfahren tatsächlich unterschiedliche Konsistenzordnungen aufweisen. Das Runge-Verfahren aus Beispiel 4.3 erreicht nur dann eine Konsistenzordnung von 2, falls f zweimal stetig differenzierbar ist. Sollte f nur einmal stetig differenzierbar sein, können wir nur eine Konsistenzordnung von 1 erwarten, so dass die Abschätzungen, auf denen unser Verfahren basiert, nicht mehr gelten.

Bemerkung 5.2 (Euler-Verfahren) *Die Annahme (5.4) hängt ebenfalls von der Differenzierbarkeit der Lösung y ab.*

Für das explizite Euler-Verfahren etwa haben wir

$$\begin{aligned} y(t+h) &= y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y^{(3)}(t_+) \\ &= y(t) + hf(t, y(t)) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y^{(3)}(t_+) \\ &= \eta(t+h; t, y(t)) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y^{(3)}(t_+) \end{aligned}$$

für einen Zwischenpunkt $t_+ \in [t, t+h]$, also folgt

$$\begin{aligned} \|\tau(t, y(t), h) - hy''(t)/2\| &= \frac{1}{h}\|y(t+h) - \eta(t+h; t, y(t)) - h^2 y''(t)/2\| \\ &= \frac{1}{h}\frac{h^3}{6}\|y^{(3)}(t_+)\| \leq \frac{h^2}{6}\max\{\|y^{(3)}(s)\| \ : \ s \in [t, t+h]\} \end{aligned}$$

so dass wir (5.4) einfach mit

$$c_\tau := y''(t)/2, \qquad C_c := \frac{1}{6}\max\{\|y^{(3)}(s)\| \ : \ s \in [t, t+h]\}$$

erfüllen können.

In die Herleitung der Formel für die Schrittweitensteuerung geht die Konstante C_c nur in der Form ein, dass h „klein genug“ sein muss, um den Term $(C_c + C_2)h$ ignorieren zu können. Diese Konstante muss also zur Durchführung des Verfahrens nicht unbedingt bekannt sein, stattdessen genügt es, sie geeignet abzuschätzen oder experimentell zu bestimmen.

Wir sind selbstverständlich daran interessiert, die genauere Näherung $\eta_2(t+h; t, \eta(t))$ mit möglichst geringem zusätzlichen Aufwand zu bestimmen. Einen guten Ansatz hierzu bieten *eingebettete Runge-Kutta-Verfahren*, bei denen zwei Näherungslösungen unterschiedlicher Genauigkeit mit Hilfe einer einzigen zusätzlichen Stufe berechnet werden.

Ein Beispiel dafür das das folgende Runge-Kutta-Fehlberg-Verfahren, das die Konsistenzordnungen 4 und 5 besitzt:

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	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Das erweiterte Butcher-Tableau ist wie folgt zu interpretieren: Die beiden letzten Zeilen geben Vektoren $b^\top$ und $\hat{b}^\top$ an, mit denen die beiden Näherungslösungen

$$\eta_1(t+h;t,x) = x + h\sum_{j=1}^{s-1} b_j k_j, \qquad \eta_2(t+h;t,x) = x + h\sum_{j=1}^{s} b_j k_j$$

berechnet werden können. Wichtig ist, dass beide Ergebnisse auf denselben Vektoren $k_1, \ldots, k_s$ basieren, die Approximation niedrigerer Konsistenzordnung allerdings den letzten Vektor k_s nicht verwendet. Durch diesen Ansatz lassen sich mit denselben Funktionsauswertungen zwei unterschiedlich gute Näherungslösungen gewinnen, indem lediglich unterschiedliche Linearkombinationen der Hilfsvektoren verwendet werden.

Man kann beweisen, dass ein explizites Runge-Kutta-Verfahren der Konsistenzordnung 5 mindestens 6 Stufen benötigt. Das hier vorgestellte Runge-Kutta-Fehlberg-Verfahren benötigt 7 Stufen, allerdings lässt sich durch den sogenannten *Fehlberg-Trick* die 7. Stufe quasi „kostenlos“ gewinnen: Wie man sieht, stimmen die letzte Spalte von A und die Zeile von $b^\top$ für die Konsistenzordnung 4 überein, es gilt also

$$k_7 = f\left(t+h, x+h\sum_{j=1}^{6} a_{ij}k_j\right) = f\left(t+h, x+h\sum_{j=1}^{6} b_j k_j\right) = f(t+h, \eta(t+h)).$$

Die letzte Auswertung von f in einem Schritt entspricht also der ersten Auswertung im folgenden Schritt, so dass für die gesamte Berechnung lediglich $6n+1$ Auswertungen statt der bei naivem Vorgehen erforderlichen $7n$ benötigt werden.

Dieser Trick funktioniert natürlich nur, falls die angepasste Schrittweite $\tilde{h}$ der für die Schrittweitensteuerung verwendeten Schrittweite h entspricht, es empfiehlt sich also, allzu häufige Änderungen der Schrittweite zu vermeiden.

5.3 Kombination durch Extrapolation

Die Steuerung der Schrittweite mit Hilfe zweier Verfahren unterschiedlicher Konsistenzordnung ist relativ elegant, falls solche Verfahren für den jeweiligen Anwendungsfall zur Verfügung stehen. Man kann allerdings auch eine Steuerung der Schrittweite erzielen, ohne auf ein separates zweites Verfahren zurückzugreifen, indem man stattdessen dieses zweite Verfahren durch Extrapolation aus dem Ausgangsverfahren konstruiert. Wir berechnen also eine Näherung

$$\eta_1(t+h;t,y(t)) = y(t) + h\Phi(t,y(t),h)$$

mit dem üblichen Einschrittverfahren und eine zweite Näherung

$$\begin{aligned}\eta_2(t+h/2;t,y(t)) &= y(t) + \frac{h}{2}\Phi(t,y(t),h/2),\\ \eta_2(t+h;t,y(t)) &= \eta_2(t+h/2;t,y(t)) + \frac{h}{2}\Phi(t,\eta_2(t+h/2;t,y(t)),h/2),\end{aligned}$$

die hoffentlich genauer als die erste sein wird. Wegen Satz 3.2 dürfen wir erwarten, dass für hinreichend kleine Schrittweiten h die Abschätzungen

$$\|y(t+h;t,y(t)) - \eta_1(t+h;t,y(t))\| \le Ch^{p+1}, \tag{5.5a}$$

$$\|y(t+h;t,y(t)) - \eta_2(t+h;t,y(t))\| \le Ch\left(\frac{h}{2}\right)^p \tag{5.5b}$$

gelten (wegen $e^{Lh}/L \approx h$). Wenn wir wieder davon ausgehen, dass sich der Fehler nach h entwickeln lässt, dass also die Ungleichungen

$$\begin{aligned}\|y(t+h;t,y(t)) - \eta_1(t+h;t,y(t)) - c_\tau h^{p+1}\| &\le C_c h^{p+2},\\ \|y(t+h;t,y(t)) - \eta_2(t+h;t,y(t)) - c_\tau h(h/2)^p\| &\le C_c\left(\frac{h}{2}\right)^{p+2}\end{aligned}$$

für einen geeigneten Vektor c_τ erfüllt sind, folgt

$$\begin{aligned}&\eta_2(t+h;t,y(t)) - \eta_1(t+h;t,y(t))\\ &\qquad = \eta_2(t+h;t,y(t)) - y(t+h) + y(t+h) - \eta_1(t+h;t,y(t))\\ &\qquad \approx -c_\tau h\left(\frac{h}{2}\right)^p + c_\tau h^{p+1} = c_\tau(1-2^{-p})h^{p+1}\end{aligned}$$

Präziser erhalten wir

$$\left\| c_\tau - \frac{\eta_2(t+h;t,y(t)) - \eta_1(t+h;t,y(t))}{(1-2^{-p})h^{p+1}} \right\| \le \frac{C_c}{1-2^{-p}}h,$$

für hinreichend kleines h ist also

$$c'_\tau := \frac{\eta_1(t+h;t,y(t)) - \eta_2(t+h;t,y(t))}{(1-2^{-p})h^{p+1}}$$

eine gute Approximation von c_τ, und wir können mit

$$h\tau(t, y(t), h) = y(t + h; t, y(t)) - \eta_1(t + h; t, y(t)) \approx c_\tau h^{p+1}$$

die Formel

$$\tilde{h} := \sqrt[p]{\frac{\tau_0}{\|c_\tau\|}} = h \left(\frac{(1 - 2^{-p})h}{\|\eta_1(t + h; t, y(t)) - \eta_2(t + h; t, y(t))\|} \right)^{1/p}$$

zur Bestimmung einer guten Schrittweite $\tilde{h}$ gewinnen.

Bemerkung 5.3 (Praktische Belange) *In der Praxis wird man mit Hilfe einer der beiden vorgestellten Regeln eine geeignete Schrittweite $\tilde{h}$ verwenden, um einen Näherungswert im Zeitpunkt $t + \tilde{h}$ zu bestimmen. Im nächsten Schritt bietet es sich an, die letzte Schrittweite $\tilde{h}$ als Ausgangswert für die Bestimmung der nächsten Schrittweite zu verwenden. Allerdings kann diese Strategie nur dann funktionieren, wenn $\tilde{h}$ nicht zu groß ist, denn dann würden unsere Formeln, die ja nur für „hinreichend kleine" Schrittweiten sinnvoll sind, zu unzuverlässige Ergebnisse ermitteln. Es empfiehlt sich also, eine obere Schranke für die Schrittweite vorzugeben.*

Falls die automatisch gewählte Schrittweite $\tilde{h}$ deutlich kleiner als h ist, müssen wir davon ausgehen, dass die Lösung y nicht sehr glatt ist. Unter diesen Bedingungen dürfte auch schon die Schrittweite h, die wir für die Heuristik verwenden, zu zu ungenauen Ergebnissen führen. In diesem Fall empfiehlt es sich, mit $h := 2\tilde{h}$ einen zweiten Versuch zur Schrittweitenanpassung zu unternehmen.

Dabei muss man zusätzlich darauf achten, dass h und $\tilde{h}$ nicht zu klein werden. Spätestens sobald man in den Bereich der Maschinengenauigkeit kommt, könnte das Näherungsverfahren sonst unendlich lange rechnen.

6 Steife Differentialgleichungen

6.1 Motivation des Begriffs

Bisher waren die von uns analysierten numerischen Verfahren überwiegend explizit, und wir haben gesehen, dass bei einer hinreichend kleinen Schrittweite diese Verfahren brauchbare Approximationen der Lösung eines Anfangswertproblems bestimmen.

Die Bedeutung von „hinreichend klein“ hängt dabei im Wesentlichen davon ab, wie groß die Lipschitz-Konstante der rechten Seite f ist: Je größer sie ist, desto „weniger glatt“ ist die Lösung, und desto kleiner müssen wir die Schrittweite wählen, desto höher wird also der Rechenaufwand.

Es gibt Situationen, in denen eine große Lipschitz-Konstante nicht unbedingt eine geringe Schrittweite erforderlich macht, weil der Term, der die Konstante in die Höhe treibt, nur geringen Einfluss auf die exakte Lösung des Problems hat. In diesen Situationen sind wir daran interessiert, Verfahren zu entwickeln, die diese Eigenschaft erben, also mit einer größeren Schrittweite trotzdem eine gute Approximation bestimmen können.

Als Beispiel befassen wir uns mit dem linearen Anfangswertproblem

$$y(0) = y_0, \qquad y'(t) = Ay(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

mit einem Startvektor $y_0 \in \mathbb{R}^2$ und der Matrix

$$A := \frac{1}{2}\begin{pmatrix} \alpha+\beta & \alpha-\beta \\ \alpha-\beta & \alpha+\beta \end{pmatrix}$$

mit Koeffizienten $\alpha, \beta \in \mathbb{R}$.

Zunächst bestimmen wir die bestmögliche Lipschitz-Konstante für die korrespondierende rechte Seite $f(t,x) = Ax$. Da A symmetrisch ist, bietet es sich an, die Eigenwerte zu bestimmen: Mit

$$Q := \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

erhalten wir $Q^\top Q = I$ und

$$Q^\top A Q = \frac{1}{4}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} \alpha+\beta & \alpha-\beta \\ \alpha-\beta & \alpha+\beta \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix},$$

also folgt sofort $\|A\|_2 = L := \max\{|\alpha|, |\beta|\}$ und damit

$$\|f(t,x) - f(t,z)\|_2 = \|A(x-z)\|_2 \leq \|A\|_2 \|x-z\|_2 = L\|x-z\|_2,$$

und da wir $x - z$ als Eigenvektor zum betragsgrößeren der beiden Eigenwerte wählen können, folgt die Optimalität von L.

Zur Analyse des Verhaltens des Lösung y führen wir die Hilfsfunktion $\hat{y}$ mit

$$\hat{y}(t) := Q^\top y(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

ein. Wegen $y(t) = Q\hat{y}(t)$ muss

$$\hat{y}'(t) = Q^\top y'(t) = Q^\top A y(t) = Q^\top A Q \hat{y}(t) = \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix} \hat{y}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

gelten, die Hilfsfunktion löst also das Anfangswertproblem

$$\hat{y}(0) = \hat{y}_0 := Q^\top y_0, \qquad \hat{y}'(t) = \begin{pmatrix} \alpha & \\ & \beta \end{pmatrix} \hat{y}(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}.$$

Nun sind die beiden Komponenten von $\hat{y}$ entkoppelt und die Lösung ist explizit durch

$$\hat{y}_1(t) = \hat{y}_{0,1} e^{\alpha t}, \qquad \hat{y}_2(t) = \hat{y}_{0,2} e^{\beta t} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

gegeben. Aus $y(t) = Q\hat{y}(t)$ folgt direkt

$$y(t) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \hat{y}_{0,1} e^{\alpha t} \\ \hat{y}_{0,2} e^{\beta t} \end{pmatrix} = a e^{\alpha t} + b e^{\beta t} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

mit den Hilfsvektoren

$$a := \frac{\hat{y}_{0,1}}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad b := \frac{\hat{y}_{0,2}}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Untersuchen wir nun das Verhalten eines Näherungsverfahrens. Der Einfachheit halber beschränken wir uns auf das explizite Euler-Verfahren, das bei fester Schrittweite h die Näherungslösung $\eta(t)$ mit

$$\eta(t+h) = \eta(t) + hA\eta(t) = (1 + hA)\eta(t) \qquad \text{für alle } t \in [0, \infty)_h$$

berechnet. Wir führen wieder die (diesmal diskrete) Hilfsfunktion $\hat{\eta}$ mit

$$\hat{\eta}(t) := Q^\top \eta(t) \qquad \text{für alle } t \in [0, \infty)_h$$

ein und erhalten

$$\begin{aligned} \hat{\eta}(t+h) = Q^\top \eta(t+h) = Q^\top (I + hA)\eta(t) = Q^\top (I + hA) Q \hat{\eta}(t) \\ = \begin{pmatrix} 1 + h\alpha & \\ & 1 + h\beta \end{pmatrix} \hat{\eta}(t) \qquad \text{für alle } t \in [0, \infty)_h. \end{aligned}$$

Per Induktion folgt

$$\hat{\eta}(t) = \begin{pmatrix} (1 + h\alpha)^{t/h} \hat{y}_{0,1} \\ (1 + h\beta)^{t/h} \hat{y}_{0,2} \end{pmatrix} \qquad \text{für alle } t \in [0, \infty)_h \tag{6.1}$$

und somit wegen $\eta(t) = Q\hat{\eta}(t)$ auch

$$\eta(t) = a(1+h\alpha)^{t/h} + b(1+h\beta)^{t/h} \qquad \text{für alle } t \in [0,\infty)_h.$$

Wir vergleichen die exakte und die genäherte Lösung:

$$\begin{array}{rcll} y(t) & = & ae^{\alpha t} + & be^{\beta t} \\ \eta(t) & = & a(1+h\alpha)^{t/h} + & b(1+h\beta)^{t/h} \end{array}$$

Falls $\alpha, \beta > 0$ gilt, verhalten sich die beiden Funktionen ähnlich: Für $t \to \infty$ streben sie exponentiell gegen unendlich.

Anders sieht es im Fall $\alpha, \beta < 0$ aus: Jetzt konvergiert $y(t)$ gegen Null, wenn wir t gegen unendlich gehen lassen, aber für $\eta(t)$ gilt das nur, wenn die Schrittweite h klein genug ist, wenn also

$$|1+h\alpha| < 1, \qquad |1+h\beta| < 1$$

gilt. Wegen $\alpha, \beta < 0$ und $L = \max\{|\alpha|, |\beta|\}$ ist das gerade für $h < 2/L$ sichergestellt, die Schrittweite wird also tatsächlich durch den betragsgrößten Eigenwert bestimmt.

Falls $\beta \ll \alpha < 0$ gilt, ist für das langfristige Verhalten die Lösung nur der von α abhängige Term relevant, weil $e^{\beta t}$ sehr schnell gegen null streben wird. Trotzdem müssen wir in unserem Näherungsverfahren die Schrittweite so wählen, dass $h < 2/L$ mit $L = |\beta|$ gilt, wir müssen also etwas approximieren, das uns eigentlich gar nicht interessiert.

Ein Anfangswertproblem, bei dem zwar die Lösung für lange Zeiträume gegen null strebt, bei dem aber verschiedene Anteile der Lösung das mit sehr unterschiedlichen Geschwindigkeiten tun, bezeichnet man als *steif*: Bis zu einer gewissen Grenzschrittweite (in unserem Beispiel $2/L$) berechnet das Näherungsverfahren unbrauchbare Lösungen, sobald diese Schrittweite unterschritten wird, funktioniert plötzlich alles.

Den Begriff „steife Differentialgleichung“ kann man dadurch motivieren, dass die Gleichung unseren numerischen Lösungsversuchen „Widerstand“ entgegensetzt, bis wir genug „Kraft“ (also Rechenaufwand) investiert haben, um den „Widerstand zu brechen“.

6.2 Einsatz impliziter Verfahren

Da steife Anfangswertprobleme in vielen Anwendungen auftreten, stellt sich die Frage, ob man Verfahren finden kann, die nicht auf die sehr restriktive Bedingung $h < 2/L$ angewiesen sind.

In dieser Hinsicht sehr erfolgreich sind implizite Verfahren. Als Beispiel stellen wir dem expliziten Euler-Verfahren das implizite Euler-Verfahren gegenüber, das in unserem Beispiel die Form

$$\eta(t+h) = \eta(t) + hA\eta(t+h) \qquad \text{für alle } t \in [0,\infty)_h$$

annimmt. Da unsere Gleichung linear ist, können wir $\eta(t+h)$ direkt berechnen, indem wir ein lineares Gleichungssystem lösen: Es gilt

$$(I - hA)\eta(t+h) = \eta(t) \qquad \text{für alle } t \in [0,\infty)_h.$$

Indem wir wieder zu $\hat{\eta}(t) = Q^\top \eta(t)$ wechseln, erhalten wir

$$(I - hA)Q\hat{\eta}(t+h) = Q\hat{\eta}(t), \qquad Q^\top(I - hA)Q\hat{\eta}(t+h) = \hat{\eta}(t),$$
$$\begin{pmatrix} 1-h\alpha & \\ & 1-h\beta \end{pmatrix} \hat{\eta}(t+h) = \hat{\eta}(t),$$

so dass wir die Darstellung

$$\hat{\eta}(t+h) = \begin{pmatrix} \frac{1}{1-h\alpha} & \\ & \frac{1}{1-h\beta} \end{pmatrix} \hat{\eta}(t)$$

bewiesen haben. Per Induktion folgt

$$\hat{\eta}(t) = \begin{pmatrix} (1-h\alpha)^{-t/h}\hat{y}_{0,1} \\ (1-h\beta)^{-t/h}\hat{y}_{0,2} \end{pmatrix} \qquad \text{für alle } t \in [0,\infty)_h, \tag{6.2}$$

und per Rücktransformation erhalten wir schließlich

$$\eta(t) = a(1-h\alpha)^{-t/h} + b(1-h\beta)^{-t/h} \qquad \text{für alle } t \in [0,\infty)_h.$$

Diese Näherungslösung besitzt andere Eigenschaften als die, die wir im Falle des expliziten Verfahrens erhalten haben: Sie könnte nur dann divergieren, wenn $|1-h\alpha| < 1$ oder $|1-h\beta| < 1$ gilt, aber dank $\alpha, \beta < 0$ ist das ausgeschlossen.

Das implizite Euler-Verfahren wird also eine Lösung berechnen, die für *beliebige* Schrittweiten abklingt. Falls $\beta \ll \alpha < 0$ gilt, ist $1 - h\beta \gg 1 - h\alpha > 0$, der von β abhängige Anteil der Lösung wird also auch wesentlich schneller als der von α abhängige abklingen, die numerisch bestimmte Näherungslösung verhält sich also zumindest in dieser Hinsicht wie die exakte Lösung.

Insbesondere genügt es in dieser Situation, die Schrittweite h so zu wählen, dass $(1 - h\alpha)^{-1/h} \approx e^\alpha$ gilt, dass sie also für die entscheidende Komponente ausreicht, während wir die schnell abklingende Komponente bei der Wahl der Schrittweite nicht mehr zu berücksichtigen brauchen.

6.3 Stabilitätsgebiete

Zur näheren Analyse dieses Phänomens untersuchen wir wieder die Stabilitätsfunktion: Wie schon bei der Untersuchung der maximalen Konsistenzordnung von expliziten Runge-Kutta-Verfahren beschränken wir uns auf das einfache Anfangswertproblem (4.5), das durch

$$y_\lambda(0) = 1, \qquad y'_\lambda(t) = \lambda y_\lambda(t) \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

für ein $\lambda \in \mathbb{C}$ gegeben ist.

Definition 6.1 (Stabilitätsfunktion) *Sei ein Näherungsverfahren für das Anfangswertproblem (4.5) gegeben, und sei η_λ die von ihm berechnete Näherungslösung. Falls eine Funktion $g : \mathbb{C} \to \mathbb{C}$ existiert, die*

$$\eta_\lambda(t+h) = g(\lambda h)\eta_\lambda(t) \qquad \text{für alle } h \in \mathbb{R}_{>0}, \lambda \in \mathbb{C}, t \in [0,\infty)_h \tag{6.3}$$

erfüllt, bezeichnen wir sie als Stabilitätsfunktion *des Verfahrens.*

Wie man an (6.1) leicht ablesen kann, besitzt das explizite Euler-Verfahren die Stabilitätsfunktion

$$g_{\mathrm{ex}}(z) = 1 + z, \qquad \text{für alle } z \in \mathbb{C},$$

während wir aus (6.2) folgern können, dass das implizite Euler-Verfahren die Stabilitätsfunktion

$$g_{\mathrm{im}}(z) = \frac{1}{1-z} \qquad \text{für alle } z \in \mathbb{C}$$

besitzt. Wir haben bereits in Lemma 4.5 gesehen, dass die Konsistenzordnung damit zusammenhängt, wie gut die Stabilitätsfunktion die Exponentialfunktion in $z = 0$ approximiert. Man kann die Stabilitätsfunktion aber auch verwenden, um zu charakterisieren, für welche Schrittweiten die Näherungslösung abklingen wird: Durch Induktion folgt aus (6.3) direkt

$$\eta_\lambda(t) = g(\lambda h)^{t/h} y_0 \qquad \text{für alle } t \in [0,\infty)_h,$$

wir können also nur dann ein Abklingen erwarten, wenn $|g(\lambda h)| < 1$ gilt.

Definition 6.2 (Stabilitätsgebiet) *Die Menge*

$$S_g := \{z \in \mathbb{C} \ : \ |g(z)| < 1\}$$

heißt Stabilitätsgebiet *zu der Stabilitätsfunktion g.*

Das Näherungsverfahren wird also zu einem sinnvollen asymptotischen Verhalten der Lösung führen, wenn wir $\lambda h \in S_g$ sicherstellen können.

Für das explizite Euler-Verfahren erhalten wir

$$S_{g_{\mathrm{ex}}} = \{|1+z| < 1 \ : \ z \in \mathbb{C}\} = K(-1,1),$$

das Stabilitätsgebiet ist also eine offene Kreisscheibe um -1, und in unserem Fall ist -2 der Punkt, an dem die reelle Achse in das Stabilitätsgebiet eintritt, wir müssen also $-2 < h\lambda < 0$ sicherstellen. Das entspricht dem Kriterium, dass wir für unser Modellproblem bewiesen haben.

Für das implizite Euler-Verfahren finden wir

$$S_{g_{\mathrm{im}}} = \{1/|1-z| < 1 \ : \ z \in \mathbb{C}\} = \{|1-z| > 1 \ : \ z \in \mathbb{C}\} = \mathbb{C} \setminus \overline{K(1,1)},$$

das Stabilitätsgebiet ist also die gesamte komplexe Ebene mit Ausnahme einer abgeschlossenen Kreisscheibe um 1. In unserem Modellproblem ist $h\lambda$ für alle Schrittweiten h negativ, also immer im Stabilitätsgebiet enthalten. Diese Eigenschaft ist natürlich besonders nützlich.

Definition 6.3 (A-Stabilität) *Ein Näherungsverfahren mit*

$$\{z \in \mathbb{C} \ : \ \operatorname{Re} z < 0\} \subseteq S_g$$

heißt A-stabil.

Bei einem A-stabilen Verfahren dürfen wir also erwarten, dass es sich besonders gut für steife Anfangswertprobleme eignet. Offensichtlich ist das implizite Euler-Verfahren A-stabil, während das explizite Euler-Verfahren es nicht ist.

Wir haben bereits gesehen, dass bei expliziten Runge-Kutta-Verfahren die Stabilitätsfunktion g ein Polynom ist, und da für alle nicht-konstanten Polynome

$$\lim_{z \to -\infty} |g(z)| = \infty$$

gilt, folgt sofort, dass derartige Verfahren niemals A-stabil sein können.

Eine Chance auf A-Stabilität haben wir also nur dann, wenn wir Verfahren mit nichtpolynomialer Stabilitätsfunktion untersuchen. Im Falle des impliziten Euler-Verfahrens beispielsweise ist die Stabilitätsfunktion rational.

Wir untersuchen allgemeine Runge-Kutta-Verfahren, die durch ein Gleichungssystem der Form

$$k_i = f\left(t + c_i h, x + h \sum_{j=1}^{s} a_{ij} k_j\right) \qquad \text{für alle } i \in \{1, \dots, s\}$$

gegeben sind. Für unser Modellproblem (4.5) erhalten wir daraus

$$k_i = \lambda x + \lambda h \sum_{j=1}^{s} a_{ij} k_j,$$

$$k_i - \lambda h \sum_{j=1}^{s} a_{ij} k_j = \lambda x \qquad \text{für alle } i \in \{1, \dots, s\},$$

und wenn wir die Zwischenergebnisse k_i in einem Vektor $k \in \mathbb{R}^s$ zusammenfassen und den konstanten Vektor $e := (1)_{i=1}^{s}$ einführen, ergibt sich

$$(I - \lambda h A)k = \lambda e x, \qquad k = (I - \lambda h A)^{-1} \lambda e x,$$

sofern die Matrix invertierbar (und damit das Verfahren überhaupt durchführbar) ist. Die Inkrementfunktion ist durch

$$\Phi(t, x, h) = \sum_{i=1}^{s} b_i k_i = \langle b, k \rangle_2 = b^\top (I - \lambda h A)^{-1} \lambda e x$$

gegeben, die nächste Iterierte durch

$$\eta(t + h) = \eta(t) + h\Phi(t, \eta(t), h) = \eta(t) + b^\top (I - \lambda h A)^{-1} \lambda e h \eta(t)$$

$$= (1 + b^\top (I - \lambda h A)^{-1} e \lambda h)\eta(t),$$

also muss die Stabilitätsfunktion gerade

$$g(z) = 1 + b^\top (I - zA)^{-1} ez \qquad \text{für alle } z \in \mathbb{C} \tag{6.4}$$

sein. Diese Funktion ist rational, und ihre Singularitäten sind gerade die Kehrwerte der Eigenwerte von A.

Im Falle eines expliziten Verfahrens ist $I - zA$ eine untere Dreiecksmatrix mit konstanten Diagonaleinträgen, also immer invertierbar. Durch Vorwärtseinsetzen können wir direkt das Resultat aus Lemma 4.4 gewinnen.

Interessanter ist natürlich die Anwendung auf implizite Verfahren. Als Beispiel untersuchen wir die implizite Trapezregel, deren Butcher-Schema

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

ist. Einsetzen in (6.4) führt zu

$$\begin{aligned} g_{\mathrm{tr}}(z) &= 1 + \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -z/2 & 1 - z/2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} z \\ &= 1 + \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{z/2}{1-z/2} & \frac{1}{1-z/2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} z \\ &= 1 + \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ \frac{1+z/2}{1-z/2} \end{pmatrix} z = 1 + \frac{1/2 - z/4 + 1/2 + z/4}{1 - z/2} z \\ &= 1 + \frac{z}{1 - z/2} = \frac{1 + z/2}{1 - z/2}. \end{aligned}$$

Um das Stabilitätsgebiet zu bestimmen, müssen wir diejenigen $z \in \mathbb{C}$ finden, für die $|g_{\mathrm{tr}}(z)| < 1$ gilt. Wir wählen ein $z \in \mathbb{C}$ und stellen es durch seinen Realteil $z_r \in \mathbb{R}$ und seinen Imaginärteil $z_i \in \mathbb{R}$ dar, also durch $z = z_r + i z_i$. Es gilt

$$\begin{aligned} |g_{\mathrm{tr}}(z)| < 1 &\iff |1 + z/2| < |1 - z/2| \iff |2 + z| < |2 - z| \\ &\iff |2 + z|^2 < |2 - z|^2 \iff (2 + z_r)^2 + z_i^2 < (2 - z_r)^2 + z_i^2 \\ &\iff 4 + 4z_r + z_r^2 < 4 - 4z_r + z_r^2 \iff 4z_r < -4z_r \iff 8z_r < 0, \end{aligned}$$

also ist $|g_{\mathrm{tr}}(z)| < 1$ äquivalent zu $z_r < 0$ und das Stabilitätsgebiet ist gegeben durch

$$S_{g_{\mathrm{tr}}} = \{z \in \mathbb{C} \ : \ \operatorname{Re} z < 0\}.$$

Wir sehen, dass das Stabilitätsgebiet der impliziten Trapezregel deutlich kleiner als das des impliziten Euler-Verfahrens ist, dass es aber immer noch die linke komplexe Halbebene enthält, so dass auch die implizite Trapezregel A-stabil ist.

7 Mehrschrittverfahren

Wir haben gesehen, dass bei Einschrittverfahren eine hohe Konsistenzordnung sehr erstrebenswert ist: Während eine Halbierung der Schrittweite bei einem Verfahren erster Ordnung nur zu einer Halbierung des Fehlers führt, bewirkt sie bei einem Verfahren vierter Ordnung schon eine Reduktion um einen Faktor von 1/16. Die mit der Halbierung der Schrittweite in Kauf genommene Verdoppelung des Rechenaufwands lohnt sich also bei einer höheren Ordnung wesentlich mehr als bei einer niedrigen.

Bei Einschrittverfahren gibt es verschiedene Ansätze, um eine hohe Konsistenzordnung zu erzielen: Explizite Runge-Kutta-Verfahren verwenden Hilfswerte in Zwischenpunkten, ihre Konstruktion ist für Ordnungen über 6 sehr aufwendig. Implizite Runge-Kutta-Verfahren lassen sich wesentlich einfacher konstruieren und erreichen auch höhere Ordnungen, das erforderliche Auflösen eines nichtlinearen Gleichungssystems in jedem Schritt führt allerdings zu einem hohen Rechenaufwand. Extrapolationsverfahren können im Prinzip beliebig hohe Ordnungen erreichen, sofern die Lösung hinreichend glatt ist, benötigen aber die aufwendige Berechnung einer Anzahl von Hilfslösungen.

Ein Schritt eines expliziten Runge-Kutta-Verfahrens der Ordnung s erfordert mindestens s Auswertungen der rechten Seite f (siehe Lemma 4.5), bei einem Extrapolationsverfahren muss man sogar mit $s(s+1)/2$ Auswertungen rechnen, für die schnellere Konvergenz ist also ein unter Umständen hoher Preis zu zahlen.

In diesem Kapitel geht es um eine Klasse von Approximationsverfahren, die eine Konsistenzordnung von s mit lediglich *einer* einzigen Auswertung von f erzielen können. Die Idee besteht darin, Näherungslösungen in mehreren Zeitpunkten zu kombinieren, um eine Näherung für einen weiteren Zeitpunkt zu gewinnen.

7.1 Adams-Bashforth-Verfahren

Wir untersuchen als einführendes Beispiel eine Klasse von expliziten Mehrschrittverfahren, die sich besonders einfach mit Hilfe der Integraldarstellung des Anfangswertproblems motivieren lässt.

Herleitung

Wie wir in Lemma 2.2 bereits gesehen haben, ist eine Funktion y genau dann die Lösung des Anfangswertproblems (2.1), wenn sie die Integralgleichung

$$y(t) = y_0 + \int_a^t f(s, y(s))\,ds \qquad \text{für alle } t \in [a, b]$$

erfüllt. Wenn wir das Zeitintervall $[a, b]$ äquidistant zerlegen, also

$$t_i = a + ih, \qquad h = \frac{b-a}{n} \qquad \text{für } n \in \mathbb{N}, i \in \{0, \ldots, n\}$$

setzen, erhalten wir die Darstellung

$$\begin{aligned} y(t_{m+1}) &= y_0 + \int_a^{t_m} f(s, y(s))\, ds + \int_{t_m}^{t_{m+1}} f(s, y(s))\, ds \\ &= y(t_m) + \int_{t_m}^{t_{m+1}} f(s, y(s))\, ds \qquad \text{für } m \in \{0, \ldots, n-1\}. \end{aligned}$$

Wir approximieren das Integral mit einer Newton-Côtes-artigen Quadraturformel, die die $k+1$ Quadraturpunkte $t_{m-k}, \ldots, t_m$ verwendet. Dazu definieren wir die Lagrange-Polynome

$$\mathcal{L}_{m,i}(s) := \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{s - t_{m-j}}{t_{m-i} - t_{m-j}} \qquad \text{für alle } m \in \{k, \ldots, n-1\}, i \in \{0, \ldots, k\},$$

ersetzen den Integranden

$$g(s) := f(s, y(s))$$

durch seinen Lagrange-Interpolanten

$$\tilde{g}_{m,k}(s) := \sum_{i=0}^{k} f(t_{m-i}, y(t_{m-1}))\mathcal{L}_{m,i}(s) \qquad \text{für alle } s \in \mathbb{R},$$

und integrieren den Interpolanten, um schließlich

$$y(t_{m+1}) \approx \tilde{y}(t_{m+1}) := y(t_m) + h \sum_{i=0}^{k} w_{k,i} f(t_{m-i}, y(t_{m-i})) \qquad \text{für } m \in \{k, \ldots, n-1\} \tag{7.1}$$

mit den durch

$$\begin{aligned} w_{k,i} &:= \frac{1}{h} \int_{t_m}^{t_{m+1}} \mathcal{L}_{m,i}(s)\, ds = \frac{1}{h} \int_{t_m}^{t_{m+1}} \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{s - t_{m-j}}{t_{m-i} - t_{m-j}}\, ds \\ &= \int_0^1 \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{a + h(m+s) - (a + h(m-j))}{(a + h(m-i)) - (a + h(m-j))}\, ds \\ &= \int_0^1 \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{h(s+j)}{h(j-i)}\, ds = \int_0^1 \prod_{\substack{j=0 \\ j \neq i}}^{k} \frac{s+j}{j-i}\, ds \qquad \text{für } i \in \{0, \ldots, k\} \end{aligned}$$

gegebenen Quadraturgewichten zu erhalten.

Fehlerabschätzung

Die übliche Fehlerabschätzung für die Lagrange-Interpolation besagt, dass es zu jedem $s \in [t_{m+1}, t_{m-k}]$ ein $\xi \in [t_{m+1}, t_{m-k}]$ so gibt, dass

$$g(s) - \tilde{g}_{m,k}(s) = \frac{\omega(s)}{(k+1)!} g^{(k+1)}(\xi)$$

gilt, wobei das Stützstellenpolynom zu den Punkten $t_m, \ldots, t_{m-k}$ durch

$$\omega(s) = (s - t_m) \ldots (s - t_{m-k}) \qquad \text{für alle } s \in [t_{m-k}, t_{m+1}]$$

gegeben ist. Es gilt

$$\begin{aligned} |s - t_{m-i}| &= |s - t_m| + |t_m - t_{m-i}| \\ &\leq h + hi = (i+1)h \qquad \text{für alle } s \in [t_m, t_{m+1}], i \in \{0, \ldots, k\}, \end{aligned}$$

also folgt

$$|\omega(s)| \leq h(2h) \ldots (k+1)h = h^{k+1}(k+1)! \qquad \text{für alle } s \in [t_m, t_{m+1}],$$

und wir haben

$$\|g(s) - \tilde{g}_{m,k}(s)\| \leq h^{k+1} \sup\{\|g^{(k+1)}(\xi)\| \ : \ \xi \in [t_{m+1}, t_{m-k}]\} \quad \text{für alle } s \in [t_m, t_{m+1}] \tag{7.2}$$

bewiesen. Indem wir die Konstante

$$C_k := \sup\{\|g^{(k+1)}(\xi)\| \ : \ \xi \in [a, b]\}$$

einführen erhalten wir die kompaktere Form

$$\|g(s) - \tilde{g}_{m,k}(s)\| \leq C_k h^{k+1} \qquad \text{für alle } s \in [t_m, t_{m+1}].$$

Durch Integration dieser Fehlerabschätzung erhalten wir direkt

$$\|y(t_{m+1}) - \tilde{y}(t_{m+1})\| \leq C_k h^{k+2} \qquad \text{für alle } m \in \{k, \ldots, n-1\},$$

also eine Abschätzung, die uns eine (geeignet verallgemeinerte) Konsistenzordnung von $k+1$ vermuten lässt.

Algorithmus

Indem wir auf der rechten Seite der Formel (7.1) die exakten Funktionswerte durch ihre Näherungen ersetzen, erhalten wir das folgende Näherungsverfahren:

Berechne Näherungen $\eta_0, \ldots, \eta_k$ von $y_0, \ldots, y_k$
for $i \in \{0, \ldots, k\}$ **do** $f_i \leftarrow f(t_i, \eta_i)$
for $m := k$ **to** $n - 1$ **do begin**
    $\eta_{m+1} \leftarrow \eta_m + h \sum_{i=0}^{k} w_{k,i} f_{m-i}$
    $f_{m+1} \leftarrow f(t_{m+1}, \eta_{m+1})$
end

Wir können sehen, dass für einen Schritt des Verfahrens tatsächlich nur eine Auswertung von f erforderlich ist. Wir haben also ein Verfahren erhalten, das mit sehr wenigen Auswertungen der rechten Seite eine Konsistenzordnung von k verspricht.

Bei der Implementierung des Verfahrens ist es völlig ausreichend, die letzten $k+1$ Werte von f zu speichern, der Wert f_{m+1} kann also beispielsweise den Wert f_{m-k} überschreiben. Demzufolge benötigt das neue Verfahren lediglich $k+1$ Hilfsvektoren und ist damit nicht viel speicheraufwendiger als ein Runge-Kutta-Verfahren.

Weil bei der Berechnung der Näherung im Zeitpunkt t_{m+1} die $k+1$ Näherungen in den Zeitpunkten $t_{m-k}, \dots, t_m$ eingehen, bezeichnen wir das neue Verfahren als $(k+1)$-Schritt-Verfahren.

7.2 Adams-Moulton-Verfahren

Wir haben bereits gesehen, dass bei steifen Differentialgleichungen ein implizites Verfahren sehr viel bessere Eigenschaften als ein explizites aufweisen kann. Also stellt sich die Frage nac impliziten Mehrschrittverfahren.

Herleitung

Wir gehen wieder von der Formel

$$y(t_{m+1}) = y(t_m) + \int_{t_m}^{t_{m+1}} f(s, y(s))\, ds \qquad \text{für } m \in \{0, \dots, n-1\}$$

aus, approximieren den Integranden g aber diesmal in den $k+2$ Punkten $t_{m-k}, \dots, t_{m+1}$. Die entsprechenden Lagrange-Polynome sind durch

$$\mathcal{L}_{m,i}(s) := \prod_{\substack{j=-1 \\ j \neq i}}^{k} \frac{s - t_{m-j}}{t_{m-i} - t_{m-j}} \qquad \text{für alle } m \in \{k, \dots, n-1\}, i \in \{-1, \dots, k\}$$

gegeben, der Interpolant durch

$$\tilde{g}_{m,k}(s) := \sum_{i=-1}^{k} f(t_{m-i}, y(t_{m-i}))\mathcal{L}_{m,i}(s) \qquad \text{für alle } s \in \mathbb{R}.$$

Wir integrieren den Interpolanten, um die Formel

$$y(t_{m+1}) \approx \tilde{y}(t_{m+1}) := y(t_m) + h \sum_{i=-1}^{k} w_{k,i} f(t_{m-i}, y(t_{m-i})) \qquad \text{für } m \in \{k, \dots, n-1\}$$

zu erhalten. Die Quadraturgewichte sind nun durch

$$w_{k,i} := \frac{1}{h} \int_{t_m}^{t_{m+1}} \mathcal{L}_{m,i}(s)\, ds = \int_0^1 \prod_{\substack{j=-1 \\ j \neq i}}^{k} \frac{s+j}{j-i}\, ds \qquad \text{für } i \in \{-1, \dots, k\}$$

gegeben.

Wenn wir die Näherung $\tilde{y}(t_{m+1})$ berechnen wollen, stellen wir fest, dass sie von der exakten Lösung $y(t_{m+1})$ abhängt. Um zu einem brauchbaren Verfahren zu gelangen, ersetzen wir die exakte Lösung durch die Näherung und erhalten die nichtlineare Gleichung

$$\tilde{y}(t_{m+1}) - hw_{k,-1}f(t_{m+1}, \tilde{y}(t_{m+1})) = y(t_m) + h\sum_{i=0}^{k} w_{k,i}f(t_{m-i}, y(t_{m-i}))$$

zur Bestimmung von $\tilde{y}(t_{m+1})$. Wie schon bei den impliziten Einschrittverfahren können wir nachweisen, dass diese Gleichung für eine hinreichend kleine Schrittweite h eindeutig lösbar ist: Mit

$$\Phi(x) := y(t_m) + hw_{k,-1}f(t_{m+1}, x) + h\sum_{i=0}^{k} w_{k,i}f(t_{m-i}, y(t_{m-i}))$$

gilt $\tilde{y}(t_{m+1}) = \Phi(\tilde{y}(t_{m+1}))$, und falls f Lipschitz-stetig mit der Lipschitz-Konstanten L ist, erhalten wir

$$\|\Phi(x) - \Phi(z)\| = hw_{k,-1}\|f(t_{m+1}, x) - f(t_{m+1}, z)\| \leq hw_{k,-1}L\|x - z\| \quad \text{für alle } x, z \in V,$$

aus $h < 1/|w_{k,-1}L|$ folgt also per Fixpunktsatz 2.1 die Existenz einer Lösung $\tilde{y}(t_{m+1})$.

Fehlerabschätzung

Da wir im impliziten Fall $k+2$ Stützstellen verwenden, nimmt die Fehlerdarstellung die Form

$$g(s) - \tilde{g}_{m,k}(s) = \frac{\omega(s)}{(k+2)!}g^{(k+2)}(\xi)$$

an, wobei $\xi \in [t_{m-k}, t_{m+1}]$ ein von $s \in [t_m, t_{m+1}]$ abhängender Zwischenpunkt ist und das Stützstellenpolynom durch

$$\omega(s) = (s - t_{m+1})(s - t_m)\dots(s - t_{m-k}) \qquad \text{für alle } s \in [t_{m-k}, t_{m+1}]$$

gegeben ist. Mit derselben Argumentation wie im Falle des Adams-Bashforth-Verfahrens erhalten wir

$$|\omega(s)| \leq h^{k+2}(k+1)! \qquad \text{für alle } s \in [t_m, t_{m+1}]$$

und folgern

$$\|y(t_{m+1}) - \tilde{y}(t_{m+1})\| \leq C_k h^{k+3} \qquad \text{für alle } m \in \{k, \dots, n-1\}$$

mit der Konstanten

$$C_k := \frac{1}{k+2}\sup\{\|g^{(k+2)}(\xi)\| \ : \ \xi \in [a, b]\}.$$

Obwohl das implizite Verfahren genausoviele „Werte aus der Vergangenheit“ wie das explizite Verfahren verwendet, erzielt es also ein besseres Fehlerverhalten.

Algorithmus

Wir können wieder die exakten Funktionswerte durch Näherungen ersetzen und erhalten das folgende Verfahren:

Berechne Näherungen $\eta_0, \ldots, \eta_k$ von $y_0, \ldots, y_k$
for $i \in \{0, \ldots, k\}$ **do** $f_i \leftarrow f(t_i, \eta_i)$
for $m := k$ **to** $n-1$ **do begin**
    Finde η_{m+1} mit $\eta_{m+1} = \eta_m + hw_{k,-1}f(t_{m+1}, \eta_{m+1}) + h\sum_{i=0}^{k} w_{k,i}f_{m-i}$
    $f_{m+1} \leftarrow f(t_{m+1}, \eta_{m+1})$
end

Bei diesem impliziten Verfahren stellt sich natürlich die Frage nach einem effizienten Lösungsverfahren für die nichtlineare Gleichung, durch die η_{m+1} definiert ist.

Ein nützliche Ansatz besteht darin, das explizite Verfahren zur Berechnung einer Ausgangsnäherung $\tilde{\eta}_{m+1}$ zu verwenden und dann mit einer Fixpunktiteration oder einem Newton-Verfahren die gesuchte Lösung η_{m+1} zu approximieren.

Falls die exakte Lösung y hinreichend glatt ist, dürfen wir davon ausgehen, dass $\tilde{\eta}_{m+1}$ bereits eine gute Näherung von η_{m+1} ist, so dass nur wenige Schritte des Iterationsverfahrens erforderlich sind.

Allgemein bezeichnet man solche Kombinationen aus expliziten und impliziten Verfahren als *Prädiktor-Korrektor-Verfahren*: Das Prädiktor-Verfahren (in diesem Fall das explizite) trifft eine „Vorhersage“ für die nächste Iterierte, das Korrektor-Verfahren (in diesem Fall das implizite) verbessert diese Ausgangsnäherung. Besonders attraktiv ist es natürlich, wenn beide Verfahren ein ähnliches Konvergenzverhalten besitzen, denn dann genügt unter Umständen ein einziger Korrektor-Schritt, um eine hinreichend gute Lösung zu finden, so dass das implizite Mehrschrittverfahren nur zwei Auswertungen der rechten Seite f pro Zeitschritt erfordert.

7.3 Stabilität expliziter Mehrschrittverfahren

Das explizite Adams-Bashforth-Verfahren hat die Form

$$\eta_{m+1} = \eta_m + h\sum_{i=0}^{k} w_{k,i}f_{m-i},$$

die Näherungslösung ergibt sich also aus einer Linearkombination von alten Näherungslösungen und den f-Werten.

Ein allgemeines explizites r-Schritt-Verfahren wird üblicherweise in der Form

$$\eta_{m+r} + a_{r-1}\eta_{m+r-1} + \ldots + a_0\eta_m = h\Phi(t_m, \eta_m, \ldots, \eta_{m+r-1}, h) \tag{7.3}$$

geschrieben, wobei η_{m+r} die nächste zu berechnende Näherung und $\eta_{m+r-1}, \ldots, \eta_m$ die dafür verwendeten vorangehenden Näherungen sind.

Während es bei Einschrittverfahren ausreichte, die lokalen Fehler unter Kontrolle zu halten, können sich bei Mehrschrittverfahren lokale Fehler von einem Schritt zum

nächsten vererben und dabei sogar verstärken. Als Beispiel untersuchen wir das Zweischrittverfahren

$$\eta_{m+2} + 4\eta_{m+1} - 5\eta_m = h\left(4f(t_{m+1}, \eta_{m+1}) + 2f(t_m, \eta_m)\right).$$

Wir wenden es auf das besonders einfache Anfangswertproblem

$$y(0) = 1, \qquad y'(t) = 0 \qquad \text{für alle } t \in \mathbb{R}_{\geq 0} \tag{7.4}$$

und erhalten die Gleichung

$$\eta_{m+2} + 4\eta_{m+1} - 5\eta_m = 0 \qquad \text{für alle } m \in \mathbb{N}_0. \tag{7.5}$$

Derartige *Differenzengleichungen* lassen sich mit Hilfe des Ansatzes $\eta_m = z^m$ lösen: Wir erhalten

$$z^{m+2} + 4z^{m+1} - 5z^m = (z^2 + 4z - 5)z^m = 0$$

und finden neben der trivialen Lösung $z_0 = 0$ wegen

$$z^2 + 4z - 5 = z^2 + 4z + 4 - 9 = (z+2)^2 - 9$$

auch die Lösungen $z_1 = 1$ und $z_2 = -5$. Offenbar erfüllt auch jede Linearkombination

$$\hat{\eta}_m = \alpha z_1^m + \beta z_2^m$$

die Gleichung (7.5). Also können wir α und β so wählen, dass $\hat{\eta}_0$ und $\hat{\eta}_1$ mit den Startwerten η_0 und η_1 übereinstimmen:

$$\begin{aligned} \eta_0 &= \alpha + \beta, & \eta_1 &= \alpha + \beta z_2, \\ \eta_0 &= \alpha + \beta, & \eta_1 &= \alpha - 5\beta, \\ \alpha &= \frac{1}{6}(5\eta_0 + \eta_1), & \beta &= \frac{1}{6}(\eta_0 - \eta_1). \end{aligned}$$

Aus $\eta_0 = \hat{\eta}_0$ und $\eta_1 = \hat{\eta}_1$ folgt dann aus der Definition von η_m direkt $\eta_m = \hat{\eta}_m$ für alle $m \in \mathbb{N}_0$, wir haben also eine explizite Darstellung für die Näherungslösungen gewonnen.

Wir gehen davon aus, dass $\eta_0 = 1$ exakt berechnet wurde, dass aber bei der Bestimmung von $\eta_1 = 1 + \epsilon$ ein Approximationsfehler $\epsilon \in \mathbb{R}$ aufgetreten ist. Daraus ergibt sich

$$\alpha = \frac{1}{6}(5 + 1 + \epsilon) = 1 + \frac{\epsilon}{6}, \qquad \beta = \frac{1}{6}(1 - (1 + \epsilon)) = -\frac{\epsilon}{6},$$

und wir erhalten die Darstellung

$$\eta_m = \hat{\eta}_m = \left(1 + \frac{\epsilon}{6}\right) - \frac{\epsilon}{6}(-5)^m \qquad \text{für alle } m \in \mathbb{N}_0.$$

Obwohl das Anfangswertproblem sehr einfach und die exakte Lösung sehr glatt ist, wird also die Näherungslösung sehr ausgeprägte Oszillationen aufweisen. Diese Oszillationen

haben nichts mit der tatsächlichen Lösung der Differentialgleichung zu tun, sondern entstehen durch die ungeschickte Wahl des Näherungsverfahrens.

Ausschlaggebend ist offenbar das Verhalten der Nullstellen des Polynoms $z^2 + 4z - 5$, das die Lösungsdarstellung $\hat{\eta}_m$ motiviert: Die Nullstelle $z_2 = -5$, deren Betrag größer als eins ist, kann zu starken Oszillationen führen, falls nicht gerade der zugehörige Koeffizient β verschwindet.

Für ein allgemeines Mehrschrittverfahren müssen wir also die Nullstellen des *charakteristischen Polynoms*

$$\varrho(z) := z^r + a_{r-1}z^{r-1} + \ldots + a_0$$

untersuchen: Falls es eine Nullstelle z_* mit $|z_*| > 1$ geben sollte, ist die Folge $\hat{\eta}_m := z_*^m$ eine Lösung von

$$\hat{\eta}_{m+r} + a_{r-1}\hat{\eta}_{m+r-1} + \ldots + a_0\hat{\eta}_m = \varrho(z_*)z_*^m = 0 \qquad \text{für alle } m \in \mathbb{N}_0, \tag{7.6}$$

und wegen $|z_*| > 1$ wächst diese Lösung exponentiell, demzufolge können kleine Fehler in den Anfangsdaten zu inakzeptablen Störungen führen.

Falls es eine Nullstelle z_* mit $|z_*| = 1$ gibt, tritt kein exponentielles Wachstum auf, falls allerdings z_* eine mehrfache Nullstelle von ϱ ist, kann sich immer noch eine störende Fehlerfortpflanzung ergeben: Falls $\varrho'(z_*) = 0$ gilt, erhalten wir für die Folge $\hat{\eta}_m := (m+1)z_*^m$ die Gleichung

$$\begin{aligned}
&\hat{\eta}_{m+r} + a_{r-1}\hat{\eta}_{m+r-1} + \ldots + a_0\hat{\eta}_m \\
&\quad = (m+r+1)z_*^{m+r} + a_{r-1}(m+r)z_*^{m+r-1} + \ldots + a_0(m+1)z_*^m \\
&\quad = \frac{\partial}{\partial z_*}(z_*^{m+r+1} + a_{r-1}z_*^{m+r} + \ldots + a_0 z_*^{m+1}) \\
&\quad = \frac{\partial}{\partial z_*}\left((z_*^r + a_{r-1}z_*^{r-1} + \ldots + a_0)z_*^{m+1}\right) \\
&\quad = \frac{\partial}{\partial z_*}(\varrho(z_*)z_*^{m+1}) = \varrho'(z_*)z_*^{m+1} + \varrho(z_*)(m+1)z_*^m = 0,
\end{aligned}$$

also ist auch diese Folge eine Lösung der homogenen Gleichung. Die Folge $(\hat{\eta}_m)_{m\in\mathbb{N}_0}$ divergiert offenbar, wenn auch nicht exponentiell.

Wenn wir sicherstellen wollen, dass ein Mehrschrittverfahren wenigstens für das triviale Problem (7.4) sinnvoll funktioniert, müssen wir also verhindern, dass ϱ Nullstellen hat, die divergierende Lösungen verursachen:

Definition 7.1 (Stabilitätsbedingung) *Seien $r \in \mathbb{N}$, $a_{r-1}, \ldots, a_0 \in \mathbb{R}$ die Koeffizienten eines r-Schritt-Verfahrens in der Form (7.3). Die Koeffizienten erfüllen die* Stabilitätsbedingung*, falls für jede Nullstelle $z_* \in \mathbb{C}$ des charakteristischen Polynoms*

$$\varrho(z_*) = z_*^r + a_{r-1}z_*^{r-1} + \ldots + a_0 \qquad \text{für } z_* \in \mathbb{C} \tag{7.7}$$

die Bedingung

$$|z_*| < 1 \qquad \text{oder} \qquad (|z_*| = 1 \text{ und } \varrho'(z_*) \neq 0)$$

erfüllt ist, falls also alle Nullstellen im Einheitskreis um Null liegen und eventuelle Nullstellen auf seinem Rand einfach sind.

Mit Hilfe dieser Bedingung können wir nun die Konvergenz von Mehrschrittverfahren analysieren. Dazu bedienen wir uns eines einfachen Tricks: Indem wir die Näherungen von r Zeitschritten in einem Vektor (oder Blockvektor, falls V nicht eindimensional ist) zusammenfassen, können wir die Analyse ähnlich wie für Einschrittverfahren durchführen.

Wir setzen also

$$\hat{\eta}_i := \begin{pmatrix} \eta_i \\ \vdots \\ \eta_{i+r-1} \end{pmatrix} \qquad \text{für alle } i \in \{0, \dots, n-r+1\}.$$

Gemäß (7.3) gilt dann

$$\hat{\eta}_{i+1} = \begin{pmatrix} (\hat{\eta}_i)_2 = \eta_{i+1} \\ \vdots \\ (\hat{\eta}_i)_r = \eta_{i+r-1} \\ -a_0(\hat{\eta}_i)_1 - \ldots - a_{r-1}(\hat{\eta}_i)_r + h\Phi(t_i, \hat{\eta}_i, h) \end{pmatrix} \qquad \text{für alle } i \in \{0, \dots, n-r\}.$$

Indem wir die Hilfsmatrix $A \in \mathbb{R}^{r \times r}$ und den Hilfsvektor $\delta \in \mathbb{R}^r$ mit

$$A = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & -a_1 & \ldots & -a_{r-1} \end{pmatrix}, \qquad \delta = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

und ihre *Tensorprodukte* mit der Identität auf V

$$A \otimes I := \begin{pmatrix} A_{11}I & \ldots & A_{1r}I \\ \vdots & \ddots & \vdots \\ A_{r1}I & \ldots & A_{rr}I \end{pmatrix}, \qquad \delta \otimes I := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ I \end{pmatrix}$$

einführen, nimmt diese Gleichung die Form

$$\hat{\eta}_{i+1} = (A \otimes I)\hat{\eta}_i + h(\delta \otimes I)\Phi(t_i, \hat{\eta}_i, h) \qquad \text{für alle } i \in \{0, \dots, n-r\} \tag{7.8}$$

an. Im skalarwertigen Fall vereinfacht sich diese Formel zu

$$\hat{\eta}_{i+1} = A\hat{\eta}_i + h\delta\Phi(t, \hat{\eta}_i, h) \qquad \text{für alle } i \in \{0, \dots, n-r\}.$$

Abgesehen von der Matrix A entspricht diese Formel gerade der Gleichung (3.3), die wir bei der Definition des Einschrittverfahrens kennengelernt haben.

Das Vorhandensein der Matrix A macht die Stabilitätsbedingung erforderlich, und sie steht in unmittelbarem Zusammenhang mit dem charakteristischen Polynom ϱ des Mehrschrittverfahrens:

Lemma 7.2 *Seien $r \in \mathbb{N}$, $a_{r-1}, \ldots, a_0 \in \mathbb{R}$ gegeben, und sei ϱ durch (7.7) definiert. Dann gilt*

$$\varrho(z) = \det(zI - A) \qquad \text{für alle } z \in \mathbb{C},$$

ϱ ist also das charakteristische Polynom von A.

Beweis. Wir untersuchen zunächst ein verwandtes Problem: Sei $(b_i)_{i\in\mathbb{N}_0}$ eine Folge in $\mathbb{C}$. Wir definieren Matrizen

$$B_{z,i} := \begin{pmatrix} z & -1 & & \\ & \ddots & \ddots & \\ & & z & -1 \\ b_0 & b_1 & \ldots & b_{i-1} \end{pmatrix} \qquad \text{für alle } z \in \mathbb{C}, i \in \mathbb{N}.$$

Wir fixieren $z \in \mathbb{C}$ und $i \in \mathbb{N}_{>1}$ und stellen per Entwicklung nach der letzten Spalte fest, dass

$$\det B_{z,i} = b_{i-1} \det \begin{pmatrix} z & -1 & \\ & \ddots & \ddots \\ & & z \end{pmatrix} + \det \begin{pmatrix} z & -1 & & \\ & \ddots & \ddots & \\ & & z & -1 \\ b_0 & b_1 & \ldots & b_{i-2} \end{pmatrix} = b_{i-1} z^{i-1} + \det B_{z,i-1}$$

gilt. Durch eine einfache Induktion erhalten wir

$$\det B_{z,i} = b_{i-1} z^{i-1} + \ldots + b_1 z + b_0 \qquad \text{für alle } z \in \mathbb{C}, i \in \mathbb{N}.$$

Nun können wir uns wieder dem ursprünglichen Problem zuwenden. Wir haben

$$\det(zI - A) = \det \begin{pmatrix} z & -1 & & \\ & \ddots & \ddots & \\ & & z & -1 \\ a_0 & a_1 & \ldots & z + a_{r-1} \end{pmatrix}$$

und können die Hilfsaussage auf $b_0 = a_0, \ldots, b_{r-2} = a_{r-2}, b_{r-1} = z + a_{r-1}$ anwenden, um $\det(zI - A) = \varrho(z)$ zu erhalten. ∎

Die Nullstellen von ϱ sind also die Eigenwerte von A, also trifft die Stabilitätsbedingung eine Aussage über das Spektrum von A. Für den Beweis der gesuchten Stabilitätsaussage genügt uns so eine Aussage nicht, wir benötigen eine Abschätzung einer geeigneten Norm von A.

Naheliegend wäre die durch

$$\|x\|_\infty := \max\{\|x_i\| \ : \ i \in \{1, \ldots, r\}\} \qquad \text{für alle } x \in V^r$$

definierte *Maximumnorm*, die es uns allerdings im Allgemeinen nicht erlaubt, die Stabilitätsbedingung optimal auszunutzen. Deshalb führen wir eine für diese Bedingung maßgeschneiderte Norm ein:

Lemma 7.3 (Norm zum Spektralradius) *Seien $r \in \mathbb{N}$, $a_{r-1}, \dots, a_0 \in \mathbb{R}$ die Koeffizienten eines r-Schritt-Verfahrens in der Form (7.3), die die Stabilitätsbedingung erfüllen. Dann existiert eine reguläre Matrix $R \in \mathbb{C}^{r\times r}$ so, dass die von*

$$\|\cdot\|_R : \mathbb{C}^r \to \mathbb{R}_{\geq 0}, \qquad x \mapsto \|Rx\|_\infty$$

induzierte Matrixnorm

$$\|X\|_R := \sup\left\{ \frac{\|Xx\|_R}{\|x\|_R} \;:\; x \in \mathbb{R}^r \setminus \{0\} \right\} \qquad \text{für alle } X \in \mathbb{R}^{r\times r}$$

die Ungleichung $\|A\|_R \leq 1$ erfüllt.

Beweis. Die von der Maximumnorm

$$\|x\|_\infty := \max\{|x_i| \;:\; i \in \{1,\dots,r\}\} \qquad \text{für alle } x \in \mathbb{R}^r$$

induzierte Matrixnorm

$$\|X\|_\infty := \sup\left\{ \frac{\|Xx\|_\infty}{\|x\|_\infty} \;:\; x \in \mathbb{R}^r \setminus \{0\} \right\} \qquad \text{für alle } X \in \mathbb{R}^{r\times r}$$

erfüllt die Gleichung

$$\|X\|_\infty = \max\left\{ \sum_{j=1}^{r} |X_{ij}| \;:\; i \in \{1,\dots,r\} \right\} \qquad \text{für alle } X \in \mathbb{R}^{r\times r},$$

wenn wir also die Summen der Beträge der Zeilen einer Matrix beschränken können, ist auch diese Norm beschränkt.

Seien $\lambda_1, \dots, \lambda_s \in \mathbb{C}$ die Eigenwerte von A, und seien $\mu_1, \dots, \mu_s \in \mathbb{N}$ die entsprechenden algebraischen Vielfachheiten. Es gibt eine reguläre Matrix $T \in \mathbb{C}^{r\times r}$, die A in die Jordan-Normalform überführt, mit der also

$$T^{-1}AT = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_s \end{pmatrix}, \qquad J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & \\ & & \ddots & \ddots \\ & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{\mu_i\times\mu_i}$$

gilt. Die Zeilensummen der Jordan-Blöcke J_i lassen sich direkt ablesen, und durch eine geeignete Skalierung können wir dafür sorgen, dass sie die gewünschte Größe annehmen: Wir setzen

$$\epsilon := \max\{1 - |\lambda_i| \;:\; i \in \{1,\dots,r\}, |\lambda_i| \neq 1\} > 0$$

und führen die Diagonalmatrix

$$D := \begin{pmatrix} 1 & & & \\ & \epsilon & & \\ & & \ddots & \\ & & & \epsilon^{r-1} \end{pmatrix} \in \mathbb{R}^{r\times r}$$

ein. Eine Ähnlichkeitstransformation mit dieser Matrix ergibt

$$D^{-1}T^{-1}ATD = \begin{pmatrix} \widetilde{J}_1 & & \\ & \ddots & \\ & & \widetilde{J}_s \end{pmatrix}, \qquad \widetilde{J}_i = \begin{pmatrix} \lambda_i & \epsilon & & \\ & \lambda_i & \epsilon & \\ & & \ddots & \ddots \\ & & & \lambda_i \end{pmatrix}.$$

Für ein λ_i mit $|\lambda_i| < 1$ gilt nach Definition $|\lambda_i| + \epsilon \leq 1$, und damit sind die Zeilensummen in $\widetilde{J}_i$ durch eins beschränkt.

Für ein λ_i mit $|\lambda_i| = 1$ gilt nach Voraussetzung $\mu_i = 1$, der Jordan-Block $\widetilde{J}_i$ ist also trivial und besitzt damit insbesondere auch eine Zeilensumme von eins.

Da Eigenwerte λ_i mit $|\lambda_i| > 1$ nach Voraussetzung ausgeschlossen sind, erhalten wir insgesamt

$$\|D^{-1}T^{-1}ATD\|_\infty \leq 1.$$

Nun müssen wir lediglich noch die gesuchte Matrix definieren: Mit

$$R := D^{-1}T^{-1}, \qquad R^{-1} = TD$$

erhalten wir

$$\|RAR^{-1}\|_\infty \leq 1 \tag{7.9}$$

und damit auch

$$\begin{aligned} \|A\|_R &= \sup\left\{ \frac{\|Ax\|_R}{\|x\|_R} \ : \ x \in \mathbb{C}^r \right\} = \sup\left\{ \frac{\|RAx\|_\infty}{\|Rx\|_\infty} \ : \ x \in \mathbb{C}^r \right\} \\ &= \sup\left\{ \frac{\|RAR^{-1}y\|_\infty}{\|y\|_\infty} \ : \ y \in \mathbb{C}^r \right\} = \|RAR^{-1}\|_\infty \leq 1. \end{aligned}$$

Es ist zu beachten, dass die Konstruktion auf eine *spezielle* Matrix A zugeschnitten ist und die Norm $\|\cdot\|_\varrho$ deshalb nicht unbedingt für die Matrizen zu anderen stabilen Mehrschrittverfahren dieselbe Abschätzung erfüllt. ∎

Um auch den vektorwertigen Fall behandeln zu können, müssen wir dieses Resultat auch auf Blockmatrizen übertragen:

Lemma 7.4 *Seien $r \in \mathbb{N}$, $a_{r-1}, \ldots, a_0 \in \mathbb{R}$ die Koeffizienten eine r-Schritt-Verfahrens in der Form (7.3), die die Stabilitätsbedingung erfüllen. Dann existiert eine Norm $\|\cdot\|_\varrho : V^r \to \mathbb{R}_{\geq 0}$ so, dass*

$$\|(A \otimes I)x\|_\varrho \leq \|x\|_\varrho \qquad \text{für alle } x \in V^r$$

gilt. Das impliziert insbesondere, dass Fehler in den Startwerten durch den Berechnungsschritt (7.8) nicht verstärkt werden.

Beweis. Sei $R \in \mathbb{C}^{r\times r}$ die Matrix aus Lemma 7.3. Wir definieren die Norm $\|\cdot\|_\varrho$ durch

$$\|x\|_\varrho := \max\{\|R_{i1}x_1 + \ldots + R_{ir}x_r\| \ : \ i \in \{1, \ldots, r\}\} \qquad \text{für alle } x \in V^r.$$

Sei $x \in V^r$, und sei

$$y := (A \otimes I)x = \begin{pmatrix} A_{11}x_1 + \ldots + A_{1r}x_r \\ \vdots \\ A_{r1}x_1 + \ldots + A_{rr}x_r \end{pmatrix}.$$

Durch einfaches Einsetzen erhalten wir

$$\begin{pmatrix} R_{11}y_1 + \ldots + R_{1r}y_r \\ \vdots \\ R_{r1}y_1 + \ldots + R_{rr}y_r \end{pmatrix} = \begin{pmatrix} (RA)_{11}x_1 + \ldots + (RA)_{1r}y_r \\ \vdots \\ (RA)_{r1}x_1 + \ldots + (RA)_{rr}y_r \end{pmatrix}.$$

Wir führen den weiteren Hilfsvektor

$$z := \begin{pmatrix} R_{11}x_1 + \ldots + R_{1r}x_r \\ \vdots \\ R_{r1}x_1 + \ldots + R_{rr}x_r \end{pmatrix}$$

ein und stellen fest, dass

$$\begin{pmatrix} (R^{-1})_{11}z_1 + \ldots + (R^{-1})_{1r}z_r \\ \vdots \\ (R^{-1})_{r1}z_1 + \ldots + (R^{-1})_{rr}z_r \end{pmatrix} = \begin{pmatrix} (R^{-1}R)_{11}x_1 + \ldots + (R^{-1}R)_{1r}x_r \\ \vdots \\ (R^{-1}R)_{r1}x_1 + \ldots + (R^{-1}R)_{rr}x_r \end{pmatrix} = x$$

gilt. Damit haben wir schließlich

$$\begin{aligned} \|y\|_\varrho &= \max\{\|R_{i1}y_1 + \ldots + R_{ir}y_r\| \ : \ i \in \{1, \ldots, r\}\} \\ &= \max\{\|(RA)_{i1}x_1 + \ldots + (RA)_{ir}x_r\| \ : \ i \in \{1, \ldots, r\}\} \\ &= \max\{\|(RAR^{-1})_{i1}z_1 + \ldots + (RAR^{-1})z_r\| \ : \ i \in \{1, \ldots, r\}\} \end{aligned}$$

erhalten und können für jedes $i \in \{1, \ldots, r\}$ die Abschätzung

$$\begin{aligned} \|(RAR^{-1})_{i1}z_1 + \ldots + (RAR^{-1})_{ir}z_r\| &\leq |(RAR^{-1})_{i1}| \, \|z_1\| + \ldots + |(RAR^{-1})_{ir}| \, \|z_r\| \\ &\leq \|RAR^{-1}\|_\infty \max\{\|z_j\| \ : \ j \in \{1, \ldots, r\}\} \end{aligned}$$

ausnutzen, um schließlich

$$\begin{aligned} \|y\|_\varrho &\leq \|RAR^{-1}\|_\infty \max\{\|z_j\| \ : \ j \in \{1, \ldots, r\}\} \\ &= \|RAR^{-1}\|_\infty \max\{\|R_{j1}x_1 + \ldots + R_{jr}x_r\| \ : \ j \in \{1, \ldots, r\}\} \\ &= \|RAR^{-1}\|_\infty \|x\|_\varrho \end{aligned}$$

zu beweisen. Aus (7.9) folgt jetzt das gewünschte Ergebnis. ∎

Die etwas ungewohnte Norm $\|\cdot\|_\varrho$ können wir mit Hilfe einfacher Argumente in Beziehung zu der Maximumnorm setzen:

Lemma 7.5 (Normäquivalenz) *Mit der Konstanten*

$$C_\varrho := \max\{|R_{i1}| + \ldots + |R_{ir}|, |(R^{-1})_{i1}| + \ldots + |(R^{-1})_{ir}| \ : \ i \in \{1, \ldots, r\}\} \geq 1$$

gelten

$$\|x\|_\varrho \leq C_\varrho \|x\|_\infty, \qquad \|x\|_\infty \leq C_\varrho \|x\|_\varrho \qquad \text{für alle } x \in V^r.$$

Beweis. Sei $x \in V^r$. Die erste Abschätzung erhalten wir aus

$$\begin{aligned}
\|x\|_\varrho &= \max\{\|R_{i1}x_1 + \ldots + R_{ir}x_r\| \ : \ i \in \{1, \ldots, r\}\} \\
&\leq \max\{|R_{i1}| \, \|x_1\| + \ldots + |R_{ir}| \, \|x_r\| \ : \ i \in \{1, \ldots, r\}\} \\
&\leq \max\{|R_{i1}| + \ldots + |R_{ir}| \ : \ i \in \{1, \ldots, r\}\} \max\{\|x_j\| \ : \ j \in \{1, \ldots, r\}\} \\
&\leq C_\varrho \|x\|_\infty.
\end{aligned}$$

Für die zweite Abschätzung führen wir

$$z := \begin{pmatrix} R_{11}x_1 + \ldots + R_{1r}x_r \\ \vdots \\ R_{r1}x_1 + \ldots + R_{rr}x_r \end{pmatrix}$$

ein und stellen fest, dass $\|x\|_\varrho = \|z\|_\infty$ und

$$\begin{pmatrix} (R^{-1})_{11}z_1 + \ldots + (R^{-1})_{1r}z_r \\ \vdots \\ (R^{-1})_{r1}z_1 + \ldots + (R^{-1})_{rr}z_r \end{pmatrix} = \begin{pmatrix} (R^{-1}R)_{11}x_1 + \ldots + (R^{-1}R)_{1r}x_r \\ \vdots \\ (R^{-1}R)_{r1}x_1 + \ldots + (R^{-1}R)_{rr}x_r \end{pmatrix} = x$$

gelten, so dass wir schließlich

$$\begin{aligned}
\|x\|_\infty &= \max\{\|(R^{-1})_{i1}z_1 + \ldots + (R^{-1})_{ir}z_r\| \ : \ i \in \{1, \ldots, r\}\} \\
&\leq \max\{|(R^{-1})_{i1}| \, \|z_1\| + \ldots + |(R^{-1})_{ir}| \, \|z_r\| \ : \ i \in \{1, \ldots, r\}\} \\
&\leq \max\{|(R^{-1})_{i1}| + \ldots + |(R^{-1})_{ir}| \ : \ i \in \{1, \ldots, r\}\} \max\{\|z_j\| \ : \ j \in \{1, \ldots, r\}\} \\
&\leq C_\varrho \|z\|_\infty = C_\varrho \|x\|_\varrho
\end{aligned}$$

erhalten. Indem wir beide Abschätzungen kombinieren erhalten wir

$$\|x\|_\infty \leq C_\varrho \|x\|_\varrho \leq C_\varrho^2 \|x\|_\infty,$$

also muss $C_\varrho^2 \geq 1$ und damit auch $C_\varrho \geq 1$ gelten. ∎

In der richtigen Norm gemessen bleibt also der Einfluss der Matrix A unter Kontrolle, so dass wir uns nun der Analyse der Inkrementfunktion zuwenden können. Wie im Einschrittverfahren genügt es, die Lipschitz-Stetigkeit von Φ vorauszusetzen:

Definition 7.6 (Stabiles Mehrschrittverfahren) *Seien $r \in \mathbb{N}$, $a_0, \dots, a_{r-1} \in \mathbb{R}$ und Φ die definierenden Größen eines Mehrschrittverfahrens in der Form (7.3). Wir bezeichnen das Verfahren als* stabil, *falls die Koeffizienten $a_0, \dots, a_{r-1}$ die Stabilitätsbedingung erfüllen und ein $L \in \mathbb{R}_{\geq 0}$ existiert, dass*

$$\|\Phi(t,x,h) - \Phi(t,z,h)\| \leq L\|x-z\|_\infty \qquad \text{für alle } x,z \in V^r$$

erfüllt, falls also die Inkrementfunktion Φ des Verfahrens Lipschitz-stetig ist.

Für ein stabiles Mehrschrittverfahren gilt eine Variante der aus Lemma 3.1 bekannten Abschätzung für die Fehlerverstärkung des Näherungsverfahrens:

Lemma 7.7 (Störungen im Mehrschrittverfahren) *Sei ein stabiles Mehrschrittverfahren in der Form (7.3) und Startwerte $x, z \in V^r$ gegeben. Seien $\hat{\eta}_i^x, \hat{\eta}_i^z \in V^r$ die durch (7.8) für diese Startwerte definierten Näherungslösungen des Mehrschrittverfahrens. Dann gilt*

$$\|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\infty \leq C_\varrho^2 e^{C_\varrho^2 L(t_i - a)}\|x-z\|_\infty \qquad \text{für alle } i \in \{0, \dots, n-r+1\},$$

Beweis. Wir zeigen zunächst

$$\|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\varrho \leq e^{C_\varrho^2 L(t_i - a)}\|x-z\|_\varrho \qquad \text{für alle } i \in \{0, \dots, n-r+1\}.$$

Für $i = 0$ ist die Aussage trivial.

Sei nun $i \in \{0, \dots, n-r\}$ so gewählt, dass die Aussage gilt. Mit Hilfe der Dreiecksungleichung und des Lemmas 7.4 folgt aus (7.8) die Abschätzung

$$\begin{aligned}
\|\hat{\eta}_{i+1}^x - \hat{\eta}_{i+1}^z\|_\varrho &\leq \|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\varrho + h \max\{\|R_{ir}(\Phi(t_i, \hat{\eta}_i^x, h) - \Phi(t_i, \hat{\eta}_i^z, h))\| \ : \ i \in \{1, \dots, r\}\} \\
&\leq \|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\varrho + h \max\{|R_{ir}| \ : \ i \in \{1, \dots, r\}\}\|\Phi(t_i, \hat{\eta}_i^x, h) - \Phi(t_i, \hat{\eta}_i^z, h)\| \\
&\leq \|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\varrho + hC_\varrho L\|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\infty \leq (1 + hC_\varrho^2 L)\|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\varrho \\
&\leq e^{hC_\varrho^2 L}\|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\varrho.
\end{aligned}$$

Per Induktionsvoraussetzung folgt daraus

$$\|\hat{\eta}_{i+1}^x - \hat{\eta}_{i+1}^z\|_\varrho \leq e^{h(i+1)C_\varrho^2 L}\|x-z\|_\varrho = e^{C_\varrho^2 L(t_{i+1} - a)}\|x-z\|_\varrho,$$

und die Induktion ist vollständig.

Mit Hilfe von Lemma 7.5 erhalten wir daraus

$$\|\hat{\eta}_i^x - \hat{\eta}_i^z\|_\infty \leq C_\varrho^2 e^{C_\varrho^2 L(t_i - a)}\|x-z\|_\infty \qquad \text{für alle } i \in \{0, \dots, n-r+1\},$$

und das ist die gewünschte Abschätzung. ∎

Beispiel 7.8 (Adams-Bashforth) *Im Falle des Adams-Bashforth-Verfahrens hat die Matrix A eine besonders einfache Struktur: Wegen $\eta_{m+r} = \eta_{m+r-1} + h\Phi(\ldots)$ gilt $a_{r-1} = -1$, $a_{r-2} = 0, \ldots, a_0 = 0$, also insbesondere*

$$A = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 1 \end{pmatrix}.$$

Daraus können wir direkt ablesen, dass A eine einfache Nullstelle bei 1 und eine $(r-1)$-fache Nullstelle bei 0 besitzt und damit die Stabilitätsbedingung erfüllt.

In diesem besonderen Fall gilt $\|A\|_\infty = 1$, so dass wir auf die Transformation mit einer Matrix R verzichten können und deshalb $C_\varrho = 1$ erhalten. Damit entspricht die Abschätzung aus Lemma 7.7 der von Lemma 3.1 für Einschrittverfahren.

7.4 Konvergenz von Mehrschrittverfahren

Mit Hilfe des Stabilitätsresultats aus Lemma 7.7 können wir uns nun der Analyse der Konvergenz expliziter Mehrschrittverfahren widmen.

Diese Analyse können wir ähnlich wie im Falle des Einschrittverfahrens durchführen, indem wir untersuchen, wie sich die in den einzelnen Schritten hinzukommenden lokalen Fehler auf die Gesamtlösung auswirken.

Wie schon bei der Stabilitätsanalyse fassen wir wieder r Näherungslösungen zu Zeitpunkten $t_i, \ldots, t_{i+r-1}$ zu einem Vektor zusammen. Offenbar ist das nur für die Indizes $i \in \{0, \ldots, n-r+1\}$ sinnvoll, also für $t_i \leq \hat{b} := t_{n-r+1} = b - h(r-1)$.

Zu einem gegebenen Vektor $\hat{\eta}_j \in V^r$ von Startwerten zu einem Startzeitpunkt t_j definieren wir die Näherungslösung

$$\hat{\eta}(\cdot; t_j, \hat{\eta}_j) : [t_j, \hat{b}]_h \to V^r$$

in Anlehnung an (7.8) durch die induktive Gleichung

$$\hat{\eta}(t_i; t_j, \hat{\eta}_j) := \begin{cases} (A \otimes I)\hat{\eta}(t_{i-1}; t_j, \hat{\eta}_j) + h(\delta \otimes I)\Phi(t_{i-t}, \hat{\eta}(t_{i-1}; t_j, \hat{\eta}_j), h) & \text{falls } t_i > t_j, \\ \hat{\eta}_j & \text{ansonsten} \end{cases}$$

für alle $t_i \in [t_j, \hat{b}]_h$. Diese Definition impliziert bereits die Fortsetzungseigenschaft

$$\hat{\eta}(t_i; t_j, \hat{\eta}_j) = \hat{\eta}(t_i; t_k, \hat{\eta}(t_k; t_j, \hat{\eta}_j)) \qquad \text{für alle } t_i, t_k \in [t_j, \hat{b}] \text{ mit } t_i > t_k, \tag{7.10}$$

die im Beweis der Konvergenz des Verfahrens eine zentrale Rolle spielen wird.

Wie im Falle des Einschrittverfahrens ist es unser Ziel, die Näherungslösungen mit den exakten Lösungen zu vergleichen, die wir zu Vektoren

$$\hat{y}(t_i; t_j, y_j) := \begin{pmatrix} y(t_i; t_j, y_j) \\ \vdots \\ y(t_{i+r-1}; t_j, y_j) \end{pmatrix} \qquad \text{für alle } t_i \in [t_j, \hat{b}]_h$$

zusammenfassen. Wie im Falle der Einschrittverfahren interessiert uns die *Konsistenz* von Mehrschrittverfahren, also Abschätzungen darüber, wie gut ein einzelner Schritt des Verfahrens, angewandt auf *exakte* Startwerte, die exakte Lösung approximiert.

Definition 7.9 (Diskretisierungsfehler) *Wir definieren den* lokalen Diskretisierungsfehler *zu einem in der Form (7.3) gegebenen Mehrschrittverfahren durch*

$$\hat{\tau}(t,\hat{x},h) := \frac{\hat{y}(t+h;t,\hat{x}_0) - \hat{\eta}(t+h;t,\hat{x})}{h} \qquad \text{für alle } h \in \mathbb{R}_{>0}, t \in [a, b-rh], x \in V^r.$$

Für Einschrittverfahren stimmt diese Definition mit Definition 3.3 überein. Wie schon im Fall der Einschrittverfahren bezeichnen wir ein Mehrschrittverfahren als konsistent mit einem Anfangswertproblem, falls der Diskretisierungsfehler entlang dessen Lösungsgraphen konvergiert.

Definition 7.10 (Konsistenz) *Sei ein Mehrschrittverfahren in der Form (7.3) gegeben. Es heißt* konsistent mit dem Anfangswertproblem (2.1)*, falls*

$$\lim_{h\to 0} \sup\{\|\hat{\tau}(t,\hat{y}(t),h)\|_\infty \ : \ t \in [a, b-rh]\} = 0 \tag{7.11}$$

gilt. Das Verfahren heißt von der Ordnung p konsistent mit dem Problem *für ein $p \in \mathbb{N}$, falls es Konstanten $C_y, h_y \in \mathbb{R}_{>0}$ so gibt, dass*

$$\sup\{\|\hat{\tau}(t,\hat{y}(t),h)\|_\infty \ : \ t \in [a, b-rh]\} \le C_y h^p \qquad \text{für alle } h \in (0, h_y) \tag{7.12}$$

gilt. Offenbar impliziert diese Bedingung bereits, dass das Verfahren auch konsistent ist.

Für ein konsistentes und stabiles Mehrschrittverfahren können wir den Konvergenzbeweis von Satz 3.2 anpassen, um eine Fehlerabschätzung zu erhalten:

Satz 7.11 (Konvergenz) *Sei ein stabiles und konsistentes Mehrschrittverfahren gegeben. Sei $y : [a,b] \to V$ eine Lösung des Anfangswertproblems (2.1). Sei $h \in \mathbb{R}_{>0}$ und*

$$K := \sup\{\|\hat{\tau}(t,\hat{y}(t),h)\|_\infty \ : \ t \in [a, b-rh]\}.$$

Dann gilt

$$\|\hat{y}(t) - \hat{\eta}(t)\|_\infty \le \begin{cases} \frac{K}{L}(e^{C_\varrho^2 L(t-a)} - 1) & \text{falls } L > 0, \\ KC_\varrho^2(t-a) & \text{ansonsten} \end{cases} \qquad \text{für alle } t \in [a, \hat{b}]_h.$$

Beweis. Sei $t \in [a,\hat{b}]_h$ gegeben. Der Fall $t = a$ ist trivial, wir untersuchen deshalb nur die Fälle $t = t_i$ mit $i \in \{1, \dots, n-r\}$.

Die zentrale Abschätzung des Beweises lautet

$$\|\hat{y}(t;t_j,y_j) - \hat{\eta}(t;t_j,\hat{y}_j)\|_\infty \le KhC_\varrho^2 \sum_{\ell=j+1}^{i} e^{C_\varrho^2 Lh(i-\ell)} \quad \text{für alle } j \in \{0,\dots,i-1\} \tag{7.13}$$

und wird durch (endliche, absteigende) Induktion über j bewiesen.

Für den Induktionsanfang $j = i - 1$ haben wir

$$\|\hat{y}(t; t_j, y_j) - \hat{\eta}(t; t_j, \hat{y}_j)\|_\infty = \|\hat{y}(t_i; t_{i-1}, y_{j-1}) - \hat{\eta}(t_i; t_{i-1}, \hat{y}_{i-1})\|_\infty \le Kh$$

nach Definition von K, und wegen $C_\varrho \ge 1$ impliziert diese Abschätzung bereits (7.13) für $j = i - 1$.

Sei nun $j \in \{1, \dots, i-1\}$ so gegeben, dass (7.13) gilt. Dank der Fortsetzungseigenschaft (7.10) und $\hat{y}_j = \hat{y}(t_j; t_{j-1}, y_{j-1})$ gilt

$$\begin{aligned}
\|\hat{y}(t; t_{j-1}, y_{j-1}) - \hat{\eta}(t; t_{j-1}, \hat{y}_{j-1})\|_\infty &= \|\hat{y}(t; t_j, y_j) - \hat{\eta}(t; t_j, \hat{\eta}(t_j; t_{j-1}, \hat{y}_{j-1}))\|_\infty \\
&\le \|\hat{y}(t; t_j, y_j) - \hat{\eta}(t; t_j, y_j)\|_\infty \\
&\quad + \|\hat{\eta}(t; t_j, \hat{y}(t_j; t_{j-1}, y_{j-1})) - \hat{\eta}(t; t_j, \hat{\eta}(t_j; t_{j-1}, \hat{y}_{j-1}))\|_\infty \\
&\le \|\hat{y}(t; t_j, y_j) - \hat{\eta}(t; t_j, y_j)\|_\infty \\
&\quad + C_\varrho^2 e^{C_\varrho^2 L(t-t_j)} \|\hat{y}(t_j; t_{j-1}, y_{j-1}) - \hat{\eta}(t_j; t_{j-1}, \hat{y}_{j-1})\|_\infty \\
&\le \|\hat{y}(t; t_j, y_j) - \hat{\eta}(t; t_j, y_j)\|_\infty + C_\varrho^2 e^{C_\varrho^2 Lh(i-j)} Kh.
\end{aligned}$$

Den ersten Term können wir mit Hilfe der Induktionsvoraussetzung abschätzen, so dass schließlich

$$\begin{aligned}
\|\hat{y}(t; t_{j-1}, y_{j-1}) - \hat{\eta}(t; t_{j-1}, \hat{y}_{j-1})\|_\infty &\le KhC_\varrho^2 \sum_{\ell=j+1}^{i} e^{C_\varrho^2 Lh(i-\ell)} + KhC_\varrho^2 e^{C_\varrho^2 Lh(i-j)} \\
&= KhC_\varrho^2 \sum_{\ell=j}^{i} e^{C_\varrho^2 Lh(i-j)}
\end{aligned}$$

bewiesen ist. Das ist die zu zeigende Abschätzung (7.13) für $j - 1$, also ist die Induktion vollständig.

Wir wenden (7.13) auf $j = 0$ an und erhalten

$$\|\hat{y}(t) - \hat{\eta}(t)\|_\infty = \|\hat{y}(t; t_0, y_0) - \hat{\eta}(t; t_0, y_0)\|_\infty \le KhC_\varrho^2 \sum_{\ell=1}^{i} e^{C_\varrho^2 Lh(i-\ell)}. \tag{7.14}$$

Für $L = 0$ ist die Summe gleich i, also $(t - a)/h$, so dass wir die gewünschte Schranke erhalten. Für $L > 0$ gilt $e^{C_\varrho^2 Lh} > 1$ und wir erhalten

$$\begin{aligned}
\sum_{\ell=1}^{i} e^{C_\varrho^2 Lh(i-\ell)} &= e^{C_\varrho^2 Lhi} \sum_{\ell=1}^{i} e^{-C_\varrho^2 Lh\ell} = e^{C_\varrho^2 L(t-a)} \sum_{\ell=1}^{i} (e^{-C_\varrho^2 Lh})^\ell \\
&= e^{C_\varrho^2 L(t-a)} \frac{e^{-C_\varrho^2 Lh} - e^{-C_\varrho^2 Lh(i+1)}}{1 - e^{-C_\varrho^2 Lh}} = e^{C_\varrho^2 L(t-a)} \frac{1 - e^{-C_\varrho^2 Lhi}}{e^{C_\varrho^2 Lh} - 1} \\
&\le e^{C_\varrho^2 L(t-a)} \frac{1 - e^{-C_\varrho^2 Lhi}}{1 + C_\varrho^2 Lh - 1} = \frac{e^{C_\varrho^2 L(t-a)} - 1}{C_\varrho^2 Lh},
\end{aligned}$$

so dass (7.14) die Form

$$\|\hat{y}(t) - \hat{\eta}(t)\|_\infty \leq \frac{K}{L} e^{C_\varrho^2 L(t-a)}$$

annimmt und der Beweis vollständig ist. ∎

Auch dieser Satz ist eine Verallgemeinerung des entsprechenden Resultats (nämlich Satz 3.2) für das Einschrittverfahren: Für $r = 1$ ist $C_\varrho = 1$, $\hat{y} = y$ und $\hat{\eta} = \eta$, so dass beide Fehlerabschätzungen übereinstimmen.

Korollar 7.12 (Konvergenzordnung) *Sei ein stabiles Mehrschrittverfahren gegeben, das konsistent von p-ter Ordnung mit dem Anfangswertproblem (2.1) ist.*

Dann existiert ein $C_K \in \mathbb{R}_{\geq 0}$ mit

$$\|\hat{y}(t) - \hat{\eta}(t)\|_\infty \leq C_K h^p \qquad \text{für alle } h \in (0, h_y), t \in [a, \hat{b}]_h.$$

Beweis. Wir setzen

$$C_K := \begin{cases} \frac{C_y}{L} e^{C_\varrho^2 L(b-a)-1} & \text{falls } L > 0, \\ C_y C_\varrho^2 (b-a) & \text{ansonsten,} \end{cases}$$

und erhalten durch Kombination von Satz 7.11 mit Definition 7.10 die gewünschte Abschätzung mit $K \leq C_y h^p$. ∎

Bemerkung 7.13 (Genäherte Startwerte) *In der Praxis können wir das Näherungsverfahren zur Berechnung von $\hat{\eta}$ häufig nicht mit exakten Startwerten $\hat{y}_0$ versorgen, müssen also auf Approximationen zurückgreifen, die beispielsweise mit Hilfe eines Einschrittverfahrens berechnet werden.*

Sei ein stabiles Mehrschrittverfahren gegeben, das konsistent von p-ter Ordnung mit dem Anfangswertproblem (2.1) ist. Sei die Näherungslösung $\hat{\eta}$ mit genäherten Startwerten $\hat{\eta}_0 \in V^r$ berechnet worden. Wir kombinieren Korollar 7.12 mit Lemma 7.7, um

$$\begin{aligned} \|\hat{y}(t) - \hat{\eta}(t)\|_\infty &\leq \|\hat{y}(t) - \hat{\eta}(t; t_0, \hat{y}_0)\|_\infty + \|\hat{\eta}(t; t_0, \hat{y}_0) - \hat{\eta}(t; t_0, \hat{\eta}_0)\|_\infty \\ &\leq C_K h^p + C_\varrho^2 e^{C_\varrho^2 L(b-a)} \|\hat{y}_0 - \hat{\eta}_0\|_\infty \end{aligned}$$

zu erhalten. Solange also die Startwerte ebenfalls mit einer Genauigkeit von h^p berechnet werden, wird auch der Gesamtfehler weiterhin mit dieser Ordnung konvergieren.

Wenden wir uns nun der Analyse des lokalen Diskretisierungsfehlers $\tau(t, \hat{x}, h)$ zu. Wir sind lediglich an $\tau(t, \hat{y}(t), h)$ interessiert, und in diesem Fall gelten

$$\hat{y}(t + h; t, \hat{y}(t)) = \begin{pmatrix} y(t+h) \\ \vdots \\ y(t + (r-1)h) \\ y(t + rh) \end{pmatrix},$$

$$\hat{\eta}(t+h;t,\hat{y}(t)) = \begin{pmatrix} y(t+h) \\ \vdots \\ y(t+(r-1)h) \\ -a_{r-1}y(t+(r-1)h) - \ldots - a_0 y(t) + h\Phi(t,\hat{y}(t),h) \end{pmatrix},$$

also unterscheiden sich lediglich die letzten Komponenten der beiden Vektoren, und es folgt die Gleichung

$$\|\hat{\tau}(t,\hat{y}(t),h)\|_\infty = \frac{1}{h}\|y(t+rh) + a_{r-1}y(t+(r-1)h) + \ldots + a_0 y(t) - h\Phi(t,\hat{y}(t),h)\|.$$

Mit ihrer Hilfe können wir nun die Konsistenz gängiger Mehrschrittverfahren analysieren.

Beispiel 7.14 (Adams-Bashforth) *Im Adams-Bashforth-Verfahren ist $a_{r-1} = -1$ und $a_{r-2} = \ldots = a_0 = 0$, während die Inkrementfunktion Φ durch eine Quadraturformel definiert ist: Infolge des Fundamentalsatzes der Differential- und Integralrechnung gilt*

$$y(t_{i+r}) = y(t_{i+r-1}) + \int_{t_{i+r-1}}^{t_{i+r}} y'(s)\,ds = y(t_{i+r-1}) + \int_{t_{i+r-1}}^{t_{i+r}} f(s,y(s))\,ds.$$

Wir ersetzen den Integranden

$$g(s) := f(s,y(s))$$

durch seinen Interpolanten

$$\tilde{g}(s) := \sum_{j=0}^{r-1} \mathcal{L}_{i+j}(s) g(t_{i+j})$$

in den r Punkten $t_i,\ldots,t_{i+r-1}$ mit den zugehörigen Lagrange-Polynomen $\mathcal{L}_i,\ldots,\mathcal{L}_{i+r-1}$ und erhalten so

$$y(t_{i+r}) \approx \eta(t_{i+r};t_i,\hat{y}(t_i)) = y(t_{i+r-1}) + \sum_{j=0}^{r-1} g(t_{i+j}) \int_{t_{i+r-1}}^{t_{i+r}} \mathcal{L}_{i+j}(s)\,ds.$$

Der Approximationsfehler ist durch

$$\begin{aligned} \|y(t_{i+r}) - \eta(t_{i+r};t_i,\hat{y}(t_i))\| &\le \int_{t_{i+r-1}}^{t_{i+r}} \|g(s) - \tilde{g}(s)\|\,ds \\ &\le h \sup\{\|g(s) - \tilde{g}(s)\| \ : \ s \in [t_{i+r-1},t_{i+r}]\} \end{aligned}$$

beschränkt, und wir haben bereits in (7.2) nachgewiesen, dass

$$\|g(s) - \tilde{g}(s)\| \le h^r \sup\{\|g^{(r)}(s_+)\| \ : \ s_+ \in [t_i,t_{i+r}]\} \qquad \text{für alle } s \in [t_{i+r-1},t_{i+r}]$$

gilt, so dass wir schließlich

$$\|y(t_{i+r}) - \eta(t_{i+r};t_i,\hat{y}(t_i))\| \le h^{r+1} \sup\{\|g^{(r)}(s_+)\| \ : \ s_+ \in [t_i,t_{i+r}]\}$$

erhalten. Mit

$$C_y := \max\{\|g^{(r)}(s_+)\| \ : \ s_+ \in [a,b]\} = \max\{\|y^{(r+1)}(s_+)\| \ : \ s_+ \in [a,b]\}$$

folgt daraus sofort

$$\|\hat{\tau}(t,\hat{y}(t),h)\| \le C_y h^r \qquad \text{für alle } h \in \mathbb{R}_{>0}, t \in [a, b-rh],$$

also besitzt das r-schrittige Adams-Bashforth-Verfahren eine Konsistenzordnung von r, falls die $(r+1)$-te Ableidung der Lösung y stetig und beschränkt ist.

Beispiel 7.15 (Leapfrog-Verfahren) *Ein weiteres populäres Mehrschrittverfahren ist das* Leapfrog-Verfahren*, ein einfaches Zweischritt-Verfahren.*

Motiviert wird es wieder durch den Fundamentalsatz der Differential- und Integralrechnung, allerdings mit im Vergleich zum Adams-Bashforth-Verfahren etwas anders gewählten Integrationsintervallen:

$$y(t_{i+2}) = y(t_i) + \int_{t_i}^{t_{i+2}} y'(s)\,ds = y(t_i) + \int_{t_i}^{t_{i+2}} f(s,y(s))\,ds.$$

Wir approximieren das Integral mit der Mittelpunktregel, also durch

$$\int_{t_i}^{t_{i+2}} f(s,y(s))\,ds \approx 2hf(t_{i+1},y(t_{i+1}))$$

und erhalten so das Näherungsverfahren

$$y(t_{i+2}) \approx y(t_i) + 2hf(t_{i+1},y(t_{i+1})).$$

Wir setzen wieder $g(s) := f(s,y(s)) = y'(s)$ und schätzen den Approximationsfehler des Integrals mit Hilfe der Taylor-Entwicklung um t_{i+1} durch

$$\begin{aligned}
\left\| \int_{t_i}^{t_{i+2}} g(s) - g(t_{i+1})\,ds \right\| &= \left\| \int_{t_i}^{t_{i+2}} g(s) - g(t_{i+1}) - (s-t_{i+1})g'(t_{i+1})\,ds \right\| \\
&\le \int_{t_i}^{t_{i+2}} \|g(s) - g(t_{i+1}) - (s-t_{i+1})g'(t_{i+1})\|\,ds \\
&\le 2h \sup\{\|g(s) - g(t_{i+1}) - (s-t_{i+1})g'(t_{i+1})\| \ : \ s \in [t_i, t_{i+2}]\} \\
&\le 2h\frac{h^2}{2} \sup\{\|g''(s_+)\| \ : \ s_+ \in [t_i, t_{i+2}]\} \\
&= h^3 \sup\{\|g''(s_+)\| \ : \ s_+ \in [t_i, t_{i+2}]\}
\end{aligned}$$

ab, können also feststellen, dass auch das Leapfrog-Verfahren eine Konsistenzordnung von 2 aufweist.

Da das charakteristische Polynom des Verfahrens durch $\varrho(z) = z^2 - 1$ gegeben ist, also nur die einfachen Nullstellen 1 und -1 besitzt, ist das Leapfrog-Verfahren nicht nur konsistent, sondern auch stabil.

8 Randwertaufgaben

Bisher haben wir uns auf typische Anfangswertprobleme konzentriert: Der Ausgangszustand eines Systems ist vollständig beschrieben, und wir wollen das Verhalten des Systems in der Zukunft vorhersagen. In diesem Kapitel beschäftigen wir uns mit einer anderen Klasse von Problemen, bei der nicht nur der Ausgangs-, sondern auch der Endzustand vorgegeben sind, bei denen also Bedingungen an *beiden* Rändern des Zeitintervalls $[a, b]$ vorkommen.

8.1 Beispiele

Ein einfaches Beispiel stammt aus der Ballistik: Wenn wir mit einer Kanonenkugel einen bestimmten Punkt treffen wollen, stehen sowohl der Ausgangspunkt der Kugel fest als auch der Endpunkt. Damit wird aus dem Anfangswertproblem (2.1) das Randwertproblem

$$y(a) = y_a, \qquad y(b) = y_b, \qquad y'(t) = f(t, y(t)) \qquad \text{für alle } t \in [a, b], \tag{8.1}$$

wobei $y_a \in V$ und $y_b \in V$ den Anfangs- und Endpunkt beschreiben. Allgemeiner kann man lineare Randbedingungen

$$Ay(a) + By(b) = c, \qquad y'(t) = f(t, y(t)) \qquad \text{für alle } t \in [a, b] \tag{8.2}$$

zulassen, wobei A und B lineare Operatoren mit demselben Bildraum und c ein Vektor aus diesem Bildraum sind. Am allgemeinsten sind nichtlineare Randbedingungen

$$r(y(a), y(b)) = 0, \qquad y'(t) = f(t, y(t)) \qquad \text{für alle } t \in [a, b] \tag{8.3}$$

mit einer beliebigen Funktion r, die von den Werten der Lösung im linken und rechten Randpunkt abhängt.

Beispiel 8.1 (Kanonenschuss) *Betrachten wir das Beispiel des Kanonenschusses etwas genauer. Wenn wir die Position der Kanonenkugel durch eine Funktion $k : \mathbb{R}_{\geq 0} \to \mathbb{R}^2$ darstellen, erhalten wir aus den Newton'schen Bewegungsaxiomen die Gleichungen*

$$k(0) = 0, \qquad k'(0) = \begin{pmatrix} v_0 \cos\alpha \\ v_0 \sin\alpha \end{pmatrix}, \qquad k''(t) = \begin{pmatrix} 0 \\ -\gamma \end{pmatrix},$$

wobei v_0 die Startgeschwindigkeit, α der Abschusswinkel und γ die geeignet skalierte Beschleunigung durch die Erdanziehung ist.

Wir fassen k und k' mit dem Parameter α zu einem Vektor zusammen, um

$$y(t) := \begin{pmatrix} k(t) \\ k'(t) \\ \alpha \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

mit den Gleichungen

$$y(0) = \begin{pmatrix} 0 \\ 0 \\ v_0 \cos\alpha \\ v_0 \sin\alpha \\ \alpha \end{pmatrix}, \qquad y'(t) = \begin{pmatrix} y_3(t) \\ y_4(t) \\ 0 \\ -\gamma \\ 0 \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

zu erhalten. Weil wir wollen, dass die Kanonenkugel in einer Entfernung $z \in \mathbb{R}_{\geq 0}$ vom Nullpunkt wieder den Boden berührt, müssen wir eine geeignete Randbedingung einführen. Da uns die Zeit, die die Kugel für ihren Flug benötigt, nicht wirklich interessiert, ihre Entfernung von der Kanone (und damit die Nähe zum Ziel) dagegen sehr, eliminieren wir die Zeit, indem wir stattdessen die Entfernung als Variable einführen. Wegen der ersten und dritten Spalte der definierenden Gleichungen ist die Entfernung d von der Kanone durch

$$d(t) = ct, \qquad c := v_0 \cos\alpha \qquad \text{für alle } t \in \mathbb{R}_{\geq 0}$$

gegeben, also können wir $t = d/c$ substituieren und erhalten

$$\hat{y}(0) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ \tan(\alpha) \\ \alpha \end{pmatrix}, \qquad \hat{y}'(d) = \begin{pmatrix} \hat{y}_3(d) \\ \hat{y}_4(d) \\ 0 \\ -\gamma/c^2 \\ 0 \end{pmatrix} \qquad \text{für alle } d \in \mathbb{R}_{\geq 0},$$

und die Bedingung, dass die Kugel in der Entfernung $d = z$ wieder den Boden erreichen soll, kann einfach durch die Bedingung

$$\hat{y}_2(z) = 0$$

in die Gleichung eingefügt werden. Wir eliminieren die überflüssigen ersten und dritten Komponenten und erhalten das Randwertproblem

$$\check{y}_1(0) = \check{y}_1(z) = 0, \qquad \check{y}_2(0) = \tan\check{y}_3(0), \qquad \check{y}'(d) = \begin{pmatrix} \check{y}_2(d) \\ -\gamma/(v_0 \cos\check{y}_3(d))^2 \\ 0 \end{pmatrix}$$

für alle $d \in [0, z]$. Wegen des Auftretens des Tangens in der Randbedingung und des Cosinus in der rechten Seite ist dieses Problem nichtlinear.

Die Lösbarkeit von Randwertproblemen lässt sich nicht so einfach wie im Falle von Anfangswertproblemen analysieren. Schon im Falle des Kanonenschusses lässt sich die Bahnkurve zwar beliebig fortsetzen, aber es ist nicht möglich, Ziele zu treffen, die zu weit von der Kanone entfernt sind. Für große Werte von z existiert also keine Lösung des Randwertproblems.

Beispiel 8.2 (Trigonometrisch) *Wir untersuchen die Gleichung*

$$y''(t) = -y(t) \qquad \text{für alle } t \in \mathbb{R}.$$

Es lässt sich einfach nachrechnen, dass y die Form

$$y(t) = \alpha \sin(t) + \beta \cos(t) \qquad \text{für alle } t \in \mathbb{R}$$

besitzen muss. Je nachdem, welche Art von Randbedingungen wir stellen, ändern sich Existenz und Eindeutigkeit der Lösungen:

- $y(0) = y(\pi/2) = 0$*: Aus* $y(0) = 0$ *folgt* $\beta = 0$*, aus* $y(\pi/2) = 0$ *folgt* $\alpha = 0$*, also ist* $y \equiv 0$ *die eindeutig bestimmte Lösung des Randwertproblems.*
- $y(0) = y(\pi) = 0$*: Wir haben wieder* $\beta = 0$*, können aber* α *beliebig wählen. Das Randwertproblem ist also lösbar, aber die Lösung ist nicht eindeutig.*
- $y(0) = y(\pi) = \epsilon$ *mit* $\epsilon > 0$*: Aus* $y(0) = \epsilon$ *folgt* $\beta = \epsilon$*. Um die Bedingung* $y(\pi) = \epsilon$ *erfüllen zu können, müsste dann* $\alpha \sin(\pi) = 2\epsilon$ *gelten, und das ist wegen* $\sin(\pi) = 0$ *ausgeschlossen. Also besitzt dieses Randwertproblem keine Lösung.*

Wie man sieht können also auch kleine Änderungen an den Randbedingungen einen erheblichen Einfluss auf die Lösbarkeit von Randwertproblemen haben.

8.2 Einfache Schießverfahren

Eine einfache Methode, um dafür zu sorgen, dass Kanonenkugeln ihr Ziel erreichen, besteht im sogenannten *Dreipunktschießen*: Angenommen, ein erster Schuss ging zu weit und ein zweiter zu kurz. Dann stellt man den Winkel für den dritten Schuss so ein, dass er zwischen den Winkeln der ersten beiden liegt, und der dritte Schuss deshalb zwischen die ersten beiden treffen wird.

Mathematisch ist dieser Ansatz ein schlichtes Bisektionsverfahren: Wir wissen, dass der optimale Winkel in einem Intervall liegt, und wir unterteilen das Intervall, um uns ihm anzunähern. Offenbar ist es theoretisch möglich, beliebig nahe an das gewünschte Ziel heran zu kommen, sofern genügend viel Munition und Zeit zur Verfügung stehen.

Dieser Ansatz lässt sich auf allgemeine Randwertprobleme ausweiten: Unser Ziel ist es, Anfangswerte $y_0 \in V$ so zu finden, dass die zugehörige Lösung $y(t; a, y_0)$ die Randbedingung erfüllt, dass also

$$r(y(a; a, y_0), y(b; a, y_0)) = r(y_0, y(b; a, y_0)) = 0$$

gilt. Sobald y_0 gefunden ist, können wir die Lösung $y(t; a, y_0)$ an beliebigen Punkten des Intervall $[a, b]$ auswerten.

Wir stehen also vor der Aufgabe, eine Nullstelle der Funktion

$$F(y_0) := r(y_0, y(b; a, y_0))$$

zu finden. Falls y_0 einem eindimensionalen Raum entstammt, lässt sich das Problem mit einem Bisektionsverfahren behandeln: Wir approximieren $y(b; a, y_0)$ durch eine diskrete Näherung $\eta(b; a, y_0)$, werten

$$\widetilde{F}(y_0) := r(y_0, \eta(b; a, y_0))$$

aus und passen dann y_0 geeignet an. Die diskrete Näherung $\eta(b; a, y_0)$ kann mit Hilfe der in den vorangehenden Kapiteln vorgestellten Verfahren effizient und hinreichend genau bestimmt werden, sofern das Anfangswertproblem gut konditioniert ist.

Im allgemeinen Fall bietet es sich an, das Nullstellenproblem $F(y_0) = 0$ mit Hilfe eines Newton-Verfahrens zu lösen. Dazu benötigen wir neben der Auswertung von F auch eine Möglichkeit, die Jacobi-Matrix F' zu berechnen. Natürlich müssen beide Operationen in der Praxis durch Näherungen ersetzt werden: F können wir durch $\widetilde{F}$ ersetzen, F' kann gemäß

$$\partial_p F(y_0) \approx \frac{F(y_0 + p) - F(y_0)}{\|p\|} \qquad \text{oder} \qquad \partial_p F(y_0) \approx \frac{F(y_0 + p) - F(y_0 - p)}{2\|p\|}$$

durch Differenzenquotienten approximiert werden. Für jeden Richtungsvektor p erfordert die Auswertung dieses Quotienten die Berechnung von Hilfslösungen $\eta(b; a, y_0 + p)$ (und gegebenenfalls auch $\eta(b; a, y_0 - p)$), und um die vollständige Matrix F' anzunähern sind $\dim V$ Differenzenquotienten zu berechnen. Offenbar kann die Durchführung des Newton-Verfahrens relativ aufwendig werden.

Der Einsatz von Differenzenquotienten kann zu numerischen Instabilitäten führen: Falls $\|p\|$ groß ist, ist der Differenzenquotient im Allgemeinen nur eine schlechte Approximation der Ableitung. Falls $\|p\|$ klein ist, kann bei der Berechnung von $y_0 + p$ und der Differenz $F(y_0 + p) - F(y_0)$ Auslöschung auftreten, so dass wieder nur eine schlechte Approximationsgüte erzielt wird. In der Praxis kann es deshalb sinnvoll sein, die Qualität der Approximation mit Hilfe von Extrapolationsverfahren zu verbessern. Der Preis für die Verbesserung der Genauigkeit ist dann die Notwendigkeit, F für weitere Argumente auswerten zu müssen. Da jede Auswertung das Lösen eines Anfangswertproblems erfordert, erhöht sich der Rechenaufwand wesentlich.

Alternativ lässt sich die Matrix F' direkt als Lösung eines matrixwertigen Anfangswertproblems darstellen, auf dessen rechter Seite die Jacobi-Matrix $\partial_x f$ der rechten Seite des Randwertproblems auftritt. Falls diese Jacobi-Matrix exakt zur Verfügung steht, kann F' mit den Verfahren aus den vorangehenden Kapiteln direkt berechnet werden, ansonsten muss $\partial_x f$ wieder durch Differenzenquotienten und Extrapolation angenähert werden („internes Differenzieren“ im Gegensatz zum „externen Differenzieren“, bei dem direkt F' durch Differenzenquotienten angenähert wird).

8.3 Mehrzielverfahren

Wir stoßen bei einfachen Schießverfahren auf Schwierigkeiten, falls die bei ihrer Durchführung auftretenden Anfangswertprobleme schlecht konditioniert sind.

Beispiel 8.3 (Definitionsbereich) *Wir untersuchen die Anfangswertprobleme*

$$y_\xi(0) = \xi, \qquad y'_\xi(t) = y_\xi(t)^2 \qquad \text{für alle } t \in [0, 1/\xi)$$

für Parameter $\xi \in \mathbb{R}_{>0}$. Für jedes $\xi \in \mathbb{R}_{>0}$ ist die Lösung durch

$$y_\xi(t) = \frac{\xi}{1 - \xi t} \qquad \text{für alle } t \in [0, 1/\xi)$$

gegeben, und da sie offenbar eine Singularität bei $1/\xi$ besitzt, kann sie nicht beliebig weit fortgesetzt werden.

Das bedeutet, dass wir bei einem Schießverfahren nicht beliebige Anfangswerte zulassen dürfen, sondern nur solche, bei denen der Endpunkt b noch im Definitionsbereich $[0, 1/\xi)$ liegt, bei denen also $b\xi < 1$ gilt.

Falls die Randbedingungen zu einer Lösung gehören, für die der Anfangswert ξ sehr nahe bei $1/b$ liegt, kann das Berechnen einer Näherung von y_ξ sehr schlecht konditioniert sein und zum Scheitern eines einfachen Schießverfahrens führen.

Das Problem ist der zu große Definitionsbereich: Auch wenn die Lösung des Randwertproblems im Prinzip gutartig ist, kann durch die Reduktion auf Anfangswerte ein schlecht konditioniertes Problem entstehen, weil Fehler in den Anfangswerten entsprechend Lemma 2.5 bzw. Lemma 3.1 exponentiell verstärkt werden.

Ein naheliegender Lösungsansatz besteht darin, die Definitionsintervall künstlich zu verkleinern: Statt das Intervall $[a, b]$ insgesamt zu betrachten, wählen wir Zwischenpunkte $a = t_0 < t_1 < \ldots < t_m = b$ und untersuchen Lösungen von m Anfangswertproblemen zu den Startzeitpunkten $t_0, \ldots, t_{m-1}$: Für jedes t_i geben wir einen Anfangswert y_i vor, der dann eine Lösung $y(t; t_i, y_i)$ für $t \in [t_i, t_{i+1}]$ definiert, sofern das Intervall $[t_i, t_{i+1}]$ nicht zu groß ist.

Die durch

$$\tilde{y}(t; y_0, \ldots, y_{m-1}) := y(t; t_i, y_i) \qquad \text{für } i \in \{0, \ldots, m-1\}, t \in [t_i, t_{i+1})$$

definierte zusammengesetzte Lösung erfüllt die ursprüngliche Differentialgleichung nur innerhalb jedes Intervalls, zwischen Intervallen kann es zu Sprüngen kommen, die wir durch zusätzliche Bedingungen

$$y(t_{i+1}; t_i, y_i) = y_{i+1} \qquad \text{für alle } i \in \{0, \ldots, m-1\}$$

ausschließen müssen.

Statt also nur einen Startwert y_0 vorzugeben und so wählen zu müssen, dass die Randbedingung erfüllt wird, stehen uns nun m Startwerte (wir können immer $y_m :=$

$y(t_m; t_{m-1}, y_{m-1})$ setzen) zur Verfügung, die durch $m-1$ zusätzliche Bedingungen aneinander gekoppelt sind.

Diese zusätzlichen Bedingungen können wir in einer entsprechend verallgemeinerten Randbedingung r_m verstecken: Falls wir mit

$$r_m(y_0, \ldots, y_{m-1}) := \begin{pmatrix} y(t_1; t_0, y_0) - y_1 \\ \vdots \\ y(t_{m-1}; t_{m-2}, y_{m-2}) - y_{m-1} \\ r(y_0, y(t_m; t_{m-1}, y_{m-1})) \end{pmatrix}$$

das Nullstellenproblem $r_m(y_0, \ldots, y_{m-1}) = 0$ lösen, ist die zusammengesetzte Funktion stetig (und damit auch Lösung der Differentialgleichung) und erfüllt außerdem auch die Randbedingung.

Dieses erweiterte Nullstellenproblem lässt sich mit ähnlichen Techniken wie im Falle des einfachen Schießverfahrens behandeln: Um das Newton-Verfahren durchführen zu können, ersetzen wir die exakten Lösungen $y(t_{i+1}; t_i, y_i)$ durch Näherungslösungen $\eta(t_{i+1}; t_i, y_i)$ und approximieren die Jacobi-Matrix mit Hilfe geeigneter Differenzenquotienten oder Extrapolationstechniken.

Den resultierenden Algorithmus bezeichnet man als *Mehrzielverfahren*, weil simultan mehrere Lösungen zu mehreren Startwerten berechnet werden, die mehrere Endwerte treffen sollen.

Durch das Hinzufügen weiterer Zwischenpunkte kann man die Stabilität eines derartigen Verfahrens verbessern (denn damit reduziert sich der Einfluss von Anfangsfehlern), aber natürlich führt diese Maßnahme auch zu einer deutlichen Erhöhung des Rechenaufwands für die Durchführung des Newton-Verfahrens. Es ist allerdings möglich, die spezielle Struktur von r_m auszunutzen, um die Durchführung des Newton-Verfahrens effizienter zu gestalten.

8.4 Globale Diskretisierungsverfahren

Die Idee der Mehrzielverfahren lässt sich weiter treiben: Statt nur eine kleinen Anzahl m von Zwischenpunkten zu verwenden, um Stabilitätsprobleme zu vermeiden, können wir auch gleich die gesamte Lösung y auf einem Gitter $a = t_0 < t_1 < \ldots < t_n = b$ approximieren und Näherungswerte η_i für $y(t_i)$ simultan zu berechnen versuchen.

Ein einfacher Algorithmus wird beispielsweise von den *Differenzenverfahren* zur Verfügung gestellt: Als Beispiel betrachten wir das Randwertproblem

$$y(a) = y_a, \qquad y(b) = y_b, \qquad -y''(t) + q(t, y(t)) = 0 \qquad \text{für alle } t \in [a, b]$$

mit einer differenzierbaren Funktion q. Der Einfachheit halber nehmen wir an, dass $t_i - t_{i-1} = h$ für alle $i \in \{1, \ldots, n\}$ mit einer Schrittweite $h \in \mathbb{R}_{>0}$ gilt.

Aus der Taylor-Entwicklung von y folgen

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y'''(t) + \frac{h^4}{24}y^{(4)}(t_+),$$

$$y(t-h) = y(t) - hy'(t) + \frac{h^2}{2}y''(t) - \frac{h^3}{6}y'''(t) + \frac{h^4}{24}y^{(4)}(t_-)$$

mit Zwischenpunkten $t_+ \in [t, t+h]$ und $t_- \in [t-h, t]$. Addition dieser Gleichungen und Division durch h^2 führt zu

$$\frac{y(t+h) - 2y(t) + y(t-h)}{h^2} = y''(t) + \frac{h^2}{24}(y^{(4)}(t_+) + y^{(4)}(t_-)),$$

der Differenzenquotient auf der linken Seite wird also wir h^2 gegen die zweite Ableitung von y konvergieren, wenn wir die Schrittweite gegen Null gehen lassen.

Indem wir die exakte zweite Ableitung durch diesen Differenzenquotienten ersetzen, wird aus dem kontinuierlichen Randwertproblem das diskrete Problem

$$\eta_0 = y_a, \quad \eta_n = y_b, \quad \frac{2\eta_i - \eta_{i-1} - \eta_{i+1}}{h^2} + q(t_i, \eta_i) = 0 \qquad \text{für alle } i \in \{1, \ldots, n-1\}.$$

Unter geeigneten Voraussetzungen lässt sich beweisen, dass die Näherungswerte η_i gegen $y(t_i)$ konvergieren und der Fehler sich wie h^2 verhält.

Zur Berechnung des Vektors $(\eta_i)_{i=0}^n$ können, wie schon im Falle der Mehrzielverfahren, wieder Newton-Verfahren zum Einsatz kommen. Das approximative Lösen von zwischengeschalteten Anfangswertproblemen ist bei diesem Ansatz nicht mehr erforderlich.

Literaturverzeichnis

[1] F. Bornemann and P. Deuflhard. *Numerische Mathematik II.* de Gruyter, 2002.

[2] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler.* Springer, 2006.

[3] Wikipedia. Grönwallsche Ungleichung — Wikipedia, Die freie Enzyklopädie, 2007. [Online; Stand 22. Oktober 2007].

[4] Wikipedia. Satz von Picard-Lindelöf — Wikipedia, Die freie Enzyklopädie, 2007. [Online; Stand 22. Oktober 2007].