NUMERIK GEWÖHNLICHER DIFFERENTIALGLEICHUNGEN

S. VOLKWEIN

Date: 3. November 2010.

Das Manuscript entstand während einer zweistündigen Vorlesung im Sommersemester 2010 am Fachbereich Mathematik und Statistik der Universität Konstanz. Ein besonderer Dank gilt Frau R. Mancini und Herrn O. Lass für die Unterstützung bei der Aufstellung der Übungsbeispiele und für die Korrekturhilfe. Ich möchte mich ferner bei Frau K. Borgmeyer für die Hinweise auf Fehler bedanken.

Inhaltsverzeichnis

1. Anfangswertprobleme

In diesem Abschnitt wollen wir uns der numerischen Lösung von Anfangswertaufgaben widmen. Diese Probleme werden oft auch *dynamische Systeme* genannt und spielen zum Beispiel in der Mechanik oder Biologie eine wichtige Rolle.

Der Abschnitt beruht wesentlich auf dem elften Kapitel im Buch [1] sowie auf Ausschnitten aus dem Buch [3].

1.1. **Einleitung.** Die Problemstellung lautet wie folgt: Gesucht wird eine Funktion $y = y(t)$ einer (Zeit-)Variablen t, die der Gleichung

$$y'(t) = f(t, y(t)), \quad t \in (t_\circ, T], \tag{1.1a}$$

und der Anfangsbedingung

$$y(t_\circ) = y^\circ \tag{1.1b}$$

genügen soll. Wir nennen (1.1a) eine *gewöhnliche Differentialgleichung erster Ordnung*. Das Problem (1.1) heißt *Anfangswertproblem*.

Beispiel 1.1. Wir betrachten das Anfangswertproblem

$$y'(t) = 2ty^2(t) \text{ für } t > 0, \quad y(0) = 1.$$

Offenbar ist $f(t, y) = 2ty^2$ in t und y stetig partiell differenzierbar, also insbesondere lokal Lipschitz-stetig in y. Nach dem Satz von Picard-Lindelöf (siehe Satz 1.4) existiert eine eindeutige Lösung lokal um $t_\circ = 0$. Durch Einsetzen sehen wir, dass $y(t) = 1/(1 - t^2)$ die Lösung des Anfangswertproblems ist. Offensichtlich existiert die Lösung aber nur für $t \in [0, 1)$. ◊

Allgemeiner hat man es mit Systemen von n gewöhnlichen Differentialgleichungen erster Ordnung

$$\begin{aligned} y_1'(t) &= f_1(t, y_1(t), \dots, y_n(t)), && t \in (t_\circ, T], \\ &\vdots \qquad\qquad \vdots \\ y_n'(t) &= f_n(t, y_1(t), \dots, y_n(t)), && t \in (t_\circ, T], \end{aligned}$$

zu tun. Anfangsbedingungen dazu sind

$$y_i(t_\circ) = y_i^\circ, \quad 1 \le i \le n.$$

Setzt man

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} : [t_\circ, T] \to \mathbb{R}^n, \; f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} : [t_\circ, T] \times \mathbb{R}^n \to \mathbb{R}^n, \; y^\circ = \begin{pmatrix} y_1^\circ \\ \vdots \\ y_n^\circ \end{pmatrix} \in \mathbb{R}^n,$$

so können wir das Anfangswertproblem kompakt in der Standardform

$$y'(t) = f(t, y(t)), \quad t \in (t_\circ, T], \quad y(t_\circ) = y^\circ \tag{1.2}$$

schreiben.

Beispiel 1.2. Die Entwicklung der Temperatur $T = T(x, t)$ an der Stelle x eines Stabes zur Zeit t ergibt sich als Lösung der *Wärmeleitungsgleichung*

$$\frac{\partial T}{\partial t}(x, t) = \kappa \frac{\partial^2 T}{\partial x^2}(x, t), \quad t > 0, \; x \in (0, \ell).$$

Die Anfangsbedingung lautet $T(x,0) = \phi(x)$ für $x \in (0,\ell)$. Die Randwerte sind $T(0,t) = T(\ell,t) = 0$ (homogene Dirichlet-Randbedingungen). Mit Hilfe der so genannten *Linienmethode* kann man die gesuchte Lösung $T(x,t)$ mit Hilfe eines Systems gewöhnlicher Differentialgleichungen erster Ordnung annähern. Dazu sei für $n_x \in \mathbb{N}$ die Schrittweite in x-Richtung durch $h_x = \ell/n_x$ gegeben. Wir approximieren die zweite partielle Ableitung nach x durch einen zentralen Differenzenquotienten:

$$\kappa \frac{\partial^2 T}{\partial x^2}(x,t) \approx \frac{\kappa}{h_x^2}\big(T(x+h_x,t) - 2T(x,t) + T(x-h_x,t)\big), \quad x \in [h_x, \ell - h_x],\ t > 0.$$

Wir bezeichnen mit $y_i(t)$, $1 \le i \le n_x - 1$, die numerischen Näherungen für $T(x_i,t)$, wobei $x_i = ih_x$ für $i = 0,\ldots,n_x$ gilt. Aufgrund der Randbedingungen ist die Temperatur an $x_0 = 0$ und $x_{n_x} = \ell$ bekannt. Daher genügt es, Näherungen für die Temperatur T an den inneren Gitterpunkten x_i, $i = 1,\ldots,n_x-1$, zu berechnen. Die Bestimmungsgleichungen ergeben sich aus der obigen Differenzenapproximation für die zweite Ableitung nach x wie folgt:

$$y_i(t) = \frac{\kappa}{h_x^2}\big(y_{i+1}(t) - 2y_i(t) + y_{i-1}(t)\big), \quad 1 \le i \le n_x - 1.$$

Setzen wir $n = n_x - 1$ und

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} : [0,\infty) \to \mathbb{R}^n, \quad y^\circ = \begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_{n_x-1}) \end{pmatrix} \in \mathbb{R}^n$$

$$A = \frac{\kappa}{h_x^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

so erhalten wir das (lineare) Anfangswertproblem

$$y'(t) = Ay(t), \quad t > 0, \quad y(0) = y^\circ.$$

Der Begriff Linienmethode reflektiert hier die Bestimmung der Funktionen $y(t)$ entlang der zur Zeitachse parallelen Linien durch die Ortsgitterpunkte. Mit Hilfe numerischer Verfahren zur Lösung des Anfangswertproblems — also über eine nachfolgende Diskretisierung der Zeit — erhalten wir insgesamt ein numerisches Verfahren zur Behandlung partieller Differentialgleichungen vom Typ der Wärmeleitungsgleichung (ein Beispiel einer sogenannten parabolischen Differentialgleichung). ◊

Übungsaufgabe 1. Wir betrachten die positiv-definite, symmetrische Matrix

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Zeigen Sie, dass die Eigenwerte und die dazugehörenden Eigenvektoren von A durch

$$\lambda_k = 4\sin^2\left(\frac{k\pi}{2(n+1)}\right) \quad k = 1,\ldots,n$$

und

$$v_k = \left(\sin\left(\frac{k\pi}{n+1}\right), \sin\left(\frac{2k\pi}{n+1}\right), \ldots, \sin\left(\frac{nk\pi}{n+1}\right)\right)^T \in \mathbb{R}^n.$$

gegeben sind.

Gewöhnliche Differentialgleichungen n-ter Ordnung sind Gleichungen, in denen Ableitungen der gesuchten Funktion bis zur n-ten Ordnung auftreten.

Beispiel 1.3. Das mathematische Pendel wird durch die gewöhnliche Differentialgleichung zweiter Ordnung

$$\varphi''(t) = -\frac{g}{\ell}\sin(\varphi(t)), \quad t > 0,$$

beschrieben. Hierbei bezeichnen $g = 9.80665\, m/s^2$ die Fallbeschleunigung und $\ell > 0$ die Pendellänge. Anfangsbedingungen lauten $\varphi(0) = \varphi_\circ$ und $\varphi'(0) = 0$ mit einer Anfangsauslenkung $\varphi_\circ \in \mathbb{R}$. ◊

Die (explizite) *Anfangswertaufgabe n-ter Ordnung* lautet wie folgt: Bestimme eine skalare Funktion $y = y(t)$, so dass

$$y^{(n)}(t) = g(t, y(t), y'(t), \ldots, y^{(n-1)}(t)), \quad t \in (t_\circ, T], \tag{1.3a}$$

$$y(t_\circ) = z_0, \ldots, y^{(n-1)}(t_\circ) = z_{n-1} \tag{1.3b}$$

mit einer gegebenen Funktion $g : [t_\circ, T] \times \mathbb{R}^m \to \mathbb{R}$ und mit Anfangswerten z_i, $0 \le i \le n-1$. Das Anfangswertproblem (1.3) kann man als ein System von n Differentialgleichungen erster Odnung mit entsprechenden Anfangsbedingungen umformulieren. Es reicht deshalb, numerische Verfahren für die Standardform (1.2) zu entwickeln. Setzen wir nämlich

$$\begin{aligned} y_1(t) &= y(t), \\ y_2(t) &= y'(t) = y_1'(t), \\ &\vdots \\ y_n(t) &= y^{(n-1)}(t) = y_{n-1}'(t), \end{aligned}$$

so folgt aus der Gleichung (1.3) die Beziehung

$$y_n'(t) = g(t, y_1(t), \ldots, y_n(t)).$$

Also erhalten wir das Differentialgleichungssytem erster Ordnung

$$y'(t) = \begin{pmatrix} y_1'(t) \\ \vdots \\ y_n'(t) \end{pmatrix} = \begin{pmatrix} y_2(t) \\ \vdots \\ y_n(t) \\ g(t, y_1(t), \ldots, y_n(t)) \end{pmatrix}, \quad t \in (t_\circ, T],$$

mit der Anfangsbedingung

$$y(t_\circ) = \begin{pmatrix} z_0 \\ \vdots \\ z_{n-1} \end{pmatrix} =: y^\circ.$$

1.2. **Einige theoretische Grundlagen.** Ein Problem heißt *korrekt gestellt*, falls

1) eine Lösung existiert,
2) diese Lösung eindeutig ist und
3) stetig von den Daten abhängt.

Aus der Theorie gewöhnlicher Differentialgleichungen ist folgendes Resultat bekannt.

Satz 1.4 (Picard-Lindelöf)**.** *Sei $f : [t_\circ, T] \times \mathbb{R}^n \to \mathbb{R}^n$, $(t, y) \mapsto f(t, y)$ eine Funktion, die stetig in (t, y) ist und darüber hinaus im folgenden Sinne Lipschitz-stetig in y ist:*

$$\|f(t, y) - f(t, z)\| \le L \|y - z\|$$

für alle $t \in [t_\circ, t_\circ + \varepsilon]$ mit $\varepsilon > 0$ und für alle $y, z \in \mathbb{R}^n$ in einer Umgebung $U_\circ \subset \mathbb{R}^n$ vom Anfangswert y°, wobei $\|\cdot\|$ eine beliebige Norm auf $\mathbb{R}^n$ bezeichnet. Dann existiert eine eindeutige Lösung y von (1.2) *in einer Umgebung von $t_\circ$. Diese Umgebung hängt von ε, f und $U_\circ$ ab.*

Beispiel 1.5. Beim mathematischen Pendel aus Beispiel 1.3 ergibt sich das System

$$y'(t) = \begin{pmatrix} y_1'(t) \\ y_2'(t) \end{pmatrix} = \begin{pmatrix} y_2(t) \\ -\frac{g}{\ell} \sin(y_1(t)) \end{pmatrix} =: f(t, y(t)), \quad t \in (t_\circ, T],$$

mit der Anfangsbedingung

$$y(0) = y^\circ = \begin{pmatrix} \phi_\circ \\ 0 \end{pmatrix} =: y^\circ.$$

Wir wählen als Norm in $\mathbb{R}^2$ die Maximumnorm $\|\cdot\|_\infty$. Für das mathematische Pendel ergibt sich mit $c = -g/\ell$ die Abschätzung

$$\begin{aligned} \|f(t, y) - f(t, z)\|_\infty &= \left\| \begin{pmatrix} y_2 - z_2 \\ c \sin y_1 - c \sin z_1 \end{pmatrix} \right\|_\infty = \left\| \begin{pmatrix} y_2 - z_2 \\ c \cos \xi (y_1 - z_1) \end{pmatrix} \right\|_\infty \\ &= \max \big\{ |y_2 - z_2|, |c| \, |\cos \xi| \, |y_1 - z_1| \big\} \le \max \big\{ 1, |c| \big\} \|y - z\|_\infty, \end{aligned}$$

also gilt die Lipschitz-Bedingung mit $L = \max\{1, g/\ell\}$ für alle $t \in [t_\circ, T]$ und $y = (y_1, y_2)$, $z = (z_1, z_2) \in \mathbb{R}^2$. ◊

Die Forderung 3) bei korrekt gestellten Problemen ist natürlich für numerische Zwecke besonders essentiell. Im vorliegenden Zusammenhang betrachten wir den einfachsten Fall, dass unter "Daten" lediglich die Anfangswerte y° verstanden werden.

Satz 1.6. *Die Funktion f erfülle die Voraussetzungen aus Satz* 1.4*. Seien y und z Lösungen zu* (1.2) *mit den Anfangsdaten y° beziehungsweise z°. Dann gilt für alle t aus einer Umgebung von $t_\circ$ die Abschätzung*

$$\|y(t) - z(t)\| \le e^{L|t - t_\circ|} \|y^\circ - z^\circ\|.$$

Bemerkung 1.7. Wir können den Begriff der Korrektheit in einen Konditionsrahmen einfügen. Die eindeutige Lösbarkeit bedeutet, dass die Zuordnung $S : \mathbb{R}^n \to C^1([t_\circ, T]; \mathbb{R}^n)$, $y^\circ \mapsto S(y^\circ)(t) := y(t)$ zumindest in einer Umgebung von $t_\circ$ eine wohldefinierte Abbildung ist. Die Lösung des Anfangswertproblems ist dann gerade die Auswertung des Lösungsoperators an der Stelle y°. Die Abschätzung aus Satz 1.6 quantifiziert dann die absolute Kondition des Lösungsoperators S, das heißt, des Anfangswertproblems, bezüglich Störungen in den Anfangsdaten. Für

längere Integrationsintervalle wird die Abschätzung allerdings exponentiell schlechter. ◊

Beispiel 1.8 (Skalarer Fall $n = 1$)**.** Es gelte

$$y'(t) = Ly(t),\ t \in (t_\circ, T],\ y(t_\circ) = y^\circ, \quad z'(t) = Lz(t),\ t \in (t_\circ, T],\ z(t_\circ) = z^\circ$$

mit $L > 0$, wobei sich die Lösungen für $t \in [t_\circ, T]$ als

$$y(t) = e^{L(t-t_\circ)}y^\circ, \quad z(t) = e^{L(t-t_\circ)}z^\circ$$

exakt angeben lassen. Daher erhalten wir

$$y(t) - z(t) = e^{L(t-t_\circ)}(y^\circ - z^\circ), \quad t \in [t_\circ, T],$$

das heißt, wir haben "=" in der Abschätzung von Satz 1.6. Wegen

$$\left|\frac{y(t) - z(t)}{y(t)}\right| = \frac{e^{L|t-t_\circ|}|y^\circ - z^\circ|}{e^{L|t-t_\circ|}|y^\circ|} = \left|\frac{y^\circ - z^\circ}{y^\circ}\right|, \quad t \in [t_\circ, T],$$

ist in diesem Fall die relative Kondition für alle Werte von y° und L gut, während die absolute Kondition für $L \gg 1$ schlecht ist. ◊

Im allgemeinen Fall erhalten wir unter den Voraussetzungen wie in Satz 1.6 für die relative Kondition die Abschätzung

$$\frac{\|y(t) - z(t)\|}{\|y(t)\|} = \frac{\|y^\circ\|}{\|y(t)\|}e^{L|t-t_\circ|}\frac{\|y^\circ - z^\circ\|}{\|y^\circ\|} = \kappa_{\mathrm{rel}}(t)\frac{\|y^\circ - z^\circ\|}{\|y^\circ\|}$$

mit $\kappa_{\mathrm{rel}}(t) = e^{L|t-t_\circ|}\|y^\circ\|/\|y(t)\|$, wobei $\|y(t)\| \neq 0$ und $\|y^\circ\| \neq 0$ vorausgesetzt seien. Die relative Konditionszahl drückt also das Verhältnis zwischen dem Wachstum der Lösung, $\|y(t)\|/\|y^\circ\|$, und dem Faktor $e^{L|t-t_\circ|}$ aus.

1.3. **Einfache Einschrittverfahren.** Wir verwenden $y^1 = y^0 + hf(t_\circ, y^0)$ mit $(t_0, y^0) = (t_\circ, y^\circ)$ als Näherung für y an $t_1 = t_0 + h$. Das führt auf das *Eulerverfahren.*

Algorithmus 1 (Eulerverfahren)**.**

1: Wähle Schrittweite $h = (T - t_\circ)/k$ mit $k \in \mathbb{N}$.
2: **for** $j = 0, \ldots, n-1$ **do**
3: $\quad t_{j+1} = t_j + h$;
4: $\quad y^{j+1} = y^j + hf(t_j, y^j)$;
5: **end for**

Das Eulerverfahren zur Lösung des Systems (1.2) ist identisch mit Algorithmus 1, wobei dann natürlich y^j und $f(t_j, y^j)$ Vektoren in $\mathbb{R}^n$ sind. Die Wahl der konstanten Schrittweite $h = (T - t_\circ)/k$ ist unwesentlich. Ein Vorteil des durch das Eulerverfahren repräsentierten Verfahrenstyp liegt gerade in der flexiblen Anpassung der Schrittweite, so dass $h = h_j$ variieren kann.

Übungsaufgabe 2. Wenden Sie das Eulerverfahren auf das Anfangswertproblem

$$y'(t) = -y(t), \quad t \in (0, 1], \quad y(0) = 1 \tag{1.4}$$

mit Schrittweite $h = 1/4$ an. Vergleichen Sie Ihre numerischen Resultate mit der exakten Lösung. Wiederholen Sie Ihre Rechnungen für die Schrittweiten $h/2$ und $h/4$.

Bemerkung 1.9. Sei die Anfangsbedingung $(t_j, y^j) \in [t_\circ, T) \times \mathbb{R}^n$ gegeben. Die Funktion $\tilde{y}(t)$ löst das Anfangswertproblem

$$\tilde{y}'(t) = f(t, \tilde{y}(t)), \quad t \in (t_j, T], \quad \tilde{y}(t_j) = y^j$$

genau dann, wenn

$$\tilde{y}(t) = y^j + \int_{t_j}^{t} f(s, \tilde{y}(s))\,\mathrm{d}t, \quad t \in [t_j, T],$$

gilt. Insbesondere folgt

$$\tilde{y}(t_{j+1}) = y^j + \int_{t_j}^{t_{j+1}} f(s, \tilde{y}(s))\,\mathrm{d}s \tag{1.5}$$

für $t = t_{j+1}$. ◊

Eine Näherung für $\tilde{y}(t)$ — als auch für die gesuchte Lösung $y(t)$ von (1.2) — im Intervall $[t_j, t_{j+1}]$ ergibt sich nun, wenn man das Integral in (1.5) durch eine Quadraturformel ersetzt. Das Eulerverfahren erhalten wir dann durch

$$\tilde{y}(t_{j+1}) = y^j + \int_{t_j}^{t_{j+1}} f(s, \tilde{y}(s))\,\mathrm{d}s \approx y^j + \int_{t_j}^{t_{j+1}} f(t_j, y^j)\,\mathrm{d}s =: y^{j+1}.$$

Dies entspricht der sogenannten *Rechteckregel* bei der numerischen Integration. Statt der Rechteckregel können wir auch die *Mittelpunktsregel*

$$\int_{t_j}^{t_{j+1}} f(s, \tilde{y}(s))\,\mathrm{d}s = \int_{t_j}^{t_{j+1}} g(s)\,\mathrm{d}s \approx hg(t_{j+1/2}) = hf(t_{j+1/2}, \tilde{y}(t_{j+1/2}))$$

mit $t_{j+1/2} = t_j + h/2$ verwenden. Der Wert $f(t_{j+1/2}, \tilde{y}(t_{j+1/2}))$ ist nicht bekannt. Diesen Wert kann man aber durch $f(t_{j+1/2}, y^{j+1/2})$ annähern, wobei

$$y^{j+1/2} = y^j + \frac{h}{2} f(t_j, y^j)$$

gilt. Damit erhalten wir das *verbesserte Eulerverfahren*.

Algorithmus 2 (Verbessertes Eulerverfahren)**.**

1: Wähle Schrittweite $h = (T - t_\circ)/k$ mit $k \in \mathbb{N}$.
2: **for** $j = 0, \ldots, k-1$ **do**
3: $\quad t_{j+1/2} = t_j + h/2$;
4: $\quad y^{j+1/2} = y^j + hf(t_j, y^j)/2$;
5: $\quad t_{j+1} = t_j + h$;
6: $\quad y^{j+1} = y^j + hf(t_{j+1/2}, y^{j+1/2})$;
7: **end for**

Wie beim Algorithmus 1 lässt sich Algorithmus 2 auch auf ein System von Differentialgleichungen erster Ordnung anwenden.

Falls wir das Integral mit der Trapezregel

$$\int_{t_j}^{t_{j+1}} g(s)\,\mathrm{d}s \approx \frac{h}{2}\big(g(t_j) + g(t_{j+1})\big)$$

annähern, ergibt sich die *Trapezmethode*.

Algorithmus 3 (Trapezmethode)**.**

1: Wähle Schrittweite $h = (T - t_\circ)/k$ mit $k \in \mathbb{N}$.
2: **for** $j = 0, \ldots, k-1$ **do**

3: $t_{j+1} = t_j + h$;
4: $y^{j+1} = y^j + h(f(t_j, y^j) + f(t_{j+1}, y^{j+1}))/2$;
5: **end for**

Auch hier ändern sich die Formeln nicht, wenn man die Trapezmethode auf ein System von Differentialgleichungen anwendet. Wenn f Lipschitz-stetig in y und h hinreichend klein sind, hat die Gleichung

$$y^{j+1} = y^j + \frac{h}{2}\big(f(t_j, y^j) + f(t_{j+1}, y^{j+1})\big) \tag{1.6}$$

eine eindeutige Lösung. Dies folgt aus dem Banachschen Fixpunktsatz mit der Fixpunktabbildung $\Phi_{j,h} : \mathbb{R}^n \to \mathbb{R}^n$,

$$y \mapsto \Phi_{j,h}(y) := y^j + \frac{h}{2}\big(f(t_j, y^j) + f(t_{j+1}, y)\big).$$

Ist f Lipschitz-stetig mit einer Lipschitz-Konstanten $L \geq 0$, so erhalten wir

$$\|\Phi_{j,h}(y_1) - \Phi_{j,h}(y_2)\| = \frac{h}{2}\|f(t_{j+1}, y_1) - f(t_j, y_2)\| \leq \frac{Lh}{2}\|y_1 - y_2\|$$

für alle $y_1, y_2 \in \mathbb{R}^n$. Ist $h < 2/L$, so folgt $Lh/2 < 1$, so dass $\Phi_{j,h}$ eine Kontraktion auf $\mathbb{R}^n$ ist.

Die Trapezmethode ist ein Beispiel einer *impliziten Methode*, denn der neu zu berechnende Schritt y^{j+1} taucht in (1.6) auch auf der rechten Seite auf. Ein Schritt dieses Verfahrens erfordert die Lösung eines (eventuell nichtlinearen) Gleichungssystems. Das (verbesserte) Eulerverfahren ist hingegen ein Beispiel einer *expliziten Methode*.

Übungsaufgabe 3. Formulieren Sie die Trapezmethode für das lineare System aus Beispiel 1.2. Beschreiben Sie den Rechenaufwand pro Iterationsschritt.

Bei allen bisher eingeführten Verfahren hat man eine Vorschrift

$$\Psi_f : [t_\circ, T] \times \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n, \quad (t_j, y^j, h_j) \mapsto y^{j+1} = \Psi_f(t_j, y^j, h_j)$$

mit $\mathbb{R}^+ = (0, \infty) \subset \mathbb{R}$. Bei der Trapezmethode wird Ψ_f nicht durch eine explizite Funktion beschrieben. Diese Verfahren heißen daher auch *implizit*. Da nur die Näherung y^j an t_j verwendet wird, um y^{j+1} zu berechnen, heißen die Verfahren *Einschrittverfahren*. Eine andere Form ist

$$y^{j+1} = \Psi_f(t_j, y^j, h_j) = y^j + h_j\left(\frac{\Psi_f(t_j, y^j, h_j) - y^j}{h_j}\right) = y^j + h_j\Phi_f(t_j, y^j, h_j).$$

Die Abbildung Φ_f heißt *Verfahrens-* oder *Inkrement-Vorschrift*. Bei expliziten Verfahren kann Φ_f durch eine explizit bekannte Funktion beschrieben werden. Bei impliziten Verfahren hingegen wird Φ_f nicht durch eine explizite Funktion beschrieben, sondern steht für eine Vorschrift, deren Ausführung die Lösung eines (nichtlinearen) Gleichungssystems erfordert.

Übungsaufgabe 4. Berechnen Sie eine numerische Lösung für (1.4) mit Hilfe des (expliziten) *Heunverfahrens*

$$y^{j+1} = y^j + \frac{h}{2}\big(f(t_j, y^j) + f(t_{j+1}, y^j + hf(t_j, y^j))\big),$$

indem Sie die Schrittweiten $h = 1/4$, $h/2$ und $h/4$ verwenden. Was erwarten Sie im Vergleich zu den numerischen Resultaten von der Übungsaufgabe 2?

Programm 1. Wir betrachten das lineare Differentialgleichungssystem aus Beispiel 1.2:

$$y'(t) = Ay(t), \quad t > 0, \quad y(0) = y^\circ \tag{1.7}$$

mit den Werten $\kappa = 1$, $\ell = 1$ und der Anfangsbedingung $y^\circ = (\phi(x_1), \ldots, \phi(x_n))^T$ mit $\phi(x) = \sin(\pi x)$. Wähle $n = 60$ bei den numerischen Tests.

a) Implementieren Sie das verbesserte Eulerverfahren

$$y_{k+1} = y_k + hf\left(t_k + \frac{h}{2}, y_{k+1/2}\right), \quad y_{k+1/2} = y_k + \frac{h}{2} f(t_k, y_k),$$

zur Lösung von (1.7) mit einer Schrittweite $h = 1/n$ und $n = 60$. Stellen Sie das Ergebnis grafisch dar und vergleichen Sie Ihre Ergebnisse für unterschiedliche Schrittweiten h. Was beobachten Sie? Wie klein sollte die Schrittweite h gewählt werden, um "gute" Resultate zu erhalten? Was passiert, wenn h falsch gewählt wird?

Hinweis: Das verbesserte Eulerverfahren ist nur dann *stabil*, wenn $|h\lambda_n(A)| < 2$ gilt, wobei $\lambda_n(A)$ der größte Eigenwert von A ist (siehe Übungsaufgabe 1.21).

b) Implementieren Sie die Trapezmethode für (1.7). Was beobachten Sie bei der Trapezmethode, wenn Sie die Schrittweite variieren?

1.4. **Fehlerbetrachtung für Einschrittverfahren.** In diesem Abschnitt werden wir uns mit den wichtigen Begriffen *Konsistenz* und *Konvergenz* beschäftigen.

1.4.1. *Globaler Diskretisierungsfehler und Konvergenz.* Es ist oft hilfreich, die Folge $\{y^j\}_{j=0}^k$ als Funktion auf dem Gitter $G_h = \{t_j\}_{j=0}^k$ zu betrachten. Diese Gitterfunktion wird mit y_h bezeichnet: $y_h(t_j) = y^j$ für $j = 0, \ldots, k$. Der *globale (Diskretisierungs-)Fehler* ist dann definiert als $e_h(t_j) = y(t_j) - y_h(t_j)$ für $j = 0, \ldots, k$. Man ist letztlich am Verhalten von

$$\max_{j=0,\ldots,k} \|e_h(t_j)\| =: \|y - y_h\|_\infty$$

für $h \to 0$ interessiert. Um dies quantifizieren zu können, benutzen wir den Begriff der Konvergenz(ordnung) eines Verfahrens.

Definition 1.10. *Ein Verfahren heißt* konvergent von der Ordnung $p > 0$*, falls*

$$\|e_h\|_\infty = \mathcal{O}(h^p) \quad \textit{für } h \to 0$$

gilt.

Der globale Fehler $e_h(t_j)$, $j \in \{1, \ldots, k\}$, entsteht durch eine Akkumulation von lokalen Fehlern an den Stellen $t_0, \ldots, t_{j-1}$.

1.4.2. *Lokaler Abbruchfehler und Konsistenz.* Der *lokale Abbruchfehler* gibt an, wie sehr der durch das numerische Verfahren gelieferte Wert nach einem Schritt von der exakten Lösung abweicht. Seien $t_a \in [t_\circ, T)$, $y^a \in \mathbb{R}^n$ und $y(t; t_a, y^a)$ die Lösung der Problemstellung

$$y'(t) = f(t, y(t)), \quad t \in (t_a, T], \quad y(t_a) = y^a,$$

und

$$y_h(t_a + h; t_a, y^a) = \Psi_f(t_a, y^a, h) = y^a + h\Phi_f(t_a, y^a, h)$$

das Resultat, das das Einschrittverfahren nach einem Schritt zum Schritt (t_a, y^a) liefert. Dann wird die Differenz

$$\delta(t_a, y^a, h) = y(t_a + h; t_a, y^a) - y_h(t_a + h; t_a, y^a)$$

der *lokale Abbruchfehler im Intervall* $[t_a, t_a + h]$ genannt. Man wählt oft

$$(t_a, y^a) = (t_j, y(t_j)). \tag{1.8}$$

Dann gilt $y(t_j + h; t_j, y(t_j)) = y(t_{j+1})$. Es ergibt sich

$$\begin{aligned}\delta_{j,h} &:= \delta(t_j, y(t_j), h) = y(t_{j+1}) - y_h(t_{j+1}; t_j, y(t_j)) \\ &= y(t_{j+1}) - y(t_j) - h\Phi_f(t_j, y(t_j), h).\end{aligned}$$

Der lokale Abbruchfehler ist also für diese Wahl die Differenz zwischen dem exakten Wert $y(t_{j+1})$ und dem berechneten Wert, wenn an der Stelle t_j vom exakten Wert $y(t_j)$ der globalen Lösung von (1.2) ausgegangen wird.

Eine ebenfalls häufige Wahl ist

$$(t_a, y^a) = (t_j, y^j),$$

das heißt, als Referenzpunkt wird ein Punkt der diskreten Näherungslösung gewählt. Für den lokalen Fehler ergibt sich dann

$$\tilde{\delta}_{j,h} := \delta(t_j, y^j, h) = y(t_{j+1}; t_j, y^j) - y^{j+1} = y(t_{j+1}; t_j, y^j) - y^j - h\Phi_f(t_j, y^j, h).$$

Für eine theoretische Konvergenzanalyse ist die Größe $\delta_{j,h}$ sehr bequem, während für Schätzungen des lokalen Abbruchfehlers in der Praxis die Größe $\tilde{\delta}_{j,h}$ besser geeignet ist.

Definition 1.11. *Unter* Konsistenzfehler *verstehen wir die Größe*

$$\tau(t_a, y^a, h) := \frac{\delta(t_a, y^a, h)}{h} = \frac{y(t_a + h; t_a, y^a) - y_h(t_a + h; t_a, y^a)}{h}.$$

Im Fall (1.8) *schreiben wir*

$$\tau_{j,h} = \frac{\delta_{j,h}}{h} = \frac{y(t_{j+1}) - y_h(t_{j+1}; t_j, y(t_j))}{h}.$$

Der folgende Begriff der *Konsistenz(ordnung)* als Maß für die Größe des lokalen Abbruchfehlers stellt ein wesentliches Qualifikationskriterium eines Verfahrens dar.

Definition 1.12. *Ein Einschrittverfahren heißt mit* (1.2) konsistent von der Ordnung $p > 0$ *(oder hat die* Konsistenzordnung *p), falls*

$$\|\tau(t_a, y^a, h)\| = \mathcal{O}(h^p) \quad \textit{für } h \to 0$$

für alle Punkte (t_a, y^a) *in einer Umgebung des Lösungsgraphen* $\{(t, y(t)) \,|\, t \in [t_\circ, T]\}$ $\subset \mathbb{R}^{n+1}$ *von* (1.2) *erfüllt ist.*

Die Konsistenzordnung quantifiziert auch, wie gut das diskrete Verfahren im folgenden Sinne das kontinuierliche Problem approximiert: Lassen wir in

$$\begin{aligned}\tau(t_a, y^a, h) &= \frac{1}{h}\left(y(t_a + h; t_a, y^a) - y_h(t_a + h; t_a, y^a)\right) \\ &= \frac{1}{h}\left(y(t_a + h; t_a, y^a) - y^a\right) - \Phi_f(t_a, y^a, h))\end{aligned}$$

die Schrittweite h gegen Null gehen, strebt der Differenzenquotient gegen den Grenzwert $y'(t_a; t_a, y^a) = f(t_a, y^a)$. Folglich ergibt sich für ein konsistentes Verfahren die Bedingung

$$\lim_{h\to 0} \Phi_f(t, v, h) = f(t, v), \quad (t, v) \in [t_\circ, T] \times \mathbb{R}^n,$$

das heißt, die Verfahrensvorschrift approximiert für $h \to 0$ die Funktion f.

Wir kommen nun zur Bestimmung der Konsistenzordnung, indem wir eine allgemeine Strategie für explizite Einschrittverfahren vorstellen. Wir betrachten dazu für festes (t_a, y^a) die Abbildung $\Phi(h) := \Phi_f(t_a, y^a, h)$ als Funktion der Schrittweite h. Wir entwickeln $\tilde{y}(t_a + h) := y(t_a + h; t_a, y^a)$ und $\Phi(h)$ gemäß des Satzes von Taylor nach der Variablen h um den Entwicklungspunkt $h = 0$. Mit $\tilde{y}(t_a) = y^a$ erhalten wir

$$\begin{aligned}
\tau(t_a, y^a, h) &= \frac{1}{h}\left(\tilde{y}(t_a) + \tilde{y}'(t_a)h + \frac{1}{2!}\tilde{y}''(t_a)h^2 + \frac{1}{3!}\tilde{y}'''(t_a)h^3 + \ldots + \mathcal{O}(h^p) - y^a\right) \\
&\quad - \left(\Phi(0) + \Phi'(0)h + \frac{1}{2!}\Phi''(0)h^2 + \ldots + \mathcal{O}(h^p)\right) \\
&= \left(\tilde{y}'(t_a) - \Phi'(0)\right) + \frac{h}{2!}\left(\tilde{y}''(t_a) - 2\Phi(0)\right) + \frac{h^2}{3!}\left(\tilde{y}'''(t_a) - 3\Phi''(0)\right) + \\
&\quad \ldots + \frac{h^{p-1}}{p!}\left(\tilde{y}^{(p)}(t_a) - p\Phi^{(p-1)}(0)\right) + \mathcal{O}(h^p).
\end{aligned}$$

Um Konsistenz der Ordnung p zu erreichen, müssen die Gleichungen

$$\tilde{y}^{(j)}(t_a) = j\Phi^{(j-1)}(0), \quad j = 1, \ldots, p,$$

gelten.

Satz 1.13. *Sei die Funktion f p-mal stetig partiell differenzierbar. Dann hat das Einschrittverfahren*

$$y^{j+1} = y^j + h\Phi_f(t_j, y^j, h_j), \quad j = 0, \ldots, n-1,$$

die Konsistenzordnung (mindestens) p, falls

$$\frac{\mathrm{d}^j}{\mathrm{d}t^j} f(t, \tilde{y}(t))\Big|_{t=t_a} = (j+1)\Phi_f^{(j)}(t_a, y^a, 0), \quad j = 0, \ldots, p-1, \tag{1.9}$$

gilt.

Beispiel 1.14. Wir betrachten das Eulerverfahren für (1.2) mit $n = 1$. Wegen $\Phi_f(t, v, h) = f(t, v)$, $(t, v) \in [t_\circ, T] \times \mathbb{R}^n$, folgen

$$\Phi(h) = \Phi_f(t_a, y^a, h) = f(t_a, y^a) = \Phi(0) \quad \text{und} \quad \Phi'(0) = 0.$$

Folglich gilt (1.9) nur für $p = 1$. Das Eulerverfahren hat somit die Konsistenzordnung $p = 1$. ◊

Bemerkung 1.15. Das verbesserte Eulerverfahren $y^{j+1} = y^j + hf(t_{j+1/2}, y^j + hf(t_j, y^j)/2))$, $j \geq 0$, hat die Konsistenzordnung (mindestens) $p = 2$ und verdient aus diesem Grund seinen Namen (siehe Übungsaufgabe 5). ◊

Übungsaufgabe 5. Zeigen Sie, dass das verbesserte Eulerverfahren konsistent von der Ordnung $p = 2$ ist. Setzen Sie dabei für y und f die notwendige Differenzierbarkeit voraus.

Übungsaufgabe 6. In der Übungsaufgabe 2 ist das Heunverfahren eingeführt worden. Welche Konsistenzordnung hat das Heunverfahren, wenn Sie für y und f die notwendige Differenzierbarkeit voraussetzen?

Das Kiterium (1.9) ist auch für implizite Einschrittverfahren gültig, aber schwieriger zu handhaben, weil bei impliziten Verfahren für die Verfahrensvorschrift Φ_f keine explizit bekannte Funktion zur Verfügung steht. Bei der Konsistenzanalyse für implizite Einschrittverfahren ist der Ausgangspunkt die Definition

$$\delta(t_a, y^a, h) = y(t_a + h; t_a, y^a) - y_h(t_a + h; t_a, y^a)$$

des lokalen Abbruchfehlers.

Beispiel 1.16 (Trapezmethode, vergleiche (1.6))**.** Wir setzen $\tilde{y}(t) := y(t; t_a, y^a)$, $y_h(t) := y_h(t; t_a, y^a)$ und $\delta := \delta(t_a, y^a, h) = \tilde{y}(t_a+h) - y_h(t_a+h)$. Für die Trapezregel gilt bekanntlich

$$\begin{aligned}\frac{h}{2}\big(f(t_a, \tilde{y}(t_a)) + f(t_a + h, \tilde{y}(t_a + h))\big) &= \int_{t_a}^{t_a+h} f(s, \tilde{y}(s))\,\mathrm{d}s + \mathcal{O}(h^3)\\ &= \int_{t_a}^{t_a+h} \tilde{y}'(s)\,\mathrm{d}s + \mathcal{O}(h^3) = \tilde{y}(t_a + h) - \tilde{y}(t_a) + \mathcal{O}(h^3).\end{aligned}$$

Hiermit ergibt sich wegen $\tilde{y}(t_a) = y^a$ die Beziehung

$$\begin{aligned}y_h(t_a + h) &= y^a + \frac{h}{2}\big(f(t_a, y^a) + f(t_a + h, y_h(t_a + h))\big)\\ &= \tilde{y}(t_a) + \frac{h}{2}\big(f(t_a, y^a) + f(t_a + h, \tilde{y}(t_a + h) - \delta)\big)\\ &= \tilde{y}(t_a) + \frac{h}{2}\big(f(t_a, y^a) + f(t_a + h, \tilde{y}(t_a + h))\big) - \frac{h}{2}\frac{\partial f}{\partial y}(t_a + h, \xi)\delta\\ &= \tilde{y}(t_a) + \tilde{y}(t_a + h) - \tilde{y}(t_a) + \mathcal{O}(h^3) - \frac{h}{2}\frac{\partial f}{\partial y}(t_a + h, \xi)\delta\\ &= \tilde{y}(t_a + h) + \mathcal{O}(h^3) - \frac{h}{2}\frac{\partial f}{\partial y}(t_a + h, \xi)\delta\end{aligned}$$

mit einer Zwischenstelle zwischen $\tilde{y}(t_a + h) - \delta$ und $\tilde{y}(t_a + h)$. Also erhalten wir

$$\delta = \tilde{y}(t_a + h) - y_h(t_a + h) = \frac{h}{2}\frac{\partial f}{\partial y}(t_a + h, \xi)\delta + \mathcal{O}(h^3) \quad \text{für } h \to 0.$$

Ist $\partial f/\partial y$ stetig, so schließen wir

$$\big(1 - \mathcal{O}(h)\big)\delta = \mathcal{O}(h^3) \quad \text{für } h \to 0.$$

Ist h hinreichend klein, so dass $|\mathcal{O}(h)| < 1$ gilt, so folgern wir

$$|\tau(t_a, y^a, h)| = \frac{|\delta|}{h} = \mathcal{O}(h^2) \quad \text{für } h \to 0.$$

Die Trapezmethode hat also die Konsistenzordnung $p = 2$. ◊

1.4.3. *Zusammenhang zwischen Konsistenz und Konvergenz.* Eigentlich sind wir an der Abschätzung des globalen Diskretisierungsfehlers

$$\|y - y_h\|_\infty = \max_{j=0,\dots,k} \|e(t_j)\| = \max_{j=0,\dots,k} \|y(t_j) - y_h(t_j)\|$$

interessiert. Der folgende Satz zeigt, dass es sich in unserem Fall lohnt, den lokalen Abbruchfehler $\delta(t_a, y^a, h)$ beziehungsweise den Konsistenzfehler $\tau(t_a, y^a, h)$ zu kennen. Einen Beweis der Aussage findet man zum Beispiel in dem Buch [7].

Satz 1.17. *Falls $f(t,y)$ und $\Phi_f(t,y,h)$ beide eine Lipschitz-Bedingung in y erfüllen, so gilt für das Einschrittverfahren $y^{j+1} = y^j + h_j\Phi_f(t_j, y^j, h_j)$, $j \geq 0$, folgende Aussage:*

$$\textit{Konsistenz der Ordnung } p \quad \Longleftrightarrow \quad \textit{Konvergenz der Ordnung } p.$$

Insbesondere gilt die Abschätzung

$$\max_{j=0,\ldots,k} \|y(t_j) - y^j\| \leq e^{\bar{L}(T-t_\circ)} \left(\|y(t_\circ) - y^0\| + \sum_{j=0}^{k-1} \|\delta_{j,h}\| \right), \tag{1.10}$$

wobei $\bar{L} \geq 0$ die Lipschitzkonstante für die Verfahrensvorschrift Φ_f ist.

Bemerkung 1.18. In (1.10) werden bereits inexakte Anfangswerte $y(t_\circ) \neq y^0$ berücksichtigt. Die Abschätzung ähnelt der Abschätzung

$$\|y(t) - z(t)\| \leq e^{L(t-t_\circ)} \|y(t_\circ) - z(t_\circ)\|, \quad t \in (t_\circ, T],$$

für die stetige Abhängigkeit der exakten Lösung von (1.2) von den Anfangsbedingungen (vergleiche Satz 1.6). Hinzu kommen die jeweiligen lokalen Abbruchfehler, die sich allerdings schlimmstenfalls aufsummieren. ◊

Bemerkung 1.19. Für den Fall exakter Anfangswerte $y^0 = y^\circ = y(t_\circ)$ ergibt sich aufgrund der Definition des Konsistenzfehlers und wegen $T - t_\circ = \sum_{j=0}^{k-1} h_j$ die Ungleichung

$$\sum_{j=0}^{k-1} \|\delta_{j,h}\| = \sum_{j=0}^{k-1} h_j \|\tau_{j,h}\| \leq (T - t_\circ) \max_{j=0,\ldots,k-1} \|\tau_{j,h}\|.$$

Also erhalten wir die Abschätzung

$$\max_{j=0,\ldots,k} \|y(t_j) - y^j\| \leq e^{\bar{L}(T-t_\circ)} (T - t_\circ) \max_{j=0,\ldots,k-1} \|\tau_{j,h}\|.$$

Damit erhalten wir die in Satz 1.17 behauptete Entsprechung von Konsistenz- und Konvergenzordnung. ◊

Bemerkung 1.20. Satz 1.17 impliziert, dass sich eine Sicherung der gewünschten Genauigkeit der numerischen Lösung im gesamten Integrationsintervall auf eine im Allgemeinen viel einfachere (da lokale) Konsistenzbehandlung reduzieren lässt. ◊

Wir wollen den Beweisgang von Satz 1.17 kurz skizzieren. Wegen

$$y^{j+1} = y^j + h_j\Phi_f(t_j, y^j, h_j)$$

und

$$\delta_{j,h} = \delta(t_j, y(t_j), h_j) = y(t_{j+1}) - y(t_j) - h_j\Phi_f(t_j, y(t_j), h_j),$$

das heißt,

$$y(t_{j+1}) = y(t_j) + h_j\Phi_f(t_j, y(t_j), h_j) + \delta_{j,h},$$

erhalten wir die Abschätzung

$$\begin{aligned}
\|y(t_{j+1}) - y^{j+1}\| &= \|y(t_j) - y^j + h_j(\Phi_f(t_j, y(t_j), h_j) - \Phi_f(t_j, y^j, h_j)) + \delta_{j,h}\| \\
&\leq \|y(t_j) - y^j\| + h_j \|\Phi_f(t_j, y(t_j), h_j) - \Phi_f(t_j, y^j, h_j)\| + \|\delta_{j,h}\| \\
&\leq \|y(t_j) - y^j\| + h_j\bar{L} \|y(t_j) - y^j\| + \|\delta_{j,h}\|,
\end{aligned}$$

wenn Φ_f eine Lipschitz-Bedingung im zweiten Argument erfüllt. Mit den nichtnegativen Größen $e_j = \|y(t_j) - y^j\|$, $d_j = \|\delta_{j,h}\|$, $b_j = h_j\bar{L}$ erhalten wir die rekursive Ungleichung

$$e_{j+1} \le (1 + b_j)e_j + d_j, \quad j = 0, 1, \ldots.$$

Wie wir daraus eine explizite Ungleichung für die Fehler e_j erhalten, illustrieren wir an dem folgenden Beispiel.

Beispiel 1.21. Wir betrachten für $\lambda > 0$ das skalare Anfangswertproblem

$$y'(t) = \lambda y(t) + g(t), \quad t \in (0, T], \quad y(0) = y^\circ,$$

wobei g eine gegebene, stetige Funktion ist. Es wird das Eulerverfahren

$$y^{j+1} = y^j + h(\lambda y^j + g(t_j)), \quad j = 0, \ldots, k-1 = \frac{T}{h} - 1$$

mit konstanter Schrittweite $h = T/k$ untersucht. Wegen

$$\delta_{j,h} = y(t_{j+1}) - y(t_j) - hf(t_j, y(t_j))$$

erhalten wir

$$y(t_{j+1}) = y(t_j) + h(\lambda y(t_j) + g(t_j)) + \delta_{j,h}.$$

Für den globalen Diskretisierungfehler $e_j := y(t_j) - y^j$ ergibt sich die Rekursion

$$e_{j+1} = (1 + h\lambda)e_j + \delta_{j,h}, \quad j = 0, \ldots, k-1.$$

Es gilt $e_0 = 0$. Daher bekommen wir

$$\begin{aligned}
e_1 &= (1 + h\lambda)e_0 + \delta_{0,h} = \delta_{0,h},\\
e_2 &= (1 + h\lambda)e_1 + \delta_{1,h} = (1 + h\lambda)\delta_{0,h} + \delta_{1,h},\\
e_3 &= (1 + h\lambda)e_2 + \delta_{2,h} = (1 + h\lambda)^2\delta_{0,h} + (1 + h\lambda)\delta_{1,h} + \delta_{2,h},\\
&\vdots\\
e_k &= (1 + h\lambda)e_{n-1} + \delta_{k-1,h} = \sum_{i=0}^{k-1}(1 + h\lambda)^i\delta_{k-1-i,h}.
\end{aligned}$$

Mit Hilfe der Ungleichung $\ln(1+x) \le x$ für $x > -1$ und wegen $1 + h\lambda > 1$ erhalten wir

$$(1 + h\lambda)^i \le (1 + h\lambda)^k = e^{k\ln(1+h\lambda)} \le e^{kh\lambda} = e^{\lambda T} \quad \text{für } 0 \le i \le k.$$

Für den globalen Fehler ergibt sich wegen $|\delta_{j,h}| \le Ch^2$ mit einer von h unabhängigen Konstante $C > 0$ die Abschätzung

$$|e_k| \le \sum_{i=0}^{k-1}(1 + h\lambda)^i|\delta_{k-1-i,h}| \le e^{\lambda T}\sum_{i=0}^{k-1} Ch^2 = e^{\lambda T}kCh^2 = e^{\lambda T}CTh = \mathcal{O}(h).$$

Das bestätigt die Konvergenzordnung $p = 1$ für das Eulerverfahren. ◊

Im obigen allgemeinen Fall gehen wir zunächst ähnlich vor:

$$\begin{aligned}
e_{j+1} &\le e_j + b_je_j + d_j \le e_{j-1} + b_{j-1}e_{j-1} + d_{j-1} + b_je_j + d_j\\
&\le \ldots \le e_0 + \sum_{i=0}^{j} d_i + \sum_{i=0}^{j} b_ie_i = e_0 + \sum_{i=0}^{j} d_i + \sum_{i=0}^{j} h_i\bar{L}e_i.
\end{aligned}$$

Wir definieren

$$C := e_0 + \sum_{i=0}^{k-1} d_i, \quad u(s) = \bar{L}, \quad v|_{[t_j, t_{j+1})} := e_j, \quad j = 0, \ldots, k-1.$$

Dann sind v und u stückweise stetige, nichtnegative Funktionen auf $[t_\circ, T]$. Damit erhalten wir

$$v(t_{j+1}) \le C + \sum_{i=0}^{j} h_i u(t_i) v(t_i) = C + \int_{t_\circ}^{t_{j+1}} u(s) v(s) \, \mathrm{d}s$$

Da hier

$$\int_{t_\circ}^{t_{j+1}} u(s) \, \mathrm{d}s = \bar{L}(t_{j+1} - t_\circ) \le \bar{L}(t_n - t_\circ)$$

gilt, liefert das Gronwall-Lemma die Abschätzung

$$\max_{j=0,\ldots,k} e_j \le e^{\bar{L}(T-t_\circ)} \left(e_0 + \sum_{i=0}^{k-1} d_i \right) = e^{\bar{L}(T-t_\circ)} \left(\|y(t_0) - y^j\| + \sum_{j=0}^{k-1} \|\delta_{j,h}\| \right).$$

Wir wollen noch die Auswirkungen von Fehlern bei der Auswertung von Φ_f betrachten. Sei

$$\tilde{y}^{j+1} = \tilde{y}^j + h_j \Phi_f(t_j, \tilde{y}^j, h_j) + r_j, \quad j = 0, 1, \ldots,$$

wobei r_j der bei der Auswertung von Φ_f auftretende Fehler ist. Dieser Fehler kann Rundungsfehler enthalten, aber auch zum Beispiel Fehler, die bei einem impliziten Verfahren entstehen, wenn die in der Vorschrift $\Phi_f(t_j, \tilde{y}^j, h_j)$ auftretenden (nichtlinearen) Gleichungssysteme nur angenähert gelöst werden. Obige Argumentation bleibt dann völlig unverändert, wobei lediglich $d_j = \|\delta_{j,h}\|$ durch $d_j = \|\delta_{j,h} - r_j\|$ ersetzt wird. Es gilt dann

$$\max_{j=0,\ldots,k} \|y(t_j) - \tilde{y}^j\| \le e^{\bar{L}(T-t_\circ)} \left(\|y(t_\circ) - y^0\| + \sum_{j=0}^{k-1} \left(\|\delta_{j,h}\| + \|r_j\| \right) \right).$$

Dieses Resultat zeigt die kontrollierte Fehlerfortpflanzung (nämlich höchstens Aufsummierung) sowohl von Konsistenzfehlern als auch von anderen Störungen. In diesem Sinn ist jedes Einschrittverfahren stabil.

Bemerkung 1.22 (Stabilität von Einschrittverfahren)**.** Datenfehler ($\|y(t_\circ) - y^0\|$), Konsistenzfehler und Störungen bei der Durchführung der Vorschrift $\Phi_f(t_j, y^j, h_j)$ werden kontrolliert, das heißt, höchstens aufsummiert. ◊

Übungsaufgabe 7. Wir betrachten das skalare Anfangswertproblem

$$y'(t) = y(t) - 2\sin t, \quad t \in [0, 4], \quad y(0) = 1. \tag{1.11}$$

Zeigen Sie, dass $y(t) = \sin t + \cos t$ eine Lösung von (1.11) ist. Lösen Sie das Anfangswertproblem (1.11), indem Sie das Eulerverfahren und das verbesserte Eulerverfahren verwenden. Schreiben Sie die Näherungslösungen y_h für beide Verfahren auf. Füllen Sie für beide Verfahren eine Tabelle der Form aus und dokumentieren Sie Ihre Ergebnisse für beide Methoden.

Übungsaufgabe 8. Wenden Sie die Trapezmethode zur numerischen Lösung von (1.11) an. Schreiben Sie die numerischen Näherungen y_h auf und füllen Sie eine Tabelle wie in der Übungsaufgabe 7 aus. Was beobachten Sie für Unterschiede zu den Ergebnissen in der Übungsaufgabe 7?

h	$\lvert y_h(1) - y(1)\rvert$	$\lvert y_h(2) - y(2)\rvert$	$\lvert y_h(4) - y(4)\rvert$
2^{-4}			
2^{-5}			
2^{-6}			
2^{-7}			

TABELLE 1.1. Form der Tabelle für Übungsaufgabe 7.

1.5. **Runge-Kutta-Einschrittverfahren.** Man sucht eine Quadraturformel

$$\int_{t_j}^{t_{j+1}} f(s, y(s))\,\mathrm{d}s \approx h \sum_{i=1}^{m} \gamma_i k_i,$$

wobei γ_i geeignete Gewichte sind und $k_i \approx f(s_i, y(s_i))$, $i = 1, \dots, m$, gilt. Dies führt zu m-stufigen Runge-Kutta-Verfahren der Form

$$y^{j+1} = y^j + h \sum_{i=1}^{m} \gamma_i k_i, \quad j = 0, 1, \dots \ . \tag{1.12}$$

Es geht dabei darum, zu gegebenem m geeignete "Hilfsrichtungen" k_i zu konstruieren, so dass

- eine möglichst hohe Konsistenzordnung p erreicht wird und
- die resultierende Verfahrensvorschrift $\Phi_f(t_j, y^j, h) := \sum_{i=1}^{m} \gamma_i k_i$ eine Lipschitzbedingung in y erfüllt.

Für $m = 1$, $\gamma_1 = 1$ und $k_1 = f(t_j, y^j)$ erhalten wir mit (1.12) das Eulerverfahren als einfachstes Beispiel. Die Wahl $m = 2$, $\gamma_1 = 1$, $\gamma_2 = 0$, $k_2 = f(t_j + h_j/2, y^j + h_j f(t_j, y^j)/2)$ identifiziert das verbesserte Eulerverfahren als zweistufiges Runge-Kutta-Verfahren.

Das folgende *klassisches Runge-Kutta-Verfahren* wird in der Praxis häufig angewendet.

Algorithmus 4 (Klassisches Runge-Kutta-Verfahren)**.**

1: Wähle Schrittweiten $\{h_j\}_{j=0}^{k-1}$ mit $\sum_{j=0}^{k-1} h_j = T - t_\circ$.
2: **for** $j = 0, \dots, k-1$ **do**
3: $\quad t_{j+1/2} = t_j + h/2$ (mit $h = h_j$);
4: $\quad t_{j+1} = t_j + h$;
5: $\quad k_1 = f(t_j, y^j)$;
6: $\quad k_2 = f(t_{j+1/2}, y^j + hk_1/2)$;
7: $\quad k_3 = f(t_{j+1/2}, y^j + hk_2/2)$;
8: $\quad k_4 = f(t_{j+1}, y^j + hk_3)$;
9: $\quad y^{j+1} = y^j + h(k_1 + 2(k_2 + k_3) + k_4)/6$;
10: **end for**

Das klassische Runge-Kutta-Verfahren hat die Konsistenzordnung $p = 4$.

Beispiel 1.23. Sei y die Lösung des skalaren Problemstellung

$$y'(t) = \lambda y(t), \quad t \in (t_\circ, T], \quad y(t_\circ) = y^\circ.$$

Wir vergleichen jeweils einen Schritt der bisher betrachteten expliziten Verfahren, wobei für y^j der exakte Wert $y^j = y(t_j)$ genommen wird.

- Eulerverfahren:

$$y^{j+1} = y^j + h\lambda y^j = (1 + h\lambda)y(t^j).$$

- verbessertes Eulerverfahren:

$$y^{j+1} = y^j + h\lambda\Big(y^j + \frac{h\lambda}{2}\, y^j\Big) = \Big(1 + h\lambda + \frac{(h\lambda)^2}{2!}\Big)y(t_j).$$

- klassisches Runge-Kutta-Verfahren:

$$\begin{aligned}
k_1 &= \lambda y^j,\\
k_2 &= \lambda\Big(y^j + \frac{h}{2}\,\lambda y^j\Big) = \Big(\lambda + \frac{h\lambda^2}{2}\Big)y^j,\\
k_3 &= \lambda\Big(y^j + \frac{h}{2}\Big(\lambda + \frac{h\lambda^2}{2}\Big)y^j\Big) = \Big(\lambda + \frac{h\lambda^2}{2} + \frac{h^2\lambda^3}{4}\Big)y^j,\\
k_4 &= \lambda\Big(y^j + h\Big(\lambda + \frac{h\lambda^2}{2} + \frac{h^2\lambda^3}{4}\Big)y^j\Big)\\
&= \Big(\lambda + h\lambda^2 + \frac{h^2\lambda^3}{2} + \frac{h^3\lambda^4}{4}\Big)y^j.
\end{aligned}$$

Damit erhalten wir

$$\begin{aligned}
y^{j+1} &= y^j + \frac{h}{6}\Big(\lambda y^j + 2\Big(\lambda + \frac{h\lambda^2}{2}\Big)y^j + 2\Big(\lambda + \frac{h\lambda^2}{2} + \frac{h^2\lambda^3}{4}\Big)y^j\\
&\qquad + \Big(\lambda + h\lambda^2 + \frac{h^2\lambda^3}{2} + \frac{h^3\lambda^4}{4}\Big)y^j\Big)\\
&= \Big(1 + h\lambda + \frac{(h\lambda)^2}{2!} + \frac{(h\lambda)^3}{3!} + \frac{(h\lambda)^4}{4!}\Big)y(t_j).
\end{aligned}$$

Die exakte Lösung ist $y(t) = e^{\lambda(t-t_\circ)}y^\circ$. Wegen $y(t_{j+1}) = y(t_j + h) = e^{h\lambda}y(t_j)$ und

$$e^{h\lambda} = 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \ldots = \sum_{i=0}^{\infty} \frac{(h\lambda)^i}{i!}$$

erhalten wir für den lokalen Abbruchfehler

$$\delta_{j,h} = y(t_{j+1}) - y^{j+1} = \Big(\sum_{i=2}^{\infty} \frac{(h\lambda)^i}{i!}\Big)y(t_j) = \mathcal{O}(h^2), \quad h \to 0,$$

beim Eulerverfahren,

$$\delta_{j,h} = y(t_{j+1}) - y^{j+1} = \Big(\sum_{i=3}^{\infty} \frac{(h\lambda)^i}{i!}\Big)y(t_j) = \mathcal{O}(h^3), \quad h \to 0,$$

beim verbesserten Eulerverfahren,

$$\delta_{j,h} = y(t_{j+1}) - y^{j+1} = \Big(\sum_{i=5}^{\infty} \frac{(h\lambda)^i}{i!}\Big)y(t_j) = \mathcal{O}(h^5), \quad h \to 0,$$

beim klassischen Runge-Kutta-Verfahren. Das klassische Runge-Kutta-Verfahren hat damit in diesem Beispiel die Konsistenzordnung $p = 4$. Wegen Satz 1.17 ist das Verfahren konvergent von der Ordnung $p = 4$. $\Diamond$

Programm 2. Wir betrachten das folgende nichtlineare Problem

$$\begin{aligned} \dot{x}(t) &= \alpha x(t) + \beta x(t)y(t), \quad t > 0, \\ \dot{y}(t) &= \gamma y(t) + \delta x(t)y(t), \quad t > 0, \end{aligned} \tag{1.13}$$

mit den Anfangsbedingungen

$$x(0) = x_\circ \quad \text{und} \quad y(0) = y_\circ,$$

wobei $\alpha > 0$, $\beta < 0$, $\gamma < 0$ und $\delta > 0$. Ein stationärer Punkt für (1.13) ist durch

$$x_s = -\frac{\gamma}{\delta} \quad \text{und} \quad y_s = -\frac{\alpha}{\beta}.$$

gegeben. In einer Umgebung von (x_s, y_s) ähnelt die Lösungskurve $\{(x(t), y(t))\}_{t>0}$ einer Ellipse. Wir wählen

$$\alpha = \frac{1}{4}, \quad \beta = -\frac{1}{100}, \quad \gamma = -1, \quad \delta = \frac{1}{100},$$

und $x_\circ = 80$, $y_\circ = 30$.

a) Berechnen Sie eine numerische Lösung für (1.13) mit dem Euler- und Heunverfahren, wobei Sie als Schrittweiten $h = 1$, $h = 0.5$ und $h = 0.25$ verwenden. Stellen Sie Ihre Ergebnisse grafisch dar.

b) Wie klein muss die Schrittweite gewählt werden, so dass Sie — in der grafischen Darstellung — eine geschlossene Lösungskurve erhalten?

c) Wiederholen Sie Teil a) unter Verwendung des klassischen Runge-Kutta-Verfahrens

$$z_{k+1} = z_k + \frac{h}{6}(F_1 + 2(F_2 + F_3) + F_4), \quad k \geq 0$$

d) Testen Sie die Stabilität der Lösung des Problems (1.13) bezüglich Störungen in den Anfangsbedingungen, indem Sie $x_\circ = 80$, $y_\circ = 30$ um eins in jede Richtung — das heißt, Sie haben vier unterschiedliche Fälle — ändern. Wiederholen Sie die Rechnungen von Teil c) für die "gestörten" Anfangsdaten.

Das (verbesserte) Eulerverfahren und das klassische Runge-Kutta-Verfahren sind Spezialfälle der *m-stufigen Runge-Kutta-Verfahren.*

Algorithmus 5 (m-stufige Runge-Kutta-Verfahren)**.**

1: Wähle Gewichte $\alpha_i, \gamma_i, \beta_{i,\ell}$, $1 \leq i, \ell \leq m$, Schrittweiten $\{h_j\}_{j=0}^{k-1}$ mit $\sum_{j=0}^{k-1} h_j = T - t_\circ$.
2: **for** $j = 0, \ldots, k-1$ **do**
3: Berechne

$$\begin{aligned} &t_{j+1} = t_j + h \quad (\text{mit } h = h_j) \\ &k_i = f\big(t_j + \alpha_i h, y^j + h\textstyle\sum_{\ell=1}^m \beta_{i,\ell} k_\ell\big), \\ &y^{j+1} = y^j + h\textstyle\sum_{\ell=1}^m \gamma_\ell k_\ell \end{aligned} \tag{1.14}$$

4: **end for**

Die Gewichte in Algorithmus 5 sind dabei so zu wählen, dass das Verfahren möglichst eine hohe Genauigkeit liefert. Üblicherweise ordnet man die Gewichte in einer Tabelle an, dem sogenannten *Butcher-Tableau* (siehe Tabelle 1.2). Da die k_i in (1.14) in der Regel von allen übrigen k_ℓ, $\ell = 1, \ldots, m$, abhängen, ist (1.14) als (im Allgemeinen) nichtlineares Gleichungssystem zu verstehen. Die k_i müssen

$$\begin{array}{c|ccc} \alpha_1 & \beta_{1,1} & \dots & \beta_{1,m} \\ \vdots & \vdots & \ddots & \\ \alpha_m & \beta_{m,1} & \dots & \beta_{m,m} \\ \hline & \gamma_1 & \dots & \gamma_m \end{array}$$

TABELLE 1.2. Butcher-Tableau für m-stufige Runge-Kutta-Verfahren.

dann näherungsweise mit Hilfe eines iterativen Verfahrens ermittelt werden. Die Lösung von Gleichungssystemen in (1.14) entfällt, falls die k_i nur von $k_1, \dots, k_{i-1}$ abhängen, das heißt, wenn $\beta_{i,\ell} = 0$ für $\ell = i, \dots, m$ gilt. Man spricht dann von expliziten Runge-Kutta-Verfahren.

$$\begin{array}{c|ccccc} \alpha_1 & & & & & \\ \alpha_2 & \beta_{2,1} & & & & \\ \alpha_3 & \beta_{3,1} & \beta_{3,2} & & & \\ \vdots & \vdots & & \ddots & & \\ \alpha_m & \beta_{m,1} & \dots & \dots & \beta_{m,m-1} & \\ \hline & \gamma_1 & \dots & \dots & \gamma_{m-1} & \gamma_m \end{array}$$

TABELLE 1.3. Butcher-Tableau für explizite Runge-Kutta-Verfahren.

Wir erhalten zum Beispiel folgende Tabellen:

- Für das Eulerverfahren sind $m = 1$, $\alpha_1 = 0$, $\gamma_1 = 1$, $k_1 = f(t_j + 0, y^j + 0)$ und $y^{j+1} = y^j + hk_1$. In Tabelle 1.4 ist das Butcher-Tableau angegeben.

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

TABELLE 1.4. Butcher-Tableau für das Eulerverfahren.

- Beim verbesserten Eulerverfahren gelten $m = 2$, $\alpha_1 = 0$, $\alpha_2 = 1/2$, $\beta_{2,1} = 1/2$, $\gamma_1 = 0$, $\gamma_2 = 1$, $k_1 = f(t_j + 0, y^j + 0)$, $k_2 = f(t_j + h/2, y^j + hk_1/2)$. Das Butcher-Tableau ist in Tabelle 1.5 präsentiert.

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

TABELLE 1.5. Butcher-Tableau für das verbesserte Eulerverfahren.

- Für das klassische Runge-Kutta-Verfahren sind $m = 4$, $k_1 = f(t_j+0, y^j+0)$, $k_2 = f(t_j + h/2, y^j + hk_1/2)$, $k_3 = f(t_j + h/2, y^j + hk_2/2)$, $k_4 = f(t_j + h, y^j + hk_3)$ und $y^{j+1} = h(k_1 + 2(k_2 + k_3) + k_4)/6$. In Tabelle 1.6 ist das Butcher-Tableau angegeben.

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

TABELLE 1.6. Butcher-Tableau für das klassische Runge-Kutta-Verfahren.

Übungsaufgabe 9. Zeigen Sie, dass das die Trapezmethode und das in Übungsaufgabe 4 eingeführte Heunverfahren als m-stufige Runge-Kutta-Verfahren gesehen werden können. Geben Sie für beide Verfahren das Butcher-Tableau an.

Übungsaufgabe 10. Wir betrachten die folgende Differentialgleichung zweiter Ordnung

$$y''(t) - 10y'(t) - 11y(t) = 0, \quad t > 0, \tag{1.15}$$

mit den Anfangsbedingungen

$$y(0 = 1, \quad y'(0) = -1.$$

Zeigen Sie, dass die Lösung des Anfangswertproblems $y(t) = e^{-t}$ ist. Betrachten Sie nun (1.15) mit "gestörten" Anfangsbedingungen

$$y(0) = 1 + \varepsilon, \quad y'(0) = -1$$

mit $0 < \varepsilon \ll 1$. Berechnen Sie die analytische Lösung des gestörten Problems und erzeugen Sie einen grafischen Vergleich der erhaltenen Ergebnisse. Was sind die Beobachtungen hinsichtlich der Stabilität der Lösung? Welche Auswirkung hat das auf die Anwendung von numerischen Verfahren zur näherungsweisen Lösung von (1.15)?

1.6. **Schrittweitensteuerung.** In einem effizienten numerischen Verfahren geht es darum, eine gewünschte Genauigkeit mit möglichst wenigen Integrationsschritten zu realisieren. Das kann man zum Beispiel dadurch erreichen, wenn der lokale Abbruchfehler im Laufe der Rechnung kontrolliert wird und die Schrittweite h dementsprechend angepasst, also adaptiv verändert wird. Hier liegt ein wesentlicher Vorteil der Einschrittverfahren (im Vergleich zu den Mehrschrittverfahren), dass eine Änderung der Schrittweite sehr leicht zu implementieren ist. Das Vorgehen ist wie folgt.

- Für ein Integrationsintervall $[t_\circ, T]$ sei $y_h(T) = y^n$ die numerische Lösung. Dann ist eine Gesamtfehlertoleranz

$$\|y(T) - y_h(T)\| \leq (T - t_\circ)\varepsilon \tag{1.16}$$

  mit $\varepsilon > 0$ vorgegeben. Aufgrund der Abschätzung (1.10) aus Satz 1.17 verwenden wir den Ansatz, dass die Summe aller lokalen Abbruchfehler $\|\tilde{\delta}_{j,h}\|$, $j = 0, \ldots, k-1$, eine brauchbare Schätzung für den globalen Fehler an $t_k = T$ ist

$$\|y(T) - y_h(T)\| \leq \sum_{j=0}^{k-1} \|\tilde{\delta}_{j,h}\|,$$

wobei wir die (von $T - t_\circ$) abhängige Konstante der Einfachheit halber unterdrücken. Wir erhalten eine günstige Schrittweitenwahl, wenn in jedem Schritt etwa der gleiche Konsistenzfehler entsteht. Gilt für den lokalen Abbruchfehler

$$\|\tilde{\delta}_{j,h}\| \leq (t_{j+1} - t_j)\varepsilon,$$

so erhalten wir

$$\sum_{j=0}^{k-1} \|\tilde{\delta}_{j,h}\| \leq \sum_{j=0}^{k-1} (t_{j+1} - t_j)\varepsilon = (T - t_\circ)\varepsilon,$$

also (1.16).
Die Strategie lautet also: Ist $h = t_{j+1} - t_j$ die momentane Schrittweite, so darf der lokale Fehler $\|\tilde{\delta}_{j,h}\|$ im Schritt von j nach $j+1$ höchstens $h\varepsilon$ sein. Auf der anderen Seite sollte diese Spannweite möglichst ausgeschöpft werden.

- Es ist also eine gute Schätzung des lokalen Abbruchfehlers in jedem Zeitschritt gefragt. Hier schätzen wir $\tilde{\delta}_{j,h}$ anstatt $\delta_{j,h}$. Wir gehen wie folgt vor. Ausgehend von t_j und y^j berechne
  1) einen Schritt mit der Schrittweite h und bezeichne das Resultat mit y^{j+1};
  2) zwei Schritte mit der Schrittweite $h/2$ und nenne das Resultat $\hat{y}^{j+1}$.

  Hat das Verfahren die Konsistenzordnung p, so ist der lokale Abbruchfehler $\tilde{y}(t_{j+1}) - y^{j+1}$ (mit $\tilde{y}(t_{j+1}) = y(t_{j+1}; t_j, y^j)$) proportional zu h^{p+1}. Es folgen daher

$$\tilde{y}(t_{j+1}) - y^{j+1} \approx c(t_j)h^{p+1}, \quad \tilde{y}(t_{j+1}) - \hat{y}^{j+1} \approx 2c(t_j)\left(\frac{h}{2}\right)^{p+1}.$$

  Wir erhalten

$$\begin{aligned}\hat{y}^{j+1} - y^{j+1} &\approx c(t_j)h^{p+1} - 2c(t_j)\left(\frac{h}{2}\right)^{p+1} = c(t_j)h^{p+1}(1 - 2^{-p}) \\ &= 2c(t_j)h^{p+1}(2^{-1} - 2^{-p-1}) = 2c(t_j)\left(\frac{h}{2}\right)^{p+1}(2^p - 1) \\ &\approx (2^p - 1)(\tilde{y}(t_{j+1}) - \hat{y}^{j+1})\end{aligned}$$

  und somit

$$\tilde{y}(t_{j+1}) - \hat{y}^{j+1} \approx \frac{1}{2^p - 1}\left(\hat{y}^{j+1} - y^{j+1}\right) =: s(h)$$

  für den lokalen Abbruchfehler der Annäherung $\hat{y}^{j+1}$.

Oft wird folgende Vorgangsweise gewählt. Sei $h = h_j$ die aktuelle Schrittweite zum Zeitpunkt t_j, für die wir $s(h)$ berechnet haben. Bei einem Verfahren der Ordnung p gilt $\tilde{\delta}_{j,h} \approx s(h) \approx ch^{p+1}$ mit einer von h unabhängigen Konstante c. Wir setzen $q(h) := |s(h)|/(\varepsilon h)$, wobei ε die Toleranz aus (1.16) ist. Ist $q(h) \leq 1$ erfüllt, so wird der Zeitschritt akzeptiert und wir gehen zum neuen Zeitpunkt t_{j+1} mit einer neuen Schrittweite h_{neu}. Im Fall $q(h) > 1$ wird h verkleinert und eine neue Näherung berechnet. Dabei soll h_{neu} der Beziehung

$$\frac{ch_{\text{neu}}^{p+1}}{\varepsilon h_{\text{neu}}} \approx 1$$

genügen. Wegen

$$\frac{ch_{\text{neu}}^{p+1}}{\varepsilon h_{\text{neu}}} = \frac{ch^{p+1}}{\varepsilon h}\left(\frac{ch_{\text{neu}}}{h}\right)^p \approx \frac{s(h)}{\varepsilon h}\left(\frac{ch_{\text{neu}}}{h}\right)^p = q(h)\left(\frac{ch_{\text{neu}}}{h}\right)^p$$

erhalten wir $h_{\text{neu}} \approx q(h)^{-1/p}h$. In der Praxis führt man oft noch Sicherheitsfaktoren $\alpha_{\max} \in [1.5, 2]$, $\alpha_{\min} \in [0.2, 0.5]$, $\beta \in [0.9, 0.95]$ ein, die verhindern, dass die Schrittweiten zu groß werden oder die Fehler knapp über der Toleranz liegen:

$$\begin{aligned} q(h) \le 1 \quad &\longrightarrow \quad h_{\text{neu}} = \beta \min\left\{\alpha_{\max}, q(h)^{-1/p}\right\}h, \\ q(h) > 1 \quad &\longrightarrow \quad h_{\text{neu}} = \beta \max\left\{\alpha_{\min}, q(h)^{-1/p}\right\}h. \end{aligned}$$

Programm 3. Implementieren Sie eine Funktion

```
[y,t] = odesolve4(fun,y0,t,epsilon)
```

zur Lösung von einem Differentialgleichngssystem erster Ordnung mit einer Schrittweitensteuerung, wobei `fun` eine Funktion für die rechte Seite der Differentialgleichung, `y0` die Anfangsbedingng, `t = [t0,T]` das betrachtete Zeitintervall mit der Anfangszeit `t0` und der Endzeit `T` sowie `epsilon` eine gegebene Toleranz sind. Die Ausgabewerte der Funktion sind die numerische Lösung `y` und das verwendete Zeitgitter `t` für die Schrittweitenwahl.

Zur Lösung der Differentialgleichung soll das klassische Runge-Kutta-Verfahren verwendet werden. Für die Schrittweitenwahl wenden wir das Verfahren einmal mit Schrittweite h und dann mit Schrittweite $h/2$ an. Aus der Differenz der erhaltenen numerischen Näherungen lassen sich der Fehlerschätzer $s(h) \approx \tilde{\delta}_{j,h}$ und das Verhältnis $q(h)$ berechnen. Mit diesen Größen lässt sich entscheiden, ob eine Schrittweite h vergrößert oder verkleinert wird und ob ein Schritt akzeptiert wird oder nicht. Zur Implementation verwenden Sie das Flussdiagramm in Abbildung 1.1. Als Parameter wählen Sie:

$$h_{\min} = 10^{-4}, \quad h_{\max} = 0.5, \quad \alpha_{\min} = 0.2, \quad \alpha_{\max} = 2, \quad \beta = 0.95.$$

Initialisieren Sie die Schrittweite h mit $h_{\max}$. Verwenden Sie `epsilon`$= 10^{-4}$ und testen Sie folgende drei Beispiele:

$$\begin{aligned} &y'(t) = -y(t), \quad t \in (0,5], \quad y(0) = 1, \\ &\begin{pmatrix} y_1'(t) \\ y_2'(t) \end{pmatrix} = \begin{pmatrix} \frac{1}{4}y_1(t) - \frac{1}{100}y_1(t)y_2(t) \\ -y_1(t) + \frac{1}{100}y_1(t)y_2(t) \end{pmatrix}, \ t \in (0,12], \quad \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 80 \\ 30 \end{pmatrix}, \\ &y''(t) = 8(1 - y(t)^2)y'(t) - y(t), \quad t \in (0,30], \quad y(0) = 2, \quad y'(0) = 0. \end{aligned}$$

Visualisieren Sie sowohl Ihre numerische Lösung als auch die Schrittweite für jeden Iterationsschritt. Dabei soll die Funktion `odesolve4` keine grafische Ausgabe enthalten. Dokumentieren Sie Ihren Code. Eine Beschreibung der implementierten Funktionen unter Verwendung des MATLAB Befehls `help` soll möglich sein.

1.7. **Mehrschrittverfahren.** Mehrschrittverfahren bieten die Möglichkeit, höhere Konsistenzordnung bei relativ wenigen Funktionsauswertungen zu erreichen. Wir greifen dabei nicht nur auf eine bereits berechnete, sondern auf mehrere berechnete Näherungen zurück. Der einfacheren Darstellung wegen sei $h = (T - t_\circ)/k$.

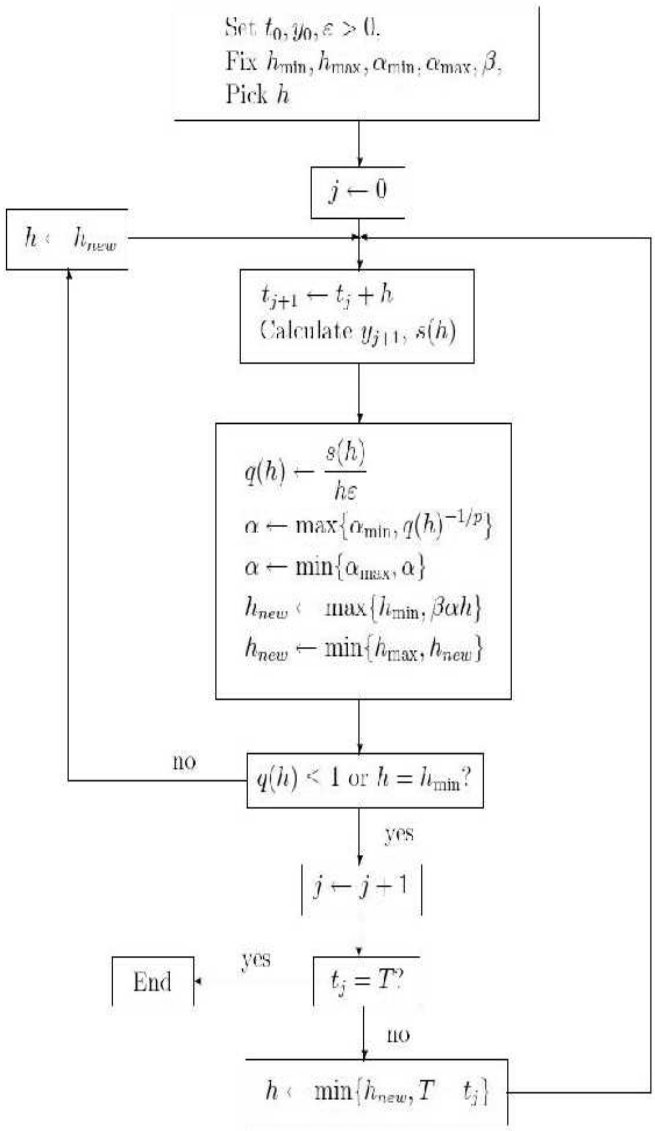

ABBILDUNG 1.1. Flußdiagramm für die Schrittweitensteuerung bei einem Verfahren der Ordnung $p > 0$; siehe [1, Abb. 11.8].

1.7.1. *Adams-Bashforth-Verfahren.* Wir diskretisieren die Integralgleichung

$$y(t_{j+k}) = y(t_{j+k-1}) + \int_{t_{j+k-1}}^{t_{j+k}} f(s, y(s))\,\mathrm{d}s \tag{1.17}$$

an den Stützstellen $t_{j+k-1}, t_{j+k-2}, \ldots, t_j$ mit Hilfe von Newton-Cotes-Formeln. Das Verfahren wird in Algorithmus 6 beschrieben.

Algorithmus 6 (m-Schritt-Adams-Bashforth-Verfahren)**.**

1: Wähle Schrittweite $h = (T - t_\circ)/k$ mit $k \in \mathbb{N}$, Koeffizienten $b_{m,\ell}$ $(0 \le \ell \le m-1)$ und Startwerte $y^0, \dots, y^{m-1} \in \mathbb{R}^n$.
2: **for** $j = 0, \dots, k - m$ **do**
3: $\quad t_{j+m} = t_{j+m-1} + h$;
4: $\quad y^{j+m} = y^{j+m-1} + h \sum_{\ell=0}^{m-1} b_{m,\ell} f(t_{j+\ell}, y^{j+\ell})$;
5: **end for**

Wir erhalten folgende Tabelle.

m	ℓ	0	1	2	3	4	Konsistenzordnung
1	$b_{1,\ell}$	1					1
2	$2b_{2,\ell}$	-1	3				2
3	$12b_{3,\ell}$	5	-16	23			3
4	$24b_{4,\ell}$	-9	37	-59	55		4
5	$720b_{5,\ell}$	251	-1274	2616	-2774	1901	5

TABELLE 1.7. Adams-Bashforth-Formeln.

Beispiel 1.24. Wir wollen den Fall $m = 2$ herleiten. Das Integral

$$\int_{t_{j+1}}^{t_{j+2}} f(s, y(s))\,\mathrm{d}s =: \int_{t_{j+1}}^{t_{j+2}} g(s)\,\mathrm{d}s =: I(g)$$

wird mit Newton-Cotes-Formeln zu den Stützstellen t_{j+1} und t_j angenähert. Das lineare Interpolationspolynom von g lautet

$$P(g|t_j, t_{j+1})(s) = \frac{t_{j+1} - s}{h}\, g(t_j) + \frac{s - t_j}{h}\, g(t_{j+1}).$$

Damit erhalten wir die Approximation

$$\begin{aligned} I(g) \approx I_1(g) &:= \int_{t_{j+1}}^{t_{j+2}} P(g|t_j, t_{j+1})(s)\,\mathrm{d}s \\ &= g(t_j) \int_{t_{j+1}}^{t_{j+2}} \frac{t_{j+1} - s}{h}\,\mathrm{d}s + g(t_{j+1}) \int_{t_{j+1}}^{t_{j+2}} \frac{s - t_j}{h}\,\mathrm{d}s \\ &= h\left(-\frac{1}{2}\, g(t_j) + \frac{3}{2}\, g(t_{j+1}) \right) = h \sum_{\ell=0}^{1} b_{2,\ell} f(t_{j+\ell}, y(t_{j+\ell})). \end{aligned}$$

Es gelten also $b_{2,0} = -1/2$, $b_{2,1} = 3/2$. Wegen des Interpolationsfehlers

$$g(s) - P(g|t_j, t_{j+1})(s) = (s - t_j)(s - t_{j+1}) \frac{g''(\xi)}{2}$$

bekommen wir

$$I(g) - I_1(g) = \frac{g''(\xi)}{2} \int_{t_{j+1}}^{t_{j+2}} (s - t_j)(s - t_{j+1})\,\mathrm{d}s = \frac{5g''(\xi)}{12}\, h^3.$$

Damit lässt sich der lokale Abbruchfehler berechnen:

$$\delta_{j+1,h} = y(t_{j+2}) - y(t_{j+1}) - h\sum_{\ell=0}^{1} b_{2,\ell} f(t_{j+\ell}, y(t_{j+\ell})) = \int_{t_{j+1}}^{t_{j+2}} y'(s)\,\mathrm{d}s - I_1(g)$$
$$= \int_{t_{j+1}}^{t_{j+2}} f(s, y(s))\,\mathrm{d}s - I_1(g) = I(g) - I_1(g) = \frac{5g''(\xi)}{12}\,h^3.$$

Das Verfahren hat also — bei glattem f — die Konsistenzordnung $p = 2$. $\Diamond$

Die Adams-Bashforth-Verfahren sind explizit. Pro Iterationsschritt ist nur eine einzige Funktionsauswertung $f(t_{j+m-1}, y^{j+m-1})$ erfoderlich, da die vorangehenden Werte $f(t_{j+m-2}, y^{j+m-2}), \ldots, f(t_j, y^j)$ bereits berechnet worden sind. Allerdings ist eine Schrittweitenänderung bei Mehrschrittverfahren problematisch, da die Berechnung zusätzlicher Punkte der Lösungskurve bei Schrittweitenänderung erforderlich ist.

Übungsaufgabe 11. Zeigen Sie, dass das Adams-Bashforth-Verfahren vierter Ordnung zur Lösung von (1.2) gegeben ist durch

$$y^{j+1} = y^j + \frac{h}{24}\left(55f_j - 59f_{j-1} + 37f_{j-2} - 9f_{j-3}\right)$$

mit $f_j = f(t_j, y^j)$.

1.7.2. *Adams-Moulton-Verfahren.* Bei diesen Verfahren wird neben den bekannten Näherungen für $f(s, y(s))$ an $t_{j+m-1}, \ldots, t_j$ auch noch der unbekannte Wert $f(t_{j+m}, y^{j+m})$ an t_{j+m} verwendet.

Algorithmus 7 (m-Schritt-Adams-Moulton-Verfahren)**.**

1: Wähle Schrittweite $h = (T - t_\circ)/k$ mit $k \in \mathbb{N}$, Koeffizienten $b_{m,\ell}$ $(0 \le \ell \le m)$ und Startwerte $y^0, \ldots, y^{m-1} \in \mathbb{R}^n$.
2: **for** $j = 0, \ldots, k - m$ **do**
3: $\quad t_{j+m} = t_{j+m-1} + h$;
4: $\quad y^{j+m} = y^{j+m-1} + h\sum_{\ell=0}^{m} b_{m,\ell} f(t_{j+\ell}, y^{j+\ell})$;
5: **end for**

Wir erhalten folgende Tabelle.

m	ℓ	0	1	2	3	4	Konsistenzordnung
1	$1b_{1,\ell}$	1	1				2
2	$12b_{2,\ell}$	-1	8	5			3
3	$24b_{3,\ell}$	1	-5	19	9		4
4	$720b_{4,\ell}$	-19	106	-264	646	251	5

Tabelle 1.8. Adams-Moulton-Formeln.

Beispiel 1.25. Wir wollen den Fall $m = 1$ herleiten. Dabei wird

$$\int_{t_j}^{t_{j+1}} f(s, y(s))\,\mathrm{d}s = \int_{t_j}^{t_{j+1}} g(s)\,\mathrm{d}s =: I_1(g)$$

mit Hilfe der Newton-Cotes-Formeln an t_j und t_{j+1} angenähert. Wir erhalten gerade die Trapezmethode

$$I_1(g) = \frac{h}{2}\big(g(t_j) + g(t_{j+1})\big).$$

Offensichtlich wird das Integral über f durch

$$\frac{h}{2}\big(f(t_j, y^j) + f(t_{j+1}, y^{j+1})\big)$$

approximiert. In der obigen Tabelle haben wir daher $b_{1,0} = b_{1,1} = 1/2$. ◊

1.7.3. *Lineare Mehrschrittverfahren.* Die allgemeine Form eines *m-Schrittverfahrens* lautet

$$y^{j+k} = \Phi_h(t_{j+k-1}, y^j, \ldots, y^{j+k}), \quad j = 0, \ldots, n-k.$$

Am häufigsten werden *lineare Mehrschrittverfahren* eingesetzt. Ein lineares m-Schrittverfahren hat die Gestalt

$$\sum_{\ell=0}^{m} a_\ell y^{j+\ell} = h \sum_{\ell=0}^{m} b_\ell f(t_{j+\ell}, y^{j+\ell}), \quad j = 0, \ldots, n-m, \tag{1.18}$$

wobei die a_ℓ, b_ℓ fest gewählte Koeffizienten sind und stets $t_j = t_\circ + jh$ gilt. Zur Ausführung von (1.18) benötigen wir die Werte $y^0, \ldots, y^{m-1}$. Diese werden in der Regel mit einem Einschrittverfahren berechnet. Ohne Beschränkung der Allgemeinheit gelte $a_m = 1$ dann ergibt sich folgender Algorithmus.

Algorithmus 8 (Lineares m-Schrittverfahren)**.**

1: Wähle Schrittweite $h = (T - t_\circ)/k$ mit $k \in \mathbb{N}$, Koeffizienten a_ℓ $(0 \le \ell \le m-1)$ und b_ℓ $(0 \le \ell \le m)$ und Startwerte $y^0, \ldots, y^{m-1} \in \mathbb{R}^n$.
2: **for** $j = 0, \ldots, k-m$ **do**
3: Berechne

$$y^{j+m} = -\sum_{\ell=0}^{m-1} a_\ell y^{j+\ell} + h \sum_{\ell=0}^{m} b_\ell f(t_{j+\ell}, y^{j+\ell}); \tag{1.19}$$

4: **end for**

Gilt in (1.19) die Beziehung $b_m \neq 0$, so ist die Lösung eines im Allgemeinen nichtlinearen Gleichungssystems erforderlich. Das Mehrschrittverfahren ist dann implizit. Ein explizites Verfahren erhalten wir im Fall von $b_m = 0$.

Bemerkung 1.26.

1) Für $a_0 = \ldots = a_{m-2} = 0$, $a_{m-1} = -1$ und $b_k = 0$ erhalten wir das Adams-Bashforth-Verfahren:

$$y^{j+m} = y^{j+m-1} + h \sum_{\ell=0}^{m-1} b_\ell f(t_{j+\ell}, y^{j+\ell}).$$

2) Bei der Wahl $a_0 = \ldots = a_{m-2} = 0$, $a_{m-1} = -1$ und $b_m \neq 0$ ergibt sich das Adams-Moulton-Verfahrens

$$y^{j+m} = y^{j+m-1} + h \sum_{\ell=0}^{m} b_\ell f(t_{j+\ell}, y^{j+\ell}).$$

3) Mit $b_0 = \ldots, b_{m-1} = 0$ und $b_m = 0$ erhalten wir das Verfahrens

$$y^{j+m} = -\sum_{\ell=0}^{m-1} a_\ell y^{j+\ell} + h b_m f(t_{j+m}, y^{j+m}),$$

welches *Rückwärtsmethode* genannt wird. ◊

Der lokale Abbruchfehler in $[t_{j+m-1}, t_{j+m}]$ ist

$$\delta_{j+m-1} := y(t_{j+m}) - y_h(t_{j+m}),$$

wobei $y_h(t_{j+m})$ das Ergebnis aus (1.19) mit den Werten $y^{j+\ell} = y(t_{j+\ell})$, $\ell = 0, \ldots, m-1$ ist. Bei expliziten Verfahren erhalten wir den Fehler durch Betrachtung von

$$\delta_{j+m-1} := y(t_{j+m}) + \sum_{\ell=0}^{m-1} a_\ell y(t_{j+\ell}) - h \sum_{\ell=0}^{m} b_\ell f(t_{j+\ell}, y(t_{j+\ell})).$$

Als Konsistenzfehler $\tau_{j+m-1,h}$ definieren wir den Quotienten $\delta_{j+m-1,h}/h$. Mit Hilfe von Taylorentwicklungen lassen sich Konsistenzbedingungen angeben.

Beispiel 1.27. Wir betrachten das Mehrschrittverfahren

$$y^{j+1} = y^{j-1} + 2hf(t_j, y^j), \quad j \geq 1, \tag{1.20}$$

und wenden das Verfahren auf

$$y'(t) = -2y(t) + 1, \quad t > 0, \quad y(0) = 1 \tag{1.21}$$

an. Das Verfahren (1.20) hat die Konsistenzordnung $p = 2$ (siehe Übungsaufgabe 12). Die Lösung von (1.21) lautet $y(t) = (e^{-2t} + 1)/2$. Diese Lösung ist stabil, da $\tilde{y}(t) = ((1 + 2\varepsilon)e^{-2x} + 1)/2$ die Lösung zu der "gestörten" Anfangsbedingung $\tilde{y}(0) = 1 + \varepsilon$ ist. Wir wenden nun (1.20) auf (1.21) an, wobei wir als Startwerte die exakten Werte $y^0 = 1$ und $y^1 = (e^{-2h} + 1)/2 = y(h)$ verwenden:

$$y^{j+1} = y^{j-1} + 2h(-2y^j + 1) = -4hy^j + y^{j-1} + 2h, \quad j \geq 1.$$

Es ergibt sich, dass

$$|y^j| \to \infty \quad \text{für } j \to \infty \tag{1.22}$$

(siehe Übungsaufgabe 13). Damit wird nicht das Verhalten der Lösung für $t \to \infty$ dargestellt. Das Verfahren ist also instabil. ◊

Übungsaufgabe 12. Zeigen Sie, dass das Verfahren (1.20) die Konsistenzordnung $p = 2$ besitzt.

Zur Untersuchung der Instabilität verwenden wir die Theorie der linearen Differenzengleichungen der Ordnung m mit konstanten Koeffizienten:

$$y_{j+1} = a_m y_j + \ldots + a_1 y_{j-m+1} + a_0, \quad j = m-1, \ldots \tag{1.23}$$

mit gegebenen Koeffizienten $a_0, \ldots, a_m \in \mathbb{R}$. Der homogene Teil von (1.23) ist:

$$y_{j+1} = a_m y_j + \ldots + a_1 y_{j-m+1}, \quad j = m-1, \ldots . \tag{1.24}$$

Wir suchen Lösungen für (1.24) von der Gestalt $y_i = \lambda^i$ mit unbekanntem λ. Offenbar löst λ^i die Gleichung (1.24), wenn

$$\lambda^{j+1} = a_m \lambda^j + \ldots + a_1 \lambda^{j-m+1}$$

beziehungsweise

$$\lambda^m - a_m \lambda^{m-1} - \ldots - a_1 = 0 \tag{1.25}$$

gilt. Die Beziehung (1.25) heißt *charakteristische Gleichung* von (1.24). Sind $\lambda_1, \ldots, \lambda_m$ paarweise verschiedene Wurzeln von (1.25), dann ist $\{\lambda_1^j, \ldots, \lambda_m^j\}$ ein Fundamentalsystem von (1.24). Wir erhalten die allgemeine Lösungen

$$y_j = \sum_{\ell=1}^{m} c_\ell \lambda_\ell^j, \quad c_\ell \in \mathbb{R}.$$

Ferner ist $y_\ell = a_0/(1 - a_1 - \ldots - a_m)$, $0 \le \ell \le j$, eine partikuläre Lösung von (1.23), denn

$$\begin{aligned} \frac{a_0}{1 - a_1 - \ldots - a_m} &= a_m y_j + \ldots + a_1 y_{j-m+1} + a_0 \\ &= (a_1 + \ldots + a_m) \frac{a_0}{1 - a_1 - \ldots - a_m} + a_0 \end{aligned}$$

impliziert

$$a_0 = (a_1 + \ldots + a_m)a_0 + (1 - a_1 - \ldots - a_m)a_0,$$

was offenbar erfüllt ist. Die allgemeine Lösung von (1.23) ist dann gegeben durch

$$y_j = \sum_{\ell=1}^{m} c_\ell \lambda_\ell^j + \frac{a_0}{1 - a_1 - \ldots - a_m}, \quad j = m, m+1, \ldots.$$

Aus den Startwerten $y_0, \ldots, y_{m-1}$ lassen sich die c_ℓ's über die Gleichungen

$$\sum_{\ell=1}^{m} c_\ell \lambda_\ell^j + \frac{a_0}{1 - a_1 - \ldots - a_m} = y_j, \quad j = 0, \ldots, m-1,$$

bestimmen.

Übungsaufgabe 13. Zeigen Sie mit Hilfe der Theorie der linearen Differenzengleichungen, dass für das Zweischrittverfahren (1.20) die Eigenschaft (1.22) gilt.

Übungsaufgabe 14. Lösen Sie die lineare Differenzengleichung

$$y^{j+1} = \frac{5}{2} y^j + y^{j-1}, \quad y_0 = y_1 = 1,$$

mit Hilfe der Theorie der linearen Differenzengleichungen. Diskutieren Sie das Verhalten der Folge $\{y^j\}_{j \in \mathbb{N}}$ für $j \to \infty$.

1.7.4. *Prädiktor-Korrektor-Verfahren.* Um bei impliziten Verfahren die im Allgemeinen iterative Bestimmung von y^{j+1} durch eine Fixpunktiteration oder das Newtonverfahren zu vermeiden, wählt man häufig ein sogenanntes *Prädiktor-Korrektor-Verfahren*, wobei y^{j+1} zunächst durch ein explizites Verfahren "geschätzt" wird:

- *Prädiktor*: Bestimme einen Startwert $y^{j+1,0}$ mit einem m_1-Schritt-Adams-Bashforth-Verfahren

$$y^{j+1,0} = y^j + h \sum_{\ell=0}^{m_1-1} b_{m_1-1-\ell} f(t_{j-\ell}, y^{j-\ell}).$$

- *Korrektor*: Berechne y^{j+1} iterativ über M Iterationen einer Fixpunktabbildung unter Verwendung eines m_2-Schritt-Adams-Moulton-Verfahrens.:

  for $i = 0, \dots, M$ **do**

  $$y^{j+1,i+1} = y^j + h\tilde{b}_{m_2} f(t_{j+1}, y^{j+1,i}) + h \sum_{\ell=0}^{m_2-1} \tilde{b}_{m_2-1-\ell} f(t_{j-\ell}, y^{j-\ell});$$

  end for

  $y^{j+1} = y^{j+1,M+1}$;

Es lässt sich zeigen, dass dieses Prädiktor-Korrektor-Verfahren die Konsistenzordnung $\min\{m_1 + 1 + M, m_2 + 1\}$ hat. Daher wählt man oft $m_1 = m_2$ und $M = 0$.

Die Kombination einer m-Schritt-Adams-Bashforth-Methode (Ordnung m) mit einem m-Schritt-Adams-Moulton-Verfahren (Ordnung $m + 1$) hat bei nur einer Fixpunktiteration ($M = 0$) folgende Eigenschaften:

- Konsistenzordnung $m + 1$.
- zwei Funktionsauswertungen erforderlich,
- bessere Stabilitätseigenschaften als ein m + 1-Schritt-Adams-Bashforth-Verfahren.

Programm 4. Implementieren Sie die beiden Funktionen

`[y,t] = odesolveABM3(fun,y0,t,h)` und `[y,t] = odesolveAB4(fun,y0,t,h)`

zur Lösung von einem System von Differentialgleichungen erster Ordnung, wobei `fun` für eine Funktion der Bauart `my_fun(t,y)` steht, die die rechte Seite der Differentialgleichung an `(t,y)` berechnet, `y0` die Anfangsbedingung, `t = [t0,T]` das Zeitintervall mit Anfangszeit `t0` und Endzeit `T` sowie `h` die Schrittweite sind. Die Ausgabewerte sind die numerische Näherung `y` und das Zeitgitter `t`.

- `odesolveABM3` ist ein Prädiktor-Korrektor-Verfahren basierend auf einem Adams-Bashforth-Verfahren dritter Ordnung (als Prädiktor) und einem Adams-Moulton-Verfahren vierter Ordnung (als Korrektor):

  $$\begin{aligned} y^{j+1,0} &= y^j + \frac{h}{12}\big(23f(t_j, y^j) - 16f(t_{j-1}, y^{j-1}) + 5f(t_{j-2}, y^{j-2})\big) \\ y^{j+1,1} &= y^j + \frac{h}{24}\big(9f(t_{j+1}, y^{j+1,0}) + 19f(t_j, y^j) - 5f(t_{j-1}, y^{j-1}) \\ &\qquad + f(t_{j-2}, y^{j-2})\big) \\ y^{j+1} &:= y^{j+1,1}. \end{aligned}$$

- `odesolveAB4` ist ein Adams Bashforth-Verfahren vierter Ordnung

  $$\begin{aligned} y^{j+1} = y^j + \frac{h}{24}\big(&55f(t_j, y^j) - 59f(t_{j-1}, y^{j-1}) + 37f(t_{j-2}, y^{j-2}) \\ &- 9f(t_{j-3}, y^{j-3})\big). \end{aligned}$$

Die Startwerte y^1, y^2 und y^3 sollten mit einem Adams-Bashforth-Verfahren vierter Ordnung berechnet werden. Wenden Sie die beiden Löser auf das Beispiel

$$y'(t) = \lambda y(t) - (\lambda + 1)e^{-t}, \quad t \in (0, 2], \quad y(0) = 1,$$

an mit einer Konstanten $\lambda < 0$. Die Lösung des Anfangswertproblems ist gegeben durch $y(t) = e^{-t}$ (also unabhängig von λ). Füllen Sie die Tabelle 1.9 mit $|y_h(2) - e^{-2}| = |y^{2/h} - e^{-2}|$ aus. Generieren Sie eine grafische Ausgabe und interpretieren Sie Ihre Ergebnisse. Was können Sie über die Konsistenzordnung sagen, wenn Sie

	$\lambda = -2, \lvert y^{2/h} - e^{-2}\rvert$		$\lambda = -20, \lvert y^{2/h} - e^{-2}\rvert$	
h	`odesolveAB4`	`odesolveABM3`	`odesolveAB4`	`odesolveABM3`
2^{-3}				
2^{-4}				
2^{-5}				
2^{-6}				
2^{-7}				

TABELLE 1.9. Vergleich der numerischen Lösung $y^{2/h}$ mit der exakten Lösung e^{-2}.

sich Ihre numerischen Ergebnisse anschauen? Die Funktionen `odesolveABM3` und `odesolveAB4` sollen keine grafische Ausgabe enthalten. Dokumentieren Sie Ihren Code. Ferner soll es möglich sein, Beschreibungen für die von Ihnen implementierten Funktionen mit Hilfe des MATLAB-Befehls `help` zu erhalten.

1.8. **Steife Systeme.** Steife Systeme von Differentialgleichungen kommen bei Prozessen mit stark unterschiedlichen Abklingzeiten vor. Beispiele sind Diffusions-Wärmeleitungsvorgänge und chemische Reaktionsgleichungen.

1.8.1. *Einleitung.* Charakteristisch für steife Differentialgleichungen sind Prozesse mit stark unterschiedlichen Abklingzeiten. Wir betrachten dazu das Problem

$$z'(t) = Az(t) + b, \quad t > 0, \quad z(0) = z^\circ \tag{1.26}$$

mit $A \in \mathbb{R}^{n\times n}$ und $z^\circ, b \in \mathbb{R}^n$. Ist A digonalisierbar, das heißt, gibt es ein $S \in \mathbb{R}^{n\times n}$ mit $\det S \neq 0$ und

$$S^{-1}AS = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) =: \Lambda \in \mathbb{R}^{n\times n},$$

so folgt aus (1.26) die Identität

$$S^{-1}z'(t) = S^{-1}ASS^{-1}z(t) + S^{-1}b,$$

und mit $y(t) := S^{-1}z$ daher

$$y'(t) = \Lambda y(t) + S^{-1}b, \quad t > 0. \tag{1.27}$$

Die Gleichung (1.27) führt auf entkoppelte Gleichungen der Form

$$y'(t) = \lambda y(t) + c, \quad t > 0.$$

Da das asymptotische Verhalten von (1.27) durch das homogene Problem bestimmt wird, betrachten wir oft

$$y'(t) = \lambda y(t), \quad t > 0, \quad y(0) = y^\circ \in \mathbb{R}^n \tag{1.28}$$

oder auch spezieller

$$y'(t) = \lambda y(t), \quad t > 0, \quad y(0) = y^\circ \in \mathbb{R}.$$

Das System (1.26) heißt *steif*, falls alle Komponenten von der Lösung für $t \to \infty$ abklingen, dies jedoch mit sehr unterschiedlicher Geschwindigkeit. Bei (1.28) bedeutet dieses für die Eigenwerte $\lambda_i \in \mathbb{C}$, $1 \le i \le n$, dass

$$\Re e(\lambda_i) < 0 \quad \text{und} \quad \max_{1\le i,j\le n} \frac{|\lambda_i|}{|\lambda_j|} \gg 1.$$

Übungsaufgabe 15. Zeigen Sie, dass das Differentialgleichungssystem aus Beispiel 1.2 steif ist. Verwenden Sie dazu Übungsaufgabe 1.

Wir haben bereits anhand von Programm 1 gesehen, dass explizite Methoden in der Regel für steife Probleme ungeeignet sind. Dieses Phänomen lässt sich wie folgt erklären. Die Anwendung eines expliziten Einschrittverfahren auf

$$y'(t) = \lambda y(t) \quad \text{mit } \lambda < 0 \tag{1.29}$$

führt auf die Rekursion

$$y^{j+1} = g(h\lambda)y^j, \quad j = 0, 1, \dots,$$

wobei die *Stabilitätsfunktion* g vom Verfahren abhängt. In Beispiel 1.23 haben wir gesehen, dass

- $g(z) = 1 + z$ beim Eulerverfahren,
- $g(z) = 1 + z + z^2/2!$ beim verbesserten Eulerverfahren und
- $g(z) = 1+z+z^2/2!+z^3/3!+z^4/4!$ beim klassischen Runge-Kutta-Verfahren

gelten. In diesem Beispiel ist also g eine abgebrochene Potenzreihe von e^z, also ein Polynom, welches wir mit p bezeichnen. Es lässt sich zeigen (siehe [1, Bemerkung 11.32]), dass die Stabilitätsfunktion eines m-stufigen Runge-Kutta-Verfahrens ein Polynom m-ten Grades in $h\lambda$ ist.

Wegen $e^z \to 0$ für $z \to \infty$ ($z = h\lambda$ mit $\lambda < 0$), aber $p(z) \to \pm\infty$ für $z \to \infty$ ergibt sich ein Problem bei den Einschrittverfahren. Damit folgt $e^z \approx p(z)$ nur für negative z mit $|z|$ hinreichend klein. Wegen $z = h\lambda$ muss daher die Schrittweite h sehr klein gewählt, insbesondere für betragsmäßig große λ-Werte.

Übungsaufgabe 16. Wie lauten das implizite Eulerverfahren und die Trapezmethode für das Problem (1.29)? Geben Sie jeweils die Stabilitätsfunktion an. Vergleichen und diskutieren Sie Ihr Ergebnis.

Bemerkung 1.28. Explizite Einschrittverfahren sind zur Behandlung steifer Probleme ungeeignet. Diese Aussage gilt auch für explizite Mehrschrittverfahren. Implizite Methoden haben dagegen ein besseres Stabilitätsverhalten. ◊

Übungsaufgabe 17. Gegeben sei das Anfangswertproblem

$$y'''(t) + y'(t) = ty(t), \quad t > 2, \quad y(2) = 0, \quad y'(2) = 2, \quad y''(2) = 2. \tag{1.30}$$

a) Transformieren Sie (1.30) in ein System von Differentialgleichungen erster Ordnung.
b) Berechnen Sie eine numerische Lösung für das unter Teil a) erhaltene System an $t = 2.5$ mit einem Schritt des impliziten Eulerverfahrens.

Programm 5. Schreiben Sie ein Programm zur numerischen Lösung des nichtlinearen Anfangswertproblems

$$y'(t) = t\,\sin(y(t)), \quad t \in [0, 5], \quad y(0) = y_\circ,$$

indem Sie die Trapezmethode verwenden. Zur Lösung der nichtlinearen Probleme setzen Sie das Newtonverfahren ein. Als Abbruchkriterium für das Newtonverfahren verwenden Sie die Toleranz 10^{-6}, das heißt $|\delta| < 10^{-6}$, wobei δ die Newtonkorrektur ist.

Verfahren	Stabilitätsintervall
Eulerverfahren	(-2,0)
Verbessertes Eulerverfahren	(-2,0)
Klassisches Runge-Kutta-Verfahren	(-2.78,0)
2-Schritt-Adams-Bashforth-Verfahren	(-1,0)
4-Schritt-Adams-Bashforth-Verfahren	(-0.3,0)
3-Schritt-Adams-Moulton-Verfahren	(-3,0)
Implizites Eulerverfahren	$(-\infty,0)$
Trapezmethode	$(-\infty,0)$

TABELLE 1.10. Stabilitätsintervalle von ausgewählten Verfahren.

a) Wählen Sie als Anfangsbedingung $y_\circ$ die Werte $0.5, 1, \ldots, 4.5$ und für die Schrittweite $h = 0.1$. Stellen Sie alle numerischen Lösungen grafisch in einer Abbildung dar. Vergessen Sie dabei nicht, die Achsen zu beschriften, einen Titel hinzuzufügen und eine Legende zum besseren Verständnis anzugeben. Beschreiben Sie Ihre Beobachtungen.
b) Verifizieren Sie die von Ihnen erhaltenen numerischen Ergebnisse, indem Sie mit Hilfe des MATLAB-Lösers `ode45` eine numerische Vergleichslösung berechnen. Stellen Sie die Lösung grafisch dar. Beschriften Sie dabei die Achsen, fügen Sie einen Titel für die Grafik ein und ergänzen Sie die Grafik durch eine Legende zum besseren Verständnis.
c) Erstellen Sie eine Dokumentation für Ihre Ergebnisse. Erläutern Sie darin, wie Sie den nächsten Schritt bei der Trapezmethode konkret berechnen und dabei das Newtonverfahren einsetzen.

1.8.2. *Stabilitätsintervalle.* Um das Dämpfungsverhalten von Verfahren bei steifen Prolemen besser zu beschreiben, definieren wir Stabilitätsintervalle (oder Stabilitätsgebiete in der komplexen Ebene). Allgemein haben wir bei Einschrittverfahren die Rekursion

$$y^{j+1} = g(h\lambda)y^j, \quad \lambda < 0, \tag{1.31}$$

wobei g die Stabilitätsfunktion des verwendeten Verfahrens bezeichnet. Es soll $|g(z)| < 1$ für möglichst große Bereiche von negativen z-Werten gelten.

Definition 1.29. *Sei g die Stabilitätsfunktion zu einem gegebenen Einschrittverfahren. Das größte Intervall $I = (-a, 0) \subset \mathbb{R}$, für das gilt*

$$z \in I \quad \Longrightarrow \quad |g(z)| < 1,$$

heißt Stabilitätsintervall *des Verfahrens.*

Die Größe des Intervalls I ist dann in Maß für die Stabilität des Verfahrens bei Anwendung auf steife Systeme. In Tabelle 1.10 sind die Stabilitätsintervalle einiger ausgewählter Verfahren dargestellt.

Wenden wir auf (1.29) ein lineares Mehrschrittverfahren an, so ist die Charakterisierung der Stabilitätsintervalle komplizierter als bei Einschrittverfahren. Eine

Darstellung in der Form (1.31) ist dann nämlich nicht mehr möglich. Unter Verwendung der Theorie der Differenzengleichung lässt sich aber auch hier eine Analyse durchführen. Wir verweisen zum Beispiel auf [1, Abschnitt 11.9.2].

1.8.3. *Rückwärtsdifferenzenmethoden.* Diese Mehrschrittverfahren lassen sich wie folgt beschreiben. Dabei steht 'BDF' für *Backward-Differentiation-Formula*.

Algorithmus 9 (m-Schritt-BDF-Verfahren)**.**

1: Wähle Schrittweite $h = (T - t_\circ)/k$ mit $k \in \mathbb{N}$, Koeffizienten a_ℓ $(0 \le \ell \le m)$ und Startwerte $y^0, \dots, y^{m-1} \in \mathbb{R}^n$.
2: **for** $j = 0, \dots, k - m$ **do**
3: Berechne y^{j+m} aus

$$\sum_{\ell=0}^{m} a_\ell y^{j+\ell} = h f(t_{j+m}, y^{j+m});$$

4: **end for**

Das Verfahren ist implizit. Der Name ist dadurch begründet, dass die Formel als Differenzenquotient aufgefasst werden kann:

$$\frac{a_0 y^j + a_1 y^{j+1} + \ldots + a_k y^{j+k}}{h} = f(t_{j+k}, y^{j+k}) \approx f(t_{j+k}, y(t_{j+k})) = y'(t_{j+k}).$$

Zur Herleitung werden Interpolationsformeln verwendet. Sei $p_m \in \mathbb{P}_m$ das Lagrange-Interpolationspolynom, das die Werte

$$(t_j, y^j), \dots, (t_{j+m}, y^{j+m})$$

interpoliert. Hierbei bezeichnet $\mathbb{P}_m$ die Menge aller auf $\mathbb{R}$ definierten Polynome vom Grad kleiner oder gleich m. Sei also

$$p_m(t) = \sum_{\ell=0}^{m} y^{j+\ell} L_{\ell m}(t),$$

wobei wir mit $L_{\ell m}$, $0 \le \ell \le m$, die Lagrange-Fundamentalpolynome zu den Stützstellen $t_j, \dots, t_{j+m}$ bezeichnen. Der Ansatz ist dann

$$p_m'(t) = f(t_{j+m}, y^{j+m}), \tag{1.32}$$

um die Koeffizienten des m-Schritt-BDF-Verfahrens zu bestimmen.

Beispiel 1.30. Das Interpolationspolynom für $m = 2$ ist durch

$$p_2(t) = \frac{(t - t_{j+1})(t - t_{j+2})}{2h^2}\, y^j + \frac{(t - t_j)(t - t_{j+2})}{-h^2}\, y^{j+1} + \frac{(t - t_j)(t - t_{j+1})}{2h^2}\, y^{j+2}.$$

gegeben. Wegen

$$p_2'(t_{j+2}) = \frac{1}{h}\left(\frac{1}{2}\, y^j - 2y^{j+1} + \frac{3}{2}\, y^{j+2}\right)$$

erhalten wir mit (1.32) die Identität

$$\frac{3}{2}\, y^{j+2} - 2y^{j+1} + \frac{1}{2}\, y^j = h f(t_{j+2}, y^{j+2})$$

für $j = 0, \dots, k - 2$. ◊

Die Rückwärtsdifferenzenmethoden sind für steife Systeme gut geeignet.

2. Randwertaufgaben gewöhnlicher Differentialgleichungen

Nun wollen wir uns Randwertaufgaben für gewöhnliche Differentialgleichungen zuwenden. Dabei konzentrieren wir uns vor allem auf lineare Differentialgleichungen zweiter Ordnung.

2.1. **Grundlegende Aussagen aus der Analysis.** Eine allgemeine Differentialgleichung zweiter Ordnung hat die Form

$$F(x, u(x), u'(x), u''(x)) = 0, \quad x \in (a, b) \subset \mathbb{R}.$$

Von großer Bedeutung sind die *linearen Differentialgleichungen zweiter Ordnung*:

$$(Lu)(x) = -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad x \in (a, b). \tag{2.1}$$

Bei nichtlinearen Differentialgleichungen wird die Gleichung

$$-u''(x) + b(x)u'(x) + f(x, u(x)) = 0, \quad x \in (a, b), \tag{2.2}$$

eine *semilineare Differentialgleichung zweiter Ordnung* und die Gleichung

$$-u''(x) + f(x, u(x), u'(x)) = 0, \quad x \in (a, b), \tag{2.3}$$

eine *quasilineare Differentialgleichung zweiter Ordnung* genannt.

Bemerkung 2.1. Für (2.1) gibt es eine geschlossene Theorie, die sich unter geeigneten Voraussetzungen auch noch für (2.1) verwenden lässt, indem lokale Argumente verwendet werden. Problem (2.3) ist im Allgemeinen schwierig. ◊

Bei Randwertproblemen werden zusätzlich zur gegebene Differentialgleichung Randbedingungen gestellt. Wir beschränken uns hier auf lineare, entkoppelte Randbedingungen. Es wird wie folgt unterschieden:

- $u(a) = \alpha$ und $u(b) = \beta$ (*Dirichlet-Bedingung* oder *Randbedingung erster Art*),
- $u'(a) = \alpha$ und $u'(b) = \beta$ (*Neumann-Bedingung* oder *Randbedingung zweiter Art*)
- $a_1u(a) + a_2u'(a) = \alpha$ und $a_3u(b) + a_4u'(b) = \beta$ (*Robin-Bedingung* oder *Randbedingung dritter Art*),

wobei α, β, $a_1, \ldots, a_4$ reelle Zahlen sind. Werden verschidene Typen von Randbedingungen gemischt (also zum Beispiel eine Dirichlet-Bedingung an $x = a$ und eine Neumann-Bedingung an $x = b$, so heißen die Randebedingungen *gemischt*.

Bemerkung 2.2 (Homogenisierung)**.** Der Ansatz

$$u(x) = v(x) + \alpha \frac{x - b}{b - a} + \beta \frac{x - a}{b - a}$$

führt zu einem Randwertproblem für v mit $v(a) = v(b) = 0$, sofern $u(a) = \alpha$ und $u(b) = \beta$ gelten. Wir betrachten daher homogene Randbedingungen. ◊

Bemerkung 2.3 (Intervalltransformation)**.** Durch die Transformation $[0, 1] \ni \xi \mapsto x(\xi) = (b - a)\xi + a \in [a, b]$ können wir vom Intervall $(0, 1)$ ausgehen. ◊

Wir betrachten im Weiteren das Randwertproblem:

$$(Lu)(x) = -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad x \in (0, 1), \tag{2.4a}$$

$$u(0) = u(1) = 0 \tag{2.4b}$$

mit $b, c, f \in C(\overline{\Omega})$.

Beispiel 2.4. Gegeben sei das Randwertproblem $u''(x) + u(x) = 0$. Die allgemeine Lösung ist gegeben durch $u(x) = c_1 \sin x + c_2 \cos x$ mit $c_1,\, c_2 \in \mathbb{R}$.

a) Bei den Randbedingungen $u(0) = u(1) = 1$ erhalten wir die Koeffizienten $c_1 = -\cos(1)/\sin(1)$ und $c_2 = 1$. Es existiert also eine eindeutige Lösung.
b) Aus den Dirichlet-Bedingungen $u(0) = 1$ und $u(\pi) = -2$ ergibt sich der Widerspruch $c_2 = 1$ und $c_2 = 2$. Es gibt also keine Lösung.
c) Gelten $u(0) = 1$ und $u(\pi) = -1$, so ist $c_1 \in \mathbb{R}$ beliebig und $c_2 = 1$. Es gibt in diesem Fall also keine eindeutige Lösung. ◊

Der folgende Satz folgt aus der Theorie gewöhnlicher Differentialgleichungen.

Satz 2.5. *Besitzt das homogene Problem nur die triviale Lösung, so ist* (2.4) *eindeutig lösbar. Gilt* $c(x) \geq 0$ *für alle* $x \in [0,1]$*, dann hat das zu* (2.4) *gehörende homogene Problem nur die triviale Lösung.*

Mithilfe der Variation der Konstanten ergibt sich eine Lösungsdarstellung

$$u(x) = \int_0^1 G(x,t) f(t)\,\mathrm{d}t$$

mit der Greenschen Funktion G. Die Funktion kann wie folgt charakterisiert werden:

a) $G = G(x,t)$ genügt als Funktion von x für $x \neq t$ der homogenen Differentialgleichung.
b) $G = G(x,t)$ genügt als Funktion von x den Randbedingungen.
c) G ist stetig, $\frac{\partial G}{\partial x}$ hat für $x = t$ eine Sprungstelle der "Größe" $1/$(Koeffizient der zweiten Ableitung).

Beispiel 2.6. Zu $Lu = -u''$ mit $u(0) = u(1) = 0$ lautet die Greensche Funktion

$$G(x,t) = \begin{cases} x(1-t), & 0 \leq x \leq t \leq 1, \\ t(1-x), & 0 \leq t \leq x \leq 1. \end{cases}$$

Damit lautet die Lösung von $-u'' = f$ in $(0,1)$ und $u(0) = u(1) = 0$ wie folgt:

$$u(x) = \int_0^x t(1-x) f(t)\,\mathrm{d}t + \int_x^x (1-t) f(t)\,\mathrm{d}t.$$

Die stückweise Auswertung des Integrals entspricht der Eigenschaft c) der Greenschen Funktion. ◊

Satz 2.7 (Vergleichsprinzip)**.** *Es gelte* $c(x) \geq 0$ *für* $x \in [0,1]$.

a) *Gilt* $f \leq 0$ *in* $[0,1]$*, so ist die Lösung von* (2.4) *nichtpositiv.*
b) *Gelten* $Lv \leq Lw$ *in* $[0,1]$*,* $v(0) \leq w(0)$ *und* $v(1) \leq w(1)$*, so folgt* $v \leq w$ *in* $[0,1]$.

Bemerkung 2.8. a) Wenn $c > 0$ in $(0,1)$ gilt, so folgt Teil a) aus Satz 2.7 sofort daraus, dass u nicht in $(0,1)$ ein positives Maximum annehmen kann: Ist $x_0 \in (0,1)$ eine Maximalstelle, so erhalten wir den Widerspruch

$$0 \overset{(2.4)}{=} \underbrace{-u''(x_0)}_{\geq 0} + b(x_0) \underbrace{u'(x_0)}_{=0} + \underbrace{c(x_0)}_{>0} \underbrace{u(x_0)}_{>0} - \underbrace{f(x_0)}_{\leq 0} > 0$$

Für einen Beweis von Satz 2.7 verweisen wir zum Beispiel auf das Buch [5].

b) Die Eigenschaft b) aus Satz 2.7 heißt *inverse Monotonie* oder *Isotonie* von L. Man sagt auch, dass der Differentialoperator L einem Maximumprinzip genügt.

Existenzaussagen bei nichtlinearen Differentialgleichungen werden meistens unter der Verwendung von Fixpunktsätzen bewiesen.

2.2. **Das klassische Differenzenverfahren.** Das Ziel ist es, das Randwertproblem (2.4) mit Hilfe einer Diskretisierung mit finiten Differenzen auf ein endlichdimensionales Problem (und damit endlichdimensionales Gleichungssystem zu bringen.

Das Vorgehen ist wie folgt:

1) Ersetze das kontinuierliche Gebiet $\Omega = (0,1) \subset \mathbb{R}$ durch eine diskrete Menge von Gitterpunkten:

Schrittweite:	$h = 1/N > 0$ mit $N \in \mathbb{N}$,
Gitterpunkte:	$x_i = ih \in \overline{\Omega}$ für $i = 0, \dots, N$,
Menge der inneren Gitterpunkte:	$\Omega_h = \{x_1, \dots, x_{N-1}\}$,
Menge der Randgitterpunkte:	$\Gamma_h = \{x_0, x_N\}$,
Gesamtgitter:	$\overline{\Omega}_h = \Omega_h \cup \Gamma_h$.

2) In jedem Gitterpunkt x_i wird Lu durch einen auf $\overline{\Omega}_h$ definierten Differenzenquotienten ersetzt. Approximationen für die erste Ableitung sind dabei
   - die *Vorwärtsdifferenz*: $(D^+u)(x) = (u(x+h) - u(x))/h$,
   - die *Rückwärtsdifferenz*: $(D^-u)(x) = (u(x) - u(x-h))/h$,
   - die *symmetrische Differenz*: $(D^\circ u)(x) = (u(x+h) - u(x-h))/(2h)$.

   Für die zweite Ableitung wird häufig der *zentrale Differenzenquotient zweiter Ordnung*

$$(D^+D^-u)(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

   verwendet.

Wir erhalten auf diese Weise für das Problem (2.4)

$$-D^+D^-u_i + b_iD^\circ u_i + c_iu_i = f_i, \qquad i = 1, \dots, N-1, \tag{2.5a}$$

$$u_0 = u_N = 0, \tag{2.5b}$$

wobei u_i, $i = 0, \dots, N$ Näherungen für u an den Gitterpunkten x_i bezeichnen und $b_i = b(x_i)$, $c_i = c(x_i)$, $f_i = f(x_i)$ gelten. Problem (2.5) führt auf ein lineares Gleichungssystem zur Bestimmung der unbekannten Näherungen $u_1, \dots, u_{N-1}$ an den inneren Gitterpunkten. Die Werte von u an den Randpunkten sind wegen (2.5b) bekannt. Wir erhalten die Gleichungen:

$$r_iu_{i-1} + s_iu_i + t_iu_{i+1} = f_i, \quad i = 1, \dots, N-1,$$

wobei $r_i = -1/h^2 - b_i/(2h)$ für $2 \le i \le N-1$, $s_i = 2/h^2 + c_i$ für $1 \le i \le N-1$ und $t_i = -1/h^2 + b_i/(2h)$ für $1 \le i \le N_2$ gelten.

Übungsaufgabe 18. Betrachte das nichtlineare Randwertproblem

$$u''(x) = 3u(x) + x^2 + 10u^3(x), \quad x \in (0,1), \quad u(0) = u(1) = 0. \tag{2.6}$$

a) Diskretisieren Sie (2.6) unter der Verwendung des zentralen Differenzenquotienten zweiter Ordnung für die Approximation von $u''(x)$. Schreiben Sie das nichtlineare Gleichungssystem für die Näherungen u_i der Lösung u an den inneren Gitterpunkten $x_i = i/N$, $1 \le i \le N-1$.

b) Zur Lösung des nichtlinearen Gleichungssystems soll das Newtonverfahren eingesetzt werden. Wie lautet die Jacobimatrix für das Newtonverfahren?
c) Schreiben Sie einen Algorithmus für die Verwendung des Newtonverfahrens zur Lösung der nichtlinearen Gleichung in der Form eines Pseudocodes auf.

Programm 6. Implementieren Sie das Newtonverfahren zur Lösung des nichtlinearen Gleichungssystems aus Übungsaufgabe 18. Verwenden Sie als Abbruchkriterium die Toleranz 10^{-6} für die Norm der Newtonkorrektur.

Analog zu (2.4) schreiben wir (2.5) in der Form

$$L_h u_h = f_h, \quad u\big|_{\Gamma_h} = 0 \tag{2.7}$$

mit $u_h = (u_1, \dots, u_{N-1})^T \in \mathbb{R}^{N-1}$.

Um die Lösungen von (2.4) und (2.7) miteinander vergleichen zu können, führen wir mit $R_h v$ die Restriktion einer auf Ω gegebenen stetigen Funktion v auf das Gitter ein, das heißt, die Abbildung $R_h : C(\overline{\Omega}) \to \mathbb{R}^{N+1}$ ist gegeben durch $(R_h v)_i = v(x_i)$ für $0 \le i \le N$.

Definition 2.9. *Das Differenzenverfahren heißt* konvergent von der Ordnung $p > 0$ *(in der ∞-Norm), wenn*

$$\|R_h u - u_h\|_\infty = \mathcal{O}(h^p), \quad h \to 0,$$

gilt.

Der Fehler $R_h u - u_h$ genügt der folgenden Gleichung

$$L_h(R_h u - u_h) = L_h R_h u - L_h u_h = L_h R_h u - f_h = L_h R_h u - R_h(Lu).$$

Definition 2.10. *Das Differenzenverfahren heißt* konsistent von der Ordnung $p > 0$ *(in der ∞-Norm), wenn*

$$\|L_h(R_h u) - R_h(Lu)\|_\infty = \max_{1 \le i \le N-1} \big|(L_h(R_h u) - R_h(Lu))_i\big| = \mathcal{O}(h^p), \quad h \to 0,$$

Definition 2.11. *Folgt aus $L_h v_h = f_h$ und $v_h|_{\Gamma_h} = 0$ die Ungleichung*

$$\|v_h\|_\infty \le C \|f_h\|_\infty$$

für eine Konstante $C \ge 0$, die nicht von der Schrittweite h abhängt, so heißt das Differenzenverfahren stabil.

Bemerkung 2.12.
1) *Ist ein Verfahren konsistent, so bedeutet dieses, dass der Differenzenoperator gut approximiert wird.*
2) *Ist ein Differenzenverfahren konsistent von der Ordnung p und stabil, so folgt*

$$\begin{aligned}\|R_h u - u_h\|_\infty &= \|L_h^{-1}(L_h(R_h u - u_h))\|_\infty \le C \|L_h(R_h u) - L_h u_h\|_\infty \\ &= C \|L_h(R_h u) - f_h\|_\infty = C \|L_h(R_h u) - R_h(Lu)\|_\infty = \mathcal{O}(h^p).\end{aligned}$$

Aufgrund der Stabilität ist also der Schluss vom Defekt $L_h(R_h u) - R_h(Lu)$ auf die Differenz $R_h u - u_h$ möglich.
3) *Für die Randwertaufgabe* (2.4) *folgt wegen der Beschränktheit der Greenschen Funktion aus der Darstellung*

$$u(x) = \int_0^1 G(x,t) f(t)\, \mathrm{d}t$$

die Stabilitätsrelation

$$\|u\|_\infty = \max_{x\in[0,1]} |u(x)| \le C\,\|f\|_\infty.$$

Darüberhinaus folgt sogar

$$\|u\|_\infty \le C\int_0^1 |f(t)|\,\mathrm{d}t = C\,\|f\|_{L^1(\Omega)}.$$

*Man spricht von einer L^∞-L^∞-*Stabilität *beziehungsweise von einer L^1-L^∞-*Stabilität ◊

Wie wir bereits bei den Anfangswertproblemen gesehen haben, erfolgen Konsistenzuntersuchungen mit Hilfe von Taylorentwicklungen.

Lemma 2.13. *Ist $u \in C^3(\overline{\Omega})$, so folgt*

$$(D^\circ u)(x) = u'(x) + g(h) \quad \textit{mit } |g(h)| \le \frac{h^2}{6}\,\|u'''\|_\infty.$$

Im Fall von $u \in C^4(\overline{\Omega})$ erhalten wir

$$(D^+D^-u)(x) = u''(x) + g(h) \quad \textit{mit } |g(h)| \le \frac{h^2}{12}\,\|u^{(4)}\|_\infty.$$

Beweis. Wir verwenden die Taylorentwicklungen

$$u(x\pm h) = u(x) \pm u'(x)h + u''(x)\frac{h^2}{2} \pm u'''(\xi_\pm)\frac{h^3}{6},$$
$$u(x\pm h) = u(x) \pm u'(x)h + u''(x)\frac{h^2}{2} \pm u'''(x)\frac{h^3}{6} + u^{(4)}(\eta_\pm)\frac{h^4}{24}$$

mit Zwischenstellen $\xi_+, \xi_-, \eta_+, \eta_- \in \Omega$. Damit folgen die beiden Beziehungen

$$(D^\circ u)(x) = \frac{u(x+h)-u(x-h)}{2h} = u'(x) + \big(u'''(\xi_+) + u'''(\xi_-)\big)\frac{h^2}{12},$$
$$(D^+D^-u)(x) = \frac{u(x+h)-2u(x)+u(x-h)}{h^2} = u''(x) + \big(u^{(4)}(\eta_+) + u^{(4)}(\eta_-)\big)\frac{h^2}{24}.$$

Wegen $u \in C^3(\overline{\Omega})$ beziehungsweise $u \in C^4(\overline{\Omega})$ folgt die Behauptung. □

Basierend auf Lemma 2.13 lässt sich der Konsistenzfehler abschätzen. Es gilt für $i = 1, \dots, N-1$:

$$\begin{aligned}\big(L_h(R_hu) - R_h(Lu)\big)_i = &-\frac{u(x_i-h)-2u(x_i)+u(x_i-h)}{h^2}\\ &+ b(x_i)\,\frac{u(x_i+h)-u(x_i-h)}{2h} + c(x_i)u(x_i)\\ &- \big(-u''(x_i) + b(x_i)u'(x_i) + c(x_i)u(x_i)\big).\end{aligned}$$

Daraus folgt

$$\big|(L_h(R_hu) - R_h(Lu))_i\big| \le \frac{h^2}{12}\,\|u^{(4)}\|_{C(\overline{\Omega})} + \frac{h^2}{6}\,\|b\|_{C(\overline{\Omega})}\,\|u'''\|_{C(\overline{\Omega})}$$

für $1 \le i \le N-1$, das heißt, an den inneren Gitterpunkten.

Satz 2.14. *Für $u \in C^4(\overline{\Omega})$ ist das klassische Differenzenverfahren konsistent von der Ordnung $p = 2$.*

Bemerkung 2.15. Ist die Lösung von (2.4) nicht aus $C^4(\overline{\Omega})$, so zeigt eine genauere Untersuchung der Restglieder die folgende Konsistenzordnung:

$$\|L_h(R_h u) - R_h L u\|_\infty \leq C_\alpha \begin{cases} h^\alpha & \text{für } u \in C^{2,\alpha}(\overline{\Omega}), \\ h^{1+\alpha} & \text{für } u \in C^{3,\alpha}(\overline{\Omega}), \end{cases}$$

wobei

$$C^{k,\alpha}(\overline{\Omega}) = \left\{ \varphi \in C^k(\overline{\Omega}) \,\middle|\, \sup_{x,y\in\overline{\Omega}, x\neq y} \frac{|\varphi^{(k)}(y) - \varphi^{(k)}(x)|}{\|x-y\|^\alpha} < \infty \right\}$$

für $\alpha \in [0,1]$ und $k \in \mathbb{N}$ gilt. ◊

Übungsaufgabe 19. Gegeben sei die Randwertaufgabe

$$-u''(x) = \frac{\lambda}{2} e^{u(x)}, \quad x \in (0,1), \quad u(0) = u(1) = 0$$

mit $\lambda \geq 0$.

a) Wie lautet das nichtlineare Gleichungssystem, wenn wir die Randwertaufgabe mittels des zentralen Differenzenquotienten zweiter Ordnung diskretisieren? Verwenden Sie dazu die Schrittweite $h = 1/N$, $N \in \mathbb{N}$, die Gitterpunkte $x_i = ih$, $0 \leq i \leq N$, und bezeichnen Sie die Näherungen für u and x_i mit u_i.

b) Schreiben Sie das in a) erhaltene nichtlineare Gleichungssystem als Nullstellenproblem in der Form $F(u_h) = 0$ mit $u_h = (u_1, \ldots, u_{N-1})^T \in \mathbb{R}^{N-1}$. Wie lautet die Funktionalmatrix $F'(u_h)$? Welches direkte Verfahren ist zur Lösung der Gleichungen

$$F'(u_h) v_h = -F(u_h)$$

geeignet?

c) Formulieren Sie ausführlich das Newtonverfahren zur Lösung der Gleichung $F(u_h) = 0$.

Nun kommen wir zur Untersuchung der Stabilität. Eliminieren wir in (2.7) die bekannten Größen u_0 und u_N, so erhalten wir das lineares Gleichungssystem

$$\hat{L}_h u_h = f_h \tag{2.8}$$

mit $u_h = (u_1, \ldots, u_{N-1})^T \in \mathbb{R}^{N-1}$, $f_h = (f_1, \ldots, f_{N-1})^T \in \mathbb{R}^{N-1}$ und

$$\hat{L}_h = \begin{pmatrix} s_1 & t_1 & & & \\ r_2 & s_2 & t_2 & & \\ & \ddots & \ddots & \ddots & \\ & & r_{N-2} & s_{N-2} & t_{N-2} \\ & & & r_{N-1} & s_{N-1} \end{pmatrix} \in \mathbb{R}^{(N-1)\times(N-1)}. \tag{2.9}$$

Die Stabilitätsbedingung ist dann äquivalent dazu, dass es eine von der Schrittweite h unabhängige Konstante $C \geq 0$ gibt mit

$$\|\hat{L}_h^{-1}\|_\infty \leq C. \tag{2.10}$$

In (2.10) bezeichnet $\|\cdot\|_\infty$ die Zeilensummennorm. Die Elemente der Matrix $\hat{L}_h$ haben folgende Eigenschaften:

$$\left.\begin{aligned} s_i &= \frac{2}{h^2} + c_i > 0 \\ r_i &= -\frac{1}{h^2} - \frac{b_i}{2h} \\ t_i &= -\frac{1}{h^2} + \frac{b_i}{2h} \end{aligned}\right\} < 0 \quad \begin{aligned} &\text{für alle } h > 0, \\ &\text{für alle } h \le h_\circ = \frac{2}{\|b\|_{C(\overline{\Omega})}} \end{aligned}$$

Das motiviert die folgende Definition.

Definition 2.16. *Sei $A \in \mathbb{R}^{n\times n}$.*

a) *Die Matrix A heißt* L_0-Matrix, *wenn $a_{ij} \le 0$ gilt für $1 \le i, j \le n$ mit $i \ne j$.*
b) *Die Matrix A heißt* L-Matrix, *wenn A eine L_0-Matrix ist und $a_{ii} > 0$ gilt.*
c) *Eine L_0-Matrix A, für die A^{-1} existiert mit $A^{-1} \ge 0$ (das heißt, alle Elemente der Matrix A^{-1} sind nichtnegativ), heißt* M-Matrix.

Übungsaufgabe 20. Wir betrachten die folgende Differentialgleichung in Divergenzform:

$$\big(a(x)u'(x)\big)' = f(x), \quad x \in \Omega = (0,1), \quad u(0) = \alpha,\ u(1) = \beta, \tag{2.11}$$

wobei $a \in C^1(\overline{\Omega})$ mit $a(x) \ge \underline{a} > 0$ für alle $x \in \Omega$, $f \in C(\overline{\Omega})$ und $\alpha, \beta \in \mathbb{R}$ gelten. Ziel ist es , (2.11) so zu diskretisieren, dass wir eine symmetrische Koeffizientenmatrix erhalten.

a) Diskretisieren Sie das Intervall Ω für $N \in \mathbb{N}$ mit der äquidistanten Schrittweite $h = 1/N$ und Gitterpunkten $x_i = ih$, $i = 0, \dots, N$. Approximieren Sie die äußere Ableitung von $(au')'$ am inneren Gitterpunkt x_i, $1 \le i \le N-1$, durch die symmetrische Differenz, wobei Sie die Hilfsgitterpunkte $x_{i\pm1/2} = x_i \pm h/2$ verwenden. Benutzen Sie die Bezeichnungen $a_i = a(x_i)$, $f_i = f(x_i)$, $i = 0, \dots, N$, und $a_{i\pm1/2} = a(x_i \pm h/2)$, $i = 1, \dots, N-1$.
b) Diskretisieren Sie die ersten Ableitungen von u in der symmetrischen Differenz für $(au')'$ durch symmetrische Differenzen mit Schrittweite $h/2$. Wie lauten die Differenzengleichungen für das Problem (2.11)? Stellen Sie die Koeffizientenmatrix auf.
c) In den Differenzengleichungen kommen die Größen $a_{i\pm1/2}$ vor. Welcher Diskretisierungsfehler liegt vor, wenn Sie $a_{i\pm1/2}$ durch das Mittel $(a_i + a_{i\pm1})/2$ ersetzen?

Analog zur inversen Monotonie bei Differentialoperatoren (vergleiche Satz 2.7 und Bemerkung 2.8) heißt A *inversmonoton*, wenn

$$Ax \le Ay \quad \Longrightarrow \quad x \le y.$$

Dies ist äquivalent damit, dass A^{-1} existiert mit $A^{-1} \ge 0$. Daher sagt man, dass eine M-Matrix eine inversmonotone L_0-Matrix ist.

Übungsaufgabe 21. Seien $P \in \mathbb{R}^{n\times n}$ und $\|\cdot\|$ die einer Vektornorm zugeordnete Grenzennorm. Nehmen Sie an, dass

$$\sum_{j=0}^{\infty} \|P\| < \infty. \tag{2.12}$$

Zeigen Sie:

a) $I - T$ ist bijektiv.
Hinweis: Zeigen Sie, dass $I - P$ injektiv ist, zum Beispiel durch einen Widerspruchsbeweis.
b) $(I - P)^{-1} = \sum_{j=0}^{\infty} P^j$.
Hinweis: Betrachten Sie die Partialsummen $S_n = \sum_{j=0}^{n} P^j$. Konvergiert die Folge $\{S_n\}_{n\in\mathbb{N}}$? Wenn ja, ist der Grenzwert gleich $(I - P)^{-1}$?
c) Gilt $\|P\| < 1$, so ist (2.12) erfüllt. Zeigen Sie, dass

$$\|(I - P)^{-1}\| \leq \frac{1}{1 - \|P\|}$$

gilt.

Satz 2.17 (M-Kriterium)**.** *Sei $A \in \mathbb{R}^{n\times n}$ eine L_0-Matrix mit Elementen a_{ij}, $1 \leq i, j \leq n$. Dann ist A genau dann inversmonoton, wenn es einen Vektor $e > 0$ gibt mit $Ae > 0$. In diesem Fall gilt*

$$\|A^{-1}\|_\infty \leq \frac{\|e\|_\infty}{\min\limits_{1\leq i\leq n} (Ae)_i}, \tag{2.13}$$

mit $(Ae)_i = \sum_{j=1}^{n} a_{ij} e_j$.

Beweis. "⇒": Ist A inversmonoton, so können wir als majorisierendes Element $e := A^{-1}w$ mit $w = (1, \ldots, 1)^T \in \mathbb{R}^n$ wählen. Wegen $A^{-1} \geq 0$ folgt dann $e > 0$ und $Ae = w > 0$.
"⇐": Sei $e > 0$ ein Vektor mit $Ae > 0$. Dann erhalten wir

$$\sum_{j=1}^{n} a_{ij} e_j > 0, \quad i = 1, \ldots, n.$$

Nach Voraussetzung ist A eine L_0-Matrix. Es gilt also $a_{ij} \leq 0$ für $i \neq j$. Wegen $e_j > 0$, $1 \leq j \leq n$, folgt $a_{ii} > 0$ für $1 \leq i \leq n$. Damit ist $A_D = \operatorname{diag}(a_{11}, \ldots, a_{nn}) \in \mathbb{R}^{n\times n}$ invertierbar. Wir setzen $P := A_D^{-1}(A_D - A) \in \mathbb{R}^{n\times n}$. Dann ist $A = A_D(I - P)$, wobei $I \in \mathbb{R}^{n\times n}$ die Einheitsmatrix bezeichnet. Aus $A_D^{-1} \geq 0$ und $A_D - A \geq 0$ erhalten wir $P \geq 0$. Wegen $A_D^{-1} \geq 0$ und $Ae > 0$ schließen wir aus

$$(I - P)e = (I - I + A_D^{-1}A)e = A_D^{-1}Ae > 0$$

die Ungleichung $Pe < e$. Nun führen wir die folgende Norm ein:

$$\|x\|_e = \max_{1\leq i\leq n} \frac{|x_i|}{e_i} \quad \text{für } x \in \mathbb{R}^n.$$

Die zughörige Matrixnorm bezeichnen wir mit

$$\|P\|_e = \max\left\{\|Px\|_e \,\middle|\, \|x\|_e = 1\right\}.$$

Es folgt $\|e\|_e = 1$. Ist $\|x\|_e = 1$, so gilt $|x_i| \leq e_i$ für $i = 1, \ldots, n$, das heißt, $x \leq e$. Mit $P \geq 0$ erhalten wir $Px \leq Pe$ für alle $x \in \mathbb{R}^n$ mit $\|x\|_e = 1$. Damit ergibt sich — wieder mit $P \geq 0$ und mit $Pe < e$ — für die Norm von P:

$$\|P\|_e = \max_{\|x\|_e=1} \|Px\|_e = \|Pe\|_e = \max_{1\leq i\leq n} \frac{|(Pe)_i|}{e_i} = \max_{1\leq i\leq n} \frac{(Pe)_i}{e_i} < 1.$$

Damit existiert die Matrix $(I-P)^{-1}$ und es gilt

$$(I-P)^{-1} = \sum_{j=0}^{\infty} P^j. \tag{2.14}$$

Also ist $A = A_D(I-P)$ invertierbar. Aus (2.14) und $P \geq 0$ folgt $(I-P)^{-1} \geq 0$. Da auch $A_D \geq 0$ gilt, ist $A^{-1} \geq 0$. Damit haben wir gezeigt, dass A inversmonoton ist.
Nun wollen wir Ungleichung (2.13) beweisen. Sei $Av = w$. Offenbar gilt

$$w \leq \begin{pmatrix} \|w\|_\infty \\ \vdots \\ \|w\|_\infty \end{pmatrix} = \|w\|_\infty \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

mit $\|w\|_\infty = \max_{1\leq i\leq n} |w_i|$. Daher erhalten wir

$$\pm v = \pm A^{-1} w \leq \|w\|_\infty A^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Aus

$$\min_{1\leq i\leq n} (Ae)_i \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \leq Ae$$

folgt

$$0 \leq A^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \leq \frac{e}{\min\limits_{1\leq i\leq n} (Ae)_i}.$$

Somit ergibt sich insgesamt

$$\begin{aligned} \|v\|_\infty = \|A^{-1}w\|_\infty &\leq \left\| \|w\|_\infty A^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_\infty = \|w\|_\infty \left\| A^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_\infty \\ &\leq \|w\|_\infty \left\| \frac{e}{\min\limits_{1\leq i\leq n} (Ae)_i} \right\|_\infty = \frac{\|w\|_\infty \|e\|_\infty}{\min\limits_{1\leq i\leq n} (Ae)_i}, \end{aligned}$$

das heißt,

$$\|A^{-1}\|_\infty = \max_{\|w\|_\infty = 1} \|A^{-1}w\|_\infty \leq \frac{\|e\|_\infty}{\min\limits_{1\leq i\leq n} (Ae)_i},$$

was zu zeigen war.

□

Bemerkung 2.18. Oft gelingt es, einen Vektor e — ein majorisierendes Element für die Matrix A — zu finden und damit $\|A^{-1}\|_\infty$ abzuschätzen. Mit dieser Vorgangsweise können wir die Stabilität zeigen. ◊

Definition 2.19. *Sei $A \in \mathbb{R}^{n\times n}$ mit den Elementen a_{ij}, $1 \leq i,j \leq n$, gegeben.*

a) *Die Matrix A heißt* streng/strikt diagonaldominant*, wenn*

$$|a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \tag{2.15}$$

für alle $i \in \{1, \dots, n\}$ gilt. Die Matrix A heißt (schwach) diagonaldominant*, wenn in* (2.15) *das Zeichen $\geq$ steht.*

b) *Die Matrix A heißt* reduzibel*, falls disjunkte, nichtleere Mengen $\mathcal{I},\ \mathcal{J} \subset \{1, \dots, n\}$ mit folgenden Eigenschaften existieren:*

$$\mathcal{I} \cup \mathcal{J} = \{1, \dots, n\} \quad \textit{und} \quad a_{ij} = 0 \textit{ für alle } i \in \mathcal{I},\ j \in \mathcal{J}, \tag{2.16}$$

Andernfalls heißt die Matrix irreduzibel*.*

c) *Die Matrix A besitzt die* Ketteneigenschaft*, wenn für beliebige Indizes $i, j \in \{1, \dots, n\}$ eine Folge von Nichtnull-Elementen der Form*

$$a_{ii_1}, a_{i_1 i_2}, a_{i_2 i_3}, \dots, a_{i_m j}$$

existiert.

d) *Die Matrix A heißt* irreduzibel diagonaldominant*, wenn A schwach diagonaldominant ist, in mindestens einer Zeile aber die strikte Ungleichung erfüllt und A irreduzibel ist.*

Beispiel 2.20. Die Matrix

$$\begin{pmatrix} 1 & 2 & 0 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

ist reduzibel: Wir wählen $\mathcal{I} = \{1, 2\}$ und $\mathcal{J} = \{3\}$. ◊

Bemerkung 2.21.

a) Sei $A \in \mathbb{R}^{n \times n}$ mit Elementen $a_{ij} \neq 0$, $1 \leq i, j, \leq n$. Dann ist A irreduzibel.

b) In dem Buch [4, Bemerkung 10.11] finden wir eine Begründung für den Begriff reduzibel. Sei $A \in \mathbb{R}^{n \times n}$ eine reguläre reduzible Matrix mit Elementen a_{ij}, $1 \leq i, j \leq n$. Wir betrachten das lineare Gleichungssystem $Ax = b$ mit einem gegebenem Vektor $b = (b_1, \dots, b_n)^T \in \mathbb{R}^n$. Gemäß Definition 2.19 existieren disjunkte, nichtleere Mengen $\mathcal{I},\ \mathcal{J} \subset \{1, \dots, n\}$ mit (2.16). Dann folgt für alle Zeilen $i \in \mathcal{I}$

$$\sum_{j=1}^{n} a_{ij} x_j = \sum_{j \in \mathcal{I}} a_{ij} x_j = b_i \quad \text{für alle } i \in \mathcal{I}.$$

Daher lassen sich zunächst alle Komponenten x_i von x mit $i \in \mathcal{I}$ bestimmen. Dann betrachten wir die Komponenten x_i mit $i \in \mathcal{J}$:

$$\sum_{j \in \mathcal{J}} a_{ij} x_j = b_i - \sum_{j \in \mathcal{I}} a_{ij} x_j \quad \text{für alle } i \in \mathcal{J}.$$

Das Gleichungssystem lässt sich also in zwei kleinere Teilaufgaben zerlegen. Das bedeutet, dass es eine Permutationsmatrix $P \in \mathbb{R}^{n \times n}$ mit

$$PAP^T = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}$$

existiert, wobei $B_{11} \in \mathbb{R}^{k \times k}$, $B_{12} \in \mathbb{R}^{k \times (n-k)}$ und $B_{22} \in \mathbb{R}^{(n-k) \times (n-k)}$ gelten. ◊

Proposition 2.22. *Eine Matrix $A \in \mathbb{R}^{n\times n}$ ist genau dann irreduzibel, wenn A die Ketteneigenschaft besitzt.*

Beweis. Wenn eine Folge

$$a_{ii_1}, a_{i_1i_2}, \ldots, a_{i_mj} \quad \text{mit } a_{ii_1}, a_{i_1i_2}, \ldots, a_{i_mj} \neq 0$$

existiert, dann sagen wir, dass i mit j verbunden ist.

"$\Leftarrow$": Angenommen, A ist reduzibel. Dann existieren $\emptyset \neq \mathcal{I}, \mathcal{J} \subset \{1, \ldots, n\}$ mit (2.16). Sei $i \in \mathcal{I}$ und $j \in \mathcal{J}$. Aus $a_{ik} \neq 0$ folgt wegen $i \in \mathcal{I}$ und (2.16) die Bedingung $k \in \mathcal{I}$. Dann kann aber i nicht mit j verbunden sein, denn in

$$a_{i,i_1}, a_{i_1,i_2}, \ldots, a_{i_m,j}$$

folgt aus $a_{i,i_1} \neq 0$ sofort $i_\ell \in \mathcal{I}$ für $\ell = 2, \ldots, m$. Aus (2.16) erhalten wir dann aber $a_{i_mj} = 0$. Das ergibt einen Widerspruch zur Ketteneigenschaft. Also muß A irreduzibel sein.

"$\Rightarrow$" Seien A irreduzibel und $i \in \{1, \ldots, n\}$ beliebig gewählt. Definiere die Indexmenge

$$\mathcal{K} = \big\{k \in \{1, \ldots, n\} \,\big|\, i \text{ ist mit } k \text{ verbunden}\big\}.$$

Dann ist $\mathcal{K}$ nichtleer: Angenommen, es gilt $a_{ik} = 0$ für alle $k \in \{1, \ldots, n\}$. Dann können wir disjunkte, nichtleere Mengen $\mathcal{I} = \{i\}$ und $\mathcal{J} = \{1, \ldots, n\} \setminus \{i\}$ wählen mit (2.16). In diesem Fall ist A reduzibel, was im Widerspruch zur Voraussetzung steht. Also gilt $\mathcal{K} \neq \emptyset$.
Wir zeigen, dass $\mathcal{K} = \{1, \ldots, n\}$ gilt. Angenommen, für ein $j \in \{1, \ldots, n\}$ ist i nicht mit j verbunden, das heißt, $j \notin \mathcal{K}$ und $\mathcal{K} \neq \{1, \ldots, n\}$. Dann existiert ein $\ell \in \{1, \ldots, n\}$ mit

$$a_{k\ell} = 0 \quad \text{für alle } k \in \mathcal{K} \text{ und } \ell \notin \mathcal{K} \tag{2.17}$$

(sonst wäre k mit ℓ verbunden). Wegen (2.16) ist damit A reduzibel. Das ist aber ein Widerspruch zur Voraussetzung. Es bleibt (2.17) zu zeigen. Für jedes $k \in \mathcal{K}$ ist i mit k verbunden. Es gibt also eine Folge

$$a_{i,i_1}, \ldots, a_{i_m,k} \neq 0.$$

Ist auch $a_{k\ell} \neq 0$, so ist offenbar i mit ℓ vebunden, das heißt, es folgt $\ell \in \mathcal{K}$. Das widerspricht (2.17), und damit folgt die Behauptung.

□

Beispiel 2.23. Wir betrachten die $n \times n$-Matrizen

$$A_1 = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$

oder

$$A_2 = \begin{pmatrix} B & -I & & & \\ -I & B & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & B & -I \\ & & & -I & B \end{pmatrix} \text{ mit } B = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}.$$

Es gelte $1 \leq i < j \leq n$. Dann folgt für die Elemente der Matrix A_1:

$$a_{i,i+1}, a_{i+1,i+2}, \ldots, a_{j-1,j} \neq 0.$$

Im Fall von $1 \leq j < i \leq n$ folgt

$$a_{i,i-1}, a_{i-1,i-2}, \ldots, a_{j+1,j} \neq 0.$$

Damit ist A_1 irreduzibel beziehungsweise besitzt die Ketteneigenschaft. Für die Matrix A_2 lässt sich ebenfalls die Ketteneigenschaft nachweisen. Allerdings ist dieser Zugang nicht so schnell durchzuführen. Mit Hilfe der Theorie der Differenzengleichungen wird ein Beweis in dem Buch [6] geführt. ◊

In [4, Bsp. 10.12 und Lem. 10.14] sind folgenden Aussagen bewiesen.

Proposition 2.24. *Sei $A \in \mathbb{R}^{n\times n}$ mit den Elementen a_{ij}, $1 \leq i, j \leq n$, gegeben.*

a) *Sei A eine Tridiagonalmatrix. Dann ist A genau dann irreduzibel, wenn jede ihrer Nebendiagonaleinträge von null verschieden ist.*
b) *Ist A irreduzibel, so ist auch $A + D$ irreduzibel für jede Diagonalmatrix $D \in \mathbb{R}^{n\times n}$.*
c) *Ist A irreduzibel und sind $b_{ij} \in \mathbb{R}$ mit $b_{ij} \neq 0$ für $i \neq j$, so ist auch die Matrix mit den Elementen $b_{ij}a_{ij}$, $1 \leq i, j \leq n$ irreduzibel.*

Satz 2.25. *Sei $A \in \mathbb{R}^{n\times n}$ strikt diagonaldominant oder irreduzibel diagonaldominant. Dann ist A invertierbar.*

Beweis. a) Sei A strikt diagonaldominant. Angenommen, es gibt ein $x \neq 0$ mit $Ax = 0$. Sei $i \in \{1, \ldots, n\}$ mit $|x_i| = \|x\|_\infty > 0$. Wegen $Ax = 0$ gilt dann

$$\begin{aligned} |a_{ii}|\,|x_i| &= \Big| -\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j \Big| \\ &\leq |x_i|\Big(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \Big), \end{aligned} \tag{2.18}$$

was der strikten Diagonaldominanz widerspricht. Damit muss A invertierbar sein.

b) Sei A irreduzibel diagonaldominant. Angenommen, es gibt ein $x \neq 0$ mit $Ax = 0$. Wir wählen $i \in \{1, \ldots, n\}$ mit

$$|a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|, \tag{2.19}$$

und setzen

$$\mathcal{J} = \Big\{ k \in \{1, \ldots, n\} \,\Big|\, |x_k| \geq |x_i| \text{ für } 1 \leq i \leq n, \quad |x_k| > |x_j| \text{ für ein } j \in \{1, \ldots, n\} \Big\}.$$

Dann ist $\mathcal{J}$ nichtleer: Angenommen, $\mathcal{J}$ ist leer. Dann gilt $|x_1| = \ldots = |x_n| \neq 0$. Aus (2.18) erhalten wir

$$|a_{ii}| \leq \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

was aber (2.19) widerspricht. Also gilt $\mathcal{J} \neq \emptyset$.
Sei $k \in \mathcal{J}$ beliebig gewählt. Dann folgt

$$|a_{kk}| \leq \sum_{j=1}^{k-1} |a_{kj}| \frac{|x_j|}{|x_k|} + \sum_{j=k+1}^{n} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{j=1}^{k-1} |a_{kj}| + \sum_{j=k+1}^{n} |a_{kj}|.$$

Wegen der Diagonaldominanz muss $a_{kj} = 0$ gelten bei $|x_k| > |x_j|$. Offenbar folgt damit

$$a_{kj} = 0 \quad \text{für alle } k \in \mathcal{J} \text{ und } j \notin \mathcal{J}.$$

Damit ist A reduzibel, was der Annahme widerspricht.

□

Lemma 2.26. *Sei $B \in \mathbb{R}^{n\times n}$ mit $B \geq 0$. Dann existiert die Matrix $(I-B)^{-1}$ mit nichtnegativen Elementen genau dann, wenn $\varrho(B) < 1$ gilt.*

Nun können wir folgendes Resultat beweisen.

Satz 2.27. *Sei $A \in \mathbb{R}^{n\times n}$ mit $a_{ij} \leq 0$ für $i \neq j$, das heißt, A ist eine L_0-Matrix. Dann ist A genau dann eine M-Matrix, wenn folgende Bedingungen erfüllt sind:*

a) $a_{ii} > 0$ *für alle* $i \in \{1, \ldots, n\}$ *(A ist L-Matrix);*
b) $\varrho(I - A_D^{-1}A) < 1$ *für* $A_D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$.

Beweis. "⇐": Es gelte $\varrho(B) < 1$ mit $B = I - A_D^{-1}A$. Dann folgt

$$B = -\begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \cdots & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & \frac{a_{23}}{a_{22}} & \cdots & \frac{a_{2n}}{a_{11}} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \frac{a_{n-1,n}}{a_{n-1,n-1}} \\ \frac{a_{n1}}{a_{nn}} & \cdots & \cdots & \frac{a_{n-1,n}}{a_{nn}} & 0 \end{pmatrix} \geq 0.$$

Wegen Lemma 2.26 existiert die Inverse $(I-B)^{-1} = (A_D^{-1}A)^{-1}$ und ist nichtnegativ. Da A_D^{-1} regulär ist mit nichtnegativer Inverse, muss auch A invertierbar und nichtnegativ sein; denn $A = A_D(I-B)$ und $A^{-1} = (I-B)^{-1}A_D^{-1} \geq 0$. Also ist A eine M-Matrix.

"⇒": Sei A eine M-Matrix. Angenommen, es gilt $a_{ii} \leq 0$ für ein $i \in \{1, \ldots, n\}$. Damit ist A keine L-Matrix, und die i-te Spalte von $a_i \in \mathbb{R}^n$ von A ist nicht positiv. Da A eine M-Matrix ist, gilt $e_i = A^{-1}a_i \leq 0$. Wegen

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \leftarrow i\text{-te Zeile}$$

ergibt sich ein Widerspuch zu der Annahme, dass $a_{ii} \leq 0$ gilt. Also folgt, dass $A_D \geq 0$ gilt und A_D^{-1} existiert. Da A eine M-Matrix ist, ist A regulär

mit nichtnegativer Inverser. Also ist auch $I-B = A_D^{-1}A$ invertierbar. Wegen $a_{ij} \leq 0$ für alle $i \neq j$ gilt $B \geq 0$. Aus

$$(I-B)^{-1} = \big(I - (I - A_D^{-1}A)\big)^{-1} = A^{-1}A_D \geq 0$$

folgt die Aussage aus Lemma 2.26.

□

In [4, Th. 4.39 und Th. 4.43] ist die folgende Aussage bewiesen.

Satz 2.28. *Sei $B \in \mathbb{R}^{n\times n}$ mit der Spektralnorm*

$$\varrho(B) = \max\{|\lambda| \,:\, \lambda \in \mathbb{C} \text{ ist ein Eigenwert von } B\}.$$

a) *Für jede durch eine Vektornorm indizierte Matrixnorm $\|\cdot\| : \mathbb{R}^{n\times n} \to [0,\infty)$ gilt $\varrho(B) \leq \|B\|$.*
b) *Ist B symmetrisch, so folgt*

$$\varrho(B) = \|B\|_2 = \max\big\{\|Bx\|_2 \,\big|\, \|x\|_2 = 1\big\},$$

wobei hier $\|Bx\|_2$ und $\|x\|_2$ die Euklidischen Normen der Vektoren Bx beziehungsweise x bezeichnen.

Wir haben nun folgendes hinreichende Kriterium für eine M-Matrix.

Satz 2.29. *Sei die Matrix $A \in \mathbb{R}^{n\times n}$ entweder strikt diagonaldominant oder irreduzibel diagonaldominant mit $a_{ij} \leq 0$ für $i \neq j$ und $a_{ii} > 0$ für alle $i \in \{1,\dots,n\}$. Dann ist A eine M-Matrix.*

Beweis. Wir setzen $B = I - A_D^{-1}A$ mit $A_D = \operatorname{diag}(a_{11},\dots,a_{nn}) \in \mathbb{R}^{n\times n}$. Gilt $\varrho(B) < 1$, so folgt die Aussage aus Satz 2.27. Nach Voraussetzung ist A diagonaldominant. Dann folgt für die Elemente b_{ij}, $1 \leq i,j \leq n$, von B die Abschätzung

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{j=1}^{i-1} \frac{|a_{ij}|}{a_{ii}} + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{a_{ii}} \leq 1. \tag{2.20}$$

a) Sei A strikt digonaldominant. Dann folgt aus (2.20)

$$\|B\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |b_{ij}| < 1.$$

Also gilt $\varrho(B) \leq \|B\|_\infty < 1$ wegen Satz 2.28-a) und wir können Satz 2.27 anwenden.
b) Sei A irreduzibel diagonaldominant. Dann gilt (2.20) für alle $i \in \{1,\dots,n\}$, und es gibt ein $k \in \{1,\dots,n\}$ mit

$$\sum_{j=1}^{n} |b_{kj}| < 1.$$

Damit gilt $\varrho(B) \leq \|B\|_\infty \leq 1$ wegen Satz 2.28-a). Angenommen, es gilt $\varrho(B) = 1$. Sei λ ein Eigenwert von B mit $|\lambda| = 1$. Dann ist $\lambda I - B$ singulär. Wegen (2.20) gilt

$$|\lambda - b_{ii}| \geq |\lambda| - |b_{ii}| = 1 - |b_{ii}| \geq \sum_{j=1}^{i-1} |b_{ij}| + \sum_{j=i+1}^{n} |b_{ij}|, \quad 1 \leq i \leq n,$$

und die strikte Ungleichung ist mindestens für ein $k \in \{1, \dots, n\}$ erfüllt. Damit ist $\lambda I - B$ diagonaldominant. Die Matrix $\lambda I - B$ hat dieselbe Besetztheitsstruktur wir A, ist daher irreduzibel und damit irreduzibel diagonaldominant. Mit Satz 2.25 ist damit $\lambda I - B$ invertierbar, was einen Widerspruch zu der Annahme ist, dass λ ein Eigenwert von B ist.

□

Wir haben nun Hilfsmittel zur Hand, um die Stabilität von Differenzenverfahren zu untersuchen. Wir betrachten dazu die Matrix $\hat{L}_h \in \mathbb{R}^{(N-1)\times(N-1)}$; vergleiche (2.9).

a) Für $h \in (0, h_\circ]$ ist $\hat{L}_h$ eine L-Matrix, das heißt,

$$r_i \le 0,\ s_i > 0,\ t_i \le 0 \quad \text{für alle } i \in \{1, \dots, n\} \text{ und } h \in (0, h_\circ]$$

mit $r_0 = t_{N-1} = 0$.

b) Wir überprüfen die Diagonaldominanz: Wegen $r_i \le 0$ und $t_i \le 0$ erhalten wir

$$|s_i| - |r_i| - |t_i| = s_i + r_i + t_i \ge c_i \ge 0, \quad i = 2, \dots, N-2,$$

$$|s_1| - |t_1| = s_1 + t_1 = c_1 + \underbrace{\frac{1}{h^2} + \frac{b_1}{2h}}_{=-r_1 \ge 0} \ge c_1 \ge 0,$$

$$|s_{N-1}| - |r_{N-1}| = s_{N-1} + r_{N-1} = c_{N-1} + \underbrace{\frac{1}{h^2} - \frac{b_1}{2h}}_{=-t_{N-1} \ge 0} \ge c_{N-1} \ge 0.$$

c) Für $c(x) \ge \underline{c} > 0$ ist $\hat{L}_h$ strikt diagonaldominant, also nach Satz 2.29 eine M-Matrix. Wir wählen $e = (1, \dots, 1)^T \in \mathbb{R}^{N-1}$ als majorisierendes Element. Dann ergibt (2.13) aus Satz 2.17 die Abschätzung

$$\|\hat{L}_h^{-1}\|_\infty \le \frac{1}{\min\limits_{1 \le i \le N-1} s_i + r_i + t_i} = \frac{1}{\min\limits_{1 \le i \le N-1} c_i} \le \frac{1}{\underline{c}},$$

wobei $r_1 = t_{N-1} = 0$ gesetzt wird.

d) Im Fall von $c \ge 0$ in $\overline{\Omega}$ ist die Matrix $\hat{L}_h$ schwach diagonaldominant, wenn die Schrittweite hinreichend klein ist. Die Matrix $\hat{L}_h$ besitzt aber die Ketteneigenschaft (vergleiche Beispiel 2.23). Damit ist $\hat{L}_h$ irreduzibel diagonaldominant. Satz 2.29 ergibt wieder, dass $\hat{L}_h$ eine M-Matrix ist.

Bemerkung 2.30. Für $c \ge 0$ in $\overline{\Omega}$ bleibt die Stabilitätskonstante unklar. Wir betrachten das Randwertproblem

$$-v''(x) + b(x)v'(x) = 1, \quad x \in \Omega, \quad v(0) = v(1) = 0.$$

Aus dem Maximumprinzip folgt (vergleiche Satz 2.7), dass $v > 0$ in Ω gilt. Da L_h eine konsistente Differenzenapproximation ist, folgt aus $(R_h L w)_i = 1, 1 \le i \le N-1$, dass $(L_h(R_h v))_i \ge 1/2$ für alle $h \in (0, h_*]$. Also gilt

$$\hat{L}_h \begin{pmatrix} v(x_1) \\ \vdots \\ v(x_{N-1}) \end{pmatrix} \ge \frac{1}{2} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Wir wenden nun Satz 2.17 mit $e = R_h v$ an. Dann erhalten wir

$$\|\hat{L}_h^{-1}\|_\infty \leq \frac{\|R_h v\|_\infty}{\min\limits_{1\leq i\leq N-1}(\hat{L}_h(R_h v))_i} \leq \frac{\|v\|_{C(\overline{\Omega})}}{\frac{1}{2}} = 2\,\|v\|_{C(\overline{\Omega})} =: C.$$

Die erhaltene Schranke $C \geq 0$ ist unabhängig von der Schrittweite. Damit ist das Differenzenverfahren stabil. ◊

Satz 2.31. *Seien die Lösung von* (2.4) *aus* $C^4(\overline{\Omega})$ *und die Schrittweite* h *hinreichend klein. Dann ist das klassische Differenzenverfahren stabil und hat die Konvergenzordnung* $p = 2$ *(in der* ∞*-Norm), das heißt, es gilt*

$$\|R_h u - u_h\|_\infty = \mathcal{O}(h^2), \quad h \to 0.$$

Übungsaufgabe 22. Vorgelegt sei das lineare Randwertproblem

$$\begin{cases} -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), & x \in (0,1), \\ u(0) = \alpha,\ u(1) = \beta. \end{cases} \tag{2.21}$$

wobei b, c, f auf $[0,1]$ stetige Funktionen mit $c(x) > 0$ für alle $x \in [0,1]$ und α, β reelle Zahlen sind. Diskretisieren Sie das Problem (2.21), indem Sie die Schrittweite $h = 1/N$, $N \in \mathbb{N}$, verwenden und indem Sie $u''(x)$ durch die zentrale Differenz zweiter Ordnung ersetzen und $u'(x)$ wie folgt diskretisieren:

$$\begin{cases} \dfrac{u(x+h) - u(x)}{h}, & \text{falls } b(x) < 0 \text{ gilt}, \\ \dfrac{u(x) - u(x-h)}{h}, & \text{falls } b(x) \geq 0 \text{ gilt}. \end{cases} \tag{2.22}$$

Welches lineare Gleichungssystem erhalten Sie? Wann ist die Koeffizientenmatrix des linearen Gleichungssytems strikt diagonaldominant?
Die Differenzenapproximation (2.22) heißt *Aufwind-(Upwind-)Differenzenquotient.* Diese Technik ist insbesondere bei *singulär gestörten Problemen*

$$-\varepsilon u''(x) + b(x)u(x) + c(x)u(x) = f(x), \quad x \in (0,1),$$

mit $0 < \varepsilon \ll 1$ sehr hilfreich.

2.3. **Andere Randbedingungen.** In diesem Abschnitt wollen wir kurz auf andere Randbedingungen eingehen.

a) Neumann-Randbedingungen: $u'(0) = \alpha$ und $u'(1) = \beta$. Die Differenzengleichungen für $i = 1$ lauten dann

$$-u_2 + 2u_1 - u_0 + \frac{b_1 h}{2}(u_2 - u_0) + c_1 h^2 u_1 = f_1 h^2. \tag{2.23}$$

Hier ist u_0 unbekannt. Nun gibt es zwei Strategien:

- Wir verwenden die Differenzengleichung auch für $i = 0$. Das entspricht einer Diskretisierung des Randwertproblems (2.4) an $x = 0$. Wir erhalten

$$-u_1 + 2u_0 - u_{-1} + \frac{b_0 h}{2}(u_1 - u_{-1}) + c_0 h^2 u_1 = f_0 h^2. \tag{2.24}$$

Die Randbedingung $u'(0) = \alpha$ approximieren wir durch die symmetrische Differenz

$$\alpha = u'(0) = \frac{u(h) - u(-h)}{2h} + \mathcal{O}(h^2) \approx \frac{u_1 - u_{-1}}{2h}.$$

Damit können wir die unbekannte Größe u_{-1} in (2.24) durch $u_{-1} = u_1 + 2\alpha h$ eliminieren.

- Wir approximieren die Randbedingung an $x = 0$ durch die Vorwärtsdifferenz:

$$\alpha = u'(0) = \frac{u(h) - u(0)}{h} + \mathcal{O}(h) \approx \frac{u_1 - u_0}{h}.$$

Nun lässt sich in (2.23) die unbekannte Größe u_0 durch $u_0 = u_1 - \alpha h$ ersetzen.

b) Periodische Randbedingungen: $u(0) = u(1)$. Nun sind u_0 und u_N unbekannt, die aber beide übereinstimmen. Es ist daher nur eine zusätzliche Gleichung notwendig, zum Beispiel für $i = 0$:

$$-u_1 + 2u_0 - u_{-1} + \frac{b_0 h}{2}(u_1 - u_{-1}) + c_0 h^2 u_1 = f_0 h^2.$$

Da die Lösung periodisch ist, können wir u_{-1} durch u_{N-1} ersetzen.

Literatur

[1] W. Dahmen und A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. zweite korrigierte Auflage, Springer, Berlin, 2008

[2] P. Deuflhard and F. Bornemann. *Numerische Mathematik 2. Gewöhnliche Differentialgleichungen*. 3. Auflage, Walter de Gruyter, Berlin, 2008

[3] G.H. Golub und J.M. Ortega. *Wissenschaftliches Rechnen und Differentialgleichungen. Eine Einführung in die Numerische Mathemetik*. In deutscher Sprache herausgegeben von R.D. Grigorieff. Berliner Studienreihe zur Mathematik, Band 6. Herdermann-Verlag, Berlin, 1992

[4] R. Plato. *Numerische Mathemtik kompakt*. Vieweg+Teubner Verlag, Wiesbaden, 2010

[5] M.H. Protter and H.F. Weinberger. *Maximum Principles in Differential Equations*. Prentice-Hall, Eaglewood Cliffs, 1976

[6] J.M. Ortega and W.C. Rheinboldt. Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York, 1970

[7] J. Stoer und R. Bulirsch. *Numerische Mathematik 2*. 3. Auflage, Springer, Berlin, 1990

S. Volkwein, Fachbereich Mathematik und Statistik, Universität Konstanz, Universitätsstrasse 10, D-78457 Konstanz, Deutschland

E-mail address: Stefan.Volkwein@uni-konstanz.de