Numerik der Differentialgleichungen

Skript zur Vorlesung
von
Helmut Harbrecht
im
Herbstsemester 2020

Author
Helmut Harbrecht

Stand: 21. November 2022

Vorwort

Diese Mitschrift kann und soll nicht ganz den Wortlaut der Vorlesung wiedergeben. Sie soll das Nacharbeiten des Inhalts der Vorlesung erleichtern. Darstellung und Umfang des Stoffes entsprechen denen der Standardwerke der Numerischen Mathematik, wie beispielsweise:

- M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner-Verlag
- J. Stoer und R. Bulirsch: *Numerische Mathematik I+II*, Springer-Verlag

Als vertiefende Literatur zur Numerik der gewöhnlichen Differentialgleichungen ist folgendes Buch empfehlenswert:

- K. Strehmel und R. Weiner: *Numerik gewöhnlicher Differentialgleichungen*, Teubner-Verlag

Dahingegen seien dem Leser zur Vertiefung des Stoffes zur Numerik der partiellen Differentialgleichungen folgende zwei Bücher nahegelegt:

- D. Braess: *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*, Springer-Verlag
- W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner-Verlag

Inhaltsverzeichnis

1 Gewöhnliche Differentialgleichungen 4
1.1 Motivation . . . 4
1.2 Theoretische Grundlagen . . . 6

2 Einschrittverfahren 10
2.1 Eulersches Polygonzug-Verfahren . . . 10
2.2 Konsistenz . . . 11
2.3 Konvergenz von Einschrittverfahren . . . 13
2.4 Runge-Kutta-Verfahren . . . 15
2.5 Schrittweitensteuerung . . . 21
2.6 Steife Differentialgleichungen . . . 26

3 Mehrschrittverfahren 30
3.1 Definition . . . 30
3.2 Explizite k-Schrittverfahren . . . 32
3.3 Implizite k-Schrittverfahren . . . 34
3.4 Stabilität . . . 37
3.5 Konvergenz von Mehrschrittverfahren . . . 40
3.6 Schrittweitensteuerung . . . 44
3.7 Steife Differentialgleichungen . . . 48

4 Partielle Differentialgleichungen 51
4.1 Beispiele . . . 51
4.2 Charakterisierung . . . 53
4.3 Maximumprinzip . . . 55

5 Finite-Differenzen-Verfahren 58
5.1 Poisson-Gleichung . . . 58
5.2 Beliebige elliptische Differentialoperatoren . . . 62
5.3 Diskretes Maximumprinzip . . . 63
5.4 Konvergenz . . . 64
5.5 Iterative Lösung . . . 67
5.5.1 Verfahren der konjugierten Gradienten . . . 67
5.5.2 GMRES-Verfahren . . . 71
5.5.3 MINRES-Verfahren* . . . 72
5.6 Vorkonditionierung . . . 75
5.7 Mehrgitterverfahren* . . . 79
5.8 Parabolische Differentialgleichungen . . . 84

1. Gewöhnliche Differentialgleichungen

1.1 Motivation

Viele Aufgabenstellungen in den Naturwissenschaften führen auf *gewöhnliche Differentialgleichungen*. Im einfachsten Fall ist dabei eine differenzierbare Funktion $y = y(x)$ gesucht, deren Ableitung $y'(x)$ einer Gleichung der Form $y'(x) = f\big(x, y(x)\big)$ genügt, kurz

$$y' = f(x, y). \tag{1.1}$$

Im allgemeinen besitzt (1.1) unendlich viele Lösungen y, weshalb *Anfangsbedingungen* vorgegeben werden:

$$y(x_0) = y_0. \tag{1.2}$$

Beispiel 1.1 Für jedes $c \in \mathbb{R}$ ist die Funktion $y(x) = ce^x$ eine Lösung der gewöhnlichen Differentialgleichung $y' = y$. Die Anfangsbedingung $y(0) = c_0$ sichert die Eindeutigkeit. △

Allgemeiner betrachtet man Systeme von n gewöhnlichen Differentialgleichungen

$$\begin{aligned}
y_1'(x) &= f_1\big(x, y_1(x), y_2(x), \ldots, y_n(x)\big)\\
y_2'(x) &= f_2\big(x, y_1(x), y_2(x), \ldots, y_n(x)\big)\\
&\vdots\\
y_n'(x) &= f_n\big(x, y_1(x), y_2(x), \ldots, y_n(x)\big)
\end{aligned}$$

für n gesuchte Funktionen $y_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2, \ldots, n$. Solche Systeme schreibt man analog zu (1.1) vektoriell als

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}' := \begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{bmatrix}, \quad \mathbf{f}(x, \mathbf{y}) := \begin{bmatrix} f_1\big(x, y_1, y_2, \ldots, y_n\big) \\ f_2\big(x, y_1, y_2, \ldots, y_n\big) \\ \vdots \\ f_n\big(x, y_1, y_2, \ldots, y_n\big) \end{bmatrix}. \tag{1.3}$$

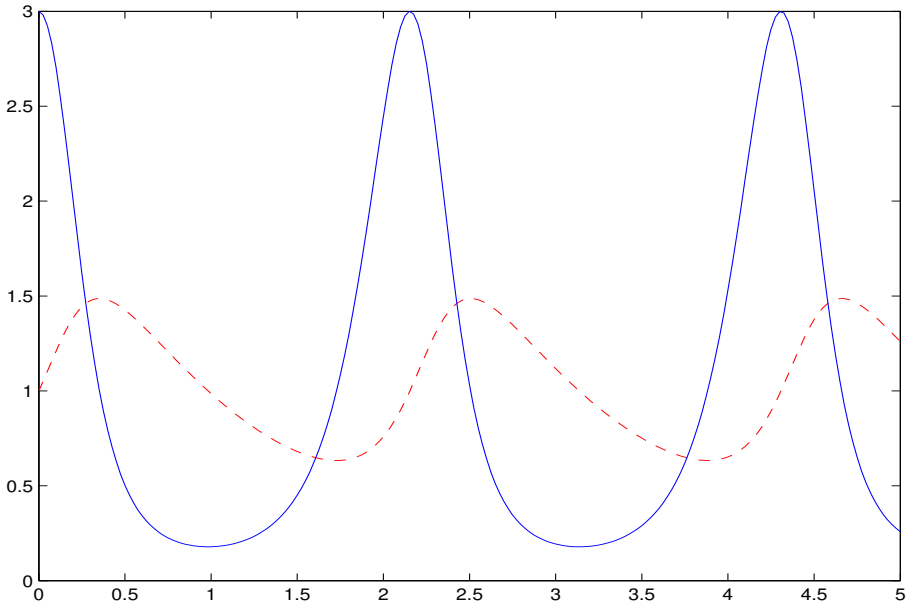
Die Anfangsbedingung (1.2) lautet nun

$$\mathbf{y}(x_0) = \mathbf{y}_0 = \begin{bmatrix} y_{0,1} \\ y_{0,2} \\ \vdots \\ y_{0,n} \end{bmatrix}.$$

Beispiel 1.2 (Räuber-Beute-Modell von Lotka-Volterra) Seien $y_1(t)$ bzw. $y_2(t)$ die Population von Beutetieren bzw. Raubtieren zum Zeitpunkt t, wobei sich die Raubtiere ausschließlich von den Beutetieren ernähren mögen. Ausgehend vom Gleichgewichtszustand $y_1(t) \equiv y_2(t) \equiv 1$ erhalten wir die Gleichungen

$$\begin{aligned} y_1'(t) &= \alpha y_1(t)\big(1 - y_2(t)\big), \\ y_2'(t) &= y_2(t)\big(y_1(t) - 1\big). \end{aligned}$$

Gegeben sei die Population zum Zeitpunkt $t = 0$. Das nachfolgende Bild gibt den Verlauf der beiden Funktionen $y_1(t)$ (durchgezogene Linie) und $y_2(t)$ (gestrichelte Line) wider:



△

Neben *gewöhnlichen Differentialgleichungen 1. Ordnung* (1.1) und (1.3), in denen nur Ableitungen erster Ordnung der unbekannten Funktion vorkommen, gibt es *gewöhnliche Differentialgleichungen m-ter Ordnung* der Form

$$y^{(m)}(x) = f\big(x, y(x), y'(x), \ldots, y^{(m-1)}(x)\big). \tag{1.4}$$

Man kann diese jedoch mit Hilfe von Hilfsfunktionen

$$\begin{aligned} z_1(x) &:= y(x) \\ z_2(x) &:= y'(x) \\ &\vdots \\ z_m(x) &:= y^{(m-1)}(x) \end{aligned}$$

stets in ein äquivalentes System gewöhnlicher Differentialgleichungen 1. Ordnung transformieren:

$$\mathbf{z}' := \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_{m-1}' \\ z_m' \end{bmatrix} = \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_m \\ f(x, z_1, z_2, \ldots, z_m) \end{bmatrix}.$$

Unter einem Anfangswertproblem für die Differentialgleichungen m-ter Ordnung versteht man die Aufgabe, eine m-mal stetig differenzierbare Funktion $y(x)$ zu finden, so dass (1.4) und

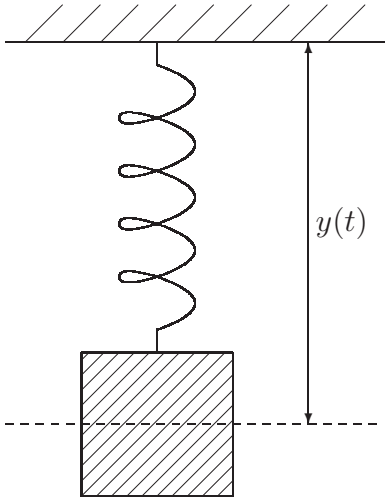
$$y^{(i)}(x_0) = y_{0,i}, \quad i = 0, 1, \ldots, m-1$$

erfüllt sind.

Beispiel 1.3 (elastische Schwingung) Ein Federpendel sei am oberen Ende fest eingespannt, während am unteren Ende ein Körper der Masse m befestigt ist. Bezüglich der Auslenkung $y(t)$ führt die Aufstellung des Kräftegleichgewichts auf die Gleichung

$$\underbrace{my''(t)}_{\text{Trägheitskraft}} + \underbrace{ry'(t)}_{\text{Reibungskraft}} + \underbrace{D\big(y(t)-\ell\big)}_{\text{Dehnungskraft}} = \underbrace{g(t)}_{\text{äußere Kraft}}.$$

Mit vorgegebenen Werten $y(0)$ und $y'(0)$, die der Lage und der Geschwindigkeit im Zeitpunkt $t = 0$ entsprechen, gelangen wir zu einem Anfangswertproblem.



△

1.2 Theoretische Grundlagen

Die Existenz und Eindeutigkeit der Lösung von Anfangswertproblemen sichert der folgende Satz:

Satz 1.4 (Picard-Lindelöf) Auf dem Streifen $S := \{(x, \mathbf{y}) : a \le x \le b, \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ stetig und genüge der Lipschitz-Bedingung

$$\|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \mathbf{z})\| \le L\|\mathbf{y} - \mathbf{z}\| \tag{1.5}$$

für alle $(x, \mathbf{y}), (x, \mathbf{z}) \in S$. Dann ist für jedes $x_0 \in [a, b]$ das Anfangswertproblem

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

eindeutig lösbar.

Beweis. Für $\mathbf{u} \in C([a,b],\mathbb{R}^n)$ sei

$$\Phi\big(\mathbf{u}(x)\big) := \mathbf{y}_0 + \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{u}(t)\big)\,\mathrm{d}t, \quad x \in [a,b].$$

Jeder Fixpunkt $\mathbf{y}(x) = \Phi(\mathbf{y}(x))$ löst das Anfangswertproblem wegen

$$\mathbf{y}'(x) = \Phi'\big(\mathbf{y}(x)\big) = \mathbf{f}\big(x,\mathbf{y}(x)\big), \qquad \mathbf{y}(x_0) = \Phi\big(\mathbf{y}(x_0)\big) = \mathbf{y}_0.$$

Umgekehrt ist jede Lösung $\mathbf{y}$ des Anfangswertproblems auch ein Fixpunkt von Φ, denn es gilt

$$\mathbf{y}(x) = \mathbf{y}(x_0) + \int_{x_0}^{x} \mathbf{y}'(t)\,\mathrm{d}t = \mathbf{y}_0 + \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{y}(t)\big)\,\mathrm{d}t = \Phi(\mathbf{y}(x)).$$

Um den Banachschen Fixpunktssatz anwenden zu können, führen wir auf $C([a,b],\mathbb{R}^n)$ eine geeignete Norm

$$|||\mathbf{u}||| := \sup_{x\in[a,b]} \big\{e^{-2Lx}\|\mathbf{u}(x)\|\big\}$$

ein. Für diese gilt

$$\begin{aligned}
e^{-2Lx}\big\|\Phi\big(\mathbf{u}(x)\big) - \Phi\big(\mathbf{v}(x)\big)\big\| &= e^{-2Lx}\left\| \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{u}(t)\big) - \mathbf{f}\big(t,\mathbf{v}(t)\big)\,\mathrm{d}t\right\| \\
&\le e^{-2Lx} \int_{x_0}^{x} \big\|\mathbf{f}\big(t,\mathbf{u}(t)\big) - \mathbf{f}\big(t,\mathbf{v}(t)\big)\big\|\,\mathrm{d}t \\
&\le Le^{-2Lx} \int_{x_0}^{x} \|\mathbf{u}(t) - \mathbf{v}(t)\|\,\mathrm{d}t \\
&= Le^{-2Lx} \int_{x_0}^{x} e^{2Lt} \underbrace{e^{-2Lt}\|\mathbf{u}(t) - \mathbf{v}(t)\|}_{\le |||\mathbf{u}-\mathbf{v}|||}\,\mathrm{d}t \\
&\le Le^{-2Lx}|||\mathbf{u} - \mathbf{v}||| \underbrace{\int_{x_0}^{x} e^{2Lt}\,\mathrm{d}t}_{=(e^{2Lx}-e^{2Lx_0})/(2L)} \\
&\le \frac{1}{2}|||\mathbf{u} - \mathbf{v}|||,
\end{aligned}$$

das heißt, Φ ist eine Kontraktion:

$$|||\Phi(\mathbf{u}) - \Phi(\mathbf{v})||| \le \frac{1}{2}|||\mathbf{u} - \mathbf{v}|||.$$

Der Banachsche Fixpunktsatz liefert uns die Existenz und Eindeutigkeit eines Fixpunktes von Φ und damit vom Anfangswertproblem. □

Bemerkungen

1. Die Funktion $f(x,y) = \sqrt{|y|}$ ist zum Beispiel nicht Lipschitz-stetig in $(0,0)$.
2. Falls f nicht Lipschitz-stetig ist, dann kann die Lösung trotzdem eindeutig sein.
3. Für $k \in \mathbb{N}_0$ und eine Menge $\Omega \subset \mathbb{R}^m$ bezeichne

$$C^k(\Omega) := \left\{\mathbf{g} : \Omega \to \mathbb{R}^n \,\middle|\, \sup_{\mathbf{x}\in\Omega} \|\partial_{\mathbf{x}}^{\boldsymbol{\alpha}}\mathbf{g}(\mathbf{x})\| < \infty \text{ für alle } |\boldsymbol{\alpha}| \le k\right\}$$

den Raum der auf Ω k-mal stetig differenzierbaren Funktionen. Dann erfüllt f im Fall $f \in C^1(S)$ die Lipschitz-Bedingung (1.5), denn es gilt

$$\|\mathbf{f}(x,\mathbf{u}) - \mathbf{f}(x,\mathbf{v})\| \leq \sup_{(x,\mathbf{y})\in S} \|\mathbf{f}_{\mathbf{y}}(x,\mathbf{y})\|\|\mathbf{u}-\mathbf{v}\|.$$

4. Aus $f \in C^k(S)$ folgt $\mathbf{y} \in C^{k+1}([a,b])$. Dies sieht man anhand von Differentiation der Gleichung $\mathbf{y}' = \mathbf{f}(x,\mathbf{y})$ nach x:

$$\begin{aligned}
\mathbf{y}'' &= \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}' = \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{f}(x,\mathbf{y}) = \mathbf{f}_x(x,\mathbf{y}) + \mathbf{f}_{\mathbf{y}}(x,\mathbf{y})\mathbf{y}' \\
\mathbf{y}''' &= \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}'' = \frac{\mathrm{d}}{\mathrm{d}x}\Big[\mathbf{f}_x(x,\mathbf{y}) + \mathbf{f}_{\mathbf{y}}(x,\mathbf{y})\mathbf{y}'\Big] \\
&= \mathbf{f}_{xx}(x,\mathbf{y}) + 2\mathbf{f}_{x\mathbf{y}}(x,\mathbf{y})\mathbf{y}' + \big(\mathbf{y}'\big)^T\mathbf{f}_{\mathbf{yy}}(x,\mathbf{y})\mathbf{y}' + \mathbf{f}_{\mathbf{y}}(x,\mathbf{y})\mathbf{y}'' \\
&\vdots
\end{aligned}$$

△

Wesentlich für die Stabilitätsuntersuchung numerischer Verfahren ist das nachfolgende Resultat, das zeigt, dass die Lösung stetig von den Anfangsdaten abhängt.

Satz 1.5 (Grönwall) Auf dem Streifen $S := \{(x,\mathbf{y}) : a \leq x \leq b,\ \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ stetig und genüge der Lipschitz-Bedingung (1.5) für alle $(x,\mathbf{y}), (x,\mathbf{z}) \in S$. Dann erfüllt für jedes $x_0 \in [a,b]$ die Lösung des Anfangswertproblems

$$\mathbf{y}' = \mathbf{f}(x,\mathbf{y}), \quad \mathbf{y}(x_0,\mathbf{u}) = \mathbf{u}$$

die Abschätzung

$$\|\mathbf{y}(x,\mathbf{u}) - \mathbf{y}(x,\mathbf{v})\| \leq e^{L|x-x_0|}\|\mathbf{u}-\mathbf{v}\|.$$

Beweis. Aus der Integraldarstellung

$$\mathbf{y}(x,\mathbf{u}) = \mathbf{u} + \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{y}(t,\mathbf{u})\big)\,\mathrm{d}t, \quad x \in [a,b]$$

folgt

$$\begin{aligned}
\|\mathbf{y}(x,\mathbf{u}) - \mathbf{y}(x,\mathbf{v})\| &= \left\|\mathbf{u}-\mathbf{v} + \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{y}(t,\mathbf{u})\big) - \mathbf{f}\big(t,\mathbf{y}(t,\mathbf{v})\big)\,\mathrm{d}t\right\| \\
&\leq \|\mathbf{u}-\mathbf{v}\| + \int_{x_0}^{x} \big\|\mathbf{f}\big(t,\mathbf{y}(t,\mathbf{u})\big) - \mathbf{f}\big(t,\mathbf{y}(t,\mathbf{v})\big)\big\|\,\mathrm{d}t \\
&\leq \|\mathbf{u}-\mathbf{v}\| + L\int_{x_0}^{x} \|\mathbf{y}(t,\mathbf{u}) - \mathbf{y}(t,\mathbf{v})\|\,\mathrm{d}t.
\end{aligned}$$

Setzen wir

$$\Phi(x) := \int_{x_0}^{x} \|\mathbf{y}(t,\mathbf{u}) - \mathbf{y}(t,\mathbf{v})\|\,\mathrm{d}t,$$

dann ergibt sich

$$\Phi'(x) = \|\mathbf{y}(x,\mathbf{u}) - \mathbf{y}(x,\mathbf{v})\|.$$

Mit anderen Worten, für $x \geq x_0$ gilt

$$\Psi(x) := \Phi'(x) - L\Phi(x) \leq \|\mathbf{u} - \mathbf{v}\|. \tag{1.6}$$

Wir betrachten nun das Anfangswertproblem

$$\Phi'(x) = \Psi(x) + L\Phi(x), \quad \Phi(x_0) = 0.$$

Es besitzt für $x \geq x_0$ die Lösung

$$\Phi(x) = e^{L(x-x_0)} \int_{x_0}^{x} \Psi(t) e^{-L(t-x_0)} \, \mathrm{d}t.$$

Wegen (1.6) folgt so für alle $x \geq x_0$ die Abschätzung

$$0 \leq \Phi(x) \leq e^{L(x-x_0)} \|\mathbf{u} - \mathbf{v}\| \int_{x_0}^{x} e^{-L(t-x_0)} \, \mathrm{d}t = \frac{1}{L} \|\mathbf{u} - \mathbf{v}\| \big(e^{L(x-x_0)} - 1\big).$$

Es ergibt sich schließlich das verlangte Resultat für $x \geq x_0$:

$$\|\mathbf{y}(x, \mathbf{u}) - \mathbf{y}(x, \mathbf{v})\| = \Phi'(x) = \Psi(x) + L\Phi(x) \leq \|\mathbf{u} - \mathbf{v}\| e^{L(x-x_0)}.$$

Ähnlich geht man im Fall $x < x_0$ vor. □

2. Einschrittverfahren

2.1 Eulersches Polygonzug-Verfahren

Zu einer ersten numerischen Methode zur Lösung des Anfangswertproblems

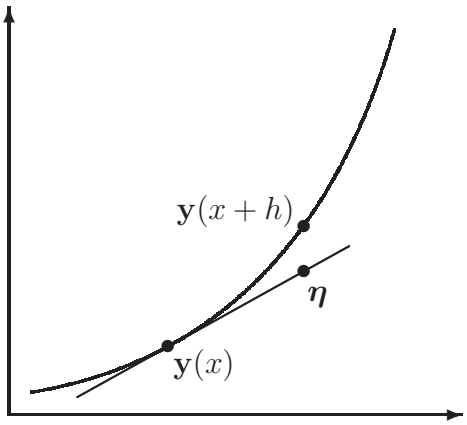
$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0 \tag{2.1}$$

kommt man durch eine einfache Überlegung: Da $\mathbf{f}\big(x, \mathbf{y}(x)\big)$ gerade die Steigung $\mathbf{y}'(x)$ der exakten Lösung $\mathbf{y}(x)$ von (2.1) ist, gilt näherungsweise für $h > 0$

$$\frac{\mathbf{y}(x+h) - \mathbf{y}(x)}{h} \approx \mathbf{y}'(x) = \mathbf{f}\big(x, \mathbf{y}(x)\big),$$

das heißt,

$$\mathbf{y}(x+h) \approx \mathbf{y}(x) + h\mathbf{f}\big(x, \mathbf{y}(x)\big).$$



Nach Wahl einer Schrittweite $h > 0$ erhält man, ausgehend von den gegebenen Anfangswerten x_0, $\mathbf{y}_0 = \mathbf{y}(x_0)$, an den äquidistanten Stützstellen $x_i := x_0 + hi$, $i = 1, 2, \ldots$, Näherungswerte $\boldsymbol{\eta}_i$ für die Werte $\mathbf{y}_i := \mathbf{y}(x_i)$ der exakten Lösung $\mathbf{y}(x)$:

Algorithmus 2.1 (Eulersches Polygonzug-Verfahren)
input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0$ und Schrittweite h
output: Approximation $\{\boldsymbol{\eta}_i\}$
① Initialisierung: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$
② für alle $i = 0, 1, \ldots$ berechne

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_i + h\mathbf{f}(x_i, \boldsymbol{\eta}_i), \qquad x_{i+1} := x_i + h$$

Definition 2.2 Ein Verfahren der Form

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_i + h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, h), \qquad x_{i+1} := x_i + h \tag{2.2}$$

heißt **Einschrittverfahren**. Die Funktion Φ heißt dabei **Inkrementfunktion**. Falls Φ nicht von $\boldsymbol{\eta}_{i+1}$ abhängt, so sprechen wir von einem **expliziten**, andernfalls von einem **impliziten** Verfahren.

Beispiel 2.3 Das Eulersche Polygonzug-Verfahren ist ein explizites Verfahren, weshalb es auch *explizites Euler-Verfahren* genannt wird. Das *implizite Euler-Verfahren* erhält man, wenn man, nicht wie beim expliziten Euler-Verfahren, den rechtsseitigen Differenzenquotienten nimmt, sondern den linksseitigen:

$$\frac{\mathbf{y}(x+h) - \mathbf{y}(x)}{h} \approx \mathbf{y}'(x+h) = \mathbf{f}\big(x+h, \mathbf{y}(x+h)\big).$$

Dies liefert die Verfahrensvorschrift

$$\boldsymbol{\eta}_{i+1} = \boldsymbol{\eta}_i + h\mathbf{f}(x_{i+1}, \boldsymbol{\eta}_{i+1}).$$

△

Bemerkung Bei expliziten Verfahren lässt sich $\boldsymbol{\eta}_{i+1}$ direkt durch Einsetzen vorher berechneter Größen ermitteln. Bei impliziten Verfahren muss hingegen ein im allgemeinen nichtlineares Gleichungssystem gelöst werden. △

2.2 Konsistenz

Um die Genauigkeit von Einschrittverfahren zu beurteilen, vergleichen wir die approximierte Lösung $\boldsymbol{\eta}_{i+1}$ aus (2.2) mit der lokal exakten Lösung $\mathbf{z}(x_{i+1})$ aus

$$\mathbf{z}' = \mathbf{f}(x, \mathbf{z}), \quad \mathbf{z}(x_i) = \boldsymbol{\eta}_i. \tag{2.3}$$

Die *Abweichung* von der lokal exakten Lösung wird beschrieben durch

$$h\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h) = \mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_{i+1} = \mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i - h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, h).$$

Definition 2.4 Der **lokale Diskretisierungsfehler** des Einschrittverfahrens zur Inkrementfunktion Φ ist definiert als

$$\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h) = \frac{\mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i}{h} - \Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, h),$$

wobei $\mathbf{z}$ die lokal exakte Lösung (2.3) bezeichnet. Das Einschrittverfahren heißt **konsistent** von der Ordnung p, falls $\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\| = \mathcal{O}(h^p)$.

Bemerkung Ist die Inkrementfunktion $\Phi(x, \mathbf{y}, \mathbf{z}, h)$ eines impliziten Verfahrens Lipschitz-stetig in $\mathbf{z}$ mit Lipschitz-Konstante L, dann genügt es, die exakte Lösung in das Verfahren einzusetzen, um die Konsistenzordnung zu untersuchen. Denn gilt

$$\big\|\widetilde{\boldsymbol{\tau}}(x_i, \boldsymbol{\eta}_i, h)\big\| = \left\|\frac{\mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i}{h} - \Phi\big(x_i, \boldsymbol{\eta}_i, \mathbf{z}(x_{i+1}), h\big)\right\| = \mathcal{O}(h^p),$$

so folgt aufgrund der Lipschitz-Stetigkeit

$$\begin{aligned}\big\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h) - \widetilde{\boldsymbol{\tau}}(x_i, \boldsymbol{\eta}_i, h)\big\| &= \big\|\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, h) - \Phi\big(x_i, \boldsymbol{\eta}_i, \mathbf{z}(x_{i+1}), h\big)\big\| \\ &\le L\|\boldsymbol{\eta}_{i+1} - \mathbf{z}(x_{i+1})\| = Lh\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\|.\end{aligned}$$

Dies impliziert aber

$$(1 - Lh)\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\| \le \big\|\widetilde{\boldsymbol{\tau}}(x_i, \boldsymbol{\eta}_i, h)\big\|,$$

also auch $\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\| = \mathcal{O}(h^p)$. Oftmals wird daher der lokale Diskretisierungsfehler auch einfach als $\widetilde{\boldsymbol{\tau}}(x_i, \boldsymbol{\eta}_i, h)$ definiert. △

Beispiel 2.5 Eine Taylor-Entwicklung liefert für das explizite Euler-Verfahren

$$\frac{\mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i}{h} = \underbrace{\mathbf{z}'(x_i)}_{=\mathbf{f}(x_i, \mathbf{z}(x_i))} + \underbrace{\boldsymbol{\tau}\big(x_i, \mathbf{z}(x_i), h\big)}_{\|\cdot\| = \mathcal{O}(h)} = \mathbf{f}(x_i, \boldsymbol{\eta}_i) + \boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)$$

mit $\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\| = \mathcal{O}(h)$. Gleiches gilt auch für das implizite Euler-Verfahren. Folglich ist das Euler-Verfahren konsistent von erster Ordnung. △

Um Verfahren höherer Ordnung zu konstruieren, ist es naheliegend, die Taylor-Entwicklung weiterzutreiben. Es gilt

$$\begin{aligned}\frac{\mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i}{h} &= \mathbf{z}'(x_i) + \frac{h}{2}\mathbf{z}''(x_i) + \underbrace{\boldsymbol{\tau}\big(x_i, \mathbf{z}(x_i), h\big)}_{\|\cdot\| = \mathcal{O}(h^2)} \\ &= \mathbf{f}\big(x_i, \mathbf{z}(x_i)\big) + \frac{h}{2}\Big\{\mathbf{f}_x\big(x_i, \mathbf{z}(x_i)\big) + \mathbf{f}_{\mathbf{y}}\big(x_i, \mathbf{z}(x_i)\big)\mathbf{z}'(x_i)\Big\} + \boldsymbol{\tau}\big(x_i, \mathbf{z}(x_i), h\big) \\ &= \mathbf{f}(x_i, \boldsymbol{\eta}_i) + \frac{h}{2}\Big\{\mathbf{f}_x(x_i, \boldsymbol{\eta}_i) + \mathbf{f}_{\mathbf{y}}(x_i, \boldsymbol{\eta}_i)\mathbf{f}(x_i, \boldsymbol{\eta}_i)\Big\} + \boldsymbol{\tau}\big(x_i, \boldsymbol{\eta}_i, h\big) \qquad (2.4)\end{aligned}$$

mit $\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\| = \mathcal{O}(h^2)$. Die Inkrementfunktion

$$\Phi(x, \mathbf{y}, h) := \mathbf{f}(x, \mathbf{y}) + \frac{h}{2}\Big\{\mathbf{f}_x(x, \mathbf{y}) + \mathbf{f}_{\mathbf{y}}(x, \mathbf{y})\mathbf{f}(x, \mathbf{y})\Big\}$$

definiert demnach ein explizites Verfahren zweiter Ordnung. In der Praxis kann man die auf diese Art gewonnenen Verfahren jedoch meist nicht brauchen, weil die Ableitungen von $\mathbf{f}$ explizit eingehen.

Einfachere Verfahren höherer Ordnung erhält man über den Ansatz

$$\Phi(x, \mathbf{y}, h) := a\mathbf{f}(x, \mathbf{y}) + b\mathbf{f}\big(x + hc, \mathbf{y} + hd\mathbf{f}(x, \mathbf{y})\big),$$

indem man die Konstanten a, b, c, d so bestimmt, dass die Taylor-Entwicklung des Fehlers $\boldsymbol{\tau}(x, \mathbf{y}, h)$ mit möglichst hoher Ordnung anfängt. Taylor-Entwicklung liefert

$$\begin{aligned}\boldsymbol{\tau}(x,\mathbf{y},h) &= \frac{\mathbf{y}(x+h)-\mathbf{y}(x)}{h} - \Phi(x,\mathbf{y},h)\\ &= \underbrace{\mathbf{y}'(x)}_{=\mathbf{f}(x,\mathbf{y})} + \frac{h}{2}\underbrace{\mathbf{y}''(x)}_{=\mathbf{f}_x(x,\mathbf{y})+\mathbf{f}_\mathbf{y}(x,\mathbf{y})\mathbf{f}(x,\mathbf{y})} - (a+b)\mathbf{f}(x,\mathbf{y})\\ &\qquad - hb\big\{c\mathbf{f}_x(x,\mathbf{y}) + d\mathbf{f}_\mathbf{y}(x,\mathbf{y})\mathbf{f}(x,\mathbf{y})\big\} + \underbrace{\boldsymbol{\varepsilon}(h)}_{\|\cdot\|=\mathcal{O}(h^2)}\\ &= (1-a-b)\mathbf{f}(x,\mathbf{y}) + h\left(\frac{1}{2}-bc\right)\mathbf{f}_x(x,\mathbf{y}) + h\left(\frac{1}{2}-bd\right)\mathbf{f}_\mathbf{y}(x,\mathbf{y})\mathbf{f}(x,\mathbf{y}) + \boldsymbol{\varepsilon}(h).\end{aligned}$$

Die Fehlerordnung $\|\boldsymbol{\varepsilon}(h)\| = \mathcal{O}(h^2)$ erhält man nun, wenn

$$a+b=1, \quad bc=\frac{1}{2}, \quad bd=\frac{1}{2}$$

erfüllt ist. Die Lösung $a = b = 1/2$, $c = d = 1$ liefert das *Verfahren von Heun*

$$\Phi(x,\mathbf{y},h) := \frac{1}{2}\Big\{\mathbf{f}(x,\mathbf{y}) + \mathbf{f}\big(x+h, \mathbf{y}+h\mathbf{f}(x,\mathbf{y})\big)\Big\},$$

die Lösung $a = 0$, $b = 1$, $c = d = 1/2$ ist das *Euler-Collatz-Verfahren*

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

In Abschnitt 2.4 werden wir dieses Konstruktionsprinzip weiter beleuchten.

2.3 Konvergenz von Einschrittverfahren

Wir wollen nun die Konvergenz der Näherungslösung $\boldsymbol{\eta}(x, h)$ für $h \to 0$ untersuchen, wobei $\boldsymbol{\eta}(x_i, h) = \boldsymbol{\eta}_i$ gilt. Wir interessieren uns für den *globalen Diskretisierungsfehler*

$$e(x,h) := \|\boldsymbol{\eta}(x,h) - \mathbf{y}(x)\|$$

bei festem x für

$$H_x := \left\{\frac{x-x_0}{n} : n \in \mathbb{N}_0\right\} \ni h \to 0.$$

Da $e(x, h)$ wie $\boldsymbol{\eta}(x, h)$ nur für $h \in H_x$ definiert ist, bedeutet dies die Untersuchung der Konvergenz von

$$e(x,h_n), \quad h_n := \frac{x-x_0}{n}, \quad n \to \infty.$$

Definition 2.6 Das Einschrittverfahren zur Inkrementfunktion Φ heißt **konvergent** von der Ordnung p, falls gilt

$$e(x,h_n) = \mathcal{O}(h_n^p).$$

Satz 2.7 Gegeben sei für $x_0 \in [a,b]$ das Anfangswertproblem

$$\mathbf{y}' = \mathbf{f}(x,\mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

mit der exakten Lösung $\mathbf{y}(x)$. Das explizite Einschrittverfahren zur Inkrementfunktion Φ sei konsistent von der Ordnung p. Ferner sei Φ stetig auf

$$G := \{(x,\mathbf{y},h) : a \le x \le b,\ \mathbf{y} \in \mathbb{R}^n, |h| \le h_0\}$$

und genüge der Lipschitz-Bedingung

$$\|\Phi(x,\mathbf{u},h) - \Phi(x,\mathbf{v},h)\| \le M\|\mathbf{u}-\mathbf{v}\|$$

für alle $(x,\mathbf{u},h),(x,\mathbf{v},h) \in G$. Dann gilt für den globalen Diskretisierungsfehler

$$e(x,h_n) := \|\boldsymbol{\eta}(x,h_n) - \mathbf{y}(x)\| \le \left\{ e(x_0,h_n) + |x-x_0| \sup_{\substack{a\le x\le b\\ \mathbf{y}\in\mathbb{R}^n}} \|\boldsymbol{\tau}(x,\mathbf{y},h_n)\| \right\} e^{M|x-x_0|}$$

für alle $x \in [a,b]$ und $h_n = (x-x_0)/n$, $n \in \mathbb{N}$, mit $|h_n| \le h_0$.

Beweis. Sei $\mathbf{y}_i = \mathbf{y}(x_i)$. Das durch Φ erzeugte Einschrittverfahren liefert Näherungswerte $\boldsymbol{\eta}_i = \boldsymbol{\eta}(x_i,h)$ mit

$$\mathbf{y}_{i+1} - \boldsymbol{\eta}_{i+1} = \mathbf{y}_i - \boldsymbol{\eta}_i + h\big\{\Phi(x_i,\mathbf{y}_i,h) - \Phi(x_i,\boldsymbol{\eta}_i,h)\big\} + h\boldsymbol{\tau}(x_i,\mathbf{y}_i,h).$$

Setzen wir $e_i := \|\mathbf{y}_i - \boldsymbol{\eta}_i\|$ und $\boldsymbol{\tau}_i = \boldsymbol{\tau}(x_i,\mathbf{y}_i,h)$, dann folgt hieraus aufgrund der Lipschitz-Bedingung

$$\begin{aligned} e_{i+1} &\le e_i + |h| \underbrace{\|\Phi(x_i,\mathbf{y}_i,h) - \Phi(x_i,\boldsymbol{\eta}_i,h)\|}_{\le M\|\mathbf{y}_i-\boldsymbol{\eta}_i\|} + |h|\|\boldsymbol{\tau}_i\| \\ &\le (1+M|h|)e_i + N|h| \quad \text{mit} \quad N := \sup_{\substack{a\le x\le b\\ \mathbf{y}\in\mathbb{R}^n}} \|\boldsymbol{\tau}(x,\mathbf{y},h)\|. \end{aligned}$$

Rekursiv erhalten wir demnach

$$\begin{aligned} e_i &\le (1+M|h|)e_{i-1} + N|h| \\ &\le (1+M|h|)^2 e_{i-2} + (1+M|h|)N|h| + N|h| \\ &\ \ \vdots \\ &\le (1+M|h|)^i e_0 + \big\{(1+M|h|)^{i-1} + \cdots + (1+M|h|) + 1\big\}N|h| \\ &= (1+M|h|)^i e_0 + \frac{(1+M|h|)^i - 1}{M|h|} N|h|. \end{aligned}$$

Mit Hilfe der Abschätzung $0 < 1 + M|h| \le e^{M|h|}$ erhalten wir demnach

$$e_i \le e^{iM|h|} e_0 + \frac{e^{iM|h|}-1}{M} N.$$

Da $iM|h|$ nichtnegativ ist, gilt die Ungleichung

$$\frac{e^{iM|h|}-1}{M} N \le e^{iM|h|} iN|h|.$$

Diese oben eingesetzt liefert

$$e_i \le e^{iM|h|}\big\{e_0 + iN|h|\big\}, \quad i \in \mathbb{N}_0.$$

Sei nun $x_0 \neq x \in [a,b]$ fest gewählt und $h := h_n = (x - x_0)/n$, $n \in \mathbb{N}$. Dann ist $x_n = x_0 + nh = x$ und es folgt

$$e(x, h_n) = e_n \le e^{M|x-x_0|}\big\{e_0 + |x - x_0|N\big\},$$

das ist die Behauptung. □

Bemerkung Aus diesem Satz folgt, dass ein konsistentes Verfahren der Ordnung p auch mit der Ordnung p konvergiert, vorausgesetzt die Funktion $\mathbf{f}$ ist hinreichend glatt. △

2.4 Runge-Kutta-Verfahren

Runge-Kutta-Verfahren basieren auf der Integraldarstellung der Lösung

$$\mathbf{y}(x_{i+1}) = \mathbf{y}(x_i) + \int_{x_i}^{x_{i+1}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\,\mathrm{d}t.$$

Ein diskretes Verfahren erhalten wir, wenn wir eine Quadraturformel mit Stützstellen $\{\alpha_j\}_{j=1}^m$ und Gewichten $\{\gamma_j\}_{j=1}^m$ einsetzen:

$$\mathbf{y}(x_{i+1}) \approx \mathbf{y}(x_i) + h\sum_{j=1}^{m} \gamma_j \mathbf{f}\big(x_i + h\alpha_j, \mathbf{y}(x_i + h\alpha_j)\big).$$

Da $\mathbf{y}(x_i + h\alpha_j)$ auch nicht bekannt ist, approximiert man weiter

$$\mathbf{f}\big(x_i + h\alpha_j, \mathbf{y}(x_i + h\alpha_j)\big) \approx \mathbf{k}_j\big(x_i, \mathbf{y}(x_i), h\big), \quad j = 1, 2, \ldots, m,$$

und erhält folglich

$$\mathbf{y}(x_i + h\alpha_j) = \mathbf{y}(x_i) + \int_{x_i}^{x_i + h\alpha_j} \mathbf{f}\big(t, \mathbf{y}(t)\big)\,\mathrm{d}t \approx \mathbf{y}(x_i) + h\sum_{\ell=1}^{m} \beta_{j,\ell}\mathbf{k}_\ell\big(x_i, \mathbf{y}(x_i), h\big)$$

für geeignete Quadraturgewichte $\{\beta_{j,\ell}\}_{j,\ell=1}^m$.

Definition 2.8 Gegeben seien reelle Gewichte $\alpha_\ell, \beta_{j,\ell}, \gamma_\ell$, $j, \ell = 1, 2, \ldots, m$. Ein allgemeines ***m*-stufiges Runge-Kutta-Verfahren** besitzt dann die Form: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$ und berechne für alle $i = 0, 1, 2, \ldots$

$$\begin{aligned} x_{i+1} &:= x_i + h \\ \mathbf{k}_j(x_i, \boldsymbol{\eta}_i, h) &:= \mathbf{f}\left(x_i + h\alpha_j, \boldsymbol{\eta}_i + h\sum_{\ell=1}^{m} \beta_{j,\ell}\mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h)\right), \quad j = 1, 2, \ldots, m \qquad (2.5) \\ \boldsymbol{\eta}_{i+1} &:= \boldsymbol{\eta}_i + \underbrace{h\sum_{\ell=1}^{m} \gamma_\ell \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h)}_{=:\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, h)}. \end{aligned}$$

Die Gewichte sind dabei so zu wählen, dass eine möglichst hohe Genauigkeit erreicht wird. Der Übersichtlichkeit halber werden sie üblicherweise im *Butcher-Tableau* erfasst:

$$\begin{array}{c|ccc|c} \alpha_1 & \beta_{1,1} & \cdots & \beta_{1,m-1} & \beta_{1,m} \\ \alpha_2 & \beta_{2,1} & \cdots & \beta_{2,m-1} & \beta_{2,m} \\ \vdots & \vdots & & \vdots & \vdots \\ \alpha_m & \beta_{m,1} & \cdots & \beta_{m,m-1} & \beta_{m,m} \\ \hline & \gamma_1 & \cdots & \gamma_{m-1} & \gamma_m \end{array}$$

Beispiel 2.9 (Butcher-Tableaus)

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

Eulersches Polygonzug-Verfahren

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

Euler-Collatz-Verfahren

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

Verfahren von Heun

△

Zunächst werden wir explizite Runge-Kutta-Verfahren betrachten. Dann gilt $\alpha_1 = 0$ und $\beta_{j,\ell} = 0$ für alle $\ell \geq j$, das heißt, (2.5) lautet

$$\begin{aligned} \mathbf{k}_1(x, \mathbf{y}, h) &= \mathbf{f}(x, \mathbf{y}) \\ \mathbf{k}_2(x, \mathbf{y}, h) &= \mathbf{f}\big(x + h\alpha_2, \mathbf{y} + h\beta_{2,1}\mathbf{k}_1(x, \mathbf{y}, h)\big) \\ \mathbf{k}_3(x, \mathbf{y}, h) &= \mathbf{f}\big(x + h\alpha_3, \mathbf{y} + h\beta_{3,1}\mathbf{k}_1(x, \mathbf{y}, h) + h\beta_{3,2}\mathbf{k}_2(x, \mathbf{y}, h)\big) \\ &\vdots \\ \mathbf{k}_m(x, \mathbf{y}, h) &= \mathbf{f}\big(x + h\alpha_m, \mathbf{y} + h\beta_{m,1}\mathbf{k}_1(x, \mathbf{y}, h) + \cdots + h\beta_{m,m-1}\mathbf{k}_{m-1}(x, \mathbf{y}, h)\big). \end{aligned}$$

Eine Taylor-Entwicklung der Inkrementfunktion liefert

$$\begin{aligned} \Phi(x, \mathbf{y}, h) &= \sum_{i=1}^{m} \gamma_i \mathbf{k}_i(x, \mathbf{y}, h) \\ &= \sum_{i=1}^{m} \gamma_i \mathbf{f}\left(x + h\alpha_i, \mathbf{y} + h\sum_{j=1}^{i-1} \beta_{i,j}\mathbf{k}_j(x, \mathbf{y}, h)\right) \\ &= \sum_{i=1}^{m} \gamma_i \left(\mathbf{f}(x, \mathbf{y}) + h\alpha_i \mathbf{f}_x(x, \mathbf{y}) + h\sum_{j=1}^{i-1} \beta_{i,j}\mathbf{f}_{\mathbf{y}}(x, \mathbf{y}) \underbrace{\mathbf{k}_j(x, \mathbf{y}, h)}_{=\mathbf{f}(x,\mathbf{y})+\cdots} + \cdots\right) \\ &= \sum_{i=1}^{m} \gamma_i \left(\mathbf{f}(x, \mathbf{y}) + h\left[\alpha_i \mathbf{f}_x(x, \mathbf{y}) + \sum_{j=1}^{i-1} \beta_{i,j}\mathbf{f}_{\mathbf{y}}(x, \mathbf{y})\mathbf{f}(x, \mathbf{y})\right] + \cdots\right). \end{aligned}$$

Hierin sind die vernachlässigten Terme alle von der Ordnung $\mathcal{O}(h^2)$. Vergleicht man diese Entwicklung mit (2.4), so sieht man, dass etwa die Wahl

$$\sum_{i=1}^{m} \gamma_i = 1, \qquad \sum_{i=1}^{m} \gamma_i \alpha_i = \frac{1}{2}, \qquad \alpha_i = \sum_{j=1}^{i-1} \beta_{i,j}, \quad i = 1, 2, \ldots, m$$

ein Verfahren von mindestens zweiter Ordnung liefert. Treibt man die Taylor-Entwicklung weiter, so erhält man zum Beispiel folgende Runge-Kutta-Verfahren vierter Ordnung:

Beispiel 2.10 (explizite Runge-Kutta-Verfahren vierter Ordnung)

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

klassisches Runge-Kutta-Verfahren

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

Gills-Formel

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

3/8-Regel

△

Algorithmus 2.11 (explizites Runge-Kutta-Verfahren)

input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0$ und Schrittweite h

output: Approximation $\{\boldsymbol{\eta}_i\}$

① Initialisierung: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$ und $i := 0$

② berechne

$$\begin{aligned} \mathbf{k}_1 &:= \mathbf{f}(x_i, \boldsymbol{\eta}_i) \\ \mathbf{k}_2 &:= \mathbf{f}(x_i + h\alpha_2, \boldsymbol{\eta}_i + h\beta_{2,1}\mathbf{k}_1) \\ \mathbf{k}_3 &:= \mathbf{f}(x_i + h\alpha_3, \boldsymbol{\eta}_i + h\beta_{3,1}\mathbf{k}_1 + h\beta_{3,2}\mathbf{k}_2) \\ &\vdots \\ \mathbf{k}_m &:= \mathbf{f}(x_i + h\alpha_m, \boldsymbol{\eta}_i + h\beta_{m,1}\mathbf{k}_1 + \cdots + h\beta_{m,m-1}\mathbf{k}_{m-1}) \end{aligned}$$

③ setze

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_i + h\sum_{\ell=1}^{m} \gamma_\ell \mathbf{k}_\ell, \qquad x_{i+1} := x_i + h$$

④ erhöhe $i := i + 1$ und gehe zu ②

Oftmals sind implizite Runge-Kutta-Verfahren den expliziten vorzuziehen. Das Butcher-Tableau besitzt in diesem Fall nichttriviale Koeffzienten $\beta_{j,\ell}$ für $\ell \geq j$. Daher muss in jedem Schritt i das nichtlineare Gleichungssystem

$$\begin{aligned} \mathbf{k}_1(x,\mathbf{y},h) &= \mathbf{f}\big(x + h\alpha_1, \mathbf{y} + h\beta_{1,1}\mathbf{k}_1(x,\mathbf{y},h) + \cdots + h\beta_{1,m}\mathbf{k}_m(x,\mathbf{y},h)\big) \\ \mathbf{k}_2(x,\mathbf{y},h) &= \mathbf{f}\big(x + h\alpha_2, \mathbf{y} + h\beta_{2,1}\mathbf{k}_1(x,\mathbf{y},h) + \cdots + h\beta_{2,m}\mathbf{k}_m(x,\mathbf{y},h)\big) \\ &\vdots \\ \mathbf{k}_m(x,\mathbf{y},h) &= \mathbf{f}\big(x + h\alpha_m, \mathbf{y} + h\beta_{m,1}\mathbf{k}_1(x,\mathbf{y},h) + \cdots + h\beta_{m,m}\mathbf{k}_m(x,\mathbf{y},h)\big) \end{aligned} \tag{2.6}$$

in mn Unbekannten $\mathbf{k}_j = \mathbf{k}_j(x,\mathbf{y},h)$, $j = 1,2,\ldots,m$, gelöst werden.

Satz 2.12 Auf dem Streifen $S := \{(x,\mathbf{y}) : a \leq x \leq b, \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ Lipschitz-stetig bezüglich $\mathbf{y}$ mit Lipschitz-Konstante L. Die Schrittweite h sei

klein genug, so dass

$$q := Lh \max_{i=1}^{m} \sum_{j=1}^{m} |\beta_{i,j}| < 1$$

ist. Dann existieren für alle $(x, \mathbf{y}) \in S$ eindeutige Lösungsvektoren $\mathbf{k}_i = \mathbf{k}_i(x, \mathbf{y}, h)$, $i = 1, 2, \ldots, m$, von (2.6).

Beweis. Setze

$$\mathbf{K} = \big[\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_m\big] := \big[\mathbf{k}_1(x, \mathbf{y}, h), \mathbf{k}_2(x, \mathbf{y}, h), \ldots, \mathbf{k}_m(x, \mathbf{y}, h)\big] \in \mathbb{R}^{n \times m}$$

und

$$\Psi : \mathbb{R}^{n \times m} \to \mathbb{R}^{n \times m}, \quad \Psi(\mathbf{K}) = \big[\boldsymbol{\psi}_1(\mathbf{K}), \boldsymbol{\psi}_2(\mathbf{K}), \ldots, \boldsymbol{\psi}_m(\mathbf{K})\big]$$

mit

$$\boldsymbol{\psi}_j(\mathbf{K}) := \mathbf{f}\bigg(x + h\alpha_j, \mathbf{y} + h \sum_{\ell=1}^{m} \beta_{j,\ell} \mathbf{k}_\ell\bigg).$$

Jeder Fixpunkt der Iteration

$$\mathbf{K}^{(r+1)} = \Psi\big(\mathbf{K}^{(r)}\big), \quad r = 0, 1, 2, \ldots$$

löst dann das nichtlineare Gleichungssystem (2.6). Wir zeigen nun, dass Ψ eine Kontraktion bezüglich der Norm $|\!|\!| \mathbf{K} |\!|\!| := \max_{i=1}^{m} \|\mathbf{k}_i\|$ ist:

$$\begin{aligned}
|\!|\!| \Psi(\mathbf{K}) - \Psi(\mathbf{L}) |\!|\!| &= \max_{i=1}^{m} \|\boldsymbol{\psi}_i(\mathbf{K}) - \boldsymbol{\psi}_i(\mathbf{L})\| \\
&\le L \max_{i=1}^{m} \left\| \mathbf{y} + h \sum_{j=1}^{m} \beta_{i,j} \mathbf{k}_j - \mathbf{y} - h \sum_{j=1}^{m} \beta_{i,j} \boldsymbol{\ell}_j \right\| \\
&\le Lh \max_{i=1}^{m} \left(\sum_{j=1}^{m} |\beta_{i,j}| \|\mathbf{k}_j - \boldsymbol{\ell}_j\| \right) \\
&\le Lh \left(\max_{i=1}^{m} \sum_{j=1}^{m} |\beta_{i,j}| \right) \left(\max_{p=1}^{m} \|\mathbf{k}_p - \boldsymbol{\ell}_p\| \right) \\
&= \underbrace{Lh \left(\max_{i=1}^{m} \sum_{j=1}^{m} |\beta_{i,j}| \right)}_{=q<1} |\!|\!| \mathbf{K} - \mathbf{L} |\!|\!|.
\end{aligned}$$

Damit folgen Existenz und Eindeutigkeit der Lösung von (2.6) aus dem Banachschen Fixpunktsatz. □

Bemerkung Am schnellsten löst man das nichtlineare Gleichungssystem (2.6) mit dem Newton-Verfahren, wobei dann wieder erste Ableitungen von $\mathbf{f}$ eingehen. △

Beispiel 2.13 (implizite Runge-Kutta-Verfahren)

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

implizite Mittelpunktsregel (Ordnung 2)

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

SDIRK2-Verfahren (single-diagonal implizites Runge-Kutta-Verfahren, Ordnung 2)

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Radau3-Verfahren (basiert auf der Radau-Quadraturformel, Ordnung 3)

△

Algorithmus 2.14 (implizites Runge-Kutta-Verfahren)

input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0$ und Schrittweite h

output: Approximation $\{\boldsymbol{\eta}_i\}$

① Initialisierung: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$ und $i := 0$

② löse das nichtlineare Gleichungssystem

$$\begin{aligned} \mathbf{k}_1 &:= \mathbf{f}(x_i + h\alpha_1, \boldsymbol{\eta}_i + h\beta_{1,1}\mathbf{k}_1 + h\beta_{1,2}\mathbf{k}_2 + \cdots + h\beta_{1,m}\mathbf{k}_m) \\ \mathbf{k}_2 &:= \mathbf{f}(x_i + h\alpha_2, \boldsymbol{\eta}_i + h\beta_{2,1}\mathbf{k}_1 + h\beta_{2,2}\mathbf{k}_2 + \cdots + h\beta_{2,m}\mathbf{k}_m) \\ &\vdots \\ \mathbf{k}_m &:= \mathbf{f}(x_i + h\alpha_m, \boldsymbol{\eta}_i + h\beta_{m,1}\mathbf{k}_1 + h\beta_{m,2}\mathbf{k}_2 + \cdots + h\beta_{m,m}\mathbf{k}_m) \end{aligned}$$

③ setze

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_i + h\sum_{\ell=1}^{m} \gamma_\ell \mathbf{k}_\ell, \qquad x_{i+1} := x_i + h$$

④ erhöhe $i := i+1$ und gehe zu ②

Satz 2.15 Die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ sei auf dem Streifen $S := \{(x, \mathbf{y}) : a \le x \le b, \mathbf{y} \in \mathbb{R}^n\}$ Lipschitz-stetig bezüglich $\mathbf{y}$ mit Lipschitz-Konstante L. Wählt man zu paarweise verschiedenen $\alpha_i \in [0,1]$, $i = 1, 2, \ldots, m$, Parameter γ_i, $\beta_{i,j}$, $i, j = 1, 2, \ldots, m$ derart, dass für ein $r \in \mathbb{N}$ gilt

$$\sum_{j=1}^{m} \beta_{i,j}\alpha_j^\ell = \frac{\alpha_i^{\ell+1}}{\ell+1}, \quad i = 1, 2, \ldots, m, \quad \ell = 0, 1, 2, \ldots, r-1, \tag{2.7}$$

und

$$\sum_{j=1}^{m} \gamma_j \alpha_j^\ell = \frac{1}{\ell+1}, \quad \ell = 0, 1, 2, \ldots, r-1, \tag{2.8}$$

gilt, dann ist die entsprechende Runge-Kutta-Formel konsistent von der Ordnung r.

Beweis. Gemäß (2.7) und (2.8) folgt

$$\frac{\alpha_i^{\ell+1}}{\ell+1} = \int_0^{\alpha_i} t^\ell \,\mathrm{d}t = \sum_{j=1}^m \beta_{i,j}\alpha_j^\ell, \quad i = 1, 2, \ldots, m, \quad \ell = 0, 1, 2, \ldots, r-1,$$

und

$$\frac{1}{\ell+1} = \int_0^1 t^\ell \,\mathrm{d}t = \sum_{j=1}^m \gamma_j \alpha_j^\ell, \quad \ell = 0, 1, 2, \ldots, r-1.$$

Dies bedeutet, die Quadraturformeln mit den Knoten $\{\alpha_j\}_{j=1}^m$ und Gewichten $\{\beta_{i,j}\}_{j=1}^m$ sind auf dem Intervall $[0, \alpha_i]$ exakt von der Ordnung r, ebenso die Quadraturformel mit den Gewichten $\{\gamma_j\}_{j=1}^m$ auf $[0, 1]$. Daher folgt

$$\begin{aligned}
\mathbf{y}(x+h) - \mathbf{y}(x) &= \int_x^{x+h} \mathbf{f}\big(t, \mathbf{y}(t)\big) \,\mathrm{d}t \\
&= h \int_0^1 \mathbf{f}\big(x + hs, \mathbf{y}(x+hs)\big) \,\mathrm{d}s \\
&= h \sum_{i=1}^m \gamma_i \mathbf{f}\big(x + h\alpha_i, \mathbf{y}(x+h\alpha_i)\big) + \mathcal{O}(h^{r+1}).
\end{aligned}$$

Zusammen mit (2.6) ergibt sich daher

$$\begin{aligned}
\big\|\widetilde{\boldsymbol{\tau}}(x, \mathbf{y}, h)\big\| &= \left\| \frac{\mathbf{y}(x+h) - \mathbf{y}(x)}{h} - \sum_{i=1}^m \gamma_i \mathbf{k}_i(x, \mathbf{y}, h) \right\| \\
&= \left\| \sum_{i=1}^m \gamma_i \mathbf{f}\big(x + h\alpha_i, \mathbf{y}(x+h\alpha_i)\big) \right. \\
&\qquad \left. - \sum_{i=1}^m \gamma_i \mathbf{f}\left(x + h\alpha_i, \mathbf{y}(x) + h\sum_{j=1}^m \beta_{i,j}\mathbf{k}_j(x, \mathbf{y}, h)\right) \right\| + \mathcal{O}(h^r).
\end{aligned}$$

In Anbetracht der Lipschitz-Stetigkeit von $\mathbf{f}$ erhalten wir somit die Abschätzung

$$\big\|\widetilde{\boldsymbol{\tau}}(x, \mathbf{y}, h)\big\| \le L \sum_{i=1}^m |\gamma_i| \left\| \mathbf{y}(x + h\alpha_i) - \mathbf{y}(x) - h\sum_{j=1}^m \beta_{i,j}\mathbf{k}_j(x, \mathbf{y}, h) \right\| + \mathcal{O}(h^r).$$

Auf die gleiche Weise schließen wir unter Ausnutzung der Quadraturformeln auf $[0, \alpha_i]$ und (2.6), dass

$$\begin{aligned}
A_i &:= \left\| \mathbf{y}(x + h\alpha_i) - \mathbf{y}(x) - h\sum_{j=1}^m \beta_{i,j}\mathbf{k}_j(x, \mathbf{y}, h) \right\| \\
&= \left\| h\int_0^{\alpha_i} \mathbf{f}\big(x + hs, \mathbf{y}(x+hs)\big) \,\mathrm{d}s - h\sum_{j=1}^m \beta_{i,j}\mathbf{k}_j(x, \mathbf{y}, h) \right\| \\
&= h \left\| \sum_{j=1}^m \beta_{i,j}\mathbf{f}\big(x + h\alpha_j, \mathbf{y}(x+h\alpha_j)\big) \right. \\
&\qquad \left. - \sum_{j=1}^m \beta_{i,j}\mathbf{f}\left(x + h\alpha_j, \mathbf{y}(x) + h\sum_{\ell=1}^m \beta_{j,\ell}\mathbf{k}_\ell(x, \mathbf{y}, h)\right) \right\| + \mathcal{O}(h^{r+1}),
\end{aligned}$$

was wir aufgrund der Lipschitz-Stetigkeit von $\mathbf{f}$ abschätzen können gemäß

$$A_i \leq Lh \sum_{j=1}^{m} |\beta_{i,j}| \underbrace{\left\| \mathbf{y}(x+h\alpha_j) - \mathbf{y}(x) - h \sum_{\ell=1}^{m} \beta_{j,\ell} \mathbf{k}_\ell(x,\mathbf{y},h) \right\|}_{=A_j} + \mathcal{O}(h^{r+1})$$

$$\leq Lh \Bigg(\underbrace{\max_{j=1}^{m} A_j}_{=:A} \Bigg) \Bigg(\underbrace{\max_{j=1}^{m} \sum_{\ell=1}^{m} |\beta_{j,\ell}|}_{=:B} \Bigg) + \mathcal{O}(h^{r+1}).$$

Hieraus ergibt sich

$$A(1 - LhB) = \mathcal{O}(h^{r+1}),$$

dies bedeutet, $A = \mathcal{O}(h^{r+1})$. Schließlich folgt die Behauptung aus

$$\left\| \widetilde{\boldsymbol{\tau}}(x,\mathbf{y},h) \right\| \leq L \sum_{i=1}^{m} |\gamma_i| A_i + \mathcal{O}(h^r) \leq LA \sum_{i=1}^{m} |\gamma_i| + \mathcal{O}(h^r) = \mathcal{O}(h^r).$$

□

Bemerkung Setzt man

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \dots & \alpha_m \\ \alpha_1^2 & \alpha_2^2 & \dots & \alpha_m^2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^{r-1} & \alpha_2^{r-1} & \dots & \alpha_m^{r-1} \end{bmatrix} \in \mathbb{R}^{r \times m}, \qquad \mathbf{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & & \vdots \\ \beta_{m,1} & \beta_{m,2} & \dots & \beta_{m,m} \end{bmatrix} \in \mathbb{R}^{m \times m},$$

und

$$\mathbf{C} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_m \\ \alpha_1^2/2 & \alpha_2^2/2 & \dots & \alpha_m^2/2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^r/r & \alpha_2^r/r & \dots & \alpha_m^r/r \end{bmatrix} \in \mathbb{R}^{r \times m}, \qquad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix} \in \mathbb{R}^m, \qquad \mathbf{c} = \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/r \end{bmatrix} \in \mathbb{R}^r,$$

dann lassen sich (2.7) und (2.8) kurz schreiben als

$$\mathbf{A}\mathbf{B}^\star = \mathbf{C}, \qquad \mathbf{A}\boldsymbol{\gamma} = \mathbf{c}.$$

Da $\mathbf{A}$ eine Vandermonde-Matrix ist, besitzt sie im Fall $r \leq m$ den vollen Rang r und beide Gleichungssysteme sind lösbar. Insbesondere folgt die Eindeutigkeit im Fall $r = m$. Damit eine Exaktheit $r > m$ erreicht werden kann, müssen die Knoten $\{\alpha_j\}_{j=1}^m$ speziell gewählt werden, beispielsweise als Gauß-Quadraturpunkte. △

2.5 Schrittweitensteuerung

Bisher haben wir stets eine feste Schrittweite h betrachtet. Um den numerischen Aufwand gering zu halten, versucht man, die Schrittweite größtmöglich zu wählen. Allerdings

bedeutet eine große Schrittweite eine geringere Rechengenauigkeit, so dass die Schrittweite nur in denjenigen Bereichen groß gewählt werden sollte, wo die Lösung glatt ist. Da die Lösung jedoch unbekannt ist, ist in der Praxis ein geeigneter Fehlerschätzer vonnöten. Solche Fehlerschätzer beruhen meist auf einem zweiten Verfahren, dem *Kontrollverfahren*, welches in der Regel genauer ist als das Ausgangsverfahren.

Bezeichnen $\boldsymbol{\eta}_i$ die Lösung des Ausgangsverfahrens und $\widehat{\boldsymbol{\eta}}_i$ die Lösung des Kontrollverfahrens, dann erwarten wir nach Konstruktion $\|\widehat{\boldsymbol{\eta}}_i - \mathbf{y}_i\| \ll \|\boldsymbol{\eta}_i - \mathbf{y}_i\|$, so dass

$$\delta := \|\boldsymbol{\eta}_i - \widehat{\boldsymbol{\eta}}_i\| \lesssim \|\boldsymbol{\eta}_i - \mathbf{y}_i\| \pm \|\widehat{\boldsymbol{\eta}}_i - \mathbf{y}_i\| \approx \|\boldsymbol{\eta}_i - \mathbf{y}_i\|. \tag{2.9}$$

Natürlich soll das Kontrollverfahren den Aufwand nicht zu sehr in die Höhe treiben. Daher verwendet man *eingebettete Runge-Kutta-Verfahren*, die auf demselben Knotenvektor $\{\alpha_j\}$ und derselben Verfahrensmatrix $\{\beta_{i,j}\}$, aber unterschiedlichen Gewichten $\{\gamma_i\}$ und $\{\widehat{\gamma}_i\}$ beruhen.

Ausgehend vom Butcher-Tableau

$$\begin{array}{c|ccccc}
0 & & & & & \\
\alpha_2 & \beta_{2,1} & & & & \\
\alpha_3 & \beta_{3,1} & \beta_{3,2} & & & \\
\vdots & \vdots & & \ddots & & \\
\alpha_m & \beta_{m,1} & \beta_{m,2} & \cdots & \beta_{m,m-1} & \\
\hline
 & \gamma_1 & \gamma_2 & \cdots & \gamma_{m-1} & \gamma_m \\
\hline
 & \widehat{\gamma}_1 & \widehat{\gamma}_2 & \cdots & \widehat{\gamma}_{m-1} & \widehat{\gamma}_m
\end{array}$$

erhalten wir die beiden Verfahren

$$\begin{aligned}
\boldsymbol{\eta}_{i+1} &= \boldsymbol{\eta}_i + h \sum_{\ell=1}^{m} \gamma_\ell \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h), \\
\widehat{\boldsymbol{\eta}}_{i+1} &= \boldsymbol{\eta}_i + h \sum_{\ell=1}^{m} \widehat{\gamma}_\ell \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h).
\end{aligned}$$

Besitzt das erste die Konsistenzordung p und das zweite die Konsistenzordung $p+1$, dann folgt

$$\mathbf{y}_{i+1} - \boldsymbol{\eta}_{i+1} = h^{p+1}\mathbf{w}_i + \mathcal{O}(h^{p+2}), \quad \mathbf{y}_{i+1} - \widehat{\boldsymbol{\eta}}_{i+1} = \mathcal{O}(h^{p+2}) \tag{2.10}$$

für ein unbekanntes $\mathbf{w}_i \in \mathbb{R}^n$.

Beispiel 2.16 Im Fall der Differentialgleichung $y'(x) = f\big(y(x)\big)$ mit $y(x_i) = \eta_i$ folgt für das zweistufige Runge-Kutta-Verfahren

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

die Entwicklung

$$\begin{aligned}
\eta_{i+1} &= y_i + \frac{h}{2} f(\eta_i) + \frac{h}{2} \underbrace{f\big(\eta_i + h f(\eta_i)\big)}_{= f(\eta_i) + h f_y(\eta_i) f(\eta_i) + \frac{h^2}{2} f_{yy}(\eta_i) f^2(\eta_i) + \mathcal{O}(h^3)} \\
&= \eta_i + h f(\eta_i) + \frac{h^2}{2} f_y(\eta_i) f(\eta_i) + \frac{h^3}{4} f_{yy}(\eta_i) f^2(\eta_i) + \mathcal{O}(h^4).
\end{aligned}$$

Hingegen ergibt sich für die exakte Lösung mit

$$\begin{aligned} y'(x) &= f\big(y(x)\big) \\ y''(x) &= f_y\big(y(x)\big) f\big(y(x)\big) \\ y'''(x) &= f_{yy}\big(y(x)\big) f^2\big(y(x)\big) + f_y^2\big(y(x)\big) f\big(y(x)\big) \end{aligned}$$

die Taylor-Entwicklung

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + hf\big(y(x_i)\big) + \frac{h^2}{2} f_y\big(y(x_i)\big) f\big(y(x_i)\big) \\ &\quad + \frac{h^3}{6}\big[f_{yy}\big(y(x_i)\big) f^2\big(y(x_i)\big) + f_y^2\big(y(x_i)\big) f\big(y(x_i)\big)\big] + \mathcal{O}(h^4) \\ &= \eta_i + hf(\eta_i) + \frac{h^2}{2} f_y(\eta_i) f(\eta_i) + \frac{h^3}{6}\big[f_{yy}(\eta_i) f^2(\eta_i) + f_y^2(\eta_i) f(\eta_i)\big] + \mathcal{O}(h^4). \end{aligned}$$

Hieraus schließen wir auf die Fehlerentwicklung

$$y(x_{i+1}) - \eta_{i+1} = h^3 \Bigg[\underbrace{-\frac{1}{12} f_{yy}(\eta_i) f^2(\eta_i) + \frac{1}{6} f_y^2(\eta_i) f(\eta_i)}_{=w_i} \Bigg] + \mathcal{O}(h^4).$$

△

Wegen (2.10) lautet der Fehlerschätzer (2.9) demnach

$$\delta_i = \delta_i(h) = \|\boldsymbol{\eta}_{i+1} - \widehat{\boldsymbol{\eta}}_{i+1}\| = \|h^{p+1}\mathbf{w}_i + \mathcal{O}(h^{p+2})\| \approx h^{p+1}\|\mathbf{w}_i\|.$$

Mit der Schrittweite $\widehat{h}$ anstelle von h ergibt sich entsprechend

$$\delta_i(\widehat{h}) \approx \widehat{h}^{p+1}\|\mathbf{w}_i\| = \left(\frac{\widehat{h}}{h}\right)^{p+1} h^{p+1}\|\mathbf{w}_i\| \approx \left(\frac{\widehat{h}}{h}\right)^{p+1} \delta_i(h) \overset{!}{\leq} \varepsilon.$$

Folglich ist

$$\widehat{h} := \tau \left(\frac{\varepsilon}{\delta_i(h)}\right)^{1/(p+1)} h$$

mit einem festen Parameter τ die größtmögliche Schrittweite für die noch $\delta_i(\widehat{h}) \lesssim \varepsilon$ gilt. Die Schrittweite $\widehat{h}$ liefert den optimalen Kompromiss aus Genauigkeit und (zukünftigen) Rechenaufwand. In der Praxis wird $\tau < 1$ gewählt, etwa $\tau = 0.8$ oder $\tau = 0.9$.

Bemerkung Im Prinzip hat man zwei Möglichkeiten: Es kann $\boldsymbol{\eta}_{i+1}$ als Approximation verwendet werden und der genauere Wert $\widehat{\boldsymbol{\eta}}_{i+1}$ geht nur in die Schätzung der Schrittweite ein. Nach Konstruktion wird dann die Schrittweite optimal gewählt. Andererseits kann man auch mit dem genauerem Wert $\widehat{\boldsymbol{\eta}}_{i+1}$ als Approximation weiterrechnen. Hier sind allerdings die Schrittweiten tendenziell zu klein, denn es gilt $\|\mathbf{y}_{i+1} - \widehat{\boldsymbol{\eta}}_{i+1}\| \ll \varepsilon$. △

Algorithmus 2.17 (Schrittweitensteuerung durch eingebettetes Runge-Kutta-Verfahren)

input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0$, Anfangsschrittweite h_0 und gewünschte Genauigkeit ε

output: Approximation $\{(x_i, \boldsymbol{\eta}_i)\}$

① Initialisierung: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$, $\delta_0 := \varepsilon$ und $i := 0$

② wiederhole

$$h_i := \tau \left(\frac{\varepsilon}{\delta_i} \right)^{1/(p+1)} h_i$$

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_i + h_i \sum_{\ell=1}^{m} \gamma_\ell \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h_i)$$

$$\delta_i := h_i \left\| \sum_{\ell=1}^{m} (\gamma_\ell - \widehat{\gamma}_\ell) \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h_i) \right\|$$

solange bis $\delta_i \leq \varepsilon$ ist

③ setze

$$x_{i+1} := x_i + h_i, \quad h_{i+1} := h_i, \quad \delta_{i+1} := \delta_i$$

④ erhöhe $i := i + 1$ und gehe nach ②

Beispiel 2.18 (Eingebettete Runge-Kutta-Verfahren)

0			
1	1		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
$p=2$	$\frac{1}{2}$	$\frac{1}{2}$	0
$p=3$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{4}{6}$

Runge-Kutta-Fehlberg-Verfahren RKF2(3), wobei das Kontrollverfahren auf der Simpson-Regel beruht

0					
$\frac{1}{2}$	$\frac{1}{2}$				
$\frac{1}{2}$	0	$\frac{1}{2}$			
1	0	0	1		
1	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	
$p=3$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{6}$
$p=4$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	0

klassisches Runge-Kutta-Verfahren als Kontrollverfahren

△

Ein Kontrollverfahren höherer Ordnung kann man auch mittels *Richardson-Extrapolation* konstruieren. Ausgehend von der numerischen Lösung $\boldsymbol{\eta}_i$ im Punkt x_i berechnet man die Lösung im Punkt $x_{i+1} = x_i + h$ mit der Schrittweite h und der Schrittweite $h/2$. Die zugehörigen Lösungen werden mit $\boldsymbol{\eta}_{i+1}^h$ und $\boldsymbol{\eta}_{i+1}^{h/2}$ bezeichnet. Ferner bezeichne $\overline{\boldsymbol{\eta}}_{i+1}^{h/2}$ die mit einem Schritt des Verfahrens mit halber Schrittweite berechnete Lösung, wobei in $x_i + h/2$ von der exakten Lösung $\mathbf{y}_{i+1/2} = \mathbf{y}(x_i + h/2)$ ausgegangen werde. Es gilt demnach

$$\overline{\boldsymbol{\eta}}_{i+1}^{h/2} = \mathbf{y}_{i+1/2} + \frac{h}{2} \Phi \left(x_i + \frac{h}{2}, \mathbf{y}_{i+1/2}, \frac{h}{2} \right).$$

Während $\boldsymbol{\eta}_{i+1}^h$ und $\boldsymbol{\eta}_{i+1}^{h/2}$ tatsächlich berechnet werden, ist $\overline{\boldsymbol{\eta}}_{i+1}^{h/2}$ nur eine Hilfsgröße für die theoretischen Untersuchungen.

Wir gehen wieder von der Fehlerentwicklung

$$\mathbf{y}_{i+1} - \boldsymbol{\eta}_{i+1}^h = h^{p+1} \mathbf{w}_i + \mathcal{O}(h^{p+2}) \tag{2.11}$$

aus. Andererseits gilt

$$\begin{aligned}\mathbf{y}_{i+1} - \boldsymbol{\eta}_{i+1}^{h/2} &= \mathbf{y}_{i+1} - \overline{\boldsymbol{\eta}}_{i+1}^{h/2} + \overline{\boldsymbol{\eta}}_{i+1}^{h/2} - \boldsymbol{\eta}_{i+1}^{h/2} \\ &= \mathbf{w}_{i+1/2}\left(\frac{h}{2}\right)^{p+1} + \mathcal{O}(h^{p+2}) + \mathbf{y}_{i+1/2} + \frac{h}{2}\Phi\left(x_i + \frac{h}{2}, \mathbf{y}_{i+1/2}, \frac{h}{2}\right) \\ &\quad - \boldsymbol{\eta}_{i+1/2}^{h/2} - \frac{h}{2}\Phi\left(x_i + \frac{h}{2}, \boldsymbol{\eta}_{i+1/2}^{h/2}, \frac{h}{2}\right).\end{aligned}$$

Für den Fehler nach einem Schritt mit der Schrittweite $h/2$ gilt

$$\mathbf{y}_{i+1/2} - \boldsymbol{\eta}_{i+1/2}^{h/2} = \mathbf{w}_i\left(\frac{h}{2}\right)^{p+1} + \mathcal{O}(h^{p+2}).$$

Kombiniert mit $\mathbf{w}_{i+1/2} = \mathbf{w}_i + \mathcal{O}(h)$ und der Lipschitz-Stetigkeit von Φ folgt schließlich

$$\mathbf{y}_{i+1} - \boldsymbol{\eta}_{i+1}^{h/2} = 2\mathbf{w}_i\left(\frac{h}{2}\right)^{p+1} + \mathcal{O}(h^{p+2}). \tag{2.12}$$

Subtraktion von (2.11) und (2.12) liefert nun

$$2\mathbf{w}_i = \frac{\boldsymbol{\eta}_{i+1}^{h/2} - \boldsymbol{\eta}_{i+1}^{h}}{2^p - 1}\left(\frac{h}{2}\right)^{-p-1} + \mathcal{O}(h).$$

Dies bedeutet nach (2.12)

$$\mathbf{y}_{i+1} - \boldsymbol{\eta}_{i+1}^{h/2} = \frac{\boldsymbol{\eta}_{i+1}^{h/2} - \boldsymbol{\eta}_{i+1}^{h}}{2^p - 1} + \mathcal{O}(h^{p+2}).$$

Speziell sehen wir, dass das durch

$$\widehat{\boldsymbol{\eta}}_{i+1} := \frac{2^p\boldsymbol{\eta}_{i+1}^{h/2} - \boldsymbol{\eta}_{i+1}^{h}}{2^p - 1}$$

definierte Verfahren die Konsistenzordnung $p+1$ besitzt und daher als Kontrollverfahren anwendebar ist.

Algorithmus 2.19 (Schrittweitensteuerung durch Richardson-Extrapolation)
input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0$, Anfangsschrittweite h_0 und gewünschte Genauigkeit ε
output: Approximation $\{(x_i, \boldsymbol{\eta}_i)\}$
① Initialisierung: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$, $\delta_0 := \varepsilon$ und $i := 0$

② wiederhole

$$h_i := \tau\left(\frac{\varepsilon}{\delta_i}\right)^{1/(p+1)} h_i$$

$$\boldsymbol{\eta}_{i+1/2} := \boldsymbol{\eta}_i + \frac{h_i}{2}\sum_{\ell=1}^{m}\gamma_\ell \mathbf{k}_\ell\left(x_i, \boldsymbol{\eta}_i, \frac{h_i}{2}\right)$$

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_{i+1/2} + \frac{h_i}{2}\sum_{\ell=1}^{m}\gamma_\ell \mathbf{k}_\ell\left(x_i + \frac{h_i}{2}, \boldsymbol{\eta}_{i+1/2}, \frac{h_i}{2}\right)$$

$$\delta_i := \frac{2^p}{2^p-1}\left\|\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i - h_i\sum_{\ell=1}^{m}\gamma_\ell \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h_i)\right\|$$

solange bis $\delta_i \le \varepsilon$ ist

③ setze

$$x_{i+1} := x_i + h_i, \quad h_{i+1} := h_i, \quad \delta_{i+1} := \delta_i$$

④ erhöhe $i := i + 1$ und gehe nach ②

Bemerkung Bei der Schrittweitensteuerung durch Richardson-Extrapolation muss nur ein Einschrittverfahren implementiert werden. Dafür sind aber mehr Funktionsauswertungen nötig als bei der Schrittweitensteuerung durch eingebettete Runge-Kutta-Verfahren entsprechender Ordung. △

2.6 Steife Differentialgleichungen

Beispiel 2.20 Gegeben sei das System

$$\left.\begin{aligned} y_1' &= \frac{\lambda_1+\lambda_2}{2}y_1 + \frac{\lambda_1-\lambda_2}{2}y_2 \\ y_2' &= \frac{\lambda_1-\lambda_2}{2}y_1 + \frac{\lambda_1+\lambda_2}{2}y_2 \end{aligned}\right\} \quad \lambda_1, \lambda_2 < 0. \tag{2.13}$$

Die allgmeine Lösung für $x \ge 0$ lautet

$$\begin{aligned} y_1(x) &= C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x} \\ y_2(x) &= C_1 e^{\lambda_1 x} - C_2 e^{\lambda_2 x} \end{aligned} \tag{2.14}$$

mit $C_1, C_2 \in \mathbb{R}$.
Berechnet man (2.13) mit dem Eulerschen Polygonzug-Verfahren, so lässt sich die numerische Näherung geschlossen darstellen

$$\begin{aligned} \eta_{1,i} &= C_1(1+h\lambda_1)^i + C_2(1+h\lambda_2)^i \\ \eta_{2,i} &= C_1(1+h\lambda_1)^i - C_2(1+h\lambda_2)^i. \end{aligned}$$

Offensichtlich konvergieren die Lösungen nur dann, falls die Schrittweite h so klein gewählt wird, dass

$$|1+h\lambda_1| < 1, \quad |1+h\lambda_2| < 1. \tag{2.15}$$

Es sei nun $|\lambda_2|$ groß gegen $|\lambda_1|$. Wegen $\lambda_2 < 0$ ist dann in (2.14) der Einfluss der Komponente $e^{\lambda_2 x}$ gegenüber $e^{\lambda_1 x}$ vernachlässigbar klein. Leider gilt das nicht für das numerische Verfahren. Wegen (2.15) muss nämlich die Schrittweite so klein gewählt werden, dass

$$h < \frac{2}{|\lambda_2|}.$$

Für den Fall $\lambda_1 = -1$ und $\lambda_2 = -1000$ bedeutet dies $h \leq 0.002$. Obwohl also e^{-1000x} zur Lösung praktisch nichts beiträgt, bestimmt der Faktor 1000 im Exponenten die Schrittweite. Dieses Verhalten bezeichnet man als *steif*. △

Definition 2.21 Die lineare Differentialgleichung

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(x), \quad \mathbf{A} \in \mathbb{R}^{n\times n}, \quad \mathbf{b}(x) \in \mathbb{R}^n$$

heißt **steif**, wenn $\operatorname{Re}(\lambda_i) < 0$ für alle Eigenwerte λ_i von $\mathbf{A}$ und $\min_{i=1}^n\{|\operatorname{Re}(\lambda_i)|\} \ll \max_{i=1}^n\{|\operatorname{Re}(\lambda_i)|\}$. Der Quotient

$$q := \frac{\max_{i=1}^n\{|\operatorname{Re}(\lambda_i)|\}}{\min_{i=1}^n\{|\operatorname{Re}(\lambda_i)|\}}$$

heißt **Steifigkeitsquotient**.

Das Anwenden eines Einschrittverfahrens auf das Problem

$$y' = \lambda y \quad \text{mit} \quad \lambda < 0$$

führt im allgemeinen auf eine Rekursion

$$\eta_{i+1} = g(h\lambda)\eta_i, \quad i = 0, 1, 2, \ldots,$$

wobei die *Stabilitätsfunktion* g vom jeweiligen Verfahren abhängt.

Beispiel 2.22 Beim Eulerschen Polygonzug-Verfahren folgt gemäß Beispiel 2.20 $g(z) = 1 + z$. Beim klassischen Runge-Kutta-Verfahren gilt

$$\begin{aligned} k_1 &= \eta_i z \\ k_2 &= \eta_i(z + z^2/2) \\ k_3 &= \eta_i(z + z^2/2 + z^3/4) \\ k_4 &= \eta_i(z + z^2 + z^3/2 + z^4/4) \\ \eta_{i+1} &= \eta_i + k_1/6 + k_2/3 + k_3/3 + k_4/6, \end{aligned}$$

dies bedeutet,

$$g(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}.$$

Beim impliziten Euler-Verfahren erhält man aus

$$\eta_{i+1} = \eta_i + hf(x_{i+1}, \eta_{i+1}) = \eta_i + z\eta_{i+1}$$

dass

$$g(z) = \frac{1}{1-z}.$$

Die Trapez-Methode

$$\eta_{i+1} = \eta_i + \frac{h}{2}\{f(x_i, \eta_i) + f(x_{i+1}, \eta_{i+1})\} = \eta_i\left(1 + \frac{z}{2}\right) + \frac{z}{2}\eta_{i+1}$$

induziert schließlich

$$g(z) = \frac{1 + z/2}{1 - z/2} = \frac{2+z}{2-z}.$$

△

Für $\lambda < 0$ folgt nun $e^{\lambda x} \overset{x\to\infty}{\longrightarrow} 0$, während $\eta_i \to 0$ genau dann gilt, wenn $|g(z)| < 1$. Da bei den beiden expliziten Verfahren $g(z)$ ein Polynom ist, gilt

$$|g(z)| \to \infty \quad \text{für} \quad z \to -\infty.$$

Daher sind diese Verfahren nur stabil, falls h klein genug ist.
Demgegenüber ist bei den beiden impliziten Verfahren $g(z)$ eine rationale Funktion mit der Eigenschaft

$$|g(z)| < 1 \quad \text{für} \quad z < 0.$$

Allerdings sieht man, dass die Dämpfung beim impliziten Euler-Verfahren für betragsgroße Werte $z = h\lambda$ sehr viel stärker als bei der Trapez-Regel ist. Bei letzterer nähert man sich der Stabilitätsgrenze $|g(z)| \overset{z\to-\infty}{\longrightarrow} 1$.

Definition 2.23 Der Bereich

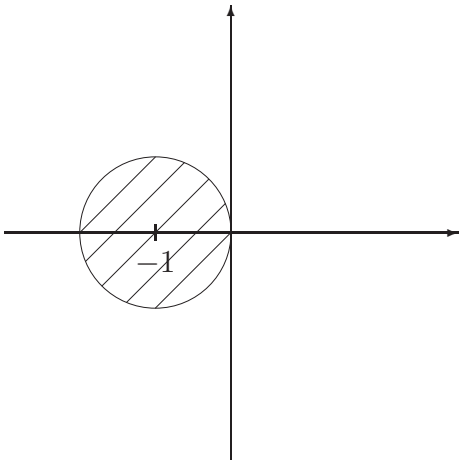
$$\mathcal{M} := \{z \in \mathbb{C} : |g(z)| < 1\}$$

heißt **Stabilitätsgebiet**. Ein Verfahren heißt **absolut stabil** oder ***A*-stabil**, falls

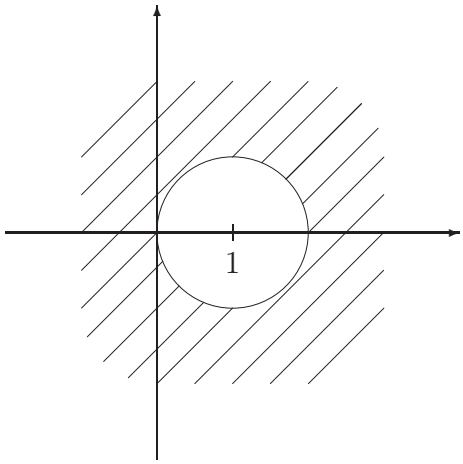
$$\{z \in \mathbb{C} : \mathrm{Re}(z) < 0\} \subset \mathcal{M},$$

wenn also $|g(z)| < 1$ gilt für alle $\mathrm{Re}(z) < 0$.

Beispiel 2.24



Stabilitätsgebiet des
expliziten Euler-Vefahren



Stabilitätsgebiet des
impliziten Euler-Vefahren

△

Das A-stabile implizite Euler-Verfahren besitzt uneingeschränkte Dämpfung für $z = h\lambda \to -\infty$, während die Stabilitätsfunktion der A-stabilen Trapez-Methode betragsmäßig gegen 1 strebt, also letztlich die Dämpfung verliert, was sich mit Rundungseffekten durchaus stark auswirken kann.

Definition 2.25 Ein Verfahren heißt **L-stabil**, falls es A-stabil ist und zusätzlich gilt

$$g(-\infty) = 0.$$

Bemerkungen

1. Explizite Einschrittverfahren sind niemals A-stabil.
2. Die Trapez-Methode ist A-stabil, das implizite Euler-Verfahren sogar L-stabil.
3. Das SDIRK2- und das Radau3-Verfahren aus Beispiel 2.13 sind beide L-stabil.

△

3. Mehrschrittverfahren

3.1 Definition

Genereller Nachteil von Einschrittverfahren ist die hohe Anzahl von Funktionsauswertungen im Fall von Verfahren höherer Ordnung. Ursache dafür ist, dass jeder Schritt so behandelt wird, als wäre es der erste.

Definition 3.1 Sei $x_i := x_0 + hi$ für $i \in \mathbb{N}$. Ein Verfahren der Form

$$\boldsymbol{\eta}_{i+k} := \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h) \tag{3.1}$$

mit festen Koeffizienten α_ℓ heißt ***k*-Schrittverfahren**. Hängt die Inkrementfunktion Φ nicht von $\boldsymbol{\eta}_{i+k}$ ab, dann ist das Mehrschrittverfahren **explizit**, andernfalls **implizit**. Das Mehrschrittverfahren heißt **linear**, falls es von der Form

$$\boldsymbol{\eta}_{i+k} := \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}) \tag{3.2}$$

mit festen Koeffizienten α_ℓ, β_ℓ ist.

Bemerkungen

1. Ein lineares Mehrschrittverfahren ist genau dann implizit, falls $\beta_k \neq 0$ gilt.
2. Ein k-Schrittverfahren benötigt immer k *Anlaufwerte*

   $$\boldsymbol{\eta}_0 = \mathbf{y}_0,\ \boldsymbol{\eta}_1 \approx \mathbf{y}(x_1),\ \boldsymbol{\eta}_2 \approx \mathbf{y}(x_2),\ \ldots,\ \boldsymbol{\eta}_{k-1} \approx \mathbf{y}(x_{k-1}).$$

   In der Praxis besorgt man sich diese aus einem hinreichend genauen Einschrittverfahren.

△

Algorithmus 3.2 (lineares Mehrschrittverfahren)

input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ und Schrittweite h

output: Approximation $\{\boldsymbol{\eta}_i\}$

① Initialisierung: setze $x_i := x_0 + ih$ und $\boldsymbol{\eta}_i := \mathbf{y}_i$ für alle $i = 0, 1, \ldots, k-1$

② für alle $i = 0, 1, \ldots$ berechne

$$\boldsymbol{\eta}_{i+k} := \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}\big), \qquad x_{i+k} := x_{i+k-1} + h$$

Bemerkung Ist das Mehrschrittverfahren explizit, so wird pro Iterationsschritt nur eine Funktionsauswertung von $\mathbf{f}$ benötigt. △

Beispiel 3.3 Folgende Spezialfälle linearer Mehrschrittverfahren sind von Bedeutung:

- Beim *Adams-Bashforth-Verfahren* gilt
$$\alpha_0 = \alpha_1 = \cdots = \alpha_{k-2} = 0, \quad \alpha_{k-1} = 1, \quad \beta_k = 0.$$
- Beim *Nyström-Verfahren* ist
$$\alpha_0 = \alpha_1 = \cdots = \alpha_{k-3} = 0, \quad \alpha_{k-2} = 1, \quad \alpha_{k-1} = 0, \quad \beta_k = 0.$$
- Beim *Adams-Moulton-Verfahren* ist
$$\alpha_0 = \alpha_1 = \cdots = \alpha_{k-2} = 0, \quad \alpha_{k-1} = 1, \quad \beta_k \neq 0.$$
- Beim *Milne-Simpson-Verfahren* ist
$$\alpha_0 = \alpha_1 = \cdots = \alpha_{k-3} = 0, \quad \alpha_{k-2} = 1, \quad \alpha_{k-1} = 0, \quad \beta_k \neq 0.$$

△

Definition 3.4 Der **lokale Diskretisierungsfehler** des Mehrschrittverfahrens ist

$$\boldsymbol{\tau}(x_i, \mathbf{y}, h) = \frac{1}{h}\left(\mathbf{y}(x_{i+k}) - \sum_{\ell=0}^{k-1} \alpha_\ell \mathbf{y}(x_{i+\ell})\right) - \Phi\big(x_i, \mathbf{y}(x_i), \ldots, \mathbf{y}(x_{i+k}), h\big).$$

Das Mehrschrittverfahren heißt **konsistent** von der Ordnung p, falls gilt

$$\|\boldsymbol{\tau}(x, \mathbf{y}, h)\| = \mathcal{O}(h^p).$$

Im Gegensatz zu Runge-Kutta-Verfahren kann man für lineare Mehrschrittverfahren sehr einfache Konsistenzbedingungen formulieren.

Satz 3.5 Sei $\mathbf{f}$ hinreichend glatt. Dann ist ein lineares Mehrschrittverfahren der Form

$$\sum_{\ell=0}^{k} \alpha_\ell \boldsymbol{\eta}_{i+\ell} = h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}), \quad i = 0, 1, 2, \ldots \tag{3.3}$$

konsistent von der Ordnung p genau dann, wenn die folgenden $p+1$ Bedingungen erfüllt sind:

$$\sum_{\ell=0}^{k} \alpha_\ell = 0, \qquad \sum_{\ell=0}^{k} \big(\ell^q \alpha_\ell - q\ell^{q-1}\beta_\ell\big) = 0, \quad q = 1, 2, \ldots, p. \tag{3.4}$$

Beweis. Taylor-Entwicklung von $\mathbf{y}$ liefert

$$\begin{aligned}
\alpha_k \boldsymbol{\tau}(x, \mathbf{y}, h) &= \frac{1}{h} \sum_{\ell=0}^{k} \alpha_\ell \mathbf{y}(x + h\ell) - \sum_{\ell=0}^{k} \beta_\ell \mathbf{y}'(x + h\ell) \\
&= \frac{1}{h} \sum_{\ell=0}^{k} \alpha_\ell \left(\sum_{q=0}^{p} \frac{(h\ell)^q}{q!} \mathbf{y}^{(q)}(x) \right) - \sum_{\ell=0}^{k} \beta_\ell \left(\sum_{q=0}^{p-1} \frac{(h\ell)^q}{q!} \mathbf{y}^{(q+1)}(x) \right) + \mathcal{O}(h^p) \\
&= \frac{1}{h} \left(\sum_{\ell=0}^{k} \alpha_\ell \mathbf{y}(x) \right) + \left(\sum_{q=1}^{p} \frac{h^{q-1}}{q!} \mathbf{y}^{(q)}(x) \sum_{\ell=0}^{k} \ell^q \alpha_\ell \right) \\
&\qquad - \left(\sum_{q=1}^{p} \frac{h^{q-1}}{q!} \mathbf{y}^{(q)}(x) \sum_{\ell=0}^{k} q\ell^{q-1} \beta_\ell \right) + \mathcal{O}(h^p) \\
&= \frac{1}{h} \left(\sum_{\ell=0}^{k} \alpha_\ell \mathbf{y}(x) \right) + \sum_{q=1}^{p} \frac{h^{q-1}}{q!} \mathbf{y}^{(q)}(x) \sum_{\ell=0}^{k} \big(\ell^q \alpha_\ell - q\ell^{q-1}\beta_\ell\big) + \mathcal{O}(h^p).
\end{aligned}$$

Gelten die Bedingungen (3.4), dann ist das Verfahren von der Konsistenzordnung p. Damit umgekehrt das Verfahren die Konsistenzordnung p besitzt, müssen die Koeffizienten von h^q verschwinden. Da es Fälle mit $\mathbf{y}^{(q)}(x) \neq 0$ gibt, müssen die Bedingungen (3.4) gelten. □

3.2 Explizite k-Schrittverfahren

Die Klasse der Adams-Bashforth-Verfahren beruht auf der Diskretisierung der Integralgleichung

$$\mathbf{y}(x_{i+k}) = \mathbf{y}(x_{i+k-1}) + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\, \mathrm{d}t \tag{3.5}$$

an den Stützstellen $x_i, x_{i+1}, \ldots, x_{i+k-1}$ mit Hilfe von *Newton-Cotes-Quadraturformeln*.

Lemma 3.6 (Newton-Cotes-Quadratur) Zu den Quadraturpunkten $\xi_0 < \xi_1 < \cdots < \xi_m$ seien die Quadraturgewichte $\{\omega_i\}_{i=0}^m$ gegeben durch

$$\omega_i := \int_a^b L_i(x)\, \mathrm{d}x, \qquad L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{x - \xi_j}{\xi_i - \xi_j}, \qquad i = 0, 1, \ldots, m, \tag{3.6}$$

wobei $L_0(x) \equiv 1$ gelte im Fall $m = 0$. Dann berechnet die Quadraturformel

$$Q[f] = \sum_{i=0}^{m} \omega_i f(\xi_i)$$

das Integral

$$I[f] = \int_a^b f(x)\,\mathrm{d}x$$

exakt für alle Polynome bis zum Grad m.

Beweis. Sei $p \in \Pi_m$. Offensichtlich interpoliert p sich in den Stützstellen $\{\xi_i\}_{i=0}^m$ selbst. Wegen der Eindeutigkeit des Interpolationspolynoms gilt daher

$$p(x) = \sum_{i=0}^m p(\xi_i) L_i(x).$$

Daraus folgt

$$\begin{aligned} I[p] &= \int_a^b p(x)\,\mathrm{d}x = \int_a^b \sum_{i=0}^m p(\xi_i) L_i(x)\,\mathrm{d}x \\ &= \sum_{i=0}^m p(\xi_i) \int_a^b L_i(x)\,\mathrm{d}x \overset{(3.6)}{=} \sum_{i=0}^m \omega_i\, p(\xi_i) = Q[p], \end{aligned}$$

was zu zeigen war. □

Wir wenden dieses Lemma nun an, um das Integral in (3.5) zu approximieren. Da die Stützstellen $x_i, x_{i+1}, \ldots, x_{i+k-1}$ außerhalb des Integrationsintervalls $[x_{i+k-1}, x_{i+k}]$ liegen, erhalten wir folgende Gewichte, die nicht mit denen der üblichen Newton-Cotes-Quadraturformeln übereinstimmen:

Adams-Bashforth-Verfahren $\alpha_0 = \alpha_1 = \cdots = \alpha_{k-2} = 0, \quad \alpha_{k-1} = 1, \quad \beta_k = 0$						
k	β_0	β_1	β_2	β_3	β_4	Konsistenzordnung
1	1					1
2	$-1/2$	$3/2$				2
3	$5/12$	$-4/3$	$23/12$			3
4	$-3/8$	$37/24$	$-59/24$	$55/24$		4
5	$251/720$	$-637/360$	$327/90$	$-1387/360$	$1901/720$	5

Da die entsprechenden Quadraturformeln gemäß Lemma 3.6 exakt für alle Polynome vom Grad $k-1$ sind, folgt die Abschätzung

$$\mathbf{y}(x_{i+k}) - \mathbf{y}(x_{i+k-1}) = \int_{x_{i+k-1}}^{x_{i+k}} \underbrace{\mathbf{f}\big(t, \mathbf{y}(t)\big)}_{\mathbf{y}'(t)}\,\mathrm{d}t = h \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big) + \mathcal{O}(h^{k+1}).$$

Dies bedeutet, dass das so konstruierte k-Schritt-Verfahren die Konsistenzordnung k besitzt:

$$\|\boldsymbol{\tau}(x_i, \mathbf{y}, h)\| = \left\| \frac{\mathbf{y}(x_{i+k}) - \mathbf{y}(x_{i+k-1})}{h} - \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big) \right\| = \mathcal{O}(h^k).$$

Algorithmus 3.7 (Adams-Bashforth-Verfahren)
input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ und Schrittweite h
output: Approximation $\{\boldsymbol{\eta}_i\}$
① Initialisierung: setze $x_i := x_0 + ih$ und $\boldsymbol{\eta}_i := \mathbf{y}_i$ für alle $i = 0, 1, \ldots, k-1$
② für alle $i = 0, 1, \ldots$ berechne

$$\boldsymbol{\eta}_{i+k} := \boldsymbol{\eta}_{i+k-1} + h \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}\big), \qquad x_{i+k} := x_{i+k-1} + h$$

Um die Klasse der Nyström-Verfahren zu erhalten, verändern wir den Ansatz (3.5) leicht gemäß

$$\mathbf{y}(x_{i+k}) = \mathbf{y}(x_{i+k-2}) + \int_{x_{i+k-2}}^{x_{i+k}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\, \mathrm{d}t. \tag{3.7}$$

Für dieselbe Stützstellenwahl wie beim Adams-Bashforth-Verfahren liefert Lemma 3.6 nun die Werte:

Nyström-Verfahren $\alpha_0 = \alpha_1 = \cdots = \alpha_{k-3} = 0, \quad \alpha_{k-2} = 1, \quad \alpha_{k-1} = 0, \quad \beta_k = 0$							
k	β_0	β_1	β_2	β_3	β_4	β_5	Konsistenzordnung
2	0	2					2
3	1/3	−2/3	7/3				3
4	−1/3	4/3	−5/3	8/3			4
5	29/90	−73/45	147/45	−133/45	269/90		5
6	−14/45	169/90	−213/45	287/45	−203/45	33/10	6

Analog zum Adams-Bashforth-Verfahren folgert man anhand der Abschätzung

$$\mathbf{y}(x_{i+k}) - \mathbf{y}(x_{i+k-2}) = \int_{x_{i+k-2}}^{x_{i+k}} \underbrace{\mathbf{f}\big(t, \mathbf{y}(t)\big)}_{\mathbf{y}'(t)}\, \mathrm{d}t = h \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big) + \mathcal{O}(h^{k+1}),$$

dass das Nyström-k-Schritt-Verfahren die Konsistenzordnung k besitzt.

Bemerkung Das Adams-Bashforth-Verfahren mit $k = 1$ ist das Eulersche Polygonzugverfahren. Das Nyström-Verfahren mit $k = 2$ wird auch *Mittelpunktsregel* genannt. △

3.3 Implizite k-Schrittverfahren

Bei der Klasse der Adams-Moulton-Verfahren wird das Integral in (3.5) durch Newton-Cotes-Formeln in den Knoten $x_i, x_{i+1}, \ldots, x_{i+k-1}$ und zusätzlich x_{i+k} approximiert. Nach Lemma 3.6 führt dies auf die Koeffizienten:

Adams-Moulton-Verfahren $\alpha_0 = \alpha_1 = \cdots = \alpha_{k-2} = 0, \quad \alpha_{k-1} = 1, \quad \beta_k \neq 0$							
k	β_0	β_1	β_2	β_3	β_4	β_5	Konsistenzordnung
1	$1/2$	$1/2$					2
2	$-1/12$	$2/3$	$5/12$				3
3	$1/24$	$-5/24$	$19/24$	$3/8$			4
4	$-19/720$	$53/360$	$-11/30$	$323/360$	$251/720$		5
5	$27/1440$	$-173/1440$	$241/720$	$-399/720$	$1427/1440$	$95/288$	6

Da die Newton-Cotes-Quadraturformeln gemäß Lemma 3.6 exakt für alle Polynome vom Grad k sind, ergibt sich die Abschätzung

$$\mathbf{y}(x_{i+k}) - \mathbf{y}(x_{i+k-1}) = \int_{x_{i+k-1}}^{x_{i+k}} \underbrace{\mathbf{f}\big(t, \mathbf{y}(t)\big)}_{\mathbf{y}'(t)} \,\mathrm{d}t = h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big) + \mathcal{O}(h^{k+2}).$$

Hieraus folgt, dass das Adams-Moulton-k-Schrittverfahren die Konsistenzordnung $k+1$ besitzt:

$$\|\boldsymbol{\tau}(x_i, \mathbf{y}, h)\| = \left\| \frac{\mathbf{y}(x_{i+k}) - \mathbf{y}(x_{i+k-1})}{h} - \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big) \right\| = \mathcal{O}(h^{k+1}).$$

Wie beim Adams-Moulton-Verfahren können wir auch das Integral in (3.7) durch Newton-Cotes-Formeln in den Knoten $x_i, x_{i+1}, \ldots, x_{i+k-1}$ und x_{i+k} approximieren. Dies führt auf die Klasse der Milne-Simpson-Verfahren. Gemäß Lemma 3.6 erhalten wir dann die folgenden Koeffizienten:

Milne-Simpson-Verfahren $\alpha_0 = \alpha_1 = \cdots = \alpha_{k-3} = 0, \quad \alpha_{k-2} = 1, \quad \alpha_{k-1} = 0, \quad \beta_k \neq 0$							
k	β_0	β_1	β_2	β_3	β_4	β_5	Konsistenzordnung
3	0	$1/3$	$4/3$	$1/3$			4
4	$-1/90$	$2/45$	$12/45$	$62/45$	$29/90$		5
5	$1/90$	$-3/45$	$7/45$	$-7/45$	$129/90$	$14/90$	6

Aufgrund des Exaktheitsgrads k der jeweiligen Newton-Cotes-Quadraturformel, schließen wir wie beim Adams-Moulton-k-Schrittverfahren, dass das Milne-Simpson-k-Schrittverfahren ebenfalls die Konsistenzordnung $k+1$ besitzt:

$$\|\boldsymbol{\tau}(x_i, \mathbf{y}, h)\| = \left\| \frac{\mathbf{y}(x_{i+k}) - \mathbf{y}(x_{i+k-2})}{h} - \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big) \right\| = \mathcal{O}(h^{k+1}).$$

Eine andere Möglichkeit, Mehrschrittverfahren zu gewinnen, besteht darin, anstelle der Approximation des Integrals in (3.5) beziehungsweise (3.7) eine Approximation der Ableitung $\mathbf{y}'(x_{i+k})$ durch numerische Differentiation aus $\mathbf{y}(x_{i+k})$ und den bisher berechneten Werten $\mathbf{y}(x_i), \mathbf{y}(x_{i+1}), \ldots, \mathbf{y}(x_{i+k-1})$ zu verwenden. Man spricht daher von *Rückwärts-Differentiationsformeln*, kurz *BDF-Verfahren* (von backward differentiation formulas). BDF-Verfahren werden vor allem bei steifen Problemen eingesetzt.

Ausgehend vom Interpolationspolynom in den Punkten $\{(x_{i+\ell}, \mathbf{y}(x_{i+\ell}))\}_{\ell=0}^{k}$, gegeben durch

$$\mathbf{p}(x) = \sum_{\ell=0}^{k} \mathbf{y}(x_{i+\ell}) L_\ell(x), \qquad L_\ell(x) = \prod_{\substack{j=0 \\ j\neq\ell}}^{k} \frac{x - x_{i+j}}{x_{i+\ell} - x_{i+j}}, \quad \ell = 0, 1, \ldots, k,$$

fordert man

$$\mathbf{p}'(x_{i+k}) = \sum_{\ell=0}^{k} \mathbf{y}(x_{i+\ell}) L'_\ell(x_{i+k}) \stackrel{!}{=} \mathbf{f}\big(x_{i+k}, \mathbf{y}(x_{i+k})\big).$$

Dies führt auf das Verfahren

$$\sum_{\ell=0}^{k} \alpha_\ell \boldsymbol{\eta}_{i+\ell} = h\mathbf{f}(x_{i+k}, \boldsymbol{\eta}_{i+k}), \qquad \alpha_\ell = hL'_\ell(x_{i+k}), \quad \ell = 0, 1, \ldots, k.$$

Nach der Normalisierung gemäß der Definition (3.2) erhalten wir die Koeffizienten:

BDF-Verfahren $\beta_0 = \beta_1 = \cdots = \beta_{k-1} = 0, \quad \alpha_k = 1$							
k	β_k	α_0	α_1	α_2	α_3	α_4	Konsistenzordnung
1	1	-1					1
2	2/3	1/3	$-4/3$				2
3	6/11	$-2/11$	9/11	$-18/11$			3
4	12/25	3/25	$-16/25$	36/25	$-48/25$		4
5	60/137	$-12/137$	75/137	$-200/137$	300/137	$-300/137$	5

Dass das k-Schritt-BDF-Verfahren ein Verfahren der Konsistenzordnung k ist, überprüft man leicht mit Hilfe von Satz 3.5.

Bemerkung Das Adams-Moulton-Verfahren mit $k = 1$ heißt auch *Trapezregel*. Die Milne-Simpson-Verfahren für $k = 2$ und $k = 3$ stimmen überein und werden *Simpson-Regel* genannt. Das BDF-Verfahren mit $k = 1$ ist gerade das implizite Euler-Verfahren. △

In jedem Schritt eines der hier vorgestellten impliziten Verfahren muss das nichtlineare Gleichungssystem

$$\boldsymbol{\eta}_{i+k} = \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell})$$

gelöst werden. Wie schon bei den impliziten Einschrittverfahren kann nachgewiesen werden, dass diese Gleichung für hinreichend kleine Schrittweiten lösbar ist, falls $\mathbf{f}$ einer Lipschitz-Bedingung in der $\mathbf{y}$-Variablen genügt. Setzen wir

$$\Psi(\mathbf{u}) := h\beta_k \mathbf{f}(x_{i+k}, \mathbf{u}) + \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}),$$

dann folgt

$$\|\Psi(\mathbf{u}) - \Psi(\mathbf{v})\| = h|\beta_k| \|\mathbf{f}(x_{i+k}, \mathbf{u}) - \mathbf{f}(x_{i+k}, \mathbf{v})\| \leq h|\beta_k| L \|\mathbf{u} - \mathbf{v}\|.$$

Ist h so klein, dass $h|\beta_k|L < 1$ gilt, dann liefert der Banachsche Fixpunktsatz die Existenz und Eindeutigkeit der gesuchten Lösung

$$\boldsymbol{\eta}_{i+k} = \Psi(\boldsymbol{\eta}_{i+k}) = \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}).$$

Bemerkung Eine gute Startnäherung $\widetilde{\boldsymbol{\eta}}_{i+k}$ des nichtlinearen Lösers berechnet man üblicherweise durch ein explizites Verfahren — dies ist der Prädiktorschritt. Mit Hilfe einer Fixpunktiteration oder einem Newton-Verfahren verbessert man die Näherung $\widetilde{\boldsymbol{\eta}}_{i+k}$ und erhält die gesuchte Lösung $\boldsymbol{\eta}_{i+k}$ — dies ist der Korrektorschritt. Man spricht hier von *Prädiktor-Korrektor-Verfahren*. Oftmals reicht in einem solchen Prädiktor-Korrektor-Verfahren sogar eine feste Anzahl an Iterationsschritten im nichtlinearen Löser aus. Berechnet man beispielsweise den Prädiktor mit dem Adams-Bashforth-Verfahren der Stufe k, so besitzt dieser den Konsistenzfehler $\mathcal{O}(h^{k+1})$. Ausgehend von diesem Prädiktor führt man nun eine Fixpunktiteration zur Bestimmung der Lösung des Adams-Moulton-Verfahrens der Stufe k aus. Da die Kontraktionskonstante $L|\beta_k|h$ ist, liefert bereits ein Iterationsschritt einen Korrektor mit dem Konsistenzfehler

$$\big(L|\beta_k|h\big)\mathcal{O}(h^{k+1}) = \mathcal{O}(h^{k+2}),$$

vorausgesetzt, die Schrittweite h ist hinreichend klein. △

3.4 Stabilität

Einschrittverfahren sind mit $k = 1$ natürlich formal in der Klasse der Mehrschrittverfahren enthalten. Allerdings impliziert im Gegensatz zu Einschrittverfahren die Konsistenz nicht die Konvergenz von Mehrschrittverfahren im Fall einer Schrittzahl $k > 1$.

Beispiel 3.8 Wir betrachten das Anfangswertproblem

$$y'(x) = y(x) - 2\sin x, \quad x \in [0, 4], \qquad y(0) = 1,$$

mit der Lösung $y(x) = \sin(x) + \cos(x)$. Zur numerischen Berechnung dieser Lösung verwenden wir das explizite Zweischrittverfahren

$$\eta_{i+2} = -4\eta_{i+1} + 5\eta_i + h\big(2f(x_i, \eta_i) + 4f(x_{i+1}, \eta_{i+1})\big), \quad i = 0, 1, 2, \ldots, \tag{3.8}$$

welches die Konsistenzordnung 3 besitzt. Führen wir ausgehend von $\eta_0 = 1$, $\eta_1 = \sin h + \cos h$ für verschiedene Schrittweiten das Verfahren durch, erhalten wir:

h	$\lvert\eta_{1/h} - y(1)\rvert$	$\lvert\eta_{2/h} - y(2)\rvert$	$\lvert\eta_{4/h} - y(4)\rvert$
2^{-2}	0.0094	2.87	$3.3 \cdot 10^5$
2^{-3}	0.280	$6.1 \cdot 10^4$	$2.7 \cdot 10^{15}$
2^{-4}	$6.4 \cdot 10^3$	$5.4 \cdot 10^{14}$	$3.7 \cdot 10^{36}$

Die Resultate deuten an, dass keine Konvergenz vorliegt. Man beobachtet sogar ein starkes Anwachsen des Fehlers, wenn die Schrittweite h kleiner wird.

Die im Verfahren (3.8) auftretende Instabiliät lässt sich wie folgt erklären. Wir betrachten die Aufgabe

$$y'(x) = 0, \quad x \in [0, b], \qquad y(0) = 1,$$

mit der Lösung $y(x) = 1$. Das Zweischrittverfahren hierauf angewandt liefert die *lineare homogene Differenzengleichung*

$$\eta_{i+2} + 4\eta_{i+1} - 5\eta_i = 0, \quad i = 0, 1, 2, \ldots. \tag{3.9}$$

Der Ansatz $\eta_i = z^i$ führt auf $z^i(z^2 + 4z - 5) = 0$ für alle i, also $z = 0$, $z = 1$ oder $z = -5$. Folglich ist $\eta_i = 1^i \cdot A + (-5)^i \cdot B$ für jedes $A, B \in \mathbb{R}$ eine Lösung von (3.9). Sind nun gestörte Startwerte $\eta_0 = 1$ und $\eta_1 = 1 + \delta$ vorgegeben, so sind $A = 1 + \delta/6$ und $B = -\delta/6$ eindeutig bestimmt und es folgt

$$\eta_i = 1 + \frac{\delta}{6} - \frac{\delta}{6}(-5)^i, \quad i = 0, 1, 2, \ldots.$$

Man sieht, dass eine kleine Störung $\delta \neq 0$ im Startwert η_1 exponentiell wächst. △

Definition 3.9 Eine Gleichung der Form

$$\sum_{\ell=0}^{k} \alpha_\ell \eta_{i+\ell} = 0, \quad i = 0, 1, 2, \ldots. \tag{3.10}$$

mit $\alpha_k \neq 0$ heißt **lineare homogene Differenzengleichung k-ter Ordnung**.

Zu jedem k-Schrittverfahren der Form

$$\sum_{\ell=0}^{k} \alpha_\ell \boldsymbol{\eta}_{i+\ell} = h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h), \qquad i = 0, 1, 2, \ldots, \tag{3.11}$$

gehört eine lineare homogene Differenzengleichung (3.10) der Ordnung k, welche offensichtlich k linear unabhängige Lösungen hat. Hingegen besitzt die approximierte Differentialgleichung nur eine Lösung. Dieser einen Lösung entspricht aber nur eine Lösung der Differenzengleichung. Daher kann man im Fall $k > 1$ nur dann Konvergenz einer Mehrschrittmethode erwarten, wenn alle Lösungen der zugehörigen homogenen Differenzengleichung beschränkt bleiben, da auch diese durch Rundungsfehler eingeschleppt werden. Aus diesem Grund untersuchen wir zunächst die Stabilität der Differenzengleichung (3.10).

Definition 3.10 Die lineare homogene Differenzengleichung (3.10) wird **stabil** genannt, wenn alle ihre Lösungen beschränkt sind.

Die Frage, wann eine lineare homogene Differenzengleichung stabil ist, beantwortet der folgende Satz.

Satz 3.11 Die lineare homogene Differenzengleichung (3.10) ist stabil genau dann, falls sie die Dahlquistsche Wurzelbedingung erfüllt, das heißt, falls die Nullstellen λ des charakteristischen Polynoms

$$p(\lambda) = \sum_{\ell=0}^{k} \alpha_\ell \lambda^\ell \tag{3.12}$$

alle $|\lambda| \leq 1$ erfüllen, und im Falle $|\lambda| = 1$ zusätzlich einfach sind.

Beweis. Zunächst sei angemerkt, dass jede Lösung der linearen Differenzengleichung (3.10) eindeutig bestimmt ist durch die k Startwerte $\eta_0, \eta_1, \ldots, \eta_{k-1}$. Die nachfolgenden Terme $\eta_k, \eta_{k+1}, \ldots$ sind dann rekursiv durch (3.10) bestimmt.
Sei A der Differentialoperator definiert durch $(Af)(\lambda) = \lambda f'(\lambda)$. Dann erfüllt die Folge

$$\eta_i = i^n \lambda^i, \quad i = 0, 1, 2, \ldots, \tag{3.13}$$

für jedes i die Beziehung

$$\begin{aligned}\sum_{\ell=0}^{k} \alpha_\ell \eta_{i+\ell} &= \sum_{\ell=0}^{k} \alpha_\ell \underbrace{(i+\ell)^n}_{=\sum_{m=0}^{n}\binom{n}{m} i^m \ell^{n-m}} \lambda^{i+\ell} \\ &= \lambda^i \sum_{m=0}^{n} \binom{n}{m} i^m \sum_{\ell=0}^{k} \alpha_\ell \ell^{n-m} \lambda^\ell \\ &= \lambda^i \sum_{m=0}^{n} \binom{n}{m} i^m (A^{n-m} p)(\lambda).\end{aligned}$$

Falls λ eine Nullstelle des charakteristischen Polynoms (3.12) mit der Vielfachheit s ist, so folgt $(A^m p)(\lambda) = 0$ für alle $m < s$. Daher löst die Folge (3.13) die Differenzengleichung (3.10) für alle $n < s$.
Nun nehmen wir an, dass $\lambda_1, \lambda_2, \ldots, \lambda_n$ die Nullstellen des charakteristischen Polynoms (3.12) sind und die Vielfachheiten $s_1, s_2, \ldots, s_n$ besitzen, dies bedeutet

$$p(\lambda) = \prod_{m=1}^{n} (\lambda - \lambda_m)^{s_m}.$$

Wir wollen nun zeigen, dass alle Lösungen der linearen homogenen Differenzengleichung (3.10) die Form

$$\eta_i = \sum_{m=1}^{n} \sum_{s=0}^{s_m - 1} \gamma_{m,s} i^s \lambda_m^i \tag{3.14}$$

mit beliebigen Koeffizienten $\gamma_{m,s} \in \mathbb{R}$ besitzen. Dazu müssen wir zeigen, dass die Koeffizienten $\gamma_{m,s}$ so gewählt werden können, dass beliebig vorgegebene Anfangsbedingungen

$$\sum_{m=1}^{n} \sum_{s=0}^{s_m - 1} \gamma_{m,s} i^s \lambda_m^i = \eta_i, \quad i = 0, 1, \ldots, k-1, \tag{3.15}$$

erfüllt sind. Das homogene adjungierte Gleichungssystem zu (3.15) lautet

$$\sum_{i=0}^{k-1} \rho_i i^s \lambda_m^i = 0, \quad s = 0, 1, \ldots, s_m - 1, \quad m = 1, 2, \ldots, n.$$

Angenommen, $\{\rho_i\}_{i=0}^{k-1}$ ist eine Lösung dieses Systems, dann besitzt das Polynom

$$q(\lambda) := \sum_{i=0}^{k-1} \rho_i \lambda^i \in \Pi_{k-1}$$

die Nullstellen $\lambda_1, \lambda_2, \ldots, \lambda_n$ mit Vielfachheiten $s_1, s_2, \ldots, s_n$. Da dies k Nullstellen sind, muss $\rho_0 = \rho_1 = \cdots = \rho_{k-1} = 0$ gelten, dies bedeutet, das Gleichungssystem (3.15) ist eindeutig lösbar.

Anhand der Darstellung (3.14) folgt schließlich die Äquivalenz von Wurzelbedingung und Stabilität. □

Definition 3.12 Das Mehrschrittverfahren (3.11) ist **nullstabil**, falls die Dahlquistsche Wurzelbedingung erfüllt ist, wenn also für alle Nullstellen λ des charakteristischen Polnoms (3.12) gilt $|\lambda| \leq 1$ und darüberhinaus, falls diese Nullstelle sogar mehrfach ist, $|\lambda| < 1$.

Bemerkungen

- Bei jedem Einschrittverfahren gilt $p(\lambda) = \lambda - 1$, das heißt, $\lambda = 1$ ist einzige Nullstelle von p. Damit ist die Wurzelbedingung erfüllt und das Einschrittverfahren immer nullstabil.
- Bei den Adams-Bashforth- und Adams-Moulton-Verfahren hat man $p(\lambda) = \lambda^{k-1}(\lambda - 1)$, also eine einfache Nullstelle bei $\lambda = 1$ und eine $(k-1)$-fache Nullstelle bei $\lambda = 0$. Beide Verfahrensklassen sind folglich nullstabil.
- Die gleiche Argumentation greift auch für die Nyström- und die Milne-Simpson-Verfahren. Hier ist $p(\lambda) = \lambda^{k-2}(\lambda^2 - 1)$, also je eine einfache Nullstelle bei $\lambda = \pm 1$ und eine $(k-2)$-fache Nullstelle bei $\lambda = 0$.
- Bei dem BDF-Verfahren mit $k = 2$ gilt

$$p(\lambda) = \lambda^2 - \frac{4}{3}\lambda + \frac{1}{3}.$$

  Als Nullstellen von $p(\lambda)$ ergeben sich $\lambda = 1$ und $\lambda = 1/3$, woraus die Nullstabilität folgt. Den Nachweis der Nullstabilität der BDF-Verfahren mit $k \geq 3$ überlassen wir dem Leser als Übungaufgabe.

△

3.5 Konvergenz von Mehrschrittverfahren

Zur Konvergenzuntersuchung müssen wir den Einfluss von Rundungsfehlern beachten. Daher betrachten wir zur näherungsweisen Lösung der Anfangswertaufgabe (2.1) das *gestörte* Mehrschrittverfahren

$$\begin{aligned} \sum_{\ell=0}^{k} \alpha_\ell \boldsymbol{\eta}_{i+\ell} &= h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h) + h\boldsymbol{\varepsilon}_{i+k}(h), \;\; x_i = x_0 + hi, \quad i = 0, 1, \ldots \\ \boldsymbol{\eta}_i &= \mathbf{y}(x_i) + \boldsymbol{\varepsilon}_i(h), \quad i = 0, 1, \ldots, k-1. \end{aligned} \tag{3.16}$$

Im folgenden wollen wir dabei stets annehmen, dass $\alpha_k = 1$ ist.

Definition 3.13 Für alle Startwerte und Störungen gelte $\|\boldsymbol{\eta}_i - \mathbf{y}(x_i)\| = \mathcal{O}(h^p)$ und $\|\boldsymbol{\varepsilon}_i(h)\| = \mathcal{O}(h^p)$. Dann heißt das Mehrschrittverfahren (3.16) **konvergent** von der Ordnung p, falls der globale Diskretisierungsfehler zur Näherungslösung $\boldsymbol{\eta}(x_i, h_n) = \boldsymbol{\eta}_i$ für jedes feste x der Abschätzung

$$e(x, h_n) = \|\boldsymbol{\eta}(x, h_n) - \mathbf{y}(x)\| = \mathcal{O}(h_n^p), \quad h_n = \frac{x - x_0}{n} \to 0$$

genügt.

Satz 3.14 Es gelte die Bedingung

$$\mathbf{f} \equiv \mathbf{0} \implies \Phi(x, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k, h) \equiv \mathbf{0} \tag{3.17}$$

für alle $x \in [a, b]$, h und $\boldsymbol{\eta}_i$. Dann folgt aus der Konvergenz des Mehrschrittverfahrens (3.16) stets seine Nullstabilität.

Beweis. Wählt man die spezielle, skalare Anfangswertaufgabe

$$y'(x) = 0, \quad x \in [0, b], \qquad y(0) = 1,$$

dann folgt wegen (3.17), dass alle Näherungslösungen des Mehrschrittverfahrens im Fall $\varepsilon_i(h) = 0$, $i \geq k$, die homogene Differenzengleichung (3.10) erfüllen. Aus der Konvergenz des Mehrschrittverfahrens folgt, dass die Lösungen für alle Startwerte beschränkt sind. Daraus ergibt sich mit Satz 3.11 das Behauptete. □

Lemma 3.15 Das Polynom

$$p(\lambda) = \lambda^k + \alpha_{k-1}\lambda^{k-1} + \cdots + \alpha_1\lambda + \alpha_0$$

erfülle die Dahlquistsche Wurzelbedingung und sei $\mathbf{A}$ die zugehörige transponierte Frobenius-Begleitmatrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \cdots & -\alpha_{k-2} & -\alpha_{k-1} \end{bmatrix} \in \mathbb{C}^{k\times k}.$$

Dann existiert eine reguläre Matrix $\mathbf{R} \in \mathbb{C}^{k\times k}$, so dass für die von der Vektornorm $\|\mathbf{x}\|_{\mathbf{R}} := \|\mathbf{R}\mathbf{x}\|_\infty$ induzierte Matrixnorm gilt

$$\|\mathbf{A}\|_{\mathbf{R}} = \max_{\|\mathbf{x}\|_{\mathbf{R}}=1} \|\mathbf{A}\mathbf{x}\|_{\mathbf{R}} = \max_{\|\mathbf{R}\mathbf{x}\|_\infty=1} \|\mathbf{R}\mathbf{A}\mathbf{x}\|_\infty \leq 1.$$

Beweis. Seien $\lambda_1, \lambda_2 \ldots, \lambda_n$ die Nullstellen von p mit den Vielfachheiten $s_1, s_2, \ldots, s_n$. Da diese wegen $\det(\lambda\mathbf{I} - \mathbf{A}) = p(\lambda)$ mit den Eigenwerten von $\mathbf{A}$ zusammenfallen, existiert

eine reguläre Matrix $\mathbf{T} \in \mathbb{C}^{k\times k}$, die $\mathbf{A}$ in die Jordan-Normalform überführt:

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} \mathbf{J}_1 & & & \\ & \mathbf{J}_2 & & \\ & & \ddots & \\ & & & \mathbf{J}_n \end{bmatrix}, \qquad \mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{s_i\times s_i}.$$

Falls es eine Nullstelle λ_i gibt mit $|\lambda_i| < 1$, so setzen wir

$$\varepsilon := \min_{\substack{i=1,2\ldots,n \\ |\lambda_i|<1}} \{1 - |\lambda_i|\} > 0,$$

ansonsten sei $\varepsilon := 1$. Mit

$$\mathbf{D} = \begin{bmatrix} 1 & & & \\ & \varepsilon & & \\ & & \ddots & \\ & & & \varepsilon^{k-1} \end{bmatrix} \in \mathbb{R}^{k\times k}$$

ergibt sich dann

$$\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{D} = \begin{bmatrix} \widetilde{\mathbf{J}}_1 & & & \\ & \widetilde{\mathbf{J}}_2 & & \\ & & \ddots & \\ & & & \widetilde{\mathbf{J}}_n \end{bmatrix}, \qquad \widetilde{\mathbf{J}}_i = \begin{bmatrix} \lambda_i & \varepsilon & & \\ & \ddots & \ddots & \\ & & \lambda_i & \varepsilon \\ & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{s_i\times s_i},$$

wobei nach Konstruktion gilt

$$\left\|\widetilde{\mathbf{J}}_i\right\|_\infty \le |\lambda_i| + \varepsilon \le 1, \quad \text{falls } |\lambda_i| < 1,$$

und

$$\left\|\widetilde{\mathbf{J}}_i\right\|_\infty = 1, \quad \text{falls } |\lambda_i| = 1.$$

Letzteres gilt, da der Jordan-Block im Fall $|\lambda_i| = 1$ aufgrund der Einfachheit der Nullstelle λ_i trivial ist. Es ergibt sich $\|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{D}\|_\infty \le 1$ und weiter mit $\mathbf{R} = \mathbf{D}^{-1}\mathbf{T}^{-1}$ schließlich

$$\|\mathbf{A}\|_{\mathbf{R}} = \max_{\|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{x}\|_\infty = 1} \|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{x}\|_\infty = \max_{\|\mathbf{y}\|_\infty = 1} \|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{D}\mathbf{y}\|_\infty \le 1.$$

□

Satz 3.16 Das (3.16) zugrundeliegende Mehrschrittverfahren sei konsistent von der Ordnung p und für die Störungen $\boldsymbol{\varepsilon}_i(h)$, $i = 0, 1, \ldots$, gelte stets $\|\boldsymbol{\varepsilon}_i(h)\| = \mathcal{O}(h^p)$. Im Streifen $S := \{(x, \mathbf{y}) : a \le x \le b, \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f}$ stetig, und die Inkrementfunktion erfülle (3.17) und genüge der Lipschitz-Bedingung

$$\|\Phi(x, \mathbf{u}_0, \ldots, \mathbf{u}_k, h) - \Phi(x, \mathbf{v}_0, \ldots, \mathbf{v}_k, h)\| \le L \sum_{\ell=0}^{k} \|\mathbf{u}_\ell - \mathbf{v}_\ell\|$$

für alle $x \in [a, b]$, h und $\mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^n$. Dann ist das Mehrschrittverfahren (3.16) genau dann konvergent von der Ordung p, wenn es nullstabil ist.

Beweis. Die Notwendigkeit der Nullstabilität wurde schon in Satz 3.14 gezeigt.
Zum Beweis der Hinlänglichkeit betrachte den Fehler $\mathbf{e}_i = \boldsymbol{\eta}_i - \mathbf{y}(x_i)$. Nach Voraussetzung gilt

$$\|\mathbf{e}_i\| = \mathcal{O}(h^p), \quad i = 0, 1, \ldots, k-1$$

und für alle $i \geq k$

$$\sum_{\ell=0}^{k} \alpha_\ell \mathbf{e}_{i+\ell} = h\Big(\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h) - \Phi\big(x_i, \mathbf{y}(x_i), \mathbf{y}(x_{i+1}), \ldots, \mathbf{y}(x_{i+k}), h\big) + \boldsymbol{\varepsilon}_{i+k}(h) - \boldsymbol{\tau}(x_i, \mathbf{y}, h)\Big) =: \mathbf{g}_i. \quad (3.18)$$

Aus der Lipschitz-Bedingung von Φ folgt

$$\|\mathbf{g}_i\| \leq h\bigg(L \sum_{\ell=0}^{k} \|\mathbf{e}_{i+\ell}\| + \|\boldsymbol{\varepsilon}_{i+k}(h)\| + \|\boldsymbol{\tau}(x_i, \mathbf{y}, h)\|\bigg). \quad (3.19)$$

Mit Hilfe der Matrizen

$$\mathbf{E}_i = \left[\ \mathbf{e}_i \ \middle|\ \mathbf{e}_{i+1} \ \middle|\ \cdots \ \middle|\ \mathbf{e}_{i+k-1}\ \right]^T \in \mathbb{R}^{k\times n}, \quad \mathbf{G}_i = \left[\ \mathbf{0} \ \middle|\ \cdots \ \middle|\ \mathbf{0} \ \middle|\ \mathbf{g}_i\ \right]^T \in \mathbb{R}^{k\times n}$$

und der transponierten Frobenius-Begleitmatrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \cdots & -\alpha_{k-2} & -\alpha_{k-1} \end{bmatrix} \in \mathbb{R}^{k\times k}$$

können wir (3.18) schreiben als

$$\mathbf{E}_{i+1} = \mathbf{A}\mathbf{E}_i + \mathbf{G}_i, \quad i = 0, 1, \ldots.$$

Für Matrizen $\mathbf{B} = [\mathbf{b}_1|\mathbf{b}_2|\cdots|\mathbf{b}_n] \in \mathbb{R}^{k\times n}$ sei die Norm $|\!|\!|\mathbf{B}|\!|\!|$ gegeben durch

$$|\!|\!|\mathbf{B}|\!|\!| := \max_{i=1}^{n} \|\mathbf{b}_i\|_{\mathbf{R}},$$

wobei $\|\cdot\|_{\mathbf{R}}$ die in Lemma 3.15 eingeführte Norm bezeichnet. Es folgt

$$|\!|\!|\mathbf{E}_{i+1}|\!|\!| \leq |\!|\!|\mathbf{A}\mathbf{E}_i|\!|\!| + |\!|\!|\mathbf{G}_i|\!|\!| \leq \|\mathbf{A}\|_{\mathbf{R}} |\!|\!|\mathbf{E}_i|\!|\!| + |\!|\!|\mathbf{G}_i|\!|\!|.$$

Aus (3.19) schließen wir wegen der Äquivalenz von Normen in endlichdimensionalen Räumen

$$|\!|\!|\mathbf{G}_i|\!|\!| \leq \widetilde{c}h\bigg(L \sum_{\ell=0}^{k} \|\mathbf{e}_{i+\ell}\| + \underbrace{\|\boldsymbol{\varepsilon}_{i+k}(h)\| + \|\boldsymbol{\tau}(x_i, \mathbf{y}, h)\|}_{=\mathcal{O}(h^p)}\bigg) \leq ch\big(|\!|\!|\mathbf{E}_i|\!|\!| + |\!|\!|\mathbf{E}_{i+1}|\!|\!| + h^p\big).$$

Da für genügend kleines h gilt $1 - ch > 0$, dürfen wir die letzten beiden Abschätzungen kombinieren gemäß

$$|\!|\!|\mathbf{E}_{i+1}|\!|\!| \leq \frac{1+ch}{1-ch} |\!|\!|\mathbf{E}_i|\!|\!| + \frac{ch^{p+1}}{1-ch}, \quad i = 0, 1, \ldots.$$

Wegen

$$\frac{1+x}{1-x} = 1 + \frac{2x}{1-x} \le 1 + 4x, \quad 0 \le x \le \frac{1}{2}$$

folgt deshalb

$$\begin{aligned}
\|\mathbf{E}_{i+1}\| &\le \underbrace{(1+4ch)}_{=:M} \|\mathbf{E}_i\| + \underbrace{\frac{ch^{p+1}}{1-ch}}_{=:N} \\
&\le M^2 \|\mathbf{E}_{i-1}\| + MN + N \\
&\vdots \\
&\le M^{i+1} \|\mathbf{E}_0\| + \left\{M^i + \cdots + M + 1\right\} N \\
&= M^{i+1} \|\mathbf{E}_0\| + \frac{M^{i+1}-1}{M-1} N.
\end{aligned}$$

Mit $0 < 1 + x \le e^x$ für $x > -1$ ergibt sich schließlich

$$\|\mathbf{E}_i\| \le e^{4cih} \|\mathbf{E}_0\| + \frac{e^{4cih}-1}{4ch} \frac{ch^{p+1}}{1-ch} \le e^{4cih} \left(\|\mathbf{E}_0\| + \frac{ch^p}{4(1-ch)} \right),$$

was für festes $x \ge x_0$ bedeutet

$$\|\mathbf{E}_{|x-x_0|/h}\| \le e^{4c|x-x_0|} \left(\underbrace{\|\mathbf{E}_0\|}_{=\mathcal{O}(h^p)} + \frac{ch^p}{4(1-ch)} \right) = \mathcal{O}(h^p).$$

□

3.6 Schrittweitensteuerung

Zur Konstruktion von Mehrschrittverfahren sind wir von einer äquidistanten Stützstellenwahl ausgegangen. Praxistaugliche Verfahren benötigen allerdings eine Schrittweitensteuerung, weshalb auf die Äquidistanz der Stützstellen $\{x_i\}$ verzichtet werden muss. Wir wollen das Vorgehen anhand der Adams-Verfahren erläutern.

Wir knüpfen dazu an die Integralgleichung (3.5) an, die durch formale Integration von $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ erzielt wurde:

$$\mathbf{y}(x_{i+k}) = \mathbf{y}(x_{i+k-1}) + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\, \mathrm{d}t.$$

Wir ersetzen den Integranden durch das Interpolationspolynom $\mathbf{p}_i^{(0)} \in \Pi_{k-1}$ in den k Stützstellen $\big\{\big(x_{i+\ell}, \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell})\big)\big\}_{\ell=0}^{k-1}$:

$$\boldsymbol{\eta}_{i+k}^{(0)} = \boldsymbol{\eta}_{i+k-1} + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{p}_i^{(0)}(t)\, \mathrm{d}t.$$

Um leicht an der Zahl der Stützstellen spielen zu können, verwenden wir die Newtonsche Interpolationsformel. Damit finden wir

$$\boldsymbol{\eta}_{i+k}^{(0)} = \boldsymbol{\eta}_{i+k-1} + \sum_{\ell=1}^{k} \mathbf{f}[x_{i+k-1}, x_{i+k-2}, \ldots, x_{i+k-\ell}] \int_{x_{i+k-1}}^{x_{i+k}} N_{[x_{i+k-1}, x_{i+k-2}, \ldots, x_{i+k-\ell}]}(t)\, \mathrm{d}t \quad (3.20)$$

mit den Newtonschen Basispolynomen

$$N_{[x_{i_1}]}(x) = 1 \quad \text{und} \quad N_{[x_{i_1},x_{i_2},\dots,x_{i_n}]}(x) = \prod_{\ell=1}^{n-1}(x - x_{i_\ell}) \in \Pi_{n-1}.$$

Hierbei bezeichnen wie üblich $\mathbf{f}[x_{i_1}, x_{i_2}, \dots, x_{i_n}]$ die dividierten Differenzen, die sich rekursiv in einem Dreiecks-Tableau berechnen lassen gemäß

$$\mathbf{f}[x_{i_\ell}] = \mathbf{f}(x_{i_\ell}, \boldsymbol{\eta}_{i_\ell}) \quad \text{und} \quad \mathbf{f}[x_{i_\ell}, \dots, x_{i_m}] = \frac{\mathbf{f}[x_{i_{\ell+1}}, \dots, x_{i_m}] - \mathbf{f}[x_{i_\ell}, \dots, x_{i_{m-1}}]}{x_{i_m} - x_{i_\ell}}.$$

Nach Konstruktion stimmt auf einem äquidistanten Gitter die Lösung $\boldsymbol{\eta}_{i+k}^{(0)}$ mit der des Adams-Bashforth-k-Schrittverfahrens überein.
Haben wir den Wert $\boldsymbol{\eta}_{i+k}^{(0)}$ berechnet, dann können wir das Interpolationspolynom $\mathbf{p}_i^{(1)} \in \Pi_{k-1}$ in den k Stützstellen $\big\{\big(x_{i+\ell}, \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell})\big)\big\}_{\ell=1}^{k-1}$ und $\big(x_{i+k}, \mathbf{f}(x_{i+k}, \boldsymbol{\eta}_{i+k}^{(0)})\big)$ verwenden, um eine weitere Näherungslösung zu berechnen:

$$\begin{aligned}
\boldsymbol{\eta}_{i+k}^{(1)} &= \boldsymbol{\eta}_{i+k-1} + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{p}_i^{(1)}(t)\,\mathrm{d}t \\
&= \boldsymbol{\eta}_{i+k-1} + \sum_{\ell=0}^{k-1} \mathbf{f}[x_{i+k}, x_{i+k-1}, \dots, x_{i+k-\ell}] \int_{x_{i+k-1}}^{x_{i+k}} N_{[x_{i+k},x_{i+k-1},\dots,x_{i+k-\ell}]}(t)\,\mathrm{d}t \\
&= \boldsymbol{\eta}_{i+k}^{(0)} + \int_{x_{i+k-1}}^{x_{i+k}} \big\{\mathbf{p}_i^{(1)}(t) - \mathbf{p}_i^{(0)}(t)\big\}\,\mathrm{d}t.
\end{aligned}$$

Dabei verschwindet das Polynom $\mathbf{p}_i^{(1)}(t) - \mathbf{p}_i^{(0)}(t) \in \Pi_{k-1}$ in den $k-1$ Punkten $x_{i+1}, x_{i+2}, \dots, x_{i+k-1}$, während im Punkt x_{i+k} gilt

$$\mathbf{p}_i^{(1)}(x_{i+k}) - \mathbf{p}_i^{(0)}(x_{i+k}) = \mathbf{f}\big(x_{i+k}, \boldsymbol{\eta}_{i+k}^{(0)}\big) - \mathbf{p}_i^{(0)}(x_{i+k}).$$

Daher folgt die Gleichung

$$\begin{aligned}
\mathbf{p}_i^{(1)}(t) - \mathbf{p}_i^{(0)}(t) &= \big\{\mathbf{f}\big(x_{i+k}, \boldsymbol{\eta}_{i+k}^{(0)}\big) - \mathbf{p}_i^{(0)}(x_{i+k})\big\} \prod_{\ell=1}^{k-1} \frac{t - x_{i+\ell}}{x_{i+k} - x_{i+\ell}} \\
&= \big\{\mathbf{f}\big(x_{i+k}, \boldsymbol{\eta}_{i+k}^{(0)}\big) - \mathbf{p}_i^{(0)}(x_{i+k})\big\} \frac{N_{[x_{i+k-1},x_{i+k-2},\dots,x_i]}(t)}{N_{[x_{i+k-1},x_{i+k-2},\dots,x_i]}(x_{i+k})}
\end{aligned}$$

und wir schließen

$$\boldsymbol{\eta}_{i+k}^{(1)} - \boldsymbol{\eta}_{i+k}^{(0)} = \big\{\mathbf{f}\big(x_{i+k}, \boldsymbol{\eta}_{i+k}^{(0)}\big) - \mathbf{p}_i^{(0)}(x_{i+k})\big\} \int_{x_{i+k-1}}^{x_{i+k}} \frac{N_{[x_{i+k-1},x_{i+k-2},\dots,x_i]}(t)}{N_{[x_{i+k-1},x_{i+k-2},\dots,x_i]}(x_{i+k})}\,\mathrm{d}t. \tag{3.21}$$

Als nächstes betrachten wir dasjenige Polynom $\mathbf{q}_i \in \Pi_k$, das die $k+1$ Stützstellen $\big\{\big(x_{i+\ell}, \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell})\big)\big\}_{\ell=0}^{k-1}$ und $\big(x_{i+k}, \mathbf{f}(x_{i+k}, \boldsymbol{\eta}_{i+k}^{(0)})\big)$ interpoliert. Wir nehmen an, dass $\mathbf{y}(x_{i+\ell}) = \boldsymbol{\eta}_{i+\ell}$ für alle $\ell = 0, 1, \dots, k-1$ ist und dass $\mathbf{f}\big(t, \mathbf{y}(t)\big)$ durch das Polynom $\mathbf{q}_i$ exakt dargestellt wird. Folglich gilt

$$\mathbf{y}(x_{i+k}) = \boldsymbol{\eta}_{i+k-1} + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\,\mathrm{d}t = \boldsymbol{\eta}_{i+k-1} + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{q}_i(t)\,\mathrm{d}t$$

und wir erhalten

$$\begin{aligned}\mathbf{y}(x_{i+k}) - \boldsymbol{\eta}_{i+k}^{(1)} &= \boldsymbol{\eta}_{i+k-1} + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{q}_i(t)\,\mathrm{d}t - \boldsymbol{\eta}_{i+k-1} - \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{p}_i^{(1)}(t)\,\mathrm{d}t \\ &= \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{q}_i(t) - \mathbf{p}_i^{(1)}(t)\,\mathrm{d}t.\end{aligned}$$

Einsetzen der Newtonschen Interpolationsformel liefert

$$\begin{aligned}\mathbf{q}_i(t) - \mathbf{p}_i^{(1)}(t) &= \sum_{\ell=0}^{k} \mathbf{f}[x_{i+k}, x_{i+k-1}, \ldots, x_{i+k-\ell}] N_{[x_{i+k}, x_{i+k-1}, \ldots, x_{i+k-\ell}]}(t) \\ &\quad - \sum_{\ell=0}^{k-1} \mathbf{f}[x_{i+k}, x_{i+k-1}, \ldots, x_{i+k-\ell}] N_{[x_{i+k}, x_{i+k-1}, \ldots, x_{i+k-\ell}]}(t) \\ &= \mathbf{f}[x_{i+k}, x_{i+k-1}, \ldots, x_i] N_{[x_{i+k}, x_{i+k-1}, \ldots, x_i]}(t).\end{aligned}$$

Dies bedeutet, es gilt

$$\mathbf{y}(x_{i+k}) - \boldsymbol{\eta}_{i+k}^{(1)} = \mathbf{f}[x_{i+k}, x_{i+k-1}, \ldots, x_i] \int_{x_{i+k-1}}^{x_{i+k}} N_{[x_{i+k}, x_{i+k-1}, \ldots, x_i]}(t)\,\mathrm{d}t. \tag{3.22}$$

Liegen die Stützstellen $x_i, x_{i+1}, \ldots, x_{i+k}$ äquidistant mit Abstand h, so folgt unter den gemachten Annahmen für den Fehler[1)]

$$\delta_i := \left\| \mathbf{y}(x_{i+k}) - \boldsymbol{\eta}_{i+k}^{(1)} \right\| = \mathcal{O}(h^{k+1}).$$

Wie im Abschnitt 2.5 wird nun ein Schritt zur Schrittweite h_{alt} akzeptiert, falls der Fehler δ_i kleiner einer vorgegebenen Genauigkeit ε ist. Ansonsten wird

$$h_{neu} = \tau \left(\frac{\varepsilon}{\delta_i} \right)^{1/(k+1)} h_{alt} \tag{3.23}$$

mit einem $\tau < 1$ gewählt.
In den obigen Berechnungen tauchen Integrale der Form

$$g_{m,n} := \int_{x_{i+k-1}}^{x_{i+k}} N_{[x_{i+k-1}, x_{i+k-2}, \ldots, x_{i+k-m}]}(t)(t - x_{i+k})^n\,\mathrm{d}t, \quad m \geq 1, n \geq 0$$

auf. Mit der Initialisierung

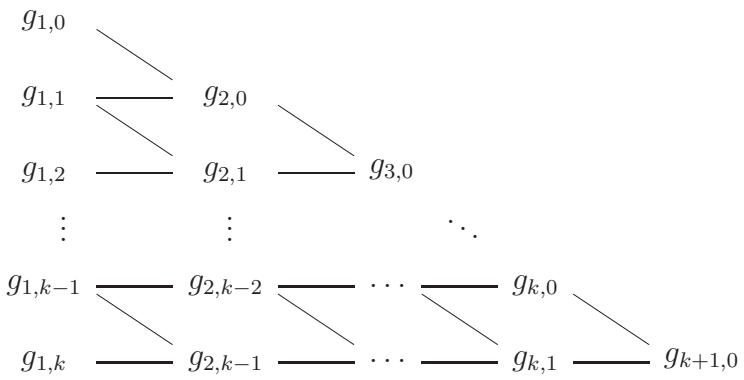
$$g_{1,n} = \int_{x_{i+k-1}}^{x_{i+k}} (t - x_{i+k})^n\,\mathrm{d}t = \frac{(-1)^n}{n+1}(x_{i+k} - x_{i+k-1})^{n+1}$$

[1)] Diese Annahmen sind vernünftig und machen uns das Leben nur leichter. Lässt man sie fallen, dann besitzt die Größe $\mathbf{y}(x_{i+k})$ selbst den Konsistenzfehler $\mathcal{O}(h^{k+2})$, während $\boldsymbol{\eta}_{i+k}^{(1)}$ weiterhin den Konsistenzfehler $\mathcal{O}(h^{k+1})$ besitzt. Dies folgt sofort aus der Bemerkung zu Prädiktor-Korrektor-Verfahren am Ende von Abschnitt 3.3.

gilt für diese die Rekursion

$$\begin{aligned} g_{m,n} &= \int_{x_{i+k-1}}^{x_{i+k}} N_{[x_{i+k-1},x_{i+k-2},\ldots,x_{i+k-m}]}(t)(t-x_{i+k})^n \,\mathrm{d}t \\ &= \int_{x_{i+k-1}}^{x_{i+k}} (t - x_{i+k} + x_{i+k} - x_{i+k-m+1}) N_{[x_{i+k-1},x_{i+k-2},\ldots,x_{i+k-m+1}]}(t)(t-x_{i+k})^n \,\mathrm{d}t \\ &= \int_{x_{i+k-1}}^{x_{i+k}} \Big\{ (x_{i+k} - x_{i+k-m+1}) N_{[x_{i+k-1},x_{i+k-2},\ldots,x_{i+k-m+1}]}(t)(t-x_{i+k})^n \\ &\qquad\qquad + N_{[x_{i+k-1},x_{i+k-2},\ldots,x_{i+k-m+1}]}(t)(t-x_{i+k})^{n+1} \Big\} \,\mathrm{d}t \\ &= (x_{i+k} - x_{i+k-m+1}) g_{m-1,n} + g_{m-1,n+1}. \end{aligned}$$

Mit Hilfe dieser Rekursionsformeln können alle benötigten Integrale durch das folgende Dreiecksschema bestimmt werden:



Wir fassen nun die Berechnungen in (3.20)–(3.23) in einem Algorithmus zusammen:

Algorithmus 3.17 (Schrittweitensteuerung)
input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$, Anfangsschrittweite h_0 und gewünschte Genauigkeit ε
output: Approximation $\{(x_i, \boldsymbol{\eta}_i)\}$
① Initialisierung: setze $i := 0$, $\delta_0 := \varepsilon$ und

$$x_\ell := x_0 + \ell h_0 \quad \text{und} \quad \boldsymbol{\eta}_\ell := \mathbf{y}_\ell \quad \text{für alle } \ell = 0, 1, \ldots, k-1$$

② wiederhole

$$h_i := \tau\left(\frac{\varepsilon}{\delta_i}\right)^{1/(k+1)} h_i, \qquad x_{i+k} := x_{i+k-1} + h_i$$

$$\boldsymbol{\eta}_{i+k}^{(0)} := \boldsymbol{\eta}_{i+k-1} + \sum_{\ell=1}^{k} \mathbf{f}[x_{i+k-1}, x_{i+k-2}, \ldots, x_{i+k-\ell}] g_{\ell,0}$$

$$\mathbf{p}_i^{(0)}(x_{i+k}) := \sum_{\ell=1}^{k} \mathbf{f}[x_{i+k-1}, x_{i+k-2}, \ldots, x_{i+k-\ell}] N_{[x_{i+k-1}, x_{i+k-2}, \ldots, x_{i+k-\ell}]}(x_{i+k})$$

$$\boldsymbol{\eta}_{i+k} := \boldsymbol{\eta}_{i+k}^{(0)} + \frac{\mathbf{f}\big(x_{i+k}, \boldsymbol{\eta}_{i+k}^{(0)}\big) - \mathbf{p}_i^{(0)}(x_{i+k})}{N_{[x_{i+k-1}, x_{i+k-2}, \ldots, x_i]}(x_{i+k})} g_{k,0}$$

$$\delta_i := \big\| \mathbf{f}[x_{i+k}, x_{i+k-1}, \ldots, x_i] g_{k,1} \big\|$$

solange bis $\delta_i \le \varepsilon$ ist

③ setze

$$h_{i+1} := h_i, \quad \delta_{i+1} := \delta_i$$

④ erhöhe $i := i + 1$ und gehe nach ②

Bemerkung Diese Strategie lässt sich auch leicht verbinden mit einer Änderung der Ordnung k des Verfahrens. Berechnet man im Punkt x_{i+k} den Fehlerschätzer $\delta_i = \delta_i(k)$ zusätzlich noch mit dem $(k-1)$- und dem $(k+1)$-Schrittverfahren, dann kann man anhand der Größen

$$\left(\frac{\varepsilon}{\delta_i(k-1)}\right)^{1/k}, \quad \left(\frac{\varepsilon}{\delta_i(k)}\right)^{1/(k+1)}, \quad \left(\frac{\varepsilon}{\delta_i(k+1)}\right)^{1/(k+2)}$$

die Ordnung k

- um eins verringern, falls der erste Term maximal ist,
- beibehalten, falls der zweite Term maximal ist,
- um eins erhöhen, falls der dritte Term maximal ist.

Ein solches Verfahren ist selbststartend, das heißt, es werden keine Anlaufwerte benötigt. Ausgehend von $k = 1$, erhöht das Verfahren automatisch sukzessive die Ordnung k. △

3.7 Steife Differentialgleichungen

Zur Untersuchung der A-Stabilität von Einschrittverfahren haben wir für $\lambda < 0$ die gewöhnliche Differentialgleichung $y' = \lambda y$ betrachtet. Verwenden wir ein lineares Mehrschrittverfahren der Form (3.3) zu ihrer Diskretisierung, so erhalten wir

$$\sum_{\ell=0}^{k} \alpha_\ell \eta_{i+\ell} = h\lambda \sum_{\ell=0}^{k} \beta_\ell \eta_{i+\ell}, \quad i = 0, 1, 2, \ldots$$

beziehungsweise

$$\sum_{\ell=0}^{k} (\alpha_\ell - h\lambda\beta_\ell)\eta_{i+\ell} = 0, \quad i = 0, 1, 2, \ldots.$$

Dies ist eine homogene Differenzengleichung, deren Lösungen offenbar alle beschränkt sein müssen, und zwar für jeden Wert des Parameters $\rho := h\lambda$ mit $\mathrm{Re}(\rho) < 0$. Die Beschränktheit der Folge $\{\eta_i\}$ bei beliebigen Anfangswerten $\eta_0, \ldots, \eta_{k-1}$ ist analog zur Nullstabilität nach Satz 3.11 äquivalent zur Dahlquistschen Wurzelbedingung für das Polynom

$$p_z(\xi) := q(\xi) - \rho r(\xi) \quad \text{mit} \quad q(\xi) := \sum_{\ell=0}^{k} \alpha_\ell \xi^\ell \quad \text{und} \quad r(\xi) := \sum_{\ell=0}^{k} \beta_\ell \xi^\ell. \tag{3.24}$$

Dies bedeutet, alle Nullstellen ξ des Polynoms müssen betragsmäßig kleiner oder gleich 1 sein, wobei sie im Fall $|\xi| = 1$ nur einfach sein dürfen.

Definition 3.18 Ein lineares Mehrschrittverfahren der Form (3.3) heißt **absolut stabil** oder **A-stabil** für ein $z \in \mathbb{C} \setminus \{0\}$, wenn die Nullstellen des zugehörigen **Stabilitätspolynoms** der Dahlquistschen Wurzelbedingung genügen. Die Menge

$$\mathcal{M} := \{z \in \mathbb{C} : \text{Verfahren ist stabil für } z\}$$

wird **Stabilitätsgebiet** des Mehrschrittverfahrens genannt.

Beispiel 3.19 Beim Milne-Simpson-Verfahren für $k = 2$ gilt

$$q(\xi) = \xi^2 - 1, \quad r(\xi) = \frac{1}{3}(\xi^2 + 4\xi + 1)$$

und folglich

$$p_z(\xi) = \left(1 - \frac{z}{3}\right)\xi^2 - \frac{4}{3}z\xi - \frac{z}{3} - 1.$$

Die Nullstellen ξ von $p_z(\xi)$ in Abhängigkeit von z zu finden, ist nicht ganz einfach. Daher wollen wir einmal annehmen, dass die Nullstellen in differenzierbarer Art und Weise von z abhängen. Dies motiviert eine Entwicklung der Form

$$\xi = a_0 + a_1 z + \mathcal{O}(z^2).$$

Der Koeffizient a_0 ergibt sich aus den Nullstellen im Fall $z = 0$, das heißt aus der Gleichung $p_z(\xi) = \xi^2 - 1 = 0$. Wir erhalten $a_0 = \pm 1$. Setzen wir $a_0 = 1$ in die Gleichung $p_z(\xi) = 0$ ein, so erhalten wir

$$\left(1 - \frac{z}{3}\right)\big(1 + a_1 z + \mathcal{O}(z^2)\big)^2 - \frac{4}{3}z\big(1 + a_1 z + \mathcal{O}(z^2)\big) - \frac{z}{3} - 1 = 0.$$

Betrachten wir die bezüglich z linearen Terme dieser Gleichung, so folgt

$$2a_1 z - \frac{z}{3}a_1 - \frac{4}{3}z - \frac{z}{3} = 0,$$

also $a_1 = 1$. Für $a_0 = -1$ ergibt sich entsprechend $a_1 = 1/3$. Die beiden Nullstellen haben demnach die Form

$$\xi_1 = 1 + z + \mathcal{O}(z^2), \quad \xi_2 = -1 + \frac{z}{3} + \mathcal{O}(z^2).$$

Für sehr kleine Schrittweite h ist z betragsmäßig klein und der Einfluss von des quadratischen Terms $\mathcal{O}(z^2)$ kann vernachlässigt werden, so dass die Nullstellen $\xi_1 \approx 1 + h\lambda$ und $\xi_2 = -1 + h\lambda/3$ lauten. Da die Bedingung $|\xi_2| \leq 1$ offenbar nur für $h = 0$ erreicht werden kann, schließen wir $\mathcal{M} = \{0\}$. Dies bedeutet, das Milne-Simpson-Verfahren ist nicht A-stabil. △

Anhand dieses Beispiels sehen wir, dass die Bestimmung des Stabilitätsgebiets im allgemeinen nicht einfach ist. Man kann aber zur praktischen Bestimmung des Stabilitätsgebietes eines linearen Mehrschrittverfahrens ausnutzen, dass dessen Rand dadurch charakterisiert ist, dass (mindestens) eine Nullstelle den Betrag 1 hat. Man stellt die Gleichung $q(\xi) - zr(\xi) = 0$ nach z um und bestimmt für $|\xi| = 1$, also $\xi = e^{i\varphi}$, die zugehörigen z. Dies ergibt die sogenannte *Wurzelortskurve*

$$\Gamma = \left\{ z \in \mathbb{C} : z = \frac{r(e^{i\varphi})}{q(e^{i\varphi})},\ \varphi \in [0, 2\pi) \right\}.$$

Für den Rand $\partial\mathcal{M}$ des Stabilitätsgebiets gilt dann offensichtlich $\partial\mathcal{M} \subset \Gamma$.

Beispiel 3.20 Für $k = 2$ lautet das charakteristische Polynom des Nyström-Verfahrens $p_z(\xi) = \xi^2 - 2z\xi - 1$ mit den Nullstellen $\xi_{1,2} = z \pm \sqrt{z^2 + 1}$. Offensichtlich ist für reelle $z \neq 0$ eine Nullstelle stets betragsmäßig größer als 1. Wegen

$$z = \frac{r(e^{i\varphi})}{q(e^{i\varphi})} = \frac{e^{2i\varphi} - 1}{2e^{i\varphi}} = \frac{1}{2}(e^{i\varphi} - e^{-i\varphi}) = i \sin\varphi$$

ist die Wurzelortskurve gegeben durch

$$\Gamma = \{ z \in \mathbb{C} : z = iy,\ y \in [-1, 1] \}.$$

Für $z \in \Gamma$ besitzt $p_z(\xi)$ zwei verschiedene Nullstellen vom Betrag 1, die für $y = 1$ beziehungsweise $y = -1$ zur doppelten Nullstelle $\xi_{1,2} = i$ beziehungsweise $\xi_{1,2} = -i$ werden. Das Stabilitätsgebiet der expliziten Mittelpunktregel besteht folglich nur aus dem Intervall $(-i, i)$ auf der imaginären Achse. △

Bemerkungen

1. Explizite Mehrschrittverfahren sind niemals A-stabil.
2. Ein A-stabiles Mehrschrittverfahren besitzt höchstens die Konvergenzordnung $p = 2$. Dieses Resultat ist als *zweite Dahlquist-Schranke* bekannt geworden.
3. Die Trapezregel (Adams-Moulton-Verfahren für $k = 2$) ist ein A-stabiles, lineares Mehrschrittverfahren mit Konvergenzordnung $p = 2$.

△

4. Partielle Differentialgleichungen

4.1 Beispiele

Potentialgleichung: Es sei $\Omega \subset \mathbb{R}^2$ ein *Gebiet*, das ist eine offene, zusammenhängende Menge, und $\Gamma := \partial\Omega$ der Rand. Der Graph der Funktion $g : \Gamma \to \mathbb{R}$ beschreibe eine Drahtschlinge, die eine Seifenhaut aufspannt. Diese Seifenhaut lässt sich als Funktion $u : \overline{\Omega} \to \mathbb{R}$ beschreiben, deren Form minimale Oberfläche besitzt

$$\int_\Omega \sqrt{1 + u_x^2 + u_y^2}\,\mathrm{d}x\,\mathrm{d}y \to \min.$$

Wegen $\sqrt{1+z} = 1 + \frac{z}{2} + \mathcal{O}(z^2)$ kann man den Integranden für kleine Werte von u_x und u_y ersetzen durch

$$F(u) := \frac{1}{2}\int_\Omega u_x^2 + u_y^2\,\mathrm{d}x\,\mathrm{d}y \to \min.$$

Ist $u \in C^2(\Omega) \cap C(\overline{\Omega})$ mit $u|_\Gamma = g$ Lösung dieser Minimierungsaufgabe, dann folgt für beliebiges $v \in C^1(\Omega) \cap C(\overline{\Omega})$ mit $v|_\Gamma = 0$, dass

$$0 = \lim_{\varepsilon\to 0} \frac{F(u+\varepsilon v) - F(u)}{\varepsilon} = \int_\Omega u_x v_x + u_y v_y\,\mathrm{d}x\,\mathrm{d}y = \int_\Omega \langle \nabla u, \nabla v\rangle\,\mathrm{d}\mathbf{x}. \tag{4.1}$$

Für $\mathbf{f} := \nabla u v$ liefert der Gaußsche Integralsatz die Identität

$$\int_\Omega \Delta u v\,\mathrm{d}\mathbf{x} + \int_\Omega \langle \nabla u, \nabla v\rangle\,\mathrm{d}\mathbf{x} = \int_\Omega \operatorname{div}\mathbf{f}\,\mathrm{d}\mathbf{x} = \int_\Gamma \langle \mathbf{f}, \mathbf{n}\rangle\,\mathrm{d}\sigma = \int_\Gamma \underbrace{v}_{=0}\frac{\partial u}{\partial \mathbf{n}}\,\mathrm{d}\sigma = 0,$$

wobei $\Delta u = u_{xx} + u_{yy}$ den *Laplace-Operator* bezeichnet. Dies eingesetzt in (4.1) ergibt für u die Bedingung

$$0 = \int_\Omega \Delta u v\,\mathrm{d}\mathbf{x}$$

für alle $v \in C^1(\Omega) \cap C(\overline{\Omega})$ mit $v|_\Gamma = 0$. Daher muss die Funktion u der *Potential-* oder *Laplace-Gleichung*

$$\Delta u(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \tag{4.2}$$

genügen.

Wärmeleitungsgleichung: In einem offenen, beschränkten Gebiet $\Omega \subset \mathbb{R}^d$ beschreibe die Funktion $u : \mathbb{R}_{\geq 0} \times \overline{\Omega} \to \mathbb{R}$ die Temperaturverteilung. Zum Zeitpunkt $t = 0$ liege die Anfangsverteilung $u(0, \mathbf{x}) = u_0(\mathbf{x}) \in C(\overline{\Omega})$ vor. Zusätzlich seien im Gebiet Ω die

Wärmequelle $f \in C(\mathbb{R}_{>0} \times \Omega)$ und an dessen Rand $\Gamma = \partial\Omega$ die Temperaturverteilung $g \in C\big(\mathbb{R}_{>0} \times \Gamma\big)$ vorgegeben.
Aus dem Erhaltungssatz folgt nun für jedes Kontrollvolumen $V \subset \Omega$

$$\underbrace{\int_V \frac{\partial}{\partial t} u(t,\mathbf{x})\,\mathrm{d}\mathbf{x}}_{\text{Wärmegehalt in } V} = \underbrace{-\int_{\partial V} \langle \mathbf{q}(t,\mathbf{x}), \mathbf{n}(\mathbf{x})\rangle\,\mathrm{d}\sigma}_{\text{Wärmefluss von außen}} + \underbrace{\int_V f(t,\mathbf{x})\,\mathrm{d}\mathbf{x}}_{\text{Wärmequelle}}.$$

Dem Materialgesetz gemäß genügt der Wärmefluss der Beziehung

$$\mathbf{q}(t,\mathbf{x}) = -c(\mathbf{x})\nabla u(t,\mathbf{x})$$

mit der materialabhängigen Wärmeleitkonstante $c(\mathbf{x}) \geq c_0 > 0$. Eingesetzt in den Gaußschen Integralsatz folgt daher

$$-\int_{\partial V} \langle \mathbf{q}(t,\mathbf{x}), \mathbf{n}(\mathbf{x})\rangle\,\mathrm{d}\sigma = -\int_V \operatorname{div} \mathbf{q}(t,\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_V \operatorname{div}\big(c(\mathbf{x})\nabla u(t,\mathbf{x})\big)\,\mathrm{d}\mathbf{x}.$$

Für die Temperaturverteilung folgt somit für alle Kontrollvolumen V die Gleichung

$$\int_V \left\{ \frac{\partial}{\partial t} u(t,\mathbf{x}) - \operatorname{div}\big(c(\mathbf{x})\nabla u(t,\mathbf{x})\big) \right\} \mathrm{d}\mathbf{x} = \int_V f(t,\mathbf{x})\,\mathrm{d}\mathbf{x},$$

dies bedeutet

$$\frac{\partial}{\partial t} u(t,\mathbf{x}) - \operatorname{div}\big(c(\mathbf{x})\nabla u(t,\mathbf{x})\big) = f(t,\mathbf{x}), \quad (t,\mathbf{x}) \in \mathbb{R}_{>0} \times \Omega.$$

Ist c konstant, etwa $c = 1$, so genügt die Temperaturverteilung $u \in C^2(\mathbb{R}_{>0}\times\Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$ der Gleichung

$$\frac{\partial}{\partial t} u(t,\mathbf{x}) - \Delta u(t,\mathbf{x}) = f(t,\mathbf{x}), \quad (t,\mathbf{x}) \in \mathbb{R}_{\geq 0} \times \Omega \tag{4.3}$$

mit dem d-dimensionalen Laplace-Operator $\Delta = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_d^2}$.

Poisson-Gleichung: Sind die Daten f und g der Wärmeleitungsgleichung nicht zeitabhängig, dann stellt sich für $t \to \infty$ ein Gleichgewichtszustand ein. Dies bedeutet, es gilt $\partial u/\partial t = 0$ und (4.3) geht über in die *Poisson-Gleichung*

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{4.4}$$

Diese Gleichung hat in der Elektrostatik ebenfalls eine immense Bedeutung: ist in Ω die Ladungsdichte $f : \Omega \to \mathbb{R}$ bekannt, so genügt die Spannung u dort der Poisson-Gleichung.

Wellengleichung: Die Bewegung in einem idealen Gas wird durch drei Gesetze bestimmt. Wie üblich wird die Geschwindigkeit mit $\mathbf{v}$, die Dichte mit ρ und der Druck mit p bezeichnet.

1. *Kontinuitätsgleichung:*

$$\frac{\partial \rho}{\partial t} = -\rho_0 \operatorname{div} \mathbf{v}.$$

Wegen der Massenerhaltung ist die Änderung der Masse in einem Kontrollvolumen V gleich dem Fluss durch die Oberfläche, das ist $\int_{\partial V} \rho\langle \mathbf{v}, \mathbf{n}\rangle\,\mathrm{d}\sigma$. Aus dem Gaußschen Integralsatz folgt daraus die Gleichung $\partial\rho/\partial t = -\operatorname{div}(\rho\mathbf{v})$. Die Approximation von ρ durch eine konstante, zeitlich unabhängige Dichte ρ_0 ergibt dann die obige Gleichung.

2. *Newtonsches Gesetz:*
$$\rho_0 \frac{\partial \mathbf{v}}{\partial t} = -\nabla p.$$
Der Druckgradient induziert ein Kraftfeld, das die Beschleunigung der Teilchen bewirkt.
3. *Zustandsgleichung:*
$$p = c^2 \rho.$$

In idealen Gasen ist der Druck bei konstanter Temperatur proportional zur Dichte. Aus den drei Gesetzen folgt

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 \rho}{\partial t^2} = -c^2 \frac{\partial}{\partial t}(\rho_0 \operatorname{div} \mathbf{v}) = -c^2 \operatorname{div}\left(\rho_0 \frac{\partial \mathbf{v}}{\partial t}\right) = c^2 \operatorname{div}(\nabla p) = c^2 \Delta p.$$

Andere Beispiele für die *Wellengleichung*

$$\frac{\partial^2 p}{\partial t^2}(t, \mathbf{x}) = c^2 \Delta p(t, \mathbf{x}), \quad (t, \mathbf{x}) \in \mathbb{R}_{>0} \times \Omega \tag{4.5}$$

ergeben sich in zwei Raumdimensionen für eine schwingende Membran oder in einer Raumdimension für eine schwingende Saite.

4.2 Charakterisierung

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet und $\mathcal{L} : C^2(\Omega) \to C(\Omega)$ ein allgemeiner linearer Differentialoperator zweiter Ordnung

$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}) + \sum_{i=1}^{d} b_i(\mathbf{x}) \frac{\partial}{\partial x_i} u(\mathbf{x}) + c(\mathbf{x}) u(\mathbf{x}), \tag{4.6}$$

wobei $\mathbf{A} = [a_{i,j}]_{i,j=1}^d \in [C(\Omega)]^{d \times d}$, $\mathbf{b} = [b_i]_{i=1}^d \in [C(\Omega)]^d$ und $c \in C(\Omega)$. Die zugehörige Differentialgleichung lautet dann

$$(\mathcal{L}u)(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{4.7}$$

Da für $u \in C^2(\Omega)$ die zweiten Ableitungen symmetrisch sind, also $\partial^2 u/(\partial x_i \partial x_j) = \partial^2 u/(\partial x_j \partial x_i)$, kann ohne Beschränkung der Allgemeinheit $a_{i,j} = a_{j,i}$ angenommen werden. Demnach ist die Matrix $\mathbf{A}$ symmetrisch und besitzt nur reelle Eigenwerte. Der Differentialoperator $-\sum_{i,j=1}^d a_{i,j} \partial^2/(\partial x_i \partial x_j)$ ist der *Hauptteil* von $\mathcal{L}$.

Definition 4.1 Der Differentialoperator (4.6) heißt

- **elliptisch in $\mathbf{x}$**, falls die Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind,
- **parabolisch in $\mathbf{x}$**, falls $d-1$ Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind und ein Eigenwert verschwindet, aber $\operatorname{rang}\big([\mathbf{A}(\mathbf{x}), \mathbf{b}(\mathbf{x})]\big) = d$ ist,
- **hyperbolisch in $\mathbf{x}$**, falls $d-1$ Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind und ein Eigenwert negatives Vorzeichen besitzt.

Der Differentialoperator (4.6) heißt **elliptisch/parabolisch/hyperbolisch (in Ω)**, falls er elliptisch/parabolisch/hyperbolisch ist für alle $\mathbf{x} \in \Omega$. Entsprechend wird die Differentialgleichung (4.7) elliptisch/parabolisch/hyperbolisch genannt, wenn der zugehörige Differentialoperator diese Eigenschaft besitzt.

Beispiel 4.2 Die Potentialgleichung (4.2) und die Poisson-Gleichung (4.4) sind elliptisch, die Wärmeleitungsgleichung (4.3) ist parabolisch, während die Wellengleichung (4.5) hyperbolisch ist. △

Zusätzlich zur Differentialgleichung (4.7) müssen noch geeignete Anfangs- oder Randbedingungen gefordert werden, um eine sachgemäße Aufgabenstellung zu ergeben.

Definition 4.3 Ein Problem heißt **sachgemäß gestellt**, wenn eine Lösung existiert, diese eindeutig ist und stetig von den vorgegebenen Daten abhängt. Andernfalls heißt das Problem **schlecht gestellt**.

Die Unterscheidung partieller Differentialgleichungen in verschiedene Typen ergäbe keinen Sinn, wenn nicht jeder Typ grundlegend andere Eigenschaften hätte.

1. *Elliptische Differentialgleichungen:* Bei elliptischen Problemen werden Randbedingungen vorgegeben: für gegebenes $f \in C(\Omega)$ und $g \in C(\Gamma)$ suche $u \in C^2(\Omega) \cap C(\overline{\Omega})$, so dass
$$\mathcal{L}u = f \text{ in } \Omega, \quad u = g \text{ auf } \Gamma. \tag{4.8}$$
Diese Randbedingungen heißen *Dirichlet-Randbedingungen*. In der Praxis treten oft auch *Neumann-Randbedingungen*, $\partial u/\partial \mathbf{n} = g$ auf Γ, auf. Lösungen elliptischer Differentialgleichungen erfüllen das Maximumprinzip (siehe nächster Abschnitt).
2. *Parabolische Differentialgleichungen:* Parabolische Differentialgleichungen beschreiben Diffusionsvorgänge. Die ausgezeichnete Koordinatenrichtung ist in der Regel die Zeit, so dass man oftmals die Differentialgleichung auf die Form $u_t + \mathcal{L}u = f$ bringen kann, wobei $\mathcal{L}$ ein elliptischer Differentialoperator ist. Zusätzlich werden Anfangsrandwerte vorgegeben: für gegebenes $f \in C(\mathbb{R}_{>0} \times \Omega)$, $g \in C(\mathbb{R}_{>0} \times \Gamma)$ und $u_0 \in C(\overline{\Omega})$ suche $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$, so dass
$$\begin{aligned} u_t + \mathcal{L}u = f &\quad \text{in } \mathbb{R}_{>0} \times \Omega \\ u = g &\quad \text{auf } \mathbb{R}_{>0} \times \Gamma \quad \text{(Randbedingung)} \\ u(0,\cdot) = u_0 &\quad \text{auf } \overline{\Omega} \quad \text{(Anfangsbedingung)} \end{aligned}$$
3. *Hyperbolische Differentialgleichungen:* Hier ist ebenfalls eine Koordinate ausgezeichnet, die wieder als Zeit interpretiert werden kann. Daher lässt sich die Differentialgleichung oft schreiben als $u_{tt} + \mathcal{L}u = f$ mit einem elliptischen Differentialoperator $\mathcal{L}$. Hyperbolische Gleichungen beschreiben physikalisch gesehen Schwingungsvorgänge. Sinnvolle Probleme erhält man mit Anfangsbedingungen: für gegebenes $f \in C(\mathbb{R}_{>0} \times \Omega)$, $g \in C(\mathbb{R}_{>0} \times \Gamma)$ und $u_0, u_1 \in C(\overline{\Omega})$ suche $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$, so dass
$$\begin{aligned} u_{tt} + \mathcal{L}u = f &\quad \text{in } \mathbb{R}_{>0} \times \Omega \\ u = g &\quad \text{auf } \mathbb{R}_{>0} \times \Gamma \quad \text{(Randbedingung)} \\ u(0,\cdot) = u_0,\ u_t(0,\cdot) = u_1 &\quad \text{auf } \overline{\Omega} \quad \text{(Anfangsbedingungen)} \end{aligned}$$

Wenn der Differentialoperator invariant gegenüber Bewegungen ist (also gegenüber Translation und Drehung), dann hat der elliptische Anteil $\mathcal{L}$ die Form

$$\mathcal{L}u = -a\Delta u + cu.$$

4.3 Maximumprinzip

Bei der Analyse von Differenzenverfahren spielt das diskrete Analogon des Maximumprinzips eine wichtige Rolle. Deshalb betrachten wir vorab eine einfache Fassung des Prinzips. Dazu seien $\Omega \subset \mathbb{R}^d$ stets ein beschränktes Gebiet und der elliptische Differentialoperator von der Form

$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) u_{x_i,x_j}(\mathbf{x}). \tag{4.9}$$

Satz 4.4 (Maximumprinzip) Die Funktion $u \in C^2(\Omega) \cap C(\overline{\Omega})$ genüge der elliptischen Differentialgleichung $\mathcal{L}u = f \leq 0$ in Ω. Dann nimmt u sein Maximum auf dem Rand Γ an.

Beweis. (*i*) Wir führen den Beweis zunächst unter der stärkeren Voraussetzung $f < 0$. Angenommen, es sei $\mathbf{y} \in \Omega$ mit

$$u(\mathbf{y}) = \sup_{\mathbf{x}\in\Omega} u(\mathbf{x}) > \max_{\mathbf{x}\in\Gamma} u(\mathbf{x}).$$

Bei einer linearen Koordinatentransformation $\mathbf{x} \mapsto \boldsymbol{\xi} = \mathbf{U}\mathbf{x}$ lautet der Differentialoperator in den neuen Koordinaten

$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} \left[\mathbf{U}\mathbf{A}(\mathbf{x})\mathbf{U}^T\right]_{i,j} u_{\xi_i,\xi_j}(\mathbf{x}),$$

wobei $\mathbf{A}(\mathbf{x}) = [a_{i,j}(\mathbf{x})]_{i,j=1}^d$ die Koeffizientenmatrix ist. Wegen der Symmetrie von $\mathbf{A}(\mathbf{x})$ können wir eine orthogonale Matrix $\mathbf{U}$ wählen, mit der $\mathbf{U}\mathbf{A}(\mathbf{y})\mathbf{U}^T$ diagonal wird. Aus der positiven Definitheit schließen wir, dass die Diagonalelemente positiv sind. Weil $\mathbf{y}$ Extremalpunkt ist, gilt

$$\nabla u(\mathbf{y}) = \mathbf{0}, \quad u_{\xi_i,\xi_i}(\mathbf{y}) \leq 0.$$

Dies bedeutet

$$(\mathcal{L}u)(\mathbf{y}) = -\sum_{i,j=1}^{d} \left[\mathbf{U}\mathbf{A}(\mathbf{y})\mathbf{U}^T\right]_{i,j} u_{\xi_i,\xi_j}(\mathbf{y}) \geq 0$$

im Widerspruch zu $(\mathcal{L}u)(\mathbf{y}) = f(\mathbf{y}) < 0$.

(*ii*) Sei nun $f \leq 0$ angenommen und es gebe ein $\mathbf{y} \in \Omega$ mit $u(\mathbf{y}) > \max_{\mathbf{x}\in\Gamma} u(\mathbf{x})$. Die Hilfsfunktion

$$h(\mathbf{x}) := \|\mathbf{x} - \mathbf{y}\|_2^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_d - y_d)^2$$

ist auf Γ beschränkt. Wenn $\delta > 0$ hinreichend klein gewählt wird, nimmt also auch die Funktion

$$w = u + \delta h$$

ihr Maximum in einem Punkt $\mathbf{z}$ im Innern an. Wegen $h_{x_i,x_j} = 2\delta_{i,j}$ ist

$$(\mathcal{L}w)(\mathbf{x}) = (\mathcal{L}u)(\mathbf{x}) + \delta(\mathcal{L}h)(\mathbf{x}) = f(\mathbf{x}) - 2\delta \sum_{i=1}^{d} a_{i,i}(\mathbf{x}) < 0$$

für alle $\mathbf{x} \in \Omega$. Wie im ersten Teil des Beweises ergibt sich daraus ein Widerspruch. □

Folgerungen:

1. *Minimumprinzip:* Ist $\mathcal{L}u = f \geq 0$ in Ω, so nimmt u sein Minimum auf dem Rand Γ an.

   Beweis. Man wende auf $v := -u$ das Maximumprinzip an. □

2. *Vergleichsprinzip:* Wenn für $u, v \in C^2(\Omega) \cap C(\overline{\Omega})$ gilt

   $$\mathcal{L}u \leq \mathcal{L}v \text{ in } \Omega, \quad u \leq v \text{ auf } \Gamma,$$

   so folgt $u \leq v$ in Ω.

   Beweis. Für $w := v - u$ ist nach Voraussetzung $\mathcal{L}w = \mathcal{L}v - \mathcal{L}u \geq 0$ und auf Γ auch $w \geq 0$. Nach dem Minimumprinzip folgt $\inf_{\mathbf{x}\in\Omega} w(\mathbf{x}) \geq 0$ und folglich $v(\mathbf{x}) \geq u(\mathbf{x})$ für alle $\mathbf{x} \in \Omega$. □

3. *Eindeutigkeit der Lösung:* Die Lösung des Dirichlet-Problems (4.8) ist eindeutig.

   Beweis. Seien u_1 und u_2 zwei Lösungen von (4.8), dann erfüllt $v = u_1 - u_2$ die Gleichung

   $$\mathcal{L}v = 0 \text{ in } \Omega, \quad v = 0 \text{ auf } \Gamma.$$

   Minimum- und Maximumprinzip implizieren

   $$0 = \inf_{\mathbf{z}\in\Omega} v(\mathbf{z}) \leq v(\mathbf{x}) \leq \sup_{\mathbf{z}\in\Omega} v(\mathbf{z}) = 0, \quad \mathbf{x} \in \Omega.$$

   □

4. *Stetige Abhängigkeit von den Randdaten:* Die Lösung des Dirichlet-Problems (4.8) hängt stetig von den Randdaten ab. Sind u_1 und u_2 Lösungen zu verschiedenen-Randwerten, so ist

   $$\max_{\mathbf{x}\in\overline{\Omega}} |u_1(\mathbf{x}) - u_2(\mathbf{x})| = \max_{\mathbf{x}\in\Gamma} |u_1(\mathbf{x}) - u_2(\mathbf{x})|.$$

   Beweis. Für $v := u_1 - u_2$ ist $\mathcal{L}v = 0$. Aus dem Maximumprinzip folgt

   $$v(\mathbf{x}) \leq \max_{\mathbf{z}\in\Gamma} v(\mathbf{z}) \leq \max_{\mathbf{z}\in\Gamma} |v(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

   Ebenso liefert das Minimumprinzip die Aussage

   $$v(\mathbf{x}) \geq \min_{\mathbf{z}\in\Gamma} v(\mathbf{z}) \geq -\max_{\mathbf{z}\in\Gamma} |v(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

   □

Definition 4.5 Ein linearer Differentialoperator $\mathcal{L}$ zweiter Ordnung heißt **gleichmäßig elliptisch**, wenn ein $\alpha > 0$ existiert, so dass die Koeffizientenmatrix $\mathbf{A}(\mathbf{x}) = [a_{i,j}(\mathbf{x})]_{i,j=1}^d$ der Abschätzung

$$\boldsymbol{\xi}^T \mathbf{A}(\mathbf{x}) \boldsymbol{\xi} \geq \alpha \|\boldsymbol{\xi}\|_2^2$$

für alle $\boldsymbol{\xi} \in \mathbb{R}^d$ und $\mathbf{x} \in \Omega$ genügt. Die Zahl α wird als **Elliptizitätskonstante** bezeichnet.

5. *Stetige Abhängigkeit von der rechten Seite:* Der Operator $\mathcal{L}$ der Form (4.9) sei gleichmäßig elliptisch in Ω. Dann gibt es eine nur von Ω und der Elliptizitätskonstante α abhängige Zahl c, so dass für jedes $u \in C^2(\Omega) \cap C(\overline{\Omega})$ gilt

$$|u(\mathbf{x})| \le \max_{\mathbf{z}\in\Gamma} |u(\mathbf{z})| + c \sup_{\mathbf{z}\in\Omega} |(\mathcal{L}u)(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

Beweis. Sei $\Omega \subset \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < R\}$ und setze

$$w(\mathbf{x}) = R^2 - \sum_{i=1}^{d} x_i^2.$$

Im Hinblick auf $w_{x_i,x_j} = -2\delta_{i,j}$ ist offensichtlich

$$\mathcal{L}w \ge 2\alpha \quad \text{und} \quad 0 \le w \le R^2 \quad \text{in } \Omega,$$

wobei α die Elliptizitätskonstante ist. Für

$$v(\mathbf{x}) := \max_{\mathbf{z}\in\Gamma} |u(\mathbf{z})| + w(\mathbf{x}) \frac{1}{2\alpha} \sup_{\mathbf{z}\in\Omega} |(\mathcal{L}u)(\mathbf{z})|$$

ist nach Konstruktion $\mathcal{L}v \ge |\mathcal{L}u|$ in Ω und $v \ge |u|$ auf Γ. Das Vergleichsprinzip liefert $-v(\mathbf{x}) \le u(\mathbf{x}) \le v(\mathbf{x})$ für alle $\mathbf{x} \in \Omega$ und wegen $w \le R^2$ erhalten wir die gewünschte Abschätzung mit $c = R^2/(2\alpha)$. □

6. *Elliptische Operatoren mit Term der Ordnung* 0*:* Für den allgemeineren Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = c(\mathbf{x})u(\mathbf{x}) - \sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) u_{x_i,x_j}(\mathbf{x}) \quad \text{mit} \quad c(\mathbf{x}) \ge 0$$

gilt ein abgeschwächtes Maximumprinzip. Aus $\mathcal{L}u \le 0$ folgt

$$\max_{\mathbf{x}\in\overline{\Omega}} u(\mathbf{x}) \le \max\{0, \max_{\mathbf{x}\in\Gamma} u(\mathbf{x})\}.$$

Beweis. Ein Beweis ist nur für $\mathbf{y} \in \Omega$ und $u(\mathbf{y}) = \sup_{\mathbf{x}\in\Omega} u(\mathbf{x}) > 0$ erforderlich. Dann ist $(\mathcal{L}u)(\mathbf{y}) - c(\mathbf{y})u(\mathbf{y}) \le (\mathcal{L}u)(\mathbf{y}) \le 0$. Außerdem ist durch den Hauptteil $\mathcal{L}u - cu$ ein elliptischer Operator der Form (4.9) definiert. Deshalb kann der Beweis wie für Satz 4.4 vollzogen werden. □

5. Finite-Differenzen-Verfahren

5.1 Poisson-Gleichung

Im folgenden wollen wir uns auf die Poisson-Gleichung beschränken. Dazu seien $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet, $f \in C(\Omega)$ und $g \in C(\Gamma)$. Gesucht ist $u \in C^2(\Omega) \cap C(\overline{\Omega})$, so dass

$$-\Delta u = f \text{ in } \Omega, \quad u = g \text{ auf } \Gamma.$$

Definition 5.1 Eine Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ der Poisson-Gleichung ist eine **klassische Lösung**. Gilt speziell $f = 0$, das heißt, ist $\Delta u = 0$ in Ω, so ist u **harmonisch**.

Wir werden mit Lösung stets die klassische Lösung meinen. Um diese zu berechnen, benötigen wir finite Differenzen:

Definition 5.2 Für $u \in C(\mathbb{R}^d)$ und eine Richtung $1 \leq j \leq d$ definieren wir die **Vorwärts-** oder **rechtsseitige Differenz** durch

$$(\partial_j^{+h} u)(\mathbf{x}) := \frac{u(\mathbf{x} + h\mathbf{e}_j) - u(\mathbf{x})}{h},$$

die **Rückwärts-** oder **linksseitige Differenz** durch

$$(\partial_j^{-h} u)(\mathbf{x}) := \frac{u(\mathbf{x}) - u(\mathbf{x} - h\mathbf{e}_j)}{h}$$

und die **symmetrische** oder **zentrale Differenz** durch

$$(\partial_j^{h} u)(\mathbf{x}) := \frac{u(\mathbf{x} + h\mathbf{e}_j) - u(\mathbf{x} - h\mathbf{e}_j)}{2h}.$$

Lemma 5.3 Ist $\{\mathbf{x}+th\mathbf{e}_j : |t|\le 1\}\subset\overline{\Omega}$ und $u\in C^4(\overline{\Omega})$, dann gilt

$$\frac{\partial u}{\partial \mathbf{e}_j}(\mathbf{x}) = (\partial_j^{\pm h}u)(\mathbf{x}) + R_1^{\pm}, \quad |R_1^{\pm}| \le \frac{h}{2}\|u\|_{C^2(\overline{\Omega})},$$
$$\frac{\partial u}{\partial \mathbf{e}_j}(\mathbf{x}) = (\partial_j^{h}u)(\mathbf{x}) + R_2, \qquad |R_2| \le \frac{h^2}{6}\|u\|_{C^3(\overline{\Omega})},$$

und

$$\begin{aligned}\frac{\partial^2 u}{\partial \mathbf{e}_j^2}(\mathbf{x}) &= (\partial_j^{-h}\partial_j^{+h}u)(\mathbf{x}) + R_3\\ &= \frac{u(\mathbf{x}+h\mathbf{e}_j) - 2u(\mathbf{x}) + u(\mathbf{x}-h\mathbf{e}_j)}{h^2} + R_3, \quad |R_3| \le \frac{h^2}{12}\|u\|_{C^4(\overline{\Omega})}.\end{aligned}$$

Beweis. Es genügt, die Behauptung im Eindimensionalen zu beweisen. Taylor-Entwicklung von u liefert

$$u(x\pm h) = u(x) \pm hu'(x) + \frac{h^2}{2}u''(\xi), \quad \xi\in(x, x\pm h),$$

woraus sofort die erste Aussage folgt. Subtrahieren wir ferner

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(\xi_1), \quad \xi_1\in(x-h,x),$$
$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(\xi_2), \quad \xi_2\in(x,x+h),$$

so folgt die zweite Aussage

$$u(x+h) - u(x-h) = 2hu'(x) + \frac{h^3}{6}\big(u'''(\xi_2) + u'''(\xi_1)\big).$$

Schließlich folgt aus Addition der drei Gleichungen

$$\begin{aligned}u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_1), \quad \xi_1\in(x-h,x),\\ -2u(x) &= -2u(x),\\ u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_2), \quad \xi_2\in(x,x+h),\end{aligned}$$

dass

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + \frac{h^2}{24}\big(u^{(4)}(\xi_1) + u^{(4)}(\xi_2)\big).$$

□

Zur Diskretisierung wird über das Gebiet Ω ein *Gitter* mit Maschenweite h gelegt

$$\Omega_h := \big\{\mathbf{x}\in\Omega : \mathbf{x} = h\mathbf{k} \text{ mit } \mathbf{k}\in\mathbb{Z}^d\big\},$$
$$\Gamma_h := \{\mathbf{x}\in\Gamma : \exists 1\le i\le d \text{ mit } x_i = hk,\ k\in\mathbb{Z}\}.$$

In Anlehnung an $\overline{\Omega} = \Omega\cup\Gamma$ setzen wir $\overline{\Omega}_h := \Omega_h\cup\Gamma_h$. Punkte aus Γ_h werden *Randpunkte* genannt. Ein Gitterpunkt $\mathbf{x}\in\Omega_h$, der einen Nachbarn aus Γ_h besitzt, heißt *randnah*. Alle

anderen Punkte aus Ω_h sind *randfern*. Ist $\overline{\Omega}$ die Vereinigung von Würfeln der Kantenlänge h, so sprechen wir von einem *Würfelgebiet*. In diesem Fall besitzen dann auch alle Rand- und randnahen Punkte immer den Abstand h zu ihren Nachbarn.

In den Randpunkten $\mathbf{x}$ aus Γ_h ist $u(\mathbf{x})$ durch die Randwerte $g(\mathbf{x})$ vorgegeben. Hingegen erhält man für jeden Punkt $\mathbf{x}$ aus Ω_h eine Gleichung für $u(\mathbf{x})$, indem man die Poisson-Gleichung durch Differenzenquotienten approximiert. In jedem randfernen Gitterpunkt $\mathbf{x}$ diskretisieren wir den Laplace-Operator durch $(\Delta_h u)(\mathbf{x}) := \sum_{i=1}^{d} (\partial_i^{-h} \partial_i^{+h} u)(\mathbf{x})$, wobei sich

$$(\Delta u)(\mathbf{x}) = \sum_{i=1}^{d} (\partial_i^{-h} \partial_i^{+h} u)(\mathbf{x}) + \mathcal{O}(h^2)$$

ergibt. Für $d = 2$ erhält man den sogenannten *5-Punkte-Differenzenstern*

$$\begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NO} \\ \alpha_W & \alpha_Z & \alpha_O \\ \alpha_{SW} & \alpha_S & \alpha_{SO} \end{bmatrix}_\star = \frac{1}{h^2} \begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}_\star .$$

Dieser ist für Würfelgebiete ausreichend.

In beliebigen Gebieten muss der Differenzenquotient für randnahe Punkte entsprechend modifiziert werden. Für $u \in C^3(\overline{\Omega})$ erhält man durch Taylor-Entwicklung in einer Dimension

$$u_{xx} = \frac{2}{h_O(h_O + h_W)} u_O - \frac{2}{h_O h_W} u_Z + \frac{2}{h_W(h_O + h_W)} u_W + \mathcal{O}(h)$$

und in zwei Dimensionen

$$\begin{aligned} \Delta u = \Delta_h u + \mathcal{O}(h) = {} & \frac{2}{h_O(h_O + h_W)} u_O + \frac{2}{h_W(h_O + h_W)} u_W + \frac{2}{h_S(h_S + h_N)} u_S \\ & + \frac{2}{h_N(h_S + h_N)} u_N - \left(\frac{2}{h_O h_W} + \frac{2}{h_S h_N} \right) u_Z + \mathcal{O}(h). \end{aligned}$$

Hierbei bezeichnet h die jeweils größte Schrittweite, das heißt, $h := \max\{h_W, h_O\}$ beziehungsweise $h := \max\{h_W, h_O, h_S, h_N\}$. Diese Diskretisierung des Laplace-Operators wird auch *Shortley-Weller-Approximation* genannt.

Beispiel 5.4 *Eindimensionaler Fall:* Sei

$$-u_{xx} = f \text{ in } (a, b), \quad u(a) = \alpha, \quad u(b) = \beta.$$

Für $h = (b-a)/n$ und $x_i = a + hi$, $i = 1, \ldots, n-1$, setzen wir $u_i = u(x_i)$ und $f_i = f(x_i)$. Dann erhalten wir das lineare Gleichungssystem

$$\frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} f_1 + \alpha/h^2 \\ f_2 \\ \vdots \\ f_{n-2} \\ f_{n-1} + \beta/h^2 \end{bmatrix}.$$

Zweidimensionales Würfelgebiet: Zur Lösung der Poisson-Gleichung im Einheitsquadrat

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma$$

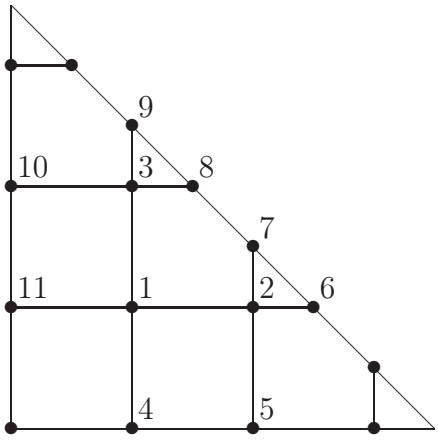
überziehen wir Ω mit einem Gitter der Maschenweite $h = 1/n$. Das entstehende Gleichungssystem wird übersichtlicher bei Benutzung von Doppelindizes $u_{i,j} = u(ih, jh)$, $1 \le i, j < n$. Es ergibt sich das Gleichungssystem

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = f_{i,j}, \quad 1 \le i, j < n$$

mit $f_{i,j} = h^2 f(ih, jh)$. Die Terme mit Indizes 0 oder n gelten als nicht geschrieben.
Im Fall von Würfelgebieten ist die Systemmatrix immer symmetrisch. Ordnet man die Indizes schachbrettartig an (zum Beispiel durch abwechselndes rotes und schwarzes Einfärben der zugehörigen Gitterpunkte), so erhält man ein Gleichungssystem der Form

$$\left[\begin{array}{cc|cc} 4 & & & \\ & \ddots & & \mathbf{A} \\ & & 4 & \\ \hline & & 4 & \\ & \mathbf{A}^T & & \ddots \\ & & & 4 \end{array}\right] \mathbf{u} = \mathbf{f}.$$

Beliebiges zweidimensionales Gebiet: Sei Ω ein rechtwinkliges gleichschenkliges Dreieck mit Katheten der Länge 7:



Zu lösen sei die Laplace-Gleichung mit Dirichlet-Randbedingungen. Für $h = 2$ enthält Ω_h drei Punkte. Es entsteht ein lineares Gleichungssystem für u_1, u_2 und u_3:

$$\begin{aligned} u_1 - \frac{u_2}{4} - \frac{u_3}{4} &= \frac{u_4}{4} + \frac{u_{11}}{4} \\ -\frac{u_1}{6} + u_2 &= \frac{u_5}{6} + \frac{u_6}{3} + \frac{u_7}{3} \\ -\frac{u_1}{6} + u_3 &= \frac{u_8}{3} + \frac{u_9}{3} + \frac{u_{10}}{6}. \end{aligned}$$

Man beachte, dass das System unsymmetrisch ist! △

Zusammengefasst erhalten wir demnach ein *Differenzenverfahren* für die näherungsweise Lösung des Poisson-Problems: suche eine *Gitterfunktion* $u_h : \overline{\Omega}_h \to \mathbb{R}$, so dass

$$\begin{aligned} -(\Delta_h u_h)(\mathbf{x}) &= f(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \Omega_h, \\ u_h(\mathbf{x}) &= g(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \Gamma_h. \end{aligned} \tag{5.1}$$

Sammelt man alle Unbekannten im Vektor $\mathbf{u}_h$, so führt (5.1) auf ein Gleichungsystem $\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h$.

5.2 Beliebige elliptische Differentialoperatoren

Der allgemeine elliptische Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} a_{i,j}(\mathbf{x})\frac{\partial^2}{\partial x_i \partial x_j}u(\mathbf{x}) + \sum_{i=1}^{d} b_i(\mathbf{x})\frac{\partial}{\partial x_i}u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x})$$

wird diskretisiert durch

$$(\mathcal{L}_h u)(\mathbf{x}) = \left[-\sum_{i=1}^{d} a_{i,i}(\mathbf{x})\partial_i^{-h}\partial_i^{+h} - \sum_{\substack{i,j=1 \\ i\neq j}}^{d} a_{i,j}(\mathbf{x})\partial_i^h\partial_j^h + \sum_{i=1}^{d} b_i(\mathbf{x})\partial_i^h + c(\mathbf{x}) \right] u(\mathbf{x}).$$

Falls $u \in C^4(\overline{\Omega})$ ist, dann gilt $|(\mathcal{L}u)(\mathbf{x}) - (\mathcal{L}_h u)(\mathbf{x})| = \mathcal{O}(h^2)$.

Beispiel 5.5 Im Zweidimensionalen ergibt sich

$$\begin{aligned}(\mathcal{L}_h u)(\mathbf{x}) &= \big[-a_{1,1}(\mathbf{x})\partial_1^{-h}\partial_1^{+h} - 2a_{1,2}(\mathbf{x})\partial_1^h\partial_2^h - a_{2,2}(\mathbf{x})\partial_2^{-h}\partial_2^{+h}\big]u(\mathbf{x}) \\ &\qquad + \big[b_1(\mathbf{x})\partial_1^h + b_2(\mathbf{x})\partial_2^h\big]u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x}) \\ &= \frac{1}{2h^2}\begin{bmatrix} a_{1,2}(\mathbf{x}) & -2a_{2,2}(\mathbf{x}) & -a_{1,2}(\mathbf{x}) \\ -2a_{1,1}(\mathbf{x}) & 4[a_{1,1}(\mathbf{x}) + a_{2,2}(\mathbf{x})] & -2a_{1,1}(\mathbf{x}) \\ -a_{1,2}(\mathbf{x}) & -2a_{2,2}(\mathbf{x}) & a_{1,2}(\mathbf{x}) \end{bmatrix}_\star u(\mathbf{x}) \\ &\qquad + \frac{1}{2h}\begin{bmatrix} 0 & b_2(\mathbf{x}) & 0 \\ -b_1(\mathbf{x}) & 0 & b_1(\mathbf{x}) \\ 0 & -b_2(\mathbf{x}) & 0 \end{bmatrix}_\star u(\mathbf{x}) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & c(\mathbf{x}) & 0 \\ 0 & 0 & 0 \end{bmatrix}_\star u(\mathbf{x}).\end{aligned}$$

△

So schön dieser Stern auch ist, so lässt sich dennoch im allgemeinen keine Stabilität nachweisen. Dies liegt an der Diskretisierung der gemischten Ableitung $\partial^2/(\partial x_1 \partial x_2)$, die wir in Abhängigkeit vom Vorzeichen von $a_{1,2}(\mathbf{x})$ wie folgt modifizieren. Wir wählen

$$\frac{1}{2h^2}\begin{bmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 0 \end{bmatrix}_\star \text{ falls } a_{1,2}(\mathbf{x}) \geq 0 \text{ bzw. } \frac{1}{2h^2}\begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}_\star \text{ falls } a_{1,2}(\mathbf{x}) < 0.$$

Mit $a_{1,2}^+ := \max\{a_{1,2}, 0\}$ und $a_{1,2}^- := \min\{a_{1,2}, 0\}$ erhalten wir den Siebenpunktstern

$$\begin{aligned}&(\mathcal{L}_h u)(\mathbf{x}) \\ &= \frac{1}{h^2}\begin{bmatrix} a_{1,2}^-(\mathbf{x}) & |a_{1,2}(\mathbf{x})| - a_{2,2}(\mathbf{x}) & -a_{1,2}^+(\mathbf{x}) \\ |a_{1,2}(\mathbf{x})| - a_{1,1}(\mathbf{x}) & 2[a_{1,1}(\mathbf{x}) + a_{2,2}(\mathbf{x}) - |a_{1,2}(\mathbf{x})|] & |a_{1,2}(\mathbf{x})| - a_{1,1}(\mathbf{x}) \\ -a_{1,2}^+(\mathbf{x}) & |a_{1,2}(\mathbf{x})| - a_{2,2}(\mathbf{x}) & a_{1,2}^-(\mathbf{x}) \end{bmatrix}_\star u(\mathbf{x}) \\ &\qquad + \frac{1}{2h}\begin{bmatrix} 0 & b_2(\mathbf{x}) & 0 \\ -b_1(\mathbf{x}) & 0 & b_1(\mathbf{x}) \\ 0 & -b_2(\mathbf{x}) & 0 \end{bmatrix}_\star u(\mathbf{x}) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & c(\mathbf{x}) & 0 \\ 0 & 0 & 0 \end{bmatrix}_\star u(\mathbf{x}).\end{aligned}$$

Diese Diskretisierung ist ebenfalls konsistent von zweiter Ordnung, das heißt, es ist $|(\mathcal{L}u)(\mathbf{x}) - (\mathcal{L}_h u)(\mathbf{x})| = \mathcal{O}(h^2)$ falls $u \in C^4(\overline{\Omega})$. Unter der Bedingung

$$|a_{1,2}(\mathbf{x})| \leq \min\{a_{1,1}(\mathbf{x}), a_{2,2}(\mathbf{x})\},$$

lässt sich nun für den Hauptteil von $\mathcal{L}$ das Sternlemma 5.6 anwenden.
Ist Ω ein Würfelgebiet, so führt das Randwertproblem

$$\begin{aligned}(\mathcal{L}u)(\mathbf{x}) &= f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega,\\ u(\mathbf{x}) &= g(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Gamma,\end{aligned}$$

unter Verwendung der hier vorgestellten Diskretiserung auf das Differenzenverfahren

$$\begin{aligned}(\mathcal{L}_h u_h)(\mathbf{x}) &= f(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \Omega_h,\\ u_h(\mathbf{x}) &= g(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \Gamma_h.\end{aligned} \tag{5.2}$$

Dies ist gleichbedeutend mit einem linearen Gleichungssystem $\mathbf{A}_h\mathbf{u}_h = \mathbf{f}_h$ für die unbekannten Gitterwerte $\mathbf{u}_h$. Die Systemmatrix $\mathbf{A}_h$ ist symmetrisch, falls $\mathbf{b} = \mathbf{0}$ gilt.

5.3 Diskretes Maximumprinzip

Alle verwendeten Differenzensterne entsprechen einer gewichteten Mittelung von Nachbarwerten. Daher kann offensichtlich kein Wert größer sein als das Maximum über alle Nachbarwerte. Dies ist der Spezialfall der Theorie der Differenzensterne, deren Koeffizienten ein bestimmtes Vorzeichenverhalten aufweisen.

Lemma 5.6 (Sternlemma) Sei $k > 1$. Für die Zahlen α_ℓ und p_ℓ, $0 \le \ell \le k$, gelte $\alpha_\ell < 0$ für alle $\ell = 1, 2, \ldots, k$ und

$$\sum_{\ell=0}^{k} \alpha_\ell \ge 0, \qquad \sum_{\ell=0}^{k} \alpha_\ell p_\ell \le 0.$$

Ferner sei $p_0 \ge 0$ oder $\sum_{\ell=0}^{k} \alpha_\ell = 0$. Dann folgt aus $p_0 \ge \max_{1\le\ell\le k}\{p_\ell\}$ die Gleichheit

$$p_0 = p_1 = \cdots = p_k.$$

Beweis. Aus den Voraussetzungen folgt

$$\sum_{\ell=1}^{k} \alpha_\ell(p_\ell - p_0) = \sum_{\ell=0}^{k} \alpha_\ell(p_\ell - p_0) = \sum_{\ell=0}^{k} \alpha_\ell p_\ell - p_0 \sum_{\ell=0}^{k} \alpha_\ell \le 0.$$

In der links stehenden Summe sind alle Summanden wegen $\alpha_\ell < 0$ und $p_\ell - p_0 \le 0$ nicht negativ. Also hat jeder Summand den Wert 0. Aus $\alpha_\ell \ne 0$ folgt die Behauptung. □

Definition 5.7 Das Gebiet Ω_h heißt **(diskret) zusammenhängend**, wenn zu jedem Punktepaar $\mathbf{x}, \mathbf{y} \in \Omega_h$ auch ein Verbindungsweg existiert, der entlang der Gitterlinien und ganz in Ω_h verläuft.

Bemerkung Für genügend kleines h ist das Gebiet Ω_h diskret zusammenhängend. △

Satz 5.8 (Diskretes Maximumprinzip) Sei u_h die Lösung der diskreten Differentialgleichung

$$(\mathcal{L}_h u_h)(\mathbf{x}) = f(\mathbf{x}) \leq 0 \quad \text{für alle } \mathbf{x} \in \Omega_h,$$

die von der Diskretisierung der elliptischen Differentialgleichung $\mathcal{L}u = f \leq 0$ in Ω herrührt. Der Differenzenstern zu jedem Gitterpunkt in Ω_h genüge folgenden drei Bedingungnen:

1. Alle Koeffizienten, abgesehen vom Zentrum, sind nicht positiv.
2. Der Koeffizient in Ostrichtung sei negativ: $\alpha_O < 0$.
3. Die Summe aller Koeffizienten ist nicht negativ.

Dann ist

$$\max_{\mathbf{x}\in\overline{\Omega}_h} u_h(\mathbf{x}) \leq \max\{\max_{\mathbf{x}\in\Gamma_h} u_h(\mathbf{x}), 0\}.$$

Wenn das Maximum im Innern angenommen wird, die Koeffizienten in den Hauptrichtungen (also in zwei Dimensionen $\alpha_O, \alpha_W, \alpha_S, \alpha_N$) negativ sind und Ω_h zusammenhängend ist, dann ist u_h konstant.

Beweis. Wenn das Maximum im Punkt $\mathbf{z} \in \Omega_h$, also im Innern, angenommen wird, dann setze $p_0 := u_h(\mathbf{z}) > 0$ und identifiziere $p_1, p_2, \ldots, p_k$ mit den Werten in allen Nachbarpunkten, die im Differenzenstern auftreten. Wegen $\sum_{\ell=0}^{k} \alpha_\ell p_\ell = f(\mathbf{z}) \leq 0$ impliziert das Sternlemma $p_0 = p_1 = \cdots = p_k$, das heißt, $u_h(\mathbf{z})$ stimmt mit allen Nachbarn überein.

Nun marschieren wir zum Rand: wir wiederholen dieses Argument solange im jeweils östlichen Nachbarn, bis wir am Rand angekommen sind.

Wenn Ω_h zusammenhängt, können wir gemäß der Voraussetzung das obige Argument solange in alle Hauptrichtungen anwenden, bis alle Punkte von $\overline{\Omega}_h$ erreicht sind. □

Bemerkung Wenn man als dritte Bedingung sogar verlangt, dass die Summe aller Koeffizienten des Sterns 0 ergibt, dann folgt das strenge Maximumprinzip $\max_{\mathbf{x}\in\overline{\Omega}_h} u_h(\mathbf{x}) \leq \max_{\mathbf{x}\in\Gamma_h} u_h(\mathbf{x})$. △

Aus dem diskreten Maximumprinzip kann man genau dieselben Folgerungen schließen wie aus dem kontinuierlichen Maximumprinzip. Insbesondere sei auf das Vergleichsprinzip und die stetige Abhängigkeit von den Daten f und g hingewiesen. Eine weitere wollen wir explizit benennen:

Proposition 5.9 Wenn das diskrete Maximumprinzip gilt, ist das Gleichungssystem $\mathbf{A}_h\mathbf{u}_h = \mathbf{f}_h$ eindeutig lösbar.

5.4 Konvergenz

Auf Ω_h und $\overline{\Omega}_h$ definieren wir die Maximumnorm durch

$$\|v_h\|_{\Omega_h} := \max_{\mathbf{x}\in\Omega_h} |v_h(\mathbf{x})|, \qquad \|v_h\|_{\overline{\Omega}_h} := \max_{\mathbf{x}\in\overline{\Omega}_h} |v_h(\mathbf{x})|.$$

Definition 5.10 Das Differenzenverfahren (5.2) heißt

- **konvergent** mit der Ordnung p, wenn

$$\|u - u_h\|_{\overline{\Omega}_h} = \mathcal{O}(h^p),$$

- **konsistent** mit der Ordnung p, wenn

$$\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^p),$$

- **stabil** (bzgl. der rechten Seite), wenn eine Konstante $C_s > 0$ existiert, so dass für alle Gitterfunktionen v_h mit $v_h = 0$ am Rand gilt

$$\|v_h\|_{\overline{\Omega}_h} \leq C_s \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

Beispiel 5.11 Auf Würfelgebieten sind wegen $\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^2)$ die Differenzenverfahren (5.1) und (5.2) konsistent mit der Ordnung 2. Da die Shortley-Weller-Approximation nur von der Ordnung 1 ist, ist hingegen das Verfahren (5.1) für allgemeine Gebiete nur konsistent mit der Ordnung 1. △

Bemerkung Stabilität bedeutet nichts anderes, als dass $\|\mathbf{A}_h^{-1}\|_\infty \leq C_s$ unabhängig von der Maschenweite h ist. Das sieht man wie folgt: Bezeichnen $\mathbf{v}_h$ und $\mathbf{w}_h$ die Vektoren der Werte der Gitterfunktion $v_h|_{\Omega_h}$ und $\mathcal{L}_h v_h$, dann folgt $\mathbf{w}_h = \mathbf{A}_h \mathbf{v}_h$. Die Stabilitätsbedingung kann nun übersetzt werden gemäß

$$\|v_h\|_{\overline{\Omega}_h} = \|\mathbf{v}_h\|_\infty = \|\mathbf{A}_h^{-1}\mathbf{w}_h\|_\infty \leq C_s \|\mathbf{w}_h\|_\infty = C_s \|\mathbf{A}_h \mathbf{v}_h\|_\infty = C_s \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

Hieraus folgt das Behauptete, da diese Ungleichung für beliebige Vektoren $\mathbf{w}_h$ gilt. △

Satz 5.12 Ist ein Differenzenverfahren stabil und konsistent mit der Ordnung p, dann ist es auch konvergent mit der Ordnung p.

Beweis. Es gilt

$$\|u - u_h\|_{\overline{\Omega}_h} \leq C_s \|\mathcal{L}_h(u - u_h)\|_{\Omega_h} = C_s \|\mathcal{L}_h u - \mathcal{L}_h u_h\|_{\Omega_h}.$$

Wegen $(\mathcal{L}_h u_h)(\mathbf{x}) = f(\mathbf{x}) = (\mathcal{L}u)(\mathbf{x})$ für alle $\mathbf{x} \in \Omega_h$, folgt

$$\underbrace{\|u - u_h\|_{\overline{\Omega}_h}}_{\text{Diskretisierungsfehler}} \leq C_s \underbrace{\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h}}_{\text{Konsistenzfehler}} = \mathcal{O}(h^p).$$

□

Lemma 5.13 Sei Ω in der Menge $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < R\}$ enthalten. Die Gitterfunktion v_h sei Lösung der Gleichung

$$\begin{aligned} -(\Delta_h v_h)(\mathbf{x}) &= 1 \quad \text{für alle } \mathbf{x} \in \Omega_h, \\ v_h(\mathbf{x}) &= 0 \quad \text{für alle } \mathbf{x} \in \Gamma_h. \end{aligned}$$

Dann gilt

$$0 \le v_h(\mathbf{x}) \le \frac{1}{2d}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h.$$

Beweis. Man betrachte die Funktion $w(\mathbf{x}) = (R^2 - \|\mathbf{x}\|_2^2)/(2d)$. Da w ein Polynom zweiten Grades ist, verschwinden die bei der Bildung des Differenzensterns vernachlässigten Ableitungen, das heißt, es gilt $-\Delta_h w = -\Delta w = 1$ in Ω_h. Außerdem ist $w \ge 0$ auf Γ_h. Aus dem diskreten Vergleichsprinzip folgt daher $v_h(\mathbf{x}) \le w(\mathbf{x})$ für alle $\mathbf{x} \in \overline{\Omega}_h$. □

Dieses Lemma impliziert die Stabilität: Sei w_h mit $w_h = 0$ auf Γ_h beliebig, dann folgt

$$-\frac{(\Delta_h w_h)(\mathbf{x})}{\|\Delta_h w_h\|_{\Omega_h}} \le 1 = -(\Delta_h v_h)(\mathbf{x}), \quad \mathbf{x} \in \Omega_h.$$

Das diskrete Vergleichsprinzip liefert sofort

$$\frac{w_h(\mathbf{x})}{\|\Delta_h w_h\|_{\Omega_h}} \le v_h(\mathbf{x}) \le \frac{1}{2d}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h,$$

dies bedeutet

$$\|w_h\|_{\overline{\Omega}_h} \le \frac{R^2}{2d}\|\Delta_h w_h\|_{\Omega_h}.$$

Korollar 5.14 Die Lösung der Poisson-Gleichung erfülle $u \in C^4(\overline{\Omega})$. Dann konvergiert das Differenzenverfahren (5.1) und es gilt

$$\|u - u_h\|_{\overline{\Omega}_h} = \mathcal{O}(h^p)$$

mit $p = 2$ im Falle von Würfelgebieten und $p = 1$ im Falle von allgemeinen Gebieten.

Bemerkung Ist der Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} a_{i,j}(\mathbf{x})\frac{\partial^2}{\partial x_i \partial x_j}u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x}), \quad c(\mathbf{x}) \ge 0$$

gleichmäßig elliptisch mit Elliptizitätskonstante $\alpha > 0$, dann folgt für $w(\mathbf{x}) = R^2 - \|\mathbf{x}\|_2^2$, dass $(\mathcal{L}_h w)(\mathbf{x}) = (\mathcal{L}w)(\mathbf{x}) \ge 2\alpha$ für alle $\mathbf{x} \in \Omega_h$ (vergleiche fünfte Folgerung des kontinuierlichen Maximumprinzips). Erfüllt der zugehörige Differenzenstern das diskrete Maximumprinzip (und damit das Vergleichsprinzip), so gilt daher Lemma 5.13 mit

$$0 \le v_h(\mathbf{x}) \le \frac{1}{2\alpha}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h.$$

Enthält hingegen $\mathcal{L}$ zusätzlich Terme erster Ordnung und ist $c > 0$, dann kann mittels einem Störargument für genügend kleine Maschenweite h Stabilität nachweisen werden. Auf Würfelgebieten erhalten wir auf diese Weise ebenfalls eine quadratische Konvergenzordnung für allgemeine (elliptische) Differentialoperatoren. △

5.5 Iterative Lösung

Die Diskretisierung von Differentialgleichungen führt auf sehr große, dünnbesetzte Gleichungssysteme $\mathbf{Ax} = \mathbf{b}$. Die effizientesten Löser für dünnbesetzte Gleichungssysteme sind Krylov-Raum-Verfahren.

Besitzt der Differenzenstern die Eigenschaft

$$\sum_{\ell=0}^{k} \alpha_\ell \geq 0 \quad \text{und} \quad a_\ell \leq 0 \quad \text{für alle } 1 \leq \ell \leq k,$$

dann folgt aus dem Satz von Gerschgorin, dass jeder Eigenwert einen positiven Realteil besitzt, denn

$$0 \neq \lambda \in K = \left\{ z \in \mathbb{C} : |z - \alpha_0| \leq \underbrace{\sum_{\ell=1}^{k} |\alpha_\ell|}_{\leq \alpha_0} \right\}.$$

Ist die Systemmatrix $\mathbf{A}$ zusätzlich symmetrisch, dann kann das CG-Verfahren zur iterativen Lösung benutzt werden. Dies tritt gemäß Abschnitt 5.2 beispielsweise auf Würfelgebieten ein bei der Diskretisierung des Differentialoperators

$$(\mathcal{L}u)(\mathbf{x}) = c(\mathbf{x})u(\mathbf{x}) - \sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}), \qquad c(\mathbf{x}) \geq 0. \tag{5.3}$$

5.5.1 Verfahren der konjugierten Gradienten

Das Verfahren der konjugierten Gradienten von Hestenes und Stiefel (1952), welches auch als CG-Verfahren (vom englischen *conjugate gradient method*) bekannt ist, ist wohl das effektivste Verfahren zur Lösung großer linearer Gleichungssysteme $\mathbf{Ax} = \mathbf{b}$ mit symmetrischer und positiv definiter Matrix $\mathbf{A}$.

Definition 5.15 Ist $\mathbf{A} \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, dann definiert

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

ein Skalarprodukt. Die induzierte Norm

$$\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}}$$

wird **Energienorm** bezüglich $\mathbf{A}$ genannt. Zwei Vektoren $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ heißen **konjugiert** bezüglich $\mathbf{A}$, falls

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y} = 0.$$

Lemma 5.16 Seien $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ bezüglich $\mathbf{A}$ konjugierte Richtungen. Für jedes $\mathbf{x}_0 \in \mathbb{R}^n$ liefert die durch

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

mit

$$\alpha_i = \frac{\mathbf{d}_i^T \mathbf{r}_i}{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i}, \qquad \mathbf{r}_i = \mathbf{b} - \mathbf{A}\mathbf{x}_i$$

für $i \geq 0$ erzeugte Folge nach (höchstens) n Schritten die Lösung $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Beweis. Mit dem Ansatz

$$\mathbf{x} - \mathbf{x}_0 = \sum_{j=0}^{n-1} \alpha_j \mathbf{d}_j$$

erhalten wir wegen den Orthogonalitätsrelationen

$$\mathbf{d}_i^T \mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{d}_i^T \mathbf{A}\left(\sum_{j=0}^{n-1} \alpha_j \mathbf{d}_j \right) = \sum_{j=0}^{n-1} \alpha_j \underbrace{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_j}_{=0 \text{ falls } i \neq j} = \alpha_i \mathbf{d}_i^T \mathbf{A} \mathbf{d}_i$$

die Beziehung

$$\alpha_i = \frac{\mathbf{d}_i^T \mathbf{A}(\mathbf{x} - \mathbf{x}_0)}{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i}.$$

Weil $\mathbf{d}_i$ zu den anderen Richtungen konjugiert ist, gilt

$$\mathbf{d}_i^T \mathbf{A}(\mathbf{x}_i - \mathbf{x}_0) = \mathbf{d}_i^T \mathbf{A}\left(\sum_{j=0}^{i-1} \alpha_j \mathbf{d}_j \right) = \sum_{j=0}^{i-1} \alpha_j \underbrace{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_j}_{=0} = 0.$$

Deshalb ist

$$\alpha_i = \frac{\mathbf{d}_i^T (\overbrace{\mathbf{A}\mathbf{x}}^{=\mathbf{b}} - \mathbf{A}\mathbf{x}_i + \mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_0)}{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i} = \frac{\mathbf{d}_i^T (\mathbf{b} - \mathbf{A}\mathbf{x}_i)}{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i} + \underbrace{\frac{\mathbf{d}_i^T \mathbf{A}(\mathbf{x}_i - \mathbf{x}_0)}{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i}}_{=0} = \frac{\mathbf{d}_i^T \mathbf{r}_i}{\mathbf{d}_i^T \mathbf{A} \mathbf{d}_i}.$$

□

Bemerkung Der Vektor $\mathbf{r}_i = \mathbf{b} - \mathbf{A}\mathbf{x}_i$ wird *Residuum* genannt. Seine Norm $\|\mathbf{r}_i\|_2$ ist ein Maß für den Fehler im i-ten Schritt. Gilt $\|\mathbf{r}_i\|_2 = 0$, so stimmt $\mathbf{x}_i$ mit der Lösung $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ überein. △

Beim Verfahren der konjugierten Gradienten werden die Richtungen $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ nicht von vornherein gewählt, sondern aus dem jeweils aktuellen Residuum $\mathbf{r}_i$ durch Addition einer Korrektur ermittelt. Das Verfahren der konjugierten Gradienten lautet:

Algorithmus 5.17 (CG-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, rechte Seite $\mathbf{b} \in \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

① Initialisierung: setze $\mathbf{d}_0 = \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$ und $i := 0$

② berechne

$$\alpha_i := \frac{\mathbf{d}_i^T \mathbf{r}_i}{\mathbf{d}_i^T \mathbf{A}\mathbf{d}_i} \tag{5.4}$$

$$\mathbf{x}_{i+1} := \mathbf{x}_i + \alpha_i \mathbf{d}_i \tag{5.5}$$

$$\mathbf{r}_{i+1} := \mathbf{r}_i - \alpha_i \mathbf{A}\mathbf{d}_i \tag{5.6}$$

$$\beta_i := \frac{\mathbf{d}_i^T \mathbf{A}\mathbf{r}_{i+1}}{\mathbf{d}_i^T \mathbf{A}\mathbf{d}_i} \tag{5.7}$$

$$\mathbf{d}_{i+1} := \mathbf{r}_{i+1} - \beta_i \mathbf{d}_i \tag{5.8}$$

③ falls $\|\mathbf{r}_{i+1}\|_2 \neq 0$ erhöhe $i := i + 1$ und gehe nach ②

Dass die durch das Verfahren produzierten Richtungen $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{i+1}$ tatsächlich bezüglich $\mathbf{A}$ konjugiert sind, folgt mit der Hilfe des folgenden Satzes.

Satz 5.18 Solange $\mathbf{r}_i \neq \mathbf{0}$ ist, gelten die folgenden Aussagen:

1. Es ist $\mathbf{d}_j^T \mathbf{r}_i = 0$ für alle $j < i$.
2. Es gilt $\mathbf{r}_j^T \mathbf{r}_i = 0$ für alle $j < i$.
3. Die Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_i$ sind paarweise konjugiert.

Beweis. Wir bemerken zunächst, dass gilt

$$\mathbf{d}_i^T \mathbf{r}_{i+1} \overset{(5.6)}{=} \mathbf{d}_i^T (\mathbf{r}_i - \alpha_i \mathbf{A}\mathbf{d}_i) = \mathbf{d}_i^T \mathbf{r}_i - \underbrace{\alpha_i}_{\overset{(5.4)}{=} \frac{\mathbf{d}_i^T \mathbf{r}_i}{\mathbf{d}_i^T \mathbf{A}\mathbf{d}_i}} \mathbf{d}_i^T \mathbf{A}\mathbf{d}_i = 0. \tag{5.9}$$

Wir führen nun den Beweis vermittels Induktion.

Im Falle $i = 1$ folgen die ersten beiden Aussagen direkt aus (5.9), während sich die dritte wegen

$$\mathbf{d}_0^T \mathbf{A}\mathbf{d}_1 \overset{(5.8)}{=} \mathbf{d}_0^T \mathbf{A}\mathbf{r}_1 - \underbrace{\beta_i}_{\overset{(5.7)}{=} \frac{\mathbf{d}_0^T \mathbf{A}\mathbf{r}_1}{\mathbf{d}_0^T \mathbf{A}\mathbf{d}_0}} \mathbf{d}_0^T \mathbf{A}\mathbf{d}_0 = 0$$

ergibt.

Für den Induktionsschritt $i \mapsto i + 1$ nehmen wir an, dass alle drei Aussagen für i gelten. Nun ergibt sich die erste Aussage für $i + 1$ und $j = i$ wieder aus (5.9) während sie für $j < i$ mit Hilfe der Induktionsannahme aus

$$\mathbf{d}_j^T \mathbf{r}_{i+1} \overset{(5.6)}{=} \mathbf{d}_j^T (\mathbf{r}_i - \alpha_i \mathbf{A}\mathbf{d}_i) = \underbrace{\mathbf{d}_j^T \mathbf{r}_i}_{=0} - \alpha_i \underbrace{\mathbf{d}_j^T \mathbf{A}\mathbf{d}_i}_{=0} = 0$$

folgt. Wegen

$$\mathbf{r}_j \overset{(5.8)}{=} \mathbf{d}_j + \beta_{j-1} \mathbf{d}_{j-1}, \qquad 1 \leq j \leq i$$

ergibt sich die zweite Aussage direkt aus der ersten.
Weiter sind $\mathbf{d}_i$ und $\mathbf{d}_{i+1}$ konjugiert, da

$$\mathbf{d}_i^T \mathbf{A}\mathbf{d}_{i+1} \overset{(5.8)}{=} \mathbf{d}_i^T \mathbf{A}\mathbf{r}_{i+1} - \underbrace{\beta_i}_{\overset{(5.7)}{=} \frac{\mathbf{d}_i^T \mathbf{A}\mathbf{r}_{i+1}}{\mathbf{d}_i^T \mathbf{A}\mathbf{d}_i}} \mathbf{d}_i^T \mathbf{A}\mathbf{d}_i = 0.$$

Für $\mathbf{d}_j$ mit $j < i$ folgt aufgrund der Induktionsannahme

$$\mathbf{d}_j^T \mathbf{A}\mathbf{d}_{i+1} \overset{(5.8)}{=} \mathbf{d}_j^T \mathbf{A}(\mathbf{r}_{i+1} - \beta_i \mathbf{d}_i) = \mathbf{d}_j^T \mathbf{A}\mathbf{r}_{i+1} - \beta_i \underbrace{\mathbf{d}_j^T \mathbf{A}\mathbf{d}_i}_{=0} = \mathbf{d}_j^T \mathbf{A}\mathbf{r}_{i+1},$$

und damit wegen der schon bewiesenen zweiten Aussage

$$\alpha_j \mathbf{d}_j^T \mathbf{A}\mathbf{d}_{i+1} = \alpha_j \mathbf{d}_j^T \mathbf{A}\mathbf{r}_{i+1} \overset{(5.6)}{=} (\mathbf{r}_j - \mathbf{r}_{j+1})^T \mathbf{r}_{i+1} = \mathbf{r}_j^T \mathbf{r}_{i+1} - \mathbf{r}_{j+1}^T \mathbf{r}_{i+1} = 0.$$

Es bleibt nur noch zu zeigen, dass α_j nicht Null werden kann. Angenommen $\alpha_j = 0$, dann folgt

$$\mathbf{d}_j^T \mathbf{r}_j = 0,$$

und wegen (5.8) ergibt sich

$$0 = (\mathbf{r}_j - \beta_{j-1}\mathbf{d}_{j-1})^T \mathbf{r}_j = \mathbf{r}_j^T \mathbf{r}_j - \beta_{j-1} \underbrace{\mathbf{d}_{j-1}^T \mathbf{r}_j}_{=0} = \|\mathbf{r}_j\|_2^2.$$

Dies steht jedoch im Widerspruch zur Voraussetzung $\mathbf{r}_j \neq 0$. □

Bemerkung

1. Äquivalent zu (5.4) und (5.7), aber effizienter und numerisch stabiler, hat sich die Wahl
$$\alpha_i = \frac{\mathbf{r}_i^T \mathbf{r}_i}{\mathbf{d}_i^T \mathbf{A}\mathbf{d}_i}, \qquad \beta_i = -\frac{\mathbf{r}_{i+1}^T \mathbf{r}_{i+1}}{\mathbf{r}_i^T \mathbf{r}_i}$$
erwiesen.
2. Das Verfahren der konjugierten Gradienten wird generell als Iterationsverfahren verwendet, das heißt, man bricht die Iteration ab, falls $\|\mathbf{r}_i\|_2$ klein ist. Pro Iterationsschritt wird nur eine Matrix-Vektor-Multiplikation benötigt. Die Konvergenz des Verfahrens hängt dabei stark von der Kondition der Matrix ab. Genauer, es gilt die Fehlerabschätzung
$$\|\mathbf{x} - \mathbf{x}_i\|_{\mathbf{A}} \leq 2\left(\frac{\sqrt{\operatorname{cond}_2 \mathbf{A}} - 1}{\sqrt{\operatorname{cond}_2 \mathbf{A}} + 1}\right)^i \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}}, \tag{5.10}$$
siehe zum Beispiel J. Stoer und R. Bulirsch *Numerische Mathematik II*.
3. Es bezeichne $\mathcal{K}_i(\mathbf{A}, \mathbf{r}_0)$ den sogenannten *Krylov-Raum*
$$\mathcal{K}_i(\mathbf{A}, \mathbf{r}_0) := \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \mathbf{A}^2\mathbf{r}_0, \ldots \mathbf{A}^{i-1}\mathbf{r}_0\} = \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{i-1}\}.$$
Man kann zeigen, dass die Iterierte $\mathbf{x}_i$ des i-ten Schritts die Funktion
$$\Phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} - \mathbf{b}^T \mathbf{x}$$
unter allen $\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_i(\mathbf{A}, \mathbf{r}_0)$ minimiert.

△

5.5.2 GMRES-Verfahren

Das *GMRES-Verfahren* (kurz für *Generalized Minimal Residual-Verfahren*) ist ein sehr schnell konvergierendes Verfahren für nichtsymmetrische Gleichungssysteme, das auf dem Arnoldi-Verfahren beruht. Die Systemmatrix ist unsymmetrisch bei Differentialoperatoren mit Termen erster Ordnung, aber auch bei der Shortley-Weller-Approximation.

Beim GMRES-Verfahrens bestimmt man eine Orthonormalbasis $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ des Krylov-Raums $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{z})$. Das Gram-Schmidtsche Orthogonalisierungsverfahren führt auf

$$
\begin{aligned}
\widetilde{\mathbf{v}}_1 &:= \mathbf{z}, & \mathbf{v}_1 &:= \frac{\widetilde{\mathbf{v}}_1}{\|\widetilde{\mathbf{v}}_1\|_2},\\
\widetilde{\mathbf{v}}_2 &:= \mathbf{A}\mathbf{v}_1 - \underbrace{\mathbf{v}_1^T\mathbf{A}\mathbf{v}_1}_{=:h_{1,1}}\mathbf{v}_1, & \mathbf{v}_2 &:= \frac{\widetilde{\mathbf{v}}_2}{\underbrace{\|\widetilde{\mathbf{v}}_2\|_2}_{=:h_{2,1}}},\\
\widetilde{\mathbf{v}}_3 &:= \mathbf{A}\mathbf{v}_2 - \underbrace{\mathbf{v}_1^T\mathbf{A}\mathbf{v}_2}_{=:h_{1,2}}\mathbf{v}_1 - \underbrace{\mathbf{v}_2^T\mathbf{A}\mathbf{v}_2}_{=:h_{2,2}}\mathbf{v}_2, & \mathbf{v}_3 &:= \frac{\widetilde{\mathbf{v}}_3}{\underbrace{\|\widetilde{\mathbf{v}}_3\|_2}_{=:h_{3,2}}},\\
\widetilde{\mathbf{v}}_4 &:= \mathbf{A}\mathbf{v}_3 - \underbrace{\mathbf{v}_1^T\mathbf{A}\mathbf{v}_3}_{=:h_{1,3}}\mathbf{v}_1 - \underbrace{\mathbf{v}_2^T\mathbf{A}\mathbf{v}_3}_{=:h_{2,3}}\mathbf{w}_2 - \underbrace{\mathbf{v}_3^T\mathbf{A}\mathbf{v}_3}_{=:h_{3,3}}\mathbf{v}_3, & \mathbf{v}_4 &:= \frac{\widetilde{\mathbf{w}}_4}{\underbrace{\|\widetilde{\mathbf{v}}_4\|_2}_{=:h_{4,3}}},\\
&\vdots & &\vdots
\end{aligned}
$$

Demnach erhalten wir nach k Schritten die Faktorisierung

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\mathbf{H}_k,$$

wobei die Spalten von $\mathbf{V}_{k+1} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k, \mathbf{v}_{k+1}]$ gerade die Orthonormalbasis des Krylov-Raums $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{z})$ ist und

$$\mathbf{H}_k = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \cdots & h_{1,k} \\ h_{2,1} & h_{2,2} & h_{2,3} & \cdots & h_{2,k} \\ 0 & h_{3,2} & h_{3,3} & \cdots & h_{3,k} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{k,k-1} & h_{k,k} \\ 0 & \cdots & 0 & 0 & h_{k+1,k} \end{bmatrix} \in \mathbb{R}^{(k+1)\times k}$$

eine obere Hessenberg-Matrix ist.

Sei $\mathbf{x}_0$ eine Startnäherung und $\mathbf{z} := \mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$. Wegen $\mathbf{v}_1 = \mathbf{r}_0/\|\mathbf{r}_0\|_2$ folgt dann

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathbf{x}_0+\mathcal{K}_k(\mathbf{A},\mathbf{r}_0)} \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2 &= \min_{\mathbf{y}\in\mathbb{R}^k} \|\mathbf{b}-\mathbf{A}(\mathbf{x}_0+\mathbf{V}_k\mathbf{y})\|_2\\
&= \min_{\mathbf{y}\in\mathbb{R}^k} \|\mathbf{r}_0 - \mathbf{V}_{k+1}\mathbf{H}_k\mathbf{y}\|_2\\
&= \min_{\mathbf{y}\in\mathbb{R}^k} \big\|\mathbf{e}_1\|\mathbf{r}_0\|_2 - \mathbf{H}_k\mathbf{y}\big\|_2.
\end{aligned}
$$

Dies bedeutet, die Lösung $\mathbf{y}$ eines $\big((k+1)\times k\big)$-Ausgleichsproblems liefert die neue Iterierte gemäß $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k\mathbf{y}$. Dieses Ausgleichsproblem lässt sich schnell durch die k Givens-Rotationen $\mathbf{G}(2,1,\theta_1), \mathbf{G}(3,2,\theta_2), \ldots, \mathbf{G}(k+1,k,\theta_k)$ lösen. Die Norm des Residuums $\mathbf{r}_k =$

$\mathbf{b} - \mathbf{A}\mathbf{x}_k$ lässt sich dann offensichtlich aus

$$\mathbf{G}(k+1,k,\theta_k)\cdots\mathbf{G}(3,2,\theta_2)\mathbf{G}(2,1,\theta_1)\big(\mathbf{e}_1\|\mathbf{r}_0\|_2 - \mathbf{H}_k\mathbf{y}\big) = \left[\frac{\mathbf{p}_k}{\rho_{k+1}}\right] - \left[\frac{\mathbf{R}_k}{0\cdots 0}\right]\mathbf{y}$$

ablesen und lautet $|\rho_{k+1}|$. Der Übergang $(\mathbf{R}_{k-1}, \mathbf{p}_{k-1}, \rho_{k-1}) \mapsto (\mathbf{R}_k, \mathbf{p}_k, \rho_k)$ ist ein einfaches Update unter Ausnutzen von der Identitäten

$$\big[\mathbf{R}_k\big]_{i,j=1}^{k-1} = \mathbf{R}_{k-1} \quad \text{und} \quad \big[\mathbf{p}_k\big]_{i=1}^{k-1} = \mathbf{p}_{k-1},$$

das nur das Erzeugen der neuen Givens-Rotation benötigt.

Algorithmus 5.19 (GMRES-Verfahren)

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$

output: Näherungslösung $\mathbf{x}_k$ mit $\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 \le \varepsilon$

① Initialisierung: setze $\mathbf{w}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\rho_0 = h_{1,0} := \|\mathbf{w}_0\|_2$ und $k := 0$

② solange $(|\rho_{k+1}| > \varepsilon)$ berechne

$$\begin{gathered} k := k+1 \\ \mathbf{v}_k := \mathbf{w}_{k-1}/h_{k,k-1} \\ \mathbf{w}_k := \mathbf{A}\mathbf{v}_k \\ h_{1,k} := \mathbf{v}_1^T\mathbf{w}_k, \ldots, h_{k,k} := \mathbf{v}_k^T\mathbf{w}_k \\ \mathbf{w}_k := \mathbf{w}_k - \sum_{i=1}^{k} h_{i,k}\mathbf{v}_i \\ h_{k+1,k} := \|\mathbf{w}_k\|_2 \\ (\mathbf{R}_{k-1}, \mathbf{p}_{k-1}, \rho_{k-1}) \mapsto (\mathbf{R}_k, \mathbf{p}_k, \rho_k) \end{gathered}$$

③ löse

$$\mathbf{R}_k\mathbf{y} = \mathbf{p}_k$$

durch Rückwärtssubstitution und setze $\mathbf{x}_k := \mathbf{x}_0 + \mathbf{V}_k\mathbf{y}$

5.5.3 MINRES-Verfahren*

Ist $c(\mathbf{x})$ in (5.3) beliebig, so kann die Matrix indefinit symmetrisch sein. Wendet man dann das GMRES-Verfahren auf das Gleichungssystem an, dann ist die obere Hessenberg-Matrix $[\mathbf{H}_k]_{i,j=1}^k$ aus Symmetriegründen eine symmetrische Tridiagonalmarix und der Arnoldi-Prozess wird zum Lanzcos-Prozess. Dies führt auf das *MINRES-Verfahren* (kurz für *Minimal Residual-Verfahren*). Wir stellen hier jedoch eine Variante vor, bei der direkt die Iterierte und das Residuum aufdatiert wird.

Lemma 5.20 Seien $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ bezüglich $\mathbf{A}^2$ konjugierte Richtungen. Für jedes $\mathbf{x}_0 \in \mathbb{R}^n$ liefert die durch

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k, \qquad k = 0, 1, \ldots, n-1,$$

mit

$$\alpha_k = \frac{\mathbf{d}_k^T \mathbf{A}\mathbf{r}_k}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k}, \qquad \mathbf{r}_k = \mathbf{b}_k - \mathbf{A}\mathbf{x}_k$$

erzeugte Folge nach (höchstens) n Schritten die Lösung $\mathbf{x}_n = \mathbf{A}^{-1}\mathbf{b}$.

Beweis. Mit dem Ansatz

$$\mathbf{x} - \mathbf{x}_0 = \sum_{\ell=0}^{n-1} \alpha_\ell \mathbf{d}_\ell$$

erhalten wir wegen den Orthogonalitätsrelationen

$$\mathbf{d}_k^T \mathbf{A}^2(\mathbf{x} - \mathbf{x}_0) = \mathbf{d}_k^T \mathbf{A}^2 \left(\sum_{\ell=0}^{n-1} \alpha_\ell \mathbf{d}_\ell \right) = \sum_{\ell=0}^{n-1} \alpha_\ell \underbrace{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_\ell}_{=0 \text{ falls } k \neq \ell} = \alpha_k \mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k$$

die Beziehung

$$\alpha_k = \frac{\mathbf{d}_k^T \mathbf{A}^2(\mathbf{x} - \mathbf{x}_0)}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k}.$$

Weil $\mathbf{d}_k$ zu den anderen Richtungen konjugiert ist, gilt

$$\mathbf{d}_k \mathbf{A}^2(\mathbf{x}_k - \mathbf{x}_0) = \mathbf{d}_k^T \mathbf{A}^2 \left(\sum_{\ell=0}^{k-1} \alpha_\ell \mathbf{d}_\ell \right) = \sum_{\ell=0}^{k-1} \alpha_\ell \underbrace{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_\ell}_{=0} = 0.$$

Deshalb ist

$$\alpha_k = \frac{\mathbf{d}_k^T \mathbf{A}(\overbrace{\mathbf{A}\mathbf{x}}^{=\mathbf{b}} - \mathbf{A}\mathbf{x}_k + \mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}_0)}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k} = \frac{\mathbf{d}_k^T \mathbf{A}(\mathbf{b} - \mathbf{A}\mathbf{x}_k)}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k} + \underbrace{\frac{\mathbf{d}_k^T \mathbf{A}^2(\mathbf{x}_k - \mathbf{x}_0)}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k}}_{=0}$$

$$= \frac{\mathbf{d}_k^T \mathbf{A}\mathbf{r}_k}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k}.$$

□

Beim MINRES-Verfahren werden die Richtungen $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ nicht von vornherein gewählt, sondern sukzessive vermittels einer Dreitermrekursion ermittelt. Das vollständige Verfahren lautet:

Algorithmus 5.21 (MINRES-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, rechte Seite $\mathbf{b} \in \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$
① Initialisierung: setze $\mathbf{d}_{-1} := \mathbf{0}$, $\mathbf{d}_0 = \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$ und $k := 0$

② berechne

$$\alpha_k := \frac{\mathbf{d}_k^T \mathbf{A}\mathbf{r}_k}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k}$$
$$\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$$
$$\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{d}_k$$
$$\beta_k := \frac{\mathbf{d}_k^T \mathbf{A}^3 \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k} \tag{5.11}$$
$$\gamma_k := \frac{\mathbf{d}_k^T \mathbf{A}^3 \mathbf{d}_{k-1}}{\mathbf{d}_{k-1}^T \mathbf{A}^2 \mathbf{d}_{k-1}} \tag{5.12}$$
$$\mathbf{d}_{k+1} := \mathbf{A}\mathbf{d}_k - \beta_k \mathbf{d}_k - \gamma_k \mathbf{d}_{k-1} \tag{5.13}$$

③ falls $\|\mathbf{r}_{k+1}\|_2 > \varepsilon$ erhöhe $k := k+1$ und gehe nach ②

Bemerkung Mithilfe einer zusätzlichen Variablen $\mathbf{s}_k = \mathbf{A}\mathbf{d}_k$ lässt sich das Verfahren so umformulieren, dass nur eine Matrix-Vektor-Multiplikation pro Iteration benötigt wird. △

Satz 5.22 Solange $\mathbf{r}_k \neq \mathbf{0}$ ist, produziert das MINRES-Verfahren von Null verschiedene, paarweise $\mathbf{A}^2$ konjugierte Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$ und es gilt $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k\}$.

Beweis. Zunächst bemerken wir, dass aus $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0) = \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ die Existenz von Zahlen γ_ℓ folgt mit

$$\mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{r}_0 = \mathbf{A} \sum_{\ell=0}^{k-1} \gamma_\ell \mathbf{A}^\ell \mathbf{r}_0.$$

Dies bedeutet, dass dann $\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ und folglich $\mathbf{r}_k = \mathbf{0}$ gelten würde. Daher besitzt $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)$ stets die Dimension $k+1$.

Wir wollen nun den Beweis vermittels Induktion führen. Wegen

$$\mathbf{d}_1^T \mathbf{A}^2 \mathbf{d}_0 \overset{(5.13)}{=} \mathbf{d}_0^T \mathbf{A}^3 \mathbf{d}_0 - \beta_0 \mathbf{d}_0^T \mathbf{A}^2 \mathbf{d}_0 \overset{(5.11)}{=} 0$$

ist die Behauptung für $k = 1$ klar.

Für den Induktionsschritt $k \mapsto k+1$ nehmen wir an, dass die Behauptung für $k \geq 1$ gilt. Dann folgt aus (5.13)

$$\mathbf{0} \neq \mathbf{d}_{k+1} = \underbrace{\mathbf{A}\mathbf{d}_k}_{\notin \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)} - \underbrace{\beta_k \mathbf{d}_k - \gamma_k \mathbf{d}_{k-1}}_{\in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)} \in \mathcal{K}_{k+2}(\mathbf{A}, \mathbf{r}_0),$$

dies bedeutet

$$\operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k+1}\} = \mathcal{K}_{k+2}(\mathbf{A}, \mathbf{r}_0).$$

Für alle $\ell \leq k$ folgt außerdem sofort

$$\mathbf{d}_{k+1}^T \mathbf{A}^2 \mathbf{d}_\ell \overset{(5.13)}{=} \mathbf{d}_k^T \mathbf{A}^3 \mathbf{d}_\ell - \underbrace{\beta_k}_{\overset{(5.11)}{=} \frac{\mathbf{d}_k^T \mathbf{A}^3 \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_k}} \mathbf{d}_k^T \mathbf{A}^2 \mathbf{d}_\ell - \underbrace{\gamma_k}_{\overset{(5.12)}{=} \frac{\mathbf{d}_k^T \mathbf{A}^3 \mathbf{d}_{k-1}}{\mathbf{d}_{k-1}^T \mathbf{A}^2 \mathbf{d}_{k-1}}} \mathbf{d}_{k-1}^T \mathbf{A}^2 \mathbf{d}_\ell,$$

was für $\ell = k-1, k$ offensichtlich gleich Null ist. Für $\ell < k-1$ sind alle Terme Null aufgrund der Induktionsannahme, wobei man dies im Fall $\mathbf{d}_k^T \mathbf{A}^3 \mathbf{d}_\ell = 0$ wie folgt einsieht:

$$\mathbf{d}_k^T \perp_{\mathbf{A}^2} \mathcal{K}_{k-1}(\mathbf{A}, \mathbf{r}_0) \supset \mathbf{A}\mathcal{K}_{k-2}(\mathbf{A}, \mathbf{r}_0) \ni \mathbf{A}\mathbf{d}_\ell.$$

□

Lemma 5.23 Die Iterierten $\mathbf{x}_k$ aus dem MINRES-Verfahren erfüllen

$$\min_{\mathbf{y} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} \|\mathbf{b} - \mathbf{A}\mathbf{y}\|_2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2.$$

Beweis. Für ein beliebiges $\mathbf{y} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ folgt

$$\begin{aligned}
\|\mathbf{b} - \mathbf{A}\mathbf{y}\|_2^2 &= \|\mathbf{x} - \mathbf{y}\|_{\mathbf{A}^2}^2 \\
&= \left\| \sum_{\ell=0}^{n-1} \alpha_\ell \mathbf{d}_\ell - (\mathbf{y} - \mathbf{x}_0) \right\|_{\mathbf{A}^2}^2 \\
&\geq \left\| \sum_{\ell=k}^{n-1} \alpha_\ell \mathbf{d}_\ell \right\|_{\mathbf{A}^2}^2 \\
&= \sum_{\ell=k}^{n-1} |\alpha_\ell|^2 \|\mathbf{d}_\ell\|_{\mathbf{A}^2}^2,
\end{aligned}$$

wobei nach Konstruktion das Minimum für $\mathbf{x}_k$ angenommen wird. □

Bemerkung Man kann (5.11)–(5.13) auch ersetzen durch

$$\beta_k := -\frac{\mathbf{d}_k^T \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{A}\mathbf{r}_k}, \qquad \mathbf{d}_{k+1} := \mathbf{r}_{k+1} + \beta_k \mathbf{d}_k.$$

In diesem Fall kann jedoch die Suchrichtung $\mathbf{d}_{k+1} = \mathbf{0}$ werden ohne dass $\mathbf{r}_{k+1} = \mathbf{0}$ gilt. Dies bedeutet, der Algorithmus kann zusammenbrechen, bevor die exakte Lösung gefunden ist. △

5.6 Vorkonditionierung

Die Stabiltätsanalyse liefert $\|\mathbf{A}^{-1}\|_\infty \leq C_s$, während wir aus der Definition des Differenzenstern schließen, dass $\|\mathbf{A}\|_\infty \leq C_c/h^2$. Hieraus erhalten wir für die Kondition der Systemmatrix $\mathbf{A}$ die Abschätzung

$$\operatorname{cond}_\infty \mathbf{A} = \|\mathbf{A}\|_\infty \left\|\mathbf{A}^{-1}\right\|_\infty \leq \frac{C_s C_c}{h^2} \xrightarrow{h \to 0} \infty.$$

Dass diese Abschätzung scharf ist, zeigt das folgende Beispiel.

Beispiel 5.24 Wie man leicht nachrechnet, besitzt die bei der Diskretisierung der eindimensionalen Poisson-Gleichung auftretende Systemmatrix $\mathbf{A} \in \mathbb{R}^{(n-1)\times(n-1)}$ (vergleiche Beispiel 5.4) die Eigenwerte

$$\lambda_k = \frac{2}{h^2}\left[1 - \cos\left(\frac{k\pi}{n}\right)\right], \quad k = 1, 2, \ldots, n-1$$

und die Eigenvektoren $\mathbf{v}_k$ mit

$$[\mathbf{v}_k]_\ell = \sin\left(\frac{k\ell\pi}{n}\right), \quad \ell = 1, 2, \ldots, n-1.$$

Wegen

$$\lambda_{\min} = \frac{2}{h^2}\left[1 - \left(+1 - \frac{h^2\pi^2}{2} + \mathcal{O}(h^4)\right)\right] = \pi^2 + \mathcal{O}(h^2),$$
$$\lambda_{\max} = \frac{2}{h^2}\left[1 - \left(-1 + \frac{h^2\pi^2}{2} + \mathcal{O}(h^4)\right)\right] = \frac{4}{h^2} + \mathcal{O}(1),$$

folgt schließlich

$$\operatorname{cond}_2 \mathbf{A} = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{4}{h^2\pi^2} + \mathcal{O}(1).$$

△

Um eine Genauigkeit $\varepsilon > 0$ mit dem CG-Verfahren zu erzielen, benötigt man gemäß (5.10)

$$K \geq \frac{\log\left(\varepsilon/(2\|\mathbf{x} - \mathbf{x}_0\|_2)\right)}{\log q}$$

Iterationen, wobei

$$q = \frac{\sqrt{\operatorname{cond}_2 \mathbf{A}} - 1}{\sqrt{\operatorname{cond}_2 \mathbf{A}} + 1} \approx \frac{2 - h\pi}{2 + h\pi} = 1 - \frac{2h\pi}{2 + h\pi}.$$

Wegen

$$\log q \approx \log\left(1 - \frac{2h\pi}{2 + h\pi}\right) \approx -h\pi$$

folgt also, dass man $K = \mathcal{O}(1/h)$ Iterationen benötigt, falls man als Startnäherung $\mathbf{x}_0$ etwa immer den Nullvektor nimmt.

Bemerkung Bezeichnet man die Anzahl der Unbekannten mit n, dann folgt für eine Differentialgleichung auf dem d-dimensionalen Würfel der Zusammenhang $h \approx n^{-1/d}$. Folglich besitzt das Berechnen einer Lösung die Gesamtkomplexität $\mathcal{O}(n^{(d+1)/d})$. △

Man versucht daher, durch eine geeignete *Vorkonditionierung* die Konvergenz zu beschleunigen. Dazu wählt man eine symmetrische und positiv definite Matrix $\mathbf{M}$, so dass $\operatorname{cond}_2\left(\mathbf{M}^{-1}\mathbf{A}\right) = \operatorname{cond}_2\left(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}\right)$ klein wird. Das CG-Verfahren wird nun auf das zu $\mathbf{A}\mathbf{u} = \mathbf{f}$ äquivalente System

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{u} = \mathbf{M}^{-1}\mathbf{f}$$

angewendet. Da $\mathbf{M}^{-1}\mathbf{A}$ im allgmeinen nicht mehr symmetrisch ist, führt man das CG-Verfahren im modifiziertem Innenprodukt $\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{M}} := \mathbf{x}^T\mathbf{M}\mathbf{y}$ durch, denn hier gilt

$$\langle \mathbf{M}^{-1}\mathbf{A}\mathbf{x}, \mathbf{y}\rangle_{\mathbf{M}} = \mathbf{x}^T(\mathbf{M}^{-1}\mathbf{A})^T\mathbf{M}\mathbf{y} = \mathbf{x}^T\mathbf{M}(\mathbf{M}^{-1}\mathbf{A})\mathbf{y} = \langle \mathbf{x}, \mathbf{M}^{-1}\mathbf{A}\mathbf{y}\rangle_{\mathbf{M}}.$$

Man beachte, dass wegen

$$\begin{aligned}\left\|\mathbf{M}^{-1}\mathbf{A}\right\|_{\mathbf{M}}^2 &= \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{M}^{-1}\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{M}\mathbf{x}} \overset{\mathbf{y}:=\mathbf{M}^{1/2}\mathbf{x}}{=} \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}^T\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-1/2}\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \\ &= \left\|\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}\right\|_2^2,\end{aligned}$$

und analog für die Inverse $\mathbf{A}^{-1}$, in der Fehlerabschätzung des vorkonditionierten CG-Verfahrens tatsächlich die Konditionszahl $\operatorname{cond}_2\left(\mathbf{M}^{-1}\mathbf{A}\right)$ auftritt. Das fertige Verfahren nennt man *PCG-Verfahren* (kurz für *preconditioned conjugate gradients*).

Algorithmus 5.25 (PCG-Verfahren)

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, Vorkonditionierer $\mathbf{M} \in \mathbb{R}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$

output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

① Initialisierung: setze $\mathbf{s}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\mathbf{d}_0 = \mathbf{r}_0 := \mathbf{M}^{-1}\mathbf{s}_0$ und $k := 0$

② berechne

$$\begin{aligned}\alpha_k &:= \frac{\mathbf{s}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} \\ \mathbf{x}_{k+1} &:= \mathbf{x}_k + \alpha_k\mathbf{d}_k \\ \mathbf{s}_{k+1} &:= \mathbf{s}_k - \alpha_k\mathbf{A}\mathbf{d}_k \\ \mathbf{r}_{k+1} &:= \mathbf{M}^{-1}\mathbf{s}_{k+1} \\ \beta_k &:= \frac{\mathbf{s}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{s}_k^T\mathbf{r}_k} \\ \mathbf{d}_{k+1} &:= \mathbf{r}_{k+1} + \beta_k\mathbf{d}_k\end{aligned}$$

③ falls $\|\mathbf{s}_{k+1}\|_2 > \varepsilon$ erhöhe $k := k+1$ und gehe nach ②

Bemerkung Optimal ist eine Vorkonditionierung mit einer Konditionszahl $\operatorname{cond}_2\left(\mathbf{M}^{-1}\mathbf{A}\right)$, welche klein und möglichst unabhängig von der Maschenweite h ist. Dies bedeutet, dass $\mathbf{M}$ möglichst "nahe" an $\mathbf{A}$ gewählt werden sollte. Zusätzlich muss $\mathbf{M}^{-1}$ jedoch auch billig anzuwenden sein. △

Eine einfache Vorkonditionierung bietet die ILU-Zerlegung (kurz für *inclomplete LU-decomposition*). Da bei einer gewöhnlichen Cholesky-Zerlegung $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ die Dreicksmatrix $\mathbf{L}$ weit mehr Nichtnullelemente besitzt als $\mathbf{A}$, rechnet man die Cholesky-Zerlegung nur auf einer Untermenge $E \subseteq \{(i,j) : 1 \le i,j \le n\}$ aus. Geeignete Mengen E kann man über Sternmuster definieren, zum Beispiel

$$\begin{bmatrix} 0 & \star & 0 \\ \star & \star & \star \\ 0 & \star & 0 \end{bmatrix}_\star, \qquad \begin{bmatrix} \star & \star & 0 \\ \star & \star & \star \\ 0 & \star & \star \end{bmatrix}_\star, \qquad \begin{bmatrix} \star & \star & \star \\ \star & \star & \star \\ \star & \star & \star \end{bmatrix}_\star.$$

Die Diagonalelemente sind dann gegeben durch

$$\ell_{j,j} = \sqrt{a_{j,j} - \sum_{\substack{k<j \\ (j,k)\in E}} \ell_{j,k}^2} > 0, \quad 1 \le j \le n,$$

und alle Nebendiagonalelemente aus E durch

$$\ell_{i,j} = \frac{1}{\ell_{j,j}}\Bigg(a_{i,j} - \sum_{\substack{k<j \\ (i,k),(j,k)\in E}} \ell_{i,k}\ell_{j,k}\Bigg), \quad j < i \le n \text{ mit } (i,j) \in E.$$

Insgesamt erhält man also eine Zerlegung $\mathbf{A} = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^T + \mathbf{R}$ mit einer Restmatrix $\mathbf{R} \neq \mathbf{0}$.

Beispiel 5.26 Wir wollen eine ILU-Zerlegung vorstellen, die *beweisbar* die Konditionszahl verbessert. Dazu bemerken wir zunächst, dass offensichtlich die Systemmatrizen zum Differentialoperator $-\Delta_h$ und $-\Delta_h + \zeta$ für alle $\zeta \ge 0$ spektral äquivelent sind. Weiterhin nehmen wir der Einfachheit halber an, dass auf den horizontalen Gitterlinien stets gleich viele und zwar m Knoten liegen. Der Nachbar des Knotens i nach Süden trägt dann den Index $i-m$ und der nach Westen den Index $i-1$. Als Matrizen $\widetilde{\mathbf{L}}$ und $\mathbf{R}$ machen wir den Ansatz

$$\begin{bmatrix} & 0 & \\ b_{i-1} & a_i & 0 \\ & c_{i-m} & \end{bmatrix}_\star \quad \text{und} \quad \begin{bmatrix} -r_i & & \\ & r_i + r_{i-m+1} & \\ & & -r_{i-m+1} \end{bmatrix}_\star .$$

Die Beziehung $\mathbf{A} = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^T + \mathbf{R}$ entspricht dann in der Sternnotation der Gleichung

$$\begin{bmatrix} & -\gamma_i & \\ -\beta_{i-1} & \alpha_i & -\beta_i \\ & -\gamma_{i-m} & \end{bmatrix}_\star = \begin{bmatrix} b_{i-1}c_{i-1} & & a_ic_i \\ a_{i-1}b_{i-1} & a_i^2 + b_{i-1}^2 + c_{i-m}^2 & a_ib_i \\ & c_{i-m}a_{i-m} & b_{i-m}c_{i-m} \end{bmatrix}_\star + \begin{bmatrix} -r_i & & \\ & r_i + r_{i-m+1} & \\ & & -r_{i-m+1} \end{bmatrix}_\star .$$

Hieraus ergeben sich die Gleichungen

$$a_i^2 = \alpha_i - b_{i-1}^2 - c_{i-m}^2 - r_i - r_{i-m+1}$$
$$b_i = -\beta_i/a_i, \quad c_i = -\gamma_i/a_i, \quad r_i = b_{i-1}c_{i-1}.$$

Setzen wir konkret $\alpha_i = 4 + 8h^2$ und $\beta_i = \gamma_i = 1$ für alle inneren Punkte und $\beta_i = \gamma_i = 0$ für alle Randpunkte, dann folgt durch vollständige Induktion:

$$a_i \ge \sqrt{2}(1+h), \quad 0 < r_i \le \frac{1}{2(1+h)^2}.$$

Mit der Formel $(x+y)^2 \le 2(x^2+y^2)$ können wir $\mathbf{x}^T\mathbf{R}\mathbf{x}$ abschätzen:

$$\begin{aligned} 0 \le \mathbf{x}^T\mathbf{R}\mathbf{x} &= \sum_i r_i(x_i - x_{i+m-1})^2 \\ &\le \sum_i \frac{1}{(1+h)^2}\big\{(x_i - x_{i-1})^2 + (x_{i-1} - x_{i+m-1})^2\big\} \\ &\le \frac{1}{(1+h)^2}\mathbf{x}^T\mathbf{A}\mathbf{x}. \end{aligned}$$

Aus $\mathbf{A} = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^T + \mathbf{R}$ folgt

$$\mathbf{x}^T\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^T\mathbf{x} \le \mathbf{x}^T\mathbf{A}\mathbf{x} \le \frac{(1+h)^2}{h(2+h)}\mathbf{x}^T\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^T\mathbf{x}$$

und damit

$$\operatorname{cond}_2\big((\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^T)^{-1}\mathbf{A}\big) \le \frac{1+h}{h} = \mathcal{O}(h^{-1}).$$

△

5.7 Mehrgitterverfahren*

Definition 5.27 Ist $\mathbf{A} \in \mathbb{R}^{m\times n}$ und $\mathbf{B} \in \mathbb{R}^{p\times q}$, so ist das **Tensorprodukt** oder **Kronecker-Produkt** $\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{mp\times nq}$ definiert durch

$$\mathbf{C} = \left[a_{i,j}\mathbf{B}\right]_{i=1,\dots,n,j=1,\dots,m} = \begin{bmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,m}\mathbf{B} \\ \vdots & & \vdots \\ a_{n,1}\mathbf{B} & \cdots & a_{n,m}\mathbf{B} \end{bmatrix}.$$

Lemma 5.28 Für $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}, \mathbf{C} \in \mathbb{R}^{q\times r}, \mathbf{D} \in \mathbb{R}^{r\times s}$ gilt

$$(\mathbf{A} \otimes \mathbf{C}) \cdot (\mathbf{B} \otimes \mathbf{D}) = \mathbf{AB} \otimes \mathbf{CD}.$$

Beweis. Wegen

$$\mathbf{A} \otimes \mathbf{C} = [a_{i,j}\mathbf{C}]_{i,j}, \qquad \mathbf{B} \otimes \mathbf{D} = [b_{i,j}\mathbf{D}]_{i,j}$$

besteht $\mathbf{F} = (\mathbf{A} \otimes \mathbf{C}) \cdot (\mathbf{B} \otimes \mathbf{D})$ aus den Blockmatrizen

$$\mathbf{F}_{i,j} := \sum_{k=1}^{n} (a_{i,k}\mathbf{C})(b_{k,j}\mathbf{D}) = \sum_{k=1}^{n} a_{i,k}b_{k,j} \cdot \mathbf{C} \cdot \mathbf{D} = \left(\sum_{k=1}^{n} a_{i,k}b_{k,j}\right) \cdot \mathbf{CD},$$
$$1 \le i \le m, \quad 1 \le j \le p.$$

Andererseits besteht $\mathbf{G} = \mathbf{AB} \otimes \mathbf{CD}$ aus den Blockmatrizen

$$\mathbf{G}_{i,j} := g_{i,j} \cdot \mathbf{CD} = \left(\sum_{k=1}^{n} a_{i,k}b_{k,j}\right) \cdot \mathbf{CD}, \quad 1 \le i \le m, \quad 1 \le j \le p,$$

woraus sich die Behauptung ergibt. □

Folgerungen:

1. Es gilt $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.
2. Sind $(\lambda, \mathbf{v})$ und $(\mu, \mathbf{w})$ Eigenpaare von $\mathbf{A}$ bzw. $\mathbf{B}$, dann ist $(\lambda\mu, \mathbf{v}\otimes\mathbf{w})$ ein Eigenpaar von $\mathbf{A} \otimes \mathbf{B}$.
3. Seien $\mathbf{A} \in \mathbb{R}^{n\times n}$ bzw. $\mathbf{B} \in \mathbb{R}^{m\times m}$ und $(\lambda, \mathbf{v})$ bzw. $(\mu, \mathbf{w})$ zugehörige Eigenpaare. Dann ist $(\mu + \lambda, \mathbf{v} \otimes \mathbf{w})$ ein Eigenpaar von $\mathbf{A} \otimes \mathbf{I}^{(m\times m)} + \mathbf{I}^{(n\times n)} \otimes \mathbf{B}$.

Wir betrachten die Diskretisierung der Poisson-Gleichnung auf dem Einheitsquadrat durch den 5-Punkte-Stern zur Schrittweite $h = 1/n$ (vergleiche Beispiel 5.4). Bezeichnet

$$\mathbf{L} = \begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}$$

die Diskretisierung des eindimensionalen Laplace-Operators und $\mathbf{I}$ die Einheitmatrix im $\mathbb{R}^{n-1}$, dann folgt bei zeilenweiser Numerierung der Unbekannten das lineare Gleichungssystem

$$\mathbf{A}\mathbf{u} = (\mathbf{L} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{L})\mathbf{u} = \mathbf{f}.$$

Wir wollen dieses Gleichungssystem mit der *Richardson-Iteration*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \omega(\mathbf{f} - \mathbf{A}\mathbf{x}_k), \quad k = 0, 1, 2, \ldots \tag{5.14}$$

lösen.

Bemerkung Die Richardson-Iteration ist das Gradientenverfahren zur Minimierung von

$$\Phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{f} \to \min$$

mit fester Schrittweite ω. Für das Modellproblem und $\omega = 1/4$ stimmt die Richardson-Iteration mit dem *Gesamtschrittverfahren*

$$\mathbf{x}_{k+1} = \mathbf{D}^{-1}[\mathbf{f} + (\mathbf{L} + \mathbf{R})\mathbf{x}_k], \quad k = 0, 1, 2, \ldots$$

überein, wobei $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{R}$. △

Die Eigenpaare von $\mathbf{L}$ sind gemäß Beispiel 5.24

$$\lambda_i = 2 - 2\cos\left(\frac{i\pi}{n}\right), \qquad \mathbf{v}_i = \left[\sin\left(\frac{ik\pi}{n}\right)\right]_{k=1}^{n-1}, \qquad i = 1, 2, \ldots, n-1.$$

Folglich besitzt die Iterationsmatrix $\mathbf{I} - \omega\mathbf{A}$ die Eigenwerte

$$\begin{aligned}\mu_{i,j} &= 1 - \omega\Bigg[\underbrace{2 - 2\cos\left(\frac{i\pi}{n}\right)}_{=4\sin^2(i\pi/(2n))} + \underbrace{2 - 2\cos\left(\frac{j\pi}{n}\right)}_{=4\sin^2(j\pi/(2n))}\Bigg] \\ &= 1 - 4\omega\left[\sin^2\left(\frac{i\pi}{2n}\right) + \sin^2\left(\frac{j\pi}{2n}\right)\right]\end{aligned}$$

mit zugehörigen Eigenvektoren $\mathbf{w}_{i,j} := \mathbf{v}_i \otimes \mathbf{v}_j$. Für $0 < \omega \leq 1/4$ sind alle Eigenwerte betragsmäßig kleiner als Eins und die Richardson-Iteration konvergiert. Aber wegen

$$\|\mathbf{I} - \omega\mathbf{A}\|_2 = \mu_{1,1} = 1 - \mathcal{O}(h^2)$$

wird die Konvergenz für $h \to 0$ immer langsamer. Hingegen ist die Konvergenz für alle Eigenvektoren mit hohem Frequenzanteil in mindestens in einer Richtung unabhängig von h gegeben durch

$$\max\{|\mu_{i,j}| : i \geq n/2 \text{ oder } j \geq n/2\} \leq 1 - 4\omega\left[\sin^2\left(\frac{\pi}{4}\right) + \sin^2(0)\right] = 1 - 2\omega < 1.$$

Somit haben wir folgendes Ergebnis gezeigt:

Lemma 5.29 Sei

$$V_{\text{osc}} := \operatorname{span}\left\{\mathbf{w}_{i,j} : 1 \le i,j < n,\ \max\{i,j\} \ge n/2\right\}$$

der Unterraum der hohen Frequenzen. Für den Startwert $\mathbf{x}_0$ der Richardson-Iteration (5.14) gelte $\mathbf{e}_0 := \mathbf{x} - \mathbf{x}_0 \in V_{\text{osc}}$. Dann liegen alle weiteren Fehler $\mathbf{e}_k := \mathbf{x} - \mathbf{x}_k$ ebenfalls in V_{osc} und es gilt

$$\|\mathbf{e}_{k+1}\|_2 \le (1-2\omega)\|\mathbf{e}_k\|_2.$$

Korollar 5.30 Der Startfehler $\mathbf{e}_0$ sei aufgeteilt in

$$\mathbf{e}_0 = \mathbf{e}_{0,\text{osc}} + \mathbf{e}_{0,\text{glatt}}, \qquad \mathbf{e}_{0,\text{osc}} \in V_{\text{osc}}, \qquad \mathbf{e}_{0,\text{glatt}} \in V_{\text{glatt}} := V_{\text{osc}}^{\perp}.$$

Dann gilt

$$\mathbf{e}_k = \mathbf{e}_{k,\text{osc}} + \mathbf{e}_{k,\text{glatt}}$$

mit

$$\mathbf{e}_{k,\text{osc}} = (\mathbf{I} - \omega\mathbf{A})^k \mathbf{e}_{0,\text{osc}} \in V_{\text{osc}}, \qquad \mathbf{e}_{k,\text{glatt}} = (\mathbf{I} - \omega\mathbf{A})^k \mathbf{e}_{k,\text{glatt}} \in V_{\text{glatt}}$$

und

$$\|\mathbf{e}_{k,\text{osc}}\|_2 \le (1-2\omega)^k \|\mathbf{e}_{0,\text{osc}}\|_2.$$

Folglich werden die hochfrequenten Fehleranteile $\mathbf{e}_{0,\text{osc}}$ mit der Richardson-Iteration schnell gedämpft, während die tieffrequenten Fehleranteile $\mathbf{e}_{0,\text{glatt}}$ sehr langsam reduziert werden. Die Idee des Mehrgitterverfahrens ist es nun, diese tieffrequenten Lösungsanteile auf einem gröberen Gitter zu approximieren.

Zutaten eines Mehrgitterverfahrens:

1. Wir benötigen eine Hierarchie von Gittern

   $$\Omega_0 \subset \Omega_1 \subset \Omega_2 \subset \cdots, \qquad \Omega_j := \Omega_{h_j},$$

   wobei $h_j = 2^{-j} h_0$. Auf jedem Gitter erhalten wir ein Gleichungssystem

   $$\mathbf{A}_j \mathbf{u}_j = \mathbf{f}_j, \quad \mathbf{u}_j \in V_j := \mathbb{R}^{|\Omega_j|}.$$

2. Für den Gittertransfer $\Omega_{j-1} \mapsto \Omega_j$ benötigen wir eine *Prolongation*

   $$\mathbf{P}_j : V_{j-1} \to V_j,$$

   welche eine Gitterfunktion $\mathbf{u}_{j-1} \in V_{j-1}$ auf das feinere Gitter Ω_j transferiert. Naheliegend ist die 9-Punkt-Prolongation

   $$\begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}_{\star},$$

   die einer bilinearen Interpolation von $\mathbf{u}_{j-1}$ in allen Gitterpunkten von Ω_j entspricht.

3. Umgekehrt benötigen wir für den Gittertransfer $\Omega_j \mapsto \Omega_{j-1}$ eine *Restriktion*
$$\mathbf{R}_j : V_j \to V_{j-1},$$
welche einer Gitterfunktion $\mathbf{u}_j \in V_j$ auf das gröbere Gitter Ω_{j-1} einschränkt. Viel besser als die triviale Restriktion
$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_\star$$
ist die 9-Punkt-Restriktion
$$\frac{1}{4}\begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}_\star,$$
da diese wegen $4\mathbf{R}_j^T = \mathbf{P}_j$ die Symmetrie bewahrt. Insbesondere gilt dann $\mathbf{A}_{j-1} = \mathbf{R}_j\mathbf{A}_j\mathbf{P}_j$.
4. Als *Glätter* verwendet man ein Iterationsverfahren wie die Richardson-Iteration, das Gesamtschrittverfahren oder das *Einzelschrittverfahren*
$$\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{L})^{-1}(\mathbf{b} + \mathbf{R}\mathbf{x}_k), \quad k = 0, 1, 2, \ldots.$$
5. Auf dem gröbsten Gitter Ω_0 wird ein direkter *Grobgitterlöser* für das Gleichungssystem $\mathbf{A}_0\mathbf{x}_0 = \mathbf{f}_0$ eingesetzt.

Bemerkung Obwohl Restriktion und Prolongation mit demselben Stern beschrieben werden, ist ihre Bedeutung komplett unterschiedlich: bei der Restriktion werden die Daten gemittelt, während sie bei der Prolongation verteilt werden. △

Algorithmus 5.31 (Zweigitterverfahren)
input: Diskretisierungslevel j, Vorglättungsschritte ν_1, Nachglättungsschritte ν_2, Startnäherung $\mathbf{x}_{\text{in}} \in \mathbb{R}^{n_j}$, Systemmatrizen $\mathbf{A}_{j-1}, \mathbf{A}_j$, rechte Seite $\mathbf{f}_j \in \mathbb{R}^{n_j}$
output: approximative Lösung $\mathbf{x}_{\text{out}}$
① wende ν_1-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_{\text{in}}$ an und erhalte $\mathbf{x}_j$
② bilde Residuum $\mathbf{r}_j := \mathbf{f}_j - \mathbf{A}_j\mathbf{x}_j$
③ bilde $\mathbf{f}_{j-1} := \mathbf{R}_j\mathbf{r}_j$
④ löse $\mathbf{A}_{j-1}\mathbf{x}_{j-1} = \mathbf{f}_{j-1}$
⑤ bilde $\mathbf{x}_j := \mathbf{x}_j + \mathbf{P}_j\mathbf{x}_{j-1}$
⑥ wende ν_2-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_j$ an und erhalte $\mathbf{x}_{\text{out}}$

Benutzen wir das Richardson-Verfahren als Glätter und setzen wir $\mathbf{e}_{\text{in}} := \mathbf{A}_j^{-1}\mathbf{f}_j - \mathbf{x}_{\text{in}}$, dann ist der Fehler nach Schritt ① genau $\mathbf{e}_{①} = (\mathbf{I} - \omega\mathbf{A}_j)^{\nu_1}\mathbf{e}_{\text{in}}$. Die *Grobgitterkorrektur*, die in den Schritten ②–⑤ berechnet wird, ist gerade $\mathbf{d} = \mathbf{P}_j\mathbf{A}_{j-1}^{-1}\mathbf{R}_j\mathbf{A}_j\mathbf{e}_{①}$. Beachtet man noch das Nachglätten, dann folgt
$$\mathbf{e}_{\text{out}} := \mathbf{A}_j^{-1}\mathbf{f}_j - \mathbf{x}_{\text{out}} = (\mathbf{I} - \omega\mathbf{A}_j)^{\nu_2}(\mathbf{I} - \mathbf{P}_j\mathbf{A}_{j-1}^{-1}\mathbf{R}_j\mathbf{A}_j)(\mathbf{I} - \omega\mathbf{A}_j)^{\nu_1}\mathbf{e}_{\text{in}}.$$
Die rekursive Anwendung des obigen Algorithmus liefert das Mehrgitterverfahren:

Algorithmus 5.32 (Mehrgitterverfahren)

input: Diskretisierungslevel j, Rekursionparameter γ,
Vorglättungsschritte ν_1, Nachglättungsschritte ν_2,
Startnäherung $\mathbf{x}_{\text{in}} \in \mathbb{R}^{n_j}$, Systemmatrizen $\mathbf{A}_0, \ldots, \mathbf{A}_j$, rechte Seite $\mathbf{f}_j \in \mathbb{R}^{n_j}$

output: approximative Lösung $\mathbf{x}_{\text{out}}$

① falls $j = 0$, dann $\mathbf{x}_{\text{out}} = \mathbf{A}_j^{-1}\mathbf{f}_j$ und stop

② wende ν_1-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_{\text{in}}$ an und erhalte $\mathbf{x}_j$

③ bilde Residuum $\mathbf{r}_j := \mathbf{f}_j - \mathbf{A}_j\mathbf{x}_j$

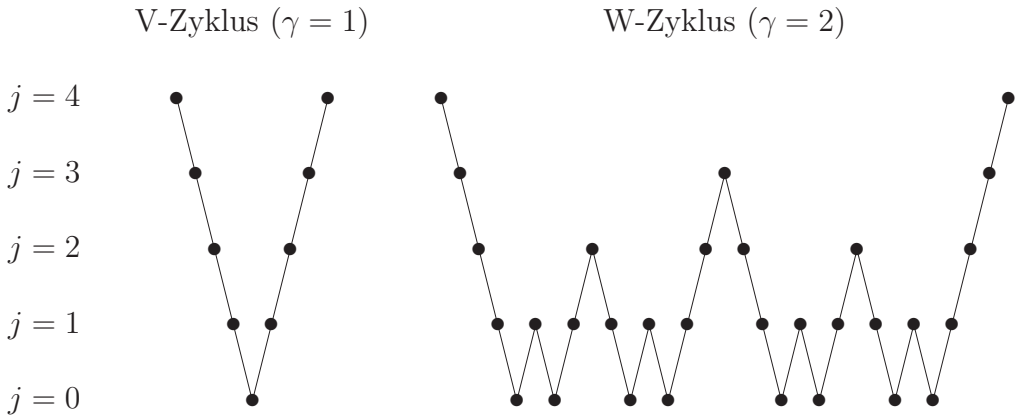
④ bilde $\mathbf{f}_{j-1} := \mathbf{R}_j\mathbf{r}_j$

⑤ wende γ-mal das Mehrgitterverfahren mit Startwert $\mathbf{0}$ zur Lösung von $\mathbf{A}_{j-1}\mathbf{u}_{j-1} = \mathbf{f}_{j-1}$ an und erhalte $\mathbf{x}_{j-1}$

⑥ bilde $\mathbf{x}_j := \mathbf{x}_j + \mathbf{P}_j\mathbf{x}_{j-1}$

⑦ wende ν_2-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_j$ an und erhalte $\mathbf{x}_{\text{out}}$

Je nach Wahl des Rekursionsparamters γ erhalten wir folgendes Schema:



Lemma 5.33 Für $\gamma < 2^d$ ist der Aufwand des Mehrgitterverfahrens 5.32 linear.

Beweis. Für die gewählten Parameter (ν_1, ν_2, γ) bezeichne $T(j)$ den Aufwand des Mehrgitterverfahrens zum Level j. Die Zahl der Unbekannten n_j im Level j skaliert wegen $h_j = 2^{-j}h_0$ wie 2^{dj}. Da das Matrix-Vektor-Produkt linear skaliert, besitzt ein Glättungsschritt die Komplexität $c_G n_j$. Ebenso skalieren die Prolongation und die Restriktion linear, $c_R n_j$ beziehungsweise $c_P n_j$. Damit erhalten wir die Rekursion

$$T(j) \leq \underbrace{\nu_1 c_G n_j}_{\text{vorglätten}} + \underbrace{c_R n_j}_{\text{Restriktion}} + \gamma T(j-1) + \underbrace{c_P n_j}_{\text{Prolongation}} + \underbrace{\nu_2 c_G n_j}_{\text{nachglätten}} \leq 2^{jd}C + \gamma T(j-1),$$

wobei $T(0) = c_L n_0^3$ gilt. Rekursiv erhalten wir

$$T(j) \leq \gamma^j T(0) + \sum_{k=1}^{j} 2^{kd} C \gamma^{j-k} \leq 2^{jd} \max\{C, T(0)\} \sum_{k=0}^{j} \Bigg(\underbrace{\frac{\gamma}{2^d}}_{<1} \Bigg)^k = \mathcal{O}(n_j).$$

□

5.8 Parabolische Differentialgleichungen

Wir betrachten die Wärmeleitungsgleichung

$$\frac{\partial}{\partial t}u(t,\mathbf{x}) - \Delta u(t,\mathbf{x}) = f(t,\mathbf{x}), \quad (t,\mathbf{x}) \in [0,T] \times \Omega$$

für ein Gebiet $\Omega \subset \mathbb{R}^d$ und einen Endzeitpunkt T. Der Einfachheit halber geben wir uns homogene Dirichlet-Randwerte vor

$$u(\cdot,\mathbf{x}) = 0 \quad \text{auf } \Gamma = \partial\Omega.$$

Um diese Gleichung numerisch zu lösen, führen wir zunächst eine Semidiskretisierung im Raum durch, indem wir wie in Abschnitt 5.1 vorgehen. Dazu sei $\Omega_h \cup \Gamma_h$ ein Gitter von $\overline{\Omega}$ und

$$u_h(t,\mathbf{x}) : [0,T] \times \overline{\Omega}_h \to \mathbb{R}$$

eine zugehörige Gitterfunktion. Die semidiskrete Gleichung

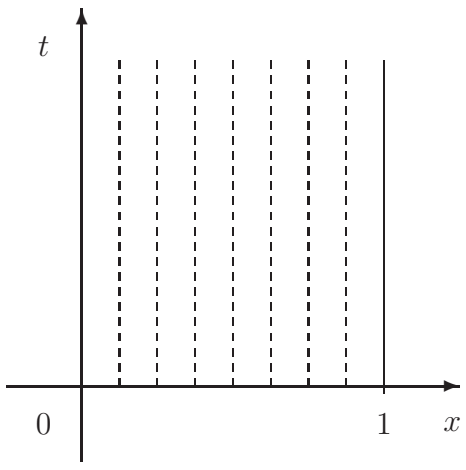
$$\begin{aligned} \frac{\partial}{\partial t}u_h(t,\mathbf{x}) - \big(\Delta_h u_h\big)(t,\mathbf{x}) &= f(t,\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega_h, \\ u_h(t,\mathbf{x}) &= 0 \qquad\quad \text{für alle } \mathbf{x} \in \Gamma_h, \end{aligned}$$

mit der Anfangsbedingung $u_h(0,\mathbf{x}) = g(\mathbf{x})$ für alle $\mathbf{x} \in \Omega_h$ entspricht einem System gewöhnlicher Differentialgleichungen

$$\frac{\partial}{\partial t}\mathbf{u}(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{f}(t), \quad \mathbf{u}(0) = \mathbf{g}. \tag{5.15}$$

Dabei ist die Matrix $\mathbf{A}$ unabhängig von der Zeit und die Elliptizität des Laplace-Operators impliziert, dass alle Eigenwerte von $\mathbf{A}$ einen positiven Realteil besitzen. Folglich ist die Differentialgleichung (5.15) steif.

Bemerkung Aus anschaulichen Gründen wird die Semidiskretisierung (5.15) *Linienmethode* genannt. Für alle $t \geq 0$ enthält nämlich die vektorwertige Funktion $\mathbf{u}(t)$ die Funktionswerte der Approximation $u_h(t,\mathbf{x})$ in den Gitterpunkten. △



Linienmethode für $\Omega = (0,1)$

Wir benötigen noch eine geeignete Zeitdiskretisierung. Das explizite Euler-Verfahren führt auf

$$\frac{\mathbf{u}_{i+1}-\mathbf{u}_i}{\Delta t}+\mathbf{A}\mathbf{u}_i=\mathbf{f}(t_i)\quad\text{bzw.}\quad \mathbf{u}_{i+1}=(\mathbf{I}-\Delta t\,\mathbf{A})\mathbf{u}_i+\Delta t\,\mathbf{f}(t_i). \tag{5.16}$$

Hingegen liefert das implizite Euler-Verfahren

$$\frac{\mathbf{u}_{i+1}-\mathbf{u}_i}{\Delta t}+\mathbf{A}\mathbf{u}_{i+1}=\mathbf{f}(t_{i+1})\quad\text{bzw.}\quad (\mathbf{I}+\Delta t\,\mathbf{A})\mathbf{u}_{i+1}=\mathbf{u}_i+\Delta t\,\mathbf{f}(t_{i+1}). \tag{5.17}$$

Der Startwert lautet in beiden Fällen $\mathbf{u}_0=\mathbf{g}$. Kombinieren wir (5.16) und (5.17), so erhalten wir das *θ-Schema*

$$\frac{\mathbf{u}_{i+1}-\mathbf{u}_i}{\Delta t}+(1-\theta)\mathbf{A}\mathbf{u}_i+\theta\mathbf{A}\mathbf{u}_{i+1}=(1-\theta)\mathbf{f}(t_i)+\theta\mathbf{f}(t_{i+1}),$$

beziehungsweise

$$(\mathbf{I}+\Delta t\,\theta\,\mathbf{A})\mathbf{u}_{i+1}=\big(\mathbf{I}-\Delta t\,(1-\theta)\,\mathbf{A}\big)\mathbf{u}_i+\Delta t\big\{(1-\theta)\mathbf{f}(t_i)+\theta\mathbf{f}(t_{i+1})\big\}.$$

Dabei gilt

$$\theta=\begin{cases}0, & \text{explizites Euler-Verfahren}\\ 1/2, & \text{Trapez-Methode}\\ 1, & \text{implizites Euler-Verfahren}\end{cases}$$

wobei die Trapez-Methode im Zusammenhang mit parabolischen Differentialgleichungen auch *Crank-Nicolson-Verfahren* genannt wird. Das θ-Schema ist konsistent von erster Ordnung, im Falle $\theta=0.5$ sogar von zweiter Ordnung. Es ist A-stabil für alle $\theta\in[0.5,1]$ und L-stabil für $\theta=1$.

Bemerkung Da das Crank-Nicolson-Verfahren das einfachste Verfahren von zweiter Ordnung ist, ist es unheimlich populär. Aufgrund der fehlenden L-Stabilität, kann es allerdings unphysikalische Oszillationen produzieren und sollte daher gemieden werden. Am besten wählt man $\theta=1/2+\xi$ mit einem geeigneten $\xi>0$, durch das die Größe der Dämpfung gesteuert wird. △

Ein geeignetes, weil L-stabiles, Verfahren höherer Ordnung ist das Radau3-Verfahren aus Beispiel 2.13. Mit $t_{i+1/2}:=t_i+\Delta t/3$ lautet es

$$\begin{bmatrix}\mathbf{I}+\frac{5\Delta t}{12}\mathbf{A} & -\frac{\Delta t}{12}\mathbf{A}\\ \frac{3\Delta t}{4}\mathbf{A} & \mathbf{I}+\frac{\Delta t}{4}\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{k}_1\\ \mathbf{k}_2\end{bmatrix}=\begin{bmatrix}\mathbf{f}(t_{i+1/2})-\mathbf{A}\mathbf{u}_i\\ \mathbf{f}(t_{i+1})-\mathbf{A}\mathbf{u}_i\end{bmatrix},\quad \mathbf{u}_{i+1}=\mathbf{u}_i+\frac{3\Delta t}{4}\mathbf{k}_1+\frac{\Delta t}{4}\mathbf{k}_2.$$

Jedoch muss hier in jedem Schritt ein Gleichungssystem gelöst werden, das doppelt so viele Unbekannte wie Ortsunbekannte enthält.

Letzteres ist der Grund für die Klasse der SDIRK-Verfahren, vergleiche Beispiel 2.13. Das SDIRK2-Verfahren führt auf

$$\begin{aligned}(\mathbf{I}+\gamma\Delta t\mathbf{A})\mathbf{k}_1&=\mathbf{f}(t_{i+1/2})-\mathbf{A}\mathbf{u}_i\\ (\mathbf{I}+\gamma\Delta t\mathbf{A})\mathbf{k}_2&=\mathbf{f}(t_{i+1})-\mathbf{A}\big(\mathbf{u}_i+(1-\gamma)\Delta t\mathbf{k}_1\big)\\ \mathbf{u}_{i+1}&=\mathbf{u}_i+(1-\gamma)\Delta t\mathbf{k}_1+\gamma\Delta t\mathbf{k}_2,\end{aligned}$$

wobei $t_{i+1/2}:=t_i+\gamma\Delta t$ und $\gamma:=(2-\sqrt{2})/2$ ist. Es müssen nun in jedem Schritt zwei Gleichungssysteme gelöst werden, die allerdings nur soviel Unbekannte wie Ortsunbekannte besitzen. Speziell ist die Systemmatrix immer dieselbe, weshalb beispielsweise nur eine einzige Cholesky-Zerlegung berechnet werden muss.

Index

3/8-Regel, 17
θ-Schema, 85

Adams-Bashforth-Verfahren, 31, 34
Adams-Moulton-Verfahren, 31, 34
Algorithmus
    Adams-Bashforth-Verfahren, 34
    CG-Verfahren, 68
    Eulersches Polygonzug-Verfahren, 10
    explizites Runge-Kutta-Verfahren, 17
    GMRES-Verfahren, 72
    implizites Runge-Kutta-Verfahren, 19
    lineares Mehrschrittverfahren, 30
    Mehrgitterverfahren, 83
    MINRES-Verfahren, 73
    PCG-Verfahren, 77
    Schrittweitensteuerung, 23, 25, 47
    Zweigitterverfahren, 82
Anfangsbedingung, 4
Anlaufwerte, 30

BDF-Verfahren, 35
Butcher-Tableau, 16

CG-Verfahren, 68
Crank-Nicolson-Verfahren, 85

Dahlquistsche Wurzelbedingung, 38
Differentialgleichung
    elliptische, 53
    gewöhnliche, 4
    hyperbolische, 53
    parabolische, 53
    steife, 27
Differentialoperator, 53
    elliptischer, 53
    hyperbolischer, 53
    parabolischer, 53
Differenz
    linksseitige, 58
    rechtsseitige, 58
    zentrale, 58
Differenzengleichung
    lineare, 38
Differenzenstern, 60
    5-Punkte-Stern, 60
    für beliebigen Differentialoperator, 62
Differenzenverfahren, 61, 63
Diskretisierungsfehler
    globaler, 13
    lokaler, 11, 31

Einschrittverfahren, 11
    explizites, 11
    implizites, 11
Einzelschrittverfahren, 82
Energienorm, 67
eq:Stabilitaetspolynom, 49
Euler-Collatz-Verfahren, 13
Euler-Verfahren
    explizites, 11
    implizites, 11
Eulersches Polygonzug-Verfahren, 10
Eulersches Polygonzugverfahren, 34

Funktion
    harmonische, 58

Gebiet, 51
    diskret zusammenhängend, 63
Gesamtschrittverfahren, 80
Gills-Formel, 17
Gitter, 59
Gitterfunktion, 61
Gitterpunkt, 59
    randferner, 59
    randnaher, 59
Gleichung
    Laplace-, 51
    Poisson-, 52
    Potential-, 51

Wärmeleitungs-, 51, 84
Wellen-, 53
GMRES-Verfahren, 71
Grobgitter
-korrektur, 82
-löser, 82

Inkrementfunktion, 11

Konistenz, 65
Konsistenz, 11, 31
Kontrollverfahren, 22
Konvergenz, 13, 41, 65
Kronecker-Produkt, 79
Krylov-Raum, 70

Lösung
klassische, 58
Laplace-Gleichung, 51
Laplace-Operator, 51, 52
Linienmethode, 84

Maximumprinzip, 55
diskretes, 64
Mehrgitterverfahren, 83
Mehrschrittverfahren, 30
explizites, 30
implizites, 30
lineares, 30
Milne-Simpson-Verfahren, 31, 35
Minimumprinzip, 56
MINRES-Verfahren, 72
Mittelpunktsregel, 18, 34

Newton-Cotes-Quadraturformeln, 32
Nyström-Verfahren, 31, 34

Operator
Laplace-, 51, 52

PCG-Verfahren, 77
Poisson-Gleichung, 52
Potentialgleichung, 51
Prädiktor-Korrektor-Verfahren, 37
Problem
sachgemäß gestelltes, 54
schlecht gestelltes, 54
Prolongation, 81

Rückwärts-Differentiationsformeln, 35
Radau3-Verfahren, 18, 85
Randbedingung
Dirichlet-, 54
Neumann-, 54
Randpunkt, 59
Regel
3/8-, 17
Mittelpunkts-, 34
Simpson-, 36
Trapez-, 36
Residuum, 68
Restriktion, 82
Richardson-Extrapolation, 24
Richardson-Iteration, 80
Runge-Kutta-Verfahren, 15
eingebettetes, 22
explizites, 16, 17
implizites, 17, 19
klassisches, 17

Schema
θ-, 85
SDIRK2-Verfahren, 18
Shortley-Weller-Approximation, 60
Simpson-Regel, 36
Stabilität, 65
A-, 28, 49
L-, 29
Stabilitätsgebiet, 28, 49
Steifigkeitsquotient, 27

Tensorprodukt, 79
Trapezregel, 36

V-Zyklus, 83
Vektoren
konjugierte, 67
Verfahren
Adams-Bashforth-, 31, 34
Adams-Moulton-, 31, 34
BDF-, 35
Crank-Nicolson-, 85
der konjugierten Gradienten, 67, 68
der konjugierten Gradienten mit Vorkonditionierung, 77
Differenzen-, 61, 63
Einschritt-, 11
Einzelschritt-, 82
Euler-
explizites, 11

implizites, 11
Euler-Collatz-, 13
Eulersches Polygonzug-, 10, 34
Gesamtschritt-, 80
GMRES-, 71
Mehrgitter-, 83
Mehrschritt-, 30
lineares, 30
Milne-Simpson-, 31, 35
MINRES-, 72
Nyström-, 31, 34
Prädiktor-Korrektor-, 37
Radau3-, 18, 85
Runge-Kutta-, 15
eingebettetes, 22
explizites, 16, 17
implizites, 17, 19
klassisches, 17
SDIRK2-, 18
von Heun, 13
Zweigitter-, 82
Vergleichsprinzip, 56
Vorkonditionierung, 76

W-Zyklus, 83
Wärmeleitungsgleichung, 51, 84
Wellengleichung, 53
Wurzelbedingung
von Dahlquist, 38
Wurzelortskurve, 50

Zweigitterverfahren, 82