Numerik I

Gewöhnliche Differentialgleichungen und Differenzenverfahren für partielle Differentialgleichungen

Vorlesungsskriptum Wintersemester 2018/19

R. Verfürth
Fakultät für Mathematik, Ruhr-Universität Bochum

Inhaltsverzeichnis

KAPITEL I

Anfangswertprobleme für gewöhnliche Differentialgleichungen

In diesem Kapitel betrachten wir numerische Verfahren zur Lösung von Anfangswertproblemen für gewöhnliche Differentialgleichungen, d.h. von Problemen der Form: Finde eine differenzierbare Funktion y einer reellen Veränderlichen t mit $y'(t) = f(t, y(t))$ und $y(t_0) = y_0$.

Zuerst stellen wir in §I.1 einige theoretische Ergebnisse für derartige Probleme bereit und zeigen unter anderem die eindeutige Lösbarkeit dieses Problems und die stetige und differenzierbare Abhängigkeit der Lösung von dem Anfangswert y_0.

In den §§I.2 – I.4 untersuchen wir dann Einschrittverfahren, die ausgehend von einer Näherung η_i für $y(t_i)$ eine neue Näherung η_{i+1} für $y(t_i + h)$ berechnen. Wichtigste Vertreter dieser Verfahrensklasse sind das explizite und das implizite Euler-Verfahren, die Trapezregel und die Runge-Kutta-Verfahren. In §I.3 zeigen wir, dass diese Verfahren unter geeigneten Annahmen an die Funktion f gegen die Lösung y des Anfangswertproblemes konvergieren und dass der Fehler bestimmt wird von demjenigen eines einzelnen Schrittes. In §I.4 betrachten wir dann einige Aspekte zur praktischen Durchführung dieser Verfahren.

In den §§I.5 – I.7 betrachten wir lineare Mehrschrittverfahren. Bei diesen Verfahren werden zur Berechnung der Näherung η_{i+1} für $y(t_i+h)$ mehrere alte Näherungen η_i, …, η_{i-k} für $y(t_i)$, …, $y(t_i - kh)$ herangezogen. Wichtigste Vertreter dieser Verfahrensklasse sind die Adams-Bashforth-, Adams-Moulton-, Nyström- und BDF-Verfahren. Wir zeigen in §I.6 die Konvergenz der Verfahren und untersuchen in §I.7 Aspekte ihrer praktischen Implementierung.

Die Konvergenzresultate der §§I.3 und I.6 sind asymptotischer Natur, d.h. sie treffen eine Aussage über das Verhalten des Fehlers im Grenzfall $h \to 0$. Tatsächlich rechnet man aber natürlich mit einer festen endlichen Schrittweite. Verfahren, die sich im Grenzfall $h \to 0$ gleich verhalten, können für endliches h völlig unterschiedliches Verhalten aufweisen. Die genaue Analyse dieses Effektes ist Gegenstand des §I.8.

Im abschließenden §I.9 betrachten wir kurz Algebro-Differentialgleichungen, bei denen Differentialgleichungen mit algebraischen Zwangsbedingungen gekoppelt sind.

I.1. Einige theoretische Ergebnisse

In diesem Paragraphen stellen wir einige für die Numerik wichtige theoretische Ergebnisse über Anfangswertprobleme für gewöhnliche Differentialgleichungen zusammen. Die Ergebnisse werden in der Regel nicht bewiesen. Stattdessen verweisen wir auf [1, Kapitel II] und [8, Kapitel XI].

Im Folgenden bezeichnen stets $I \subset \mathbb{R}$ ein nicht leeres offenes Intervall und $U \subset \mathbb{R}^n$, $n \in \mathbb{N}^*$, eine nicht leere offene Menge; $\|\cdot\|$ ist eine beliebige Norm auf $\mathbb{R}^n$.

Definition I.1.1 (Gewöhnliche Differentialgleichungen, Anfangswertprobleme). (1) Seien $k \in \mathbb{N}^*$, $V \subset \mathbb{R}^{kn}$ eine nicht leere offene Menge und $f : I \times V \to \mathbb{R}^n$ eine Funktion. Dann heißt das Problem: Finde $y \in C^k(I, \mathbb{R}^n)$ mit

$$y^{(k)}(t) = f(t, y(t), \dots, y^{(k-1)}(t)) \tag{I.1.1}$$

in I eine *gewöhnliche Differentialgleichung k-ter Ordnung* auf I. Ist speziell $k = 1$, so sprechen wir einfach von einer *gewöhnlichen Differentialgleichung (gDgl)*.
(2) Seien $t_0 \in I$ und $v_0 = (v_{0,0}, \dots, v_{0,k-1}) \in V$. Dann heißt (I.1.1) zusammen mit den Bedingungen

$$y(t_0) = v_{0,0}, \quad \dots \quad , y^{(k-1)}(t_0) = v_{0,k-1}$$

ein *Anfangswertproblem (AWP)*.
(3) Eine gDgl bzw. ein AWP heißen *autonom*, wenn die Funktion f nicht von der Variablen t abhängt.

Bemerkung I.1.2 (Ordnungsreduktion, Autonomisierung). (1) Eine gDgl k-ter Ordnung, $k \geq 2$, kann stets in eine äquivalente gDgl erster Ordnung auf $\mathbb{R}^{nk}$ umformuliert werden. Um dies einzusehen, setze

$$z(t) = (z_0(t), \dots, z_{k-1}(t)) = (y(t), y'(t), \dots, y^{(k-1)}(t)) \in \mathbb{R}^{nk}$$

und

$$F(t, z(t)) = (z_1(t), \dots, z_{k-1}(t), f(t, z_0(t), \dots, z_{k-1}(t))) \in \mathbb{R}^{nk}.$$

Dann ist offensichtlich y genau dann eine Lösung von (I.1.1), wenn z eine Lösung von

$$z'(t) = F(t, z(t))$$

ist. Aus diesem Grunde betrachten wir im Folgenden nur gewöhnliche Differentialgleichungen erster Ordnung.
(2) Eine gDgl kann stets in eine äquivalente autonome gDgl umformuliert werden. Um dies einzusehen, setze

$$z(s) = (y(s), s) \in \mathbb{R}^{n+1}$$

und

$$F(z) = F((y(s), s)) = (f(s, y(s)), 1) \in \mathbb{R}^{n+1}.$$

Dann ist y genau dann eine Lösung von (I.1.1), wenn z eine Lösung von

$$z'(s) = F(z(s))$$

ist.

BEISPIEL I.1.3 (Populationsdynamik). Die Funktion $y(t)$ beschreibe die Größe einer Population zur Zeit $t > 0$. Die relative zeitliche Änderung der Populationsgröße sei eine bekannte Funktion r der Zeit und der aktuellen Populationsgröße. Diese Annahmen führen auf die gDgl

$$y'(t) = r(t, y(t))y(t).$$

Wichtige Spezialfälle sind die des unbeschränkten Wachstums

$$r(t, y(t)) = \alpha > 0$$

und des beschränkten Wachstums

$$r(t, y(t)) = \alpha(y^* - y(t))$$

mit $\alpha > 0$, $y^* > 0$. Die Zahl y^* spielt die Rolle einer Grenzpopulation, deren Überschreiten zum Absterben der Population führt.

BEISPIEL I.1.4 (Federschwingung). Die vertikale Schwingung $z(t)$ einer Feder mit Masse $m > 0$ wird nach den Newtonschen Kraftgesetzen durch die gDgl 2. Ordnung

$$mz''(t) = -kz(t) - rz'(t)$$

beschrieben. Dabei ist $-kz(t)$, $k > 0$, die Rückstellkraft der Feder und $-rz'(t)$, $r \geq 0$, die Reibungskraft. Die äquivalente gDgl 1. Ordnung lautet

$$\begin{aligned} z'(t) &= v(t) \\ v'(t) &= -\frac{k}{m}z(t) - \frac{r}{m}v(t). \end{aligned}$$

Dabei ist $v(t)$ die Geschwindigkeit der Auslenkung.

Wir stellen uns zunächst die Frage nach der Existenz und Eindeutigkeit von Lösungen von AWPen. Dazu benötigen wir einige Vorbereitungen.

DEFINITION I.1.5 (Lipschitz-Stetigkeit). Die Funktion $f \in C(I \times U, \mathbb{R}^n)$ heißt *gleichmäßig Lipschitz-stetig* auf $I \times U$ bzgl. U wenn es ein $L \in \mathbb{R}_+$, die sog. *Lipschitz-Konstante*, gibt mit

$$\|f(t, x) - f(t, y)\| \leq L \|x - y\|$$

für alle $t \in I$, $x, y \in U$. Die Funktion f heißt *Lipschitz-stetig* auf $I \times U$ bzgl. U, wenn es zu jedem $(t_0, x_0) \in I \times U$ eine Umgebung $J \times V \in \mathcal{U}((t_0, x_0))$ gibt, derart dass f auf $J \times V$ gleichmäßig Lipschitz-stetig ist bzgl. V.

BEMERKUNG I.1.6 (Lipschitz-Stetigkeit und Differenzierbarkeit). (1) Ist $f \in C(I \times U, \mathbb{R}^n)$ bzgl. der Variablen y differenzierbar und $D_y f \in C(I \times U, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n))$, so ist f auf $I \times U$ bzgl. U Lipschitz-stetig.
(2) Ist $f \in C(I \times U, \mathbb{R}^n)$ Lipschitz-stetig auf $I \times U$ bzgl. U und $J \times K \subset I \times U$ kompakt, so ist f auf $J \times K$ gleichmäßig Lipschitz-stetig bzgl. K.
(3) Ist $f \in C^1(U, \mathbb{R}^n)$, so ist f Lipschitz-stetig (auf U bzgl. U). Ist $f \in C(U, \mathbb{R}^n)$ Lipschitz-stetig und $K \subset U$ kompakt, so ist f gleichmäßig Lipschitz-stetig auf K.

SATZ I.1.7 (Lemma von Gronwall). *Seien $\alpha, \beta, u \in C(I, \mathbb{R}_+)$ und $t_0 \in I$ mit*

$$u(t) \le \alpha(t) + \left| \int_{t_0}^{t} \beta(s) u(s) \mathrm{d}s \right|$$

für alle $t \in I$. Dann gilt für alle $t \in I$

$$u(t) \le \alpha(t) + \left| \int_{t_0}^{t} \alpha(s)\beta(s) e^{\left|\int_s^t \beta(\sigma)\mathrm{d}\sigma\right|} \mathrm{d}s \right|.$$

BEWEIS. [1, Lemma 6.1] und [8, Satz XI.1.6].
Beweisidee: Betrachte $t \in I$ mit $t > t_0$ und setze

$$v(t) = \int_{t_0}^{t} \beta(s) u(s) \mathrm{d}s, \quad \gamma(t) = e^{-\int_{t_0}^{t} \beta(s)\mathrm{d}s}.$$

Dann folgt

$$\begin{aligned}(\gamma v)'(t) &= \gamma'(t)v(t) + \gamma(t)v'(t) = -\beta(t)\gamma(t)v(t) + \gamma(t)\beta(t)u(t) \\ &\le \gamma(t)\beta(t)\alpha(t)\end{aligned}$$

und somit

$$\begin{aligned}\int_{t_0}^{t} \beta(s)u(s)\mathrm{d}s = v(t) &\le \gamma(t)^{-1} \int_{t_0}^{t} \alpha(s)\beta(s)\gamma(s)\mathrm{d}s \\ &= \int_{t_0}^{t} \alpha(s)\beta(s) e^{\int_s^t \beta(\sigma)\mathrm{d}\sigma} \mathrm{d}s.\end{aligned}$$

Hieraus folgt die Behauptung für $t > t_0$. Der Fall $t < t_0$ wird analog behandelt. □

SATZ I.1.8 (Satz von Picard-Lindelöf). *Sei $f \in C(I \times U, \mathbb{R}^n)$ auf $I \times U$ bzgl. U Lipschitz-stetig. Dann gibt es zu jedem $(t_0, y_0) \in I \times U$ ein $\varepsilon > 0$, so dass das AWP*

$$y' = f(t, y) \text{ auf } (t_0 - \varepsilon, t_0 + \varepsilon), \quad y(t_0) = y_0$$

eine eindeutige Lösung $y \in C^1((t_0 - \varepsilon, t_0 + \varepsilon), \mathbb{R}^n)$ hat.

BEWEIS. [1, Satz 7.4] und [8, Satz XI.1.7].
Beweisidee: Wähle $\varepsilon_0 > 0$ und $R > 0$ so, dass $[t_0 - \varepsilon_0, t_0 + \varepsilon_0] \times \overline{B(y_0, R)} \subset I \times U$ ist und f auf $[t_0 - \varepsilon_0, t_0 + \varepsilon_0] \times \overline{B(y_0, R)}$ gleichmäßig Lipschitz-stetig ist bzgl. $\overline{B(y_0, R)}$. Bezeichne die Lipschitz-Konstante

mit L. Sei M das Maximum von $\|f\|$ auf $[t_0-\varepsilon_0, t_0+\varepsilon_0] \times \overline{B(y_0, R)}$ und setze $\varepsilon = \min\{\varepsilon_0, \frac{R}{M}\}$. Sei $X = C([t_0-\varepsilon, t_0+\varepsilon], \mathbb{R}^n)$ und $K = B_{\|\cdot\|_\infty}(\overline{y}_0, R)$. Dabei bezeichnet $\|\cdot\|_\infty$ die Maximumsnorm auf X und $\overline{y}_0$ die konstante Funktion mit Wert y_0. Definiere die Abbildung $\Phi : X \to X$ durch

$$\Phi y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))\mathrm{d}s$$

für alle $|t-t_0| < \varepsilon$. Dann ist $y \in X$ genau dann ein Fixpunkt von Φ, wenn es das AWP löst. Mit Hilfe der Definition von Φ und der Lipschitz-Stetigkeit von f zeigt man, dass Φ eine Kontraktion auf K ist mit Kontraktionsrate $\kappa = 1 - e^{-L\varepsilon}$. Dabei wird X versehen mit der Norm

$$\|y\|_L = \max_{|t-t_0|\le\varepsilon} \left\|e^{-L|t-t_0|} y(t)\right\| .$$

Dann folgt die Behauptung aus dem Banachschen Fixpunktsatz. □

Bemerkung I.1.9 (Nicht eindeutige und nicht globale Lösbarkeit, Autonomisierung). (1) Auf die Lipschitz-Bedingung kann man in Satz I.1.8 nicht verzichten. Betrachte dazu das autonome AWP

$$y' = \sqrt{|y|}, \quad y(0) = 0.$$

Die Funktion $\sqrt{|y|}$ ist in keiner Umgebung von 0 Lipschitz-stetig. Dieses AWP hat offensichtlich die Lösungen

$$y(t) = 0 \quad \text{und} \quad y(t) = \frac{1}{4}t^2.$$

Für beliebiges $a, b \in \mathbb{R}_+^* \cup \{+\infty\}$ sind weitere Lösungen gegeben durch

$$y_{a,b}(t) = \begin{cases} -\frac{1}{4}(t+a)^2 & \text{für } t \le -a, \\ 0 & \text{für } -a < t < b, \\ \frac{1}{4}(t-b)^2 & \text{für } b \le t. \end{cases}$$

(2) Unter den Voraussetzungen von Satz I.1.8 kann man i.a. nur die lokale Existenz der Lösung des AWP erwarten. Betrachte z.B. das autonome AWP

$$y' = y^2, \quad y(0) = 1.$$

Die eindeutige Lösung lautet

$$y(t) = \frac{1}{1-t}$$

für alle $t < 1$. Sie explodiert für $t \to 1$.
(3) Satz I.1.8 zeigt, dass es nicht immer ratsam ist, eine nicht autonome gDgl in die äquivalente autonome gDgl umzuformen. Für die nicht autonome gDgl benötigen wir nämlich nur Lipschitz-Stetigkeit bzgl. y, für die äquivalente autonome gDgl dagegen Lipschitz-Stetigkeit bzgl. y und t.

SATZ I.1.10 (Globaler Existenz- und Eindeutigkeitssatz). *Sei* $f \in C(I \times U, \mathbb{R}^n)$ *auf* $I \times U$ *bzgl.* U *Lipschitz-stetig. Dann existiert für jedes* $(t_0, y_0) \in I \times U$ *genau eine nicht fortsetzbare Lösung* $y(\cdot; t_0, y_0) \in C^1(I(t_0, y_0), U)$ *des AWP*

$$y' = f(t, y), \quad y(t_0) = y_0. \tag{I.1.2}$$

Das maximale Existenzintervall $I(t_0, y_0)$ *ist offen, d.h.*

$$I(t_0, y_0) = (t^-(t_0, y_0), t^+(t_0, y_0)),$$

und es gilt entweder

$$t^- = t^-(t_0, y_0) = \inf I \quad \textit{bzw.} \quad t^+ = t^+(t_0, y_0) = \sup I$$

oder

$$\lim_{t \to t^\pm \mp 0} \min \left\{ \operatorname{dist}(y(t; t_0, y_0), \partial U), \|y(t; t_0, y_0)\|^{-1} \right\} = 0.$$

Dabei ist $\operatorname{dist}(z, \emptyset) = +\infty$ *und* $\operatorname{dist}(z, A) = \inf\{\|z - y\| : y \in A\}$ *für eine nicht leere abgeschlossene Menge* A.

BEWEIS. [1, Satz 7.6] und [8, Satz XI.1.9].
Beweisidee: Sei $(t_0, y_0) \in I \times U$ beliebig. Wegen Satz I.1.8 gibt es ein $\varepsilon_1 > 0$, so dass (I.1.2) eine eindeutige Lösung u auf $I_1 = [t_0 - \varepsilon_1, t_0 + \varepsilon_1]$ besitzt. Wegen Satz I.1.8 gibt es weiter ein $\varepsilon_2 > 0$, so dass das AWP

$$y' = f(t, y), \quad y(t_0 + \varepsilon_1) = u(t_0 + \varepsilon_1)$$

eine eindeutige Lösung v auf $I_{1,2} = [t_0 + \varepsilon_1 - \varepsilon_2, t_0 + \varepsilon_1 + \varepsilon_2]$ besitzt. Wegen der Eindeutigkeit der Lösung gilt $u = v$ auf $I_1 \cap I_{1,2}$. Folglich ist u_1 mit

$$u_1 = \begin{cases} u & \text{auf } I_1 \\ v & \text{auf } I_{1,2} \end{cases}$$

eine Lösung von (I.1.2), die u echt fortsetzt. Wegen dieses Fortsetzungsargumentes ist folgende Definition sinnvoll

$$t^+ = t^+(t_0, y_0) = \sup\{\beta \in I : \text{(I.1.2) besitzt eine Lösung auf } [t_0, \beta]\},$$
$$t^- = t^-(t_0, y_0) = \inf\{\beta \in I : \text{(I.1.2) besitzt eine Lösung auf } [\beta, t_0]\}.$$

Dann existert genau eine nicht fortsetzbare Lösung $u \in C^1((t^-, t^+), U)$ von (I.1.2). Das maximale Existenzintervall ist offen, da wir sonst das obige Fortsetzungsargument anwenden können.
Nehme nun an, dass $t^+ < \sup I$ und

$$\lim_{t \to t^+ - 0} \min \left\{ \operatorname{dist}(u(t), \partial U), \|u(t)\|^{-1} \right\} > 0$$

ist. Dann kann man zeigen, dass es eine Folge $(t_n)_{n \in \mathbb{N}} \subset I$ mit $\lim_{n \to \infty} t_n = t^+$ und $t_n < t^+$ für alle $n \in \mathbb{N}$ und Zahlen $\varepsilon > 0$, $\delta > 0$ gibt, so dass für alle $n \in \mathbb{N}$ und alle $s \in [0, \min\{\delta, t^+ - t_n\}]$ gilt

$$\|u(t_n + s)\| < \frac{1}{\varepsilon} \quad \text{und} \quad \operatorname{dist}(u(t_n + s), \partial U) > \varepsilon.$$

Hieraus folgt dann, dass es eine Konstante $M > 0$ gibt mit

$$\|u(t) - u(s)\| \leq M\,|t - s|$$

für alle $s, t \in [t^+ - \delta, t^+)$. Daher ist $(u(t'_n))_{n\in\mathbb{N}}$ eine Cauchy-Folge, wenn $(t'_n)_{n\in\mathbb{N}}$ eine beliebige Folge mit $t'_n < t^+$ für alle $n \in \mathbb{N}$ und $\lim\limits_{n\to\infty} t'_n = t^+$ bezeichnet. Hieraus folgt aber, dass die maximale Lösung u von (I.1.2) auf $[t_0, t^+]$ fortgesetzt werden kann im Widerspruch zur Definition von t^+. Ganz analog beweist man die Aussage für t^-. □

Bemerkung I.1.11 (Maximale Trajektorien). Die Bezeichnungen und Voraussetzungen seien wie in Satz I.1.10. Definiere

$$\gamma^+(t_0, y_0) = \left\{u(t; t_0, y_0) : t \in [t_0, t^+(t_0, y_0))\right\},$$
$$\gamma^-(t_0, y_0) = \left\{u(t; t_0, y_0) : t \in (t^-(t_0, y_0), t_0]\right\}.$$

Dann gilt:

(1) Ist $\gamma^\pm$ beschränkt, so ist $t^+ = \sup I$ bzw. $t^- = \inf I$ oder es gilt $\operatorname{dist}(u(t; t_0, y_0), \partial U) \to 0$ für $t \to t^\pm \mp 0$. D.h., die Lösung existiert entweder für alle Zeiten oder sie läuft gegen den Rand von U.
(2) Ist $\gamma^\pm$ in einer kompakten Menge enthalten, so ist $t^+ = \sup I$ bzw. $t^- = \inf I$.

Bemerkung I.1.12 (Linear beschränkte Differentialgleichungen). Die Funktion f sei linear beschränkt, d.h. es gebe $\alpha, \beta \in C(I, \mathbb{R}_+)$ mit

$$\|f(t, y)\| \leq \alpha(t)\,\|y\| + \beta(t)$$

für alle $(t, y) \in I \times U$. Dann folgt aus dem Lemma von Gronwall, Satz I.1.7, dass jede Lösung von $y' = f(t, y)$ beschränkt ist auf beschränkten Intervallen. Ist insbesondere $U = \mathbb{R}^n$, so besitzt das AWP $y' = f(t, y)$, $y(t_0) = y_0$ für alle $t_0 \in I$, $y_0 \in \mathbb{R}^n$ eine eindeutige globale Lösung.

Insbesondere für die Behandlung von Randwertproblemen in Kapitel II ist die stetige und differenzierbare Abhängigkeit der Lösungen von AWPen von den Anfangswerten von fundamentaler Bedeutung. Sei hierzu $f \in C(I \times U, \mathbb{R}^n)$ auf $I \times U$ Lipschitz-stetig bzgl. U. Wegen Satz I.1.8 wissen wir, dass das AWP (I.1.2) für jedes $(t_0, y_0) \in I \times U$ in einer Umgebung von t_0 eine eindeutige Lösung $y = y(\cdot; t_0, y_0)$ besitzt. Wir wollen zeigen, dass $y(\cdot; t_0, y_0)$ stetig und – unter zusätzlichen Voraussetzungen an f – differenzierbar von y_0 abhängt. Dazu benutzen wir zunächst, dass aus dem Beweis von Satz I.1.8 folgt, dass es zu jedem $(t_0, y_0) \in I \times U$ zwei Umgebungen $J = J(t_0) \in \mathcal{U}(t_0)$ und $V = V(y_0) \in \mathcal{U}(y_0)$ mit den folgenden Eigenschaften gibt:

(1) f ist gleichmäßig Lipschitz-stetig auf $J \times V$ bzgl. V.
(2) Zu jedem $y_1 \in V$ besitzt das AWP

$$y' = f(t, y(t)), \quad y(t_0) = y_1$$

eine eindeutige Lösung $y(\cdot; t_0, y_1) \in C^1(J, V)$ auf J.

Daher können wir durch $\varphi(z) = y(\cdot; t_0, z)$ eine Abbildung $\varphi : V \to C^1(J, V)$ definieren. Damit können wir unser Ziel wie folgt formulieren: $\varphi \in C(V, C(J, V))$ bzw. – unter zusätzlichen Bedingungen an f – $\varphi \in C^1(V, C(J, V))$. Dazu nehmen wir zur Vereinfachung an, dass J beschränkt ist.

SATZ I.1.13 (Stetige Abhängigkeit von den Anfangswerten). *Die Funktion φ ist gleichmäßig Lipschitz-stetig auf V:*

$$\|\varphi(z_1) - \varphi(z_2)\|_{C(J,V)} \leq e^{Ld} \|z_1 - z_2\|_{\mathbb{R}^n}$$

für alle $z_1, z_2 \in V$. Dabei ist $d = \sup\{|s - t| : s, t \in J\}$ die Länge von J und L die Lipschitz-Konstante von f.

BEWEIS. [1, Satz 8.3] und [8, Satz XI.3.1].
Beweisidee: Aus der Darstellung

$$\varphi(z)(t) = y(t; t_0, z) = z + \int_{t_0}^{t} f(s, y(s; t_0, z))\mathrm{d}s$$

folgt

$$\varphi(z_2)(t) - \varphi(z_1)(t) = z_2 - z_1 + \int_{t_0}^{t} [f(s, y(s; t_0, z_2)) - f(s, y(s; t_0, z_1))]\,\mathrm{d}s$$

und somit wegen der Lipschitz-Stetigkeit von f

$$\begin{aligned}
&\|\varphi(z_2)(t) - \varphi(z_1)(t)\| \\
&\quad \leq \|z_2 - z_1\| + \left| \int_{t_0}^{t} L \|y(s; t_0, z_2)) - y(s; t_0, z_1))\|\,\mathrm{d}s \right| \\
&\quad = \|z_2 - z_1\| + \left| \int_{t_0}^{t} L \|\varphi(z_2)(s) - \varphi(z_1)(s)\|\,\mathrm{d}s \right|.
\end{aligned}$$

Damit folgt die Behauptung aus dem Lemma von Gronwall, Satz I.1.7. □

SATZ I.1.14 (Differenzierbare Abhängigkeit von den Anfangswerten). *Die Funktion $f(t, y)$ sei zusätzlich auf $J \times V$ stetig differenzierbar bzgl. der Variablen y und $D_y f$ sei Lipschitz-stetig bzgl. V. Dann ist φ stetig differenzierbar:*

$$(D\varphi(z)w)(t) = Z(t; t_0, z)w$$

für alle $t \in J$, $z \in V$, $w \in \mathbb{R}^n$. Dabei ist $Z(\cdot; t_0, z) \in C^1(J, \mathbb{R}^{n\times n})$ die eindeutige Lösung des linearen AWP

$$Z' = D_y f(t, y(t; t_0, z))Z, \quad Z(t_0) = Id_{\mathbb{R}^n}.$$

BEWEIS. [1, Satz 9.2] und [8, Satz XI.3.2].
Beweisidee: Aus der Definition von φ und Z und dem Hauptsatz der

Differential- und Integralrechnung folgt

$$\begin{aligned}
&\varphi(z_2)(t) - \varphi(z_1)(t) - Z(t;t_0,z_1)(z_2-z_1)\\
&\quad = y(t;t_0,z_2) - y(t;t_0,z_1) - Z(t;t_0,z_1)(z_2-z_1)\\
&\quad = \int_{t_0}^{t} [y'(s;t_0,z_2) - y'(s;t_0,z_1) - Z'(s;t_0,z_1)(z_2-z_1)]\,\mathrm{d}s\\
&\quad = \int_{t_0}^{t} [f(s,y(s;t_0,z_2)) - f(s,y(s;t_0,z_1)) -\\
&\qquad\qquad D_y f(s,y(s;t_0,z_1))Z(s;t_0,z_1)(z_2-z_1)]\,\mathrm{d}s.
\end{aligned}$$

Erneute Anwendung des Hauptsatzes der Differential- und Integralrechnung ergibt

$$\begin{aligned}
&f(s,y(s;t_0,z_2)) - f(s,y(s;t_0,z_1))\\
&\qquad - D_y f(s,y(s;t_0,z_1))Z(s;t_0,z_1)(z_2-z_1)\\
&\quad = \int_0^1 [D_y f(s,y(s;t_0,z_1) + \theta(y(s;t_0,z_2) - y(s;t_0,z_1)))\\
&\qquad\qquad \cdot [y(s;t_0,z_2) - y(s;t_0,z_1)]\\
&\qquad\qquad - D_y f(s,y(s;t_0,z_1))Z(s;t_0,z_1)(z_2-z_1)]\,\mathrm{d}\theta\\
&= \int_0^1 D_y f(s,y(s;t_0,z_1) + \theta(y(s;t_0,z_2) - y(s;t_0,z_1)))\\
&\qquad\qquad \cdot [y(s;t_0,z_2) - y(s;t_0,z_1) - Z(s;t_0,z_1)(z_2-z_1)]\,\mathrm{d}\theta\\
&\qquad + \int_0^1 [D_y f(s,y(s;t_0,z_1) + \theta(y(s;t_0,z_2) - y(s;t_0,z_1)))\\
&\qquad\qquad - D_y f(s,y(s;t_0,z_1))]\, Z(s;t_0,z_1)(z_2-z_1)d\theta.
\end{aligned}$$

Aus diesen Gleichungen und der Lipschitz-Stetigkeit von $D_y f$ folgt mit dem Lemma von Gronwall, Satz I.1.7,

$$\lim_{\|z_2-z_1\|\to 0} \frac{1}{\|z_2-z_1\|}\ \|\varphi(z_2) - \varphi(z_1) - Z(\cdot;t_0,z_1)(z_2-z_1)\|_{C(J,V)} = 0$$

und damit die Behauptung. □

Für die Anwendungen in Kapitel II ist es wichtig, eine Abschätzung über das Wachstum der Funktion Z aus Satz I.1.14 zu haben. Hierzu bezeichne $\|\cdot\|_{\mathcal{L}}$ eine beliebige Matrixnorm auf $\mathbb{R}^{n\times n}$.

Satz I.1.15 (Wachstum von Lösungen linearer Differentialgleichungen). *Sei $T \in C([a,b],\mathbb{R}^{n\times n})$ und $k(t) = \|T(t)\|_{\mathcal{L}}$. Dann gilt für die Lösung Y des AWP*

$$Y' = T(t)Y(t), \quad Y(a) = I$$

die Abschätzung

$$\|Y(t) - I\|_{\mathcal{L}} \le \exp\left\{\int_a^t k(s)\mathrm{d}s\right\} - 1.$$

BEWEIS. Offensichtlich gilt für alle $t \in [a,b]$

$$Y(t) = I + \int_a^t T(s)Y(s)\mathrm{d}s.$$

Mit $u(t) = \|Y(t) - I\|_{\mathcal{L}}$, folgt

$$u(t) \leq \int_a^t \|T(s)\|_{\mathcal{L}} \, \|Y(s)\|_{\mathcal{L}} \, \mathrm{d}s \leq \int_a^t k(s)[u(s)+1]\mathrm{d}s.$$

Diese Abschätzung und das Lemma von Gronwall, Satz I.1.7, mit $\alpha(t) = \int_a^t k(s)\mathrm{d}s$ und $\beta(t) = k(t)$ beweisen die Behauptung. □

Die Sätze I.1.13 – I.1.15 liefern sinnvolle Aussagen über das Kurzzeitverhalten von Lösungen des AWP (I.1.2) bei Störungen des Anfangswertes. Für Aussagen über das Langzeitverhalten, d.h. $|t - t_0| \gg 1$, sind sie unbrauchbar, da ihre Abschätzungen Faktoren enthalten, die exponentiell mit $|t - t_0|$ wachsen. Andererseits zeigt das Beispiel $f(y) = -y$ mit der Lösung $y(t; t_0, y_0) = y_0 e^{-(t-t_0)}$, dass u.U. für alle Anfangswerte y_1, y_2 gilt

$$\lim_{t\to\infty} \|y(t; t_0, y_1) - y(t; t_0, y_2)\| = 0.$$

Derartige Aussagen über das Langzeitverhalten von Lösungen von Anfangswertproblemen sind mit den Begriffen *Ljapunov-Stabilität* bzw. *asymptotische Stabilität* verbunden. Aus Zeitgründen können wir hier auf diesen Problemkreis nicht näher eingehen und verweisen stattdessen auf [1, §§13, 15] und [8, §XI.4].

Wir schließen diesen einführenden, theoretischen Paragraphen mit einem Regularitätssatz für Lösungen von gDglen, der mit der Kettenregel durch vollständige Induktion folgt. Aufgrund dieses Regularitätssatzes unterscheiden sich gDglen grundsätzlich von partiellen Differentialgleichungen.

SATZ I.1.16 (Regularitätssatz). *Sei $f \in C^k(I \times U, \mathbb{R}^n)$. Dann gilt für jede Lösung y der gDgl $y' = f(t,y)$ auf ihrem Definitionsbereich J die Regularitätsaussage $y \in C^{k+1}(J, \mathbb{R}^n)$.*

I.2. Einschrittverfahren

Im Folgenden wollen wir die Lösung des AWP (I.1.2) (S. 10) mit Lipschitz-stetigem f numerisch approximieren. Dazu betrachten wir Gitterpunkte $t_0 < t_1 < \ldots$ und bezeichnen die numerische Approximation mit η_i oder $\eta(t_i, h_i)$. Dabei bezeichnet $h_i = t_i - t_{i-1}$ die Schrittweite. Häufig ist sie konstant, d.h., die Gitterpunkte sind äquidistant. Für gegebenes $t > t_0$ sind wir an der Konvergenz(-geschwindigkeit) von $\eta(t, (t-t_0)/n)$ für $n \to \infty$ gegen die exakte Lösung $y(t)$ interessiert. Dazu bezeichnet $\|\cdot\|$ eine beliebige Norm auf $\mathbb{R}^n$ und $\|\cdot\|_{\mathcal{L}}$ die zugehörige Matrixnorm.

Bevor wir eine allgemeine Definition von Einschrittverfahren geben, wollen wir drei der gebräuchlichsten Verfahren, an denen sich die prinzipiellen Strukturen besonders gut studieren lassen, aus geometrischen und analytischen Überlegungen herleiten.

Angenommen wir kennen die Lösung y des AWP (I.1.2) (S. 10) im Punkte t. Wegen $y'(t) = f(t, y(t))$ gibt $f(t, y(t))$ die Steigung der Tangente an die Lösungskurve im Punkt t an. Daher sollte $y(t)+hf(t, y(t))$ eine passable Näherung an $y(t + h)$ sein. Diese Überlegung führt auf:

DEFINITION I.2.1 (Explizites Euler-Verfahren). Das *explizite Euler-Verfahren* ist gegeben durch:

$$\begin{aligned} \eta_0 &= y_0 \\ \eta_{i+1} &= \eta_i + hf(t_i, \eta_i) \\ t_{i+1} &= t_i + h. \end{aligned}$$

Genauso gut hätten wir eine Näherung η_{i+1} für $y(t + h)$ aus der Bedingung herleiten können, dass die Tangente durch die Lösungskurve im Punkte $t_{i+1} = t_i + h$ durch den Punkt (t_i, y_i) gehen soll, d.h.

$$y_i = \eta_{i+1} - hf(t_i + h, \eta_{i+1}).$$

Dies führt auf:

DEFINITION I.2.2 (Implizites Euler-Verfahren). Das *implizite Euler-Verfahren* ist gegeben durch:

$$\begin{aligned} \eta_0 &= y_0 \\ \eta_{i+1} &= \eta_i + hf(t_{i+1}, \eta_{i+1}) \\ t_{i+1} &= t_i + h. \end{aligned}$$

Im Gegensatz zu I.2.1 muss man in I.2.2 in jedem Schritt ein nichtlineares Gleichungssystem der Form $u = z + hf(t, u)$ mit bekannten Größen z und t lösen. Falls f stetig differenzierbar ist, folgt aus dem Satz über implizite Funktionen, dass dieses Gleichungssystem für hinreichend kleines h eine Lösung in der Nähe von z hat. Auf die praktische Berechnung dieser Lösung gehen wir in §I.4 (S. 28) ein.

Das dritte Verfahren erhalten wir durch folgende Überlegung. Für die exakte Lösung von (I.1.2) (S. 10) gilt

$$y(t + h) = y(t) + \int_t^{t+h} f(s, y(s))\mathrm{d}s. \tag{I.2.1}$$

Das Integral in (I.2.1) wird beim expliziten Euler-Verfahren durch

$$\int_t^{t+h} f(s, y(s))\mathrm{d}s \approx hf(t, y(t))$$

approximiert und beim impliziten Euler-Verfahren durch

$$\int_t^{t+h} f(s, y(s))\mathrm{d}s \approx hf(t + h, y(t + h)).$$

Ebenso könnten wir es aber auch durch die Trapezregel

$$\int_t^{t+h} f(s, y(s))\mathrm{d}s \approx \frac{h}{2}\left[f(t, y(t)) + f(t+h, y(t+h))\right]$$

annähern. Dies führt zu:

DEFINITION I.2.3 (Trapezregel, Crank-Nicolson-Verfahren). Die *Trapezregel* bzw. das *Verfahren von Crank-Nicolson* ist gegeben durch:

$$\begin{aligned}\eta_0 &= y_0\\ \eta_{i+1} &= \eta_i + \frac{h}{2}\left[f(t_i, \eta_i) + f(t_{i+1}, \eta_{i+1})\right]\\ t_{i+1} &= t_i + h.\end{aligned}$$

Bei den drei vorgestellten Verfahren wird die Näherung η_{i+1} für $y(t_{i+1})$ ausschließlich durch die letzte Näherung η_i für $y(t_i)$ bestimmt. Daher spricht man von einem *Einschrittverfahren* kurz *ESV*. Die allgemeine Definition lautet:

DEFINITION I.2.4 (Einschrittverfahren, Verfahrensfunktion). Ein allgemeines *Einschrittverfahren* kurz *ESV* zur Lösung des AWP (I.1.2) (S. 10) hat die Form

$$\begin{aligned}\eta_0 &= y_0\\ \eta_{i+1} &= \eta_i + h\Phi(t_i, \eta_i, h; f)\\ t_{i+1} &= t_i + h.\end{aligned}$$

Die Funktion Φ heißt die *Verfahrensfunktion* des ESV.

Der Einfachheit halber lassen wir im Folgenden stets das Argument f bei der Verfahrensfunktion fort.

Für das explizite Euler-Verfahren ist offensichtlich

$$\Phi(t, y, h) = f(t, y).$$

Die Verfahrensfunktionen für das implizite Euler-Verfahren und die Trapezregel sind komplizierter. Aus der Definition der Verfahren und dem Satz über implizite Funktionen erhalten wir für hinreichend kleines h für das implizite Euler-Verfahren

$$\Phi(t, y, h) = \frac{1}{h}\left[\left[Id - hf(t+h, \cdot)\right]^{-1} y - y\right]$$

und für die Trapezregel

$$\Phi(t, y, h) = \frac{1}{h}\left[\left[Id - \frac{h}{2}f(t+h, \cdot)\right]^{-1}\left[y + \frac{h}{2}f(t, y)\right] - y\right],$$

wobei g^{-1} die Umkehrfunktion der Funktion g bezeichnet.

An dieser Stelle sei betont, dass die Verfahrensfunktion der theoretischen Analyse der Verfahren dient. Für praktische Rechnungen arbeitet man mit den Definitionen I.2.1 – I.2.3.

Die Verfahrensfunktion eines ESV ist gleich dem Differenzenquotienten der Näherungslösung. Für vernünftige Verfahren sollte sie den Differenzenquotienten der exakten Lösung möglichst gut approximieren. Dies führt auf:

DEFINITION I.2.5 (Lokaler Verfahrensfehler, Konsistenz, Ordnung). Bezeichne für $x \in [a,b]$ und $y \in \mathbb{R}^n$ mit z die Lösung des AWP

$$z' = f(t, z(t)), \quad z(x) = y.$$

Dann heißt die Funktion

$$\tau(x,y,h) = \frac{1}{h}\{z(x+h) - y\} - \Phi(x,y,h), \quad h > 0,$$

der *lokale Verfahrensfehler* des ESV mit Verfahrensfunktion Φ. Das Verfahren heißt *konsistent*, falls gilt

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

für alle $x \in [a,b]$, $y \in \mathbb{R}^n$ und $f \in C^1((a,b)\times\mathbb{R}^n, \mathbb{R}^n)$. Es hat die *Ordnung* p, falls gilt

$$\limsup_{h\to 0} h^{-p} \|\tau(x,y,h)\| < \infty$$

für alle $x \in [a,b]$, $y \in \mathbb{R}^n$ und $f \in C^{[p]}((a,b)\times\mathbb{R}^n, \mathbb{R}^n)$.

BEMERKUNG I.2.6 (Fehler eines Schrittes und lokaler Verfahrensfehler). Setzen wir $\eta_0 = y$, ist

$$\tau(x,y,h) = \frac{1}{h}\left[z(x+h) - \eta_1\right],$$

d.h., *der lokale Verfahrensfehler ist der durch die Schrittweite dividierte Fehler eines Schrittes des ESV*. Die Division durch die Schrittweite wird durch folgende heuristische Überlegung motiviert: Wir wollen ein AWP auf dem Intervall $[t_0, t_1]$ lösen und führen hierzu n Schritte des ESV mit Schrittweite $h = \frac{t_1 - t_0}{n}$ aus. Dann erwarten wir, dass der Fehler zur Endzeit t_1 das n-fache des Fehlers eines einzelnen Schrittes ist. Der Einzelfehler wird also mit dem Faktor $n = \frac{h}{t_1 - t_0}$ verstärkt. Satz I.3.1 (S. 27) zeigt, dass diese heuristische Überlegung korrekt ist.

Für das explizite Euler-Verfahren erhalten wir durch Taylor-Entwicklung wegen $f(x,y) = z'(x)$

$$\begin{aligned}\tau(x,y,h) &= \frac{1}{h}\{z(x+h) - y\} - f(x,y)\\ &= z'(x) + \frac{1}{2}hz''(x) + \mathrm{O}(h^2) - f(x,y)\\ &= \frac{1}{2}hz''(x) + \mathrm{O}(h^2).\end{aligned}$$

Also hat das explizite Euler-Verfahren die Ordnung 1.
Für das implizite Euler-Verfahren und die Trapezregel bestimmen wir

die Ordnung am besten durch die Analyse des Fehlers eines einzelnen Schrittes. Für das implizite Euler-Verfahren erhalten wir wegen $\eta_0 = y = z(x)$

$$\begin{aligned} z(x+h) - \eta_1 &= z(x+h) - \eta_0 - (\eta_1 - \eta_0) \\ &= \int_x^{x+h} z'(t)\mathrm{d}t - hf(x+h, \eta_1) \\ &= \int_x^{x+h} f(t, z(t))\mathrm{d}t - hf(x+h, \eta_1) \\ &= \int_x^{x+h} \left[f(t, z(t)) - f(x+h, z(x+h))\right] \mathrm{d}t \\ &\quad + h\left[f(x+h, z(x+h)) - f(x+h, \eta_1)\right]. \end{aligned}$$

Bezeichnet L die Lipschitz-Konstante von f und K eine Schranke für $\|D_y f\|_{\mathcal{L}}$, so folgt hieraus

$$\|z(x+h) - \eta_1\| \leq Kh^2 + hL\|z(x+h) - \eta_1\|.$$

Also ist für hinreichend kleines h

$$\|z(x+h) - \eta_1\| \leq \frac{Kh^2}{1-hL}.$$

Also hat das implizite Euler-Verfahren ebenfalls die Ordnung 1.
Für die Trapezregel erhalten wir analog

$$\begin{aligned} z(x+h) - \eta_1 &= z(x+h) - \eta_0 - \frac{h}{2}\left[f(x, \eta_0) + f(x+h, \eta_1)\right] \\ &= \int_x^{x+h} f(t, z(t))\mathrm{d}t - \frac{h}{2}\left[f(x, z(x)) + f(x+h, z(x+h))\right] \\ &\quad + \frac{h}{2}\left[f(x+h, z(x+h)) - f(x+h, \eta_1)\right]. \end{aligned}$$

Bezeichnet $p(t)$ das lineare Interpolationspolynom von $f(t, z(t))$ in den Punkten x und $x+h$, ist

$$\frac{h}{2}\left[f(x, z(x)) + f(x+h, z(x+h))\right] = \int_x^{x+h} p(t)\mathrm{d}t$$

und daher

$$\begin{aligned} &\left\| \int_x^{x+h} f(t, z(t))\mathrm{d}t - \frac{h}{2}\left[f(x, z(x)) + f(x+h, z(x+h))\right] \right\| \\ &= \left\| \int_x^{x+h} \left[f(t, z(t)) - p(t)\right] \mathrm{d}t \right\| \\ &\leq \frac{1}{2} M h^3, \end{aligned}$$

wobei M eine obere Schranke für $\|D^2 f\|_{\mathcal{L}^2}$ ist. Hieraus folgt wie beim impliziten Euler-Verfahren

$$\|z(x+h) - \eta_1\| \leq \frac{\frac{1}{2}Mh^3}{1 - \frac{1}{2}hL}.$$

Also hat die Trapezregel die Ordnung 2.

Beispiel I.2.7 (Schwingung). Wir führen jeweils 100 Schritte der beiden Euler-Verfahren und der Trapezregel für das AWP

$$y' = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} y, \quad y(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

mit $h = 0.13$ aus. Die exakte Lösung lautet $y(t) = (\cos t, \sin t)$ mit $0 \leq t \leq 13$. Abbildung I.2.1 zeigt die entsprechenden Lösungskurven. Das explizite Euler-Verfahren explodiert. Dies ist eine Folge seiner fehlenden Stabilität (s. §I.8). Das implizite Euler-Verfahren dagegen dämpft die Lösung zu stark. Die Trapezregel schließlich liefert eine Lösungskurve, die auf der exakten Lösungskurve, dem Kreis um den Ursprung mit Radius 1, liegt.

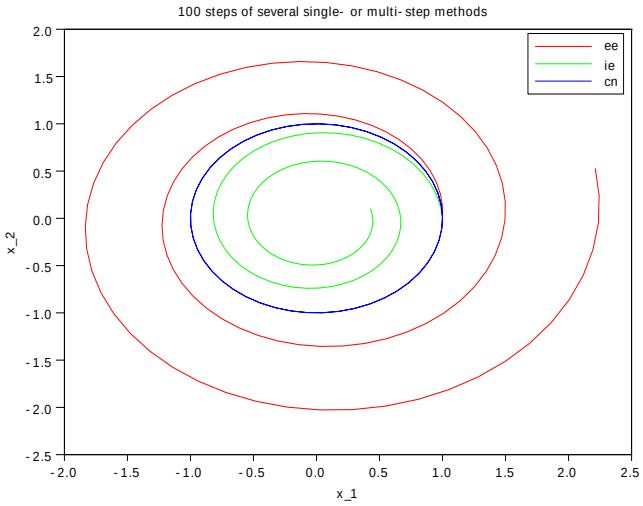

Abbildung I.2.1. Lösungskurven der beiden Euler-Verfahren und der Trapezregel für das AWP aus Beispiel I.2.7

Beispiel I.2.8 (Räuber-Beute-Modell). Wir führen jeweils 1000 Schritte der Euler-Verfahren und der Trapezregel für das AWP

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 4x - 8xy \\ -0.3y + 0.6xy \end{pmatrix}, \quad \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} 0.9 \\ 0.1 \end{pmatrix}$$

mit $h = 0.01$ aus. Dieses AWP beschreibt ein Räuber-Beute-Modell. Die exakte Lösung ist eine geschlossene Kurve [8, Beispiel XI.2.2]. Abbildung I.2.2 zeigt die Lösungskurven der drei Verfahren. Das explizite Euler-Verfahren explodiert. Das implizite Euler-Verfahren dämpft die Lösung zu stark. Die Trapezregel dagegen liefert qualitativ gute Ergebnisse.

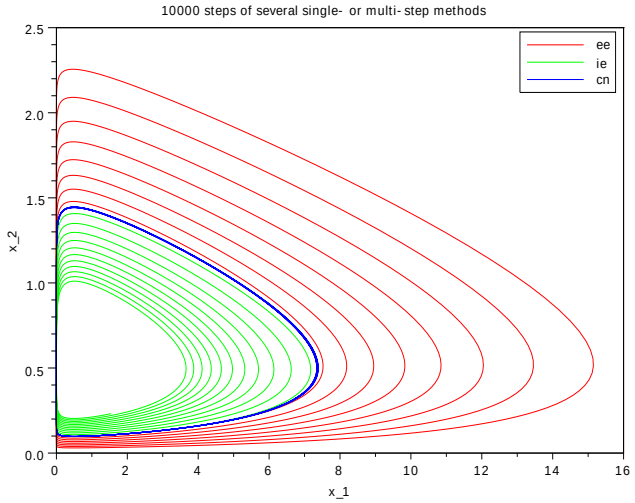

ABBILDUNG I.2.2. Lösungskurven der beiden Euler-Verfahren und der Trapezregel für das AWP aus Beispiel I.2.8

Die bisher betrachteten Verfahren sind Spezialfälle einer größeren Verfahrensklasse, den sog. *Runge-Kutta-Verfahren*. Um diese Verfahren zu motivieren betrachten wir die beiden Euler Verfahren unter einem anderen Blickwinkel. Seien dazu η_i^{EE} und η_i^{IE} die mit dem expliziten und dem impliziten Euler-Verfahren berechneten Näherungen für die Lösung des AWP (I.1.2) (S. 10). Sei $u \in \mathbb{P}_1$ das lineare Polynom, das in den Punkten t_i und t_{i+1} die Werte η_i^{EE} und η_{i+1}^{EE} annimmt. Dann gilt offensichtlich

$$\begin{aligned} \eta_i^{EE} &= u(t_i) \\ \eta_{i+1}^{EE} &= u(t_i + h) \\ u'(t_i) &= \frac{1}{h}\left\{\eta_{i+1}^{EE} - \eta_i^{EE}\right\} = f(t_i, \eta_i^{EE}) \\ &= f(t_i, u(t_i)). \end{aligned}$$

Analog sei $v \in \mathbb{P}_1$ das lineare Polynom, das in den Punkten t_i und t_{i+1}, die Werte η_i^{IE} und η_{i+1}^{IE} annimmt. Dann gilt offensichtlich

$$\begin{aligned} \eta_i^{IE} &= v(t_i) \\ \eta_{i+1}^{IE} &= v(t_i + h) \\ v'(t_i + h) = \frac{1}{h}\left\{\eta_{i+1}^{IE} - \eta_i^{IE}\right\} &= f(t_i + h, \eta_{i+1}^{IE}) \\ &= f(t_i + h, v(t_i + h)). \end{aligned}$$

Die beiden Euler-Verfahren haben also folgende allgemeine Struktur: Wähle $r \geq 1$ Punkte $0 \leq c_1 < \ldots < c_r \leq 1$, bestimme das Polynom $w \in \mathbb{P}_r$ mit

(1) $w(t_i) = \eta_i$,
(2) $w'(t_i + c_\ell h) = f(t_i + c_\ell h, w(t_i + c_\ell h)),\ 1 \leq \ell \leq r$,

und setze

$$\eta_{i+1} = w(t_i + h).$$

Die Bedingung (2) heißt auch *Kollokationsbedingung*. Bei den beiden Euler-Verfahren ist $r = 1$. Beim expliziten Euler-Verfahren ist $c_1 = 0$; beim impliziten Euler-Verfahren ist $c_1 = 1$.
Um aus den Bedingungen (1) und (2) das Polynom w zu bestimmen, setzen wir

$$k_{i,j} = w'(t_i + c_j h), \quad 1 \leq j \leq r$$

und bezeichnen mit $\lambda_1, \ldots, \lambda_r$ die Lagrangeschen Grundpolynome zu den Knoten $c_1, \ldots, c_r$ mit $\lambda_i \in \mathbb{P}_{r-1}$ und $\lambda_i(c_j) = \delta_{ij}$ für $1 \leq i, j \leq r$. Dann gilt

$$w'(t_i + \theta h) = \sum_{j=1}^{r} k_{i,j} \lambda_j(\theta).$$

Durch Integration folgt für $1 \leq \ell \leq r$ mit $a_{\ell j} = \displaystyle\int_0^{c_\ell} \lambda_j(\theta) \mathrm{d}\theta$

$$\begin{aligned} w(t_i + c_\ell h) &= w(t_i) + h \int_0^{c_\ell} w'(t_i + \theta h) \mathrm{d}\theta \\ &= \eta_i + h \sum_{j=1}^{r} k_{i,j} \int_0^{c_\ell} \lambda_j(\theta) \mathrm{d}\theta \\ &= \eta_i + h \sum_{j=1}^{r} k_{i,j} a_{\ell j} \end{aligned}$$

und mit $b_j = \int_0^1 \lambda_j(\theta)\mathrm{d}\theta$

$$\begin{aligned}\eta_{i+1} = w(t_i + h) &= w(t_i) + h\int_0^1 w'(t_i + \theta h)d\theta \\ &= \eta_i + h\sum_{j=1}^{r} k_{i,j}\int_0^1 \lambda_j(\theta)\mathrm{d}\theta \\ &= \eta_i + h\sum_{j=1}^{r} k_{i,j}b_j.\end{aligned}$$

Einsetzen der Darstellung von $w(t_i + c_\ell h)$ in die Kollokationsbedingungen (2) liefert schließlich die äquivalenten Bedingungen

$$\begin{aligned}k_{i,\ell} = w'(t_i + c_\ell h) &= f(t_i + c_\ell h, w(t_i + c_\ell h)) \\ &= f\left(t_i + c_\ell h, \eta_i + h\sum_{j=1}^{r} k_{i,j}a_{\ell j}\right).\end{aligned}$$

Insgesamt erhalten wir somit die folgende Verfahrensvorschrift:

Definition I.2.9 (Runge-Kutta-Verfahren). Ein *r-stufiges Runge-Kutta-Verfahren* ist gegeben durch:

$$\begin{aligned}\eta_0 &= y_0 \\ k_{i,\ell} &= f\left(t_i + c_\ell h, \eta_i + h\sum_{j=1}^{r} a_{\ell j}k_{i,j}\right), \quad 1 \le \ell \le r, \\ \eta_{i+1} &= \eta_i + h\sum_{j=1}^{r} b_j k_{i,j} \\ t_{i+1} &= t_i + h\end{aligned}$$

mit $0 \le c_1 < \ldots < c_r \le 1$.
Der Übersichtlichkeit halber fasst man die Zahlen c_k, $a_{\ell j}$, b_k in einem Schema der Form

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \ldots & a_{1r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_r & a_{r1} & a_{r2} & \ldots & a_{rr} \\ \hline & b_1 & b_2 & \ldots & b_r \end{array}$$

zusammen. Außerdem definiert man die $r \times r$ Matrizen A und C und die r-Vektoren b und e durch

$$A = \begin{pmatrix} a_{11} & \dots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \dots & a_{rr} \end{pmatrix}, \quad C = \begin{pmatrix} c_1 & & 0 \\ & \ddots & \\ 0 & & c_r \end{pmatrix},$$

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_r \end{pmatrix}, \qquad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Ein Runge-Kutta-Verfahren heißt:

- *explizit*, wenn A eine strikte untere Dreiecksmatrix ist,
- *diagonal-implizit*, wenn A eine untere Dreiecksmatrix mit nicht verschwindender Diagonale ist,
- *implizit*, wenn A eine allgemeine Matrix ist,
- *linear-implizit* oder *stark diagonal-implizit*, wenn es diagonal implizit ist und die Diagonalelemente von A alle gleich sind.

Bei einem diagonal-impliziten Verfahren müssen in jedem Schritt r nichtlineare Gleichungssysteme mit n Unbekannten gelöst werden; bei einem impliziten Verfahren ist es ein nichtlineares Gleichungssystem mit rn Unbekannten pro Schritt. Eine besonders effiziente numerische Lösung der nichtlinearen Gleichungssysteme erlauben die linear-impliziten Runge-Kutta-Verfahren.

Bemerkung I.2.10 (Alternative Darstellung, Invarianz unter Autonomisierung). (1) Eine äquivalente Darstellung von Runge-Kutta-Verfahren erhält man durch Einführen der Größen

$$\eta_{i,\ell} = \eta_i + h \sum_{j=1}^{r} a_{\ell j} k_{i,j}$$

Damit lautet das Verfahren

$$\begin{aligned} \eta_0 &= y_0 \\ \eta_{i,\ell} &= \eta_i + h \sum_{j=1}^{r} a_{\ell j} f(t_i + c_j h, \eta_{i,j}), \quad 1 \le \ell \le r, \\ \eta_{i+1} &= \eta_i + h \sum_{j=1}^{r} b_j f(t_i + c_j h, \eta_{i,j}) \\ t_{i+1} &= t_i + h \end{aligned}$$

Diese Darstellung ist für die Ordnungsberechnung von Vorteil.
(2) Eine wünschenswerte Eigenschaft eines Runge-Kutta-Verfahrens ist seine Invarianz unter Autonomisierung: Das Verfahren angewandt auf das äquivalente autonome AWP sollte das gleiche Ergebnis liefern wie

das Verfahren angewandt auf das ursprüngliche AWP. Wie man leicht nachrechnet ist dies der Fall, falls für alle $1 \leq k \leq r$ gilt

$$Ae = Ce \quad \text{oder äquivalent} \quad c_k = \sum_{j=1}^{r} a_{kj}.$$

Die Verfahrensfunktion eines r-stufigen Runge-Kutta-Verfahren ist gegeben durch

$$\Phi(t, y, h) = \sum_{j=1}^{r} b_j f(t + c_j h, u_j(h)) \tag{I.2.2}$$

mit

$$u_\ell(h) = y + h \sum_{j=1}^{r} a_{\ell j} f(t + c_j h, u_j(h)), \quad 1 \leq \ell \leq r. \tag{I.2.3}$$

Alle bisher betrachteten Verfahren sind Runge-Kutta-Verfahren. Dem expliziten Euler-Verfahren entspricht das Schema

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \quad r = 1.$$

Dem impliziten Euler-Verfahren entspricht das Schema

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \quad r = 1.$$

Der Trapezregel entspricht schließlich das Schema

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

Für $r = 2$ und das Schema

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

erhalten wir das sog. *modifizierte Euler-Verfahren*:

$$\begin{aligned} \eta_0 &= y_0 \\ \eta_{i+1} &= \eta_i + hf\left(t_i + \frac{1}{2}h, \eta_i + \frac{1}{2}hf(t_i, \eta_i)\right) \\ t_{i+1} &= t_i + h. \end{aligned}$$

Für $r = 2$ und das Schema

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

erhalten wir das *Verfahren von Heun*, auch *modifizierte Trapezregel* genannt:

$$\begin{aligned}\eta_0 &= y_0\\ \eta_{i+1} &= \eta_i + \frac{h}{2} f(t_i, \eta_i) + \frac{h}{2} f\left(t_i + h, \eta_i + h f(t_i, \eta_i)\right)\\ t_{i+1} &= t_i + h.\end{aligned}$$

Das sog. *klassische Runge-Kutta-Verfahren* ist schließlich gegeben durch das Schema

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0\\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0\\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0\\ 1 & 0 & 0 & 1 & 0\\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array} \qquad r = 4.$$

Die entsprechende Verfahrensvorschrift lautet:

$$\begin{aligned}\eta_0 &= y_0\\ \eta_{i,1} &= \eta_i\\ \eta_{i,2} &= \eta_i + \frac{h}{2} f(t_i, \eta_{i,1})\\ \eta_{i,3} &= \eta_i + \frac{h}{2} f\left(t_i + \frac{h}{2}, \eta_{i,2}\right)\\ \eta_{i,4} &= \eta_i + h f\left(t_i + \frac{h}{2}, \eta_{i,3}\right)\\ \eta_{i+1} &= \eta_i + \frac{h}{6}\left\{ f\left(t_i, \eta_{i,1}\right) + 2f\left(t_i + \frac{h}{2}, \eta_{i,2}\right)\right.\\ &\qquad \left. + 2f\left(t_i + \frac{h}{2}, \eta_{i,3}\right) + f\left(t_i + h, \eta_{i,4}\right)\right\}\\ t_{i+1} &= t_i + h.\end{aligned}$$

Die Bestimmung der Ordnung eines Runge-Kutta-Verfahrens erfolgt durch Taylor-Entwicklung von (I.2.2), (I.2.3). So folgt z. B. aus (I.2.2), (I.2.3):

$$u_\ell(h) = u_\ell + h\dot{u}_\ell + \mathrm{O}(h^2), \quad 1 \le \ell \le r$$

mit

$$u_\ell = y, \quad 1 \le \ell \le r$$

und

$$\dot{u}_\ell = \sum_{j=1}^{r} a_{\ell j} f(t, y), \quad 1 \le \ell \le r,$$

sowie

$$\Phi(t, y, h) = \Phi_0(t, y) + \Phi_1(t, y)h + \mathrm{O}(h^2)$$

mit

$$\Phi_0(t,y) = \sum_{j=1}^{r} b_j f(t,y) = b^T e z'(t)$$

und

$$\begin{aligned}\Phi_1(t,y) &= \sum_{j=1}^{r} b_j \left\{ c_j f_t(t,y) + f_y(t,y) f(t,y) \sum_{\ell=1}^{r} a_{j\ell} \right\} \\ &= b^T C e f_t(t,y) + b^T A e f_y(t,y) f(t,y).\end{aligned}$$

Hieraus folgt für den lokalen Verfahrensfehler:

$$\begin{aligned}\tau(x,y,h) &= (1 - b^T e) z'(x) \\ &\quad + h \left[\frac{1}{2} z''(x) - b^T C e f_x(x,y) - b^T A e f_y(x,y) f(x,y) \right] \\ &\quad + \mathrm{O}(h^2) \\ &= (1 - b^T e) z'(x) \\ &\quad + h \left[\left(\frac{1}{2} - b^T C e \right) f_x(x,y) + \left(\frac{1}{2} - b^T A e \right) f_y(x,y) f(x,y) \right] \\ &\quad + \mathrm{O}(h^2).\end{aligned}$$

Also hat das Verfahren die Ordnung 1, falls gilt

$$b^T e = 1$$

und die Ordnung 2, falls zusätzlich gilt

$$b^T C e = b^T A e = \frac{1}{2}.$$

Ganz analog beweist man den folgenden Satz [2, Lemme 5.10, 5.11].

SATZ I.2.11 (Ordnung eines Runge-Kutta-Verfahrens). *Ein r-stufiges Runge-Kutta-Verfahren hat die Ordnung:*

$$\begin{aligned}1 &\iff b^T e = 1, \\ 2 &\iff \textit{Ordnung } 1 \textit{ plus } b^T C e = b^T A e = \frac{1}{2}.\end{aligned}$$

Falls $Ae = Ce$ ist, hat es die Ordnung

$$\begin{aligned}3 &\iff \textit{Ordnung } 2 \textit{ plus } b^T C^2 e = \frac{1}{3}, b^T A C e = \frac{1}{6}, \\ 4 &\iff \textit{Ordnung } 3 \textit{ plus} \\ &\qquad b^T C^3 e = \frac{1}{4}, b^T A C^2 e = \frac{1}{12}, b^T A^2 C e = \frac{1}{24}, b^T C A C e = \frac{1}{8}.\end{aligned}$$

Aus Satz I.2.11 folgt zusammenfassend:

KOROLLAR I.2.12 (Ordnung einiger wichtiger Runge-Kutta-Verfahren). *(1) Das explizite und das implizite Euler-Verfahren haben die Ordnung* 1.
(2) Die Trapezregel, das modifizierte Euler-Verfahren und die modifizierte Trapezregel haben die Ordnung 2.
(3) Das klassische Runge-Kutta-Verfahren hat die Ordnung 4.

I.3. Konvergenz von Einschrittverfahren

In diesem Paragraphen sei stets f Lipschitz-stetig auf $(a,b)\times\mathbb{R}^n$ bzgl. $\mathbb{R}^n$, $t_0\in(a,b)$, $y_0\in\mathbb{R}^n$ und $y=y(\cdot;t_0,y_0)$ die nicht fortsetzbare Lösung des AWP (I.1.2) (S. 10). Wir wollen die Konvergenz eines allgemeinen ESV mit Verfahrensfunktion Φ untersuchen. Dazu bezeichnen wir für $t\geq t_0$ mit $\eta(t;h)=\eta_i$ falls $t=t_0+ih$ die numerische Näherungslösung und mit $e(t;h)=\eta(t;h)-y(t)$ den *globalen Diskretisierungsfehler*. Der folgende Satz zeigt, dass Verfahren der Ordnung p, $p>0$, konvergent sind und dass für $t\geq t_0$ die Fehlerabschätzung $\|e(t;h_n)\|=\mathrm{O}(h_n^p)$ mit $h_n=\frac{t-t_0}{n}$ gilt.

SATZ I.3.1 (Globale Fehlerbschätzung für Einschrittverfahren). *Die Funktion f sei gleichmäßig Lipschitz-stetig mit Lipschitz-Konstante L. Weiter gebe es ein $h_0>0$, so dass die Verfahrensfunktion Φ des ESV für alle $t\in[a,b]$, alle $z\in\mathbb{R}^n$ und alle $0<h\leq h_0$ die Ordnungsbedingung*

$$\|\tau(t,z,h)\|\leq Kh^p$$

erfüllt. Dann gilt für den globalen Diskretisierungsfehler des ESV für alle $t\geq t_0$ und alle n mit $h_n\leq h_0$ die Abschätzung

$$\|e(t;h_n)\|\leq h_n^p K\frac{e^{L(t-t_0)}-1}{L}.$$

BEWEIS. Offensichtlich ist $e(t_0)=0$. Für $i\geq 0$ bezeichne z_i die Lösung des AWP

$$z'=f(t,z(t)),\quad z(t_i)=\eta_i.$$

Dann ist

$$\begin{aligned}\|e(t_{i+1})\| &= \|\eta_{i+1}-y(t_{i+1})\|\\ &\leq \|\eta_{i+1}-z_i(t_{i+1})\|+\|z_i(t_{i+1})-y(t_{i+1})\|\,.\end{aligned}$$

Da $\eta_{i+1}-z_i(t_{i+1})=h\tau(t_i,\eta_i,h)$ ist, folgt aus der Ordnungsbedingung

$$\|\eta_{i+1}-z_i(t_{i+1})\|\leq Kh^{p+1}.$$

Da y auch das AWP

$$x'=f(t,x(t)),\quad x(t_i)=y(t_i)$$

löst, folgt aus Satz I.1.13 (S. 12)

$$\|z_i(t_{i+1})-y(t_{i+1})\|\leq e^{Lh}\,\|\eta_i-y(t_i)\|=e^{Lh}\,\|e(t_i)\|\,.$$

Also gilt für alle $i \geq 0$

$$\|e(t_{i+1})\| \leq e^{Lh} \|e(t_i)\| + Kh^{p+1}.$$

Hieraus folgt durch Induktion für alle $i \geq 0$

$$\|e(t_i)\| \leq e^{iLh} \|e(t_0)\| + Kh^{p+1} \sum_{j=0}^{i-1} e^{jLh} = Kh^{p+1} \frac{e^{iLh} - 1}{e^{Lh} - 1}.$$

Wegen $e^{Lh} - 1 \geq Lh$ und $e^{iLh} = e^{(t_i - t_0)L}$ folgt hieraus die Behauptung. □

Bemerkung I.3.2. (1) Die Einschränkung der Schrittweite ist für implizite Verfahren erforderlich, da die Lösbarkeit der auftretenden nichtlinearen Probleme nur für hinreichend kleine Schrittweiten gewährleistet ist.
(2) Die Voraussetzungen von Satz I.3.1 können dahin gehend abgeschwächt werden, dass die Funktion f nur Lipschitz-stetig sein muss in einem Streifen um die Lösung von (I.1.2) (S. 10) und dass die Ordnungsbedingung für die Verfahrensfunktion Φ auch nur in einem solchen Streifen gelten muss.

Der folgende Satz ist für eine effiziente Implementierung der ESV von Bedeutung. Wir verzichten hier auf den sehr technischen Beweis und verweisen stattdessen auf [5, S. 176 ff].

Satz I.3.3 (Asymptotische Fehlerentwicklung für Einschrittverfahren). *Sei $f \in C^{K+2}((a,b) \times \mathbb{R}^n, \mathbb{R}^n)$ und $\eta(t;h)$ eine mit einem ESV der Ordnung $p \leq K$ gewonnene Näherung für die Lösung $y(t)$ des AWP (I.1.2) (S. 10). Dann besitzt $\eta(t;h)$ die asymptotische Entwicklung*

$$\eta(t;h) = y(t) + e_p(t)h^p + e_{p+1}(t)h^{p+1} + \ldots + \\ + e_K(t)h^K + E_{K+1}(t;h)h^{K+1}$$

für alle $t \in [x_0, b)$, $h = \frac{t-t_0}{n}$ mit $e_k(t_0) = 0$, $p \leq k \leq K$. Die Funktionen e_k, $k \leq p \leq K$, sind stetig und von h unabhängig. Die Funktion $E_{K+1}(t,.)$ ist für jedes t beschränkt.

I.4. Implementierung von Einschrittverfahren

In diesem Paragraphen befassen wir uns mit zwei Aspekten der praktischen Implementierung von ESV:

- dem Lösen der nichtlinearen Gleichungssysteme,
- der Schrittweitensteuerung.

Bei impliziten ESV wie dem impliziten Euler-Verfahren oder der Trapezregel müssen in jedem Schritt ein (oder mehrere) nichtlineare Gleichungssysteme gelöst werden. Dies kann mit Hilfe des Newton-Verfahrens geschehen. Die Näherungslösung η_i am Punkt t_i ist dabei ein guter Startwert für das Newton-Verfahren zur Berechnung von η_{i+1}. Wegen der quadratischen Konvergenz des Newton-Verfahrens reichen in

der Regel wenige Iterationen aus. Bei der praktischen Implementierung beschränkt man sich häufig sogar auf eine Newton-Iteration pro Schritt.

Beim Newton-Verfahren muss in jeder Iteration eine Jacobi-Matrix berechnet und ein lineares Gleichungssystem mit N Unbekannten gelöst werden. Da dies häufig recht aufwändig ist, benutzt man in der Praxis als Alternative sog. *Prädiktor-Korrektor-Verfahren*. Dabei berechnet man ausgehend von η_i zunächst mit einem expliziten Verfahren eine Näherung $\eta_{i+1}^{(0)}$ und benutzt diese dann als Startwert für eine Fixpunktiteration angewandt auf das implizite Verfahren, wobei man häufig nur einen Iterationsschritt durchführt. Damit das Gesamtverfahren die gleiche Ordnung hat wie das benutzte implizite Verfahren, sollte das explizite Verfahren die gleiche Ordnung haben wie das implizite. Dieses Vorgehen wird an folgenden zwei Beispielen deutlich.

Beispiel I.4.1 (Implizites Euler-Verfahren mit Prädiktor-Korrektor). Wir betrachten das explizite Euler-Verfahren als Prädiktor und das implizite Euler-Verfahren als Korrektor. Dies liefert

$$\begin{aligned}\eta_{i+1}^{(0)} &= \eta_i + hf(t_i, \eta_i)\\ \eta_{i+1} &= \eta_i + hf(t_i + h, \eta_{i+1}^{(0)})\end{aligned}$$

oder direkt

$$\eta_{i+1} = \eta_i + hf\left(t_i + h, \eta_i + hf(t_i, \eta_i)\right).$$

Dies ist ein explizites zweistufiges Runge-Kutta-Verfahren mit dem Schema

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

Aus Satz I.2.11 (S. 26) folgt, dass es die Ordnung 1 hat.

Beispiel I.4.2 (Trapezregel mit Prädiktor-Korrektor). Wir nehmen wieder das explizite Euler-Verfahren als Prädiktor, aber die Trapezregel als Korrektor. Dies liefert

$$\begin{aligned}\eta_{i+1}^{(0)} &= \eta_i + hf(t_i, \eta_i)\\ \eta_{i+1} &= \eta_i + \frac{h}{2}f(t_i, \eta_i) + \frac{h}{2}f(t_i + h, \eta_{i+1}^{(0)})\end{aligned}$$

oder direkt

$$\eta_{i+1} = \eta_i + \frac{h}{2}f(t_i, \eta_i) + \frac{h}{2}f\left(t_i + h, \eta_i + hf(t_i, \eta_i)\right).$$

Wir erhalten also die modifizierte Trapezregel, die die Ordnung 2 hat.

Wir wenden uns nun dem Problem der Schrittweitensteuerung zu. Ziel ist es, ausgehend vom Punkt t_0 in möglichst wenigen Schritten eine Näherung der Lösung $y(t)$ im Punkt t mit einer gegebenen Genauigkeit

ε zu berechnen. Typischerweise ist $\varepsilon = K \cdot \text{eps}$, wobei eps die Maschinengenauigkeit und K eine obere Schranke für $\|y(t)\|$ ist. Es gibt zwei Strategien zur Schrittweitensteuerung:

- Vergleich verschiedener Schrittweiten,
- Vergleich verschiedener Verfahren.

Zur Darstellung der ersten Strategie bezeichne mit $\eta(t;h)$ die mit einem ESV der Ordnung p und Schrittweite h gewonnene Näherung für $y(t)$. Dann ist unter Vernachlässigung von Termen höherer Ordnung gemäß Satz I.3.3 (S. 28)

$$\begin{aligned} \eta(t;h) &= y(t) + e_p(t)h^p \\ \eta\left(t;\frac{h}{2}\right) &= y(t) + e_p(t)\left(\frac{h}{2}\right)^p \end{aligned}$$

und somit

$$e_p(t) = \frac{1}{h^p(1-2^{-p})}\left\{\eta(t;h) - \eta\left(t;\frac{h}{2}\right)\right\}.$$

Also gilt bei Fortführen des Verfahrens mit der neuen Schrittweite H unter Vernachlässigung von Termen höherer Ordnung

$$\begin{aligned} e(t+H;H) &= \eta(t+H;H) - y(t+H) \\ &= e_p(t+H)H^p \\ &= e_p(t)H^p \\ &= \frac{1}{1-2^{-p}}\left\{\eta(t;h) - \eta\left(t;\frac{h}{2}\right)\right\}\left(\frac{H}{h}\right)^p. \end{aligned}$$

Mithin erhalten wir als optimale Schrittweite

$$H = h\left\{(1-2^{-p})\frac{\varepsilon}{\left\|\eta(t;h) - \eta\left(t;\frac{h}{2}\right)\right\|}\right\}^{\frac{1}{p}}.$$

Dies führt auf Algorithmus I.4.1 zur Schrittweitensteuerung.

Algorithmus I.4.1 Schrittweitensteuerung durch Halbieren

Gegeben: Gegeben t_0, y_0, Startschrittweite h und Toleranz ε
Gesucht: Neue Werte t_0, y_0 und h für nächsten Zeitschritt

1: Berechne $\eta(t_0+h;h)$, $\eta\left(t_0+h;\frac{h}{2}\right)$.

2: $H \leftarrow h\left\{(1-2^{-p})\dfrac{\varepsilon}{\left\|\eta(t_0+h;h)-\eta\left(t_0+h;\frac{h}{2}\right)\right\|}\right\}^{\frac{1}{p}}$

3: **if** $H \le \frac{h}{4}$ **then**

4: $\quad h \leftarrow 2H$

5: $\quad$ Gehe zu 1 zurück.

6: **end if**

7: $t_0 \leftarrow t_0+h$, $y_0 \leftarrow \eta\left(t_0+h;\frac{h}{2}\right)$, $h \leftarrow 2H$

Bei der Durchführung von Algorithmus I.4.1 muss man im Normalfall drei Schritte des ESV pro Iteration durchführen: einen mit Schrittweite h und zwei mit Schrittweite $\frac{h}{2}$. Aus Satz I.3.3 (S. 28) folgt außerdem, dass

$$\frac{1}{2^p-1}\left\{2^p\eta\left(t_0+h;\frac{h}{2}\right)-\eta(t_0+h;h)\right\}$$

eine Näherung der Ordnung h^{p+1} für $y(t_0+h)$ ist. Dies ist das einfachste Beispiel für ein Extrapolationsverfahren.

Beispiel I.4.3 (Schwingung). Wir wenden die beiden Euler-Verfahren und die Trapezregel mit der Schrittweitensteuerung von Algorithmus I.4.1 auf das AWP aus Beispiel I.2.7 (S. 19) an. Alle Verfahren liefern die exakte Lösungskurve, den Kreis um den Ursprung mit Radius 1. Wie aus Tabelle I.4.1 ersichtlich ist, unterscheiden sie sich aber deutlich bei der erreichten Endzeit T, der maximalen Schrittweite $h_{\max}$ und der Rechenzeit. Die Anfangszeit und die zu erreichende Endzeit ist jeweils 0 bzw. 13, die erste Schrittweite ist immer 0.013 und die Toleranz ist stets $\varepsilon = 10^{-6}$.

Tabelle I.4.1. Euler-Verfahren und Trapezregel mit Schrittweitensteuerung durch Halbieren für die Schwingungsgleichung aus Beispiel I.2.7; angegeben sind die erreichte Endzeit T, die Zahl der Schritte N, die maximale Schrittweite $h_{\max}$ und die Rechenzeit in Millisekunden auf einem MacBook Pro

	T	N	$h_{\max}$	msec
expliziter Euler	1.299	1000	0.00615	247
impliziter Euler	1.317	1000	0.00615	525
Trapezregel	10.987	1000	0.04590	423

Beispiel I.4.4 (Räuber-Beute-Modell). Wir wenden die beiden Euler-Verfahren und die Trapezregel mit der Schrittweitensteuerung von Algorithmus I.4.1 auf das Räuber-Beute-Modell aus Beispiel I.2.8 an. Anfangs- und Endzeit sind jeweils 0 bzw. 100, die erste Schrittweite ist immer 0.1 und die Toleranz ist stets $\varepsilon = 10^{-6}$. Alle Verfahren liefern die gleiche Lösungskurve. Die erreichten Endzeiten, benötigten Schritte, minimalen Schrittweiten und benötigten Rechenzeiten sind in Tabelle I.4.2 angegeben.

Tabelle I.4.2. Euler-Verfahren und Trapezregel mit Schrittweitensteuerung durch Halbieren für das Räuber-Beute-Modell aus Beispiel I.2.8; angegeben sind die erreichte Endzeit T, die Zahl der Schritte N, die maximale Schrittweite $h_{\max}$ und die Rechenzeit in Millisekunden auf einem MacBook Pro

	T	N	$h_{\max}$	msec
expliziter Euler	0.959	5000	0.00390	1148
impliziter Euler	0.889	5000	0.00451	2298
Trapezregel	43.938	5000	0.24362	2299

Zur Darstellung der zweiten Strategie bezeichnen wir mit $\widetilde{\eta}(t;h)$ die mit einem Verfahren der Ordnung q, $q > p$, gewonnenen Näherungen. Dann ist bis auf Terme höherer Ordnung

$$\begin{aligned} \eta(t;h) &= y(t) + e_p(t)h^p \\ \widetilde{\eta}(t;h) &= y(t) + \widetilde{e}_q(t)h^q \end{aligned}$$

und somit

$$e_p(t) = h^{-p}\left\{\eta(t;h) - \widetilde{\eta}(t;h)\right\}.$$

Mithin ist für das Verfahren der Ordnung p die optimale Schrittweite gegeben durch

$$H = h\left\{\frac{\varepsilon}{\|\eta(t;h) - \widetilde{\eta}(t;h)\|}\right\}^{\frac{1}{p}}.$$

Dies führt zu Algorithmus I.4.2 zur Schrittweitensteuerung:

Algorithmus I.4.2 Schrittweitensteuerung durch Ordnungsvergleich

Gegeben: Gegeben t_0, y_0, Startschrittweite h und Toleranz ε
Gesucht: Neue Werte t_0, y_0 und h für nächsten Zeitschritt

1: Berechne $\eta(t_0+h;h)$, $\widetilde{\eta}(t_0+h;h)$.
2: $H \leftarrow h \left\{ \dfrac{\varepsilon}{\|\eta(t_0+h;h) - \widetilde{\eta}(t_0+h;h)\|} \right\}^{\frac{1}{p}}$.
3: **if** $H \leq \frac{h}{2}$ **then**
4: $\quad h \leftarrow H$
5: $\quad$ Gehe zu 1 zurück.
6: **end if**
7: $t_0 \leftarrow t_0 + h$, $y_0 \leftarrow \eta\,(t_0+h;h)$, $h \leftarrow H$

Bei den *Runge-Kutta-Fehlberg-Verfahren* kann man die Näherung $\widetilde{\eta}$ ohne großen zusätzlichen Aufwand berechnen.

Beispiel I.4.5 (Runge-Kutta-Fehlberg-Verfahren). (1) Das Schema

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{27}{40} & -\frac{189}{800} & \frac{729}{800} & 0 & 0 \\ 1 & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & 0 \\ \hline & \frac{533}{2106} & 0 & \frac{800}{1053} & -\frac{1}{78} \end{array}$$

beschreibt ein Runge-Kutta-Fehlberg-Verfahren (RKF2) der Ordnung 2. Dann ist

$$\begin{aligned} \eta_1 = \eta_0 &+ \frac{214}{891} h f(t_0, \eta_{0,1}) + \frac{1}{33} h f\left(t_0 + \frac{1}{4}h, \eta_{0,2}\right) \\ &+ \frac{650}{891} h f\left(t_0 + \frac{27}{40}h, \eta_{0,3}\right) \end{aligned}$$

eine Näherung der Ordnung $p = 2$ und

$$\begin{aligned} \widetilde{\eta}_1 = \eta_0 &+ \frac{533}{2106} h f(t_0, \eta_{0,1}) + \frac{800}{1053} h f\left(t_0 + \frac{27}{40}h, \eta_{0,3}\right) \\ &- \frac{1}{78} h f(t_0 + h, \eta_1) \end{aligned}$$

eine Näherung der Ordnung $q = 3$.
(2) Das Schema

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

beschreibt ein Runge-Kutta-Fehlberg-Verfahren (RKF3) der Ordnung 3. Dann ist

$$\eta_1 = \eta_0 + \frac{1}{6}hf(t_0, \eta_{0,1}) + \frac{1}{3}hf\left(t_0 + \frac{1}{2}h, \eta_{0,2}\right) + \frac{1}{3}hf\left(t_0 + \frac{1}{2}h, \eta_{0,3}\right) + \frac{1}{6}hf(t_0 + h, \eta_{0,4})$$

eine Näherung der Ordnung $p = 3$ und

$$\widetilde{\eta}_1 = \eta_0 + \frac{1}{6}hf(t_0, \eta_{0,1}) + \frac{1}{3}hf\left(t_0 + \frac{1}{2}h, \eta_{0,2}\right) + \frac{1}{3}hf\left(t_0 + \frac{1}{2}h, \eta_{0,3}\right) + \frac{1}{6}hf(t_0 + h, \eta_1)$$

eine Näherung der Ordnung $q = 4$.
(3) Das Schema

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

beschreibt ein Runge-Kutta-Fehlberg-Verfahren (RKF4) der Ordnung 4. Dann ist

$$\eta_1 = \eta_0 + \frac{35}{384}hf(t_0, \eta_{0,1}) + \frac{500}{1113}hf\left(t_0 + \frac{3}{10}h, \eta_{0,3}\right) + \frac{125}{192}hf\left(t_0 + \frac{8}{10}h, \eta_{0,4}\right) - \frac{2187}{6784}hf\left(t_0 + \frac{8}{9}h, \eta_{0,5}\right) + \frac{11}{84}hf(t_0 + h, \eta_{0,6})$$

eine Näherung der Ordnung $p = 4$ und

$$\begin{aligned}\widetilde{\eta}_1 = \eta_0 &+ \frac{5179}{57600} hf(t_0, \eta_{0,1}) + \frac{7571}{16695} hf\left(t_0 + \frac{3}{10}h, \eta_{0,3}\right) \\ &+ \frac{393}{640} hf\left(t_0 + \frac{8}{10}h, \eta_{0,4}\right) - \frac{92097}{339200} hf\left(t_0 + \frac{8}{9}h, \eta_{0,5}\right) \\ &+ \frac{187}{2100} hf(t_0 + h, \eta_{0,6}) + \frac{1}{40} hf(t_0 + h, \eta_1)\end{aligned}$$

eine Näherung der Ordnung $q = 5$.

Beispiel I.4.6 (Schwingung). Wir wenden die Runge-Kutta-Fehlberg-Verfahren aus Beispiel I.4.5 mit der Schrittweitensteuerung von Algorithmus I.4.2 auf das AWP aus Beispiel I.2.7 (S. 19) an. Alle Verfahren liefern die exakte Lösungskurve, den Kreis um den Ursprung mit Radius 1. Die entsprechenden Ergebnisse sind in Tabelle I.4.3 angegeben. Dabei sind die Bezeichnungen wie in Beispiel I.4.3.

Tabelle I.4.3. Runge-Kutta-Fehlberg-Verfahren mit Schrittweitensteuerung durch Ordnungsvergleich für die Schwingungsgleichung aus Beispiel I.2.7; angegeben sind die erreichte Endzeit T, die Zahl der Schritte N, die maximale Schrittweite $h_{\max}$ und die Rechenzeit in Millisekunden auf einem MacBook Pro

	T	N	$h_{\max}$	msec
RKF2	13.000	260	0.25664	31
RKF3	13.000	400	0.13323	199
RKF4	13.000	151	0.35478	67

Beispiel I.4.7 (Räuber-Beute-Modell). Wir wenden die Runge-Kutta-Fehlberg-Verfahren aus Beispiel I.4.5 mit der Schrittweitensteuerung von Algorithmus I.4.2 auf das Räuber-Beute-Modell aus Beispiel I.2.8 (S. 19) an. Alle Verfahren liefern die gleiche Lösungskurve. Die entsprechenden Ergebnisse sind in der folgenden Tabelle angegeben. Dabei sind die Bezeichnungen wie in Beispiel I.4.4.

I.5. Lineare Mehrschrittverfahren

Bei den bisher betrachteten ESV wird die neue Näherung η_{j+1} direkt aus der alten Näherung η_j berechnet. In diesem und den nächsten Paragraphen wollen wir *lineare Mehrschrittverfahren (MSV)* betrachten, bei denen zur Berechnung von η_{j+1} alte Näherungen η_j, η_{j-1}, …, η_{j-k} benötigt werden. Bevor wir eine allgemeine Definition geben, betrachten wir einige Beispiele.

TABELLE I.4.4. Runge-Kutta-Fehlberg-Verfahren mit Schrittweitensteuerung durch Ordnungsvergleich für das Räuber-Beute-Modell aus Beispiel I.2.8, angegeben sind die erreichte Endzeit T, die Zahl der Schritte N, die maximale Schrittweite $h_{\max}$ und die Rechenzeit in Millisekunden auf einem MacBook Pro

	T	N	$h_{\max}$	msec
RKF2	100.000	2633	1.35645	215
RKF3	100.000	4011	0.60405	1615
RKF4	100.000	1593	1.45927	869

BEISPIEL I.5.1 (Formeln von Adams-Bashforth, Adams-Moulton und Nyström). Wir wählen äquidistante Punkte $t_i = t_0 + ih$, $i \in \mathbb{N}$. Sei $p > 0$. Wir nehmen an, dass wir die Lösung y des AWP (I.1.2) (S. 10) in den Punkten $t_0, \ldots, t_p$ kennen. Sei $0 \le s \le p$. Dann folgt

$$\begin{aligned} y(t_{p+1}) - y(t_{p-s}) &= \int_{t_{p-s}}^{t_{p+1}} y'(\tau)\mathrm{d}\tau \\ &= \int_{t_{p-s}}^{t_{p+1}} f(t, y(\tau))\mathrm{d}\tau. \end{aligned} \tag{I.5.1}$$

Sei nun $k \in \{0, 1\}$ und $q \ge 0$ so, dass $p + k - q \ge 0$ ist. Dann approximieren wir den Integranden in (I.5.1) durch das Interpolationspolynom in den Punkten $t_{p+k}, t_{p+k-1}, \ldots, t_{p+k-q}$ und erhalten

$$\begin{aligned} &y(t_{p+1}) - y(t_{p-s}) \\ &\quad\approx \int_{t_{p-s}}^{t_{p+1}} \sum_{i=0}^{q} f(t_{p+k-i}, y(t_{p+k-i})) \prod_{\substack{j=0 \\ j\neq i}}^{q} \frac{\tau - t_{p+k-j}}{t_{p+k-i} - t_{p+k-j}}\mathrm{d}\tau \\ &\quad= h \sum_{i=0}^{q} f(t_{p+k-i}, y(t_{p+k-i})) \int_{-s}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{q} \frac{z - k + j}{j - i}\mathrm{d}z \\ &\quad= h \sum_{i=0}^{q} f(t_{p+k-i}, y(t_{p+k-i}))\beta_{qi}. \end{aligned}$$

Dies liefert das MSV

$$\eta_{p+1} = \eta_{p-s} + h \sum_{i=0}^{q} \beta_{qi} f(t_{p+k-i}, \eta_{p+k-i}).$$

Je nach Wahl von s und k erhalten wir drei häufig benutzte Klassen von Verfahren:

Adams-Bashforth-Formeln: $s = 0$, $k = 0$

$$\eta_{p+1} = \eta_p + h \sum_{i=0}^{q} \beta_{qi}^{AB} f(t_{p-i}, \eta_{p-i}).$$

Die Koeffizienten sind in folgender Tabelle zusammengefasst:

q	β_{qi}^{AB}				
0	1				
1	$\frac{3}{2}$	$-\frac{1}{2}$			
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$		
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	
4	$\frac{1901}{720}$	$-\frac{2774}{720}$	$\frac{2616}{720}$	$-\frac{1274}{720}$	$\frac{251}{720}$

Für $q = 0$ erhalten wir also das explizite Euler-Verfahren zurück.
Adams-Moulton-Formeln: $s = 0$, $k = 1$

$$\eta_{p+1} = \eta_p + h \sum_{i=0}^{q} \beta_{qi}^{AM} f(t_{p+1-i}, \eta_{p+1-i}).$$

Die Koeffizienten sind in folgender Tabelle zusammengefasst:

q	β_{qi}^{AM}				
0	1				
1	$\frac{1}{2}$	$\frac{1}{2}$			
2	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$		
3	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	
4	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

Die Formeln von Adams-Moulton sind also implizit. Für $q = 0$ erhalten wir das implizite Euler-Verfahren, und für $q = 1$ erhalten wir die Trapezregel.
Nyström-Formeln: $s = 1$, $k = 0$

$$\eta_{p+1} = \eta_{p-1} + h \sum_{i=0}^{q} \beta_{qi}^{N} f(t_{p-i}, \eta_{p-i}).$$

Die Koeffizienten sind in folgender Tabelle zusammengefasst:

q	β_{qi}^N				
1	2	0			
2	$\frac{7}{3}$	$-\frac{2}{3}$	$\frac{1}{3}$		
3	$\frac{8}{3}$	$-\frac{5}{3}$	$\frac{4}{3}$	$-\frac{1}{3}$	
4	$\frac{269}{90}$	$-\frac{266}{90}$	$\frac{294}{90}$	$-\frac{146}{90}$	$\frac{29}{90}$

Für $q = 1$ erhält man speziell die häufig benutzte *Mittelpunktsregel*

$$\eta_{p+1} = \eta_{p-1} + 2hf(t_p, \eta_p).$$

Beispiel I.5.2 (Rückwärtige Differentiation, BDF-Formeln). Die Idee ist, das Interpolationspolynom ζ zu den Knoten t_{p+1}, t_p, …, t_{p-m} und den Daten η_{p+1}, η_p, …, η_{p-m} zu bestimmen und $\zeta'(t_{p+1})$ gleich $f(t_{p+1}, \eta_{p+1})$ zu setzen. Dies liefert nach Normierung des Koeffizienten von η_{p+1} ein MSV der Form

$$\sum_{k=0}^{m+1} a_{mk}\eta_{p+1-k} = hb_m f(t_{p+1}, \eta_{p+1})$$

mit den folgenden Koeffizienten

m	b_m	a_{mk}					
0	1	1	-1				
1	$\frac{2}{3}$	1	$-\frac{4}{3}$	$\frac{1}{3}$			
2	$\frac{6}{11}$	1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$		
3	$\frac{12}{25}$	1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$	
4	$\frac{60}{137}$	1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$

.

Die betrachteten Verfahren sind alle von der Form:
Linearkombination von mehreren aufeinanderfolgenden Näherungen = Linearkombination von mehreren aufeinanderfolgenden f-Werten.

Dies führt auf folgende allgemeine Definition, wobei der Bequemlichkeit halber die Indizierung geändert ist.

Definition I.5.3 (Lineares Mehrschrittverfahren, MSV). Ein *lineares r-Schritt-Verfahren* kurz *MSV* zur Lösung des AWP (I.1.2) (S. 10) hat die Form

$$\eta_i = y(t_i) + \varepsilon_i, \quad 0 \le i \le r-1 \tag{I.5.2a}$$

$$\sum_{k=0}^{r} a_k\eta_{j+k} = h\sum_{k=0}^{r} b_k f(t_{j+k}, \eta_{j+k}), \quad j = 0, 1, \ldots \tag{I.5.2b}$$

mit $a_r = 1$.
Die *Startfehler* ε_i, $0 \le i \le r-1$, fasst man zu der *Fehlerfunktion* $\varepsilon(t;h)$ mit

$$\varepsilon(t_i, h) = \varepsilon_i$$

zusammen. Für die Lösung von (I.5.2a), (I.5.2b) schreibt man $\eta(t;\varepsilon,h)$ mit

$$\eta(t_i;\varepsilon,h) = \eta_i.$$

Die Polynome

$$\psi(z) = \sum_{k=0}^{r} a_k z^k, \quad \chi(z) = \sum_{k=0}^{r} b_k z^k$$

nennt man die *charakteristischen Polynome* des Verfahrens.
Der *lokale Verfahrensfehler* $\tau(x,y;h)$ des Verfahrens ist definiert als

$$\tau(x,y;h) = \frac{1}{h}\left\{\sum_{k=0}^{r} a_k z(x+kh) - h\sum_{k=0}^{r} b_k f(x+kh, z(x+kh))\right\}$$

wobei z die Lösung des AWP

$$z' = f(t, z(t)), \quad z(x) = y$$

ist.
Das Verfahren heißt *konsistent*, wenn es eine Funktion $\sigma(h)$ gibt mit

$$\|\tau(x,y;h)\| \le \sigma(h)$$

für alle $x \in [a,b]$, $y \in \mathbb{R}^n$ und

$$\lim_{h\to 0} \sigma(h) = 0.$$

Es hat die *Ordnung* p, $p > 0$, wenn es konsistent ist und wenn für die Funktion σ gilt

$$\limsup_{h\to 0} h^{-p}\sigma(h) < \infty$$

für alle $f \in C^p((a,b)\times\mathbb{R}^n, \mathbb{R}^n)$. Das Verfahren heißt schließlich *konvergent*, wenn gilt

$$\lim_{n\to\infty} \eta\left(t;\varepsilon,\frac{t-t_0}{n}\right) = y(t)$$

für alle $t \in [t_0, b]$, $y_0 \in \mathbb{R}^n$, $f \in C^1((a,b)\times\mathbb{R}^n, \mathbb{R}^n)$ und alle Fehlerfunktionen ε, zu denen es eine Funktion $\zeta(h)$ gibt mit

$$\|\varepsilon(t;h)\| \le \zeta(h)$$

für alle $t \in [a,b]$ und

$$\lim_{h\to 0} \zeta(h) = 0.$$

BEMERKUNG I.5.4 (Berechnung der Startwerte). Ein r-Schritt Verfahren benötigt offensichtlich r Startwerte $\eta_0, \ldots, \eta_{r-1}$. Falls $r > 1$ ist, werden diese mit einem ESV berechnet. Die Größen ε_i, $0 \le i \le r-1$, berücksichtigen die dabei auftretenden Fehler. Sinnvollerweise benutzt man zur Berechnung der Startwerte ein ESV, das die gleiche Ordnung hat wie das MSV.

Die Konsistenz und Ordnung eines MSV lässt sich besonders leicht mit Hilfe der charakteristischen Polynome berechnen.

SATZ I.5.5 (Bedingungen für Konsistenz und Ordnung eines MSV). *Ein MSV ist genau dann konsistent, wenn gilt*

$$\psi(1) = 0 \quad \textit{und} \quad \psi'(1) - \chi(1) = 0.$$

Außerdem sind die folgenden Aussagen äquivalent:

(1) *Das MSV hat die Ordnung $p > 0$.*
(2) $\displaystyle\sum_{k=0}^{r} \left\{ a_k \frac{k}{\mu!} - b_k \frac{1}{(\mu-1)!} \right\} k^{\mu-1} = 0$ *für alle* $1 \le \mu \le p$.
(3) *Die Funktion* $\varphi(z) = \frac{\psi(z)}{\ln z} - \chi(z)$ *hat in* $z = 1$ *eine p-fache Nullstelle.*

BEWEIS. Seien $x \in [a,b]$, $y \in \mathbb{R}^n$, $f \in C^m((a,b) \times \mathbb{R}^n, \mathbb{R}^n)$ und z Lösung des AWP

$$z' = f(t, z(t)), \quad z(x) = y.$$

Dann ist $z \in C^{m+1}((a,b))$. Taylor-Entwicklung des lokalen Verfahrensfehlers um x liefert

$$\begin{aligned}
\tau(x,y;h) &= \frac{1}{h} \left\{ \sum_{k=0}^{r} a_k z(x+kh) - h \sum_{k=0}^{r} b_k z'(x+kh) \right\} \\
&= \frac{1}{h} \left\{ \sum_{k=0}^{r} a_k \left[\sum_{\mu=0}^{m} \frac{1}{\mu!} z^{(\mu)}(x) k^\mu h^\mu \right] \right. \\
&\qquad \left. - \sum_{k=0}^{r} b_k \left[\sum_{\mu=0}^{m-1} \frac{1}{\mu!} z^{(\mu+1)}(x) k^\mu h^{\mu+1} \right] \right\} + \mathrm{O}(h^m) \\
&= \frac{1}{h} z(x) \sum_{k=0}^{r} a_k + z'(x) \left\{ \sum_{k=0}^{r} k a_k - \sum_{k=0}^{r} b_k \right\} \\
&\quad + \sum_{\mu=1}^{m-1} h^\mu \left\{ \sum_{k=0}^{r} \left[\frac{1}{(\mu+1)!} a_k k^{\mu+1} - b_k \frac{1}{\mu!} k^\mu \right] z^{(\mu+1)}(x) \right\} \\
&\quad + \mathrm{O}(h^m).
\end{aligned}$$

Also ist das Verfahren konsistent genau dann, wenn gilt

$$\psi(1) = \sum_{k=0}^{r} a_k = 0$$

und

$$\psi'(1) - \chi(1) = \sum_{k=0}^{r} (ka_k - b_k) = 0.$$

Außerdem hat es die Ordnung $p > 0$ genau dann, wenn Bedingung (2) erfüllt ist. Daher müssen wir nur noch die Äquivalenz von (2) und (3) zeigen. Taylor-Entwicklung von $\varphi(e^h)$ um $h = 0$ liefert

$$\begin{aligned}
\varphi(e^h) &= \frac{1}{h}\psi(e^h) - \chi(e^h) \\
&= \frac{1}{h}\left[\psi(e^h) - h\chi(e^h)\right] \\
&= \frac{1}{h}\left\{\sum_{k=0}^{r} a_k e^{kh} - h\sum_{k=0}^{r} b_k e^{kh}\right\} \\
&= \frac{1}{h}\left\{\sum_{k=0}^{r} a_k \left[\sum_{\mu=0}^{m} \frac{1}{\mu!}k^\mu h^\mu\right] - \sum_{k=0}^{r} b_k \left[\sum_{\mu=0}^{m-1} \frac{1}{\mu!}k^\mu h^{\mu+1}\right]\right\} \\
&\quad + \mathrm{O}(h^m) \\
&= \frac{1}{h}\sum_{k=0}^{r} a_k + \sum_{\mu=0}^{m-1} h^\mu \left\{\sum_{k=0}^{r}\left[a_k \frac{1}{(\mu+1)!}k^{\mu+1} - b_k \frac{1}{\mu!}k^\mu\right]\right\} \\
&\quad + \mathrm{O}(h^m).
\end{aligned}$$

Also ist (2) äquivalent dazu, dass $\varphi(e^h)$ in $h = 0$ eine p-fache Nullstelle hat. Da die Funktion e^h mitsamt allen Ableitungen in $h = 0$ nicht verschwindet, ist dies äquivalent zu (3). □

Bemerkung I.5.6 (Konsistenz und Ordnung 1 eines MSV). Wegen

$$\sum_{k=0}^{r} (a_k k - b_k) = \psi'(1) - \chi(1)$$

hat ein konsistentes MSV immer mindestens die Ordnung 1.

Bemerkung I.5.7 (Lineare 1-Schritt-Verfahren, θ-Schema). Die allgemeine Form eines linearen 1-Schritt-Verfahrens, d.h. eines MSV mit $r = 1$ lautet

$$\eta_{j+1} + a_0\eta_j = h\left(b_1 f(t_{j+1}, \eta_{j+1}) + b_0 f(t_j, \eta_j)\right).$$

Die charakteristischen Polynome sind

$$\psi(z) = z + a_0, \quad \chi(z) = b_1 z + b_0.$$

Wegen Satz I.5.5 ist es genau dann konsistent, wenn $a_0 = -1$ und $b_1 + b_0 = 1$ ist. Daher lautet die allgemeine Form eines konsistenten linearen 1-Schritt-Verfahrens

$$\eta_{j+1} = \eta_j + h\left(\theta f(t_{j+1}, \eta_{j+1}) + (1-\theta) f(t_j, \eta_j)\right)$$

mit $\theta \in \mathbb{R}$. Wegen Satz I.5.5 hat es für alle θ mindestens die Ordnung 1 und die Ordnung 2 genau dann, wenn $\theta = \frac{1}{2}$ ist. Für $\theta = 0$, $\theta = 1$ und $\theta = \frac{1}{2}$ erhalten wir offensichtlich das explizite Euler-Verfahren, das implizite Euler-Verfahren und die Trapezregel.

Aus Satz I.5.5 folgt mit einiger Rechnung:

KOROLLAR I.5.8 (Ordnung der Adams-Verfahren und der Nyström- und BDF-Formeln). *Die Adams-Verfahren und die Nyström-Formeln haben die Ordnung $q+1$. Die BDF-Formeln mit $m+1$ Schritten haben die Ordnung $m+1$.*

Satz I.5.5 zeigt auch, wie man ein MSV möglichst hoher Ordnung konstruieren kann. Zu gegebenem ψ bestimmt man χ so, dass die Terme h, ..., h^r in der Taylor-Entwicklung von φ verschwinden. Dann kann man zudem versuchen, ψ so zu bestimmen, dass weitere Terme in der Taylor-Entwicklung von φ verschwinden. Wir werden im nächsten Paragraphen sehen, dass dies nicht ohne weiteres möglich ist und dass man weitere Bedingungen bei der Konstruktion und Analyse von MSV beachten muss.

I.6. Konvergenz von linearen Mehrschrittverfahren

In diesem Paragraphen bezeichnen

$$\psi(z) = \sum_{k=0}^{r} a_k z^k \quad \text{und} \quad \chi(z) = \sum_{k=0}^{r} b_k z^k$$

die charakteristischen Polynome eines linearen r-Schritt Verfahrens. Wir wollen Bedingungen für die Konvergenz dieses Verfahrens herleiten. Wir werden sehen, dass wir zusätzlich zur Konsistenz eine weitere Bedingung, die für ESV automatisch erfüllt ist, benötigen.

Zur Motivation wenden wir das MSV auf das AWP

$$y' = \lambda y, \quad y(0) = y_0 \tag{I.6.1}$$

mit $\lambda \in \mathbb{C}$ an.
Dieses AWP spielt eine herausragende Rolle bei der Analyse von numerischen Verfahren zur Lösung gewöhnlicher Differentialgleichungen, da sich jede Differentialgleichung nach Linearisierung und Koordinatentransformation lokal wie (I.6.1) verhält, wobei λ ein Eigenwert der Jacobi Matrix ist.
Wir erhalten dann für $j = 0, 1, \ldots$ die Gleichungen

$$\sum_{k=0}^{r} a_k \eta_{j+k} = h\lambda \sum_{k=0}^{r} b_k \eta_{j+k}$$

oder mit $\mu = \lambda h$

$$\sum_{k=0}^{r} (a_k - \mu b_k)\eta_{j+k} = 0 \quad j = 0, 1, \ldots. \tag{I.6.2}$$

Gleichung (I.6.2) ist eine *homogene Differenzengleichung.*
Zunächst charakterisieren wir die Lösungen von (I.6.2).

SATZ I.6.1 (Lösungen homogener Differenzengleichungen). *Jede Lösung* $(u_n)_{n\in\mathbb{N}}$ *der homogenen Differenzengleichung*

$$\sum_{k=0}^{r} c_k u_{j+k} = 0 \quad j = 0, 1, \dots$$

mit $c_k \in \mathbb{C}$, $c_r \neq 0$ *ist von der Form*

$$u_n = \sum_{i=1}^{\ell} \sum_{j=0}^{m_i - 1} \alpha_{ij} n^j z_i^n \quad n \in \mathbb{N}$$

mit $\alpha_{ij} \in \mathbb{C}$. *Dabei sind* $z_1, \dots, z_\ell \in \mathbb{C}$ *die Nullstellen mit Vielfachheit* $m_1, \dots, m_\ell$, $\sum m_i = r$, *des charakteristischen Polynoms*

$$\rho(z) = \sum_{k=0}^{r} c_k z^k.$$

BEWEIS. Zum besseren Verständnis betrachten wir zunächst den Fall, dass ρ nur einfache Nullstellen besitzt, d.h. $\ell = r$, $m_1 = \ldots = m_r = 1$. Für $1 \leq i \leq r$ definieren wir die Folge $(u_n^{(i)})_{n\in\mathbb{N}}$ durch $u_n^{(i)} = z_i^n$ für alle $n \in \mathbb{N}$. Dann folgt für $j \in \mathbb{N}$ und $1 \leq i \leq r$

$$\sum_{k=0}^{r} c_k u_{j+k}^{(i)} = \sum_{k=0}^{r} c_k z_i^{j+k} = z_i^j \rho(z_i) = 0.$$

Also sind die Folgen $u^{(1)}$, ..., $u^{(r)}$ Lösungen der Differenzengleichung (I.6.2). Da der Lösungsraum offensichtlich r-dimensional ist, reicht es zu zeigen, dass diese r Folgen linear unabhängig sind. Angenommen, das Gegenteil sei der Fall. Dann sind $(u_{j-1}^{(i)})_{1\leq j\leq r}$, $1 \leq i \leq r$, r linear abhängige Vektoren im $\mathbb{R}^r$. Also sind die Spalten der Matrix A mit $A_{ki} = u_{k-1}^{(i)}$, $1 \leq k, i \leq r$, linear abhängig. Dann gilt gleiches für die Zeilen, und es gibt r Zahlen $\gamma_1, \dots, \gamma_r$ mit $\sum |\gamma_k|^2 \neq 0$ und

$$\sum_{k=1}^{r} \gamma_k A_{ki} = \sum_{k=1}^{r} \gamma_k z_i^{k-1} = \sum_{m=0}^{r-1} \gamma_{m+1} z_i^m = 0$$

für alle $1 \leq i \leq r$. Also hat das Polynom $\sigma(z) = \sum_{m=0}^{r-1} \gamma_{m+1} z^m$ einen Grad $< r$ und besitzt die r verschiedenen Nullstellen $z_1, \dots, z_r$. Dies ist ein Widerspruch.
Nun betrachten wir den allgemeinen Fall. Wir definieren Polynome $\rho_0, \dots, \rho_{r-1}$ vom Grade r durch

$$\rho_0(z) = \rho(z), \quad \rho_{i+1}(z) = z\rho_i'(z), \ i = 0, \dots, r-2.$$

Dann folgt durch Induktion

$$\rho_k(z) = \sum_{\mu=0}^{r} c_\mu \mu^k z^\mu, \quad k = 0, \dots, r-1.$$

Sei nun z^* eine ℓ^*-fache Nullstelle von ρ, $\ell^* \geq 1$. Dann folgt durch Induktion $\rho_k(z^*) = 0$ für $k = 0, \dots, \ell^* - 1$. Damit erhalten wir für $0 \leq m \leq \ell^* - 1$ und $n \in \mathbb{N}$

$$\begin{aligned}\sum_{k=0}^{r} c_k (n+k)^m \left(z^*\right)^{n+k} &= \sum_{k=0}^{r} \sum_{\mu=0}^{m} \binom{m}{\mu} c_k n^{m-\mu} k^\mu \left(z^*\right)^{n+k} \\ &= \sum_{\mu=0}^{m} \binom{m}{\mu} n^{m-\mu} \left(z^*\right)^n \rho_\mu(z^*) \\ &= 0.\end{aligned}$$

Also sind die Folgen $(u_n^{(i,j)})_{n\in\mathbb{N}} = (n^j z_i^n)_{n\in\mathbb{N}}$, $1 \leq i \leq \ell$, $0 \leq j \leq m_i - 1$, Lösungen der homogenen Differenzengleichung. Daher reicht es zu zeigen, dass diese r Folgen linear unabhängig sind. Wir nehmen an, das Gegenteil sei der Fall. Dann sind die Spalten der Matrix A mit

$$A_{\mu\nu} = (\nu - 1)^j z_i^{\nu-1} \quad 1 \leq \mu, \nu \leq r, \mu = \sum_{\alpha=1}^{i-1} m_\alpha + j$$

linear abhängig. Also sind auch die Zeilen linear abhängig. Also gibt es Zahlen $\gamma_0, \dots, \gamma_{r-1} \in \mathbb{C}$ mit $\sum |\gamma_i|^2 \neq 0$ und $\sum_{\mu=0}^{r-1} \gamma_\mu \mu^j z_i^\mu = 0$ für $1 \leq i \leq \ell$, $0 \leq j \leq m_i - 1$. Wie oben folgt hieraus, dass das Polynom $\sigma(z) = \sum_{\mu=0}^{r-1} \gamma_\mu z^\mu$ die Nullstellen $z_1, \dots, z_\ell$ mit Vielfachheiten $m_1, \dots, m_\ell$ hat. Also ist σ das Nullpolynom. Dies ist ein Widerspruch. □

Aus Satz I.6.1 folgt unmittelbar:

KOROLLAR I.6.2 (Stabilitätsbedingung). *Folgende Aussagen sind äquivalent:*

(1) *Für jede Lösung $(u_n)_{n\in\mathbb{N}}$ der homogenen Differenzengleichung*

$$\sum_{k=0}^{r} c_k u_{j+k} = 0 \quad j = 0, 1, \dots$$

gilt

$$\sup_{n\in\mathbb{N}} |u_n| < \infty.$$

(2) *Für jede Nullstelle z^* des charakteristischen Polynoms*

$$\rho(z) = \sum_{k=0}^{r} c_k z^k$$

gilt:
$|z^*| \leq 1$ und z^* hat die Vielfachheit 1, falls $|z^*| = 1$ ist.

Wählen wir in Gleichung (I.6.1) speziell $\lambda = 0$, dann ist ψ das charakteristische Polynom der Differenzengleichung (I.6.2). Da die exakte Lösung des AWP konstant ist, sollten für ein vernünftiges MSV in diesem Fall die Lösungen von (I.6.2) beschränkt bleiben. Also sollte ψ die Bedingung (2) von Korollar I.6.2 erfüllen. Dies führt auf folgende, von Dahlquist stammende Definition.

DEFINITION I.6.3 (Asymptotische Stabilität). Ein lineares r-Schritt Verfahren heißt *asymptotisch stabil* oder kurz *stabil*, wenn das zugehörige Polynom ψ die Stabilitätsbedingung (2) von Korollar I.6.2 erfüllt.

Wir wollen zeigen, dass ein konsistentes und stabiles MSV konvergent ist. Dazu benötigen wir ein Hilfsergebnis.

LEMMA I.6.4 (Stabilitätsbedingung und Matrixnormen). *Das charakteristische Polynom der $n \times n$ Matrix A erfülle die Stabilitätsbedingung (2) von Korollar I.6.2. Dann gibt es eine Vektornorm $\|\cdot\|_A$, so dass für die zugehörige Matrixnorm $\|\cdot\|_{\mathcal{L}}$ gilt $\|A\|_{\mathcal{L}} \leq 1$.*

BEWEIS. Seien $\lambda_1, \ldots, \lambda_\ell$ die Eigenwerte von A mit Vielfachheiten $m_1, \ldots, m_\ell$, $\sum m_i = n$. Nach Voraussetzung gilt $|\lambda_i| \leq 1$ für $1 \leq i \leq \ell$ und $m_i = 1$ falls $|\lambda_i| = 1$. Weiter gibt es eine reguläre Matrix U, so dass UAU^{-1} Jordansche Normalform hat:

$$UAU^{-1} = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_\ell \end{pmatrix} \quad \text{mit} \quad J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix}.$$

Sei $\delta = 1 - \max\{|\lambda_i| : m_i > 1\}$. Nach Voraussetzung ist $\delta > 0$. Definiere

$$D = \begin{pmatrix} D_1 & & \\ & \ddots & \\ & & D_\ell \end{pmatrix} \quad \text{mit} \quad D_i = \begin{pmatrix} \delta^{m_i - 1} & & & \\ & \ddots & & \\ & & \delta & \\ & & & 1 \end{pmatrix}.$$

Dann folgt

$$DUAU^{-1}D^{-1} = \begin{pmatrix} \widetilde{J}_1 & & \\ & \ddots & \\ & & \widetilde{J}_\ell \end{pmatrix}$$

mit

$$\widetilde{J}_i = D_i J_i D_i^{-1} = \begin{pmatrix} \lambda_i & \delta & & \\ & \ddots & \ddots & \\ & & \ddots & \delta \\ & & & \lambda_i \end{pmatrix}.$$

Also ist

$$\left\| DUAU^{-1}D^{-1} \right\|_{\mathcal{L},\infty} \leq 1,$$

wobei $\|\cdot\|_{\mathcal{L},\infty}$ die zur Maximumsnorm $\|\cdot\|_\infty$ zugehörige Zeilensummennorm ist. Definiere

$$\|x\|_A = \|DUx\|_\infty .$$

Dann folgt

$$\begin{aligned}\|A\|_{\mathcal{L}} &= \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_A}{\|x\|_A} = \sup_{x\neq 0} \frac{\|DUAx\|_\infty}{\|DUx\|_\infty} \\ &= \sup_{y\neq 0} \frac{\|DUAU^{-1}D^{-1}y\|_\infty}{\|y\|_\infty} = \left\|DUAU^{-1}D^{-1}\right\|_{\mathcal{L},\infty} \\ &\leq 1.\end{aligned}$$

□

Wir kommen nun zum wesentlichen Ergebnis dieses Paragraphen.

Satz I.6.5 (Konvergenz von Mehrschrittverfahren). *Folgende Aussagen sind äquivalent:*

(1) *Das MSV ist konsistent und asymptotisch stabil.*
(2) *Das MSV konvergiert für jede Funktion $f \in C((a,b)\times\mathbb{R}^n, \mathbb{R}^n)$, die gleichmäßig Lipschitz-stetig ist bzgl. $\mathbb{R}^n$.*

Beweis. (2) $\Longrightarrow$ (1): Wende das MSV auf

$$y' = 0, \quad y(t_0) = y_0$$

an. Dann folgt die asymptotische Stabilität aus Korollar I.6.2.
Wähle speziell $y_0 = 1$ und $\eta_0 = \ldots = \eta_{r-1} = 1$. Dann folgt

$$\sum_{k=0}^{r} a_k \eta_{j+k} = 0,\ j = 0, 1, \ldots \quad \text{und} \quad \lim_{n\to\infty} \eta_n = 1.$$

Also gilt

$$0 = \lim_{j\to\infty} \left\{ \sum_{k=0}^{r} a_k \eta_{j+k} \right\} = \sum_{k=0}^{r} a_k = \psi(1).$$

Aus der asymptotischen Stabilität folgt weiter $\psi'(1) \neq 0$. Daher ist $K = \frac{\chi(1)}{\psi'(1)}$ wohldefiniert. Wende das MSV auf

$$y' = 1, \quad y(t_0) = 0$$

an mit den Startwerten $\eta_i = ihK$, $0 \leq i \leq r-1$. Dann folgt für $\eta_j = jhK$, $j \geq 0$,

$$\begin{aligned}\sum_{k=0}^{r} a_k \eta_{j+k} - h\sum_{k=0}^{r} b_k &= jhK \sum_{k=0}^{r} a_k + \sum_{k=0}^{r} a_k khK - h\sum_{k=0}^{r} b_k \\ &= jhK\psi(1) + hK\psi'(1) - h\chi(1) \\ &= 0.\end{aligned}$$

Aus der Konvergenz des Verfahrens folgt daher

$$t - t_0 = y(t) = \lim_{n\to\infty} \eta\left(t; \frac{t-t_0}{n}\right) = (t-t_0)K$$

und somit $K = 1$. Damit ist die Konsistenz des Verfahrens gezeigt.
(1) $\Longrightarrow$ (2): Sei f auf $(a,b) \times \mathbb{R}^n$ gleichmäßig Lipschitz-stetig bzgl. $\mathbb{R}^n$ mit Lipschitz-Konstante L. Definiere $e_j = \eta_j - y(t_j)$. Dann gilt

$$\begin{aligned}
e_{j+r} &= \eta_{j+r} - y(t_{j+r}) \\
&= -\sum_{k=0}^{r-1} a_k \eta_{j+k} + h \sum_{k=0}^{r} b_k f(t_{j+k}, \eta_{j+k}) - y(t_{j+r}) \\
&= -\sum_{k=0}^{r-1} a_k e_{j+k} - \left\{ \sum_{k=0}^{r} a_k y(t_{j+k}) - h \sum_{k=0}^{r} b_k f(t_{j+k}, y(t_{j+k})) \right\} \\
&\quad + h \sum_{k=0}^{r} b_k \left[f(t_{j+k}, \eta_{j+k}) - f(t_{j+k}, y(t_{j+k})) \right] \\
&= -\sum_{k=0}^{r-1} a_k e_{j+k} - h\tau(t_j, y(t_j); h) \\
&\quad + h \sum_{k=0}^{r} b_k \left[f(t_{j+k}, \eta_{j+k}) - f(t_{j+k}, y(t_{j+k})) \right].
\end{aligned}$$

Definiere

$$E_j = \begin{pmatrix} e_j \\ \vdots \\ e_{j+r-1} \end{pmatrix}, \; T_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -\tau(t_j, y(t_j); h) \end{pmatrix}, \; A = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & \ddots & & \ddots \\ & & & 0 \quad 1 \\ -a_0 & \dots & \dots & -a_{r-1} \end{pmatrix}$$

$$\Lambda_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sum_{k=0}^{r} b_k [f(t_{j+k}, \eta_{j+k}) - f(t_{j+k}, y(t_{j+k}))] \end{pmatrix}.$$

Dann folgt aus obiger Rechnung

$$E_{j+1} = AE_j + hT_j + h\Lambda_j, \quad j = 0, 1, \dots \; .$$

Gemäß Lemma I.6.4 gibt es eine Vektornorm $\|\cdot\|_A$, so dass für die zugehörige Matrixnorm $\|\cdot\|_{\mathcal{L}}$ gilt $\|A\|_{\mathcal{L}} \leq 1$. (Hier geht die asymptotische Stabilität ein, da ψ das charakteristische Polynom von A ist.) Da auf dem $\mathbb{R}^r$ alle Normen äquivalent sind, gibt es eine Konstante $c > 0$ mit

$$\frac{1}{c} \|u\|_\infty \leq \|u\|_A \leq c \|u\|_\infty$$

für alle $u \in \mathbb{R}^r$. Damit folgt

$$\begin{aligned}
\|E_{j+1}\|_A &\le \|E_j\|_A + h\,\|T_j\|_A + h\,\|\Lambda_j\|_A \\
&\le \|E_j\|_A + ch\,\|T_j\|_\infty + ch\,\|\Lambda_j\|_\infty \\
&\le \|E_j\|_A + ch\,\|\tau(t_j, y(t_j); h)\| \\
&\quad + ch \sum_{k=0}^{r} |b_k|\,\|f(t_{j+k}, \eta_{j+k}) - f(t_{j+k}, y(t_{j+k}))\| \\
&\le \|E_j\|_A + ch\sigma(h) + ch \sum_{k=0}^{r} |b_k|\,L\,\|e_{j+k}\| \\
&\le \|E_j\|_A + ch\sigma(h) + chL \left(\sum_{k=0}^{r-1} |b_k| \right) \|E_j\|_\infty \\
&\quad + chL\,|b_r|\,\|E_{j+1}\|_\infty \\
&\le \left(1 + c^2 hL \sum_{k=0}^{r-1} |b_k| \right) \|E_j\|_A + ch\sigma(h) + c^2 hL\,|b_r|\,\|E_{j+1}\|_A .
\end{aligned}$$

Definiere $B = \sum_{k=0}^{r} |b_k|$. Falls $c^2 hLB \le \frac{1}{2}$ ist, folgt aus obiger Abschätzung

$$\begin{aligned}
\|E_{j+1}\|_A &\le \frac{1 + c^2 hLB}{1 - c^2 hLB}\,\|E_j\|_A + \frac{ch\sigma(h)}{1 - c^2 hLB} \\
&\le (1 + 4c^2 hLB)\,\|E_j\|_A + 2ch\sigma(h).
\end{aligned}$$

Hieraus folgt durch Induktion für $n \ge 1$

$$\|E_n\|_A \le e^{nh4c^2 LB}\,\|E_0\|_A + \frac{e^{nh4c^2 LB} - 1}{4c^2 LBh} 2ch\sigma(h)$$

und somit für $n \ge r$

$$\begin{aligned}
\|e_n\| &\le \|E_n\|_\infty \\
&\le c\,\|E_n\|_A \\
&\le c^2 e^{nh4c^2 LB}\,\|E_0\|_\infty + \frac{e^{nh4c^2 LB} - 1}{2LB}\sigma(h) \\
&= c^2 e^{(t-t_0)4c^2 LB} \max_{0 \le j \le r-1} \|\eta_j - y(t_j)\| + \frac{e^{(t-t_0)4c^2 LB} - 1}{2LB}\sigma(h) \\
&= c^2 e^{(t-t_0)4c^2 LB} \left\| \varepsilon\left(t; \frac{t - t_0}{n} \right) \right\| + \frac{e^{(t-t_0)4c^2 LB} - 1}{2LB}\sigma(h).
\end{aligned}$$

Dies beweist die Konvergenz des Verfahrens. □

Bemerkung I.6.6. (1) Ein lineares 1-Schritt-Verfahren ist immer asymptotisch stabil.
(2) Mit technischem Mehraufwand kann man in Satz I.6.5 (2) die gleichmäßige Lipschitz-Stetigkeit durch Lipschitz-Stetigkeit und $\mathbb{R}^n$ durch eine offene Teilmenge U ersetzen.

Aus dem Beweis von Satz I.6.5 folgt unmittelbar:

KOROLLAR I.6.7 (Globale Fehlerabschätzung). *Das MSV sei asymptotisch stabil und habe die Ordnung p. Für die Fehlerfunktion ε gelte*

$$\|\varepsilon(t;h)\| = \mathrm{O}(h^p).$$

Dann gilt für den Fehler des MSV die Abschätzung

$$\left\| y(t) - \eta\left(t; \frac{t-t_0}{n}\right)\right\| = \mathrm{O}\left(\left|\frac{t-t_0}{n}\right|^p\right).$$

Die Stabilitätsbedingung begrenzt die maximal erreichbare Ordnung eines MSV. Dies zeigt folgendes Beispiel.

BEISPIEL I.6.8 (Stabilität begrenzt die maximal erreichbare Ordnung eines MSV). Wir betrachten ein 2-Schritt Verfahren der Form

$$\eta_{j+2} + a_1\eta_{j+1} + a_0\eta_j = hb_1 f(t_{j+1}, \eta_{j+1}) + hb_0 f(t_j, \eta_j)$$

Aus Satz I.5.5 (S. 40) folgen die Bestimmungsgleichungen

$$\begin{aligned} 1 + a_1 + a_0 &= 0 \\ 2 + a_1 - b_1 - b_0 &= 0 \\ 2 + \frac{1}{2}a_1 - b_1 &= 0 \\ \frac{4}{3} + \frac{1}{6}a_1 - \frac{1}{2}b_1 &= 0, \end{aligned}$$

die die maximale Ordnung 3 liefern. Die Lösung lautet

$$a_1 = 4, a_0 = -5, b_1 = 4, b_0 = 2.$$

Das entsprechende Polynom $\psi = z^2 + 4z - 5$ hat die Nullstellen

$$\lambda_{1,2} = -2 \pm \sqrt{4+5} = \begin{cases} 1 \\ -5 \end{cases}.$$

Also ist das Verfahren nicht stabil und damit auch nicht konvergent.

Die maximale Ordnung eines stabilen linearen r-Schritt Verfahrens wurde 1959 von Dahlquist bestimmt:

SATZ I.6.9 (Dahlquist, maximale Ordnung eines stabilen MSV). *Ein stabiles, lineares r-Schritt Verfahren hat höchstens die Ordnung $r+1$, falls r ungerade ist, bzw. $r+2$, falls r gerade ist.*

BEISPIEL I.6.10 (Trapezregel, Simpson-Regel). Die Trapezregel ist ein stabiles lineares 1-Schritt-Verfahren der Ordnung 2. Das Verfahren

$$\eta_{j+2} - \eta_j = \frac{h}{3}f(t_{j+2}, \eta_{j+2}) + \frac{4h}{3}f(t_{j+1}, \eta_{j+1}) + \frac{h}{3}f(t_j, \eta_j)$$

ist ein stabiles 2-Schritt-Verfahren der Ordnung 4. Man erhält es durch Anwenden der Simpson-Regel auf die Identität

$$y(t+2h) - y(t) = \int_t^{t+2h} f(s, y(s))\mathrm{d}s.$$

I.7. Implementierung linearer Mehrschrittverfahren

Bei den impliziten linearen MSV wie z.B. den Verfahren von Adams-Moulton oder den BDF-Formeln muss in jedem Schritt ein nichtlineares Gleichungssystem gelöst werden. Dies kann wie bei ESV mittels einiger Iterationen des Newton-Verfahrens oder mit einer Prädiktor-Korrektor-Formel geschehen. Für die Adams-Moulton-Verfahren benutzt man dabei das entsprechende Adams-Bashforth-Verfahren als Prädiktor. Dabei ist zu beachten, dass in der Notation von §I.5 das Adams-Moulton-Verfahren mit $q+1$ dem Adams-Bashforth-Verfahren mit q in dem Sinne entspricht, dass beide auf die gleiche Zahl alter Näherungen zurückgreifen. Für $r \geq 1$ Schritte ist das entsprechende Prädiktor-Korrektor-Verfahren dann gegeben durch

$$\begin{aligned}\eta_{p+1}^{(0)} &= \eta_p + h\sum_{i=0}^{r-1} \beta_{r-1,i}^{AB} f(t_{p-i}, \eta_{p-i}) \\ \eta_{p+1} &= \eta_p + h\sum_{i=1}^{r} \beta_{r,i}^{AM} f(t_{p+1-i}, \eta_{p+1-i}) + h\beta_{r,0}^{AM} f(t_{p+1}, \eta_{p+1}^{(0)}).\end{aligned}$$

Die Schrittweitensteuerung geschieht bei MSV wie bei ESV im Prinzip durch Vergleich von Rechnungen mit einem Verfahren und zwei verschiedenen Schrittweiten oder durch Vergleich von Rechnungen mit zwei verschiedenen Verfahren und identischer Schrittweite. Wegen der Kosten der Anlaufrechnung ist die zweite Möglichkeit häufig der ersten überlegen. Bei den Adams-Verfahren kann man dabei wieder das Adams-Moulton-Verfahren mit $q+1$ zur Steuerung des Adams-Bashforth-Verfahrens mit q benutzen. Sei dazu für $r \geq 1$

$$\begin{aligned}\tau_r^{AB} = \frac{1}{h}&[y(t+rh) - y(t+(r-1)h)] \\ &- \sum_{i=0}^{r-1} \beta_{r-1,i}^{AB} f(t+(r-1-i)h, y(t+(r-1-i)h))\end{aligned}$$

und

$$\begin{aligned}\tau_r^{AM} = \frac{1}{h}&[y(t+rh) - y(t+(r-1)h)] \\ &- \sum_{i=0}^{r} \beta_{r,i}^{AM} f(t+(r-i)h, y(t+(r-i)h))\end{aligned}$$

der lokale Verfahrensfehler der beiden Verfahren und η_i die mit dem Adams-Bashforth-Verfahren berechnete Näherungslösung. Dann folgt

aus Korollar I.5.8 (S. 42) und dem Beweis von Satz I.6.5 (S. 46)

$$
\begin{aligned}
&- \sum_{i=0}^{r-1} \beta_{r-1,i}^{AB} f(t+(r-1-i)h, \eta_{r-1-i}) + \sum_{i=0}^{r} \beta_{r,i}^{AM} f(t+(r-i)h, \eta_{r-i}) \\
&\quad = - \sum_{i=0}^{r-1} \beta_{r-1,i}^{AB} [f(t+(r-1-i)h, \eta_{r-1-i}) \\
&\qquad\qquad - f(t+(r-1-i)h, y(t+(r-1-i)h))] \\
&\qquad + \sum_{i=0}^{r} \beta_{r,i}^{AM} [f(t+(r-i)h, \eta_{r-i}) \\
&\qquad\qquad - f(t+(r-i)h, y(t+(r-i)h)] \\
&\qquad + \tau_r^{AB} - \tau_r^{AM} \\
&\quad = c_r^{AB} h^r - c_r^{AM} h^{r+1} + \mathrm{O}(h^{r+1}).
\end{aligned}
$$

Damit folgt für die Fehlerkonstante des Adams-Bashforth-Verfahrens in erster Ordnung

$$
\begin{aligned}
c_r^{AB} &= h^{-r} \left\{ \sum_{i=0}^{r} \beta_{r,i}^{AM} f(t+(r-i)h, \eta_{r-i}) \right. \\
&\qquad \left. - \sum_{i=0}^{r-1} \beta_{r-1,i}^{AB} f(t+(r-1-i)h, \eta_{r-1-i}) \right\} \\
&= h^{-r-1} [\eta_r - \eta_r^{(0)}].
\end{aligned}
$$

Hieraus kann die Schrittweite dann wie bei ESV gesteuert werden.

Bei Änderung der Schrittweite benötigt ein r-Schritt Verfahren r neue Startwerte. Diese können im Prinzip durch eine neue Anlaufrechnung bestimmt werden. Da dies sehr aufwändig ist, bestimmt man die neuen Startwerte häufig durch Interpolation von $p+1$ bereits berechneten Näherungen, wobei p die Ordnung des Verfahrens ist. Die neuen Startwerte haben dann die gleiche Genauigkeit wie die bisher berechneten Näherungen. Eine andere Möglichkeit ist die Verwendung von Adams-artigen Formeln mit variabler Schrittweite. Wir gehen aus Zeitgründen hierauf aber nicht näher ein.

I.8. Stabilität von Ein- und Mehrschrittverfahren

Die Ordnung eines ESV oder MSV, die wir bisher betrachtet haben, erlaubt einen asymptotischen Vergleich der Verfahren, d.h. für hinreichend kleine Schrittweiten h. Danach ist bei gleicher Ordnung ein explizites Verfahren vorzuziehen, da es pro Schritt einen geringeren Aufwand erfordert als ein implizites Verfahren. In der Praxis rechnet man aber mit einer endlichen Schrittweite, die u.U. viel größer ist als diejenige, für die die asymptotischen Aussagen gelten. Daher sind in dieser Situation die Verfahren u.U. gänzlich anders zu beurteilen, und

man benötigt andere Kriterien als die Ordnung. Ein Kriterium für die Qualität eines Verfahrens bei endlicher Schrittweite ist die *Stabilität*. Bevor wir eine Definition dieses Begriffes geben, betrachten wir ein einfaches Beispiel.

BEISPIEL I.8.1 (Wärmeleitungsgleichung). Wir betrachten einen Stab der Länge 1, der zur Zeit $t = 0$ eine gegebene Temperatur $u_0(x)$ mit $u_0(x) > 0$ für $0 < x < 1$ hat und dessen Enden auf die Temperatur 0 gekühlt werden. Physikalisch erwarten wir, dass für jedes $x \in (0,1)$ die Temperatur $u(x,t)$ eine streng monoton fallende Funktion der Zeit ist und für $t \to \infty$ gegen 0 strebt. Mathematisch wird die Temperatur durch die *Wärmeleitungsgleichung*

$$\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} &= 0 && \text{in } (0,1)\times(0,\infty) \\ u(x,t) &= 0 && \text{für } x \in \{0,1\}, t > 0 \\ u(x,0) &= u_0(x) && \text{für } x \in (0,1) \end{aligned} \tag{I.8.1}$$

beschrieben (vgl. Beispiel III.1.5 (S. 84)). Dabei haben wir der Einfachheit halber die Gleichung so skaliert, dass der Wärmeleitwert gleich 1 ist.
Zur numerischen Lösung (vgl. §III.4 (S. 109)) wählen wir $N \in \mathbb{N}^*$, setzen $\Delta x = \frac{1}{N+1}$ und approximieren $\frac{\partial^2 u}{\partial x^2}$ in den Gitterpunkten $x_i = i\Delta x$, $1 \le i \le N$, durch den zweiten zentralen Differenzenquotienten

$$\frac{\partial^2 u}{\partial x^2}(x_i,t) = \frac{1}{(\Delta x)^2}[u(x_{i-1},t) - 2u(x_i,t) + u(x_{i+1},t)] + \mathrm{O}((\Delta x)^2).$$

Mit der Notation

$$y(t) = \begin{pmatrix} u(x_1,t) \\ \vdots \\ u(x_N,t) \end{pmatrix}, y_0 = \begin{pmatrix} u_0(x_1) \\ \vdots \\ u_0(x_N) \end{pmatrix},$$

$$A = \frac{1}{(\Delta x)^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{\mathbb{N}\times N}$$

geht dann (I.8.1) über in das AWP

$$y' = Ay, \quad y(0) = y_0. \tag{I.8.2}$$

Definiere die Vektoren $z^{(k)} \in \mathbb{R}^N$ und die Zahlen $\lambda_k \in \mathbb{R}$, $1 \le k \le N$, durch

$$\begin{aligned} z_i^{(k)} &= \sin k\pi x_i && 1 \le i \le N, 1 \le k \le N, \\ \lambda_k &= \frac{2}{(\Delta x)^2}(1 - \cos k\pi\Delta x) && 1 \le k \le N. \end{aligned}$$

Mit Hilfe der Additionstheoreme rechnet man leicht nach, dass die Vektoren $z^{(k)}$ die Eigenvektoren von A zu den Eigenwerten $-\lambda_k$ sind. Daher

lautet die exakte Lösung von (I.8.2)

$$y(t) = \sum_{k=1}^{N} \alpha_k e^{-\lambda_k t} z^{(k)} \tag{I.8.3}$$

mit

$$\sum_{k=1}^{N} \alpha_k z^{(k)} = y_0.$$

Wendet man nun das explizite Euler-Verfahren mit konstanter Schrittweite Δt auf (I.8.2) an, so erhält man die Näherungslösung

$$\eta(i\Delta t; \Delta t) = \sum_{k=1}^{N} \alpha_k \left(1 - \lambda_k \Delta t\right)^i z^{(k)}. \tag{I.8.4}$$

Damit (I.8.4) für festes Δt und $i \to \infty$ das gleiche asymptotische Verhalten aufweist wie (I.8.3) für $t \to \infty$ muss offensichtlich gelten $|1 - \lambda_k \Delta t| < 1$ für alle $1 \le k \le N$. Wie man leicht nachrechnet ist

$$\begin{aligned} 0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n, \\ \lambda_1 = \frac{4}{(\Delta x)^2} \sin^2 \frac{\pi \Delta x}{2} = \pi^2 + \mathrm{O}((\Delta x)^2), \\ \lambda_N = \frac{2}{(\Delta x)^2}(1 + \cos \pi \Delta x) = \mathrm{O}((\Delta x)^{-2}). \end{aligned}$$

Also muss gelten (vgl. Lemma III.4.6 (S. 113) und Satz III.4.7 (S. 115))

$$\Delta t < 2\lambda_N^{-1} = \mathrm{O}((\Delta x)^2).$$

Dies ist eine gravierende Einschränkung an den Zeitschritt, die von der uninteressantesten, da am schnellsten abklingenden Lösungskomponente herrührt.
Wenden wir hingegen das implizite Euler-Verfahren mit konstanter Schrittweite Δt auf (I.8.2) an, so erhalten wir die Näherung

$$\eta(i\Delta t; \Delta t) = \sum_{k=1}^{N} \alpha_k \left(1 + \lambda_k \Delta t\right)^{-i} z^{(k)}. \tag{I.8.5}$$

Wegen $\lambda_k > 0$, $1 \le k \le N$, hat (I.8.5) für *jede* Schrittweite $\Delta t > 0$ das gleiche asymptotische Verhalten wie die exakte Lösung (I.8.3). Wir erhalten also selbst für große Schrittweiten eine zumindest qualitativ richtige Lösung.

Das AWP (I.8.2) ist ein Beispiel für eine *steife Dgl*: Die Komponenten der Lösung haben ein sehr stark unterschiedliches Abklingverhalten, und die Komponente, die am schnellsten abklingt und die damit am uninteressantesten ist, stellt die stärksten Einschränkungen an die Schrittweite der Diskretisierung. Das obige Beispiel zeigt auch, dass implizite Verfahren trotz des zusätzlichen Aufwandes für die Lösung der (nicht-) linearen Gleichungssysteme expliziten Verfahren gleicher

Ordnung weit überlegen sein können. Es legt das folgende *Stabilitätskriterium* zur Beurteilung der Qualität eines numerischen Verfahrens für AWPe nahe:
Bei Anwendung auf das AWP (I.6.1) (S. 42) mit $\lambda \in \mathbb{C}$ und $\operatorname{Re}\lambda < 0$ sollte das numerische Verfahren für möglichst große Schrittweiten h noch das gleiche qualitative Verhalten haben wie die exakte Lösung von (I.6.1) (S. 42).

Wir betrachten nun zunächst ein lineares r-Schritt Verfahren mit den charakteristischen Polynomen

$$\psi(z) = \sum_{k=0}^{r} a_k z^k, \quad \chi(z) = \sum_{k=0}^{r} b_k z^k.$$

Wenn wir dieses MSV auf (I.6.1) (S. 42) anwenden, erhalten wir die Differenzengleichung

$$\sum_{k=0}^{r} a_k \eta_{j+k} = \mu \sum_{k=0}^{r} b_k \eta_{j+k} \quad j = 0, 1, \ldots$$

mit $\mu = h\lambda$ und dem charakteristischen Polynom

$$\rho(\mu; z) = \psi(z) - \mu\chi(z).$$

Das obige Kriterium und Korollar I.6.2 (S. 44) führen dann auf folgende Definition.

Definition I.8.2 (Stabilitätsgebiet eines linearen MSV). Das *Stabilitätsgebiet* S eines linearen MSV mit charakteristischen Polynomen ψ und χ ist definiert durch

$$\begin{aligned} S = \{\mu \in \mathbb{C} : &\rho(\mu; z) = \psi(z) - \mu\chi(z) \\ &\text{hat nur Nullstellen mit } |z| \le 1, \\ &\text{ist außerdem } z \text{ eine Nullstelle mit } |z| = 1, \\ &\text{so ist } z \text{ einfache Nullstelle}\}. \end{aligned}$$

Beispiel I.8.3 (Stabilitätsgebiete der Euler-Verfahren und der Trapezregel). Für das explizite Euler-Verfahren ist

$$\rho(\mu; z) = z - 1 - \mu$$

und damit

$$S = \{\mu \in \mathbb{C} : |\mu + 1| \le 1\}.$$

Für das implizite Euler-Verfahren ist

$$\rho(\mu; z) = z - 1 - \mu z = (1 - \mu)z - 1$$

und damit

$$S = \left\{\mu \in \mathbb{C} : \frac{1}{|1 - \mu|} \le 1\right\} = \{\mu \in \mathbb{C} : |1 - \mu| \ge 1\}.$$

Für die Trapezregel erhalten wir schließlich

$$\rho(\mu; z) = z - 1 - \frac{\mu}{2}(z+1) = \left(1 - \frac{\mu}{2}\right) z - \left(1 + \frac{\mu}{2}\right)$$

und

$$S = \left\{ \mu \in \mathbb{C} : \left| \frac{1 + \frac{\mu}{2}}{1 - \frac{\mu}{2}} \right| \leq 1 \right\} = \{ \mu \in \mathbb{C} : \operatorname{Re} \mu \leq 0 \} = H_-.$$

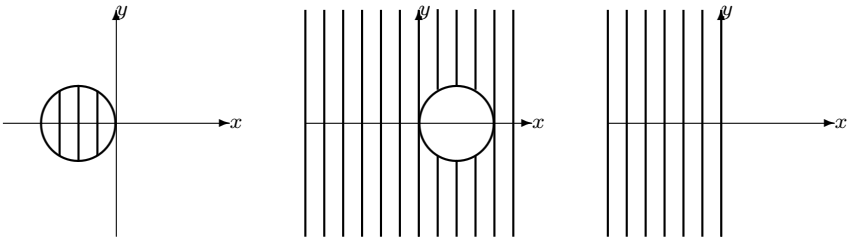

ABBILDUNG I.8.1. Stabilitätsgebiete des expliziten und impliziten Euler-Verfahrens (links und Mitte) und der Trapezregel (rechts)

DEFINITION I.8.4 (A-Stabilität). Ein lineares MSV mit Stabilitätsgebiet S heißt *absolut-stabil* oder *A-stabil*, wenn gilt $H_- = \{\mu \in \mathbb{C} : \operatorname{Re} \mu \leq 0\} \subset S$.

Ein A-stabiles MSV angewandt auf (I.6.1) (S. 42) liefert also für *jede* Schrittweite h eine zumindest qualitativ richtige Näherungslösung.

Aus Beispiel I.8.3 folgt:

SATZ I.8.5 (A-Stabilität des impliziten Euler-Verfahrens und der Trapezregel). *Das implizite Euler-Verfahren und die Trapezregel sind A-stabil.*

Wegen der günstigen qualitativen Eigenschaften A-stabiler Verfahren ist man daran interessiert, A-stabile Verfahren möglichst hoher Ordnung zu konstruieren. Es gilt jedoch:

SATZ I.8.6 (Dahlquist, maximale Ordnung A-stabiler linearer Mehrschrittverfahren). *Ein A-stabiles lineares MSV hat höchstens die Ordnung* 2.

Wegen dieses negativen Resultates hat man verschiedene Stabilitätsbegriffe, die die A-Stabilität abschwächen, eingeführt. Zwei der wichtigsten geben wir in folgender Definition an.

DEFINITION I.8.7 (A(α)-stabil, steif-stabil). (1) Ein lineares MSV heißt *A(α)-stabil* mit $\alpha \geq 0$, falls für das Stabilitätsgebiet S gilt

$$S_\alpha = \{\mu \in \mathbb{C} : \mu = Re^{i\omega}, R \geq 0, \omega \in \mathbb{R}, |\omega - \pi| \leq \alpha\} \subset S.$$

(2) Ein lineares MSV heißt *steif-stabil*, wenn es Zahlen $D \geq 0$ und $R > 0$ gibt, so dass für das Stabilitätsgebiet S gilt

$$S_{D,R} = \{\mu \in \mathbb{C} : \operatorname{Re}\mu \leq -D\} \cup \{\mu \in \mathbb{C} : -D \leq \operatorname{Re}\mu \leq 0, |\operatorname{Im}\mu| \leq R\} \subset S.$$

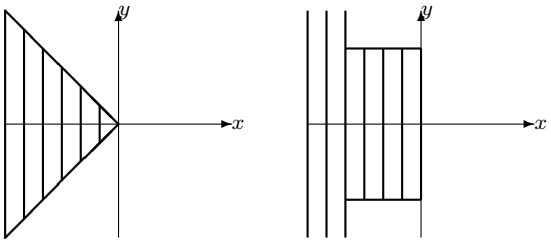

ABBILDUNG I.8.2. Die Mengen $S_{\frac{\pi}{4}}$ (links) und $S_{1,1}$ (rechts)

Um bei Anwendung auf (I.8.2) eine qualitativ richtige Lösung zu erhalten, reicht offensichtlich die A(0)-Stabilität aus.

BEISPIEL I.8.8 (Stabilitätsgebiete der Adams-Verfahren). Für die Verfahren von Adams-Bashforth (AB) und Adams-Moulton (AM) mit r-Schritten gilt $[-q_r, 0] \subset S$. Die Zahlen q_r sind in Abhängigkeit von r in Tabelle I.8.1 zusammengefasst.

TABELLE I.8.1. Abschnitte $[-q_r, 0]$ im Stabilitätsgebiet der Adams-Verfahren

r	1	2	3	4	5
AB	2	1	$\frac{6}{11}$	$\frac{3}{10}$	$\frac{90}{551}$
AM	∞	6	3	$\frac{90}{49}$	$\frac{45}{38}$

BEISPIEL I.8.9 (Stabilitätsgebiete der BDF-Formeln). Die BDF-Formeln mit m Schritten sind A(α_m)-stabil. Außerdem gilt $\{\mu \in \mathbb{C} : \operatorname{Re}\mu \leq -D_m\} \subset S$. Die Zahlen α_m und D_m sind in Abhängigkeit von m in Tabelle I.8.2 zusammengefasst. Sie zeigt, dass diese Formeln für $m = 3, 4$ ein befriedigendes Stabilitätsverhalten haben.

TABELLE I.8.2. Öffnungswinkel und Abschnitte auf der negativen Halbachse im Stabilitätsgebiet der BDF-Formeln

m	3	4	5	6
α_m	88°2′	73°21′	51°50′	17°50′
D_m	0.083	0.667	2.327	6.075

Zum Abschluss gehen wir noch kurz auf die Stabilität der Runge-Kutta-Verfahren ein. Anwenden des r-stufigen Runge-Kutta-Verfahrens

$$\begin{aligned}
\eta_0 &= y_0 \\
\eta_{i,j} &= \eta_i + h\sum_{k=1}^{r} a_{jk} f(t_i + c_k h, \eta_{i,k}) \quad 1 \le j \le r \\
\eta_{i+1} &= \eta_i + h\sum_{k=1}^{r} b_k f(t_i + c_k h, \eta_{i,k}) \\
t_{i+1} &= t_i + h \quad i = 0, 1, \ldots
\end{aligned}$$

auf das AWP (I.6.1) (S. 42) liefert mit $\mu = \lambda h$

$$\begin{aligned}
\eta_0 &= y_0 \\
\eta_{i,j} &= \eta_i + \mu\sum_{k=1}^{r} a_{j,k}\eta_{i,k} \\
\eta_{i+1} &= \eta_i + \mu\sum_{k=1}^{r} b_k \eta_{i,k}.
\end{aligned}$$

Mit den Vektoren

$$b = (b_1, \ldots, b_r)^T, \quad e = (1, \ldots, 1)^T, \quad \widetilde{\eta}_i = (\eta_{i,1}, \ldots, \eta_{i,r})^T$$

und der Matrix

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \ldots & a_{1r} \end{pmatrix}$$

folgt

$$\widetilde{\eta}_i = \eta_i e + \mu A\widetilde{\eta}_i, \quad \eta_{i+1} = \eta_i + \mu b^T \widetilde{\eta}_i$$

und somit

$$\eta_{i+1} = [1 + \mu b^T (I - \mu A)^{-1} e]\eta_i, \tag{I.8.6}$$

wobei $I - \mu A$ für $|\mu|$ hinreichend klein sicherlich regulär ist. Aus (I.8.6) folgt:

DEFINITION I.8.10 (Stabilitätsgebiet und absolute Stabilität eines Runge-Kutta-Verfahrens). (1) Das *Stabilitätsgebiet* S eines r-stufigen Runge-Kutta-Verfahrens ist definiert durch

$$S = \left\{\mu \in \mathbb{C} : I - \mu A \text{ ist regulär und } \left|1 + \mu b^T (I - \mu A)^{-1} e\right| \leq 1\right\}.$$

(2) Das r-stufige Runge-Kutta-Verfahren heißt *absolut-stabil* oder *A-stabil*, wenn gilt $H_- \subset S$.

Man überlegt sich leicht, dass die Funktion

$$g(z) = 1 + z b^T (I - zA)^{-1} e$$

rational ist und dass die Zähler- und Nennerpolynome den Grad $\leq r$ haben. Insbesondere ist für explizite Runge-Kutta-Verfahren g ein Polynom vom Grade $\leq r$. Hieraus folgt:

KOROLLAR I.8.11 (Stabilitätsgebiet expliziter Runge-Kutta-Verfahren). *Das Stabilitätsgebiet eines expliziten Runge-Kutta-Verfahrens ist beschränkt.*

Man kann jedoch zeigen, dass es implizite, absolut stabile Runge-Kutta-Verfahren beliebiger Ordnung gibt. Statt eines Beweises geben wir zwei Beispiele an.

BEISPIEL I.8.12 (A-stabile Runge-Kutta-Verfahren). (1) Das zweistufige Runge-Kutta-Verfahren

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

ist absolut stabil und hat die Ordnung 4.
(2) Das dreistufige Runge-Kutta-Verfahren

$$\begin{array}{c|ccc} \frac{5-\sqrt{15}}{10} & \frac{5}{36} & \frac{10-3\sqrt{15}}{45} & \frac{25-6\sqrt{15}}{180} \\ \frac{1}{2} & \frac{10+3\sqrt{15}}{72} & \frac{2}{9} & \frac{10-3\sqrt{15}}{72} \\ \frac{5+\sqrt{15}}{10} & \frac{25+6\sqrt{15}}{180} & \frac{10+3\sqrt{15}}{45} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

ist absolut stabil und hat die Ordnung 6.

Am Ende von § I.1 haben wir kurz die Bedeutung des Langzeitverhaltens von Lösungen eines AWP unter Störungen der Anfangswerte und den damit verbundenem Begriff der Ljapunov-Stabilität erwähnt. Offensichtlich ist es wünschenswert, dass sich die Ljapunov-Stabilität eines AWP auf das numerische Lösungsverfahren überträgt. Aus Zeitgründen können wir hier auf diesen Punkt nicht näher eingehen und verweisen stattdessen auf [4, §6]. Grob gesprochen, sind unter diesem

Aspekt L-stabile bzw. L(α)-stabile MSV oder Runge-Kutta-Verfahren besonders günstig. Dabei heißt ein MSV oder Runge-Kutta-Verfahren *L-stabil* bzw. *L(α)-stabil*, wenn es A-stabil bzw. A(α)-stabil ist und der unendlich ferne Punkt im Innern des Stabilitätsgebietes liegt. Das implizite Euler-Verfahren ist z.B. L-stabil; bei der Trapezregel dagegen liegt der unendlich ferne Punkt auf dem Rand des Stabilitätsgebietes.

I.9. Algebro-Differentialgleichungen

BEISPIEL I.9.1 (Schwingendes Pendel). Die Bewegungsgleichungen eines schwingenden Pendels der Masse m und der Länge L lauten in kartesischen Koordinaten (x_1, x_2):

$$mx_1'' = -\lambda m x_1 \tag{I.9.1a}$$

$$mx_2'' = -\lambda m x_2 - mg \tag{I.9.1b}$$

$$L^2 = x_1^2 + x_2^2. \tag{I.9.1c}$$

Die linke Seite von (I.9.1a), (I.9.1b) beschreibt die beschleunigende Kraft, der λ-freie Term auf der rechten Seite von (I.9.1b) ist die Gravitationskraft, und die λ-Terme auf der rechten Seite von (I.9.1a), (I.9.1b) beschreiben die Kräfte, die aufgebracht werden müssen, um die Zwangsbedingung (I.9.1c) zu erfüllen.
Problem (I.9.1) ist eine sogenannte *Algebro-Differentialgleichung* und stellt eine Kombination aus gDgl (I.9.1a), (I.9.1b) und aus algebraischer Gleichung (I.9.1c) dar. Dieses Problem passt offensichtlich nicht in den bisher behandelten Rahmen. Durch geeignete Umformungen kann allerdings (I.9.1) in den bisherigen Rahmen gepresst werden. Dazu differenzieren wir (I.9.1a), (I.9.1b) einmal und (I.9.1c) dreimal und erhalten die Beziehungen

$$mx_1''' = -\lambda m x_1' - \lambda' m x_1 \tag{I.9.2a}$$

$$mx_2''' = -\lambda m x_2' - \lambda' m x_2 \tag{I.9.2b}$$

und

$$0 = 2x_1x_1' + 2x_2x_2' \tag{I.9.2c}$$

$$0 = x_1x_1'' + x_1'^2 + x_2x_2'' + x_2'^2 \tag{I.9.2d}$$

$$0 = x_1x_1''' + 3x_1'x_1'' + x_2x_2''' + 3x_2'x_2''. \tag{I.9.2e}$$

Multiplikation von (I.9.2a) mit x_1 und von (I.9.2b) mit x_2, Addition der Gleichungen, Einsetzen in (I.9.2e) und Ausnutzen von (I.9.2a), (I.9.2b)

und (I.9.1a), (I.9.1b) liefert

$$\begin{aligned}
0 &= \frac{1}{m}\left\{x_1 m x_1''' + x_2 m x_2'''\right\} + 3x_1'x_1'' + 3x_2'x_2'' \\
&= \frac{1}{m}\left\{\underbrace{-\lambda m x_1 x_1' - \lambda m x_2 x_2'}_{=0} \underbrace{-\lambda' m x_1^2 - \lambda' m x_2^2}_{=-\lambda' m L^2}\right\} + 3x_1'x_1'' + 3x_2'x_2'' \\
&= -\lambda' L^2 + \frac{3}{m}\left\{x_1' m x_1'' + x_2' m x_2''\right\} \\
&= -\lambda' L^2 + \frac{3}{m}\left\{\underbrace{-\lambda m x_1 x_1' - \lambda m x_2 x_2'}_{=0} - m g x_2'\right\} \\
&= -\lambda' L^2 - 3g x_2'.
\end{aligned}$$

Insgesamt erhalten wir also die gDgl 2. Ordnung

$$\begin{aligned}
m x_1'' &= -\lambda m x_1 \\
m x_2'' &= -\lambda m x_2 - mg \\
L^2\lambda' &= -3g x_2'.
\end{aligned} \tag{I.9.3}$$

Nach Übergang zu dem äquivalenten System 1. Ordnung ist Problem (I.9.3) offensichtlich von dem bisher betrachteten Typ. Der Übergang von (I.9.1) zu (I.9.3) ist wegen der Differentiation allerdings mit einem Informationsverlust verbunden: Die Gleichung (I.9.1c) enthält offensichtlich mehr Information als die daraus durch Differentiation gewonnene Gleichung (I.9.2c).

Um einen kurzen Einblick in die numerische Behandlung von Algebro-Differentialgleichungen zu geben, betrachten wir im folgenden das Problem

$$x' = f(t, x(t), y(t)) \tag{I.9.4a}$$

$$0 = g(t, x(t), y(t)) \tag{I.9.4b}$$

$$(x(t_0), y(t_0)) = (x_0, y_0). \tag{I.9.4c}$$

Dabei sind $x : I \to U$, $y : I \to V$, $f : I \times U \times V \to \mathbb{R}^m$, $g : I \times U \times V \to \mathbb{R}^n$ hinreichend glatte Funktionen auf geeigneten offenen Mengen $I \subset \mathbb{R}$, $U \subset \mathbb{R}^m$, $V \subset \mathbb{R}^n$. Damit (I.9.4) lösbar ist müssen die Anfangswerte (x_0, y_0) offensichtlich die *Konsistenzbedingung*

$$0 = g(t_0, x_0, y_0)$$

erfüllen. Um technische Schwierigkeiten zu vermeiden, setzen wir voraus, dass g stetig differenzierbar ist und dass gilt

$$D_y g(t_0, x_0, y_0) \in \mathrm{Isom}(\mathbb{R}^n, \mathbb{R}^n). \tag{I.9.5}$$

Wenn diese Voraussetzung erfüllt ist, sagt man auch, die Algebro-Differentialgleichung (I.9.4) hat den *Index* 1.

Beispiel I.9.1 erfüllt Bedingung (I.9.5) nicht; es ist ein Problem vom Index 3. Allerdings folgt aus (I.9.2d) und (I.9.1)

$$\begin{aligned} 0 &= \frac{1}{m}\{x_1 m x_1'' + x_2 m x_2''\} + x_1'^2 + x_2'^2 \\ &= \frac{1}{m}\left\{\underbrace{-\lambda m x_1^2 - \lambda m x_2^2}_{=-\lambda m L^2} - mgx_2\right\} + x_1'^2 + x_2'^2 \\ &= -\lambda L^2 - gx_2 + x_1'^2 + x_2'^2. \end{aligned}$$

Ersetzen wir (I.9.1c) durch diese Bedingung, so erhalten wir ein Problem vom Index 1.

Aus (I.9.5) und dem Satz über implizite Funktionen folgt, dass es eine Umgebung $\widetilde{I}\times\widetilde{U}$ von (t_0, x_0) und eine stetig differenzierbare Funktion $\varphi : \widetilde{I} \times \widetilde{U} \to V$ gibt mit

$$\varphi(t_0, x_0) = y_0$$

und

$$g(t, x, y) = 0, (t, x, y) \in \widetilde{I} \times \widetilde{U} \times V \iff y = \varphi(t, x).$$

Also ist (I.9.5) äquivalent zu dem AWP

$$x' = f(t, x(t), \varphi(t, x(t))), \quad x(t_0) = x_0.$$

Auf dieses Problem können die bisher betrachteten numerischen Verfahren angewandt werden. Außer in trivialen Spezialfällen ist diese Vorgehensweise allerdings nicht praktikabel, da die Funktion φ nicht explizit bekannt ist.

Differentiation von (I.9.4b) liefert

$$0 = D_t g(t, x, y) + D_x g(t, x, y)x' + D_y g(t, x, y)y'.$$

Wegen (I.9.5) ist in einer Umgebung $\widetilde{I} \times \widetilde{U} \times \widetilde{V}$ von (t_0, x_0, y_0) die Ableitung $D_y g \in \mathrm{Isom}(\mathbb{R}^n, \mathbb{R}^n)$. Daher können wir auf dieser Umgebung (I.9.4a), (I.9.4b) in die gDgl

$$\begin{aligned} x' &= f(t, x, y) \\ y' &= -D_y g(t, x, y)^{-1}\left[D_t g(t, x, y) + D_x(t, x, y)\underbrace{f(t, x, y)}_{=x'}\right] \end{aligned}$$

transformieren. Auf dieses Problem können wieder die bisher betrachteten Verfahren angewandt werden. Allerdings müssen wir wegen der Differentiation von (I.9.4b) mit einem Informationsverlust rechnen.

Die besten Ergebnisse erzielt man in der Praxis, indem man eines der gebräuchlichen Verfahren auf den gDgl-Anteil (I.9.4a) anwendet und in jedem Schritt auf die Nebenbedingung (I.9.4b) projiziert. Wir wollen diese Vorgehensweise an Hand des impliziten Euler-Verfahrens

und der Trapezregel erläutern. Dazu bezeichnen wir die Approximationen an $x(t_i)$ bzw. $y(t_i)$ mit η_i bzw. μ_i. Dann lautet das implizite Euler-Verfahren für (I.9.4)

$$\begin{aligned}
\eta_0 &= x_0, \\
\mu_0 &= y_0 \\
\eta_{i+1} &= \eta_i + h_i f(t_{i+1}, \eta_{i+1}, \mu_{i+1}) \\
0 &= g(t_{i+1}, \eta_{i+1}, \mu_{i+1}) \\
t_{i+1} &= t_i + h_i
\end{aligned}$$

und die Trapezregel

$$\begin{aligned}
\eta_0 &= x_0, \\
\mu_0 &= y_0 \\
\eta_{i+1} &= \eta_i + \frac{1}{2} h_i \left[f(t_{i+1}, \eta_{i+1}, \mu_{i+1}) + f(t_i, \eta_i, \mu_i) \right] \\
0 &= g(t_{i+1}, \eta_{i+1}, \mu_{i+1}) \\
t_{i+1} &= t_i + h_i.
\end{aligned}$$

Bei beiden Verfahren ist in jedem Schritt ein nichtlineares Gleichungssystem der Form

$$\begin{aligned}
\eta + \theta h f(t, \eta, \mu) &= F \\
g(t, \eta, \mu) &= 0
\end{aligned}$$

zu lösen. Dabei ist für das implizite Euler-Verfahren $\theta = 1$ und für die Trapezregel $\theta = \frac{1}{2}$. Die Funktionalmatrix hat die Struktur

$$\begin{pmatrix} I + \theta h D_x f & \theta h D_y f \\ D_x g & D_y g \end{pmatrix}.$$

Wegen (I.9.5) ist sie für hinreichend kleines h regulär, so dass beide Verfahren durchführbar sind.

KAPITEL II

Randwertprobleme für gewöhnliche Differentialgleichungen

Wir betrachten in diesem Kapitel Randwertprobleme. Wie im vorigen Kapitel ist dabei eine Lösung einer gewöhnlichen Differentialgleichung gesucht. Aber anders als dort muss diese Lösung nicht einen vorgegebenen Wert an *einem* vorgegebenem Punkt annehmen (die Anfangsbedingung), sondern einen vorgegebenen funktionalen Zusammenhang zwischen den Werten an *zwei* verschiedenen Punkten erfüllen (die Randbedingung). Die einfachsten Beispiele für Randwertprobleme sind die Schwingung einer eingespannten Saite und die Flugbahn eines Basketballes bei einem Freiwurf.

In §II.1 beschreiben wir die Klasse der betrachteten Randwertprobleme zunächst genauer und zeigen dann, dass auch andere Probleme wie z.B. Eigenwertprobleme oder freie Randwertprobleme in die beschriebene Form passen. Danach stellen wir einige theoretische Ergebnisse für Randwertprobleme vor und zeigen, dass für diese Probleme kein allgemeiner Existenz- und Eindeutigkeitsatz wie der Satz von Picard-Lindelöf, Satz I.1.8 (S. 8), für Anfangswertprobleme existiert.

In §II.2 und §II.3 beschreiben wir zwei Verfahren zur Lösung von Randwertproblemen. Beide beruhen auf dem Ansatz, Anfangswertprobleme zu lösen und den Wert der Lösung am zweiten Randpunkt mit der Randbedingung abzugleichen. Dies führt auf ein nichtlineares Gleichungssystem, das mit dem Newton-Verfahren gelöst wird. Zur Berechnung der Jacobi-Matrix braucht man dann insbesondere die differenzierbare Abhängigkeit der Lösung eines Anfangswertproblemes vom Anfangswert, Satz I.1.14 (S. 12). Die beiden Verfahren unterscheiden sich darin, dass bei demjenigen aus §II.3 das Intervall, über das die Differentialgleichung zu lösen ist, in kleine Teilintervalle unterteilt wird und in jedem Teilintervall separat ein Anfangswertproblem gelöst wird. Dies erlaubt einerseits die Parallelisierung und Beschleunigung des Verfahrens und andererseits seine Stabilisierung, da gemäß Satz I.1.13 (S. 12) Lösungen zu verschiedenen Anfangswerten exponentiell mit der Länge des Lösungsintervalls auseinander laufen.

In §II.4 betrachten wir schließlich Differenzenverfahren für eine spezielle Klasse von Randwertproblemen, die Sturm-Liouville-Probleme. Diese Probleme können auch mit den Verfahren der §§II.2 und II.3 gelöst werden, aber die Differenzenverfahren sind auf ihre spezielle

Struktur zugeschnitten und daher effizienter. Außerdem bereitet dieser Abschnitt die Techniken und Ergebnisse von §III.3 über Differenzenverfahren für elliptische Diferentialgleichungen vor, da die Sturm-Liouville-Probleme deren eindimensionale Analoga sind.

II.1. Einige theoretische Ergebnisse

Wir betrachten in diesem Kapitel *Randwertprobleme* kurz *RWP* für gewöhnliche Differentialgleichungen der folgenden Form:
Gegeben sind zwei Zahlen $a, b, \in \mathbb{R}$ mit $a < b$, eine stetige Funktion $f : (a, b) \times \mathbb{R}^n \to \mathbb{R}^n$ und eine stetige Funktion $r : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. Gesucht ist eine Funktion $y : [a, b] \to \mathbb{R}^n$, die die gewöhnliche Differentialgleichung

$$y' = f(t, y(t)) \tag{II.1.1a}$$

in (a, b) löst und die Randbedingung

$$r(y(a), y(b)) = 0 \tag{II.1.1b}$$

erfüllt.

Ein Spezialfall sind *lineare RWP*. In diesem Fall ist

$$r(y(a), y(b)) = Ay(a) + By(b) - c$$

mit $A, B \in \mathbb{R}^{n\times n}$ und $c \in \mathbb{R}^n$. Häufig sind die Randbedingungen *entkoppelt*, d.h. es ist

$$A = \begin{pmatrix} A_1 \\ 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ B_2 \end{pmatrix}$$

mit $A_1 \in \mathbb{R}^{m\times n}$, $B_2 \in \mathbb{R}^{(n-m)\times n}$.

Eine Reihe anderer Probleme lassen sich formal als RWP schreiben, so dass die Methoden dieses Kapitels angewandt werden können.

Beispiel II.1.1 (Eigenwertproblem). Gesucht sind eine Funktion $u : [a, b] \to \mathbb{R}^m$ und eine Zahl $\lambda \in \mathbb{R}$ mit

$$\begin{aligned} u' &= \widetilde{f}(t, u(t)) \quad (\widetilde{f} : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m) \\ \widetilde{r}(u(a), u(b), \lambda) &= 0 \qquad\quad (\widetilde{r} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^{m+1}). \end{aligned}$$

Dieses *Eigenwertproblem* lässt sich in der Form (II.1.1) schreiben mit $n = m + 1$ und

$$\begin{aligned} y_i &= u_i \quad 1 \le i \le m \\ y_{m+1} &= \lambda \\ f(t, y) &= (\widetilde{f}_1(t, \pi y), \dots, \widetilde{f}_m(t, \pi y), 0)^T \\ r(v, w) &= \widetilde{r}(\pi v, w) \\ \pi &: \mathbb{R}^n \to \mathbb{R}^m \quad y \mapsto (y_1, \dots, y_m)^T. \end{aligned}$$

BEISPIEL II.1.2 (Freies Randwertproblem). Zu gegebenem a ist eine Zahl b mit $b > a$ und eine Funktion $u : [a, b] \to \mathbb{R}^m$ mit

$$\begin{aligned} u' &= \widetilde{f}(s, u(s)) \quad (\widetilde{f} : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m) \\ \widetilde{r}(u(a), u(b)) &= 0 \qquad\qquad (\widetilde{r} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^{m+1}) \end{aligned}$$

gesucht. Dieses *freie Randwertproblem* lässt sich in der Form (II.1.1) schreiben mit $n = m + 1$ und

$$\begin{aligned} y_i &= u_i \quad 1 \le i \le m \\ y_{m+1} &= b - a \\ t &= \frac{s - a}{y_{m+1}} \\ f(t, y) &= (y_n \widetilde{f}_1(a + t y_n, \pi y), \ldots, y_n \widetilde{f}_m(a + t y_n, \pi y), 0)^T \\ r(u, v) &= \widetilde{r}(\pi u, \pi v) \\ \pi &: \mathbb{R}^n \to \mathbb{R}^m \quad y \mapsto (y_1, \ldots, y_m)^T. \end{aligned}$$

Die Frage nach der Existenz und Eindeutigkeit einer Lösung des RWP (II.1.1) lässt sich nicht so leicht beantworten wie beim AWP. Dies zeigt das folgende Beispiel.

BEISPIEL II.1.3 (Schlecht gestelltes RWP). Betrachte das RWP

$$y' = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} y, \quad A y(a) + B y(b) - c = 0.$$

Die allgemeine Lösung der Differentialgleichung lautet

$$y = \begin{pmatrix} \gamma_1 \cos x + \gamma_2 \sin x \\ -\gamma_1 \sin x + \gamma_2 \cos x \end{pmatrix}.$$

Für

$$a = 0, b = \pi, A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, c = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

erhalten wir die widersprüchlichen Bedingungen $\gamma_1 = 0$ und $\gamma_1 = 1$, so dass das RWP keine Lösung hat.
Für

$$a = 0, b = \pi, A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

erhalten wir die Bedingung $\gamma_1 = 0$, so dass alle Funktionen $y = \left(\begin{smallmatrix} \gamma \sin x \\ \gamma \cos x \end{smallmatrix}\right)$ Lösung des RWP sind.

Beispiel II.1.3 zeigt, dass es keinen allgemeinen Existenz- und Eindeutigkeitssatz für RWP geben kann. Wir geben daher einen solchen Satz nur unter speziellen Voraussetzungen an und gehen nicht auf mögliche Abschwächungen der Voraussetzungen ein.

SATZ II.1.4 (Existenz- und Eindeutigkeitssatz für RWP). *Die folgenden Voraussetzungen seien erfüllt:*

(1) *f ist in $C^1((a,b)\times\mathbb{R}^n,\mathbb{R}^n)$ und r ist stetig differenzierbar.*
(2) *Die Matrix*

$$P(u,v) = D_u r(u,v) + D_v r(u,v)$$

besitzt für alle $u,v\in\mathbb{R}^n$ eine Zerlegung der Form

$$P(u,v) = P_0[I + M(u,v)]$$

mit einer von u,v unabhängigen, regulären Matrix P_0 und es gibt zwei Konstanten $\widetilde{m}\neq 0$ und $0\le\mu<1$ mit

$$\|M(u,v)\|_{\mathcal{L}} \le \mu, \quad \left\|P_0^{-1}D_v r(u,v)\right\|_{\mathcal{L}} \le \widetilde{m}$$

für alle $u,v\in\mathbb{R}^n$.
(3) *Es gibt eine Zahl λ mit $0<\lambda<1-\mu$ und eine Funktion $k\in C([a,b],\mathbb{R}_+)$ mit*

$$\|D_y f(t,y)\|_{\mathcal{L}} \le k(t)$$

für alle $t\in[a,b]$ und $y\in\mathbb{R}^n$ und

$$\int_a^b k(t)\mathrm{d}x \le \ln\left(1+\frac{\lambda}{\widetilde{m}}\right).$$

Dann besitzt das RWP (II.1.1) eine eindeutige Lösung.

Beweis. Für $s\in\mathbb{R}^n$ bezeichne mit $y(t;s)$ und $z(t;s)$ die eindeutigen Lösungen der AWP

$$\begin{aligned} y' &= f(t,y(t;s)), & y(a;s) &= s\\ Z' &= D_y f(t,y(t;s))z(t;s), & Z(a;s) &= I. \end{aligned}$$

Gemäß Satz I.1.14 (S. 12) ist $Z(t;s) = D_s y(t;s)$.
Definiere die Funktionen $F:\mathbb{R}^n\to\mathbb{R}^n$ und $\Phi:\mathbb{R}^n\to\mathbb{R}^n$ durch

$$F(s) = r(s,y(b;s)), \quad \Phi(s) = s - P_0^{-1}F(s).$$

Offensichtlich ist $y(\cdot;\alpha)$ genau dann eine Lösung des RWP (II.1.1), wenn α ein Fixpunkt von Φ ist. Daher genügt es zu zeigen, dass Φ eine Kontraktion ist. Mit der Kettenregel folgt

$$\begin{aligned} D\Phi(s) &= I - P_0^{-1}[D_u r(s,y(b;s)) + D_v r(s,y(b;s))D_s y(b;s)]\\ &= I - P_0^{-1}[D_u r(s,y(b;s)) + D_v r(s,y(b;s))Z(b;s)]\\ &= I - P_0^{-1}[D_u r(s,y(b;s)) + D_v r(s,y(b;s))]\\ &\quad - P_0^{-1}D_v r(s,y(b;s))[Z(b;s)-I]\\ &= -M(s,y(b;s)) - P_0^{-1}D_v r(s,y(b;s))[Z(b;s)-I]. \end{aligned}$$

Aus den Voraussetzungen und Satz I.1.15 (S. 13) folgt

$$\begin{aligned}\|D\Phi(s)\|_{\mathcal{L}} &\le \|M(s,y(b;s))\|_{\mathcal{L}} + \left\|P_0^{-1}D_v r(s,y(b;s))\right\|_{\mathcal{L}} \|Z(b;s)-I\|_{\mathcal{L}} \\ &\le \mu + \widetilde{m}\left\{\exp\left[\int_a^b k(t)\mathrm{d}t\right]-1\right\} \\ &\le \mu + \widetilde{m}\left\{1+\frac{\lambda}{\widetilde{m}}-1\right\} \\ &= \lambda+\mu \\ &< 1.\end{aligned}$$

Hieraus folgt die Behauptung. □

Bemerkung II.1.5 (Schlecht gestelltes RWP). Für das RWP aus Beispiel II.1.3 ist $P(u,v) = \left(\begin{smallmatrix}1 & 0\\ 1 & 0\end{smallmatrix}\right)$. Wie man leicht nachprüft sind die Voraussetzungen von Satz II.1.4 nicht erfüllt.

II.2. Das Schießverfahren

Wir betrachten das RWP (II.1.1) (S. 64). Für $s \in \mathbb{R}^n$ sei $y(\cdot;s)$ die Lösung des AWP

$$y' = f(t,y(t;s)), \quad y(a;s) = s. \tag{II.2.1}$$

Definiere die Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$ durch

$$F(s) = r(s,y(b;s)).$$

Offensichtlich ist $y(\cdot;s)$ genau dann eine Lösung von (II.1.1), wenn s eine Nullstelle von F ist. Wegen Satz I.1.14 (S. 12) ist die Jacobi-Matrix von F gegeben durch

$$DF(s) = D_u r(s,y(b;s)) + D_v r(s,y(b;s))Z(b;s),$$

wobei $Z(t;s)$ das AWP

$$Z' = D_y f(t,y(t;s))Z(t;s), \quad Z(a;s) = I \tag{II.2.2}$$

löst. Die Idee des Schießverfahrens besteht darin, die AWP (II.2.1) und (II.2.2) mit einem der Verfahren aus Kapitel I näherungsweise zu lösen, wobei in (II.2.2) statt y die berechnete Näherung für y benutzt wird, damit F und DF zu approximieren und ein Newton-Verfahren zur Berechnung der Nullstelle von F anzuwenden. Dies führt auf Algorithmus II.2.1.

Bemerkung II.2.1 (Aufwand des Schießverfahrens). (1) In jeder Iteration von Algorithmus II.2.1 müssen zwei AWP der Größe n bzw. n^2 gelöst werden.
(2) Das Newton-artige Verfahren in Algorithmus II.2.1 kann mit den Methoden von [9, §III.3] modifiziert werden. Insbesondere kann die Matrix $D^{(i)}$ während mehrerer Schritte festgehalten werden, was wegen der Größe des AWP in Schritt (2) den Aufwand erheblich vermindert,

Algorithmus II.2.1 Schießverfahren

Gegeben: Startwert $s \in \mathbb{R}^n$, Toleranz ε, maximale Iterationszahl N
Gesucht: Näherung $\eta(\cdot; h)$ für die Lösung des RWP (II.1.1) (S. 64)

1: $e \leftarrow \infty$, $i \leftarrow 0$
2: **while** $e > \varepsilon$ und $i < N$ **do**
3: Berechne eine Näherung $\eta(\cdot; h)$ für die Lösung des AWP
$$y' = f(t, y(t)), \quad y(a) = s.$$
4: $F \leftarrow r(s, \eta(b; h))$, $e \leftarrow \|F\|$
5: Berechne eine Näherung $\zeta(\cdot; h)$ für die Lösung des AWP
$$Z' = D_y f(t, \eta(t; h))Z(t), \quad Z(a) = I.$$
6: $D \leftarrow D_u r(s, \eta(b; h)) + D_v r(s, \eta(b; h))\zeta(b; h)$
7: Löse das lineare Gleichungssystem $D\Delta s = -F$.
8: $s \leftarrow s + \Delta s$, $i \leftarrow i + 1$
9: **end while**

und es kann und sollte eine Dämpfung der Newton-Schritte eingebaut werden.

Unabhängig von den Schwierigkeiten, die beim Newton-Verfahren auftreten, ergeben sich bei der Durchführung von Algorithmus II.2.1 Probleme, die an der Natur der Differentialgleichungen liegen und die seine praktische Anwendbarkeit stark einschränken. Dies zeigt das folgende Beispiel.

Beispiel II.2.2 (Schlecht konditioniertes RWP). Betrachte das RWP

$$y' = \begin{pmatrix} 0 & 1 \\ 110 & 1 \end{pmatrix} y, \quad \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} y(0) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} y(10) - \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0.$$

Die allgemeine Lösung der Differentialgleichung lautet

$$y(t) = c_1 e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + c_2 e^{11t} \begin{pmatrix} 1 \\ 11 \end{pmatrix}.$$

Hieraus ergeben sich für die Lösung des RWP die Bedingungen

$$c_1 + c_2 = 1, \quad c_1 e^{-100} + c_2 e^{110} = 1$$

und somit

$$c_1 = \frac{e^{110} - 1}{e^{110} - e^{-100}}, \quad c_2 = \frac{1 - e^{-100}}{e^{110} - e^{-100}}.$$

Analog ergibt sich für $s = (s_1, s_2)^T$ als Lösung des AWP

$$y' = \begin{pmatrix} 0 & 1 \\ 110 & 1 \end{pmatrix} y, \quad y(0) = s$$

die Funktion

$$y(t; s) = \frac{11s_1 - s_2}{21} e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + \frac{10s_1 + s_2}{21} e^{11t} \begin{pmatrix} 1 \\ 11 \end{pmatrix}.$$

Der zur Lösung des RWP gehörige Parameter s^* ist

$$s^* = \begin{pmatrix} 1 \\ -10 + 21 \cdot \frac{1-e^{-100}}{e^{110}-e^{-100}} \end{pmatrix}.$$

Bei 10-stelliger Arithmetik erhalten wir im Rahmen der Rechnergenauigkeit

$$\mathrm{rd}(s^*) = \begin{pmatrix} 1(1+\varepsilon_1) \\ -10(1+\varepsilon_2) \end{pmatrix}$$

mit $|\varepsilon_i| \leq 10^{-10}$, $i = 1, 2$. Sei insbesondere $\varepsilon_1 = 0$, $\varepsilon_2 = -10^{-10}$. Dann erhalten wir für den Anfangswert

$$\widetilde{s} = \begin{pmatrix} 1 \\ -10 + 10^{-9} \end{pmatrix}$$

die Lösung

$$y(t; \widetilde{s}) = \frac{21 - 10^{-9}}{21} e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + \frac{10^{-9}}{21} e^{11t} \begin{pmatrix} 1 \\ 11 \end{pmatrix}$$

und somit

$$y_1(10; \widetilde{s}) \approx \frac{1}{21} 10^{-9} e^{110} \approx 2.8 \cdot 10^{37}.$$

Das Ergebnis von Beispiel II.2.2 ist nicht erstaunlich, wenn wir uns an Satz I.1.13 (S. 12) erinnern. Dort hatten wir für zwei Lösungen eines AWP zu verschiedenen Startwerten die Abschätzung

$$\|y(t; s_1) - y(t; s_2)\| \leq e^{L|t-t_0|} \, \|s_1 - s_2\| \tag{II.2.3}$$

gezeigt. In Beispiel II.2.2 ist $L = 110$ und $|t - t_0| = 10$, so dass wir mit einer Verstärkung des Anfangsfehlers um den Faktor e^{1100} rechnen müssen. Die Abschätzung (II.2.3) zeigt aber auch, dass wir das Phänomen aus Beispiel II.2.2 vermeiden können, wenn es uns gelingt, die Intervalllänge $|t - t_0|$ zu reduzieren. Diese Beobachtung ist Grundlage der Modifizierung des Schießverfahrens, die wir im folgenden Paragraphen betrachten.

II.3. Die Mehrzielmethode

Wir betrachten wieder das RWP (II.1.1) (S. 64), wählen m Punkte $t_1, \ldots, t_m$ mit $a = t_1 < t_2 < \ldots < t_m = b$ und bezeichnen für $s_1, \ldots, s_m \in \mathbb{R}^n$ mit $y(\cdot; t_k, s_k)$, $1 \leq k \leq m-1$, die Lösung des AWP

$$y' = f(t, y(t; t_k, s_k)), \quad y(t_k; t_k, s_k) = s_k. \tag{II.3.1}$$

Dann ist

$$\widetilde{y}(t) = \begin{cases} y(t; t_k, s_k) & \text{falls } t_k \leq t < t_{k+1}, 1 \leq k \leq m-1 \\ s_m & \text{falls } t = t_m = b \end{cases}$$

eine stückweise stetige Funktion. Konstruktionsgemäß ist die Funktion $\widetilde{y}$ genau dann eine Lösung von (II.1.1), wenn sie global stetig ist und die Randbedingungen erfüllt, d.h., wenn gilt

$$\begin{aligned} y(t_{k+1}; t_k, s_k) &= s_{k+1} \quad 1 \le k \le m-1 \\ r(s_1, s_m) &= 0. \end{aligned}$$

Also ist (II.1.1) dazu äquivalent, eine Nullstelle der Funktion $F : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ mit

$$F(s_1, \ldots, s_m) = \begin{pmatrix} y(t_2; t_1, s_1) - s_2 \\ \vdots \\ y(t_m; t_{m-1}, s_{m-1}) - s_m \\ r(s_1, s_m) \end{pmatrix}$$

zu finden. Bezeichne für $1 \le k \le m-1$ mit $Z(\cdot; t_k, s_k)$ die Lösungen der AWP

$$Z' = D_y f(t, y(t; t_k, s_k)) Z(t; t_k, s_k), \quad Z(t_k; t_k, s_k) = I. \tag{II.3.2}$$

Dann ist gemäß Satz I.1.14 (S. 12) $Z(t; t_k, s_k) = D_s y(t; t_k, s_k)$. Damit folgt

$$DF = \begin{pmatrix} Z(t_2; t_1, s_1) & -I & & & \\ & Z(t_3; t_2, s_2) & -I & 0 & \\ & \ddots & \ddots & & \\ & & & Z(t_m; t_{m-1}, s_{m-1}) & -I \\ D_u r(s_1, s_m) & 0 & \ldots & 0 & D_v r(s_1, s_m) \end{pmatrix}.$$

Wie beim Schießverfahren besteht die Idee der Mehrzielmethode darin, die AWP (II.3.1) und (II.3.2) mit einem der Verfahren aus Kapitel I zu lösen, wobei in (II.3.2) $y(\cdot; t_k, s_k)$ durch die berechnete Näherung ersetzt wird, damit die Funktionen F und DF zu approximieren und ein Newton-Verfahren zur Berechnung der Nullstelle von F zu verwenden. Dabei muss in jedem Schritt ein lineares Gleichungssystem der Form

$$DF \begin{pmatrix} \Delta s_1 \\ \vdots \\ \Delta s_m \end{pmatrix} = -F(s_1, \ldots, s_m) \tag{II.3.3}$$

gelöst werden. Dies ist ein System mit nm Gleichungen und Unbekannten. Wegen der speziellen Form von DF kann dies aber auf das Lösen eines linearen Gleichungssystems im $\mathbb{R}^n$ reduziert werden. Um dies einzusehen, schreiben wir zur Abkürzung

$$\begin{aligned} G_k &= Z(t_{k+1}; t_k, s_k), \quad 1 \le k \le m-1, \\ A &= D_u r(s_1, s_m) \\ B &= D_v r(s_1, s_m) \end{aligned}$$

Dann hat das Gleichungssystem (II.3.3) die Form

$$\begin{aligned} G_1\Delta s_1 - \Delta s_2 &= -F_1 \\ G_2\Delta s_2 - \Delta s_3 &= -F_2 \\ \vdots \qquad & \qquad \vdots \\ G_{m-1}\Delta s_{m-1} - \Delta s_m &= -F_{m-1} \\ A\Delta s_1 + B\Delta s_m &= -F_m. \end{aligned}$$

Hieraus folgt sukzessiv

$$\begin{aligned} \Delta s_2 &= F_1 + G_1\Delta s_1 \\ \Delta s_3 &= F_2 + G_2\Delta s_2 \\ &= F_2 + G_2F_1 + G_2G_1\Delta s_1 \\ \vdots \qquad & \qquad \vdots \\ \Delta s_m &= F_{m-1} + G_{m-1}\Delta s_{m-1} \\ &= \sum_{j=1}^{m-1} \left(\prod_{i=j+1}^{m-1} G_i \right) F_j + G_{m-1} \cdot \ldots \cdot G_1\Delta s_1 \end{aligned}$$

und

$$(A + BG_{m-1} \ldots G_1)\Delta s_1 = -F_m - B \sum_{j=1}^{m-1} \left(\prod_{i=j+1}^{m-1} G_i \right) F_j.$$

Das folgende Lemma zeigt, dass

$$H = A + BG_{m-1} \cdot \ldots \cdot G_1$$

unter geeigneten Annahmen an das RWP regulär ist.

LEMMA II.3.1 (Invertierbarkeit von H). *Die Voraussetzungen seien wie in Satz II.1.4 (S. 65). Dann ist H ist regulär.*

BEWEIS. Es ist

$$\begin{aligned} I - P_0^{-1}H = I - P_0^{-1} \, [&A + B + B(G_{m-1} - I) + BG_{m-1}(G_{m-2} - I) \\ &+ \ldots + BG_{m-1}G_{m-2} \cdot \ldots \cdot G_2(G_1 - I)] \\ = -M(s_1, s_m) - P_0^{-1}D_v r(s_1, s_m) &\sum_{j=1}^{m-1} \left(\prod_{i=j+1}^{m-1} G_i \right) (G_j - I). \end{aligned}$$

Aus den Voraussetzungen von Satz II.1.4 (S. 65) folgt daher

$$\left\| I - P_0^{-1}H \right\|_{\mathcal{L}} \leq \mu + \widetilde{m} \sum_{j=1}^{m-1} \left(\prod_{i=j+1}^{m-1} \|G_i\|_{\mathcal{L}} \right) \|G_j - I\|_{\mathcal{L}}.$$

Wegen Satz I.1.15 (S. 13) ist andererseits

$$\|G_i\|_{\mathcal{L}} \leq \exp \left\{ \int_{t_i}^{t_{i+1}} k(s)\mathrm{d}s \right\}$$

und

$$\|G_j - I\|_{\mathcal{L}} \leq \left[\exp\left\{\int_{t_j}^{t_{j+1}} k(s)\mathrm{d}s\right\} - 1\right].$$

Hieraus folgt

$$\begin{aligned}
&\left\|I - P_0^{-1}H\right\|_{\mathcal{L}} \\
&\leq \mu + \widetilde{m}\sum_{j=1}^{m-1} \exp\left\{\int_{t_{j+1}}^{t_m} k(s)\mathrm{d}s\right\}\left[\exp\left\{\int_{t_j}^{t_{j+1}} k(s)\mathrm{d}s\right\} - 1\right] \\
&= \mu + \widetilde{m}\sum_{j=1}^{m-1}\left[\exp\left\{\int_{t_j}^{t_m} k(s)\mathrm{d}s\right\} - \exp\left\{\int_{t_{j+1}}^{t_m} k(s)\mathrm{d}s\right\}\right] \\
&= \mu + \widetilde{m}\left\{\exp\left\{\int_a^b k(s)\mathrm{d}s\right\} - 1\right\} \\
&\leq \mu + \lambda \\
&< 1.
\end{aligned}$$

Also ist $I - (I - P_0^{-1}H) = P_0^{-1}H$ und damit auch H regulär. □

Zusammenfassend erhalten wir Algorithmus II.3.1 zur Lösung des RWP (II.1.1) (S. 64).

Bemerkung II.3.2 (Aufwand und Durchführung der Mehrzielmethode). (1) Bei der gleichen Anzahl von Gitterpunkten auf dem Gesamtintervall $[a, b]$ haben die Schritte (1) und (2) der Algorithmen II.2.1 (S. 68) und II.3.1 die gleiche Komplexität. Die AWP in den Schritten (1) und (2) von Algorithmus II.3.1 können für alle Teilintervalle $(t_k, t_{k+1}), 1 \leq k \leq m-1$, simultan gelöst werden, so dass sich auf einem Parallelrechner mit $m-1$ Prozessoren der Aufwand um den Faktor $m-1$ reduziert.
(2) Für die Durchführung des Newton-Verfahrens gelten die Bemerkungen II.2.1(2) (S. 67) analog.
(3) Falls keine zusätzlichen Informationen über die Lösung von (II.1.1) (S. 64) bekannt sind, kann man die Punkte $t_1, \ldots, t_m$ äquidistant wählen, d.h. $t_i = a + \frac{i-1}{m-1}(b-a)$.

II.4. Differenzenverfahren

Wir betrachten das *Sturm-Liouville-Problem*

$$\begin{aligned}
-(py')' + qy &= f \quad \text{in } (a,b) \\
y(a) &= \alpha \\
y(b) &= \beta
\end{aligned} \tag{II.4.1}$$

mit $p \in C^1([a,b],\mathbb{R})$, $q \in C([a,b],\mathbb{R})$ und

$$\underline{p} = \min_{a\leq x\leq b} p(x) > 0, \quad \underline{q} = \min_{a\leq x\leq b} q(x) > 0.$$

Algorithmus II.3.1 Mehrzielmethode

Gegeben: Punkte $a = t_1 < \ldots < t_m = b$, Vektoren $s_1, \ldots, s_m \in \mathbb{R}^n$, Toleranz ε, maximale Iterationszahl N

Gesucht: Näherung $\eta(\cdot; h)$ für die Lösung des RWP (II.1.1) (S. 64)

1: $e \leftarrow \infty$, $i \leftarrow 0$
2: **while** $e > \varepsilon$ und $i < N$ **do**
3: Berechne Näherungen $\eta_j(\cdot; h)$ für die Lösungen der AWP
$$y_j' = f(t, y_j(t)),\ y_j(t_j) = s_j,\ (j = 1, \ldots, m-1).$$
4: $F_j \leftarrow \eta_j(t_{j+1}; h) - s_{j+1}$, $(j = 1, \ldots, m-1)$
5: $F_m \leftarrow r(s_1, s_m)$, $e \leftarrow \sum_{j=1}^{m} \|F_j\|$
6: Berechne Näherungen $\zeta_j(\cdot; h)$ für die Lösungen der AWP
$$Z_j' = D_y f(t, \eta_j(t; h)) Z_j(t),\ Z_j(t_j) = I,\ (j = 1, \ldots, m-1).$$
7: $G_j \leftarrow \zeta_j(t_{j+1}; h)$, $(j = 1, \ldots, m-1)$
8: $A \leftarrow D_u r(s_1, s_m)$, $B \leftarrow D_v r(s_1, s_m)$
9: $H = A + BG_{m-1} \cdot \ldots \cdot G_1$, $\varphi = -F_m - B \sum_{j=1}^{m-1} \left(\prod_{l=j+1}^{m-1} G_l \right) F_j$
10: Löse das lineare Gleichungssystem $H \Delta s_1 = \varphi$.
11: $\Delta s_{j+1} \leftarrow G_j \Delta s_j + F_j$, $(j = 1, \ldots, m-1)$
12: $s_k \leftarrow s_k + \Delta s_k$ $(k = 1, \ldots, m)$, $i \leftarrow i + 1$
13: **end while**

Wir überlegen uns zuerst, dass wir

$$a = 0\ ,\ b = 1\ ,\ \alpha = 0\ ,\ \beta = 0$$

annehmen können. Dazu setzen wir

$$y(x) = \alpha + \frac{\beta - \alpha}{b - a}(x - a) + z(x)$$

mit

$$z(a) = z(b) = 0.$$

Dies liefert

$$\begin{aligned} -(pz')' + qz &= -\left(p \left[y' - \frac{\beta - \alpha}{b - a} \right] \right)' + q \left(y - \alpha - \frac{\beta - \alpha}{b - a}(x - a) \right) \\ &= f + \frac{\beta - \alpha}{b - a} p' - q \left[\alpha + \frac{\beta - \alpha}{b - a}(x - a) \right] \\ &= g \end{aligned}$$

in (a,b). Definiere weiter t, u, P, Q und F durch

$$t = \frac{x-a}{b-a}, \qquad z(x) = u\left(\frac{x-a}{b-a}\right),$$
$$p(x) = (b-a)^2 P\left(\frac{x-a}{b-a}\right), \quad q(x) = Q\left(\frac{x-a}{b-a}\right),$$
$$g(x) = F\left(\frac{x-a}{b-a}\right)$$

und bezeichne die Ableitung nach t mit $\cdot$. Dann folgt

$$\begin{aligned} F\left(\frac{x-a}{b-a}\right) &= g(x) \\ &= q(x)z(x) - (p(x)z'(x))' \\ &= Q\left(\frac{x-a}{b-a}\right) u\left(\frac{x-a}{b-a}\right) \\ &\quad - \left(\frac{1}{(b-a)^2}(b-a)^2 P\left(\frac{x-a}{b-a}\right) \dot{u}\left(\frac{x-a}{b-a}\right)\right)^{\cdot} \end{aligned}$$

also

$$\begin{aligned} -(P\dot{u})^{\cdot} + Qu &= F \quad \text{in } (0,1) \\ u(0) &= 0 \\ u(1) &= 0. \end{aligned} \tag{II.4.2}$$

Man kann zeigen, dass das Problem (II.4.2) eine eindeutige Lösung u besitzt und dass $u \in C^4([0,1],\mathbb{R})$ ist, sofern F, P und Q hinreichend oft differenzierbar sind. In der Vorlesung „Numerik II“ werden wir zeigen, dass (II.4.2) eine eindeutige Lösung in einem abgeschwächten Sinn besitzt, sofern P und Q stetig und F quadrat-integrierbar sind.

Nach Übergang zum äquivalenten System erster Ordnung nimmt (II.4.2) die Form (II.1.1) (S. 64) an und kann mit den Methoden der vorigen beiden Abschnitte gelöst werden. Stattdessen betrachten wir im Folgenden ein spezielles Verfahren zur Lösung von (II.4.2), das dessen spezielle Struktur ausnutzt und daher effizienter ist. In §III.3 werden die Methoden dieses Abschnittes auf elliptische partielle Differentialgleichungen übertragen, für die das Sturm-Liouville-Problem (II.4.2) das eindimensionale Analogon ist.

Zur Diskretisierung von (II.4.2) erinnern wir daran, dass für den *zentralen Differenzenquotienten*

$$\partial_h \varphi(t) = \frac{1}{h}\left[\varphi\left(t + \frac{h}{2}\right) - \varphi\left(t - \frac{h}{2}\right)\right]$$

und $\varphi \in C^3([0,1],\mathbb{R})$ gilt

$$\begin{aligned}\partial_h\varphi(t) &= \varphi'(t) + \frac{h^2}{48}\left[\varphi'''\left(t+\theta_1\frac{h}{2}\right) + \varphi'''\left(t+\theta_2\frac{h}{2}\right)\right]\\ &= \varphi'(t) + \frac{h^2}{24}\varphi'''(t+\theta h)\end{aligned}$$

mit $\theta_1,\theta_2 \in (0,1)$ und $\theta \in (-\frac{1}{2},\frac{1}{2})$.
Wir wählen nun ein $n \in \mathbb{N}^*$ und setzen $h = \frac{1}{n+1}$, $t_i = ih$, $0 \le i \le n+1$. Wir fordern die Gültigkeit von (II.4.2) nur noch in den Punkten t_i mit $1 \le i \le n$ und ersetzen die Ableitungen durch den Differenzenquotienten ∂_h. Mit der Notation

$$F_i = F(t_i),\ Q_i = Q(t_i),\ P_{i\pm\frac{1}{2}} = P\left(t_i \pm \frac{1}{2}h\right),\ u_i = u(t_i)$$

liefert dies

$$\begin{aligned}F_i &\approx Q_iu_i - \partial_h(P\partial_h u)_i\\ &= Q_iu_i - \frac{1}{h^2}\left[P_{i+\frac{1}{2}}(u_{i+1}-u_i) - P_{i-\frac{1}{2}}(u_i-u_{i-1})\right].\end{aligned}$$

Wir approximieren also (II.4.2) durch die *Differenzengleichung*

$$\begin{aligned}F_i &= (L_hu)_i\\ &= -\frac{1}{h^2}P_{i-\frac{1}{2}}u_{i-1} + \left(\frac{1}{h^2}[P_{i-\frac{1}{2}} + P_{i+\frac{1}{2}}] + Q_i\right)u_i\\ &\quad - \frac{1}{h^2}P_{i+\frac{1}{2}}u_{i+1} \quad (1 \le i \le n)\\ u_0 &= 0\\ u_{n+1} &= 0.\end{aligned} \tag{II.4.3}$$

Man beachte, dass (II.4.3) ein lineares Gleichungssystem mit einer symmetrischen, tridiagonalen $n \times n$ Matrix ist. Man kann zeigen, dass diese Matrix positiv definit ist. Daher kann (II.4.3) durch eine Cholesky-Zerlegung oder ein CG-Verfahren gelöst werden.

Im Folgenden identifizieren wir den $\mathbb{R}^n$ mit $\{0\} \times \mathbb{R}^n \times \{0\}$ und versehen ihn mit der Maximum-Norm $\|\cdot\|_\infty$.

LEMMA II.4.1 (Stabilität). *Für alle $u \in \mathbb{R}^n$ gilt*

$$\underline{q}\,\|u\|_\infty \le \|L_hu\|_\infty. \tag{II.4.4}$$

Insbesondere besitzt das LGS (II.4.3) stets eine eindeutige Lösung.

BEWEIS. Seien $u \in \mathbb{R}^n$, $M = \|u\|_\infty$ und $i \in \mathbb{N}_n^*$ mit $|u_i| = M$. Dann folgt

$$\begin{aligned}
\|L_h u\|_\infty &\geq |(L_h u)_i| \\
&\geq \left[\frac{1}{h^2}(P_{i-\frac{1}{2}} + P_{i+\frac{1}{2}}) + Q_i\right] |u_i| \\
&\quad - \frac{1}{h^2} P_{i-\frac{1}{2}} |u_{i-1}| - \frac{1}{h^2} P_{i+\frac{1}{2}} |u_{i+1}| \\
&\geq Q_i M \\
&\geq \underline{q} M \\
&= \underline{q} \|u\|_\infty .
\end{aligned}$$

Wegen (II.4.4) besitzt das LGS $L_h u = 0$ nur die triviale Lösung. Hieraus folgt die eindeutige Lösbarkeit. □

SATZ II.4.2 (Fehlerabschätzung). *Bezeichne mit u und u_h die eindeutigen Lösungen der Probleme (II.4.2) und (II.4.3). Zusätzlich gelte $P \in C^3([0,1], \mathbb{R})$ und $u \in C^4([0,1], \mathbb{R})$. Dann gilt die Fehlerabschätzung*

$$\max_{1 \leq i \leq n} |u(t_i) - u_{h,i}| \leq \frac{5}{12} h^2 \left(1 + \frac{1}{24} h^2\right) \underline{q}^{-1} \|P\|_{C^3} \|u\|_{C^4}$$

mit $\|\varphi\|_{C^k} = \max\limits_{0 \leq l \leq k} \max\limits_{0 \leq t \leq 1} \left|\varphi^{(l)}(t)\right|$.

BEWEIS. Aus (II.4.4) folgt

$$\max_{1 \leq i \leq n} |u(t_i) - u_{h,i}| \leq \underline{q}^{-1} \max_{1 \leq i \leq n} |L_h(u - u_h)_i| .$$

Sei $i \in \mathbb{N}_n^*$ beliebig. Dann ist

$$|L_h(u - u_h)_i| = |L_h u(t_i) - F_i| = |L_h u(t_i) - L u(t_i)|$$

und

$$\begin{aligned}
&L_h u(t_i) - Lu(t_i) \\
&\quad = \frac{1}{h^2}\left[-P\left(t_i - \frac{1}{2}h\right)u(t_i - h)\right. \\
&\qquad\quad + \left[P\left(t_i - \frac{1}{2}h\right) + P\left(t_i + \frac{1}{2}h\right)\right]u(t_i) \\
&\qquad\quad \left. -P\left(t_i + \frac{1}{2}h\right)u(t_i + h)\right] + (Pu')'(t_i) \\
&\quad = \frac{1}{2}\left[P\left(t_i - \frac{1}{2}h\right) + P\left(t_i + \frac{1}{2}h\right)\right] \\
&\qquad\quad \cdot \frac{1}{h^2}\left[2u(t_i) - u(t_i - h) - u(t_i + h)\right] \\
&\qquad - \frac{1}{2}\left[P\left(t_i + \frac{1}{2}h\right) - P\left(t_i - \frac{1}{2}h\right)\right] \\
&\qquad\quad \cdot \frac{1}{h^2}\left[u(t_i + h) - u(t_i - h)\right] \\
&\qquad + P(t_i)u''(t_i) + P'(t_i)u'(t_i).
\end{aligned}$$

Taylor-Entwicklung um t_i liefert mit $\theta_1, \ldots, \theta_4 \in (-1, 1)$

$$\begin{aligned}
&\frac{1}{2}\left[P\left(t_i - \frac{1}{2}h\right) + P\left(t_i + \frac{1}{2}h\right)\right]\frac{1}{h^2}\left[2u(t_i) - u(t_i - h) - u(t_i + h)\right] \\
&\quad = -\left[P(t_i) + \frac{h^2}{8}P''\left(t_i + \frac{1}{2}\theta_1 h\right)\right]\left[u''(t_i) + \frac{h^2}{12}u^{(4)}\left(t_i + \theta_2 h\right)\right]
\end{aligned}$$

und

$$\begin{aligned}
&-\frac{1}{2}\left[P\left(t_i + \frac{1}{2}h\right) - P\left(t_i - \frac{1}{2}h\right)\right]\frac{1}{h^2}\left[u(t_i + h) - u(t_i - h)\right] \\
&\quad = -\left[P'(t_i) + \frac{h^2}{24}P'''\left(t_i + \frac{1}{2}\theta_3 h\right)\right]\left[u'(t_i) + \frac{h^2}{6}u'''\left(t_i + \theta_4 h\right)\right].
\end{aligned}$$

Hieraus folgt

$$|L_h(u - u_h)_i| \leq \frac{5}{12}h^2\left(1 + \frac{1}{24}h^2\right)\|P\|_{C^3}\|u\|_{C^4}. \qquad \square$$

Bemerkung II.4.3. (1) Die Konstante in der Fehlerabschätzung von Satz II.4.2 kann noch verbessert werden. Die Voraussetzung $\underline{q} > 0$ ist überflüssig. Man benötigt nur $\underline{p} > 0$. Ähnliche Abschätzungen gelten auch für die ℓ^2-Norm.
(2) Die Regularitätsannahme $u \in C^4$ ist besonders im Hinblick auf partielle Differentialgleichungen zu restriktiv. Die Methoden der Vorlesung „Numerik II“ kommen mit schwächeren Regularitätsannahmen aus.

Beispiel II.4.4 (Einfluss der Diffusion P). Abbildung II.4.1 zeigt die Lösung der Differenzendiskretisierung (II.4.3) des Sturm-Liouville-Problems (II.4.2) mit $F = 1$, $Q = 1$, $P \in \{1, 0.01, 0.001\}$ und $h \in \{0.1, 0.01\}$.

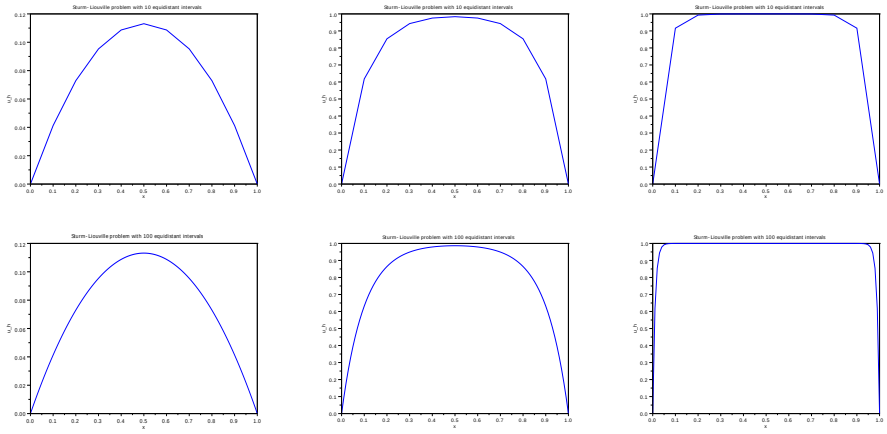

Abbildung II.4.1. Diskrete Lösung aus Beispiel II.4.4 mit $P = 1$ (linke Spalte), $P = 0.01$ (mittlere Spalte), $P = 0.0001$ (rechte Spalte) und $h = 0.1$ (obere Zeile), $h = 0.01$ (untere Zeile)

Beispiel II.4.5 (Innere Grenzschicht). Wir betrachten das Sturm-Liouville-Problem (II.4.1) mit $a = -1$, $b = 1$, $\alpha = -1$, $\beta = 1$, $p = \varepsilon$, $q = 1 + 2\tanh\left(\frac{x}{\varepsilon}\right)^2$, $f = 3\frac{\tanh\left(\frac{x}{\varepsilon}\right)}{\tanh\left(\frac{1}{\varepsilon}\right)}$ und exakter Lösung $u = \frac{\tanh\left(\frac{x}{\varepsilon}\right)}{\tanh\left(\frac{1}{\varepsilon}\right)}$. Abbildung II.4.2 zeigt die exakte Lösung (grün), die diskrete Lösung (blau) und den Fehler (rot) für $\varepsilon = 0.1$ (linke Spalte), $\varepsilon = 0.01$ (mittlere Spalte), $\varepsilon = 0.001$ (rechte Spalte) und die Gitterweiten $h = 0.1$ (erste Zeile), $h = 0.01$ (zweite Zeile), $h = 0.001$ (dritte Zeile), $h = 0.0001$ (vierte Zeile). Tabelle II.4.1 gibt für diese Werte von ε und h den Fehler $\|u - u_h\|_\infty$ in der Maximumnorm und die geschätzte Konvergenzordnung $\frac{\ln\|u-u_{10h}\|_\infty - \ln\|u-u_h\|_\infty}{\ln 10}$ an. Sie unterstreicht die asymptotische Natur der Fehlerabschätzung von Satz II.4.2.

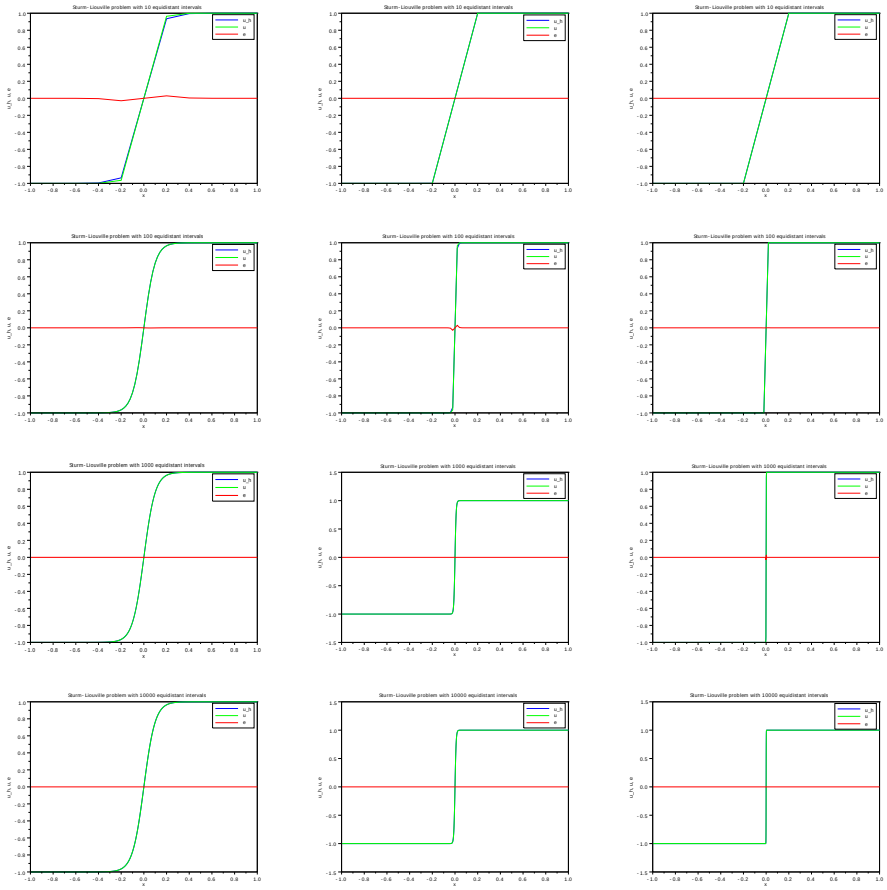

ABBILDUNG II.4.2. Exakte Lösung (grün), diskrete Lösung (blau) und Fehler (rot) für Beispiel II.4.5 mit $\varepsilon = 0.1$ (linke Spalte), $\varepsilon = 0.01$ (mittlere Spalte), $\varepsilon = 0.001$ (rechte Spalte) und $h = 0.1$ (erste Zeile), $h = 0.01$ (zweite Zeile), $h = 0.001$ (dritte Zeile), $h = 0.0001$ (vierte Zeile)

TABELLE II.4.1. Fehler $\|u - u_h\|_\infty$ und geschätzte Konvergenzordnung $\frac{\ln\|u-u_{10h}\|_\infty - \ln\|u-u_h\|_\infty}{\ln 10}$ für Beispiel II.4.5

	$\varepsilon = 0.1$		$\varepsilon = 0.01$		$\varepsilon = 0.001$	
h	Fehler	Ord.	Fehler	Ord.	Fehler	Ord.
0.1	0.028854		0.000832		0.000008	
0.01	0.001627	1.25	0.028854		0.000832	
0.001	0.000016	2.01	0.001627	1.25	0.028854	
0.0001	0.000000	2.02	0.000016	2.01	0.001627	1.25

KAPITEL III

Differenzenverfahren für partielle Differentialgleichungen

In diesem Kapitel geben wir einen elementaren Überblick über die numerische Lösung partieller Differentialgleichungen. Die betrachteten Differenzenverfahren benötigen nur einen geringen technischen Apparat – im wesentlichen die mehrdimensionale Taylor-Entwicklung und die Cauchy-Schwarzsche Ungleichung für Summen – und erlauben einen schnellen Einblick in die entscheidenden Probleme bei der Lösung partieller Differentialgleichungen, die diese wesentlich von gewöhnlichen Differentialgleichungen unterscheiden. Dafür sind diese Verfahren für die Lösung komplexer praktischer Probleme nicht geeignet, da sie unrealistische Differenzierbarkeitsannahmen an die Lösung der Differentialgleichung erfordern und zu wenig flexibel sind, indem sie keine lokale adaptive Anpassung an die Struktur der Lösung erlauben. Verfahren, die dieses Manko beheben und dafür wesentlich anspruchsvollere mathematische Techniken erfordern, sind Gegenstand weiterführender Vorlesungen wie z.B. „Numerik II“ und „Finite Element Methoden für die Navier-Stokes Gleichungen“.

In §III.1 geben wir zunächst einen Überblick über die wichtigsten partiellen Diffentialgleichungen und die zugrunde liegenden physikalischen Probleme. Wir beschreiben die drei wichtigsten Haupttypen elliptische, parabolische und hyperbolische Gleichungen zweiter Ordnung zusammen mit den entsprechenden Modellproblemen Diffusions-Reaktions-Gleichung, Wärmeleitungsgleichung und Wellengleichung. Ein wesentlicher Aspekt ist hierbei der Einfluss des Randes auf die Differenzierbarkeit der Lösung der Differentialgleichung. Insbesondere bei den für die Numerik besonders wichtigen Polyedergebieten besitzen die Differentialgleichungen selbst bei beliebig oft differenzierbaren rechten Seiten häufig keine klassische Lösung, geschweige denn eine beliebig oft differenzierbare Lösung.

In §III.2 stellen wir einen allgemeinen Rahmen für die Konvergenzanalyse von Diskretisierungsverfahren vor. Wichtigstes Ergebnis ist die Äquivalenz von Konsistenz und Stabilität des Verfahrens zu seiner Konvergenz.

Anschließend wenden wir in den §§III.3 – III.5 die allgemeinen Ergebnisse von §III.2 auf elliptische, parabolische und hyperbolische Differentialgleichungen mit der Diffusions-Reaktions-Gleichung, der Wärmeleitungsgleichung und der Wellengleichung als wichtigsten Modellproblemen an. Ein wichtiger Aspekt bei den zeitabhängigen Problemen ist dabei die Tatsache, dass die Orts- und Zeitschrittweite der Diskretisierung nicht unabhängig von einander gewählt werden können.

Die Verfahren der §§III.3 – III.5 erfordern die Lösung großer, dünn besetzter, schlecht konditionierter linearer Gleichungssysteme. Wegen der Größe und der dünnen Besetzung sind direkte Lösungsverfahren wie die Gauß-Elimination nicht effizient genug [9, Beispiel IV.6.1]. Wegen der schlechten Kondition scheiden klassische iterative Verfahren wie die Jacobi- oder Gauß-Seidel-Relaxation ebenfalls aus [9, Beispiel IV.6.5]. Für Probleme mittlerer Größe bietet sich ein geeignet vorkonditioniertes CG-Verfahren [9, Algorithmen IV.7.10, IV.7.14] an, das wir als erstes in §III.6 aufgreifen. Im Zentrum dieses Abschnittes stehen Mehrgitterverfahren, die die effizientesten Verfahren zur Lösung der bei der Diskretisierung partieller Differentialgleichung entstehenden diskreten Probleme sind.

III.1. Beispiele partieller Differentialgleichungen

In diesem Paragraphen geben wir einige Beispiele partieller Differentialgleichungen kurz pDGl und ihrer Anwendungen, ohne jedoch auf die physikalische Herleitung näher einzugehen. Weiter beschreiben wir die verschiedenen Typen pDGlen und ihre wesentlichen Eigenschaften.

Beispiel III.1.1 (Membran-, Poisson-Gleichung). Sei $\Omega \subset \mathbb{R}^2$ ein offenes, beschränktes, zusammenhängendes Gebiet, das von einer dünnen Membran (z. B. Trommelfell) in ihrer Ruhelage eingenommen wird. Auf die Membran wirke eine äußere Kraft f. Diese bewirkt eine Auslenkung $u = u(x) = u(x_1, x_2)$ in vertikaler Richtung. Unter der Annahme, dass die Membran nicht dehnbar und die Auslenkung klein ist, folgt aus dem Prinzip von „Actio = Reactio“, dass die Auslenkung beschrieben wird durch

$$f = -\Delta u = -\sum_{i=1}^{2} \frac{\partial^2 u}{\partial x_i^2} \quad \text{in } \Omega. \tag{III.1.1}$$

Dies ist eine pDGl, die sog. *Membran-* oder *Poisson-Gleichung*. Falls die Membran am Rand Γ von Ω eingespannt ist, gilt dort

$$u = 0 \quad \text{auf } \Gamma. \tag{III.1.2}$$

Dies ist eine Randbedingung, die sog. *homogene Dirichlet-Randbedingung*. Falls die Membran am Rand dagegen frei gelagert ist, gilt dort

$$\frac{\partial u}{\partial n} = 0 \quad \text{auf } \Gamma. \tag{III.1.3}$$

Dies ist die sog. *homogene Neumann-Randbedingung*. Man kann zeigen, dass das Problem (III.1.1) zusammen mit der Randbedingung (III.1.2) oder mit der Randbedingung (III.1.3) plus der Normierungsbedingung

$$\int_\Omega u = 0$$

in einem geeigneten schwachen Sinn stets eine eindeutige Lösung hat. Diese ist bestimmt durch die Bedingung

$$J(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega fu \to \min \text{ in } V, \tag{III.1.4}$$

wobei V ein geeigneter Funktionenraum ist. Dabei bezeichnet ∇u den Gradienten von u.

Beispiel III.1.2 (Platten-, biharmonische Gleichung). Wir ersetzen die Membran aus Beispiel III.1.1 durch eine dünne, starre Platte und bezeichnen mit u die Auslenkung der Mittelebene der Platte. Dann folgt aus den gleichen physikalischen Prinzipien, dass u bestimmt wird durch

$$f = \Delta^2 u = \Delta(\Delta u) = \frac{\partial^4 u}{\partial x_1^4} + 2\frac{\partial^4 u}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4 u}{\partial x_2^4} \quad \text{in } \Omega. \tag{III.1.5}$$

Dies ist die sog. *Platten-* oder *biharmonische Gleichung*. Falls der Rand der Platte fest eingespannt ist, gilt zusätzlich die Randbedingung

$$u = 0 \quad \text{und} \quad \frac{\partial u}{\partial n} = 0 \quad \text{auf } \Gamma, \tag{III.1.6}$$

ist der Rand dagegen frei gelagert, so gilt die Randbedingung

$$u = 0 \quad \text{und} \quad \Delta u = 0 \quad \text{auf } \Gamma. \tag{III.1.7}$$

Man kann wieder zeigen, dass Problem (III.1.5) mit den Randbedingungen (III.1.6) oder (III.1.7) in einem geeigneten schwachen Sinn stets eine eindeutige Lösung besitzt. Diese ist bestimmt durch die Bedingung

$$J(u) = \frac{1}{2}\int_\Omega |\Delta u|^2 - \int_\Omega fu \to \min \text{ in } V, \tag{III.1.8}$$

wobei V wieder ein geeigneter Funktionenraum ist.

Beispiel III.1.3 (Minimalflächen). Sei Ω wie in Beispiel III.1.1. Wir suchen unter allen Flächen der Form

$$S = \left\{x \in \mathbb{R}^3 : x' = (x_1, x_2) \in \Omega\,,\ x_3 = u(x')\,,\ u(x') = u_0(x')\ \forall x' \in \Gamma\right\}$$

diejenige, die einen minimalen Flächeninhalt hat. Dabei ist u_0 eine gegebene Funktion. Dies führt auf das Variationsproblem

$$J(u) = \int_\Omega \sqrt{1 + |\nabla u|^2}\,\mathrm{d}x' \to \min \text{ in } V, \tag{III.1.9}$$

wobei V eine konvexe Teilmenge eines geeigneten Funktionenraumes ist. Man kann zeigen, dass dieses Problem eine nicht notwendig eindeutige Lösung hat, die die pDGl

$$-\nabla \cdot \left(\left\{ 1 + |\nabla u|^2 \right\}^{-\frac{1}{2}} \nabla u \right) = 0 \quad \text{in } \Omega$$

mit der Randbedingung

$$u = u_0 \quad \text{auf } \Gamma$$

erfüllt. Dabei bezeichnet $\nabla \cdot \varphi = \sum_{i=1}^{n} \frac{\partial \varphi_i}{\partial x_i}$ die Divergenz des Vektorfeldes $\varphi : \mathbb{R}^n \to \mathbb{R}^n$.

Beispiel III.1.4 (Gasgleichung). Wir betrachten die rotationsfreie Strömung eines idealen, kompressiblen Gases. Aus der Rotationsfreiheit folgt für die Geschwindigkeit $\mathbf{V}$ des Gases

$$\mathbf{V} = \nabla u$$

für ein skalares Potential u. Aus der Massenerhaltung folgt

$$\nabla \cdot (\rho \mathbf{V}) = 0,$$

wobei $\rho = \rho(\mathbf{V})$ die Dichte des Gases ist. Da das Gas ideal ist, gilt die Zustandsgleichung

$$\rho(\mathbf{V}) = \left[1 - \frac{\gamma - 1}{2} |\mathbf{V}|^2 \right]^{\frac{1}{\gamma - 1}},$$

wobei $\gamma > 1$ der spezifische Wärmekoeffizient ist. Insgesamt erfüllt somit das Potential u die pDGl

$$-\nabla \cdot \left[\left(1 - \frac{\gamma - 1}{2} |\nabla u|^2 \right)^{\frac{1}{\gamma - 1}} \nabla u \right] = 0 \quad \text{in } \Omega$$

zusammen mit der Randbedingung

$$u = u_0 \quad \text{auf } \Gamma,$$

wobei u_0 eine gegebene Funktion ist.

Beispiel III.1.5 (Wärmeleitungsgleichung). Sei $\Omega \subset \mathbb{R}^3$ ein offenes, beschränktes, zusammenhängendes Gebiet. Die Funktion $u(x,t) : \Omega \times \mathbb{R}_+ \to \mathbb{R}$ beschreibe die Temperatur zur Zeit t im Punkt x des Körpers. Wenn dieser einer äußeren Wärmequelle f ausgesetzt ist, wird der zeitliche Verlauf der Temperatur durch die sog. *Wärmeleitungsgleichung*

$$\frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } \Omega \times \mathbb{R}_+ \tag{III.1.10}$$

beschrieben. Diese pDGl ist zu ergänzen durch eine Information über die anfängliche Temperaturverteilung, d.h. durch eine *Anfangsbedingung*

$$u(x,0) = u_0(x) \quad \text{in } \Omega, \tag{III.1.11}$$

wobei $u_0 : \Omega \to \mathbb{R}$ eine bekannte Funktion ist. Wenn der Rand des Körpers künstlich auf einer bestimmten, zeitlich nicht notwendig konstanten Temperatur gehalten wird, gilt zusätzlich die Randbedingung

$$u(x,t) = g_D(x,t) \quad \text{auf } \Gamma \times \mathbb{R}_+^*. \tag{III.1.12}$$

Ist der Rand des Körpers dagegen isoliert, d.h. findet dort kein Wärmefluss statt, gilt die Randbedingung

$$\frac{\partial u}{\partial n}(x,t) = g_N(x,t) \quad \text{auf } \Gamma \times \mathbb{R}_+^*, \tag{III.1.13}$$

wobei g_N eine bekannte Funktion ist, z.B. $g_N = 0$. Man beachte, dass f Wärmequellen im Innern des Körpers beschreibt, wogegen g_N die Wärmezufuhr von außen modelliert.
Sei nun u eine hinreichend glatte Lösung von (III.1.10) zu $f = 0$ mit der Randbedingung (III.1.12) mit $g_D = 0$. Multipliziere (III.1.10) mit u und integriere über Ω. Dann folgt mit dem Integralsatz von Gauß

$$0 = \int_\Omega \left\{ \frac{\partial u}{\partial t} - \Delta u \right\} u = \frac{d}{dt} \left[\frac{1}{2} \int_\Omega u^2 \right] + \int_\Omega |\nabla u|^2 .$$

Also ist die Energie $E : \mathbb{R}_+ \to \mathbb{R}$ mit

$$E(t) = \frac{1}{2} \int_\Omega u(x,t)^2 \mathrm{d}x \tag{III.1.14}$$

monoton fallend.
Man kann zeigen, dass die Gleichung (III.1.10) mit der Anfangsbedingung (III.1.11) und der Randbedingung (III.1.12) bzw. (III.1.13) in einem geeigneten schwachen Sinn stets eine eindeutige Lösung hat. Falls die Funktionen f und g_D oder g_N für $t \to +\infty$ gegen zeitliche konstante Funktionen $\overline{f}$ und $\overline{g}_D$ bzw. $\overline{g}_N$ konvergieren, kann man zusätzlich zeigen, dass die Lösung $u(x,t)$ für $t \to +\infty$ gegen die Lösung $\overline{u} = \overline{u}(x)$ der *stationären Gleichung*

$$-\Delta \overline{u} = \overline{f} \quad \text{in} \quad \Omega$$
$$\overline{u} = \overline{g}_D \quad \text{auf } \Gamma \quad \text{bzw.} \quad \frac{\partial \overline{u}}{\partial n} = \overline{g}_N \quad \text{auf } \Gamma$$

konvergiert.

Beispiel III.1.6 (Transport-Diffusions-Gleichung). Die Funktion $u(x,t)$ bezeichne die räumliche und zeitliche Verteilung einer Flüssigkeit, z.B. Grundwasser, in einem porösen Körper $\Omega \subset \mathbb{R}^3$, z.B. Erde. Die Funktion u ist Lösung der sog. *Transport-Diffusions-Gleichung*

$$\frac{\partial u}{\partial t} - \nabla \cdot (D(x,u)\nabla u) + \mathbf{k}(x,u) \cdot \nabla u = f \quad \text{in } \Omega \times \mathbb{R}_+. \tag{III.1.15}$$

Diese ist zu ergänzen durch eine Anfangsbedingung der Form (III.1.11) und Randbedingungen der Form (III.1.12) oder (III.1.13). Der *Quellterm* f bezeichnet die Zufuhr (Quellen) oder Entnahme (Brunnen) von Flüssigkeit. Die *Diffusivität* $D(x,u) : \Omega \times \mathbb{R} \to \mathbb{R}^{3\times 3}$ und die *Konduktivität* $\mathbf{k}(x,u) : \Omega \times \mathbb{R} \to \mathbb{R}^3$ beschreiben spezifische Eigenschaften des Körpers (Ton, Lehm, Sand usw.).
Unter geeigneten Voraussetzungen besitzt (III.1.15) zusammen mit Anfangs- und Randbedingungen eine eindeutige Lösung, die unter geeigneten zusätzlichen Voraussetzungen gegen einen stationären Zustand $\overline{u}$ konvergiert. Dieser ist Lösung der sog. *Konvektions-Diffusions-Gleichung*

$$-\nabla \cdot (D(x,\overline{u})\nabla\overline{u}) + \mathbf{k}(x,\overline{u}) \cdot \nabla\overline{u} = \overline{f} \quad \text{in } \Omega$$

mit Dirichlet- oder Neumann-Randbedingungen.

Beispiel III.1.7 (Wellengleichung). Wir betrachten wie in Beispiel III.1.1 eine dünne, elastische Membran, versetzen sie aber diesmal durch eine zeitlich veränderliche äußere Kraft in Schwingung. Falls die Auslenkung klein ist, tritt keine Dämpfung durch innere Reibung auf. Die Auslenkung $u(x,t)$ am Ort x zur Zeit t wird dann durch die sog. *Wellengleichung*

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \quad \text{in } \Omega \times \mathbb{R}_+ \tag{III.1.16}$$

beschrieben. Zusätzlich gelten auf $\Gamma \times \mathbb{R}_+$ die Randbedingungen (III.1.2) oder (III.1.3), je nachdem ob die Membran eingespannt oder frei gelagert ist. Schließlich muss noch der Anfangszustand des Systems angegeben werden. Dies geschieht durch die *Anfangsbedingungen*

$$u(x,0) = u_0(x) \quad \text{und} \quad \frac{\partial u}{\partial t}(x,0) = u_1(x) \quad \text{in } \Omega. \tag{III.1.17}$$

Man kann wiederum zeigen, dass die pDGl (III.1.16) mit den Anfangsbedingungen (III.1.17) und den Randbedingungen (III.1.2) oder (III.1.3) in einem geeigneten Sinne stets eine schwache Lösung besitzt. Betrachte nun speziell Gleichung (III.1.16) mit $f = 0$ und der Randbedingung (III.1.2). Sei u eine hinreichend glatte Lösung. Multiplikation von (III.1.16) mit $\frac{\partial u}{\partial t}$, Integration über Ω und der Integralsatz von Gauß liefern dann

$$\begin{aligned} 0 &= \int_\Omega \left\{ \frac{\partial^2 u}{\partial t^2} - \Delta u \right\} \frac{\partial u}{\partial t} \mathrm{d}x = \int_\Omega \frac{\partial}{\partial t} \left[\frac{1}{2} \left(\frac{\partial u}{\partial t} \right)^2 + \frac{1}{2} |\nabla u|^2 \right] \mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \left[\frac{1}{2} \int_\Omega \left\{ \left(\frac{\partial u}{\partial t} \right)^2 + |\nabla u|^2 \right\} \mathrm{d}x \right]. \end{aligned}$$

Also bleibt die Energie

$$E(t) = \frac{1}{2} \int_\Omega \left\{ \left(\frac{\partial u}{\partial t} \right)^2 + |\nabla u|^2 \right\} \mathrm{d}x \tag{III.1.18}$$

erhalten.

DEFINITION III.1.8 (Differentialoperator, Ordnung, Typ, Differentialgleichung). (1) Ein *Differentialoperator m-ter Ordnung*, $m \geq 1$, hat die Form

$$u \mapsto \mathcal{D}(x, u(x), Du(x), \ldots, D^m u(x))$$

mit $x \in \Omega \subset \mathbb{R}^n$. Er heißt *quasilinear*, wenn er darstellbar ist als

$$\begin{aligned}\mathcal{D}(x, u(x), \ldots, D^m u(x)) = {} & A(x, u(x), \ldots, D^m u(x)) \\ & + B(x, u(x), \ldots, D^{m-1} u(x))\end{aligned}$$

mit

$$A(x, u(x), \ldots, D^m u(x)) = \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = m}} a_\alpha(x, u(x), \ldots, D^{m-1} u(x)) D^\alpha u(x).$$

Der Operator A heißt dann der *Hauptteil* des Differentialoperators. Der Differentialoperator $\mathcal{D}$ heißt *semilinear*, wenn er quasilinear ist und die Koeffizienten a_α des Hauptteils nur von x, aber nicht von u oder Ableitungen von u abhängen. Er heißt *linear*, wenn zusätzlich gilt

$$B(x, u(x), \ldots, D^{m-1} u(x)) = \sum_{\substack{\beta \in \mathbb{N}^n \\ |\beta| \leq m-1}} b_\beta(x) D^\beta u(x).$$

Der lineare Differentialoperator $\mathcal{D}$ hat *konstante Koeffizienten*, wenn die Funktionen a_α und b_β konstant sind.
(2) Ein quasilinearer Differentialoperator m-ter Ordnung heißt *elliptisch*, wenn für alle $x \in \Omega$, alle hinreichend oft differenzierbaren Funktionen $u : \Omega \to \mathbb{R}$ und alle $z \in \mathbb{R}^n \setminus \{0\}$ gilt

$$\sum_{|\alpha| = m} a_\alpha(x, u(x), \ldots, D^{m-1} u(x)) z^\alpha \neq 0.$$

Dabei ist $z^\alpha = z_1^{\alpha_1} \cdot \ldots \cdot z_n^{\alpha_n}$.
(3) Eine *partielle Differentialgleichung m-ter Ordnung*, kurz pDgl ist ein Ausdruck der Form

$$\mathcal{D}(u) = f$$

mit einer gegebenen Funktion $f : \Omega \to \mathbb{R}$ und einem Differentialoperator m-ter Ordnung. Sie heißt *quasilinear*, *semilinear*, *linear*, *linear mit konstanten Koeffizienten* bzw. *elliptisch*, wenn gleiches für den Differentialoperator $\mathcal{D}$ gilt.

BEMERKUNG III.1.9 (Typisierung der Beispiele). Die biharmonische Gleichung ist eine pDGl 4-ter Ordnung. Die anderen pDGlen in obigen Beispielen sind von 2-ter Ordnung. Alle betrachteten Gleichungen sind quasilinear. Die pDGlen der Beispiele III.1.1, III.1.2, III.1.5 und III.1.7 sind linear mit konstanten Koeffizienten. Die Konvektions-Diffusions- bzw. Transport-Diffusions-Gleichung aus Beispiel III.1.6 ist semilinear, wenn die Diffusität D nur von x abhängt.

Für die Hauptteile der linearen und semilinearen Differentialoperatoren der Beispiele III.1.1, III.1.2, III.1.5, III.1.6 und III.1.7 ergibt sich:

1.1 $A(u) = -\Delta u$,
$\sum_{|\alpha|=2} a_\alpha z^a = -\sum_{i=1}^n z_i^2 < 0$ für alle $z \in \mathbb{R}^n \setminus \{0\}$,
elliptisch;

1.2 $A(u) = \Delta^2 u$,
$\sum_{|\alpha|=4} a_\alpha z^\alpha = z_1^4 + 2z_1^2 z_2^2 + z_2^4 = (z_1^2 + z_2^2)^2 > 0$, für alle $z \in \mathbb{R}^2 \setminus \{0\}$,
elliptisch;

1.5 $A(u) = -\Delta u$,
$\sum_{\substack{\alpha\in\mathbb{N}^4 \\ |\alpha|=2}} a_\alpha z^\alpha = z_1^2 + z_2^2 + z_3^2 + 0 \cdot z_4^2$,
nicht elliptisch;

1.6 $A(u) = -\sum_{i,j} D_{ij}(x,u) \dfrac{\partial^2 u}{\partial x_i \partial x_j}$,
nicht elliptisch im zeitabhängigen Fall, elliptisch im stationären Fall, falls $D(x,u)$ definit;

1.7 $A(u) = \dfrac{\partial^2 u}{\partial t^2} - \Delta u$,
$\sum_{\substack{\alpha\in\mathbb{N}^4 \\ |\alpha|=2}} a_\alpha z^\alpha = z_1^2 + z_2^2 + z_3^2 - z_4^2$,
nicht elliptisch.

Die nicht semilinearen Differentialoperatoren der Beispiele III.1.3 und III.1.4 haben die Struktur

$$\mathcal{D}(u, Du, D^2u) = -\operatorname{div}\left(a\left(|\nabla u|^2\right)\nabla u\right) \tag{III.1.19}$$

mit $a(\eta) = (1+\eta)^{-\frac{1}{2}}$ für die Minimalflächengleichung III.1.3 und $a(\eta) = (1 - \frac{\gamma-1}{2}\eta)^{\frac{1}{\gamma-1}}$ für die Gasgleichung III.1.4. Wegen

$$\begin{aligned}
-\operatorname{div}\left(a\left(|\nabla u|^2\right)\nabla u\right) &= -\sum_i D_i\left(a\left(\sum_j (D_j u)^2\right) D_i u\right) \\
&= -\sum_i a\left(|\nabla u|^2\right) D_{ii} u \\
&\quad -\sum_{i,j} 2a'\left(|\nabla u|^2\right) D_j u\, D_{ij} u\, D_i u \\
&= -a\left(|\nabla u|^2\right)\Delta u - 2a'\left(|\nabla u|^2\right)\sum_{i,j} D_j u\, D_{ij} u\, D_i u
\end{aligned}$$

ergibt sich für den Hauptteil der Minimalflächengleichung III.1.3

$$A(u) = -\left\{1+|\nabla u|^2\right\}^{-\frac{3}{2}}\left\{\left(1+|\nabla u|^2\right)\Delta u - \sum_{i,j} D_i u D_j u D_{ij} u\right\}$$

und für den Hauptteil der Gasgleichung III.1.4

$$A(u) = -\left\{1-\frac{\gamma-1}{2}|\nabla u|^2\right\}^{\frac{1}{\gamma-1}}\left\{\Delta u - \left(1-\frac{\gamma-1}{2}|\nabla u|^2\right)^{-1}\sum_{i,j} D_i u D_j u D_{ij} u\right\}.$$

Weiter ist mit $\eta = |\nabla u|^2$

$$\begin{aligned}\sum_\alpha a_\alpha z^\alpha &= -a(\eta)\sum_i z_i^2 - 2a'(\eta)\sum_{i,j} D_i u D_j u\, z_i z_j\\ &= z^t A z\end{aligned}$$

mit

$$A = -a(\eta) I - 2a'(\eta)\nabla u \otimes \nabla u.$$

Jeder Vektor $w \in \mathbb{R}^n$ mit $w \perp \nabla u$ ist ein Eigenvektor von A zum Eigenwert $-a(\eta)$. Ebenso ist ∇u ein Eigenvektor von A zum Eigenwert $-a(\eta)-2a'(\eta)\eta$. Daher ist die Differentialoperator (III.1.19) genau dann elliptisch, wenn für alle $\eta > 0$ gilt $a(\eta) \neq 0$ und $1 + 2\frac{a'(\eta)}{a(\eta)}\eta > 0$.
Für die Minimalflächengleichung III.1.3 ergibt sich

$$\begin{aligned}a(\eta) &= (1+\eta)^{-\frac{1}{2}} \neq 0,\\ a'(\eta) &= \frac{1}{2}(1+\eta)^{-\frac{3}{2}},\\ 1 + 2\frac{a'(\eta)}{a(\eta)}\eta &= 1 - \frac{\eta}{1+\eta} = \frac{1}{1+\eta} > 0.\end{aligned}$$

Also ist die Minimalflächengleichung elliptisch.
Für die Gasgleichung III.1.4 ergibt sich

$$\begin{aligned}a(\eta) &= \left(1-\frac{\gamma-1}{2}\eta\right)^{\frac{1}{\gamma-1}},\\ a'(\eta) &= \frac{1}{\gamma-1}\left(-\frac{\gamma-1}{2}\right)\left(1-\frac{\gamma-1}{2}\eta\right)^{\frac{1}{\gamma-1}-1},\\ 1 + 2\frac{a'(\eta)}{a(\eta)}\eta &= 1 - \frac{\eta}{1-\frac{\gamma-1}{2}\eta}\\ &= \frac{1-\frac{\gamma+1}{2}\eta}{1-\frac{\gamma-1}{2}\eta}.\end{aligned}$$

Der erste Ausdruck ist ungleich Null für $\eta < \frac{2}{\gamma-1}$; der letze Ausdruck ist positiv für $\eta \leq \frac{2}{\gamma+1}$. Wegen $\frac{2}{\gamma+1} < \frac{2}{\gamma-1}$ ist daher die Gasgleichung genau dann elliptisch, wenn $\eta \leq \frac{2}{\gamma+1}$ ist. Dies entspricht einer *Unterschallströmung*.

Partielle Differentialgleichungen 2. Ordnung kann man genauer charakterisieren als in Definition III.1.8. Wir beschränken uns hier auf den linearen Fall, der allgemeine quasilineare geht aber ganz analog. Der allgemeine lineare Differentialoperator 2. Ordnung hat die Form

$$\mathcal{D}(u) = A(x) : D^2u + a(x) \cdot \nabla u + \alpha(x)u \tag{III.1.20}$$

mit $\alpha : \Omega \to \mathbb{R}$, $a : \Omega \to \mathbb{R}^n$, $A : \Omega \to \mathbb{R}^{n\times n}$ und

$$A : B = \sum_{1\leq i,j\leq n} A_{ij}B_{ij}.$$

Wegen der Symmetrie von D^2u kann man dabei o.E. annehmen, dass $A(x)$ für alle x symmetrisch ist.

Definition III.1.10 (Charakterisierung von Differentialoperatoren 2. Ordnung). Der Differentialoperator (III.1.20) heißt

- *elliptisch* in x, wenn alle Eigenwerte von $A(x)$ von Null verschieden sind und gleiches Vorzeichen haben,
- *hyperbolisch* in x, wenn alle Eigenwerte von $A(x)$ von Null verschieden sind und genau ein Eigenwert das entgegengesetzte Vorzeichen der restlichen Eigenwerte hat,
- *parabolisch* in x, wenn genau ein Eigenwert von $A(x)$ gleich Null ist, die restlichen Eigenwerte gleiches Vorzeichen haben und die $n \times (n + 1)$ Matrix $[A, a]$ maximalen Rang hat.

Er heißt *elliptisch*, *hyperbolisch*, bzw. *parabolisch*, wenn er in jedem Punkt elliptisch, hyperbolisch, bzw. parabolisch ist.

Bemerkung III.1.11 (Eigenschaften der Differentialgleichungstypen). (1) Die Definition von elliptisch aus Definition III.1.10 stimmt mit derjenigen aus Definition III.1.8 überein.
(2) Die Wellengleichung aus Beispiel III.1.7 ist hyperbolisch. Die Wärmeleitungsgleichung aus Beispiel III.1.5 ist parabolisch. Die Transport-Diffusions-Gleichung aus Beispiel III.1.6 ist parabolisch, sofern $D(x)$ positiv definit ist.
(3) Elliptische, parabolische und hyperbolische Gleichungen beschreiben verschiedene physikalische Situationen. Diese kann man grob wie folgt charakterisieren:

- Eine elliptische Gleichung beschreibt ein *Variations-* oder *Minimierungsproblem*, (vgl. Funktionale (III.1.4), (III.1.8) und (III.1.9)).
- Eine parabolische Gleichung beschreibt ein *Dissipationsphänomen*, d.h. ein System, in dem eine Energie gedämpft wird (vgl. Gleichung (III.1.14)).

- Eine hyperbolische Gleichung beschreibt einen *Erhaltungssatz*, d.h. ein System, in dem eine Energie erhalten wird (vgl. Gleichung (III.1.18)).

Eine wesentliche Eigenschaft gewöhnlicher Differentialgleichungen ist die Tatsache, dass jede Lösung so regulär ist, wie es die rechte Seite zulässt (vgl. Satz I.1.16 (S. 14)). Ein solches Ergebnis können wir für partielle Differentialgleichungen nicht erwarten, wie das folgende Beispiel zeigt. Eigenschaften des Randes Γ spielen eine wesentliche Rolle.

BEISPIEL III.1.12 (Einfluss des Randes auf die Regularität der Lösung einer pDgl). Sei $0 < \alpha < 2\pi$ und Ω_α das Kreissegment

$$\Omega_\alpha = \left\{ x \in \mathbb{R}^2 : x = (r\cos\varphi, r\sin\varphi), 0 < r < 1, 0 < \varphi < \alpha \right\}.$$

Definiere die Funktion $v : \Omega_\alpha \to \mathbb{R}$ durch

$$v(x) = r^{\frac{\pi}{\alpha}} \sin\left(\frac{\pi}{\alpha}\varphi\right)$$

mit $x = (r\cos\varphi, r\sin\varphi)$. Dann gilt für jedes $x \in \Omega_\alpha$

$$\Delta v(x) = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 v}{\partial \varphi^2} = 0.$$

Sei $w \in C_0^\infty(\mathbb{R}^2, \mathbb{R})$ mit $\operatorname{supp} w \subset B(0, \frac{2}{3})$ und $w = 1$ auf $\overline{B(0, \frac{1}{3})}$. Definiere

$$u = wv, \quad f = \Delta[(1-w)v].$$

Dann gilt

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega_\alpha \\ u &= 0 \quad \text{auf } \partial\Omega_\alpha. \end{aligned}$$

Offensichtlich ist $(1-w)v \in C^\infty(\mathbb{R}^2, \mathbb{R})$ und somit $f \in C^\infty(\overline{\Omega}_\alpha)$. Ebenso ist $u \in C^\infty(\Omega_\alpha)$. Wegen $u = v$ in $B(0, \frac{1}{3})$ gilt aber $u \notin C^\infty(\overline{\Omega}_\alpha)$. Wie man leicht nachrechnet, gilt $u \in C^k(\overline{\Omega}_\alpha)$ mit $k \geq 1$ genau dann, wenn $0 < \alpha \leq \frac{\pi}{k}$ ist, und $D^k u \in L^2(\Omega_\alpha)$ mit $k \geq 2$ genau dann, wenn $0 < \alpha < \frac{\pi}{k-1}$ ist. Wir können also bei gegebenem α i.a. keine Abschätzung der Form

$$\|u\|_{C^{k+2}(\overline{\Omega}_\alpha)} \leq c_k \|f\|_{C^k(\overline{\Omega}_\alpha)}$$

oder

$$\left\{ \sum_{|\beta| \leq k+2} \left\| D^\beta u \right\|^2_{L^2(\Omega_\alpha)} \right\}^{\frac{1}{2}} \leq c_k' \left\{ \sum_{|\beta| \leq k} \left\| D^\beta f \right\|^2_{L^2(\Omega_\alpha)} \right\}^{\frac{1}{2}}$$

erwarten, wie sie für gewöhnliche Differentialgleichungen gelten würde.

III.2. Konvergenz von Diskretisierungsverfahren

Wir betrachten in diesem Paragraphen folgende abstrakte Situation. Gegeben sind zwei Banach-Räume $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$, ein Element $f \in Y$ und eine lineare, nicht notwendig stetige Abbildung $L : X \to Y$. Gesucht ist eine Lösung $u \in X$ der Gleichung

$$Lu = f. \tag{III.2.1}$$

Zur Diskretisierung von (III.2.1) betrachten wir zwei Familien $(X_h, \|\cdot\|_{X_h})$, $(Y_h, \|\cdot\|_{Y_h})$ endlich dimensionaler Banach-Räume, $h > 0$, eine Familie $f_h \in Y_h$ von diskreten Approximationen von f und eine Familie linearer Abbildungen $L_h : X_h \to Y_h$ von Diskretisierungen von L. Dann ersetzen wir (III.2.1) durch die diskreten Probleme

$$L_h u_h = f_h \tag{III.2.2}$$

mit unbekannten Elementen $u_h \in X_h$. Die Räume X und Y sind mit ihren Diskretisierungen X_h, Y_h durch lineare, nicht notwendig stetige Operatoren $R_{X_h} : X \to X_h$ und $R_{Y_h} : Y \to Y_h$ verbunden. Die Situation kann man anschaulich in folgendem Schema zusammenfassen.

$$\begin{array}{ccc} X & \xrightarrow{L} & Y \\ {\scriptstyle R_{X_h}}\downarrow & & \downarrow{\scriptstyle R_{Y_h}} \\ X_h & \xrightarrow[L_h]{} & Y_h \end{array}$$

Bemerkung III.2.1 (Gitterweite). Der Parameter $h > 0$ entspricht einer Gitterweite und wir sind an dem Verhalten für $h \to 0$ interessiert. Die Dimension der Räume X_h und Y_h wächst für $h \to 0$.

Definition III.2.2 (Diskretisierungsverfahren, Konsistenz, Stabilität, Konvergenz). (1) Das durch (III.2.2) gegebene *Diskretisierungsverfahren* heißt *konsistent* mit (III.2.1), wenn für alle $u \in X$ gilt

$$\|L_h R_{X_h} u - R_{Y_h} L u\|_{Y_h} \xrightarrow[h\to 0]{} 0.$$

(2) Das Diskretisierungsverfahren (III.2.2) heißt *stabil*, wenn gilt

$$\sup_{h>0} \left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h, X_h)} < \infty.$$

(3) Das Diskretisierungsverfahren (III.2.2) heißt *konvergent*, wenn für jedes $f \in Y$, jede Lösung u von (III.2.1) und jede Familie $f_h \in Y_h$ mit

$$\lim_{h\to 0} \|R_{Y_h} f - f_h\|_{Y_h} = 0$$

gilt

$$\lim_{h\to 0} \|R_{X_h} u - u_h\|_{X_h} = 0,$$

wobei u_h eine Lösung von (III.2.2) ist.

Das folgende Beispiel zeigt, dass das Differenzenverfahren aus §II.4 in den abstrakten Rahmen obiger Definition passt.

Beispiel III.2.3 (Sturm-Liouville-Problem). Für das Sturm-Liouville-Problem

$$\begin{aligned} -u'' + u &= f \quad \text{in } (0,1) \\ u(0) &= 0 \\ u(1) &= 0 \end{aligned}$$

mit $f \in C([0,1],\mathbb{R})$ setzen wir

$$\begin{aligned} Y &= C([0,1],\mathbb{R}), & \|\varphi\|_Y &= \max_{0\le x\le 1} |\varphi(x)|\,, \\ X &= C^2([0,1],\mathbb{R}) \cap C_0([0,1],\mathbb{R}), & \|\varphi\|_X &= \max_{0\le x\le 1} \max_{0\le k\le 2} \left|\varphi^{(k)}(x)\right|\,, \\ Lu &= -u'' + u. \end{aligned}$$

Für $n \in \mathbb{N}^*$ betrachten wir das Gitter $\{ih : 1 \le i \le n\}$ mit $h = \frac{1}{n+1}$. Wir setzen

$$\begin{aligned} Y_h &= \mathbb{R}^n, & \|\varphi\|_{Y_h} &= \max_{1\le i\le n} |\varphi_i|\,, \\ X_h &= \{0\} \times \mathbb{R}^n \times \{0\}, & \|\varphi\|_{X_h} &= \max_{1\le i\le n} |\varphi_i|\,, \\ R_{Y_h} f &= (f(ih))_{1\le i\le n}, & R_{X_h} u &= (u(ih))_{0\le i\le n+1}, \\ L_h u &= \left(\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} + u_i\right)_{1\le i\le n}. \end{aligned}$$

Mittels Taylor-Entwicklung folgt für jedes $u \in X$

$$\lim_{h\to 0} \max_{1\le i\le n} |-u''(ih) + u(ih) - (L_h R_{X_h} u)_i| = 0,$$

d.h. Konsistenz. Für $u \in C^4([0,1],\mathbb{R})$ gilt sogar

$$\|R_{Y_h} Lu - L_h R_{X_h} u\|_{Y_h} \le ch^2 \|u\|_{C^4}\,.$$

Gemäß Lemma II.4.1 (S. 75) gilt für alle $u_h \in X_h$ und $h > 0$

$$\|u_h\|_{X_h} \le \|L_h u_h\|_{Y_h}\,,$$

d.h.

$$\sup_{h>0} \left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h,X_h)} \le 1.$$

Dies ist die Stabilität. Die Konvergenz des Verfahrens haben wir in Satz II.4.2 (S. 76) gezeigt.

Im Folgenden wollen wir notwendige und hinreichende Bedingungen für die Konvergenz eines Diskretisierungsverfahrens angeben.

Satz III.2.4 (Konsistenz und Stabilität implizieren Konvergenz). *Das Diskretisierungsverfahren* (III.2.2) *sei konsistent mit* (III.2.1) *und stabil. Dann ist es konvergent.*

BEWEIS. Aus der Stabilität folgt insbesondere $L_h \in \mathrm{Isom}(X_h, Y_h)$ für alle $h > 0$. Seien $f \in Y$ beliebig, $u \in X$ eine Lösung von (III.2.1), $f_h \in Y_h$ eine Familie von Diskretisierungen von f mit

$$\lim_{h\to 0} \|R_{Y_h} f - f_h\|_{Y_h} = 0$$

und $u_h \in X_h$ die eindeutige Lösung von (III.2.2). Dann folgt

$$\begin{aligned}\|R_{X_h} u - u_h\|_{X_h} &\le \left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h,X_h)} \|L_h R_{X_h} u - L_h u_h\|_{Y_h} \\ &\le \left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h,X_h)} \Big\{ \|L_h R_{X_h} u - R_{Y_h} L u\|_{Y_h} \\ &\qquad + \|R_{Y_h} f - f_h\|_{Y_h} \Big\} \\ &\to 0 \quad \text{für } h \to 0.\end{aligned}$$

□

BEMERKUNG III.2.5. (1) Das Ersetzen von $R_{Y_h} f$ durch f_h ist für die Praxis hilfreich, da die exakte Berechnung von $R_{Y_h} f$ unter Umständen recht aufwändig ist.
(2) Wie wir in späteren Paragraphen sehen werden, ist der Nachweis der Stabilität häufig das Hauptproblem.
(3) Aus dem Beweis des Satzes III.2.4 folgt, dass man die Voraussetzungen wie folgt abschwächen kann

$$\begin{aligned}\left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h,X_h)} &\le c_1 h^{-\alpha}, \quad \alpha > 0, \quad \forall h > 0 \\ \|L_h R_{X_h} u - R_{Y_h} L u\|_{Y_h} &\le c_2 h^{\beta}, \quad \beta > \alpha, \quad \forall h > 0 \\ \|R_{Y_h} f - f_h\|_{Y_h} &\le c_3 h^{\beta} \qquad\qquad \forall h > 0.\end{aligned}$$

Dann gilt immer noch

$$\|R_{X_h} u - u_h\|_{X_h} \le c_1(c_2 + c_3) h^{\beta-\alpha} \xrightarrow[h\to 0]{} 0.$$

Die folgenden beiden Sätze zeigen, dass die Voraussetzungen von Satz III.2.4 im wesentlichen notwendig sind.

SATZ III.2.6 (Konvergenz impliziert Stabilität). *Das Diskretisierungsverfahren (III.2.2) sei konvergent. Dann ist es stabil.*

BEWEIS. Angenommen es ist $\sup_{h>0} \left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h,X_h)} = \infty$. Dann existieren $\widetilde{f}_h \in Y_h$ mit $\left\|\widetilde{f}_h\right\|_{Y_h} = 1$ für alle $h > 0$ und $\lim_{h\to 0} \left\|L_h^{-1} \widetilde{f}_h\right\|_{X_h} = \infty$. Setze

$$a(h) = \left\|L_h^{-1} \widetilde{f}_h\right\|_{X_h} \quad \text{und} \quad f_h = a(h)^{-\frac{1}{2}} \widetilde{f}_h.$$

Dann folgt $\lim_{h\to 0} \|f_h\|_{Y_h} = 0$. Betrachte (III.2.1) mit $f = 0$ und Lösung $u = 0$. Aus der Konvergenz des Verfahrens folgt

$$a(h)^{\frac{1}{2}} = \left\|L_h^{-1} f_h\right\|_{Y_h} \xrightarrow[h\to 0]{} 0.$$

Dies ist ein Widerspruch. □

SATZ III.2.7 (Konvergenz impliziert Konsistenz). *Das Diskretisierungsverfahren* (III.2.2) *sei konvergent und es gelte* $\sup_{h>0} \|L_h\|_{\mathcal{L}(X_h,Y_h)} < \infty$. *Dann ist es konsistent mit* (III.2.1).

BEWEIS. Sei $u \in X$ beliebig, $f = Lu$ und $f_h = R_{Y_h} f$. Dann folgt aus den Voraussetzungen

$$\|R_{Y_h} Lu - L_h R_{X_h} u\|_{Y_h} \le \|L_h\|_{\mathcal{L}(X_h,Y_h)} \left\|R_{X_h} u - L_h^{-1} f_h\right\|_{X_h} \xrightarrow[h\to 0]{} 0. \qquad \square$$

Bisher haben wir nur Konvergenz in den diskreten Räumen X_h bzgl. der evtl. von h abhängigen Normen $\|\cdot\|_{X_h}$. Für viele praktische Anwendungen ist es aber wünschenswert, Konvergenz gegen u in einem festen Banach-Raum zu zeigen. Dazu müssen die diskreten Lösungen u_h fortgesetzt werden. Da dies in der Topologie von X häufig sehr schwierig ist, betrachten wir einen Banach-Raum $(\widetilde{X}, \|\cdot\|_{\widetilde{X}})$ derart, dass $X \subset \widetilde{X}$ bzgl. $\|\cdot\|_{\widetilde{X}}$ dicht ist, und eine Familie $I_{X_h} : X_h \to \widetilde{X}$ linearer Operatoren. Da $\dim X_h < \infty$ ist, sind diese automatisch stetig. Der folgende Satz stellt dann eine Beziehung zwischen den R_{X_h} und den I_{X_h} her.

SATZ III.2.8 (Zusammenhang zwischen R_{X_h} und I_{X_h}). *Folgende Aussagen sind äquivalent:*

(1) *Aus* $\lim_{h\to 0} \|u_h - R_{X_h} u\|_{X_h} = 0$ *folgt* $\lim_{h\to 0} \|I_{X_h} u_h - u\|_{\widetilde{X}} = 0$.

(2) *Es gilt* $\sup_{h>0} \|I_{X_h}\|_{\mathcal{L}(X_h,\widetilde{X})} < \infty$ *und* $\|u - I_{X_h} R_{X_h} u\|_{\widetilde{X}} \xrightarrow[h\to 0]{} 0$ *für alle* $u \in X$.

BEWEIS. (1) $\Longrightarrow$ (2): Der zweite Teil von (2) folgt aus (1), indem wir $u_h = R_{X_h} u$ setzen.
Angenommen es ist $\sup_{h>0} \|I_{X_h}\|_{\mathcal{L}(X_h,\widetilde{X})} = \infty$. Dann gibt es eine Familie $u_h \in X_h$ mit $\|u_h\|_{X_h} = 1$ für alle $h > 0$ und $a(h) = \|I_{X_h} u_h\|_{\widetilde{X}} \xrightarrow[h\to 0]{} \infty$. Setze $v_h = a(h)^{-\frac{1}{2}} u_h \in X_h$. Dann folgt

$$\lim_{h\to 0} \|v_h\|_{X_h} = \lim_{h\to 0} \|v_h - R_{X_h} 0\|_{X_h} = 0$$

und wegen (1)

$$a(h)^{\frac{1}{2}} = \|I_{X_h} v_h\|_{\widetilde{X}} = \|I_{X_h} v_h - 0\|_{\widetilde{X}} \xrightarrow[h\to 0]{} 0.$$

Dies ist ein Widerspruch.
(2) $\Longrightarrow$ (1): Folgt aus

$$u - I_{X_h} u_h = [u - I_{X_h} R_{X_h} u] + I_{X_h}[R_{X_h} u - u_h]. \qquad \square$$

BEISPIEL III.2.9 (Sturm-Liouville-Problem). Wir wählen in Beispiel III.2.3

$$\widetilde{X} = C_0([0,1], \mathbb{R}), \quad \|\varphi\|_{\widetilde{X}} = \max_{0\le x\le 1} |\varphi(x)|$$

und für I_{X_h} die stetige, stückweise lineare Interpolierende in den Knoten $ih, 0 \le i \le n+1$. Dann folgt sofort

$$\|I_{X_h}\|_{\mathcal{L}(X_h,\widetilde{X})} = 1$$

und

$$\|u - I_{X_h} R_{X_h} u\|_{\widetilde{X}} \le ch^2 \|u\|_{C^2}$$

für alle $u \in X$.

III.3. Elliptische Differentialgleichungen

Wir betrachten in diesem Paragraphen die Reaktions-Diffusions-Gleichung

$$\begin{aligned} -\nabla \cdot (A\nabla u) + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma = \partial\Omega. \end{aligned} \tag{III.3.1}$$

Dabei ist $\Omega \subset \mathbb{R}^n$ ein beschränktes, zusammenhängendes, offenes Gebiet und $\alpha \in C(\overline{\Omega}, \mathbb{R}_+)$, $A \in C^1(\Omega, \mathbb{R}^{n\times n}) \cap C(\overline{\Omega}, \mathbb{R}^{n\times n})$ mit $A_{ij}(x) = A_{ji}(x)$ für alle $x \in \overline{\Omega}$ und $1 \le i, j \le n$ und strikt positivem kleinsten Eigenwert, d.h.

$$\lambda_0 = \inf_{x\in\Omega} \min_{z\in\mathbb{R}^n\setminus\{0\}} \frac{z^T A(x) z}{z^T z} > 0. \tag{III.3.2}$$

Im Rahmen von §III.2 setzen wir

$$\begin{aligned} X &= C^2(\overline{\Omega}, \mathbb{R}) \cap C_0(\overline{\Omega}, \mathbb{R}), \quad \|\varphi\|_X = \sup_{x\in\Omega} \max_{\substack{\alpha\in\mathbb{N}^n \\ |\alpha|\le 2}} |D^\alpha \varphi(x)|, \\ Y &= C(\overline{\Omega}, \mathbb{R}), \qquad\qquad\quad \|\varphi\|_Y = \sup_{x\in\Omega} |\varphi(x)|, \\ L\varphi &= -\nabla \cdot (A\nabla\varphi) + \alpha\varphi. \end{aligned}$$

Für die Räume X_h, Y_h benötigen wir den Begriff des Gitters.

Definition III.3.1 (Äquidistantes Gitter). Sei $h > 0$.
(1) Die Menge

$$G_h = \{x = \underline{i}h : \underline{i} \in \mathbb{Z}^n\}$$

heißt *Gitter* der *Maschenweite* h.
(2) Definiere (vgl. Abb. III.3.1)

$$\begin{aligned} \overline{\Omega}_h &= \overline{\Omega} \cap G_h, \\ \Gamma_h &= \{x \in \overline{\Omega}_h : \ \operatorname{dist}(x, \Gamma) < h\}, \\ \Omega_h &= \overline{\Omega}_h \setminus \Gamma_h. \end{aligned}$$

Dabei ist

$$\operatorname{dist}(x, \Gamma) = \inf_{y\in\Gamma} \|x - y\|.$$

Die Menge Γ_h heißt *diskreter Rand* von Ω_h.
(3) Definiere die Räume X_h und Y_h durch

$$X_h = \{u_h : \overline{\Omega}_h \to \mathbb{R} : u_h(x) = 0 \text{ für alle } x \in \Gamma_h\}$$
$$Y_h = \{v_h : \Omega_h \to \mathbb{R}\}.$$

(4) Die Operatoren $R_{X_h} : X \to X_h$ und $R_{Y_h} : Y \to Y_h$ sind definiert durch

$$(R_{X_h}u)(x) = \begin{cases} u(x) & \text{für alle } x \in \Omega_h \\ 0 & \text{für alle } x \in \Gamma_h \end{cases}$$
$$(R_{Y_h}v)(x) = v(x) \quad \text{für alle } x \in \Omega_h.$$

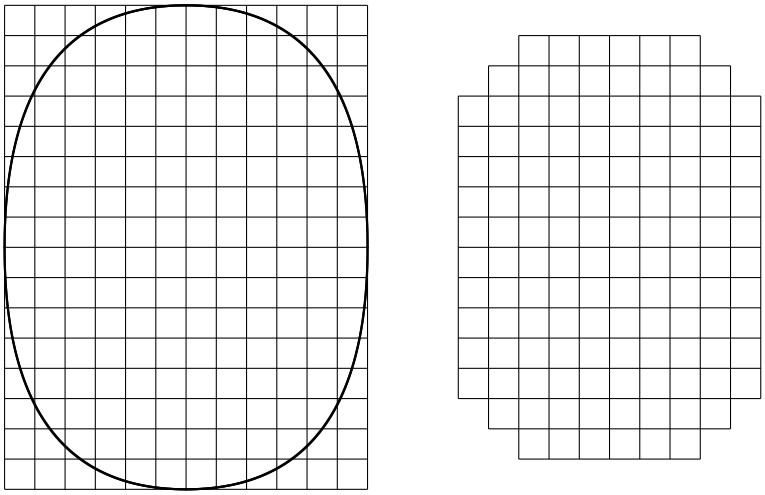

ABBILDUNG III.3.1. Gebiet Ω mit äquidistantem Gitter G_h (links) und Approximation $\overline{\Omega}_h$ (rechts)

BEMERKUNG III.3.2 (Ergänzungen). (1) Die Normen von X_h und Y_h legen wir später fest.
(2) Sei $N_h = \#\Omega_h$. Dann ist $X_h \cong \mathbb{R}^{N_h}$. Da Ω beschränkt ist, gibt es eine nur von Ω abhängige Konstante c_Ω mit $N_h \le c_\Omega h^{-n}$.
(3) Man kann auch nicht äquidistante Gitter einführen (vgl. Abb. III.3.2). Betrachte dazu für $1 \le k \le n$ streng monoton wachsende Folgen $(x_{k,m})_{m\in\mathbb{Z}}$ mit

$$\lim_{m\to-\infty} x_{k,m} = -\infty, \lim_{m\to+\infty} x_{k,m} = +\infty$$

für alle $1 \le k \le n$. Dann wird durch

$$G_h = \{(x_{1,m_1}, \dots, x_{n,m_n}) : m_k \in \mathbb{Z}, 1 \le k \le n\}$$

ein nicht äquidistantes Gitter definiert. Die Definitionen von $\overline{\Omega}_h, \Gamma_h$ und Ω_h übertragen sich analog. Das richtige Analogon zu h ist

$$\max_{1\le k\le n} \sup_{m\in\mathbb{Z}} |x_{k,m} - x_{k,m+1}| .$$

Die Abschätzung für $N_h = \#\Omega_h$ bleibt gültig, wenn gilt

$$\min_{1\le k\le n} \inf_{m\in\mathbb{Z}} |x_{k,m} - x_{k,m+1}| > 0.$$

Die Beschränkung auf äquidistante Gitter ist nicht wesentlich, vereinfacht aber im Folgenden die Argumentation.

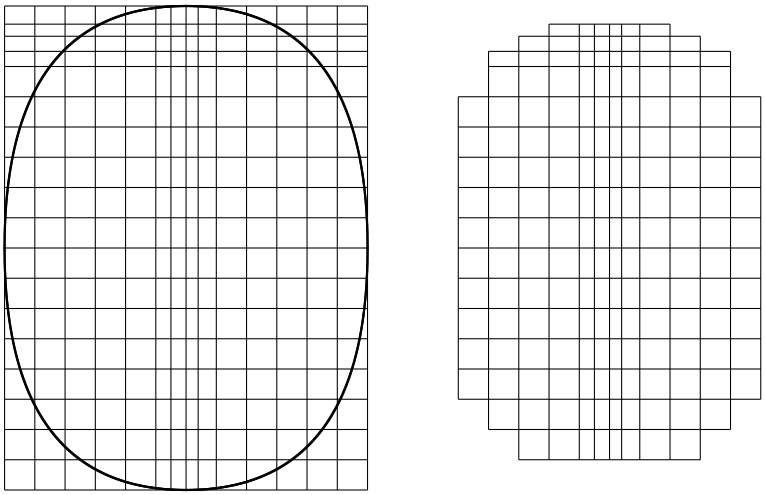

ABBILDUNG III.3.2. Gebiet Ω mit nicht äquidistantem Gitter G_h (links) und Approximation $\overline{\Omega}_h$ (rechts)

Für die Definition von L_h benötigen wir die vorwärts und rückwärts Differenzenquotienten.

DEFINITION III.3.3 (Differenzenquotienten). Seien $h > 0$, $1 \le k \le n$ und $x \in \mathbb{R}^n$. Dann heißen

$$\partial^{\pm}_{h,k} u(x) = \pm\frac{1}{h}[u(x \pm he_k) - u(x)]$$

der k-te *vorwärts* $(^+)$ bzw. *rückwärts* $(^-)$ *Differenzenquotient*. Dabei ist e_k der k-te Einheitsvektor.

LEMMA III.3.4 (Fehlerabschätzung für Differenzenquotienten). *(1) Für $u \in C^1(\overline{\Omega}, \mathbb{R})$ und $1 \le k \le n$ gilt*

$$\lim_{h\to 0} \max_{x\in\Omega_h} \left|\partial^{\pm}_{h,k} u(x) - \partial_k u(x)\right| = 0.$$

(2) Sei $m \in \mathbb{N}^*$ und $u \in C^{m+1}(\Omega, \mathbb{R})$. Dann gibt es für alle $x \in \Omega_h$ und alle $1 \le k \le n$ ein $\theta \in (0,1)$ mit

$$\partial^{\pm}_{h,k} u(x) = \partial_k u(x) + \sum_{l=1}^{m-1} \frac{1}{(l+1)!} (\pm h)^l \partial_k^{l+1} u(x) + \frac{1}{(m+1)!} (\pm h)^m \partial_k^{m+1} u(x \pm \theta h e_k).$$

(3) Sei $u \in C^4(\Omega, \mathbb{R})$. Dann gibt es für alle $x \in \Omega_h$ und alle $1 \le k \le n$ ein $\theta \in (-1, 1)$ mit

$$\begin{aligned}\partial^+_{h,k}(\partial^-_{h,k}u)(x) &= \partial^-_{h,k}(\partial^+_{h,k}u)(x)\\ &= \partial^2_k u(x) + \frac{1}{12}h^2\partial^4_k u(x + \theta h e_k).\end{aligned}$$

Beweis. ad (1): Folgt direkt aus der Definition der partiellen Ableitungen und der gleichmäßigen Stetigkeit von $\partial_k u$ (Beachte: $\overline{\Omega}$ ist kompakt!).
ad (2): Folgt direkt aus der Taylor-Formel.
ad (3): Folgt aus der Taylor-Formel und

$$\partial^+_{h,k}(\partial^-_{h,k}u)(x) = \partial^-_{h,k}(\partial^+_{h,k}u)(x) = \frac{1}{h^2}[u(x+he_k)-2u(x)+u(x-he_k)]. \quad \square$$

Lemma III.3.4 (3) und die Ergebnisse von §III.2 legen folgende Definition von L_h nahe.

Definition III.3.5 (Differenzendiskretisierung). (1) Die Differenzendiskretisierung $L_h : X_h \to Y_h$ des Differentialoperators L aus Gleichung (III.3.1) ist für alle $x \in \Omega_h$ definiert durch

$$(L_h u)(x) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \partial^-_{h,i}(A_{ij}\partial^+_{h,j}u)(x) + \alpha u(x). \tag{III.3.3}$$

(2) Die Differenzendiskretisierung von (III.3.1) ist gegeben durch: Finde $u_h \in X_h$ mit

$$L_h u_h(x) = f(x) \quad \text{für alle } x \in \Omega_h. \tag{III.3.4}$$

Bemerkung III.3.6 (Eigenschaften). (1) Man kann in (III.3.3) die Rollen von $\partial^+_{h,j}$ und $\partial^-_{h,i}$ vertauschen.
(2) Gleichung (III.3.4) ist ein LGS mit N_h Gleichungen und N_h Unbekannten. Aus den Ergebnissen dieses Paragraphen folgt, dass die entsprechende Matrix symmetrisch positiv definit ist. Außerdem ist sie *dünn besetzt*: Das Element μ, ν ist höchstens dann von Null verschieden, wenn für die zu diesen Indizes gehörenden Gitterpunkte x und y gilt $\|x - y\|_\infty \le h$, wobei $\|\cdot\|_\infty$ die Maximum-Norm auf $\mathbb{R}^n$ ist.
Wir werden uns in §III.6 ausführlich mit den Eigenschaften dieses LGS und seiner numerischen Lösung beschäftigen.

Für die Konvergenzanalyse müssen wir Normen auf X_h und Y_h spezifizieren.

Definition III.3.7 (Diskretes Skalarprodukt und Normen). Das *diskrete L^2-Skalarprodukt* $(\cdot,\cdot)_h$, die *diskrete L^2-Norm* $\|\cdot\|_{0,h}$ und die

diskrete H^1-Norm $\|\cdot\|_{1,h}$ auf Ω_h sind definiert durch

$$(u\,,\,v)_h = h^n \sum_{x\in\Omega_h} u(x)v(x),$$
$$\|u\|_{0,h} = (u\,,\,u)_h^{\frac{1}{2}}\,,$$
$$\|u\|_{1,h} = \left\{ \sum_{k=1}^{n} \left\|\partial_{h,k}^{+} u\right\|_{0,h}^2 \right\}^{\frac{1}{2}}.$$

Bemerkung III.3.8 (Skalierung, Wahl der Differenzenquotienten).
(1) Man prüft leicht nach, dass $(\cdot\,,\,\cdot)_h$ und $\|\cdot\|_{0,h}$ ein Skalarprodukt und eine Norm sind. Um einzusehen, dass $\|\cdot\|_{1,h}$ auf X_h eine Norm ist, beachte man, dass aus $\|u\|_{1,h} = 0$ folgt $u =$ konstant.
(2) Die Skalierung von $\|\cdot\|_{0,h}$ ist so gewählt, dass $\|1\|_{0,h} \approx 1$ ist und es eine von h unabhängige Konstante c gibt mit $\|u\|_{0,h} \le c\,\|u\|_\infty$ für alle $u \in X_h$.
(3) In der Definition von $\|\cdot\|_{1,h}$ können die vorwärts Differenzenquotienten $\partial_{h,k}^{+}$ durch die rückwärts Differenzenquotienten $\partial_{h,k}^{-}$ ersetzt werden.

Im Folgenden setzen wir

$$\|\cdot\|_{X_h} = \|\cdot\|_{1,h}\,, \quad \|\cdot\|_{Y_h} = \|\cdot\|_{0,h}\,. \tag{III.3.5}$$

Lemma III.3.9 (Konsistenz). *Das Diskretisierungsverfahren der Gleichungen* (III.3.3), (III.3.4) *ist bzgl. der Normen* (III.3.5) *konsistent. Für alle* $u \in C^3(\overline{\Omega}, \mathbb{R})$ *gilt zudem*

$$\|L_h R_{X_h} u - R_{Y_h} L u\|_{0,h} \le ch\,\|u\|_{C^3(\overline{\Omega},\mathbb{R})}\,\|A\|_{C^2(\overline{\Omega},\mathbb{R})}\,.$$

Ist zusätzlich A eine konstante Diagonalmatrix, so gilt für alle $u \in C^4(\overline{\Omega}, \mathbb{R})$

$$\|L_h R_{X_h} u - R_{Y_h} L u\|_{0,h} \le c' h^2\,\|u\|_{C^4(\overline{\Omega},\mathbb{R})}\,.$$

Beweis. Die Behauptung folgt aus Lemma III.3.4. □

Für die Stabilität brauchen wir zwei Hilfsergebnisse. Zur Motivation des ersten Hilfsergebnisses betrachte eine Funktion $u \in C^1(\Omega, \mathbb{R}) \cap C(\overline{\Omega}, \mathbb{R})$ mit $u = 0$ auf Γ. Da Ω beschränkt ist, gibt es ein $R > 0$ mit $\overline{\Omega} \subset [-R, R]^n$. Setze u durch Null auf $[-R, R]^n$ fort. Dann gilt für die $L^2(\Omega)$-Norm $\|\cdot\|$

$$\|u\|^2 = \int_\Omega |u|^2 = \int_{[-R,R]^n} |u|^2\,, \quad \|\partial_1 u\|^2 = \int_\Omega |\partial_1 u|^2 = \int_{[-R,R]^n} |\partial_1 u|^2\,.$$

Betrachte einen beliebigen Punkt $x \in [-R, R]^n$ und schreibe ihn in der Form $x = (x_1, x')$ mit $x' \in [-R, R]^{n-1}$. Wegen $u(R, x') = 0$ und der Cauchy-Schwarzen Ungleichung ist dann

$$|u(x)|^2 = \left| \int_{x_1}^{R} \partial_1 u(t, x')\mathrm{d}t \right|^2 \le 2R \int_{-R}^{R} |\partial_1 u(t, x')|^2\,\mathrm{d}t.$$

Integration bzgl. x liefert wegen des Satzes von Fubini

$$\int_{[-R,R]^n} |u|^2 \leq 4R^2 \int_{[-R,R]^n} |\partial_1 u|^2 .$$

Insgesamt ergibt sich somit die *Friedrichsche Ungleichung*

$$\|u\| \leq \frac{2R}{\sqrt{n}} \|\nabla u\| .$$

LEMMA III.3.10 (Diskrete Friedrichsche Ungleichung). *Es gibt eine nur von Ω abhängige Konstante c_Ω, so dass für alle $u \in X_h$ gilt*

$$\|u\|_{0,h} \leq c_\Omega \|u\|_{1,h} .$$

BEWEIS. Der Beweis imitiert obige Vorüberlegung. Seien $R > 0$ und $N \in \mathbb{N}^*$ so, dass $\overline{\Omega} \subset [-R,R]^n$ und $R \leq Nh < R + h$ ist. Setze $D_h = G_h \cap [-R,R]^n$. Funktionen aus X_h setzen wir durch Null konstant auf D_h fort. Jedes $x \in \Omega_h$ können wir schreiben als $x = (ih, x')$ mit $-N \leq i \leq N$ und $x' \in \mathbb{R}^{n-1}$. Wegen $u(Nh, x') = 0$ folgt aus der Cauchy-Schwarzschen Ungleichung für Summen

$$\begin{aligned}
|u(x)| &= \left| \sum_{\ell=i}^{N-1} [u((\ell+1)h, x') - u(\ell h, x')] \right| \\
&\leq \sum_{\ell=i}^{N-1} h \left| \partial^+_{h,1} u(\ell h, x') \right| \\
&\leq \left\{ \sum_{\ell=i}^{N-1} h^2 \right\}^{\frac{1}{2}} \left\{ \sum_{\ell=i}^{N-1} \left| \partial^+_{h,1} u(\ell h, x') \right|^2 \right\}^{\frac{1}{2}} \\
&\leq \left\{ 2Nh^2 \right\}^{\frac{1}{2}} \left\{ \sum_{\ell=-N}^{N-1} \left| \partial^+_{h,1} u(\ell h, x') \right|^2 \right\}^{\frac{1}{2}}
\end{aligned}$$

und damit

$$\begin{aligned}
\|u\|_{0,h} &\leq \left\{ h^n 2Nh^2 \sum_{x \in \Omega_h} \sum_{\ell=-N}^{N} \left| \partial^+_{h,1} u(\ell h, x') \right|^2 \right\}^{\frac{1}{2}} \\
&\leq \left\{ h^n 4N^2 h^2 \sum_{x \in D_h} \left| \partial^+_{h,1} u(x) \right|^2 \right\}^{\frac{1}{2}} \\
&\leq c_\Omega \|u\|_{1,h} .
\end{aligned}$$

□

LEMMA III.3.11 (Diskrete partielle Integration). *Für alle $u, v \in X_h$ und alle $1 \leq k \leq n$ gilt*

$$\left(\partial^-_{h,k} u \, , \, v \right)_h = - \left(u \, , \, \partial^+_{h,k} v \right)_h \quad \textit{und} \quad \left(\partial^+_{h,k} u \, , \, v \right)_h = - \left(u \, , \, \partial^-_{h,k} v \right)_h .$$

BEWEIS. Wegen der Symmetrie des Skalarproduktes folgt die zweite Identität aus der ersten durch Vertauschen von u und v. Zum Nachweis der ersten Identität können wir uns auf den Fall $k = 1$ beschränken, da der allgemeine Fall aus diesem durch Umnummerieren der Koordinaten folgt. Mit den gleichen Notationen wie im Beweis von Lemma III.3.10 folgt

$$\begin{aligned}\left(\partial^-_{h,1}u\,,\,v\right)_h &= h^n \sum_{x'} \sum_{i=-N+1}^{N} \partial^-_{h,1}u(ih,x')v(ih,x')\\ \left(u\,,\,\partial^+_{h,1}v\right)_h &= h^n \sum_{x'} \sum_{j=-N}^{N-1} u(jh,x')\partial^+_{h,1}v(jh,x')\end{aligned}$$

und

$$\begin{aligned}&\sum_{i=-N+1}^{N} \partial^-_{h,1}u(ih,x')v(ih,x')\\ &\quad= \sum_{i=-N+1}^{N} h^{-1}[u(ih,x') - u((i-1)h,x')]v(ih,x')\\ &\quad= \sum_{i=-N+1}^{N} h^{-1}u(ih,x')v(ih,x') - \sum_{j=-N}^{N-1} h^{-1}u(jh,x')v((j+1)h,x')\\ &\quad= \sum_{j=-N}^{N-1} u(jh,x')\partial^+_{h,1}v(jh,x')\end{aligned}$$

□

LEMMA III.3.12 (Stabilität). *Das Diskretisierungsverfahren der Gleichungen* (III.3.3), (III.3.4) *ist bzgl. der Normen* (III.3.5) *stabil.*

BEWEIS. Sei $u \in X_h \setminus \{0\}$ beliebig. Dann folgt aus Lemma III.3.11

$$\begin{aligned}(L_h u\,,\,u)_h &= \left(-\sum_{i=1}^{n}\sum_{j=1}^{n} \partial^-_{h,i}(A_{ij}\partial^+_{h,j}u) + \alpha u\,,\,u\right)_h\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \left(A_{ij}\partial^+_{h,j}u\,,\,\partial^+_{h,i}u\right)_h + (\alpha u\,,\,u)_h\,.\end{aligned}$$

Da α nicht negativ ist, ist $(\alpha u\,,\,u)_h \geq 0$. Andererseits folgt aus (III.3.2)

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} \left(A_{ij}\partial^+_{h,j}u\,,\,\partial^+_{h,i}u\right)_h &= h^n \sum_{x\in\Omega_h}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}(x)\partial^+_{h,j}u(x)\partial^+_{h,i}u(x)\\ &\geq \lambda_0 h^n \sum_{x\in\Omega_h}\sum_{k=1}^{n} \left|\partial^+_{h,k}u(x)\right|^2\\ &= \lambda_0 \left\|u\right\|^2_{1,h}\,.\end{aligned}$$

Lemma III.3.10 impliziert

$$(L_h u\,,\, u)_h \leq \|L_h u\|_{0,h}\, \|u\|_{0,h} \leq c_\Omega\, \|L_h u\|_{0,h}\, \|u\|_{1,h}\,.$$

Also ist

$$\lambda_0\, \|u\|_{1,h} \leq c_\Omega\, \|L_h u\|_{0,h}$$

und somit

$$\left\|L_h^{-1}\right\|_{\mathcal{L}(Y_h, X_h)} \leq c_\Omega \lambda_0^{-1}. \qquad \square$$

Satz III.3.13 (Fehlerabschätzung). *Für die Lösung u von (III.3.1) gelte $u \in C^3(\overline{\Omega}, \mathbb{R})$. Dann gilt für die Lösung u_h von (III.3.4) die Fehlerabschätzung*

$$\|u - u_h\|_{0,h} \leq c_\Omega\, \|u - u_h\|_{1,h} \leq ch\, \|u\|_{C^3}\, \|A\|_{C^2}\,.$$

Ist zusätzlich A eine konstante Diagonalmatrix und gilt $u \in C^4(\overline{\Omega}, \mathbb{R})$, so gilt sogar

$$\|u - u_h\|_{0,h} \leq c_\Omega\, \|u - u_h\|_{1,h} \leq c' h^2\, \|u\|_{C^4}\,.$$

Beweis. Die Behauptung folgt aus Satz III.2.4 (S. 93), Lemma III.3.9, Bemerkung III.3.8 (2), Lemma III.3.10 und Lemma III.3.12. $\square$

Bemerkung III.3.14 (Fehlerabschätzung bzgl. der L^2-Norm). Um Satz III.2.8 (S. 95) anzuwenden, wählen wir für I_{X_h} die n-lineare Interpolierende zwischen den Gitterpunkten und $\widetilde{X} = L^2(\Omega, \mathbb{R})$, $\|\cdot\|_{\widetilde{X}} = \|\cdot\|_{L^2}$.

Beispiel III.3.15 (Poisson-Gleichung). Wir betrachten die Poisson-Gleichung, d.h. (III.3.1) mit $A = I$ und $\alpha = 0$, auf dem Quadrat $\Omega = [-1, 1]^2$ mit homogenen Randbedingungen und glatter exakter Lösung $u(x, y) = (1 - x^2)(1 - y^2)$ und auf dem L-förmigen Gebiet $\Omega = [-1, 1]^2 \setminus [0, 1] \times [-1, 0]$ mit inhomogenen Randbedingungen und singulärer exakter Lösung $u(r\cos\varphi, r\sin\varphi) = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\varphi\right)$. Abbildung III.3.3 zeigt die Konturlinien der beiden diskreten Lösungen zu $h = \frac{1}{64}$. Tabelle III.3.1 gibt den Fehler $\|u - u_h\|_{1,h}$ und die geschätzte Konvergenzordnung $\frac{\ln(\|u-u_{2h}\|_{1,h}) - \ln(\|u-u_h\|_{1,h})}{\ln 2}$ für die beiden Beispiele an. Sie zeigt deutlich den Einfluss der Regularität der Lösung u auf die Konvergenzgeschwindigkeit der Diskretisierung.

Im Rest dieses Paragraphen behandeln wir in Ergänzung der obigen Ergebnisse die Behandlung inhomogener Dirichlet-Randbedingungen, die Diskretisierung Neumannscher Randbedingungen und die Behandlung von Ableitungen erster Ordnung der Form $a \cdot \nabla u$.

Beispiel III.3.16 (Inhomogene Dirichlet-Randbedingungen). Für die pDGl (III.3.1) mit der inhomogenen Dirichlet-Randbedingung $u =$

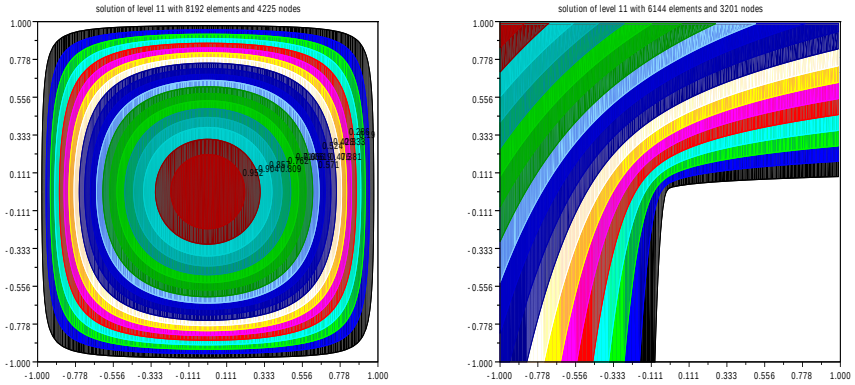

ABBILDUNG III.3.3. Konturlinien der diskreten Lösungen aus Beispiel III.3.15 zu $h = \frac{1}{64}$

TABELLE III.3.1. Fehler $\|u - u_h\|_{1,h}$ und geschätzte Konvergenzordnung $\frac{\ln(\|u-u_{2h}\|_{1,h})-\ln(\|u-u_h\|_{1,h})}{\ln 2}$ für Beispiel III.3.15

	glatte Lösung		singuläre Lösung	
h^{-1}	Fehler	Ordnung	Fehler	Ordnung
4	0.3130		0.0932	
8	0.1614	0.96	0.0667	0.48
16	0.0815	0.99	0.0435	0.62
32	0.0410	0.99	0.0278	0.65
64	0.0205	1.00	0.0176	0.66

g auf Γ ersetzen wir X, Y, X_h ,Y_h, L und L_h durch

$$X = C^2(\Omega, \mathbb{R}) \cap C(\overline{\Omega}, \mathbb{R}), \qquad Y = C(\Omega, \mathbb{R}) \times C(\Gamma, \mathbb{R}),$$
$$X_h = \{u : \overline{\Omega}_h \to \mathbb{R}\}, \qquad Y_h = \{u : \Omega_h \to \mathbb{R}\} \times \{\varphi : \Gamma_h \to \mathbb{R}\},$$
$$Lu = \left(-\nabla \cdot (A\nabla u) + \alpha u, u\big|_\Gamma\right)$$

und

$$L_h u_h(x) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \partial^-_{h,i}(A_{ij}\partial^+_{h,j}u_h)(x) + \alpha(x)u_h(x) \quad \text{für } x \in \Omega_h$$
$$L_h u_h(x) = u_h(x) \quad \text{für } x \in \Gamma_h.$$

Das LGS (III.3.4) geht über in

$$\begin{aligned} L_h u_h(x) &= f(x) \quad \text{für alle } x \in \Omega_h \\ u_h(x) &= g_h(x) \quad \text{für alle } x \in \Gamma_h. \end{aligned} \tag{III.3.6}$$

Man beachte, dass dieses LGS äquivalent dazu ist, dass man auf der linken Seite von (III.3.6) u_h in allen diskreten Randpunkten gleich g_h

setzt und die entsprechenden Terme auf die rechte Seite schafft. Daher ist (III.3.6) wie (III.3.4) ein LGS mit $N_h = \#\Omega_h$ Gleichungen und Unbekannten. Die Funktion g_h ist wie folgt definiert: Für $x \in \Gamma_h$ sei $\pi x \in \Gamma$ gegeben durch $\|x - \pi x\| = \inf_{y\in\Gamma} \|x - y\|$. Falls h hinreichend klein ist, ist πx eindeutig definiert. Dann ist $g_h(x) = g(\pi x)$. Man kann zeigen, dass mit diesen Modifikationen Satz III.3.13 gültig bleibt.

Beispiel III.3.17 (Inhomogene Neumann-Randbedingungen). Wir betrachten die pDGl (III.3.1) mit der Neumannschen Randbedingung $\nu \cdot \nabla u = g$ auf Γ. Dabei ist $\|\nu\| = 1$. Zusätzlich gebe es ein $\varepsilon > 0$ mit $x - t\nu \in \Omega$ für alle $0 < t < \varepsilon$ und $x \in \Gamma$, d.h. ν zeige nach außen. Besonders wichtige Spezialfälle sind $\nu = \mathbf{n}$ und $\nu = \|\mathbf{n} \cdot A\|^{-1} \mathbf{n} \cdot A$, wobei $\mathbf{n}$ die äußere Normale an Γ ist. Wir ersetzen jetzt X, Y, X_h, Y_h, L und L_h durch

$$\begin{aligned} X &= C^2(\Omega, \mathbb{R}) \cap C^1(\overline{\Omega}, \mathbb{R}), & Y &= C(\Omega, \mathbb{R}) \times C(\Gamma, \mathbb{R}), \\ X_h &= \{u : \overline{\Omega}_h \to \mathbb{R}\}, & Y_h &= \{u : \overline{\Omega}_h \to \mathbb{R}\}, \\ Lu &= \left(-\nabla \cdot (A\nabla u) + \alpha u, \nu \cdot \nabla u\big|_\Gamma\right) \end{aligned}$$

und

$$\begin{aligned} L_h u_h(x) &= -\sum_{i=1}^{n} \sum_{j=1}^{n} \partial^-_{h,i}(A_{ij} \partial^+_{h;i} u_h)(x) + \alpha(x) u_h(x) && \text{für } x \in \Omega_h \\ L_h u_h(x) &= \partial_{h,\nu} u_h(x) && \text{für } x \in \Gamma_h. \end{aligned}$$

Das LGS (III.3.4) geht über in

$$\begin{aligned} L_h u_h(x) &= f(x) && \text{für alle } x \in \Omega_h \\ \partial_{h,\nu} u_h &= g_h(x) && \text{für alle } x \in \Gamma_h. \end{aligned} \tag{III.3.7}$$

Man beachte, dass (III.3.7) ein LGS mit $\overline{N}_h = \#\overline{\Omega}_h$ Gleichungen und Unbekannten ist. Die Funktion g_h ist wie in Bemerkung III.3.16 definiert. Der Differenzenquotient $\partial_{h,\nu}$ wird wie folgt konstruiert (vgl. Abb. III.3.4): Sei $x \in \Gamma_h$ und $\gamma_x = \{y \in \mathbb{R}^n : \|x - y\|_\infty = h\}$. Bezeichne mit $\underline{x}$ den Schnittpunkt der Geraden durch x in Richtung $-\nu$ mit γ_x. Falls h hinreichend klein ist, gilt $\underline{x} \in \Omega$. Setze $\partial_{h,\nu} u_h(x) = \|x - \underline{x}\|^{-1} [u_h(x) - \underline{u}_h(\underline{x})]$. Dabei ist $\underline{u}_h(\underline{x})$ eine geeignete Interpolierende in einem oder mehreren Gitterpunkten, die einen Abstand $\le h$ von $\underline{x}$ haben: $\underline{u}_h(\underline{x}) = u_h(x_i)$ mit $\|\underline{x} - x_i\| = \min_{j=0,1} \|\underline{x} - x_j\|$ oder $\underline{u}_h(\underline{x}) = \theta u_h(x_0) + (1-\theta) u_h(x_1)$ mit $\underline{x} = \theta x_0 + (1-\theta) x_1$. Man kann wiederum zeigen, dass Satz III.3.13 gültig bleibt, sofern h hinreichend klein ist.

Bemerkung III.3.18 (Gemischte Randbedingungen). Die Aussagen von Bemerkung III.3.16 und III.3.17 bleiben gültig, wenn die entsprechenden Randbedingungen nur auf Teilen des Randes gelten.

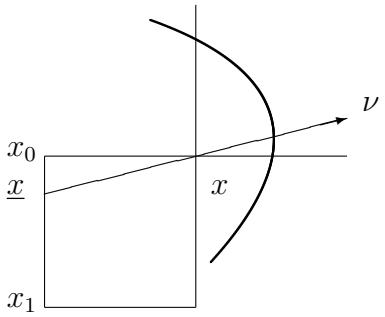

ABBILDUNG III.3.4. Konstruktion von $\partial_{h,\nu} u_h(x)$

Zum Abschluss betrachten wir die elliptische pDGl

$$\begin{aligned} -\nabla \cdot (A\nabla u) + a \cdot \nabla u + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned} \tag{III.3.8}$$

mit A und α wie in (III.3.1) und $a \in C(\overline{\Omega}, \mathbb{R}^n)$. Wir behalten X, Y, X_h, Y_h, R_{X_h} und R_{Y_h} bei und definieren

$$\begin{aligned} Lu &= -\nabla \cdot (A\nabla u) + a \cdot \nabla u + \alpha u \\ L_h u &= -\sum_{i=1}^{n} \sum_{j=1}^{n} \partial_{h,i}^{-} (A_{ij} \partial_{h,j}^{+} u) + \sum_{i=1}^{n} a_i \partial_{h,i}^{+} u + \alpha u. \end{aligned} \tag{III.3.9}$$

SATZ III.3.19 (Ladyzhenskaya-Bedingung). *Der Differentialoperator L aus (III.3.9) erfülle die sog.* Ladyzhenskaya-Bedingung*, d.h., es gibt ein $\mu_0 > 0$ mit*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} z_i A_{ij}(x) z_j + \sum_{i=1}^{n} a_i(x) z_i z_0 + \alpha(x) z_0^2 \geq \mu_0 \sum_{i=1}^{n} z_i^2$$

für alle $x \in \Omega$ und $(z_0, \ldots, z_n) \in \mathbb{R}^{n+1}$. Dann besitzt das LGS

$$L_h u_h(x) = f(x) \quad \text{für alle } x \in \Omega_h$$

mit L_h wie in (III.3.9) in X_h eine eindeutige Lösung. Gilt für die Lösung u von (III.3.8) $u \in C^3(\overline{\Omega}, \mathbb{R})$, so ist

$$\|u - u_h\|_{0,h} \leq \|u - u_h\|_{1,h} \leq ch \|u\|_{C^3} \left\{ \|A\|_{C^2} + \|a\|_{C^0} \right\}.$$

BEWEIS. Aus der Ladyzhenskaya-Bedingung und Lemma III.3.11 folgt mit $z_0 = u(x)$ und $z_i = \partial^+_{h,i}u(x)$

$$\begin{aligned}(L_h u\,,\, u)_h &= h^n \sum_{x\in\Omega_h} \left\{ \sum_{i=1}^n \sum_{j=1}^n A_{ij}(x)\partial^+_{h,i}u(x)\partial^+_{h,j}u(x) \right. \\ &\qquad \left. + \sum_{i=1}^n a_i(x)\partial^+_{h,i}u(x)u(x) + \alpha(x)u(x)^2 \right\} \\ &\geq \mu_0 h^n \sum_{x\in\Omega_h} \sum_{k=1}^n \left|\partial^+_{h,k}u(x)\right|^2 \\ &= \mu_0 \left\|u\right\|^2_{1,h}\,.\end{aligned}$$

Hieraus ergibt sich wie in Lemma III.3.12 die Stabilität des Diskretisierungsverfahrens bzgl. der Normen (III.3.5). Die Konsistenz folgt aus Lemma III.3.4 und Bemerkung III.3.8 (2). Damit ergibt sich die Behauptung aus Satz III.2.4 (S. 93). □

Das folgende Beispiel zeigt, dass die Ladyzhenskaya Bedingung nicht immer erfüllt ist.

BEISPIEL III.3.20 (Verletzte Ladyzhenskaya-Bedingung). In $\mathbb{R}^2$ sei

$$Lu = -\Delta u + a\partial_x u + \alpha u$$

mit $a, \alpha \in \mathbb{R}$, $\alpha > 0$. Dann ist die Ladyzhenskaya-Bedingung erfüllt, wenn für alle $(z_0, z_1, z_2) \in \mathbb{R}^3$ gilt

$$z_1^2 + z_2^2 + az_1z_0 + \alpha z_0^2 \geq \mu_0(z_1^2 + z_2^2)$$

mit $\mu_0 > 0$. Quadratische Ergänzung der linken Seite und etwas Rechnen liefert, dass eine solche Ungleichung nur gelten kann, wenn $a^2 < 4\alpha$ ist.

Das folgende eindimensionale Beispiel zeigt, dass die Diskretisierung (III.3.9) schlecht ist, falls die Ladyzhenskaya Bedingung nicht erfüllt und h nicht hinreichend klein ist.

BEISPIEL III.3.21 (Sturm-Liouville-Problem mit verletzter Ladyzhenskaya-Bedingung). Betrachte das Sturm-Liouville-Problem

$$\begin{aligned} -u'' + au' &= 0 \quad \text{in } (0,1) \\ u(0) &= 0 \\ u(1) &= 1 \end{aligned}$$

mit $a > 0$. Die Ladyzhenskaya Bedingung lautet jetzt $z_1^2 + az_1z_0 \geq \mu_0 z_1^2$ für alle $(z_0, z_1) \in \mathbb{R}^2$. Die Wahl $z_0 = 1$, $z_1 = -\frac{a}{2}$ zeigt, dass sie nicht erfüllt werden kann. Die exakte Lösung des RWP lautet

$$u(x) = \frac{e^{ax} - 1}{e^a - 1}.$$

Mit $h = \frac{1}{N+1}$, $x_i = ih$, $0 \le i \le N+1$, und $\mu = ah$ liefert (III.3.9) das LGS

$$\begin{aligned} -(1-\mu)u_{i+1} + (2-\mu)u_i - u_{i-1} &= 0 \quad 1 \le i \le N \\ u_0 &= 0 \\ u_{N+1} &= 1. \end{aligned} \tag{III.3.10}$$

Dieses ist für $\mu = 1$ offensichtlich nicht lösbar. Aus der Formel für die allgemeine Lösung einer homogenen Differenzengleichung, Satz I.6.1 (S. 43), folgt, dass für $\mu \neq 1$ die Lösung von (III.3.10) gegeben ist durch

$$u_i = \frac{(1-\mu)^{-i} - 1}{(1-\mu)^{-N-1} - 1} \qquad 0 \le i \le N+1.$$

Für $\mu > 2$ erhält man also eine stark oszillierende numerische Lösung. Erst für $\mu < 1$ ist die numerische Lösung qualitativ richtig. Man erhält im Prinzip das gleiche Ergebnis, wenn man $\partial_h^+ u$ durch den zentralen Differenzenquotienten $\frac{1}{2}(\partial_h^+ u + \partial_h^- u)$ ersetzt. Ersetzt man dagegen $\partial_h^+ u$ in (III.3.9) durch $\partial_h^- u$, so erhält man statt (III.3.10) das LGS

$$\begin{aligned} -u_{i+1} + (2+\mu)u_i - (1+\mu)u_{i-1} &= 0 \quad 1 \le i \le N \\ u_0 &= 0 \\ u_{N+1} &= 1. \end{aligned} \tag{III.3.11}$$

Dieses hat für alle μ die eindeutige Lösung

$$\widetilde{u}_i = \frac{(1+\mu)^i - 1}{(1+\mu)^{N+1} - 1} \qquad 0 \le i \le N+1.$$

Diese hat für alle Werte von μ das gleiche qualitative Verhalten wie die exakte Lösung des RWP.

Eine genaue Analyse von Beispiel III.3.21 zeigt, dass der wesentliche Unterschied zwischen den verschiedenen Diskretisierungen darin liegt, dass das LGS (III.3.11) eine Koeffizientenmatrix mit strikt positiven Diagonalelementen und nicht positiven Außerdiagonalelementen besitzt. Dies führt auf folgende Diskretisierung von (III.3.8).

DEFINITION III.3.22 (Upwind Diskretisierung). Die *upwind Diskretisierung* von (III.3.8) ist gegeben durch

$$L_h^u u = -\sum_{i=1}^n \sum_{j=1}^n \partial_{h,i}^-(A_{ij}\partial_{h,j}^+ u) + \sum_{i=1}^n a_i \partial_{h,i}^u u + \alpha u. \tag{III.3.12}$$

Dabei ist der i-te *upwind Differenzenquotient* $\partial_{h,i}^u$ gegeben durch

$$\partial_{h,i}^u u(x) = -\operatorname{sgn}(a_i) h^{-1} \left[u(x - \operatorname{sgn}(a_i) h e_i) - u(x) \right].$$

BEMERKUNG III.3.23 (Eigenschaften der upwind Diskretisierung). (1) Sei M die Matrix des LGS

$$L_h^u u_h(x) = f(x) \quad \text{für alle } x \in \Omega_h \tag{III.3.13}$$

mit L_h^u wie in (III.3.12). Dann gilt $M_{\mu\mu} > 0$, $M_{\mu\nu} \leq 0$ für alle $1 \leq \mu, \nu \leq N_h$, $\mu \neq \nu$. Diese Bedingung ist für die eindeutige Lösbarkeit von (III.3.13) und für die Konvergenzanalyse wesentlich.
(2) Man kann zeigen, dass (III.3.13) stets eine eindeutige Lösung besitzt und dass die Fehlerabschätzungen von Satz III.3.13 gilt.
(3) Die Bemerkungen III.3.16 – III.3.18 übertragen sich direkt auf den Operator L und seine Approximation L_h^u.
(4) Für den Nachweis der eindeutigen Lösbarkeit von (III.3.8) ist das folgende *Maximumprinzip* wesentlich: *Gilt* $Lu \leq 0$ *in* Ω*, so besitzt* u *kein positives Maximum in* Ω. Der Operator L_h^u erfüllt das folgende diskrete Analogon: *Gilt* $L_h^u u_h(x) \leq 0$ *für alle* $x \in \Omega_h$*, so besitzt* u_h *kein positives Maximum in* Ω_h. Dieses ist für eine Konvergenzaussage der Maximumsnorm wesentlich. Außerdem impliziert es, dass im Gegensatz zur ersten Diskretisierung aus Beispiel III.3.21 keine Oszillationen der Lösung u_h auftreten können.

III.4. Parabolische Differentialgleichungen

Wir betrachten in diesem Paragraphen die parabolische pDGl

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla \cdot (A \nabla u) + \alpha u &= f \quad \text{in } \Omega \times (0,T] \\ u &= 0 \quad \text{auf } \Gamma \times (0,T] \\ u(\cdot,0) &= u_0 \quad \text{in } \Omega. \end{aligned} \tag{III.4.1}$$

Dabei sind die Bezeichnungen und Voraussetzungen wie in §III.3. Allerdings ist jetzt f von x und t abhängig. $T > 0$ und u_0 sind gegeben. Man beachte, dass dagegen A und α von t unabhängig sind. Der Fall zeitabhängiger Koeffizienten A und α kann mit technischem Mehraufwand genau so behandelt werden wie der hier betrachtete Fall, sofern man $\min\limits_{(x,t)\in\overline{\Omega}\times[0,T]} \alpha(x,t) \geq 0$ und $\min\limits_{(x,t)\in\Omega\times(0,T]} \min\limits_{z\in\mathbb{R}^n\setminus\{0\}} \dfrac{z^T A(x,t) z}{z^T z} = \gamma_0 > 0$ voraussetzt.

Im Rahmen von §III.2 setzen wir

$$\begin{aligned} X &= C^1([0,T], C^2(\overline{\Omega},\mathbb{R}) \cap C_0(\overline{\Omega},\mathbb{R})), \\ Y &= C(\overline{\Omega}\times[0,T],\mathbb{R}) \times C(\overline{\Omega},\mathbb{R}), \\ L\varphi &= \left(\frac{\partial \varphi}{\partial t} - \nabla \cdot (A \nabla \varphi) + \alpha \varphi \ , \ \varphi(\cdot,0) \right). \end{aligned}$$

Zur weiteren Motivation betrachten wir die Differenzendiskretisierung

$$(\mathcal{L}_h u)(x) = -\sum_{i=1}^n \sum_{j=1}^n \partial_{h,i}^- (A_{ij} \partial_{h,j}^+ u)(x) + \alpha(x) u(x) \quad \forall x \in \Omega_h$$

des Differentialoperators $\mathcal{L}u = -\nabla \cdot (A \nabla u) + \alpha u$ aus Definition III.3.5 (S. 99). Wenn wir wie in §III.3 Funktionen auf dem Gitter Ω_h mit Vektoren in $\mathbb{R}^{N_h}$, $N_h = \#\Omega_h$, identifizieren und die Ortsableitungen

durch $\mathcal{L}_h$ approximieren, können wir (III.4.1) durch die folgende gDGl in $\mathbb{R}^{N_h}$ approximieren

$$\dot{u}_h = f_h - \mathcal{L}_h u_h \text{ in } (0,T], \quad u_h(0) = u_{h,0} \tag{III.4.2}$$

mit $f_h(t) = (f(x,t))_{x\in\Omega_h}$ und $u_{h,0} = (u_0(x))_{x\in\Omega_h}$. Als nächstes können wir (III.4.2) mit einem der Verfahren auf Kapitel I näherungsweise lösen und erhalten so eine Approximation an die Lösung von (III.4.1). Für Fehlerabschätzungen könnten wir solche für numerische Verfahren angewandt auf (III.4.2) mit Fehlerabschätzungen für $u - u_h$ kombinieren. Dieses Vorgehen ist nicht effizient, da wir dafür Regularitätsaussagen über u_h als Funktion von t benötigen, über die wir nicht verfügen. Wir werden daher Fehlerabschätzungen mit Hilfe des abstrakten Rahmens aus §III.2 herleiten. Dennoch liefern uns (III.4.2) und die Ergebnisse aus Kapitel I, insb. §I.8 (S. 51), nützliche Informationen über sinnvolle Diskretisierungen der Zeitableitung. So lehrt uns Beispiel I.8.1 (S. 52), dass Kenntnisse über das Spektrum von $\mathcal{L}$ für die Wahl der Zeitdiskretisierung besonders wichtig sind.

Lemma III.4.1 (Inverse Abschätzung, Spektrum von $\mathcal{L}$). *Für alle $v, w : \Omega_h \to \mathbb{R}$ gilt*

$$\begin{aligned} \|v\|_{1,h} &\le 2\sqrt{n} h^{-1} \|v\|_{0,h} \\ (\mathcal{L}_h v\,,\, v)_h &\ge \lambda_0 \|v\|_{1,h}^2 \\ (\mathcal{L}_h v\,,\, w)_h &\le K \|v\|_{1,h} \|w\|_{1,h} \end{aligned}$$

mit λ_0 aus Gleichung (III.3.2) (S. 96), $K = \|A\|_{C^0} + c_\Omega^2 \|\alpha\|_{C^0}$ und c_Ω aus Lemma III.3.10 (S. 101).

Beweis. Für alle $x \in \Omega_h$ und alle $k \in \{1,\dots,n\}$ ist

$$\left|\partial_{h,k}^+ v(x)\right| \le h^{-1} \left\{|v(x)| + |v(x + he_k)|\right\}.$$

Zusammen mit $\|v(\cdot + he_k)\|_{0,h} = \|v\|_{0,h}$ und der Definition von $\|\cdot\|_{1,h}$ folgt hieraus die erste Ungleichung.
Die zweite Ungleichung wurde im ersten Schritt des Beweises von Lemma III.3.12 (S. 102) gezeigt.
Aus Lemma III.3.11 (S. 101) folgt schließlich

$$(\mathcal{L}_h v\,,\, w)_h \le \|A\|_{C(\Omega,\mathbb{R}^{n\times n})} \|v\|_{1,h} \|w\|_{1,h} + \|\alpha\|_{C(\Omega,\mathbb{R})} \|v\|_{0,h} \|w\|_{0,h}.$$

Zusammen mit Lemma III.3.10 (S. 101) beweist dies die dritte Ungleichung. □

Aus Lemma III.4.1 folgt, dass die zu $-\mathcal{L}_h$ gehörige Matrix negativ definit ist und dass ihre Eigenwerte in einem Intervall $[-c_1 h^{-2}, -c_2]$ mit $0 < c_2 \le c_1$ liegen. Das Verfahren zur numerischen Lösung von (III.4.2) sollte also mindestens A_0-stabil sein. Mithin kommen BDF- und implizite Runge-Kutta-Verfahren infrage. Wir beschränken uns hier auf das

θ-Schema aus Beispiel I.5.7 (S. 41). Wegen Beispiel I.5.7 und Definition I.8.2 (S. 54) ist das Stabilitätsgebiet

$$S = \left\{ \mu \in \mathbb{C} : \left| \frac{1+\mu(1-\theta)}{1-\mu\theta} \right| \leq 1 \right\}.$$

Wegen

$$\lim_{\substack{\mu \to -\infty \\ \mu \in \mathbb{R}}} \left| \frac{1+\mu(1-\theta)}{1-\mu\theta} \right| = \frac{1-\theta}{\theta} > 1 \quad \text{für } \theta < \frac{1}{2}$$

kann das θ-Schema für $0 \leq \theta < \frac{1}{2}$ nicht A_0-stabil sein. Mit Hilfe des Riemannschen Abbildungssatzes der Funktionentheorie kann man zeigen, dass es für $\theta \geq \frac{1}{2}$ A-stabil ist. Der Wert $\theta = \frac{1}{2}$ ist von besonderem Interesse, da er mit der Trapezregel auf ein Verfahren zweiter Ordnung führt. Andererseits ist die Wahl $\theta = 0$ auch interessant, da sie auf das explizite Euler-Verfahren führt.

Definition III.4.2 (Differenzendiskretisierung). Seien $h > 0$ und $M \in \mathbb{N}^*$ beliebig, $\tau = \frac{T}{M}$ und $I_\tau = \{k\tau : 0 \leq k \leq M\}$. Dann ist die Differenzendiskretisierung von (III.4.1) gegeben durch

$$\begin{aligned}
X_h &= \{u : \Omega_h \times I_\tau \to \mathbb{R}\}, \\
\|u\|_{X_h} &= \max_{0 \leq k \leq M} \|u(\cdot, k\tau)\|_{0,h}, \\
Y_h &= X_h, \\
\|\cdot\|_{Y_h} &= \|\cdot\|_{X_h}, \\
(L_h u)(x,t) &= \begin{cases} \partial_\tau^- u(x,t) + \theta \mathcal{L}_h u(x,t) & \\ \quad + (1-\theta)\mathcal{L}_h u(x, t-\tau) & \text{für } t > 0, \\ u(x,0) & \text{für } t = 0, \end{cases} \\
(R_{X_h} u)(x,t) &= u(x,t) \quad \text{für alle } (x,t) \in \Omega_h \times I_\tau \\
\left(R_{Y_h} \left(\begin{smallmatrix} f \\ u_0 \end{smallmatrix}\right)\right)(x,t) &= \begin{cases} \theta f(x,t) + (1-\theta) f(x, t-\tau) & \text{für } t > 0, \\ u_0(x) & \text{für } t = 0. \end{cases}
\end{aligned}$$

Dabei ist

$$\begin{aligned}
\partial_\tau^- u(x,t) &= \frac{1}{\tau}[u(x,t) - u(x, t-\tau)], \\
\mathcal{L}_h u(x,t) &= -\sum_{i=1}^n \sum_{j=1}^n \partial_{h,i}^- (A_{ij} \partial_{h,j}^+ u)(x,t) + \alpha(x) u(x,t)
\end{aligned}$$

und $\theta \in [0,1]$.

Bemerkung III.4.3 (Realisierung). Setze $u^k(x) = u(x, k\tau)$ für alle $x \in \Omega_h$ und alle $k \in \{0, \ldots, M\}$. Dann ist $L_h u = R_Y \left(\begin{smallmatrix} f \\ u_0 \end{smallmatrix}\right)$ äquivalent zu $u^0 = u_0$ in Ω_h und für $1 \leq k \leq M$

$$\frac{1}{\tau}[u^k - u^{k-1}] + \theta \mathcal{L}_h u^k + (1-\theta)\mathcal{L}_h u^{k-1} = \theta f^k + (1-\theta) f^{k-1}$$

in Ω_h. Daher kann u^k für $k = 1, \ldots, M$ sukzessiv aus u^{k-1} durch Lösen des LGS

$$\left[\frac{1}{\tau} I + \theta \mathcal{L}_h\right] u^k = g^k \tag{III.4.3}$$

mit $g^k = \theta f^k + (1-\theta) f^{k-1} - (1-\theta)\mathcal{L}_h u^{k-1} + \frac{1}{\tau} u^{k-1}$ berechnet werden. Für $\theta = 0$ ist das Lösen von (III.4.3) trivial. Aus Lemma III.4.1 folgt, dass die Matrix auf der linken Seite von (III.4.3) für alle $\theta \in [0,1]$ symmetrisch positiv definit ist. Für $\theta > 0$ entspricht (III.4.3) der Diskretisierung aus §III.3 der elliptischen pDGl

$$\begin{aligned} -\nabla \cdot (A \nabla v) + \left(\alpha + \frac{1}{\theta\tau}\right) v &= \widetilde{g} \quad \text{in } \Omega \\ v &= 0 \quad \text{auf } \Gamma \end{aligned}$$

mit geeigneter Funktion $\widetilde{g}$.

Für die Konsistenz- und Konvergenzabschätzung benötigen wir:

Definition III.4.4 (Hölder-Räume). Für $\alpha, \beta \in \mathbb{N}$ sei

$$\begin{aligned} C^{\alpha,\beta}(\overline{\Omega} \times [0,T], \mathbb{R}) = \big\{ & \varphi \in C(\overline{\Omega} \times [0,T], \mathbb{R}) : \\ & \partial_t^\mu \varphi \in C(\overline{\Omega} \times [0,T], \mathbb{R}), 0 \le \mu \le \beta, \\ & D_x^\gamma \varphi \in C(\overline{\Omega} \times [0,T], \mathbb{R}), 0 \le |\gamma| \le \alpha \big\} \end{aligned}$$

mit $D_x^\gamma \varphi = \frac{\partial^{|\gamma|}\varphi}{\partial x_1^{\gamma_1} \ldots \partial x_n^{\gamma_n}}$ und $|\gamma| = \gamma_1 + \ldots + \gamma_n$ sowie

$$\|\varphi\|_{C^{\alpha,\beta}} = \max\left\{ \max_{0 \le \mu \le \beta} \|\partial_t^\mu \varphi\|_{C(\overline{\Omega}\times[0,T],\mathbb{R})}, \max_{0 \le |\gamma| \le \alpha} \|D_x^\gamma \varphi\|_{C(\overline{\Omega}\times[0,T],\mathbb{R})} \right\}.$$

Lemma III.4.5 (Konsistenz). *Das Diskretisierungsverfahren aus Definition III.4.2 ist konsistent. Für $u \in C^{3+\rho,2+\sigma}(\overline{\Omega} \times [0,T], \mathbb{R})$ gilt*

$$\|L_h R_{X_h} u - R_{Y_h} L u\|_{Y_h} \le c \left[\tau^{1+\sigma} + h^{1+\rho}\right] \|u\|_{C^{3+\rho,2+\sigma}} \max\{1, \|A\|_{C^{2+\rho}}\}$$

mit

$$\rho = \begin{cases} 1 & \text{falls } A \text{ konstante Diagonalmatrix,} \\ 0 & \text{sonst,} \end{cases} \qquad \sigma = \begin{cases} 1 & \text{falls } \theta = \frac{1}{2}, \\ 0 & \text{sonst.} \end{cases}$$

Beweis. Für $(x,t) \in \Omega_h \times \{I_\tau \setminus \{0\}\}$ gilt

$$\begin{aligned} &(L_h R_{X_h} u)(x,t) - (R_{Y_h} L u)(x,t) \\ &\quad = \frac{1}{\tau}[u(x,t) - u(x,t-\tau)] - \theta \partial_t u(x,t) - (1-\theta)\partial_t u(x,t-\tau) \\ &\qquad + \theta\left[\mathcal{L}_h(x,t) + \nabla \cdot (A \nabla u)(x,t) - \alpha(x) u(x,t)\right] \\ &\qquad + (1-\theta)\left[\mathcal{L}_h u(x,t-\tau) + \nabla \cdot (A \nabla u)(x,t-\tau) - \alpha(x) u(x,t-\tau)\right] \end{aligned}$$

und

$$\frac{1}{\tau}[u(x,t)-u(x,t-\tau)]-\theta\partial_t u(x,t)-(1-\theta)\partial_t u(x,t-\tau)$$
$$=\frac{1}{2}(1-2\theta)\tau\partial_t^2 u(x,t)+\frac{1}{6}(-2+3\theta)\tau^2\partial_t^3 u(x,t)+\mathrm{O}(\tau^3).$$

Hieraus und aus Lemma III.3.9 (S. 100) folgt die Behauptung. □

LEMMA III.4.6 (Stabilität). *Es sei $\theta \in \{0\} \cup [\frac{1}{2}, 1]$. Für $\theta = 0$ gelte zusätzlich die* Courant-Friedrichs-Levy-Bedingung *(kurz* CFL-Bedingung*) $4nK\tau \le h^2$. Dann ist das Diskretisierungsverfahren aus Definition III.4.2 stabil.*

BEWEIS. Sei $u \in X_h$ beliebig und $v = L_h u$. Dann ist definitionsgemäß

$$\|u(\cdot,0)\|_{0,h} = \|L_h u(\cdot,0)\|_{0,h} \le \|L_h u\|_{Y_h}.$$

Daher reicht es zu zeigen, dass für alle $1 \le m \le M$ gilt

$$\|u(\cdot,m\tau)\|_{0,h} \le c\,\|L_h u\|_{Y_h} \tag{III.4.4}$$

mit einer von u, h und τ unabhängigen Konstanten c.
Setze zur Abkürzung für alle $0 \le k \le M$

$$u^k = u(\cdot,k\tau), \quad v^k = v(\cdot,k\tau).$$

Wir betrachten zunächst den Fall $\theta = 0$. Aus $v = L_h u$ folgt für alle $1 \le k \le M$

$$\frac{1}{\tau}[u^k - u^{k-1}] = v^{k-1} - \mathcal{L}_h u^{k-1} \quad \text{in } \Omega_h.$$

Bildet man das $(\cdot\,,\cdot)_h$-Skalarprodukt dieser Gleichung mit u^k, ergibt sich

$$\frac{1}{\tau}\left\{\|u^k\|_{0,h}^2 - \left(u^{k-1}\,,\,u^k\right)_h\right\} = \left(v^{k-1}\,,\,u^k\right)_h - \left(\mathcal{L}_h u^{k-1}\,,\,u^k\right)_h. \tag{III.4.5}$$

Für die linke Seite von (III.4.5) folgt

$$\frac{1}{\tau}\left\{\|u^k\|_{0,h}^2 - \left(u^{k-1}\,,\,u^k\right)_h\right\}$$
$$= \frac{1}{2\tau}\left\{\|u^k\|_{0,h}^2 - \|u^{k-1}\|_{0,h}^2 + \|u^{k-1}-u^k\|_{0,h}^2\right\}.$$

Wegen

$$2\left(\mathcal{L}_h u^{k-1}\,,\,u^k\right)_h = \left(\mathcal{L}_h u^{k-1}\,,\,u^{k-1}\right)_h + \left(\mathcal{L}_h u^k\,,\,u^k\right)_h$$
$$- \left(\mathcal{L}_h(u^k-u^{k-1})\,,\,u^k-u^{k-1}\right)_h$$

erhalten wir mit der Cauchy-Schwarschen Ungleichung für die rechte Seite von (III.4.5)

$$\begin{aligned}&\left(v^{k-1}\,,\,u^k\right)_h-\left(\mathcal{L}_hu^{k-1}\,,\,u^k\right)_h\\&\quad\le\left\|v^{k-1}\right\|_{0,h}\left\|u^k\right\|_{0,h}-\frac12\left(\mathcal{L}_hu^{k-1}\,,\,u^{k-1}\right)_h\\&\qquad-\frac12\left(\mathcal{L}_hu^k\,,\,u^k\right)_h+\frac12\left(\mathcal{L}_h(u^k-u^{k-1})\,,\,u^k-u^{k-1}\right)_h.\end{aligned}$$

Wegen

$$\left\|v^{k-1}\right\|_{0,h}\left\|u^k\right\|_{0,h}\le\frac{c_\Omega^2}{2\lambda_0}\left\|v^{k-1}\right\|_{0,h}^2+\frac{\lambda_0}{2c_\Omega^2}\left\|u^k\right\|_{0,h}^2$$

folgt hieraus mit Lemma III.4.1

$$\begin{aligned}&\left(v^{k-1}\,,\,u^k\right)_h-\left(\mathcal{L}_hu^{k-1}\,,\,u^k\right)_h\\&\quad\le\frac{c_\Omega^2}{2\lambda_0}\left\|v^{k-1}\right\|_{0,h}^2+\frac{\lambda_0}{2c_\Omega^2}\left\|u^k\right\|_{0,h}^2-\frac{\lambda_0}{2}\left\|u^k\right\|_{1,h}^2+\frac{K}{2}\left\|u^k-u^{k-1}\right\|_{1,h}^2\\&\quad\le\frac{c_\Omega^2}{2\lambda_0}\left\|v^{k-1}\right\|_{0,h}^2+\frac12 4nKh^{-2}\left\|u^k-u^{k-1}\right\|_{0,h}^2.\end{aligned}$$

Wegen der CFL-Bedingung $\tau Kh^{-2}4n\le1$ folgt aus den Abschätzungen für (III.4.5)

$$\left\|u^k\right\|_{0,h}^2-\left\|u^{k-1}\right\|_{0,h}^2\le\frac{c_\Omega^2}{\lambda_0}\tau\left\|v^{k-1}\right\|_{0,h}^2.$$

Summation von $k=1$ bis m liefert

$$\begin{aligned}\left\|u^m\right\|_{0,h}^2&\le\left\|u^0\right\|_{0,h}+\frac{c_\Omega^2}{\lambda_0}\tau\sum_{k=1}^m\left\|v^{k-1}\right\|_{0,h}^2\\&\le\left\|u^0\right\|_{0,h}^2+\frac{c_\Omega^2}{\lambda_0}m\tau\max_{0\le l\le m-1}\left\|v^k\right\|_{0,h}^2\\&\le\left(1+\frac{c_\Omega^2}{\lambda_0}T\right)\|v\|_{Y_h}^2.\end{aligned}$$

Dies beweist (III.4.4) mit $c=\left[1+\frac{c_\Omega^2}{\lambda_0}T\right]^{\frac12}$.
Betrachte nun den Fall $\theta\in[\frac12,1]$. Aus $v=L_hu$ folgt für alle $1\le k\le M$

$$\frac1\tau[u^k-u^{k-1}]=\theta v^k+(1-\theta)v^{k-1}-\theta\mathcal{L}_hu^k-(1-\theta)\mathcal{L}_hu^{k-1}\quad\text{in }\Omega_h.$$

Bildet man das $(\cdot\,,\,\cdot)_h$-Skalarprodukt dieser Gleichung mit $w=\theta u^k+(1-\theta)u^{k-1}$, ergibt sich

$$\text{(III.4.6)}\quad\frac1\tau\left(u^k-u^{k-1}\,,\,w\right)_h=\left(\theta v^k+(1-\theta)v^{k-1}\,,\,w\right)_h-\left(\mathcal{L}_hw\,,\,w\right)_h.$$

Für die linke Seite von (III.4.6) erhalten wir

$$\begin{aligned}
&\frac{1}{\tau}\left(u^k - u^{k-1}\,,\, w\right)_h \\
&\quad= \frac{1}{\tau}\left\{\theta\left\|u^k\right\|_{0,h}^2 - (1-\theta)\left\|u^{k-1}\right\|_{0,h}^2 + (1-2\theta)\left(u^k\,,\, u^{k-1}\right)_h\right\} \\
&\quad= \frac{1}{\tau}\left\{\theta\left\|u^k\right\|_{0,h}^2 - (1-\theta)\left\|u^{k-1}\right\|_{0,h}^2\right. \\
&\qquad\qquad \left. +\frac{1}{2}(2\theta-1)\left[\left\|u^k - u^{k-1}\right\|_{0,h}^2 - \left\|u^k\right\|_{0,h}^2 - \left\|u^{k-1}\right\|_{0,h}^2\right]\right\} \\
&\quad= \frac{1}{2\tau}\left\{\left\|u^k\right\|_{0,h}^2 - \left\|u^{k-1}\right\|_{0,h}^2 + (2\theta-1)\left\|u^k - u^{k-1}\right\|_{0,h}^2\right\} \\
&\quad\geq \frac{1}{2\tau}\left\{\left\|u^k\right\|_{0,h}^2 - \left\|u^{k-1}\right\|_{0,h}^2\right\}.
\end{aligned}$$

Für die rechte Seite von (III.4.6) folgt mit Lemma III.4.1

$$\begin{aligned}
&\left(\theta v^k + (1-\theta)v^{k-1}\,,\, w\right)_h - \left(\mathcal{L}_h w\,,\, w\right)_h \\
&\quad\leq \left\|\theta v^k + (1-\theta)v^{k-1}\right\|_{0,h}\|w\|_{0,h} - \lambda_0\|w\|_{1,h}^2 \\
&\quad\leq \frac{c_\Omega^2}{4\lambda_0}\left\|\theta v^k + (1-\theta)v^{k-1}\right\|_{0,h}^2 + \frac{\lambda_0}{c_\Omega^2}\|w\|_{0,h}^2 - \lambda_0\|w\|_{1,h}^2 \\
&\quad\leq \frac{c_\Omega^2}{4\lambda_0}\max_{l=k-1,k}\left\|v^l\right\|_{0,h}^2 \\
&\quad\leq \frac{c_\Omega^2}{4\lambda_0}\|v\|_{Y_h}^2.
\end{aligned}$$

Kombinieren der letzten beiden Abschätzungen und Summation von $k=1$ bis $k=m$ liefert wieder

$$\|u^m\|_{0,h}^2 \leq \left\|u^0\right\|_{0,h}^2 + \frac{c_\Omega^2}{2\lambda_0}\tau m\|v\|_{Y_h}^2 \leq \left(1 + \frac{c_\Omega^2}{2\lambda_0}T\right)\|v\|_{Y_h}^2. \qquad \square$$

Satz III.4.7 (Fehlerabschätzung). *Seien u die Lösung von (III.4.1) und u_h die mit dem Diskretisierungsverfahren aus Definition III.4.2 berechnete Näherungslösung. Die Voraussetzungen von Lemma III.4.6 seien erfüllt. Dann gilt*

$$\begin{aligned}
&\max_{0\leq k\leq M}\|u(\cdot,k\tau) - u_h(\cdot,k\tau)\|_{0,h} \\
&\qquad\leq c\left[\tau^{1+\sigma} + h^{1+\rho}\right]\|u\|_{C^{3+\rho,2+\sigma}}\max\left\{1, \|A\|_{C^{2+\rho}}\right\}
\end{aligned}$$

mit ρ und σ wie in Lemma III.4.5.

Beweis. Folgt aus Satz III.2.4 (S. 93) und Lemmata III.4.5 und III.4.6. $\square$

Bemerkung III.4.8. (1) Bei gegebenem h ist die optimale Schrittweite τ in Satz III.4.7 gegeben durch $h^{\frac{1+\rho}{1+\sigma}}$. Für $\theta = 0$ und allgemeines A ist diese wesentlich größer als die durch die CFL-Bedingung von

Lemma III.4.6 vorgeschriebene Schrittweite. Die Trapezregel $\theta = \frac{1}{2}$ ist allen anderen Verfahren vorzuziehen, weil es die größte optimale Zeitschrittweite τ erlaubt.
(2) Die Bemerkungen von §III.3 zur Behandlung anderer Randbedingungen und von Termen der Form $a \cdot \nabla u$ übertragen sich sinngemäß auf die entsprechenden parabolischen Probleme.

Beispiel III.4.9 (Eindimensionale Wärmeleitungsgleichung). Wir betrachten die Differenzendiskretisierung mit $\tau = 0.2$ und $h = 0.02$ der eindimensionalen Wärmeleitungsgleichung (III.4.1) mit $\Omega = (0, 1)$, $T = 1$, $A = 1$, $\alpha = 0$, $f = 0$, $u_0 = \sin(\pi x)$ und exakter Lösung $u = e^{-\pi^2 t} \sin(\pi x)$. Abbildung III.4.1 zeigt die diskreten Lösungen zu den Zeiten 0 (rot), 0.2 (grün), 0.4 (blau), 0.6 (gelb), 0.8 (türkis), 1 (lila) und den Werten $\theta = 0$ (links), $\theta = 0.5$ (Mitte), $\theta = 1$ (rechts). Sie belegt die mangelnde Stabilität des expliziten Euler-Verfahrens $\theta = 0$ und die höhere Genauigkeit der Trapezregel $\theta = 0.5$.

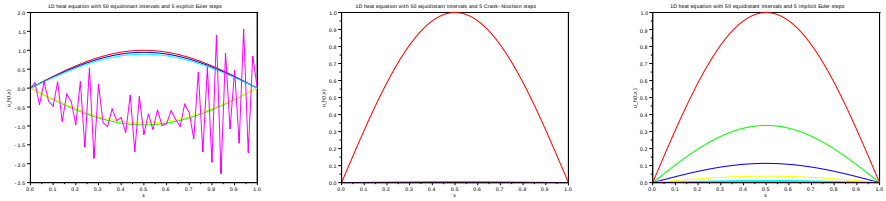

Abbildung III.4.1. Diskrete Lösungen aus Beispiel III.4.9 zu den Zeiten 0 (rot), 0.2 (grün), 0.4 (blau), 0.6 (gelb), 0.8 (türkis), 1 (lila) und den Werten $\theta = 0$ (links), $\theta = 0.5$ (Mitte), $\theta = 1$ (rechts)

III.5. Hyperbolische Differentialgleichungen

In diesem Paragraphen betrachten wir die hyperbolische pDGl

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - \nabla \cdot (A \nabla u) + \alpha u &= f \quad \text{in } Q = \Omega \times (0, T) \\ u &= 0 \quad \text{auf } S = \Gamma \times (0, T) \\ u(\cdot, 0) &= u_0 \quad \text{in } \Omega \\ \frac{\partial}{\partial t} u(\cdot, 0) &= u_1 \quad \text{in } \Omega. \end{aligned} \tag{III.5.1}$$

Dabei benutzen wir die gleichen Bezeichnungen und Voraussetzungen wie in den §§III.3 und III.4. Ähnlich wie in §III.4 könnte man mit technischem Mehraufwand den Fall zeitabhängiger Koeffizienten A und α behandeln. Wir beschränken uns hier aber auf den zeitlich konstanten Fall.

Im Rahmen von §III.2 setzen wir

$$\begin{aligned} X &= C^2(\overline{Q},\mathbb{R}) \cap C([0,T],C_0(\overline{\Omega},\mathbb{R})),\\ Y &= C(\overline{Q},\mathbb{R}) \times C(\overline{\Omega},\mathbb{R}) \times C(\overline{\Omega},\mathbb{R})\\ L\varphi &= \left(\frac{\partial^2\varphi}{\partial t^2} - \nabla\cdot(A\nabla\varphi) + \alpha\varphi\ ,\ \frac{\partial\varphi}{\partial t}(\cdot,0)\ ,\ \varphi(\cdot,0)\right). \end{aligned}$$

Den Ortsteil des Differentialoperators diskretisieren wir wie in §III.3. Die Zeitableitungen behandeln wir gleichberechtigt, d.h., wir diskretisieren sie durch symmetrische Differenzenquotienten in Zeitrichtung.

DEFINITION III.5.1 (Differenzendiskretisierung). Seien $h > 0$ und $M \in \mathbb{N}^*$ beliebig, $\tau = \frac{T}{M}$, $I_\tau = \{k\tau : 0 \le k \le M\}$ und $Q_{h,\tau} = \Omega_h \times I_\tau$. Dann ist die Differenzendiskretisierung von (III.5.1) gegeben durch

$$\begin{aligned} X_h &= \{u : Q_{h,\tau} \to \mathbb{R}\},\\ \|u\|_{X_h} &= \Bigg\{ \|u(\cdot,0)\|_{1,h}^2 \\ &\qquad + \tau \sum_{k=1}^{M} \left[\|u(\cdot,k\tau)\|_{1,h}^2 + \left\|\partial_\tau^- u(,k\tau)\right\|_{0,h}^2 \right] \Bigg\}^{\frac{1}{2}},\\ Y_h &= \{f : Q_{h,\tau} \to \mathbb{R}\} \times \{u_1 : \Omega_h \to \mathbb{R}\}\\ &\qquad \times \{u_0 : \Omega_h \to \mathbb{R}\},\\ \|(f,u_1,u_0)\|_{Y_h} &= \left\{ \tau \sum_{k=1}^{M} \|f(\cdot,k\tau)\|_{0,h}^2 + \|u_1\|_{0,h}^2 + \|u_0\|_{1,h}^2 \right\}^{\frac{1}{2}},\\ (L_h u)(x,t) &= \begin{cases} \partial_\tau^+\partial_\tau^- u(x,t-\tau) + \mathcal{L}_h u(x,t-\tau) & 2\tau \le t \le T,\\ \partial_\tau^- u(x,t) & t = \tau,\\ u(x,t) & t = 0, \end{cases}\\ (R_{X_h} u)(x,t) &= u(x,t) \quad (x,t) \in Q_{h,\tau}\\ \left(R_{Y_h}\begin{pmatrix} f\\ u_1\\ u_0 \end{pmatrix}\right)(x,t) &= \begin{cases} f(x,t-\tau) & (x,t) \in Q_{h,\tau}, t \ge \tau,\\ u_1(x,0) & (x,t) \in Q_{h,\tau}, t = 0,\\ u_0(x,0) & (x,t) \in Q_{h,\tau}, t = 0. \end{cases} \end{aligned}$$

Dabei ist

$$\begin{aligned} \partial_\tau^\pm u(x,t) &= \pm\frac{1}{\tau}[u(x,t\pm\tau) - u(x,t)],\\ \mathcal{L}_h u(x,t) &= -\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{h,i}^-(A_{ij}\partial_{h,j}^+ u)(x,t) + \alpha(x)u(x,t). \end{aligned}$$

BEMERKUNG III.5.2 (Realisierung). Mit $u^k(x) = u(x, k\tau)$ für alle $x \in \Omega_h$ und $0 \le k \le M$ ist

$$L_h u = \varphi = \begin{pmatrix} f \\ u_1 \\ u_0 \end{pmatrix} \tag{III.5.2}$$

äquivalent zu

$$\begin{aligned} u^0(x) &= u_0(x) \\ u^1(x) &= u^0(x) + \tau u_1(x) \\ u^{k+1}(x) &= 2u^k(x) - u^{k-1}(x) - \tau^2 \mathcal{L}_h u^k(x) + \tau^2 f^k(x). \end{aligned} \tag{III.5.3}$$

Die Lösung von (III.5.2) kann also durch das sukzessive Auswerten der Gleichungen von (III.5.3) berechnet werden. Das Verfahren aus Definition III.5.1 ist also wie das Verfahren zu $\theta = 0$ aus §III.4 explizit. Wir erwarten daher, dass die Stabilität des Verfahrens nur dann gewährleistet ist, wenn eine CFL Bedingung erfüllt ist.

LEMMA III.5.3 (Konsistenz). *Das Diskretisierungsverfahren aus Definition III.5.1 ist konsistent. Für $u \in C^3(\overline{Q}, \mathbb{R})$ gilt*

$$\|R_{Y_h} L u - L_h R_{X_h} u\|_{Y_h} \le c[\tau + h] \, \|u\|_{C^3} ,$$

wobei c von $\|A\|_{C^2}$ abhängt.

BEWEIS. Folgt aus Lemmata III.3.4 (S. 98) und III.3.9 (S. 100). □

LEMMA III.5.4 (Stabilität). *Es gelte die CFL-Bedingung*

$$\tau \le \min\left\{ \frac{\lambda_0}{4}, \frac{\lambda_0}{K 2\sqrt{n}} h \right\},$$

wobei λ_0 und K wie in Lemma III.4.1 (S. 110) sind. Dann ist das Diskretisierungsverfahren aus Definition III.5.1 stabil.

BEWEIS. Sei $u \in X_h$ beliebig und $L_h u = (v, w, z)$. Dann gilt

$$\begin{aligned} \|u(\cdot, 0)\|_{1,h} = \|z\|_{1,h} &\le \|L_h u\|_{Y_h} \\ \left\|\partial_\tau^- u(\cdot, \tau)\right\|_{0,h} = \|w\|_{0,h} &\le \|L_h u\|_{Y_h} \end{aligned}$$

und für $1 \le k \le M - 1$

$$\partial_\tau^+ \partial_\tau^- u^k + \mathcal{L}_h u^k = v^k \quad \text{in } \Omega_h.$$

Bilden wir das $(\cdot\,,\,\cdot)_h$ Skalarprodukt dieser Gleichung mit $\partial_\tau^+ u^k + \partial_\tau^- u^k$ und beachten, dass $\partial_\tau^- u^{k+1} = \partial_\tau^+ u^k$ ist, erhalten wir

$$\begin{aligned} &\left(v^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h \\ &\quad = \left(\partial_\tau^+ \partial_\tau^- u^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h + \left(\mathcal{L}_h u^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h \\ &\quad = \frac{1}{\tau}\left(\partial_\tau^+ u^k - \partial_\tau^- u^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h + \left(\mathcal{L}_h u^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h \\ &\quad = \frac{1}{\tau}\left\|\partial_\tau^+ u^k\right\|_{0,h}^2 - \frac{1}{\tau}\left\|\partial_\tau^- u^k\right\|_{0,h}^2 + \left(\mathcal{L}_h u^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h \\ &\quad = \frac{1}{\tau}\left\|\partial_\tau^- u^{k+1}\right\|_{0,h}^2 - \frac{1}{\tau}\left\|\partial_\tau^- u^k\right\|_{0,h}^2 + \left(\mathcal{L}_h u^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h . \end{aligned}$$

Sei $2 \le p \le M-1$. Dann liefert Summation von $k=1$ bis $p-1$

$$\begin{aligned}
&\tau \sum_{k=1}^{p-1} \left(v^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h \\
&\quad = \left\|\partial_\tau^- u^p\right\|_{0,h}^2 - \left\|\partial_\tau^- u^1\right\|_{0,h}^2 + \sum_{k=1}^{p-1} \left(\mathcal{L}_h u^k \,,\, u^{k+1} - u^{k-1}\right)_h \\
&\quad = \left\|\partial_\tau^- u^p\right\|_{0,h}^2 - \left\|\partial_\tau^- u^1\right\|_{0,h}^2 \\
&\qquad + \sum_{k=1}^{p-1} \left(\mathcal{L}_h u^k \,,\, u^{k+1}\right)_h - \sum_{k=1}^{p-1} \left(\mathcal{L}_h u^k \,,\, u^{k-1}\right)_h .
\end{aligned}$$

Wegen

$$\left(\mathcal{L}_h u^k \,,\, u^{k-1}\right)_h = \left(\mathcal{L}_h u^{k-1} \,,\, u^k\right)_h$$

erhalten wir

$$\begin{aligned}
&\tau \sum_{k=1}^{p-1} \left(v^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h \\
&\quad = \left\|\partial_\tau^- u^p\right\|_{0,h}^2 - \left\|\partial_\tau^- u^1\right\|_{0,h}^2 + \left(\mathcal{L}_h u^{p-1} \,,\, u^p\right)_h - \left(\mathcal{L}_h u^1 \,,\, u^0\right)_h \\
&\quad = \left\|\partial_\tau^- u^p\right\|_{0,h}^2 - \left\|\partial_\tau^- u^1\right\|_{0,h}^2 + \left(\mathcal{L}_h u^p \,,\, u^p\right)_h - \left(\mathcal{L}_h u^0 \,,\, u^0\right)_h \\
&\qquad - \left(\mathcal{L}_h (u^p - u^{p-1}) \,,\, u^p\right)_h - \left(\mathcal{L}_h (u^1 - u^0) \,,\, u^0\right)_h .
\end{aligned}$$

Wegen $\partial_\tau^- u^1 = \partial_\tau^+ u^0$ folgt hieraus mit Lemma III.4.1 (S. 110)

$$\begin{aligned}
&\left\|\partial_\tau^- u^p\right\|_{0,h}^2 + \lambda_0 \left\|u^p\right\|_{1,h}^2 \\
&\quad \le \left\|\partial_\tau^- u^p\right\|_{0,h} + \left(\mathcal{L}_h u^p \,,\, u^p\right)_h \\
&\quad = \left\|\partial_\tau^+ u^0\right\|_{0,h}^2 + \left(\mathcal{L}_h u^0 \,,\, u^0\right)_h \\
&\qquad + \left(\mathcal{L}_h (u^p - u^{p-1}) \,,\, u^p\right)_h + \left(\mathcal{L}_h (u^1 - u^0) \,,\, u^0\right)_h \\
&\qquad + \tau \sum_{k=1}^{p-1} \left(v^k \,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h .
\end{aligned}$$

Wegen Lemma III.4.1 ist

$$\left(\mathcal{L}_h u^0 \,,\, u^0\right)_h \le K \left\|u^0\right\|_{1,h}^2 .$$

Wiederum wegen Lemma III.4.1 und $K 2\sqrt{n} h^{-1} \tau \le \lambda_0$ ist

$$\begin{aligned}
\left(\mathcal{L}_h (u^p - u^{p-1}) \,,\, u^p\right)_h &\le K \left\|u^p - u^{p-1}\right\|_{1,h} \left\|u^p\right\|_{1,h} \\
&\le K 2\sqrt{n} h^{-1} \tau \left\|\partial_\tau^- u^p\right\|_{0,h} \left\|u^p\right\|_{1,h} \\
&\le \lambda_0 \left\|\partial_\tau^- u^p\right\|_{0,h} \left\|u^p\right\|_{1,h} \\
&\le \frac{\lambda_0}{2} \left\|\partial_\tau^- u^p\right\|_{0,h}^2 + \frac{\lambda_0}{2} \left\|u^p\right\|_{1,h}^2 .
\end{aligned}$$

Wegen $\partial_\tau^- u^1 = \partial_\tau^+ u^0$ folgt mit dem gleichen Argument

$$\left(\mathcal{L}_h(u^1 - u^0)\,,\, u^0\right)_h \leq \frac{\lambda_0}{2}\left\|\partial_\tau^+ u^0\right\|_{0,h}^2 + \frac{\lambda_0}{2}\left\|u^0\right\|_{1,h}^2.$$

Aus der Cauchy-Schwarzschen Ungleichung, der Dreiecksungleichung und der Abschätzung $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ folgt schließlich

$$\begin{aligned}\left(v^k\,,\, \partial_\tau^+ u^k + \partial_\tau^- u^k\right)_h &\leq \left\|v^k\right\|_{0,h}\left[\left\|\partial_\tau^+ u^k\right\|_{0,h} + \left\|\partial_\tau^- u^k\right\|_{0,h}\right] \\ &\leq \left\|v^k\right\|_{0,h}^2 + \frac{1}{2}\left\|\partial_\tau^+ u^k\right\|_{0,h}^2 + \frac{1}{2}\left\|\partial_\tau^- u^k\right\|_{0,h}^2.\end{aligned}$$

Insgesamt erhalten wir somit

$$\begin{aligned}&\left\|\partial_\tau^- u^p\right\|_{0,h}^2 + \lambda_0\left\|u^p\right\|_{1,h}^2 \\ &\quad\leq \left(1 + \frac{\lambda_0}{2}\right)\left\|\partial_\tau^+ u^0\right\|_{0,h}^2 + \left(K + \frac{\lambda_0}{2}\right)\left\|u^0\right\|_{1,h}^2 \\ &\qquad + \frac{\lambda_0}{2}\left\|\partial_\tau^- u^p\right\|_{0,h}^2 + \frac{\lambda_0}{2}\left\|u^p\right\|_{1,h}^2 \\ &\qquad + \tau\sum_{k=1}^{p}\left\|v^k\right\|_{0,h}^2 + \tau\sum_{k=1}^{p}\left\|\partial_\tau^- u^k\right\|_{0,h}^2.\end{aligned}$$

Da wir o.E. $0 < \lambda_0 \leq 1 \leq K$ annehmen können, folgt hieraus

$$\begin{aligned}&\frac{\lambda_0}{2}\left\{\left\|\partial_\tau^- u^p\right\|_{0,h}^2 + \left\|u^p\right\|_{1,h}^2\right\} \\ &\quad\leq \frac{3}{2}K\left\{\left\|\partial_\tau^+ u^0\right\|_{0,h}^2 + \left\|u^0\right\|_{1,h}^2\right\} + \tau\sum_{k=1}^{p}\left\|v^k\right\|_{0,h}^2 + \tau\sum_{k=1}^{p}\left\|\partial_\tau^- u^k\right\|_{0,h}^2.\end{aligned}$$

Definiere

$$\begin{aligned}\alpha_p &= \tau\sum_{k=1}^{p}\left\{\left\|\partial_\tau^- u^k\right\|_{0,h}^2 + \left\|u^k\right\|_{1,h}^2\right\}, \\ \beta_p &= \frac{3}{2}K\left\{\left\|\partial_\tau^+ u^0\right\|_{0,h}^2 + \left\|u^0\right\|_{1,h}^2\right\} + \tau\sum_{k=1}^{p}\left\|v^k\right\|_{0,h}^2.\end{aligned}$$

Dann haben wir gezeigt

$$\frac{\alpha_p - \alpha_{p-1}}{\tau} \leq \frac{2}{\lambda_0}\alpha_p + \frac{2}{\lambda_0}\beta_p.$$

Wegen $\tau \leq \frac{\lambda_0}{4}$ folgt hieraus

$$\alpha_p \leq \frac{1}{1 - \frac{2}{\lambda_0}\tau}\alpha_{p-1} + \frac{\frac{2}{\lambda_0}\tau}{1 - \frac{2}{\lambda_0}\tau}\beta_p$$

für alle $2 \le p \le M-1$. Mittels Induktion liefert dies die Abschätzung

$$\begin{aligned}
\alpha_p &\le \left(1+\frac{\frac{2}{\lambda_0}\tau}{1-\frac{2}{\lambda_0}\tau}\right)^{p-1}\left[\alpha_1+\frac{2}{\lambda_0}\tau\sum_{k=2}^{p}\beta_k\left(1+\frac{\frac{2}{\lambda_0}\tau}{1-\frac{2}{\lambda_0}\tau}\right)^{2-k}\right]\\
&\le \left(1+\frac{4}{\lambda_0}\tau\right)^{p-1}\left[\alpha_1+\frac{2}{\lambda_0}\tau p\beta_p\right]\\
&\le e^{\frac{4}{\lambda_0}T}\left\{\tau\left\|\partial_\tau^- u^1\right\|_{0,h}^2+\left\|u^1\right\|_{1,h}^2\right.\\
&\qquad\left.+\frac{2}{\lambda_0}T\left[\frac{3}{2}K\left\{\left\|\partial_\tau^+ u^0\right\|_{0,h}^2+\left\|u^0\right\|_{0,h}^2\right\}+\tau\sum_{k=1}^{p}\left\|v^k\right\|_{0,h}^2\right]\right\}.
\end{aligned}$$

Wegen

$$\begin{aligned}
\left\|\partial_\tau^- u^1\right\|_{0,h} &= \left\|\partial_\tau^+ u^0\right\|_{0,h} = \|w\|_{0,h}\\
\left\|u^0\right\|_{1,h} &= \|z\|_{1,h}\\
\left\|u^1\right\|_{1,h} &\le \tau\left\|\partial_\tau^- u^1\right\|_{1,h}+\left\|u^0\right\|_{1,h}\\
&\le 2\sqrt{n}\tau h^{-1}\left\|\partial_\tau^- u^1\right\|_{0,h}+\left\|u^0\right\|_{1,h}\\
&\le \frac{\lambda_0}{K}\|w\|_{0,h}+\|z\|_{1,h}
\end{aligned}$$

folgt hieraus

$$\|u\|_{X_h}^2 \le c(n,\lambda_0,K)e^{\frac{4}{\lambda_0}T}\max\{1,T\}\|(v,w,z)\|_{Y_h}^2. \qquad \square$$

Satz III.5.5 (Fehlerabschätzung). *Sei u die Lösung von (III.5.1) und u_h die durch das Diskretisierungsverfahren von Definition III.5.1 gelieferte Näherungslösung. Es gelte die CFL-Bedingung von Lemma III.5.4 und $u \in C^3(\overline{Q},\mathbb{R})$. Dann gilt die Fehlerabschätzung*

$$\|u-u_h\|_{X_h} \le c[\tau+h]\|u\|_{C^3(\overline{Q},\mathbb{R})}$$

mit $c = c(n,\lambda_0,K)\|A\|_{C^2(\Omega,\mathbb{R}^{n\times n})}e^{\frac{4}{\lambda_0}T}\max\{1,T\}$.

Beweis. Folgt aus Satz III.2.4 (S. 93) und Lemmata III.5.3 und III.5.4. $\square$

Bemerkung III.5.6. (1) Die CFL-Bedingung $\tau \le ch$ aus Lemma III.5.4 ist wesentlich schwächer als die CFL-Bedingung $\tau \le ch^2$ aus Lemma III.4.6 (S. 113) für den Fall $\theta = 0$. Wegen der Fehlerabschätzung ist zudem die Wahl $\tau \approx h$ optimal.
(2) Falls A eine konstante Diagonalmatrix und $u \in C^4(\overline{Q},\mathbb{R})$ ist, erhält man sogar die Abschätzung

$$\|u-u_h\|_{X_h} \le c[\tau^2+h^2]\|u\|_{C^4(\overline{Q},\mathbb{R})}.$$

(3) Andere Randbedingungen und Terme der Form $a\cdot\nabla u$ können wie in §III.3 behandelt werden.

III.6. Numerische Lösung der diskreten Probleme

Die Ergebnisse der §§III.4 und III.5 zeigen, dass man bei der numerischen Lösung parabolischer und hyperbolischer pDGl mit impliziten Zeitschrittverfahren in jedem Zeitschritt ein diskretes Problem lösen muss, das einer Differenzendiskretisierung einer elliptischen pDGl entspricht. Da die Zahl der Zeitschritte groß ist, muss dieser Lösungsprozess sehr effizient sein.

Wir betrachten daher in diesem Paragraphen numerische Verfahren zur Lösung der Differenzengleichung

$$L_h u_h = f_h \quad \text{in } \Omega_h \tag{III.6.1a}$$

mit

$$(L_h u_h)(x) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \partial^-_{h,i}(A_{ij}\partial^+_{h,j}u_h)(x) + \alpha(x)u_h(x) \tag{III.6.1b}$$

aus §III.3. Wir benutzen die gleichen Bezeichnungen und Voraussetzungen wie dort.

Offensichtlich ist (III.6.1) ein LGS mit $N_h = \#\Omega_h$ Gleichungen und Unbekannten. Da Ω beschränkt ist, gilt $N_h = \mathrm{O}(h^{-n})$. Wegen der Voraussetzungen an A und α ist die Matrix $\mathcal{A}$ des LGS (III.6.1) symmetrisch positiv definit. Sei $x_1, \ldots, x_{N_h}$ irgendeine Abzählung der Gitterpunkte in Ω_h. Dann folgt aus (III.6.1) $\mathcal{A}_{ij} = 0$ falls $\|x_i - x_j\|_\infty > h$. Dabei bezeichnet $\|\cdot\|_\infty$ die Maximum-Norm auf $\mathbb{R}^n$. Die Matrix $\mathcal{A}$ ist also *dünn besetzt*. Die Zahl der von Null verschiedenen Elemente ist $e_h = \mathrm{O}(3^n h^{-n})$. Dies ist auch die Zahl der arithmetischen Operationen für eine Matrix-Vektor Multiplikation. Der Anteil der von Null verschiedenen Elemente an der Gesamtzahl der Elemente von $\mathcal{A}$ ist $p_h = \mathrm{O}((3h)^n)$ und nimmt somit für $h \to 0$ ab. Sei $b_h = \max\{|i-j| : \mathcal{A}_{ij} \neq 0\}$ die maximale Bandbreite von $\mathcal{A}$. Mit graphentheoretischen Methoden kann man zeigen, dass bei geeigneter Aufzählung der Gitterpunkte gilt $b_h = \mathrm{O}(h^{-n+1})$ und dass diese Größenordnung optimal ist. Daher benötigt ein direktes Lösungsverfahren für (III.6.1) wie z.B. Gauß-Elimination, LR-Zerlegung oder Cholesky-Zerlegung $s_h = N_h b_h = \mathrm{O}(h^{-2n+1})$ Speicherplätze und $z_h = N_h b_h^2 = \mathrm{O}(h^{-3n+2})$ arithmetische Operationen. Der Gesamtaufwand verhält sich also wie $N_h^{3-\frac{2}{n}}$ und wächst somit superlinear mit der Zahl der Unbekannten. Daher sind direkte Lösungsverfahren für das LGS (III.6.1) sowohl vom Speicher- als auch vom Rechenbedarf nicht effizient.

Dies wird eindrücklich durch Tabelle III.6.1 belegt. Ein Giga-Flop-Rechner würde z.B. zur Lösung der Poisson-Gleichung auf dem Einheitswürfel mit der Schrittweite $h = \frac{1}{128}$ mehr als 6 Tage und einen Speicherplatz von 33 Gigabyte erfordern. Man beachte, dass in diesem

TABELLE III.6.1. Speicherbedarf und Rechenaufwand für die Cholesky-Zerlegung des LGS (III.6.1) mit $A = I$, $\alpha = 0$, $\Omega = [0,1]^n$

n	h	N_h	e_h	b_h	s_h	z_h
2	$\frac{1}{16}$	225	$1.1 \cdot 10^3$	15	$3.3 \cdot 10^3$	$7.6 \cdot 10^5$
	$\frac{1}{32}$	961	$4.8 \cdot 10^3$	31	$2.9 \cdot 10^4$	$2.8 \cdot 10^7$
	$\frac{1}{64}$	$3.9 \cdot 10^3$	$2.0 \cdot 10^4$	63	$2.5 \cdot 10^5$	$9.9 \cdot 10^8$
	$\frac{1}{128}$	$1.6 \cdot 10^4$	$8.0 \cdot 10^4$	127	$2.0 \cdot 10^6$	$3.3 \cdot 10^{10}$
3	$\frac{1}{16}$	$3.3 \cdot 10^3$	$2.4 \cdot 10^4$	225	$7.6 \cdot 10^5$	$1.7 \cdot 10^8$
	$\frac{1}{32}$	$3.0 \cdot 10^4$	$2.1 \cdot 10^5$	961	$2.8 \cdot 10^7$	$2.8 \cdot 10^{10}$
	$\frac{1}{64}$	$2.5 \cdot 10^5$	$1.8 \cdot 10^6$	$3.9 \cdot 10^3$	$9.9 \cdot 10^8$	$3.9 \cdot 10^{12}$
	$\frac{1}{128}$	$2.0 \cdot 10^6$	$1.4 \cdot 10^7$	$1.6 \cdot 10^4$	$3.3 \cdot 10^{10}$	$5.3 \cdot 10^{14}$

Beispiel zur Speicherung eines Vektors „nur“ 2 Megabyte und zur Speicherung der von Null verschiedenen Matrixelemente „nur“ 14 Megabyte erforderlich sind.

Wegen dieser Überlegungen benutzt man in der Praxis fast ausschließlich iterative Verfahren zur Lösung von (III.6.1). Deren Effizienz wird wesentlich beeinflusst durch folgende Überlegungen:

- Die exakte Lösung von (III.6.1) approximiert die Lösung der entsprechenden pDGl, an der wir eigentlich interessiert sind, mit einem Fehler von $O(h)$ oder $O(h^2)$. Daher reicht es vollkommen aus, das LGS (III.6.1) auch nur mit einem entsprechenden Fehler zu lösen.
- Ist $\widetilde{u}_{h_1}$ eine Näherungslösung von (III.6.1) zur Schrittweite h_1, die obigem Kriterium genügt, so ist eine geeignete Interpolierende von $\widetilde{u}_{h_1}$ eine gute Startnäherung für jeden iterativen Lösungsalgorithmus zu LGS (III.6.1) mit einer Schrittweite $h_2 < h_1$, sofern $\frac{h_2}{h_1}$ nicht zu klein ist.

Diese Überlegungen liegen Algorithmus III.6.1 und Satz III.6.1 zugrunde. Dabei betrachten wir eine Folge $h_0 > h_2 > \ldots > h_R$ kleiner werdender Schrittweiten und zugehörige Differenzenapproximationen

$$L_{h_k} u_{h_k} = f_{h_k}, \quad 0 \le k \le R.$$

Zur Abkürzung ersetzen wir überall den Index h_k durch k.

Algorithmus III.6.1 Geschachtelte Iteration

Gegeben: Rechte Seiten $f_0, \dots, f_R$.
Gesucht: Näherungslösungen $\widetilde{u}_0, \dots, \widetilde{u}_R$ für die Systeme $L_k u_k = f_k$, $k = 1, \dots, R$

1: $\widetilde{u}_0 \leftarrow u_0 = L_0^{-1} f_0$
2: **for** $k = 0, \dots, R$ **do**
3: Berechne eine Näherungslösung $\widetilde{u}_k$ für $u_k = L_k^{-1} f_k$ durch Anwenden von m_k Iterationen eines iterativen Lösungsverfahrens auf das entsprechende LGS (III.6.1) zu h_k mit Startwert $I_{k-1,k}\widetilde{u}_{k-1}$, wobei $I_{k-1,k} : X_{k-1} \to X_k$ ein gegebener Interpolationsoperator ist.
4: **end for**

Satz III.6.1 (Fehlerabschätzung für die geschachtelte Iteration). *Bezeichne mit δ_k die Konvergenzrate des im k-ten Schritt von Algorithmus III.6.1 benutzten iterativen Lösungsverfahrens. Die folgenden Voraussetzungen seien erfüllt:*

(1) $\|u_k - R_k u\| \le c_0(u) h_k^\alpha$ *für alle* $0 \le k \le R$, *wobei* $c_0(u)$ *nur von* u *abhängt und* $\alpha > 0$ *von* k *und* u *unabhängig ist.*
(2) $\|I_{k-1,k}\|_{\mathcal{L}(X_{k-1}, X_k)} \le c_1$ *für alle* $1 \le k \le R$, *wobei* c_1 *nicht von* k *abhängt.*
(3) $\|I_{k-1,k} R_{k-1} u - R_k u\|_{X_k} \le c_2 c_0(u) h_k^\alpha$ *für alle* $1 \le k \le R$, *wobei* c_2 *nicht von* k *und* u *abhängt.*
(4) $\frac{h_{k-1}}{h_k} \le c_3$ *für alle* $1 \le k \le R$.
(5) $\delta_k^{m_k} \le [1 + c_2 + 2c_1 c_3^\alpha]^{-1}$.

Dann gilt für alle $0 \le k \le R$

$$\|\widetilde{u}_k - R_k u\|_{X_k} \le 2 c_0(u) h_k^\alpha,$$

d.h., $\widetilde{u}_k$ ist eine ebenso gute Approximation an u wie u_k.

Beweis. Für $k = 0$ folgt die Behauptung aus (1). Sei also $k \ge 1$. Dann folgt aus der Definition von δ_k und (1)

$$\begin{aligned}\|\widetilde{u}_k - R_k u\|_{X_k} &\le \|\widetilde{u}_k - u_k\|_{X_k} + \|u_k - R_k u\|_{X_k} \\ &\le \delta_k^{m_k} \|I_{k-1,k}\widetilde{u}_{k-1} - u_k\|_{X_k} + c_0(u) h_k^\alpha.\end{aligned}$$

Wegen (2) – (4) ist

$$\begin{aligned}\|I_{k-1,k}\widetilde{u}_{k-1} - u_k\|_{X_k} &\le \|I_{k-1,k}(\widetilde{u}_{k-1} - R_{k-1}u)\|_{X_k} \\ &\quad + \|I_{k-1,k} R_{k-1} u - R_k u\|_{X_k} + \|R_k u - u_k\|_{X_k} \\ &\le 2c_1 c_0(u) h_{k-1}^\alpha + c_2 c_0(u) h_k^\alpha + c_0(u) h_k^\alpha \\ &\le c_0(u) h_k^\alpha \left\{2c_1 c_3^\alpha + c_2 + 1\right\}.\end{aligned}$$

Damit folgt aus (5)

$$\|\widetilde{u}_k - R_k u\|_{X_k} \le c_0(u) h_k^\alpha \left\{1 + \delta_k^{m_k} \left[2c_1 c_3^\alpha + c_2 + 1\right]\right\} \le 2 c_0(u) h_k^\alpha. \quad \square$$

BEMERKUNG III.6.2 (Anwendungen). (1) Im k-ten Schritt von Algorithmus III.6.1 muss der Anfangsfehler nur um einen von k unabhängigen Faktor reduziert werden. Falls δ_k von k unabhängig ist, kann dies mit einem zu N_{h_k} proportionalen Aufwand geschehen. Der Gesamtaufwand von Algorithmus III.6.1 ist dann zu N_{h_R} proportional. (2) Für (III.6.1) ist $\alpha = 1$ oder $\alpha = 2$, $c_0(u) = \|u\|_{C^{2+\alpha}}$, $c_3 = 2$, $I_{k-1,k}$ die lineare Interpolierende zwischen Ω_{k-1} und Ω_k, $c_1 = 1$, $c_2 = 1$ und daher

$$1 + c_2 + 2c_1 c_3^\alpha = \begin{cases} 6 & \text{für } \alpha = 1, \\ 10 & \text{für } \alpha = 2. \end{cases}$$

Es reicht also, den Anfangsfehler um einen Faktor 10 zu reduzieren.

Um die Frage beantworten zu können, welches geeignete Iterationsverfahren zur Lösung des LGS (III.6.1) sind, müssen wir zunächst den minimalen und maximalen Eigenwert der Matrix $\mathcal{A}$ abschätzen.

LEMMA III.6.3 (Spektrum von $\mathcal{A}$). *Seien $\lambda_{\min}$ und $\lambda_{\max}$ der minimale bzw. maximale Eigenwert der Matrix $\mathcal{A}$ des LGS (III.6.1). Dann gibt es zwei Konstanten c_1 und c_2, die nur von Ω, A und α abhängen, mit $c_1 \le \lambda_{\min} \le \lambda_{\max} \le c_2 h^{-2}$. Diese Schranken sind scharf.*

BEWEIS. Wir identifizieren Vektoren im $\mathbb{R}^{N_h}$ mit den entsprechenden Gitterfunktionen auf Ω_h. Da $\mathcal{A}$ symmetrisch ist, besitzt es einen vollständigen Satz von Eigenvektoren. Diese können bzgl. des Skalarproduktes $(\cdot\,,\,\cdot)_h$ auf $\mathbb{R}^{N_h}$ orthonormiert werden. Da wir o.E. die Eigenwerte der Größe nach anordnen können, gilt somit

$$L_h u_k = \lambda_k u_k, \quad (u_k\,,\,u_\ell)_h = \delta_{k\ell}, \quad \lambda_1 \le \ldots \le \lambda_{N_h}.$$

Aus Lemmata III.3.10 (S. 101) und III.3.11 (S. 101) und dem Beweis von Lemma III.3.12 (S. 102) folgt für $1 \le k \le N_h$

$$\begin{aligned} \lambda_k &= \lambda_k\,(u_k\,,\,u_k)_h = (L_h u_k\,,\,u_k)_h \ge \lambda_0 \|u_k\|_{1,h}^2 \ge \lambda_0 c_\Omega^{-2} \|u_k\|_{0,h}^2 \\ &= \lambda_0 c_\Omega^{-2} \end{aligned}$$

und

$$\begin{aligned} \lambda_k &= (L_h u_k\,,\,u_k)_h \\ &\le \|A\|_{C(\Omega,\mathbb{R}^{n\times n})} \sum_{i=1}^n \sum_{j=1}^n \left\|\partial_{h,j}^+ u_k\right\|_{0,h} \left\|\partial_{h,i}^+ u_k\right\|_{0,h} + \|\alpha\|_{C(\Omega,\mathbb{R})} \|u_k\|_{0,h}^2 \\ &\le \left\{ n^2 \|A\|_{C(\Omega,\mathbb{R}^{n\times n})} + \|\alpha\|_{C(\Omega,\mathbb{R})}\, c_\Omega^2 \right\} \|u_k\|_{1,h}^2 . \end{aligned}$$

Gemäß Lemma III.4.1 (S. 110) gilt aber

$$\|u_k\|_{1,h} \le h^{-1} 2\sqrt{n}\, \|u_k\|_{0,h} = h^{-1} 2\sqrt{n}. \qquad \square$$

Aus Lemma III.6.3 folgt, dass die Richardson-Iteration den Fehler gemessen in der $\|\cdot\|_{0,h}$-Norm pro Iteration um den Faktor $1 - \frac{\lambda_{\min}}{\lambda_{\max}} = 1 - ch^2$ reduziert. Mit etwas Mehraufwand kann man zeigen, dass

Gleiches mit anderer Konstante c für das Jacobi und Gauß-Seidel-Verfahren gilt. Die genannten Verfahren benötigen daher einen Aufwand von $\mathrm{O}(h^{-2})$ Iterationen und $\mathrm{O}(h^{-2}N_h) = \mathrm{O}(h^{-n-2})$ Operationen, um den Anfangsfehler um einen festen Faktor zu reduzieren. Sie sind somit nicht geeignet, um in Schritt 2 von Algorithmus III.6.1 angewandt zu werden. Besser geeignet sind das konjugierte Gradienten- und das vorkonditionierte konjugierte Gradienten-Verfahren [9, Algorithmen IV.7.10, IV.7.14]. Wegen ihrer Bedeutung geben wir diese Verfahren hier nochmals an.

Algorithmus III.6.2 Vorkonditioniertes konjugiertes Gradienten-Verfahren, PCG-Verfahren

Gegeben: Matrix, $\mathcal{A}$, rechte Seite b, Startnäherung x, Toleranz ε, Vorkonditionierungsmatrix $\mathcal{C}$, Maximalzahl für Iterationen N
Gesucht: Näherungslösung x mit $\|\mathcal{A}x - b\| \leq \varepsilon$

1: $r \leftarrow b - \mathcal{A}x$, $z \leftarrow \mathcal{C}^{-1}r$, $d \leftarrow z$, $\gamma \leftarrow (r\,,\, z)$, $n \leftarrow 0$
2: **while** $\gamma > \varepsilon^2$ und $n \leq N$ **do**
3: $\quad s \leftarrow \mathcal{A}d$, $\alpha \leftarrow \dfrac{\gamma}{(d\,,\, s)}$, $x \leftarrow x + \alpha d$, $r \leftarrow r - \alpha s$
4: $\quad z \leftarrow \mathcal{C}^{-1}r$, $\beta \leftarrow \frac{(r\,,\, z)}{\gamma}$, $\gamma \leftarrow (r\,,\, z)$, $d \leftarrow z + \beta d$, $n \leftarrow n + 1$
5: **end while**

Bemerkung III.6.4. (1) $(\cdot\,,\,\cdot)$ bezeichnet in Algorithmus III.6.2 das euklidische Skalarprodukt. Der Algorithmus ist invariant unter Skalierungen des Skalarproduktes.
(2) $\mathcal{C}$ beschreibt die Vorkonditionierung. $\mathcal{C} = I$ entspricht dem konjugierten Gradienten-Verfahren.

Satz III.6.5 (Konvergenzrate des PCG-Verfahrens). *Sei e_i der Fehler der i-ten Iterierten von Algorithmus III.6.2 und $\|x\|_{\mathcal{A}} = (\mathcal{A}x\,,\, x)^{\frac{1}{2}}$. Dann gilt*

$$\|e_i\|_{\mathcal{A}} \leq 2\left[\frac{\sqrt{\kappa(\mathcal{C}^{-1}\mathcal{A})} - 1}{\sqrt{\kappa(\mathcal{C}^{-1}\mathcal{A})} + 1}\right]^i \|e_0\|_{\mathcal{A}}\,,$$

wobei $\kappa(\mathcal{C}^{-1}\mathcal{A})$ die Kondition von $\mathcal{C}^{-1}\mathcal{A}$ bzgl. der euklidischen Norm ist.

Beweis. [9, Satz IV.7.12] □

Bemerkung III.6.6 (Anwendung auf (III.6.1)). (1) Bei Anwendung auf das LGS (III.6.1) ist $\|\cdot\|_{\mathcal{A}}$ äquivalent zu $\|\cdot\|_{1,h}$, wobei die entsprechenden Konstanten nicht von h abhängen.
(2) Aus Lemma III.6.3 folgt $\kappa(\mathcal{A}) = \frac{c_2}{c_1}h^{-2}$. Algorithmus III.6.2 mit $\mathcal{C} = I$ reduziert also den Fehler gemessen in der $\|\cdot\|_{1,h}$-Norm pro Iteration um den Faktor $1 - ch$.

Algorithmus III.6.3 SSOR-Vorkonditionierung

Gegeben: Matrix $\mathcal{A}$, Vektor r, Relaxationsparameter $\omega \in (0,2)$
Gesucht: $z = \mathcal{C}^{-1}r$

1: $z \leftarrow 0$
2: **for** $i = 1, \dots, n$ **do**
3: $\quad z_i \leftarrow z_i + \dfrac{\omega}{\mathcal{A}_{ii}} \left\{ r_i - \sum_{j=1}^{n} \mathcal{A}_{ij} z_j \right\}$
4: **end for**
5: **for** $i = n, n-1, \dots, 1$ **do**
6: $\quad z_i \leftarrow z_i + \dfrac{\omega}{\mathcal{A}_{ii}} \left\{ r_i - \sum_{j=1}^{n} \mathcal{A}_{ij} z_j \right\}$
7: **end for**

Algorithmus III.6.3 liefert für das LGS (III.6.1) eine ordentliche Vorkonditionierung.

SATZ III.6.7 (Darstellung von $\mathcal{C}^{-1}r$, Kondition von $\mathcal{C}^{-1}\mathcal{A}$). *(1) Die Matrix $\mathcal{A}$ sei symmetrisch, positiv definit und besitze eine Zerlegung der Form $\mathcal{A} = \mathcal{D} - \mathcal{L} - \mathcal{L}^T$ mit einer Diagonalmatrix $\mathcal{D}$ und einer strikten unteren Dreiecksmatrix $\mathcal{L}$. Dann gilt für das Ergebnis z von Algorithmus III.6.3 $z = \mathcal{C}^{-1}r$ mit $\mathcal{C} = \frac{1}{\omega(2-\omega)}(\mathcal{D} - \omega\mathcal{L})\mathcal{D}^{-1}(\mathcal{D} - \omega\mathcal{L}^T)$. (2) Zusätzlich gebe es zwei Konstanten $0 < \gamma \le \Gamma$ mit*

$$\gamma\,(x\,,\,\mathcal{D}x) \le (x\,,\,\mathcal{A}x)\,, \quad \left(x\,,\,\Delta\mathcal{D}^{-1}\Delta^T x\right) \le \frac{1}{4}\Gamma\,(x\,,\,\mathcal{A}x)$$

für alle $x \in \mathbb{R}^{N_h}$, wobei $\Delta = \frac{1}{2}\mathcal{D} - \mathcal{L}$ ist. Dann gilt mit $\mu = \frac{2-\omega}{2\omega}$

$$\left[\frac{\mu}{2\gamma} + \frac{1}{2} + \frac{\Gamma}{8\mu}\right]^{-1} \le \lambda_{\min}(\mathcal{C}^{-1}\mathcal{A}) \le \lambda_{\max}(\mathcal{C}^{-1}\mathcal{A}) \le 2.$$

Insbesondere ist $\kappa(\mathcal{C}^{-1}\mathcal{A})$ optimal für $\omega_{op} = \frac{2}{1+\sqrt{\gamma\Gamma}}$. Für $\omega = \omega_{op}$ gilt

$$\kappa(\mathcal{C}^{-1}\mathcal{A}) = 1 + \sqrt{\frac{\Gamma}{\gamma}}.$$

BEWEIS. ad (1): Bezeichne mit $\widetilde{z}$ das Ergebnis von Schritt 2 von Algorithmus III.6.3. Dann folgt wegen Schritt 1

$$\widetilde{z} = \omega\mathcal{D}^{-1}\{r + \mathcal{L}\widetilde{z}\}$$

und

$$\begin{aligned} z &= \widetilde{z} + \omega\mathcal{D}^{-1}\{r - \mathcal{D}\widetilde{z} + \mathcal{L}\widetilde{z} + \mathcal{L}^T z\} \\ &= \widetilde{z} + \omega\mathcal{D}^{-1}\{r + \mathcal{L}\widetilde{z}\} - \omega\widetilde{z} + \omega\mathcal{D}^{-1}\mathcal{L}^T z \\ &= (2-\omega)\widetilde{z} + \omega\mathcal{D}^{-1}\mathcal{L}^T z. \end{aligned}$$

Hieraus folgt

$$\begin{aligned} z &= (I - \omega\mathcal{D}^{-1}\mathcal{L}^T)^{-1}(2-\omega)\widetilde{z} \\ &= (I - \omega\mathcal{D}^{-1}\mathcal{L}^T)^{-1}(2-\omega)(I - \omega\mathcal{D}^{-1}\mathcal{L})^{-1}\omega\mathcal{D}^{-1}r \\ &= (\mathcal{D} - \omega\mathcal{L}^T)^{-1}\omega(2-\omega)\mathcal{D}(\mathcal{D} - \omega\mathcal{L})^{-1}r. \end{aligned}$$

ad (2): Aus Teil (1) und der Definition von μ und Δ folgt

$$\begin{aligned} \mathcal{C} &= (\mathcal{D} - \omega\mathcal{L})\frac{1}{\omega(2-\omega)}\mathcal{D}^{-1}(\mathcal{D} - \omega\mathcal{L}^T) \\ &= \left(\frac{1}{\omega}\mathcal{D} - \mathcal{L}\right)\left[(2-\omega)\mathcal{D}\right]^{-1}\omega\left(\frac{1}{\omega}\mathcal{D} - \mathcal{L}^T\right) \\ &= \left(\frac{1}{\omega}\mathcal{D} - \mathcal{L}\right)\left[\left(\frac{2}{\omega} - 1\right)\mathcal{D}\right]^{-1}\left(\frac{1}{\omega}\mathcal{D} - \mathcal{L}^T\right) \\ &= (\mu\mathcal{D} + \Delta)[2\mu\mathcal{D}]^{-1}(\mu\mathcal{D} + \Delta^T) \\ &= (\mu\mathcal{D} + \Delta)\left(\frac{1}{2}I + \frac{1}{2\mu}\mathcal{D}^{-1}\Delta^T\right) \\ &= \frac{\mu}{2}\mathcal{D} + \frac{1}{2}\Delta^T + \frac{1}{2}\Delta + \frac{1}{2\mu}\Delta\mathcal{D}^{-1}\Delta^T \\ &= \frac{\mu}{2}\mathcal{D} + \frac{1}{2}\mathcal{A} + \frac{1}{2\mu}\Delta\mathcal{D}^{-1}\Delta^T. \end{aligned}$$

Da $\mathcal{D}$ und $\Delta\mathcal{D}^{-1}\Delta^T$ positiv definit sind, folgt hieraus für alle $x \in \mathbb{R}^{N_h}$

$$(x\,,\,\mathcal{C}x) \geq \frac{1}{2}\,(x\,,\,\mathcal{A}x)$$

oder äquivalent

$$\sup_{x\in\mathbb{R}^{N_h}\setminus\{0\}} \frac{(x\,,\,\mathcal{A}x)}{(x\,,\,\mathcal{C}x)} \leq 2.$$

Die linke Seite dieser Ungleichung ist aber gerade der maximale Eigenwert von $\mathcal{C}^{-1}\mathcal{A}$.
Aus den Voraussetzungen an $\mathcal{D}$ und Δ folgt weiter für alle $x \in \mathbb{R}^{N_h}$

$$(x\,,\,\mathcal{C}x) \leq \left[\frac{\mu}{2\gamma} + \frac{1}{2} + \frac{\Gamma}{8\mu}\right](x\,,\,\mathcal{A}x)$$

oder äquivalent

$$\inf_{x\in\mathbb{R}^{N_h}\setminus\{0\}} \frac{(x\,,\,\mathcal{A}x)}{(x\,,\,\mathcal{C}x)} \geq \left[\frac{\mu}{2\gamma} + \frac{1}{2} + \frac{\Gamma}{8\mu}\right]^{-1}.$$

Die linke Seite dieser Ungleichung ist aber gerade der kleinste Eigenwert von $\mathcal{C}^{-1}\mathcal{A}$.
Dies beweist die behauptete Abschätzung der Eigenwerte von $\mathcal{C}^{-1}\mathcal{A}$. Insbesondere gilt stets

$$\kappa(\mathcal{C}^{-1}\mathcal{A}) \leq 2\left[\frac{\mu}{2\gamma} + \frac{1}{2} + \frac{\Gamma}{8\mu}\right].$$

Definiere $\varphi \in C^1(\mathbb{R}_+^*, \mathbb{R}_+^*)$ durch $\varphi(x) = \frac{x}{\gamma} + 1 + \frac{\Gamma}{4x}$. Offensichtlich gilt

$$\lim_{x\to 0+} \varphi(x) = +\infty, \quad \lim_{x\to+\infty} \varphi(x) = +\infty$$

und

$$\varphi'(x) = \frac{1}{\gamma} - \frac{\Gamma}{4x^2} = 0 \quad \Longleftrightarrow \quad x = \frac{1}{2}\sqrt{\Gamma\gamma}.$$

Also ist

$$\inf_{x\in\mathbb{R}_+^*} \varphi(x) = \varphi\left(\frac{1}{2}\sqrt{\Gamma\gamma}\right) = 1 + \sqrt{\frac{\Gamma}{\gamma}}.$$

Wegen

$$\frac{1}{\omega} - \frac{1}{2} = \mu = \frac{1}{2}\sqrt{\Gamma\gamma} \quad \Longleftrightarrow \quad \omega = \frac{2}{1+\sqrt{\Gamma\gamma}}$$

folgt hieraus die Aussage über die optimale Wahl von ω und den entsprechenden Wert von $\kappa(\mathcal{C}^{-1}\mathcal{A})$. □

LEMMA III.6.8 (Anwendung auf (III.6.1)). *Die Matrix $\mathcal{A}$ des LGS (III.6.1) erfüllt die Voraussetzungen von Satz III.6.7 mit $\gamma = c_1 h^2$ und $\Gamma = c_2$, wobei c_1 und c_2 nicht von h abhängen.*

BEWEIS. Sei x der k-te Gitterpunkt in Ω_h. Dann folgt

$$\mathcal{D}_{kk} = h^{-2} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} A_{ij}(x) + A_{ii}(x - he_i) \right\} + \alpha(x).$$

Wenn wir wieder Vektoren des $\mathbb{R}^{N_h}$ mit entsprechenden Gitterfunktionen identifizieren, folgt hieraus

$$\begin{aligned}(u\,,\,\mathcal{D}u) &= h^{-n}\,(u\,,\,\mathcal{D}u)_h \\ &\le \left\{ n(n+1)\,\|A\|_{C(\Omega,\mathbb{R}^{n\times n})} + \|\alpha\|_{C(\Omega,\mathbb{R})} \right\} h^{-2} h^{-n}\,\|u\|_{0,h}^2\end{aligned}$$

und

$$(u\,,\,\mathcal{D}u) = h^{-n}\,(u\,,\,\mathcal{D}u)_h \ge \lambda_0 h^{-n}\,\|u\|_{0,h}^2\, h^{-2}.$$

Andererseits folgt aus dem Beweis von Lemma III.3.12 (S. 102)

$$(u\,,\,\mathcal{A}u) = h^{-n}\,(u\,,\,L_h u)_h \ge \lambda_0 h^{-n}\,\|u\|_{1,h}^2 \ge \lambda_0 c_\Omega^{-2} h^{-n}\,\|u\|_{0,h}^2\,.$$

Also ist für alle $u \in \mathbb{R}^{N_h}$

$$(u\,,\,\mathcal{A}u) \ge \left\{ n(n+1)\,\|A\|_C + \|\alpha\|_C \right\}^{-1} h^2 \lambda_0 c_\Omega^{-2}\,(u\,,\,\mathcal{D}u)\,.$$

Dies beweist die Aussage über γ.
Zum Beweis der Aussage über Γ sei $v = \Delta^T u$. Dann folgt

$$\begin{aligned}\left(v\,,\,\mathcal{D}^{-1}v\right) &= h^{-n}\left\|\mathcal{D}^{-\frac{1}{2}}v\right\|_{0,h}^2 \le \lambda_0^{-1}h^2\left(\mathcal{D}^{-\frac{1}{2}}v\,,\,\mathcal{D}\mathcal{D}^{-\frac{1}{2}}v\right) \\ &= \lambda_0^{-1}h^2\,(v\,,\,v) = \lambda_0^{-1}h^2 h^{-n}\,\|v\|_{0,h}^2\,.\end{aligned}$$

Ähnlich wie in Beweis von Lemma III.3.12 (S. 102) folgt

$$\|v\|_{0,h} \le \left\{ \|A\|_{C(\Omega,\mathbb{R}^{n\times n})} + \|\alpha\|_{C(\Omega,\mathbb{R})} \right\}^{\frac{1}{2}} \|u\|_{1,h}\,.$$

Zusammen mit Lemma III.4.1 (S. 110) liefert dies

$$\begin{aligned}\left(u\,,\,\Delta\mathcal{D}^{-1}\Delta^T u\right) &\le \lambda_0^{-1}h^2h^{-n}c(A,\alpha)\,\|u\|_{1,h}^2 \le \lambda_0^{-1}h^{-n}c\,\|u\|_{0,h}^2\\ &\le \lambda_0^{-2}c_\Omega^2 c\,(u\,,\,\mathcal{A}u)\,.\end{aligned}$$ □

Aus Satz III.6.5, Satz III.6.7 und Lemma III.6.8 folgt zusammenfassend:

Satz III.6.9 (Konvergenzraten des CG- und PCG-Vefahrens für Problem (III.6.1)). *Das CG-Verfahren angewandt auf das LGS (III.6.1) reduziert den Fehler, gemessen in der $\|\cdot\|_{1,h}$-Norm, pro Iteration um einen Faktor $1-c_1h$. Das PCG-Verfahren mit SSOR-Vorkonditionierung reduziert den Fehler, gemessen in der $\|\cdot\|_{1,h}$-Norm, pro Iteration um einen Faktor $1-c_2h^{\frac{1}{2}}$. Dabei hängen c_1 und c_2 nur von A, α und Ω ab.*

In Tabelle III.6.2 stellen wir die Zahl der Iterationen zusammen, die das Gauß-Seidel-, das CG- und das PCG-Verfahren mit SSOR Vorkonditionierung benötigen, um für Problem (III.6.1) mit $A=I$, $\alpha=0$, $\Omega=[0,1]^n$ den Fehler um den Faktor 10 zu reduzieren.

Tabelle III.6.2. Zahl der Iterationen, die verschiedene Verfahren benötigen, um bei Problem (III.6.1) mit $A=I$, $\alpha=0$, $\Omega=[0,1]^n$ den Fehler um den Faktor 10 zu reduzieren

h	Gauß-Seidel	CG	SSOR-PCG
$\frac{1}{16}$	236	12	4
$\frac{1}{32}$	954	23	5
$\frac{1}{64}$	3820	47	7
$\frac{1}{128}$	15287	94	11

Pro Iteration und Gitterpunkt benötigen das Gauß-Seidel-, CG- und SSOR-PCG-Verfahren $2n+1$, $2n+6$, bzw. $5n+8$ Operationen.

Tabelle III.6.3 stellt die Zahl der Operationen zusammen, die benötigt werden um den Fehler um den Faktor 10 zu reduzieren. Bei Verwendung der Verfahren innerhalb Algorithmus III.6.1 sind diese Werte mit einem Faktor 2 zu multiplizieren, um eine Approximation mit der Genauigkeit des Diskretisierungsfehlers zu erhalten. Zum Vergleich sind die Operationen aufgeführt, die das Cholesky-Verfahren benötigt. Die Tabelle zeigt deutlich die Überlegenheit der iterativen Verfahren, besonders in Verbindung mit Algorithmus III.6.1.

Das PCG-Verfahren ist 3 – 4 Zehnerpotenzen schneller als das Cholesky-Verfahren oder andere direkte Verfahren. Zudem benötigt es wesentlich weniger Speicherplatz.

TABELLE III.6.3. Zahl der Operationen, die verschiedene Verfahren benötigen, um bei Problem (III.6.1) mit $A = I$, $\alpha = 0$, $\Omega = [0,1]^n$ den Fehler um den Faktor 10 zu reduzieren

n	h	Cholesky	Gauß-Seidel	CG	SSOR-PCG
2	$\frac{1}{16}$	759′375	265′500	27′000	16′200
	$\frac{1}{32}$	$2.8 \cdot 10^7$	4′583′970	221′030	86′490
	$\frac{1}{64}$	$9.9 \cdot 10^8$	$7.6 \cdot 10^7$	1′865′430	500′094
	$\frac{1}{128}$	$3.3 \cdot 10^{10}$	$1.2 \cdot 10^9$	$1.5 \cdot 10^7$	3′193′542
3	$\frac{1}{16}$	$1.7 \cdot 10^8$	5′575′500	486′000	310′500
	$\frac{1}{32}$	$2.8 \cdot 10^{10}$	$2.0 \cdot 10^8$	8′222′316	3′425′965
	$\frac{1}{64}$	$3.9 \cdot 10^{12}$	$6.7 \cdot 10^9$	$1.4 \cdot 10^8$	$4.0 \cdot 10^7$
	$\frac{1}{128}$	$5.3 \cdot 10^{14}$	$2.2 \cdot 10^{11}$	$2.3 \cdot 10^9$	$5.2 \cdot 10^8$

Offensichtlich liefert das PCG-Verfahren mit SSOR-Vorkonditionierung gute Ergebnisse. Dennoch wächst sein Aufwand wie $O(N_h^{1+\frac{1}{2n}})$. Wir wollen nun ein Verfahren angeben, dessen Aufwand immer proportional zu N_h ist. Zur Motivation betrachten wir ein Beispiel.

BEISPIEL III.6.10 (Fünf-Punkte-Stern). Betrachte das Problem (III.6.1) mit $A = I$, $\alpha = 0$, $\Omega = [0,1]^2$ und Gitterweite $h = \frac{1}{n+1}$. Die Differenzengleichung lautet dann

$$f(x) = L_h(x) = \frac{1}{h^2}[4u(x) - u(x + he_1) - u(x - he_1) \\ - u(x + he_2) - u(x - he_2)]$$

für alle $x \in \Omega_h$. Für $1 \le k, \ell \le n$ definiere $u_{k,\ell}(x) = \sin(k\pi x_1)\sin(\ell\pi x_2)$. Aus den Additionstheoremen folgt dann für alle $x \in \Omega_h$

$$L_h u_{k,\ell}(x) = \lambda_{k,\ell} u_{k,\ell}(x) \quad \text{mit} \quad \lambda_{k,\ell} = \frac{2}{h^2}[2 - \cos(k\pi h) - \cos(\ell\pi h)].$$

Der kleinste Eigenwert ist

$$\lambda_{\min} = \min_{1 \le k,\ell \le n} \lambda_{k,\ell} = \lambda_{1,1} = \frac{4}{h^2}[1 - \cos(\pi h)] \\ = \frac{8}{h^2}\sin^2\left(\frac{\pi h}{2}\right);$$

der größte Eigenwert ist wegen $(n+1)h = 1$

$$\begin{aligned}\lambda_{\max} &= \max_{1\le k,\ell\le n} \lambda_{k,\ell} = \lambda_{n,n} = \frac{4}{h^2}[1-\cos(n\pi h)] = \frac{4}{h^2}[1+\cos(\pi h)] \\ &= \frac{8}{h^2}\cos^2\left(\frac{\pi h}{2}\right).\end{aligned}$$

Aus den Additionstheoremen folgt außerdem für $1 \le k,\ell,\mu,\nu \le n$

$$\begin{aligned}&(u_{k,\ell}\,,\, u_{\mu,\nu})_h \\ &= h^2 \left\{\sum_{i=1}^{n} \sin(k\pi ih)\sin(\mu\pi ih)\right\}\left\{\sum_{j=1}^{n} \sin(l\pi jh)\sin(\nu\pi jh)\right\} \\ &= \frac{1}{4}h^2 \left\{\sum_{i=1}^{n} [\cos((k-\mu)\pi ih) - \cos((k+\mu)\pi ih)]\right\} \\ &\qquad \cdot \left\{\sum_{j=1}^{n} [\cos((\ell-\nu)\pi jh) - \cos((\ell+\nu)\pi jh)]\right\} \\ &= \frac{1}{4}h^2 n^2 \delta_{k\mu}\delta_{l\nu}.\end{aligned}$$

Die Funktion $u_{k,\ell}$ sind also paarweise orthogonal bzgl. $(\cdot\,,\,\cdot)_h$.
Wir lösen das LGS (III.6.1) mit der Richardson-Iteration und Dämpfungsparameter $\omega = \frac{h^2}{8}$. Dann sind die $u_{k,\ell}$ auch Eigenfunktionen der Iterationsmatrix. Seien e^0 bzw. e^1 der Fehler vor bzw. nach einem Iterationsschritt. Mit

$$e^0 = \sum_{1\le k,\ell\le n} c_{k,\ell} u_{k,\ell}$$

gilt dann

$$e^1 = \sum_{1\le k,\ell\le n} c_{k,\ell}\left(1 - \frac{h^2}{8}\lambda_{k,\ell}\right) u_{k,\ell}$$

und somit

$$\begin{aligned}\left\|e^0\right\|_{0,h} &= \frac{1}{2}\frac{n}{n+1}\left\{\sum_{1\le k,\ell\le n} c_{k,\ell}^2\right\}^{\frac{1}{2}}, \\ \left\|e^1\right\|_{0,h} &= \frac{1}{2}\frac{n}{n+1}\left\{\sum_{1\le k,\ell\le n}\left(1-\frac{h^2}{8}\lambda_{k,\ell}\right)^2 c_{k,\ell}^2\right\}^{\frac{1}{2}}.\end{aligned}$$

Offensichtlich gilt

$$\max_{1\le k,\ell\le n}\left|1-\frac{h^2}{8}\lambda_{k,\ell}\right| = \left|1-\frac{h^2}{8}\lambda_{1,1}\right| = 1-\sin^2\left(\frac{\pi h}{2}\right) = 1-\mathrm{O}(h^2)$$

und

$$\max_{\max\{k,\ell\}\ge\frac{n}{2}}\left|1-\frac{h^2}{8}\lambda_{k,\ell}\right| = \max_{\max\{k,\ell\}\ge\frac{n}{2}}\left|\frac{1}{2}+\frac{1}{4}\cos(k\pi h)+\frac{1}{4}\cos(\ell\pi h)\right| \le \frac{3}{4}.$$

Der Fehler wird also pro Iteration nur um einen Faktor $1 - \mathrm{O}(h^2)$ gedämpft. Die Komponenten des Fehlers zu den Eigenfunktionen $u_{k,\ell}$ mit $\max\{k,\ell\} \geq \frac{n}{2}$ werden dagegen mindestens um den Faktor $\frac{3}{4}$ gedämpft. Die schlechte Konvergenz des Verfahrens rührt also von den langsam schwingenden Fehleranteilen her. Diese könnten aber gut auf einem Gitter mit doppelter Gitterweite $2h$ approximiert werden.

Beispiel III.6.10 führt auf die Mehrgitteridee:
Führe einige wenige Schritte eines stationären Iterationsverfahrens aus, um die schnell schwingenden Fehleranteile zu reduzieren. Löse danach ein analoges Problem auf dem groben Gitter zur Schrittweite $2h$, um die langsam schwingenden Fehleranteile zu reduzieren. Das Grobgitterproblem wird dabei mit dem gleichen Algorithmus nur näherungsweise gelöst.

Zur Beschreibung des Mehrgitteralgorithmus benötigen wir folgende Notationen:

$h_k = 2^{-k}h_0$	$0 \leq k \leq R$, Gitterweiten,
$L_k = L_{h_k}$	Diskretisierung auf Ω_{h_k},
f_k	rechte Seite,
$f_R = f_{h_R}$,	$f_k, 0 \leq k \leq R-1$, wird rekursiv bestimmt,
u_k	Näherungslösung auf Ω_{h_k},
$I_{k-1,k}$	Interpolationsoperator von $\Omega_{h_{k-1}}$ nach Ω_{h_k},
$R_{k,k-1}$	Restriktionsoperator von Ω_{h_k} nach $\Omega_{h_{k-1}}$,
M_k, N_k	beschreiben ein stationäres Iterationsverfahren für $L_k u_k = f_k$,
$L_k = M_k - N_k$,	M_k regulär.

Bemerkung III.6.11 (Bausteine des Mehrgitterverfahrens). (1) Algorithmus III.6.4 ist rekursiv. Für Programmiersprachen wie Fortran, die keine Rekursionen zulassen, kann diese Rekursion explizit aufgelöst werden. Besonders einfach ist dies für den Fall $\mu = 1$.
(2) Der Parameter μ bestimmt wesentlich die Komplexität des Mehrgitteralgorithmus. Gebräuchlichsind $\mu = 1$, genannt *V-Zyklus*, und $\mu = 2$, genannt *W-Zyklus*. Der Ablauf von Algorithmus III.6.4 ist in Abbildung III.6.1 für $\mu = 1$ und $R = 2$, d.h. drei Gitter, schematisch dargestellt.
(3) Zur Glättung kann fast jedes stationäre Iterationsverfahren verwendet werden. Besonders beliebt sind die Richardson-, Jacobi- und Gauß-Seidel-Iterationen.
(4) Es gibt verschiedene Möglichkeiten für die Restriktion und Interpolation. Sehr häufig verwendet wird eine lineare Interpolation. Es hat sich als theoretisch und praktisch vorteilhaft erwiesen, wenn die Restriktion zur Interpolation adjungiert ist. Die lineare Interpolation und die dazu adjungierte Restriktion sind in den Abbildungen III.6.2 und

Algorithmus III.6.4 $\texttt{MG}(k, \mu, \nu_1, \nu_2, L_k, f, u)$ eine Iteration des Mehrgitteralgorithmus auf dem k-ten Gitter Ω_{h_k}

Gegeben: Gitterzahl k, Parameter μ, ν_1, ν_2, Matrix L_k, rechte Seite f, Approximation M_k für L_k^{-1}, Startwert u.
Gesucht: Verbesserte Näherung u.

1: **if** $k = 0$ **then**
2: $\quad u \leftarrow L_0^{-1} f$, `stop`
3: **end if**
4: **for** $i = 1, \ldots, \nu_1$ **do** ▷ Vor-Glättung
5: $\quad u \leftarrow u + M_k(f - L_k u)$
6: **end for**
7: $f \leftarrow R_{k,k-1}(f - L_k u)$, $v \leftarrow 0$ ▷ Grobgitterkorrektur
8: Führe μ Iterationen von $\texttt{MG}(k-1, \mu, \nu_1, \nu_2, L_{k-1}, f, v)$ aus; Ergebnis v.
9: $u \leftarrow u + I_{k-1,k} v$
10: **for** $i = 1, \ldots, \nu_2$ **do** ▷ Nach-Glättung
11: $\quad u \leftarrow u + M_k(f - L_k u)$
12: **end for**

III.6.3 schematisch dargestellt. Analoge Vorschriften gelten in drei Dimensionen.
(5) Algorithmus III.6.4 ist besonders effizient in Verbindung mit Algorithmus III.6.1. Man spricht dann von einem *vollen Mehrgitteralgorithmus* oder auch *full multigrid algorithm*.
(6) Das exakte Lösen in Schritt (1) von Algorithmus III.6.4 kann durch einige Iterationen eines stationären Iterationsverfahrens ersetzt werden.
(7) Falls in Schritt (3) von Algorithmus III.6.4 das Grobgitterproblem exakt gelöst wird, spricht man von einem *Zweigitteralgorithmus*. Dieser ist nur für die Konvergenzanalyse (s. Beweis von Satz III.6.14) von Interesse.

Als nächstes schätzen wir den Aufwand von Algorithmus III.6.4 ab.

Satz III.6.12 (Aufwand des Mehrgitteralgorithmus). *Es gebe von k unabhängige Konstanten g, d, r, p, so dass ein Glättungsschritt, die Berechnung von $L_k u_k$, die Berechnung von $R_{k,k-1} u_k$ und die Berechnung von $I_{k-1,k} u_{k-1}$ jeweils gN_k, dN_k, rN_k bzw. pN_k Operationen erfordert. Weiter sei $1 \le \mu < 2^n$. Bezeichne mit m_k den Aufwand für eine Mehrgitteriteration auf dem k-ten Gitter. Dann gilt*

$$m_k \le \mu^k m_0 + \frac{2^n}{2^n - \mu}[(\nu_1 + \nu_2)g + d + r + p]N_k.$$

Asymptotisch, d.h. für großes k, benötigt Algorithmus III.6.4 pro Iteration und pro Gitterpunkt $\frac{2^n}{2^n-\mu}[(\nu_1 + \nu_2)g + d + r + p]$ Operationen.

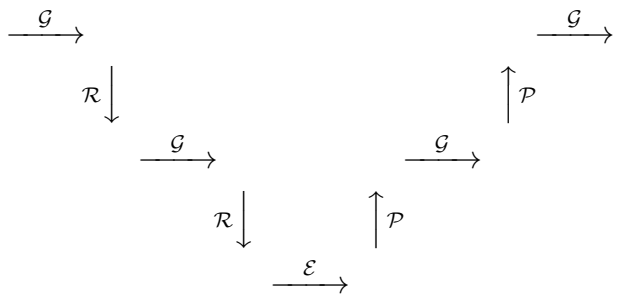

ABBILDUNG III.6.1. Schematischer Verlauf eines Mehrgitterverfahrens mit V-Zyklus und drei Gittern. Es bedeuten: $\mathcal{G}$ Glätten, $\mathcal{R}$ Restringieren, $\mathcal{P}$ Interpolieren, $\mathcal{E}$ exaktes Lösen.

$$\begin{array}{ccccccc}
\times\frac{1}{2} & & & & \times\frac{1}{4} & & \times\frac{1}{4} \\
\circ & & \times\frac{1}{2} \quad \circ & & \times\frac{1}{2} & & \circ \\
\times\frac{1}{2} & & & & \times\frac{1}{4} & & \times\frac{1}{4}
\end{array}$$

ABBILDUNG III.6.2. Interpolation im Feingitterpunkt $\circ$; die Zahlen geben die Gewichte der Grobgitterpunkte $\times$ an.

$$\begin{array}{ccc}
\circ\frac{1}{4} & \circ\frac{1}{2} & \circ\frac{1}{4} \\
\circ\frac{1}{2} & \times 1 & \circ\frac{1}{2} \\
\circ\frac{1}{4} & \circ\frac{1}{2} & \circ\frac{1}{4}
\end{array}$$

ABBILDUNG III.6.3. Restriktion im Grobgitterpunkt $\times$; die Zahlen geben die Gewichte der Gitterpunkte an.

BEWEIS. Aus den Voraussetzungen folgt

$$m_k = \mu m_{k-1} + [(\nu_1 + \nu_2)g + d + r + p]N_k.$$

Wegen $N_{k-1} \leq 2^{-n} N_k$ folgt hieraus durch Induktion

$$\begin{aligned}
m_k &\leq \mu^k m_0 + [(\nu_1 + \nu_2)g + d + r + p]N_k \sum_{l=0}^{k-1} (\mu 2^{-n})^l \\
&\leq \mu^k m_0 + [(\nu_1 + \nu_2)g + d + r + p]N_k \frac{1}{1 - \mu 2^{-n}}.
\end{aligned}$$

Hieraus folgt die erste Behauptung. Die zweite folgt aus

$$\mu^k m_0 N_k^{-1} = [\mu \cdot 2^{-n}]^k m_0 N_0^{-1} \underset{k\to\infty}{\longrightarrow} 0. \qquad \square$$

BEMERKUNG III.6.13 (Aufwand für (III.6.1)). Für das Problem (III.6.1) und Richardson-, Jacobi- oder Gauß-Seidel-Iteration als Glätter gelten die Voraussetzungen von Satz III.6.12 mit $g = d = 2n+1$ und $p = r = n$. Bei Verwendung eines V-Zyklus ($\mu = 1$) mit je einer Vor- und Nachglättung, d.h. $\nu_1 = \nu_2 = 1$, benötigt Algorithmus III.6.4 daher für $n = 2$ und $n = 3$ asymptotisch höchstens 26 Operationen pro Iteration und Gitterpunkt. Die entsprechenden Zahlen für den PCG Algorithmus mit SSOR Vorkonditionierung waren 18 ($n = 2$) bzw. 23 ($n = 3$) Operationen pro Iteration und Gitterpunkt. Da bei den genannten Parametern zwei Iterationen von Algorithmus III.6.4 mit Gauß-Seidel-Glättung ausreichen, um den Fehler gemessen in der $\|\cdot\|_{0,h}$-Norm um den Faktor 10 zu reduzieren, ergibt sich für Problem (III.6.1) mit $A = I$, $\alpha = 0$, $\Omega = [0,1]^n$ die in Tabelle III.6.4 angegebenen Werte für den Aufwand. Bei den genannten Gitterweiten benötigt der Mehrgitteralgorithmus also nur $20-50\%$ des Aufwandes für den PCG-Algorithmus.

TABELLE III.6.4. Aufwand des PCG-Verfahrens mit SSOR-Vorkonditionierung und des Mehrgitterverfahrens mit V-Zylus und je einer Vor- und Nach-Glättung für Problem (III.6.1) mit $A = I$, $\alpha = 0$, $\Omega = [0,1]^n$

n	h	PCG-SSOR	Mehrgitter
2	$\frac{1}{16}$	16′200	11′700
	$\frac{1}{32}$	86′490	48′972
	$\frac{1}{64}$	500′094	206′988
	$\frac{1}{128}$	3′193′542	838′708
3	$\frac{1}{16}$	310′500	175′500
	$\frac{1}{32}$	3′425′965	1′549′132
	$\frac{1}{64}$	$4.0 \cdot 10^7$	$1.3 \cdot 10^7$
	$\frac{1}{128}$	$5.2 \cdot 10^8$	$1.1 \cdot 10^8$

Aus Zeitgründen können wir hier keinen vollständigen Konvergenzbeweis für Algorithmus III.6.4 bringen. Die folgenden Sätze und Bemerkungen sollen einen Eindruck über die zugrunde liegende Idee liefern.

SATZ III.6.14 (Konvergenz des Mehrgitteralgorithmus). *Bezeichne mit $|||\cdot|||_k$ die zur $\|\cdot\|_{0,h_k}$-Norm gehörende Operaternorm. Folgende Voraussetzungen seien erfüllt:*

(1) (Glättungseigenschaft) *Es gibt eine monoton fallende Funktion* $\eta : [1, \infty) \to \mathbb{R}_+^*$ *mit* $\lim_{z\to\infty} \eta(z) = 0$, *ein* $\alpha \in \mathbb{R}_+^*$ *und ein* $c_1 \in \mathbb{R}_+^*$ *mit*

$$\left|\left|\left| L_k(I - M_k^{-1}L_k)^\nu \right|\right|\right|_k \leq c_1\eta(\nu)h_k^{-\alpha}$$

für alle $k \geq 0$.

(2) (Approximationseigenschaft) *Es gibt ein* $c_2 \in \mathbb{R}_+^*$ *mit*

$$\left|\left|\left| L_k^{-1} - I_{k-1,k}L_{k-1}^{-1}R_{k,k-1} \right|\right|\right|_k \leq c_2 h_k^\alpha$$

für alle $k \geq 0$, *wobei* α *wie in (1) ist.*

(3) *Es ist*

$$R_{k,k-1}I_{k-1,k} = Id_{k-1},$$

und es gibt $c_3, c_4 \in \mathbb{R}_+^*$ *mit*

$$|||I_{k-1,k}|||_k = \sup_{\|u_{k-1}\|_{0,h_{k-1}}=1} \|I_{k-1,k}u_{k-1}\|_{0,h_k} \leq c_3$$

für alle $k \geq 1$ *und*

$$|||R_{k,k-1}|||_k = \sup_{\|u_k\|_{0,h_k}=1} \|R_{k,k-1}u_k\|_{0,h_{k-1}} \leq c_4$$

für alle $k \geq 1$.

(4) $\left|\left|\left| I - M_k^{-1}L_k \right|\right|\right|_k \leq 1$ *für alle* $k \geq 0$.

Dann gilt:

(i) *Die Konvergenzrate* δ_{TG} *des Zweigitteralgorithmus gemessen in der* $\|\cdot\|_{0,h_k}$*-Norm ist beschränkt durch*

$$\delta_{TG} \leq c_1c_2\eta(\nu_1).$$

(ii) *Es gibt ein* $\nu_0 \in \mathbb{N}$, *so dass der Mehrgitteralgorithmus mit* $\mu \geq 2$, $\nu_1 \geq \nu_0$ *und* $\nu_2 \geq 0$ *konvergiert. Für seine Konvergenzrate* δ_k *gemessen in der* $\|\cdot\|_{0,h_k}$*-Norm gilt*

$$\delta_k \leq 2\delta_{TG} \leq \frac{1}{2}.$$

BEWEIS. Bezeichne mit TG_k und MG_k die Iterationsmatrizen des Zwei- bzw. Mehrgitteralgorithmus auf dem k-ten Gitter. Dann gilt

$$TG_k = [I - M_k^{-1}L_k]^{\nu_2}[I - I_{k-1,k}L_{k-1}^{-1}R_{k,k-1}L_k][I - M_k^{-1}L_k]^{\nu_1},$$
$$MG_0 = 0$$

und

$$MG_k = [I - M_k^{-1}L_k]^{\nu_2}[I - I_{k-1,k}L_{k-1}^{-1}R_{k,k-1}L_k + I_{k-1,k}MG_{k-1}^\mu L_{k-1}^{-1}R_{k,k-1}L_k][I - M_k^{-1}L_k]^{\nu_1}.$$

Damit folgt aus (1), (2) und (4)

$$\delta_{TG} = |||TG_k|||_k \leq \left|\left|\left| [L_k^{-1} - I_{k-1,k}L_{k-1}^{-1}R_{k,k-1}]L_k(I - M_k^{-1}L_k)^{\nu_1} \right|\right|\right|_k \leq c_1c_2\eta(\nu_1)$$

Dies beweist (i).
Wähle $\nu_0 \in \mathbb{N}$, so dass

$$c_1 c_2 \eta(\nu_0) \leq \min\left\{(5c_3c_4)^{-1}, \frac{1}{4}\right\}$$

ist. Wegen (1) ist dies möglich. Aus (1) – (4) folgt dann für $\mu \geq 2$, $\nu_1 \geq \nu_2$ und $\nu_2 \geq 0$

$$\begin{aligned}
\delta_k &= |||MG_k|||_k \\
&\leq \delta_{TG} + \left|\left|\left|[I_{k-1,k} MG_{k-1}^{\mu} L_{k-1}^{-1} R_{k,k-1} L_k] \cdot [I - M_k^{-1} L_k]^{\nu_1}\right|\right|\right|_k \\
&\leq \delta_{TG} + c_3 \delta_{k-1}^{\mu} \left|\left|\left| L_{k-1}^{-1} R_{k,k-1} L_k [I - M_k^{-1} L_k]^{\nu_1}\right|\right|\right|_k \\
&\leq \delta_{TG} + c_3 c_4 \delta_{k-1}^{\mu} \left\{\left|\left|\left|[I - M_k^{-1} L_k]^{\nu_1}\right|\right|\right|_k \right. \\
&\quad \left. + \left|\left|\left| L_k^{-1} - I_{k-1,k} L_{k-1}^{-1} R_{k,k-1}\right|\right|\right|_k \left|\left|\left| L_k [I - M_k^{-1} L_k]^{\nu_1}\right|\right|\right|_k \right\} \\
&\leq \delta_{TG} + c_3 c_4 \delta_{k-1}^{\mu} \left\{1 + c_1 c_2 \eta(\nu_1)\right\} \\
&\leq \delta_{TG} + \frac{5}{4} c_3 c_4 \delta_{k-1}^{\mu}.
\end{aligned}$$

Offensichtlich gilt $\delta_0 = 0 \leq 2\delta_{TG}$. Nehme also an, dass (ii) für $k = 1$ bewiesen ist. Dann folgt aus obiger Abschätzung

$$\begin{aligned}
\delta_k &\leq \delta_{TG} + \frac{5}{4} c_3 c_4 \delta_{k-1}^2 \leq \delta_{TG} + \frac{5}{4} c_3 c_4 4 \delta_{TG}^2 = \delta_{TG}(1 + 5 c_3 c_4 \delta_{TG}) \\
&\leq 2\delta_{TG}.
\end{aligned}$$

□

Bemerkung III.6.15. (1) Die Voraussetzung (4) von Satz III.6.14 sei erfüllt, d.h., das als Glätter verwendete Iterationsverfahren ist konvergent. Dann kann man zeigen, dass

$$\left|\left|\left| L_k \left(I - \frac{1}{2} M_k^{-1} L_k\right)^{\nu} \right|\right|\right|_k \leq \frac{c}{\sqrt{\nu}} |||M_k|||_k$$

ist mit einer von k unabhängigen Konstanten c. Für die Richardson-, Jacobi- und Gauß-Seidel-Iteration angewandt auf das LGS (III.6.1) kann man zudem die Abschätzung $|||M_k|||_k \leq c' h_k^{-2}$ beweisen.
(2) Mit einigem technischen Mehraufwand kann man zeigen, dass für Problem (III.6.1) die Approximationseigenschaft mit $\alpha = 2$ erfüllt ist, sofern Ω konvex ist. Diese Einschränkung rührt daher, dass die Approximationseigenschaft äquivalent ist zu einer Regularitätsaussage für die Lösung der pDGl, die nur für konvexe Gebiete und Gebiete mit glattem Rand erfüllt ist.
(3) Für die Interpolation und Restriktion aus Beispiel III.6.11 (4) ist die Voraussetzung (3) von Satz III.6.14 mit $c_3 = c_4 = 1$ erfüllt.
(4) Mit einer anderen Technik kann man ohne Regularitätsvoraussetzungen an die Lösung der pDGl zeigen, dass die Konvergenzrate von Algorithmus III.6.4 mit $\mu = 1$, $\nu_1 = \nu_2 = \nu \geq 1$ und Gauß-Seidel-Iteration als Glättung durch $\frac{c}{c+\nu}$ beschränkt ist mit einer von k unabhängigen Konstanten c.

(5) Für den unter (4) angegebenen Mehrgitteralgorithmus erhält man in der Praxis in Abhängigkeit von der Glattheit der Koeffizienten A und α Konvergenzraten zwischen 0.1 und 0.5.

Index

Literaturverzeichnis

[1] H. Amann, *Gewöhnliche Differentialgleichungen*, de Gruyter Lehrbuch, Walter de Gruyter & Co., Berlin, 1983.

[2] M. Crouzeix and A. L. Mignot, *Analyse numérique des équations différentielles*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1984.

[3] W. Dahmen and A. Reusken, *Numerik für Ingenieure und Naturwissenschaftler*, Springer, Berlin, 2006.

[4] P. Deuflhard and F. Bornemann, *Numerische Mathematik 2*, de Gruyter Lehrbuch, Walter de Gruyter & Co., Berlin, 2008.

[5] R. D. Grigorieff, *Numerik gewöhnlicher Differentialgleichungen, Band 1: Einschrittverfahren*, B. G. Teubner, Stuttgart, 1972.

[6] A. Iserles, *A first course in the numerical analysis of differential equations*, second ed., Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2009.

[7] L. R. Scott, *Numerical Analysis*, Princeton University Press, Princeton, 2011.

[8] R. Verfürth, *Analysis I-III*, Skriptum, Ruhr-Universität Bochum, Bochum, Oktober 2006, 479 Seiten,
www.rub.de/num1/files/lectures/Analysis.pdf.

[9] ———, *Einführung in die Numerische Mathematik*, Skriptum, Ruhr-Universität Bochum, Bochum, März 2013, 137 Seiten,
www.rub.de/num1/files/lectures/EinfNumerik.pdf.