10 Differenzenverfahren

10.1 Systeme erster Ordnung

Wir konzentrieren uns im Folgenden auf die Betrachtung *linearer* RWA im $\mathbb{R}^d$, obwohl die meisten Aussagen sinngemäß auch für die Approximation isolierter Lösungen nichtlinearer Probleme übertragen werden können. Zur Abkürzung der Notation führen wir Operatoren $L : C^1(I) \to C(I)$ und $R : C(I) \to \mathbb{R}$ ein:

$$\begin{aligned} (Lu)(t) &:= u'(t) - A(t)u(t) = f(t)\,, \quad t \in I = [a,b], \\ Ru &:= B_a u(a) + B_b u(b) = g, \end{aligned} \tag{10.1.1}$$

Wir nehmen an, dass diese RWA eine eindeutige Lösung besitzt, bzw. dass die zugehörige Matrix $B_a + B_b Y(b)$ regulär ist.

Grundsätzlich kann jedes Differenzenverfahren zur Lösung von AWA,

$$u'(t) - A(t)u(t) = f(t)\,, \quad t \geq a\,, \quad u(a) = \alpha\,, \tag{10.1.2}$$

auch zur Lösung der RWA (10.1.1) verwendet werden. Dazu wählt man etwa ein äquidistantes Gitter auf $[a,b]$,

$$t_n = a + nh\,, \quad 0 \leq n \leq N\,, \quad h = (b-a)/N\,,$$

auf dem Approximationen y_n zu $u(t_n)$ bestimmt werden sollen. Anwendung der Differenzenformeln auf eine Gitterfunktion $y^h := (y_n)_{n=0}^N$ ergibt dann ein System von N Differenzengleichungen

$$(L_h y^h)_n := \sum_{j=0}^{N} C_{nj}(h) y_j = F_n(h; f)\,, \quad n = 1, \ldots, N\,. \tag{10.1.3}$$

Eine $(N+1)$-te Gleichung erhält man durch exakte Übernahme der Randbedingung

$$R_h y^h := B_a y_0 + B_b y_N = g\,. \tag{10.1.4}$$

Das diskrete Operatorpaar $\{L_h, R_h\}$ wird als Approximation von $\{L, R\}$ betrachtet. Durch die Beziehungen (10.1.3), (10.1.4) mit der Koeffizientenmatrix $(C_{nj}(h))_{n,j=1}^N$ und dem inhomogenen Vektor $(F_n(h;f))_{n=1}^N$ wird eine sehr allgemeine Klasse von Differenzenapproximationen der RWA (10.1.1) erfasst; alle im ersten Teil des Textes beschriebenen Einschritt-, Mehrschritt- und Extrapolationsverfahren fallen darunter. Man erhält so ein lineares Gleichungssystem

$$\mathcal{A}_h y^h = \beta^h \tag{10.1.5}$$

für den diskreten Lösungsvektor $y^h = (y_0, \ldots, y_N)^T \in \mathbb{R}^{(N+1)d}$ mit

$$\mathcal{A}_h = \begin{bmatrix} B_a & 0 & \ldots & B_b \\ C_{10}(h) & C_{11}(h) & \ldots & C_{1N}(h) \\ \vdots & \vdots & \ddots & \vdots \\ C_{N0}(h) & C_{N1}(h) & \ldots & C_{NN}(h) \end{bmatrix}, \quad \beta_h = \begin{bmatrix} g \\ F_1(h;f) \\ \vdots \\ F_N(h;f) \end{bmatrix}.$$

Dies ist eine $(N+1)\times(N+1)$-Blockmatrix, deren Blockelemente wir mit $\mathcal{A}_{h;nm}$ bezeichnen. Im Folgenden wird die einheitliche Bezeichnung y^h sowohl für die diskrete Gitterfunktion $y^h : \{t_0, ..., t_N\} \to \mathbb{R}^d$ wie für den zugehörigen Block-Vektor der Gitterwerte $y^h := (y_0, ..., y_N)^T$ verwenden. Verwechslungen sollten ausgeschlossen sein.

Beispiel 10.1: Eins der einfachsten Differenzenverfahren ist das sog. „Box-Schema“, welches auf der Trapezregel als AWA-Löser basiert:

$$L_h y_n = h^{-1}(y_n - y_{n-1}) - \tfrac{1}{2}A_{n-1/2}(y_n + y_{n-1}) = f_{n-1/2}\,,$$

wobei $t_{n-1/2} := \frac{1}{2}(t_n + t_{n-1})$ und $A_{n-1/2} := A(t_{n-1/2})$ gesetzt ist. Das zugehörige Gleichungssystem (10.1.5) hat hier die Gestalt

$$\begin{bmatrix} B_a & & 0 & & B_b \\ -I - \frac{1}{2}hA_{1/2} & & I - \frac{1}{2}hA_{1/2} & & 0 \\ & \ddots & & \ddots & \\ 0 & & -I - \frac{1}{2}hA_{N-1/2} & & I - \frac{1}{2}hA_{N-1/2} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} g \\ hf_{1/2} \\ \vdots \\ hf_{N-1/2} \end{bmatrix}.$$

Die eindeutige Lösbarkeit dieses Systems (für hinreichend kleines h) wird unten im Rahmen eines sehr viel allgemeineren Resultats gezeigt werden.

Zur Untersuchung der Konvergenz der Differenzenapproximation (10.1.3), (10.1.4) führen wir wieder den Abschneidefehler τ^h als Gitterfunktion ein durch

$$\tau^h := L_h u^h - F^h(h; f)\,.$$

Dabei bezeichnet u^h die durch Restriktion der Lösung u auf das Punktgitter $\{t_o, ..., t_N\}$ gewonnene Gitterfunktion. Im Falle einer AWA hatten wir gesehen, dass hinreichend für die Konvergenz der Differenzenapproximation ihre *Konsistenz* (mit Ordnung $m \geq 1$),

$$\|y_0 - u_0\| + \max_{1\leq n\leq N} \|\tau_n\| = O(h^m)\,, \tag{10.1.6}$$

und ihre *Stabilität*,

$$\max_{0\leq n\leq N} \|y_n\| \leq K\big\{\|y_0\| + \max_{1\leq n\leq N} \|(L_h y^h)_n\|\big\}, \tag{10.1.7}$$

für Gitterfunktionen $y^h = (y_n)_{0\leq n\leq N}$ ist. Die Stabilitätsbedingung (10.1.7) ist lediglich eine Umformulierung der Aussagen der Stabilitätssätze 2.2 und 4.2 für Einschritt- bzw. für lineare Mehrschrittverfahren, wobei $K \sim e^{L(b-a)}$ ist. Wird nun das Differenzenverfahren zur Lösung der RWA (10.1.1) angewendet, so definiert man sinngemäß *Konsistenz* (mit der Ordnung m) durch

$$\|R_h u^h - g\| + \max_{1\leq n\leq N} \|(L_h u^h)_n - F_n(h; f)\| = O(h^m) \tag{10.1.8}$$

und *Stabilität* durch

$$\max_{0\le n\le N}\|y_n\| \le K\big\{\|R_h y^h\| + \max_{1\le n\le N}\|(L_h y^h)_n\|\big\} \tag{10.1.9}$$

für Gitterfunktionen $y^h = (y_n)_{0\le n\le N}$ und hinreichend kleines h.

Die Stabilitätseigenschaft (10.1.9) des Differenzenoperators $\{L_h, R_h\}$ lässt sich äquivalent durch eine Stabilitätseigenschaft der zugehörigen Matrix $\mathcal{A}_h$ ausdrücken. Dazu führen wir die von der Vektornorm

$$\|y^h\|_\infty := \max_{0\le n\le N}\|y_n\|, \quad y^h \in \mathbb{R}^{(N+1)d},$$

erzeugte natürliche Matrizennorm ein:

$$\|\mathcal{A}_h\|_\infty := \sup_{y^h\in\mathbb{R}^{(N+1)d}} \frac{\|\mathcal{A}_h y^h\|_\infty}{\|y^h\|_\infty}, \quad \mathcal{A}_h \in \mathbb{R}^{(N+1)d\times(N+1)d}.$$

Diese Matrizennorm ist gerade die sog. „Maximale-Blockzeilensummen-Norm“

$$\|\mathcal{A}_h\|_\infty = \max_{n=0,\dots,N}\sum_{m=0}^{N}\|\mathcal{A}_{h;n,m}\|,$$

wobei wiederum $\|\mathcal{A}_{h;n,m}\|$ die von der euklidischen Vektornorm des $\mathbb{R}^d$ erzeugte natürliche Matrizennorm des $\mathbb{R}^{d\times d}$ ist.

Hilfssatz 10.1 (Stabilität linearer Differenzenverfahren): *Die Stabilität des Differenzenschemas (10.1.3), (10.1.4) ist äquivalent dazu, dass die zugehörigen Matrizen $\mathcal{A}_h$ regulär sind mit*

$$\sup_{h>0}\|\mathcal{A}_h^{-1}\|_\infty < \infty. \tag{10.1.10}$$

Beweis: Für eine Gitterfunktion $y^h = \{y_n\}_{0\le n\le N}$ bzw. Gitterwertvektor $y^h = (y_n)_{n=0}^N$ ist $\mathcal{A}_h y^h = 0$ äquivalent mit $(L_h y^h)_n = 0\,(1\le n\le N)$ und $R_h y^h = 0$. Aus (10.1.9) folgt also notwendig die Regularität von $\mathcal{A}_h$ sowie $\|\mathcal{A}_h^{-1}\|_\infty \le c$ (gleichmäßig bzgl. h) und umgekehrt:

$$\|\mathcal{A}_h^{-1}\|_\infty := \sup_{y^h\in\mathbb{R}^{(N+1)d}} \frac{\|y^h\|_\infty}{\|A_h y^h\|_\infty} \le cK.$$

Q.E.D.

In Analogie zu den Konvergenzsätzen für Diskretisierungen von AWA haben wir nun den folgenden

Satz 10.1 (Konvergenzsatz): *Ist das Differenzenschema (10.1.3), (10.1.4) konsistent mit der Ordnung $m \ge 1$ und stabil, so ist es auch konvergent mit der Ordnung m:*

$$\max_{t_n\in I}\|y_n - u(t_n)\| = O(h^m) \quad (h\to 0).$$

Beweis: Für die Fehlerfunktion $e^h := y^h - u^h$ gilt

$$L_h e^h = L_h y^h - L_h u^h = F^h(h; f) - L_h u^h = -\tau^h \,.$$

Aufgrund der Konsistenz folgt

$$\|R_h e^h\| + \max_{1 \le n \le N} \|\tau_n\| = O(h^m) \,,$$

so dass die Stabilität direkt die gewünschte Konvergenzaussage impliziert. Q.E.D.

Das Hauptproblem bei der Analyse von Differenzendiskretisierungen der RWA (10.1.1) besteht also im Nachweis der Stabilität. Überraschenderweise gibt es aber dafür ein sehr allgemeines Kriterium:

Satz 10.2 (Äquivalenzsatz): *Das Differenzenschema (10.1.3), (10.1.4) ist konsistent (mit Ordnung m) und stabil für die RWA (10.1.1) genau dann, wenn es konsistent (mit Ordnung m) und stabil für die AWA (10.1.2) ist.*

Beweis: (i) Die Äquivalenz der Konsistenz mit Ordnung m ist klar, da die Randbedingung bzw. Anfangsbedingung exakt erfüllt werden. Die AWA kann als spezielle RWA mit $B_a = I$ und $B_b = 0$ aufgefasst werden. Zum Beweis des Satzes betrachten wir nun zwei beliebige Randwertaufgaben RWA(0) bzw. RWA(1) für das System

$$Lu(t) = u'(t) - A(t)u(t) = f(t) \,, \quad t \in I \,, \tag{10.1.11}$$

zu den Randbedingungen

$$R^{(i)}u = B_a^{(i)}u(a) + B_b^{(i)}u(b) = g \,, \quad i = 0, 1, \tag{10.1.12}$$

und nehmen an, dass beide eindeutig lösbar sind. Dies ist gleichbedeutend damit, dass die zugehörigen Matrizen $B_a^{(i)} + B_b^{(i)}Y(b)$ beide regulär sind. Hierbei bezeichnet $Y(t)$ wieder die Fundamentalmatrix des Systems (10.1.11), d. h. die Lösung der Matrix-AWA $Y'(t) - A(t)Y(t) = 0,\ t \in [a, b],\ Y(a) = I$. Es ist dann zu zeigen, dass die Stabilität des Differenzenschemas für RWA(0) auch die für RWA(1) impliziert. Die zugehörigen Verfahrensmatrizen sind

$$\mathcal{A}_h^{(i)} = \begin{bmatrix} B_a^{(i)} & 0 & \dots & 0 & B_b^{(i)} \\ C_{10}(h) & & \dots & & C_{1N}(h) \\ \vdots & & & & \vdots \\ C_{N0}(h) & & \dots & & C_{NN}(h) \end{bmatrix} \,, \quad i = 0, 1 \,.$$

Sei nun das Schema stabil für RWA(0), d. h.: Nach Hilfssatz 10.1 ist $\mathcal{A}_h^{(0)}$ regulär und $\sup_{h>0} \|A_h^{(0)-1}\|_\infty < \infty$. Mit der Differenz

$$D_h := A_h^{(1)} - A_h^{(0)} = \begin{bmatrix} B_a^{(1)} - B_a^{(0)} & 0 & \dots & 0 & B_b^{(1)} - B_b^{(0)} \\ 0 & & \dots & & 0 \\ \vdots & & & & \vdots \\ 0 & & \dots & & 0 \end{bmatrix}$$

schreiben wir

$$\mathcal{A}_h^{(1)} = (I + D_h\,\mathcal{A}_h^{(0)-1})\,\mathcal{A}_h^{(0)}\,. \tag{10.1.13}$$

Im Hinblick auf die angenommene Regularität der Matrizen $\mathcal{A}_h^{(0)}$ und der gleichmäßigen beschränktheit ihrer Inversen muss jetzt die Matrix $I + D_h\,\mathcal{A}_h^{(0)-1}$ untersucht werden.

(ii) Wir wollen jetzt zeigen, dass die Matrix $I + D_h\,\mathcal{A}_h^{(0)-1}$ regulär und ihre Inverse gleichmäßig in h beschränkt ist. Dazu schreiben wir die Inverse von $\mathcal{A}_h^{(0)}$ in der Blockform

$$\mathcal{A}_h^{(0)-1} = \begin{bmatrix} Z_{00}^{(0)} & \dots & Z_{0N}^{(0)} \\ \vdots & & \vdots \\ Z_{N0}^{(0)} & \dots & Z_{NN}^{(0)} \end{bmatrix}, \quad Z_{jk}^{(0)} \in \mathbb{R}^{d\times d}\,.$$

Durch zeilenweise Auswertung der Beziehung $\mathcal{A}_h^{(0)}\mathcal{A}_h^{(0)-1} = I$ ergeben sich dann die Beziehungen

$$B_a^{(0)} Z_{00}^{(0)} + B_b^{(0)} Z_{N0}^{(0)} = I \tag{10.1.14}$$

$$B_a^{(0)} Z_{0k}^{(0)} + B_b^{(0)} Z_{Nk}^{(0)} = 0\,, \quad k = 1, ..., N\,, \tag{10.1.15}$$

und

$$\sum_{l=0}^{N} C_{nl} Z_{l0}^{(0)} = 0\,, \quad n = 1, ..., N\,. \tag{10.1.16}$$

Mit diesen Bezeichnungen können wir schreiben:

$$\begin{aligned} I + D_h A_h^{(0)-1} &= \begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix} + \begin{bmatrix} B_a^{(1)} - B_a^{(0)} & \dots & B_b^{(1)} - B_b^{(0)} \\ & & \\ 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} Z_{00}^{(0)} & \dots & Z_{0N}^{(0)} \\ \vdots & & \vdots \\ Z_{N0}^{(0)} & \dots & Z_{NN}^{(0)} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q_{h,0}} & Q_{h,1} & \dots & Q_{h,N} \\ & I & & \\ & & \ddots & \\ & & & I \end{bmatrix}, \end{aligned}$$

mit den Matrizen

$$\begin{aligned} Q_{h,0} &:= I + [B_a^{(1)} - B_a^{(0)}]\, Z_{00}^{(0)} + [B_b^{(1)} - B_b^{(0)}] Z_{N0}^{(0)} \\ &= I + [B_a^{(1)} Z_{00}^{(0)} + B_b^{(1)} Z_{N0}^{(0)}] - \underbrace{[B_a^{(0)} Z_{00}^{(0)} + B_b^{(0)} Z_{N0}^{(0)}]}_{=\,I}, \\ Q_{h,k} &:= [B_a^{(1)} - B_a^{(0)}]\, Z_{0k}^{(0)} + [B_b^{(1)} - B_b^{(0)}] Z_{Nk}^{(0)} \\ &= [B_a^{(1)} Z_{0k}^{(0)} + B_b^{(1)} Z_{Nk}^{(0)}] - \underbrace{[B_a^{(0)} Z_{0k}^{(0)} + B_b^{(0)} Z_{Nk}^{(0)}]}_{=\,0}. \end{aligned}$$

(iii) Als nächstes wollen wir zeigen, dass der erste Diagonalblock $Q_{h,0} = B_a^{(1)} Z_{00}^{(0)} + B_b^{(1)} Z_{N0}^{(0)}$ regulär ist. Aus (10.1.14) und (10.1.16) entnehmen wir, dass der „Blockvektor“ $Z_0^{(0)} := (Z_{k0}^{(0)})_{k=0,\ldots,N}$ die Beziehungen $L_h Z_0^{(0)} = 0$ und $R_h^{(0)} Z_0^{(0)} = I$ erfüllt. Er ist somit die Differenzenapproximation zur Lösung $Z^{(0)}(t)$ der Matrix-RWA

$$Z'(t) - A(t)Z(t) = 0\,, \quad t \in [a,b]\,, \quad R^{(0)}Z = I. \tag{10.1.17}$$

Nach dem Konvergenzsatz 10.1 gilt dabei für den Fehler

$$\max_{0\le n\le N} \|Z_{n0}^{(0)} - Z^{(0)}(t_n)\| = O(h^m) \quad (h \to 0)\,,$$

und folglich

$$\| \underbrace{B_a^{(1)} Z_{00}^{(0)} + B_b^{(1)} Z_{N0}^{(0)}}_{=Q_{h,0}} - \underbrace{\{B_a^{(1)} Z^{(0)}(a) + B_b^{(1)} Z^{(0)}(b)\}}_{=R^{(1)}Z^0} \| = O(h^m)\,. \tag{10.1.18}$$

Die mit der Fundamentallösung $Y(t) \in \mathbb{R}^{d\times d}$ des Systems (10.1.11) gebildete Matrix $Q^{(0)} := R^{(0)}Y = B_a^{(0)} + B_b^{(0)} Y(b)$ ist nach Voraussetzung regulär. Wegen

$$\begin{aligned}[Y Q^{(0)-1}]'(t) - A(t)[Y(t)Q^{(0)-1}] &= [Y'(t) - A(t)Y(t)]Q^{(0)-1} = 0, \quad t \in [a,b],\\ R^{(0)}[Y(t)Q^{(0)-1}] &= R^{(0)}Y(t)Q^{(0)-1} = Q^{(0)}Q^{(0)-1} = I,\end{aligned}$$

ist dann auch die Matrixfunktion $Y(t)\,Q^{(0)-1}$ Lösung der RWA (10.1.17). Wegen der Eindeutigkeit dieser Lösung folgt die Gleichung

$$Z^{(0)}(t) = Y(t)\,Q^{(0)-1}.$$

Aufgrund der vorausgesetzten Lösbarkeit der Randwertaufgabe RWA(1) ist die Matrix $Q^{(1)} := R^{(1)}Y = B_a^{(1)} + B_b^{(1)} Y(b)$ und damit auch die Matrix

$$R^{(1)}Z^{(0)} = R^{(1)}Y\,Q^{(0)-1} = Q^{(1)}Q^{(0)-1}$$

regulär. Wegen (10.1.18) gibt es für ein beliebiges, aber festes $\rho \in (0,1)$ ein $h_\rho > 0$, so dass gilt:

$$\|Q_{h,0} - Q^{(1)}Q^{(0)-1}\| \le \frac{\rho}{\|Q^{(0)}Q^{(1)-1}\|}\,, \quad 0 < h \le h_\rho\,.$$

Für solche h ist dann auch $Q_{h,0}$ regulär (Übungsaufgabe), und es gilt

$$\|Q_{h,0}^{-1}\| \le \frac{\|Q^{(0)}Q^{(1)-1}\|}{1 - \|Q^{(0)}Q^{(1)-1}\|\,\|Q_{h,0} - Q^{(1)}Q^{(0)-1}\|}\,,$$

d. h.: $\sup_{0<h\le h_0} \|Q_{h,0}^{-1}\| < \infty$.

(iv) Mit $Q_{h,0}$ ist nun auch $I + D_h\,\mathcal{A}_h^{(0)-1}$ und schließlich auch $\mathcal{A}_h^{(1)}$ regulär. Man verifiziert leicht, dass

$$(I - D_h\,\mathcal{A}_h^{(0)-1})^{-1} = \begin{bmatrix} Q_{h,0}^{-1} & -Q_{h,0}^{-1}Q_{h,1} & \ldots & -Q_{h,0}^{-1}Q_{h,N} \\ & I & & \\ & & \ddots & \\ & & & I \end{bmatrix}.$$

Damit folgt die Abschätzung

$$\begin{aligned}\|\mathcal{A}_h^{(1)-1}\|_\infty &= \|\mathcal{A}_h^{(0)-1}(I + D_h\mathcal{A}_h^{(0)-1})^{-1}\|_\infty \le \|\mathcal{A}_h^{(0)-1}\|_\infty \, \|(I + D_h\mathcal{A}_h^{(0)-1})^{-1}\|_\infty \\ &\le c\|\mathcal{A}_h^{(0)-1}\|_\infty \, \max\Big\{1, \|Q_{h,0}^{-1}\|\Big\{1 + \sum_{k=1}^{N} \|Q_{h,k}\|\Big\}\Big\} \le K.\end{aligned}$$

Die Beschränktheit von $\sum_{k=1}^{N} \|Q_{h,k}\|$ erschließen wir wie folgt: Aus der (angenommenen) Beschränktheit der Normen (maximale Block-Zeilensumme) $\|\mathcal{A}_h^{(0)-1}\|_\infty$ folgt insbesondere die Beschränktheit der Blocksummen $\sum_{k=0}^{N} \|Z_{0k}^{(0)}\|$ und $\sum_{k=0}^{N} \|Z_{Nk}^{(0)}\|$, und damit schließlich auch die von

$$\sum_{k=1}^{N} \|Q_{h,k}\| = \sum_{k=1}^{N} \|B_a^{(1)} Z_{0k}^{(0)} + B_b^{(1)} Z_{Nk}^{(0)}\| \le \|B_a^{(1)}\| \sum_{k=1}^{N} \|Z_{0k}^{(0)}\| + \|B_b^{(1)}\| \sum_{k=1}^{N} \|Z_{Nk}^{(0)}\|.$$

Das Differenzenschema ist also auch stabil für AWA(1). Q.E.D.

Als Konsequenz des fundamentalen Satzes 10.2 sehen wir, dass alle bisher betrachteten konvergenten Differenzenverfahren für AWA auch zur Lösung von RWA verwendet werden können, und hier auch dieselbe Konvergenzordnung besitzen.

Beispiel 10.2: Das oben eingeführte Box-Schema hat als Abkömmling der Trapezregel die Konsistenzenordnung $m = 2$ und ist stabil für die AWA (10.1.2). Gemäß Satz 10.2 impliziert dies also die Regularität der zugehörigen Matrizen $\mathcal{A}_h$ für hinreichend kleines h und die Konvergenz der diskreten Lösungen

$$\max_{t_n \in I} \|y_n - u(t_n)\| = O(h^2) \quad (h \to 0)\,.$$

Für den Diskretisierungsfehler lassen sich auch wieder asymptotische Entwicklungen nach Potenzen von h^2 (wegen der Symmetrie der Differenzenformel) nachweisen:

$$y_n - u(t_n) = \sum_{i=1}^{R} h^{2i} e_i(t_n) + O(h^{2R+2})\,.$$

Dabei ist wieder $u \in C^{2R+2}(I)$ angenommen, und die Fehlerfunktionen $e_i(t)$ sind unabhängig von h. Auf der Basis dieser Entwicklung lässt sich für das Box-Schema die Richardson-Extrapolation auf $h = 0$ anwenden. Durch $(R+1)$-malige Auswertung des Gleichungssystems $\mathcal{A}_h y^h = \beta^h$ mit den Gitterweiten $2^{-p}h, p = 0, 1, 2, \ldots, R$, erhält man so eine Näherung $\bar{y}^h$ $(2R+2)$-ter Ordnung:

$$\max_{t_n \in I} \|\bar{y}_n - u(t_n)\| = O(h^{2R+2})\,.$$

10.2 Sturm-Liouville-Probleme

Für die in der Praxis häufig auftretenden Sturm-Liouville-Probleme (skalare RWA 2-ter Ordnung) verwendet man meist spezielle Differenzenverfahren, welche nicht den Umweg über die Transformation auf ein System 1-ter Ordnung gehen. Wir wollen das hier anhand des Spezialfalls mit *Dirichlet*-Randbedingungen

$$\begin{aligned} &Lu(t) := -[pu']'(t) + q(t)u'(t) + r(t)u(t) = f(t)\,, \quad t \in I = [a,b], \\ &u(a) = \alpha\,, \quad u(b) = \beta, \end{aligned} \tag{10.2.19}$$

studieren. Dieses Problem stellt einen eindimensionalen Modellfall für eine große Klasse von höherdimensionalen Differentialgleichungsproblemen dar. Obwohl die im Folgenden betrachtete Diskretisierung von (10.2.19) i. Allg. praktisch nicht konkurrenzfähig (da nur von zweiter Ordnung) ist, erlaubt ihre Analyse doch schon Rückschlüsse auf das Verhalten analoger Verfahren für die sehr viel schwierigeren mehrdimensionalen Probleme. Dieser sowie der nächste Abschnitt über Variationsmethoden sind also im wesentlichen als Vorbereitung für die Untersuchung von Diskretisierungsverfahren bei *partiellen* Differentialgleichungen zu sehen.

Unter den Standardvoraussetzungen

$$p, q \in C^1(I)\,, \quad r, f \in C(I)\,, \quad p(t) \geq \rho > 0\,, \quad t \in I\,,$$

und zusätzlich

$$\rho + (b-a)^2 \min_{t \in I} \{r(t) - \tfrac{1}{2}q'(t)\} \; > \; 0 \tag{10.2.20}$$

besitzt Problem (10.2.19) nach Satz 8.3 eine eindeutige Lösung $u \in C^2(I)$. Im Falle $p \in C^3(I)$, $q, r, f \in C^2(I)$ ist sogar $u \in C^4(I)$.

Zur Diskretisierung von (10.2.19) sei (zur Vereinfachung) ein äquidistantes Punktgitter

$$a = t_0 < \ldots t_n < \ldots < t_{N+1} = b, \quad I_n := [t_{n-1}, t_n], \quad t_n - t_{n-1} = h = \frac{b-a}{N+1},$$

zugrunde gelegt. Das Differenzenanalogon von Problem (10.2.19) lautet dann

$$\begin{aligned} &L_h y_n := -\Delta_{h/2}[p_n \Delta_{h/2} y_n] + q_n \Delta_h y_n + r_n y_n = f_n\,, \quad 1 \leq n \leq N\,, \\ &\quad y_0 = \alpha\,, \quad y_{N+1} = \beta, \end{aligned} \tag{10.2.21}$$

mit dem *zentralen* Differenzenquotienten 2-ter Ordnung

$$\Delta_h y(t) = \frac{y(t+h) - y(t-h)}{2h}$$

und der Schreibweise $g_n = g(t_n)$ für die Gitterwerte einer stetigen Funktion.

Dies ist äquivalent zu einem linearen $(N+2) \times (N+2)$-Gleichungssystem für den Gittervektor $\bar{y}^h = (y_0, y_1, \ldots, y_N, y_{N+1})^T$,

$$\bar{A}_h \bar{y}^h = \bar{b}^h, \tag{10.2.22}$$

wobei die Matrix $\bar{A}_h$ und die rechte Seite $\bar{b}_h$ aus den Gleichungen $y_0 = \alpha$, $y_{N+1} = \beta$ und

$$-[p_{n-1/2}y_{n-1} - (p_{n-1/2} + p_{n+1/2})y_n + p_{n+1/2}y_{n+1}] + \tfrac{1}{2}hq_n(y_{n+1} - y_{n-1}) + \\ + h^2 r_n y_n = h^2 f_n\,, \quad 1 \le n \le N\,,$$

abgelesen werden können. Durch Elimination der bekannten Randwerte $y_0 = \alpha$, $y_{N+1} = \beta$ wird dieses System auf ein $N \times N$-System

$$A_h y_h = b_h \tag{10.2.23}$$

für den Vektor $y^h = (y_1, \ldots, y_N)^T$ reduziert mit der $N \times N$-Matrix

$$A_h = \frac{1}{h^2}\begin{bmatrix} p_{1/2} + p_{3/2} + h^2 r_1 & -p_{3/2} + \frac{1}{2}hq_1 & & \\ \ddots & \ddots & \ddots & \\ -p_{n-1/2} - \frac{1}{2}hq_n & p_{n-1/2} + p_{n+1/2} + h^2 r_n & -p_{n+1/2} + \frac{1}{2}hq_n & \\ & \ddots & \ddots & \ddots \\ & & -p_{N-1/2} - \frac{1}{2}hq_N & p_{N-1/2} + p_{N+1/2} + h^2 r_N \end{bmatrix}$$

und dem N-Vektor

$$b^h = (f_1 + h^{-2}p_{1/2}\alpha + \tfrac{1}{2}h^{-1}q_1\alpha,\, f_2, \ldots, f_{N-1}, f_N + h^{-2}p_{N+1/2}\beta - \tfrac{1}{2}h^{-1}q_N\beta)^T.$$

Die Genauigkeit dieser Differenzenapproximation wird wieder mit Hilfe des Abschneidefehlers

$$\tau_n := (L_h u^h)_n - f_n, \quad 1 \le n \le N,$$

für die Gitterfunktion $u^h := (u_0, \ldots, u_{N+1})^T$ beschrieben. Für den zentralen Differenzenquotienten einer Funktion $z \in C^3(I)$ gilt

$$\Delta_h z(t) = z'(t) + \tfrac{1}{6}h^2 z'''(\xi_t), \quad \xi_t \in [t-h, t+h].$$

Damit folgt für $u \in C^4(I)$:

$$\begin{aligned}(L_h u^h)_n - f_n &= -\Delta_{h/2}[p_n \Delta_{h/2} u_n] + q_n \Delta_h u_n + r_n u_n - f_n \\ &= -\Delta_{h/2}[p_n u_n' + \tfrac{1}{24}h^2 p_n u'''(\xi_n)] + q_n u_n' + \tfrac{1}{6}h^2 q_n u'''(\eta_n) + r_n u_n - f_n \\ &= -[pu']_n' - \tfrac{1}{24}h^2 (pu')'''(\zeta_n) - \tfrac{1}{24}h^2 \Delta_{h/2}[p_n u'''(\xi_n)] \\ &\quad + \tfrac{1}{6}h^2 q_n u'''(\eta_n) + r_n u_n - f_n \\ &= \underbrace{-[pu']_n' + q_n u_n' + r_n u_n - f_n}_{=0} + h^2 O\big(\max_{I_n}|u^{(iv)}| + \max_{I_n}|u'''|\big),\end{aligned}$$

d. h.: Die obige Differenzenapproximation ist von *2. Ordnung* in h. Aus der Darstellung des Abschneidefehlers folgt, dass die Differenzenapproximation *exakt* ist für quadratische Polynome (im Fall $q \equiv 0$ sogar für kubische Polynome). Dies wird im Folgenden an entscheidender Stelle für den Nachweis der Konvergenz des Verfahrens verwendet werden.

Der Nachweis der Konvergenz erfolgt nun auf analogem Wege wie vorher bei den Differenzenapproximationen von Systemen 1. Ordnung. Grundlage ist der Nachweis der Stabilität der Differenzenapproximation im Sinne, dass

$$\max_{1\le n\le N} |y_n^h| \le K\big\{|y_0^h| + |y_{N+1}^h| + \max_{1\le n\le N} |(L_h y^h)_n|\big\}\,, \tag{10.2.24}$$

für jede Gitterfunktion $y^h = (y_n^h)_{0\le n\le N+1}$ mit einer h-unabhängigen Konstante K. Für die Fehlerfunktion $e^h := y^h - u^h$ gilt nun definitionsgemäß

$$L_h e_n^h = f_n - L_h u_n^h = -\tau_n^h,\ \ n = 1,\dots,N, \quad e_0 = e_{N+1} = 0,$$

woraus mit der Stabilitätsungleichung (10.2.24) direkt die Konvergenz des Verfahrens sowie eine optimale a priori Fehlerabschätzung folgen:

$$\max_{1\le n\le N} |e_n^h| \le K \max_{1\le n\le N} |\tau_n^h| = O(h^2). \tag{10.2.25}$$

Der schwierige Teil der Analyse ist also wieder der Nachweis der Stabilität. Diese könnte z. B. durch Rückführung auf den allgemeinen Stabilitätssatz für Systeme erster Ordnung gewonnen werden. Wir werden aber hier einen anderen Weg beschreiten, der die speziellen algebraischen Eigenschaften der Diskretisierung ausnutzt und sich auch bei partiellen Differentialgleichungen anwenden lässt.

a) Der symmetrische Fall ($q \equiv 0$):

Wir wollen einige spezielle Eigenschaften der tridiagonalen Matrix $A_h = (a_{ij})_{i,j=1}^N$ ableiten. Zunächst wird der Fall betrachtet, dass auf I gilt: $p > 0$, $q \equiv 0$, $r \ge 0$. Dann ist A_h offensichtlich symmetrisch und hat zusätzlich die Eigenschaften

- *stark diagonal dominant*: Es gilt für mindestens ein $s \in \{1,\dots,N\}$:

$$\sum_{j\ne i} |a_{ij}| \le |a_{ii}|\,, \quad 1 \le i \le N, \quad \sum_{j\ne s} |a_{sj}| < |a_{ss}|.$$

- *irreduzibel*: Zu je zwei Indizes $i, k \in \{1,\dots,n\}$ gibt es eine Folge von Indizes $j_1,\dots,j_m$, so dass $a_{j_1 i} \ne 0, a_{j_2 j_1} \ne 0, \dots, a_{j_m j_{m-1}} \ne 0, a_{j j_m} \ne 0$.

Diese Eigenschaften bedeuten, dass die Matrix A_h dem sog. „schwachen Zeilensummenkriterium“ genügt (hinreichend für die Konvergenz des Jacobi-Verfahrens). Die Lösbarkeit des Gleichungssystems (10.2.22) wird dann durch folgenden Hilfssatz sichergestellt:

Hilfssatz 10.2 (Diagonal-dominante Matrizen): *Für eine stark diagonal-dominante, irreduzible Matrix $A \in R^{N\times N}$ gilt:*

i) A ist regulär.

ii) Im Falle $A = A^T$ und $a_{ii} > 0$ $(i = 1,\dots,N)$ ist A positiv definit.

iii) Im Falle $a_{ii} > 0$ und $a_{ij} \le 0$ für $i \ne j$ $(i,j = 1,\dots,N)$ ist A eine sog. M-Matrix, d. h.: $A^{-1} \ge 0$ (elementweise).

Beweis: Wir benötigen einige Vorbereitungen. Die Irreduzibilität von A bedingt

$$\sum_{j=1}^{N} |a_{ij}| > 0\,, \quad i = 1, \ldots, N\,,$$

und die Diagonaldominanz dann auch $|a_{ii}| > 0$, $i = 1, \ldots, N$. Wir zerlegen A gemäß

$$A = \underbrace{\begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{N,N} \end{bmatrix}}_{=:\,D} + \underbrace{\begin{bmatrix} 0 & & 0 \\ & \ddots & \\ a_{ij} & & 0 \end{bmatrix}}_{=:\,L} + \underbrace{\begin{bmatrix} 0 & & a_{ij} \\ & \ddots & \\ 0 & & 0 \end{bmatrix}}_{=:\,R}$$

und bilden die sog. *Jacobi-Matrix* $J := -D^{-1}(L + R)$. Wir wollen zunächst zeigen, dass $\mathrm{spr}(J) < 1$ ist. Seien $\mu \in \mathbb{C}$ irgendein Eigenwert von J und $w \in \mathbb{C}^{N-1}$ ein zugehöriger Eigenvektor mit

$$|w_r| = \max_{1 \le i \le N} |w_i| := \|w\|_\infty = 1\,.$$

Es gilt dann

$$\mu w_i = a_{ii}^{-1} \sum_{j \neq i} a_{ij} w_j, \quad 1 \le i \le N. \tag{10.2.26}$$

Hieraus folgt zunächst mit Hilfe der Diagonaldominanz

$$|\mu| = |\mu w_r| \le |a_{rr}|^{-1} \sum_{j \neq r} |a_{rj}|\, \|w\|_\infty \le 1\,.$$

Angenommen, es ist $|\mu| = 1$. Wegen der Irreduzibilität existieren Indizes $i_1, \ldots, i_m$, so dass $a_{i_1 s} \neq 0, \ldots, a_{r i_m} \neq 0$. Durch sukzessive Anwendung von (10.2.26) folgt damit der Widerspruch

$$\begin{aligned}
|w_s| &= |\mu w_s| \le |a_{ss}|^{-1} \sum_{j \neq s} |a_{sj}|\, \|w\|_\infty < \|w\|_\infty \\
|w_{i_1}| &= |\mu w_{i_1}| \le |a_{i_1 i_1}|^{-1} \Big\{ \sum_{j \neq i_1, s} |a_{i_1 j}|\, \|w\|_\infty + \underbrace{|a_{i_1 s}|}_{\neq 0} |w_s| \Big\} < \|w\|_\infty \\
&\vdots \\
|w_{i_m}| &= |\mu w_{i_m}| \le |a_{i_m i_m}|^{-1} \Big\{ \sum_{j \neq i_m, i_{m-1}} |a_{i_m j}|\, \|w\|_\infty + \underbrace{|a_{i_m i_{m-1}}|}_{\neq 0} |w_{i_{m-1}}| \Big\} < \|w\|_\infty \\
\|w\|_\infty = |w_r| &= |\mu w_r| \le |a_{rr}|^{-1} \Big\{ \sum_{j \neq r, i_m} |a_{rj}|\, \|w\|_\infty + \underbrace{|a_{r i_m}|}_{\neq 0} |w_{i_m}| \Big\} < \|w\|_\infty\,.
\end{aligned}$$

Also ist $\mathrm{spr}(J) < 1$.

i) Wäre A irregulär, so gäbe es ein $w \in \mathbb{R}^N$ mit $w \neq 0$ und $Aw = 0$. Aus der Identität

$$0 = Aw = Dw + (L+R)w = D(w - Jw)$$

folgte dann, dass $\mu = 1$ Eigenwert von J wäre im Widerspruch zu $\mathrm{spr}(J) < 1$.

ii) Seien $\lambda \in \mathbb{C}$ irgendein Eigenwert von A und $w \in C^N$ ein zugehöriger Eigenvektor mit $|w_r| = \|w\|_\infty = 1$. Dann gilt

$$\lambda w_r = \sum_{j=1}^{N} a_{rj} w_j = \sum_{j \neq r}^{N} a_{rj} w_j + a_{rr} w_r$$

bzw.

$$|\lambda - a_{rr}| \underbrace{|w_r|}_{=1} \leq \sum_{j \neq r}^{N} |a_{rj}| \underbrace{|w_j|}_{\leq 1} \leq |a_{rr}| .$$

Im Falle $a_{rr} > 0$ folgt hieraus zunächst $Re\lambda \geq 0$ und dann wegen der Regularität von A auch $Re\lambda > 0$. Für symmetrisches A ist $\lambda \in \mathbb{R}$ und somit $\lambda > 0$, d. h.: A ist positiv definit.

iii) Im Falle $a_{ii} > 0$ und $a_{ij} \leq 0, i \neq j$ $(i, j = 1, \ldots, N)$ ist $D \geq 0$ und $L + R \leq 0$, d. h.: $J \geq 0$ (elementweise). Wegen $\mathrm{spr}(J) < 1$ konvergiert die Reihe

$$0 \leq \sum_{k=0}^{\infty} J^k = (I - J)^{-1}.$$

Hieraus folgt schließlich wegen

$$(I - J)^{-1} = (I + D^{-1}[L + R])^{-1} = (D^{-1}[D + L + R])^{-1} = A^{-1} D$$

und $D \geq 0$ notwendig $A^{-1} \geq 0$. Q.E.D.

Stark diagonal dominante und irreduzible Matrizen haben besonders angenehme Eigenschaften. Zum einen besitzen sie stets Darstellungen $A = LR$ als das Produkt von Dreiecksmatrizen, welche sich mit Hilfe des Gaußschen Eliminationsverfahrens (ohne Pivotierung!) berechnen lassen, zum anderen konvergieren für sie die einfachen Iterationsverfahren wie z. B. das Jacobi-Verfahren, das Gauß-Seidel-Verfahren oder das SOR-Verfahren. Die zusätzliche Eigenschaft von A_h, M-Matrix zu sein, erlaubt es schließlich, die angestrebte Stabilität der Differenzenapproximation zu zeigen.

Satz 10.3 (Stabilität von Differenzenverfahren): *Im Falle $q = 0$, $r \geq 0$ auf I ist die Differenzenapproximation stabil, d. h.: Für jede Gitterfunktion $\bar{z}^h = (z_n^h)_{0 \leq n \leq N+1}$ gilt die Stabilitätsabschätzung*

$$\max_{1 \leq n \leq N} |z_n^h| \leq K \left\{ |z_0^h| + |z_{N+1}^h| + \max_{1 \leq n \leq N} |L_h z_n^h| \right\},$$

mit einer h-unabhängigen Konstante K. Ferner gilt die Konvergenzaussage

$$\max_{t_n \in I} |u(t_n) - y_n^h| = O(h^2) \quad (h \to 0) .$$

Beweis: Wir betrachten der Übersichtlichkeit halber nur den Spezialfall $p = 1$. Sei $(z_n^h)_{0\le n\le N+1}$ eine beliebige Gitterfunktion. Die mit der linearen Funktion

$$l(t) = \frac{(b-t)z_0^h + (t-a)z_{N+1}^h}{b-a},$$

gebildete Gitterfunktion $\tilde{z}^h = z^h - l^h$ hat dann homogene Randwerte $\tilde{z}_0^h = \tilde{z}_{N+1}^h = 0$. Sei

$$M_h := \max_{1\le n\le N} |L_h \tilde{z}_n^h|.$$

Für die quadratische Funktionen $w(t) := \frac{1}{2}M_h(b-t)(t-a) > 0$ gilt dann wegen der Exaktheit der Differenzenapproximation für quadratische Polynome

$$(L_h w^h)_n = Lw(t_n) = M_h + r_n w_n^h \ge M_h, \quad 1 \le n \le N,$$

bzw. $A_h w^h \ge M_h$ komponentenweise mit $w^h := (w_n)_{0\le n\le N+1}$. Damit ist ebenfalls komponentenweise

$$A_h[\pm\tilde{z}^h - w^h]_n = \pm A_h \tilde{z}_n^h - A_h w_n^h \le M_h - M_h = 0.$$

Wegen $A_h^{-1} \ge 0$ ergibt sich dann $\pm\tilde{z}^h - w^h \le 0$ bzw.

$$\max_{1\le n\le N} |\tilde{z}_n^h| \le \max_{1\le n\le N} |w_n^h| \le \tfrac{1}{2}(b-a)^2 M_h\,.$$

Mit der Definition von $\tilde{z}_n^h$ und M_h folgt mit $K_0 = \frac{1}{2}(b-a)^2$:

$$\begin{aligned}\max_{1\le n\le N} |z_n^h| &\le \max_{1\le n\le N} |l_n^h| + K_0 \max_{1\le n\le N} |(L_h\tilde{z}^h)_n| \\ &\le |z_0^h| + |z_N^h| + K_0 \max_{1\le n\le N} |(L_h z^h)_n| + K_0 \max_{1\le n\le N} |(L_h l^h)_n|\end{aligned}$$

und weiter wegen $|L_h l_n^h| = |Ll_n| = |r_n l_n| \le \max_{1\le n\le N} |r_n|\{|z_0^h| + |z_N^h|\}$:

$$\max_{1\le n\le N} |z_n^h| \le K\big\{|z_0^h| + |z_N^h| + \max_{1\le n\le N} |(L_h z^h)_n|\big\}$$

mit $K := K_0(1 + \max_{1\le n\le N} |r_n|)$. Q.E.D.

b) Der unsymmetrische Fall ($q \not\equiv 0$):

Wir betrachten nun den Fall $q \not\equiv 0$. Die Systemmatrix A_h erhält dann die Gestalt

$$A_h = \frac{1}{h^2}\begin{bmatrix} p_{1/2}+p_{3/2}+h^2 r_1 & -p_{3/2}+\frac{1}{2}hq_1 & & \\ \ddots & \ddots & \ddots & \\ -p_{n-1/2}-\frac{1}{2}hq_n & p_{n-1/2}+p_{n+1/2}+h^2 r_n & -p_{n+1/2}+\frac{1}{2}hq_n & \\ \ddots & \ddots & \ddots & \\ & -p_{N-1/2}-\frac{1}{2}hq_N & p_{N-1/2}+p_{N+1/2}+h^2 r_N & \end{bmatrix}.$$

Diese Matrix ist offensichtlich unsymmetrisch und nur unter der Bedingung

$$h \le 2 \min_{1\le n\le N} \left\{ \frac{\min\{p_{n-1/2}, p_{n+1/2}\}}{|q_n|} \right\} \tag{10.2.27}$$

diagonal dominant. In diesem Fall genügt A_h auch der Vorzeichenbedingung und ist damit eine M-Matrix. Die auf dieser Eigenschaft aufbauende obige Konvergenzanalyse kann mit etwas mehr Aufwand hierfür übertragen werden, und wir erhalten wieder die Lösbarkeit der Differenzengleichungen sowie die Konvergenz des Verfahrens mit der Fehlerordnung $O(h^2)$. Der Fall $|q_n| \gg |p_n|$ ist aber kritisch, da oft die Gitterweite h aus Kapazitätsgründen nicht gemäß (10.2.27) klein genug gewählt werden kann. Die in diesem Zusammenhang auftretenden Phänomene wollen wir zunächst anhand eines einfachen Modellproblems diskutieren.

Beispiel 10.3: Wir setzen $I = [0,1]$ sowie $f \equiv 0\,,\ q \equiv 1\,, r \equiv 0$ und $0 < p \equiv: \epsilon << 1$. Die *singulär gestörte* RWA

$$L^\epsilon u(t) := -\epsilon u''(t) + u'(t) = 0\,, \quad x \in I, \qquad u(0) = 1,\ u(1) = 0\,,$$

hat die (eindeutige) Lösung

$$u^\epsilon(t) = \frac{e^{1/\epsilon} - e^{t/\epsilon}}{e^{1/\epsilon} - 1}.$$

Im betrachteten Fall $\epsilon \ll 1$ nennt man dies eine *Grenzschichtlösung* (s. Bild), denn für $t = 1 - \delta$ und $\delta > \epsilon$ ist

$$u^\epsilon(1-\delta) = \frac{e^{1/\epsilon}}{e^{1/\epsilon} - 1}\left(1 - e^{-\delta/\epsilon}\right) \approx 1\,, \qquad \max_I |u^{\epsilon\prime\prime}| \approx \epsilon^{-2}\,.$$

Für $\epsilon = 0$ ergibt sich die Grenzlösung $u^0 \equiv 1$, welche die Randbedingung bei $t = 1$ nicht erfüllt.

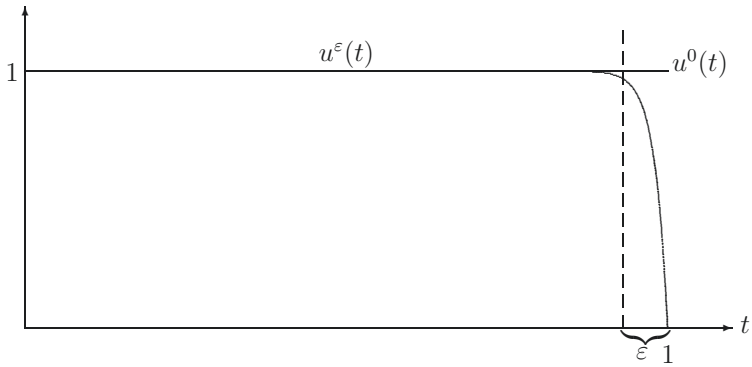

Abbildung 10.1: Lösung des singulär gestörten Problems für $\epsilon = .1$

Die Approximation dieses Problems mit dem obigen Differenzenverfahren zur (äquidistanten) Schrittweite $h = 1/(N+1)$ ergibt

$$-(\epsilon + \tfrac{1}{2}h)y_{n-1} + 2\epsilon y_n - (\epsilon - \tfrac{1}{2}h)y_{n+1} = 0\,, \quad 1 \le n \le N\,, \qquad y_0 = 1,\ y_{N+1} = 0\,.$$

Die zugehörige Systemmatrix ist offensichtlich diagonal dominant und von nicht-negativen Typ für $h \le 2\varepsilon$. Diese Bedingung ist aber für sehr kleines $\varepsilon \ll 1$ in der Praxis nur schwer erfüllbar.

Für die Lösung dieser Differenzengleichungen machen wir wieder einen Ansatz der Form $y_n = \lambda^n$. Die möglichen Werte für λ sind gerade die Wurzeln $\lambda_\pm$ der quadratischen Gleichung

$$\lambda^2 + \frac{2\epsilon}{\frac{1}{2}h - \epsilon}\lambda - \frac{\frac{1}{2}h + \epsilon}{\frac{1}{2}h - \epsilon} = 0\,.$$

Berücksichtigung der Randbedingungen $y_0 = 1$ und $y_N = 0$ in dem Ansatz

$$y_n = c_+\lambda_+^n + c_-\lambda_-^n$$

ergibt für die Koeffizienten die Beziehungen $c_+ + c_- = 1$, $c_+\lambda_+^{N+1} + c_-\lambda_-^{N+1} = 0$ und folglich

$$c_- = \frac{\lambda_+^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}}\,, \qquad c_+ = 1 - \frac{\lambda_+^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}} = -\frac{\lambda_-^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}}\,.$$

Die Lösung hat also die Gestalt

$$y_n = \frac{\lambda_+^{N+1}\lambda_-^n - \lambda_-^{N+1}\lambda_+^n}{\lambda_+^{N+1} - \lambda_-^{N+1}}\,. \tag{10.2.28}$$

Im vorliegenden Fall sind die Wurzeln gegeben durch

$$\lambda_{+,-} = \frac{-\epsilon \pm \sqrt{\epsilon^2 + (\frac{1}{2}h + \epsilon)(\frac{1}{2}h - \epsilon)}}{\frac{1}{2}h - \epsilon} = \frac{\epsilon \mp \frac{1}{2}h}{\epsilon - \frac{1}{2}h}\,, \qquad \lambda_+ = 1,\ \ \lambda_- = \frac{\epsilon + \frac{1}{2}h}{\epsilon - \frac{1}{2}h}\,.$$

Für $\varepsilon \ll \frac{1}{2}h$ ist $\lambda_- \sim -1$. In diesem Fall wird also eine oszillierende Lösung erzeugt,

$$y_n = \frac{\lambda_-^n - \lambda_-^{N+1}}{1 - \lambda_-^{N+1}},$$

welche qualitativ nicht den richtigen Lösungsverlauf wiedergibt.

Zur Unterdrückung dieses Defekts gibt es verschiedene Strategien, die im Folgenden skizziert werden.

i) Upwind-Diskretisierung: Zunächst kann der Term erster Ordnung $u'(t)$ in der Differentialgleichung statt mit dem zentralen mit einem der einseitigen Differenzenquotienten

$$\Delta_h^+ u(t) = \frac{u(t+h) - u(t)}{h}\,, \quad \Delta_h^- u(t) = \frac{u(t) - u(t-h)}{h}$$

approximiert werden. Bei Wahl des *rückwärtigen* Differenzenquotienten Δ_h^- wird dem physikalischen Vorgang eines Informationstransports in positive t-Richtung Rechnung getragen (vergl. die Form der Grenzlösung $u^0(t)$). Dies führt auf die Differenzengleichungen

$$(-\epsilon + h)y_{n-1} + (2\epsilon + h)y_n - \epsilon y_{n+1} = 0\,.$$

Die zugehörige Matrix A_h ist dann für beliebiges $h > 0$ wieder diagonal dominant und genügt der Vorzeichenbedingung, d. h.: Sie ist eine M-Matrix. Der Lösungsansatz $y_n = \lambda^n$ führt in diesem Fall auf die Gleichung

$$\lambda^2 - \frac{2\epsilon + h}{\epsilon}\lambda + \frac{\epsilon + h}{\epsilon} = 0\,,$$

mit den Wurzeln

$$\lambda_{+,-} = \frac{2\varepsilon + h}{2\varepsilon} \pm \sqrt{(2\varepsilon + h)^2 - 4\varepsilon(\varepsilon + h)} = \frac{2\varepsilon + h \pm h}{2\epsilon}, \qquad \lambda_+ = \frac{\varepsilon + h}{\varepsilon},\ \ \lambda_- = 1\,.$$

Die kritische Wurzel λ_+ ist hier stets positiv, so dass in der diskreten Lösung gemäß (10.2.28),

$$y_n = \frac{\lambda_+^{N+1} - \lambda_+^n}{\lambda_+^{N+1} - 1},$$

keine ungewollten Oszillationen in der Näherungslösung entstehen. Diese spezielle Art der einseitigen Diskretisierung des Terms $u'(t)$ nennt man „Rückwärtsdiskretisierung“ oder auch englisch „upwind discretization“. Da der verwendete einseitige Differenzenquotienten aber nur die Approximationsordnung $O(h)$ hat, ist auch das Gesamtverfahren nur von erster Ordnung genau. Dies limitiert die Approximationsgenauigkeit in Bereichen, in denen die Lösung glatt ist, selbst wenn die Gitterweite in der Grenzschicht ausreichend fein gemäß $h \approx \epsilon$ gewählt wird.

ii) Künstliche (numerische) Diffusion: Unter Beibehaltung der zentralen Diskretisierung des Terms $u'(t)$ wird der *Diffusionskoeffizient* ϵ für feste Schrittweite h auf einen größeren Wert $\hat\epsilon := \epsilon + \delta h$ gesetzt. Dies führt auf die Differenzengleichungen

$$-(\hat\epsilon + h/2)y_{n-1} + 2\hat\epsilon y_n - (\hat\epsilon - h/2)y_{n+1} = 0\,, \quad 1 \le n \le N\,.$$

Für die zugehörige Lösung erhält man wieder durch einen Potenzansatz die Darstellung

$$y_n = \frac{\lambda_+^{N+1} - \lambda_+^n}{\lambda_+^{N+1} - 1}, \qquad \lambda_+ = \frac{\hat\epsilon + h/2}{\hat\epsilon - h/2}\,.$$

Offenbar ist in diesem Fall $\lambda_+ > 0$ für $\epsilon + \delta h > h/2$, d. h. für die Wahl $\delta \ge 1/2$. Mit diesem Ansatz erhält man also ebenfalls wieder eine M-Matrix und somit eine stabile Diskretisierung. Allerdings wird nun die Grenzschicht stark verschmiert auf das Intervall $[1 - \hat\epsilon, 1] \approx [1 - h, 1]$, und die globale Approximationsgüte ist aufgrund der Störung des Differentialoperators ebenfalls lediglich $O(h)$.

Beim allgemeinen Sturm-Liouville-Problem mit variablem $q(t)$ muss das „Upwinding“ abhängig vom Vorzeichen von q_n angesetzt werden. Die einseitigen Differenzenquotienten werden gemäß der folgenden Schaltvorschrift angesetzt:

$$\mathrm{sgn}(q_n) = \begin{cases} +1 & : \quad \Delta_h^- \\ -1 & : \quad \Delta_h^+\,. \end{cases}$$

Dies führt dann wieder auf eine für alle $h > 0$ diagonal dominante Matrix A_h.

Anhand des obigen einfachen Beispiels haben wir gesehen, dass bei *singulär gestörten* Problemen die einfachen Dämpfungsstrategien *Rückwärtsdiskretisierung* oder *künstliche Diffusion* zwar auf stabile Diskretisierungen führen, die Approximationsordnung aber auf $O(h)$ reduzieren. Die Frage nach einer sicheren Dämpfungsstrategie höherer Ordnung zur Diskretisierung von *Transporttermen* ist noch nicht vollständig geklärt. Versuche in diese Richtung bedienen sich z. B. einseitiger Differenzenquotienten höherer Ordnung (beim *Upwinding*) oder künstlicher Diffusionsterme der Form $\delta h^2 u^{(iv)}$. Allerdings kann die starke M-Matrixeigenschaft nur mit Diskretisierungen erster Ordnung erreicht werden. Einen anderen Ansatz werden wir im nächsten Abschnitt im Zusammenhang mit Galerkin-Verfahren für Sturm-Liouville-Probleme kennenlernen.

10.2.1 Konditionierung

Zum Abschluss dieses Abschnittes wollen wir noch die Konditionierung der Systemmatrizen A_h in Abhängigkeit von der Gitterweite h untersuchen. Dazu betrachten wir den einfachen Modellfall $p \equiv 1, q \equiv 0, r \equiv 0$, d. h.:

$$A_h = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}, \qquad h = \frac{1}{N+1}.$$

Die Eigenwerte und zugehörigen Eigenvektoren dieser Matrix sind, wie man leicht nachrechnet:

$$\lambda_k = \frac{1}{h^2}\{2 - 2\cos(kh\pi)\}, \quad w^k = \big(\sin(ikh\pi)\big)_{i=1}^{N}, \quad k = 1, \ldots, N.$$

Also ist

$$\begin{aligned} \lambda_{\max} &= \frac{1}{h^2}\{2 - 2\cos((1-h)\pi)\} = \frac{4}{h^2} + \mathcal{O}(1), \\ \lambda_{\min} &= \frac{1}{h^2}\{2 - 2\cos(h\pi)\} = \frac{1}{h^2}\{2 - 2(1 - \tfrac{1}{2}h^2\pi^2) + \mathcal{O}(h^4)\} = \pi^2 + \mathcal{O}(h^2), \end{aligned}$$

und folglich

$$\operatorname{cond}_2(A_h) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{4}{\pi^2 h^2} + \mathcal{O}(1).$$

Die Konditionierung der Systemmatrix A_h wird also mit kleiner werdender Gitterweite, d. h. mit zunehmender Diskretisierungsgenauigkeit, wie $\mathcal{O}(h^{-2})$ schlechter. Dabei wird der Exponent -2 offensichtlich durch die Ordnung des Differentialoperators L bestimmt (und nicht etwa durch die Ordnung der Differenzenapproximation!).

10.3 Übungsaufgaben

Aufgabe 10.1: Zur Lösung der d-dimensionalen linearen RWA 1. Ordnung

$$u'(t) - A(t)u(t) = f(t), \quad t \in [a,b], \qquad B_a u(a) + B_b u(b) = g,$$

kann die Polygonzugmethode verwendet werden (a) als direktes Differenzenschema, oder (b) als AWA-Löser im Zuge des Schießverfahrens. Beide Ansätze liefern Näherungen der Ordnung $\mathcal{O}(h)$. Man vergleiche den jeweiligen numerischen Aufwand, d. h. die Anzahl der Auswertungen von $A(t)$ und $f(t)$, bei gleicher Gitterweite h, sowie den Aufwand zur Lösung der auftretenden Gleichungssysteme.

Aufgabe 10.2: Man betrachte die Trapezregel zur Diskretisierung der RWA 2. Ordnung

$$-u''(t) + u'(t) = 1, \quad t \in [0,1], \quad u(0) = u(1) = 0,$$

auf einem äquidistanten Gitter mit Gitterweite $h = 1/N$. Ist die RWA überhaupt eindeutig lösbar? Wie lautet das zugehörige Gleichungssystem und von welcher Ordnung ist dieses Differenzenschema?

Aufgabe 10.3: Man betrachte das Sturm-Liouville-Problem

$$-u''(t) + 100u'(t) = 1, \quad t \in [0,1], \quad u(0) = u(1) = 0.$$

a) Ist die RWA eindeutig lösbar?

b) Man approximiere die RWA auf einem äquidistanten Gitter mit Gitterweite $h = 1/N$ unter Verwendung des zentralen Differenzenquotienten zweiter Ordnung

$$u''(t) \approx h^{-2}\big\{u(t+h) - 2u(t) + u(t-h)\big\},$$

für die zweiten Ableitungen und des zentralen Differenzenquotienten erster Ordnung

$$u'(t) \approx (2h)^{-1}\big\{u(t+h) - u(t-h)\big\}.$$

für die erste Ableitung. Wie lauten die zugehörigen Gleichungssysteme? Unter welcher Bedingung an die Gitterweite h sind die Systemmatrizen (strikt) diagonal dominant?

c) Man verwende zur Approximation der ersten Ableitung anstelle des zentralen den rückwärtsgenommenen Differenzenquotienten erster Ordnung

$$u'(t) \approx h^{-1}\big\{u(t) - u(t-h)\big\}$$

Unter welcher Bedingung an die Gitterweite h ist die zugehörige Systemmatrix (strikt) diagonal dominant?

Aufgabe 10.4: Man betrachte das „singulär gestörte“ Strurm-Liouville-Problem

$$L^\epsilon u(t) := -\epsilon u''(t) + u'(t) = 0\,, \quad t \in [0,1], \qquad u(0) = 1,\ u(1) = 0\,,$$

mit einem $\epsilon \ll 1$. Ein zum „upwinding“ alternativer Ansatz zur Umgehung der Schrittweitenbedingung $h \leq 2\epsilon$ ist die Verwendung „künstlicher Diffusion“, d. h. die Ersetzung von ϵ durch $\hat{\epsilon} := \epsilon + \frac{1}{2}h$ unter Beibehaltung der Approximation des Ableitungsterms $u'(t)$ durch den zentralen Differenzenquotienten zweiter Ordnung. Dies führt auf die Differenzengleichung

$$-(\hat{\epsilon} + \tfrac{1}{2}h)y_{n-1} + 2\hat{\epsilon}y_n - (\hat{\epsilon} - \tfrac{1}{2}h)y_{n+1} = 0\,, \quad 1 \leq n \leq N\,.$$

Mit Hilfe des Potenzansatzes aus der Vorlesung bestimme man deren Lösung und zeige, dass diese wie beim „upwinding“ keine oszillierende Komponente besitzt. Mit diesem Ansatz erhält man ebenfalls wieder eine M-Matrix und somit eine stabile Diskretisierung. Allerdings ist deren Konvergenzordnung aufgrund der Störung des Differentialoperators ebenfalls lediglich $O(h)$.