Numerik für partielle Differentialgleichungen: instationäre Probleme

Michael Feischl und Markus Melenk

7. Mai 2025

Inhaltsverzeichnis

Kapitel 1

Eigenwertprobleme

Ausgangspunkt: zahlreiche technische und naturwissenschaftliche Probleme führen auf Eigenwertproblem, z.B. in der Strukturmechanik ("Eigenfrequenzen eines elastischen Körpers"), Quantenmechanik,...

Ziel: numerische Approximation der Eigenwerte und Eigenfunktionen

Modellproblem: Finde Funktion u und $\lambda \in \mathbb{R}$, so dass

$$-\Delta u = \lambda u, \quad \text{auf } \Omega, \qquad u|_{\partial\Omega} = 0. \tag{1.1}$$

Bemerkung 1.1 Das EWP (1.1) hat physikalische Bedeutung: Die Eigenwerte λ in (1.1) sind tatsächlich positiv, und die Werte $\sqrt{\lambda}$ entsprechen den Eigenfrequenzen einer eingespannten Membran. ∎

Das Problem (1.1) werden wir variationell verstehen:

$$\text{Finde } (u,\lambda) \in H_0^1(\Omega) \setminus \{0\} \times \mathbb{R} \text{ s.d.} \quad \int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega uv \qquad \forall v \in H_0^1(\Omega). \tag{1.2}$$

Beispiel 1.2 Sei $\Omega = (0,\pi)$ und betrachte

$$-u'' = \lambda u \quad \text{auf } \Omega, \qquad u(0) = u(\pi) = 0.$$

Die allg. Lösung der Differentialgleichung $-u'' - \lambda u = 0$ ist $u(x) = C_1 \sin(\sqrt{\lambda}x) + C_2 \cos(\sqrt{\lambda}x)$. Damit es nichttriviale Lösungen gibt, muss $\sqrt{\lambda} = n$, $n \in \mathbb{N}$ sein. Damit sind die Eigenpaare (u_n, λ_n)

$$u_n(x) = \sin(\sqrt{\lambda_n}x), \qquad \lambda_n = n^2, \qquad n = 1, 2, \ldots,$$

Beobachtung/Nachrechnen: die Eigenfunktionen erfüllen 2 Orthogonalitätsbeziehungen:

$$\begin{aligned} (u_n, u_m)_{L^2(\Omega)} &= 0 \qquad n \neq m, \\ (u_n, u_m)_{H_0^1(\Omega)} := \int_\Omega u_n' u_m' &= 0 \qquad n \neq m. \end{aligned}$$

Weiters ist $(u_n)_n$ sogar eine Orthogonalbasis von $L^2(\Omega)$ (und, wie wir später sehen werden, auch eine von $H_0^1(\Omega)$.) ∎

Wir betrachten ein etwas allgemeineres Setting im vorliegenden Kapitel:

- V, H Hilberträume (über $\mathbb{R}$)
- $V \subset H$ dicht, kompakt
- $a : V \times V \to \mathbb{R}$ symmetrisch, stetig, bilinear, koerziv
- $(\cdot,\cdot)_H$ = Skalarprodukt auf H

- $a(\cdot,\cdot)$ erzeugt ein Innenprodukt auf V, welches äquivalent zum Innenprodukt auf V ist.

Wir betrachten:

$$\text{Finde } (u,\lambda) \in V \setminus \{0\} \times \mathbb{R} \text{ s.d.} \qquad a(u,v) = \lambda (u,v)_H \qquad \forall v \in V. \tag{1.3}$$

Für die Lösbarkeitstheorie verwenden wir die Theorie kompakter Operatoren. Sei $T : H \to V$ der Lösungsoperator für die Variationsaufgabe:

$$\text{Finde } u \in V \text{ s.d.} \qquad a(u,v) = (f,v)_H \qquad \forall v \in V, \tag{1.4}$$

d.h. $Tf \in V$ ist charakterisiert durch

$$a(Tf,v) = (f,v)_H \qquad \forall v \in V. \tag{1.5}$$

Weil $V \subset H$ können wir T auf V einschränken und als Operator auf V auffassen. Dann gilt:

Satz 1.3 *(i)* $T : V \to V$ *ist kompakt*

(ii) $T : V \to V$ *ist selbstadjungiert bzgl.* $a(\cdot,\cdot)$

(iii) "T *ist positiv auf* V*":* $a(Tf,f) > 0$ *für alle* $f \in V \setminus \{0\}$.

Beweis: *ad (i):* Der Operator $T : V \to V$ ist eine Verkettung der kompakten Einbettung $V \subset H$ mit dem stetigen Operator $T : H \to V$ (Lax-Milgram!):

$$T : V \hookrightarrow H \to V$$

Damit ist $T : V \to V$ kompakt als Verkettung eines stetigen und eines kompakten Operators.
ad (ii): Seien f, $v \in V$. D.g.:

$$a(Tf,v) = (f,v)_H = (v,f)_H = a(Tv,f) \overset{\text{a sym.}}{=} a(f,Tv)$$

ad (iii):

- für $f \in V \setminus \{0\}$ ist $a(Tf,f) = (f,f)_H \geq 0$.
- Weil die Einbettung $V \subset H$ injektiv ist, würde aus $\|f\|_H = 0$ folgen, dass $f = 0$ (als Element von V).

□

Das variationell formulierte EWP (1.3) ist äquivalent zu einem EWP für T:

Lemma 1.4 $(u,\lambda) \in V \setminus \{0\} \times \mathbb{R} \setminus \{0\}$ *löst (1.3)* $\iff$ $(u,\lambda) \in V \setminus \{0\} \times \mathbb{R} \setminus \{0\}$ *löst*

$$Tu = \frac{1}{\lambda} u. \tag{1.6}$$

Beweis: "$\Longrightarrow$": Sei (u,λ) Lösung von (1.3). Dann gilt:

$$a(u,v) = \lambda (u,v)_H = \lambda a(Tu,v) = a(\lambda Tu, v) \qquad \forall v \in V.$$

Weil $a(\cdot,\cdot)$ ein Skalarprodukt ist, folgt

$$u = \lambda Tu.$$

"$\Longleftarrow$": Gelte (1.6). Dann folgt für alle $v \in V$

$$a(\lambda^{-1} u, v) = a(Tu,v) = (u,v)_H,$$

d.h. (1.3). □

Lemma 1.4 besagt, dass die gesuchten Eigenpaare (u,λ) von (1.3) durch die Eigenpaare $(u, 1/\lambda)$ von T gegeben sind. Da T kompakt und selbstadjungiert (bzgl. $a(\cdot,\cdot)$) ist, kann die Spektraltheorie kompakter Operatoren eingesetzt werden. Es gilt:

Satz 1.5 (Spektralsatz für kompakte, selbstadj. Operatoren) *Sei X ein Hilbertraum über $\mathbb{C}$ und $A : X \to X$ kompakt und selbstadjungiert. Dann gilt:*

(i) Das Spektrum $\sigma(A) = \{\mu \in \mathbb{C} \,|\, \mu - A : X \to X$ ist nicht stetig invertierbar$\}$ ist eine abzählbare Menge mit einzig möglichem Häufungspunkt 0. Die Elemente $\mu \in \sigma(A) \setminus \{0\}$ heissen Eigenwerte.

(ii) $\sigma(A) \subset \mathbb{R}$. Ist A nichtnegativ, d.h. $(Ax, x)_X \geq 0$ für alle x, dann ist $\sigma(A) \subset [0, \|A\|_X] \subset [0, \infty)$.

(iii) Für jedes $\mu \in \sigma(A) \setminus \{0\}$ ist $\mathrm{Ker}(\mu - A)$ endlichdimensional. Die Zahl $\dim \mathrm{Ker}(\mu - A) \in \mathbb{N}$ heisst Vielfachheit[1] *des Eigenwertes μ. Der Raum $\mathrm{Ker}(\mu - A)$ heisst* Eigenraum *zum Eigenwert μ.*

(iv) Für μ_1, $\mu_2 \in \sigma(A) \setminus \{0\}$ mit $\mu_1 \neq \mu_2$ und $u_1 \in \mathrm{Ker}(\mu_1 - A)$ und $u_2 \in \mathrm{Ker}(\mu_2 - A)$ gilt $(u_1, u_2)_X = 0$.

(v) X hat eine ONB aus Eigenvektoren von A. Genauer: Sei die abzählbare Menge $\sigma(A) \setminus \{0\}$ als Folge $\mu_1, \mu_2, \ldots,$ geschrieben, wobei jeder Eigenwert gemäss seiner Vielfachheit ggf. mehrfach aufgeführt wird. Dann existiert eine Folge $(e_n)_n \subset X$ mit

$$(e_n, e_m)_X = \delta_{n,m}$$

und $X = \mathrm{Ker}\, A \oplus_X \mathrm{span}\{e_1\} \oplus_X \mathrm{span}\{e_2\} \oplus_X \cdots$.

(vi) Jedes $x \in X$ kann als "Fourierreihe" geschrieben werden:

$$x = a + \sum_{n=1}^{\infty} (x, e_n)_X e_n,$$

wobei $a \in \mathrm{Ker}\, A$. Weiters ist

$$Ax = \sum_{n=1}^{\infty} \mu_n (x, e_n)_X e_n$$

Beweis: Eine reelle Version dieses Satzes wird in den Übungen bewiesen. □

Eine reelle Version von Satz 1.5 kann auch durch "Komplexifizieren" aus der komplexen Version rausgeholt werden:

Übung 1.6 Sei X ein Hilbertraum über $\mathbb{R}$ und $A : X \to X$ ein symmetrischer Operator. Die Komplexifizierung von X ist der Raum

$$\widetilde{X} := \{x + \mathbf{i}y \,|\, x, y \in X\}.$$

Zeigen Sie: $\widetilde{X}$ ist ein Hilbertraum über $\mathbb{C}$, wenn man das Skalarprodukt als

$$((x + \mathbf{i}y), (x' + \mathbf{i}y'))_{\widetilde{X}} := ((x, x')_X + (y, y')_X) + \mathbf{i}\,((y, x')_X - (x, y')_X)$$

wählt. Der Raum X ist in natürlicher Weise in $\widetilde{X}$ eingebettet. Die Komplexifizierung von $\widetilde{A}$ ist entsprechend definiert durch $\widetilde{A}(x + \mathbf{i}y) := Ax + \mathbf{i}Ay$. Zeigen Sie: $\widetilde{A}$ ist selbstadjungiert. Zeigen Sie: falls A kompakt ist, dann ist $\widetilde{A}$ kompakt. ∎

Einschub

Lemma 1.7 *Sei A selbstadjungiert auf dem Hilbertraum X (über $\mathbb{C}$ oder $\mathbb{R}$). Sei R_A der Rayleighquotient:*

$$R_A(x) := \frac{(Ax, x)_X}{(x, x)_X}, \qquad x \in X \setminus \{0\}$$

Dann gilt:

$$\|A\|_X = \sup_x |R_A(x)|$$

Für kompakte Operatoren A wird das Supremum sogar angenommen und der Maximierer ist ein Eigenvektor. Insbesondere ist dann $\|A\|_X$ ein Eigenwert.

[1] genauer: $r_g := \dim \mathrm{Ker}(\mu - A)$ ist die *geometrische* Vielfachheit von μ. Man nennt die *kleinste* Zahl $\alpha \in \mathbb{N}$ mit $\mathrm{Ker}(\mu - A)^\alpha = \mathrm{Ker}(\mu - A)^{\alpha+1}$ den *Azent* von μ. Die *algebraische* Vielfachheit des Eigenwertes μ ist dann definiert als $r_a := \dim \mathrm{Ker}(\mu - A)^\alpha$. Offensichtlich ist $r_a \geq r_g$. Im Fall von selbstadjungierten Operatoren gilt $\alpha = 1$, so dass $r_a = r_g$. Denn: wäre $\alpha \geq 2$, so gäbe es ein $x \neq 0$ mit $x \in \mathrm{Ker}(\mu - A)^\alpha$ und $x \notin \mathrm{Ker}(\mu - A)^{\alpha-1}$. Wegen $\alpha \geq 2$ können wir dann sinnvoll $(\mu - A)^{\alpha-2}x$ betrachten, und es folgt mit der Selbstadjungiertheit von $\mu - A$: $0 = ((\mu - A)^\alpha x, (\mu - A)^{\alpha-2}x)_X = ((\mu - A)^{\alpha-1}x, (\mu - A)^{\alpha-1}x)_X = \|(\mu - A)^{\alpha-1}x\|_X^2 > 0$, ein Widerspruch.

Beweis: Dass für kompakte Operatoren das Supremum von einem Eigenvektor angenommen wird, wird in den Übungen gezeigt.

Sei

$$s := \sup_x |R_A(x)| \leq \|A\|_X,$$

wobei die Aussage $s \leq \|A\|_X$ einfach zu sehen ist (hier wird die Selbstadjungiertheit nicht benötigt). Um $\|A\|_X \leq s$ zu sehen, müssen wir die Selbstadjungiertheit benutzen. Seien x, $y \in X$ beliebig. Dann gilt:

$$\begin{aligned} 2\left|(Ax,y)_X + (Ay,x)_X\right| &= \left|(A(x+y),(x+y))_X - (A(x-y),(x-y))_X\right| \\ &\leq s\left(\|x+y\|_X^2 + \|x-y\|_X^2\right) = 2s\left(\|x\|_X^2 + \|y\|_X^2\right). \end{aligned} \tag{1.7}$$

Wir wählen nun $y = tAx$ mit einen $t > 0$, welches wir später geeignet wählen. Dann folgt wegen der Selbstadjungiertheit von A

$$4t\|Ax\|_X^2 \leq 2s\left(\|x\|_X^2 + t^2\|Ax\|_X^2\right)$$

d.h.

$$\left(2t - st^2\right)\|Ax\|_X^2 \leq s\|x\|_X^2.$$

Für eine scharfe Abschätzung wählen wir $t > 0$ so, dass die linke Seite möglichst gross wird. Dies führt auf die Wahl $t = 1/s > 0$. Es ergibt sich also

$$\frac{1}{s}\|Ax\|_X^2 \leq s\|x\|_X^2.$$

Weil x beliebig war, ergibt sich $\|A\|_X \leq s$.

Strikt genommen haben wir im Beweis $s > 0$ angenommen. Im degenerierten Fall $s = 0$ jedoch können wir genauso vorgehen, indem wir in (1.7) s durch durch $s + \varepsilon$ für beliebiges $\varepsilon > 0$ ersetzen. Dann ergibt sich $\|A\|_X \leq (s + \varepsilon)$ für beliebiges $\varepsilon > 0$. □

Korollar 1.8 *Es gibt eine Folge $(u_n, \lambda_n)_n \subset V \setminus \{0\} \times \mathbb{R}$ mit folgenden Eigenschaften:*

(i) $a(u_n, v) = \lambda_n (u,v)_H \quad \forall v \in V$

(ii) Die Folge $(\lambda_n)_n$ erfüllt $0 < \lambda_1 \leq \lambda_2 \leq \cdots$ und $\lim_{n\to\infty} \lambda_n = \infty$. Insbesondere hat jeder Eigenwert von (1.3) nur endliche Vielfachheit.

(iii) $(u_n)_n$ ist ONB in H. Insbesondere lässt sich jedes $u \in H$ darstellen als

$$u = \sum_{n=1}^{\infty} (u, u_n)_H u_n.$$

(iv) $(\lambda_n^{-1/2} u_n)_n$ ist ONB in $(V, a(\cdot,\cdot))$ Insbesondere lässt sich jedes $u \in V$ darstellen als

$$u = \sum_{n=1}^{\infty} a(u, \lambda_n^{-1/2} u_n)\lambda_n^{-1/2} u_n.$$

Beweis: Wir verwenden Satz 1.5 für den kompakten Operator T und $X = (V, a(\cdot,\cdot))$.

1. Schritt: Wir behaupten $\operatorname{Ker} T = \{0\}$. Um dies zu sehen, sei $u \in \operatorname{Ker} T$. Dann ist

$$0 = a(0,v) = a(Tu, v) = (u,v)_H \qquad \forall v \in V.$$

Weil V dicht in H liegt, folgt also $u = 0$ (als Element von H und dann wegen der Injektivität der Einbettung $V \subset H$ auch als Element von V).

2. Schritt: Seien $(e_n, \mu_n)_n$ die (orthonormierten) Eigenpaare des Operators T wie in Satz 1.5 bestimmt. D.h.

a) $Te_n = \mu_n e_n$ für alle $n \in \mathbb{N}$

b) $(e_n)_n$ ist eine ONB von $(V, a(\cdot,\cdot))$.

Mit $\lambda_n := 1/\mu_n$ ergibt sich dann (vgl. Lemma 1.4):

$$a(e_n, v) = \lambda_n (e_n, v)_H \qquad \forall v \in V.$$

3. Schritt: Wir behaupten, dass die $(e_n)_n$ auch paarweise orthogonal bzgl. $(\cdot,\cdot)_H$ sind:

$$(e_n, e_m)_H = \lambda_n (e_n, \lambda_n^{-1} e_m)_H = a(e_n, \lambda_n^{-1} e_m) = \lambda_n^{-1} \delta_{n,m}$$

4. Schritt: Wir definieren die im Satz angegeben Funktionen u_n:

$$u_n := \sqrt{\lambda_n} e_n$$

Dann gilt:

- $(u_n, u_m)_H = \delta_{n,m}$
- $a(u_n, u_m) = \lambda_n \delta_{n,m}$
- $(\lambda_n^{-1/2} u_n)_n$ ist ONB von $(V, a(\cdot,\cdot))$.

Es bleibt zu zeigen, dass $(u_n)_n$ den Raum H aufspannt. Sei $\Pi_N : H \to \operatorname{span}\{u_1, \ldots, u_N\}$ die Orthogonalprojektion (in H). Wir behaupten:

$$\lim_{N\to\infty} \|u - \Pi_N u\|_H = 0.$$

Für jedes $u \in V$ gilt

$$u = \sum_{n=1}^{\infty} a(u, \lambda_n^{-1/2} u_n) \lambda_n^{-1/2} u_n = \sum_{n=1}^{\infty} (u, u_n)_H u_n$$
$$\Pi_N u = \sum_{n=1}^{N} (u, u_n)_H u_n = \sum_{n=1}^{N} a(u, \lambda_n^{-1/2} u_n) \lambda_n^{-1/2} u_n.$$

D.h.: die Projektion $\Pi_N u$ stimmt mit der abgebrochenen (Orthogonal-)Entwicklung von u im Raum $(V, a(\cdot,\cdot))$ überein! Damit gilt

$$\lim_{N\to\infty} \|u - \Pi_N u\|_V = 0 \qquad \forall u \in V.$$

Ein Dichtheitsargument dehnt die Aussage nun auf H aus. Sei $u \in H$. Sei $\varepsilon > 0$. Wähle $u_\varepsilon \in V$ mit $\|u - u_\varepsilon\|_H \le \varepsilon$. Dann gilt

$$\|u - \Pi_N u\|_H \le \underbrace{\|u - u_\varepsilon\|_H}_{\le \varepsilon} + \underbrace{\|\Pi_N (u - u_\varepsilon)\|_H}_{\le \|(u-u_\varepsilon)\|_H \le \varepsilon} + \underbrace{\|u_\varepsilon - \Pi_N u_\varepsilon\|_H}_{\le C\|u_\varepsilon - \Pi_N \varepsilon\|_V \to 0 \text{ für } N\to\infty}$$

□

finis 1.Stunde

Für selbstadjungierte, kompakte Operatoren A ist der *Rayleighquotient*

$$R_A(x) := \frac{(Ax, x)_X}{(x, x)_X}, \qquad x \in X \setminus \{0\}$$

ein wichtiges Hilfsmittel, um die Eigenwerte zu charakterisieren[2] Für $A = T$ und $X = (V, a(\cdot,\cdot))$ entsteht

$$\frac{(x, x)_H}{a(x, x)}.$$

Wir betrachten den Kehrwert

$$R(x) := \frac{a(x, x)}{(x, x)_H} \qquad x \in V \setminus \{0\}. \tag{1.8}$$

Es gilt

[2]Für selbstadjungierte Operatoren ist der Rayleighquotient auch ein wichtiges numerisches Hilfsmittel wie z.B. bei Ritz-Verfahren.

Satz 1.9 (Minimumprinzip) *Seien die Eigenwerte λ_n der Grösse nach sortiert: $\lambda_1 \leq \lambda_2 \leq \cdots$ und entsprechend ihrer Vielfachheit aufgeführt. Seien die u_n die zugehörigen Eigenvektoren. Dann gilt:*

(i) $R(u_n) = \lambda_n$ *für alle* n

(ii) $\lambda_1 = \min_{u \in V} R(u)$

(iii) mit

$$V_m := \operatorname{span}\{u_1, \ldots, u_m\}, \tag{1.9}$$

$$V_m^\perp := \{v \in V \,|\, a(v,w) = 0 \quad \forall w \in V_m\} = \{v \in V \,|\, (v,w)_H = 0 \quad \forall w \in V_m\} \tag{1.10}$$

folgt

$$\lambda_m = \min_{v \in V_{m-1}^\perp} R(v), \qquad m = 2, 3, \ldots. \tag{1.11}$$

Beweis: *ad (i):* ist unmittelbar.

ad (ii): Sei $v \in V$. Dann gilt wegen $a(u_n, v) = \lambda_n (u_n, v)_H$ und den Orthogonalitäten, die die Funktionen u_n erfüllen:

$$v = \sum_{n=1}^{\infty} (v, u_n)_H u_n = \sum_{n=1}^{\infty} a(v, \lambda_n^{-1/2} u_n) \lambda_n^{-1/2} u_n$$

$$\|v\|_H^2 = \sum_{n=1}^{\infty} |(v, u_n)_H|^2$$

$$a(v,v) = \sum_{n=1}^{\infty} |a(v, \lambda_n^{-1/2} u_n)|^2 = \sum_{n=1}^{\infty} \lambda_n |(v, u_n)_H|^2$$

D.h.:

$$R(v) = \frac{a(v,v)}{\|v\|_H^2} = \frac{\sum_{n=1}^{\infty} \lambda_n |(v, u_n)_H|^2}{\sum_{n=1}^{\infty} |(v, u_n)_H|^2}$$

Weil $\lambda_n \uparrow$, folgt

$$\min_{v \in V} R(v) = \lambda_1.$$

ad (iii): Sei $v \in V_{m-1}^\perp$. Dann hat v die Darstellung

$$v = \sum_{n=1}^{\infty} (v, u_n)_H u_n = \sum_{n=m}^{\infty} (v, u_n)_H u_n.$$

Damit ergibt sich

$$R(v) = \frac{a(v,v)}{\|v\|_H^2} = \frac{\sum_{n=m}^{\infty} \lambda_n |(v, u_n)_H|^2}{\sum_{n=m}^{\infty} |(v, u_n)_H|^2}$$

und aus der Monotonie $\lambda_n \uparrow$ folgt wieder

$$\min_{v \in V_{m-1}^\perp} R(v) = \lambda_m.$$

□

Der explizite Bezug auf die Eigenvektoren u_n für die Charakterisierung der Eigenwerte λ_m, $m \geq 2$ in Satz 1.9 ist oft unhandlich, da die Eigenvektoren nicht explizit bekannt sind. Man kann das umgehen:

Satz 1.10 (Minimax-Prinzip)

$$\lambda_m = \min_{\substack{E_m \subset V \\ \dim E_m = m}} \max_{v \in E_m} R(v), \qquad m = 1, 2, \ldots.$$

Beweis: "$\geq$": Sei $E_m := V_m = \text{span}\{u_1, \ldots, u_m\}$. Dann hat jedes $v \in V_m$ die Darstellung $v = \sum_{n=1}^m \alpha_n u_n$ mit $\alpha_n = (v, u_n)_H$ und

$$R(v) = \frac{\sum_{n=1}^m \lambda_n \alpha_n^2}{\sum_{n=1}^m \alpha_n^2};$$

Maximieren über alle $(\alpha_n)_{n=1}^m \in \mathbb{R}^m$ liefert als Maximum λ_m wegen der Monotonie $\lambda_n \uparrow$, d.h.

$$\max_{v \in V_m} R(v) = \lambda_m. \tag{1.12}$$

"$\leq$": Sei $E_m \subset V$ mit $\dim E_m = m$. Wähle $v \in E_m \setminus \{0\}$ so, dass

$$(v, u_n)_H = 0, \qquad n = 1, \ldots, m-1.$$

(das geht, weil nur $m-1$ *lineare* Bedingungen gestellt werden). Dann ist $v \in V_{m-1}^\perp \cap E_m$ und somit nach Satz 1.9

$$\lambda_m = \min_{w \in V_{m-1}^\perp} R(w) \leq R(v) \leq \max_{w \in E_m} R(w).$$

□

1.0.1 FEM Diskretisierung

Die Diskretisierung des variationell gestellten Eigenwertproblems (1.3) erfolgt durch Wahl eines abgeschlossenen Unterraums $V_h \subset V$ und Betrachten von

$$\text{Finde } (u_h, \lambda_h) \in V_h \setminus \{0\} \times \mathbb{R} \text{ s. d.} \qquad a(u_h, v) = \lambda_h (u_h, v)_H \qquad \forall v \in V_h. \tag{1.13}$$

Das (diskrete) Eigenwertproblem (1.13) ist äquivalent zu einem algebraischen (verallgemeinerten) Eigenwertproblem: Wählt man eine Basis $\{\varphi_1, \ldots, \varphi_N\}$ von V_h, so ist (1.13) äquivalent zu:

$$\text{Finde } (\mathbf{u}_h, \lambda_h) \in \mathbb{R}^N \setminus \{0\} \times \mathbb{R} \text{ s. d.} \qquad \mathbf{A}\mathbf{u}_h = \lambda_h \mathbf{M}\mathbf{u}_h, \qquad \mathbf{A}_{ij} = a(\varphi_j, \varphi_i), \qquad \mathbf{M}_{ij} = (\varphi_j, \varphi_i)_H. \tag{1.14}$$

Bemerkung 1.11 Für moderate Problemgrössen N kann man das Matrix-Eigenwertproblem (1.14) mit Standardmethoden der numerischen linearen Algebra lösen, z.B. dem QR-Algorithmus (wenn $\mathbf{M} = \text{Id}$ oder wenn man $\mathbf{M}^{-1}\mathbf{A}$ betrachtet) oder Varianten wie dem QZ-Algorithmus (für allgemeines $\mathbf{M}$). Dies liefert alle Eigenwerte und Eigenvektoren mit Aufwand $O(N^3)$. Für grosse N arbeitet man mit iterativen Verfahren, die nur einen kleinen Teil des Spektrums liefern. Dies aus zwei Gründen: 1) die Kosten sind nicht tragbar und QR-artige Algorithmen können sehr schlecht die Besetzungsstruktur von $\mathbf{A}$, $\mathbf{M}$ ausnutzen (typischerweise sind dies *dünn* besetzte Matrizen); 2) man ist ohnehin nur an einem kleinen Teil des Spektrums interessiert, weil die numerischen Approximationen eines grossen Teils des Spektrums sehr schlecht sind ($\rightarrow$ später). ∎

Übung 1.12 Überlegen Sie sich, dass das Minimax-Prinzip auch für das das endlichdimensionale Problem (1.13) gilt, d.h.

$$\lambda_{h,m} = \min_{\substack{E_m \subset V_h \\ \dim E_m = m}} \max_{v \in E_m} R(v) \tag{1.15}$$

∎

Wir nutzen Übung 1.12, um zu zeigen, dass die Konvergenz der diskreten Eigenwerte $\lambda_{h,m}$ "von oben" erfolgt:

Satz 1.13 *Es ist*

$$\lambda_m \leq \lambda_{h,m}, \qquad m = 1, \ldots, N.$$

Beweis: Anwenden des diskreten Minimaxprinzips (1.15) und des kontinuierlichen Minimax-Prinzips liefert wegen $V_h \subset V$:

$$\lambda_{h,m} = \min_{\substack{E_m \subset V_h \\ \dim E_m = m}} \max_{v \in E_m} R(v) \geq \min_{\substack{E_m \subset V \\ \dim E_m = m}} \max_{v \in E_m} R(v) = \lambda_m$$

□

Übung 1.14 Falls die Approximationsräume V_h geschachtelt sind, dann ist die Konvergenz sogar monoton: Seien $V_h \subset V_{h'} \subset V$. Dann gilt

$$\lambda_m \leq \lambda_{h',m} \leq \lambda_{h,m}, \qquad m = 1, \ldots, N.$$

■

1.0.2 Konvergenz der Eigenwerte

Wir definieren den *Ritzprojektor* $P_h : V \to V_h$ als die Orthogonalprojektion auf V_h im $a(\cdot,\cdot)$-Innenprodukt, d.h. $P_h u \in V_h$ ist charakterisiert durch

$$a(u - P_h u, v) = 0 \qquad v \in V_h. \tag{1.16}$$

Wir erinnern an folgende Fakten:

- P_h ist linear
- Bestapproximationseigenschaft: $\|u - P_h u\|_V \leq C \inf_{v \in V_h} \|u - v\|_V$
- Orthogonalprojektion in $a(\cdot,\cdot)$: $\|P_h\|_E \leq 1$, wobei $\|\cdot\|_E$ die Energienorm ist.

Es ist naheliegend, dass für die Konvergenz der Eigenwerte und Eigenvektoren die Frage geklärt werden sollte, in welcher Beziehung der Raum V_m (aufgespannt durch die ersten m Eigenvektoren) zu seiner Projektion $P_h V_m$ in den Approximationsraum V_h steht. Es gilt:

Lemma 1.15 *Definiere*

$$\sigma_{h,m} := \inf_{v \in V_m} \frac{\|P_h v\|_H}{\|v\|_H}, \qquad 1 \leq m \leq N. \tag{1.17}$$

Dann gilt: Falls (für ein $m \in \{1, \ldots, N\}$), $\sigma_{h,m} > 0$, so gilt (für dieses m)

$$\lambda_m \leq \lambda_{h,m} \leq \sigma_{h,m}^{-2} \lambda_m. \tag{1.18}$$

Beweis: (*Bemerkung:* $\sigma_{h,m}$ beschreibt die Norm des Inversen der Operators $P_h : V_m \to P_h V_m$, d.h. $1 - \sigma_{h,m}^{-1}$ ist ein Mass dafür, wie nahe V_m und $P_h V_m$ beieinander sind.) Wir verwenden wieder das Minimax-Prinzip, diesmal mit dem Raum $E_m = P_h V_m$. Hierzu müssen wir aus der Voraussetzung $\sigma_{h,m} > 0$ sehen, dass $\dim E_m = m$: wäre $\dim E_m = P_h V_m < m$, so gäbe es $v \in V_m \setminus \{0\}$ mit $P_h v = 0$, was aber $\sigma_{h,m} > 0$ widerspricht.

Das Minimax-Prinzip (1.15) liefert nun für $E_m = P_h V_m$

$$\begin{aligned} \lambda_{h,m} &\leq \max_{v \in E_m} R(v) = \max_{v \in E_m} \frac{a(v,v)}{\|v\|_H^2} = \max_{v \in V_m} \frac{a(P_h v, P_h v)}{\|P_h v\|_H^2} \leq \max_{v \in V_m} \frac{a(v,v)}{\|P_h v\|_H^2} \\ &= \max_{v \in V_m} \frac{a(v,v)}{\|v\|_H^2} \frac{\|v\|_H^2}{\|P_h v\|_H^2} \leq \lambda_m \sigma_{h,m}^{-2}, \end{aligned}$$

wobei wir im letzten Schritt die Beobachtung (1.12) verwendet haben. □

finis 2.Stunde

Lemma 1.15 zeigt, dass wir für festes m zeigen wollen:[3]

$$\lim_{h \to 0} \sigma_{h,m} = 1.$$

Das folgende Lemma zeigt, dass dies möglich ist:

[3] obwohl die Räume $V_h \subset V$ bis jetzt beliebig waren, stellen wir uns natürlich vor, dass der Raum $S^{p,1}(\mathcal{T}_h)$ für ein Gitter der Maschenweite h gemeint ist

Lemma 1.16

$$\sigma_{h,m}^2 \geq 1 - \frac{2\|a\|\sqrt{m}}{\lambda_1} \sup_{v \in V_m} \frac{\|v - P_h v\|_V^2}{\|v\|_H^2}.$$

Beweis: *Vorbemerkung:* Dreiecksungleichung liefert $\|P_h v\|_H \geq \|v\|_H - \|v - P_h v\|_H \geq \|v\|_H - C\|v - P_h v\|_V$. Der Punkt des Lemmas ist deshalb, dass $\|v - P_h v\|_V^2$ erscheint.

Sei $v \in V_m$ mit $\|v\|_H = 1$ von der Form

$$v = \sum_{n=1}^{m} \alpha_n u_n, \qquad \sum_{n=1}^{m} |\alpha_n|^2 = 1.$$

Wegen $\|v\|_H = 1$ gilt

$$\begin{aligned} 1 - \|P_h v\|_H^2 &= (v, v)_H - (P_h v, P_h v)_H = (v - P_h v, v + P_h v)_H = (v - P_h v, 2v + P_h v - v)_H \\ &= \underbrace{-\|v - P_h v\|_H^2}_{\leq 0} + 2(v - P_h v, v)_H \end{aligned}$$

und damit

$$\|P_h v\|_H^2 = 1 - 2(v - P_h v, v)_H + \|v - P_h v\|_H^2 \geq 1 - 2(v - P_h v, v)_H.$$

Mit $a(z, u_n) = \lambda_n (z, u_n)_H$ für alle $z \in V$ und $a(v - P_h v, w) = 0$ für alle $w \in V_h$ berechnen wir:

$$\begin{aligned} (v - P_h v, v)_H &= \sum_{n=1}^{m} \alpha_n (v - P_h v, u_n)_H \\ &= \sum_{n=1}^{m} \frac{\alpha_n}{\lambda_n} a(v - P_h v, u_n) \\ &= \sum_{n=1}^{m} \frac{\alpha_n}{\lambda_n} a(v - P_h v, u_n - P_h u_n) \\ &\leq \sum_{n=1}^{m} \frac{|\alpha_n|}{\lambda_n} \|a\| \|v - P_h v\|_V \|u_n - P_h u_n\|_V \\ &\leq \|v - P_h v\|_V \frac{\|a\|}{\lambda_1} \underbrace{\sqrt{\sum_{n=1}^{m} |\alpha_n|^2}}_{=1} \underbrace{\sqrt{\sum_{n=1}^{m} \|u_n - P_h u_n\|_V^2}}_{\leq \sqrt{m} \sup\limits_{\substack{w \in V_m \\ \|w\|_H = 1}} \|w - P_h w\|_V} \\ &\leq \frac{\|a\|}{\lambda_1} \sqrt{m} \sup_{\substack{w \in V_m \\ \|w\|_H = 1}} \|w - P_h w\|_V^2 \end{aligned}$$

und damit folgt

$$\|P_h v\|_H^2 \geq 1 - 2\frac{\|a\|}{\lambda_1} \sqrt{m} \sup_{\substack{w \in V_m \\ \|w\|_H = 1}} \|w - P_h w\|_V^2.$$

Wegen $\|v\|_H = 1$ folgt die Behauptung. □

Lemma 1.16 zeigt, dass wir weiterkommen, wenn wir eine Approximationseigenschaft für die Räume $(V_h)_{h>0}$ fordern. Wir werden nun fordern:

$$\forall v \in V \colon \lim_{h \to 0} \inf_{w \in V_h} \|v - w\|_V = 0. \tag{1.19}$$

Satz 1.17 *Es gelte (1.19). Dann gilt: Für jedes $m \in \mathbb{N}$ existiert ein $h_0 > 0$ und ein $C_m > 0$, so dass für alle $h < h_0$ gilt:*

$$0 \le \lambda_{h,m} - \lambda_m \le C_m \sup_{v \in V_m} \inf_{w \in V_h} \frac{\|v - w\|_V^2}{\|v\|_H^2}.$$

Beweis: Wir können wieder $v \in V_m$ mit $\|v\|_H = 1$ betrachten. Schreiben wir $v = \sum_{n=1}^m \alpha_n u_n$ mit $\sum_{n=1}^m |\alpha_n|^2 = 1$, so gilt

$$\|v - P_h v\|_V \le \sum_{n=1}^m |\alpha_n| \|u_n - P_h u_n\|_V \le \underbrace{\sqrt{\sum_{n=1}^m |\alpha_n|^2}}_{=1} \sqrt{\sum_{n=1}^m \|u_n - P_h u_n\|_V^2}$$

Nach Voraussetzung (1.19) können wir jede der m Funktion u_n, $n = 1, \dots, m$, beliebig gut approximieren, wenn h klein genug ist. D.h., wir haben gezeigt:

$$\forall \varepsilon > 0 \quad \exists h_0 > 0 \quad \forall h < h_0 \colon \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \|v - P_h v\|_V \le \varepsilon.$$

Wegen der elementaren Ungleichung

$$\frac{1}{1 - \varepsilon} \le 1 + 2\varepsilon \qquad \text{für } \varepsilon > 0 \text{ hinreichend klein}$$

folgt somit aus Lemma 1.16, dass für hinreichend kleine h

$$\sigma_{h,m}^{-2} \le 1 + C \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \|v - P_h v\|_V^2.$$

und wegen der Bestapproximationseigenschaft des Ritzprojektors

$$\sigma_{h,m}^{-2} \le 1 + C \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \inf_{w \in V_h} \|v - w\|_V^2.$$

Mittels Lemma 1.15 folgt dann die Behauptung. □

Bemerkung 1.18

- Satz 1.17 zeigt, dass die Konvergenz der Eigenwerte *doppelt so schnell* ist wie man es für die Eigenvektoren erwarten kann.
- Satz 1.17 zeigt, dass man für festes m Konvergenz $\lambda_m = \lim_{h \to 0} \lambda_{h,m}$ erwarten kann, wenn man mit Gittern der Maschenweite h arbeitet. ∎

1.0.3 Konvergenz der Eigenvektoren

Wir betrachten hier nur den Fall eines *einfachen* Eigenwertes λ_m; die Konvergenztheorie für Eigenwerte mit Vielfachheit $k \ge 2$ ist möglich.

Die Eigenfunktionen $(u_n)_n$ wurden bereits in Korollar 1.8 definiert. Völlig analog ergeben sich diskrete Eigenfunktionen $(u_{h,n})_{n=1}^N$. Diese bilden eine ONB von $(V_h, (\cdot,\cdot)_H)$ und die Funktionen $(\lambda_{h,n}^{-1/2} u_n)_{n=1}^N$ bilden eine ONB von $(V_h, a(\cdot,\cdot))$.

Eine wichtige Grösse in der Konvergenztheorie von Eigenwertproblemen ist, wie gut der Eigenwert λ_m vom Rest des Spektrums getrennt ist: $\min\{|\lambda_m - \lambda_i| \,|\, i \in \mathbb{N} \setminus \{m\}\} > 0$ (beachte: λ_m ist als einfacher EW gefordert). Entsprechend benötigen wir eine Grösse, die misst, inwieweit das diskrete Spektrum $\{\lambda_{h,i} \,|\, 1 \le i \le N\}$ von λ_m getrennt ist. Genauer: Die Zahl

$$\rho_{h,m} := \max_{\substack{1 \le i \le N \\ i \ne m}} \frac{\lambda_m}{|\lambda_m - \lambda_{h,i}|} \tag{1.20}$$

ist ein Mass dafür, wie weit die diskreten Eigenwerte, die *nicht* gegen λ_m konvergieren, von λ_m getrennt sind.

Übung 1.19 Zeigen Sie: Unter den Voraussetzungen von Satz 1.17 und der Annahme, dass λ_m ein einfacher EW ist, gibt es ein $h_0 > 0$ und ein $c_m > 0$, so dass für alle $h < h_0$ gilt: $\rho_{h,m} \leq c_m$. ∎

Das folgende Lemma zeigt, dass (bis auf Skalierung) die diskrete Eigenfunktion $u_{h,m}$ tatsächlich eng mit der Eigenfunktion u_m zusammenhängt:

Lemma 1.20 *Sei λ_m einfacher Eigenwert und gelte die Approximationseigenschaft (1.19). Dann existiert $h_0 > 0$, so dass für alle $h < h_0$ gilt: Es existiert ein Vorzeichen $\sigma \in \{\pm 1\}$, so dass*

$$\|u_m - \sigma u_{h,m}\|_H \leq 2(1 + \rho_{m,h})\|u_m - P_h u_m\|_H. \tag{1.21}$$

Beweis: Die kontinuierlichen Eigenfunktionen $(u_n)_n$ und die diskreten Eigenfunktionen $(u_{h,n})_n$ sind nur bis auf ein Vorzeichen eindeutig. Wir wählen nun das Vorzeichen σ von $u_{h,m}$ durch die Normierungsbedingung

$$(P_h u_m, u_{h,m})_H \geq 0. \tag{1.22}$$

Mit dieser Wahl wollen wir nun (1.21) mit $\sigma = +1$ zeigen.

Die Funktionen $(u_{h,n})_{n=1}^N$ bilden eine ONB von $(V_h, (\cdot,\cdot)_H)$. Sei $v_{h,m} \in \operatorname{span}\{u_{h,m}\}$ die H-Orthogonalprojektion von $P_h u_m$, d.h.

$$v_{h,m} := (P_h u_m, u_{h,m})_H u_{h,m}$$

Wir bemerken

$$\begin{aligned} P_h u_m - v_{h,m} &= \sum_{\substack{n=1 \\ n\neq m}}^{N} (P_h u_m, u_{h,n})_H u_{h,n} \\ \|P_h u_m - v_{h,m}\|_H^2 &= \sum_{\substack{n=1 \\ n\neq m}}^{N} |(P_h u_m, u_{h,n})_H|^2. \end{aligned}$$

Wir werden unten mit der Dreiecksungleichung

$$\|u_m - u_{h,m}\|_H \leq \|u_m - P_h u_m\|_H + \|P_h u_m - v_{h,m}\|_H + \|v_{h,m} - u_{h,m}\|_H \tag{1.23}$$

zum Ziel kommen. Wir müssen also diese drei Terme abschätzen.

Über die Eigenschaft, dass die Funktionen $(u_n)_n$ und $(u_{h,n})_n$ Eigenfunktionen sind, können wir wieder das $(\cdot,\cdot)_H$-Skalarprodukt in ein $a(\cdot,\cdot)$-Skalarprodukt umwandeln, um die Galerkinorthogonalität des Ritzprojektors P_h auszunutzen:

$$(P_h u_m, u_{h,n})_H = \lambda_{h,n}^{-1} a(P_h u_m, u_{h,n}) = \lambda_{h,n}^{-1} a(u_m, u_{h,n}) = \frac{\lambda_m}{\lambda_{h,n}} (u_m, u_{h,n})_H.$$

Damit ergibt sich

$$\begin{aligned} (\lambda_{h,n} - \lambda_m)(P_h u_m, u_{h,n})_H &= \lambda_m (u_m - P_h u_m, u_{h,n})_H \\ \Longrightarrow \|P_h u_m - v_{h,m}\|_H^2 &= \sum_{\substack{n=1 \\ n\neq m}}^{N} |(P_h u_m, u_{h,n})_H|^2 \\ &\leq \rho_{h,m}^2 \sum_{\substack{n=1 \\ n\neq m}}^{N} |(u_m - P_h u_m, u_{h,n})_H|^2 \\ &\leq \rho_{h,m}^2 \sum_{n=1}^{N} |(u_m - P_h u_m, u_{h,n})_H|^2 \leq \rho_{h,m}^2 \|u_m - P_h u_m\|_H^2. \end{aligned} \tag{1.24}$$

Damit haben wir den zweiten Term in (1.23) in der gewünschten Art abgeschätzt. Es bleibt $v_{h,m} - u_{h,m}$ abzuschätzen. Es ist:

$$\begin{aligned} v_{h,m} - u_{h,m} &= \left((P_h u_m, u_{h,m})_H - 1\right) u_{h,m} \\ \Longrightarrow \|v_{h,m} - u_{h,m}\|_H &= \left|1 - (P_h u_m, u_{h,m})_H\right|. \end{aligned}$$

Die umgekehrte Dreiecksungleichung liefert mit unserer Normierungsannahme $(P_h u_m, u_{h,m})_H \geq 0$

$$\underbrace{\|u_m\|_H}_{=1} - \|u_m - v_{h,m}\|_H \leq \underbrace{\|v_{h,m}\|_H}_{|(P_h u_m, u_{h,m})_H| = (P_h u_m, u_{h,m})_H} \leq \underbrace{\|u_m\|_H}_{=1} + \|u_m - v_{h,m}\|_H, \tag{1.25}$$

so dass wir erhalten

$$|1 - (P_h u_m, u_{h,m})_H| \leq \|u_m - v_{h,m}\|_H.$$

und damit

$$\|v_{h,m} - u_{h,m}\|_H = |1 - (P_h u_m, u_{h,m})_H| \leq \|u_m - v_{h,m}\|_H. \tag{1.26}$$

Die Dreiecksungleichung angewandt auf (1.26) liefert mit (1.24)

$$\begin{aligned}\|v_{h,m} - u_{h,m}\|_H &\leq \|u_m - v_{h,m}\|_H \leq \|u_m - P_h u_m\|_H + \|P_h u_m - v_{h,m}\|_H \\ &\leq \|u_m - P_h u_m\|_H + \rho_{h,m}\|u_m - P_h u_m\|_H\end{aligned}$$

Damit:

$$\begin{aligned}\|u_m - u_{h,m}\|_H &\leq \|u_m - P_h u_m\|_H + \|P_h u_m - v_{h,m}\|_H + \|v_{h,m} - u_{h,m}\|_H \\ &\leq \|u_m - P_h u_m\|_H + (1 + \rho_{h,m})\|u_m - P_h u_m\|_H + \rho_{h,m}\|u_m - P_h u_m\|_H,\end{aligned}$$

was die gewünschte Aussage ist. □

Lemma 1.25 liefert die Konvergenz der Eigenwerte in der (ziemlich schwachen) $\|\cdot\|_H$-Norm. Für Aussagen in der $\|\cdot\|_V$-Norm kann man folgendes Resultat verwenden, welches auch bei Fehlerschätzern für EWP Anwendung findet:

Lemma 1.21 *Sei $(u, \lambda) \in V \setminus \{0\} \times \mathbb{R}$ ein Eigenpaar. Dann gilt für beliebige $v \in V \setminus \{0\}$:*

$$a(v, v) - \lambda\|v\|_H^2 = a(u - v, u - v) - \lambda(u - v, u - v)_H$$

Beweis: Folgt durch direktes Nachrechnen und Ausnutzen von $a(u, z) = \lambda(u, z)_H$ für beliebige $z \in V$. □

Setzt man $u = u_m$ und $v = u_{h,m}$ in Lemma 1.21 ein, so erhalten wir eine Konvergenzaussage für $u_m - u_{h,m}$ aus Lemma 1.20 und Satz 1.17:

Satz 1.22 *Sei λ_m ein einfacher Eigenwert und gelte (1.19). Dann existiert ein $h_0 > 0$ und ein $C > 0$, so dass für alle $h < h_0$ gilt (hier ist das VZ von $u_{h,m}$ immer wie im Beweis von Lemma 1.20 gewählt)*

$$\begin{aligned}\|u_m - u_{h,m}\|_V &\leq C \sup_{\substack{v \in V_m \\ \|v\|_H = 1}} \inf_{v_h \in V_h} \|v - v_h\|_V \\ \|u_m - u_{h,m}\|_H &\leq C\|u_m - P_h u_m\|_H.\end{aligned}$$

Beweis: Lemma 1.21 liefert mit $u = u_m$ und $v = u_{h,m}$ mit den Normierungen $\|u_m\|_H = 1 = \|u_{h,m}\|_H$

$$\begin{aligned}\lambda_{h,m} - \lambda_m &= \lambda_{h,m}\|u_{h,m}\|_H^2 - \lambda_m\|u_{h,m}\|_H^2 = a(u_{h,m}, u_{h,m}) - \lambda_m\|u_{h,m}\|_H^2 \\ &= a(u_m - u_{h,m}, u_m - u_{h,m}) - \lambda_m\|u_m - u_{h,m}\|_H^2.\end{aligned}$$

Das liefert

$$a(u_m - u_{h,m}, u_m - u_{h,m}) = \lambda_{h,m} - \lambda_m + \lambda_m\|u_m - u_{h,m}\|_H^2,$$

woraus sich aus Lemma 1.20 und Satz 1.17 die Behauptung ergibt. □

1.0.4 Bemerkungen

- Die Konvergenzaussage in Satz 1.17 verlangt die Approximierbarkeit von $V_m = \text{span}\{u_1, \ldots, u_m\}$ und nicht einfach nur $\text{span}\{u_m\}$ (im Falle von einfachen Eigenwerten). Tatsächlich kann man die Aussagen auch dahingehend verschärfen, dass bei einfachen Eigenwerten gilt:

$$\lambda_{h,m} - \lambda_m \leq C \inf_{v \in V_h} \|u_m - v\|_V^2. \tag{1.27}$$

- Wir haben hier nur den symmetrischen Fall betrachtet, bei dem der Operator T selbstadjungiert ist. Die Konvergenztheorie bei nichtsymmetrischen Bilinearformen $a(\cdot,\cdot)$ ist möglich. Nicht alle Konvergenzaussagen aus dem symmetrischen Fall sind übertragbar, u.a. weil geometrische und algebraische Vielfachheit von Eigenwerten nicht mehr übereinstimmen. Z.B. ist dann die Verdopplung der Konvergenzrate wie in (1.27) nicht mehr gewährleistet. ∎

1.0.5 ein operatortheoretischer Zugang

Ziel: Illustration eines allgemeineren Zugangs mittels funktionalanalytischer Techniken, der auch auf nicht-symmetrische Probleme verallgemeinerbar ist und auch quantifizierbare Konvergenzaussagen liefern kann.

Eine Grundannahme wird (1.19), sein, das wir etwas anders formulieren:

$$\lim_{h \to 0} P_h u = u \qquad \forall u \in V. \tag{1.28}$$

Satz 1.23 *(i) Sei $(u_h, \lambda_h)_{h>0}$ eine beschränkte Folge von diskreten Eigenpaaren (d.h. $\|u_h\|_V \leq 1$ und $\lambda_h \leq C$ für ein $C > 0$ unabhängig von h). Dann existiert eine Teilfolge $(u_{h'}, \lambda_{h'})$ und ein Eigenpaar $(u, \lambda) \in V \setminus \{0\} \times \mathbb{R}$ von (1.2) mit $u_{h'} \to u$ (in V) und $\lambda_{h'} \to \lambda$.*

(ii) Sei λ ein Eigenwert von (1.2). Dann existiert eine Folge (u_h, λ_h) von diskreten Eigenpaaren mit $\lambda_h \to \lambda$.

Für den Beweis benötigen wir eine Operatordarstellung der diskreten Eigenwertproblem:

Lemma 1.24 *(i) Sei (u_h, λ_h) ein diskretes Eigenpaar. Dann gilt:*

$$\lambda_h P_h T u_h = u_h \tag{1.29}$$

$$\lambda_h P_h T P_h u_h = P_h u_h = u_h. \tag{1.30}$$

(ii) Jedes Eigenpaar $(u, \lambda) \in V \setminus \{0\} \times \mathbb{R} \setminus \{0\}$ von

$$P_h T P_h u = \frac{1}{\lambda} u \tag{1.31}$$

ist ein diskretes *Eigenpaar, d.h. $u = P_h u \in V_h$ und $(u, \lambda) \in V_h \times \mathbb{R}$ löst (1.13).*

Beweis: *ad (i):* Wir nutzen die Selbstadjungiertheit von T und P_h bzgl. $a(\cdot,\cdot)$ aus: Sei $(u_h, \lambda_h) \in V_h \times \mathbb{R}$. D.g.:

$$\begin{aligned} & a(u_h, v) = \lambda_h (u_h, v)_H \qquad \forall v \in V_h \\ \iff & \underbrace{a(u_h, P_h v)}_{= a(P_h u_h, v) = a(u_h, v)} = \lambda_h \underbrace{(u_h, P_h v)_H}_{a(T P_h v, u_h) = a(P_h v, T u_h) = a(v, P_h T u_h)} \qquad \forall v \in V \\ \iff & u_h = \lambda_h P_h T u_h \end{aligned}$$

Damit ist (1.29) gezeigt. (1.30) folgt aus (1.29) durch Anwenden von P_h.

ad (ii): Erfülle (u, λ) (1.31). Dann ist $u = \lambda P_h T P_h u \in V_h$ in V_h. Insbesondere sind die Umformungen aus dem Beweis von (i) möglich und zeigen, dass (u, λ) ein diskretes Eigenpaar ist. □

Beweis von Satz 1.23: entscheidendes Hilfsmittel ist die *Normkonvergenz* $P_hT \to T$. Diese folgt aus

$$P_hT - T = \underbrace{(P_h - I)}_{\to 0 \text{ punktweise}} \underbrace{T}_{\text{kompakt}}$$

ad (i): Sei (u_h, λ_h) beschränkte Folge von diskreten Eigenpaaren mit

$$\|u_h\|_E = 1 \qquad \text{und} \qquad |\lambda_h| \le C \qquad \forall h > 0.$$

Nach Übergang auf eine Teilfolge (die wir wieder mit (u_h, λ_h) bezeichnen) können wir annehmen:

$$u_h \overset{V}{\rightharpoonup} u \in V, \tag{1.32}$$

$$u_h \overset{H}{\to} u, \tag{1.33}$$

$$Tu_h \overset{V}{\to} Tu, \tag{1.34}$$

$$\lambda_h \to \lambda \in \mathbb{R}. \tag{1.35}$$

Nun schliessen wir

$$\underbrace{u_h}_{\to u} = \lambda_h P_h T u_h = \underbrace{\lambda_h}_{\to \lambda} \left(\underbrace{(P_hT - T)}_{\to 0 \text{ in Norm}} \underbrace{u_h}_{\text{beschränkt}} + \underbrace{T(u_h - u)}_{\to 0 \text{wg. } T \text{ kompakt}} + Tu \right)$$

Damit konvergiert die rechte Seite in Norm gegen λTu, so dass die (schwach konvergente) Folge auf der linken Seite auch in Norm gegen ihren schwachen Limes u konvergiert. Wir haben also

$$u = \lambda T u.$$

Es bleibt zu zeigen, dass $u \neq 0$ und $\lambda \neq 0$. Das folgt aus

$$1 = \|u_h\|_E^2 = a(u_h, u_h) = \lambda_h (u_h, u_h)_H = \lambda_h \|u_h\|_H^2 \overset{(1.33),(1.35)}{\to} \lambda \|u\|_H^2.$$

Damit folgt $\lambda \neq 0$ und $\|u\|_H \neq 0$.

ad (ii): Wir argumentieren mittels eines Störungsargumentes wie im Beweis vom Satz von Bauer-Fike. Wir schreiben

$$T = P_hTP_h + \delta_h$$

wobei $\|\delta_h\|_V \to 0$, denn

$$\delta_h = T - P_hTP_h = (I - P_h)T + P_hT(I - P_h)$$

lässt sich schreiben als Verkettung von kompakten Operatoren und punktweise gegen Null konvergierenden Operatoren. Der Operator P_hTP_h ist kompakt und selbstadjungiert (bzgl. $a(\cdot,\cdot)$) und hat deshalb eine ONB $(e_n, \mu_{h,n})$ von $(V, a(\cdot,\cdot))$ aus Eigenvektoren:

$$P_hTP_he_n = \mu_{h,n} e_n, \qquad n = 1, \dots,$$

Die Eigenvektoren zu den EW $\mu_{h,n} = 0$ spannen nach dem Spektralsatz $\operatorname{Ker} P_hTP_h$ auf. Nach Lemma 1.24, (ii) sind die Paare $(e_n, 1/\mu_{h,n})$ mit $\mu_{h,n} \neq 0$ diskrete Eigenpaare. Mit den diskreten Eigenwerten mit $\lambda_{h,i} > 0$, $i = 1, \dots, N$, haben wir also:

$$\mu_{h,n} = \lambda_{h,n}^{-1}, \qquad n = 1, \dots, N, \qquad \mu_{h,n} = 0, \qquad n = N+1, \dots,$$

Sei nun λ ein Eigenwert des kontinuierlichen Problems. Setze $\mu := \lambda^{-1}$. Wir behaupten:

$$\inf\{|\mu - \mu_{h,n}| \,|\, n = 1, \dots, N\} \to 0 \quad \text{für } h \to 0. \tag{1.36}$$

Es reicht, $\inf\{|\mu - \mu_{h,n}| \,|\, n = 1, \dots, N\} > 0$ zu betrachten. Dann ist λ kein diskreter Eigenwert und $\mu - P_hTP_h$ ist invertierbar:

$$(\mu - P_hTP_h)v = \sum_{n=1}^{\infty} (\mu - \mu_{h,n}) a(v, e_n) e_n \quad \text{und} \quad (\mu - P_hTP_h)^{-1} v = \sum_{n=1}^{\infty} \frac{1}{\mu - \mu_{h,n}} a(v, e_n) e_n.$$

Insbesondere erhalten wir

$$\|(\mu - P_h T P_h)^{-1}\|_E = \max_n \frac{1}{|\mu - \mu_{h,n}|} = \frac{1}{\min_n |\mu - \mu_{h,n}|}$$

Sei nun $0 \neq u$ ein zu λ gehöriger Eigenvektor. Dann ist $\mu u = Tu = P_h T P_h u + \delta_h u$ und damit

$$\begin{aligned} u = (\mu - P_h T P_h)^{-1} \delta_h u \qquad &\Longrightarrow \qquad \|u\|_E \leq \|(\mu - P_h T P_h)^{-1}\|_E \|\delta_h\|_E \|u\|_E \\ \Longrightarrow 1 \leq \|(\mu - P_h T P_h)^{-1}\|_E \|\delta_h\|_E &\Longrightarrow \min_n |\mu - \mu_{h,n}| \leq \|\delta_h\|_E \to 0. \end{aligned}$$

Wir haben somit erhalten, dass die diskreten Eigenwerte die kontinuierlichen approximieren. □

Kapitel 2

parabolische Probleme

Vorbemerkung: viele zeitabhängige Probleme werden durch parabolische Gleichungen beschrieben, z.B.

- die Wärmeleitungsgleichung $u_t - \Delta u = f$ in $\Omega_T = \Omega \times (0, T)$
- die Navier-Stokes Gleichungen

$$\begin{aligned} \mathbf{u}_t - \mu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p &= f \\ \operatorname{div} \mathbf{u} &= 0 \end{aligned}$$

Bemerkung: die Zeitvariable t ist ausgezeichnet: die höchste Ableitungsordnung nach t ist 1 während sie nach den Ortsvariablen 2 ist.

Wir betrachten die (lineare) Wärmeleitungsgleichung. Sei $\Omega \subset \mathbb{R}^d$ und zu gegebenem f, u_0

$$u_t - \Delta u = f \qquad \text{auf } \Omega_T := \Omega \times (0, T), \tag{2.1a}$$

$$u(x, t) = 0 \qquad \forall\, x \in \Gamma_T := \partial\Omega \times (0, T), \qquad \text{“parabolischer Rand”} \tag{2.1b}$$

$$u(\cdot, 0) = u_0 \qquad \text{auf } \Omega \tag{2.1c}$$

Bemerkung 2.1 (2.1) beschreibt z.B. die Temperaturverteilung in einem Körper Ω zum Zeitpunkt t, wobei u_0 die Anfangstemperaturverteilung ist und f die zugeführte Energie beschreibt. Die Randbedingung bedeutet, dass die Temperatur am Rand fixiert wird (“Eisbad”).

Vorbemerkungen zur Numerik:

1. das “klassische” Vorgehen ist die *Linienmethode*: man wählt eine *feste* Diskretisierung im Ort und löst das resultieren ODE-System mit einem numerischen Verfahren.
   Vorteile:
   - “time marching” speichereffizient, keine Beschränkung für den Endzeitpunkt
   - Kombination von “klassischen” Diskretisierungen (sowohl im Ort als auch in der Zeit), d.h. gut verstandene und verfügbare Bausteine

   Nachteil: feste Ortsdiskretisierung → Adaptivität im Raum?
2. *Rothemethode:* verwende Zeitdiskretisierung und löse in jedem Zeitschritt ein Ortsproblem (→ Adaptivität im Ort möglich)
3. *Space-time:* diskretisiere das volle Problem (2.1). *Nachteil:* teuer. *Vorteil:* volle Adaptivät möglich

Wir betrachten hier die Linienmethode.

2.1 Variationsformulierung im Ort

Ziel: Formulierung, die für die *Linienmethode* geeignet ist und FEM als Ortsdiskretisierung verwendet.

Aus (2.1) mit $a(u,v) := \int_\Omega \nabla u \cdot \nabla v$ ergibt sich, dass eine klassische Lösung u die Gleichung

$$\langle u_t(t,\cdot), v\rangle_{L^2} + a(u(t,\cdot), v) = \langle f(t,\cdot), v\rangle_{L^2} \qquad forall v \in H_0^1(\Omega) \tag{2.2}$$

erfüllt. Dies ist bereits sinnvoll, wenn für jedes feste t gilt: $u(t,\cdot) \in H_0^1(\Omega)$. Wir werden deshalb (zunächst) mit folgendem Lösungsbegriff für (2.1) arbeiten:

$$\left.\begin{array}{l} \textit{Finde } u \in C^1([0,T], H_0^1(\Omega)), \textit{ sodass} \\ \langle u'(t), v\rangle_{L^2} + a(u(t), v) = \langle f(t), v\rangle_{L^2} \qquad \forall\ v \in H_0^1(\Omega) \\ u(0) = u_0 \ \textit{(in } H_0^1(\Omega)) \end{array}\right\} \tag{2.3}$$

Dabei wird gefordert

- $f \in C([0,T]; L^2(\Omega))$
- $u_0 \in H_0^1(\Omega)$

Bemerkung 2.2 (Ableitungsbegriff) Sei $u : (0,T) \to (\mathcal{X}; \|\cdot\|_{\mathcal{X}})$ eine Funktion wobei $t \in (0,T)$. Ein Element $u'(t) \in \mathcal{X}$ heißt Ableitung von u an der Stelle t, falls

$$\lim_{h\to 0} \left\| \frac{u(t+h) - u(t)}{h} - u'(t) \right\|_{\mathcal{X}} = 0.$$

Bemerkung 2.3 Die Formulierung (2.3) ist nicht die schwächstmögliche (Zeitableitungen sind hier "klassisch"[1]), aber bequem, um ODE-Theorie einzusetzen. Eine Verallgemeinerung der Formulierung wird in der Vorlesung PDE betrachtet. Insbesondere kann der Lösungsbegriff so erweitert werden, dass auch $u_0 \in L^2(\Omega)$ sinnvoll behandelt werden kann.

Übung 2.4 Sei $u \in C^1((0,T); H_0^1(\Omega))$. Dann ist $u \in C^1((0,T); L^2(\Omega))$

Übung 2.5 Für $u \in C^1((0,T); H_0^1(\Omega))$ gilt:

- $t \mapsto \|u(t)\|_{L^2}^2$ ist stetig differenzierbar und
- $\frac{d}{dt}\|u(t)\|_{L^2}^2 = 2\langle u'(t), u(t)\rangle_{L^2}$

Satz 2.6 *(Energieungleichung) Es gelte für ein $\gamma > 0$*

- $\gamma\|v\|_{H^1(\Omega)}^2 \le a(v,v) \qquad \forall\ v \in H_0^1(\Omega)$
- *u löst (2.3)*

Dann gilt:

$$\|u\|_{L^2(\Omega)} \le e^{-\gamma t}\|u_0\|_{L^2(\Omega)} + \int_0^t e^{-\gamma(t-s)}\|f(s)\|_{L^2(\Omega)} ds$$

Beweis:
1. Schritt: Wir nehmen zunächst an, dass $\|u(s)\|_{L^2} > 0$ für alle $0 < s < t$ ist. Dann gilt:

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|_{L^2}^2 \overset{\text{Übung 2.5}}{=} \langle u'(t), u(t)\rangle_{L^2} \qquad und$$
$$\frac{d}{dt}\|u(t)\|_{L^2} = \frac{d}{dt}\sqrt{\|u(t)\|_{L^2}^2} = \frac{1}{2\sqrt{\|u(t)\|_{L^2}^2}}\frac{d}{dt}\|u(t)\|_{L^2}^2 = \frac{\langle u'(t), u(t)\rangle_{L^2}}{\|u(t)\|_{L^2}}$$

[1] es ist auch möglich, $u \in L^1_{loc}(0,T;\mathcal{X})$ distributionell zu differenzieren

Mit einer Testfunktion $v = u(s)$ für festes s folgt aus (2.3)

$$\begin{aligned}\langle f(s), u(s)\rangle_{L^2} = \langle f(s), v\rangle_{L^2} &\overset{(2.3)}{=} \langle u'(s), v\rangle_{L^2} + a(u(s), v) = \\ &= \underbrace{\langle u'(s), u(s)\rangle_{L^2}}_{\|u(s)\|_{L^2}\frac{d}{dt}\|u(t)\|_{L^2}\Big|_{t=s}} + a(u(s), u(s))\end{aligned}$$

$$\Rightarrow a(u(s), u(s)) + \|u(s)\|_{L^2}\frac{d}{dt}\|u(t)\|_{L^2}\Big|_{t=s} \overset{\substack{Cauchy-\\Schwarz}}{\leq} \|f(s)\|_{L^2}\|u(s)\|_{L^2}$$

Da $a(u(s), u(s)) \geq \gamma\|u(s)\|^2_{H^1} \geq \gamma\|u(s)\|^2_{L^2}$

$$\Rightarrow \gamma\|u(s)\|_{L^2} + \frac{d}{dt}\|u(t)\|_{L^2}\Big|_{t=s} \leq \|f(s)\|_{L^2} \qquad \forall\ 0 \leq s \leq t$$

Ein integrierender Faktor für die linke Seite ist $e^{\gamma t}$

$$\Rightarrow \frac{d}{dt}\left(e^{\gamma t}\|u(t)\|_{L^2}\right)\Big|_{t=s} = e^{\gamma s}\left(\gamma\|u(s)\|_{L^2} + \frac{d}{dt}\|u(t)\|_{L^2}\Big|_{t=s}\right) \leq e^{\gamma s}\|f(s)\|_{L^2}$$

Durch Integration von 0 bis t ergibt sich

$$e^{\gamma t}\|u(t)\|_{L^2} - e^{\gamma 0}\|u(0)\|_{L^2} \leq \int_0^t e^{\gamma s}\|f(s)\|_{L^2} ds$$

$$\Rightarrow \|u(t)\|_{L^2} \leq e^{-\gamma t}\|u_0\|_{L^2} + \int_0^t e^{-\gamma(t-s)}\|f(s)\|_{L^2} ds$$

2.Schritt: Wir betrachten den Fall $\|u(s)\|_{L^2} \geq 0$ auf $[0, T]$. Aus (2.3) ergibt sich mit $v = u(s)$

$$\underbrace{a(u(s), u(s))}_{\geq\gamma\|u\|^2_{H^1}\geq\gamma\|u\|^2_{L^2}} + \underbrace{\langle u'(s), u(s)\rangle_{L^2}}_{\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2}\Big|_{t=s}} = \langle f(s), u(s)\rangle_{L^2} \overset{\substack{Cauchy-\\Schwarz}}{\leq} \|f(s)\|_{L^2}\|u(s)\|_{L^2}$$

Um das Problem zu umgehen, dass $\|u(s)\|_{L^2} = 0$ sein kann, weil man ja durch $\|u(s)\|_{L^2}$ dividieren will, definiert man für ein $\varepsilon > 0$ die Funktion $h_\varepsilon(t) := \sqrt{\|u(t)\|^2_{L^2} + \varepsilon^2}$. Dann ist

- $\frac{d}{dt}h_\varepsilon(t)\Big|_{t=s} = \frac{d}{dt}\sqrt{h^2_\varepsilon(t)}\Big|_{t=s} = \frac{1}{2h_\varepsilon(s)}\frac{d}{dt}h^2_\varepsilon(t)\Big|_{t=s} = \frac{1}{h_\varepsilon(s)}\langle u'(t), u(t)\rangle_{L^2}\Big|_{t=s}$
- $\frac{d}{dt}\|u(t)\|^2_{L^2}\Big|_{t=s} = \frac{d}{dt}h^2_\varepsilon(t)\Big|_{t=s} = 2h_\varepsilon(s)\frac{d}{dt}h_\varepsilon(t)\Big|_{t=s}$

Also folgt:

$$\gamma h^2_\varepsilon(s) + h_\varepsilon(s)\frac{d}{dt}h_\varepsilon(t)\Big|_{t=s} \leq \|f(s)\|_{L^2}\underbrace{\|u(s)\|_{L^2}}_{\leq h_\varepsilon(s)} + \gamma\varepsilon^2$$

Da $\frac{\varepsilon^2}{h_\varepsilon(s)} = \frac{\varepsilon^2}{\sqrt{\|u\|^2_{L^2}+\varepsilon^2}} \leq \frac{\varepsilon^2}{\sqrt{\varepsilon^2}} = \varepsilon$, ergibt sich bei Division der Ungleichung durch $h_\varepsilon(s)$

$$\gamma h_\varepsilon(s) + \frac{d}{dt}h_\varepsilon(t)\Big|_{t=s} \leq \|f(s)\|_{L^2} + \gamma\frac{\varepsilon^2}{h_\varepsilon(s)} \leq \|f(s)\|_{L^2} + \gamma\varepsilon$$

Wie im 1. Schritt ergibt sich durch Multiplikation mit $e^{\gamma s}$ und Integration

$$e^{\gamma s}h_\varepsilon(t) - h_\varepsilon(0) \leq \int_0^t e^{\gamma s}\left[\|f(s)\|_{L^2} + \gamma\varepsilon\right] ds$$

$$\Rightarrow h_\varepsilon(t) \leq e^{-\gamma t}h_\varepsilon(0) + \int_0^t e^{-\gamma(t-s)}\left[\|f(s)\|_{L^2} + \gamma\varepsilon\right] ds \qquad \forall\ \varepsilon > 0$$

Für $\varepsilon \to 0$ folgt die Behauptung. $\square$

Übung 2.7 Satz 2.6 liefert die Eindeutigkeit der Lösung von (2.3).

Bemerkung 2.8 Für $f \equiv 0$ ist die Wärmeleitungsleichung *dissipativ* (in $L^2(\Omega)$), d.h. $\|u(t)\|_{L^2} \leq e^{-\gamma t}\|u(0)\|_{L^2}$. Dann folgt für 2 verschiedene Anfangsbedingungen $u_0,\ \tilde{u}_0$, dass die Lösungen $u(t),\ \tilde{u}(t)$ die Bedingung $\|u(t) - \tilde{u}(t)\|_{L^2} \leq e^{-\gamma t}\|u_0 - \tilde{u}_0\|_{L^2}$ erfüllen. Gute numerische Verfahren sollten dieses qualitative Verhalten widerspiegeln.

Bemerkung 2.9 (Evolutionsoperator) Für $f = 0$ kann man den Evolutionsoperator

$$E(t) : u_0 \mapsto u(t)$$

definieren. Wegen der Eindeutigkeit der Lösung ist er ein *Halbgruppe*:

- $E(t+s) = E(t) \circ E(s), \qquad t,\ s \geq 0$
- $E(0) = \mathrm{Id}$

$E(t)$ kann explizit mittels der Eigenfunktionen des Dirichlet-Laplaceoperators dargestellt werden: Seien (φ_n, λ_n) die Eigenpaare von

$$-\Delta\varphi_n = \lambda_n\varphi_n \quad \text{auf } \Omega, \qquad \varphi_n|_{\partial\Omega} = 0$$

und sei $(\varphi_n)_n$ eine ONB von $L^2(\Omega)$ (und Orthogonalbasis von $H_0^1(\Omega)$). Dann ist

$$E(t)u_0 = \sum_{n=1}^{\infty} e^{-\lambda_n t}\langle u_0, \varphi_n\rangle_{L^2}\varphi_n \tag{2.4}$$

Bemerkung 2.10 (Duhamelprinzip) für $f \neq 0$ (hinreichend glatt) gilt für die Lösung u

$$u(t) = E(t)u_0 + \int_0^t E(t-s)f(s)\,ds. \tag{2.5}$$

(Die Konvergenz der Reihen ist in $H_0^1(\Omega)$ zu verstehen, falls $u_0 \in H_0^1(\Omega)$ und $f \in C([0,T]; L^2(\Omega))$.

2.2 Semidiskretisierung im Ort („Linienmethode“)

Ziel: Approximotion von (2.3) durch ein (endliches) System von ODEs. Sei $V_h \subset H_0^1(\Omega)$ mit $\dim(V_h) = N < \infty$ und Basis $\{\varphi_i \mid i = 1, \ldots, N\}$. Sei $u_{0,h} \in V_h$ eine Approximation an u_0. Dann ist die *semidiskrete Approximation* u_h an die Lösung u gegeben durch

$$\left.\begin{array}{ll} \textit{Finde } u_h \in C^1([0,T]; V_h) \textit{ sodass} & \\ (2.6a) \quad \langle u_h'(t), v\rangle_{L^2} + a(u_h(t), v) = \langle f(t), v\rangle_{L^2} & \forall\ v \in V_h \\ (2.6b) \quad u_h(0) = u_{0,h} & \end{array}\right\} \tag{2.6}$$

(2.6) stellt ein ODE-System dar. Definiert man die Steifigkeitsmatrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ und die Massematrix $\mathbf{M} \in \mathbb{R}^{N\times N}$ durch

$$\mathbf{A}_{ij} = a(\varphi_j, \varphi_i), \qquad \mathbf{M}_{ij} = \langle \varphi_j, \varphi_i\rangle_{L^2}, \quad i,j = 1, \ldots, N \tag{2.7}$$

und $\mathbf{F}(t)$ durch

$$\mathbf{F}_i(t) = \langle f(t), \varphi_i\rangle_{L^2} \tag{2.8}$$

so ergibt sich aus dem Ansatz $u_h(t) = \sum_{i=1}^{N} \mathbf{u}_i(t)\varphi_i$, dass (2.6) äquivalent ist zum ODE-System

$$\left.\begin{array}{ll} \mathbf{M}\mathbf{u}'(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{F}(t) & t > 0 \\ \mathbf{u}(0) = \mathbf{u}_0, & \end{array}\right\} \tag{2.9}$$

wobei $u_{0,h} = \sum_{i=1}^{N} \mathbf{u}_{0,i}\,\varphi_i$ die Darstellung in der Basis $\{\varphi_i \mid i = 1, \dots, N\}$ von $u_{0,h} \in V_h$ ist.[2]

Übung 2.11 Die Matrizen $\mathbf{A}$ und $\mathbf{M}$ sind SPD. Überdies gilt für alle $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$ mit $\mathbf{v} = \sum_{i=1}^{N} \mathbf{v}_i\varphi_i$, $\mathbf{w} = \sum_{i=1}^{N} \mathbf{w}_i\varphi_i$, dass $\mathbf{v}^T\mathbf{M}\mathbf{w} = \langle w, v\rangle_{L^2}$ und $\mathbf{v}^T\mathbf{A}\mathbf{w} = a(w, v)$.

Bemerkung 2.12 Aus Übung 2.11 folgt, dass (2.9) äquivalent ist zu

$$\begin{aligned}\mathbf{u}' &= \mathbf{M}^{-1}\mathbf{F}(t) - \mathbf{M}^{-1}\mathbf{A}\mathbf{u}(t)\\ \mathbf{u}(0) &= u_0\end{aligned}$$

D.h. die Existenz und Eindeutigkeit von (2.6) ist gegeben.

Analog zu Satz 2.6 gilt

Lemma 2.13 *Seien $r \in C^0([0,T]; L^2(\Omega))$ und $w \in C^1([0,T]; V_h)$ und erfüllen*

$$\langle w', v\rangle_{L^2} + a(w, v) = \langle r(t), v\rangle_{L^2} \text{ für alle } v \in V_h.$$

Dann gilt:

$$\|w(t)\|_{L^2} \le e^{-\gamma t}\|w(0)\|_{L^2} + \int_0^t e^{-\gamma(t-s)}\|r(s)\|_{L^2} ds$$

Beweis: Der Beweis erfolgt analog zu jenem von Satz 2.6. □

Bemerkung 2.14 Analog zur Evolution E können wir die semidiskrete Evolution E_h definieren, welche für festes t $u_{0,h} \in V_h$ auf die Lösung $y_h(t) =: E_h u_{0,h}$ von (2.9) mit $f \equiv 0$ abbildet. Lemma 2.13 besagt, $\|E_h(t)v_h\|_{L^2} \le \|v_h\|_{L^2}$ für alle $v_h \in V_h$. Weiters kann durch E_h durch die diskreten Eigenfunktionen und Eigenwerte geschrieben werden (Übung). Es das diskrete Duhamelprinzip:

$$u_h(t) = E_h(t)u_{0,h} + \int_0^t E_h(t-s)\Pi^{L^2} f(s)\, ds$$

mit der L^2-Projektion $\Pi^{L^2} : L^2(\Omega) \to V_h$. ∎

Im folgenden wollen wir $\|u(t) - u_N(t)\|_{L^2}$ abschätzen, indem wir $\|u(t) - P_h u(t)\|_{L^2}$ abschätzen. P_N bezeichnet dabei wieder den Ritzprojektor $P_h : H_0^1(\Omega) \to V_h$. Dieser ist für alle $v \in V_h$ durch $a(w, v) = a(P_h w, v)$ definiert. Wir wissen bereits, dass P_h linear und beschränkt ist. Die Beschränktheit folgt aus der Existenz einer Konstante $C > 0$, sodass $\|P_h w\|_{H^1} \le C\|w\|_{H^1}$ für alle $w \in H_0^1(\Omega)$ gilt.

---

[2] Um die Äquivalenz von (2.6) mit (2.9) zu sehen, schreiben wir $u_h(t) = \sum_{i=1}^{N} \mathbf{u}_i(t)\varphi_i$. Damit gilt:

$$\begin{aligned}
&u_h \text{ löst (2.6a)}\\
&\Leftrightarrow \left\langle \sum_{j=1}^{N} \mathbf{u}_j'(t)\varphi_j, \sum_{i=1}^{N} \mathbf{v}_i\varphi_i \right\rangle_{L^2} + a\left(\sum_{j=1}^{N} \mathbf{u}_j(t)\varphi_j, \sum_{i=1}^{N} \mathbf{v}_i\varphi_i\right) = \left\langle f(t), \sum_{i=1}^{N} \mathbf{v}_i\varphi_i \right\rangle_{L^2} \qquad \forall\, \mathbf{v} \in \mathbb{R}^N\\
&\Leftrightarrow \sum_{i,j=1}^{N} \mathbf{u}_j'(t)\mathbf{v}_i\langle\varphi_j, \varphi_i\rangle_{L^2} + \sum_{i,j=1}^{N} \mathbf{u}_j(t)\mathbf{v}_i\, a(\varphi_j, \varphi_i) = \sum_{i=1}^{N} \mathbf{v}_i\langle f(t), \varphi_i\rangle_{L^2} \qquad \forall\, v \in \mathbb{R}^N\\
&\Leftrightarrow \mathbf{v}^T\mathbf{M}\mathbf{u}'(t) + \mathbf{v}^T\mathbf{A}\mathbf{u}(t) = \mathbf{v}^T\mathbf{F}(t) \qquad \forall\, \mathbf{v} \in \mathbb{R}^N\\
&\Leftrightarrow \mathbf{M}\mathbf{u}'(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{F}(t)
\end{aligned}$$

Satz 2.15 *Sei $u \in C^1([0,T]; H_0^1(\Omega))$ eine Lösung von (2.3) und u_h eine Lösung von (2.6). Dann gilt*

$$\|u(t) - u_h(t)\|_{L^2} \leq \|u(t) - P_h u(t)\|_{L^2} + \|u_h(0) - P_h u(0)\|_{L^2} e^{-\gamma t} \quad + \int_0^t e^{-\gamma(t-s)} \|u'(s) - P_h u'(s)\|_{L^2} ds$$

Beweis: Wird $u_h(t) - u(t) = \underbrace{u_h(t) - P_h u(t)}_{=:\theta(t)} + \underbrace{P_h u(t) - u(t)}_{=:\varrho(t)}$ geschrieben, dann ist

$$\|u_h(t) - u(t)\|_{L^2} \leq \|\theta(t)\|_{L^2} + \|\varrho(t)\|_{L^2}.$$

1. Schritt: Aus der Linearität und der Beschränktheit von P_h und $u \in C^1([0,T]; H_0^1(\Omega))$ folgt, dass $(P_h u)' = P_h u'$, weil

$$\begin{aligned}
&\overline{\lim_{k\to 0}} \left\| \frac{1}{k} \left(P_h u(t+k) - P_h u(t) \right) - P_h u'(t) \right\|_{H^1} \overset{P_N \text{ linear}}{=} \\
&= \overline{\lim_{k\to 0}} \left\| P_h \left(\frac{u(t+k) - u(t)}{k} - u'(t) \right) \right\|_{H^1} \overset{P_h \text{ beschränkt}}{=} \\
&\leq \overline{\lim_{k\to 0}} C \left\| \frac{u(t+k) - u(t)}{k} - u'(t) \right\|_{H^1} = 0 \qquad \text{nach Definition von } u'.
\end{aligned}$$

2. Schritt: Aus dem 1. Schritt folgt $\theta \in C^1([0,T]; V_h)$. Weiters ist für jedes $v \in V_h$

$$\begin{aligned}
\langle \theta'(t), v \rangle_{L^2} + a(\theta(t), v) &= \langle u_h', v \rangle_{L^2} + a(u_h, v) - \langle P_h u', v \rangle_{L^2} - a(P_h u, v) = \\
&\overset{(2.6)}{=} \langle f(t), v \rangle_{L^2} - \langle P_h u', v \rangle_{L^2} - a(P_h u, v) = \\
&\overset{\substack{Def. \\ v.P_h}}{=} \langle f(t), v \rangle_{L^2} - \langle P_h u', v \rangle_{L^2} - a(u, v) = \\
&\overset{(2.3)}{=} \langle u', v \rangle_{L^2} - \langle P_h u', v \rangle_{L^2} = \langle u' - P_h u'(t), v \rangle_{L^2}
\end{aligned}$$

3. Schritt: Aus Lemma 2.13 folgt für θ

$$\|\theta\|_{L^2} \leq e^{-\gamma t} \underbrace{\|\theta(0)\|_{L^2}}_{= \|u_h(0) - P_h u(0)\|_{L^2}} + \int_0^t e^{-\gamma(t-s)} \|u'(s) - P_h u'(s)\|_{L^2} ds.$$

□

Bemerkung 2.16 Satz 2.15 zeigt, dass $\|u(t) - u_h(t)\|_{L^2}$ durch den Fehler $\|u(t) - P_h u(t)\|_{L^2} + 2$ *Terme* abgeschätzt werden kann, die eine Fehlerakkumulation für die Zeiten $0 \leq s < t$ darstellen. Man spricht vom „Gedächtnis“ von parabolischen Gleichungen. ∎

Regularitätsannahmen an u erlauben Abschätzungen, die explizit in h sind.

Korollar 2.17 *Sei $V_h = S_0^1(\mathcal{T})$, wobei $\mathcal{T}$ eine formreguläre, reguläre Triangulierung sei. Erfülle die Lösung u von (2.3) die Regularitätsvoraussetzung $u \in C^3(\bar{\Omega} \times [0,T])$. Sei $u_{0,h} \in V_h$ entweder $u_{0,h} = P_h u_0$ oder $u_{0,h} = I u_0$, wobei $I u_0$ den stückweisen linearen Interpolanten bezeichne. Dann gilt:*

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch \max_{0 \leq s \leq t} \left(|u(\cdot, s)|_{H^2(\Omega)} + |u_t(\cdot, s)|_{H^2(\Omega)} \right).$$

Ist Ω sogar konvex, dann ist

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch^2 \max_{0 \leq s \leq t} \left(|u(\cdot, s)|_{H^2(\Omega)} + |u_t(\cdot, s)|_{H^2(\Omega)} \right).$$

Beweis: Es gilt für jedes $v \in H^2(\Omega)$ (für $d \in \{1,2,3\}$)

- $\|v - P_h v\|_{L^2(\Omega)} \leq \|v - P_h v\|_{H^1(\Omega)} \leq C\|v - Iv\|_{H^1(\Omega)} \leq Ch|v|_{C^2(\bar{\Omega})}$
- falls Ω konvex ist, gilt sogar mittels Aubin-Nitsche $\|v - P_h v\|_{L^2(\Omega)} \leq Ch^2|v|_{C^2(\bar{\Omega})}$

Damit folgt die Behauptung aus Satz 2.15 □

Übung 2.18 Sei $\theta \in C^1([0,T]; V_h)$ und gelte für alle $v \in V_h$ und für ein $r \in C^0([0,T]; L^2(\Omega))$, dass $\langle \theta', v\rangle_{L^2(\Omega)} + a(\theta, v) = \langle r(t), v\rangle_{L^2(\Omega)}$. Zeigen Sie:

$$|\theta(t)|^2_{H^1(\Omega)} \leq |\theta(0)|^2_{H^1(\Omega)} + \int_0^t \|r(s)\|^2_{L^2(\Omega)} ds$$

Hinweis: Betrachte $v = \theta'$.

Zeigen Sie, dass

$$|u(t) - u_h(t)|^2_{H^1} \leq 2\ |u(t) - P_h u(t)|^2_{H^1} + 2\ |u_{0,N} - P_h u_0|^2_{H^1} + 2 \int_0^t \|u'(s) - P_h u'(s)\|^2_{L^2} ds$$

2.3 Volldiskrete Verfahren

Die Semidiskretisierung führt auf das ODE-System

$$\mathbf{M}\mathbf{u}' + \mathbf{A}\mathbf{u} = \mathbf{F}, \qquad \mathbf{u}(0) = \mathbf{u}_0 \tag{2.10}$$

wobei $\mathbf{M}$ und $\mathbf{A}$ SPD sind.

Frage: Richtige Wahl der Zeitdiskretisierung?

Antwort: Verfahren für steife Differentialgleichungen, also A- oder sogar L-stabile Verfahren

Um die richtige Wahl der Zeitdiskretisierung zu motivieren, sollten wir verstehen, wie sich die Lösungen von (2.10) verhalten. Wir versuchen deshalb, (2.10) in ein entkoppeltes ODE-System umzuwandeln.

Satz 2.19 *Seien* $\mathbf{A}$ *und* $\mathbf{M} \in \mathbb{R}^{N\times N}$ *SPD. Dann gelten für das verallgemeinerte Eigenwertproblem*

$$\text{Finde } (\mathbf{v}, \lambda) \in \mathbb{R}^N \setminus \{0\} \times \mathbb{C}, \text{ sodass } \mathbf{A}\mathbf{v} = \lambda \mathbf{M}\mathbf{v} \tag{2.11}$$

folgende Aussagen:

(i) Der Eigenwert λ *erfülle* $\lambda > 0$.

(ii) Es gibt N *Eigenpaare* $(\mathbf{v}_i, \lambda_i)$, $i = 1, \ldots, N$, *die orthogonal bzgl.* $(\cdot,\cdot)_{\mathbf{A}}$ *und* $(\cdot,\cdot)_{\mathbf{M}}$ *sind, d.h.*

$$(\mathbf{v}_i, \mathbf{v}_j)_{\mathbf{M}} = \langle \mathbf{M}\mathbf{v}_i, \mathbf{v}_j\rangle_2 = 0 \qquad \forall i \neq j$$
$$(\mathbf{v}_i, \mathbf{v}_j)_{\mathbf{A}} = \langle \mathbf{A}\mathbf{v}_i, \mathbf{v}_j\rangle_2 = 0 \qquad \forall i \neq j$$

(iii) Die Matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N) \in \mathbb{R}^{N\times N}$ *diagonalisiert* $\mathbf{M}$ *und* $\mathbf{A}$ *simultan, d.h.*

$$\mathbf{V}^T\mathbf{M}\mathbf{V} = \textit{Diagonalmatrix}$$
$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \textit{Diagonalmatrix}$$

(iv) Falls die $\mathbf{v}_i$ *so normiert werden, dass* $(\mathbf{v}_i, \mathbf{v}_j)_{\mathbf{M}} = \delta_{ij}$, *dann gilt*

$$\mathbf{V}^T\mathbf{M}\mathbf{V} = \mathrm{Id}, \quad \mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{D} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}$$

Beweis: Übung. *Hinweis:* Betrachte das EWP $\mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{M}^{-\frac{1}{2}}\mathbf{x} = \lambda\mathbf{x}$ □

Definiert man nun $\widetilde{\mathbf{u}} = \mathbf{V}^{-1}\mathbf{u}$, $\widetilde{\mathbf{F}} = \mathbf{V}^T\mathbf{F}$, $\widetilde{\mathbf{u}}_0 = \mathbf{V}^{-1}\mathbf{u}_0$, dann ist (2.10) äquivalent zu

$$\widetilde{\mathbf{u}}' + \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix} \widetilde{\mathbf{u}} = \widetilde{\mathbf{F}}, \quad \widetilde{\mathbf{u}}(0) = \widetilde{\mathbf{u}}_0 \tag{2.12}$$

(2.12) stellt ein im Sinne der Vorlesung "Numerik von Differentialgleichungen" steifes ODE-System dar, falls einige EW $\lambda_i > 0$ groß sind. Das ist bei parabolischen Problemen der Fall, wie der folgende Satz zeigt.

Satz 2.20 *Sei $\mathcal{T}$ eine formreguläre, reguläre Triangulierung. Seien die Steifigkeitsmatrix $\mathbf{A}$ und die Massematrix $\mathbf{M}$ durch (2.7) definiert, wobei $\{\varphi_i \mid i = 1, \ldots, N\}$ die Basis aus Hutfunktionen von $V_h = S_0^1(\mathcal{T})$ ist. Sei $h_{min} = \min\limits_{K\in\mathcal{T}} h_K$. Dann existiert eine Konstante $C > 0$, die nur von der Formregularität von $\mathcal{T}$ abhängt, sodass*

$$C^{-1}\|u\|^2_{L^2(\Omega)} \le |u|^2_{H^1(\Omega)} \le \frac{C}{h^2_{min}}\|u\|^2_{L^2(\Omega)} \qquad \forall\ u \in S_0^1(\mathcal{T})$$

Insbesondere folgt für die EW λ_i von (2.11)

$$C^{-1} \le \min_{i=1,\ldots,N} \lambda_i \le \max_{i=1,\ldots,N} \lambda_i \le \frac{C}{h^2_{min}} \tag{2.13}$$

Beweis: siehe Übungen. Insbesondere ist die obere Abschätzung $O(h^{-2}_{min})$ scharf. □

2.3.1 das implizite Eulerverfahren

Mittels der VO "Numerik von Differentialgleichungen" zeigt Satz 2.20, dass wir ein *implizites* Verfahren zum Lösen von (2.8) verwenden sollten. Das einfachste Verfahren ist das *implizite Eulerverfahren*:

$$\left\langle \frac{u_h^{n+1} - u_h^n}{k}, v \right\rangle_{L^2(\Omega)} + a(u_h^{n+1}, v) = \langle f(t_{n+1}), v\rangle_{L^2(\Omega)} \qquad \forall\ v \in V_h, \tag{2.14}$$

wobei $k > 0$ der Zeitschritt, $t_n = nk$ und $u_h^n \approx u_h(t_n)$ ist. In Matrixschreibweise ist (2.14)

$$\frac{1}{k}\mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + \mathbf{A}\mathbf{u}^{n+1} = \mathbf{F}^{n+1} \tag{2.15}$$

d.h., in jedem Schritt muss das LGS

$$(\mathbf{M} + k\mathbf{A})\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^n + k\mathbf{F}^{n+1}$$

gelöst werden.

Bemerkung 2.21 Es bietet sich an, $\mathbf{M} + k\mathbf{A}$ einmal zu zerlegen (Choleskyzerlegung), und dann in jedem Zeitschritt eine Vorwärts- und Rückwärtssubstitution zu machen. ∎

Satz 2.22 *Sei $u \in C^1([0,T]; H_0^1(\Omega))$ Lösung von (2.3) und erfülle die Regularitätsvoraussetzung $u \in C^2([0,T]; L^2(\Omega))$. Sei $k_0 > 0$ fest gewählt. Sei λ_{min} der kleinste EW[3] des verallgemeinerten EWP*

$$\lambda\mathbf{M}x = \mathbf{A}x,$$

[3] für die Wärmleitungsgleichung ist $\lambda_{min} = O(1)$, d.h. "harmlos"

wobei $\mathbf{M}$ *und* $\mathbf{A}$ *die Massematrix und die Steifigkeitsmatrix der Ortsdiskretisierung sind. Dann gilt: Es existiert* $b > 0$*, welches nur von* k_0 *und* λ_{min} *abhängt, so dass für jeden Zeitschritt* $k \in (0, k_0]$ *gilt:*

$$\|u_h^n - u(t_n)\|_{L^2} \leq \|u(t_n) - P_h u(t_n)\|_{L^2} + e^{-bt_n}\|u_{0,h} - P_h u_0\|_{L^2} + \int_0^{t_n} e^{-b(t_n - t)} \left[\|u'(t) - P_h u'(t)\|_{L^2} + k\ \|u''(t)\|_{L^2}\right] dt, \tag{2.16}$$

$$\|u_h^n - u(t_n)\|_{L^2}^2 + \sum_{j=0}^{n-1} \|u_h^{j+1} - u_h^j\|_{L^2}^2 + k|u_h^{j+1} - u(t_{j+1})|_{H^1}^2 \leq C\Big[\|u_{0,h} - P_h u\|_{L^2}^2 + \int_0^{t_n} \|u' - P_h u'\|_{L^2}^2 + k^2\|u''\|_{L^2}^2\, dt + \|u(t_n) - P_h u(t_n)\|_{L^2}^2 + k\sum_{j=0}^{n-1} |u(t_j) - P_h u(t_j)|_{H^1}^2\Big] \tag{2.17}$$

Beweis: *Beweis von (2.16):* Das Vorgehen ist das klassische Vorgehen für ESV für ODEs:

1. Rekurrenzformel für den Fehler
2. Auflösen der Rekurrenz mittels diskreten Gronwall-Lemmas

1. Schritt:

$$u_h^n - u(t_n) = \underbrace{u_h^n - P_h u(t_n)}_{=:\theta^n} + \underbrace{P_h u(t_n) - u(t_n)}_{=:\varrho^n}$$

2. Schritt: (Rekurrenzrelation für θ^n)

$$\tfrac{1}{k}\langle u_h^{n+1} - u_h^n, v\rangle_{L^2} + a(u_h^{n+1}, v) = \langle f(t_{n+1}), v\rangle_{L^2} \quad \forall\ v \in V_h$$
$$\langle u'(t_{n+1}), v\rangle_{L^2} + a(u(t_{n+1}), v) = \langle f(t_{n+1}), v\rangle_{L^2} \quad \forall\ v \in H_0^1(\Omega)$$

Für die Differenzbildung dieser beiden Gleichungen ist es geschickt, $u'(t_{n+1})$) mittels Taylor umzuschreiben:

$$u(t_n) \overset{Taylor}{=} u(t_{n+1}) + \underbrace{(t_n - t_{n+1})}_{=-k} u'(t_{n+1}) + \int_{t_{n+1}}^{t_n} (t - t_{n+1})u''(t)dt$$

$$\Rightarrow u'(t_{n+1}) = \frac{u(t_{n+1}) - u(t_n)}{k} - \frac{1}{k}\int_{t_n}^{t_{n+1}} (t - t_{n+1})u''(t)dt =$$

$$= \frac{P_h u(t_{n+1}) - P_h u(t_n)}{k} + \frac{u(t_{n+1}) - u(t_n)}{k} - \frac{P_h u(t_{n+1}) - P_h u(t_n)}{k} - \frac{1}{k}\int_{t_n}^{t_{n+1}} (t - t_{n+1})u''(t)dt =$$

$$= \frac{P_h u(t_{n+1}) - P_h u(t_n)}{k} + \underbrace{\frac{1}{k}\int_{t_n}^{t_{n+1}} u'(t) - P_h u'(t)dt}_{=:w_1^{n+1}} - \underbrace{\frac{1}{k}\int_{t_n}^{t_{n+1}} (t - t_{n+1})u''(t)dt}_{=:w_2^{n+1}}$$

Es gilt:

$$a(u(t_{n+1}), v) = a(P_h u(t_{n+1}), v) \quad \forall v \in V_h \tag{2.18}$$

$$\frac{1}{k}\langle u_h^{n+1} - u_h^n, v\rangle_{L^2} + a(u_h^{n+1}, v) = \langle f(t_{n+1}), v\rangle_{L^2} \quad \forall v \in V_h \tag{2.19}$$

$$\frac{1}{k}\langle P_h u(t_{n+1}) - P_h u(t_n), v\rangle_{L^2(\Omega)} + a(P_h u(t_{n+1}), v) = \langle f(t_{n+1}), v\rangle_{L^2} - \langle w_1^{n+1}, v\rangle_{L^2} - \langle w_2^{n+1}, v\rangle_{L^2} \quad \forall v \in V_h \tag{2.20}$$

Differenzenbildung von (2.19) und (2.20) führt auf

$$\frac{1}{k}\langle \theta^{n+1} - \theta^n, v\rangle_{L^2} + a(\theta^{n+1}, v) = \langle w_1^{n+1}, v\rangle_{L^2} + \langle w_2^{n+1}, v\rangle_{L^2} \qquad \forall\ v \in V_h. \tag{2.21}$$

Mit $v = \theta^{n+1}$ ergibt sich

$$\|\theta^{n+1}\|_{L^2}^2 + k \underbrace{a(\theta^{n+1}, \theta^{n+1})}_{=|\theta^{n+1}|_{H^1}^2 \geq \lambda_{min}\|\theta^{n+1}\|_{L^2}^2} = \langle \theta^n, \theta^{n+1}\rangle_{L^2} + k\langle w_1^{n+1}, v\rangle_{L^2} + k\langle w_2^{n+1}, v\rangle_{L^2}. \tag{2.22}$$

Aus (2.22) ergibt sich:

$$\begin{aligned}
&\Rightarrow (1 + k\lambda_{min})\|\theta^{n+1}\|_{L^2}^2 \overset{\substack{Cauchy-\\ Schwarz}}{\leq} \\
&\qquad \leq \|\theta^n\|_{L^2}\|\theta^{n+1}\|_{L^2} + k\|w_1^{n+1}\|_{L^2}\|\theta^{n+1}\|_{L^2} + k\|w_2^{n+1}\|_{L^2}\|\theta^{n+1}\|_{L^2} \\
&\Rightarrow (1 + k\lambda_{min})\|\theta^{n+1}\|_{L^2} \leq \|\theta^n\|_{L^2} + k\|w_1^{n+1}\|_{L^2} + k\|w_2^{n+1}\|_{L^2}
\end{aligned}$$

3. Schritt: (Auflösen der Rekurrenz)

$$\|\theta^{n+1}\|_{L^2} \leq \frac{1}{1 + k\lambda_{min}}\|\theta^n\|_{L^2} + \frac{k}{1 + k\lambda_{min}}\left(\|w_1^{n+1}\|_{L^2} + \|w_2^{n+1}\|_{L^2}\right)$$

$$\begin{aligned}
\Rightarrow \|\theta^n\|_{L^2} &\leq (1 + \lambda_{min}k)^{-n}\|\theta^0\|_{L^2} + \frac{k}{1 + \lambda_{min}k}\sum_{j=1}^{n}(1 + k\lambda_{min})^{-(n-j)}\left(\|w_1^j\|_{L^2} + \|w_2^j\|_{L^2}\right) = \\
&= (1 + \lambda_{min}k)^{-n}\|\theta^0\|_{L^2} + k\sum_{j=1}^{n}(1 + \lambda_{min}k)^{-(n-(j-1))}\left(\|w_1^j\|_{L^2} + \|w_2^j\|_{L^2}\right).
\end{aligned}$$

Mit $t_n = nk$ und $t_{n-(j-1)} = k(n - (j-1))$ ergibt sich, dass man

$$\begin{aligned}
(1 + \lambda_{min}k)^{-n} &= (1 + \lambda_{min}k)^{-t_n\lambda_{min}/(\lambda_{min}k)} \\
(1 + \lambda_{min}k)^{-(n-(j-1))} &= (1 + \lambda_{min}k)^{-t_{n-j+1}\lambda_{min}/(\lambda_{min}k)}
\end{aligned}$$

abschätzen muss. Elementare Überlegungen[4] zeigen, dass die Funktion

$$x \mapsto (1 + x)^{-1/x}$$

monoton wachsend ist. Aus $k \leq k_0$ folgt somit, dass $\lambda_{min}k \leq \lambda_{min}k_0$ ist, d.h.

$$\begin{aligned}
(1 + \lambda_{min}k)^{-n} &= (1 + \lambda_{min}k_0)^{-t_n\lambda_{min}/(\lambda_{min}k_0)} \leq e^{-bt_n} \\
(1 + \lambda_{min}k)^{-(n-(j-1))} &= (1 + \lambda_{min}k_0)^{-t_{n-j+1}\lambda_{min}/(\lambda_{min}k_0)} \leq e^{-b(t_n - t_{j-1})},
\end{aligned}$$

wobei $b > 0$ durch die Beziehung $(1 + \lambda_{min}k_0)^{-\lambda_{min}/(\lambda_{min}k_0)} = e^{-b}$ definiert ist. Damit erhalten wir

$$\|\theta^n\|_{L^2} \leq e^{-bt_n}\|\theta^0\|_{L^2} + k\sum_{j=1}^{n} e^{-b(t_n - t_{j-1})}\left(\|w_1^j\|_{L^2} + \|w_2^j\|_{L^2}\right)$$

Wegen

$$\|w_1^j\|_{L^2} \leq \frac{1}{k}\int_{t_{j-1}}^{t_j}\|u'(t) - P_hu'(t)\|_{L^2}dt, \qquad \|w_2^j\|_{L^2} \leq \frac{k}{k}\int_{t_{j-1}}^{t_j}\|u''(t)\|_{L^2}dt,$$

folgt damit

$$\begin{aligned}
\|\theta^n\|_{L^2} &\leq e^{-bt_n}\|\theta^0\|_{L^2} + \sum_{j=1}^{n} e^{-b(t_n - t_{j-1})}\int_{t_{j-1}}^{t_j}\|u'(t) - P_hu'(t)\|_{L^2} + k\|u''(t)\|_{L^2}dt \leq \\
&\leq e^{-bt_n}\|\theta^0\|_{L^2} + \sum_{j=1}^{n}\int_0^{t_n} e^{-b(t_n - t)}\|u'(t) - P_hu'(t)\|_{L^2} + k\|u''(t)\|_{L^2}dt
\end{aligned}$$

[4] $g(x) := \ln\left((1+x)^{-1/x}\right) = -\frac{1}{x}\ln(1+x)$; $g'(x) = \frac{\ln(1+x)-1+1/(1+x)}{x^2} \geq 0$ da $\frac{d}{dx}(\ln(1+x) - 1 + 1/(1+x)) = \frac{1}{1+x} - \frac{1}{(1+x)^2} \geq 0$ und $\lim_{x\to 0}(\ln(1+x) - 1 + 1/(1+x)) = 0$

$$
\begin{aligned}
\theta = 0 \text{ (expliziter Euler):} &\quad \mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + k\mathbf{A}\mathbf{u}^n = k\mathbf{F}(t_n) \\
\theta = 1 \text{ (impliziter Euler):} &\quad \mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + k\mathbf{A}\mathbf{u}^{n+1} = k\mathbf{F}(t_{n+1}) \\
\theta = 1/2 \text{ (Crank-Nicolson):} &\quad \mathbf{M}(\mathbf{u}^{n+1} - \mathbf{u}^n) + \frac{k}{2}(\mathbf{A}\mathbf{u}^n + \mathbf{A}\mathbf{u}^{n+1}) = k\mathbf{F}(t_{n+\frac{1}{2}})
\end{aligned}
$$

Abbildung 2.1: Matrixform des θ-Schemas für die Wärmeleitungsgleichung

Beweis von (2.17): Ausgangspunkt ist (2.22).

$$
\begin{aligned}
\underbrace{\langle \theta^{n+1} - \theta^n, \theta^{n+1} \rangle_{L^2}}_{\overset{\text{Taschenspielertrick}}{=} \frac{1}{2}(\|\theta^{n+1}\|^2 - \|\theta^n\|^2 + \|\theta^{n+1} - \theta^n\|^2)} + k|\theta^{n+1}|_{H^1}^2 &= k\langle w_1^{n+1}, \theta^{n+1} \rangle_{L^2} + k\langle w_2^{n+1}, \theta^{n+1} \rangle_{L^2} \\
&\leq k \left(\|w_1^{n+1}\|_{L^2} + \|w_2^{n+1}\|_{L^2} \right) \|\theta^{n+1}\|_{L^2} \\
&\leq kC_P |\theta^{n+1}|_{H^1} \left(\|w_1^{n+1}\|_{L^2} + \|w_2^{n+1}\|_{L^2} \right).
\end{aligned}
$$

Damit:

$$
\begin{aligned}
\|\theta^{n+1}\|_{L^2}^2 - \|\theta^n\|_{L^2}^2 + \|\theta^{n+1} - \theta^n\|_{L^2}^2 + 2k|\theta^{n+1}|_{H^1}^2 &\leq 2kC_p |\theta^{n+1}|_{H^1} \left(\|w_1^{n+1}\|_{L^2} + \|w_2^{n+1}\|_{L^2} \right) \\
&\leq k|\theta^{n+1}|_{H^1}^2 + \frac{k}{C_P^2} \left(\|w_1^{n+1}\|_{L^2} + \|w_2^{n+1}\|_{L^2} \right)^2
\end{aligned}
$$

Subtraktion von $k|\theta^{n+1}|_{H^1}$ auf beiden Seiten und anschliessende Summation liefert

$$
\|\theta^n\|_{L^2}^2 - \|\theta^0\|_{L^2}^2 + \sum_{j=0}^{n-1} \|\theta^{j+1} - \theta^j\|_{L^2}^2 + k|\theta^{j+1}|_{H^1}^2 \leq \frac{k}{C_P^2} \sum_{j=0}^{n-1} \left(\|w_1^{j+1}\|_{L^2} + \|w_2^{j+1}\|_{L^2} \right)^2
$$

Es ist

$$
k \sum_{J=0}^{n-1} \|w_1^{j+1}\|_{L^2}^2 + \|w_2^{j+1}\|_{L^2}^2 \leq \int_0^{t_n} \|u' - P_h u'\|_{L^2}^2 + k^2 \|u''\|_{L^2}^2 \, dt
$$

so dass sich ergibt

$$
\|\theta^n\|_{L^2}^2 + \sum_{j=0}^{n-1} \|\theta^{j+1} - \theta^j\|_{L^2}^2 + k|\theta^{j+1}|_{H^1}^2 \leq \|\theta^0\|_{L^2}^2 + \int_0^{t_n} \|u' - P_h u'\|_{L^2}^2 + k^2 \|u''\|_{L^2}^2 \, dt
$$

Setzt man $\rho^j := P_h u(t_j) - u(t_j)$, so folgt

$$
\|\rho^n\|_{L^2}^2 + \sum_{j=0}^{n-1} \underbrace{\|\rho^{j+1} - \rho^j\|_{L^2}^2}_{\leq k \int_{t_{j-1}}^{t_j} \|u' - P_h u'\|_{L^2}^2} + k|\rho^{j+1}|_{H^1}^2 \leq \|\rho^n\|_{L^2}^2 + k \int_0^{t_n} \|u' - P_h u'\|_{L^2}^2 + k \sum_{j=0}^{n-1} |u(t_j) - P_h u(t_j)|_{H^1}^2
$$

und damit mittels $u_h^{j+1} - u(t_{j+1}) = \theta^{j+1} + \rho^{j+1}$ und der Dreiecksungleichung die Behauptung. □

Korollar 2.23 *Sei $u \in C^3([0,T] \times \bar{\Omega})$. Dann gilt:*

(i) $\|u_h^n - u(t_n)\|_{L^2} \leq C[h + k]$

(ii) falls Ω konvex ist, dann ist $\|u_h^n - u(t_n)\|_{L^2} \leq C[h^2 + k]$

Beweis: Der Beweis verbleibt als Übung. □

$0 \le \theta < 1/2$	Verfahren *nicht* A-stabil	Ordnung 1
$\theta = 1/2$	Verfahren A-stabil	Ordnung 2
$1/2 < \theta \le 1$	Verfahren A-stabil (sogar L-stabil)	Ordnung 1

Abbildung 2.2: Stabilität und Konvergenzordnung des θ-Schemas.

2.3.2 das θ-Schema

Das implizite Eulerverfahren ist ein Spezialfall des sog. θ-Schemas, welches für die ODE $y' = f(t, y)$ so definiert ist:

$$y^{i+1} = y^i + kf(t_i + \theta(t_{i+1} - t_i), y_i + \theta(y_{i+1} - y_i)).$$

Für $\theta = 0$ ergibt sich das explizite Eulerverfahren, für $\theta = 1$ das implizite Eulerverfahren. Für $\theta = 1/2$ das Crank-Nicolson-Verfahren[5]–siehe Fig. 2.2. Aussagen analog zu Satz 2.22 gelten auch für das θ-Schema, z.B. kann man analog zu (2.17) in Satz 2.22 kann man folgende Stabilitätsaussage zeigen:

Übung 2.24 Seien die u_h^n mit dem θ-Schema und $\theta \in (1/2, 1]$ bestimmt. Setze $u_h^{n+\theta} := u_h^n + \theta(u_h^{n+1} - u_h^n)$ und $t_{j+\theta} := t_j + \theta(t_{j+1} - t_j)$. Dann gilt:

$$\|u_h^n\|_{L^2}^2 + \sum_{j=0}^{n-1} k|u_h^{j+\theta}|_{H^1}^2 + (2\theta - 1)\|u_h^{j+1} - u_h^j\|_{L^2}^2 \le \|u_h^0\|_{L^2}^2 + C\sum_{j=0}^{n-1} k\|f(t_{j+\theta})\|_{L^2}^2.$$

(Hinweis: betrachte (2.21) mit $v = \theta^{n+\theta} = u_h(t_{n+\theta}) - P_h u(t_{n+\theta})$.)

Bemerkung 2.25 Man kann die Konvergenzaussage (2.17) aus Satz 2.22 auch auf Übung 2.24 stützen. ∎

Mit geeigneter Regularität der Lösung erhält man für das Crank-Nicolson-Verfahren bessere Ordnung in k:

Satz 2.26 (CN ist Ordnung 2) *Seien die u_h^n definiert durch*

$$\frac{1}{k}\langle u_h^{n+1} - u_h^n, v\rangle_{L^2} + a\left(\frac{u_h^{n+1} - u_h^n}{2}, v\right) = \langle f(t_n + k/2), v\rangle_{L^2} \qquad \forall v \in V_h.$$

Dann gilt:

$$\|u_h^n - u(t_n)\|_{L^2} \le \|u_{0,h} - P_h u_0\|_{L^2} + \|u(t_n) - P_h u(t_n)\|_{L^2} + \int_0^{t_n} \|u' - P_h u'\|_{L^2} + Ck^2\|u'''\|_{L^2}\, dt$$

Beweis: Analog zu Satz 2.22, (2.16). Verwene $v = (\theta^{n+1} + \theta^n)/2$ in (2.21). □

2.3.3 Stabilität des θ-Schemas—die CFL-Bedingung

Für $0 \le \theta < 1/2$ ist das θ-Schema nur bedingt stabil, d.h. es gibt eine Schrittweitenbeschränkung ("CFL"-Bedingung[6]), die in der Praxis sehr restriktiv ist. Um dies zu sehen, betrachten wir einen Schritt des Verfahrens. Es hat die Form

$$u_h^{n+1} = Pu_h^n + kF^n$$

für einen linearen Operator $P : V_h \to V_h$. Man erhält also (wir lassen die Wahl der Norm offen...)

$$\|u_h^{n+1}\| \le \|P\|\|u_h^n\| + k\|F^n\|$$

[5] John Crank, Phyllis Nicolson. Im vorliegenden Fall einer linearen PDE stimmt das CN-Verfahren mit der impliziten Mittelpunktsregel (= 1-stufiges Gaussverfahren) überein.

[6] Richard Courant, Kurt Friedrichs, Hans Lewy

Name	$\mathbf{P}$	$\sigma(\mathbf{P})$	$\varrho(\mathbf{P})$	stabil?
$\mathbf{P}_{expl}$	$\mathbf{I} - k\mathbf{M}^{-1}\mathbf{A}$	$\{1 - k\lambda \mid \lambda \in \sigma\}$	$\lvert 1 - k\lambda_{max}\rvert$	falls $k \leq \frac{2}{\lambda_{max}}$
$\mathbf{P}_{impl}$	$(\mathbf{M} + k\mathbf{A})^{-1}\mathbf{M}$	$\left\{\frac{1}{1+k\lambda} \mid \lambda \in \sigma\right\}$	$\frac{1}{\lvert 1+k\lambda_{min}\rvert} \leq 1$	für alle $k > 0$
$\mathbf{P}_{CN}$	$\left(\mathbf{M} + \frac{k}{2}\mathbf{A}\right)^{-1}\left(\mathbf{M} - \frac{k}{2}\mathbf{A}\right)$	$\left\{\frac{1-k/2\lambda}{1+k/2\lambda} \mid \lambda \in \sigma\right\}$	≤ 1	für alle $k > 0$

Abbildung 2.3: Analyse von $\mathbf{P}$ in $\mathbf{u}^{n+1} = \mathbf{P}\mathbf{u}^n + \cdots$. $\sigma = \{\lambda : \exists \mathbf{x} \neq 0 \text{ mit } \mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}\}$

Gronwall liefert dann

$$\|u_h^N\| \leq \|P\|^N \|u_h^0\| + \cdots$$

Mit $t_N = Nk = T =$ Endzeitpunkt ergibt sich

$$\|u_h^N\| \leq \|P\|^{T/k} \|u_h^0\| + \cdots$$

Dies zeigt, dass $\|P\| \leq 1 + Ck$ sein sollte, denn dann ist

$$\|P\|^{T/k} \leq (1 + Ck)^{T/k} = (1 + Ck)^{TC/(Ck)} \leq e^{CT}$$

gleichmässig in k.[7]

Um die Analyse von $\|P\|$ einfach zu halten, untersuchen wir $\|P\|$ auf dem Matrixlevel. Die Iteration ist dann in "expliziter" Form $\mathbf{u}^{n+1} = \mathbf{P}\mathbf{u}^n + \cdots$, wobei die zughörige Matrix $\mathbf{P}$ (siehe Fig. 2.1 für die Iterationsvorschrift) in Fig. 2.3 gegeben ist.

Weil für jede Matrixnorm gilt $\|\mathbf{P}\| \leq \rho(\mathbf{P})$ mit dem Spektralradius $\rho(\mathbf{P}) = \max\{|\lambda| \,|\, \lambda \in \sigma(\mathbf{P})\}$, sehen wir, dass die Bedingung $\rho(\mathbf{P}) \leq 1 + Ck$ eine sinnvolle Stabilitätsbedingung ist. Tatsächlich wird man fordern:

$$\rho(\mathbf{P}) \leq 1. \tag{2.23}$$

Diese Bedingung wird in Fig. 2.3 ausgeführt. Wir erinnern an Satz 2.20, welcher $\lambda_{max} \sim C/h^2$ liefert. Damit kann die Bedingung (2.23) für das θ-Schema umgerechnet werden in eine Beziehung zwischen der Ortsschrittweite h und der Zeitschrittweite k. Für das explizite Eulerverfahren ergibt sich

$$\frac{k}{h^2} = O(1) \qquad \text{d.h.} \qquad k \leq Ck^2$$

Der Fall des impliziten Eulerverfahrens und des CN-Verfahrens ist in Fig. 2.3 ausgeführt. Die Stabilität von allgemeineren RK-Verfahren wird später behandelt.

Übung 2.27 Auch für $\theta \in (0, 1/2)$ muss man $k \leq Ch^2$ für Stabilität fordern.

Numerisches Beispiel

Wir wählen nun den Anfangswert für die 1D Wärmeleitungsgleichung $u_0 = 1$ und $f \equiv 0$ und betrachten die graphisch dargestellten Ergebnisse aus Abbildung 2.4. In den oberen beiden Graphiken ist zu sehen, wie sich das explizite Eulerverfahren zu zwei verschiedenen Schrittweiten knapp über bzw. knapp unter der Stabilitätsschranke verhält.

Links ist $k = \frac{2.001}{\lambda_{max}} \geq \frac{2}{\lambda_{max}}$ gewählt. Die Lösung oszilliert sehr stark (die größten Werte liegen im Bereich 10^7) und geben in keinster Weise das tatsächliche Lösungsverhalten wieder. In der rechten Graphik ist $k = \frac{1.999}{\lambda_{max}} \leq \frac{2}{\lambda_{max}}$ gewählt und die numerische und die exakte Lösung sind ziemlich ähnlich.

[7] andernfalls: $\|P\| \geq 1 + \delta$ für ein $\delta > 0$ unabhängig von k liefert $\|P\|^{T/k} \geq (1 + \delta)^{T/k} \to \infty$ für $k \to 0$.

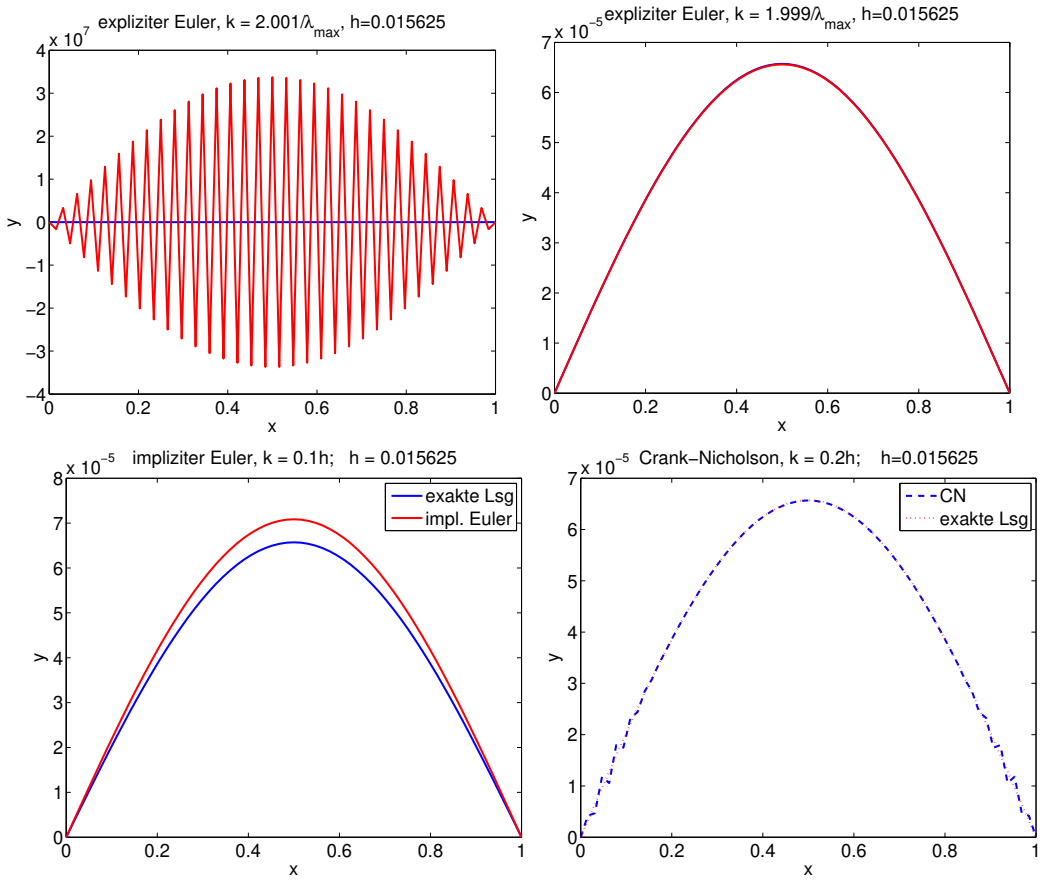


Abbildung 2.4: Vergleich zwischen exakter und numerischer Lösung

In der unteren linken Graphik von Abbildung 2.4 ist die Wärmeleitungsgleichung mit dem implizitem Eulerverfahren und in der rechten unteren Graphik mit dem Crank-Nicolson-Verfahren gelöst worden. Hier ist die numerische Lösung in beiden Fällen eine gute Approximation an die exakte Lösung.

Zum Abschluss betrachten wir noch das Konvergenzverhalten für $u(x,t) = e^{-t}x(1-x)$. In der linken Graphik von Abbildung 2.5 wird das explizite Eulerverfahren betrachtet. Wieder mit den zwei Zeitschrittweiten $k = \frac{2.001}{\lambda_{max}}$ und $k = \frac{1.999}{\lambda_{max}}$ knapp über bzw. knapp unter der Stabilitätsschranke. Der Fehler für die Lösung zur Wahl von $k = \frac{1.999}{\lambda_{max}}$ verhält sich wie $\mathcal{O}(h)$. Liegt k aber nur knapp über $\frac{2}{\lambda_{max}}$ konvergiert das explizite Eulerverfahren nicht mehr.

Bemerkung 2.28 In der Praxis will man aus Kostengründen möglichst grosse Zeitschritte machen. (Jeder Zeitschritt bedeutet ja das Lösen eines LGS!) Beim expliziten Eulerverfahren müsste man dann λ_{max} möglichst genau kennen, um das maximal zulässige k zu bestimmen. Das Beispiel zeigt, dass man k keinesfalls zu gross wählen darf. Man spricht deshalb auch von "hit or miss": also entweder hat man Glück und bestimmt k so, dass das Verfahren konvergiert, oder man liegt knapp daneben und das Verfahren konvergiert nicht. ∎

In der rechten Graphik von Abbildung 2.5 ist zu sehen, dass sich das implizite ebenso wie das explizite Eulerverfahren wie $\mathcal{O}(h)$ verhält. Das Crank-Nicolson-Verfahren hat Konvergenzordnung 2. Das waren auch unsere Erwartungen, denn der Fehler des expl. sowie des implizite Eulerverfahrens ist $\mathcal{O}(k+h^2)$ und jener des Crank-Nicolson-Verfahrens ist $\mathcal{O}(k^2+h^2)$.

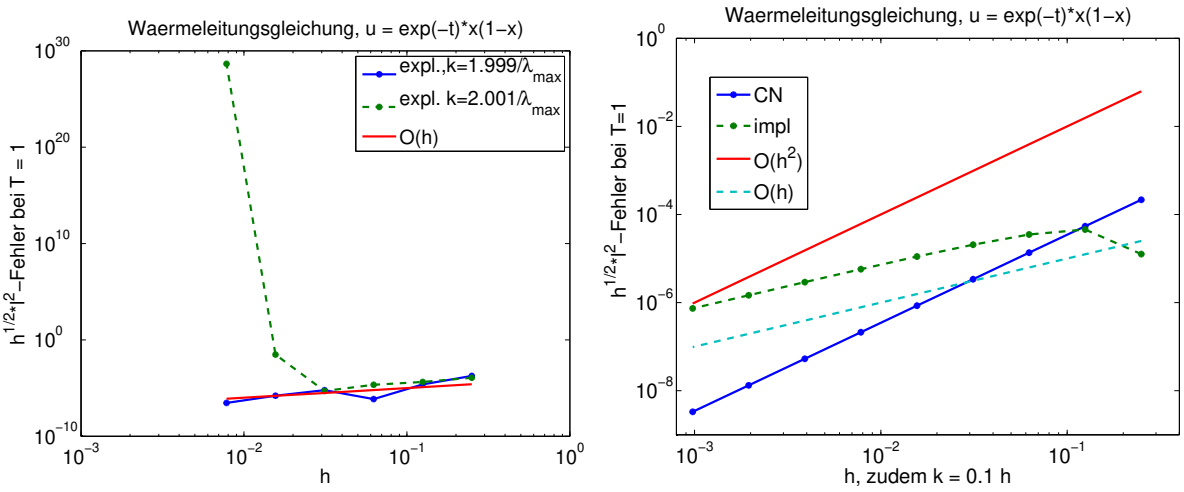


Abbildung 2.5: Konvergenzbetrachtung

2.4 Bemerkungen zum Lösungsbegriff

bisher: $u \in C^1((0,T); H_0^1(\Omega)) \cap C^0([0,T], H_0^1(\Omega))$

Beob.:

- dies fordert $u_0 \in H_0^1(\Omega)$ ("kompatible Anfangsbedingungen")
- $u_0 \in H_0^1(\Omega)$ ist unnatürlich vom Standpunkt der Anwendungen und der mathematischen Struktur (vgl. die Lösungsformeln in Bem. 2.9 und 2.10).

Notation: $(\varphi_n, \lambda_n)_{n\in\mathbb{N}} \subset H_0^1 \times \mathbb{R}$ Eigenpaare des Dirichletlaplaceoperators, $(\varphi_n)_n$ ONB von $L^2(\Omega)$. Für $v \in L^2(\Omega)$ mit Darstellung $v = \sum_n v_n \varphi_n$ ist

$$\begin{aligned}
\|v\|_{L^2(\Omega)}^2 &= \sum_n |v_n|^2, \\
|v|_{H^1(\Omega)}^2 &= \sum_n \lambda_n |v_n|^2 \\
\|v\|_{H^{-1}}^2 &:= \sup_{w\in H_0^1(\Omega)} \frac{\langle v, w\rangle_{L^2}}{|w|_{H^1}} = \sup_{(w_n)_n} \frac{\sum_n v_n w_n}{\sqrt{\sum_n \lambda_n |w_n|^2}} = \sqrt{\sum_n |v_n|^2 \lambda_n^{-1}}
\end{aligned}$$

Diese Beobachtungen realisieren folgende Isomorphismen:

$$\begin{aligned}
L^2(\Omega) &\simeq \ell^2, \\
H_0^1(\Omega) &\simeq \{(v_n)_n \mid \sum_n \lambda_n |v_n|^2 < \infty\}, \\
H^{-1}(\Omega) = (H_0^1(\Omega))^* &\simeq \{(v_n)_n \mid \sum_n \lambda_n^{-1} |v_n|^2 < \infty\}.
\end{aligned}$$

Wir erinnern an Lösungsoperator $E(t)u_0 = \sum_n e^{-\lambda_n t}\langle u_0, \varphi_n\rangle_{L^2}\varphi_n$ aus Bem. 2.9 für das Anfangswertproblem $u_t - \Delta u = 0$, $u(0) = u_0$. Man rechnet direkt nach, daß

Lemma 2.29 *Für $u_0 \in L^2(\Omega)$ gilt*

$$t \mapsto E(t)u_0 \in C^\infty((0,T); H_0^1(\Omega)), \tag{2.24}$$
$$t \mapsto E(t)u_0 \in L^2((0,T); H_0^1(\Omega)), \tag{2.25}$$
$$t \mapsto \partial_t(E(t)u_0) \in L^2((0,T); H^{-1}(\Omega)), \tag{2.26}$$
$$t \mapsto E(t)u_0 \in C([0,T]; L^2(\Omega)). \tag{2.27}$$

Die Normen in (2.25)–(2.27) können durch $C\|u_0\|_{L^2}$ abgeschätzt werden.

Beweis: Wir illustrieren (2.26):

$$\begin{aligned}
\partial_t E(t)u_0 &= \sum_n -\lambda_n e^{-\lambda_n t}\langle u_0, \varphi_n\rangle_{L^2}\varphi_n, \\
\Longrightarrow \quad \int_0^T \|\partial_t E(t)u_0\|_{H^{-1}(\Omega)}^2 \, dt &= \int_0^T \sum_n \lambda_n^2 e^{-2\lambda_n t} \frac{|\langle u_0, \varphi_n\rangle_{L^2}|^2}{\lambda_n} \, dt \le \frac{1}{2}\|u_0\|_{L^2}^2.
\end{aligned}$$

□

Motivation: Lemma 2.29 zeigt:

- $u_0 \in L^2(\Omega)$ ist sinnvoll
- man kann $u \in C([0,T]; L^2(\Omega))$ erhoffen
- man kann $u \in L^2((0,T); H_0^1(\Omega))$ und $u' \in L^2((0,T); H^{-1}(\Omega))$ erwarten

Tatsächlich ist bei richtiger (d.h. distributioneller) Interpretation[8]von ∂_t eine gute Formulierung der Wärmeleitungsgleichung:

Finde $u \in L^2((0,T);H_0^1(\Omega))$ mit $u' \in L^2((0,T);H^{-1}(\Omega))$, sodaß

$$\int_0^T \varphi(t)\left[\langle u'(t),v\rangle_{H^{-1}\times H_0^1} + a(u(t),v)\right]\,dt = \int_0^T \varphi(t)\langle f(t),v\rangle_{H^{-1}\times H_0^1} \qquad \forall v \in H_0^1(\Omega), \quad \forall \varphi \in C_0^\infty(0,T), \tag{2.28a}$$

$$u(0) = u_0 \in L^2(\Omega) \tag{2.28b}$$

wobei $u_0 \in L^2(\Omega)$ und $f \in L^2((0,T);H^{-1}(\Omega))$ gegeben sind. Die Forderung (2.28b) ist sinnvoll, denn es gilt: $u \in L^2((0,T);H_0^1)$ mit $u' \in L^2((0,T);H^{-1})$ impliziert $u \in C([0,T];L^2(\Omega))$ (vgl. PDE Vorlesung oder Buch von Evans).

Variiert man φ in (2.28), erhält man die Formulierung der Wärmeleitungsgleichung, die man oft in den Lehrbüchern findet: Finde $u \in L^2((0,T);H_0^1(\Omega))$ mit $u' \in L^2((0,T);H^{-1}(\Omega))$, sodaß

$$\langle u'(t),v\rangle_{H^{-1}\times H_0^1} + a(u(t),v) = \langle f(t),v\rangle_{H^{-1}\times H_0^1} \qquad \forall v \in H_0^1(\Omega), \quad \text{für fast alle } t \in (0,T), \tag{2.29a}$$

$$u(0) = u_0 \in L^2(\Omega) \tag{2.29b}$$

[8]siehe folgendes Kapitel. Für eine separablen Hilberträume X und eine bochnerintegrierbare Funktion $u \in L^1((0,T);X)$ ist die distributionelle Ableitung die lineare Abbildung $C_0^\infty(0,T) \to X$ geg. durch $\varphi \mapsto -\int_0^T u(t)\varphi'(t)\,dt$, wobei das Integral ein Bochnerintegral ist. Mit einer stetigen Einbettung $\iota_{X\to Y} : X \subset Y$ sagen wir $u' \in L^1((0,T);Y)$, falls es ein $v \in L^1((0,T);Y)$ gibt mit $-\iota_{X\to Y}\int_0^T u(t)\varphi'(t)\,dt = \int_0^T v(t)\varphi(t)\,dt$ für alle $\varphi \in C_0^\infty(0,T)$. Man schreibt dann kurz: $u' = v$.

2.5 nichtglatte Anfangsdaten

Ziel: Eine Konvergenztheorie für *realistische* Startwerte u_0, d.h. $u_0 \in L^2$. Die Konvergenztheorie soll auf dem Abschätzen der Fehler durch Semidiskretisierung aufbauen:

$$\|u(t_n) - u_h^n\| \le \|u(t_n) - u_h(t_n)\| + \|u_h(t_n) - u_h^k\|$$

Wir werden in Abschnitt 2.5.4 den Fehler $\|u(t_n) - u_h(t_n)\|_{L^2}$ kontrollieren und in Abschnitt 2.5.5 den Fehler $\|u_h(t_n) - u_h^n\|_{L^2}$.

Problem bei $\|u(t_n) - u_h(t_n)\|$: Die Lösungstheorie zeigt, dass mit $V = H_0^1(\Omega)$ und $H = L^2(\Omega)$ das "natürliche" setting ist: $u \in L^2(0,T;V)$ und $u' \in L^2(0,T;V')$ und $u \in C([0,T];H)$. Wir können eigentlich *nicht* erwarten, dass u (oft) differenzierbar ist *mit Werten in* V. Das war allerdings unser Startpunkt für die Analyse des Semidiskretisierungsfehlers $u - u_h$ in Satz 2.15:

$$\|u(t) - u_h(t)\|_{L^2} \le \|u_0 - P_h u_0\|_{L^2} e^{-\gamma t} + \|u(t) - P_h u(t)\|_{L^2} + \int_0^t e^{-\gamma(t-s)} \|u'(s) - P_h u'(s)\|_{L^2}\, ds;$$

für sinnvolle Abschätzungen benötigt man, dass $u_0 \in V$ (und nicht in H!) und dass $u' \in L^1((0,T);L^2(\Omega))$.

Vereinfachung: Wir werden uns im vorliegenden Abschnitt auf den Speziallfall $f \equiv 0$ konzentieren, bei dem die für die Wärmleitungsgleichung typischen Phänomene besonders ausgeprägt auftreten. Qualitativ ähnliche Aussagen gelten auch für numerische Verfahren mit hinreichend glatten $f \neq 0$.

2.5.1 Glättungseigenschaft

Typisch für parabolische Probleme ist die *Glättungseigenschaft*, d.h. die oben angedeuteten Schwierigkeiten treten nur bei $t = 0$ auf. Um zu sehen, welche Regularität man erwarten kann, betrachten wir den den Lösungsoperator $E(t) : u_0 \mapsto u(t)$, welcher die folgende Form hat[9]

$$E(t)u_0 = \sum_{n=1}^{\infty} e^{-\lambda_n t} (u_0, \varphi_n)_{L^2} \varphi_n.$$

Es stellt sich heraus, dass die Ableitungen der Lösung $E(t)u_0$ in geeignet *gewichteten* V-wertigen Räumen sind:

Lemma 2.30 *(i) Für jedes $u_0 \in L^2(\Omega)$ ist $t \mapsto E(t)u_0$ in $C^\infty((0,\infty);H_0^1(\Omega))$ und auch in $C^\infty((0,\infty);L^2(\Omega))$.*

(ii) für jedes $m \in \mathbb{N}_0$ ist $\|\frac{d^m}{dt^m} E(t)u_0\|_{H^1(\Omega)} \le C_m t^{-1/2-m} \|u_0\|_{L^2(\Omega)}$.

(iii) für jedes $m \in \mathbb{N}_0$ ist $\|\frac{d^m}{dt^m} E(t)u_0\|_{L^2(\Omega)} \le C_m t^{-m} \|u_0\|_{L^2(\Omega)}$.

(iv) $\int_0^t \|E(s)u_0\|_{H^1}^2\, ds \le C\|u_0\|_{L^2}^2$.

(v) $\int_0^t s^2 \|\frac{d}{dt} E(s)u_0\|_{H^1}^2\, ds \le C\|u_0\|_{L^2}^2$.

Beweis: Übung. Beachte, dass

$$\|z\|_{L^2(\Omega)}^2 = \sum_n |(z,\varphi_n)_{L^2}|^2, \qquad |z|_{H^1(\Omega)}^2 = \sum_n \lambda_n |(z,\varphi_n)_{L^2}|^2.$$

Daraus ergibt sich z.B. für (iv)

$$\int_0^t |E(t)u_0|_{H^1(\Omega)}^2\, ds = \int_0^t \sum_n \lambda_n |(E(s)u_0,\varphi_n)_{L^2}|^2\, ds = \int_0^t \sum_n \lambda_n e^{-2\lambda_n s} |(u_0,\varphi_n)_{L^2}|^2\, ds \le C \sum_n |(u_0,\varphi_n)_{L^2}|^2$$

Für (ii), (iii) verwendet man $\sup_{x>0} xe^{-x} < \infty$. □

[9] Erinnerung: für $f \neq 0$ liefert das Duhamelsche Prinzip auch eine Lösungsformel

Bemerkung 2.31 Falls $\partial\Omega \in C^\infty$, dann sind die Eigenfunktionen φ_n auch glatt (auf $\overline{\Omega}$), und man kann zeigen: $E(t)u_0 \in C^\infty((0,\infty); H^k(\Omega))$ für jedes k. Die Singularität bei $t = 0$ bleibt jedoch erhalten. ∎

Frage: ist es trotzdem möglich, dass $t \mapsto E(t)u_0 \in C^m([0,T]; V)$ für ein $m \in \mathbb{N}_0$?
Antwort: kompatible Anfangswerte u_0

Lemma 2.32 *Definiere für $m \in \mathbb{N}_0$ den Raum*

$$\widehat{H}^m(\Omega) := \{u \in L^2(\Omega) \mid \sum_n \lambda_n^m |(u, \varphi_n)_{L^2}|^2 < \infty\}, \qquad \|u\|^2_{\widehat{H}^m(\Omega)} := \sum_n \lambda_n^m |(u, \varphi_n)_{L^2}|^2 \tag{2.30}$$

Dann gilt:

(i) Für $u_0 \in L^2$ gilt $E(t)u_0 \in \widehat{H}^m(\Omega)$ für jedes $m \in \mathbb{N}_0$ und $t > 0$.

(ii) Falls $u_0 \in \widehat{H}^m(\Omega)$, dann gilt für $\ell, m, q \geq 0$

$$\|\frac{d^\ell}{dt^\ell} E(t)u_0\|_{\widehat{H}^q(\Omega)} \leq Ct^{-(q-m)/2-\ell} \|u_0\|_{\widehat{H}^m(\Omega)}, \qquad t > 0.$$

Beweis: Übung. Man bemerkt, dass $\sup_{x>0} x^{q-m+2\ell} e^{-2x} < \infty$. □

Bemerkung 2.33

- für $m = 0$ ist $\widehat{H}^0(\Omega) = L^2(\Omega)$
- für $m = 1$ ist $\widehat{H}^1(\Omega) = H_0^1(\Omega)$
- für $m > 1$ ist $\widehat{H}^m(\Omega) \subset H_0^1(\Omega)$ nicht einfach zu beschreiben. Für glatte $\partial\Omega$ gilt

  $$\widehat{H}^m(\Omega) = \{u \in H^m(\Omega) \mid (\Delta^j u)|_{\partial\Omega} = 0 \quad 0 \leq j < m/2\}.$$

  Insbesondere ist auf dem Raum $\widehat{H}^m(\Omega)$ die Norm $\|\cdot\|_{H^m(\Omega)}$ äquivalent zur Norm $\|\cdot\|_{\widehat{H}^m(\Omega)}$ (vgl. [1, Lemma 3.1]).
- für $m = 2$ ist $\widehat{H}^m(\Omega) = H^2(\Omega) \cap H_0^1(\Omega)$.

Man sieht, dass die Forderung $u_0 \in \widehat{H}^m(\Omega)$ eine unnatürliche *Kompatibilitäts*bedingung auf $\partial\Omega$ darstellt.

∎

Offensichtlich sind für $u_0 = 0$ *alle* Kompatibilitätsbedingungen erfüllt ($\rightarrow$ "versteckte" Regularität). Mit dem Duhamelschen Prinzip können wir deshalb folgende Regularitätsaussage erhalten:

Lemma 2.34 *Sei $u_0 = 0$ und $f \in L^2((0,T); L^2(\Omega))$. Dann erfüllt die Lösung $u(t) = \int_0^t E(t-s)f(s)\,ds$:*

$$\|u(t)\|^2_{L^2(\Omega)} \leq Ct \int_0^t \|f(s)\|^2_{L^2(\Omega)}\,ds, \qquad t \in (0,T], \tag{2.31}$$

$$\|u(t)\|^2_{H^1(\Omega)} \leq C \int_0^t \|f(s)\|^2_{L^2(\Omega)}\,ds, \qquad t \in (0,T], \tag{2.32}$$

$$\int_{s=0}^t \|u'(s)\|^2_{L^2(\Omega)}\,ds \leq C \int_0^t \|f(s)\|^2_{L^2(\Omega)}\,ds. \tag{2.33}$$

Beweis: Übung. Schreiben Sie $f_n(s) = (f(s), \varphi_n)_{L^2}$. Starten Sie von $u(t) = \sum_n \varphi_n \int_0^t f_n(s) e^{-\lambda_n(t-s)}, ds$. Für (2.32), (2.33) überlegen Sie sich, dass Sie schlussendlich $\sum_n \int_0^t \|f_n(s)\|^2\,ds \lambda_n \int_0^t e^{-2\lambda_n(t-s)}\,ds$ abschätzen müssen. □

Übung 2.35 Sei $u_{h,0} = 0$. Dann gilt für die semidiskrete Approximation $u_h(t) = \int_0^t E_h(t-s)f(s) = \int_0^t E_h(t-s)\Pi^{L^2} f(s)\,ds$

$$\begin{aligned}
\|u_h(t)\|^2_{L^2(\Omega)} &\leq Ct \int_0^t \|\Pi^{L^2} f(s)\|^2_{L^2(\Omega)}\,ds, \qquad t \in (0,T],\\
\|u_h(t)\|^2_{H^1(\Omega)} &\leq C \int_0^t \|\Pi^{L^2} f(s)\|^2_{L^2(\Omega)}\,ds, \qquad t \in (0,T],\\
\int_{s=0}^t \|u_h'(s)\|^2_{L^2(\Omega)}\,ds &\leq C \int_0^t \|\Pi^{L^2} f(s)\|^2_{L^2(\Omega)}\,ds.
\end{aligned}$$

■

2.5.2 Reduktion auf die Analyse von ρ

Ausgangspunkt der Fehleranalyse ist die Fehlergleichung: Wie im Beweis von Satz 2.15 schreiben wir

$$e_h(t) := u_h(t) - u(t) = \underbrace{u_h(t) - P_h u(t)}_{=\theta(t)} + \underbrace{P_h u(t) - u(t)}_{=\rho(t)} \tag{2.34}$$

und erinnern uns an die Fehlergleichung, die dort bewiesen wurde:

$$(\theta'(t), v)_{L^2} + a(\theta(t), v) = -(\rho'(t), v)_{L^2} \qquad \forall v \in V_h. \tag{2.35}$$

Für eine Abschätzung von θ tritt die Schwierigkeit auf, dass $\theta(0)$ nicht definiert ist. Eine der Kernbeobachtungen von Lemma 2.30 ist, dass wir u' zwar nicht in $L^2(0,T;V)$ oder $L^2(0,T;H)$ erwarten können, dass aber Schwierigkeiten nur bei $t = 0$ zu erwarten sind, d.h. man kann mit gewichteten Räumen arbeiten. So sind z.B. $\int_0^t s^2\|u'(s)\|^2_{L^2}\,ds$ und $\int_0^t s^2\|u(s)\|^2_{H^1}\,ds$ kontrollierbare Objekte. Wir verfeinern deshalb die Abschätzungen aus Satz 2.15 und Übung 9.5 dahingehend, dass wir mit gewichteten Normen arbeiten.

Lemma 2.36 *Es ist für alle $t > 0$*

$$\int_0^t \|\theta(s)\|^2_{L^2}\,ds \leq Ct\|\Pi^{L^2} e_h(0)\|^2_{L^2} + C\int_0^t \|\rho(s)\|^2_{L^2}\,ds, \tag{2.36}$$

$$\begin{aligned}
&t\|\theta(t)\|^2_{L^2} + \int_0^t s|\theta(s)|^2_{H^1}\,ds\\
&\qquad \leq C\left[\int_0^t s^2\|\rho'(s)\|^2_{L^2} + \|\rho(s)\|^2_{L^2}\,ds + \sup_{s\in(0,t)} s\|\rho(s)\|^2_{L^2} + t\|\Pi^{L^2} e_h(0)\|^2_{L^2}\right].
\end{aligned} \tag{2.37}$$

Beweis: *ad (2.36):* Der "Trick" beim Beweis von (2.36) besteht in einem "parabolischen Dualitätsargument[10]". Sei $t = t_0$ fest gewählt. Betrachte die "Rückwärtsgleichung"

$$\begin{aligned}
-z_t - \Delta z &= \theta \quad \text{auf } \Omega \times (0,t_0), &(2.38)\\
z &= 0 \quad \text{auf } \partial\Omega \times (0,t_0), &(2.39)\\
z(t_0) &= 0. &(2.40)
\end{aligned}$$

Betrachte die semidiskrete Approximation an z, d.h. die Funktion $z_h \in C^1((0,t_0);V_h) \cap C^0([0,t_0];V_h)$, welche die Gleichung

$$\begin{aligned}
-(z_h', v)_{L^2} + a(z_h, v) &= (\theta, v)_{L^2} \qquad \forall v \in V_h, \quad \text{auf } (0,t_0) &(2.41)\\
z_h(t_0) &= 0 &(2.42)
\end{aligned}$$

[10] Dualitätsargumente sind auch nur eine clevere Wahl der Testfunktion...

löst. Dann ergibt sich für $s \in (0, t_0)$

$$\begin{aligned}
\|\theta(s)\|_{L^2}^2 &= -(z_h'(s), \theta(s))_{L^2} + a(z_h(s), \theta(s)) \\
&= -\frac{d}{dt}(z_h(s), \theta(s))_{L^2} + (z_h(s), \theta'(s))_{L^2} + a(z_h(s), \theta(s)) \\
&= -\frac{d}{dt}(z_h(s), \theta(s))_{L^2} - (\rho'(s), z_h(s))_{L^2} \\
&= -\frac{d}{dt}(z_h(s), \theta(s))_{L^2} + (\rho(s), z_h'(s))_{L^2} - \frac{d}{dt}(z_h(s), \rho(s))_{L^2} \\
&= -\frac{d}{dt}(z_h(s), e_h(s))_{L^2} + (\rho(s), z_h'(s))_{L^2}.
\end{aligned}$$

Für $0 < \varepsilon < t_0$ liefert Integration und $z_h(t_0) = 0$

$$\int_\varepsilon^{t_0} \|\theta(s)\|_{L^2}^2 \leq \underbrace{(z_h(\varepsilon), e_h(\varepsilon))_{L^2}}_{\to (z_h(0), e_h(0))_{L^2} \text{ wg. } e_h \in C([0,T]; L^2)} + \underbrace{\int_\varepsilon^{t_0} \|\rho(s)\|_{L^2} \|z_h'(s)\|_{L^2}\, ds}_{\leq \sqrt{\int_0^{t_0} \|\rho\|_{L^2}^2} \sqrt{\int_0^{t_0} \|z_h'\|_{L^2}^2}}$$

Weil $z_h(0) \in V_h$ ist also $(z_h(0), e_h(0))_{L^2} = (z_h(0), \Pi^{L^2} e_h(0))_{L^2}$. Übung 2.35 liefert

$$\int_0^{t_0} \|z_h'(s)\|_{L^2}^2\, ds + t_0^{-1} \|z_h(0)\|_{L^2}^2 \leq C \int_0^{t_0} \|\theta(s)\|_{L^2}^2\, ds.$$

Damit ergibt sich

$$\int_0^{t_0} \|\theta(s)\|_{L^2}^2\, ds \leq C \left[\sqrt{\int_0^{t_0} \|\rho(s)\|_{L^2}^2\, ds} + \sqrt{t_0} \|\Pi^{L^2} e_h(0)\|_{L^2} \right] \sqrt{\int_0^{t_0} \|\theta(s)\|_{L^2}^2\, ds}$$

ad (2.37): Der Beweis verläuft wie in Satz 2.15. Die Wahl $v = t\theta(t)$ in (2.35) liefert

$$\frac{1}{2} \frac{d}{dt} (t \|\theta(t)\|_{L^2}^2) + t a(\theta(t), \theta(t)) = -t(\rho'(t), \theta(t))_{L^2} + \frac{1}{2} \|\theta(t)\|_{L^2}^2$$

Integration über (ε, t) liefert

$$\frac{1}{2} t \|\theta(t)\|_{L^2}^2 + \int_\varepsilon^t s |\theta(s)|_{H^1}^2\, ds = \frac{1}{2} \varepsilon \|\theta(\varepsilon)\|_{L^2}^2 - \underbrace{\int_\varepsilon^t (s\rho'(s), \theta(s))_{L^2}\, ds}_{\leq \sqrt{\int_0^t s^2 \|\rho'(s)\|_{L^2}^2\, ds} \sqrt{\int_0^t \|\theta(s)\|_2^2\, ds}} + \frac{1}{2} \int_\varepsilon^t \|\theta(s)\|_{L^2}^2\, ds$$

Es bleibt, eine Abschätzung für $\limsup_{\varepsilon \to 0} \varepsilon \|\theta(\varepsilon)\|_{L^2}^2$ zu finden. Wir schreiben $\theta = u_h - P_h u = u - u_h - \rho = e_h - \rho$ und erhalten wegen $e_h \in C([0,T]; L^2(\Omega))$

$$\limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|\theta(\varepsilon)\|_{L^2} \leq \limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|e_h(\varepsilon)\|_{L^2} + \limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|\rho(\varepsilon)\|_{L^2} = \limsup_{\varepsilon \to 0} \sqrt{\varepsilon} \|\rho(\varepsilon)\|_{L^2} \leq \sup_{s \in (0,t)} \sqrt{s} \|\rho(s)\|_{L^2}.$$

Insgesamt ergibt sich mit der Young-Ungleichung und (2.36):

$$t \|\theta(t)\|_{L^2}^2 + \int_0^t s |\theta(s)|_{H^1}^2\, ds \leq C \left[\int_0^t s^2 \|\rho'(s)\|_{L^2}^2 + \|\rho\|_{L^2}^2\, ds + \sup_{s \in (0,t)} s \|\rho(s)\|_{L^2}^2 + t \|\Pi^{L^2} e_h(0)\|_{L^2}^2 \right].$$

□

2.5.3 Rückblick: semidiskrete Konvergenzresultate mit kompatiblen Anfangsdaten

Wie Satz 2.15 und Lemma 2.36 zeigen, muss ρ kontrolliert werden. Satz 2.15 benötigt jedoch die Glattheit der gesamten Lösung u als Voraussetzung. Im Folgenden werden wir nur die Glattheit der kompatiblen Anfangsdaten benutzen. Damit werden die Approximationseigenschaften des Ritzprojektors P_h wichtig. Die folgende Notation fixiert den Parameter r — er ergibt sich allgemein aus dem (elliptischen) Dualitätsargument ("Nitsche-Trick"):

Notation 2.37 *Sei $r \in (0,1]$ so, dass*

$$\|v - P_h v\|_{L^2} \le Ch^r \|v - P_h v\|_{H^1} \le Ch^r \|v\|_{H^1}$$

Für konvexe oder glatt berandete Gebiete gilt $r = 1$ in Notation 2.37, falls typische FEM-Räume gewählt werden und h die Gitterweite darstellt. Um später auf diesen Fall verweisen zu können, formulieren folgende Annahme, welche für typische FEM-Räume (basierend auf stückweise linearen Ansatzfunktionen) für konvexe Gebiete oder glatte berandete Gebiete erfüllt ist:

Voraussetzung 2.38 *Es gilt Notation 2.37 mit $r = 1$ und*

$$\|v - P_h v\|_{L^2} \le Ch^2 \|v\|_{\widehat{H}^2(\Omega)} \qquad \forall v \in \widehat{H}^2(\Omega).$$

Für *kompatible* Anfangsdaten $u_0 \in \widehat{H}^2(\Omega)$ erhält man die optimale Rate:

Satz 2.39 *Es gelte Voraussetzung 2.38. Sei $f \equiv 0$ und $u_0 \in \widehat{H}^2(\Omega)$. Sei zusätzlich $u_{h,0} = \Pi^{L^2} u_0$ (es geht auch $u_{h,0} = P_h u_0$). Dann ist*

$$\|u(t) - u_h(t)\|_{L^2(\Omega)} \le Ch^2 \|u_0\|_{\widehat{H}^2(\Omega)}$$

Beweis: Wir verwenden Lemma 2.36 und die Eigenschaft aus Bemerkung 2.33, dass die volle H^2-Norm äquivalent zur $\|\cdot\|_{\widehat{H}^2}$-Norm ist. Mit Annahme 2.38 und Lemma 2.32 folgt

$$\begin{aligned}
\|e_h(0)\|_{L^2} &\le Ch^2 \|u_0\|_{\widehat{H}^2(\Omega)}, \\
\int_0^t \|\rho(s)\|_{L^2}^2 \, ds &\le Ch^4 \int_0^t \|E(s)u_0\|_{\widehat{H}^2(\Omega)}^2 \, ds \le Ch^4 \int_0^t \|u_0\|_{\widehat{H}^2(\Omega)}^2 \le Ch^4 t \|u_0\|_{\widehat{H}^2(\Omega)}^2, \\
\int_0^t s^2 \|\rho'(s)\|_{L^2}^2 \, ds &\le Ch^4 \int_0^t s^2 \|\frac{d}{dt} E(s)u_0\|_{\widehat{H}^2(\Omega)}^2 \, ds \le Ch^4 \int_0^t s^2 s^{-2} \|u_0\|_{\widehat{H}^2(\Omega)}^2 \le Ch^4 t \|u_0\|_{\widehat{H}^2(\Omega)}^2, \\
\sup_{s\in(0,t)} s\|\rho(s)\|_{L^2(\Omega)}^2 &\le Ch^4 \sup_{s\in(0,t)} s\|E(s)u_0\|_{\widehat{H}^2(\Omega)}^2 \le Ch^4 t \|u_0\|_{\widehat{H}^2(\Omega)}^2,
\end{aligned}$$

was den Beweis abschliesst. □

2.5.4 Semidiskrete Konvergenzresultate mit inkompatiblen Anfangsdaten

Satz 2.40 *Sei r wie in Notation 2.37. Sei $u_0 \in L^2(\Omega)$ und $f \equiv 0$. Sei u_h die semidiskrete Approximation mit Startwert $u_{h,0} = \Pi^{L^2} u_0$. Dann gilt:*

$$\|u(t) - u_h(t)\|_{L^2} \le Ch^r t^{-1/2} \|u_0\|_{L^2}. \tag{2.43}$$

Falls zusätzlich Voraussetzung 2.38 gilt, dann gilt sogar

$$\|u(t) - u_h(t)\|_{L^2} \le Ch^2 t^{-1} \|u_0\|_{L^2}. \tag{2.44}$$

Beweis: *ad (2.43):* Wegen $u_h(t) - u(t) = \theta(t) + \rho(t)$ und Lemma 2.36 müssen wir $\|\rho(s)\|_{L^2}$, $\|\rho'(s)\|_{L^2}$ abschätzen. Der zusätzliche Term $\Pi^{L^2} e_h(0)$ verschwindet wegen der Wahl $u_{h,0} = \Pi^{L^2} u_0$!

Für $s > 0$ ist $u(s) = E(s)u_0 \in H_0^1(\Omega)$ und nach Voraussetzung

$$\|\rho(s)\|_{L^2} \le Ch^r \|u(s)\|_{H^1}, \qquad \|\rho'(s)\|_{L^2} \le Ch^r \|u'(s)\|_{H^1}.$$

Mittels Lemma 2.30 schliessen wir

$$t\|\theta(t)\|_{L^2}^2 \quad \leq \quad \int_0^t \|\rho(s)\|_{L^2}^2\,ds + \int_0^t s^2\|\rho'(s)\|_{L^2}^2\,ds + \sup_{0<s<t} s\|\rho(s)\|_{L^2}^2 \leq Ch^{2r}\|u_0\|_{L^2}^2.$$

ad (2.44): Zuerst bemerken wir, dass die Aussage (2.43) mit $r = 1$ gilt. Zu dem Evolutionsoperator $E(t)$ und dem diskreten Evolutionsoperator $E_h(t)$ führen wir noch den Fehleroperator $F_h(t)$ ein:

$$F_h(t)v := E_h(t)\Pi^{L^2}v - E(t)v$$

(der Operator ist nichts anderes als $u(t) - u_h(t)$, wenn $u(0) = v$ und $u_h(0) = \Pi^{L^2}v$ gewählt wird). Unser Ziel ist,

$$\|F_h(t)v\|_{L^2} \leq Ch^2t^{-1}\|v\|_{L^2} \qquad \forall v \in L^2 \tag{2.45}$$

zu zeigen. Zuerst bemerken wir, dass mit der Dreiecksungleichung und elementaren Eigenschaften von $E(t)$ und $E_h(t)$ folgt:

$$\|F_h(t)v\|_{L^2} \leq 2\|v\|_{L^2} \tag{2.46}$$

Wir können uns also auf den Fall $h^2t^{-1} \leq 1$ beschränken. Wir nutzen die Halbgruppeneigenschaften von E und E_h aus, d.h. $E(t+s) = E(t) \circ E(s)$ und $E_h(t+s) = E_h(t) \circ E_h(s)$. Damit gilt:

$$F_h(t) = F_h(t/2)E(t/2) + E(t/2)F_h(t/2) + (F_h(t/2))^2, \tag{2.47}$$

denn

$$\begin{aligned}
&F_h(t/2)E(t/2) + E(t/2)F_h(t/2) + F_h(t/2)F_h(t/2)\\
&= (E_h(t/2)\Pi^{L^2} - E(t/2))E(t/2) + E(t/2)(E_h(t/2)\Pi^{L^2} - E(t/2)) + (E_h(t/2)\Pi^{L^2} - E(t/2))^2\\
&= E_h(t/2)\Pi^{L^2}E(t/2) - E(t) + E(t/2)E_h(t/2)\Pi^{L^2} - E(t) +\\
&\quad E_h(t/2)\Pi^{L^2}E_h(t/2)\Pi^{L^2} - E_h(t/2)\Pi^{L^2}E(t/2) - E(t/2)E_h(t/2)\Pi^{L^2} + E(t)\\
&= E_h(t/2)\Pi^{L^2}E_h(t/2)\Pi^{L^2} - E(t) = E_h(t/2)E_h(t/2)\Pi^{L^2} - E(t) = E_h(t)\Pi^{L^2} - E(t) = F_h(t).
\end{aligned}$$

Die Glättungseigenschaft liefert $E(t/2)u_0 \in \widehat{H}^2(\Omega)$. Damit liefern Satz 2.39 und Lemma 2.32

$$\|F_h(t/2)E(t/2)u_0\|_{L^2} \leq Ch^2\|E(t/2)u_0\|_{\widehat{H}^2(\Omega)} \leq Ch^2t^{-1}\|u_0\|_{L^2}$$

Nun sind E_h und F_h bzgl. des L^2-Innenproduktes selbstadjungiert (Übung!). Damit ist $\|E(t/2)F_h(t/2)\|_{L^2} = \|F_h(t/2)E(t/2)\|_{L^2}$ und damit

$$\|E(t/2)F_h(t/2)u_0\|_{L^2} \leq Ch^2t^{-1}\|u_0\|_{L^2}$$

Es bleibt, $\|(F_h(t/2))^2u_0\|_{L^2}$ abzuschätzen. Aus (2.43) mit $r = 1$ folgt

$$\|F_h(t/2)F_h(t/2)u_0\|_{L^2} \leq Cht^{-1/2}\|F_h(t/2)u_0\|_{L^2}$$

Insgesamt haben wir erhalten:

$$\|F_h(t)u_0\|_{L^2} \leq Ch^2t^{-1}\|u_0\|_{L^2} + Cht^{-1/2}\|F_h(t/2)u_0\|_{L^2}. \tag{2.48}$$

Einsetzen diese Abschätzung in sich selbst liefert

$$\|F_h(t)u_0\|_{L^2} \leq Ch^2t^{-1}\|u_0\|_{L^2} + Ch^2t^{-1}\|F_h(t/4)u_0\|_{L^2}.$$

Wir erinnern nun daran, dass nach (2.46) $\|F_h(t)\|_{L^2} \leq 2$ unabhängig von t. Somit ergibt sich die gewünschte Abschätzung $\|F_h(t)u_0\|_{L^2} \leq Ch^2t^{-1}\|u_0\|_{L^2}$. □

Übung 2.41 Zeigen Sie: Falls der Raum V_h die Approximationseigenschaften

$$\|v - P_h v\|_{L^2(\Omega)} \le Ch^{q+1}\|v\|_{H^{q+1}(\Omega)}, \quad \|v - P_h v\|_{H^1(\Omega)} \le Ch^q\|v\|_{H^{q+1}(\Omega)} \qquad \forall v \in H^{q+1}(\Omega) \cap H_0^1(\Omega).$$

hat, dann gilt folgende Erweiterung der Sätze 2.39, 2.40:

$$\|u(t) - u_h(t)\|_{L^2} \le Ch^{q+1}\|u_0\|_{H^{q+1}(\Omega)} \qquad \forall u_0 \in \widehat{H}^{q+1}(\Omega), \tag{2.49}$$

$$\|u(t) - u_h(t)\|_{L^2} \le Ch^{q+1}t^{-(q+1)/2}\|u_0\|_{L^2(\Omega)}. \tag{2.50}$$

Hinweis: Für (2.49) gehen Sie vor wie im Beweis von Satz 2.39 und verwenden Sie die Normäquivalenz aus Bemerkung 2.33. Für (2.50) iterieren Sie (2.48) geeignet oft. ∎

Bis jetzt war immer $f \equiv 0$ angenommen. Wir zitieren (vgl. [1, Thm. 3.6]) ein Resultat, welches mit ähnlichen Techniken gezeigt werden kann und welches zeigt, dass auch für $f \neq 0$ weg von $t = 0$ mit optimalen Raten gerechnet werden kann:

Satz 2.42 *Gelte Voraussetzung 2.38. Sei $f \in L^2((0,T); L^2(\Omega))$ und $u_{h,0} = \Pi^{L^2} u_0$. Dann ist für jedes $\ell \ge 0$, $t \ge \delta > 0$ (so dass die rechte Seite endlich ist)*

$$\|\frac{d^\ell}{dt^\ell}(u(t) - u_h(t))\|_{L^2} \le Ch^2 \left(\|u_0\|_{L^2} + \int_0^t \|f(s)\|_{L^2}\, ds + \sum_{j \le \ell+1} \int_{t-\delta}^t \|\frac{d^j}{dt^j} u(s)\|_{H^2(\Omega)}\, ds \right).$$

2.5.5 Zeitdiskretisierung

Der vorangehende Abschnitt hat den Fehler $u(t) - u_h(t)$ untersucht. Wir betrachten nun den Zeitdiskretisierungsfehler $u_h(t_n) - u_h^n$. Wir untersuchen wieder (modellhaft) den Fall $f \equiv 0$. Unser Ziel wird eine Abschätzung der Form

$$\|u_h(t_n) - u_h^n\|_{L^2} \le Ck^p t_n^{-p}\|u_{h,0}\|_{L^2}, \qquad n = 1, 2, \ldots,$$

sein, wobei p die Ordnung des Zeitschrittverfahrens ist. (Spielarten diese Abschätzungen für hinreichend glatte $f \neq 0$ und/oder kompatible Startwerte u_0 gibt es natürlich auch ...)

Vorbetrachtung

Wir betrachten Einschrittverfahren mit folgenden Eigenschaft:

1. ("Stabilitätsfunktion" R) Angewandt auf das skalare Problem $y' = \lambda y$ ist ein Schritt des Verfahrens $y_1 = R(k\lambda)y_0$ für eine Funktion R.
2. (Koordinateninvarianz des Verfahrens) Sei $\mathbf{B} \in \mathbb{R}^{N\times N} = \mathbf{V}^{-1}\mathbf{D}\mathbf{V}$ diagonalisierbar mit Diagonalmatrix $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_N)$. Dann kommutiert der Basiswechsel $\mathbf{x} \mapsto \mathbf{V}\mathbf{x}$ mit der Anwendung des Verfahrens auf $\mathbf{y}' = \mathbf{B}\mathbf{y}$, d.h. für $\widetilde{\mathbf{y}}_1 = \mathbf{V}\mathbf{y}_1$ und $\widetilde{\mathbf{y}}_0 = \mathbf{V}\mathbf{y}_0$ gilt

$$\widetilde{\mathbf{y}}_1 = \mathrm{diag}(R(kd_1), \ldots, R(kd_N))\widetilde{\mathbf{y}}_0$$

Übung 2.43 (a) Überlegen Sie sich, dass RK-Verfahren die Eigenschaft der Koordinateninvarianz haben.

(b) Die Stabilitätsfunktion R kann für RK-Verfahren explizit angegeben werden:

$$R(z) = 1 + z\mathbf{b}^\top(\mathrm{Id} - z\mathbf{A})^{-1}\mathbf{e}, \qquad \mathbf{e} = (1, 1, \ldots, 1)^\top$$

(c) Teilaufg. b) zeigt, dass für RK-Verfahren die Stabilitätsfunktion R eine rationale Funktion ist. Es ist:

- $R(z) = 1 + z$ für das explizite Eulerverfahren
- $R(z) = \frac{1}{1-z}$ für das implizite Eulerverfahren
- $R(z) = \frac{1+z/2}{1-z/2}$ für das Crank-Nicolson-Verfahren

■

Wendet man das ODE-Verfahren auf $y' = \lambda y$ an, so ist die Lösung $y(t) = e^{\lambda t} y_0$. Ist das Verfahren der Ordnung p, so gilt nach Definition der Ordnung $|y(k) - y_1| \leq Ck^{p+1}$. Damit sieht man, dass die Funktion R folgende Asymptotik bei $z = 0$ haben sollte:

$$R(z) = e^z + O(|z|^{p+1}) \tag{2.51}$$

Hier $p = 1$ für die Eulerverfahren und $p = 2$ für das Crank-Nicolson-Verfahren.

Fehleranalyse

Im Fall $f \equiv 0$ ist das ODE-System

$$\mathbf{M}\mathbf{u}' + \mathbf{A}\mathbf{u} = 0.$$

Der Startvektor $\mathbf{u}^0$ korrespondiert mit der Funktion $u_{h,0} =: u_h^0$.

Wir fassen die Eigenwerte $\lambda_{h,n}$, $n = 1, \ldots, N$, des verallgemeinerten EWP $\lambda \mathbf{M}\mathbf{x} = \mathbf{A}\mathbf{x}$ im Spektrum

$$\sigma_h := \{\lambda_{h,n} \,|, n = 1, \ldots, N\}$$

zusammen. Der exakte Lösungsoperator ist durch $E_h(t)$ beschrieben, d.h.

$$E_h(t)u_h^0 = \sum_{m=1}^{N} e^{-\lambda_{h,m} t} (u_h^0, \varphi_{h,m})_{L^2} \varphi_{h,m}$$

Ein Schritt des numerischen Verfahrens ist wegen unserer Annahmen gegeben durch

$$u_h^1 = \sum_{m=1}^{N} R(-\lambda_{h,m} k)(u_h^0, \varphi_{h,m})_{L^2} \varphi_{h,m}.$$

Entsprechend ergibt sich dann die Approximation u_h^n als

$$u_h^n = \sum_{m=1}^{N} (R(-\lambda_{h,m} k))^m (u_h^0, \varphi_{h,m})_{L^2} \varphi_{h,m}.$$

Es ist geschickt, die Funktion

$$F_n(z) := e^{-zn} - (R(-z))^n$$

einzuführen. Damit lässt sich der Fehler bei t_n schreiben als

$$\|u_h(t_n) - u_h^n\|_{L^2}^2 = \sum_{m=1}^{N} |F_n(k\lambda_{h,m})|^2 |(u_h^0, \varphi_{h,m})_{L^2}|^2 \leq \sup_{\lambda \in \sigma_h} |F_n(k\lambda)|^2 \|u_h^0\|_{L^2}^2$$

Wir müssen also die Funktion $F_n(k\lambda)$ kontrollieren. Weil $\lambda = \lambda_{h,m}$ gross sein kann, ist das Ziel, $F_n(k\lambda_{h,m})$ *uniform in* $\lambda_{h,m}$ und explizit in n zu kontrollieren.

Die Konsistenzforderung (2.51) liefert gute Kontrolle von $F_n(k\lambda_{h,m})$ (und festes n) für die $\lambda_{h,m}$ mit $k\lambda_{h,m}$ klein. Für die restlichen $\lambda_{h,m}$ müssen wir weitere Eigenschaften von R fordern. Eine Minimalforderung ist *Stabilität*

$$|R(-k\lambda)| \leq 1 \qquad \forall \lambda \in \sigma_h \tag{2.52}$$

Diese Bedingung ist für A-stabile Verfahren erfüllt. Es stellt sich heraus, dass es geschickt ist, $|R(\infty)| < 1$ zu fordern, genauer:

$$\sup_{\lambda \in \sigma_h} |R(-k\lambda)| = q < 1. \tag{2.53}$$

Für typische Verfahren wie den A-stabilen RK-Verfahren ist R auf $(-\infty, 0]$ definiert und dort sogar eine glatte Funktion. Wir betrachten zwei Fälle:

(I) R ist auf $(-\infty, 0]$ definiert und erfüllt $|R(z)| \leq 1$ für alle $z \in (-\infty, 0]$

(II) R ist auf $(-\infty, 0]$ definiert und $\forall z_0 > 0 \quad \exists q(z_0) < 1$ mit $|R(z)| \leq q(z_0) < 1$ für alle $z < -z_0$.

Die Bedingung (I) wird von A-stabilen RK-Verfahren wie dem Crank-Nicolson-Verfahren erfüllt (allgemeiner, z.B. den Gauss-Verfahren) Die stärke Bedingung (II) wird z.B. von L-stabilen RK-Verfahren wie dem impliziten Eulerverfahren erfüllt (allgemeiner, z.B. den Radau IIA-Verfahren).

Lemma 2.44 *Es gelte die Konsistenzbedingung (2.51).*

(i) Sei $c \in (0,1)$. Es gibt $z_0 > 0$ und $C > 0$, so dass

$$|F_n(z)| \leq Cnz^{p+1}e^{-cnz} \qquad 0 < z < z_0.$$

(ii) Gilt zusätzlich die Stabilitätsbedingung (I), so gilt

$$|F_n(z)| \leq Cz^p \qquad \forall z \in (0, \infty).$$

(iii) Gilt zusätzlich die Stabilitätsbedingung (II), so gibt es für jedes $z_0 > 0$ ein $c > 0$ und ein $C > 0$

$$|F_n(z)| \leq Ce^{-cn} \qquad \forall z \geq z_0.$$

Beweis: *ad (i):* Die Konsistenzbedingung (2.51) liefert für hinreichend kleines $\lambda_0 > 0$

$$|e^{-\lambda} - R(-\lambda)| \leq C\lambda^{p+1} \qquad 0 \leq \lambda \leq \lambda_0. \tag{2.54}$$

Elementare Betrachtungen[11] zeigen dann, dass für beliebiges $c \in (0,1)$ und geeignet verkleinertes z_0 gilt:

$$|R(-z)| \leq e^{-cz} \qquad 0 \leq z \leq z_0. \tag{2.55}$$

$$\begin{aligned} |F_n(z)| \qquad &= \qquad |e^{-zn} - (R(-z))^n| = \left|e^{-z} - R(-z)\right| \left|\sum_{j=0}^{n-1} (R(-z))^{n-1-j} e^{-jz}\right| \\ &\overset{(2.54),(2.55)}{\leq} \qquad Cz^{p+1} n e^{-c(n-1)z} \leq Cz^p, \end{aligned} \tag{2.56}$$

wobei wir im vorletzten Schritt $\max_{j \in \{0,\ldots,n-1\}} -jz - cz(n-1-j) = -cz(n-1)$ ausgenutzt haben und im letzten Schritt wieder $\sup_{x>0} xe^{-x} < \infty$ verwendet haben.

ad (ii): Sei z_0 wie in (i). Weil (2.56) die gewünschte Abschätzung für $z \in (0, z_0)$ liefert, reicht es, $z \geq z_0$ zu betrachten. Weil wegen der geforderten Stabilität $|R(\zeta)| \leq 1$ für $\zeta \in (-\infty, 0]$ gilt, erhalten wir für $z \geq z_0$

$$|F_n(z)| \leq |e^{-zn}| + |(R(-z))^n| \leq 2 \leq Cz_0^p \leq Cz^p$$

für geeignetes $C > 0$.

ad (iii): Es ist

$$|F_n(z)| = |e^{-nz} - |R(-z))^n| \leq e^{-nz} + |R(-z)|^n.$$

Für $z \geq z_0$ ist nach Voraussetzung $|R(-z)| \leq q < 1$. Damit folgt

$$|F_n(z)| \leq e^{-nz} + |R(-z)|^n \leq e^{-nz_0} + q^n,$$

was den Beweis abschliesst. □

Satz 2.45 *Gelte die Konsistenzbedingung (2.51) sowie die Stabilitätsbedingung (II). Dann gilt:*

$$\|u_h(t_n) - u_h^n\|_{L^2} \leq Ck^p t_n^{-p} \|u_h^0\|_{L^2}$$

[11] $R(-z) = e^{-z} + O(p^{p+1}) = 1 - z + O(z^2)$ und $e^{-cz} = 1 - cz + O(z^2)$, so dass $R(-z) \leq e^{-cz}$ für hinreichen kleine $z > 0$

Beweis: Wir müssen zeigen:

$$\sup_{\lambda\in\sigma_h} |F_n(k\lambda)| \le Ck^p t_n^{-p} = Ck^p(kn)^{-p} = Cn^{-p}. \tag{2.57}$$

Sei z_0 wie in Lemma 2.44, (i). Für $\lambda \in \sigma_h$ mit $k\lambda \le z_0$ gilt nach Lemma 2.44, (i)

$$|F_n(k\lambda)| \le Cn(k\lambda)^{p+1}e^{-cnk\lambda} = Cn^{-p}(t_n\lambda)^{p+1}e^{-c\lambda t_n} \le Cn^{-p},$$

denn $\sup_{x>0} x^{p+1}e^{-x} < \infty$. Für $\lambda \in \sigma_h$ mit $k\lambda \ge z_0$ verwenden wir Lemma 2.44, (iii):

$$|F_n(k\lambda)| \le Ce^{-cn} \le Cn^{-p}.$$

Damit ist (2.57) gezeigt. □

Satz 2.45 verlangt $|R(\infty)| < 1$. Damit ist das Crank-Nicolson-Verfahren nicht erfasst. Tatsächlich kann man es "retten", wenn man einige (implizite) Eulerschritte am Anfang macht.

Übung 2.46 Betrachten Sie den Fall $f \equiv 0$. Betrachten Sie folgendes Verfahren: Sie machen *zwei* Schritte des impliziten Eulerverfahrens und dann Crank-Nicolson-Schritte. Zeigen Sie:

$$\|u_h(t_n) - u_h^n\|_{L^2} \le Ck^2 t_n^{-2}\|u_h^0\|_{L^2}$$

Hinweis: Schreibt man R_0 und R_1 für die Stabilitätsfunktionen des Eulerverfahrens und des Crank-Nicolson-Verfahrens, so muss analog zu Lemma 2.44 die Abschätzung $|F_n(z)| \le Cn^{-2}$ gezeigt werden kann, wenn

$$F_n(z) := (R_0(-z))^2(R_1(-z))^{n-2} - e^{-nz}.$$

Gehen Sie wie folgt vor:

- $|R_0(-z))^2(R_1(-z))^{n-2}| \le Cn^{-2}$ für $z \ge z_0$ für geeignetes z_0 (überlegen Sie sich, wo das Maximum dieser Funktion für geg. $n \ge 2$ (gross) ist.
- Setzen Sie $F_n(z) = R_0(-z)^2((R_1(-z))^{n-2} - e^{-(n-2)z}) + ((R_0(-z))^2 - e^{-2z})e^{-(n-2)z}$
- Schätzen Sie $((R_1(-z))^{n-2} - e^{-(n-2)z}) \le Cn^{-2}$ für $z \le z_0$ ab.
- Schätzen Sie $(R_0(-z))^2 - e^{-2z} \le Cz^2$ für $z \le z_0$ ab.

Bemerkung 2.47 Die Abschätzungen in Satz 2.45 und Übung 2.46 sind nicht uniform in t_n: die Konstante wird schlecht für $t_n \to 0$. Will man also gute Abschätzungen "bis an $t = 0$" haben, so muss man mit Gitterverfeinerung in der Zeit (und ggf. im Ort) arbeiten. ∎

Kapitel 3

Wahrscheinlichkeitstheoretische Methoden für parabolische PDEs

3.1 Die Brownsche Bewegung

Wir betrachten einen stochastischen Prozess $X(t)$ auf dem Intervall $[0,T]$ (oder $[0,\infty)$). Grob gesagt ist das einfach eine Funktion $t \mapsto X(t)$, wobei $X(t)\colon \Omega_{\text{prob}} \to \mathbb{R}$ eine Zufallsvariable ist (also eine Funktion die einen Wahrscheinlichkeitsraum Ω_{prob} nach $\mathbb{R}$ abbildet). Man kann also X also Funktion von $[0,T] \times \Omega_{\text{prob}}$ sehen und $X(t,\omega)$ schreiben. Der Einfachheit halber, lassen wir das Argument ω meistens weg. Ein Wahrscheinlichkeitsraum braucht ein Maß welches wir mit $\mathbb{P}$ bezeichnen, und eine σ-Algebra $\mathcal{F}$. Ein stochastischer Prozess hängt von t ab, und so darf auch $\mathcal{F}$ von t abhängen. Wir definieren eine Familie von σ-Algebren $\mathcal{F}_t \subseteq \mathcal{F}$ mit $\mathcal{F}_s \subseteq \mathcal{F}_t$ für $s \leq t$ (so einen aufsteigende Folge nennt man "Filtration"). Wir nennen eine Prozess "$\mathcal{F}_t$-adaptiert", falls $X(t)$ bezüglich $\mathcal{F}_t$ messbar ist. Wir betrachten hier nur die standard Filtration $\mathcal{F}_t$ definiert als die kleinste σ-Algebra in der alle $X(s)$, $s \leq t$ messbar sind.

3.1.1 Ein-dimensionale Brownsche Bewegung

Die Brownsche Bewegung bezeichnet die vom schottischen Botaniker Robert Brown im Jahr 1827 unter dem Mikroskop entdeckte Bewegung kleiner Teilchen in Flüssigkeiten. Die physikalische Erklärung dieser Bewegung durch Wärmebewegung gelangt 1905 Albert Einstein durch die statistische Betrachtung von Kollisionen der Moleküle der Flüssigkeit mit den Teilchen.

Mathematisch ist eine Brownsche Bewegung wie folgt definiert:

Definition 3.1 *Ein stochastischer Prozess $B(t)\colon [0,T] \times \Omega_{\text{prob}} \to \mathbb{R}$ heißt Brownsche Bewegung falls*

1. *$B(0) = 0$ fast sicher.*
2. *$B(t) - B(s)$ ist unabhängig von $B(r)$ für alle $r \leq s < t$.*
3. *$B(t) - B(s) \sim N(0, t-s)$ für alle $s < t$ (das Inkrement is Normalverteilt mit Varianz $t-s$ und Mittelwert 0).*
4. *$t \mapsto B(t,\omega)$ ist stetig für fast alle $\omega \in \Omega_{\text{prob}}$.*

Der folgende Satz zeigt, dass Brownsche Bewegungen tatsächlich existieren. Es gibt dafür zahlreiche Konstruktionen, die erste stammt von Norbert Wiener (J. Math. Phys., 1923). Daher werden Brownsche Bewegungen oft auch Wiener Prozesse genannt.

Satz 3.2 *Es existiert ein Wahrscheinlichkeitsraum Ω_{prob} und ein stochastischer Prozess $B(t)$, so dass $B(t)$ eine Brownsche Bewegung ist.*

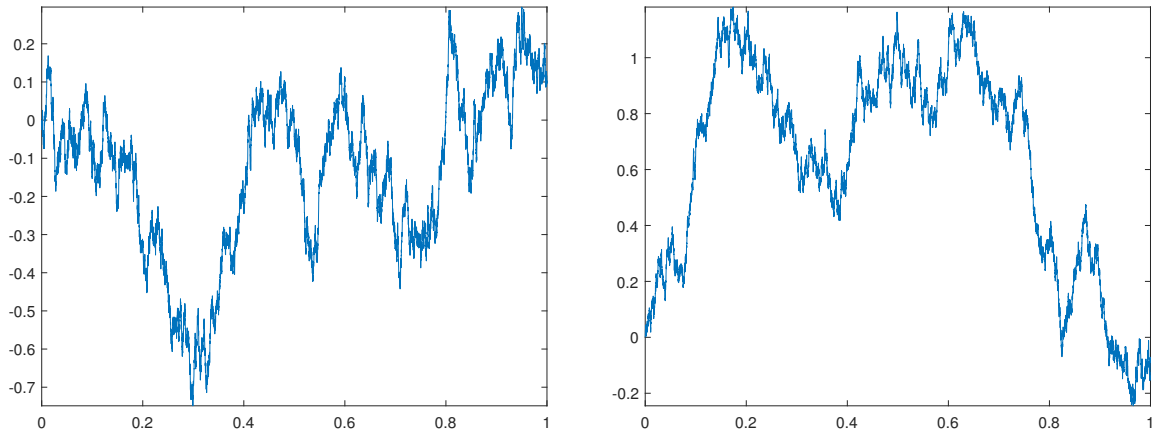


Abbildung 3.1: Zwei zufällige Stichproben einer eindimensionalen Brownsche Bewegung auf $[0,1]$ (d.h. wir plotten $t \mapsto B(t,\omega)$ für zufällig gewählte $\omega \in \Omega_{\text{prob}}$.

Bemerkung 3.3 Zur Erinnerung: Zwei Ereignisse $E_1, E_2 \subseteq \Omega_{\text{prob}}$ sind unabhängig, wenn $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$. Zwei σ-Algebren $\mathcal{F}_1$ und $\mathcal{F}_2$ sind unabhängig, wenn alle Paare $(E_1, E_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ unabhängig sind. Eine Zufallsvariable X ist unabhängig von $\mathcal{F}$, falls die kleinste σ-Algebra $\mathcal{F}_X$ in der X messbar ist, unabhängig ist von $\mathcal{F}$. Zwei Zufallsvariablen X, Y sind unabhängig, falls $\mathcal{F}_X$ und $\mathcal{F}_Y$ unabhängig sind.

Wir betrachten die weitere wichtige Eigenschaften einer Brownschen Bewegung:

- Es gilt $\text{Cov}(B(t), B(s)) = \mathbb{E}(B(t)B(s)) = \min\{s,t\}$ für alle $s,t \in [0,T]$. *Beweis:* Da $\mathbb{E}(B(t)) = \mathbb{E}(B(t) - B(0)) = 0$, haven wir $\text{Cov}(B(t), B(s)) = \mathbb{E}(B(t)B(s))$. Außerdem gilt für $s < t$

$$\begin{aligned}\mathbb{E}(B(t)B(s)) &= \mathbb{E}((B(t) - B(s) + B(s))B(s)) = \mathbb{E}(B(t) - B(s))\mathbb{E}(B(s)) + \mathbb{E}(B(s)^2) \\ &= \mathbb{E}((B(s) - B(0))^2) = s.\end{aligned}$$

- Das Inkrement $B(t) - B(s)$ ist unabhängig von allen $\mathcal{F}_s$-messbaren Zufallsvariablen ($t > s$). *Beweis:* $B(t) - B(s)$ ist unabhängig von $B(s)$ per definition. Das heißt, $B(t) - B(s)$ ist unabhängig von $\mathcal{F}_s$ und daher auch von allen Zufallsvariablen die bezüglich $\mathcal{F}_s$ messbar sind.

Zusätzlich gilt, dass Brownsche Bewegungen Hölder stetig mit Exponent $\alpha \in [0, 1/2)$ sind, d.h., für fast alle $\omega \in \Omega_{\text{prob}}$ existiert $C(\omega) < \infty$ mit

$$|B(t,\omega) - B(s,\omega)| \leq C(\omega)|t - s|^\alpha$$

für alle $s,t \in [0,T]$.

Die oben genannten Eigenschaften zeigen einen sehr einfachen Weg um Brownsche Bewegungen zu simulieren:

`Algorithmus (Brownsche Bewegung)` Input: $t_0 < t_2 < \ldots < t_m$, $B_0 \in \mathbb{R}$.
Set $B(t_0) = B_0$. For $j = 2, \ldots, m$ do:

1. Ziehe standardnormalverteilte Stichprobe z_j (zum Beispiel mit `randn()` in Matlab)
2. Berechne $B(t_j) := B(t_{j-1}) + \sqrt{t_j - t_{j-1}}z_j$.

Übung 3.4 Zeigen Sie, dass die im obige Algorithmus erzeugte Zahlenfolge alle Eigenschaften einer (diskreten) Brownschen Bewegung hat (Definition 3.1 mit $s,t \in \{t_0, \ldots, t_m\}$ natürlich ohne Stetigkeit).

Die Übung zeigt, dass der Algorithmus nicht nur eine Approximation an eine Brownsche Bewegung liefert, sondern sogar exakte Stichproben mit endlich vielen Auswertungspunkten.

3.1.2 Mehrdimensionale Brownsche Bewegung

Wir können auch Vektoren von Brownschen Bewegungen betrachten, d.h., $B(t) = (B_1(t), \dots, B_d(t))$, wobei die Komponenten $B_i(t)$ unabhängige Brownsche Bewegungen sind. Daher gilt

1. $B_i(t)$ ist normalverteilt für alle $t \in [0,T]$ und $i = 1, \dots, d$.
2. $\mathbb{E}(B_i(t)) = 0$ für alle $t \in [0,T]$ und $i = 1, \dots, d$.
3. $\mathbb{E}(B_i(t)B_j(s)) = \delta_{ij} \min\{t,s\}$ für alle $s,t \in [0,T]$ und $i,j = 1, \dots, d$.
4. $t \mapsto B(t,\omega)$ ist stetig in $\mathbb{R}^d$ für fast alle $\omega \in \Omega_{\text{prob}}$.

3.2 Das Ito Integral

Wir wollen der Differentialgleichung

$$\partial_t y(t) = g(y(t),t) + f(y(t),t)\partial_t B(t) \tag{3.1}$$

eine rigorose mathematische Definition geben. Solche Gleichungen kommen in der Finanzmathematik sehr häufig vor, wobei die Brownsche Bewegung die unvorhersehbare (zufällige) Natur der Aktienmärkte modelliert. Wir schreiben die DGL in Integralform und erhalten

$$y(t) - y(0) = \int_0^t g(y(s),s)\,ds + \int_0^t f(y(s),s)\partial_t B(s)\,ds.$$

Natürlich ist $\partial_t B(s)$ nicht definiert und wir brauchen einen neuen Integralbegriff, das Ito-Integral.

Angenommen, $f(t)$ wäre ein sogenannter Stufenprozess, also ein Prozess zu dem es eine Partition $0 = s_0 < s_1 < \ldots < s_m = t$ gibt mit $f|_{[s_i,s_{i+1}]} = \xi_i$ für eine $\mathcal{F}_{s_i}$-messbare Zufallsvariable ξ fast überall und für $i = 0, \dots, m-1$. Dann können wir das Ito-Integral definieren als

$$\int_0^t f(y(s),s)\partial_t B(s)\,ds := \sum_{i=0}^{m-1} f(s_i)(B(s_{i+1}) - B(s_i)) = \sum_{i=0}^{m-1} \xi_i (B(s_{i+1}) - B(s_i)).$$

Die (endliche) Summe ist messbar, da $B(s_{i+1})$, $B(s_i)$ auch $\mathcal{F}_{s_i}$-messbar sind.

Sei nun $\mathcal{L}^2(0,T)$ die Menge aller $\mathcal{F}$-adaptierten Prozesse die

$$\|g\|_{\mathcal{L}^2(0,T)} := \Big(\mathbb{E}\Big(\int_0^T g(t)^2\,dt\Big)\Big)^{1/2} < \infty$$

erfüllen. Jeder Stufenprozess ist natürlich ein $\mathcal{L}^2$-Prozess und (wie im gewöhnlichen L^2) gilt die Dichtheit der Stufenprozesse in $\mathcal{L}^2$.

Satz 3.5 *Die Menge aller adaptierten Stufenprozesse ist dicht in $\mathcal{L}^2(0,T)$.*

Beweis: Beweis erfolgt wie in L^2. □

Wir können nun das Ito-Integral definieren.

Definition 3.6 *Sei $g \in \mathcal{L}^2(0,T)$ und sei B eine Brownsche Bewegung. Dann definieren wir das Ito-Integral als*

$$\int_0^T g(s)\,dB(s) := \lim_{k\to\infty} \int_0^T g_k(s)\,dB(s)$$

wobei $g_k \to g$ in $\mathcal{L}^2(0,T)$ eine approximierende Folge von Stufenprozessen ist und die Konvergenz in $\mathcal{L}^2$ zu verstehen ist, d.h.,

$$\lim_{k\to\infty} \mathbb{E}\Big(\Big|\int_0^T g(s)\,dB(s) - \int_0^T g_k(s)\,dB(s)\Big|^2\Big).$$

Wir geben einige Eigenschaften das Ito-Integrals an:

1. $\int_0^T dB(s) = B(T)$
2. Das Ito-Integral ist eine lineare Funktion des Integranden, d.h.

$$\int_0^T ag(s) + bf(s)\, dB(s) = a\int_0^T g(s)\, dB(s) + b\int_0^T f(s)\, dB(s)$$

   für $a, b \in \mathbb{R}$ und $g, f \in \mathcal{L}^2(0,T)$.

3. Ito-Isometrie: Für $f, g \in \mathcal{L}^2(0,T)$ gilt

$$\mathbb{E}\Big(\int_0^T f\, dB \int_0^T g\, dB\Big) = \mathbb{E}\Big(\int_0^T fg\, ds\Big),$$

   wobei das rechte Integral ein gewöhnliches Lebesgue Integral ist.

4. $\mathbb{E}(\int_0^T g(s)\, dB(s)) = 0$.
5. Für $g \in \mathcal{L}^2(0,T)$ existiert ein stetiger Prozess G, sodass

$$G(t) = \int_0^t g(s)\, dB(s)$$

   fast überall gilt.

Beweis:[Beweisskizzen] Die Funktion $g(t) = 1$ ist ein Stufenprozess und daher folgt sofort (1). Die Linearität (2) folgt sofort aus der Definition und der Linearität des Limes. Für Stufenprozesse f, g können wir immer eine gemeinsame Partition von $[0,T]$ finden, d.h. $0 = s_0 < s_1 < \ldots < s_m = T$ mit $f|_{[s_i,s_{i+1}]} =$ konstant und $g|_{[s_i,s_{i+1}]} =$ konstant für all $i = 0, \ldots, m$. Damit erhalten wir

$$\mathbb{E}\Big(\int_0^T f\, dB \int_0^T g\, dB\Big) = \sum_{i,j=0}^{m-1} \mathbb{E}\Big(f(s_i)g(s_j)(B(s_{i+1}) - B(s_i))(B(s_{j+1}) - B(s_j))\Big)$$

Sei zunächst $i > j$: Laut Definition von Brownschen Bewegungen ist $B(s_{i+1}) - B(s_i)$ unabhängig von $\mathcal{F}_{s_i}$ und daher auch unabhängig von $B(s_{j+1}) - B(s_j)$, $f(s_i)$, und $g(s_j)$ (die ja alle bezüglich $\mathcal{F}_{s_j}$ messbar sind). Also gilt

$$\begin{aligned}\mathbb{E}\Big(f(s_i)g(s_j)(B(s_{i+1}) - B(s_i))(B(s_{j+1}) - B(s_j))\Big)\\ = \mathbb{E}\Big(f(s_i)g(s_j)(B(s_{j+1}) - B(s_j))\mathbb{E}(B(s_{i+1}) - B(s_i))\Big) = 0.\end{aligned}$$

Das gleiche gilt für $i < j$. Für $i = j$ erhalten wir

$$\mathbb{E}\Big(f(s_i)g(s_i)(B(s_{i+1}) - B(s_i))^2\Big) = \mathbb{E}(f(s_i)g(s_i))\mathbb{E}(B(s_{i+1}) - B(s_i))^2) = \mathbb{E}(f(s_i)g(s_i))(s_{i+1} - s_i).$$

Daher fallen alle Terme mit $i \neq j$ weg und es bleibt

$$\mathbb{E}\Big(\int_0^T f\, dB \int_0^T g\, dB\Big) = \sum_{i=0}^{m-1} \mathbb{E}(f(s_i)g(s_i))(s_{i+1} - s_i) = \int_0^T \mathbb{E}(f(s)g(s))\, ds.$$

Übergang zum Limes zeigt die Aussage für allgemeine Prozesse in $\mathcal{L}^2(0,T)$ und damit (3). Wir zeigen (4) wieder zuerst für Stufenprozesse. Es gilt

$$\mathbb{E}(\int_0^T g(s)\, dB(s)) = \sum_{i=0}^{m-1} \mathbb{E}(g(s_i))\mathbb{E}(B(s_{i+1}) - B(s_i)) = 0$$

da $g(s_i)$ unabhängig von $B(s_{i+1}) - B(s_i)$ ist und $B(s_{i+1}) - B(s_i) \sim N(0, s_{i+1} - s_i)$. Übergang zum Limes zeigt dann (4). □

Wir haben nun die Mittel, der stochastischen Differentialgleichung (3.1) (SDE) eine mathematische Bedeutung zu geben. Wir sagen: Für $H, K \colon \mathbb{R}^d \times [0,T] \to \mathbb{R}$ erfüllt ein Prozess $X(t)$ die SDE

$$dX(t) = H(X(t), t)\, dt + K(X(t), t)\, dB(t) \tag{3.2}$$

wenn $X(t)$ die Integralgleichung

$$X(t) - X(0) = \int_0^t H(X(s), s)\, ds + \int_0^t K(X(s), s)\, dB(s)$$

fast überall in $[0,T]$ und fast überall in Ω_{prob} erfüllt.

Satz 3.7 *Seien H, K stetig in beiden Variablen und global Lipschitzstetig in x, d.h.,*

$$|H(x,t) - H(y,t)| + |K(x,t) - K(y,t)| \le C|x - y| \quad \text{für alle } x, y \in \mathbb{R}^d,\ t \in [0,T]$$

für eine Konstante $C > 0$. Sei $x_0 \in \mathbb{R}^d$ gegeben, dann existiert eine eindeutige Lösung $X(t)$ der SDE (3.2) mit $X(0) = x_0$. Die Lösung ist stetig in t und $\mathcal{F}_t$ adaptiert.

3.2.1 Das Lemma von Ito

Das Lemma von Ito ist die Kettenregel der stochastischen Analysis. Sei $x(t)$ eine differenzierbare Funktion so dass

$$x'(t) = g(x(t), t).$$

Die Integralform dieser ODE ist

$$x(t) - x(0) = \int_0^t g(x(s), s)\, ds.$$

Die gewöhnliche Kettenregel zeigt

$$\partial_t \phi(x(t)) = \phi'(x(t)) x'(t) = \phi'(x(t)) g(x(t), t)$$

oder in Integralform

$$\phi(x(t)) - \phi(x(0)) = \int_0^t \phi'(x(s)) g(x(s), s)\, ds.$$

Falls ϕ glatt ist, so erhalten wir mit Taylor Entwicklung

$$\phi(x(t)) = \phi(x(0)) + \phi'(x(0))(x(t) - x(0)) + \frac{1}{2}\phi''(x(0))(x(t) - x(0))^2 + \mathcal{O}(|x(t) - x(0)|^3)$$

sogar wenn $x(t)$ nicht differenzierbar ist. Setzen wir nun $x(t) = B(t)$ (Brownsche Bewegung), so sehen wir, dass $\phi'(B(0))(B(t) - B(0))$ im Durchschnitt gleich Null ist (Die Inkremente von B sind $N(0,t)$-verteilt). Weiters ist $\phi''(B(0))(B(t) - B(0))^2$ im Durchschnitt gleich t. Das lässt vermuten, dass die Kettenregel für $\phi(B(t))$ die folgende Form haben sollte

$$\phi(B(t)) - \phi(B(0)) = \int_0^t \phi'(s)\, dB(s) + \frac{1}{2}\int_0^t \phi''(s)\, ds.$$

Satz 3.8 (Ito) *Sei $\phi(x,t)$ zweimal differenzierbar in x und einmal differenzierbar in t. Sei $X(t)$ ein Prozess mit*

$$dX(t) = F(t)\, dt + G(t) dB(t),$$

dann gilt für $Y(t) := \phi(X(t), t)$

$$dY(t) = \Big(\partial_x\phi(X(t),t)F(t) + \partial_t\phi(X(t),t) + \frac{1}{2}\partial_x^2\phi(X(t),t)G^2(t)\Big)\,dt + \partial_x\phi(X(t),t)G(t)\,dB(t)$$

oder in Integralform

$$Y(t) - Y(0) = \int_0^t \Big(\partial_x\phi(X(s),s)F(s) + \partial_t\phi(X(s),s) + \frac{1}{2}\partial_x^2\phi(X(s),s)G^2(s)\Big)\,ds + \int_0^t \partial_x\phi(X(s),s)G(s)\,dB(s).$$

Beweis:[Skizze] Ein exakter Beweis beschäftigt sich mit der Konvergenz von Zufallsvariablen und ist etwas aufwendiger. Wir geben nur die Idee des Beweises wieder: Sei $0 = s_0 < s_1 < \ldots < s_m = t$ eine Partition von $[0,t]$. Dann gilt mit der Taylorentwicklung von ϕ

$$\begin{aligned}
Y(t) - Y(0) &= \sum_{i=0}^{m-1} Y(s_{i+1}) - Y(s_i) \\
&= \sum_{i=0}^{m-1} \partial_t\phi(X(s_i),s_i)(s_{i+1}-s_i) + \sum_{i=0}^{m-1} \partial_x\phi(X(s_i),s_i)(X(s_{i+1})-X(s_i)) \\
&\quad + \sum_{i=0}^{m-1} \frac{1}{2}\partial_x^2\phi(X(s_i),s_i)(X(s_{i+1})-X(s_i))^2 \\
&\quad + \sum_{i=0}^{m-1} \mathcal{O}\big(s_{i+1}-s_i)(X(s_{i+1})-X(s_i))\big) + \mathcal{O}(s_{i+1}-s_i)^2) + \mathcal{O}(X(s_{i+1})-X(s_i))^3.
\end{aligned} \tag{3.3}$$

Wenn wir den Limes über die Partitionen bilden (d.h. wir finden eine Folge von Partitionen so dass $\max_{i=0,\ldots,m-1} s_{i+1} - s_i \to 0$), dann bleibt die linke Seite der Gleichung unverändert. Die erste Summe auf der rechten Seite erfüllt

$$\sum_{i=0}^{m-1} \partial_t\phi(X(s_i),s_i)(s_{i+1}-s_i) \to \int_0^t \partial_t\phi(X(s),s)\,dt$$

da ϕ regulär und X stetig ist und damit die Riemann Summe konvergiert (klassische Analysis). Für die zweite Summe erhalten wir nach Definition von X

$$\begin{aligned}
&\sum_{i=0}^{m-1} \partial_x\phi(X(s_i),s_i)(X(s_{i+1})-X(s_i)) \\
&\quad = \sum_{i=0}^{m-1} \partial_x\phi(X(s_i),s_i)\Big(\int_{s_i}^{s_{i+1}} F(s)\,ds + \int_{s_i}^{s_{i+1}} G(s)\,dB(s)\Big) \\
&\quad \to \int_0^t \partial_x\phi(X(s,s))F(s)\,ds + \int_0^t \partial_x\phi(X(s,s))G(s)\,dB(s)
\end{aligned}$$

da ϕ regulär und X stetig ist (für das erste Integral reichen klassisch Riemann Argumente, für das zweite braucht man die analogen Argumente für das Ito Integral). Die dritte Summe auf der rechten Seite von (3.3) ist die komplizierteste und wir betrachten nur den Fall $F = 0$ und $G = 1$, also $X = B$. Wir haben bereits gezeigt, dass gilt

$$\sum_{i=0}^{m-1} \partial_x^2\phi(X(s_i),s_i)(s_{i+1}-s_i) \to \int_0^t \partial_x^2\phi(X(s),s)\,ds.$$

Es reicht also

$$I := \sum_{i=0}^{m-1} \frac{1}{2}\partial_x^2\phi(X(s_i),s_i)\big((X(s_{i+1})-X(s_i))^2 - (s_{i+1}-s_i)\big) \to 0$$

zu zeigen (im $L^2(\Omega_{\text{prob}})$ Sinn). Wir definieren $\delta X_i := X(s_{i+1}) - X(s_i)$ und $\delta s_i := s_{i+1} - s_i$. Es gilt

$$\mathbb{E}(I^2) = \frac{1}{4}\sum_{i,j=0}^{m-1} \mathbb{E}\Big(\partial_x^2\phi(X(s_i),s_i)\partial_x^2\phi(X(s_j),s_j)(\delta X_i^2 - \delta s_i)(\delta X_j^2 - \delta s_j)\Big).$$

Für jeden Term dieser Summe mit $i \neq j$ gilt (wir nehmen O.B.d.A $i < j$ an)

$$\begin{aligned}\mathbb{E}\Big(\partial_x^2\phi(X(s_i),s_i)\partial_x^2\phi(X(s_j),s_j)(\delta X_i^2 - \delta s_i)(\delta X_j^2 - \delta s_j)\Big)\\ = \mathbb{E}\Big(\partial_x^2\phi(X(s_i),s_i)\partial_x^2\phi(X(s_j),s_j)(\delta X_i^2 - \delta s_i)\Big)\mathbb{E}(\delta X_j^2 - \delta s_j) = 0\end{aligned}$$

wo wir verwenden, dass δX_j unabhängig zu $\mathcal{F}_j$ und $\mathcal{F}_i$ ist (X ist Brownsche Bewegung) und es gilt $\mathbb{E}\delta X_j^2 = \delta s_j$. Damit bleiben nur die diagonalen Terme übrig und wir erhalten

$$\mathbb{E}(I^2) = \frac{1}{4}\sum_{i=0}^{m-1} \mathbb{E}(\partial_x^2\phi(X(s_i),s_i)^2)\mathbb{E}(\delta X_i^2 - \delta s_i)^2 \leq C\sum_{i=0}^{m-1} \mathbb{E}(\partial_x^2\phi(X(s_i),s_i)^2)\delta s_i^2.$$

Hier haben wir

$$\mathbb{E}(\delta X_i^2 - \delta s_i)^2 = \mathbb{E}(\delta X_i^4) - 2\mathbb{E}(\delta X_i^2)\delta s_i + \delta s_i^2 = 3\delta s_i^2 - 2\delta s_i^2 + \delta s_i^2 = 2\delta s_i^2$$

verwendet ($\delta X_i \sim N(0,\delta s_i)$ und das vierte Moment der zentrierten Normalverteilung ist $3\sigma^4$). Da $\partial_x^2\phi$ gleichmäßig beschränkt ist (stetige Funktion auf $[0,T]$) erhalten wir $\mathbb{E}(I^2) \to 0$ wenn $\max \delta s_i \to 0$.

Zuletzt müssen wir noch argumentieren, dass die Fehlerterme gegen Null gehen: Der Term

$$\sum_{i=0}^{m-1} \mathcal{O}(s_{i+1} - s_i)^2 \leq \max_{i=0,\dots,m-1} \delta s_i \sum_{i=0}^{m-1} \delta s_i = T \max_{i=0,\dots,m-1} \delta s_i$$

geht natürlich gegen Null falls die Partition unendlich fein wird.

Wir nehmen noch immer $X = B$ an. Die Summe $\sum_{i=0}^{m-1} \mathcal{O}\big(s_{i+1} - s_i)(X(s_{i+1}) - X(s_i))\big)$ erfüllt

$$\begin{aligned}\mathbb{E}\Big|\sum_{i=0}^{m-1} \mathcal{O}\big(s_{i+1} - s_i)(X(s_{i+1}) - X(s_i))\big)\Big|^2 &\lesssim \max_{i=0,\dots,m-1} \delta s_i \sum_{i,j=0}^{m-1} \mathbb{E}(X(s_{i+1}) - X(s_i))(X(s_{j+1}) - X(s_j))\\ = \max_{i=0,\dots,m-1} \delta s_i \sum_{i=0}^{m-1} \mathbb{E}(X(s_{j+1}) - X(s_j))^2 &= \max_{i=0,\dots,m-1} \delta s_i \sum_{i=0}^{m-1} \delta s_i = T \max_{i=0,\dots,m-1} \delta s_i\end{aligned}$$

und geht daher auch gegen Null falls $\max_{i=0,\dots,m-1} \delta s_i \to 0$ (Hier haben wir wieder die Unabhängigkeit von δX_i und δX_j verwendet um die Terme mit $i \neq j$ verschwinden zu lassen).

Die Summe $\sum_{i=0}^{m-1} \mathcal{O}(X(s_{i+1}) - X(s_i))^3$ erfüllt

$$\begin{aligned}&\mathbb{E}\Big|\sum_{i=0}^{m-1} \mathcal{O}(X(s_{i+1}) - X(s_i))^3\Big|^2\\ &\lesssim \sum_{i\neq j=0}^{m-1} \mathbb{E}(X(s_{i+1}) - X(s_i))^3\mathbb{E}(X(s_{j+1}) - X(s_j))^3 + \sum_{i=0}^{m-1} \mathbb{E}(X(s_{i+1}) - X(s_i))^6\\ &= 15\sum_{i=0}^{m-1} \delta s_i^3,\end{aligned}$$

wo wir wieder die Unabhängigkeit der Inkremente (X ist hier nur die Brownsche Bewegung) und die Tatsache $X(s_{i+1}) - X(s_i) \sim N(0,\delta_i)$ verwenden (das dritte zentrale Moment der Normalverteilung ist Null, das sechste Moment ist $15\sigma^6 = 15\delta s_i^3$). Genau wie zuvor sehen wir, dass die Summe gegen Null geht, falls $\max_{i=0,\dots,m-1} \delta s_i \to 0$. □

Das Lemma von Ito gilt auch für mehrdimensionale Brownsche Bewegungen. Sei $G\colon [0,T]\times\Omega_{\text{prob}}\to\mathbb{R}^{d\times d}$ ein Matrix-wertiger stochastischer Prozess und B eine d-dimensionale Brownsche Bewegung, dann können wir das Ito-Integral definieren als

$$X(t) := \int_0^t G(s)\,dB(s)$$

wobei

$$X(t)_i := \sum_{j=1}^{d}\int_0^t G(s)_{ij}\,dB(s)_j$$

für $i=1,\dots,d$. Die Brownsche Bewegung (Vektor) wird also formal einfach mit der Matrix G multipliziert und dann integriert. Das können wir wieder in Differenzialschreibweise als

$$dX(t) = G(t)dB(t)$$

schreiben.

Satz 3.9 (Ito in d Dimensionen) *Sei $\phi(x_1,\dots,x_d,t)$ eine skalare Funktion die zweimal differenzierbar ist in x_i und einmal differenzierbar in t. Sei $X\colon [0,T]\times\Omega_{\text{prob}}\to\mathbb{R}^d$ ein d-dimensionaler Prozess mit*

$$dX(t) = F(t)\,dt + G(t)dB(t),$$

für einen Matrix-wertigen Prozess $G(t)\in\mathbb{R}^{d\times d}$ und einen Vektor-wertigen Prozess $F(t)\in\mathbb{R}^d$. Dann gilt für $Y(t) := \phi(X(t),t)$

$$\begin{aligned} dY(t) = &\Big(F(t)\cdot\nabla_x\phi(X(t),t) + \partial_t\phi(X(t),t)\Big)\,dt \\ &+ \sum_{i=1}^{d}\sum_{j=1}^{d}\partial_{x_i}\phi(X(t),t)G(t)_{ij}dB(t)_j \\ &+ \frac{1}{2}\Big(\sum_{k=1}^{d}\sum_{i,j=1}^{d}\partial^2_{x_ix_j}\phi(X(t),t)G(t)_{ik}G(t)_{jk}\Big)\,dt. \end{aligned}$$

Die SDE $dX(t) = F(t)\,dt + G(t)dB(t)$ im obigen Satz ist in Integralform wie folgt zu lesen:

$$X(t)_i - X(0)_i = \int_0^t F(s)_i\,ds + \sum_{j=1}^{m}\int_0^t G(s)_{ij}\,dB(s)_j$$

für alle $i=1,\dots,d$.

3.3 Die Feynman-Kac Formel

Die Feynman-Kac Formel ist nach dem Physiker Richard Feynman (1918-1988) und dem Mathematiker Mark Kac (1914-1984) benannt.

Wir betrachten zuerst die stationäre Gleichung ohne Randbedingungen: Sei $u\in C^2(\mathbb{R}^d)$ eine Lösung von

$$-\frac{1}{2}\Delta u(x) + b(x)\cdot\nabla u(x) + c(x)u(x) = f(x) \quad \text{für } x\in\mathbb{R}^d \tag{3.4}$$

wobei wir Stetigkeit von b und c voraussetzen.

Satz 3.10 *Sei $u \in C^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ eine Lösung von (3.4) mit $0 < c_0 \le c(x) \le c_1 < \infty$ und sei $X(t) \in \mathbb{R}^d$ eine Lösung der (d-dimensionalen) stochastischen Differentialgleichung*

$$dX(t) = -b(X(t))dt + dB(t),\ t \ge 0$$

mit $X(0) = x \in \mathbb{R}^d$ fast sicher. Dann gilt

$$u(x) = \mathbb{E}\Big(\int_0^\infty \exp\Big(-\int_0^t c(X(s))\,ds\Big)\, f(X(t))\,dt\Big)$$

Beweis: Wir beweisen zunächst die Produktregel (3.5) für stochastische Prozesse: Sei g eine stückweise klassisch differenzierbare Funktion auf $[0,t]$ (die auch von ω abhängen darf) und Z ein $\mathcal{L}^2$-Prozess der $dZ = F\,dt + G dB$ erfüllt. Wir verlangen außerdem, dass g' (wo es definiert ist) gleichmäßig beschränkt ist. Dann gilt

$$\begin{aligned} g(t)Z(t) - g(s)Z(s) &= (g(t)-g(s))Z(t) + g(s)(Z(t)-Z(s)) \\ &= \int_s^t g'(r)\,dr Z(t) + g(s)\Big(\int_s^t F(r)\,dr + \int_s^t G(r)\,dB(r)\Big). \end{aligned}$$

Summieren über eine Partition $s_1, \dots, s_{m+1}$ von $[0,t]$ zeigt

$$\begin{aligned} g(t)Z(t) - g(0)Z(0) &= \sum_{i=1}^m g(s_{i+1})Z(s_{i+1}) - g(s_i)Z(s_i) \\ &= \sum_{i=1}^m \int_{s_i}^{s_{i+1}} g'(r)Z(s_{i+1})\,dr + \int_{s_i}^{s_{i+1}} g(s_i)F(r)\,dr + \int_{s_i}^{s_{i+1}} g(s_i)G(r)\,dB(r). \end{aligned}$$

Da $r \mapsto g'(r)Z(r)$ stückweise stetig und gleichmäßig beschränkt ist, konvergiert die Riemann Summe im klassischen Integral. Für das Ito Integral verwenden wir, dass $r \mapsto g(r)G(r)$ wieder ein $\mathcal{L}^2$-Prozess ist. Bilden wir den Limes über immer feinere Partitionen, erhalten wir daher

$$g(t)Z(t) - g(0)Z(0) = \int_0^t g'(r)Z(r) + g(r)F(r)\,dr + \int_0^t g(r)G(r)\,dB(r). \tag{3.5}$$

Wir wenden nun die mehrdimensionale Ito Formel auf den Prozess $u(X(t))$ an. Die Funktion $\phi(t,x) := u(x)$ ist laut Annahme zweimal stetig differenzierbar im Ort. Daher erhalten wir mit Ito's Formel

$$d(u(X(t))) = -b(X(t)) \cdot \nabla u(X(t)) + \frac{1}{2}\Delta u(X(t)\,dt + \nabla u(X(t)) \cdot dB(t).$$

Für den Prozess $H(t)$ definiert durch

$$H(t) := \exp\Big(-\int_0^t c(X(s))\,ds\Big) u(X(t))$$

gilt dann mit obiger Produktregel ($\int_0^t c(X(s))\,dx$ ist stetig differenzierbar)

$$\begin{aligned} H(t) - H(0) &= \int_0^t \exp\Big(-\int_0^s c(X(r))\,dr\Big)\Big(-b\cdot\nabla u(X(s)) + \frac{1}{2}\Delta u(X(s)) - c(X(s))u(X(s))\Big)ds \\ &\quad + \int_0^t \exp\Big(-\int_0^r c(X(r))\,dr\Big)\nabla u(X(s)) \cdot dB(s) \\ &= -\int_0^t \exp\Big(-\int_0^s c(X(r))\,dr\Big) f(X(s))ds \\ &\quad + \int_0^t \exp\Big(-\int_0^s c(X(r))\,dr\Big)\nabla u(x) \cdot dB(s). \end{aligned}$$

Wenden wir den Erwartungswert auf beiden Seiten an, so erhalten wir

$$\mathbb{E}\Big(\int_0^t \exp\Big(-\int_0^s c(X(r))\,dr\Big)\nabla u(x)\cdot dB(s)\Big)=0$$

da Ito Integrale immer Erwartungswert Null haben und

$$|\mathbb{E}(H(t))|\le \exp(-c_0 t)\|u\|_{L^\infty}\to 0 \text{ für } t\to\infty$$

nach Definition von H. Damit ist die Behauptung bewiesen. □

Die Bedingung $c(x)\ge c_0>0$ kann man vermeiden, wenn man etwas mehr Arbeit in den Beweis steckt.

Wir betrachten nun das Randwertproblem: Sei $u\in C^2(\Omega)$ eine Lösung von

$$\begin{aligned} -\frac{1}{2}\Delta u(x)+b\cdot\nabla u(x)+cu(x)&=f(x) \quad \text{für } x\in\Omega,\\ u&=g \quad \text{auf } \partial\Omega. \end{aligned} \tag{3.6}$$

Satz 3.11 *Sei $u\in C^2(\overline{\Omega})$ eine Lösung von (3.6) mit $)<c_0\le c\le c_1<\infty$ und sei X_t eine Lösung der stochastischen Differentialgleichung*

$$dX(t)=-b(X(t))dt+dB(t),\ t\ge 0$$

mit $X(0)=x\in\mathbb{R}^d$ fast sicher. Dann gilt

$$u(x)=\mathbb{E}\Big(\int_0^\gamma \exp\Big(-\int_0^t c(X(s))\,ds\Big)f(X(t))\,dt\Big)-\mathbb{E}\Big(g(X(\gamma)\exp\Big(-\int_0^\gamma c(X(s))\,ds\Big)$$

wobei $\gamma(\omega):=\inf\{t\ge 0: X(t,\omega)\in\mathbb{R}^d\setminus\Omega\}$ als die boundary hitting time *bezeichnet wird.*

Beweis: Der Beweis funktioniert zunächst genau wie der Beweis von Satz 3.10 wobei wir nun

$$H(t):=\exp\Big(-\int_0^{\min(t,\gamma)} c(X(s))\,ds\Big)u(X(\min(t,\gamma)))$$

definieren. Funktionion $t\mapsto \exp\Big(-\int_0^{\min(t,\gamma)} c(X(s))\,ds\Big)$ ist nun nur stückweise stetig differenzierbar (eventueller Knick bei γ), allerdings noch immer gleichmäßig beschränkt. Die Ito Formel und die Produktregel liefern daher wieder

$$\begin{aligned} H(t)-H(0)=&-\int_0^{\min(t,\gamma)}\exp\Big(-\int_0^s c(X(r))\,dr\Big)f(X(s))ds\\ &+\int_0^{\min(t,\gamma)}\exp\Big(-\int_0^s c(X(r))\,dr\Big)\nabla u(x)\cdot dB(s). \end{aligned}$$

Anwenden des Erwartungswerts auf beiden Seiten zeigt wie oben

$$\mathbb{E}(H(t))-\mathbb{E}(H(0))=-\mathbb{E}\Big(\int_0^{\min(t,\gamma)}\exp\Big(-\int_0^s c(X(r))\,dr\Big)f(X(s))ds\Big).$$

Da $c\ge c_0>0$, haben wir mit $\exp(-c_0 s)\|u\|_{L^\infty}$ eine integrierbare Majorante und zeigen mit dem Majorantenkriterium, dass

$$\lim_{t\to\infty}\mathbb{E}(H(t))=\mathbb{E}\Big(g(X(\gamma))\exp\Big(-\int_0^\gamma c(X(s))\,ds\Big)\Big).$$

Das beweist die Behauptung. □

Zuletzt betrachten wir die parabolische Gleichung

$$\begin{aligned} \partial_t u(t,x)-\frac{1}{2}\Delta u(t,x)+b\cdot\nabla u(t,x)+cu(t,x)&=0 \quad \text{für } (t,x)\in[0,T]\times\Omega,\\ u&=g \quad \text{auf } [0,T]\times\partial\Omega,\\ u(0)&=u_0. \end{aligned} \tag{3.7}$$

Der Einfachheit halber erweitern wir die Funktion g, so dass $g(0,x)=u_0(x)$.

Satz 3.12 *Sei u zweimal im Ort differenzierbar auf $\overline{\Omega}$ und einmal in der Zeit differenzierbar und sei u eine Lösung von (3.7) mit $0 < c_0 \leq c \leq c_1 < \infty$. Für $(t,x) \in [0,T] \times \Omega$, sei X eine Lösung der stochastischen Differentialgleichung*

$$dX(s) = -b(X(s), t-s)ds + dB(s),\ s \geq 0$$

mit $X(0) = x \in \mathbb{R}^d$ fast sicher. Dann gilt

$$u(t,x) = -\mathbb{E}\Big(g(t-\gamma_t, X(t)) \exp\Big(-\int_0^{\gamma_t} c(\gamma_t - s, X(s))\, ds\Big)$$

wobei $\gamma_t(\omega) := \min\{t, \inf\{s \geq 0 : X(s,\omega) \in \mathbb{R}^d \setminus \Omega\}\}$ nun leicht verändert definiert ist sodass $\gamma_t(\omega)$ der Zeitpunkt ist an dem der Prozess $(X(s), t-s)$ den parabolischen Rand $\Omega \times \{0\} \cup \partial\Omega \times [0,T]$ trifft.

Beweis: Wieder wenden wir die Ito Formel auf $w := s \mapsto u(t-s, X(s))$ an und erhalten

$$dw(s) = \big(-b(t-s,X(s)) \cdot \nabla_x u(t-s,X(s)) - \partial_s u(t-s,X(s))\big)\, ds + \nabla_x u(t-s,X(s)) \cdot dB(s) + \frac{1}{2}\Delta u(t-s,X(s))\, ds$$

Mit der Produktregel aus dem Beweis von Satz 3.10 gilt nun für $t, t_0 \geq 0$ und

$$H(t) := \exp\Big(-\int_0^{\min(t,\gamma_{t_0})} c(X(s), \min(t,\gamma_{t_0}) - s)\, dr\Big) u(t_0 - \min(t,\gamma_{t_0}), X(\min(t,\gamma_{t_0}))$$

die Gleichung

$$\begin{aligned}
&H(t) - H(0) = H(\min\{t,\gamma_{t_0}\}) - H(0)\\
&= \int_0^{\min\{t,\gamma_{t_0}\}} \exp\Big(-\int_0^s c(X(r), s-r)\, dr\Big)\\
&\quad \cdot \underbrace{\Big(-c(t_0-s,X(s)) - b(t_0-s,X(s)) \cdot \nabla_x u(t_0-s,X(s)) - \partial_s u(t_0-s,X(s)) + \frac{1}{2}\Delta u(t_0-s,X(s))\Big)}_{=(-\partial_t + \frac{1}{2}\Delta - b\cdot\nabla - c)u = 0} ds\\
&\quad + \int_0^{\min\{t,\gamma_{t_0}\}} \exp\Big(-\int_0^s c(X(r), s-r)\, dr\Big) \nabla u(t_0 - s, X(s)) \cdot dB(s).
\end{aligned}$$

Anwenden des Erwartungswerts auf beide Seiten zeigt wie oben

$$\mathbb{E}(H(t)) - \mathbb{E}(H(0)) = 0.$$

Mit $\lim_{t\to\infty} \min\{t, \gamma_{t_0}\} = \gamma_{t_0} \leq t_0$ erhalten wir $u(t_0 - \gamma_{t_0}, X(\gamma_{t_0})) = g(t_0 - \gamma_{t_0}, X(\gamma_{t_0}))$ und daher

$$\lim_{t\to\infty} \mathbb{E}(H(t)) = \mathbb{E}\Big(g(t_0 - \gamma_{t_0}, X(\gamma_{t_0}) \exp\Big(-\int_0^{\gamma_{t_0}} c(\gamma_{t_0} - s, X(s))\, ds\Big).$$

Zusätzlich gilt $\mathbb{E}(H(0)) = u(t_0, x)$, was die Behauptung zeigt. □

3.3.1 Numerische Approximation von SDEs

Um Die Feynman-Kac Formeln in einen implementierbaren Algorithmus zu verwandeln, müssen wir SDEs numerisch approximieren können. Die einfachste Methode hierfür ist das Euler-Maruyama Schema. Angelehnt an das explizite Eulerverfahren, ist die Approximation für die SDE

$$dX = F(X(t),t)\, dt + G(X(t),t)\, dB(t), \quad X(0) = X_0$$

gegeben auf der Partition $0 = t_0 < t_1 < \ldots < t_n = T$ durch

$$X_{i+1} = X_i + F(X_i, t_i)\delta t_i + G(X_i, t_i)\delta B_i$$

mit $X_i \approx X(t_i)$ und $\delta t_i = t_{i+1} - t_i$, $\delta B_i := B(t_{i+1}) - B(t_i) \sim N(0, \delta t_i)$. Klarerweise sind die X_i Zufallsvariablen (der Zufall kommt durch δB_i ins Spiel).

Lemma 3.13 *Angenommen F und G sind Lipschitzstetig in beiden Argumenten und $\sup_{0\le t\le T}\mathbb{E}|X(t)|^2 < \infty$. Dann konvergiert die Euler-Maruyama Methode mit starker Ordnung $1/2$, d.h.*

$$\sqrt{\mathbb{E}|X(t_i)-X_i|^2} \le C\big(\max_{i=1,\dots,n}\delta t_i\big)^{1/2}$$

für $i=1,\dots,n$ und einer Konstante $C>0$ die nur an F, G und T hängt.

Bemerkung 3.14 Es gilt natürlich mit Hölder und $\mathbb{E}1 = 1$ auch

$$\mathbb{E}|X(t_i)-X_i| \le \sqrt{\mathbb{E}1^2}\sqrt{\mathbb{E}|X(t_i)-X_i|^2} \le C\big(\max_{i=1,\dots,n}\delta t_i\big)^{1/2}.$$

Dies ist die übliche Definition von starker Ordnung. Analog zum den Runge-Kutta Methoden höherer Ordnung gibt es auch stochastische Methoden von Ordnung ≥ 1.

Beweis: Schritt 1: Wir wissen bereits, dass X existiert und stetig ist. Mit der Ito Isometrie gilt

$$\mathbb{E}|X(t)-X(s)|^2 \lesssim |t-s|^2 \max_{s\le r\le t}\mathbb{E}F(X(r),r)^2 + \int_s^t \mathbb{E}G(X(r),r)^2\,dr \lesssim |t-s|\max_{s\le r\le t}\big(\mathbb{E}F(X(r),r)^2 + \mathbb{E}G(X(r),r)^2\big).$$

Die Lipschitzstetigkeit zeigt dann

$$\mathbb{E}|X(t)-X(s)|^2 \lesssim |t-s|\max_{s\le r\le t}(1+\mathbb{E}|X(r)|^2) \lesssim |t-s|$$

laut Annahme an X.

Schritt 2: Wir definieren $Y(t) = X(t_i)$ für $t\in[t_i,t_{i+1})$ und alle $i=0,\dots,n-1$. Analog definieren wir $r(t) := t_i$ für $t\in[t_i,t_{i+1})$ und alle $i=0,\dots,n-1$. Dann gilt

$$\begin{aligned} X_j &= X_0 + \sum_{i=0}^{j-1} F(X_i,t_i)\delta t_i + G(X_i,t_i)\delta B_i \\ &= X_0 + \int_0^{t_j} F(Y,r)\,ds + \int_0^{t_j} G(Y,r)\,dB(s). \end{aligned}$$

Analog dazu gilt für die exakte Lösung

$$X(t) = X_0 + \int_0^t F(X(s),s)\,ds + \int_0^t G(X(s),s)\,dB(s).$$

Vergleichen wir nun X_j mit $X(t)$ für $t\in[t_j,t_{j+1})$, so gilt

$$\begin{aligned} \mathbb{E}|X_j-X(t)|^2 &\lesssim \mathbb{E}|X_j-X(t_j)|^2 + \mathbb{E}|X(t_j)-X(t)|^2 \\ &\lesssim \mathbb{E}\Big(\int_0^{t_j} F(Y(s),r(s)) - F(X(s),s)\,ds\Big)^2 + \mathbb{E}\Big(\int_0^{t_j} G(Y(s),r(s)) - G(X(s),s)\,dB(r)\Big)^2 \\ &\quad + \mathbb{E}|X(t_j)-X(t)|^2. \end{aligned}$$

Schritt 3: Das Lebesgue Integral erfüllt mit Lipschitzstetigkeit und Hölder

$$\begin{aligned} \mathbb{E}\Big(\int_0^{t_j} F(Y(s),r(s)) - F(X(s),s)\,ds\Big)^2 &\le t_j\int_0^{t_j}\mathbb{E}(F(Y(s),r(s)) - F(X(s),s))^2\,ds \\ &\lesssim T\int_0^{t_j}\mathbb{E}|X(s)-Y(s)|^2 + |s-r(s)|^2\,ds. \end{aligned}$$

Für das Ito-Integral folgt analog mit der Ito-Isometrie

$$\begin{aligned} \mathbb{E}\Big(\int_0^{t_j} G(Y(s),r(s)) - G(X(s),s)\,dB(s)\Big)^2 &= \int_0^{t_j}\mathbb{E}(G(Y(s),r(s)) - G(X(s),s))^2\,ds \\ &\lesssim \int_0^{t_j}\mathbb{E}|X(s)-Y(s)|^2 + |s-r(s)|^2\,ds. \end{aligned}$$

Zusammen mit Schritt 1 und der Definition $\alpha(t) := \mathbb{E}|X(t) - Y(t)|^2 + |t - r(t)|^2$ folgt nun

$$\alpha(t) \leq C \int_0^t \alpha(s)\, ds + C \max_{i=1,\dots,n-1} \delta t_i$$

da $|t - r(t)|^2 = \delta t_i^2$. Das Lemma von Gronwall zeigt nun

$$\alpha(t) \leq C e^{CT} \max_{i=1,\dots,n-1} \delta t_i$$

und damit die Behauptung. □

Es bleibt noch den Erwartungswert numerisch zu approximieren. Dazu gibt es zahlreiche Methoden die alle eine eigene Vorlesung rechtfertigen würden (sparse grids, quasi-Monte Carlo, ...). Die einfachste und älteste Methode ist allerdings die Monte Carlo Methode. Die Idee dabei ist, den Erwartungswert durch den empirischen Erwartungswert anzunähern: Sei $f \in L^2(\Omega_{\text{prob}})$ eine quadratisch integrierbare Zufallsvariable (= Funktion auf Wahrscheinlichkeitsraum). Seien $\omega_1, \dots, \omega_n \in \Omega_{\text{prob}}$ unabhängige Zufallsvariablen die bezüglich $\mathbb{P}$ gleich verteilt sind. Dann ist die MC Approximation definiert durch

$$\mathbb{E}f = \int_{\Omega_{\text{prob}}} f\, d\mathbb{P}(\omega) = Q_n(f) := \frac{1}{n} \sum_{i=1}^n f(\omega_i).$$

Bemerkung 3.15 Wir sehen wir ω_i als zufällige Elemente in Ω_{prob}. Man kann ω_i auch also Zufallsvariable $\omega_i \colon \Omega_{\text{prob}} \to \Omega_{\text{prob}}$ betrachten.

Satz 3.16 *Sei $f \in L^2(\Omega_{\text{prob}})$. Dann gilt*

$$\sqrt{\mathbb{E}_{\omega_1,\dots,\omega_n} |\mathbb{E}f - Q_n(f)|^2} \leq \frac{\|f - \mathbb{E}f\|_{L^2(\Omega_{\text{prob}})}}{\sqrt{n}}.$$

Beachte: $Q_n(f)$ ist auch eine Zufallsvariable da die $\omega_1, \dots, \omega_n$ zufällig gewählt werden. Der äußere Erwartungswert bezieht sich auf diese Variablen.

Beweis: Es gilt

$$\mathbb{E}_{\omega_1,\dots,\omega_n} |\mathbb{E}f - Q_n(f)|^2 = (\mathbb{E}f)^2 - 2\mathbb{E}f \mathbb{E}_{\omega_1,\dots,\omega_n} Q_n(f) + \mathbb{E}_{\omega_1,\dots,\omega_n}(Q_n(f)^2).$$

Die Unabhängigkeit der ω_i zeigt

$$\mathbb{E}_{\omega_1,\dots,\omega_n}(Q_n(f)^2) = n^{-2} \sum_{i \neq j = 1}^n \mathbb{E}f(\omega_i)\mathbb{E}f(\omega_j) + n^{-2} \sum_{i=1}^n \mathbb{E}(f(\omega_i)^2) = (1 - n^{-1})(\mathbb{E}f)^2 + n^{-1}\mathbb{E}(f^2).$$

Es gilt auch

$$\mathbb{E}_{\omega_1,\dots,\omega_n} Q_n(f) = \mathbb{E}f$$

und daher

$$\begin{aligned} \mathbb{E}_{\omega_1,\dots,\omega_n} |\mathbb{E}f - Q_n(f)|^2 &= (\mathbb{E}f)^2 + (1 - n^{-1})(\mathbb{E}f)^2 + n^{-1}\mathbb{E}(f^2) - 2(\mathbb{E}f)^2 \\ &= n^{-1}(\mathbb{E}(f^2) - (\mathbb{E}(f))^2). \end{aligned}$$

Mit $\|f - \mathbb{E}f\|^2_{L^2(\Omega_{\text{prob}})} = \mathbb{E}(f^2) - (\mathbb{E}(f))^2$ zeigen wir die Behauptung. □

Bemerkung 3.17 Wir haben nun alle Werkzeuge um die Feynman-Kac Formel numerisch zu approximieren. The stochastischen Prozess approximieren wir mitttels Euler-Mayurama Schema, die Erwartungswerte berechnen wir mittels Monte-Carlo Methode. Eine informelle Abschätzung des Gesamtfehlers zeigt

$$\text{Fehler} \lesssim \delta t^{1/2} + n^{-1/2}.$$

Das heißt, für eine Genauigkeit von $\varepsilon > 0$ ist der Rechenaufwand $\delta t^{-1} n^{-1} \simeq \varepsilon^{-4}$ (für jede der n Stichproben von $X(t)$ müssen wir t^{-1} Schritte des EM-Schemas berechnen). Vergleichen wir das mit der traditionellen Vorgehensweise mit Zeitschrittverfahren und P1-FEM: Für jeden Zeitschritt muss eine FEM gelöst werden. Das FEM Gitter hat $\mathcal{O}(h^{-d})$ Elemente. Das heißt, die FEM hat im besten FAll den Aufwand $\mathcal{O}(h^{-d})$. Multipliziert mit der Anzahl der Zeitschritte ergibt sich ein Gesamtaufwand von $\mathcal{O}(h^{-d}\delta t^{-1})$ für eine Genauigkeit von $\mathcal{O}(\delta t + h)$. Für eine Genauigkeit von $\varepsilon > 0$ hat man also $\mathcal{O}(\varepsilon^{-d-1})$-Aufwand.

Ist man nur an wenigen Punktauswertungen $u(t,x)$, $(t,x) \in [0,T] \times \Omega$ der Lösung interessiert, ist die probabilistische Methode für $d \geq 4$ deutlich schneller.

Kapitel 4

lineare hyperbolische Gleichungen

4.0 hyperbolische Gleichungen

Die allgemeine Form von (Systemen von) *hyperbolischen Erhaltungsgleichungen* in "Erhaltungsform" ist

$$\partial_t \mathbf{u} + \sum_{j=1}^{d} \partial_{x_j}(\mathbf{f}^j(\mathbf{u})) = 0 \qquad (x,t) \in \Omega \times (0,\infty). \tag{4.1}$$

- Die Komponenten $\mathbf{u}_j : \Omega \to \mathbb{R}$, $j = 1,\dots,s$ der Funktion $\mathbf{u} : \Omega \to \mathbb{R}^s$ heißen *Zustandgrößen*, der Vektor $\mathbf{u}$ Zustandsvektor ("state vector").
- Die d Funktionen $\mathbf{f}^j : \mathbb{R}^s \to \mathbb{R}^s$, $j = 1,\dots,d$, heißen *Flußfunktionen*. Allgemein heißt der Definitionsbereich (der Wertebereich des Zustandsvektors) der Flußfunktionen die "Zustandsmenge" (set of states).

Hyperbolische Erhaltungsgleichungen von der Form (4.1) beschreiben zahlreiche Erhaltungsgleichung. Die (auch historisch) bedeutendsten ergeben sich aus der Stömungs- und Gasdynamik, z.B. den *Eulergleichungen*. Die Zustandsgrößen sind dann z.B. Masse, Impuls, Energie, und die Gleichungen beschreiben die Erhaltung dieser Größen.

Beispiel 4.1 (Eulergleichungen) Die *Eulergleichungen* der Gasdynamik beschreiben die Strömung eines Gases ("Fluids"). Dabei gelten Masseerhaltung, Energieerhaltung und Impulserhaltung. Wenn man mit $\mathbf{v}(x,t) \in \mathbb{R}^3$ die Geschwindigkeit von Partikeln am Ort x zum Zeitpunkt t bezeichnet, mit $\rho(x,t)$ die Dichte, mit $p = p(x,t)$ den Druck und mit e die (spezifische) innere Energie ("Temperatur") so ergibt sich mit der Gesamtenergie $E = \rho e + \frac{1}{2}|\mathbf{v}|^2$ die folgenden Erhaltungsgleichungen:

$$\begin{aligned}
\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) &= 0 && \text{Massererhaltung} \\
\partial_t(\rho \mathbf{v}_i) + \nabla \cdot (\rho \mathbf{v}\mathbf{v}_i) + \partial_{x_i} p &= 0, \quad i = 1,2,3, && \text{Impulserhaltung} \\
\partial_t E + \nabla \cdot (\mathbf{v}(E+p)) &= 0 && \text{Energieerhaltung.}
\end{aligned}$$

Tatsächlich stellt dies (im $\mathbb{R}^3$) 5 Gleichungen für 6 Unbekannte Funktionen dar. Die fehlende Gleichung kann z.B. durch ein "konstitutives Gesetz" erzeugt werden. Bei "idealen Gasen" z.B. ist der Druck p eine Funktion der inneren Energie: $p = \rho(\gamma - 1)e$, wobei γ eine Konstante ist[1]. Die Eulergleichungen können tatsächlich auf die Form (4.1) gebracht werden:

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho\mathbf{v}_1 \\ \rho\mathbf{v}_2 \\ \rho\mathbf{v}_3 \\ E \end{pmatrix}, \qquad \mathbf{f}^1 = \begin{pmatrix} \rho\mathbf{v}_1 \\ p + \rho\mathbf{v}_1^2 \\ \rho\mathbf{v}_1\mathbf{v}_2 \\ \rho\mathbf{v}_1\mathbf{v}_3 \\ \mathbf{v}_1(E+p) \end{pmatrix}, \qquad \mathbf{f}^2 = \begin{pmatrix} \rho\mathbf{v}_2 \\ \rho\mathbf{v}_1\mathbf{v}_2 \\ p + \rho\mathbf{v}_2^2 \\ \rho\mathbf{v}_2\mathbf{v}_3 \\ \mathbf{v}_2(E+p) \end{pmatrix}, \qquad \mathbf{f}^3 = \begin{pmatrix} \rho\mathbf{v}_3 \\ \rho\mathbf{v}_1\mathbf{v}_3 \\ \rho\mathbf{v}_2\mathbf{v}_3 \\ p + \rho\mathbf{v}_3^2 \\ \mathbf{v}_3(E+p) \end{pmatrix},$$

■

[1] Aus der Schule kennt man diese Beziehung in der Form $pV = nRT$, wenn man zusätzlich die innere Energie e von der Form $e = cT$ annimmt

Bemerkung 4.2 Die Gleichung (4.1) drückt eine Erhaltungsgleichung aus: Für ein beliebiges "Kontrollvolumen" $D \subset \mathbb{R}^d$ ergibt sich durch Integrieren über D und vertauschen von $\int_D$ mit $\frac{d}{dt}$

$$\frac{d}{dt}\int_D \mathbf{u}\,dx + \int_{\partial D} \mathbf{F}(\mathbf{u}, n)\,ds = 0,$$

wobei n die äußere Normale an D ist und $\mathbf{F}(\mathbf{u}, \underline{\omega}) := \sum_{j=1}^d \omega_j \mathbf{f}^j(\mathbf{u})$ der Fluß in Richtung $\underline{\omega} = (\omega_1, \ldots, \omega_d)$ ist. ∎

Ein wichtiger Spezialfall ist der Fall einer skalaren Gleichung (also nur eine einzige Erhaltungsgröße):

Beispiel 4.3 Im Fall $s = 1$ sind die Funktionen $f^1, \ldots, f^d$ reellwertig. Schreibt man $\mathbf{F}(u) := (f^1, \ldots, f^d)^\top$, so ergibt sich die Erhaltungsgleichung in der Form

$$\partial_t u + \nabla \cdot (\mathbf{F}(u)) = 0. \tag{4.2}$$

Die Erhaltungsform nimmt die etwas vertrautere Form

$$\frac{d}{dt}\int_D u\,dx + \int_{\partial D} \mathbf{F}(u) \cdot n\,ds = 0$$

an. ∎

Beispiel 4.4 Betrachtet man nur eine skalare Gleichung und nimmt an, daß $\mathbf{F}(u)$ von der Form $\mathbf{b}u$ ist, so ergibt sich die *Advektionsgleichung*

$$u_t + \mathbf{b} \cdot \nabla u = 0. \tag{4.3}$$

∎

Bemerkung 4.5 (lineare hyperbolische Systeme) Fallls die Funktionen $\mathbf{f}^j(\mathbf{u})$ die Form $\mathbf{A}_j\mathbf{u}$ haben für konstante Matrizen $\mathbf{A}_j \in \mathbb{R}^{s\times s}$, so spricht man von einem linearen System. Es hat die Form

$$\partial_t \mathbf{u} + \sum_{j=1}^d \mathbf{A}_j \partial_{x_j} \mathbf{u} = 0$$

Sind die Matrizen $\mathbf{A}_j$ alle symmetrisch, so spricht man von einem symmetrischen System ("Friedrichs system"). ∎

Strikt genommen gehört zur Hyperbolizität des Systems (4.1) noch eine Bedingung der reellen Diagonalisierbarkeit der Linearisierung:

Definition 4.6 (Hyperbolizität einer Erhaltungsgleichung) *(4.1) heißt hyperbolisch, falls die Ableitung $D_{\mathbf{u}}\mathbf{F}(\mathbf{u}, \underline{\omega})$ für jeden Zustandsvektor $\mathbf{u}$ und jede Richtung $\underline{\omega} \in \mathbb{R}^d \setminus \{0\}$ reell diagonalisierbar ist. Falls die Eigenwerte (für jeden Zustand $\mathbf{u}$ und jede Richtung $\underline{\Omega}$) paarweise verschieden sind, dann heißt das System* strikt hyperbolisch.

Historisch wichtig ist der Fall $d = 1$ von Systemen:

Übung 4.7 Sei $d = 1$. Das System hat die Form

$$\partial_t \mathbf{u} + \partial_x(\mathbf{F}(\mathbf{u})) = 0 \tag{4.4}$$

Zeigen Sie: es ist hyperbolisch, falls $D\mathbf{F}(\mathbf{u})$ reell diagonalisierbar ist für alle Zustandsvektoren $\mathbf{u}$. ∎

Bemerkung 4.8 Das Problem (4.1) muß noch mit Randbedingungen (und Anfangsbedingungen) vervollständigt werden. ∎

AUSARBEITEN: FD fuer glatte Lsg wie WEllengleichung, KdV,...—FVM fuer unstetige Lsgnen, SChocks

4.1 klassische Differenzenverfahren am Beispiel der Advektionsgleichung

4.1.1 Vorbemerkungen zu finiten Differenzenverfahren

Vorbemerkung: Differenzenverfahren erzeugen numerische Verfahren, indem Ableitungen durch Differenzenquotienten ersetzt werden.

Vorteile:

- einfache, schnelle Erzeugung des Verfahrens
- einfacher Umgang mit *nichtlinearen* Gleichungen

Nachteile: Handhabung komplexer Geometrien, Randbedingungen

Fokus des vorliegenden Abschnittes: Stabilität des Verfahrens in der Zeit ($\rightarrow$ Randbedingungen werden ausgeblendet durch Betrachten eines Vollraumproblems)

Eine simultane Diskretisierung in Ort und Zeit wie wir es in Abschnitt 4.4 vorstellen werden erfolgt in der Praxis selten. Fast ausschließlich werden bei (zeitabhängigen) hyperbolischen Problemen Zeitschrittverfahren eingesetzt—meist sogar explizit in der Zeit. Eine der zentralen Fragen bei solchen Verfahren ist die der Stabilität, und der vorliegende Abschnitt ist primär dieser Frage gewidmet. Eine zweite Frage ist, insbesondere bei Differenzenverfahren, die Realisierung von Randbedingungen. Wir werden diese Frage allenfalls kursorisch behandeln. Um die Frage nach Randbedingungen zu umgehen, betrachten wir ein reines Cauchyproblem (d.h. $\Omega = \mathbb{R}^d$– die klassische Alternative ist die Untersuchung von periodischen Randbedingungen). Um die Situation noch einfacher zu gestalten, betrachten wir den räumlichen 1D-Fall.

4.1.2 FD für die Advektionsgleichung

Der einfachste Fall einer linearen hyperbolischen Gleichung ist damit die Advektionsgleichung:

$$u_t + a u_x = g \qquad \text{auf } \mathbb{R} \times (0, \infty), \qquad u(x, 0) = u_0(x), \tag{4.5}$$

wobei g und der Startwert u_0 kompakten Träger haben mögen. Für $g \equiv 0$ kann die Lösung explizit angegeben werden:

$$u(x, t) = u_0(x - at). \tag{4.6}$$

Bei *Differenzenverfahren* werden Ableitungen durch Differenzenquotienten approximiert. Wir betrachten ein regelmäßiges Gitter $x_i = ih$, $i \in \mathbb{Z}$ im Ort und ein regelmäßiges Gitter in der Zeit $t_n = nk$, $n = 0$, $1, \ldots,$. Die (zu berechnenden) Werte u_i^n sollen Approximationen an die (unbekannten) Funktionswerte $u(x_i, t_n)$ sein.

Wir führen den Begriff der "Gitterfunktion" ein: eine Folge $(U_i)_{i\in\mathbb{Z}}$ heißt Gitterfunktion–man kann sich dies als die Werte in den Knoten $x_i = ih$ vorstellen. Um den Zeitschritt n zu markieren, verwenden wir einen Superskript. Z.B. kann man sich bei der Gitterfunktion $U^n = (U_i^n)_{i\in\mathbb{Z}}$ Werte in den Punkten (x_i, t_n) vorstellen. Wird sie mit einem Superskript versehen, so beschreibt dieser den Zeitpunkt.

Wir betrachten folgende *explizite* Verfahren:

1. "forward time/backward space":

$$\frac{u_i^{n+1} - u_i^n}{k} + a \frac{u_i^n - u_{i-1}^n}{h} = g(x_i, t_n). \tag{4.7}$$

2. "forward time/forward space":

$$\frac{u_i^{n+1} - u_i^n}{k} + a \frac{u_{i+1}^n - u_i^n}{h} = g(x_i, t_n). \tag{4.8}$$

3. "Lax-Friedrichs":

$$\frac{1}{k}\left(u_i^{n+1} - \frac{1}{2}\left(u_{i+1}^n + u_{i-1}^n\right)\right) + \frac{a}{2h}\left(u_{i+1}^n - u_{i-1}^n\right) = g(x_i, t_n). \tag{4.9}$$

Faßt man $U^n := (u_i^n)_{i\in\mathbb{Z}}$ als Gitterfunktion auf, so haben die Schemata die Form

$$U^{n+1} = EU^n + kG^n, \tag{4.10}$$

wobei der *Propagationsoperator* E ein *linearer* Operator auf dem Raum der *Gitterfunktionen* ist; wir schreiben weiter G^n für die Gitterfunktion $(g(x_i, t_n))_{i\in\mathbb{Z}}$.

Wie bei linearen Problemen üblich sind die zentralen Begriffe "Konsistenz" und "Stabilität"[2]. Konsistenz ist der Fehler, den das numerische Verfahren in einem (Zeit-)Schritt macht, d.h. man vergleicht die exakte Lösung nach einem Zeitschritt mit dem Ergebnis, das man aus dem Verfahren erhält, wenn man die exakte Lösung als Anfangswert wählt. Also: Sei u eine exakte Lösung und die Gitterfunktion U_{kh}^n gegeben durch

$$U_{kh,i}^n := u(x_i, t_n)$$

dann ist der Konsistenzfehler

$$\tau_i^{n+1} = \frac{1}{k}\left[U_{kh,i}^{n+1} - ((EU_{kh}^n)_i + kG_i^n)\right].$$

Beim Berechnen des Konsistenzfehlers (in Abhängigkeit von k und h) muß man die Eigenschaft nutzen, daß u die Differentialgleichung löst—das ist für die Bestimmung der Konsistenzordnung unpraktisch. Für die vorliegenden Diskretisierungen nutzt man einfach direkt die Gleichung $u_t + au_x = g$ aus, so daß man den Konsistenzfehler τ definiert als

$$\tau_i^{n+1} = \frac{1}{k}\left(U_{kh,i}^{n+1} - (EU_{kh}^n)_i\right) - (u_t(x_i, t_n) + au_x(x_i, t_n)) \tag{4.11}$$

für jede hinreichend glatte Funktion u.

Übung 4.9 Zeigen Sie mittels Taylorentwicklung, daß für die obigen Verfahren gilt:

$$|\tau_i^n| \le C\,[k + h]\,,$$

wobei die Konstante C von u abhängt. M.a.W.: die Verfahren sind von der Ordnung $(1, 1)$.

Während der Begriff der Konsistenz den Fehler faßt, der in *einem* Zeitschritt gemacht wird, faßt der Begriff der *Stabilität* den Einfluß von Fehlern (z.B. Konsistenzfehlern) in vergangenen Zeitschritten auf den Fehler. M.a.W.: Stabilität mißt die Verstärkung von Fehlern durch das Verfahren. Definiere den *Fehler*

$$\varepsilon_i^n := u(x_i, t_n) - u_i^n = U_{kh,i}^n - u_i^n.$$

Für das Verfahren und den Konsistenzfehler gilt:

$$\begin{aligned} U^{n+1} &= EU^n + kG^n \\ U_{kh}^{n+1} &= EU_{kh}^n + kG^n + k\tau^{n+1} \end{aligned}$$

Wegen der Linearität von E ergibt sich die Rekursion

$$\varepsilon^{n+1} = E\varepsilon^n + k\tau^{n+1}.$$

Wir fixieren nun eine Norm auf dem Raum der Gitterfunktionen:

$$\|(V_i)_{i\in\mathbb{Z}}\|_{\ell_h^1} := \sum_{i\in\mathbb{Z}} h|V_i|.$$

Es ergibt sich

$$\|\varepsilon^n\|_{\ell_h^1} \le \|E^n\|_{\ell_h^1}\|\varepsilon^0\|_{\ell_h^1} + k\sum_{\ell=1}^{n}\|E^{n-\ell}\|_{\ell_h^1}\|\tau^\ell\|_{\ell_h^1}$$

[2]Für lineare, wohlgestellte Probleme besagt das Lax'sche Äquivalenzprinzip tatsächlich: Konvergenz $\Longleftrightarrow$ Konsistenz + Stabilität

Wir erkennen, daß wir für festes T und $0 \le nk \le T$ fordern sollten

$$\sup_{n:0\le nk\le T} \|E^n\|_{\ell_h^1} \le C_T, \tag{4.12}$$

für ein $C_T > 0$ unabhängig von k und h, denn dann ergibt sich:

$$\|\varepsilon^n\|_{\ell^1} \le C_T \left[\|\varepsilon^0\|_{\ell_h^1} + \sup_{\ell\le n} \|\tau^\ell\|_{\ell_h^1} \underbrace{\sum_{\ell=1}^{n} k}_{\le T} \right]$$

Für die betrachteten Verfahren ergibt sich somit

$$\sup_{n:0\le nk\le T} \|\varepsilon^n\|_{\ell_h^1} \le C_T \left[\|\varepsilon^0\|_{\ell_h^1} + C(k+h) \right].$$

Man kann sich als Startfehler $\|\varepsilon^0\|_{\ell_h^1}$ also $O(h)$ erlauben, was mit Knotenauswertung oder sogar bei Wahl als Mittelwerte über Elemente der Fall ist. Entscheidend ist die Stabilitätsbedingung (4.12). Praktisch ist sie erfüllt, falls

- $\|E\|_{\ell_h^1} \le 1$
- oder doch wenigstens $\|E\|_{\ell_h^1} \le 1 + Ck$ für eine Konstante $C > 0$, die nicht von k und h ist.

Typisch für explizite Verfahren ist, daß dieser Bedingungen nicht für alle Kombinationen von k und h erfüllt sind sondern unter der Bedingung, daß k/h hinreichend klein ist, die sog. "CFL"-Bedingung[3]

$$\frac{|ak|}{h} \le c, \tag{4.13}$$

wobei $c > 0$ eine Konstante ist, die nicht von k und h abhängt (wohl aber vom Problem und dem Verfahren).

Beispiel 4.10 (Advektionsgleichung auf Torus)

$$\begin{aligned} &u_t + u_x = 0, \qquad x \in (0,1), \quad t > 0, \qquad u(0,t) = u(1,t) \quad \forall t > 0 \\ &u(x,0) = \sin(\pi x) \end{aligned}$$

Figuren 4.1–4.4 zeigen das Verhalten der unterschiedlichen Verfahren. Insbesondere sieht man, daß die CFL-Bedingung $\lambda = \frac{k|a|}{h} \le 1$ wesentlich ist bei diesen expliziten Verfahren. ∎

Übung 4.11 Zeigen Sie, daß für das Lax-Friedrichs-Schema unter der Voraussetzung der CFL-Bedingung $|ak/h| \le 1$ die folgende Abschätzung gilt:

$$\|E\|_{\ell_h^1} \le 1$$

4.1.3 Upwinding

Wir untersuchen nun "forward time/forward space" und "forward time/backward space". Als entscheidend für die Stabilität wird sich das Vorzeichen von a herausstellen:

$$\begin{cases} \text{verwende ft/bs (4.7)} & \text{falls } a > 0 \\ \text{verwende ft/fs (4.8)} & \text{falls } a < 0\ . \end{cases} \tag{4.14}$$

Satz 4.12 (Stabilität des Upwind-Verfahrens) *Unter der CFL-Bedingung $|ak/h| \le 1$ erfüllt der Propagationsoperator E des Upwind-Verfahrens (4.14) die Stabilitätsbedingung $\|E\|_{\ell_h^1} \le 1$. Für glatte Lösungen ist damit der Fehler $O(k+h)$.*

[3] Courant-Friedrichs-Lewy-Bedingung, benannt nach Richard Courant, Kurt Friedrichs und Hans Lewy

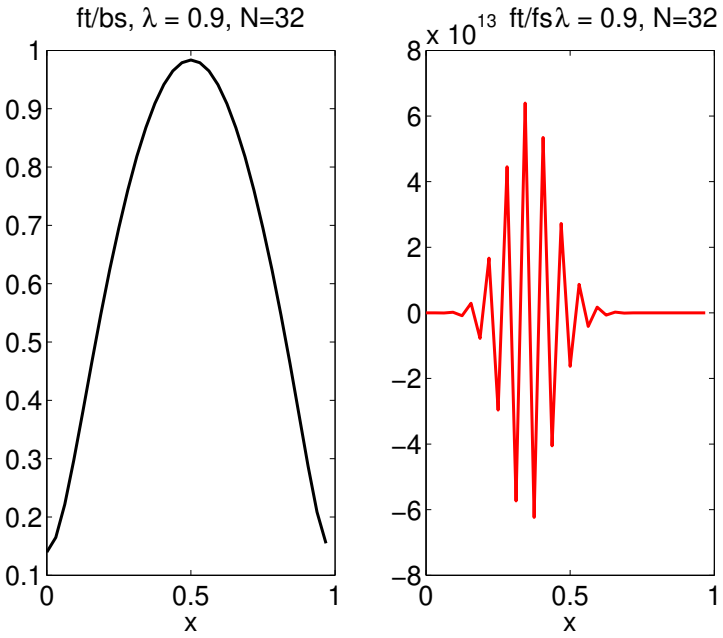


Abbildung 4.1: exakte Lösung; Vergleich ft/bs mit ft/fs

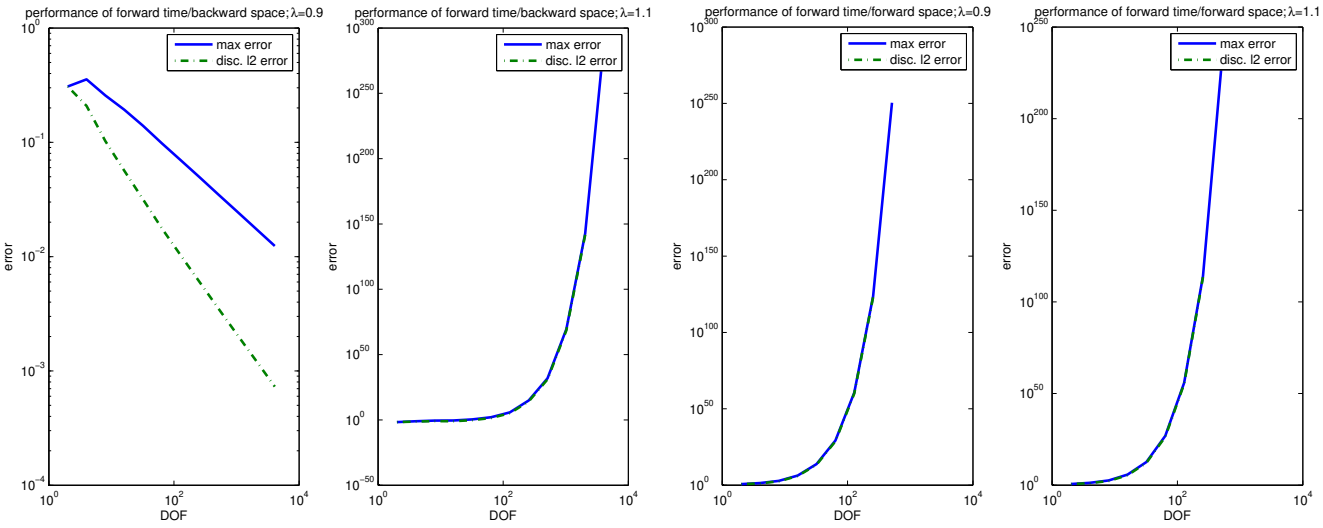


Abbildung 4.2: Einfluß der CFL-Bedingung auf ft/bs und ft/fs

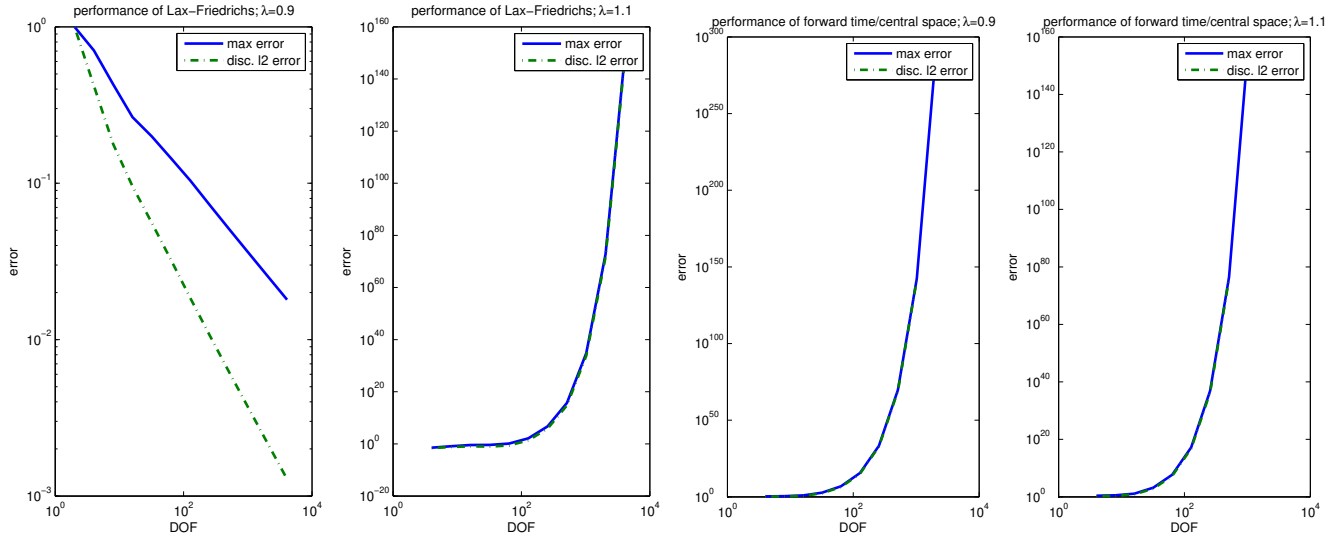


Abbildung 4.3: Einfluß der CFL-Bedingung auf Lax-Friedrichs und ft/central

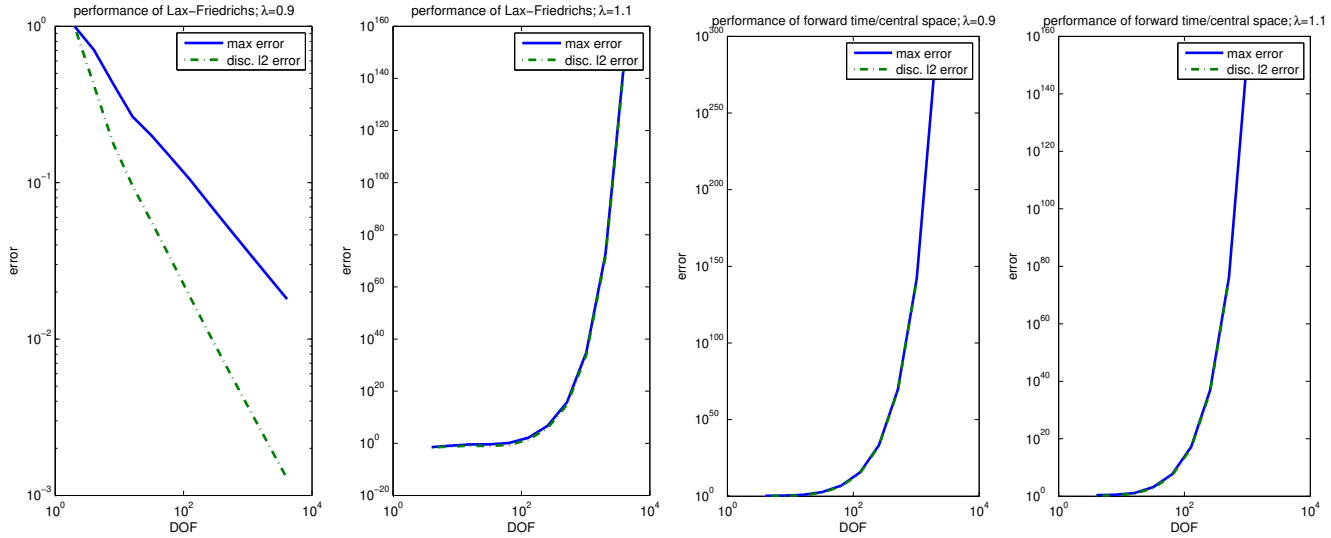


Abbildung 4.4: Einfluß der CFL-Bedingung auf das Leap Frog Verfahren

Beweis: Betrachte den Fall $a > 0$. Für eine Gitterfunktion $(V_i)_{i\in\mathbb{Z}}$ gilt damit

$$(EV)_i = V_i - \frac{ak}{h}\,(V_i - V_{i-1})$$

und damit

$$\begin{aligned}\|EV\|_{\ell^1_h} &\leq \sum_i h\left[\left|1-\frac{ka}{h}\right||V_i| + |V_{i-1}|\left|\frac{ka}{h}\right|\right]\\ &\overset{(4.13)}{=} \left(1-\frac{ak}{h}\right)\sum_i h|V_i| + \frac{ak}{h}\sum_i h|V_{i-1}| = \|V\|_{\ell^1_h}.\end{aligned}$$

□

Übung 4.13 Zeigen Sie im Setting von Satz 4.12: falls die CFL-Bedingung $|a|k/h \leq 1$ gilt, dann ist $\|E\|_{\ell^\infty\leftarrow\ell^\infty} \leq 1$, d.h. das Verfahren ist in ℓ^∞ stabil. ∎

Bemerkung 4.14 vergleiche auch: den Begriff der Monotonie des Verfahrens — siehe das Buch von Asher

Bemerkung 4.15 (physikalische Plausibilität der CFL-Bedingung und des Upwinding) Die Lösungsformel (4.6) zeigt, daß die exakte Lösung bei (x_i, T) durch $u_0(x_i - aT)$ gegeben ist. Damit ist eine notwendige Bedingung an das numerische Verfahren, daß zum Zeitpunkt $t_n = T$ die Startwerte in der Nähe von $x_i - aT$ "gesampled" werden. Betrachtet man das ft/bs-Verfahren, so "sieht" der Punkt (x_i, t_1) die Werte bei (x_i, t_0) und (x_{i-1}, t_1), der Punkt (x_i, t_2) entsprechend die Werte bei (x_{i-2}, t_0), (x_{i-1}, t_0), (x_i, t_0) etc. Bei $t_n = T$ damit die Startwerte bei (x_{i-j}, t_0), $j = 0, \ldots, n$. Falls also das Intervall $[x_{i-n}, x_i]$ *nicht* den Punkt $x_i - aT$ enthält, dann kann das Verfahren nicht funktionieren.

- Im Fall $a < 0$ kann das Verfahren also *nicht* funktionieren.
- Im Fall $a > 0$ muß zusätzlich die Bedingung $nh \geq aT$, d.h. $nh \geq akn$ gelten, was genau die CFL-Bedingung $ka/h \leq 1$ ausdrückt. ∎

Bemerkung 4.16 Upwinding kann auch für Systeme formuliert werden. Sei $\mathbf{A}$ eine konstante Matrix und betrachte

$$\mathbf{U}_t + \mathbf{A}\mathbf{U}_x = 0$$

Kann man $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ diagonalisieren, dann würde man in den neuen Variablen $\widetilde{\mathbf{U}} = \mathbf{V}^{-1}\mathbf{U}$ ft/bs für die Komponenten mit $\mathbf{D}_{ii} > 0$ machen und ft/fs für die Komponenten mit $\mathbf{D}_{ii} < 0$. Schreibe

$$\mathbf{D} = \mathbf{D}^+ + \mathbf{D}^-,$$

wobei $\mathbf{D}^+$ die positiven und $\mathbf{D}^-$ die negativen Diagonalelemente von $\mathbf{D}$ hat (formal: $\mathbf{D}^+_{ii} = \max\{\mathbf{D}_{ii}, 0\}$, $\mathbf{D}^-_{ii} = \min\{\mathbf{D}_{ii}, 0\}$). Das Upwind-Verfahren ist dann

$$\widetilde{\mathbf{U}}^{n+1}_j = \widetilde{\mathbf{U}}^n_j - \frac{k}{h}\mathbf{D}^+(\widetilde{\mathbf{U}}^n_j - \widetilde{\mathbf{U}}^n_{j-1}) - \frac{k}{h}\mathbf{D}^-(\widetilde{\mathbf{U}}^n_{j+1} - \widetilde{\mathbf{U}}^n_j);$$

Rücktransformation zur $\mathbf{U}$-Variablen führt mit

$$\mathbf{A}^+ = \mathbf{V}\mathbf{D}^+\mathbf{V}^{-1}, \qquad \mathbf{A}^- = \mathbf{V}\mathbf{D}^-\mathbf{V}^{-1}$$

auf

$$\mathbf{U}^{n+1}_j = \mathbf{U}^n_j - \frac{k}{h}\mathbf{A}^+(\mathbf{U}^n_j - \mathbf{U}^n_{j-1}) - \frac{k}{h}\mathbf{A}^-(\mathbf{U}^n_{j+1} - \mathbf{U}^n_j);$$

4.2 von Neumann-Analyse

Die klassische Stabilitätsanalyse wird nicht in $\|\cdot\|_{\ell_h^1}$ durchgeführt sondern in der Norm

$$\|(V_i)\|_{\ell_h^2} := \left(\sum_i h|V_i|^2\right)^{1/2}.$$

Entsprechend bezeichnen wir mit ℓ_h^2 den Raum der Folgen mit endlicher Norm.

Der Grund ist in erster Linie ein technischer: für lineare Differentialgleichungen mit konstanten Koeffizienten auf regelmäßigen Gittern liefert die Fourieranalyse ein sehr bequemes Werkzeug. Man definiert:

- Die "Fouriertransformation" einer Folge $(v_i)_{i\in\mathbb{Z}}$:

$$\widehat{v}(\xi) := (\mathcal{F}_h(v))(\xi) := h\sum_j e^{-\mathbf{i}\xi x_j} v_j, \qquad \xi \in [-\pi/h, \pi/h]$$

- die L_h^2-norm:

$$\|\widehat{v}\|_{L_h^2}^2 := \int_{-\pi/h}^{\pi/h} |\widehat{v}(\xi)|^2\, d\xi$$

- Die Faltung zweier Folgen $u = (u_i)_i$, $v = (v_i)_i$

$$(u * v)_i := h\sum_j u_{i-j} v_j = h\sum_j u_j v_{i-j}$$

Es gilt:

Satz 4.17 *(i) (Parseval) $\mathcal{F}_h$ ist ein Isomorphismus $\ell_h^2 \to L_h^2$:*

$$\sqrt{2\pi}\|(v_i)_{i\in\mathbb{Z}}\|_{\ell_h^2} = \|\widehat{v}\|_{L_h^2}, \qquad \widehat{v} = \mathcal{F}_h((v_i)_{i\in\mathbb{Z}}).$$

Die Inverse ist gegeben durch

$$v_j = (\mathcal{F}_h^{-1}(\widehat{v}))_j = \frac{1}{2\pi}\int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x_j}\widehat{v}(\xi)\, d\xi$$

(ii) Für $(u_i)_{i\in\mathbb{Z}} \in \ell_h^2$ und $(v_i)_{i\in\mathbb{Z}} \in \ell_h^1$ ist $u * v \in \ell_h^2$ und

$$\widehat{(u * v)}(\xi) = \widehat{u}(\xi)\widehat{v}(\xi)$$

(iii) (Translation) Für $j_0 \in \mathbb{Z}$ ist $\mathcal{F}_h((v_{j+j_0})_{j\in\mathbb{Z}})(\xi) = e^{\mathbf{i}\xi x_{j_0}}\widehat{v}(\xi)$

(iv) (Modulation) Für $\xi_0 \in \mathbb{R}$ ist $\mathcal{F}_h((e^{\mathbf{i}\xi_0 x_j} v_j)_{j\in\mathbb{Z}})(\xi) = \widehat{v}(\xi - \xi_0)$

(v) (Dilation) Für $m \in \mathbb{Z}\setminus\{0\}$ ist $\mathcal{F}_h((v_{mj})_{j\in\mathbb{Z}})(\xi) = \widehat{v}(\xi/m)/|m|$

(vi) (Konjugation) $\mathcal{F}_h((\overline{v_j})_{j\in\mathbb{Z}})(\xi) = \overline{\widehat{v}(-\xi)}$

Beweis: Übung. □

Betrachte ein Einschrittverfahren mit Propagationsoperator E der Form

$$(Ev)_i = \sum_{\ell=-r}^{s} \alpha_\ell v_{i+\ell}$$

für Koeffizienten α_ℓ. Der Operator E ist vom Faltungstyp:

$$(Ev) = a * v, \qquad a_\ell = \frac{1}{h}\alpha_{-\ell}.$$

Damit ist

$$\widehat{(Ev)}(\xi) = \widehat{a}(\xi)\widehat{v}(\xi),$$

wobei $\widehat{a}$ der *Verstärkungsfaktor* genannt wird. Es ist:

Operator	Symbol
Vorwärtsdifferenz D_+: $u_{i+1} - u_i$	$e^{\mathbf{i}\xi} - 1$
Rückwärtsdifferenz D_-: $u_i - u_{i-1}$	$1 - e^{-\mathbf{i}\xi}$
symmetrische Differenz D_0: $u_{i+1} - u_{i-1}$	$2\mathbf{i}\sin\xi$
δ: $u_{i+1/2} - u_{i-1/2}$	$2\mathbf{i}\sin(\xi/2)$
D_+D_-	$-4\sin^2(\xi/2)$

Abbildung 4.5: Symbole einiger Differenzenoperatoren

Übung 4.18

$$\|E\|_{\ell_h^2 \leftarrow \ell_h^2} = \max_{\xi\in[-\pi/h,\pi/h]} |\widehat{a}(\xi)|$$

Damit ist die Stabilitätsanalyse zurückgeführt auf die Berechnung von $\widehat{a}$. Man sagt, daß ein Verfahren die *von-Neumann-Stabilitätsbedingung* erfüllt, falls der zugehörige Verstärkungsfaktor $\widehat{a}$ die folgende Bedingung erfüllt:

$$|\widehat{a}(\xi)| \le 1 + Ck \qquad \forall \xi \in [-\pi/h, \pi/h], \tag{4.15}$$

wobei $C > 0$ unabhängig von k (und natürlich ξ) ist.

Beispiel 4.19 (Upwind-Verfahren) Das Upwind-Verfahren für die Advektionsgleichung im Fall $a = -1$ (und $g \equiv 0$) ist

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \lambda(v_{j+1}^n - v_j^n), \qquad \lambda = \frac{k}{h}.$$

Dies hat die Form $Ev^n = a * v^n$ mit

$$a_j = \frac{1}{h}\lambda\delta_{-1,j} + \frac{1}{h}(1-\lambda)\delta_{j,0}, \qquad\qquad \delta_{n,m} = \text{Kronecker } \delta$$

Die Fouriertransformierte ist

$$\widehat{a}(\xi) = h(e^{-\mathbf{i}\xi x_{-1}}a_{-1} + e^{-\mathbf{i}\xi x_0}a_0) = \lambda e^{\mathbf{i}\xi h} + (1-\lambda)$$

Man sieht, daß im Fall $0 < \lambda \le 1$ gilt:

$$|\widehat{a}(\xi)| \le 1 \qquad \forall \xi \in [-\pi/h, \pi/h],$$

d.h. das Verfahren erfüllt die von-Neumann-Bedingung (4.15). ∎

In der Praxis kürzt man die Bestimmung von $g(\xi) := \widehat{a}(\xi)$ mit einer "Rechenregel" erheblich ab: man macht den Ansatz $v_j^n = g^n e^{\mathbf{i}\xi jh}$ und setzt in das Verfahren ein, um eine Formel für $g(\xi)$ zu erhalten.

Beispiel 4.20 (Lax-Wendroff) Für die Advektionsgleichung mit $a = -1$ ist das Verfahren gegeben durch

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \frac{1}{2}\lambda(v_{j+1}^n - v_{j-1}^n) + \frac{1}{2}\lambda^2(v_{j+1}^n - 2v_j^n + v_{j-1}^n), \qquad \lambda = k/h.$$

Einsetzen des Ansatzes $v_j^n = g^n e^{\mathbf{i}\xi jh}$ und Kürzen mit $g^n e^{\mathbf{i}\xi jh}$ liefert

$$g(\xi) = 1 + \frac{1}{2}\lambda(e^{\mathbf{i}\xi h} - e^{-\mathbf{i}\xi h}) + \frac{1}{2}\lambda^2(e^{\mathbf{i}\xi h} - 2 + e^{-\mathbf{i}\xi h}) = 1 + \mathbf{i}\lambda\sin\xi h - 2\lambda^2\sin^2\frac{\xi h}{2}.$$

Eine elementare Rechnung zeigt, daß auch hier für $\lambda \in (0,1]$ die Bedingung

$$\sup_{\xi\in[-\pi/h,\pi/h]} |g(\xi)|^2 = \sup_{\xi\in[-\pi/h,\pi/h]} \left(1 - 4\lambda^2(1-\lambda^2)\sin^4(\xi h/2)\right) \overset{\lambda\in(0,1]}{\le} 1$$

erfüllt ist, d.h. die von-Neumann-Bedingung (4.15).

Die Form des Verstärkungsfaktors g zeigt auch, daß das Lax-Wendroff-Verfahren *dissipativ*[4] ist: Während $g(\xi) \approx 1$ für $\xi \approx 0$ ist für $|\xi| \approx \pi/h$ sogar $|g(\xi)| \approx 1 - 4\lambda^2(1-\lambda^2) < 1$, falls $\lambda < 1$. M.a.W.: Während die niederfrequenten Term (wegen Konsistenz!) weder verstärkt noch gedämpft werden, werden die hochfrequenten Lösungskomponenten (und damit auch die hochfrequenten Fehleranteile) gedämpft.
■

Die Bedeutung des Faktors Ck in der Bedingung (4.15) ergibt aus dem Wunsch, auch Gleichungen mit Termen niedrigerer Ordnung zu behandeln:

Übung 4.21 Betrachten Sie

$$u_t - u_x + c(x)u = 0$$

mit der Upwinddiskretisierung

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \lambda(v_{j+1}^n - v_j^n) + kc(x_j)v_j, \qquad \lambda = \frac{k}{h}. \tag{4.16}$$

Sei E_0 die Evolution, die zu $c \equiv 0$ gehört. Aus Beispiel 4.19 folgt, daß $\|E\|_{\ell_h^2} \leq 1$. Sei nun c stetig mit $\|c\|_{L^\infty} < \infty$. Überlegen Sie sich, daß dann $\|E\|_{\ell_h^2} \leq \|E_0\|_{\ell_h^2} + k\|c\|_{L^\infty}$ gilt. Schließen Sie, daß damit auch E die von-Neumann-Bedingung erfüllt. ■

Übung 4.22 betrachten Sie die Diskretisierung der Wärmeleitungsgleichung $u_t - u_{xx} = 0$ mit dem expliziten Eulerverfahren in der Zeit (und symmetrischen Differenzen im Ort):

$$u_i^{n+1} - u_i^n = \sigma\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right), \qquad \sigma := \frac{k}{h^2}$$

Zeigen Sie, daß der Verstärkungsfaktor $g(\xi) = 1 - 4\sigma\sin^2(\xi h/2)$. Was können Sie über die Stabilität des Verfahrens in Abhängigkeit von k und h aussagen? ■

Bemerkung 4.23

- die von-Neumann-Analyse kann auch für Systeme (und damit auch für Mehrschrittverfahren wie dem leap frog Verfahren) durchgeführt werden—siehe Übungen.
- Im allg. liefert es nur *notwendige* Bedingungen für Stabilität, aber nicht hinreichend. In der Praxis liefert es aber doch eine recht gute Vorstellung davon, was die CFL-Bedingung ist.
- Für Probleme mit nichtkonstanten Koeffizienten geht man typischerweise so vor, daß man in einem ersten Schritt den Wertebereich der Koeffizienten bestimmt und dann eine von-Neumann-Analyse für die Gleichung mit eingefrorenen Koeffizienten ("freezing of coefficients") durchführt. Im Prinzip kann dies auch für nichtlineare Gleichungen durchgeführt werden. Auf diese Weise erhält man natürlich nicht scharfe Abschätzungen für die CFL-Bedingung, aber oft brauchbare Schätzwerte.

■

4.2.1 Stabilitätsanalyse des Leap Frog Verfahrens

Vorbemerkung: Leap Frog ist ein wichtiger Vertreter von 2-Schritt-Verfahren. Mehrschrittverfahren können als Vektoreinschrittverfahren verstanden und analysiert werden (siehe Übung), aber eine direkte Analyse ist oft einfacher.

Für die Advektionsgleichung ist das Verfahren

$$\frac{u_i^{n+1} - u_i^{n-1}}{2k} + a\frac{u_{i+1}^n - u_{i-1}^n}{2h} = 0 \tag{4.17}$$

[4] genauer: von der Ordnung 4. Allg. ist ein Verfahren dissipativ von der Ordnung $2r$, falls der Verstärkungsfaktor g die Beziehung $|g(\xi)| \leq (1 + Ck)(1 - C|\xi h|^{2r})$, für $\xi \in (-\pi/h, \pi/h)$ erfüllt. Siehe die Diskussion in Abschnitt 4.3. LW ist strikt dissipativ, Lax-Friedrichs ist dissipativ aber nicht strikt dissipativ; siehe Diskussion in [**?**, p. 179].

Bemerkung 4.24
- Man benötigt 2 Startwerte u_0^0 und u_i^1, $i \in \mathbb{Z}$. Die Werte u_i^1 können mit einem Einschrittverfahren erzeugt werden.
- Mit dem Vektor
$$U_i^n := \begin{pmatrix} u_i^n \\ u_i^{n-1} \end{pmatrix}$$
kann das Verfahren (4.17) als Einschrittverfahren formuliert und analysiert werden.

■

Wir machen eine Stabilitätsanalyse mittels Fouriertransformation: die Transformierte $\widehat{u}^n(\xi) := \mathcal{F}_h((u_i^n)_i)(\xi)$ erfüllt

$$\widehat{u}^{n+1}(\xi) + 2\mathbf{i}a\lambda \sin(\xi h)\widehat{u}^n(\xi) - \widehat{u}^{n-1}(\xi) = 0, \qquad \lambda = \frac{k}{h}. \tag{4.18}$$

Für feste ξ, h ist dies eine Rekurrenz in n, die aufgelöst werden kann:

Übung 4.25 Seien α, $\beta \in \mathbb{C}$. Seien x_1, x_2 die Lösungen von $0 = x^2 + \alpha x + \beta$. Betrachte alle Folgen $(v_n)_n$ mit
$$v_{n+1} + \alpha v_n + \beta v_{n-1} = 0, \qquad \forall n \geq 1.$$
Dann gilt:

(a) Der Raum diese Folgen ist ein 2-dimensionaler Vektorraum
(b) Falls $x_1 \neq x_2$, dann sind die beiden Folgen $(x_1^n)_n$, $(x_2^n)_n$ eine Basis dieses Raumes.
(c) Falls $x_1 = x_2$, dann sind $(x_1^n)_n$ und $(nx_1^n)_n$ eine Basis dieses Raumes.

Seien nun $g_+(\xi)$ und $g_-(\xi)$ Lösungen von
$$g^2 + 2\mathbf{i}a\lambda \sin(\xi h)g - 1 = 0$$
d.h.
$$g_\pm(\xi h) = -\mathbf{i}a\lambda \sin(\xi h) \pm \sqrt{1 - (a\lambda)^2 \sin^2(\xi h)}.$$
Damit können wir $\widehat{u}^n$ schreiben als:
$$\begin{aligned} \widehat{u}^n(\xi) &= \gamma(\xi)(g_+(\xi h))^n + \delta(\xi)(g_-(\xi))^n \qquad \text{falls } g_+(\xi h) \neq g_-(\xi h), \\ \widehat{u}^n(\xi) &= \gamma(\xi)(g_+(\xi h))^n + \delta(\xi)n(g_+(\xi))^n \qquad \text{falls } g_+(\xi h) = g_-(\xi h). \end{aligned}$$
Die Funktionen $\gamma(\xi)$, $\delta(\xi)$ ergeben sich aus den Startwerten $\widehat{u}^0(\xi)$, $\widehat{u}^1(\xi)$. Stabilität der Diskretisierung fordert hier
$$|g_+(\xi h)| \leq 1 \quad \text{und} \quad |g_-(\xi h)| \leq 1 \qquad \forall \xi \in (-\pi/h, \pi/h), \tag{4.19}$$
$$\text{falls } g_+(\xi) = g_-(\xi) \text{ dann muss} |g_+(\xi h)| < 1 \quad \text{sein.} \tag{4.20}$$

Wir beobachten nun:

- $|a\lambda| > 1$ impliziert $|g_-(-\pi/(2h))| > 1$ woraus folgt, daß die CFL-Bedingung $|a\lambda| \leq 1$ gelten muß.
- $|a\lambda| \leq 1$ impliziert $|g_+|^2 = |g_-|^2 = 1 - (a\lambda)^2 \sin^2(\xi h) + (a\lambda \sin(\xi h))^2 = 1$, d.h. das Verfahren ist stabil, es sei denn, $g_+(\xi h) = g_-(\xi h)$. In diesem Fall muß $1 = |a\lambda \sin(\xi h)|$, d.h. $|a\lambda| = 1$ und $\xi = \pm\pi/(2h)$. Dann ist $g_+ = g_- = \pm\mathbf{i}$ d.h. (4.20) ist verletzt.

Zusammenfassung: Falls $|a\lambda| < 1$, dann ist Leap Frog stabil. Falls $|a\lambda| = 1$, dann ist es (schwach) instabil.

4.3 Dissipative Verfahren

4.3.1 Vorbetrachtung:

Ausgangspunkt:

1. für lineare Gleichungen mit konstanten Koeffizienten liefert die von-Neumann-Analyse ein einfaches Werkzeug zur Stabilitätsanalyse

2. für *Systeme*
$$\mathbf{U}^{n+1} = E\mathbf{U}^m + k\mathbf{G}^n$$
untersucht man analog die Fouriertransformierte $\widehat{\mathbf{E}}(\xi)$. Der Einfachheit halber untersucht man den Spektralradius $\rho(\widehat{\mathbf{E}})$ für $\xi \in (-\pi/h, \pi/h)$. Man sagt, das Verfahren erfüllt die von-Neumann-Bedingung, falls $\rho(\widehat{\mathbf{E}}(\xi)) \leq 1 + Ck$ für alle $\xi \in (-\pi/h, \pi/h)$.

Probleme: variable Koeffiziente? Nichtlineare Gleichungen?

"Standardvorgehen": von-Neumann-Analyse für "frozen coefficients", d.h. man macht von-Neumann-Analyse für alle (erwarteten) Werte der Koeffizienten. Oft gibt dies gute Hinweise auf Stabilität bzw. Schrittweitenbeschränkungen. Für dissipative Verfahren und gewisse Problemklassen reicht in der Tat die Methode des "frezzing of coefficients". Dies triff insbesondere auf parabolische Probleme zu.

Das folgende Beispiel zeigt, dass die Methode des "freezing of coefficients" nicht immer zuverlässig ist:

Beispiel 4.26 (Zabusky-Kruskal-Verfahren für KdV) Wir betrachten die KdV-Gleichung $u_t + (\alpha u^2 + \nu u_{xx})_x = 0$ mit der Leap frog-artigen Diskretisierung

$$u_j^{n+1} = u_j^{n-1} - \frac{2\alpha\mu}{3}\left(u_{j-1}^n + u_j^n + u_{j+1}^n\right)\left(u_{j+1}^n - u_{j-1}^n\right) - \frac{\nu\mu}{h^2}\left(u_{j+2}^n - 2u_{j+1}^n + 2u_{j-1}^n - u_{j-2}^n\right), \qquad \mu = \frac{k}{h}.$$

Die von Neumann-Analyse liefert nach "freezing of coefficients"

$$k < \frac{h}{4|\nu|/h^2 + 2|\alpha u_{max}|}$$

Wir betrachten das Verfahren mit

- $\nu = 0.022^2, \qquad \alpha = 0.5, \qquad u_0(x) = \cos(\pi x), \qquad u(0,t) = u(2,t)$
- Beobachtung: $|u| \leq 5$
- $\Longrightarrow$ Erwartung: $h = 0.01$ und $k < 0.004$ sind stabil.
- Numerik: $h = 0.01$ und $k = 0.0001$

Die Numerik in Fig. 4.6 zeigt, daß diese konservative Schrittweitenwahl nicht ausreicht! Tatsächlich kommt es bei $T > 21/\pi$ zu einem blow-up, obwohl die exakte Lösung für alle Zeiten existiert. ∎

Das Versagen des Zabusky-Kruskal-Verfahrens [**?**][5] führt man allgemein auf eine "Instabilität" zurück in dem Sinn, daß hochfrequente Lösungsanteile (d.h. große ξ im Fourierbild) durch die Nichtlinearität verstärkt werden. Ein manchmal propagiertes Mittel ist deshalb, Verfahren mit (etwas) Dissipation zu verwenden.

4.3.2 Dissipative Verfahren

Definition 4.27 *Ein Verfahren mit Propagationsoperator* $\mathbf{E}$ *heißt* dissipativ[6] *von der Ordnung* $2r$*, falls*

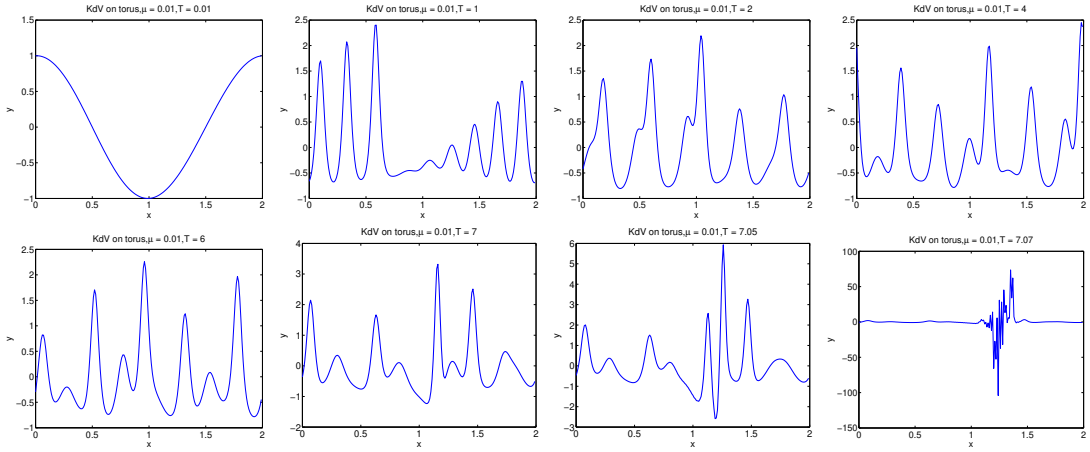
$$\rho(\widehat{\mathbf{E}}(\xi)) \leq (1 - \delta|\xi h|^{2r})(1 + Ck)$$

für Konstante C*,* $\delta > 0$ *unabhängig von* k*,* h*.*

[5] in diesem Paper wird der Begriff der Solitonen eingeführt

[6] manchmal auch: strikt dissipativ

Abbildung 4.6: das Zabusky-Kruskal-Verfahren für die KdV-Gleichung



Beispiel 4.28 Betrachte das Lax-Wendroff-Verfahren für $u_t - u_x = 0$:

$$g(\xi) = 1 + \mathbf{i}\lambda \sin(\xi h) - 2\lambda^2 \sin^2(\xi h/2)$$
$$|g(\xi)|^2 = (1 - 2\lambda^2 \sin^2(\xi h/2))^2 + \lambda^2 \sin^2(\xi h) = 1 - 4\lambda^2(1-\lambda^2)\sin^4(\xi h/2)$$

Damit ist das Verfahren dissipativ von der Ordnung $2r = 4$, *falls* $\lambda = k/h < 1$ (verwende $\sqrt{1-x} = 1 - 1/2x + O(x^2)$ für kleine x). ∎

Dissipativität tritt relativ natürlich bei Diskretisierungen von parabolischen Gleichungen auf, weil das kontinuierliche Problem ein solche Eigenschaft hat:

Beispiel 4.29 Betrachte die explizite Eulerdiskretisierung der Wärmeleitungsgleichung aus Übung 4.22 mit Verstärkungsfaktor $g(\xi)$ $g(\xi) = 1 - 4\sigma \sin^2(\xi h/2)$. Stabilität ($|g(\xi)| \leq 1$) verlangt also $\sigma \leq 1/2$. Für $\sigma \leq 1/2$ ist das Verfahren dissipativ von der Ordnung 2.

Übung: Für das implizite Eulerverfahren ist das Verstärkungsfaktor $g(\xi) = 1/(1 + 4\sigma \sin^2(\xi h/2))$. Das Verfahren ist also stabil und ebenfalls dissipativ von der Ordnung 2. ∎

Im einfachsten Fall liefert Dissipativität, daß konsistente Verfahren tatsächlich stabil sind:

Satz 4.30 (Kreiss) *Betrachte das strikt hyperbolische System*

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0$$

d.h. die konstante Matrix $\mathbf{A}$ *habe paarweise verschiedene reelle Eigenwerte. Dann ist ein (explizites) Differenzenverfahren stabil, wenn es konsistent und dissipativ von der Ordnung* $2r > 0$ *ist.*

Beweis: siehe [**?**, Thm. 5.2] und z.B. [**?**, Thm. 6.5.2]. □

Bemerkung 4.31 Das Zusammenspiel von Konsistenz und Dissipativität bereits bei parabolischen Gleichungen im Abschnitt 2.5.5 gesehen haben bei der Analyse der Funktion F_n: für "kleine $z = \lambda_n k$ nutzten wir die Konsistenz, für große $z = \lambda_n k$ benötigten wir die L-Stabilität des Einschrittverfahrens. Ähnliches liegt auch Satz 4.30 zugrunde: für konsistente Verfahren erwartet man, daß das Stabilitätsverhalten für kleine ξ sich aus dem kontinuierlichen Problem erhalten läßt (cf. Satz 4.32)), während das Stabilitätsverhalten für große ξ eine zusätzliche Eigenschaft ist.

Wie Satz 4.30 zeigt, kann Dissipativität eines Verfahrens eine nützliche Eigenschaft sein. Tatsächlich haben sehr viele Verfahren (auch für hyperbolische Probleme) ein wenig Dissipation—das Lax-Wendroff-Verfahren ist ein Beispiel. Während für parabolische Probleme wie der Wärmeleitungsgleichung Dissipation eine Eigenschaft des kontinuiertlichen Problems ist, ist es bei hyperbolischen Erhaltungsgleichungen meist nicht. Man versucht deshalb meist, die Dissipation des Verfahrens möglichst gering zu halten.

noch ueberarbeiten–streichen Aus der Konsistenz eines Verfahrens und der Wohlgestelltheit des kontinuierlichen Problems kann man $\tilde{\mathbf{E}}(\xi)$ für "kleine " ξ kontrollieren:

Satz 4.32 *Betrachte das lineare Problem $\mathbf{u}_t = P\mathbf{u}$ mit einen linearen Ortsdifferentialoperator P mit konstanten Koeffizienten. Habe das numerische Verfahren Konsistenzordnung p (in Ort und Zeit). Dann gilt:*

$$\|\widehat{\mathbf{E}}(\xi) - e^{\widehat{P}(\mathbf{i}\xi)k}\| \le C|\xi h|^{p+1},$$

wobei die Matrix $e^{\widehat{P}(\mathbf{i}\xi)k}$ den exakten Lösungsoperator beschreibt, d.h. die Lösung von $\mathbf{u}_t = P\mathbf{u}$ mit Anfangsbedingung $\mathbf{u}(\cdot,0) = e^{\mathbf{i}\xi x}\mathbf{u}_0$ ist $e^{\widehat{P}(\mathbf{i}\xi)t}\mathbf{u}_0$, wenn $\mathbf{u}_0 \in \mathbb{R}^s$ ist, mit Problemgröße $s \in \mathbb{N}$.

Beispiel 4.33 Betrachte das Lax-Wendroff-Verfahren für $u_t - u_x = 0$. Der Operator $e^{\widehat{P}(\mathbf{i}\xi)t}$ ergibt sich aus

$$e^{\widehat{P}(\mathbf{i}\xi)t}e^{\mathbf{i}\xi x} \stackrel{!}{=} u(x,t) = e^{\mathbf{i}\xi(x+t)},$$

d.h.

$$e^{\widehat{P}(\mathbf{i}\xi)t} = e^{\mathbf{i}\xi t}.$$

Der Verstärkungsfaktor $g(\xi) = \widehat{E}(\xi)$ von Lax-Wendroff ist $g(\xi) = 1 + \mathbf{i}\lambda\sin(\xi h) - 2\lambda^2\sin^2(\xi h/2)$ Damit

$$g(\xi) - e^{\widehat{P}(\mathbf{i}\xi)k} = 1 + \mathbf{i}\lambda\sin(\xi h) - 2\lambda^2\sin^2(\xi h/2)e^{\mathbf{i}\xi k} = \cdots = O((\xi h)^3),$$

was zur Konsistenzordnung $p = 2$ von Lax-Wendroff paßt. ∎

Beweis:[von Satz 4.32] Sei $\hat{\mathbf{u}} \in C_0^\infty(-\pi/h, \pi/h)$. Definiere

$$\mathbf{u}(x) = \int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x}\hat{\mathbf{u}}(\xi)\,d\xi$$

Um die Konsistenz auszunutzen, schreiben wir einen Schritt des Verfahrens:

$$\widehat{\mathbf{E}\mathbf{u}}(\xi) = \widehat{\mathbf{E}}(\xi)\hat{\mathbf{u}}(\xi)$$
$$\Longrightarrow \mathbf{u}^1(x_j) = \int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x}\widehat{\mathbf{E}}(\xi)\hat{\mathbf{u}}(\xi)\,d\xi$$

Die exakte Evolution ist

$$\mathbf{u}(x,k) = \int_{-\pi/h}^{\pi/h} e^{\widehat{P}(\mathbf{i}\xi)k}\hat{\mathbf{u}}(\xi)e^{\mathbf{i}\xi x}\,d\xi$$

Damit ist der Fehler

$$\mathbf{u}(x_j,k) - \mathbf{u}^1(x_j) = \int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x}(e^{\widehat{P}(\mathbf{i}\xi)k} - \widehat{\mathbf{E}}(\xi))\hat{\mathbf{u}}(\xi)\,d\xi$$

Für $x_j = 0$ folgt damit wegen der Konsistenzordnung p:

$$\left|\int_{-\pi/h}^{\pi/h}(e^{\widehat{P}(\mathbf{i}\xi)} - \widehat{\mathbf{E}}(\xi))\hat{\mathbf{u}}(\xi)\,d\xi\right| \le Ch^{p+1}\|u^{(p+1)}\|_{L^\infty(\mathbb{R})}$$

Läßt man $\hat{\mathbf{u}} \to \delta_{\xi'}$ laufen für ein festes $\xi' \in (-\pi/h, \pi/h)$ und nutzt, daß dann $\|u^{(p+1)}\|_{L^\infty(\mathbb{R})} \le C|\xi'|^{p+1}$, so ergibt sich die Behauptung. □

Satz 4.32 ist interessant, weil er zeigt, daß aus der Stabilität des kontinuierlichen Problems (Kontrolle von $e^{\widehat{P}(\mathbf{i}\xi)}$) und Konsistenz des Verfahrens man Kontrolle von $\widehat{\mathbf{E}}(\xi)$ für "kleine" ξ erhalten kann. Für große ξ muß dies aus dem Verfahren herrühren, z.B. durch Dissipativität.

"Herleitung" von Lax-Wendroff

Wie wir es bei parabolischen Gleichungen gemacht haben, könnte man zahlreiche Verfahren erzeugen, indem man zuerst eine Semidiskretisierung im Ort durchführt und dann ein Zeitschrittverfahren darauf anwendet. Das Lax-Wendroff-Verfahren (für $u_t + au_x = 0$) wird anders "hergeleitet": Mit den Differenzenoperatoren D_+, D_-, D_0 (z.B. $D_+u(x) = u(x+h) - u(x)$, $D_0u(x) = u(x+h) - u(x-h)$) ergibt Taylor

$$u(t+k,x) = u + ku_t + \frac{k^2}{2}u_{tt} + O(k^2)$$

Die Gleichung $u_t + au_x = 0$ liefert

$$u_t = -au_x, \qquad u_{tt} = a^2u_{xx}. \tag{4.21}$$

So können Zeitableitungen durch Ortsableitungen ersetzt werden und anschließend durch Differenzenquotienten approximiert werden:

$$u(t+k,x) = u - \frac{k}{2h}aD_0u + \frac{k^2}{2h^2}a^2D_+D_-u + kO(k^2+h^2)$$

Das zeigt, daß die Konsistenzordnung des Verfahrens

$$u_i^{n+1} = u_i^n - \frac{\lambda}{2}\left(u_{i+1}^n - u_{i-1}^n\right) + \frac{\lambda^2}{2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^m\right) = 0, \qquad \lambda = \frac{ka}{h}$$

gerade 2 (genauer: $(2,2)$) ist.

Modifizierte Gleichung

Die Fourieranalyse (d.h. die von-Neumann-Analyse) ist eine Möglichkeit, das Verhalten von Diskretisierungen zu verstehen. Eine andere Möglichkeit, das qualitative Verhalten von Verfahren zu verstehen ist, sie als "bessere" Approximation an eine andere ("modifizierte") Gleichung zu interpretieren und zu stipulieren, daß das qualitative Verhalten dieser kontinuierlichen Gleichung das numerische Verfahren beschreibt. Wir illustrieren das mit drei klassischen Beispielen:

Beispiel 4.34 (modifizierte Gleichung des upwind-Verfahrens) Wir betrachten $u_t + a u_x = 0$ mit $a < 0$. Das Verfahren ist

$$u_i^{n+1} = u_i^n - a\lambda(u_{i+1}^n - u_i^n), \qquad \lambda = \frac{k}{h}.$$

Taylor um (t, x) liefert

$$\begin{aligned}
u(t+k,x) &= u + ku_t + \frac{k^2}{2}u_{tt} + \frac{k^3}{6}u_{ttt} + \cdots, \\
u(t,x+h) &= u + hu_x + \frac{h^2}{2}u_{xx} + \frac{h^3}{6}u_{xxx} + \cdots, \\
\tau &= \frac{u(t+k,x)-u}{k} + a\frac{u(t,x+h)-u}{h} = \left[u_t + \frac{k}{2}u_{tt} + \frac{k^2}{6}u_{ttt} + \cdots\right] + a\left[u_x + \frac{h}{2}u_{xx} + \frac{h^2}{6}u_{xxx} + \cdots\right] \\
&= u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} + O(k^2+h^2)
\end{aligned}$$

Falls wir $u_t + au_x = 0$ ausnutzen, dann erhalten wir, daß das Upwindverfahren Konsistenzordnung $(1,1)$ hat. Falls wir jedoch annehmen, daß u die Gleichung

$$u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} = 0 \tag{4.22}$$

erfüllt, dann erhalten wir (formal) Konsistenzordnung $(2,2)$. Man nennt (4.22) die modifizierte Gleichung. Typischerweise formuliert man die modifizierte Gleichung nochmals um, indem man Zeitableitungen durch Ortsableitungen ausdrückt: aus (4.22) ergibt sich (formal) durch Differenzieren nach t und x:

$$u_{tt} + au_{xt} = O(k+h), \qquad u_{xt} + au_{xx} = O(k+h), \tag{4.23}$$

daß wir (4.22) schreiben können (unter Vernachlässigung von Termen $O(h^2+k^2)$)

$$\begin{aligned}
0 = u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} &\overset{(4.23)}{=} u_t + au_x + \frac{1}{2}\left[-kau_{xt} + hau_{xx} + kO(k+h)\right] \\
&\overset{(4.23)}{=} u_t + au_x + \frac{1}{2}\left[ka^2u_{xx} + hau_{xx} + kO(k+h)\right] = u_t + au_x + \frac{1}{2}\left[ka^2 + ha\right]u_{xx} + O(k^2+h^2).
\end{aligned}$$

Damit approximiert das obige (!) Upwindverfahren die Gleichung

$$u_t + au_x - \nu u_{xx} = 0, \qquad \nu := -\frac{1}{2}\left[ka^2 + ah\right] \overset{a\le 0}{=} \frac{h}{2k/h}\left[\frac{k}{h}|a| - \frac{k^2}{h^2}a^2\right] \tag{4.24}$$

mit Konsistenzordnung $(2,2)$. Wir bemerken, daß $\nu > 0$, falls die CFL-Bedingung erfüllt ist. Man kann sich also vorstellen, daß das Upwindverfahren zwar die Advektionsgleichung approximiert, aber mit höherer Genauigkeit die Wärmeleitungsgleichung (4.24) mit (kleiner) Diffusion. Da die Wärmeleitungsgleichung Dissipation hat, erwartet man, daß auch das Upwindverfahren dissipativ ist. ∎

Übung 4.35 Zeigen Sie, daß für das Lax-Friedrichs-Schema die modifizierte Gleichung

$$u_t + au_x + \nu u_{xx} = 0, \qquad \nu = \frac{h}{2\lambda}\left(1 - \lambda^2 a^2\right), \qquad \lambda = \frac{k}{h}$$

ist. Insbesondere ist die Diffusionskonstante ν für Lax-Friedrichs größer als für Upwind.

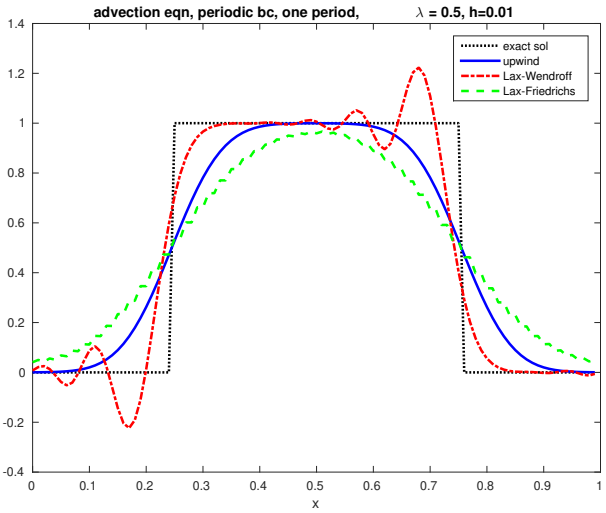


Abbildung 4.7: Upwind, LW, LF für Advektionsgleichung mit unstetigen Anfangsbedingungen

Übung 4.36 Die modifizierte Gleichung für das Lax-Wendroff-Verfahren ist gegeben durch

$$u_t + au_x + \frac{ah^2}{6}\left(1 - \frac{k^2}{h^2}a^2\right)u_{xxx} = 0$$

Beispiel 4.37 Wir betrachten die Advektionsgleichung mit periodischen Randbedingungen:

$$u_t - u_x = 0 \quad \text{auf } (0,1)\times\mathbb{R}^+, \qquad u(0,t) = u(1,t) \quad \forall t > 0. \tag{4.25}$$

Um den Einfluß der Verfahren auf die hochfrequenten Anteil der Anfangsbedingung herauszuarbeiten, ist $u_0(x) = \chi_{[0.25,0.75]}(x)$ (und damit bleibt die exakte Lösung auch eine stückweise konstante Funktion). In Fig. 4.7 wird der Ergebnis verschiedener Verfahren bei $T = 1$ und $k/h = 0.5$ gezeigt. Die modifizierte Gleichung für das Upwindverfahren und Lax-Friedrichs ist eine parabolische Gleichung mit relativ viel Dissipation (Lax-Friedrichs hat sogar noch mehr Dissipation als das Upwindverfahren). Das ist spiegelt sich in Fig. 4.7 wider. Die modifizierte Gleichung von Lax-Wendroff ist eine Gleichung 3. Ordnung, bei der Dispersion eine große Rolle spielt. Das spiegelt sich in den Oszillationen in der numerischen Lösung wider. ∎

Bemerkung 4.38 (Dispersion) Die Lösung u von

$$u_t + au_x = 0, \qquad u(0,x) = u_0(x)$$

ist

$$u(t,x) = u_0(x-at) = \frac{1}{\sqrt{2\pi}}\int_\xi \widehat{u}_0(\xi)e^{\mathbf{i}\xi(x-at)}\,d\xi$$

(diese Darstellung erhält man direkt aus der Fourierinversionsformel[7] oder indem man die Differentialgleichung fouriertransformiert und dann löst). Wir interpretieren dies so: Die Fourierkomponente $\widehat{u}_0(\xi)e^{\mathbf{i}\xi x}$ zur Frequenz ξ breitet sich mit Geschwindigkeit a (nach rechts) aus. Die Geschwindigkeit a ist *unabhängig* von ξ. Sei nun $\xi \mapsto a(\xi)$ eine Funktion von ξ. Dann stellt $v(x,t) = \frac{1}{\sqrt{2\pi}}\int_\xi \widehat{u}_0(\xi)e^{\mathbf{i}\xi(x-a(\xi)t)}\,d\xi$ eine Funktion dar mit folgender Eigenschaft: die Fourierkomponente von $v(\cdot,0)$, die zur Frequenz ξ gehört, wandern mit Geschwindigkeit $a(\xi)$. Allgemein spricht man von *Dispersion*, wenn die Fourierkomponenten sich mit ξ-abhängiger Geschwindigkeit ausbreiten.

Wir lösen die Gleichung

$$u_t + au_x + \nu u_{xxx} = 0, \qquad u(0,x) = u_0(x)$$

Fouriertransformation (in x) liefert die ODE

$$\widehat{u}(t,\xi) + a\mathbf{i}\xi\widehat{u}(t,\xi) + \nu(\mathbf{i}\xi)^3\widehat{u}(t,\xi) = 0, \qquad \widehat{u}(0,\xi) = \widehat{u}_0(\xi)$$

[7] Wir verwenden $\widehat{u}(\xi) = \frac{1}{\sqrt{2\pi}}\int_x e^{-\mathbf{i}x\xi}(x)\,dx$ mit Inversionsformel $u(x) = \frac{1}{\sqrt{2\pi}}\int_\xi e^{\mathbf{i}x\xi}\widehat{u}(\xi)\,d\xi$

mit Lösung

$$\widehat{u}(t,\xi) = \exp(-(a\mathbf{i}\xi + \nu(\mathbf{i}\xi)^3)t)\widehat{u}_0(\xi)$$

Rücktransformation liefert

$$u(t,x) = \frac{1}{\sqrt{2\pi}}\int_\xi e^{\mathbf{i}\xi x} e^{-(a\mathbf{i}\xi+\nu(\mathbf{i}\xi)^3)t}\widehat{u}_0(\xi), d\xi \frac{1}{\sqrt{2\pi}}\int_\xi e^{\mathbf{i}\xi\left[x-(a+\nu(\mathbf{i}\xi)^2)t\right]}\widehat{u}_0(\xi), d\xi.$$

Wir erkennen insbesondere, daß Ausbreitung der unterschiedlichen Fourierkomponenten der Anfangswerte von u_0 unterschiedliche Geschwindigkeit haben. Dies ist beim Lax-Wendroff-Verfahren in Fig. 4.7 erkennbar, denn es "erklärt" die Oszillationen. ∎

4.4 Raum-Zeit-DG

Wir betrachten eine etwas allgemeinere Form von (4.3):

$$u_t + \mathbf{b}(x,t)\cdot\nabla u + c(x,t)u = g. \tag{4.26}$$

Im Unterschied zu parabolischen Gleichungen hat in (4.26) die t-Variable keine besonders herausgehobene Rolle. Nennt man $x_0 = t$, so können wir auch auch folgendes, allgemeineres Problem betrachten:

$$\mathbf{b}(x)\cdot\nabla u + c(x)u = f \quad \text{auf } \Omega, \tag{4.27}$$

wobei wir jetzt $\Omega \subset \mathbb{R}^{d+1}$ zulassen. Bei dieser Gleichung kann nicht eine Randbedingung auf dem gesamten Rand $\partial\Omega$ gefordert werden. Wir definieren den "Einströmrand", den "Ausströmrand" und den "charakteristischen Rand" durch

$$\Gamma^- := \{x\in\partial\Omega\colon \mathbf{b}(x)\cdot n(x) < 0\}, \tag{4.28}$$
$$\Gamma^+ := \{x\in\partial\Omega\colon \mathbf{b}(x)\cdot n(x) > 0\}, \tag{4.29}$$
$$\Gamma^= := \{x\in\partial\Omega\colon \mathbf{b}(x)\cdot n(x) = 0\}. \tag{4.30}$$

Hier ist $n(x)$ der (äußere) Normalenvektor im Punkt $x\in\partial\Omega$. Tatsächlich können wir für (4.27) nur eine Randbedingung auf Γ^- oder Γ^+ fordern. Wir betrachten deshalb das Randwertproblem

$$\mathbf{b}(x)\cdot\nabla u + c(x)u = g \quad \text{auf } \Omega, \tag{4.31a}$$
$$u = 0 \quad \text{auf } \Gamma^-. \tag{4.31b}$$

Beispiel 4.39 Betrachten Sie das Problem

$$-u' + u = g \quad \text{auf } (0,1),$$

Überlegen Sie sich, daß man nicht sinnvollerweise $u(0) = 0 = u(1)$ fordern kann sondern nur $u(0) = 0$ *oder* $u(1) = 0$. ∎

Unser Ziel ist ein numerisches Verfahren für (4.31). Hierzu sei $\mathcal{T}$ ein Gitter, welches Γ^- auflöst, d.h. jede Randkante e erfüllt entweder $e\subset\Gamma^-$ oder $e\subset\partial\Omega\setminus\Gamma^-$. Für jedes Element K bezeichnet wir mit n_K die (äußere) Normale des Elementes K.

Zur Motivation nehmen wir an, daß die Lösung u von (4.31) hinreichend glatt ist. Sei v eine stückweise glatte Testfunktion, d.h. $v|_K$ ist glatt für jedes K. Betrachtet man ein Element K, so folgt aus (4.31a) durch Multiplikation mit v, Integration über K und partieller Integration:

$$\int_K gv = \int_K (c + \mathbf{b}\cdot\nabla u)v = \int_K u(c - u\nabla\cdot(\mathbf{b}v) + \int_{\partial K}(\mathbf{b}\cdot n_K)uv.$$

Durch Summation über alle Elemente $K\in\mathcal{T}$ ergibt sich:

$$\sum_{K\in\mathcal{T}}\int_K u\,(cv - \nabla\cdot(\mathbf{b}v)) + \int_{\partial K}(\mathbf{b}\cdot n_K)uv = \sum_{K\in\mathcal{T}}\int_K gv.$$

Wir führen den Begriff des Flußes auf dem Rand des Elementes K ein:

$$f_K = (\mathbf{b} \cdot n_K)u.$$

Es wird sich als zweckmäßig erweisen, die gemeinsame Kante/Fläche zwischen zwei Elementen K, K' zu bezeichnen; wir verwenden das Symbol

$$K|K',$$

wobei die Reihenfolge gleichzeitig eine "Orientierung" festlegt. Wir werden im Folgenden kurz "Kante" für diese Schnitte sagen, auch wenn es in höheren Dimensionen eigentlich Hyperfläche sind.

Es wird sich auch als zweckmäßig erweisen, die Nachbarelemente zu definieren:

$$\mathcal{N}(K) := \{K' \in \mathcal{T} \mid \ K \text{ und } K' \text{ teilen sich eine Manigfaltigkeit mit Kodimension } 1\}$$

Bei der Diskretisierung wird man u stückweise glatt ansetzen. Damit liegt es nahe, den Raum

$$\mathcal{S}^{p,0} := \{u \in L^2(\Omega) \colon u|_K \in \mathcal{P}_p \qquad \forall K \in \mathcal{T}\}$$

zu gegebenem $p \in \mathbb{N}_0$ zu betrachten. Für eine numerische Realisierung liegt es dann nahe, die Testfunktionen v ebenfalls aus diesem Raum zu wählen. Hierzu ist zu bemerken, daß bei unstetigem Ansatz für u und unstetigen Testfunktionen v *keine* Kopplung zwischen $u|_K$ und $u|_{K'}$ für benachbarte Elemente K und K' existiert. Diese Kopplung realisieren wir nun dadurch, daß auf der Kante $K|K'$, auf der eigentlich zwei Approximationen (nämlich $u|_K$ und $u|_{K'}$) an die exakte Lösung zur Verfügung stehen die Kopplung über einen "numerischen Fluß" $\widehat{h}_{K|K'}$ zu realisieren, d.h. den Fluß f_K wird durch einen "numerischen Fluß" $\widehat{h}_{K|K'}$ ersetzt. Eine sinnvolle Wahl des numerischen Flußes erscheint z.B.

Begriff der Konsistenz

$$\widehat{h}_{K|K'} = (\mathbf{b} \cdot n_K)\widehat{u}_{K|K'}$$

wobei es für die Wahl von $\widehat{u}$ viele Möglichkeiten gibt; plausibel erscheinen z.B.

- $\widehat{u}_{K|K'} = \frac{1}{2}(u|_K + u|_{K'})|_e$
- $\widehat{u}_{K|K'} = u|_K$
- $\widehat{u}_{K|K'} = u|_{K'}$

Für Randkanten $e \subset \Gamma^-$ wird man sinnvollerweise

$$\widehat{u}|_e = 0 \qquad \forall e \subset \Gamma^-$$

wählen.[8] Im vorliegenden Fall ist also der numerische Fluß $\widehat{h}_{K|K'}$ bereits eindeutig durch die Wahl von $\widehat{u}$ auf den Kanten festgelegt:

$$\widehat{h}_{K|K'} = \mathbf{b} \cdot n_K \widehat{u}|_{K|K'}.$$

Es ergibt sich als numerisches Verfahren:

Finde $u \in S^{p,0}(\mathcal{T})$ s.d.

$$B_{DG}^{Trans}(u,v) := \sum_{K\in\mathcal{T}} \int_K u\,(cv + \nabla\cdot(\mathbf{b}v)) + \int_{\partial K} (\mathbf{b}\cdot n_K)\widehat{u}v = l(v) := \int_\Omega vg \qquad \forall v \in S^{p,0}(\mathcal{T}) \quad (4.32)$$

Üblicherweise wird diese Formulierung wieder partiell rückintegriert, und wir erhalten:

Finde $u \in S^{p,0}(\mathcal{T})$ s.d.

$$B_{DG}^{Trans}(u,v) = \sum_{K\in\mathcal{T}} \int_K (cu + \mathbf{b}\cdot\nabla u)v) + \int_{\partial K} (\mathbf{b}\cdot n_K)(\widehat{u} - u)v = l(v) := \int_\Omega vg \qquad \forall v \in S^{p,0}(\mathcal{T}) \quad (4.33)$$

Die Wahl des numerischen Flußes beeinflußt entscheidend die Qualität des numerischen Verfahrens. Wir führen das am folgenden Beispiel vor:

[8] diese Wahl ist zwar naheliegend, aber nicht zwingend—weiterhin ist es zwar naheliegend, $\widehat{u}|_e$ nur abhängig von den Funktionswerten in den angrenzenden Elementen zu machen, aber auch das ist nicht zwingend

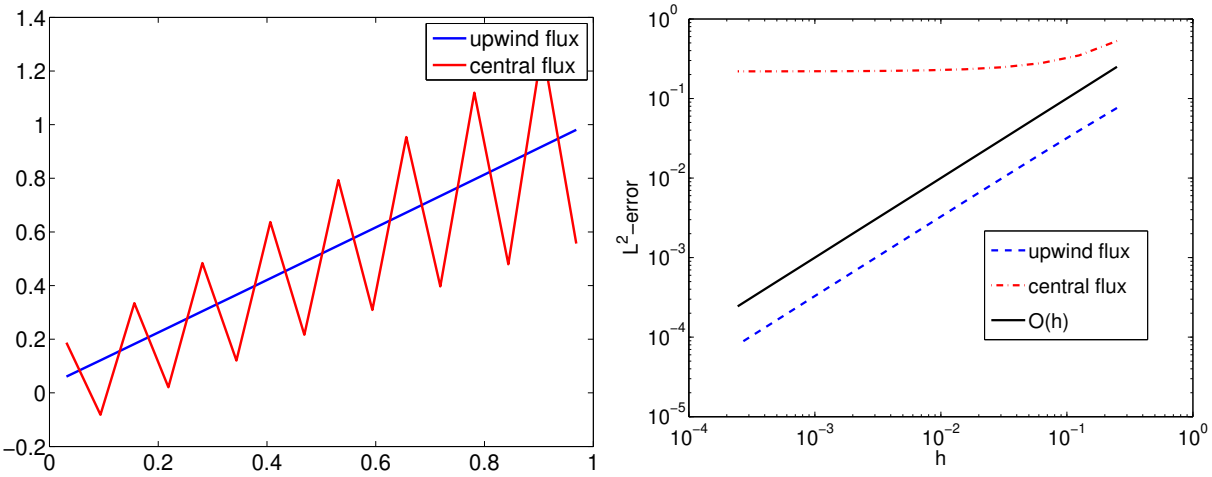


Abbildung 4.8: Links: Lösungsplot. Rechts: Konvergenzverhalten

Beispiel 4.40

$$u' + u = g \quad \text{auf } (0,1), \qquad u(0) = 0.$$

Hier ist $g(x) = 1 + x$, so daß die exakte Lsg $u(x) = x$. Sei $x_i = ih$, $i = 0, \ldots, N$ und $K_i = (x_{i-1}, x_i)$. Für die Wahl $p = 0$ (d.h. $S^{0,0}(\mathcal{T})$ besteht aus den stückweise konstanten Funktionen) schreiben wir für $u|_{K_i} = u_i$ und die Testfunktionen bestehen ebenfalls aus stückweise konstanten Funktionen, für die $v|_{K_i} = v_i$ schreiben:

$$\begin{aligned} B_{DG}^{Trans}(u,v) &= \sum_{i=1}^{N} \int_{K_i} (u' + u)v + (\widehat{u}(x_i)v_i - \widehat{u}(x_{i-1})v(x_i)) \\ &= \sum_{i=1}^{N} \int_{K_i} uv + (\widehat{u}(x_i) - \widehat{u}(x_{i-1}))v_i \\ &= \sum_{i=1}^{N} \left[hu_i + (\widehat{u}(x_i) - \widehat{u}(x_{i-1}))\right] v_i \end{aligned}$$

woraus sich als LGS ergibt:

$$\int_{K_i} g\,dx = hu_i + \widehat{u}(x_i) - \widehat{u}(x_{i-1}), \qquad i = 1, \ldots, N.$$

Wir betrachten 2 Wahlen von $\widehat{u}(x_i)$:

- *upwind flux:* $\widehat{u}(x_j) = u_j$ für $1 \le j \le N$ und $\widehat{u}(x_0) = 0$ (und $\widehat{u}(x_N) = u_N$)
- *central flux:* $\widehat{u}(x_j) = \frac{1}{2}(u_j + u_{j+1})$ für $1 \le j \le N-1$ und $\widehat{u}(x_0) = 0$ und $\widehat{u}(x_N) = u_N$.

Wir erkennen in Fig. 4.8, daß die Wahl des *upwind flux* gute Approximationen und sogar die erhoffte Konvergenz $O(h)$ liefert, während die Wahl des *central flux* zu keiner Konvergenz führt. ∎

Entscheidend für die Wahl des numerischen Flußes $\widehat{u}$ ist, daß man ihn abhängig vom Vorzeichen von $\mathbf{b} \cdot n_K$ in (4.33) wählt. Wie wir unten sehen werden, führt die richtige Wahl des numerischen Flußes auf ein Verfahren mit guten Stabilitätseigenschaften.

Der *upwind flux* $(\mathbf{b} \cdot n_K)\widehat{u}$ ist definiert durch folgende Wahl von $\widehat{u}$ auf jeder Kante:

- Sei e eine *innere* Kante von $\mathcal{T}$, welche sich K und K' teilen. Für $x \in e$ definieren wir:

$$\begin{aligned} \widehat{u}(x) &= \text{egal} && \text{falls } \mathbf{b}(x) \cdot n_K(x) = \mathbf{b}(x) \cdot n_{K'}(x) = 0 \\ \widehat{u}(x) &= u|_K(x) && \text{falls } \mathbf{b}(x) \cdot n_K(x) > 0 \\ \widehat{u}(x) &= u|_{K'}(x) && \text{falls } \mathbf{b}(x) \cdot n_{K'}(x) > 0. \end{aligned}$$

- Sei e ein Kante auf Γ^-: dann definieren wir $\widehat{u}|_e = 0$
- Sei e ein Kante auf $\partial\Omega \setminus \Gamma^-$: dann definieren wir $\widehat{u}|_e$ als Limes von u vom angrenzenden Element

Dieses Verfahren hat gut Stabilitätseigenschaften, wie wir nun zeigen. Hierzu benötigen wir den *Sprung* $[\![u]\!]$ auf einer Kanten $e = K|K'$:

$$\begin{aligned} [\![u]\!]|_e &= u|_K n_K + u|_K n_{K'} \qquad \text{falls } e = K|K' \text{ eine innere Kante ist} \\ [\![u]\!]|_e &= u|_K n_K \qquad \text{falls } e \text{ eine Randkante ist mit } e \subset \partial K. \end{aligned}$$

Satz 4.41 *Es gelte*

$$c - \frac{1}{2}\nabla \cdot \mathbf{b} \geq c_0 > 0 \qquad \text{auf } \overline{\Omega}.$$

Mit dem Sprung $[\![\cdot]\!]$ gilt für die Wahl des "upwind flux" und stückweise glatte Funktionen u:

$$B_{DG}^{Trans}(u,u) \geq \sum_{K\in\mathcal{T}} \|\sqrt{c_0}u\|^2_{L^2(K)} + \sum_{e\in\mathcal{E}} \frac{1}{2}\||\mathbf{b}\cdot n_K|^{1/2}[\![u]\!]\|^2_{L^2(e)},$$

wobei $\mathcal{E}$ die Menge aller Kanten von $\mathcal{T}$ ist. Für Randkanten ist der Sprung einfach definiert als der Wert der Spur.

Beweis: Für glatte Funktionen u ist $u\nabla u = \nabla(\frac{1}{2}u^2)$. Damit gilt für jedes Element $K \in \mathcal{T}$

$$\int_K u(c + \mathbf{b}\cdot\nabla u) = \int_K u^2 \left(c - \frac{1}{2}\nabla\cdot\mathbf{b}\right) + \int_{\partial K} \frac{1}{2}\mathbf{b}\cdot n_K u^2$$

Damit

$$B_{DG}^{Trans}(u,u) = \sum_{K\in\mathcal{T}} \int_K u^2 \underbrace{\left(c - \frac{1}{2}\nabla\cdot\mathbf{b}\right)}_{\geq c_0} + \int_{\partial K} \mathbf{b}\cdot n_K \left[\widehat{u}u - \frac{1}{2}u^2\right].$$

Die Summe $\sum_{K\in\mathcal{T}} \int_{\partial K}$ schreiben wir als Summe über Kanten. Dabei:

- Sei e eine *innere* Kante, die sich Elemente K und K' teilen. Sei $x \in e$. Sei oBdA K das Element mit $\mathbf{b}(x)\cdot n_K(x) > 0$ (der Fall $b(x)\cdot n_K(x) = 0$ ist nicht interessant). Dann rechnen wir wegen $n_K = -n_{K'}$ und der Wahl von $\widehat{u}(x)$:

$$\begin{aligned} & \mathbf{b}(x)\cdot n_K(x)\left[\widehat{u}(x)u_K(x) - \frac{1}{2}u_K(x)^2\right] + \mathbf{b}(x)\cdot n_{K'}(x)\left[\widehat{u}(x)u_{K'}(x) - \frac{1}{2}u_{K'}(x)^2\right] \\ = \;& \mathbf{b}(x)\cdot n_K(x)\left[u_K(x)^2 - \frac{1}{2}u_K(x)^2 - u_K(x)u_{K'}(x) + \frac{1}{2}u_{K'}(x)^2\right] \\ = \;& \mathbf{b}(x)\cdot n_K(x)\frac{1}{2}|u_K(x) - u_{K'}(x)|^2. \end{aligned}$$

  Diese Rechnung ist auch für den Fall $\mathbf{b}(x)\cdot n_K(x) = 0$ richtig.

- Falls e eine Randkante mit $e \subset \Gamma^-$ ist, dann ist $\widehat{u} = 0$ auf e und somit

$$b\cdot n_K\left[\widehat{u} - \frac{1}{2}u\right]u = -\mathbf{b}\cdot n_K\frac{1}{2}u^2 = \frac{1}{2}|\mathbf{b}\cdot n_K|u^2 = \frac{1}{2}|\mathbf{b}\cdot n_K||[\![u]\!]|^2,$$

  wobei wir im letzten Schritt ausgenutzt haben, wie der Sprung auf Randkanten definiert ist.

- Falls e eine Randkante mit $e \subset \partial\Omega \setminus \Gamma^-$ ist, dann ist $\mathbf{b}\cdot n_K \geq 0$ und $\widehat{u} = u$. Somit

$$b\cdot n_K\left[\widehat{u} - \frac{1}{2}u\right]u = \frac{1}{2}b\cdot n_K u^2 = \frac{1}{2}b\cdot n_K|[\![u]\!]|^2,$$

  wobei wir wieder ausgenutzt haben, wie der Sprung auf Randkanten definiert ist.

Faßt man alle Kantenbeiträge zusammen, dann hat man

$$\sum_{K\in\mathcal{T}}\int_{\partial K}\mathbf{b}\cdot n_K\left(\widehat{u}-\frac{1}{2}u\right)u=\frac{1}{2}\sum_{e\in\mathcal{T}}\||\mathbf{b}\cdot n_K|^{1/2}[\![u]\!]\|^2_{L^2(e)},$$

wobei wir leicht schlampig mit n_K einen Normalenvektor auf e bezeichnen. □

Satz 4.41 zeigt, daß die Bilinearform B koerziv ist. Damit ist insbesondere eindeutige Lösbarkeit des diskreten Verfahrens gegeben.

Die Herleitung der Variationsformulierung zeigt außerdem, daß das Verfahren konsistent ist im folgenden Sinn: Falls u eine Lösung von (**??**) ist und zudem die Regularitätsbedingung $u\in H^1(\Omega)$ gilt, dann ist

$$B_{DG}^{Trans}(u,v)=l(v)\qquad\forall v\in S^{p,0}(\mathcal{T}). \tag{4.34}$$

Damit ergibt sich die Galerkinorthogonalität

$$B_{DG}^{Trans}(u-u_N,v)=0\qquad\forall v\in S^{p,0}(\mathcal{T}). \tag{4.35}$$

4.5 DG und Finite Volumenmethoden im Ort—RK in der Zeit

Das numerische Verfahren in Abschnitt 4.4 hat ausgenutzt, daß die Zeitvariable eigentlich keine herausgehobene Rolle hat und deshalb eine Diskretisierung im Raum-Zeit-Zylinder durchgeführt. Wie bei parabolischen Verfahren sind jedoch in der Praxis Zeitschrittverfahren vorherrschend. Wir werden mit k den Zeitschritt bezeichnen.

Die Behandlung von Randbedingungen ist ein eigenes Thema bei hyperbolischen Problemen. Wir behandeln hier den einfachsten Fall eines reinen Anfangswertproblems, d.h. $\Omega=\mathbb{R}^d$. Die Anfangsbedinung u_0 wird mit kompaktem Träger vorausgesetzt.

$$u_t+\nabla\cdot\mathbf{f}(u)=0\quad\text{ auf }\Omega\times\mathbb{R}^+,\qquad u(\cdot,0)=u_0 \tag{4.36}$$

Wir gehen von einer Triangulierung $\mathcal{T}$ von Ω aus. Für Testfunktionen $v\in S^{p,0}(\mathcal{T})$ ergibt sich nach partieller Integration (und Vertauschen von Integration und $\frac{d}{dt}$)

$$\sum_K\frac{d}{dt}\int_K uv\,dt-\int_K\nabla v\cdot\mathbf{f}(u)+\int_{\partial K}n_K\cdot\mathbf{f}(u)v=0.$$

Weil die Testfunktionen unstetig sind, muß ein Kopplung über benachbarte Elemente erfolgen. Dies erfolgt mit dem zu wählenden *numerischen Fluß* $\widehat{h}=\widehat{h}(u,v,n)$. Bezeichnet man mit $\mathcal{N}(K)$ die Nachbarelemente von K, so ergibt sich als numerisches Verfahren: Finde $u\in S^{p,0}(\mathcal{T})$, so daß

$$\sum_K\frac{d}{dt}\int_K uv-\int_K\nabla v\cdot\mathbf{f}(u)+\sum_{K'\in\mathcal{N}(K)}\int_{K|K'}\widehat{h}(u_K,u_{K'},n_K)v=0.$$

Hier ist wie oben $K|K'$ eine Kurznotation für den Schnitt $\overline{K}\cap\overline{K'}$. Wir nennen ihn Kante, was aus dem 2D-Fall im Ort motiviert ist.

Definition 4.42 (numerischer Fluß) *Sei* $\mathbb{S}^{d-1}=\{x\in\mathbb{R}^d\,|\,\|x\|_2=1\}$. *Eine Funktion* $\widehat{h}:\mathbb{R}\times\mathbb{R}\times\mathbb{S}^{d-1}\to\mathbb{R}$ *heißt numerischer Fluß, wenn sie lokal lipschitz-stetig stetig ist. Der numerische Fluß heißt*

- konsistent, *wenn* $\widehat{h}(u,u,\mathbf{n})=\mathbf{f}(u)\cdot\mathbf{n}$ *für alle* u.
- konservativ, *falls* $\widehat{h}(u,v,\mathbf{n})=-\widehat{h}(v,u,-\mathbf{n})$.
- monoton, *falls* $\widehat{h}$ *monoton wachsend im ersten Argument und monoton fallend im zweiten Argument ist:* $\widehat{h}(\uparrow,\downarrow,\mathbf{n})$

Beispiel 4.43 Der Fall der Advektionsgleichung entspricht $f(u) = \mathbf{b}u$ für konstantes $\mathbf{b}$. Upwinding entspricht der folgenden Wahl des numerisches Flußes auf einer gemeinsamen Kante $K|K'$ von zwei (Orts-)Elementen K, K':

$$\widehat{h}(u_K, u_{K'}, n_K) = \begin{cases} \mathbf{b} \cdot n_K u_K & \text{falls } \mathbf{b} \cdot n_K > 0 \\ \mathbf{b} \cdot n_K u_{K'} & \text{falls } \mathbf{b} \cdot n_K < 0. \end{cases}$$

Man sieht, daß der numerische Fluß konservativ (Übung!) und natürlich auch konsistent ist.

Man kann diese Wahl des Flußes auch ohne Fallunterscheidungen schreiben:

$$\widehat{h}(u_K, u_{K'}, n_K) = \frac{1}{2}\mathbf{b} \cdot n_K \left(u_{K'} + u_K\right) - \frac{1}{2}\left|\mathbf{b} \cdot n_K\right| \left(u_{K'} - u_K\right)$$

■

Bemerkung 4.44 Die Bedingung $\widehat{h}(u, v, n) = -\widehat{h}(v, u, -n)$ drückt eine Erhaltungseigenschaft aus: Für die Testfunktion $v \equiv 1$ ergibt sich:

$$\frac{d}{dt}\int_\Omega u = \sum_K \frac{d}{dt}\int_K u = -\sum_K \sum_{K' \in \mathcal{N}(K)} \int_{K|K'} \widehat{h}(u_K, u_{K'}, n_K).$$

Schreibt man dies als Summe über alle Kanten, so ergibt sich mit der Beobachtung, daß jede Kante e von 2 Elementen K_e, K'_e geteilt wird:

$$\begin{aligned} \frac{d}{dt}\int_\Omega u &= -\sum_K \int_{K|K'} \widehat{h}(u_K, u_{K'}, n_K) = -\sum_e \int_e \widehat{h}(u_{K_e}, u_{K'_e}, n_{K_e}) + \int_e \widehat{h}(u_{K'_e}, u_{K_e}, n_{K'_e}) \\ &= -\sum_e \int_e \widehat{h}(u_{K_e}, u_{K'_e}, n_{K_e}) - \widehat{h}(u_{K_e}, u_{K'_e}, n_{K_e}) = 0; \end{aligned}$$

hier haben wir die Eigenschaft der Konservativität ausgenutzt. ■

Ein vollständiges Verfahren ergibt sich nur die Wahl eines Zeitintegrators. Im einfachsten Fall wird man ein explizites Eulerverfahren wählen. Im Fall des expliziten Eulerverfahrens gilt immer noch die (globale) Erhaltungseigenschaft aus Bemerkung 4.44:

Übung 4.45 Formulieren Sie das explizite Eulerverfahren. Bezeichne mit u^n die numerische Approximation zum Zeitpunkt t_n. Zeigen Sie:

$$\int_\Omega u^{n+1} = \int_\Omega u^n \qquad \forall n \in \mathbb{N}_0.$$

■

strong stability preserving methods

Im Prinzip gibt es viel Auswahl bei den Zeitdiskretisierungen. Meist will man jedoch zusätzliche Eigenschaften erhalten. Bei vielen hyperbolischen Erhaltungsgleichungen gelten gewisse Monotonieeigenschaften auf der kontinuierlichen Ebene, die dann ins Diskrete vererbt werden sollen. Bei skalaren Gleichungen z.B. könnte $\|u(\cdot, t)\|_{L^\infty} \leq \|u_0(\cdot)\|_{L^\infty}$ gelten. Das wesentliche Vorgehen kann man bereits auf dem ODE-Level verstehen.

Betrachte das ODE-System

$$\mathbf{y}' = \mathbf{g}(\mathbf{y}).$$

Wir nehmen an, daß das einfachste RK-Verfahren, das explizite Eulerverfahren, in einer Norm $\|\cdot\|$ die Kontraktivitätsbedingung

$$\|\mathbf{y}^n + k\mathbf{g}(\mathbf{y}^n)\| \leq \|\mathbf{y}^n\|$$

erfüllt. Dann sagen wir, daß ein (anderes) RK-Verfahren die SSP-Eigenschaft[9] hat, falls für es ebenfalls die Eigenschaft $\|\mathbf{y}^{n+1}\| \leq \|\mathbf{y}^n\|$ gilt. Das folgende Lemma 4.47 gibt hinreichende Kriterien an, unter denen ein explizites RK-Verfahren SSP ist. Zuvor eine Umformulierung des klassischen Vorgehens:

[9] strong stability preserving

Übung 4.46 Sei ein s-stufiges explizites RK-Verfahren beschrieben durch das Butcher-Tableau $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ Dann gilt für das RK-Verfahren angewandt auf $y' = g(y)$:

- Das RK-Verfahren wie folgt definiert mit Hilfe der (Zwischen-)Stufen Y_i, $i = 1, \ldots, s$:

$$\begin{aligned} Y_i &= y_0 + k \sum_{j=1}^{i-1} A_{i,j} g(Y_j), \qquad i = 1, \ldots, s \\ y_1 &= y_0 + k \sum_{j=1}^{s} b_j g(Y_j) \end{aligned}$$

- Erweitert man die Matrix A zu einer Matrix in $\mathbb{R}^{(s+1)\times s}$, indem $A_{s+1,:} := b^\top$ gesetzt wird, dann hat das Verfahren die Form

$$\begin{aligned} Y_1 &= y_0 \\ Y_{i+1} &= y_0 + k \sum_{j=1}^{i} A_{i+1,j} g(Y_j), \qquad i = 1, \ldots, s \\ y_1 &= Y_{s+1} \end{aligned}$$

- Das RK-Verfahren kann in der folgenden Form geschrieben werden mit Koeffizienten $\alpha_{i,j} \geq 0$, die zudem die folgende Bedingung erfüllen: $\alpha_{i,j} = 0$ impliziert $\beta_{i,j} = 0$.

$$\begin{aligned} Y_1 &= y_0 \\ Y_{i+1} &= \sum_{j=1}^{i} \alpha_{i,j} Y_j + k \sum_{j=1}^{i} \beta_{i,j} g(Y_j), \qquad i = 1, \ldots, s \\ y_1 &= Y_{s+1} \end{aligned}$$

  (Die Darstellung ist nicht eindeutig). Zeigen Sie, daß die Konsistenz des Verfahrens die Bedingung $\sum_{j=1}^{i} \alpha_{i,j} = 1$ (für jedes i) erwingt.

■

Lemma 4.47 *Sei ein explizites s-stufiges RK-Verfahren gegeben. Möge das explizite Eulerverfahren stabil sein unter der "CFL-Bedingung" $k \leq k_{expEuler}$. Mögen die Koeffizienten $\beta_{i,j}$ in der Darstellung aus Übung 4.46 die Bedingung $\beta_{i,j} \geq 0$ erfüllen. Dann gilt: das RK-Verfahren ist SSP, falls die Schrittweite k die Bedingung*

$$k \leq k_{expEuler} \min_{i,j} \frac{\alpha_{i,j}}{\beta_{i,j}}$$

erfüllt.

Beweis: Die Idee ist, daß das Verfahren eine Konvexkombination aus expliziten Eulerschritten ist. Der Einfachheit nehmen wir an, daß alle $\alpha_{i,j} > 0$. Dann ist

$$Y_{i+1} = \sum_{j=1}^{i} \alpha_{i,j} \left(Y_j + \frac{\beta_{i,j}}{\alpha_{i,j}} k g(Y_j) \right)$$

und damit

$$\|Y_{i+1}\| \leq \sum_{j=1}^{i} \alpha_{i,j} \|Y_j + \frac{\beta_{i,j}}{\alpha_{i,j}} k g(Y_j)\| \leq \sum_{j=1}^{i} \alpha_{i,j} \|Y_j\|,$$

wobei wir im letzten Schritt die Stabilität des expliziten Eulerverfahrens und die stringentere Schrittweitenbedingung verwendet haben. Aus der Konsistenz $\sum_j \alpha_{i,j} = 1$ folgt

$$\|Y_{i+1}\| \leq \max_{j=1,\ldots,i} \|Y_j\|.$$

Induktiv folgt damit $\|Y_{i+1}\| \leq \|Y_1\| = \|y_0\|$ für jedes i und damit $\|y_1\| = \|Y_{s+1}\| \leq \|y_0\|$. □

Übung 4.48

- Zeigen Sie, daß die explizite Trapezregel SSP ist:

0	0	0
1	1	0
	1/2	1/2

- Zeigen Sie, daß die explizite Mittelpunktsregel nicht SSP ist:

0	0	0
1/2	1/2	0
	0	1

■

Bemerkung 4.49 Wenn einige der $\beta_{i,j}$ negativ sind, dann kann man u.U. immer noch stabile Verfahren erzeugen, wenn geeignet das explizite Eulerverfahren durch das implizite Eulerverfahren ersetzt wird, [**?**].

■

Literaturverzeichnis

[1] Vidar Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1997.