Skript zur

Numerik Partieller Differentialgleichungen II

Sommersemester 2008

Johannes Schropp

Universität Konstanz

Fachbereich Mathematik und Statistik

Stand: 4. Dezember 2008

Inhaltsverzeichnis

1. Finite Elemente für elliptische Differentialgleichungen

a) Gebräuchliche Finite Elemente

Vorgelegt sei

$$\begin{aligned} -\Delta u &= g \text{ in } \Omega, \\ u &= \gamma \text{ auf } \partial\Omega. \end{aligned} \tag{1-1}$$

$\Omega \subset \mathbb{R}^2$ sei ein beschränktes Gebiet mit stückweise glattem Rand und strikter Kegeleigenschaft. Die Funktionen g und γ mögen von Ω nach $\mathbb{R}$ abbilden.

Die zu (1-1) gehörige schwache Formulierung lautet

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega g \cdot v \,\mathrm{d}x \quad \forall v \in V := H_0^1(\Omega) \tag{1-2}$$

bzw. mit

$$\begin{aligned} a : V \times V &\to \mathbb{R}, \\ a(u,v) &= \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \\ b : V &\to \mathbb{R}, \\ b(v) &= \int_\Omega g \cdot v \,\mathrm{d}x \end{aligned}$$

finden wir die sogenannte „Variationsgleichung"

$$a(u,v) = b(v) \quad \forall v \in V = H_0^1(\Omega). \tag{1-3}$$

1.1 Definition. $u \in V$ heißt eine „schwache Lösung" von (1-1), wenn u (1-3) erfüllt.

1.2 Bemerkung. Die Variationsgleichung (1-3) hat die gleichen Lösungen $u \in V$ wie die Aufgabe

$$F(v) = \frac{1}{2}a(v,v) - b(v) = \int_\Omega \frac{1}{2}|\nabla v|^2 - g \cdot v \,\mathrm{d}x \to \min_{v \in V},$$

wobei $F : V \to \mathbb{R}$. Letztere Minimierungsaufgabe heißt „Variationsproblem".

Man sucht also eine Funktion u mit $u - \gamma \in V = H_0^1(\Omega)$. Dazu transformiert man die Aufgabe (1-1) auf homogene Randbedingungen. Ist $w = u - \gamma$ eine Lösung von

$$-\Delta w = g + \Delta\gamma \text{ in } \Omega, w = 0 \text{ auf } \partial\Omega,$$

so folgt

$$\begin{aligned} -\Delta u &= -\Delta(w+\gamma) = -\Delta w - \Delta\gamma = g, \\ u|_{\partial\Omega} &= w|_{\partial\Omega} + \gamma|_{\partial\Omega} = \gamma|_{\partial\Omega}. \end{aligned}$$

Ohne Einschränkung betrachten wir im Folgenden die Aufgabe (1-1) mit $\gamma = 0$.

Das Galerkin-Verfahren zur Aufgabe (1-1) lautet dann: Sei $V_h \subset V$ ein endlich-dimensionaler Teilraum von V. Gesucht ist ein $u_h \in V_h$ mit

$$a(u_h, v) = b(v) \quad \forall v \in V_h.$$

Die Finite Elemente Methode ist dann ein Galerkin-Verfahren für einen Ansatzraum mit speziellen Eigenschaften.

Bei Ansätzen mit $V_h \subset V$ spricht man von konformen Finiten Elementen.

Sei nun $V_h \subset V$ mit $V_h = \operatorname{span}\{u_1, \ldots, u_m\}$ für gewisse Funktionen $u_i \in V$, $i = 1, \ldots, m$. Dann genügt es

$$a(u_h, v) = b(v)$$

für $v = u_i$, $i = 1, \ldots, m$, d.h. auf einer Basis von V_h, zu fordern.

Für unsere Gleichung (1-1) folgt mit

$$u_h = \sum_{i=1}^{m} c_i u_i$$

sowie der Definition von a und b sofort

$$\int_\Omega \nabla \left(\sum_{j=1}^{m} c_j \cdot u_j \right) \cdot \nabla u_i - g \cdot u_i \, \mathrm{d}(x,y) = 0, \quad i = 1, \ldots, m$$

bzw.

$$\int_\Omega \sum_{j=1}^{m} c_j \nabla u_j \cdot \nabla u_i - g u_i \, \mathrm{d}(x,y).$$

Wir erstellen also das lineare Gleichungssystem

$$Ac = r, \quad A \in \mathbb{R}^{m,m}, \quad c, r \in \mathbb{R}^m$$

mit

$$\begin{aligned} A_{ij} &= \int_\Omega \nabla u_j \cdot \nabla u_i \, \mathrm{d}(x,y), \quad 1 \le i, j \le m, \\ r_i &= \int_\Omega g \cdot u_i \, \mathrm{d}(x,y), \quad 1 \le i \le m. \end{aligned}$$

1.3 Definition. A heißt „Steifigkeitsmatrix“ und r heißt „Ladevektor“.

Wahl der Ansatzfunktionen bei Finiten Elementen

Man unterteilt das Gebiet Ω in sogenannte Finite Elemente durch eine Triangulierung. Unsere vereinfachende Annahme sei dabei, dass Ω polygonal berandet ist, d.h. der Rand $\partial\Omega$ bestehe aus endlich vielen Geradestücken.

1.4 Definition. Eine Zerlegung $\Omega_{T_h} = \{e_1, \ldots, e_M\}$ von Ω in Dreieckelemente heißt „zulässige Triangulierung“, falls Folgendes gilt:

i) $\overline{\Omega} = \bigcup_{i=1}^{M} e_i$,

ii) Besteht $e_i \cap e_j$ aus genau einem Punkt, so ist dieser Eckpunkt sowohl von e_i als auch von e_j.

iii) Besteht $e_i \cap e_j$ für $i \neq j$ aus mehr als einem Punkt, so ist $e_i \cap e_j$ eine Kante sowohl von e_i als auch von e_j.

h ist dabei die maximale auftretende Kantenlänge.

1.5 Beispiel. Auf der nachstehenden Abbildung wird ein Beispiel einer Triangulierung gegeben.

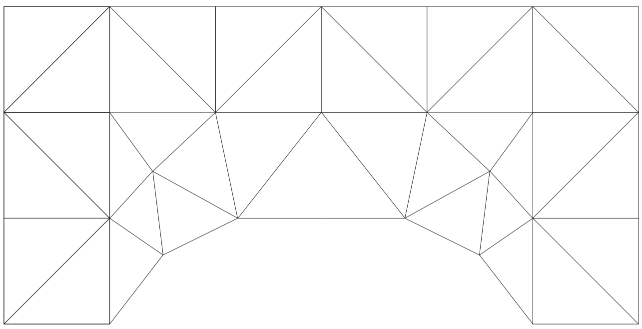


Abbildung 1: Triangulierung eines Gebietes $\Omega \subset \mathbb{R}^2$

Für die Lösung elliptischer Probleme zweiter Ordnung wählt man im Allgemeinen Finite Elemente in $H^1(\Omega)$.

1.6 Satz. *Sei $k \geq 1$ und sei $\Omega \subset \mathbb{R}^2$ beschränktes Gebiet mit zulässiger Triangulierung $\Omega_{T_h} = \{e_1, \ldots, e_M\}$. Eine Funktion $v : \overline{\Omega} \to \mathbb{R}$ mit $v|_{e_i} \in C^\infty(e_i)$, $i = 1, \ldots, M$ gehört genau dann zu $H^k(\Omega)$, wenn $v \in C^{k-1}(\overline{\Omega})$ gilt.*

Beweis: Es genügt den Fall $k = 1$ zu zeigen. Für $k > 1$ folgt die Aussage sofort aus der rekursiven Anwendung auf die partiellen Ableitungen der Ordnung $k - 1$.

„⇒“ Sei $v \in C(\overline{\Omega})$. Für $i = 1, 2$ setzen wir $w, z : \Omega \to \mathbb{R}$, $w(x,y) = \frac{\partial}{\partial x} v(x,y)$, $z(x,y) = \frac{\partial}{\partial y} v(x,y)$, wobei auf den Kanten einer der beiden Grenzwerte gewählt werden kann. Ferner sei $\varphi \in C_0^\infty(\Omega)$ beliebig.

Mit der Greenschen Formel folgt

$$\begin{aligned}
\int_\Omega \varphi w \,\mathrm{d}(x,y) &= \sum_{j=1}^{M} \int_{e_j} \varphi \frac{\partial v}{\partial x} \,\mathrm{d}(x,y) \\
&= \sum_{j=1}^{M} \left\{ - \int_{e_j} \frac{\partial \varphi}{\partial x} v \,\mathrm{d}(x,y) + \int_{\partial e_j} \varphi v n_x \,\mathrm{d}S \right\},
\end{aligned}$$

wobei $n = (n_x, n_y)$.

Da v stetig ist, heben sich die Integrale über die inneren Kanten gegenseitig auf. Außerdem verschwindet φ auf $\partial\Omega$. Dies liefert

$$\begin{aligned}
\int_\Omega \varphi w \,\mathrm{d}(x,y) &= -\sum_{e_j}^{M} \int_{e_j} \frac{\partial \varphi}{\partial x} v \,\mathrm{d}(x,y) \\
&= -\int_\Omega \frac{\partial \varphi}{\partial x} v \,\mathrm{d}(x,y).
\end{aligned}$$

Analog folgt

$$\int_\Omega \varphi z \,\mathrm{d}(x,y) = -\int_\Omega \frac{\partial \varphi}{\partial y} v \,\mathrm{d}(x,y), \quad \varphi \in C_0^\infty.$$

Dies stellt aber zusammengefasst die Definition der schwachen Differenzierbarkeit dar.

„⇐“ Sei jetzt $v \in H^1(\Omega)$. Betrachte v in der Umgebung einer Kante und drehe die Kante so um, dass sie auf der y-Achse liegt. Sie umfasse speziell das Intervall $[y_1 - \delta, y_2 + \delta]$ mit $y_1 < y_2$ und $\delta > 0$. Setze

$$\psi(x) = \int_{y_1}^{y_2} v(x,y) \,\mathrm{d}x.$$

Überdies sei nun $v \in C^\infty(\Omega)$ angenommen. Dann folgt mit Cauchy-Schwarz

$$\begin{aligned}
|\psi(x_1) - \psi(x_2)|^2 &= \left| \int_{x_1}^{x_2} \int_{y_1}^{y_2} \frac{\partial v}{\partial x} \,\mathrm{d}(x,y) \right|^2 \\
&\le \left| \int_{x_1}^{x_2} \int_{y_1}^{y_2} \,\mathrm{d}(x,y) \right| \cdot \|v\|_{H^1}^2 \\
&\le |x_1 - x_2| \cdot |y_1 - y_2| \cdot \|v\|_{H^1(\Omega)}^2.
\end{aligned}$$

Wegen der Dichtheit von $C^\infty(\Omega)$ in $H^1(\Omega)$ gilt diese Aussage auch für $v \in H^1(\Omega)$. Also ist die Funktion $x \mapsto \psi(x)$ stetig bei Null.

Da y_1 und y_2 beliebig gewählt sind und der Bedingung $y_1 < y_2$ genügen, muss die stückweise stetige Funktion v auch auf der Kante stetig sein.

□

1.7 Bemerkung. Gilt für den Ansatzraum V_h die Inklusion $V_h \subset C^k(\Omega)$, $k = 0, 1, 2$, so spricht man von C^k-Finiten Elementen oder kurz C^k-Elementen.

Es sei e_μ das μ-te Element der Zerlegung $\Omega_{T_h} = \{e_1, \ldots, e_M\}$ des Grundgebietes Ω mit den Ecken $p^i = (x_i, y_i)$, $p^j = (x_j, y_j)$, $p^k = (x_k, y_k)$.

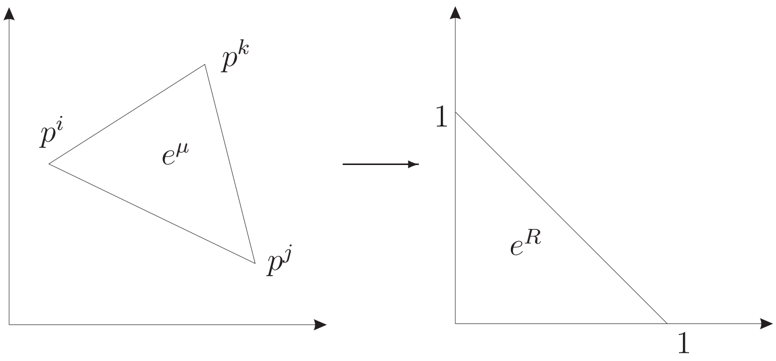


Abbildung 2: Transformation auf ein Referenzdreieck $e^R = \{(\xi, \eta) | 0 \le \xi, \eta, \xi + \eta \le 1\}$

Betrachte die Transformation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} x_j - x_i & x_k - x_i \\ y_j - y_i & y_k - y_i \end{pmatrix} \cdot \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} R_1(\xi, \eta) \\ R_2(\xi, \eta) \end{pmatrix} = R(\xi, \eta). \quad \text{(1-4)}$$

Die Abbildung R ist affin linear und invertierbar. Offensichtlich gilt $R(0,0) = p^i$, $R(1,0) = p^j$, $R(0,1) = p^k$. Diese Transformation bildet das Referenzdreieck e^R auf das Dreieck e_μ ab.

Löst man (1-4) nach $\begin{pmatrix} \xi \\ \eta \end{pmatrix}$ auf, so hat man

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = g(x, y) = \begin{pmatrix} g_1(x, y) \\ g_2(x, y) \end{pmatrix}$$

mit affin linearen Funktionen g_1, g_2 in x und y.

Isoparametrisches Prinzip

Ist $p(\xi,\eta)$ eine Basisfunktion auf e^R, so ist $p(g_1(x,y),g_2(x,y)) = p(g(x,y))$ eine Basisfunktion auf e_μ.

Unser Ziel ist es, auf jedem Dreieck e_i, $i \in \{1,\dots,M\}$ geeignete glatte Funktionen so vorzugeben, dass global auf $\Omega_{T_h} = \bigcup_{i=1}^M e_i$ eine C^k-Funktion ($k = 0,1,2,\dots$) entsteht. Präziser: man setze die Funktionen auf dem Referenzdreieck e^R mit den Eckpunkten $(0,0)$, $(0,1)$, $(1,0)$ an und benutze das „isoparametrische Prinzip“.

Konstruktion von C^0-Elementen mit Polynomen

Wählen den Ansatzraum

$$V_{T_h} = \{u \in C^0(\overline{\Omega}) \mid \; u|_{e_i} \text{ ist ein Polynom mit } \deg(u|_{e_i}) \le r, i = 1,\dots,M \text{ und } u|_{\partial\Omega} = 0\}.$$

Sei $P_r(\Gamma) = \{u : \Gamma \to \mathbb{R} \mid u \text{ Polynom mit } \deg(u) = r\}$. Dann gilt für $\Gamma \subset \mathbb{R}^2$

$$\dim P_r(\Gamma) = \sum_{i=1}^{r+1} i = \frac{(r+1)(r+2)}{2}.$$

Somit erhalten wir 3, 6 bzw. 10 Freiheitsgrade für lineare, quadratische bzw. kubische C^0-Elemente. Man benötigt also „geeignet gesetzte“ $\frac{1}{2}(r+1)(r+2)$ Knoten im Referenzdreieck bei C^0-Elementen mit Polynomen vom Grade r.

1.8 Bemerkung. Sei u ein Polynom mit $\deg(u) = r$. Wendet man eine affin lineare Transformation an und drückt u in neuen Koordinaten aus, so erhält man wieder ein Polynom vom Grad r. $P_r(\Gamma)$ ist invariant unter R, d.h. $R(P_r(\Gamma)) \subset P_r(R(\Gamma))$.

1.9 Lemma. *Es seien $\widetilde{e}$, $\hat{e}$ benachbarte Dreiecke einer Zerlegung mit einer gemeinsamen Kante K und die Polynome $\widetilde{u}$, $\hat{u}$ mit $\deg(\widetilde{u}) = \deg(\hat{u}) = r$ stimmen auf $(r+1)$-Punkten $p^1,\dots,p^{r+1} \in K$ überein, dann ist die Funktion*

$$u(x) = \begin{cases} \widetilde{u}(x), & x \in \widetilde{e}, \\ \hat{u}(x), & x \in \hat{e}\backslash K \end{cases} \tag{1-5}$$

auf $\widetilde{e} \cup \hat{e}$ stetig.

Beweis: $\widetilde{u}$, $\hat{u}$ sind Polynome vom Grad r in 2 Variablen. Also sind $\widetilde{u}|_K$ und $\hat{u}|_K$ Polynome vom Grad r in einer Variablen. Betrachte $d = \widetilde{u}|_K - \hat{u}|_K$. Es folgt $d \equiv 0$, da $d(p^i) = 0$, $i = 1,\dots,r+1$ gilt und d ein Polynom vom Grade r in einer Variablen auf K ist.

Dies motiviert eine Knotenverteilung wie in der Abbildung 3.

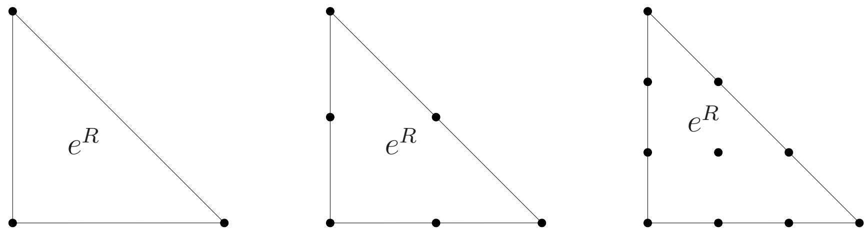


Abbildung 3: Knotierung für lineare, quadratische und kubische C^0-Elemente.

Man konstruiert nun zu diesen Knoten $p^1, \ldots, p^{N_r}$, $N_r = \frac{1}{2}(r+1)(r+2)$, eine Funktionenmenge $f_1, \ldots, f_{N_r}$ mit

$$f_i(p^j) = \delta_{ij}, \quad 1 \le i, j \le N_r \tag{1-6}$$

Man spricht von Lagrange-Elementen, wenn in (1-6) nur Funktionswerte vorgegeben sind. Man spricht von Hermite-Elementen, wenn statt (1-6) Funktions- und Ableitungsvorgaben gemacht werden.

$r = 1$: Es handelt sich hier um lineare Finite Elemente mit $N = 3$ und $e = e^R$ mit Knoten

$$p^1 = (1,0), \quad p^2 = (0,1), \quad p^3 = (0,0)$$

und Funktionen

$$f_1(\xi,\eta) = \xi, f_2(\xi,\eta) = \eta, f_3(\xi,\eta) = 1 - \xi - \eta.$$

Dabei stellen f_1, f_2, f_3 eine nodale Basis auf e^R dar.

$r = 2$: Es ist $N = 6$. Wir sprechen dabei von quadratischen Elementen auf $e = e^R$ mit Knoten

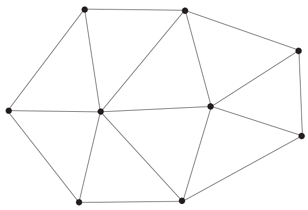
$$\begin{aligned} &p^1 = (1,0), \quad p^2 = (0,1), \quad p^3 = (0,0), \\ &p^4 = (1/2,0), \quad p^5 = (0,1/2), \quad p^6 = (1/2,1/2) \end{aligned}$$

und Funktionen

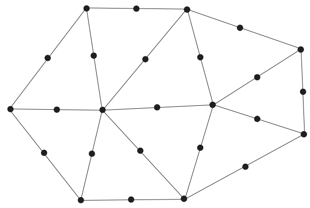
$$\begin{aligned} &f_1(\xi,\eta) = \xi(2\xi - 1), \quad f_2(\xi,\eta) = \eta(2\eta - 1), \\ &f_3(\xi,\eta) = (1-\xi-\eta)(1-2\xi-2\eta), \quad f_4(\xi,\eta) = 4\xi(1-\xi-\eta), \\ &f_5(\xi,\eta) = 4\eta(1-\xi-\eta), \quad f_6(\xi,\eta) = 4\xi\eta. \end{aligned}$$

Dabei gilt $\deg(f_i) = 2$, $i = 1, \ldots, 6$, $f_i(p^j) = \delta_{ij}$, $1 \le i, j \le 6$, d.h. $f_1, \ldots, f_6$ ist eine nodale Basis auf e^R.

($r > 2$ entsprechend): Durch eine geeignete Festlegung der Knotenstellen ist sichergestellt, dass sich die gemäß des isoparametrischen Prinzips konstruierten Funktionen auf e_μ zu einer stetigen Funktion auf $\bigcup_{i=1}^{M} e_i = \Omega$ zusammensetzen lassen. Wie bei dem Referenzdreieck treten für $r = 1$ alle Eckpunkte der Triangulierung und für $r = 2$ alle Eckpunkte und Kantenmitten als Knoten auf.

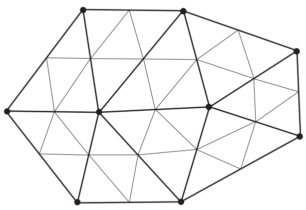


$r = 1$

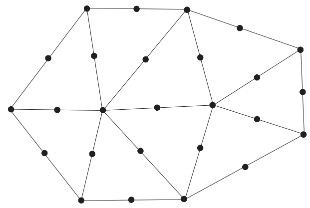


$r = 2$

Ein fairer Vergleich zwischen linearen und quadratischen Finiten Elementen benutzt aus diesem Grund die halbe Gitterweite.

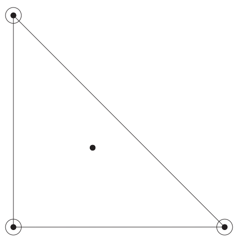


Lineare FE mit $h/2$

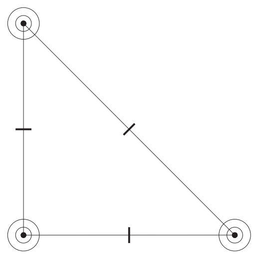


Quadratische FE mit h

Weitere Finite Elemente mit Hermite-Interpolation lassen sich wie folgt definieren.



(a) Kubisches C^0-Element ($N = 10$)



(b) Quintisches C^1-Element (Agyris-Element) ($N = 21$)

Abbildung 4: Elemente mit Hermite-Interpolation

Dabei haben wir die folgende Bezeichnung verwendet:

$\odot$ – Vorgabe der Funktionswerte und des Gradienten (d.h. 3 Vorgaben pro Punkt)

$\bullet$ – Vorgabe des Funktionswertes

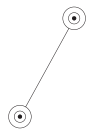
$\circledcirc$ – Vorgabe des Funktionswertes, des Gradienten und der Hesse-Matrix (6 Vorgaben pro Punkt)

$|$ – Vorgabe der Normalenableitung

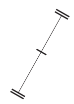
□

1.10 Bemerkung (Agyris-Element). Betrachtet man die Restriktion von $q \in P_5(e)$ auf eine Kante K_e von e, so sind in den Ecken die Werte bis zur zweiten Ableitung vorgegeben ($2 \times 3 = 6$ Vorgaben).

Somit ist die Hermitsche Interpolationsaufgabe in $P_5(K_e)$ eindeutig lösbar. Der Ansatz $h(t) = \sum_{i=0}^{5} \alpha_i t^i$ hat somit 6 Freiheitsgrade. Also hat man die Stetigkeit der Funktion auf K_e und die der tangentialen Ableitung. Ferner ist die Normalableitung ein Polynom vierten Grades.



Restriktion auf die Kante



Restriktion auf die Ecken
(5 Vorgaben)

Diese ist an den Ecken mitsamt der ersten Ableitungen und in den Seitenmitten gegeben. Da das zugeordnete eindimensionale Interpolationsproblem mit 5 Freiheitsgraden korrekt gestellt ist, hat man auch die Stetigkeit der Normalableitung.

Wir skizzieren nun kurz den weiteren Weg bei C^0-Lagrange-Elementen.

Es sei $\Omega_h = \{e_1, \ldots, e_M\}$. Nach Konstruktion auf dem Referenzelement erhält man Knotenpunkte $p^1, \ldots, p^{M_f} \in \Omega_{T_h}$ und Funktionen $u_i : \Omega_{T_h} \to \mathbb{R}$ mit

- $u_i(p^j) = \delta_{ij}$, $1 \leq i, j \leq M_f$,
- $u \in C(\Omega_{T_h})$,
- $u|_{e^j} \in P_r(e^j)$, $j = 1, \ldots, M$

mit zugehörigem Ansatzraum $V_{T_h} = \operatorname{span}\{u_1, \ldots, u_{M_f}\}$. Die Funktionenmenge $\{u_i\}_{i=1,\ldots,M_f}$ heißt „nodale Basis“.

1.11 Bemerkung. Gemäß Satz 1.6 mit $v \equiv u_l$ folgt sofort $u_l \in H^1(\Omega)$ für $l \in \{1, \ldots, M_f\}$, d.h. $V_{T_h} \subset V = H_0^1(\Omega)$.

In unserem Fall erhält man das Gleichungssystem

$$Ac = r, \quad A \in \mathbb{R}^{M_f, M_f}, \quad c, r \in \mathbb{R}^{M_f}$$

mit

$$A_{ij} = \int_\Omega \nabla u_j \cdot \nabla u_i \,\mathrm{d}(x, y), \quad 1 \leq i, j \leq M_f,$$

$$r_i \;=\; \int_\Omega g\cdot u_i\,\mathrm{d}(x,y), \quad 1\le i\le M_f.$$

Mit $g_{T_h}=\sum_{j=1}^{M_f} g(p^j)u_j$ und $\Omega_{T_h}=\bigcup_{l=1}^{M} e_l$ bleiben damit die Integrale

$$\int_e \nabla u_i\cdot\nabla u_j\,\mathrm{d}(x,y), \quad \int_e u_i\cdot u_j\,\mathrm{d}(x,y), \quad 1\le i,j\le M_f \tag{1-7}$$

auf einem Dreieck e zu bestimmen.

1.12 Beispiel ($r=2,\ N_2=6$). Es sind dann die Integrale

$$S_{jk}=\int_e \nabla u_{ij}\cdot\nabla u_{ik}\,\mathrm{d}(x,y), \quad 1\le j,k\le N_r, \quad N_r=\frac{1}{2}(r+1)(r+2)$$

von Null verschieden.

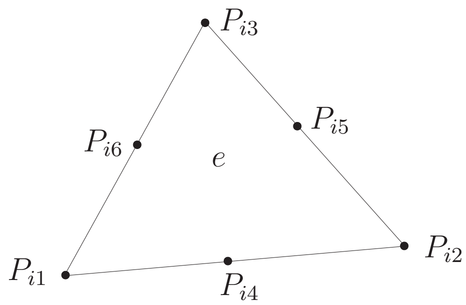


Abbildung 5: Knotennummerierung im Dreieck e

Die symmetrische Matrix $S^l=(S_{jk})_{1\le i,k\le N_r}$ ist beim Aufbau der Gesamtmatrix A auf die Untermatrix

$$\begin{pmatrix} A_{i_1,i_1} & A_{i_1,i_2} & \dots & A_{i_1,i_{N_r}} \\ A_{i_2,i_1} & A_{i_2,i_2} & \dots & A_{i_2,i_{N_r}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{i_{N_r},i_1} & A_{i_{N_r},i_2} & \dots & A_{i_{N_r},i_{N_r}} \end{pmatrix}$$

aufzuaddieren.

Bei der Berechnung des zweiten Integrals in (1-7) tritt an die Stelle von $S^l\in\mathbb{R}^{N_r,N_r}$ die Matrix $M^l\in\mathbb{R}^{N_r,N_r}$, wobei

$$(M^l)_{jk}=\int_e u_{ij}\cdot u_{ik}\,\mathrm{d}(x,y), \quad 1\le j,k\le N_r.$$

Erweiterung des Grundkonzepts auf Gebiete Ω mit glattem Rand

Sei $\Omega\subset\mathbb{R}^2$ ein beschränktes Gebiet mit glattem Rand. Bei linearen Finiten Ele-

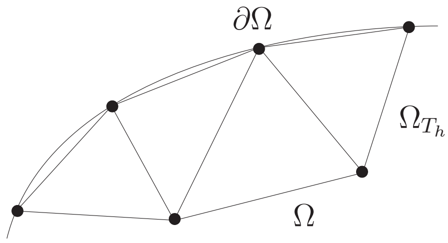


Abbildung 6: Polygonale Approximation Ω_{T_h} von Ω

menten ändert sich bei glattem Rand die Ordnung des Fehlers nicht, d.h. es gilt

$$\|u - u_h\|_{H^1} \leq C \cdot h, \text{ falls } u \in H^2(\Omega).$$

Bei quadratischen Finiten Elementen gilt jedoch

$$\|u - u_h\| \leq \begin{cases} C \cdot h^2, & \text{falls } u \in H^3(\Omega) \text{ und } \Omega \text{ polygonal berandet,} \\ C \cdot h^{3/2}, & \text{falls } u \in H^3(\Omega) \text{ und } \Omega \text{ beschränkt und glatt berandet.} \end{cases}$$

Die lineare Randapproximation führt also bei quadratischen Finiten Elementen zu einem Verlust in der Konvergenzordnung. Dieses Problem lässt sich aber über isoparametrische Elemente umgehen. Die Idee ist es, die Transformation R des Referenzdreiecks bei Randelementen geeignet zu modifizieren.

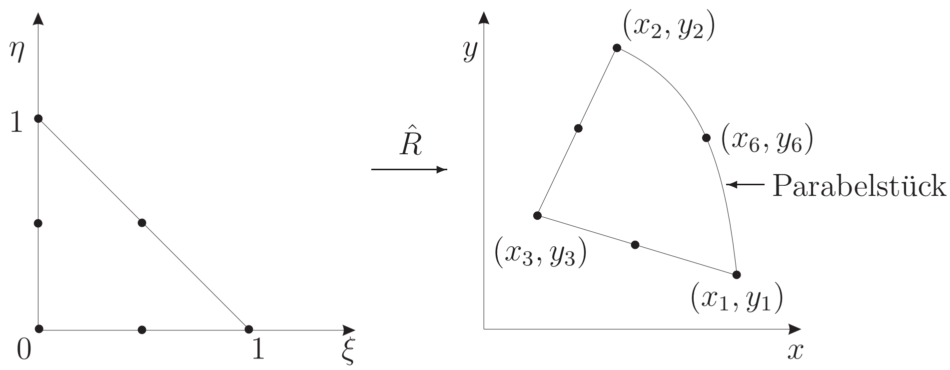


Abbildung 7: Transformation $\hat{R}$

1.13 Beispiel (Quadratische Finite Elemente, $r = 2$, $N_2 = 6$). Setze für quadratische Elemente an:

$$\begin{aligned} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} x_3 \\ y_3 \end{pmatrix} + \begin{pmatrix} x_1 - x_3 \\ y_1 - y_3 \end{pmatrix} \xi + \begin{pmatrix} x_2 - x_3 \\ y_2 - y_3 \end{pmatrix} \eta \\ &\quad + \begin{pmatrix} x_6 - (x_1 + x_2)/2 \\ y_6 - (y_1 + y_2)/2 \end{pmatrix} 4\xi\eta =: \hat{R}(\xi, \eta). \end{aligned}$$

$\hat{R}$ ist in (ξ, η) ein Polynom vom Grade 2 mit

$$\hat{R}(0,0) = \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}, \; \hat{R}(0,1) = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \; \hat{R}(1,0) = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \; \hat{R}(1/2,1/2) = \begin{pmatrix} x_6 \\ y_6 \end{pmatrix}.$$

Sind $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ und $\begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$ Randpunkte von Ω, so kann man natürlich den Rand durch eine Parabel besser approximieren als durch eine Gerade. Dies führt letzlich zu

$$\|u - u_h\|_{H^1} \leq C \cdot h^2,$$

falls $u \in H^3(\Omega)$ gilt und $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit glattem Rand ist, d.h. zur optimalen Ordnung quadratischer Finiter Elemente bei isoparametrischer Randapproximation.

b) Das Stokes-Problem

Die Gleichung von Stokes beschreibt die Bewegung einer inkompressiblen Flüssigkeit in einem Körper. Im zweidimensionalen Raum lautet sie

$$\begin{aligned} -\Delta u_1 + \frac{\partial p}{\partial x_1} &= f_1 \text{ in } \Omega, \\ -\Delta u_2 + \frac{\partial p}{\partial x_2} &= f_2 \text{ in } \Omega, \\ \operatorname{div} u = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} &= 0 \text{ in } \Omega, \\ u &= u_0 \text{ auf } \partial\Omega, \end{aligned} \tag{1-8}$$

wobei $\Omega \subset \mathbb{R}^2$ ein Gebiet ist. Es ist $u = (u_1, u_2) : \Omega \to \mathbb{R}^2$ das Geschwindigkeitsfeld und $p : \Omega \to \mathbb{R}$ der Druck.

Damit zu den Randdaten u_0 überhaupt eine divergenzfreie Strömung exisitieren kann, muss

$$\int_{\partial\Omega} u_0 \cdot n \, \mathrm{d}S = \int_{\partial\Omega} u \cdot n \, \mathrm{d}S = \int_{\Omega} \operatorname{div} u \, \mathrm{d}x = 0$$

gelten. Betrachte deshalb den Fall $u_0 = 0$, d.h. homogene Randwerte.

1.14 Definition. Die Funktionen $u \in C^2(\Omega, \mathbb{R}^2) \cap C^0(\overline{\Omega}, \mathbb{R}^2)$, $p \in C^1(\Omega)$, die (1-8) erfüllen, heißen klassische Lösung des Stokes-Problems.

1.15 Bemerkung. Der Druck p ist nur bis auf eine additive Konstante bestimmt. Man benutzt in der Regel die Normierung $\int_\Omega p \, \mathrm{d}x = 0$.

Variationelle Formulierung von (1-8)

Sei $\varphi = (\varphi_1, \varphi_2) \in C_0^\infty(\Omega, \mathbb{R}^2)$. Multipliziert man die ersten zwei Gleichungen in (1-8) mit φ skalar in $L^2(\Omega)$ und addiert die resultierten Gleichungen auf, so ergibt sich

$$\int_\Omega -\Delta u_1 \cdot \varphi_1 + \frac{\partial \varphi}{\partial x_1} \varphi_1 \, \mathrm{d}x + \int_\Omega -\Delta u_2 \cdot \varphi_2 + \frac{\partial \varphi}{\partial x_2} \varphi_2 \, \mathrm{d}x = \int_\Omega f_1 \varphi_1 + f_2 \varphi_2 \, \mathrm{d}x,$$

was zu

$$\int_\Omega \nabla u_1 \cdot \nabla \varphi_1 + \frac{\partial p}{\partial x_1}\varphi_1 \,\mathrm{d}x + \int_{\partial\Omega} \frac{\partial u_1}{\partial n} \underbrace{v_1}_{=0} \,\mathrm{d}S$$
$$+ \int_\Omega \nabla u_2 \cdot \nabla \varphi_2 + \frac{\partial p}{\partial x_2}\varphi_2 \,\mathrm{d}x + \int_{\partial\Omega} \frac{\partial u_2}{\partial n} \underbrace{v_2}_{=0} \,\mathrm{d}S = \int_\Omega f_1\varphi_1 + f_2\varphi_2 \,\mathrm{d}x$$

und daher auch zu

$$\int_\Omega \nabla u_1 \cdot \nabla \varphi_1 + \nabla u_2 \cdot \nabla \varphi_2 \,\mathrm{d}x - \int_\Omega p\frac{\partial \varphi_1}{\partial x_1} + p\frac{\partial \varphi_2}{\partial x_2} \,\mathrm{d}x$$
$$= \int_\Omega f_1\varphi_1 + f_2\varphi_2 \,\mathrm{d}x$$

äquivalent ist.

Gilt ferner $\operatorname{div}\varphi = 0$, so folgt

$$\int_\Omega \nabla u_1 \cdot \nabla \varphi_1 + \nabla u_2 \cdot \nabla \varphi_2 \,\mathrm{d}x = \int_\Omega f_1\varphi_1 + f_2\varphi_2 \,\mathrm{d}x$$

für alle $\varphi \in C_0^\infty(\Omega)$ mit $\operatorname{div}\varphi = 0$.

Sei nun $(H_0^1(\Omega))^2 = H_0^1(\Omega) \times H_0^1(\Omega)$. Da die Einbettung $C_0^\infty(\Omega, \mathbb{R}^2) \subset H_0^1(\Omega)$ dicht und das Funktional $\langle \nabla u, \nabla \cdot \rangle - \langle f, \cdot \rangle$ auf $H_0^1(\Omega)$ stetig ist, folgt

$$\int_\Omega \nabla u_1 \cdot \nabla \varphi_1 + \nabla u_2 \cdot \nabla \varphi_2 \,\mathrm{d}x = \int_\Omega f_1\varphi_1 + f_2\varphi_2 \,\mathrm{d}x$$

für alle $\varphi = (\varphi_1, \varphi_2) \in H_0^1(\Omega)$ mit $\operatorname{div}\varphi = 0$.

Schwache Formulierung

Es sei $V_0 = \{z \in H_0^1(\Omega) \,|\, \operatorname{div} z = 0\}$. Ferner seien

$$a : V_0 \times V_0 \to \mathbb{R}, \quad (u, v) \mapsto \int_\Omega \sum_{i=1}^2 \nabla u_i \cdot \nabla v_i \,\mathrm{d}x,$$
$$b : V_0 \to \mathbb{R}, \quad v \mapsto \int_\Omega \sum_{i=1}^2 f_i v_i \,\mathrm{d}x.$$

Gesucht ist ein $u_0 \in V_0$ mit

$$a(u, v) = b(v), \quad \forall v \in V_0. \tag{1-9}$$

Es lässt sich zeigen, dass die Bilinearform a stetig und V_0-elliptisch ist und b stetig ist, falls $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit hinreichend glattem Rand ist. Damit sichert der Satz von Lax und Milgram die Existenz einer schwachen Lösung von (1-9).

Das Galerkin-Verfahren zu (1-9) lautet nun: Sei $V_{0,h} \subset V_0$ ein endlich-dimensionaler Teilraum von V_0. Suche dann ein $u_h \subset V_{0,h}$ mit

$$a(u_h, v) = b(v), \quad \forall v \in V_{0,h}.$$

Es sei $\Omega_{T_h} = \{e_1, \dots, e_M\}$. Setzt man

$$V_{0,h} = \{u = (u_1, u_2) \in C(\Omega_{T_h}, \mathbb{R}^2) \mid \; u_k|_{e_i} \in P_1(e_i), k = 1, 2, i = 1, \dots, M \text{ und } \operatorname{div} u = 0\},$$

so lässt sich zeigen, dass der Raum $V_{0,h}$ nur die Nullfunktion enthält. Für die Numerik benötigt man also einen anderen Ansatz!

Variationelle Formulierung

Vorgelegt sei

$$\begin{aligned} -\Delta u_i + \frac{\partial p}{\partial x_i} &= f_i, \quad i = 1, 2, \\ \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} &= 0 \text{ in } \Omega \subset \mathbb{R}^2 \end{aligned}$$

mit homogenen Dirichlet-Randbedingungen

$$u = 0 \text{ auf } \partial\Omega.$$

Ist $\varphi = (\varphi_1, \varphi_2) \in C_0^\infty(\Omega, \mathbb{R}^2)$ und $\psi \in C_0^\infty(\Omega)$ mit $\int_\Omega \psi \,\mathrm{d}x = 0$, so folgt

$$\int_\Omega -\Delta u_1 \varphi_1 + \frac{\partial p}{\partial x_1} \varphi_1 \,\mathrm{d}x + \int_\Omega -\Delta u_2 \varphi_2 + \frac{\partial p}{\partial x_2} \varphi_2 \,\mathrm{d}x = \int_\Omega f_1 \varphi_1 + f_2 \varphi_2 \,\mathrm{d}x,$$

woraus sich

$$\sum_{i=1}^{2} \int_\Omega \nabla u_i \cdot \nabla \varphi_i - p \frac{\partial \varphi_i}{\partial x_i} \,\mathrm{d}x + \int_\Omega \frac{\partial u_i}{\partial n} \varphi_i \,\mathrm{d}S = \sum_{i=1}^{2} \int_\Omega f_i \varphi_i \,\mathrm{d}x$$

und wegen $\varphi|_{\partial\Omega} = 0$

$$\sum_{i=1}^{2} \int_\Omega \nabla u_i \cdot \nabla \varphi_i - p \frac{\partial \varphi_i}{\partial x_i} \,\mathrm{d}x = \int_\Omega \sum_{i=1}^{2} f_i \varphi_i \,\mathrm{d}x$$

ergibt. Wir multiplizieren nun die zweite Gleichung mit ψ und finden

$$\int_\Omega \underbrace{\left(\frac{\partial u_i}{\partial x_1} + \frac{\partial u_2}{\partial x_2} \right)}_{= \operatorname{div} u} \psi \,\mathrm{d}x = 0.$$

Setzt man nun

$$\begin{aligned} V &= (H_0^1(\Omega))^2 = H_0^1(\Omega) \times H_0^1(\Omega), \\ W &= L_*^2(\Omega) = \Big\{q \in L^2(\Omega) \,\Big|\, \int_\Omega q \,\mathrm{d}x = 0\Big\}. \end{aligned}$$

Aufgrund der Dichtheit von $C_0^\infty(\Omega, \mathbb{R}^2) \times D$, $D = \{v \in C_0^\infty(\Omega, \mathbb{R}) \mid \int_\Omega v \,\mathrm{d}x = 0\}$, in $(H_0^1(\Omega))^2 \times L_*^2(\Omega)$ sowie der Stetigkeit obiger Funktionale in der Topologie letzteren Raumes folgt

$$\begin{aligned} \int_\Omega \sum_{i=1}^2 \nabla u_i \cdot \nabla v_i - p \cdot \operatorname{div} v \,\mathrm{d}x &= \int_\Omega \sum_{i=1}^2 f_i v_i \,\mathrm{d}x, \quad \forall v = (v_1, v_2) \in (H_0^1(\Omega))^2 =: V, \\ -\int_\Omega \operatorname{div} u \cdot w \,\mathrm{d}x &= 0, \quad \forall w \in L_*^2(\Omega) =: W. \end{aligned}$$

Wir definieren lineare Abbildungen

$$\begin{aligned} a : V \times V &\to \mathbb{R}, \\ a(u, v) &= \int_\Omega \sum_{i=1}^2 \nabla u_i \cdot \nabla v_i \,\mathrm{d}x, \\ b : V \times W &\to \mathbb{R}, \\ b(v, q) &= -\int_\Omega \operatorname{div} u \cdot q \,\mathrm{d}x \end{aligned}$$

und erhalten die folgende Problemstellung: Gesucht ist ein $(u, p) \in V \times W$ mit

$$\begin{aligned} a(u, v) + b(v, p) &= \langle f, v\rangle \text{ für alle } v \in V, \\ b(u, q) &= 0 \text{ für alle } q \in W. \end{aligned} \tag{1-10}$$

1.16 Definition. Die Aufgabe (1-10) heißt die „Sattelpunktgleichung“ für das Stokes-Problem.

Eine Lösung (u, p) von (1-10) ist eine klassische Lösung des Stokes-Problems (1-8), falls $u \in C^2(\Omega, \mathbb{R}^2)$ und $p \in C^1(\Omega)$.

1.17 Bemerkung. Jede klassische Lösung der Sattelpunktgleichung (1-10) ist eine klassische Lösung von (1-8).

Sattelpunktprobleme

Wir wenden uns jetzt den Variationsproblemen mit Nebenbedingungen zu. Seien V und W Hilbert-Räume und

$$a : V \times V \to \mathbb{R}, \quad b : V \times W \to \mathbb{R}$$

seien stetige Bilinearformen, d.h.

$$|a(u,v)| \leq C_1 \|u\|_V \cdot \|v\|_V, \quad \forall u, v \in V$$

und

$$|b(v,q)| \leq C_2 \|v\|_V \cdot \|q\|_W, \quad \forall u \in V, q \in W.$$

Betrachtet sei das folgende Variationsproblem: Gesucht sind ein $u \in V$ sowie ein $p \in W$ mit

$$\begin{aligned} a(u,v) + b(v,p) &= F(v), \quad \forall v \in V, \\ b(u,q) &= G(p), \quad \forall q \in W \end{aligned} \tag{1-11}$$

mit $F \in V'$ und $G \in W'$, wobei V' und W' die Dualräume von V bzw. W (d.h. die Räume stetiger linearer Funktionale auf V bzw. W) bezeichnen. Ferner sei $Z = \{v \in V \,|\, b(v,q) = 0, \forall q \in W\}$. Es ist leicht zu sehen, dass $Z \subset V$ abgeschlossen ist.

Das Variationsproblem hängt eng mit folgender Extremwertaufgabe zusammen: Gesucht wird in V das Minimum von

$$R(v) = a(v,v) - 2F(v)$$

unter den Nebenbedingungen

$$b(v,y) = G(y), \quad \forall y \in W.$$

Die zur Extremwertaufgabe gehörige Lagrangefunktion $J : V \times W \to \mathbb{R}$ wird durch

$$\begin{aligned} J(v,w) &= R(v) - 2(G(w) - b(v,w)) \\ &= a(v,v) + 2b(v,w) - 2F(v) - 2G(w) \end{aligned}$$

gegeben.

1.18 Definition. Ein Punkt $(v^*, w^*) \in V \times W$ heißt ein Sattelpunkt von J, falls

$$J(v^*, w) \leq J(v^*, w^*) \leq J(v, w^*)$$

für alle $v \in V$ und $w \in W$ gilt.

Es gilt nun die folgende Charakterisierung:

1.19 Satz. *$(v^*, w^*) \in V \times W$ erfüllt*

$$\begin{aligned} a(v^*, x) + b(x, w^*) &= F(x), \quad \forall x \in V, \\ b(v^*, y) &= G(y), \quad \forall y \in W \end{aligned}$$

genau dann, wenn die Sattelpunkteigenschaft

$$J(v^*, w) \leq J(v^*, w^*) \leq J(v, w^*), \quad \forall v \in V, w \in W \tag{1-12}$$

gilt.

Beweis: s. Übungszettel. □

Wir kehren nun zur Analyse des Sattelpunktproblem (1-11) zurück und nehmen zunächst $G \equiv 0$ in (1-11) an, d.h. $u \in Z$. Dann wird u in (1-11) bestimmt durch: Gesucht ist ein $u \in Z$ mit

$$a(u, v) = F(v), \quad \forall v \in Z. \tag{1-13}$$

(1-13) ist korrekt gestellt (vgl. Satz von Lax und Milgram), falls a koerziv auf Z ist, d.h.

$$a(v, v) \geq \alpha \|v\|_V^2, \quad \forall v \in Z.$$

Im Falle des Stokes-Problems haben wir

$$a(u, v) = \sum_{i=1}^{2} \int_\Omega \nabla u_i \cdot \nabla v_i \, \mathrm{d}(x, y), \quad V = (H_0^1(\Omega))^2.$$

Da $a^i(u_i, v_i) = \int_\Omega \nabla u_i \cdot \nabla v_i \, \mathrm{d}(x, y)$, $i = 1, 2$, koerziv auf $H_0^1(\Omega)$ (vgl. Analyse der Poisson-Gleichung) ist, ist die Bilinearform sogar koerziv auf ganz V.

Der Fall $G \neq 0$ kann auf den Fall $G = 0$ zurückgeführt werden. Sei dazu $u_0 \in V$ irgendein Element mit $b(u_0, q) = G(q)$ für alle $q \in W$, d.h. $u_0 \notin Z$. Zerlege dann u aus (1-11) gemäß $u = u_1 + u_0$ und finde

$$b(u, q) = b(u_1, q) + \underbrace{b(u_0, q)}_{=G(q), \forall q \in W} = G(q), \quad \forall q \in W,$$

was mit

$$b(u_1, q) = 0, \quad \forall q \in W,$$

d.h. $u_1 \in Z$, äquivalent ist. Nun gilt

$$\begin{aligned} a(u, v) + b(v, p) &= a(u_0, v) + a(u_1, v) + b(v, p) \\ &= F(v), \quad \forall v \in V. \end{aligned}$$

Für $v \in Z$ gilt

$$b(v, p) = 0, \quad \forall v \in Z$$

und somit

$$a(u_1, v) = F(v) - a(u_0, v), \quad \forall v \in Z. \tag{1-14}$$

(1-14) ist wieder von der Form (1-13).

Wir betrachten also im Folgenden ohne Einschränkung die Aufgabe

$$\begin{aligned} a(u, v) + b(v, p) &= F(v), \quad \forall v \in V, \\ b(u, q) &= 0, \quad \forall q \in W. \end{aligned} \tag{1-15}$$

Damit ist u als die Lösung von

$$a(u, v) = F(v), \quad \forall v \in Z$$

gegeben. Wir setzen dieses $u \in Z$ in die erste Gleichung in (1-15) ein und erhalten

$$b(v, p) = -a(u, v) + F(v), \quad \forall v \in V. \tag{1-16}$$

Die Korrektgestelltheit dieses Problems erfordert einen neuen Typ von Koerzivität.

Wir betrachten die Bedingung

$$\beta \cdot \|p\|_W \leq \sup_{v \in V} \frac{b(v, p)}{\|v\|_V}, \quad \forall p \in W \tag{1-17}$$

für ein $\beta > 0$. Letztere Ungleichung heißt „Babuška-Brezzi-Bedingung“ oder kurz BB-Bedingung.

Für die Motivation betrachten wir eine koerzive Bilinearform $a : V \times V \to \mathbb{R}$, d.h.

$$a(v, v) \geq \gamma \|v\|^2, \quad \forall v \in V.$$

Dies impliziert

$$\gamma \|v\|_V \leq \frac{a(v, v)}{\|v\|_V} \leq \sup_{w \in V} \frac{a(w, v)}{\|w\|_V}.$$

Diese Ungleichung legt nun für $b : V \times W \to \mathbb{R}$ die Bedingung

$$\gamma \|p\|_W \leq \sup_{w \in V} \frac{b(w, p)}{\|w\|_V}$$

nahe.

Unser Problem (1-16) ist von der Form

$$b(v, p) = \widetilde{F}(v), \quad v \in V \tag{1-18}$$

mit $\widetilde{F}(v) = -a(u, v) + F(v)$. Ferner gilt für alle $v \in Z$:

$$\widetilde{F}(v) = -a(u, v) + F(v) = b(v, p) = 0,$$

da u die Gleichung (1-15) löst. Wir zeigen nun die Eindeutigkeit der Lösung von (1-18) unter der Bedingung (1-17). Seien $p_1, p_2 \in W$ zwei Lösungen von (1-18). Setze $\hat{p} = p_1 - p_2 \in W$. Dann gilt

$$b(v, \hat{p}) = b(v, p_1) - b(v, p_2) = \widetilde{F}(v) - \widetilde{F}(v) = 0, \quad \forall v \in V.$$

Mit der Babuška-Brezzi-Bedingung folgt

$$\beta \|\hat{p}\|_W \leq \sup_{w \in V} \frac{b(w, \hat{p})}{\|w\|_V} = 0$$

und somit $\|\hat{p}\|_W = 0$, d.h. $p_1 = p_2$. Des Weiteren lässt sich auch die Existenz einer Lösung zeigen. Man erhält also:

1.20 Lemma. *Vorgelegt sei das Variationsproblem (1-11) mit stetigen Bilinearformen a und b. Gilt die Babuška-Brezzi-Bedingung (1-17), so hat das Problem*

$$b(v,p) = -a(u,v) + F(v), \quad \forall v \in V$$

genau eine Lösung $p \in W$.

Diskrete gemischte Formulierung

Seien nun $V_h \subset V$ und $W_h \subset W$. Wir betrachten jetzt die Aufgabe: Finde ein $u_h \in V_h$ und $p_h \in W_h$ mit

$$\begin{aligned} a(u_h,v) + b(v,p_h) &= F(v) \text{ für alle } v \in V_h, \\ b(u_h,q) &= 0 \text{ für alle } q \in W_h. \end{aligned} \tag{1-19}$$

Analog zum kontinuierlichen Fall setzen wir

$$Z_h = \{v \in V_h \,|\, b(v,q) = 0 \text{ für alle } q \in W_h\}.$$

Nun ist (1-19) äquivalent zum Folgenden: Bestimme ein $u_h \in Z_h$ mit

$$a(u_h,v) = F(v), \quad \forall v \in Z_h \tag{1-20}$$

und finde dann ein $p_h \in W_h$ mit

$$b(v,p_h) = -a(u_h,v) + F(v), \quad \forall v \in V_h. \tag{1-21}$$

Wäre $Z_h \subset Z$, so könnte man Ceas Lemma anwenden, um $\|u - u_h\|_V$ aus (1-20)—(1-21) abzuschätzen. Im Allgemeinen ist dies nicht der Fall und wir benötigen eine entsprechende Verallgemeinerung von Ceas Lemma.

1.21 Lemma. *Seien V und V_h Unterräume eines Hilbertraumes H. Es sei a eine stetige (nicht notwendigerweise symmetrische) Bilinearform mit einer Konstanten $C > 0$ auf H, welche koerziv auf V_h ist, d.h.*

$$\gamma \|v\|_V^2 \le a(v,v), \quad \forall v \in V_h.$$

Ferner löse $u \in V$ die Aufgabe

$$a(u,v) = F(v), \quad \forall v \in V$$

mit $F \in H' = \{R : H \to \mathbb{R} \,|\, R$ linear und stetig $\}$ und $u_h \in V_h$ löse

$$a(u_h,v) = F(v), \quad \forall v \in V_h.$$

Dann gilt

$$\|u - u_h\|_H \le \left(1 + \frac{C}{\gamma}\right) \inf_{v \in V_h} \|u - v\|_H + \frac{1}{\gamma} \sup_{w \in V_h \setminus \{0\}} \frac{|a(u - u_h, w)|}{\|w\|_H}.$$

Beweis: Die Koerzivität von a, d.h.

$$\gamma\|v\|_H^2 \leq a(v,v), \quad \forall v \in V_h,$$

impliziert

$$\gamma\|v\|_H \leq \frac{a(v,v)}{\|v\|_H} \leq \sup_{w\in V_h\setminus\{0\}} \frac{|a(v,w)|}{\|w\|_H}, \quad \forall v \in V_h.$$

Sei nun $v \in V_h$ beliebig. Es folgt

$$\begin{aligned}
\|u-u_h\|_H &\leq \|u-v\|_H + \|v-u_h\|_H \\
&\leq \|u-v\|_H + \frac{1}{\gamma}\sup_{w\in V_h\setminus\{0\}} \frac{|a(v-u_h,w)|}{\|w\|_H} \\
&\leq \|u-v\|_H + \frac{1}{\gamma}\left(\sup_{w\in V_h\setminus\{0\}} \frac{|a(v-u,w)|}{\|w\|_H} + \sup_{w\in V_h\setminus\{0\}} \frac{|a(u-u_h,w)|}{\|w\|_H}\right) \\
&\leq \|u-v\|_H + \frac{C}{\gamma}\|v-u\|_H + \frac{1}{\gamma}\sup_{w\in V_h\setminus\{0\}} \frac{|a(u-u_h,w)|}{\|w\|_H}, \quad \forall v \in V_h,
\end{aligned}$$

woraus sich unmittelbar die Behauptung ergibt. □

In unserem Fall wenden wir das Lemma 1.21 an mit $\gamma = \alpha$, $H = (H_0^1(\Omega))^2 = V$, $Z = V$, $V_h = Z_h$ und finden

$$\|u-u_h\|_V \leq \left(1+\frac{C}{\alpha}\right)\inf_{v\in Z_h}\|u-v\|_V + \frac{1}{\alpha}\sup_{w\in Z_h\setminus\{0\}} \frac{|a(u-u_h,w)|}{\|w\|_V}, \tag{1-22}$$

falls $\alpha\|v\|_V^2 \leq a(v,v)$ für alle $v \in Z_h$ gilt. Wir analysieren nun den hinteren Term in (1-22). Ist $w \in Z_h$, so gilt

$$\begin{aligned}
a(u-u_h,w) &= a(u,w) - a(u_h,w) \\
&= a(u,w) - F(w) = -b(w,p) \\
&= -b(w,p) + \underbrace{b(w,q)}_{=0,\text{ da } q\in W_h} = -b(w,p-q), \quad \forall q \in W_h.
\end{aligned}$$

Die Stetigkeit von b liefert ferner

$$|b(w,p-q)| \leq C\cdot\|w\|_V\cdot\|p-q\|_W.$$

Da $q \in W_h$ beliebig war, folgt

$$|a(u-u_h,w)| \leq |b(w,p-q)| \leq C\cdot\|w\|_V\|p-q\|_W, \quad \forall q \in W_h,$$

d.h.

$$|a(u-u_h,w) \leq C\cdot\|w\|_V\cdot\inf_{q\in W_h}\|p-q\|_W.$$

Einsetzen in (1-22) ergibt die Abschätzung

$$\|u-u_h\|_V \leq \left(1+\frac{C}{\alpha}\right) \inf_{v\in Z_h} \|u-v\|_V + \frac{C}{\alpha}\cdot \inf_{q\in W_h} \|p-q\|_W. \tag{1-23}$$

Das Resultat (1-23) ist als „Brezzi-Theorem” bekannt.

Der Hauptpunkt in der Abschätzung (1-23) von Brezzi ist, dass die Güte der Abschätzung nur von der Approximation der Räume Z_h und W_h sowie der Babuška-Brezzi-Bedingung

$$\beta\cdot\|p\|_W \leq \sup_{v\in V} \frac{b(v,p)}{\|v\|_V}, \quad \forall p\in W$$

abhängt.

Konvergenzresultate für das Geschwindigkeitsfeld

Es sei $\Omega\subset\mathbb{R}^2$ ein beschränktes, polygonal berandetes Gebiet. Zu einer Triangulierung $\Omega_{T_h}=\{e_1,\dots,e_M\}$ von Ω seien

$$\begin{aligned}
V_{T_h}^{(r)} &= \{v=(v_1,v_2)\,|\quad v_i|_{e_k}\in P^r(e_k), i=1,2, k=1,\dots,M,\\
&\qquad\qquad\qquad\quad v_i|_{\partial\Omega_{T_h}}=0, i=1,2\},\\
W_{T_h}^{(r)} &= \{w\,|\quad w|_{e_k}\in P^{r-1}(e_k), k=1,\dots,M, \int_\Omega w\,\mathrm{d}x=0\}
\end{aligned}$$

für $r\geq 2$ mit $V_{T_h}^{(r)}\subset V=(H_0^1(\Omega))^2$ und $W_{T_h}^{(r)}\subset W=L_*^2(\Omega)$, wobei $L_*^2(\Omega)=\{u\in L^2(\Omega)\,|\,\int_\Omega u\,\mathrm{d}x=0\}$. h sei die maximale Kantenlänge der Triangulierung. Es sei nun $\{\Omega_{T_h}\}_{h>0}$ eine Familie von Triangulierungen von Ω mit maximaler Kantenlänge h, welche nicht entartet sei, d.h. die Maximalwinkelbedingung

$$\sup\{\alpha_k\in e_k\,|\,k=1,\dots,M_h\}\leq\alpha<\frac{\pi}{2}, \quad 0<h<h_0$$

sei erfüllt. Dann gilt die Abschätzung

$$\|u-u_h\|_{(H^1(\Omega))^2}\leq C\cdot h^r\left(\|u\|_{(H^{r+1}(\Omega))^2}+\|p\|_{H^r(\Omega)}\right),$$

falls $(u,p)\in(H^{r+1}(\Omega))^2\times H^r(\Omega)$.

Konvergenzresultate für den Druck

Seien $V_h\subset V$, $W_h\subset W$ und $Z_h=\{v\in V_h\,|\,b(v,q)=0,\forall q\in W_h\}$. Bestimme ein $u_h\in Z_h$ mit

$$a(u_h,v)=F(v), \quad \forall v\in Z_h$$

und berechne dann $p_h\in W_h$ mit

$$b(v,p_h)=-a(u_h,v)+F(v), \quad \forall v\in V_h. \tag{1-24}$$

Es bleibt also die Lösbarkeit von (1-24) in V_h, $h > 0$, zu analysieren. Für jedes feste $h > 0$ benötigen wir die Bedingung

$$\beta_h \cdot \|q\|_W \leq \sup_{v \in V_h} \frac{|b(v,q)|}{\|v\|_V}, \quad \forall q \in W_h \text{ mit } \beta_h \geq 0.$$

Für die Lösung von (1-24) zu $0 < h \leq h_0$ benötigt man die sogenannte „Infimum-Supremum-Bedingung"

$$0 < \beta := \inf_{q \in W_h} \sup_{v \in V_h} \frac{|b(v,q)|}{\|v\|_V \cdot \|q\|_W}, \tag{1-25}$$

um die eindeutige Lösbarkeit von (1-24) für $0 < h \leq h_0$ zu sichern. Die Norm von $(p - p_h)$ lässt sich dabei folgendermaßen abschätzen. Man diskretisiert das kontinuierliche Problem: Gesucht ist ein $(u,p) \in V \times W$ mit

$$\begin{aligned} a(u,v) + b(v,p) &= F(v), \quad \forall v \in V, \\ b(u,q) &= 0, \quad \forall q \in W \end{aligned}$$

durch ein Galerkin-Verfahren: Finde ein $(u_h,p_h) \in V_h \times W_h$ mit

$$\begin{aligned} a(u_h,v) + b(v,p_h) &= F(v), \quad \forall v \in V_h \subset V, \\ b(u_h,q) &= 0, \quad \forall q \in W_h \subset W \end{aligned}$$

und geht danach wie folgt vor.

Substitution der ersten Zeilen liefert dann

$$a(u - u_h, v) + b(v, p - p_h) = 0, \quad \forall v \in V_h,$$

was mit

$$b(v, p - p_h) = -a(u - u_h, v), \quad \forall v \in V_h \tag{1-26}$$

gleichbedeutend ist. Ist nun ein $q \in W_h$ beliebig gewählt, so folgt mit (1-25) und (1-26)

$$\begin{aligned} \beta\|q - p_h\| &\leq \sup_{v \in V_h} \frac{|b(v, q - p_h)|}{\|v\|_V} \\ &= \sup_{v \in V_h} \frac{|b(v, p - p_h) + b(v, q - p)|}{\|v\|_V} \\ &= \sup_{v \in V_h} \frac{|-a(u - u_h, v) + b(v, q - p)|}{\|v\|_V} \\ &\leq C_1\|u - u_h\|_V + C_2\|q - p\|_W \\ &\leq C\left(\|u - u_h\|_V + \|q - p\|_W\right), \quad C := \max\{C_1, C_2\} \end{aligned} \tag{1-27}$$

für alle $q \in W_h$. Wir erhalten dann für $q \in W_h$

$$\|p - p_h\|_W \leq \|p - q\|_W + \|q - p_h\|_W$$

$$\leq \quad \|p - q\|_W + \frac{C}{\beta}\left(\|u - u_h\|_V + \|q - p\|_W\right)$$

und somit die Abschätzung

$$\|p - p_h\|_W \leq \frac{C}{\beta}\|u - u_h\|_V + \left(1 + \frac{C}{\beta}\right)\inf_{q \in W_h}\|p - q\|_W. \tag{1-28}$$

Ferner lässt sich die Gültigkeit von

$$\inf_{q \in W_{T_h}^{(r)}} \|p - q\|_W \leq C \cdot h^r \|p\|_{H^r(\Omega)}$$

mit $W_{T_h}^{(r)} = \{w \mid \ w|_{e_k} \in P^{r-1}(e_k), k = 1, \dots, M, \int_\Omega w \, \mathrm{d}x = 0\}$, $r \geq 2$, zeigen.

Geeignete Finite-Elemente-Räume V_h und W_h

Das Taylor-Hood Element zu einer Triangulierung $\Omega_{T_h} = \{e_1, \dots, e_M\}$ lautet

$$\begin{aligned} V_{T_h}^{(2)} &= \{v = (v_1, v_2) \mid v_i|_{e_k} \in P^2(e_k), i = 1, 2, k = 1, \dots, M, \\ &\qquad v_i|_{\partial\Omega_{T_h}} = 0, i = 1, 2\}, \\ W_{T_h}^{(2)} &= \{w \mid w|_{e_k} \in P^1(e_k), k = 1, \dots, M, \int_\Omega w \, \mathrm{d}x = 0\} \end{aligned}$$

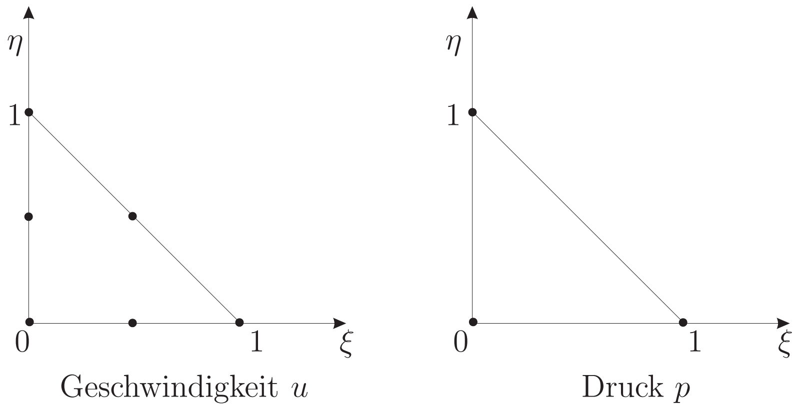


Abbildung 8: Knotenverteilung des Referenzdreiecks e^R

Für das Taylor-Hood-Element kann man die Infimum-Supremum-Bedingung

$$0 < \beta = \inf_{q \in W_h} \sup_{v \in V_h} \frac{|b(v, q)|}{\|v\|_V \cdot \|q\|_W}$$

nachweisen. Insgesamt gilt für nichtentartete Triangulierungen eines polygonal berandeten Gebietes Ω die Abschätzung

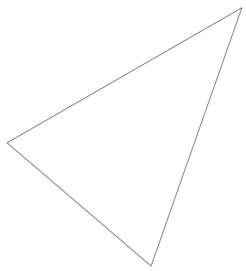
$$\|u - u_h\|_{(H^1(\Omega))^2} + \|p - p_h\|_{L^2(\Omega)} \leq C \cdot h^2 \left(\|u\|_{(H^3(\Omega))^2} + \|p\|_{H^2(\Omega)}\right),$$

falls $(u, p) \in (H^3(\Omega))^2 \times H^2(\Omega)$.

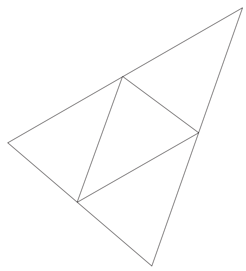
Die numerische Variante des Taylor-Hood-Elementes lautet: Es seien

$$\Omega_{T_h} = \{e_1, \ldots, e_M\}, \quad W_h = W_{T_h}^{(2)}.$$

Ferner sei $\Omega_{T_{h/2}}$ diejenige Triangulierung, welche sich aus Ω_{T_h} ergibt, wenn jedes Dreieck e durch Teilung an den Kantenmitten in vier dazu kongruente Dreiecke zerlegt wird (siehe Abbildung 9)



(a) $e \in \Omega_{T_h}$



(b) $e = \cup_{i=1}^4 e^i \in \Omega_{T_{h/2}}$

Abbildung 9: Dreiecke aus Ω_{T_h} bzw. $\Omega_{T_{h/2}}$

Setzt man $V_h = V_{T_{h/2}}^{(1)}$, so erweist sich auch diese Variante als stabil. Man verwendet also Polynome zweiter Ordnung für die Approximation des Geschwindigkeitsfeldes.

Aufstellen des Gleichungssystems

Es seien

$$\begin{aligned} V_h \subset V, &\quad V_h = \text{span}\{v_1, \ldots, v_N\}, \\ W_h \subset W, &\quad W_h = \text{span}\{w_1, \ldots, w_M\}. \end{aligned}$$

Man wählt den Ansatz $v^h = \sum_{i=1}^N c_i v_i$, $w^h = \sum_{i=1}^M d_i w_i$ zur Bestimmung der Lösung (v^h, w^h) von

$$\begin{aligned} a(v^h, x) + b(x, w^h) &= F(x) \text{ für alle } x \in V_h, \\ b(v^h, y) &= 0 \text{ für alle } y \in W_h. \end{aligned} \tag{1-29}$$

Aufgrund der Linearität des Problems genügt es, die Gültigkeit von (1-29) auf einer Basis von $V_h \times W_h$ zu sichern. Dies liefert

$$a(\sum_{k=1}^N c_k v_k, v_i) + b(v_i, \sum_{k=1}^M d_k w_k) = F(v_i), \quad i = 1, \ldots, N,$$

$$b(\sum_{k=1}^{N} c_k v_k, w_j) \quad = \quad 0, \quad j = 1, \ldots, M.$$

Setzt man

$$\begin{aligned}
c &= (c_1, \ldots, c_N), \quad d \;=\; (d_1, \ldots, d_M), \\
A &= (a_{ij})_{ij} \in \mathbb{R}^{N,N}, \quad a_{ij} \;=\; a(v_i, v_j),\ 1 \le i, j \le N, \\
B &= (b_{ij})_{ij} \in \mathbb{R}^{N,M}, \quad b_{ij} \;=\; b(v_i, w_j),\ 1 \le i \le N, 1 \le j \le M, \\
f_h &= (f_1, \ldots, f_N), \quad f_i = F(v_i),\ i = 1, \ldots, N,
\end{aligned}$$

so ergibt sich das Gleichungssystem

$$\begin{aligned}
A_h c + B_h d &= f_h, \\
B_h^T c &= 0
\end{aligned} \tag{1-30}$$

bzw.

$$\begin{pmatrix} A_h & B_h \\ B_h^T & 0 \end{pmatrix} \cdot \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} f_h \\ 0 \end{pmatrix} \in \mathbb{R}^{N+M}.$$

Ist A_h invertierbar, so folgt

$$A_h c = f_h - B_h d$$

bzw.

$$c = A_h^{-1}(f_h - B_h d)$$

und man erhält

$$\begin{aligned}
0 \;=\; B_h^T c &= B_h^T A_h^{-1}(f_h - B_h d), \\
(B_h^T A_h^{-1} B_h) d &= B_h^T A_h^{-1} f_h.
\end{aligned}$$

Das bekannteste Iterationsverfahren für Sattelpunktprobleme ist der Uzawa-Algorithmus.

Uzawa-Algorithmus: Sei $d^0 \in \mathbb{R}^M$ und $\alpha > 0$ vorgegeben. Bestimme für $k = 1, 2, \ldots$ ein Paar (c^k, d^k), welches

$$\begin{aligned}
A_h c^k &= f_h - B_h d^{k-1}, \\
d^k &= d^{k-1} + \alpha B_h^T c^k.
\end{aligned} \tag{1-31}$$

löst. Dabei wird der Schrittweitenparameter $\alpha > 0$ als genügend klein vorausgesetzt.

1.22 Lemma. *Vorgelegt sei das Gleichungssystem (1-30). Die Matrix $A_h \in \mathbb{R}^{N,N}$ sei symmetrisch und positiv definit und $B_h \in \mathbb{R}^{N,M}$ habe den Rang $\mathrm{rg}(B_h) = M$. Dann besitzt (1-30) genau eine Lösung (c^*, d^*) und diese ist durch den Uzawa-Algorithmus (1-31) berechenbar, falls $\alpha \|B_h A_h^{-1} B_h^T\|_2 < 2$ gilt.*

Beweis: Sei (c^*, d^*) die Lösung des Sattelpunktproblems (1-30). Es gilt

$$\begin{aligned} d^k - d^{k-1} &= \alpha B_h^T c^k \\ &= \alpha B_h^T A_h^{-1}(f_h - B_h d^{k-1}). \end{aligned}$$

Aus $A_h c^* + B_h d^* = f_h$ folgt

$$A_h^{-1} f_h = c^* + A_h^{-1} B_h^* d^*$$

und somit ergibt sich

$$\begin{aligned} d^k - d^{k-1} &= \alpha B_h^T A_h^{-1} f_h - \alpha B_h^T A_h^{-1} d^{k-1} \\ &= \alpha \underbrace{B_h^T c^*}_{=0} + \alpha B_h^T A_h^{-1} B_h d^* - \alpha B_h^T A_h^{-1} B_h d^{k-1} \\ &= \alpha B_h^T A_h^{-1} B_h (d^* - d^{k-1}). \end{aligned}$$

Für den durch $e^k := d^k - d^*$ definierten Fehler im k-ten Schritt erhalten wir

$$e^k - e^{k-1} = d^k - d^{k-1} = -\alpha B_h^T A_h^{-1} B_h e^{k-1},$$

d.h.

$$e^k = \left(I - \alpha B_h^T A_h^{-1} B_h\right) e^{k-1}, \quad k \in \mathbb{N}.$$

Induktiv folgt

$$e^k = (I - \alpha B_h^T A_h^{-1} B_h)^k e^0$$

bzw.

$$d^k - d^* = (I - \alpha B_h^T A_h^{-1} B_h)^k (d_0 - d^*).$$

Dies impliziert

$$\|d^k - d^0\|_2 \le \|(I - \alpha \underbrace{B_h^T A_h^{-1} B_h}_{=:C_h} \|_2^k \|d^0 - d^*\|_2. \tag{1-32}$$

Sei nun $\sigma(C_h) = \{\lambda_1, \ldots, \lambda_M\}$ mit $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_M$, wobei zu beachten ist, dass C_h eine symmetrische und positiv definite Matrix ist, da auch A_h eine solche ist und $\mathrm{rg}(B_h) = M$ gilt. Somit folgt

$$\|C_h\|_2 = \max\{|\lambda_i| \mid i = 1, \ldots, M\} = \lambda_M.$$

Ferner gilt

$$\sigma(I - \alpha C_h) = \{1 - \alpha\lambda_1, \ldots, 1 - \alpha\lambda_M\}$$

und somit

$$\begin{aligned} \|I - \alpha C_h\|_2 < 1 \quad &\Leftrightarrow \quad 1 - \alpha\lambda_M > -1 \\ &\Leftrightarrow \quad 2 > \alpha\lambda_M = \alpha\|C_h\|_2. \end{aligned}$$

Einsetzen in (1-32) liefert

$$\|d^k - d^*\|_2 \leq \|I - \alpha C_h\|_2^k \cdot \|d^0 - d^*\|_2 \to 0 \text{ für } k \to \infty.$$

Ferner folgt sofort

$$c^k = A_h^{-1}(f_h - B_h d^{k-1}) \to A_h^{-1}(f_h - B_h d^*) = c^* \text{ für } k \to \infty.$$

□

Verifikation der Brezzi-Bedingung für die Stokes-Gleichung

Ein Paar $(u,p) \in V \times W$ ist eine schwache Lösung des Stokes-Problems, falls

$$\begin{aligned} a(u,v) + b(v,p) &= F(v) \text{ für alle } v \in V, \\ b(u,q) &= 0 \text{ für alle } q \in W \end{aligned}$$

gilt, wobei

$$\begin{aligned} V &= (H_0^1(\Omega))^2, \quad W = L_*^2(\Omega), \\ a(u,v) &= \int_\Omega \sum_{i=1}^2 \nabla u_i \cdot \nabla v_i \,\mathrm{d}x, \\ b(v,q) &= -\int_\Omega \operatorname{div} u \cdot q \,\mathrm{d}x, \\ F(v) &= \int_\Omega \sum_{i=1}^2 f_i v_i \,\mathrm{d}x. \end{aligned}$$

Partielle Integration für $q \in H^1(\Omega)$ ergibt

$$\begin{aligned} b(v,q) &= -\int_\Omega \operatorname{div}(v) \cdot q \,\mathrm{d}x = -\int_\Omega \sum_{i=1}^2 \frac{\partial v_i}{\partial x_i} \cdot q \,\mathrm{d}x \\ &= \int_\Omega \sum_{i=1}^2 v \cdot \frac{\partial q}{\partial x_i} \,\mathrm{d}x - \underbrace{\int_{\partial\Omega} \sum_{i=1}^2 v_i q n_i \,\mathrm{d}S}_{=0} \\ &= \int_\Omega v \cdot \nabla q \,\mathrm{d}x = \langle v, \nabla q \rangle_{L^2(\Omega)}. \end{aligned} \tag{1-33}$$

Somit lautet die Babuška-Brezzi-Bedingung

$$\beta \|p\|_{L^2(\Omega)} \leq \sup_{v \in (H_0^1(\Omega))^2} \frac{\langle v, \nabla p \rangle_{L^2(\Omega)}}{\|v\|_{H^1}}, \quad \forall p \in L_*^2(\Omega).$$

Sobolev-Räume mit negativem Index

1.23 Definition. Es sei durch

$$(H^{-m}(\Omega))^k = \{f : (H_0^m(\Omega))^k \to \mathbb{R} \mid f \text{ linear und stetig}\}$$

der Dualraum zu $(H_0^m(\Omega))^k$ bezeichnet, versehen mit der Norm

$$\|g\|_{H^{-m}(\Omega)} = \sup_{v \in H_0^m(\Omega)} \frac{|g(v)|}{\|v\|_{H^m(\Omega)}}.$$

Es lässt sich zeigen, dass $H_0^m(\Omega) \overset{(1)}{\subset} L^2(\Omega) \overset{(2)}{\subset} H^{-m}(\Omega)$ gilt, wobei die Einbettungen (1) und (2) stetig und dicht sind. Man kann deshalb das Skalarprodukt $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$ auf $L^2(\Omega)$ zu einer bilinearen, stetigen Abbildung $H_0^m \times H_0^{-m} \to \mathbb{R}$ fortsetzen und schreibt $g(v) = \langle g, v \rangle_{L^2(\Omega)}$. Damit gilt

$$\|g\|_{H^{-m}(\Omega)} = \sup_{v \in H_0^m(\Omega)} \frac{|\langle g, v \rangle_{L^2(\Omega)}|}{\|v\|_{H^m(\Omega)}}.$$

1.24 Satz. *Gegeben sei die Abbildung*

$$\nabla : L^2(\Omega) \to H^{-1}(\Omega), \quad g \mapsto \nabla g$$

auf einem beschränkten Gebiet Ω, welches die strikte Kegeleigenschaft erfülle. Dann existiert ein $C = C(\Omega) > 0$ derart, dass

$$\begin{aligned} \|p\|_{L^2(\Omega)} &\leq C \cdot \left(\|\nabla p\|_{H^{-1}(\Omega)} + \|p\|_{H^{-1}(\Omega)}\right), \quad \forall p \in L_*^2(\Omega), \\ \|q\|_{L^2(\Omega)} &\leq C \cdot \|\nabla q\|_{H^{-1}(\Omega)} \end{aligned}$$

gilt.

Wir zeigen nun, dass die Stokes-Gleichung unter den Voraussetzungen von Satz 1.24 die Babuška-Brezzi-Bedingung erfüllt. Nach Satz 1.24 gilt

$$\|p\|_{L^2(\Omega)} \leq C \cdot \|\nabla p\|_{H^{-1}(\Omega)}, \quad \forall q \in L_*^2(\Omega) =: W$$

und

$$\|\nabla p\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)} \frac{|\langle \nabla p, v \rangle_{L^2(\Omega)}|}{\|v\|_{H^1(\Omega)}}.$$

Nach Definition des Supremums existiert eine Folge $(v_n)_{n\in\mathbb{N}} \subset H_0^1(\Omega)$ mit

$$\|\nabla p\|_{H^{-1}(\Omega)} = \lim_{n\to\infty} \frac{|\langle \nabla p, v_n \rangle_{L^2(\Omega)}|}{\|v_n\|_{H^1(\Omega)}} \geq \frac{1}{C} \cdot \|p\|_{L^2(\Omega)}.$$

Somit finden wir ein $\widetilde{v} \in (H_0^1(\Omega))^2$ mit

$$\frac{|\langle \nabla p, \widetilde{v} \rangle_{L^2(\Omega)}|}{\|\widetilde{v}\|_{H^1(\Omega)}} = \frac{\langle \nabla p, \widetilde{v} \rangle_{L^2(\Omega)}}{\|\widetilde{v}\|_{H^1(\Omega)}} \geq \frac{1}{2C} \cdot \|p\|_{L^2(\Omega)}.$$

Mit (1-33) folgt dann

$$\sup_{v \in (H_0^1(\Omega))^2} \frac{b(v,p)}{\|v\|_{H^1(\Omega)}} \geq \frac{b(\widetilde{v},p)}{\|\widetilde{v}\|_{H^1(\Omega)}} = \frac{\langle \nabla p, \widetilde{v} \rangle_{L^2(\Omega)}}{\|\widetilde{v}\|_{H^1(\Omega)}} \geq \frac{1}{2C} \cdot \|p\|_{L^2(\Omega)}$$

für alle $p \in L^2_*(\Omega)$, d.h. die Babuška-Brezzi-Bedingung ist mit $\beta = \frac{1}{2C}$ erfüllt.

2. Numerik parabolischer Differentialgleichungen

a) Finite Differenzenmodelle

Vorgelegt sei die Anfangsrandwertaufgabe

$$\begin{aligned} u_t &= u_{xx} + f(u,x,t), \quad x \in \Omega = (0,1), \quad t \in (0,T), \\ u(x,0) &= u_0(x), \quad x \in \Omega, \\ u(0,t) &= \gamma_0(t), \quad u(1,t) = \gamma_1(t), \quad 0 \le t \le T. \end{aligned} \tag{2-1}$$

Wir diskretisieren das Problem (2-1) mit der Linienmethode, d.h. die Ableitungen in der Raumvariablen $x \in (0,1)$ werden durch entsprechende Differenzenquotienten ersetzt. Dazu wählen wir ein Ortsgitter $\Omega_{\Delta x} = \{j\Delta x \,|\, j = 0, \ldots, M\}$, $\Delta x = \frac{1}{M}$. Es sei ferner

$$v(t) = (u(\Delta x,t), u(2\Delta x,t), \ldots, u(1-\Delta x,t)) = (v_1, v_2, \ldots, v_{M-1})(t), \quad 0 \le t \le T.$$

Wir ersetzen den Ausdruck $u_{xx}(x,t) + f(u(x,t), u_x(x,t), u, t)$, $u(0,t) = \gamma_0(t)$, $u(1,t) = \gamma_1(t)$ für $t \in [0,T]$ durch

$$-A^{\Delta x} v(t) + r^{\Delta x}(t) + H^{\Delta x}(v(t))$$

mit

$$A^{\Delta x} = -\frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix},$$

$$r^{\Delta x} = \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0(t) \\ 0 \\ \vdots \\ 0 \\ \gamma_1(t) \end{pmatrix},$$

$$H^{\Delta x}(v(t)) = \begin{pmatrix} f\left(v_1(t), \Delta x, t\right) \\ f\left(v_j(t), j\Delta x, t\right), \quad j = 2, \ldots, M-2 \\ f\left(v_{M-1}(t), (M-1)\Delta x, t\right) \end{pmatrix}. \tag{2-2}$$

Präziser wird $A(t)u(\cdot,t) + f(u(\cdot,t),\cdot,t)$, $u(t,\cdot) \in D(A(t))$ durch

$$-A^{\Delta x} v(t) + r^{\Delta x}(t) + H^{\Delta x}(v(t))$$

ersetzt, wobei der zeitabhängige Differentialoperator $A(t)$ mit dem Definitionsbereich

$$D(A(t)) := \{u \in C^2((0,1)) \cap C([0,1]) \,|\, u(0) = \gamma_0(t),\; u(1) = \gamma_1(t)\}$$

wie folgt definiert ist:

$$\begin{aligned} A(t) : D(A(t)) \subset C^2((0,1)) \cap C([0,1]) &\longrightarrow C((0,1)), \\ u &\longmapsto u_{xx} + f(u(\cdot,t),\cdot,t). \end{aligned}$$

Mit $v'(t) = (u_t(\Delta x, t), u_t(2\Delta x, t), \ldots, u_t(1-\Delta x, t))$ und (2-1) ergibt sich ein System gewöhnlicher Differentialgleichungen

$$\begin{aligned} v'(t) &= -A^{\Delta x} v(t) + H^{\Delta x}(v(t)) + r^{\Delta x}(t), \\ &= \quad F_{\Delta x}(v(t), t), \quad 0 \le t \le T, \\ v(0) &= \quad v^0 \;=\; (u_0(\Delta x), u_0(2\Delta x), \ldots, u_0(1-\Delta x)) \end{aligned} \tag{2-3}$$

als Ersatz für das Anfangsrandwertproblem (2-1).

2.1 Definition. Das Gleichungssystem (2-3) heißt Semidiskretisierung zu (2-1).

Wir lösen die Anfangswertaufgabe (2-3) mit dem ϑ-Verfahren. Sei $\Delta t = \frac{T}{N} > 0$. Der Wert v^j approximiere $v(j\Delta t)$, $j = 0, \ldots, N$. Das Verfahren lautet dann

$$\begin{aligned} v^{j+1} &= v^j + \Delta t \left[\vartheta F_{\Delta x}(v^{j+1}, t_{j+1}) + (1-\vartheta) F_{\Delta x}(v^j, t_j)\right], \quad j = 0, \ldots, N-1, \\ v^0 &= (u_0(\Delta x), \ldots, u_0((M-1)\Delta x)). \end{aligned}$$

Die Spezialfälle sind dabei:

$\vartheta = 0$: Euler-Cauchy Verfahren,

$\vartheta = \frac{1}{2}$: Crank-Nicholson Verfahren,

$\vartheta = 1$: implizites Euler-Cauchy Verfahren.

Fehleranalyse der Differenzenverfahren

Sei $\Gamma_h = \{(i\Delta x, j\Delta t) \,|\, i = 1, \ldots, M-1, j = 0, \ldots, N\}$, $h = (\Delta x, \Delta t)$. Wir schreiben das Differenzenverfahren in der Form

$$T^h(u) = 0, \quad T^h : \mathbb{R}^{\Gamma_h} \to \mathbb{R}^{\Gamma_h}, \quad u = (u_1^0, \ldots, u_{M-1}^0, \ldots, u_1^N, \ldots, u_{M-1}^N),$$

wobei

$$(T^h(u))_i^j = \begin{cases} u_i^0 - u_0(x_i), & \begin{array}{l} j = 0 \\ i = 1, \ldots, M-1 \end{array} \\ \begin{array}{c} \frac{1}{\Delta t}\left(u_i^j - u_i^{j-1}\right) \\ -\vartheta\left[\frac{1}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f(u_i^j, x_i, t_j)\right] \\ -(1-\vartheta)\left[\frac{1}{\Delta x^2}\left(u_{i-1}^{j-1} - 2u_i^{j-1} + u_{i+1}^{j-1}\right) + f(u^{j-1}, x_i, t_{j-1})\right] \end{array} & \begin{array}{l} j = 1, \ldots, N \\ i = 1, \ldots, M-1 \end{array} \end{cases} \tag{2-4}$$

Seien $\overline{u}$ eine Lösung von (2-1) sowie $\overline{u}_h$ deren Restriktion auf das Gitter Γ_h.

2.2 Definition. $\|T^h(\overline{u}_h)\|$ heißt der Konsistenzfehler des Modells $T^h = 0$ bezüglich einer Norm $\|\cdot\|$.

Für das ϑ-Verfahren und $\|\cdot\| = \|\cdot\|_\infty$ gilt insbesondere

$$\|T^h(\overline{u}_h)\|_\infty = \begin{cases} O(\Delta t + \Delta x^2), & \vartheta \neq \frac{1}{2}, \\ O(\Delta t^2 + \Delta x^2), & \vartheta = \frac{1}{2} \end{cases}$$

an jeder hinreichend glatten Lösung von (2-1).

2.3 Definition. Das Modell $T^h = 0$ heißt stabil bezüglich einer Norm $\|\cdot\|$, falls es ein von h unabhängiges $C > 0$ derart gibt, dass

$$\|u - v\| \leq C \cdot \|T^h(u) - T^h(v)\|$$

für alle $u, v \in \mathbb{R}^{\Gamma_h}$ und $0 < h \leq h_0$ gilt.

Wir analysieren die Stabilität der linearen Wärmeleitungsgleichung, d.h. den Fall $f \equiv 0$ in (2-1), versehen mit den von t unabhägigen Randbedingungen $u(0,t) = \gamma_0$, $u(1,t) = \gamma_1$. Unter diesen vereinfachten Annahmen lautet das ϑ-Verfahren

$$\begin{pmatrix} I & 0 & 0 & \dots & 0 & 0 & 0 \\ -B & A & 0 & \dots & 0 & 0 & 0 \\ 0 & -B & A & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A & 0 & 0 \\ 0 & 0 & 0 & \dots & -B & A & 0 \\ 0 & 0 & 0 & \dots & 0 & -B & A \end{pmatrix} \begin{pmatrix} v^0 \\ v^1 \\ v^2 \\ \vdots \\ v^{N-2} \\ v^{N-1} \\ v^N \end{pmatrix} - \begin{pmatrix} r^0 \\ r^1 \\ r^1 \\ \vdots \\ r^1 \\ r^1 \\ r^1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{2-5}$$

mit

$$\begin{aligned} r^0 &= \begin{pmatrix} u_0(x_1) \\ u_0(x_2) \\ \vdots \\ u_0(x_{M-2}) \\ u_0(x_{M-1}) \end{pmatrix}, \quad r^1 = \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0 \\ 0 \\ \vdots \\ 0 \\ \gamma_1 \end{pmatrix}, \\ A &= \frac{1}{\Delta t} I + \vartheta C, \quad B = \frac{1}{\Delta t} I - (1-\vartheta) C, \\ C &= \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix}. \end{aligned}$$

Für die Stabilität von Blockmatrizen gilt allgemein:

2.4 Lemma. *Es sei auf $\mathbb{R}^m$ eine Norm $\|\cdot\|_*$ gegeben. Die Matrizen $A(\Delta t)$, $B(\Delta t)$ mögen von $\Delta t \in (0,T)$ abhängen. $A(\Delta t)$ sei invertierbar und erfülle*

$$\|A(\Delta t)^{-1}\|_* \leq C_1 \Delta t \text{ für alle } t \in (0,T).$$

Ferner existiere ein $C_2 > 0$ mit

$$\|(A^{-1}B)^n(\Delta t)\|_* \leq C_2 \text{ für alle } n \in N \text{ mit } 0 \leq n\Delta t \leq T.$$

Dann gilt für die $(n+1)$-blockige Matrix

$$H(\Delta t) = \begin{pmatrix} I & 0 & 0 & \dots & 0 & 0 & 0 \\ -B & A & 0 & \dots & 0 & 0 & 0 \\ 0 & -B & A & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A & 0 & 0 \\ 0 & 0 & 0 & \dots & -B & A & 0 \\ 0 & 0 & 0 & \dots & 0 & -B & A \end{pmatrix}$$

mit $0 \leq n\Delta t \leq T$ die Stabilitätsungleichung

$$\|v\|_{*,\infty} \leq C_2(1 + C_1 T) \cdot \|H(\Delta t)v\|_{*,\infty}, \quad \forall v \in \mathbb{R}^{m(n+1)},$$

wobei

$$\|v\|_{*,\infty} = \|(v^0, v^1, \dots, v^N)\|_{*,\infty} = \max\{\|v^i\|_* \,|\, i = 0, \dots, n\}.$$

Daraus lässt sich die Gültigkeit des nachstehenden Lemmas folgern.

2.5 Lemma. *Unter der Bedingung*

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)}$$

ist das ϑ-Verfahren für die Wärmeleitungsgleichung bezüglich der Norm $\|v\|_{\infty,\infty} = \|v\|_\infty$ auf $\mathbb{R}^{\Gamma_h}$ stabil.

Stabilität der Wärmeleitungsgleichung bezüglich $\|\cdot\|_{2,*}$ auf $\mathbb{R}^{\Gamma_h}$

Setzt man

$$\|v\|_* = \|v\|_2 = \left(\Delta x \sum_{i=1}^{M-1} v_i^2\right)^{1/2}, v \in \mathbb{R}^{M-1},$$

so stimmt $\|\cdot\|_2$ bis auf den Fehler der Ordnung $\sqrt{\Delta x}$ mit der euklidischen Norm überein, falls v Riemann-integrierbar ist. Im Grenzwert $\Delta x \to 0$ finden wir

$$\Delta x \sum_{i=1}^{M-1} v_i^2 \xrightarrow[M(\Delta x)\to\infty]{\Delta x \to 0} \int_0^1 v^2(x)\,\mathrm{d}x = \|v\|^2_{L^2(0,1)}$$

aufgrund der Konvergenz der Riemannschen Summe gegen das Riemannsche Integral.

Ferner gilt für eine symmetrische Matrix $A \in \mathbb{R}^{M-1,M-1}$:

$$\begin{aligned} \|A\|_2 &= \sup\{\|Av\|_2 \mid v \in \mathbb{R}^{M-1}, \|v\|_2 = 1\} \\ &= \max\{|\lambda| \mid \lambda \in \sigma(A) \subset \mathbb{R}\}. \end{aligned}$$

2.6 Lemma. *Es sei*

$$C = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{M-1,M-1}, \quad \Delta x = \frac{1}{M} > 0.$$

C hat die Eigenwerte

$$\lambda_k = \frac{2}{\Delta x^2}\left(1 - \cos\left(\frac{k\pi}{M}\right)\right), \quad k = 1, \dots, M-1$$

mit den Eigenvektoren $v^k = (v_1^k, \dots, v_{M-1}^k) \in \mathbb{R}^{M-1}$,

$$v_i^k = \sin\left(\frac{ik\pi}{M}\right), \quad i = 1, \dots, M-1, \quad k = 1, \dots, M-1.$$

Beweis: Man rechne nach! □

Wir kehren nun zur Wärmeleitungsgleichung zurück und berechnen $\|A^{-1}(\Delta t)\|_2$ und $\|(A^{-1}B)(\Delta t)\|_2$ für

$$A(\Delta t) = \frac{1}{\Delta t} I + \vartheta C, \quad B(\Delta t) = \frac{1}{\Delta t} I - (1-\vartheta) C.$$

Gemäß Lemma 2.6 hat C die Eigenwerte $\lambda_k = \frac{2}{\Delta x^2}\left(1 - \cos\left(\frac{k\pi}{M}\right)\right)$, $k = 1, \dots, M-1$. Für λ_k gilt die Abschätzung

$$0 < \lambda_k < \frac{4}{\Delta x^2}, \quad k = 1, \dots, M-1.$$

Die Matrizen $A(\Delta t)$, $B(\Delta t)$, $A^{-1}(\Delta t)$ und C sind symmetrisch. Ferner gilt $(BA)(\Delta t) = (AB)(\Delta t)$ nach Definition von $A(\Delta t)$ und $B(\Delta t)$. Also folgt

$$(A^{-1}B)^T(\Delta t) = (BA^{-1})^T(\Delta t) = ((A^{-1})^T B^T)(\Delta t) = (A^{-1}B)(\Delta t),$$

d.h. $(A^{-1}B)(\Delta t)$ ist symmetrisch.

$A(\Delta t)$ hat die Eigenwerte $\frac{1}{\Delta t} + \vartheta\lambda_k$. Somit sind

$$\frac{1}{\frac{1}{\Delta t} + \vartheta\lambda_k} = \frac{\Delta t}{1 + \Delta t\vartheta\lambda_k}, \qquad k = 1, \ldots, M-1$$

die Eigenwerte von $A^{-1}(\Delta t)$, und wir finden

$$\|A^{-1}(\Delta t)\|_2 = \max\left\{\left|\frac{\Delta t}{1 + \Delta t\vartheta\lambda_k}\right| \mid k = 1, \ldots, M-1\right\} \leq \Delta t,$$

d.h. $C_1 = 1$ in Lemma 2.4.

Die Eigenwerte von $(A^{-1}B)(\Delta t) = \left(\frac{1}{\Delta t}I + \vartheta C\right)^{-1}\left(\frac{1}{\Delta t}I - (1-\vartheta)C\right)$ sind

$$\mu_k = \frac{\frac{1}{\Delta t} - (1-\vartheta)\lambda_k}{\frac{1}{\Delta t} + \vartheta\lambda_k} \leq 1, \quad k = 1, \ldots, M-1.$$

Man beachte dazu, dass $A(\Delta t)$, $B(\Delta t)$, $A^{-1}(\Delta t)$, $(A^{-1}B)(\Delta t)$ dieselbe Basis $v^1, \ldots, v^{M-1}$ aus Eigenvektoren wie C besitzen. Wir finden also

$$\|(A^{-1}B)(\Delta t)\|_2 \leq 1, \text{ falls } \mu_k \geq -1, \quad k = 1, \ldots, M-1.$$

Dies ist äquivalent zu

$$\begin{aligned}\mu_k \geq -1 \quad &\Leftrightarrow \quad \frac{1}{\Delta t} - (1-\vartheta)\lambda_k \geq -\frac{1}{\Delta t} - \vartheta\lambda_k \\ &\Leftrightarrow \quad \frac{2}{\Delta t} \geq (1-2\vartheta)\lambda_k, \quad k = 1, \ldots, M-1.\end{aligned}$$

Für $\vartheta = \frac{1}{2}$ ist dies stets erfüllt, während sich für $\vartheta < \frac{1}{2}$ die Bedingung

$$\lambda_k \leq \frac{2}{\Delta t(1-2\vartheta)}, \quad k = 1, \ldots, M-1$$

ergibt. Dies ist insbesondere abgesichert, wenn

$$\lambda_k \leq \frac{4}{\Delta x^2} \leq \frac{2}{\Delta t(1-2\vartheta)} \Leftrightarrow \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-2\vartheta)}.$$

gilt. Das Lemma 2.4 ist also mit $C_1 = C_2 = 1$ anwendbar.

2.7 Satz. *Unter der Bedingung*

$$\frac{\Delta t}{\Delta x^2} \leq \begin{cases} \infty, & \text{für } \vartheta \geq \frac{1}{2} \\ \frac{1}{2(1-2\vartheta)}, & \text{für } \vartheta < \frac{1}{2} \end{cases}$$

ist das ϑ-Verfahren für die Wärmeleitungsgleichung bezüglich der Norm

$$\|u\|_{2,\infty} = \max \left\{ \left(\Delta x \sum_{i=1}^{M-1} (u_i^j)^2 \right)^{1/2} \mid j = 0, \ldots, N \right\}, \quad u \in \mathbb{R}^{\Gamma_h}, \; h = (\Delta x, \Delta t)$$

stabil. An jeder Lösung $\overline{u}$ der linearen Wärmeleitungsgleichung mit

$$\frac{\partial^\nu \overline{u}}{\partial t^\nu} \in C([0,1] \times [0,T]), \nu = 1, 2, \quad \frac{\partial^\nu \overline{u}}{\partial x^\nu} \in C([0,1] \times [0,T]), \nu = 1, 2, 3, 4$$

liegt die Konvergenz gemäß

$$\|\overline{u}_h - u^h\|_{2,\infty} = O(\Delta t + \Delta x^2), \quad T^h(u^h) = 0$$

vor. Für das Crank-Nicholson Verfahren gilt sogar

$$\|\overline{u}_h - \overline{u}^h\|_{2,\infty} = O(\Delta t^2 + \Delta x^2),$$

falls zusätzlich $\frac{\partial^3 \overline{u}}{\partial t^3} \in C([0,1] \times [0,T])$.

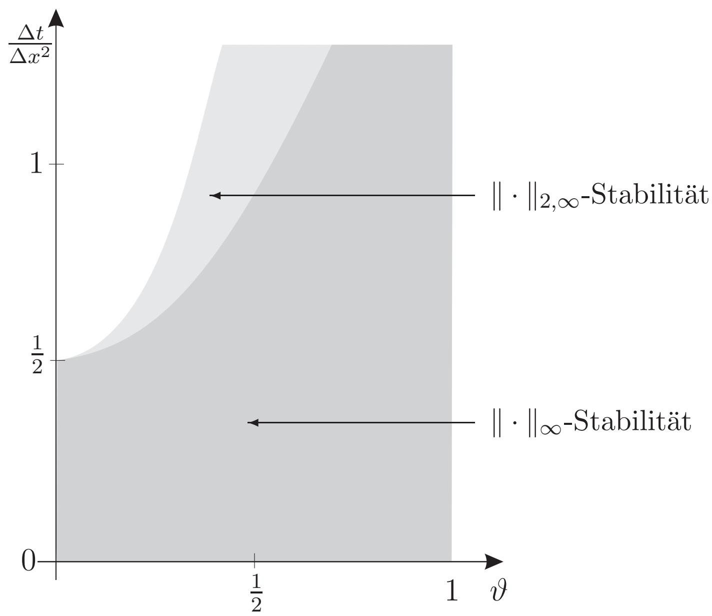


Abbildung 10: Stabilitätsbereiche

b) Finite Elemente Methoden für parabolische Differentialgleichungen

Vorgelegt sei die parabolische Anfangsrandwertaufgabe

$$\frac{\partial u}{\partial t} - \Delta u \;=\; f \text{ in } (0,T) \times \Omega, \tag{2-6}$$

$$
\begin{aligned}
u &= 0 \text{ auf } (0,T)\times\partial\Omega,\\
u(\cdot,0) &= u_0 \text{ in } \Omega,
\end{aligned}
$$

wobei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet sei. Wir wenden auf (2-6) die Theorie schwacher Lösungen an.

Es sei $t \in (0,T)$ beliebig, aber fest. Ferner sei $v \in C_0^\infty(\Omega)$ willkürlich gewählt. Dann folgt

$$\int_\Omega \frac{\partial u}{\partial t}(t,x)v(x) - \Delta u(t,x)v(x)\,\mathrm{d}x = \int_\Omega f(t,x)v(x)\,\mathrm{d}x,$$

woraus sich mit der partiellen Integration

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(t,x)v(x)\,\mathrm{d}x + \int_\Omega \nabla u(t,x)\cdot\nabla v(x)\,\mathrm{d}x + \underbrace{\int_{\partial\Omega}\frac{\partial u(t,x)}{\partial n}v(x)\,\mathrm{d}S}_{=0} = \int_\Omega f(t,x)v(x)\,\mathrm{d}x$$

ergibt, wobei u eine klassische Lösung von (2-6) sei, d.h. $u \in C(\overline{\Omega}\times[0,T])$, $u_t, u_{xx} \in C(\Omega\times(0,T))$. Also gilt

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(t,x)v(x)\,\mathrm{d}x + \int_\Omega \nabla u(t,x)\cdot\nabla v(x)\,\mathrm{d}x = \int_\Omega f(t,x)v(x)\,\mathrm{d}x$$

für alle $v \in C_0^\infty$ und $t \in (0,T)$.

Das funktionalanalytische Lösen der Aufgabe (2-6) erfordert eine besondere Behandlung der Variablen t und x:

a) Für jedes feste $t \in (0,T)$ handelt es sich bei der Abbildung

$$x \mapsto u(t,x), \text{ d.h. } u(t,\cdot)$$

um ein Element eines Sobolev-Raumes V, also $u(t,\cdot) \in V$ für $t \in (0,T)$. Hierfür schreibt man kurz $u(t) \in V$. Naheliegend ist die Wahl des Raumes $V = H_0^1(\Omega)$.

b) Variiert man nun $t \in (0,T)$, so entsteht eine Funktion $t \mapsto u(t)$ mit Werten im Banachraum V.

Für unsere weiteren Schritte benötigen wir den Begriff eines Bochner-Integrals.

Es seien (I,Σ,μ) ein Maßraum und H ein Banachraum. Das Bochner-Integral wird auf eine ähnliche Weise definiert wie ein Lebesgue-Integral.

2.8 Definition. Es sei s eine Stufenfunktion der Gestalt $s(x) = \sum_{i=1}^n \chi_{E_i}(x)b_i$, wobei $\{E_1,\dots,E_n\} \subset \Sigma$ paarweise disjunkt sowie $\{h_1,\dots,h_n\} \subset H$ paarweise verschieden sind und χ_E die charakteristische Funktion einer Menge $E \subset I$ bezeichnet.

Das Bochner-Integral von s ist dann durch

$$\int_I s \,\mathrm{d}\mu = \sum_{i=1}^{n} \mu(E_i) h_i$$

gegeben.

2.9 Definition. Eine messbare Funktion $u : I \to H$ heißt Bochner-integrierbar, wenn es eine Folge $\{s_n\}_n$ von Stufenfunktionen gibt mit

$$\lim_{n\to\infty} \int_I \|f - s_n\|_H \,\mathrm{d}\mu = 0.$$

In diesem Fall wird das Bochner-Integral vermöge

$$\int_I f \,\mathrm{d}x = \lim_{n\to\infty} \int_H s_n \,\mathrm{d}\mu$$

definiert.

Nun können wir folgende funktionalanalytische Räume definieren.

2.10 Definition. Es seien (I, Σ, μ) ein Maßraum, H ein Banachraum und $1 \leq p \leq \infty$. Der Bochner-Raum $L^p(T, X)$ ist definiert als der Kolmoroff-Quotient (bezüglich der Gleichheit fast überall) des Raumes der messbaren Funktionen $u : I \to H$, deren zugehörige Norm $\|\cdot\|_{L^p(I,H)}$ endlich ist. Dabei ist

$$\begin{aligned} \|u\|_{L^p(I,H)} &:= \left(\int_T \|u(t)\|_H^p \,\mathrm{d}\mu(t) \right)^{1/p}, \quad 1 \leq p < \infty \\ \|u\|_{L^\infty(I,H)} &:= \operatorname{ess\,sup}_{t\in I} \|u(t)\|_H. \end{aligned}$$

2.11 Definition. Ist $p = 2$, so lässt sich $L^2(I, H)$ als ein Hilbertraum mit dem Skalarprodukt

$$\langle u, v \rangle_{L^2(I,H)} := \int_I \langle u(t), v(t) \rangle_H \,\mathrm{d}t$$

auffassen.

2.12 Bemerkung. Ist $I = (0, T)$ ein offenes Intervall, Σ die Borel-σ-Algebra auf I sowie μ das dazu gehörige Lebesgue-Maß auf Σ, so gilt

$$L^p((0, T), H) = \{u : (0, T) \to H \,|\, \|u\|_{L^p(T,H)} < \infty\},$$

wobei

$$\|u\|_{L^p(T,H)} := \|\|u(\cdot)\|_H\|_{L^p((0,T))}.$$

2.13 Definition. Sei $I \subset \mathbb{R}$ ein Intervall. Zu einem gegebenen Banachraum H wird der Raum der stetigen Funktionen definiert durch

$$C^0(I,H) = \{u : I \to H \,|\, u \text{ stetig in der } \|\cdot\|_H\text{-Norm}, \|u\|_{C^0(I,H)} < \infty\}$$

mit

$$\|g\|_{C^0(I,H)} := \sup_{t\in I} \|u(t)\|_H.$$

Es seien nun $f(t), u_0 \in L^2(\Omega)$, $v, w \in H_0^1(\Omega)$. Wir definieren

$$\begin{aligned} \langle f(t), z\rangle_0 &:= \int_\Omega f(t,x)z(x)\,\mathrm{d}x, \\ a(v,w) &:= \int_\Omega \nabla v \cdot \nabla w\,\mathrm{d}x \end{aligned}$$

für alle $v, w \in H_0^1(\Omega)$. Damit finden wir unter Beachtung der Dichtheit der Einbettung $C_0^\infty(\Omega) \subset H_0^1(\Omega)$ sowie der Stetigkeit des entsprechenden Funktionals auf $H_0^1(\Omega)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), v\rangle_0 + a(u(t), v) = \langle f(t), v\rangle_0$$

für alle $v \in V$ und $t \in (0,T)$.

Lösungstheorie für parabolische Variationsgleichungen

Es seien $V = H_0^m(\Omega)$, $H = L^2(\Omega)$, $V' = \{f : V \to \mathbb{R} \,|\, f \text{ linear und stetig}\} = H^{-m}(\Omega)$ der Dualraum zu V. Da die Einbettungen $V \subset H \subset V'$ jeweils stetig und dicht sind, ist (V, H, V') ein Gelfand-Dreier.

2.14 Definition. Sei $u \in L^2((0,T), V)$. Dann heißt ein Element $w \in L^2((0,T), V')$ eine verallgemeinerte Ableitung, falls

$$\int_0^T \langle u(t), v\rangle_0 \varphi'(t)\,\mathrm{d}t = -\int_0^T \langle w(t), v\rangle_{V'\times V}\varphi(t)\,\mathrm{d}t$$

für alle $v \in V$ und $\varphi \in C_0^\infty((0,T))$ gilt. Man schreibt dazu wieder $u' = w$.

Somit erhalten wir

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), v\rangle_0 = \langle u'(t), v\rangle_{V'\times V}$$

für alle $v \in V$ und fast alle $t \in (0,T)$.

Dies motiviert folgende Definition.

2.15 Definition. Eine schwache Lösung der parabolischen Anfangsrandwertaufgabe (2-6) ist ein Element $u \in L^2((0,T), V)$, das eine schwache Ableitung $\frac{\mathrm{d}u}{\mathrm{d}t} = u' \in$

$L^2((0,T),V')$ besitzt, und die Anfangsbedingung $u(0) = u_0$ sowie die Differentialgleichung

$$\langle u'(t), v\rangle_{V'\times V} + a(u(t),v) = \langle f(t), v\rangle_0 \quad \text{für alle } v\in V \text{ und fast alle } t\in(0,T)$$

erfüllt.

Man kann nun wieder das Skalarprodukt $\langle\cdot,\cdot\rangle_0$ auf $L^2(\Omega)$ zu einer stetigen Bilinearform $H_0^m \times H^{-m} \to \mathbb{R}$ fortsetzen und $\langle u'(t), v\rangle_0$ statt $\langle u'(t), v\rangle_{V'\times V}$ schreiben.

2.16 Bemerkung. Die Bedingungen $u \in L^2((0,T),V)$ und $u' \in L^2((0,T),V')$ implizieren $u \in C([0,T],H)$. Dieses Ergebnis ist als ein Interpolationssatz bekannt.

Im Folgenden sei $a : V\times V \to \mathbb{R}$ eine stetige und auf V-koerzive Bilinearform, d.h.

$$\begin{aligned} |a(u,v)| &\le C\cdot\|u\|_V \times \|v\|_V, \quad \forall u,v\in V,\\ a(v,v) &\ge \alpha\|v\|_V^2, \quad \forall v\in V. \end{aligned}$$

Es gilt:

2.17 Lemma. *Es sei a eine V-elliptische, stetige Bilinearform. Sind $u_0\in H$ und $f\in C([0,T],H)$, dann gilt für eine Lösung u des Problems*

$$\begin{aligned} \langle u'(t), v\rangle_0 + a(u(t),v) &= \langle f(t), v\rangle_0, \quad \forall v\in V, t\in(0,T), \\ u(0) &= u_0 \end{aligned} \tag{2-7}$$

die Abschätzung

$$\|u(t)\|_0 \le \|u_0\|_0 \exp(-\alpha t) + \int_0^t \|f(s)\|_0 \exp(-\alpha(t-s))\,\mathrm{d}s$$

für alle $t\in(0,T)$.

Beweis: Wendet man (2-7) mit $v = u(t)$ an, so liefert die V-Elliptizität von a

$$\langle u'(t), u(t)\rangle_0 + \alpha\|u(t)\|_V^2 \le \langle f(t), u(t)\rangle_0.$$

Ferner folgt mit

$$\langle u'(t), u(t)\rangle_0 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle u(t), u(t)\rangle_0 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0^2 = \|u(t)\|_0 \frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0$$

sofort

$$\|u(t)\|_0 \frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0 + \alpha\|u(t)\|_V^2 \le \langle f(t), u(t)\rangle_0 \le \|f(t)\|_0\|u(t)\|_0.$$

Mit $\|u(t)\|_0 \le \|u(t)\|_V$ ergibt sich durch Division mit $\|u(t)\|_0$ dann

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0 + \alpha\|u(t)\|_0 \le \|f(t)\|_0, \quad t\in(0,T).$$

Man findet nun

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\exp(\alpha t)\cdot\|u(t)\|_0) &= \alpha\exp(\alpha t)\|u(t)\|_0+\exp(\alpha t)\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0\\ &= \exp(\alpha t)\left[\alpha\|u(t)\|_0+\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|_0\right]\\ &\le \exp(\alpha t)\|f(t)\|_0,\quad 0<t<T.\end{aligned}$$

Integration über $(0,T)$ liefert

$$\exp(\alpha t)\|u(t)\|_0-\|u(0)\|_0\le\int_0^t\exp(\alpha s)\|f(s)\|\,\mathrm{d}s,$$

was mit

$$\|u(t)\|_0-\exp(-\alpha t)\|u_0\|_0\le\int_0^t\|f(s)\|\exp(-\alpha(t-s))\,\mathrm{d}s$$

äquivalent ist. Wir bekommen schließlich

$$\|u(t)\|_0\le\|u_0\|_0\exp(-\alpha t)+\int_0^t\|f(s)\|_0\exp(-\alpha(t-s))\,\mathrm{d}s$$

für alle $t\in(0,T)$. □

2.18 Korollar. *Die Lösung u von (2-7) ist unter Voraussetzungen des Lemmas 2.17 eindeutig.*

Beweis: Sind u_1,u_2 zwei Lösungen von (2-7), so löst $\hat{u}=u_1-u_2$ die Gleichung (2-7) mit $f\equiv 0$ und $u_0=0$. Lemma 2.17 liefert dann $\hat{u}=0$, also $u_1=u_2$. □

Semidiskretisierung im Raum

Wir wenden uns hier der numerischen Behandlung der folgenden variationellen Anfangsrandwertaufgabe zu: Gesucht ist ein $u\in L^2((0,T),V)$ mit $u'\in L^2((0,T),V')$ und

$$\begin{aligned}\langle u'(t),v\rangle_0+a(u(t),v) &= \langle f(t),v\rangle_0,\quad \forall v\in V, t\in(0,T),\\ u(0) &= u_0\in H.\end{aligned}\tag{2-8}$$

Wir wählen einen endlich dimensionalen Teilraum $V_h\subset V$ und bezeichnen mit u_{0h} eine Approximation von u_0 in V_h. Das Galerkin-Problem besteht darin, ein $u_h\in L^2((0,T),V_h)$ mit $u'\in L^2((0,T),V')$ mit

$$\begin{aligned}\langle u_h'(t),v\rangle_0+a(u_h(t),v) &= \langle f(t),v\rangle_0,\quad \forall v\in V_h,\quad t\in(0,T),\\ u_h(0) &= u_{0h}.\end{aligned}\tag{2-9}$$

zu finden.

Sei nun $V_h = \operatorname{span}\{v_1, \dots, v_M\}$, $u_h(t) = \sum_{i=1}^M c_i(t) v_i$ und $u_{0h} = \sum_{i=1}^M c_{i0} v_i$. Einsetzen in (2-8) liefert

$$\Big\langle \sum_{i=1}^M c_i'(t) v_i, v \Big\rangle_0 + a\Big(\sum_{i=1}^M c_i(t) v_i, v \Big) = \langle f(t), v\rangle_0, \quad \forall v \in V_h, \quad t \in (0,T),$$

Es genügt, dies auf der Basis $\{v_1, \dots, v_M\}$ zu sichern, d.h.

$$\sum_{i=1}^M c_i'(t) \langle v_i, v_j\rangle_0 + a\Big(\sum_{i=1}^M c_i(t) v_i, v_j \Big) = \langle f(t), v_j\rangle_0, \quad j = 1, \dots, M. \tag{2-10}$$

Das Galerkin-Verfahren (2-9) zu unserer Anfangsrandwertaufgabe (2-6) ist genau dann eindeutig lösbar, wenn es Funktionen $c_i \in C^1([0,T], \mathbb{R})$, $i = 1, \dots, M$ gibt, welche (2-10) erfüllen.

Mit der Steifigkeitsmatrix

$$A = (a_{ij})_{1 \le i,j \le M}, \quad a_{ij} = a(v_i, v_j)$$

und der Massenmatrix

$$B = (b_{ij})_{1 \le i,j \le M}, \quad b_{ij} = \langle v_i, v_j \rangle$$

sowie den Vektoren

$$\begin{aligned} r &= (r_i)_{1 \le i \le M}, \quad r_i(t) = \langle f(t), v_i \rangle_0, \\ c_0 &= (c_{1,0}, c_{2,0}, \dots, c_{M,0}), \\ c(t) &= (c_1(t), c_2(t), \dots, c_M(t)) \end{aligned}$$

erhalten wir das System

$$\begin{aligned} Bc'(t) + Ac(t) &= r(t), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-11}$$

2.19 Definition. Das Problem (2-11) heißt semidiskrete Differentialgleichung des Anfangsrandwertproblems.

Ferner ist $B \in \mathbb{R}^{M,M}$ eine symmetrische, positiv definite Matrix. Die Abbildung $r : (0,T) \to \mathbb{R}^M$ ist stetig, da $f \in C([0,T], H)$ gilt und $\langle \cdot, \cdot \rangle_0$ stetig ist. Somit ist (2-11) äquivalent zur Anfangsrandwertaufgabe

$$\begin{aligned} c'(t) &= B^{-1}(r(t) - Ac(t)), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-12}$$

Das Problem (2-12) ist eine lineare inhomogene Anfangsrandwertaufgabe mit stetiger Inhomogenität $B^{-1}r(\cdot)$. Diese besitzt nach dem Existenz- und Eindeutigkeitssatz eine Lösung $c : [0, T] \to \mathbb{R}^M$. Somit ist (2-12) eindeutig lösbar und damit auch (2-8).

Semidiskrete Fehlerabschätzung

Im Folgenden soll der Term $u(t) - u_h(t)$ abgeschätzt werden.

2.20 Definition. Die elliptische Projektion oder die Ritz-Projektion $R_h : V \to V_h$ ist für eine V-elliptische, stetige Bilinearform $a : V \times V \to \mathbb{R}$ durch

$$v \mapsto R_h(v) \Longleftrightarrow (a(R_h v - v, w) = 0, \quad \forall w \in V_h)$$

definiert

2.21 Bemerkung. Eine Ritz-Projektion besitzt die folgenden Eigenschaften:

i) R_h ist als Abbildung wohldefiniert.

ii) $R_h : V \to V_h$ ist linear und stetig,

iii) R_h liefert die quasioptimale Approximation, d.h.

$$\|v - R_h v\|_V \leq \frac{C}{\alpha} \inf_{w \in V_h} \|v - w\|.$$

Beweis: Siehe Übung. □

2.22 Satz. *Es sei a eine V-elliptische, stetige Bilinearform mit Konstanten C bzw. α. Ferner gelte $f \in C([0, T], H)$, $u_0 \in V$ sowie $u_{0h} \in V_h$. Dann folgt die Abschätzung*

$$\begin{aligned} \|u_h(t) - u(t)\|_0 \quad &\leq \quad \|u_{0h} - R_h u_0\| \exp(-\alpha t) + \|(I - R_h)u(t)\|_0 \\ &\quad + \int_0^t \|(I - R_h)u'(s)\|_0 \exp(-\alpha(t - s))\, ds. \end{aligned}$$

Beweis: Es gilt

$$u_h(t) - u(t) = \underbrace{u_h(t) - R_h u(t)}_{=:\theta(t)} + \underbrace{R_h u(t) - u(t)}_{\rho(t)} = \theta(t) + \rho(t).$$

Ferner sei $w \in V_h \subset V$. Dann gilt nach Definition von R_h

$$\langle u'(t), w\rangle_0 + a(u(t), w) = \langle u'(t), w\rangle_0 + a(R_h u(t), w) = \langle f(t), w\rangle_0,$$

wobei die Gültigkeit von $a(R_h v, w) = a(v, w)$ für alle $w \in V_h$ zu beachten ist. Nun folgt

$$\langle u_h'(t), w\rangle_0 + a(u_h(t), w) = \langle f(t), w\rangle_0.$$

Substitution liefert

$$\begin{aligned} 0 &= \langle u_h'(t), w\rangle_0 - \langle u'(t), w\rangle_0 + a(\underbrace{u_h(t) - R_h u(t)}_{=\theta(t)}, w) \\ &= \langle \underbrace{u_h'(t) - u'(t)}_{=\theta'(t)+\rho'(t)}, w\rangle_0 + a(\theta(t), w), \end{aligned}$$

was mit

$$\langle \theta'(t), w\rangle_0 + a(\theta(t), w) = -\langle \rho'(t), w\rangle_0, \quad 0 < t < T$$

gleichbedeutend ist. Wendet man nun darauf das Lemma 2.17 an, so ergibt sich

$$\|\theta(t)\|_0 \le \|\theta(0)\|_0 \exp(-\alpha t) + \int_0^t \|\rho'(s)\|_0 \exp(-\alpha(t-s))\,\mathrm{d}s. \tag{2-13}$$

Weiter gilt $\rho(t) = (R_h - I)u(t)$, $0 < t < T$. Somit folgt

$$\rho'(t) = (R_h - I)u'(t), \quad 0 < t < T.$$

Die Abschätzung (2-13) impliziert nun

$$\begin{aligned} \|u_h(t) - u(t)\|_0 \le\ & \|\underbrace{u_h(0)}_{=u_{0h}} - R_h \underbrace{u(0)}_{=u_0}\|_0 \exp(-\alpha t) \\ & + \int_0^t \|(I - R_h)u'(s)\|_0 \exp(-\alpha(t-s))\,\mathrm{d}s, \end{aligned}$$

woraus sich mit der Dreiecksungleichung die Behauptung ergibt. □

Die Abschätzung des Fehlerterms $\|u_h(t) - u(t)\|_0$ im Satz 2.22 erfolgt durch:

- den Anfangsfehler, welcher in der Zeit exponentiell abfällt und nur dann auftritt, wenn u_{0h} nicht mit $R_h u_0$ identisch ist,
- den Projektionsfehler in der Norm von H der exakten Lösung u,
- den durch die Integration über $(0, T)$ mit dem Faktor $\exp(-\alpha(t-s))$ gewichteten Projektionsfehler von $u'(t)$.

2.23 Korollar. *Gilt für die elliptische Projektion eine Fehlerabschätzung der Form*

$$\|(I - R_h)u(t)\|_0 \le C \cdot h^2 \|u(t)\|_2$$

und approximiert überdies u_{0h} die elliptische Projektion $R_h u_0$ ebenfalls in dieser asymptotischen Qualität, so erhält man die Fehlerabschätzung

$$\|u_h(t) - u(t)\|_0 \le \hat{C} h^2 \|u(t)\|_2,$$

falls $u(t) \in L^2((0, T), H^2(\Omega))$.

Volldiskretes Problem

Sei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit polygonalem Rand und $\Omega_{T_h} = \{e_1, \ldots, e_m\}$ sei eine Triangulierung. Sei $V_h = \operatorname{span}\{v_1, \ldots, v_M\} \subset H_0^1(\Omega)$. Typischerweise wählt man

$$V_h = V_{T_h}^{(1)} = \{v \in C(\Omega) \mid \ v|_{e_k} \in P^1(e_k) \text{ und } v = 0 \text{ auf } \partial\Omega_{T_h}\} \tag{2-14}$$

Diese Wahl wird oft als lineare Lagrangesche Finite Elemente bezeichnet. Mit dem Ansatz

$$u_h(t) = \sum_{i=1}^{M} c_i(t) v_i, \quad u_{0h} = \sum_{i=1}^{M} c_{i0} v_i$$

erhält man das Differentialgleichungssystem

$$\begin{aligned} Bc'(t) + Ac(t) &= r(t), \quad 0 < t < T, \\ c(0) &= c_0 \end{aligned} \tag{2-15}$$

mit

$$\begin{aligned} A &= (a_{ij})_{1 \le i,j \le M}, \quad a_{ij} = a(v_i, v_j), \\ B &= (b_{ij})_{1 \le i,j \le M}, \quad b_{ij} = \langle v_i, v_j \rangle, \\ r(t) &= (r_i)_{1 \le i \le M}, \quad r_i(t) = \langle f(t), v_i \rangle_0, \\ c_0 &= (c_{i0})_{1 \le i \le M}. \end{aligned}$$

Im Sinne von Satz 2.22 erscheint nun die Wahl $u_{0h} = R_h u_0$ optimal mit dem Ritz-Projektor $R_h : V \to V_h$ definiert durch

$$v \mapsto R_h v \Longleftrightarrow a(R_h v - v, w) = 0, \quad \forall w \in V_h.$$

Die Anfangsrandwertaufgabe (2-15) kann dann wieder mit dem ϑ-Verfahren diskretisiert werden.

Sei $\Delta t = \frac{T}{N} > 0$, und sei $c^j \in \mathbb{R}^M$ die Approximation für $c(t_j)$, $t_j = j\Delta t$ und r^j bezeichne $r(t_j)$, $j = 0, \ldots, N$. Dann gilt

$$\begin{aligned} B\frac{c^{j+1} - c^j}{\Delta t} &= (1-\vartheta)(-Ac^j + r^j) + \vartheta(-Ac^{j+1} + r^{j+1}), \quad j = 0, \ldots, N, \\ c^0 &= c_0 \end{aligned}$$

mit $R_h u(0) = u_{0h} = \sum_{i=1}^{M} c_{i0} v_i$. Insgesamt erhält man die Approximation

$$u_h^j = \sum_{k=1}^{M} c_k^j \cdot v_k \in V_h$$

für $u_h(t_j)$ und Damit $u(t_j)$. Es lässt sich zeigen:

$$\max\{\|u(t_j) - u_h^j\| \mid j = 0, \dots, N\} \le C(u)(h^2 + \Delta t), \quad \frac{1}{2} < \vartheta \le 1$$

bzw.

$$\max\{\|u(t_j) - u_h^j\| \mid j = 0, \dots, N\} \le C(u)(h^2 + \Delta t^2), \quad \vartheta = \frac{1}{2},$$

falls die Triangulierungen $\{\Omega_{T_h}\}_{0<h<h_0}$ regulär sind und $u \in C^2([0,T], H^2(\Omega) \cap V)$ bzw. $u \in C^3([0,T], H^2(\Omega) \cap V)$.

c) Numerik gewöhnlicher Differentialgleichungen

Wir betrachten eine allgemeine Anfangsrandwertaufgabe

$$\begin{aligned} u'(t) &= f(t, u(t)) \text{ für } t \in [t_0, t_e], \\ u(t_0) &= \alpha \end{aligned} \tag{2-16}$$

für $f \in C^1([t_0, t_e] \times \mathbb{R}^N, \mathbb{R}^N)$. Wir nehmen an, dass (2-16) eine Lösung $\overline{u}(t)$ für $t \in [t_0, t_e]$ besitzt.

Zur numerischen Berechnung der Lösung gehen wir nun wie folgt vor. Wir wählen eine Schrittweite $h = \frac{t_e - t_0}{\sigma(h)} > 0$ und das Gitter

$$\Omega_h = \{t_j = t_0 + jh \mid j = 0, \dots, \sigma(h)\}.$$

Ferner bezeichne u^j die numerische Approximation für $\overline{u}(t_j)$. Ein Einschrittverfahren zu (2-16) hat die allgemeine Form

$$\begin{aligned} \frac{1}{h}(u_{m+1} - u_m) &= V(h, t_m, u_m), \quad m = 0, \dots, \sigma(h) - 1, \\ u^0 &= \alpha. \end{aligned} \tag{2-17}$$

2.24 Definition. Die Funktion $V : (0, h_0] \times [t_0, t_e) \times \mathbb{R}^N \to \mathbb{R}^N$ in (2-17) heißt die Verfahrensfunktion des Einschrittverfahrens.

Zur Analyse schreibt man wieder (2-17) in die Form $T^h(u) = 0$ um mit $T^h : (\mathbb{R}^N)^{\Omega_h} \to (\mathbb{R}^N)^{\Omega_h}$ definiert durch

$$T_h(u) := (u_0 - \alpha, h^{-1}(u^{j+1} - u^j) - V(h, t_j, u^j), j = 0, \dots, \sigma(h) - 1)$$

und $u = (u_0, u_1, \dots, u_{M_h}) \in (\mathbb{R}^N)^{\Omega_h}$ versehen mit der Norm

$$\|u\|_\infty = \max\{|u_i(t_j)| \mid i = 1, \dots, N,\ j = 0, \dots, \sigma(h)\}.$$

Man kann dann wieder die Begriffe der Konsistenz, Stabilität und Konvergenz definieren.

2.25 Definition. $\overline{u}_h$ bezeichne die Restriktion der wahren Lösung $\overline{u}(t)$, $t \in [t_0, t_e]$ auf das Gitter Ω_h. Dann heißt $\|T^h(\overline{u}_h)\|_\infty$ der Konsistenzfehler.

2.26 Definition. Ein numerisches Modell $T^h(u) = 0$ wird stabil genannt, falls es eine von h unabhängige Konstante $C > 0$ derart gibt, dass

$$\|u - v\|_\infty \leq C \cdot \|T^h(u) - T^h(v)\|_\infty$$

für alle $u, v \in (\mathbb{R}^N)^{\Omega_h}$ und $0 < h < h_0$ gibt.

2.27 Definition. Ist u^h die Lösung von $T^h(u) = 0$, so bezeichnet $u^h - \overline{u}_h \in (\mathbb{R}^N)^{\Omega_h}$ den Konvergenzfehler. Die Konvergenz der Ordnung p in der Maximumsnorm verlangt dann

$$\|u^h - \overline{u}_h\|_\infty = O(h^p).$$

Hinreichend hierfür ist die Stabilität des numerischen Modells $T^h(u) = 0$ und die Konsistenz der Ordnung p gemäß $\|T^h(\overline{u}_h)\|_\infty = O(h^p)$.

Die Runge-Kutta-Verfahren

Seien $h = \frac{t_e - t_0}{\sigma(h)} > 0$, $\Omega_h = \{t_j = t_0 + jh \mid j = 0, \ldots, \sigma(h)\}$. Ein s-stufiges Runge-Kutta-Verfahren für die Anfangswertaufgabe (2-16) ist gegeben durch

$$\begin{aligned} u_0 &= \alpha, && (2\text{-}18) \\ u_{m+1} &= u_m + h \sum_{i=1}^{s} b_i f(t_m + c_i h, U_i^m), \quad m = 0, \ldots, \sigma(h), \end{aligned}$$

wobei sich die sogenannten Stufenwerte $U^m = (U_1^m, U_2^m, \ldots, U_s^m)$ als Lösung des Gleichungssystems

$$U_i^m = u_m + h \sum_{j=1}^{s} a_{ij} f(t_m + c_j h, U_j^m), \quad i = 1, \ldots, s \tag{2-19}$$

ergeben.

Ein Runge-Kutta-Verfahren wird durch die Vorgabe eines Runge-Kutta-Tableaus definiert:

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array}$$

Hierfür schreibt man kurz: $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$, $c, b \in \mathbb{R}^s$, $A \in \mathbb{R}^{s,s}$.

2.28 Definition. Das Verfahren (2-18) ist explizit, falls $a_{ij} = 0$ für $j \geq i$ gilt, sonst ist es implizit.

Im Allgemeinen ist auf jedem Zeitlevel t_m das (sN)-dimensionale System (2-19) zu lösen.

Die freien Parameter c_i, b_i, $1 \leq i \leq s$, a_{ij}, $1 \leq i, j \leq s$ werden nun dazu benutzt, um dem Verfahren wünschenswerte Eigenschaften wie z.B. eine möglichst hohe Konsistenzordnung zu verleihen.

Die einfachsten Verfahren für $s = 1, 2$ sind folgende:

(i) Euler-Cauchy-Verfahren $(s = 1)$: $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$

$$\begin{aligned} u_0 &= \alpha, \\ u_{m+1} &= u_m + h \cdot 1 \cdot f(t_m, U_1^m) \\ &= u_m + hf(t_m, u_m), \quad m = 0, \ldots, \sigma(h) - 1. \end{aligned}$$

(ii) implizites Euler-Cauchy-Verfahren $(s = 1)$: $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$

(iii) ϑ-Verfahren für $\vartheta \in (0, 1)$ $(s = 2)$: $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\vartheta & \vartheta \\ \hline & 1-\vartheta & \vartheta \end{array}$

Für die Durchführbarkeit lässt sich der folgende Satz zeigen:

2.29 Satz. *f genüge einer Lipschitzbedingung*

$$\|f(t, v) - f(t, w)\|_\infty \leq L_f \cdot \|v - w\|_\infty \tag{2-20}$$

für alle $v, w \in \mathbb{R}^N$ und $t \in [t_0, t_e]$. Dann besitzt das Gleichungssystem

$$U_i^m = u_m + h \sum_{j=1}^{s} a_{ij} f(t_m + c_j h, U_j^m), \quad i = 1, \ldots, s$$

für $(t_m, u_m) \in [t_0, t_e] \times \mathbb{R}^N$ und jede Schrittweite $h \in (0, t_e - t_0)$ mit $q := hL_f \|A\|_\infty < 1$ genau eine Lösung $(U_1^m, \ldots, U_j^m) = U^m = U(t, t_m, u_m) \in \mathbb{R}^{sN}$.

Für die qualitative Untersuchung eines Runge-Kutta-Verfahrens sind die sogenannten Bedingungen von Butcher oft sehr nützlich:

$$B(p) : \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, p,$$

$$C(q) : \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad i = 1, \dots, s, \ k = 1, \dots, q,$$
$$D(m) : \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{1}{k} b_j (1 - c_j^k), \quad j = 1, \dots, s, \ k = 1, \dots, m.$$

Eines der zentralen Resultate lautet:

2.30 Satz. *Genügen die Koeffizienten b, c, A eines s-stufigen Runge-Kutta-Verfahrens den vereinfachten Bedingungen von Butcher $B(p)$, $C(q)$ und $D(m)$ mit $p \le q+m+1$, $p \le 2q+2$, so besitzt das Verfahren die Konsistenzordnung p, falls die Funktion f p-mal stetig differenzierbar in einer Umgebung der Lösung ist.*

Bezüglich der Stabilität lässt sich zeigen:

2.31 Satz. *Genügt f einer Lipschitz-Bedingung der Form (2-20), so ist das Runge-Kutta-Verfahren bzgl. der Maximumsnorm stabil.*

Man erhält dann also die Konvergenz der Ordnung p bzgl. $\|\cdot\|_\infty$, d.h.

$$\max\{\|u^h(t_j) - \overline{u}_h(t_j)\|_\infty \mid j = 0, \dots, \sigma(h)\} = O(h^p),$$

falls $f \in C^{p+1}([t_0 t_e] \times \mathbb{R}^N, \mathbb{R}^N)$ global Lipschitz-stetig ist und die Butcher-Bedingungen $B(p)$, $C(q)$ und $D(m)$ mit $p \le q+m+1$, $p \le 2q+2$ erfüllt.

Lineare Mehrschrittverfahren

Vorgelegt sei die Anfangswertaufgabe

$$\begin{aligned} u'(t) &= f(t, u(t)), \quad t_0 \le t \le t_e, \\ u(t_0) &= \alpha \in \mathbb{R}^N \end{aligned}$$

mit $f \in C^1([t_0, t_e] \times \mathbb{R}^N, \mathbb{R}^N)$. In Erweiterung zu Einschrittverfahren machen Mehrschrittverfahren nicht nur von einem, sondern von mehreren vorangegangenen Näherungswerten Gebrauch.

Zu einer Schrittweite $h = \frac{t_e - t_0}{\sigma(h)}$ überziehe man das Intervall $[t_0, t_e]$ mit einem Gitter $\Omega_h = \{t_j = t_0 + jh \mid j = 0, \dots, \sigma(h)\}$. Die allgemeine Form eines k-Schritt-Verfahrens mit Schrittweite h ist

$$\frac{1}{h}(a_0 u_j + a_1 u_{j+1} + \dots + a_k u_{j+k}) = \Phi(h, t_j, \dots, t_{j+k}, u_j, \dots, u_{j+k}) \tag{2-21}$$

mit gegebenen Koeffizienten $a_i \in \mathbb{R}$, $i = 0, \dots, k$ und einer Verfahrensfunktion $\Phi : (0, h_0] \times [t_0, t_e]^{k+1} \times \mathbb{R}^{N(k+1)} \to \mathbb{R}^N$. Hierbei steht u_i für eine Approximation von $\overline{u}(t_i)$.

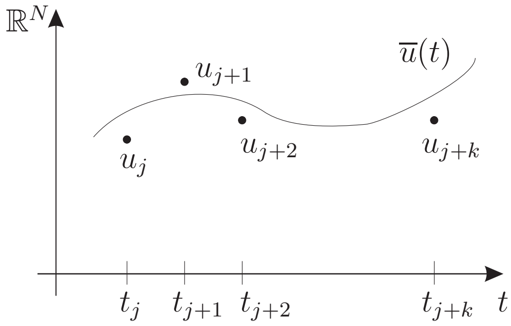


Abbildung 11: Ein Mehrschrittverfahren

2.32 Definition. Das Verfahren (2-21) heißt explizit, falls Φ nicht von u_{j+k} abhängt, sonst implizit.

Die gebräuchlisten Mehrschrittverfahren sind die linearen Mehrschrittverfahren der Gestalt

$$\begin{aligned}\Phi(h, t_j, \dots, t_{j+k}, u_j, \dots, u_{j+k}) &= b_0 f(t_j, u_j) + b_1 f(t_{j+1}, u_{j+1}) + \dots + b_k f(t_{j+k}, u_{j+k}) \\ &= \sum_{i=0}^{k} b_i f(t_{j+i}, u_{j+i})\end{aligned}$$

mit Koeffizienten $b_i \in \mathbb{R}$, $i = 0, \dots, k$. Das Verfahren lautet dann

$$\frac{1}{h} \sum_{i=0}^{k} a_i u_{j+1} = \sum_{i=0}^{k} b_i f(t_{j+i}, u_{j+i}), \quad j = 0, \dots, \sigma(h) - k \tag{2-22}$$

zu gegebenen Startwerten $u_0, \dots, u_{k-1}$. Die Parameter des Verfahrens lassen sich in Form einer Tabelle angeben:

$$\begin{array}{cccc} a_0 & a_1 & \dots & a_k \\ \hline b_0 & b_1 & \dots & b_k \end{array}$$

Letztere wird als ein Mehrschrittverfahren-Tableau bezeichnet. Für $b_k = 0$ ist das Verfahren explizit und für $b_k \neq 0$ implizit.

Wir können das lineare Mehrschrittverfahren (2-22) auf die Form $T^h(u) = 0$, $T^h : (\mathbb{R}^N)^{\Omega_h} \to (\mathbb{R}^N)^{\Omega_h}$, $u \in (\mathbb{R}^N)^{\Omega_h}$ bringen mit

$$\begin{aligned}T^h(u) = \Bigg(u_0 - \gamma_{0,h}, \dots, u_{k-1} - \gamma_{k-1,h}, h^{-1} \sum_{i=0}^{k} a_i u_{j+i} - \sum_{i=0}^{k} b_i f(t_{j+i}, u_{j+i}), \\ j = 0, \dots, \sigma(h) - k \Bigg). \end{aligned} \tag{2-23}$$

Damit sind die Begriffe der Konsistenz, Stabilität und Konvergenz direkt übertragbar.

Für ein Mehrschrittverfahren mit k-Schritten benötigt man eine Anfangsrechnung zur Bestimmung der Approximationen $\gamma_{0,h}, \ldots, \gamma_{k-1,h}$ für $\overline{u}(t_0), \ldots, \overline{u}(t_{k-1})$. Dies kann z.B. durch $(k-1)$ Schritte eines Einschrittverfahrens geschehen.

Bei impliziten linearen Mehrschrittverfahren (d.h. im Falle $b_k \neq 0$) ist in jedem Zeitschritt das Gleichungssystem

$$u_{j+k} = \frac{hb_k}{a_k} f(t_{j+k}, u_{j+k}) + \frac{1}{a_k}\left[\sum_{i=0}^{k-1} hb_i f(t_{j+i}, u_{j+i}) - a_i u_{j+i}\right]$$

nach u_{j+k} aufzulösen.

Ein Fixpunktargument sichert die Auflösbarkeit unter der Bedingung

$$q := h\frac{|b_k|}{|a_k|} L_f < 1,$$

falls

$$\|f(t,v) - f(t,w)\|_\infty \leq L_f \|v - w\|_\infty$$

für alle $v, w \in \mathbb{R}^N$ und $t \in [t_0, t_e]$ gilt.

Man erhält ein Mehrschrittverfahren der Konsistenzordnung p, d.h. $\|T^h(\overline{u}_h)\|_\infty = O(h^p)$, falls für die Koeffizienten des linearen k-Schritt-Verfahrens

$$\begin{aligned} \sum_{i=0}^{k} a_i &= 0, \quad (i^0 = 1,\ 0! = 1) \qquad (2\text{-}24)\\ \sum_{i=0}^{k} a_i \frac{i^l}{l!} - b_i \frac{i^{l-1}}{(l-1)!} &= 0 \text{ für } l = 1, \ldots, p \end{aligned}$$

gilt und f p-mal stetig differenzierbar in einer Umgebung der Lösung ist, sowie $\|\gamma_{j,h} - \overline{u}(t_j)\|_\infty = O(h^p)$ für $j = 0, \ldots, k-1$ gilt. Ferner normiert man die Koeffizienten gemäß

$$\sum_{i=0}^{k} b_i = 1. \qquad (2\text{-}25)$$

Für die praktische Anwendung geben wir hier einige Formelsätze für Mehrschrittverfahren.

(i) Adams-Bashforth-Verfahren:
$a_i = 0$, $i = 0, \ldots, k-2$, $b_k = 0$,
a_{k-1}, a_k, b_i, $i = 0, \ldots, k-1$ bestimmt aus (2-24)— (2-25)

Das Verfahren ist explizit der Ordnung $p = k$.

Beispiel für $k = 2$: $u_{j+2} - u_{j+1} = h(3/2\, f(t_{j+1}, u_{j+1}) - 1/2\, f(t_j, u_j))$

(ii) BDF-Verfahren (BDF steht für „backward differentiation formulae“):
$b_0 = b_1 = \cdots = b_{k-1} = 0$,
$a_0, \ldots, a_k, b_k$ bestimmt aus (2-24)— (2-25)

Das Verfahren ist implizit der Ordnung $p = k$

Beispiel für $k = 3$: $11/6\, u_{j+3} - 3\, u_{j+2} + 3/2\, u_{j+1} - 1/3\, u_j = hf(t_{j+3}, u_{j+3})$

Lineare Mehrschrittverfahren sind nicht mehr uneingeschränkt stabil. Zu einer linearen Mehrschrittverfahren (2-22) heißt

$$p(z) = \sum_{i=0}^{k} a_i z^i$$

das charakteristische Polynom des Verfahrens. Es gilt $\deg(p) = k$, da $a_k \neq 0$.

2.33 Satz (Dahlquist). *Das numerische Modell $T^h(u) = 0$ mit der in (2-23) definierten Verfahrensfunktion T^h eines linearen Mehrschrittverfahrens erfüllt eine Stabilitätsungleichung*

$$\|u - v\|_\infty \leq C \cdot \|T^h(u) - T^h(v)\|_\infty, \quad u, v \in (\mathbb{R}^N)^{\Omega_h},\ 0 < h \leq h_0,$$

falls f einer globalen Lipschitz-Bedingung genügt, und die folgende Wurzelbedingung gilt:

$$\text{Für jede Nullstelle } z \in \mathbb{C} \text{ des charakteristischen Polynoms gilt entweder } |z| < 1 \text{ oder } |z| = 1 \text{ und } z \text{ ist eine einfache Wurzel.} \tag{2-26}$$

2.34 Definition. Ein Verfahren, dass die Bedingung (2-26) erfüllt, heißt nullstabil.

Als Konsequenz haben wir:

2.35 Korollar. *Ein lineares Mehrschrittverfahren, dessen charakteristische Polynom die Wurzelbedingung (2-26) erfüllt, ist Konvergent der Ordnung p, falls Konsistenz der Ordnung p vorliegt und f einer globalen Lipschitz-Bedingung genügt.*

$z = 1$ ist immer eine Nullstelle von p, denn $p(1) = \sum_{i=0}^{k} a_i = 0$. Die $(k-1)$ anderen Nullstellen von p in $\mathbb{C}$ entscheiden also über die Stabilität.

2.36 Beispiel. Das charakteristische Polynom des Adams-Bashforth-Verfahrens $p(z) = z^k - z^{k-1} = z^{k-1}(z-1)$ erfüllt die Wurzelbedingung.

Die BDF-Verfahren erfüllen die Wurzelbedingung für $k = 1, 2, \ldots, 6$. Ab der Ordnung 7 sind die BDF-Verfahren instabil. Deshalb werden sie nur für $k = 1, 2, \ldots, 6$ benutzt.

d) Zeitintegration für Liniensysteme parabolischer Anfangsrandwertaufgaben

Betrachte

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta u &= 0 \text{ in } \Omega \times (0,T), \\ u &= 0 \text{ auf } \partial\Omega \times (0,T), \\ u(x,0) &= u_0(x) \text{ für } x \in \Omega, \end{aligned} \tag{2-27}$$

wobei $\Omega \subset \mathbb{R}^d$, $d = 1, 2$ ein beschränktes Gebiet ist.

Das mit klassischen Differenzenverfahren hergeleitete Liniensystem zu (2-27) erhält die Form

$$w' = \Gamma w, \quad w(0) = w^0.$$

Hierbei bekommt man für $\Omega = (0,1)$

$$\begin{aligned} \Gamma &= -\frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} = \Gamma_{\Delta x} \in \mathbb{R}^{M-1,M-1}, \quad \Delta x = \frac{1}{M}, \\ w^0 &= (u_0(\Delta x), \dots, u_0((M-1)\Delta x)) \end{aligned}$$

mit den Eigenwerten

$$\lambda_k = -\frac{2}{\Delta x^2}\left(1 - \cos\left(\frac{k\pi}{M}\right)\right), \quad k = 1, \dots, M-1.$$

Im Fall $\Omega = (0,1)^2$ ergibt sich

$$\Gamma = -\frac{1}{\Delta x^2} \begin{pmatrix} B & -C & 0 & \dots & 0 & 0 & 0 \\ -C & B & -C & \dots & 0 & 0 & 0 \\ 0 & -C & B & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & B & -C & 0 \\ 0 & 0 & 0 & \dots & -C & B & -C \\ 0 & 0 & 0 & \dots & 0 & -C & B \end{pmatrix} = \Gamma_{\Delta x} \in \mathbb{R}^{(M-1)^2,(M-1)^2},$$

wobei

$$
\begin{aligned}
B &= \begin{pmatrix} 4 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 4 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 4 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 4 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 4 \end{pmatrix} \in \mathbb{R}^{M-1,M-1}, \\
C &= I_{M-1} \in \mathbb{R}^{M-1,M-1}, \quad w = (w_0(i\Delta x, j\Delta x))_{1 \le i,j \le M-1}
\end{aligned}
$$

mit den Eigenwerten

$$\lambda_{k,l} = -\frac{2}{\Delta x^2}\left(2 - \cos\left(\frac{k\pi}{M}\right) - \cos\left(\frac{l\pi}{M}\right)\right), \quad 1 \le k, l \le M-1.$$

Im Fall Finiter Elemente erhält man das System

$$
\begin{aligned}
Bw'(t) + Aw(t) &= 0, \quad 0 \le t \le T, \\
w(0) &= w^0
\end{aligned}
$$

mit

$$
\begin{aligned}
&A = (a_{ij})_{1 \le i,j \le M} \in \mathbb{R}^{M,M}, \quad a_{ij} = a(v_i, v_j) = \int_\Omega \nabla v_i \cdot \nabla v_j \, \mathrm{d}x, \quad 1 \le i, j \le M, \\
&B = (b_{ij})_{1 \le i,j \le M}, \quad b_{ij} = \langle v_i, v_j \rangle = \int_\Omega v_i v_j \, \mathrm{d}x, 1 \le i, j \le M, \\
R_h u_0 &= \sum_{i=1}^{M} (w^0)_i v_i,
\end{aligned}
$$

wobei $R_h : V \to V_h := \mathrm{span}\{v_1, \dots, v_M\}$ der Ritz-Projektor sei. Wegen der Symmetrie und der positiven Definitheit von $B \in \mathbb{R}^{M-1,M-1}$ ist dies äquivalent zu

$$
\begin{aligned}
w' &= \Gamma w, \\
w(0) &= w^0
\end{aligned}
$$

mit $\Gamma := -B^{-1}A$.

Wir betrachten nun allgemein ein M-dimensionales System

$$w' = \Gamma w, \quad w(0) = w^0 \tag{2-28}$$

mit einer über $\mathbb{C}$ diagonalisierbaren Matrix $\Gamma \in \mathbb{R}^{M,M}$. Es existieren also linear unabhängige Vektoren $y^k \in \mathbb{C}^m$ derart, dass

$$\Gamma y^k = \lambda_k y^k, \quad k = 1, \dots, M, \tag{2-29}$$

$$Y^{-1}\Gamma Y \quad = \quad \Lambda \; = \; \operatorname{diag}(\lambda_1, \ldots, \lambda_M), \quad Y \; = \; (y^1, \ldots, y^M).$$

Die Lösung von (2-28) ergibt sich dann bekanntlich in der Form

$$w(t) = \sum_{k=1}^{M} c_k \exp(\lambda_k t) y^k, \quad t \in \mathbb{R}.$$

Dabei gilt $w(0) = w_0 = \sum_{k=1}^{M} c_k y^k$.

Wir analysieren nun Einschrittverfahren, welche angewandt mit der Zeitschrittweite Δt auf (2-28) eine Rekursion

$$w^m = g(\Delta t \Gamma) w^{m-1}, \quad m = 1, \ldots, N, \quad N \Delta t = T \tag{2-30}$$

mit einer rationalen Funktion $g(z) = \frac{p(z)}{q(z)}$, $z \in D \subset \mathbb{C}$ offen, für polynomiale p und q liefern.

2.37 Definition. Zu einem Polynom $p : D \to \mathbb{C}$, $p(z) = \sum_{l=0}^{k} \alpha_l z^l$ und einer Matrix $B \in \mathbb{R}^{M,M}$ setze

$$p(B) := \sum_{l=0}^{k} \alpha_l B^l, \quad B^0 = I.$$

Ferner setzt man zu einer rationalen Funktion $g(z) = \frac{p(z)}{q(z)}$, $z \in D \subset \mathbb{C}$ offen, $q(z) \neq 0$:

$$g(B) := (q(B))^{-1} p(B),$$

falls $q(B)$ invertierbar ist.

2.38 Bemerkung. $q(B)$ hat die Eigenwerte $q(\lambda)$, $\lambda \in \sigma(B)$. Folglich ist $q(B)$ invertierbar, falls $\sigma(B) \cap \{\hat{z} \in D \,|\, q(\hat{z}) = 0\} = \varnothing$.

Eine Darstellung der Form (2-30) gilt allgemein für Runge-Kutta-Verfahren. Betrachte zunächst den Fall $M = 1$, d.h. die Anfangswertaufgabe

$$w' = \lambda w, \quad w(0) = w^0.$$

Sei $\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$ das Tableau eines s-stufigen Runge-Kutta-Verfahrens. Man findet

$$w^{m+1} = w^m + \Delta t \sum_{i=1}^{s} b_i F(t_m + c_i \Delta t, W_i^m) = w^m + \Delta t \sum_{i=1}^{s} \lambda b_i W_i^m = w^m + \Delta t \, \lambda b^T W^m$$

mit $W^m = (W_1^m, \ldots, W_s^m)$ gegeben als die Lösung von

$$W_i^m \quad = \quad w^m + \Delta t \sum_{j=1}^{s} a_{ij} F(t_m + c_j \Delta t, W_j^m)$$

$$= \quad w^m + \Delta t \sum_{j=1}^{s} \lambda W_j^m, \quad i = 1, \ldots, s.$$

Dies ist äquivalent zu

$$\begin{pmatrix} 1 - a_{11}\Delta t\lambda & -a_{12}\Delta t\lambda & \ldots & -a_{1s}\Delta t\lambda \\ -a_{21}\Delta t & 1 - a_{22}\Delta t\lambda & \ldots & a_{2s}\Delta t\lambda \\ \vdots & \vdots & \ddots & \vdots \\ -a_{s1}\Delta t\lambda & -a_{s2}\Delta t\lambda & \ldots & 1 - a_{ss}\Delta t\lambda \end{pmatrix} \begin{pmatrix} W_1^m \\ W_2^m \\ \vdots \\ W_s^m \end{pmatrix} = \begin{pmatrix} w^m \\ w^m \\ \vdots \\ w^m \end{pmatrix},$$

d.h.

$$(I - \Delta t\lambda A)W^m = \mathbb{I}w^m$$

mit $\mathbb{I} = (1, 1, \ldots, 1)^T \in \mathbb{R}^s$. Ist die Matrix $I - \Delta t\lambda A$ invertierbar, so finden wir

$$W^m = (I - \Delta t\lambda A)^{-1}\mathbb{I}w^m$$

und daher

$$\begin{aligned} w^{m+1} &= w^m + \Delta t\lambda \cdot b^T(I - \Delta t\lambda A)^{-1}\mathbb{I}w^m \\ &= (1 + \Delta t\lambda b^T(I - \Delta t\lambda A)^{-1}\mathbb{I})w^m =: R(\Delta t\lambda)w^m \end{aligned}$$

mit

$$R(z) := 1 + zb^T(I - zA)^{-1}\mathbb{I} = g(z). \tag{2-31}$$

2.39 Definition. Die Funktion $R(z)$ in (2-31) heißt Stabilitätsfunktion des Runge-Kutta-Verfahrens.

Des Weiteren sichert die Cramersche Regel, dass $\Delta t\lambda W_i^m$, $i = 1, \ldots, s$ rationale Funktionen in $\lambda\Delta t$ sind, falls $I - \Delta t\lambda A$ invertierbar ist. Der Zählergrad und der Nennergrad überschreiten dabei nicht s. Somit hat die Funktion g aus (2-31) die Darstellung

$$g(z) = R(z) = \frac{p(z)}{q(z)},$$

wobei p und q Polynome mit $\deg(p) \leq s$ und $\deg(q) \leq s$ sind.

Wir kehren jetzt zum allgemeinen Fall (2-28) zurück. Wir nutzen die Diagonalisierbarkeit von Γ aus und setzen

$$u = Y^{-1}w$$

mit Y aus (2-29). Dann gilt

$$u' = Y^{-1}w' = Y^{-1}\Gamma w = Y^{-1}\Gamma Y u = \operatorname{diag}(\lambda_1, \ldots, \lambda_M)u.$$

Einsetzen liefert dann mit kurzer Rechnung die Iteration

$$w^m = g(\Delta t\Gamma)w^{m-1}, \quad m = 1, \ldots, N. \tag{2-32}$$

Mit (2-32) und $Y^{-1}\Gamma Y = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$ finden wir

$$\begin{aligned} w^m &= g(\Delta t\Gamma)w^{m-1} = g(\Delta t\Gamma)^m w^0 = g(\Delta t\Gamma)^m \underbrace{\left(\sum_{k=1}^{M} c_k y^k\right)}_{=w^0} \\ &= \sum_{k=1}^{M} c_k g(\Delta t\lambda_k)^m y^k, \quad m = 1, \ldots, N. \end{aligned}$$

Man beachte dabei, dass $g(\Delta t\Gamma)$ die Eigenwerte $g(\Delta t\lambda_1), \ldots, g(\Delta t\lambda_M)$ mit den Eigenvektoren $y^1, \ldots, y^M$ hat.

Vergleich der kontinuerlichen und der diskreten Lösung

Ist $\overline{w}$ die wahre Lösung von (2-28), so folgt

$$\begin{aligned} \overline{w}(m\Delta t) - w^m &= \sum_{k=1}^{M} c_k \underbrace{\exp(m\Delta t\lambda_k)}_{=\exp(\Delta t\lambda_k)^m} y^k - \sum_{k=1}^{M} c_k g(\Delta t\lambda_k)^m y^k \\ &= \sum_{k=1}^{M} (\exp(\Delta t\lambda_k)^m - g(\Delta t\lambda_k)^m) c_k y^k. \end{aligned}$$

Zu vergleichen sind die sogenannten Amplikationsfaktoren $\exp(\lambda_k\Delta t)$ für die kontinuerliche Aufgabe und $g(\Delta t\lambda_k)$ für die diskrete Aufgabe.

Seien nun die Eigenwerte von $\Gamma \in \mathbb{R}^{M,M}$ gemäß

$$\mathrm{Re}(\lambda_M) \leq \ldots \mathrm{Re}(\lambda_2) \leq \mathrm{Re}(\lambda_1)$$

angeordnet. Damit $g(\Delta t\lambda_k)$ und $\exp(\Delta t\lambda_k)$ wenigstens qualitativ übereinstimmen, ist die Gültigkeit von

$$|g(\Delta t\lambda_k)| \begin{cases} < 1, & \text{falls } \mathrm{Re}(\lambda_k) < 0, \\ > 1, & \text{falls } \mathrm{Re}(\lambda_k) > 0 \end{cases}, \quad k = 1, \ldots, M \tag{2-33}$$

zu fordern. Der für die Anwendungen interessante Fall ist nun

$$\mathrm{Re}(\lambda_M) \leq \ldots \mathrm{Re}(\lambda_2) \leq \mathrm{Re}(\lambda_1) < 0. \tag{2-34}$$

Man vergleiche dazu beispielsweise Liniensysteme für parabolische Differentialgleichungen.

2.40 Definition. Das Verhältnis $\sigma = \frac{\mathrm{Re}(\lambda_M)}{\mathrm{Re}(\lambda_1)}$ heißt im diesem Fall die Steifheit von Γ. Man sagt, $w'(t) = \Gamma w(t)$ ist eine steife Differentialgleichung, falls $\sigma = \sigma(\Gamma) \gg 1$.

2.41 Beispiel. Für die Matrix

$$\Gamma = -\frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

mit den Eigenwerten

$$\begin{aligned} \lambda_k &= -\frac{2}{\Delta x^2}(1-\cos(k\pi\Delta x)) = -\frac{2}{\Delta x^2}\big(1-\big[1-\frac{1}{2}k^2\pi^2\Delta x^2 + O(\Delta x^4)\big]\big) \\ &= -k^2\pi^2 + O(\Delta x^2), \quad k = 1,\dots,M \end{aligned}$$

für festes $M_0 \in \{1,\dots,M-1\}$ ergibt sich

$$\sigma(\Gamma) = \frac{\lambda_{M-1}}{\lambda_1} = \frac{-(M-1)^2\pi^2}{-\pi^2} + O(\Delta x^2) = O\left(\frac{1}{\Delta x^2}\right),$$

d.h. Liniensysteme parabolischer Differentialgleichungen sind für $\Delta x \ll 1$ steife Systeme.

Im Falle von (2-34) erzwingt (2-33) die Bedingung

$$|g(\Delta t\lambda_k)| < 1, \quad k = 1,\dots,M.$$

Dies motiviert:

2.42 Definition. Die Menge $S := \{z \in \mathbb{C} \mid |g(z)| < 1\}$ heißt der Bereich der absoluten Stabilität des Einschrittverfahrens mit der Funktion g.

In der Situation von (2-34), d.h. 0 ist asymptotisch ein stabiler stationärer Punkt von $w' = \Gamma w$, ist es wünschenswert, Verfahren mit einem möglichst großen Bereich der absoluten Stabilität zu haben, damit die Forderung $|g(\Delta t\lambda_k)| < 1$, $k = 1,\dots,M$ nicht zu kleinen Schrittweiten erzwingt. Im Idealfall soll gelten

$$\mathbb{C}_- := \{z \in \mathbb{C} \mid \operatorname{Re}(z) < 0\} \subset S. \tag{2-35}$$

2.43 Definition. Ein Verfahren mit der Eigenschaft (2-35) heißt absolut stabil oder kurz A-stabil.

2.44 Definition. Ein A-stabiles numerisches Verfahren heißt stark absolut stabil oder kurz L-stabil, wenn $g(z) \to 0$ für $\operatorname{Re}(z) \to -\infty$ gilt.

Die Koeffizienten $g(\Delta t\lambda_k)^j$ in der diskreten Lösung klingen dann umso stärker ab, je kleiner $\mathrm{Re}(\lambda_k)$ ist, genau wie dies die Koeffizienten $\exp(\Delta t\lambda_k)^j$ in der kontinuerlichen Lösung tun.

2.45 Beispiel (ϑ-Verfahren). Man findet $g(z) = g_\vartheta(z) = \frac{1+(1-\vartheta)z}{1-\vartheta z}$, $0 \leq \vartheta \leq 1$. Das ϑ-Verfahren ist A-stabil, d.h. $S \supset \mathbb{C}_-$, falls $1/2 \leq \vartheta \leq 1$. Ferner gilt

$$\lim_{\mathrm{Re}(z)\to-\infty} |g_\vartheta(z)| = \lim_{\mathrm{Re}(z)\to-\infty} \left|\frac{1+(1-\vartheta)z}{1-\vartheta z}\right| = \left|\frac{1-\vartheta}{\vartheta}\right|,$$

d.h. das ϑ-Verfahren ist L-stabil, nur falls $\vartheta = 1$.

2.46 Beispiel (Explizite Runge-Kutta-Verfahren). Ein explizites Runge-Kutta-Verfahren lautet

$$w^{m+1} = w^m + \Delta t \sum_{i=1}^{s} b_i \Gamma W_i^m, \quad m = 0, 1, \ldots, N-1$$

mit

$$\begin{aligned} W_1^m &= w^m, \\ W_i^m &= w^m + \Delta t \sum_{j=1}^{i-1} a_{ij} \Gamma W_j^m, \quad i = 2, \ldots, s, \end{aligned}$$

wobei $a_{ij} = 0$ für $j \geq i$ zu beachten ist. Man erhält, dass $g(z)$ ein Polynom vom höchstens s-ten Grade ist, d.h.

$$g(z) = \frac{p(z)}{1}, \quad \deg(p) \leq s.$$

Wegen des Satzes von Liouville, d.h. $|p(z)| \to \infty$ für $\mathrm{Re}(z) \to -\infty$ folgt, dass explitize Runge-Kutta-Verfahren niemals A-stabil sind.

2.47 Beispiel (Implizite Runge-Kutta-Verfahren). Nicht alle impliziten Runge-Kutta-Verfahren sind A-stabil. Aber manche sind es. Ferner findet man A-stabile Runge-Kutta-Verfahren von beliebig hoher Ordnung. So sind z.B. die Verfahren vom Gauss-Typ oder vom Radau IIA Typ A-stabil. Die Radau IIA Verfahren sind sogar L-stabil.

Absolute Stabilitätsbereiche für lineare Mehrschrittverfahren

Vorgelegt sei die Anfangswertaufgabe

$$w'(t) = \lambda w(t), \quad w(0) = w^0, \quad \lambda \in \mathbb{C}. \tag{2-36}$$

welche durch $\overline{w}(t) = \exp(\lambda t)$ gelöst wird. Wir diskretisieren das Problem (2-36) mit einem linearen Mehrschrittverfahren mit dem Tableau

$$\begin{array}{cccc} a_0 & a_1 & \dots & a_k \\ \hline b_0 & b_1 & \dots & b_k \end{array}$$

und erhalten mit der Aproximation w^l für $\overline{w}(t_l)$ das Schema

$$\sum_{i=0}^{k} a_i w^{j+i} = \Delta t \sum_{i=0}^{k} b_i \lambda w^{j+i}, \quad j = 0, \dots, N-k$$

zu vorgegebenen $w^0, \dots, w^{k-1}$. Die obige Iteration ist äquivalent zu

$$\sum_{i=0}^{j} (a_i - b_i \Delta t \lambda) w^{j+i} = 0, \quad j = 0, \dots, N-k. \tag{2-37}$$

(2-37) hat die Form

$$\sum_{i=0}^{k} g_i(\Delta t \lambda) w^{j+i} = 0$$

mit $g_i(z) = a_i - b_i z$, $i = 0, \dots, k$.

2.48 Definition. Zum Polynom

$$p(z, \xi) := \sum_{i=0}^{k} g_i(z) \xi^i \tag{2-38}$$

heißt die Menge

$$\begin{aligned} S := \{z \in \mathbb{C} \mid &\text{Für jede Nullstelle } \xi \text{ von } p(z, \cdot) \text{ gilt } |\xi| \le 1, \\ &\text{falls } \xi \text{ einfache Nullstelle oder } |\xi| < 1 \text{ sonst.}\} \end{aligned} \tag{2-39}$$

der absolute Stabilitätsbereich des linearen Mehrschrittverfahrens.

Zur Motivation von (2-38)—(2-39) betrachten wir die Lösung von (2-37) wobei wir der Einfachheit halber annehmen, dass das Polynom $p(\Delta t \lambda, \xi)$ nur einfache Nullstellen $\xi_j \in \mathbb{C}$, $j = 1, \dots, k$ hat. Wir bestimmen nun $\gamma_1, \gamma_2, \dots, \gamma_k$ eindeutig als Lösung von

$$\underbrace{\begin{pmatrix} 1 & 1 & \dots & 1 \\ \xi_1 & \xi_2 & \dots & \xi_k \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1^{k-1} & \xi_2^{k-1} & \dots & \xi_k^{k-1} \end{pmatrix}}_{=:V} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{pmatrix} = \begin{pmatrix} w^0 \\ w^1 \\ \vdots \\ w^{k-1} \end{pmatrix} \tag{2-40}$$

2.49 Bemerkung. Die Matrix $V \in \mathbb{R}^{k,k}$ ist die Vandermondesche Matrix, welche genau dann invertierbar ist, wenn $\xi_i \neq \xi_j$ für $i \neq j$.

Dann lautet die Lösung von (2-37)

$$w^j = \sum_{l=1}^{k} \gamma_l \xi_l^j, \quad j \in \mathbb{N}_0. \tag{2-41}$$

Sei $j \in \{0, \dots, k-1\}$. Nach Konstruktion von $\gamma_1, \dots, \gamma_k$ gilt

$$w^j = \sum_{l=1}^{k} \gamma_k \xi_l^j.$$

Für $j \geq k$ finden wir

$$\sum_{\kappa=0}^{k} g_\kappa(\Delta t\lambda) w^{j+\kappa} = \sum_{\kappa=0}^{k} g_\kappa(\Delta t\lambda) \sum_{l=1}^{k} \gamma_l \underbrace{\xi_l^{j+\kappa}}_{=\xi_l^j \cdot \xi_l^\kappa} = \sum_{l=1}^{k} \gamma_l \xi_l^j \underbrace{\sum_{\kappa=0}^{k} g_\kappa(\Delta t\lambda) \xi_l^\kappa}_{=p(\Delta t\lambda, \xi_l)=0} = 0.$$

Anhand der Lösungsdarstellung (2-41) sehen wir, dass im Fall $\mathrm{Re}(\lambda) < 0$ wiederum $\Delta\lambda \in S$ zu fordern ist, damit die numerische Lösung, genau wie die kontinuerliche Lösung, monoton abfällt.

2.50 Definition. Ein lineares Mehrschrittverfahren heißt absolut stabil oder A-stabil, falls $\mathbb{C}_- \subset S$ gilt. Es heißt sogar L-stabil, falls es A-stabil ist und falls zusätzlich für alle Nullstellen $\xi(z)$ von $p(z, \cdot)$

$$\lim_{\mathrm{Re}(z) \to -\infty} \xi(z) = 0$$

gilt.

Es gibt aber ein bekanntes Resultat von Dahlquist, nach dem jedes A-stabile Mehrschrittverfahren implizit ist und höchstens die Konvergenzordnung 2 hat. Dazu gehören die ϑ-Verfahren für $1/2 \leq \vartheta \leq 1$ und das BDF-Verfahren der Stufe 2 mit dem Tableau $\begin{array}{ccc} 1/2 & -2 & 3/2 \\ \hline 0 & 0 & 1 \end{array}$.

Um also lineare Mehrschrittverfahren mit höherer Konvergenzordnung als 2 zu erhalten, muss man den Begriff der A-Stabilität etwas abschwächen. Dies führt dann zu folgender Definition.

2.51 Definition. Ein lineares Mehrschrittverfahren heißt A(α)-stabil mit $0 < \alpha < \frac{\pi}{2}$, falls

$$S \supset \mathbb{C}_{-,\alpha} = \{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0, \frac{\mathrm{Im}(z)}{\mathrm{Re}(z)} < \tan(\alpha)\}.$$

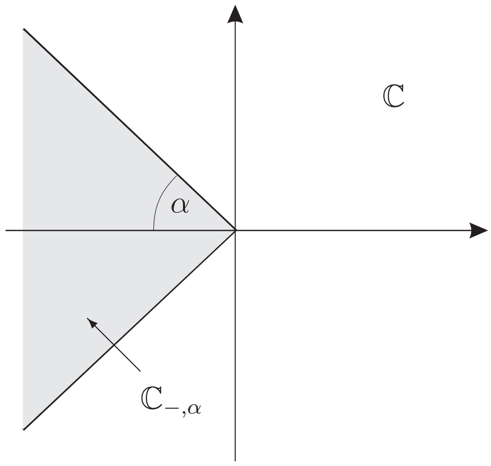


Abbildung 12: Die Menge $\mathbb{C}_{-,\alpha}$ für ein $\alpha > 0$

Man erhält, dass die BDF-Formeln bis zur Ordnung 6 (nur bis zu dieser Ordnung sind sie überhaupt stabil) A(α)-stabil sind. Die BDF-Formeln erfüllen überdies

$$\lim_{\mathrm{Re}(z)\to-\infty} \xi(z) = 0$$

für alle Nullstellen $\xi(z)$ von $p(z,\cdot)$, d.h. sind L(α)-stabil.

2.52 Bemerkung. A(α)- bzw. L(α)-Stabilität, $\alpha > 0$, ist beispielsweise völlig hinreichend für Probleme der Gestalt $w' = \Gamma w$ mit lediglich reellen Eigenwerten λ. Man beachte dabei, dass $\Delta t\lambda \in \mathbb{R}$ für alle Eigenwerte $\lambda \in \sigma(\Gamma)$ gilt. Darunter fallen unsere Liniensysteme parabolischer Anfangsrandwertprobleme unabhängig davon, ob die Diskretisierung im Raum mit Finiten Differenzen oder Finiten Elementen gemacht wurde.

Software-Pakete für steife Differentialgleichungen

a) Radau 5:
Radau IIA Verfahren der Stufe $s = 3$ und Ordnung $p = 5$ (Runge-Kutta-Verfahren), E. Hairer (Genf).

b) DASS K:
BDF-Verfahren mit variabler Ordnung durch Benutzung der Stufen $k = 2, \dots, 6$ (lineares Mehrschrittverfahren), L. Petzold (St. Barbara).

c) ode15s:
BDF-Verfahren für steife Differentialgleichungen in **Matlab**. Benutzt die Formelsätze vom BDF-Typ für $k = 1, \dots, 5$.

e) Nichtlineare parabolische Anfangsrandwertaufgaben

Es sei $\Omega = (0,1)$. Betrachte die nichtlineare Anfangsrandwertaufgabe

$$\begin{aligned} \frac{\partial u}{\partial t} &= u_{xx} + f(u,x,t) \text{ in } \Omega \times (0,T), \\ u(x,0) &= u_0(x) \text{ in } \Omega, \\ u(0,t) &= \gamma_0(t), \quad u(1,t) = \gamma_1(t) \text{ in } (0,T). \end{aligned} \tag{2-42}$$

Es wird $f \in C^1(\mathbb{R} \times [0,1] \times [0,T], \mathbb{R})$ und

$$q \leq \frac{\partial f}{\partial u}(u,x,t) \leq \mu, \quad u \in \mathbb{R}, \quad 0 \leq x \leq 1, \quad 0 \leq t \leq T \tag{2-43}$$

vorausgesetzt.

Es seien $\Delta x = \frac{1}{M} > 0$, $\Omega_{\Delta x} = \{x_j = j\Delta x \,|\, j = 1, \ldots, M-1\}$. Setze

$$v(t) = (u(x_1,t), \ldots, u(x_{M-1},t)) = (v_1(t), \ldots, v_{M-1}(t)), \quad 0 \leq t \leq T.$$

Das Liniensystem durch Diskretisierung mit dem klassichen Differenzenverfahren für den nichtlinearen Differentialoperator $u_{xx}(\cdot,t) + f(u(\cdot,t),\cdot,t)$, $u(0,t) = \gamma(0,t)$, $u(1,t) = \gamma_1(t)$, $t \in (0,T)$ lautet

$$\begin{aligned} v'(t) &= F_{\Delta x}(v(t),t), \quad 0 \leq t \leq T, \\ v(0) &= v^0 \end{aligned}$$

mit

$$\begin{aligned} F_{\Delta x}(v,t) &= -\Gamma v + E(v,t) + r^1(t), \\ \Gamma &= \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{M-1,M-1}, \\ r^{\Delta x} &= \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0(t) \\ 0 \\ \vdots \\ 0 \\ \gamma_1(t) \end{pmatrix} \in \mathbb{R}^{M-1}, \\ E(v,t) &= \begin{pmatrix} f(v_1,x_1,t) \\ \vdots \\ f(v_i,x_i,t) \\ \vdots \\ f(v_{M-1},x_{M-1},t) \end{pmatrix}, \quad v^0 = \begin{pmatrix} u_0(x_1) \\ \vdots \\ u_0(x_i) \\ \vdots \\ u_0(x_{M-1}) \end{pmatrix} \in \mathbb{R}^{M-1} \end{aligned} \tag{2-44}$$

Mit $v^j = (u_1^j, \dots, u_{M-1}^j) \in \mathbb{R}^{M-1}$, $j = 0, \dots, N$, $N\Delta t = T$, $t_j = j\Delta t$ lautet das ϑ-Verfahren

$$\begin{aligned} \frac{1}{\Delta t}(v^j - v^{j-1}) &= \vartheta F_{\Delta x}(v^j, t_j) + (1-\vartheta) F_{\Delta x}(v^{j-1}, t_{j-1}), \quad j = 1, \dots, N, \\ v^0 &= r^0 = (u_0(x_1), \dots, u_0(x_{M-1})), \end{aligned}$$

wobei $F_{\Delta x}(v,t) = -\Gamma v + E(v,t) + r^1(t)$. Mit den Abkürzungen

$$\begin{aligned} G_j(v) &= \frac{1}{\Delta t} v - \vartheta F_{\Delta x}(v, t_j), \\ H_j(v) &= \frac{1}{\Delta t} v + (1-\vartheta) F_{\Delta x}(v, t_j) \end{aligned}$$

schreiben wir das diskrete Problem in der Form $T^h(u) = 0$ mit $T^h : \mathbb{R}^{\Omega_h} \to \mathbb{R}^{\Omega_h}$ gegeben durch

$$T^h(u) = \begin{pmatrix} v^0 - r^0 \\ G_1(v^1) - H_0(v^0) \\ \vdots \\ G_N(v^N) - H_{N-1}(v^{N-1}) \end{pmatrix} = 0$$

wobei $\Omega_h = \{(x_i, t_j) \,|\, i = 1, \dots, N-1,\ j = 0, \dots, N\}$.

Gemäß Voraussetzung (2-43) gilt für $w, z \in \mathbb{R}^{M-1}$

$$\begin{aligned} F_{\Delta x}(w,t) - F_{\Delta x}(z,t) &= -F(w-z) + E(w,t) - E(z,t) \\ &= -\Gamma(w-z) + \begin{pmatrix} f(w_1, x_1, t) - f(z_1, x_1, t) \\ \vdots \\ f(w_{M-1}, x_{M-1}, t) - f(w_{M-1}, x_{M-1}, t) \end{pmatrix} \\ &= -\Gamma(w-z) + \begin{pmatrix} \frac{\partial f}{\partial u}(\xi_1, x_1, t)(w_1 - z_1) \\ \vdots \\ \frac{\partial f}{\partial u}(\xi_{M-1}, x_{M-1}, t)(w_{M-1} - z_{M-1}) \end{pmatrix} \\ &= -\Gamma(w-z) + \underbrace{\operatorname{diag}(D_{ii}(w_i, z_i), i = 1, \dots, M-1)}_{=:D(v,z)}(w-z) \\ &= -\Gamma(w-z) + D(w,z)(w-z) \end{aligned}$$

mit

$$q \le D_{ii}(w,z) \le \mu, \quad i = 1, \dots, M-1, \tag{2-45}$$

wobei

$$\begin{aligned} D_{ii}(w_i, z_i) &= \frac{\partial f}{\partial u}(\xi_i, x_i, t), \quad i = 1, \dots, M-1, \\ \xi_i &= w_i + s(z_i - w_i) \text{ für ein } s \in (0,1), \quad i = 1, \dots, M-1. \end{aligned}$$

Herleitung einer Stabilitätsungleichung

Wir wählen zwei Elemente $v = (v^0, \ldots, v^N)$, $z = (z^0, \ldots, z^N) \in \mathbb{R}^N$ und setzen $g = (g^0, \ldots, g^N) := T^h(v) - T^h(z)$, d.h.

$$\begin{aligned} g^0 &= v^0 - z^0, \\ g^j &= G_j(v^j) - G_j(z^j) - (H_{j-1}(v^{j-1}) - H_{j-1}(z^{j-1})), \quad j = 1, \ldots, N. \end{aligned}$$

Diese Rekursion lässt sich ferner wie folgt schreiben:

$$\begin{aligned} g^j &= \frac{1}{\Delta t} v^j - \vartheta F_{\Delta x}(v^j, t_j) - \frac{1}{\Delta t} z^j + \vartheta F_{\Delta x}(z^j, t_j) \\ &\quad - \left[\frac{1}{\Delta t} v^j - (1-\vartheta) F_{\Delta x}(v^j, t_j) - \frac{1}{\Delta t} z^j + (1-\vartheta) F_{\Delta x}(z^j, t_j) \right] \\ &= \underbrace{\left(\frac{1}{\Delta t} I - \vartheta(-\Gamma + D(v^j, z^j)) \right)}_{=:A_j} (v^j - z^j) \\ &\quad - \underbrace{\left(\frac{1}{\Delta t} I + (1-\vartheta)(-\Gamma + D(v^{j-1}, z^{j-1})) \right)}_{=:B_{j-1}} (v^{j-1} - z^{j-1}) \\ &= A_j(v^j - z^j) - B_{j-1}(v^{j-1} - z^{j-1}), \quad j = 1, \ldots, N-1, \\ g^0 &= v^0 - z^0. \end{aligned}$$

Überdies ist A_j jeweils eine L_0-Matrix, denn $(A_j)_{lk} = \vartheta \Gamma_{lk} \le 0$ für $l \neq k$, und es gilt

$$A_j \mathbb{I} = \frac{1}{\Delta t} \mathbb{I} + \underbrace{\vartheta \Gamma \mathbb{I}}_{\ge 0} - \vartheta \underbrace{D(v^j, z^j) \mathbb{I}}_{\le \mu \mathbb{I}} \ge \left(\frac{1}{\Delta t} - \vartheta \mu \right) \mathbb{I}.$$

Setzt man nun $\Delta t \vartheta \mu \le \kappa < 1$ voraus, so ist A_j sogar eine M-Matrix mit

$$\mathbb{I} = A_j^{-1} A_j \mathbb{I} \ge \left(\frac{1}{\Delta t} - \vartheta \mu \right) A_j^{-1} \mathbb{I},$$

d.h.

$$A_j^{-1} \mathbb{I} \le \frac{1}{\frac{1}{\Delta t} - \vartheta \mu} \mathbb{I} = \frac{\Delta t}{1 - \Delta t \vartheta \mu} \mathbb{I}. \tag{2-46}$$

Wir erhalten ferner

$$\|A_j^{-1}\|_\infty = \|A_j^{-1} \mathbb{I}\|_\infty \le \left\| \frac{\Delta t}{1 - \Delta t \vartheta \mu} \mathbb{I} \right\|_\infty = \frac{\Delta t}{1 - \Delta t \vartheta \mu} \le \frac{\Delta t}{1 - \kappa} =: C_1 \Delta t. \tag{2-47}$$

Im nächsten Schritt zeigen wir, dass $G_j : \mathbb{R}^{M-1} \to \mathbb{R}^{M-1}$ für $j = 1, \ldots, N$ bijektiv ist, wobei wir uns die Darstellung

$$G_j(v^i) - G_j(z^i) = \frac{1}{\Delta t} v^i - \frac{1}{\Delta t} z^i - \vartheta(F_{\Delta x}(v^i, t_j) - F_{\Delta x}(z^i, t_j))$$

$$= \left(\frac{1}{\Delta t} I - \vartheta(-\Gamma + D(v^i, z^i))\right)(v^i - z^i) \;=\; A(v,z)(v-z)$$

mit einer M-Matrix $A(v,z)$ zu Nutze machen. Da sich für alle $v^i, z^i \in \mathbb{R}^{M-1}$ mit $G_j(v^i) = G_j(z^i)$ die Gleichheit

$$0 = G_j(v^i) - G_j(z^i) = A_j(v^i - z^i)$$

und damit $v^i - z^i = 0$ ergibt, bekommt man die Injektivität von G_j. Um die Surjektivität von G_j nachzuweisen, nimmt man ein beliebiges $r \in \mathbb{R}^{M-1}$ und setzt $\hat{r} := r - G_j(0)$, $\hat{v} := (A(\hat{v}, 0))^{-1}\hat{r}$. Man beachte, dass $\hat{v}$ wegen (2-47) für $C_1\Delta t < 1$ gemäß des Fixpunktsatzes wohldefiniert ist. Dann folgt

$$G_j(\hat{v}) - G_j(0) = A_j(\hat{v} - 0) = \hat{r} = r - G_j(0)$$

und daher auch $G_j(\hat{v}) = r$.

Also ist das ϑ-Verfahren unter der Bedingung

$$\Delta t \vartheta \mu \leq \kappa < 1$$

durchführbar.

Diskussion von B_j

Es gilt

$$B_j = \frac{1}{\Delta t} I + (1-\vartheta)\underbrace{(-\Gamma + D(v^j, z^j))}_{\leq \mu I} \leq \frac{1}{\Delta t} I + (1-\vartheta)(-\Gamma + \mu I) =: P. \qquad (2\text{-}48)$$

Wir wollen nun die Ungleichung $|B_j| \leq P$, d.h.

$$|(B_j)_{kl}| \leq P_{kl}, \quad 1 \leq k, l \leq M-1$$

erfüllen. Dies ist äquivalent zu $B_j \leq P$ und $B_j \geq -P$. Die Formelzeile (2-48) sichert die Abschätzung $B_j \leq P$. Zu zeigen bleibt also, dass auch $B_j \geq -P$ gilt.

Für die Nebendiagonalelemente von B_j ergibt sich

$$(B_j)_{kl} = -(1-\vartheta)\Gamma_{k,l} \geq 0 \geq (1-\vartheta)\Gamma_{kl} = -P_{kl}.$$

Somit folgt $B_j \geq -P$, falls für die Hauptdiagonalelemente

$$\begin{aligned}(B_j)_{kk} &= \frac{1}{\Delta t} + (1-\vartheta)\Big(-\frac{2}{\Delta x^2} + \underbrace{(D(v^j, z^j))_{kk}}_{\geq q}\Big) \\ &\geq \frac{1}{\Delta t} + (1-\vartheta)\left(-\frac{2}{\Delta x^2} + q\right) \overset{!}{\geq} -\frac{1}{\Delta t} - (1-\vartheta)\left(-\frac{2}{\Delta x^2} + \mu\right) = -P_{kk}\end{aligned}$$

gilt. Dies ist genau dann erfüllt, wenn

$$\frac{2}{\Delta t} \geq (1-\vartheta)\left(\frac{4}{\Delta x^2} - (q+\mu)\right),$$

was mit

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)} + \frac{q+\mu}{4}\Delta t \tag{2-49}$$

äquivalent ist.

Wir setzen nun (2-49) voraus und leiten eine Stabilitätsungleichung her. Aus

$$g^j = A_j(v^j - z^j) - B_{j-1}(v^{j-1} - z^{j-1})$$

folgt

$$\begin{aligned} \|v^j - z^j\| &= \|A_j^{-1}B_{j-1}(v^{j-1}-z^{j-1}) + A_j^{-1}g^j\|_\infty \\ &\leq \|A_j^{-1}B_{j-1}\|_\infty \cdot \|v^{j-1}-z^{j-1}\|_\infty + \|A_j^{-1}\|_\infty \cdot \|g^j\|_\infty. \end{aligned} \tag{2-50}$$

Nun gilt mit $|\cdot| : \mathbb{R}^{M-1,M-1} \longrightarrow \mathbb{R}^{M-1,M-1}$ definiert durch

$$C \mapsto (|C_{ij}|)_{1\leq i,j\leq M-1}$$

sofort

$$\pm B_{j-1} \leq |B_{j-1}| \leq P \;\Rightarrow\; \pm A_j^{-1}B_{j-1} \leq A_j^{-1}|B_{j-1}| \leq A_j^{-1}P$$

und somit

$$\|A_j^{-1}B_{j-1}\|_\infty \leq \|A_j^{-1}P\|_\infty.$$

Einsetzen in (2-50) liefert

$$\|v^j - z^j\| \leq \|A_j^{-1}P\|_\infty \cdot \|v^{j-1}-z^{j-1}\|_\infty + \|A_j^{-1}\|_\infty \cdot \|g^j\|_\infty.$$

Wir benötigen jetzt eine Abschätzung für $\|A^{-1}P\|_\infty$:

$$\begin{aligned} (A_j - P)\mathbb{I} &= \left[\left(\frac{1}{\Delta t} - \vartheta(-\Gamma + D(v^j,z^j))\right) - \left(\frac{1}{\Delta t}I + (1-\vartheta)(-\Gamma+\mu I)\right)\right]\mathbb{I} \\ &= (\Gamma - \vartheta\underbrace{D(v^j,z^j)}_{\leq \mu I} - (1-\vartheta)\mu I)\mathbb{I} \\ &\geq (\Gamma - \mu I)\mathbb{I} = \underbrace{\Gamma\mathbb{I}}_{\geq 0} \geq -\mu\mathbb{I}. \end{aligned}$$

Da A_j eine M-Matrix ist, folgt

$$(I - A_j^{-1}P)\mathbb{I} = A_j^{-1}(A_j - P)\mathbb{I} \geq -\mu A_j^{-1}\mathbb{I},$$

was mit

$$A_j^{-1}P\mathbb{I} \leq (I + \mu A_j^{-1})\mathbb{I}$$

gleichbedeutend ist. Dies liefert

$$\begin{aligned}
A_j^{-1}P\mathbb{I} &\leq \mathbb{I} + \mu A_j^{-1}\mathbb{I} \;=\; \mathbb{I} + \max\{\mu, 0\} \underbrace{A_j^{-1}\mathbb{I}}_{\leq \frac{\Delta t}{1-\Delta t\vartheta\mu}\mathbb{I}} \\
&\overset{(2-45)}{\leq} \mathbb{I} + \underbrace{\max\{\mu, 0\} \cdot \frac{\Delta t}{1-\Delta t\vartheta\mu}}_{=:C_2\Delta t}\mathbb{I} \;=\; (1 + C_2\Delta t)\mathbb{I}.
\end{aligned}$$

Wegen $A_j^{-1}P \geq 0$ folgt ferner

$$\|A_j^{-1}P\|_\infty = \|A_j^{-1}P\mathbb{I}\|_\infty \leq \|(1 + C_2\Delta t)\mathbb{I}\|_\infty = 1 + C_2\Delta t. \tag{2-51}$$

Benutzen wir nun (2-47) und (2-51), so folgt

$$\begin{aligned}
\|v^j - z^j\|_\infty &\leq \|A_j^{-1}P\|_\infty \cdot \|v^{j-1} - z^{j-1}\|_\infty + \|A_j^{-1}\|_\infty \cdot \|g^j\|_\infty \\
&\leq (1 + C_2\Delta t)\|v^{j-1} - z^{j-1}\|_\infty + C_1\Delta t \cdot \|g^j\|_\infty, \quad j = 1, \ldots, N.
\end{aligned}$$

Durch Induktion erhalten wir dann

$$\|v^j - z^j\|_\infty \leq (1 + C_2\Delta t)^j\|g^0\|_\infty + \sum_{k=1}^{j}(1 + C_2\Delta t)^{j-k} \cdot C_1\Delta t \cdot \|g^k\|_\infty. \tag{2-52}$$

$j = 0:$ Offensichtlich gilt

$$\|v^0 - z^0\|_\infty \leq \|g^0\|_\infty = \|v^0 - z^0\|_\infty.$$

$j - 1 \to j:$ Es folgt

$$\begin{aligned}
\|v^j - z^j\|_\infty &\leq (1 + C_2\Delta t) \cdot \|v^{j-1} - z^{j-1}\|_\infty + C_1\Delta t\|g^j\|_\infty \\
&\leq (1 + C_2\Delta t)\big[(1 + C_2\Delta t)^{j-1}\|g^0\|_\infty \\
&\quad + \sum_{k=1}^{j-1}(1 + C_2\Delta t)^{j-1-k} \cdot C_1\Delta t \cdot \|g^k\|_\infty\big] + C_1\Delta t\|g^j\|_\infty \\
&= (1 + C_2\Delta t)^j\|g^0\|_\infty + \sum_{k=1}^{j-1}(1 + C_2\Delta t)^{j-k} \cdot C_1\Delta t \cdot \|g^k\|_\infty + C_1\Delta t \cdot \|g^j\|_\infty \\
&= (1 + C_2\Delta t)^j\|g^0\|_\infty + \sum_{k=1}^{j}(1 + C_2\Delta t)^{j-k} \cdot C_1\Delta t\|g^j\|_\infty.
\end{aligned}$$

Mit (2-52) und $1 + C_2\Delta t \leq \exp(C_2\Delta t)$ ergibt sich dann

$$\|v^j - z^j\|_\infty \leq \exp(C_2 \cdot j\Delta t) \cdot \|g^0\|_\infty + \sum_{k=1}^{j} \underbrace{(1 + C_2\Delta t)^{j-k}}_{\leq(1+C_2\Delta t)^j \leq \exp(C_2 j\Delta t)} \cdot C_1\Delta t \cdot \|g^k\|_\infty$$

$$
\begin{aligned}
&\leq \exp(C_2 \cdot \underbrace{j\Delta t}_{\leq T}) \cdot \|g^0\|_\infty \\
&\quad + C_2 \underbrace{j\Delta t}_{\leq T} \cdot \exp(C_2 \underbrace{j\Delta t}_{\leq T}) \cdot \max\{\|g^j\|_\infty \mid j = 1, \ldots, N\} \\
&\leq \exp(C_2 T)(1 + C_1 T)\|g\|_\infty \\
&= \exp(C_2 T)(1 + C_1 T)\|T^h(v) - T^h(z)\|_\infty, \quad j = 0, \ldots, N.
\end{aligned}
$$

Dies liefert die Behauptung mit der Stabilitätskonstanten $C := \exp(C_2T)(1+C_1T)$.

Fasst man nun alles zusammen, so ergibt sich der nachstehende Satz.

2.53 Satz. *Unter der Voraussetzung $f \in C^1(\mathbb{R} \times [0,1] \times [0,T], \mathbb{R})$, $q \leq \frac{\partial f}{\partial u}(u,x,t) \leq \mu$, $u \in \mathbb{R}$, $0 \leq x \leq 1$, $0 \leq t \leq T$ sowie der Schrittweitenbedingung $\Delta t \vartheta \mu \leq \kappa < 1$ ist das ϑ-Verfahren zur Anfangsrandwertaufgabe (2-42) durchführbar. Mit der zusätzlichen Bedingung*

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)} + \frac{q+\mu}{4}\Delta t$$

ist es auch bezüglich $\|\cdot\|_\infty$ stabil. Es gilt dann die Stabilitätsungleichung

$$\|u - w\|_\infty \leq \exp(C_2T)(1+C_1T)\|T^h(u) - T^h(w)\|_\infty, \quad \forall u, w \in \mathbb{R}^{\Omega_h}, \quad h = (\Delta x, \Delta t)$$

mit $C_1 = \frac{1}{1-\kappa}$, $C_2 = \max\{\mu, 0\}$.

An jeder klassischen Lösung $\overline{u}$ mit $\frac{\partial^\nu}{\partial t^\nu}\overline{u} \in C([0,1]\times[0,T])$, $\nu = 1,2$, $\frac{\partial^\nu}{\partial x^\nu}\overline{u} \in C([0,1]\times[0,T])$, $\nu = 1,2,3,4$ liegt die Konvergenz der Ordnung $O(\Delta t + \Delta x^2)$ bezüglich $\|\cdot\|_\infty$ vor. Für das Crank-Nicholson-Verfahren erhält man sogar die Konvergenz der Ordnung $O(\Delta t^2 + \Delta x^2)$ bezgl. $\|\cdot\|_\infty$, falls zusätzlich $\frac{\partial^3}{\partial t^3}\overline{u} \in C([0,1] \times [0,T])$.

f) Finite Elemente für nichtlineare Probleme

Vorgelegt sei zunächst das nichtlineare Poisson-Problem

$$
\begin{aligned}
-\Delta u &= f(u) \text{ in } \Omega, \\
u &= 0 \text{ auf } \partial\Omega
\end{aligned}
\tag{2-53}
$$

für ein beschränktes Gebiet $\Omega \subset \mathbb{R}^2$. Ist u eine klassische Lösung von (2-53), so folgt für alle $\varphi \in C_0^\infty(\Omega)$

$$
\begin{aligned}
0 = \int_\Omega -\Delta u \varphi - f(u)\varphi \,\mathrm{d}x &= \int_\Omega \nabla u \cdot \nabla \varphi - f(u)\varphi \,\mathrm{d}x - \int_{\partial\Omega} \frac{\partial u}{\partial n}\varphi \,\mathrm{d}S \\
&= \int_\Omega \nabla u \cdot \nabla \varphi - f(u)\varphi \,\mathrm{d}x.
\end{aligned}
$$

Wegen der Dichtheit von $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$ und der Stetigkeit des Funktionals $\langle \nabla u, \nabla \cdot \rangle - \langle f(u), \cdot \rangle$ auf $H_0^1(\Omega)$, folgt die variationelle Formulierung

$$\int_\Omega \nabla u \cdot \nabla v - f(u) v \,\mathrm{d}x = 0, \quad \forall v \in H_0^1(\Omega).$$

Das Galerkin-Verfahren hierzu lautet: Sei $V_h \subset V$. Finde ein $u_h \in V_h$ mit

$$\int_\Omega \nabla u_h \cdot \nabla v - f(u_h) v \,\mathrm{d}x = 0, \quad \forall v \in V_h.$$

Für $V_h = \mathrm{span}\{v_1, \ldots, v_M\}$, $u_h = \sum_{i=1}^M c_i v_i$ erhalten wir

$$\int_\Omega \sum_{j=1}^M c_j \nabla v_j \cdot \nabla v_i - f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}x = 0, \quad i = 1, \ldots, M,$$

was mit

$$\sum_{j=1}^M c_j \int_\Omega \nabla v_j \cdot \nabla v_i \,\mathrm{d}x = \int_\Omega f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}x, \quad i = 1, \ldots, M$$

äquivalent ist.

Mit

$$\begin{aligned} a : V \times V \to \mathbb{R}, \quad & a(v, w) = \int_\Omega \nabla v \cdot \nabla w \,\mathrm{d}x, \\ A = (a_{ij})_{1 \le i,j \le M} \in \mathbb{R}^{M,M}, \quad & a_{ij} = a(v_i, v_j), \quad 1 \le i, j \le M \end{aligned}$$

und

$$\begin{aligned} & G : \mathbb{R}^M \to \mathbb{R}^M, \\ & G(c) := (G_i(c))_{1 \le i \le M}, \quad G_i(c) = \int_\Omega f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}x, \quad 1 \le i \le M \end{aligned}$$

erhalten wir das nichtlineare Problem

$$T(c) := Ac - G(c) = 0.$$

Dieses löst man mit dem Newton-Verfahren, d.h. man benötigt dazu $DT(c) = A - DG(c)$. Man findet

$$\begin{aligned} \frac{\partial G_i}{\partial c_k}(c) &= \frac{\partial}{\partial c_k} \int_\Omega f\big(\sum_{j=1}^M c_j v_j\big) v_i \,\mathrm{d}x = \int_\Omega \frac{\partial}{\partial c_k} \Big(f\big(\sum_{j=1}^M c_j v_j\big) v_i \Big) \,\mathrm{d}x \\ &= \int_\Omega f'\big(\sum_{j=1}^M c_j v_j\big) \cdot v_k \cdot v_i \,\mathrm{d}x, \quad 1 \le i, k \le M. \end{aligned}$$

2.54 Bemerkung. $DG(c)$ ist nun keine Diagonalmatrix wie bei den Differenzenverfahren. Eine eventuell vorhandene L_0- bzw. M-Struktur von A wird also im Allgemeinen durch $DG(c)$ zerstört.

Dies ist unerwünscht und man geht deshalb in der Praxis auf andere Weise vor, indem man knotenorientierte Quadraturformeln zur Approximation des Vektorfeldes $G(c)$ benutzt.

Auf einer Triangulierung $\Omega_{T_h} = \{e_1, \ldots, e_r\}$ von Ω definiert man den Ansatzraum der Lagrangeschen Finiten Elemente

$$V_h = \{v \in C(\Omega_{T_h}) \,|\, v|_{e_k} \in P^r(e_k), \quad v|_{\partial\Omega_{T_h}} = 0\}.$$

Es sei ferner $V_h = \operatorname{span}\{v_1, \ldots, v_M\}$. Durch $v_1, \ldots, v_M$ und die Knoten $P_1, \ldots, P_M$ sei eine nodale Basis von V_h gegeben, d.h.

$$v_i(P_j) = \delta_{ij}, \quad 1 \le i, j \le M.$$

Wir benutzen eine Quadraturformel vom Typ

$$Q(g) = \sum_{i=1}^{M} w_i g(P_i)$$

für $g \in C(\overline{\Omega})$. Dabei erhält man eine derartige Formel durch

$$Q(g) := \int_\Omega I(g)\,\mathrm{d}x \text{ für } g \in C(\overline{\Omega})$$

mit dem Interpolationsoperator $I : C(\overline{\Omega}) \to V_h$, $I(g) := \sum_{j=1}^M g(P_j) v_j$. Man erhält dann also

$$\sum_{i=1}^{M} w_i f(P_i) = Q(f) = \int_\Omega I(f)\,\mathrm{d}x = \int_\Omega \sum_{i=1}^{M} f(P_i) v_i\,\mathrm{d}x = \sum_{i=1}^{M} f(P_i) \int_\Omega v_i\,\mathrm{d}x,$$

d.h. die Gewichte lauten $w_i = \int_\Omega v_i\,\mathrm{d}x$, $i = 1, \ldots, M$. Konkret bekommt man z.B. für lineare Finite Elemente

$$w_i = \frac{1}{3} \sum_{e \in \Omega_{T_h}, P_i \in e} \mu(e) > 0, \quad i = 1, \ldots, M.$$

Das Vektorfeld G mit $G_i(c) = \int_\Omega f\big(\sum_{j=1}^M c_j v_j\big) v_i\,\mathrm{d}x$, $1 \le i \le M$ wir dann ersetzt durch

$$\sum_{k=1}^{M} w_k \left(f\Big(\sum_{j=1}^{M} c_j v_j\Big) v_i \right)(P_k) \overset{v_i(P_k)=\delta_{ik}}{=} \sum_{k=1}^{M} w_i f\left(\Big(\sum_{j=1}^{M} c_j v_j\Big)(P_k) \right) \cdot \delta_{ik}$$

$$
\begin{aligned}
&= w_i f\left(\Big(\sum_{j=1}^{M} c_j v_j\Big)(P_i) \right) \\
&\overset{v_j(P_i)=\delta_{ji}}{=} w_i f\Big(\sum_{j=1}^{M} c_j \delta_{ji}\Big) = w_i f(c_i) =: \widetilde{G}_i(c).
\end{aligned}
$$

$\widetilde{G} : \mathbb{R}^M \to \mathbb{R}^M$, $\widetilde{G}_i(c) = w_i f(c_i)$, $i = 1, \ldots, M$ ist ein Diagonalfeld. Die Linearisierung lautet

$$
\frac{\partial(\widetilde{G}(c))_i}{\partial c_k} = \frac{\partial}{\partial c_k}(w_i f(c_i)) = w_i f'(c_i)\frac{\partial c_i}{\partial c_k} = w_i f'(c_i)\delta_{ik}, \quad 1 \le i, k \le M,
$$

d.h. $D\widetilde{G}(c) = \operatorname{diag}(w_i f'(c_i) \,|\, i = 1, \ldots, M)$ ist eine Diagonalmatrix. Zu lösen ist also das Problem

$$
\widetilde{T}(c) := Ac - \widetilde{G}(c) = 0
$$

mit der Linearisierung $D\widetilde{T}(c) = A - D\widetilde{G}(c)$, welche nun mit dem Newton-Verfahren gelöst wird. Wie bei den Differenzenverfahren, bleibt eine L_0- oder M-Struktur von A erhalten.

Nichtlineare parabolische Probleme

Nun wenden wir uns dem Lösen des nichtlinearen parabolischen Problems

$$
\begin{aligned}
u_t &= \Delta u + f(u) \text{ in } \Omega \times (0, T), \\
u(\cdot, t) &= 0 \text{ auf } \partial\Omega \times (0, T), \\
u(\cdot, 0) &= u_0 \text{ in } \Omega
\end{aligned}
\tag{2-54}
$$

zu, wobei $\Omega \subset \mathbb{R}^2$ eine beschränktes Gebiet ist. Die Nichtlinearität $f : \mathbb{R} \to \mathbb{R}$ sei Lipschitz-stetig. Für die Anfangsdaten gelte $u_0 \in L^2(\Omega)$.

Analog zum linearen Fall erhält man

$$
\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(x,t)v(x)\,\mathrm{d}x - \int_\Omega \nabla u(x,t) \cdot \nabla v(x)\,\mathrm{d}x = \int_\Omega f(u(x,t))v(x)\,\mathrm{d}x
$$

für alle $v \in C_0^\infty(\Omega)$, $t \in (0, T)$. Wir interpretieren wieder für jedes feste t die Abbildung

$$
x \mapsto u(t, x), \text{ d.h. } u(\cdot, t)
$$

als ein Element eines Sobolev-Raumes V (hier: $V = H_0^1(\Omega)$). Wir setzen

$$
\begin{aligned}
\langle f(u(t)), v\rangle &= \int_\Omega f(u(x,t))v(x)\,\mathrm{d}x, \quad 0 < t < T, \quad v \in H_0^1(\Omega), \\
a(v, w) &= \int_\Omega \nabla u(x) \cdot \nabla v(x)\,\mathrm{d}x, \quad v, w \in H_0^1(\Omega)
\end{aligned}
$$

und finden mit der Definition der verallgemeinerten Ableitung das folgende schwache Problem: Gesucht ist ein Element $u \in L^2((0,T),V)$, $V = H_0^1(\Omega)$, das eine schwache Ableitung $u' \in L^2((0,T),V')$ besitzt, mit $u(0) = u_0$ und

$$\langle u'(t), v\rangle_0 + a(u(t), v) = \langle f(u(t)), v\rangle_0, \quad \forall v \in V, \quad t \in (0,T). \tag{2-55}$$

2.55 Bemerkung. Es ist zu beachten, dass sich aus $u \in L^2((0,T),V)$, $u' \in L^2((0,T),V')$ mit dem Interpolationssatz $u \in C([0,T],H)$, $H = L^2(\Omega)$ ergibt. Mit einer Lipschitz-stetigen Abbildung $f : \mathbb{R} \to \mathbb{R}$ folgt dann, dass durch

$$\Omega \ni x \mapsto f(u(t))(x) = f(u(t,x))$$

eine Abbildung $f(u(\cdot)) \in C([0,T],H)$ definiert wird.

Das Galerkin-Verfahren zu (2-55) lautet: Sei $V_h \subset V$ endlich-dimensional und $u_{0h} \in V_h$ eine Approximation von u_0. Finde ein $u_h \in L^2((0,T),V_h)$ mit $u_h' \in L^2((0,T),V')$ und $u(0) = u_0$ sowie

$$\langle u_h'(t), v\rangle_0 + a(u_h(t), v) = \langle f(u_h(t)), v\rangle_0, \quad \forall v \in V_h, \quad t \in (0,T). \tag{2-56}$$

Unsere nächste Aufgabe ist es, die Lösbarkeit des Problems (2-56) zu diskutieren. Der numerische Ansatz zu $V_h = \operatorname{span}\{v_1, \ldots, v_M\}$ lautet

$$\begin{aligned} u_h(t) &= \sum_{i=1}^M c_i(t) v_i, \\ u_{0h}(t) &= \sum_{i=1}^M c_{i0} v_i \in V_h. \end{aligned}$$

Analog zum linearen Fall erhält man

$$\begin{aligned} \sum_{i=1}^M c_i'(t)\langle v_i, v_j\rangle_0 + \sum_{i=1}^M c_i(t) a(v_i, v_j) &= \Big\langle f\Big(\sum_{k=1}^M c_k(t) v_k\Big), v_j\Big\rangle, \quad j = 1, \ldots, M, \ 0 < t < T, \\ c(0) &= c_0. \end{aligned}$$

Mit der Steifigkeitsmatrix

$$A = (a_{ij})_{1 \le i,j \le M}, \quad a_{ij} = a(v_i, v_j), \quad 1 \le i, j \le M,$$

der Massenmatrix

$$B = (b_{ij})_{1 \le i,j \le M}, \quad b_{ij} = \langle v_i, v_j\rangle_0, \quad 1 \le i, j \le M$$

und dem Vektorfeld $G : \mathbb{R}^M \to \mathbb{R}^M$

$$G(a) := (G_j(a)), \quad G_j(a) = \int_\Omega f\Big(\sum_{k=1}^M a_k v_k\Big) v_j \,\mathrm{d}x, \quad 1 \le j \le M$$

erhalten wir das Sytem

$$\begin{aligned} Bc'(t) + Ac(t) &= G(c(t)), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-57}$$

Dabei ist $B \in \mathbb{R}^{M,M}$ symmetrisch und positiv definit, und die Abbildung $G : \mathbb{R}^M \to \mathbb{R}^M$ ist global Lipschitz-stetig, da $f : \mathbb{R} \to \mathbb{R}$ Lipschitz-stetig ist.

Das System (2-57) ist äquivalent zu

$$\begin{aligned} c'(t) &= B^{-1}(G(c(t)) - Ac(t)), \quad 0 < t < T, \\ c(0) &= c_0. \end{aligned} \tag{2-58}$$

Da die rechte Seite von (2-58) Lipschitz-stetig in c und stetig in t ist, existiert nach dem Satz von Picard und Lindelöf die Lösung c auf $[0, T]$. Somit ist das Galerkin-Problem (2-56) eindeutig lösbar.

Volldiskretes Problem

Wir diskretisieren nun (2-57) mit dem ϑ-Verfahren. Sei $\Delta t = \frac{T}{N}$, $\Omega_{T_h} = \{t_j = j\Delta t \,|\, j = 0, \ldots, N\}$ und sei c^j die Approximation für $c(t_j)$, $j = 0, \ldots, N$. Dies liefert

$$\begin{aligned} B\frac{c^{j+1} - c^j}{\Delta t} &= \vartheta(-Ac^{j+1} + G(c^{j+1})) + (1 - \vartheta)(-Ac^j + B(c^j)), \quad j = 0, \ldots, N - 1, \\ c^0 &\in \mathbb{R}^M \text{ gegeben.} \end{aligned}$$

Letztere Iteration ist gleichbedeutend mit

$$\begin{aligned} (B - \Delta t\vartheta A)c^{j+1} - \Delta t\vartheta G(c^{j+1}) &= Bc^j + \Delta t(1 - \vartheta)(-Ac^j + G(c^j)), \\ &\quad j = 0, \ldots, N - 1, \\ c^0 &\in \mathbb{R}^M \text{ gegeben.} \end{aligned} \tag{2-59}$$

(2-59) hat die Operatorform

$$T(c^{j+1}) = r^j, \quad j = 1, \ldots, N \tag{2-60}$$

mit

$$\begin{aligned} T(a) &= (B - \Delta t\vartheta A)a - \Delta t\vartheta G(a), \\ r^j &= Bc^j + \Delta t(1 - \vartheta)(-Ac^j + G(c^j)). \end{aligned}$$

Löst man (2-60) mit dem Newton-Verfahren, so wird die Linearisierung

$$DT(a) = (B - \Delta t\vartheta A) - \Delta t\vartheta DG(a)$$

benötigt. Wie im stationären Fall gilt dabei

$$\frac{\partial G_i}{\partial a_k}(a) = \int_\Omega f'\Big(\sum_{j=1}^{M} a_j v_j\Big) v_k \cdot v_i \,\mathrm{d}x, \quad 1 \le i, k \le M.$$

Somit zerstören die nichtdiagonalen Matrizen B und $DG(a)$ eine eventuell vorhandene Struktur von A. Eine Abhilfe verspricht jetzt die schon aus dem stationären Fall bekannte Technik des Mass-Lumping.

Auf einer Triangulierung $\Omega_{T_h} = \{e_1, \dots, e_r\}$ von Ω definieren wir wieder den Ansatzraum der Lagrangeschen Finiten Elemente

$$V_h = \{v \in C(\Omega_{T_h}) \,|\, v|_{e_k} \in P^r(e_k), \quad v|_{\partial\Omega_{T_h}} = 0\}.$$

Es sei ferner $V_h = \operatorname{span}\{v_1, \dots, v_M\}$. Durch $v_1, \dots, v_M$ und die Knoten $P_1, \dots, P_M$ sei eine nodale Basis von V_h gegeben, d.h.

$$v_i(P_j) = \delta_{ij}, \quad 1 \le i, j \le M.$$

Wir benutzen nun wieder die knotenorientierte Quadraturformel aus dem stationären Fall. Das Vektorfeld G mit $G_i(a) = \int_\Omega f\big(\sum_{j=1}^M a_j v_j\big) v_i \, \mathrm{d}x$, $1 \le i \le M$ wird dann durch

$$\sum_{k=1}^M w_i \big(f\big(\sum_{j=1}^M a_j v_j\big) v_i\big)(P_k) = w_i f(a_i) =: \widetilde{G}_i(a), \quad i = 1, \dots, M$$

ersetzt. Ferner wird die Matrix $B = (b_{ij})_{1 \le i,j \le M}$, $b_{ij} := \int_\Omega v_i v_j \, \mathrm{d}x$, $1 \le i, j \le M$ ersetzt durch

$$\sum_{k=1}^M w_i (v_j v_i)(P_k) \overset{v_i(P_k) = \delta_{ik}}{=} w_i v_j(P_i) = w_i \delta_{ji},$$

d.h. B wird durch $\widetilde{B} = \operatorname{diag}(w_i, i = 1, \dots, M)$ ersetzt.

Des Weiteren werden nur solche knotenorientierten Quadraturformeln verwendet, die $w_i > 0$, $i = 1, \dots, M$ sicherstellen, was im Falle linearer und quadratischer Finiter Elemente immer erfüllt ist. Benutzt man den Mass-Lumping-Prozess, so lautet das ϑ-Verfahren

$$\begin{aligned} (\widetilde{B} - \Delta t \vartheta A) c^{j+1} - \Delta t \vartheta \widetilde{G}(c^{j+1}) &= \widetilde{B} c^j + \Delta t (1 - \vartheta)(-A c^j + \widetilde{G}(c^j)), \\ & \quad\; j = 0, \dots, N-1, \qquad (2\text{-}61) \\ c^0 &\in \mathbb{R}^M \text{ gegeben.} \end{aligned}$$

Die Formelzeile (2-61) hat die Gestalt

$$\widetilde{T}(c^{j+1}) = \widetilde{r}^j, \quad j = 1, \dots, N \qquad (2\text{-}62)$$

mit

$$\begin{aligned} \widetilde{T}(a) &= (\widetilde{B} - \Delta t \vartheta A) a - \Delta t \vartheta \widetilde{G}(a), \\ \widetilde{r}^j &= \widetilde{B} c^j + \Delta t (1 - \vartheta)(-A c^j + \widetilde{G}(c^j)). \end{aligned}$$

Die Linearisierung lautet

$$D\widetilde{T}(a) = (\widetilde{B} - \Delta t\vartheta A) - \Delta t\vartheta D\widetilde{G}(a)$$

mit den Diagonalmatrizen

$$\begin{aligned} \widetilde{B} &= \operatorname{diag}(w_i, i = 1, \ldots, M), \\ D\widetilde{G}(a) &= \operatorname{diag}(w_i f_i(a), i = 1, \ldots, M). \end{aligned}$$

Offensichtlich gilt

$$D\widetilde{T}(a) = \widetilde{B} + O(\Delta t),$$

so dass für hinreichend kleine $\Delta t > 0$ die Matrix $D\widetilde{T}(a)$ immer invertierbar ist.

3. Hyperbolische Erhaltungsgleichungen

a) Theorie skalarer Gleichungen

In diesem Kapitel wenden wir uns hyperbolischen Erhaltungsgleichungen zu.

3.1 Definition. Sei $u = u(x,t) : \mathbb{R}^d \times \mathbb{R}_+ \longrightarrow \mathbb{R}^n$, und seien $F^i = F^i(u) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ für $i = 1, \ldots, d$. Dann heißt

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i}(F^i(u)) = 0 \tag{3-1}$$

ein System von n Erhaltungsgleichungen im $\mathbb{R}^d$ mit Flussfunktionen F^i, $i = 1, \ldots, d$.

3.2 Definition. Das System (3-1) heißt hyperbolisch, falls $A(u,w) = \sum_{i=1}^{d} w_i DF^i(u)$ für alle $u \in \mathbb{R}^n$ und $w \in \mathbb{R}^d \backslash \{0\}$ n reelle Eigenwerte $\lambda_i(u,w)$, $i = 1, \ldots, n$ und n linear unabhängige Eigenvektoren $v^i(u,w)$, $i = 1, \ldots, n$ hat. Sind überdies alle Eigenwerte paarweise verschieden, so heißt das System (3-1) strikt hyperbolisch.

3.3 Beispiel. Die Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u \text{ in } \mathbb{R}^d$$

lässt sich als ein System der Erhaltungsgleichungen umschreiben. Man setzt dazu

$$v_i := \frac{\partial u}{\partial x_i}, \; i = 1, \ldots, d, \quad v_{d+1} := \frac{\partial u}{\partial t}$$

und findet

$$\begin{aligned} \frac{\partial v_i}{\partial t} &= \frac{\partial^2 u}{\partial t \partial x_i} = \frac{\partial^2 u}{\partial x_i \partial t} = \frac{\partial v_{d+1}}{\partial x_i}, \quad i = 1, \ldots, d. \\ \frac{\partial v_{d+1}}{\partial t} &= \frac{\partial^2 u}{\partial t^2} = c^2 \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2} v_i. \end{aligned}$$

Mit $v = (v_1, \ldots, v_{d+1})^T$ erhalten wir dann

$$\frac{\partial v}{\partial t} = \underbrace{\begin{pmatrix} 0 & 0 & \ldots & 1 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ c^2 & 0 & \ldots & 0 \end{pmatrix}}_{=DF^1(v)} \frac{\partial v}{\partial x_1} + \underbrace{\begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & c^2 & \ldots & 0 \end{pmatrix}}_{=DF^2(v)} \frac{\partial v}{\partial x_2} + \cdots + \underbrace{\begin{pmatrix} 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ 0 & 0 & \ldots & c^2 \end{pmatrix}}_{=DF^d(v)} \frac{\partial v}{\partial x_d}$$

mit $DF^i(v) \in \mathbb{R}^{d+1,d+1}$, $i = 1,\dots,d$, d.h. ein System von $(d+1)$ Erhaltungsgleichungen in $\mathbb{R}^d$.

Die durch

$$A(v,w) := \begin{pmatrix} 0 & 0 & \dots & 0 & w_1 \\ 0 & 0 & \dots & 0 & w_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & w_d \\ w_1c^2 & w_2c^2 & \dots & w_dc^2 & 0 \end{pmatrix} \in \mathbb{R}^{d+1,d+1}.$$

definierte Matrix hat die Eigenwerte $\pm c\sqrt{\sum_{i=1}^d w_i^2}, \underbrace{0,\dots,0}_{(d-2)\text{-mal}}$ und linear unabhängige Eigenvektoren. Die Wellengleichung ist also hyperbolisch, aber für $d \geq 3$ nicht strikt hyperbolisch.

Existenz und Eindeutigkeit von Lösungen im skalaren Fall

Ein Cauchy-Problem für skalare Erhaltungsgleichungen besteht darin, eine Funktion $u : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ zu finden, welche

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) &= 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x,0) &= u_0(x), \quad x \in \mathbb{R} \end{aligned} \tag{3-2}$$

für ein vorgegebenes $f : \mathbb{R} \to \mathbb{R}$ erfüllt.

Im Fall von $f(u) = au$ spricht man von der Advektionsgleichung. Ist $f(u) = \frac{1}{2}u^2$, so handelt es sich bei (3-2) um eine Burgers' Gleichung.

Die Charakteristiken der Gleichung (3-2) findet man durch Lösen der Anfangswertaufgabe

$$x'(t) = f'(u(x(t),t)), \quad x(0) = x_0. \tag{3-3}$$

Ist $x(t)$ eine Lösung von (3-3), so ist u entlang der Kurve $(x(t),t)$ im (x,t)-Raum konstant, denn

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u(x(t),t) &= \frac{\partial u}{\partial x}(x(t),t)x'(t) + \frac{\partial u}{\partial t}(x(t),t) \\ &= \underbrace{\frac{\partial u}{\partial x}(x(t),t)f'(u(x(t),t))}_{=\frac{\partial}{\partial x}f(u(x(t),t))} + \frac{\partial u}{\partial t}(x(t),t) \;=\; 0, \end{aligned}$$

d.h. $u(x(t),t) = u(x_0,0) = u_0(x_0)$, $t \geq 0$.

Aus $x'(t) = f'(u(x(t),t)) = f'(u_0(x_0)) = \text{const}$ folgt, dass die Charakteristik $(x(t),t)$ eine Gerade im (x,t)-Raum sein muss. Diese hat die Darstellung

$$x = x(t) = x_0 + t \cdot f'(u_0(x_0)). \tag{3-4}$$

Kann man (3-4) für alle $(x,t) \in \mathbb{R} \times \mathbb{R}^+$ eindeutig nach x_0 auflösen, so hat man die Lösung

$$u(x,t) = u_0(x_0)$$

der Erhaltungsgleichung (3-2) gefunden. Dabei ist x_0 immer implizit durch

$$x = x_0 + f'(u_0(x_0)) \cdot t$$

gegeben.

3.4 Beispiel. Für die Advektionsgleichung gilt $f'(u) = a$. Es folgt $x = x_0 + at$, d.h. $x_0 = x - at$. Somit ergibt sich die Lösung $u(x,t) = u_0(x - at)$.

Leider ist dies im Allgemeinen nicht möglich, denn (3-4) kann nicht immer eindeutig aufgelöst werden, da die Charakteristiken sich nach einiger Zeit schneiden können. Wir analysieren diesen Effekt am Beispiel der Burgers' Gleichung

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}u^2\right) &= 0, \\ u(x,0) &= u_0(x) = -x. \end{aligned}$$

Mit $f(u) = \frac{u^2}{2}$ und $f'(u) = u$ ergeben sich aus (3-4) die Charakteristiken

$$x = x_0 + t\,\underbrace{u_0(x_0)}_{=-x_0} = x_0 - tx_0 = x_0(1-t), \tag{3-5}$$

woraus wir

$$\begin{aligned} t &= \frac{x_0 - x}{x_0} = 1 - \frac{x}{x_0} = -\frac{1}{x_0}x + 1, \\ x_0 &= \frac{x}{1-t} \end{aligned}$$

und somit

$$u(x,t) = u_0(x_0) = u_0\left(\frac{x}{1-t}\right) = \frac{-x}{1-t}.$$

(s. Abbildung 13) finden.

Jede Charakteristik auf der Abbildung (14) geht durch die Punkte $(x_0, 0)$ und $(0,1)$ im (x,t)-Raum. Man erhält also an den Schnittpunkten von Charakteristiken mehrwertige Lösungen der Gleichung. Eine klassische Lösung kann aber nicht mehrwertig sein. In diesem Fall bewegt sich die entstandene Unstetigkeit als schwache Lösung (Schock) fort.

3.5 Definition. Eine Funktion $u = u(x,t)$ heißt schwache Lösung des Cauchy-Problems (3-2), falls

$$\int_{\mathbb{R}} \int_0^\infty \left(u\frac{\partial \varphi}{\partial t} + f(u)\frac{\partial \varphi}{\partial x}\right) \mathrm{d}t\mathrm{d}x + \int_{\mathbb{R}} u_0(x)\varphi(x,0)\,\mathrm{d}x = 0$$

für alle $\varphi \in C_0^1(\mathbb{R} \times \mathbb{R}) = \{\varphi \in C^1(\mathbb{R} \times \mathbb{R}) \mid \operatorname{supp}\varphi \text{ kompakt}\}$ gilt.

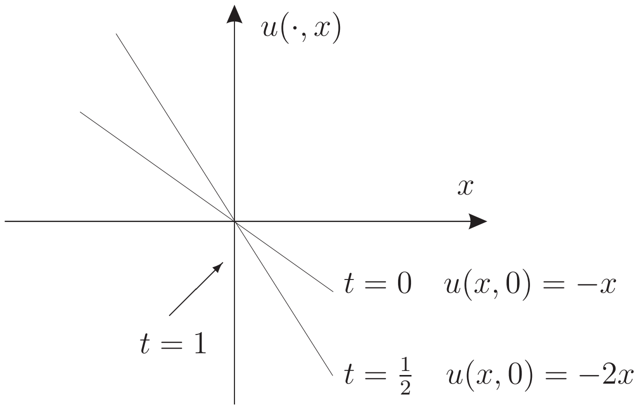


Abbildung 13: Beispiel einer Charakteristikenschar

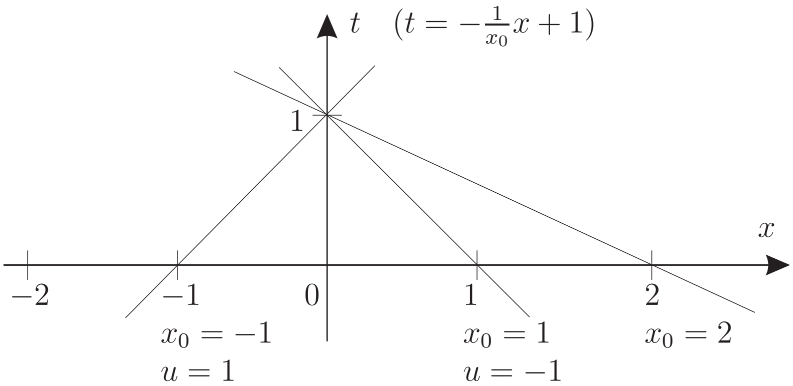


Abbildung 14: Beispiel einer Charakteristikenschar

3.6 Bemerkung. Ist u eine klassische Lösung, so auch eine schwache Lösung.

Eine typische Situation, bei der Unstetigkeiten generisch auftreten, sind Riemann-Probleme. Darunter versteht man Aufgaben, bei denen schon die Anfangsdaten unstetig sind, z.B.

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) = 0, \quad u(x,0) = u_0(x) = \begin{cases} u_l, & x < 0, \\ u_r, & x \geq 0. \end{cases}$$

Es lässt sich zeigen: Eine schwache Lösung des Riemann-Problems ist z.B. die Schockwelle

$$u(x,t) = u_0(x) = \begin{cases} u_l, & x < st, \\ u_r, & x > st. \end{cases}$$

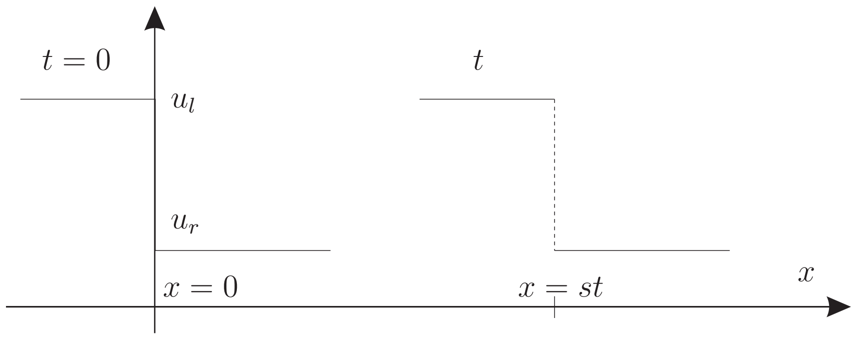
Hierbei ist s durch die Rankine-Hugoniot Bedingung

$$s = \frac{f(u_l) - f(u_r)}{u_l - u_r}$$

gegeben.

Für die Burgers' Gleichung gilt:

$$s = \frac{1}{2} \cdot \frac{u_l^2 - u_r^2}{u_l - u_r} = \frac{1}{2}(u_l + u_r).$$



Schockposition für $t = 0$ Schockposition für ein $t > 0$

Abbildung 15: Schockfortpflanzung

Eine weitere Lösung des Riemann-Problems der Burgers' Gleichung für $u_l < u_r$ ist gegeben durch die Verdünnungswelle

$$u(x,t) = \begin{cases} u_l, & \frac{x}{t} \leq u_l, \\ \frac{x}{t}, & u_l \leq \frac{x}{t} \leq u_r, \\ u_r, & \frac{x}{t} \geq u_r. \end{cases}$$

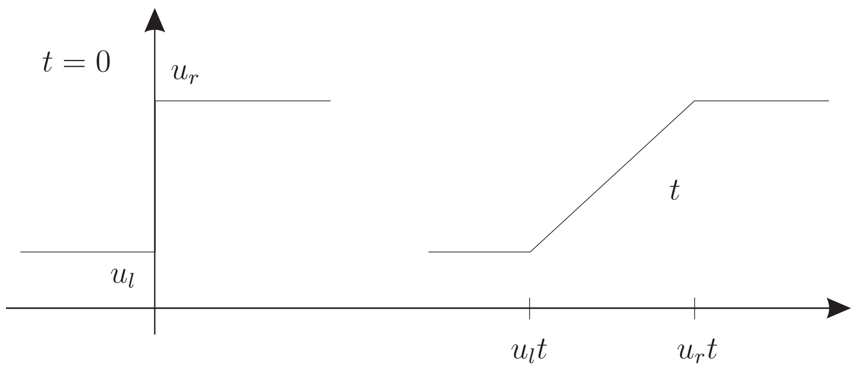


Abbildung 16: Verdünnungswelle der Burgers' Gleichung

Man braucht also weitere Bedingungen, um die eindeutige Lösbarkeit des Cauchy-Problems zu sichern. Betrachte dazu die viskose Regularisierung der Erhaltungsgleichung, d.h.

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) \quad = \quad \varepsilon \frac{\partial^2 u}{\partial x^2}, \qquad (x,t) \in \mathbb{R}_+ \times \mathbb{R}, \tag{3-6}$$

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}$$

für $0 < \varepsilon \ll 1$.

Die Aufgabe (3-6) ist nach Oleinik (1956) eindeutig in $L^\infty(\mathbb{R} \times \mathbb{R}_+)$ lösbar, falls $u_0 \in L^\infty(\mathbb{R})$ und $f \in C^1(\mathbb{R})$. Als Viskositätslösung bzw. Entropielösung der Erhaltungsgleichung bezeichnet man diejenige schwache Lösung, welche sich als Grenzwert $\varepsilon \to 0$ der Lösungen $(u_\varepsilon)_\varepsilon$ von (3-6) ergibt.

3.7 Definition. Es sei

$$L^\infty(\Omega) := \{f : \Omega \to \mathbb{R} \,|\, f \text{ Lebesgue-messbar und } \|f\|_{L^\infty} < \infty\}$$

mit

$$\|f\|_{L^\infty} = \operatorname{ess\,sup}_{x\in\Omega}|f(x)| = \inf_{\substack{\Lambda\subset\Omega\\ \mu(\Lambda)=0}} \sup_{x\in\Omega\setminus\Lambda} |f(x)|.$$

3.8 Definition. Es sei $u \in L^\infty(\Omega)$, $\Omega \subset \mathbb{R}$ offen. Dann ist die totale Variation von u durch

$$\mathrm{TV}(u) = \limsup_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_\Omega |u(x+\varepsilon) - u(x)| \,\mathrm{d}x.$$

definiert. Der Raum der Funktionen von totaler Variation ist dann

$$\mathrm{BV}(\Omega) = \{u \in L^\infty(\Omega) \,|\, \mathrm{TV}(u) < \infty\}.$$

Es gilt nun das folgende Theorem, das wir als Richtschnur zur Entwicklung numerischer Methoden nehmen werden.

3.9 Satz (Kružkov). *Das skalare Cauchy-Problem*

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) &= 0, \quad f \in C^1(\mathbb{R}), \\ u(x,0) &= u_0(x), \quad u_0 \in L^\infty(\mathbb{R}) \end{aligned}$$

hat eine eindeutige Entropielösung $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$, die die folgenden Eigenschaften hat:

i) $\|u(\cdot,t)\|_{L^\infty} \le \|u_0(\cdot)\|_{L^\infty}$, $t \ge 0$

ii) $u_0 \ge v_0 \implies u(\cdot,t) \ge v(\cdot,t)$, $t \ge 0$

iii) $u_0 \in \mathrm{BV}(\mathbb{R}) \implies u(\cdot,t) \in \mathrm{BV}(\mathbb{R})$ *und* $\mathrm{TV}(u(\cdot,t)) \le \mathrm{TV}(u_0)$, $t \ge 0$

iv) $u_0 \in L^1(\mathbb{R}) \implies \int_\mathbb{R} u(x,t)\,\mathrm{d}x = \int_\mathbb{R} u_0(x)\,\mathrm{d}x$, $t \ge 0$

3.10 Definition. Die Eigenschaften *i) — iv)* heißen L^∞-Stabilität, Monotonie, TV-Stabilität und Konservativität.

3.11 Bemerkung. Der Satz kann auf mehrere Dimensionen $x \in \mathbb{R}^d$, $d > 1$ erweitert werden, was aber leider für Systeme nicht stimmt.

b) Grundlagen und Notationen

Vorgelegt sei das Cauchy-Problem

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) &= 0, \quad x \in \mathbb{R},\, t \geq 0 \\ u(x,0) &= u_0(x), \quad x \in \mathbb{R}. \end{aligned}$$

Es sei für $\Delta x > 0$, $\Delta t > 0$ ein äquidistantes Gitter mit den Gitterpunkten (t_j, x_i), $j \in \mathbb{N}$, $i \in \mathbb{Z}$ eingeführt, wobei $t_j = j\Delta t$, $x_i = i\Delta x$.

Elementares Verfahren

Es bezeichne u_i^j die Annäherung für $u(x_i, t_j)$. Wir approximieren dann

$$\begin{aligned} \frac{\partial}{\partial x}(f(u))(x_i, t_j) &\sim \frac{f(u_{i+1}^j) - f(u_{i-1}^j)}{2\Delta x}, \\ \frac{\partial}{\partial t} u(x_i, t_j) &\sim \frac{u_i^{j+1} - u_i^j}{\Delta t}. \end{aligned}$$

Dies liefert

$$\begin{aligned} \frac{u_i^{j+1} - u_i^j}{\Delta t} + \frac{f(u_{i+1}^j) - f(u_{i-1}^j)}{2\Delta x} &= 0, \quad j \in \mathbb{N}, i \in \mathbb{Z}, \\ u_i^0 &= u_0(x_i), \quad i \in \mathbb{Z}, \end{aligned}$$

was sich als eine explizite Iteration

$$\begin{aligned} u_i^{j+1} &= u_i^j - \frac{\Delta t}{2\Delta x}\big(f(u_{i+1}^j) - f(u_{i-1}^j)\big) \\ &= H(u_{i-1}^j, u_i^j, u_{i+1}^j), \quad j \in \mathbb{N}, i \in \mathbb{Z}, \\ u_i^0 &= u_0(x_i), \quad i \in \mathbb{Z} \end{aligned}$$

mit der Verfahrensfunktion H umschreiben lässt.

In der Praxis rechnet man auf kompakten Gebieten, d.h. $x \in [0, L]$, $t \in [0, T]$ und z.B. mit periodischen Randbedingungen $u(0,t) = u(L,t)$, d.h. man wählt

$$\Delta x = \frac{L}{M}, \quad x_i = i\Delta x, \quad i = 1, \ldots, M,$$

$$\Delta t \;=\; \frac{T}{N}, \quad t_j = j\Delta t, \quad i = 0,\dots,N,$$

Es wird sich zeigen, dass das elementare Verfahren keine guten Stabilitätseigenschaften besitzt. Eine Abhilfe ist es, u_i^j durch $\frac{1}{2}(u_{i+1}^j + u_{i-1}^j)$ zu ersetzen. Dies liefert das sogenannte „Lax-Friedrichs-Verfahren"

$$\begin{aligned} u_i^{j+1} &= \frac{1}{2}(u_{i+1}^j + u_{i-1}^j) - \frac{\Delta t}{2\Delta x}\big(f(u_{i+1}^j) - f(u_{i-1}^j)\big) \\ &= H(u_{i-1}^j, u_i^j, u_{i+1}^j). \end{aligned}$$

Die Linienmethode

Betrachte

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) &= 0, \quad 0 \le x \le 1,\ t \ge 0, \\ u(0,t) &= u(1,t), \quad t \ge 0, \\ u(x,0) &= u_0(x), \quad 0 \le x \le 1. \end{aligned}$$

Wir wählen $\{x_i = i\Delta x \,|\, i = 1,\dots,M\}$, $\Delta x = \frac{1}{M}$ als Raumgitter und definieren den Gittervektor

$$v(t) = (u(t,x_1), u(t,x_2), \dots, u(t,x_M)) = (v_1(t), v_2(t), \dots, v_M(t)) \in \mathbb{R}^M$$

auf dem Zeitlevel t. Wir ersetzen nun $\frac{\partial u}{\partial t} = -\frac{\partial}{\partial x}(f(u))$ durch

$$v_i'(t) = -\left(\frac{f(v_{i+1}(t)) - f(v_{i-1}(t))}{2\Delta x}\right), \quad i = 2,\dots,M-1.$$

Mit den periodischen Randbedingungen folgt

$$\begin{aligned} v_1'(t) &= -\left(\frac{f(v_2(t)) - f(v_M(t))}{2\Delta x}\right), \quad (v_0 = v_M), \\ v_M'(t) &= -\left(\frac{f(v_1(t)) - f(v_{M-1}(t))}{2\Delta x}\right), \quad (v_{M+1} = v_1). \end{aligned}$$

Wir erhalten also ein System gewöhnlicher Differentialgleichungen der Form

$$v'(t) = F_{\Delta x}(v(t)), \quad v(0) = v^0$$

mit

$$F_{\Delta x}(v) \;=\; -\frac{1}{2\Delta x}\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & -1 \\ -1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} f(v_1) \\ f(v_2) \\ f(v_3) \\ \vdots \\ f(v_{M-2}) \\ f(v_{M-1}) \\ f(v_M) \end{pmatrix},$$

$$v^0 \quad = \quad (u_0(x_1), \ldots, u_0(x_M))^T \in \mathbb{R}^M.$$

Zur ersten Analyse der Stabilität unserer Verfahren wählen wir $f(u) = au$, d.h. die Advektionsgleichung

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$$

als Testgleichung. Dann erhalten wir als Liniensystem

$$v'(t) = A_{\Delta x} v(t), \quad v(0) = v^0 \tag{3-7}$$

mit

$$A_{\Delta x} \quad = \quad -\frac{a}{2\Delta x} \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & -1 \\ -1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & -1 & 0 \end{pmatrix} \in \mathbb{R}^{M,M}, \tag{3-8}$$

$$v^0 \quad = \quad (u_0(x_1), \ldots, u_0(x_M))^T \in \mathbb{R}^M.$$

3.12 Bemerkung. Es gilt $A_{\Delta x}^T = -A_{\Delta x}$, d.h. A ist schiefsymmetrisch, und es folgt $\sigma(A_{\Delta x}) \subset i\mathbb{R}$.

In unserem Fall kann man die Eigenschaften explizit angeben, denn $A_{\Delta x}$ hat die Eigenwerte

$$\lambda_k = \frac{ia}{\Delta x}\sin(2\pi k \Delta x), \quad 1 \leq k \leq M$$

und die Eigenvektoren w^k mit

$$(w^k)_l = \exp(2\pi i k l \Delta x), \quad 1 \leq k, l \leq M.$$

Die Eigenwerte liegen auf der imaginären Achse zwischen $-\frac{ia}{\Delta x}$ und $\frac{ia}{\Delta x}$.

Für die Zeitdisketisierung der Anfangswertaufgabe (3-7) ist also ein Verfahren empfehlenswert, für welches $\Delta t \lambda_k$ $(1 \leq k \leq M)$ im absoluten Stabilitätsbereich liegt.

Volldiskrete Modelle

a) Euler-Cauchy-Vefahren:

   Diskretisierung von (3-7) mit dem Euler-Cauchy-Verfahren liefert gerade das elementare Verfahren. Das Stabilitätsgebiet des Euler-Cauchy-Verfahrens ist gegeben durch

$$S = \{z \in \mathbb{C} \mid \; |z+1| < 1\}.$$

Man benötigt somit

$$|1 + \Delta t \lambda_k| \leq 1, \quad 1 \leq k \leq M. \tag{3-9}$$

Die Bedingung (3-9) ist jedoch wegen $\Delta t \lambda_k \subset i\mathbb{R}\backslash\{0\}$, $1 \leq k \leq M$ niemals erfüllbar.

3.13 Bemerkung. Das Verfahren kann aber konvergent sein für $\Delta t, \Delta x \to 0$, falls zusätzlich $\frac{\Delta t}{\Delta x} \to 0$ gilt. Dann folgt

$$\Delta t\, \lambda_k = \frac{\Delta t}{\Delta x} ia \sin(2\pi k \Delta x) \to 0 \text{ für } 1 \leq k \leq M$$

und die klassische Stabilität des Euler-Cauchy-Verfahrens, d.h. $0 \in S$ genügt dann eventuell, um die Konvergenz des elementaren Verfahrens zu sichern.

b) Lax-Friedrichs-Verfahren

Unter Beachtung von

$$\frac{1}{2}(u_{i+1}^j + u_{i-1}^j) = u_i^j + \frac{1}{2}(u_{i-1}^j - 2u_i^j + u_{i+1}^j)$$

ergibt sich

$$u_i^{j+1} = u_i^j - \frac{a\Delta t}{2\Delta x}(u_{i+1}^j - u_{i-1}^j) + \frac{1}{2}(u_{i-1}^j - 2u_i^j + u_{i+1}^j), \quad i = 1, \ldots, M,\ j = 0, \ldots, N$$

und $u_0^j = u_M^j$, $u_{M+1}^j = u_1^j$. Dies ist äquivalent mit

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} + a\frac{u_{i+1}^j - u_{i-1}^j}{2\Delta x} = \frac{\Delta x^2}{2\Delta t}\left(\frac{u_{i-1}^j - 2u_i^j + u_{i+1}^j}{\Delta x^2}\right) \tag{3-10}$$

für $i = 1, \ldots, M$, $j = 0, \ldots, N$. Das Schema (3-10) entsteht als eine Diskretisierung von

$$\begin{aligned} \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} &= \varepsilon\frac{\partial^2 u}{\partial x^2}, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= u_0(x), \quad 0 \leq x \leq 1, \\ u(0, t) &= u(1, t), \quad t \geq 0 \end{aligned}$$

mit $\varepsilon = \frac{\Delta x^2}{2\Delta t}$, d.h. der viskosen Regularisierung der Erhaltungsgleichung. (3-10) erlaubt auch, das Lax-Friedrichs-Verfahren in Linienform zu interpretieren.

Sei $v'(t) = B_{\Delta x} v(t)$, $v(0) = v^0$ mit

$$B_{\Delta x} = A_{\Delta x} - \varepsilon \cdot \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & -1 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ -1 & 0 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{M,M}. \tag{3-11}$$

Man erhält dann das Lax-Friedrichs-Verfahren (3-11), indem man (3-11) mit dem Euler-Cauchy-Verfahren diskretisiert. Im Fall $\varepsilon = \frac{\Delta x^2}{2\Delta t} \ll 1$ ist also die Matrix $B_{\Delta x}$ eine Störung der Matrix $A_{\Delta x}$. $B_{\Delta x}$ hat die Eigenwerte

$$\mu_k = \lambda_k - \frac{2\varepsilon}{\Delta x^2}\big(1 - \cos(2\pi k \Delta x)\big), \quad 1 \le k \le M,$$

wobei $\lambda_k = \frac{ia}{\Delta x}\sin(2\pi k\Delta x)$, $1 \le k \le M$ die Eigenwerte von $A_{\Delta x}$ sind, und die Eigenvektoren w^k, $1 \le k \le M$ von $A_{\Delta x}$. Für $\varepsilon = \frac{\Delta x^2}{2\Delta t}$ lässt sich zeigen $\Delta t\, \mu_k \in S$, d.h. $|1 + \Delta t\, \mu_k| < 1$, falls $\left|\frac{a\Delta t}{\Delta x}\right| \le 1$.

Die Erhaltungseigenschaften

Wir versuchen, die Erhaltungseigenschaft des Kružkov-Theorems auf den diskreten Fall zu übertragen.

Es seien $\Delta x, \Delta t > 0$, $x_i = i\Delta x$, $t_j = j\Delta t$, und u_i^j sei eine Approximation für die Lösung $u(x_i, t_j)$ der Erhaltungsgleichung. Alle unsere Verfahren haben die Form

$$\begin{aligned} u_i^{j+1} &= H(u_{i-1}^j, u_i^j, u_{i+1}^j), \quad j \in \mathbb{N}, i \in \mathbb{Z}, \\ u_i^0 &= u_0(x), \quad i \in \mathbb{Z}. \end{aligned}$$

3.14 Definition. Ein Differenzenverfahren $u_i^{j+1} = H(u_{i-1}^j, u_i^j, u_{i+1}^j)$ heißt in Erhaltungsform geschrieben, falls

$$H(u_{-1}, u_0, u_1) = u_0 - \frac{\Delta t}{\Delta x}(\hat{F}(u_0, u_1) - \hat{F}(u_{-1}, u_0))$$

mit einer Flussfunktion $\hat{F} : \mathbb{R}^2 \to \mathbb{R}$ gilt. Das numerische Verfahren heißt dann auch konservativ. Die numerische Flussfunktion $\hat{F}$ heißt konsistent mit der kontinuerlichen Flussfunktion f, falls $\hat{F}(u, u) = f(u)$.

3.15 Beispiel. Wir betrachten folgende Verfahren.

a) Elementares Verfahren:

$$\begin{aligned} u_i^{j+1} &= u_i^j - \frac{\Delta t}{2\Delta x}\big(f(u_{i+1}^j) - f(u_{i-1}^j)\big) \\ &= u_i^j - \frac{\Delta t}{2\Delta x}\left(\frac{1}{2}\big(f(u_i^j) + f(u_{i+1}^j)\big) - \frac{1}{2}\big(f(u_{i-1}^j) + f(u_i^j)\big)\right). \end{aligned}$$

Mithin ist

$$\hat{F}(u, v) = \frac{1}{2}(f(u) + f(v)), \quad \hat{F}(u, u) = f(u).$$

b) Lax-Friedrichs-Verfahren:

$$u_i^{j+1} = \frac{1}{2}(u_{i+1}^j + u_{i-1}^j) - \frac{\Delta t}{2\Delta x}\big(f(u_{i+1}^j) - f(u_{i-1}^j)\big).$$

Man erhält dann

$$\hat{F}(u,v) = \frac{\Delta x}{2\Delta t}(u-v) + \frac{1}{2}(f(u)+f(v)), \quad \hat{F}(u,u) = f(u).$$

Beide Verfahren sind also konservativ und konsistent.

3.16 Lemma. *Ist das Verfahren $u_i^{j+1} = H(u_{i-1}^j, u_i^j, u_{i+1}^j)$ in Erhaltungsform geschrieben, so gilt für alle $u^j = (u_i^j)_{i\in\mathbb{Z}} \in l^1$ und $\vec{H}u^j = (u_i^{j+1})_{i\in\mathbb{Z}} \in l^1$ mit $\vec{H} : l^1 \to l^1$ die Relation*

$$\sum_{i\in\mathbb{Z}} (\vec{H}u^j)_i = \sum_{i\in\mathbb{Z}} u_i^j.$$

3.17 Bemerkung. Bekanntlich ist

$$\begin{aligned} l^p &= \Big\{(v_i)_{i\in\mathbb{Z}} \,|\, \Big(\sum_{i\in\mathbb{Z}} |v_i|^p\Big)^{1/p} < \infty\Big\}, \quad 1 \le p < \infty, \\ l^\infty &= \Big\{(v_i)_{i\in\mathbb{Z}} \,|\, \sup_{i\in\mathbb{Z}} |v_i| < \infty\Big\}. \end{aligned}$$

Beweis: Es gilt

$$\begin{aligned} \sum_{i\in\mathbb{Z}} (\vec{H}u^j)_i &= \sum_{i\in\mathbb{Z}} u_i^{j+1} = \sum_{i\in\mathbb{Z}} H(u_{i-1}^j, u_i^j, u_{i+1}^j) \\ &= \sum_{i\in\mathbb{Z}} \Big[u_i^j - \frac{\Delta t}{\Delta x}\big(\hat{F}(u_i^j, u_{i+1}^j) - \hat{F}(u_{i-1}^j, u_i^j)\big)\Big] \\ &= \sum_{i\in\mathbb{Z}} u_i^j - \frac{\Delta t}{\Delta x}\underbrace{\Big(\sum_{i\in\mathbb{Z}} \hat{F}(u_i^j, u_{i+1}^j) - \sum_{i\in\mathbb{Z}} \hat{F}(u_{i-1}^j, u_i^j)\Big)}_{=0} \\ &= \sum_{i\in\mathbb{Z}} u_i^j, \quad j \in \mathbb{N}. \end{aligned}$$

□

3.18 Bemerkung. Im Fall periodischer Randbedingungen auf $[0, L]$ ist die Summe $\sum_{i\in\mathbb{Z}}$ durch $\sum_{i=1}^M$, $M \cdot \Delta x = L$ zu ersetzen.

Konsistenz von Verfahren

3.19 Definition. Ein Verfahren $u_i^{j+1} = H(u_{i-1}^j, u_i^j, u_{i+1}^j)$ hat die Konsistenzordnung p, falls gilt

$$\frac{1}{\Delta t}\big(\overline{u}(x, t+\Delta t) - H(\overline{u}(x-\Delta x, t), \overline{u}(x,t), \overline{u}(x+\Delta t.t))\big) = O(\Delta x^p) + O(\Delta t^p).$$

Dabei bezeichnet $\overline{u}$ eine klassische Lösung der Erhaltungsgleichung

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) = 0.$$

Für ein konservatives Verfahren mit Flussfunktion $\hat{F}$ und Konsistenzordnung p folgt

$$\begin{aligned}(Lu)(t,x) &= \frac{u(t+\Delta t, x) - u(t,x)}{\Delta t} + \frac{1}{\Delta x}\big(\hat{F}(u(x,t), u(x+\Delta x, t))\big)\\ &\quad -\hat{F}(u(x-\Delta x, t), u(x,t)) = O(\Delta x^p) + O(\Delta t^p).\end{aligned}$$

Für die Entwicklung des Konsistenzfehlers lässt sich zeigen:

3.20 Satz (ohne Beweis). *Sei H ein konservatives numerisches Verfahren für die Erhaltungsgleichung $\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) = 0$ mit konsistenter numerischer Flussfunktion $\hat{F}$. Sind $\hat{F} \in C^3(\mathbb{R}^2, \mathbb{R})$, $f \in C^3(\mathbb{R}, \mathbb{R})$ und $\lambda = \frac{\Delta t}{\Delta x}$ fest, so gilt für jede klassische Lösung*

$$(L\overline{u})(t,x) = -\Delta t \frac{\partial}{\partial x}\left(B(\overline{u}, \lambda)\frac{\partial \overline{u}}{\partial x}\right)(t,x) + O(\Delta t^2)$$

mit

$$B(u,\lambda) = \frac{1}{2}\left[\frac{1}{\lambda^2}\left(\left(\frac{\partial H}{\partial u_1} + \frac{\partial H}{\partial u_{-1}}\right)(u,u,u)\right) - (f'(u))^2\right].$$

Ist also H ein Verfahren erster Ordnung für $\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) = 0$, d.h. $B \neq 0$, dann ist H eine Approximation zweiter Ordnung von der parabolischen Differentialgleichung

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) = \Delta t \frac{\partial}{\partial x}\big(B(u,\lambda)\frac{\partial u}{\partial x}\big). \tag{3-12}$$

Die parabolische Gleichung (3-12) heißt „modifizierte Gleichung“. Das numerische Verhalten ist also immer näher an der modifizierten Gleichung (3-12) als an der Erhaltungsgleichung.

3.21 Definition. Der Term $\Delta t\, B$ heißt numerische Viskosität.

Aus Gründen der Stabilitätstheorie parabolischer Gleichungen sollte $\Delta t\, B > 0$ gelten und aus Konsistenzgründen sollte $\Delta t\, B$ nicht zu groß sein.

3.22 Beispiel. Wir betrachten für die Advektionsgleichung $\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$ die folgenden Verfahren.

a) Elementares Verfahren
Man findet $B(u, \lambda) = -\frac{a^2}{2} < 0$, was auf ein instabiles Verfahren hinweist.

b) Lax-Friedrichs-Verfahren
Hier gilt: $B(u, \lambda) = \frac{1}{\lambda^2}\big(1 - \big(\frac{a\Delta t}{\Delta x}\big)^2\big)$, d.h. $B(u, \lambda) > 0$, falls $\big|\frac{a\Delta t}{\Delta x}\big| < 1$. Dies liefert die Einschränkung $\frac{\Delta t}{\Delta x} < \frac{1}{|a|}$ an das numerische Gitter.

Die CFL-Bedingung

Vorgelegt sei die Advektionsgleichung $\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$. Die Charakteristiken der Advektionsgleichung werden durch

$$x'(t) = f'(u(x(t), t)) = a, \quad x(0) = x_0$$

gegeben, denn $f(u) = au$. Damit folgt $x = x(t) = x_0 + at$, d.h. $t = \frac{x - x_0}{a} = \frac{1}{a}x - \frac{x_0}{a}$. Dies stellt eine Gerade im (x, t)-Raum mit Steigung $\frac{1}{a}$ dar.

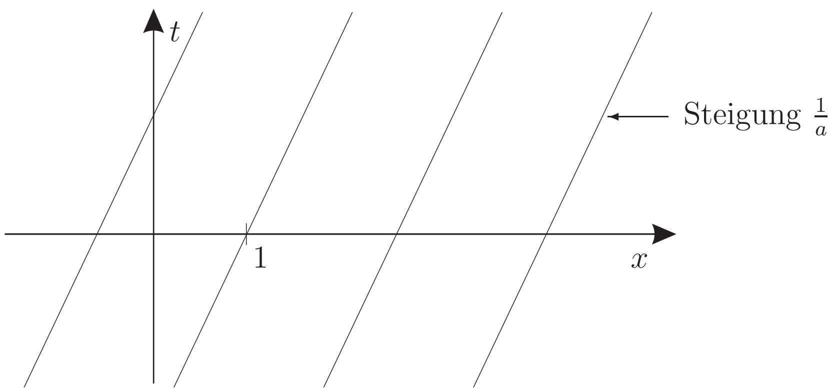


Abbildung 17: Charakteristiken der Advektionsgleichung

Die Information, gegeben durch die Anfangsbedingung, wird durch die kontinuierliche Gleichung mit Geschwindigkeit a transportiert. Im diskreten Verfahren mit Verfahrensfunktion H hat man Informationstransport mit der Geschwindigkeit $\pm\frac{\Delta x}{\Delta t}$.

Die Bedingung $\big|a\frac{\Delta t}{\Delta x}\big| < 1$, d.h. $|a| < \frac{x}{\Delta t}$, bedeutet also, dass der analytische Informationstransport langsamer ist als der diskrete Transport. Das numerische Schema kann dann Information mit der Geschwindigkeit a transportieren. Ist dies nicht der Fall, so kann das numerische Verfahren nicht konvergent sein.

Im nichtlinearen Fall $\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) = 0$ lautet die Bedingung:

$$\big|f'(v)\frac{\Delta t}{\Delta x}\big| < 1 \text{ für alle } v \in [-\|u\|_\infty, \|u\|_\infty], \tag{3-13}$$

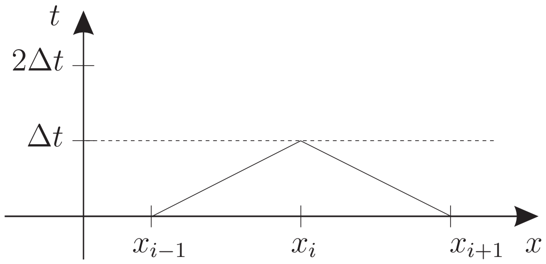


Abbildung 18: Charakteristiken der Advektionsgleichung

wobei u die wahre Lösung der Erhaltungsgleichung ist.

3.23 Definition. Die Bedingung (3-13) heißt die Courant-Friedrichs-Lewy-Bedingung oder kurz die CFL-Bedingung.

Die CFL-Bedingung ist also notwendig, aber nicht unbedingt hinreichend für die Konvergenz.

c) Konvergenztheorie für die Differenzenverfahren

Vorgelegt sei das Anfangswertproblem

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\big(f(u)\big) &= 0, \quad x \in \mathbb{R}, \quad t \geq 0, \\ u(x,0) &= u_0(x), \quad x \in \mathbb{R}. \end{aligned}$$

Zur Erstellung eines diskreten Modells wählt man die Schrittweiten $\Delta t, \Delta x > 0$. u_i^j bezeichne die numerische Approximation für $u(x_i, t_j)$ mit $x_i = i\Delta x$, $t_j = j\Delta t$ für $i \in \mathbb{Z}$ und $j \in \mathbb{N}$. Mit dieser Notation schreibt sich das Differenzenverfahren in Erhaltungsform:

$$u_i^{j+1} = H(u_{i-1}^j, u_i^j, u_{i+1}) = u_i^j - \lambda(\hat{F}(u_i^j, u_{i+1}^j) - \hat{F}(u_{i-1}^j, u_i^j)),$$

wobei $\lambda = \frac{\Delta t}{\Delta x}$ und $\hat{F}$ die numerische Flussfunktion mit $\hat{F}(u,u) = f(u)$ ist. Der Konsinstenzfehler lautet dann

$$\begin{aligned} (Lu)(t,x) &= \frac{1}{\Delta t}\big(u(t+\Delta t, x) - H(u(x-\Delta x, t), u(x,t), u(x+\Delta x, t))\big) \\ &= \frac{u(t+\Delta t, x) - u(t,x)}{\Delta t} + \frac{1}{\Delta x}\big(\hat{F}(u(x,t), u(x+\Delta x, t)) - \hat{F}(u(x-\Delta x, t), u(x,t))\big). \end{aligned}$$

Sei nun $v = (v_i)_{i\in\mathbb{Z}}$. Wir identifizieren die Folge v mit der stückweise konstanten Gitterfunktion v^c definiert durch $v^c(x) = v_i$ für $x \in \big[(i-\frac{1}{2})\Delta x, (i+\frac{1}{2})\Delta x\big)$, $i \in \mathbb{Z}$. Desweiteren setzen wir $u^c(t_j, x) = u_i^j$ für $x \in \big[(i-\frac{1}{2})\Delta x, (i+\frac{1}{2})\Delta x\big)$, $i \in \mathbb{Z}$. Ist nun $\overline{u}(t,x)$ die Lösung der Erhaltungsgleichung, so bezeichnet

$$e^{\Delta t}(t_j, x) = \overline{u}(t_j, x) - u^c(t_j, x), \quad x \in \mathbb{R}$$

den Fehler auf dem Zeitlevel $t = t_j$. Um das Differenzenvefahren in kompakter Form zu schreiben, setzen wir $u^j = (u_i^j)_{i\in\mathbb{Z}}$ und definieren $\vec{H}(u^j)$ durch $(\vec{H}(u^j))_i = H(u_{i-1}^j, u_i^j, u_{i+1}^j)$, $i \in \mathbb{Z}$. Ferner identifizieren wir u^j mit $u^c(t_j, \cdot)$ und schreiben ebenso

$$u^c(t_{j+1}, x) = \vec{H}(u^c(t_j, \cdot))(x), \quad x \in \mathbb{R}.$$

Wir untersuchen nun den Fehler für $t \in [0, T]$. Seien $\Delta t = \frac{T}{N} > 0$, $\Omega_{\Delta x} = \{t_j = j\Delta t \,|\, j = 0, \ldots, N\}$. Der Raum V aller stückweise konstanten Funktionen über x sei mit einer Norm $\|\cdot\|$ versehen. Wir betrachten zunächst den linearen Fall, d.h. die Advektionsgleichung.

3.24 Definition. Das Verfahren $u^{j+1} = \vec{H}(u^j)$ heißt stabil bezüglich $\|\cdot\|$, falls ein $C > 0$ mit

$$\|\vec{H}\| = \sup\left\{\frac{\|\vec{H}\|}{\|u\|} \,|\, u \neq 0, u \in V\right\} \leq 1 + C\Delta t$$

existiert. Es heißt stark stabil, falls $C = 0$.

Aus Definition 3.24 folgt $\|\vec{H}^j\| \leq \hat{C}$ für $j \in \mathbb{N}$ und $j\Delta t \leq T$, denn

$$\|\vec{H}^j\| \leq \|\vec{H}\|^j \leq (1 + C\Delta t)^j \leq \exp(C\Delta t)^j = \exp(Cj\Delta t) = \exp(CT) =: \hat{C}.$$

Dies wird häufig zur Definition der Stabilität eines Verfahrens zur Lösung der Advektionsgleichung herangezogen.

Für eine auf $\big[(i - \frac{1}{2})\Delta x, (i + \frac{1}{2})\Delta x\big)$, $i \in \mathbb{Z}$ stückweise konstante Funktion $v \in L^p(\mathbb{R})$ gilt

$$\|v\|_{L^p} = \left(\int_{\mathbb{R}} |v(x)|^p\right)^{1/p} = \left(\Delta x \sum_{i\in\mathbb{Z}} |v(x_i)|^p\right)^{1/p} = (\Delta x)^{1/p} \left(\sum_{i\in\mathbb{Z}} |v_i|^p\right) =: \|v\|_{l^p}.$$

Dies unterscheidet sich nur um den Faktor $(\Delta x)^{1/p} > 0$ von der klassischen Norm auf $l^p(\mathbb{Z})$ und erzeugt die selbe Normtopologie. Somit folgt $\|v\|_{L^p} = \|v\|_{l^p}$ für alle stückweise konstanten Gitterfunktionen aus $L^p(\mathbb{R})$.

3.25 Satz. *Vorgelegt sei das Differenzenverfahren $u^{j+1} = \vec{H}(u^j)$ mit einer konservativen und konsistenten Verfahrensfunktion H gegeben durch*

$$(\vec{H}(u^j))_i = H(u_{i-1}^j, u_i^j, u_{i+1}^j), \quad i \in \mathbb{Z}$$

für die Advektionsgleichung

$$\begin{aligned} \frac{\partial u}{\partial t} + a \cdot \frac{\partial u}{\partial x} &= 0, \quad x \in \mathbb{R}, \quad t \in [0, T], \\ u(x, 0) &= u_0(x), \quad x \in \mathbb{R}. \end{aligned}$$

Es gelte $\|\vec{H}\| \leq 1+C\Delta t$. *Auf dem Gitter* $\Omega_{\Delta t} = \{t_j = j\Delta t \,|\, j = 0,\ldots,N\}$, $N\Delta t = T$, *sei der Operator* $T_{\Delta t} : V^{\Omega_{\Delta t}} \longrightarrow V^{\Omega_{\Delta t}}$ *durch*

$$(T_{\Delta t})(u) = \left(u^0 - u_0, \frac{1}{\Delta t}\big(u^{j+1} - \vec{H}(u^j)\big),\ j = 0,\ldots,N\right)$$

für $u = (u^0,\ldots,u^N) \in V^{\Omega_{\Delta t}}$ *gegeben. Dann genügt* $T_{\Delta t}$ *einer Stabilitätsungleichung*

$$\|u - v\|_\infty \leq C \cdot \|T_{\Delta t}(u) - T_{\Delta t}(v)\|_\infty,$$

wobei $\|u\|_\infty = \|(u^0,\ldots,u^N)\|_\infty = \max\{\|u^j\| \,|\, j = 0,\ldots,N\}$.

Beweis: Seien $v = (v^j)_{j\in\mathbb{Z}}$, $w = (w^j)_{j\in\mathbb{Z}}$ beliebig. Dann erhalten wir

$$T^h(v) - T^h(w) = g = (g^0,\ldots,g^N)$$

mit $g^0 = v^0 - w^0$ und $g^j = \frac{1}{\Delta t}(v^j - \vec{H}(v^j) - (w^j - \vec{H}(w^j)))$, $j = 1,\ldots,N$. Letzteres ist gleichbedeutend mit $g^0 = v^0 - w^0$ und $\Delta t g^j + \vec{H}(v^{j-1}) - \vec{H}(w^{j-1}) = v^j - w^j$, $j = 1,\ldots,N$. Dies liefert

$$\|v^j - w^j\| \leq \|\underbrace{\vec{H}(v^{j-1}) - \vec{H}(w^{j-1})}_{=\vec{H}(v^{j-1}-w^{j-1})}\| + \Delta t\|g^j\| \leq (1+C\Delta t)\|v^{j-1} - w^{j-1}\| + \Delta t\|g^j\|.$$

Wie im parabolischen Fall (Vgl. Kapitel 2, Abschnitt e)) folgt dann induktiv

$$\|v^j - w^j\| \leq (1+C\Delta t)^j\|g^0\| + \sum_{k=1}^{j}(1+C\Delta t)^{j-k}\cdot\Delta t\|g^k\|$$

und unter Beachtung von $1 + C\Delta t \leq \exp(C\Delta t)$ ergibt sich ferner

$$\|v^j - w^j\| \leq \exp(CT)(1+T)\underbrace{\max\{\|g^k\| \,|\, k = 0,\ldots,N\}}_{=\|g\|_\infty}.$$

Somit ist die Stabilitätsungleichung bewiesen. □

3.26 Bemerkung. Betrachtet man nun den nichtlinearen Fall, d.h. die Erhaltungsgleichung

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\big(f(u)\big) &= 0, \quad x \in \mathbb{R}, \quad t \geq 0, \\ u(x,0) &= u_0(x), \quad x \in \mathbb{R}, \end{aligned}$$

so benötigt man für das numerische Verfahren $u^{j+1} = \vec{H}(u^j)$ die Stabilitätsungleichung

$$\|\vec{H}(v^j) - \vec{H}(w^j)\| \leq (1+C\Delta t)\|v^j - w^j\|. \tag{3-14}$$

Der Beweis von Satz 3.25 lässt sich dann auch auf den nichtlinearen Fall übertragen.

3.27 Satz (Konvergenzresultat). *Sei $u^{j+1} = \vec{H}(u^j)$ ein Differenzenverfahren für die Erhaltungsgleichung*

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u)) &= 0, \quad x \in \mathbb{R}, \quad t \geq 0, \\ u(x,0) &= u_0(x), \quad x \in \mathbb{R}, \end{aligned}$$

und es existiere eine eindeutige klassische Lösung $\overline{u}$ für $x \in \mathbb{R}$ und $t \in [0,T]$. $\vec{H}$ sei stabil bezüglich $\|\cdot\|$ im Sinne von (3-14) und konsistent der Ordnung p. Dann ist das Verfahren konvergent der Ordnung p bezüglich $\|\cdot\|$.

Die von Neumannsche Stabilität

Wir untersuchen nun die Stabilität der linearen Advektionsgleichung bezüglich der $\|\cdot\|_{L^2(\mathbb{R})}$-Norm definiert durch

$$\|v\|_{L^2} = \left(\int_{\mathbb{R}} |v(x)|^2 \,\mathrm{d}x \right)^{1/2}$$

für $v \in L^2(\mathbb{R})$. Wir wählen ein festes $\Delta x > 0$ und interpretieren

$$\begin{aligned} (\vec{H}v)(x) &:= H(v(x-\Delta x), v(x), v(x+\Delta x)) \\ &= c_{-1}v(x-\Delta x) + c_0 v(x) + c_1 v(x+\Delta x), \quad c_{-1}, c_0, c_1 \text{ fest} \end{aligned}$$

als einen Operator auf $L^2(\mathbb{R})$, d.h. $\vec{H} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$.

Im Fall der L^2-Norm kann die Stabilität leicht durch eine Fourieranalyse überprüft werden. Zu einem $v \in L^2(\mathbb{R})$ ist $\mathcal{F}(v) \in L^2(\mathbb{R})$ gegeben durch

$$\mathcal{F}(v)(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(x) \exp(-ix\xi) \,\mathrm{d}\xi, \quad \xi \in \mathbb{R},$$

die Fouriertransformierte von v. Nach dem Parsevalschen Satz gilt

$$\|\mathcal{F}(v)\|_{L^2} = \|v\|_{L^2}.$$

Außerdem lässt sich die Fouriertransformierte der Translationsabbildung $v_y(\cdot) = v(\cdot + y)$ ($y \in \mathbb{R}$ fest) leicht berechnen gemäß

$$\mathcal{F}(v_y)(\xi) = \exp(i\xi y)\mathcal{F}(v)(\xi).$$

Damit folgt für ein Verfahren in Erhaltungsform für die Advektionsgleichung, d.h. für

$$u_i^{j+1} = H(u_{i-1}^j, u_i^j, u_{i+1}^j) = c_{-1}u_{i-1}^j + c_0 u_i^j + c_1 u_{i+1}^j \tag{3-15}$$

bzw.

$$(\vec{H}v)(x) = H(v(x-\Delta x), v(x), v(x+\Delta x)) = c_{-1}v(x-\Delta x) + c_0 v(x) + c_1 v(x+\Delta x) \tag{3-16}$$

wegen der Linearität der Fouriertransformation sofort

$$\begin{aligned}
\mathcal{F}(\vec{H}v)(x) &= c_{-1}\mathcal{F}(v_{-\Delta x})(\xi) + c_0\mathcal{F}(v)(\xi) + c_1\mathcal{F}(v_{\Delta x})(\xi) \\
&= c_{-1}\exp(-i\Delta x\xi)\mathcal{F}(\xi) + c_0\mathcal{F}(v)(\xi) + c_1\exp(i\Delta x\xi)\mathcal{F}(v)(\xi) \\
&= \big(\underbrace{c_{-1}\exp(-i\Delta x\xi) + c_0 + c_1\exp(i\Delta x\xi)}_{=\rho(\xi)}\big)\mathcal{F}(v)(\xi) \\
&= \rho(\xi)\mathcal{F}(v)(\xi).
\end{aligned}$$

Der Term $\rho(\xi) = c_{-1}\exp(-i\Delta x\xi) + c_0 + c_1\exp(i + \Delta x\xi)$ heißt das Symbol von $\vec{H}$. Wegen

$$\begin{aligned}
\|\vec{H}v\|_{L^2} = \|\mathcal{F}(\vec{H}v)\|_{L^2} = \|\rho\cdot\mathcal{F}(v)\|_{L^2} &\le \sup\{|\rho(\xi)| \ \mid \xi\in\mathbb{R}\}\cdot\|\mathcal{F}(v)\|_{L^2} \\
&= \sup\{|\rho(\xi)| \ \mid \xi\in\mathbb{R}\}\cdot\|v\|_{L^2}
\end{aligned}$$

folgt

$$\|\vec{H}\|_{L^2} \le \sup\{|\rho(\xi)| \ \mid \xi\in\mathbb{R}\} = \|\rho\|_{L^\infty}.$$

3.28 Satz (von Neumann). *Das lineare Verfahren $\vec{H}$ der Form (3-15) ist L^2-stabil, falls*

$$\sup\{|\rho(\xi)| \ \mid \xi\in\mathbb{R}\} \le 1.$$

3.29 Beispiel. Wir wenden nun den von Neumannschen Satz auf die folgenden Verfahren an.

a) Elementares Verfahren:

$$u_i^{j+1} = u_i^j - \frac{\Delta t a}{2\Delta x}(u_{i+1}^j - u_{i-1}^j).$$

Man erhält $\rho(\xi) = 1 - \frac{\Delta ta}{\Delta x} - i\sin(\Delta x\xi)$ und damit

$$|\rho(\xi)|^2 = 1 + \frac{\Delta t^2 a^2}{\Delta x^2}\sin^2(\Delta x\xi) > 1.$$

b) Lax-Friedrichs-Verfahren:

$$u_i^{j+1} = \frac{u_{i+1}^j + u_{i-1}^j}{2} - \frac{\Delta t a}{2\Delta x}(u_{i+1}^j - u_{i-1}^j).$$

Es folgt $\rho(\xi) = \cos(\Delta x\xi) - i\frac{\Delta t}{\Delta x}a\sin(\Delta x\xi)$ und somit

$$\cos^2(\Delta x\xi) + \left(\frac{\Delta t}{\Delta x a}\right)^2\sin^2(\Delta x\xi).$$

Damit gilt

$$\sup\{|\rho(\xi)| \mid \ \xi\in\mathbb{R}\} \le 1,$$

falls $\left(\frac{\Delta t}{\Delta xa}\right)^2 \le 1$. Also ist das Lax-Friedrichs-Verfahren unter der CFL-Bedingung L^2-stabil.

L^∞-Stabilität und TVD

Wir kommen nun zu zwei weiteren Eigenschaften, die ein „gutes“ numerisches Verfahren für die Erhaltungsgleichungen haben sollte. Für die Entropielösung des Kružkov-Theorems gilt

$$\|u(\cdot,t)\|_{L^\infty} \le \|u_0(\cdot)\|_{L^\infty}, \quad t \ge 0 \qquad (L^\infty\text{-Stabilität})$$

und

$$\mathrm{TV}(u(\cdot,t)) \le \mathrm{TV}(u_0(\cdot)), \quad t \ge 0. \qquad (\text{TV-Stabilität})$$

3.30 Bemerkung. Ist $v(x)$ eine auf $\big[(i-\frac{1}{2})\Delta x, (i+\frac{1}{2})\Delta x\big)$, $i \in \mathbb{Z}$ stückweise konstante Funktion zu $(v_i)_{i\in\mathbb{Z}}$, so lässt sich funktionalanalytisch zeigen:

$$\mathrm{TV}(v) = \sum_{i\in\mathbb{Z}} |v_{i+1} - v_i| = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_{\mathbb{R}} |v(x+\varepsilon) - v(x)| \,\mathrm{d}x$$

und

$$\|v\|_{L^\infty} = \sup\{|v_i| \mid i \in \mathbb{Z}\}.$$

Man definiert nun die diskrete L^∞-Stabilität und TVD wie folgt.

3.31 Definition. Das durch $\vec{H}$ definierte Schema ist stark L^∞-Stabil, falls

$$\|\vec{H}v\|_{L^\infty} \le \|v\|_{L^\infty}$$

für alle auf $\big[(i-\frac{1}{2})\Delta x, (i+\frac{1}{2})\Delta x\big)$, $i \in \mathbb{Z}$ stückweise konstanten $v \in L^\infty(\mathbb{R})$.

3.32 Definition. Das durch $\vec{H}$ definierte Schema ist TVD (total variation diminishing), falls

$$\mathrm{TV}(\vec{H}v) \le \mathrm{TV}(v)$$

für alle auf $\big[(i-\frac{1}{2})\Delta x, (i+\frac{1}{2})\Delta x\big)$, $i \in \mathbb{Z}$ stückweise konstanten $v \in \mathrm{BV}(\mathbb{R})$.

3.33 Beispiel (Lax-Friedrichs-Vefahren). Betrachte das Lax-Friedrichs-Vefahren

$$u_i^{j+1} = H(u_{i+1}^j, u_i^j, u_{i-1}^j) = \frac{u_{i+1}^j + u_{i-1}^j}{2} - \lambda \frac{f(u_{i+1}^j) - f(u_{i-1}^j)}{2}$$

mit $\lambda = \frac{\Delta t}{\Delta x}$ und einem stetigen $f \in C^1(\mathbb{R})$. Es gilt

$$\begin{aligned} u_i^{j+1} &= \frac{u_{i+1}^j + u_{i-1}^j}{2} - \frac{u_{i+1}^j - u_{i-1}^j}{2} \cdot \lambda \cdot \frac{f(u_{i+1}^j) - f(u_{i-1}^j)}{u_{i+1}^j - u_{i-1}^j} \\ &= \frac{u_{i+1}^j + u_{i-1}^j}{2} - \frac{u_{i+1}^j - u_{i-1}^j}{2} \cdot \lambda \cdot f'(\xi_i^j) \end{aligned}$$

mit $\xi_i^j \in [\min\{u_{i-1}^j, u_{i+1}^j\}, \max\{u_{i-1}^j, u_{i+1}^j\}]$. Mit der CFL-Bedingung folgt $|\lambda \cdot f'(\xi_i^j)| \leq 1$, $i \in \mathbb{Z}$, $j \in \mathbb{N}$ und somit

$$u_i^{j+1} = \frac{u_{i+1}^j + u_{i-1}^j}{2} + \alpha_i^j \frac{u_{i+1}^j - u_{i-1}^j}{2}, \quad |\alpha_i^j| \leq 1,$$

woraus sich $u_i^{j+1} \in (\min\{u_{i-1}^j, u_{i+1}^j\}, \max\{u_{i+1}^j, u_{i-1}^j\})$ ergibt, d.h. wir erhalten

$$|u_i^{j+1}| \leq \max\{|u_{i-1}^j|, |u_{i+1}^j|\},$$

und es folgt

$$\begin{aligned} \|\vec{H}u^j\|_{L^\infty} &= \|u^{j+1}\|_{L^\infty} = \sup\{|u_i^{j+1}| \mid i \in \mathbb{Z}\} \\ &\leq \sup\{\max\{|u_{i-1}^j|, |u_{i+1}^j|\} \mid i \in \mathbb{Z}\} = \|u^j\|_{L^\infty}. \end{aligned}$$

Also ist das Lax-Friedrichs-Verfahren stark L^∞-stabil, falls die CFL-Bedingung gilt. Es lässt sich überdies zeigen, dass die CFL-Bedingung auch die TVD-Stabilität des Lax-Friedrichs-Verfahrens impliziert.

3.34 Beispiel (Nichlineares Upwind-Verfahren). Das nichtlineare Upwind-Verfahren hat die Form

$$\begin{aligned} u_i^{j+1} &= u_i^j - \frac{\lambda}{2}\big(f(u_{i+1}^j) - f(u_{i-1}^j)\big) \\ &\quad + \frac{\lambda}{2}\big(|a(u_i^j, u_{i+1}^j)| \cdot (u_{i+1}^j - u_i^j) - |a(u_{i-1}^j, u_i^j)| \cdot (u_i^j - u_{i-1}^j)\big) \end{aligned}$$

mit $\lambda = \frac{\Delta t}{\Delta x}$ und $a(u,v) = \frac{f(u)-f(v)}{u-v}$. Das Upwind-Verfahren ist ein Verfahren in Erhaltungsform mit der Flussfunktion

$$\hat{F}(u,v) = \frac{1}{2}(f(u) + f(v)) - \frac{|a(u,v)|}{2}(v-u),$$

d.h.

$$\hat{F}(u,v) = \begin{cases} f(u), & \text{falls } a(u,v) > 0, \\ f(v), & \text{falls } a(u,v) < 0 \end{cases}$$

sowie $\hat{F}(u,u) = \frac{1}{2}(f(u) + f(u)) = f(u)$. Es lässt sich ferner zeigen, dass das Upwind-Verfahren unter der CFL-Bedingung TVD$-$ und L^∞-stabil ist.

Konvergenz gegen schwache Lösungen

Wir untersuchen das Verhalten von konservativen Verfahren bei schwachen Lösungen.

3.35 Satz (Lax-Wendroff). *Es sei durch $u^{j+1} = \vec{H}(u^j)$ ein konsistentes und konservatives numerisches Verfahren für die Gleichung*

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\big(f(u)\big) = 0, \quad x \in \mathbb{R}, \quad t \geq 0,$$

$$u(x,0) \quad = \quad u_0(x), \quad x \in \mathbb{R}$$

gegeben. Betrachte eine Folge $\{\Omega_n\}_{n\in\mathbb{N}}$ von Gittern mit den Schrittweiten $\{(\Delta t_n, \Delta x_n)\}_{n\in\mathbb{N}}$, für die $\Delta t_n \to 0$ und $\frac{\Delta t_n}{\Delta x_n} = \lambda$, $n \in \mathbb{N}$ mit einer festen Gitterkonstanten $\lambda > 0$ gilt. Die numerische Lösung zur Schrittweite Δt sei $u^c_{\Delta t}$, d.h. $u^c_{\Delta t}(t,x) = u^j_i$, $x \in \big[(i-\frac{1}{2})\Delta x, (i+\frac{1}{2})\Delta x\big)$, $i \in \mathbb{Z}$, $t \in \big[(j-\frac{1}{2})\Delta t, (j+\frac{1}{2})\Delta t\big)$, $j \in \mathbb{N}$ sei als eine konstante Gitterfunktion aufgefasst. Angenommen, die Folge $(u^c_{\Delta t_n})_{n\in\mathbb{N}}$ erfüllt

$$\forall T > 0 \quad \exists R > 0 \quad \forall n \in \mathbb{N} \quad \mathrm{TV}(u^c_{\Delta t_n}) < R, \quad 0 \le t \le T \tag{3-17}$$

und es gibt ein $\overline{u}$ so, dass auf jedem Kompaktum $Q = [a,b] \times [0,T]$

$$\int_0^T \int_a^b |u^c_{\Delta t_n}(x,t) - \overline{u}(x,t)| \,\mathrm{d}x\mathrm{d}t \to 0 \qquad \textit{für } n \to \infty \tag{3-18}$$

gilt. Dann ist $\overline{u}$ eine schwache Lösung der Erhaltungsgleichung.

3.36 Bemerkung. Jedes TVD-Verfahren erfüllt die Voraussetzung des Satzes von Lax und Wendroff, d.h. konvergiert auch an einer schwachen Lösung. Aus (3-18) folgt dann $\|u^c_{\Delta t_n} - \overline{u}\|_{l^1(\mathbb{Q})} \to 0$ für $n \to \infty$.

Beweisskizze: Das numerische Verfahren lautet

$$u_i^{j+1} = H(u^j_{i-1}, u^j_i, u^j_{i+1}) = u^j_i - \frac{\Delta t}{\Delta x}\big[\hat{F}(u^j_i, u^j_{i+1}) - \hat{F}(u^j_{i-1}, u^j_i)\big].$$

Sei nun $\varphi \in C^1_0(\mathbb{R}\times\mathbb{R})$ beliebig, aber fest. Multipliziere die Verfahrensvorschrift mit $\varphi(x_i, t_j)$ und finde

$$\varphi(x_i,t_j)u_i^{j+1} = \varphi(x_i,t_j)u^j_i - \frac{\Delta t}{\Delta x}\varphi(x_i,t_j)\big[\hat{F}(u^j_i, u^j_{i+1}) - \hat{F}(u^j_{i-1}, u^j_i)\big], \quad i \in \mathbb{Z},\ j \in \mathbb{N}.$$

Summiert man die obige Identität über i und j, so ergibt sich

$$\sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}} \varphi(x_i,t_j)(u_i^{j+1} - u^j_i) + \frac{\Delta t}{\Delta x}\sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}} \varphi(x_i,t_j)\big[\hat{F}(u^j_i, u^j_{i+1}) - \hat{F}(u^j_{i-1}, u^j_i)\big] = 0,$$

wobei die Summen wegen der Kompaktheit von supp φ endlich sind.

Ferner benötigen wir die Formel der partiellen Summation:

$$\sum_{i=r}^{m} a_i(b_i - b_{i-1}) = a_m b_m - a_r b_{r-1} - \sum_{i=r}^{m-1}(a_{i+1} - a_i)b_i, \quad r < m.$$

deren Beweis durch Induktion erbracht wird.

$m = r$: Trivialerweise bekommt man

$$\sum_{i=r}^{r} a_i(b_i - b_{i-1}) = a_r - a_r b_{r-1} = a_r - a_r b_{r-1} - \underbrace{\sum_{i=r}^{r-1}(a_{i+1} - a_i)}_{=0}.$$

$m \to m+1$: Es ergibt sich

$$\begin{aligned}
\sum_{i=r}^{m+1} a_i(b_i - b_{i-1}) &= \sum_{i=r}^{m} a_i(b_i - b_{i-1}) + a_{m+1}b_{m+1} - a_{m+1}b_m \\
&\overset{\text{(IV)}}{=} a_m b_m - a_r b_{r-1} - \sum_{i=r}^{m}(a_{i+1} - a_i)b_i + a_{m+1}b_{m+1} - a_{m+1}b_m \\
&= a_{m+1}b_{m+1} - a_r b_{r-1} - \sum_{i=r}^{m+1}(a_{i+1} - a_i)b_i.
\end{aligned}$$

Wir wenden nun partielle Summation an und finden

$$\sum_{j=0}^{\infty} \underbrace{\varphi(x_i, t_j)}_{=:a_j}\underbrace{(u_i^{j+1} - u_i^j)}_{=:b_j - b_{j-1}} = -\varphi(x_i, t_0)u_i^0 - \sum_{j=0}^{\infty}\big(\varphi(x_i, t_{j+1}) - \varphi(x_i, t_j)\big)u_i^{j+1}$$

und

$$\sum_{i\in\mathbb{Z}} \underbrace{\varphi(x_i, t_j)}_{=:a_i}\big(\underbrace{\hat{F}(u_i^j, u_{i+1}^j)}_{=:b_i} - \underbrace{\hat{F}(u_{i-1}^j, u_i^j)}_{=:b_{i-1}}\big) = -\sum_{i\in\mathbb{Z}}\big(\varphi(x_{i+1}, t_j) - \varphi(x_i, t_j)\big)\hat{F}(u_i^j, u_{i+1}^j).$$

Somit folgt

$$\begin{aligned}
&\sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}}\varphi(x_i, t_j))(u_i^{j+1} - u_i^j) + \frac{\Delta t}{\Delta x}\sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}}\varphi(x_i, t_j)\big[\hat{F}(u_i^j, u_{i+1}^j) - \hat{F}(u_{i-1}^j, u_i^j)\big] \\
&\quad = -\sum_{i\in\mathbb{Z}}\varphi(x_i, t_j)u_i^0 - \sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}}\big(\varphi(x_i, t_{j+1}) - \varphi(x_i, t_j)\big)u_i^{j+1} \\
&\qquad - \frac{\Delta t}{\Delta x}\sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}}(\varphi(x_{i+1}, t_j) - \varphi(x_i, t_j))\hat{F}(u_i^j, u_{i+1}^j) = 0.
\end{aligned}$$

Multipliziere dies mit Δx und finde

$$\begin{aligned}
\Delta x\,\Delta t\Bigg[&\sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}}\left(\frac{\varphi(x_i, t_{j+1}) - \varphi(x_i, t_j)}{\Delta t}\right)u_i^j \\
&+\sum_{j=0}^{\infty}\sum_{i\in\mathbb{Z}}\left(\frac{\varphi(x_{i+1}, t_j) - \varphi(x_i, t_j)}{\Delta x}\right)\hat{F}(u_i^j, u_{i+1}^j)\Bigg] = -\Delta x\sum_{i\in\mathbb{Z}}\varphi(x_i, t_0)u_i^0.
\end{aligned} \tag{3-19}$$

Wir führen nun den Grenzübergang $\Delta t_n \to 0$, $\lambda = \frac{\Delta t_n}{\Delta x_n}$ konstant, für jeden Term in (3-19) separat durch. Dafür schreiben wir präziser $u^c_{\Delta t_n}(x_i, t_j)$ statt u^j_i, um die Abhängigkeit von n zu betonen. Unter Beachtung von (3-18) und der Glattheit von φ lässt sich zeigen:

$$\Delta x_n \, \Delta t_n \sum_{j=0}^{\infty} \sum_{i \in \mathbb{Z}} \left(\frac{\varphi(x_i, t_{j+1}) - \varphi(x_i, t_j)}{\Delta t_n} \right) u^c_{\Delta t_n}(x_i, t_j) \to \int_0^\infty \int_{\mathbb{R}} \frac{\partial \varphi}{\partial t} \overline{u}(x,t) \, \mathrm{d}x \, \mathrm{d}t$$

$$\Delta x_n \sum_{i \in \mathbb{Z}} \varphi(x_i, t_0) u^c_{\Delta t_n}(x_i, t_0) \to \int_{\mathbb{R}} \varphi(x, 0) \overline{u}(x, 0) \, \mathrm{d}x$$

für $n \to \infty$. Den dritten Term in (3-19) müssen wir etwas genauer behandeln. Aufgrund der Glattheit von $\hat{F}$ zeigt man, dass es zu $u \in \mathbb{R}$ ein $L = L(u) > 0$ mit

$$|\hat{F}(v, w) - \hat{F}(u, u)| \leq L \cdot \max\{|v - u|, |w - u|\} \text{ für } v, w \text{ mit } |v - u|, |w - u| < \varepsilon$$

gibt. Somit folgt

$$\begin{aligned} |\hat{F}(u^c_{\Delta t_n}(x,t), & u^c_{\Delta t_n}(x + \Delta x_n, t)) - f(u^c_{\Delta t_n}(x,t))| \\ &= |\hat{F}(u^c_{\Delta t_n}(x,t), u^c_{\Delta t_n}(x + \Delta x_n, t)) - \hat{F}(u^c_{\Delta t_n}(x,t), u^c_{\Delta t_n}(x,t))| \\ &\leq L \, |u^c_{\Delta t_n}(x + \Delta x_n, t) - u^c_{\Delta t_n}(x,t)|. \end{aligned}$$

Da nun $u^c_{\Delta t_n}$ eine Funktion mit beschränkter totaler Variation ist, kann man die Gültigkeit von

$$|u^c_{\Delta t_n} - u^c(x,t)| \to 0$$

für $n \to \infty$, $t \geq 0$ und fast alle $x \in \mathbb{R}$ zeigen. Also lässt sich die numerische Flussfunktion $\hat{F}(u^c_{\Delta t_n}(x,t), u^c_{\Delta t_n}(x + \Delta x, t))$ durch $f(u^c_{\Delta t_n}(x,t))$ approximieren mit Fehlern, welche fast überall verschwinden. Insgesamt erhält man dann

$$\begin{aligned} \Delta x_n \, \Delta t_n \sum_{j=0}^{\infty} \sum_{i \in \mathbb{Z}} & \left(\frac{\varphi(x_{i+1}, t_j) - \varphi(x_i, t_j)}{\Delta x_n} \right) \hat{F}(u^c_{\Delta t_n}(x_i, t_j), u^c_{\Delta t_n}(x_{i+1}, t_j)) \\ & \to \int_0^\infty \int_{\mathbb{R}} \frac{\partial \varphi}{\partial x}(x,t) f(\overline{u}(t,x)) \, \mathrm{d}x \, \mathrm{d}t \text{ für } n \to \infty. \end{aligned}$$

Da dies für alle Testfunktionen $\varphi \in C^1_0(\mathbb{R} \times \mathbb{R})$ gilt, ist $\overline{u}$ also eine schwache Lösung unserer Erhaltungsgleichung. $\square$