Kapitel 12

lineare hyperbolische Gleichungen

12.0 hyperbolische Gleichungen

Die allgemeine Form von (Systemen von) *hyperbolischen Erhaltungsgleichungen* in "Erhaltungsform" ist

$$\partial_t \mathbf{u} + \sum_{j=1}^{d} \partial_{x_j}(\mathbf{f}^j(\mathbf{u})) = 0 \qquad (x,t) \in \Omega \times (0,\infty). \tag{12.1}$$

- Die Komponenten $\mathbf{u}_j : \Omega \to \mathbb{R}$, $j = 1, \ldots, s$ der Funktion $\mathbf{u} : \Omega \to \mathbb{R}^s$ heißen *Zustandgrößen*, der Vektor $\mathbf{u}$ Zustandsvektor ("state vector").
- Die d Funktionen $\mathbf{f}^j : \mathbb{R}^s \to \mathbb{R}^s$, $j = 1, \ldots, d$, heißen *Flußfunktionen*. Allgemein heißt der Definitionsbereich (der Wertebereich des Zustandsvektors) der Flußfunktionen die "Zustandsmenge" (set of states).

Hyperbolische Erhaltungsgleichungen von der Form (12.1) beschreiben zahlreiche Erhaltungsgleichung. Die (auch historisch) bedeutendsten ergeben sich aus der Stömungs- und Gasdynamik, z.B. den *Eulergleichungen*. Die Zustandsgrößen sind dann z.B. Masse, Impuls, Energie, und die Gleichungen beschreiben die Erhaltung dieser Größen.

Beispiel 12.1 (Eulergleichungen) Die *Eulergleichungen* der Gasdynamik beschreiben die Strömung eines Gases ("Fluids"). Dabei gelten Masseerhaltung, Energieerhaltung und Impulserhaltung. Wenn man mit $\mathbf{v}(x,t) \in \mathbb{R}^3$ die Geschwindigkeit von Partikeln am Ort x zum Zeitpunkt t bezeichnet, mit $\rho(x,t)$ die Dichte, mit $p = p(x,t)$ den Druck und mit e die (spezifische) innere Energie ("Temperatur") so ergibt sich mit der Gesamtenergie $E = \rho e + \frac{1}{2}|\mathbf{v}|^2$ die folgenden Erhaltungsgleichungen:

$$\begin{aligned}
\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) &= 0 && \text{Massererhaltung} \\
\partial_t(\rho \mathbf{v}_i) + \nabla \cdot (\rho \mathbf{v}\mathbf{v}_i) + \partial_{x_i} p &= 0, \quad i = 1,2,3, && \text{Impulserhaltung} \\
\partial_t E + \nabla \cdot (\mathbf{v}(E+p)) &= 0 && \text{Energieerhaltung.}
\end{aligned}$$

Tatsächlich stellt dies (im $\mathbb{R}^3$) 5 Gleichungen für 6 Unbekannte Funktionen dar. Die fehlende Gleichung kann z.B. durch ein "konstitutives Gesetz" erzeugt werden. Bei "idealen Gasen" z.B. ist der Druck p eine Funktion der inneren Energie: $p = \rho(\gamma - 1)e$, wobei γ eine Konstante ist[1]. Die Eulergleichungen können tatsächlich auf die Form (12.1) gebracht werden:

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho \mathbf{v}_1 \\ \rho \mathbf{v}_2 \\ \rho \mathbf{v}_3 \\ E \end{pmatrix}, \qquad \mathbf{f}^1 = \begin{pmatrix} \rho \mathbf{v}_1 \\ p + \rho \mathbf{v}_1^2 \\ \rho \mathbf{v}_1 \mathbf{v}_2 \\ \rho \mathbf{v}_1 \mathbf{v}_3 \\ \mathbf{v}_1(E+p) \end{pmatrix}, \qquad \mathbf{f}^2 = \begin{pmatrix} \rho \mathbf{v}_2 \\ \rho \mathbf{v}_1 \mathbf{v}_2 \\ p + \rho \mathbf{v}_2^2 \\ \rho \mathbf{v}_2 \mathbf{v}_3 \\ \mathbf{v}_2(E+p) \end{pmatrix}, \qquad \mathbf{f}^3 = \begin{pmatrix} \rho \mathbf{v}_3 \\ \rho \mathbf{v}_1 \mathbf{v}_3 \\ \rho \mathbf{v}_2 \mathbf{v}_3 \\ p + \rho \mathbf{v}_3^2 \\ \mathbf{v}_3(E+p) \end{pmatrix},$$

■

[1] Aus der Schule kennt man diese Beziehung in der Form $pV = nRT$, wenn man zusätzlich die innere Energie e von der Form $e = cT$ annimmt

Bemerkung 12.2 Die Gleichung (12.1) drückt eine Erhaltungsgleichung aus: Für ein beliebiges "Kontrollvolumen" $D \subset \mathbb{R}^d$ ergibt sich durch Integrieren über D und vertauschen von $\int_D$ mit $\frac{d}{dt}$

$$\frac{d}{dt}\int_D \mathbf{u}\,dx + \int_{\partial D} \mathbf{F}(\mathbf{u}, n)\,ds = 0,$$

wobei n die äußere Normale an D ist und $\mathbf{F}(\mathbf{u}, \underline{\omega}) := \sum_{j=1}^d \omega_j \mathbf{f}^j(\mathbf{u})$ der Fluß in Richtung $\underline{\omega} = (\omega_1, \ldots, \omega_d)$ ist. ∎

Ein wichtiger Spezialfall ist der Fall einer skalaren Gleichung (also nur eine einzige Erhaltungsgröße):

Beispiel 12.3 Im Fall $s = 1$ sind die Funktionen $f^1, \ldots, f^d$ reellwertig. Schreibt man $\mathbf{F}(u) := (f^1, \ldots, f^d)^\top$, so ergibt sich die Erhaltungsgleichung in der Form

$$\partial_t u + \nabla \cdot (\mathbf{F}(u)) = 0. \tag{12.2}$$

Die Erhaltungsform nimmt die etwas vertrautere Form

$$\frac{d}{dt}\int_D u\,dx + \int_{\partial D} \mathbf{F}(u) \cdot n\,ds = 0$$

an. ∎

Beispiel 12.4 Betrachtet man nur eine skalare Gleichung und nimmt an, daß $\mathbf{F}(u)$ von der Form $\mathbf{b}u$ ist, so ergibt sich die *Advektionsgleichung*

$$u_t + \mathbf{b} \cdot \nabla u = 0. \tag{12.3}$$

∎

Bemerkung 12.5 (lineare hyperbolische Systeme) Fallls die Funktionen $\mathbf{f}^j(\mathbf{u})$ die Form $\mathbf{A}_j\mathbf{u}$ haben für konstante Matrizen $\mathbf{A}_j \in \mathbb{R}^{s\times s}$, so spricht man von einem linearen System. Es hat die Form

$$\partial_t \mathbf{u} + \sum_{j=1}^d \mathbf{A}_j \partial_{x_j} \mathbf{u} = 0$$

Sind die Matrizen $\mathbf{A}_j$ alle symmetrisch, so spricht man von einem symmetrischen System ("Friedrichs system"). ∎

Strikt genommen gehört zur Hyperbolizität des Systems (12.1) noch eine Bedingung der reellen Diagonalisierbarkeit der Linearisierung:

Definition 12.6 (Hyperbolizität einer Erhaltungsgleichung) *(12.1) heißt hyperbolisch, falls für jeden Zustandsvektor* $\mathbf{u}$ *und jede Richtung* $\underline{\omega} \in \mathbb{R}^d \setminus \{0\}$ *die Matrix*

$$\mathbf{A}(\mathbf{u}, \underline{\omega}) := \sum_{j=1}^d \omega_j D_{\mathbf{u}} \mathbf{f}^j(\mathbf{u})$$

reell diagonalisierbar ist. Falls die Eigenwerte (für jeden Zustand $\mathbf{u}$ *und jede Richtung* $\underline{\omega}$*) sogar paarweise verschieden sind, dann heißt das System* strikt hyperbolisch.

Historisch wichtig ist der Fall $d = 1$ von Systemen:

Übung 12.7 Sei $d = 1$. Das System hat die Form

$$\partial_t \mathbf{u} + \partial_x(\mathbf{F}(\mathbf{u})) = 0 \tag{12.4}$$

Zeigen Sie: es ist hyperbolisch, falls $D\mathbf{F}(\mathbf{u})$ reell diagonalisierbar ist für alle Zustandsvektoren $\mathbf{u}$. ∎

Bemerkung 12.8 Das Problem (12.1) muß noch mit Randbedingungen (und Anfangsbedingungen) vervollständigt werden. ∎

AUSARBEITEN: FD fuer glatte Lsg wie WEllengleichung, KdV,...—FVM fuer unstetige Lsgnen, SChocks

12.1 klassische Differenzenverfahren am Beispiel der Advektionsgleichung

12.1.1 Vorbemerkungen zu finiten Differenzenverfahren

Vorbemerkung: Differenzenverfahren erzeugen numerische Verfahren, indem Ableitungen durch Differenzenquotienten ersetzt werden.

Vorteile:

- einfache, schnelle Erzeugung des Verfahrens
- einfacher Umgang mit *nichtlinearen* Gleichungen

Nachteile: Handhabung komplexer Geometrien, Randbedingungen

Fokus des vorliegenden Abschnittes: Stabilität des Verfahrens in der Zeit ($\rightarrow$ Randbedingungen werden ausgeblendet durch Betrachten eines Vollraumproblems)

Eine simultane Diskretisierung in Ort und Zeit wie wir es in Abschnitt 12.5 vorstellen werden erfolgt in der Praxis selten. Fast ausschließlich werden bei (zeitabhängigen) hyperbolischen Problemen Zeitschrittverfahren eingesetzt—meist sogar explizit in der Zeit. Eine der zentralen Fragen bei solchen Verfahren ist die der Stabilität, und der vorliegende Abschnitt ist primär dieser Frage gewidmet. Eine zweite Frage ist, insbesondere bei Differenzenverfahren, die Realisierung von Randbedingungen. Wir werden diese Frage allenfalls kursorisch behandeln. Um die Frage nach Randbedingungen zu umgehen, betrachten wir ein reines Cauchyproblem (d.h. $\Omega = \mathbb{R}^d$– die klassische Alternative ist die Untersuchung von periodischen Randbedingungen). Um die Situation noch einfacher zu gestalten, betrachten wir den räumlichen 1D-Fall.

12.1.2 FD für die Advektionsgleichung

Der einfachste Fall einer linearen hyperbolischen Gleichung ist damit die Advektionsgleichung:

$$u_t + au_x = g \qquad \text{auf } \mathbb{R} \times (0, \infty), \qquad u(x, 0) = u_0(x), \tag{12.5}$$

wobei g und der Startwert u_0 kompakten Träger haben mögen. Für $g \equiv 0$ kann die Lösung explizit angegeben werden:

$$u(x, t) = u_0(x - at). \tag{12.6}$$

Bei *Differenzenverfahren* werden Ableitungen durch Differenzenquotienten approximiert. Wir betrachten ein regelmäßiges Gitter $x_i = ih$, $i \in \mathbb{Z}$ im Ort und ein regelmäßiges Gitter in der Zeit $t_n = nk$, $n = 0$, $1, \ldots,$. Die (zu berechnenden) Werte u_i^n sollen Approximationen an die (unbekannten) Funktionswerte $u(x_i, t_n)$ sein.

Wir führen den Begriff der "Gitterfunktion" ein: eine Folge $(U_i)_{i\in\mathbb{Z}}$ heißt Gitterfunktion–man kann sich dies als die Werte in den Knoten $x_i = ih$ vorstellen. Um den Zeitschritt n zu markieren, verwenden wir einen Superskript. Z.B. kann man sich bei der Gitterfunktion $U^n = (U_i^n)_{i\in\mathbb{Z}}$ Werte in den Punkten (x_i, t_n) vorstellen. Wird sie mit einem Superskript versehen, so beschreibt dieser den Zeitpunkt.

Wir betrachten folgende *explizite* Verfahren:

1. "forward time/backward space":

$$\frac{u_i^{n+1} - u_i^n}{k} + a\frac{u_i^n - u_{i-1}^n}{h} = g(x_i, t_n). \tag{12.7}$$

2. "forward time/forward space":

$$\frac{u_i^{n+1} - u_i^n}{k} + a\frac{u_{i+1}^n - u_i^n}{h} = g(x_i, t_n). \tag{12.8}$$

3. "Lax-Friedrichs":

$$\frac{1}{k}\left(u_i^{n+1} - \frac{1}{2}\left(u_{i+1}^n + u_{i-1}^n\right)\right) + \frac{a}{2h}\left(u_{i+1}^n - u_{i-1}^n\right) = g(x_i, t_n). \tag{12.9}$$

Faßt man $U^n := (u_i^n)_{i\in\mathbb{Z}}$ als Gitterfunktion auf, so haben die Schemata die Form

$$U^{n+1} = EU^n + kG^n, \tag{12.10}$$

wobei der *Propagationsoperator* E ein *linearer* Operator auf dem Raum der *Gitterfunktionen* ist; wir schreiben weiter G^n für die Gitterfunktion $(g(x_i, t_n))_{i\in\mathbb{Z}}$.

Wie bei linearen Problemen üblich sind die zentralen Begriffe "Konsistenz" und "Stabilität"[2]. Konsistenz ist der Fehler, den das numerische Verfahren in einem (Zeit-)Schritt macht, d.h. man vergleicht die exakte Lösung nach einem Zeitschritt mit dem Ergebnis, das man aus dem Verfahren erhält, wenn man die exakte Lösung als Anfangswert wählt. Also: Sei u eine exakte Lösung und die Gitterfunktion U_{kh}^n gegeben durch

$$U_{kh,i}^n := u(x_i, t_n)$$

dann ist der Konsistenzfehler

$$\tau_i^{n+1} = \frac{1}{k}\left[U_{kh,i}^{n+1} - ((EU_{kh}^n)_i + kG_i^n)\right].$$

Beim Berechnen des Konsistenzfehlers (in Abhängigkeit von k und h) muß man die Eigenschaft nutzen, daß u die Differentialgleichung löst—das ist für die Bestimmung der Konsistenzordnung unpraktisch. Für die vorliegenden Diskretisierungen nutzt man einfach direkt die Gleichung $u_t + au_x = g$ aus, so daß man den Konsistenzfehler τ definiert als

$$\tau_i^{n+1} = \frac{1}{k}\left(U_{kh,i}^{n+1} - (EU_{kh}^n)_i\right) - (u_t(x_i, t_n) + au_x(x_i, t_n)) \tag{12.11}$$

für jede hinreichend glatte Funktion u.

Übung 12.9 Zeigen Sie mittels Taylorentwicklung, daß für die obigen Verfahren gilt:

$$|\tau_i^n| \le C\,[k + h]\,,$$

wobei die Konstante C von u abhängt. M.a.W.: die Verfahren sind von der Ordnung $(1, 1)$.

Während der Begriff der Konsistenz den Fehler faßt, der in *einem* Zeitschritt gemacht wird, faßt der Begriff der *Stabilität* den Einfluß von Fehlern (z.B. Konsistenzfehlern) in vergangenen Zeitschritten auf den Fehler. M.a.W.: Stabilität mißt die Verstärkung von Fehlern durch das Verfahren. Definiere den *Fehler*

$$\varepsilon_i^n := u(x_i, t_n) - u_i^n = U_{kh,i}^n - u_i^n.$$

Für das Verfahren und den Konsistenzfehler gilt:

$$\begin{aligned} U^{n+1} &= EU^n + kG^n \\ U_{kh}^{n+1} &= EU_{kh}^n + kG^n + k\tau^{n+1} \end{aligned}$$

Wegen der Linearität von E ergibt sich die Rekursion

$$\varepsilon^{n+1} = E\varepsilon^n + k\tau^{n+1}.$$

Wir fixieren nun eine Norm auf dem Raum der Gitterfunktionen:

$$\|(V_i)_{i\in\mathbb{Z}}\|_{\ell_h^1} := \sum_{i\in\mathbb{Z}} h|V_i|.$$

Es ergibt sich

$$\|\varepsilon^n\|_{\ell_h^1} \le \|E^n\|_{\ell_h^1}\|\varepsilon^0\|_{\ell_h^1} + k\sum_{\ell=1}^{n}\|E^{n-\ell}\|_{\ell_h^1}\|\tau^\ell\|_{\ell_h^1}$$

[2] Für lineare, wohlgestellte Probleme besagt das Lax'sche Äquivalenzprinzip tatsächlich: Konvergenz $\Longleftrightarrow$ Konsistenz + Stabilität

Wir erkennen, daß wir für festes T und $0 \le nk \le T$ fordern sollten

$$\sup_{n:0\le nk\le T} \|E^n\|_{\ell^1_h} \le C_T, \tag{12.12}$$

für ein $C_T > 0$ unabhängig von k und h, denn dann ergibt sich:

$$\|\varepsilon^n\|_{\ell^1} \le C_T \left[\|\varepsilon^0\|_{\ell^1_h} + \sup_{\ell\le n} \|\tau^\ell\|_{\ell^1_h} \underbrace{\sum_{\ell=1}^{n} k}_{\le T} \right]$$

Für die betrachteten Verfahren ergibt sich somit

$$\sup_{n:0\le nk\le T} \|\varepsilon^n\|_{\ell^1_h} \le C_T \left[\|\varepsilon^0\|_{\ell^1_h} + C(k+h) \right].$$

Man kann sich als Startfehler $\|\varepsilon^0\|_{\ell^1_h}$ also $O(h)$ erlauben, was mit Knotenauswertung oder sogar bei Wahl als Mittelwerte über Elemente der Fall ist. Entscheidend ist die Stabilitätsbedingung (12.12). Praktisch ist sie erfüllt, falls

- $\|E\|_{\ell^1_h} \le 1$
- oder doch wenigstens $\|E\|_{\ell^1_h} \le 1 + Ck$ für eine Konstante $C > 0$, die nicht von k und h ist.

Typisch für explizite Verfahren ist, daß dieser Bedingungen nicht für alle Kombinationen von k und h erfüllt sind sondern unter der Bedingung, daß k/h hinreichend klein ist, die sog. "CFL"-Bedingung[3]

$$\frac{|ak|}{h} \le c, \tag{12.13}$$

wobei $c > 0$ eine Konstante ist, die nicht von k und h abhängt (wohl aber vom Problem und dem Verfahren).

Beispiel 12.10 (Advektionsgleichung auf Torus)

$$\begin{aligned} &u_t + u_x = 0, \qquad x \in (0,1), \quad t > 0, \qquad u(0,t) = u(1,t) \quad \forall t > 0 \\ &u(x,0) = \sin(\pi x) \end{aligned}$$

Figuren 12.1–12.4 zeigen das Verhalten der unterschiedlichen Verfahren. Insbesondere sieht man, daß die CFL-Bedingung $\lambda = \frac{k|a|}{h} \le 1$ wesentlich ist bei diesen expliziten Verfahren. ∎

Übung 12.11 Zeigen Sie, daß für das Lax-Friedrichs-Schema unter der Voraussetzung der CFL-Bedingung $|ak/h| \le 1$ die folgende Abschätzung gilt:

$$\|E\|_{\ell^1_h} \le 1$$

12.1.3 Upwinding

Wir untersuchen nun "forward time/forward space" und "forward time/backward space". Als entscheidend für die Stabilität wird sich das Vorzeichen von a herausstellen:

$$\begin{cases} \text{verwende ft/bs (12.7)} & \text{falls } a > 0 \\ \text{verwende ft/fs (12.8)} & \text{falls } a < 0\ . \end{cases} \tag{12.14}$$

Satz 12.12 (Stabilität des Upwind-Verfahrens) *Unter der CFL-Bedingung $|ak/h| \le 1$ erfüllt der Propagationsoperator E des Upwind-Verfahrens (12.14) die Stabilitätsbedingung $\|E\|_{\ell^1_h} \le 1$. Für glatte Lösungen ist damit der Fehler $O(k+h)$.*

[3] Courant-Friedrichs-Lewy-Bedingung, benannt nach Richard Courant, Kurt Friedrichs und Hans Lewy

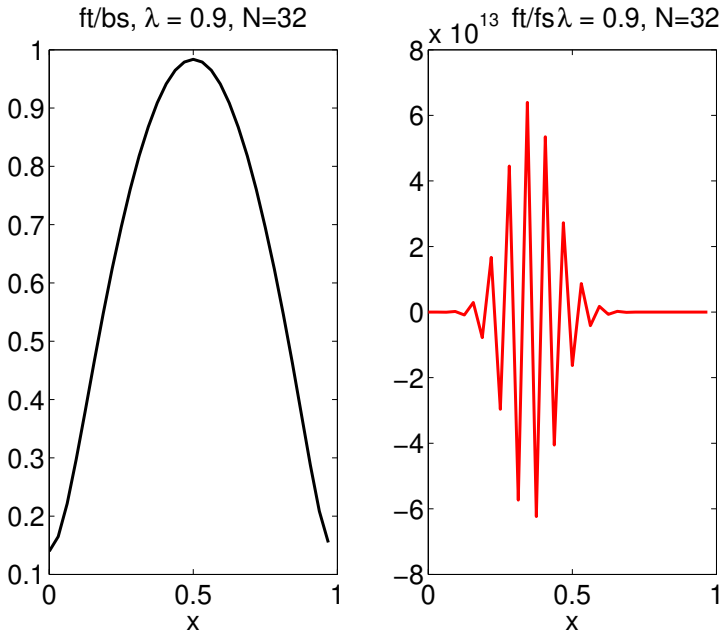


Abbildung 12.1: exakte Lösung; Vergleich ft/bs mit ft/fs

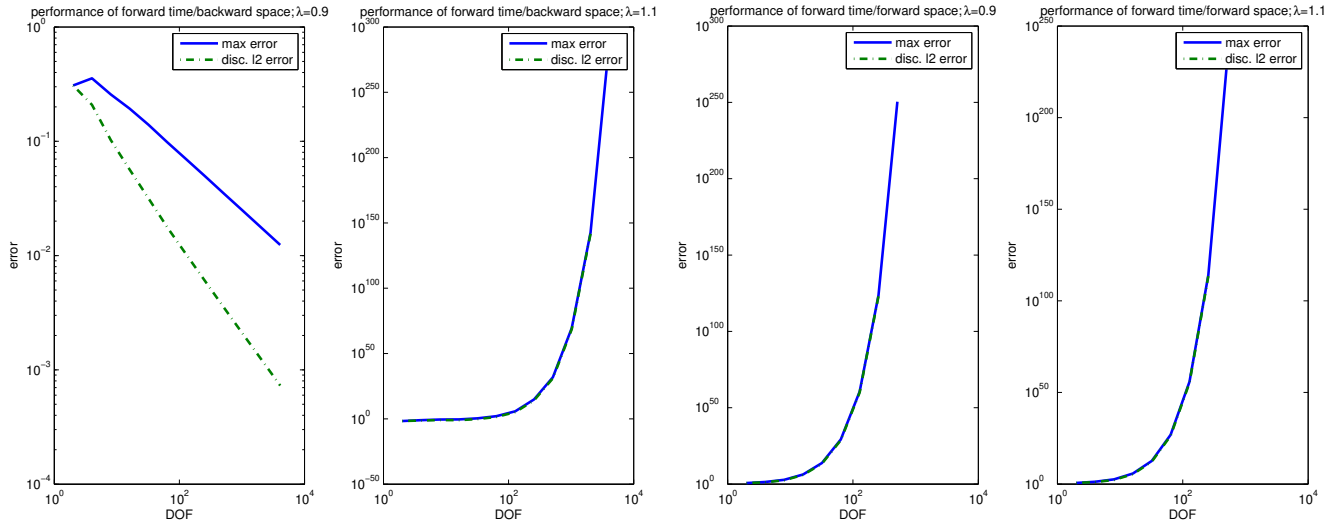


Abbildung 12.2: Einfluß der CFL-Bedingung auf ft/bs und ft/fs

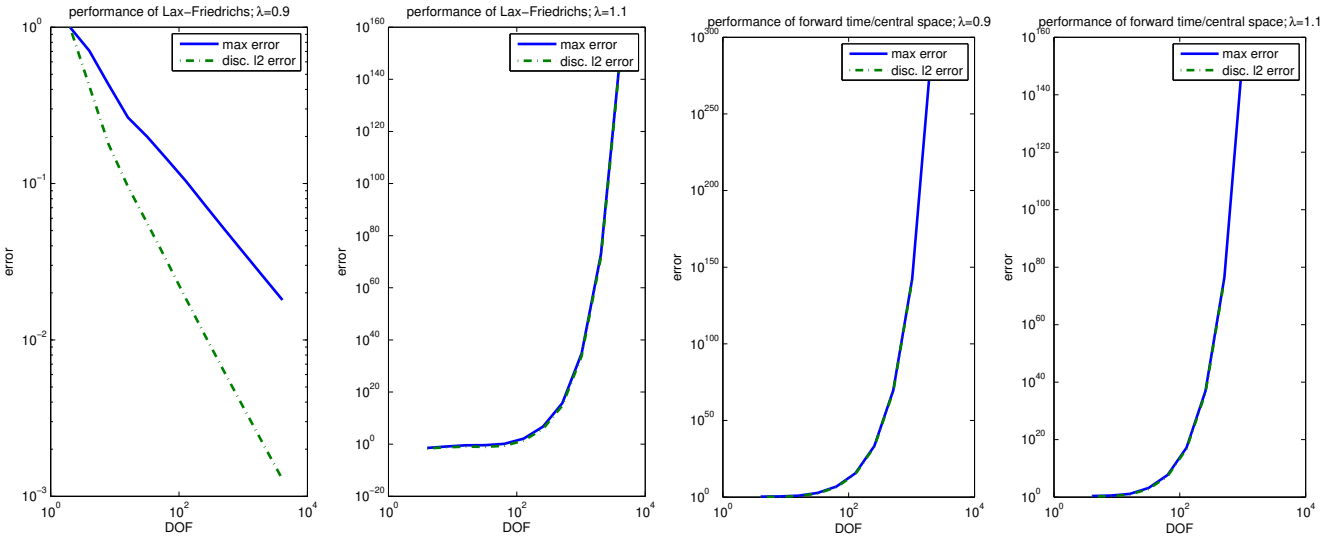


Abbildung 12.3: Einfluß der CFL-Bedingung auf Lax-Friedrichs und ft/central

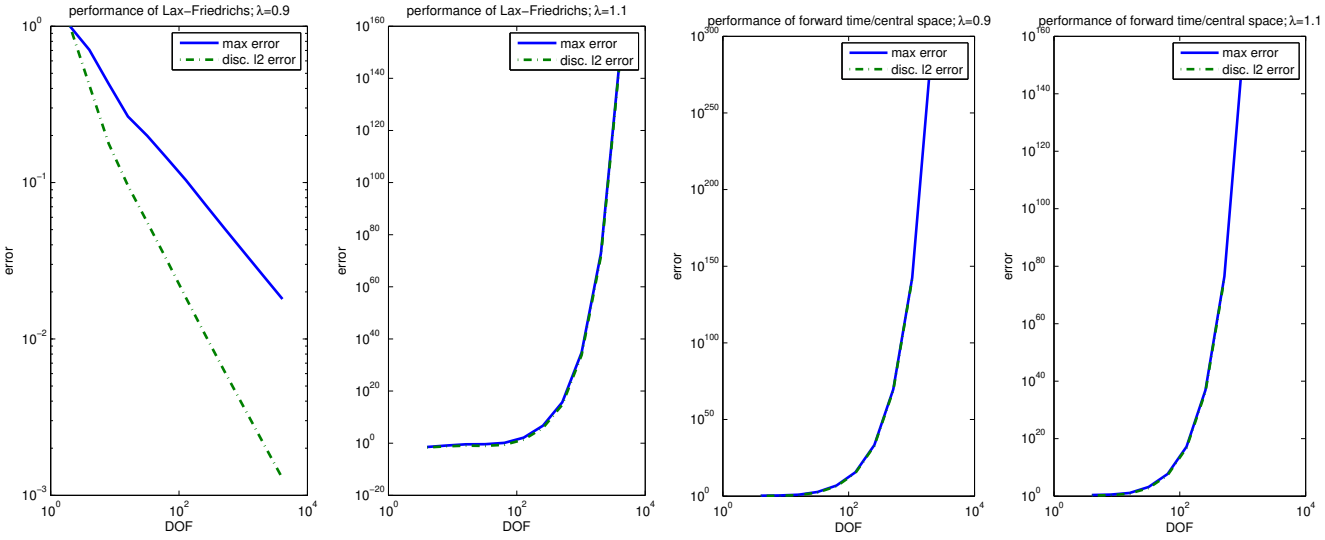


Abbildung 12.4: Einfluß der CFL-Bedingung auf das Leap Frog Verfahren

Beweis: Betrachte den Fall $a > 0$. Für eine Gitterfunktion $(V_i)_{i\in\mathbb{Z}}$ gilt damit

$$(EV)_i = V_i - \frac{ak}{h}\,(V_i - V_{i-1})$$

und damit

$$\begin{aligned}\|EV\|_{\ell^1_h} &\leq \sum_i h\left[\left|1-\frac{ka}{h}\right||V_i| + |V_{i-1}|\left|\frac{ka}{h}\right|\right] \\ &\overset{(12.13)}{=} \left(1-\frac{ak}{h}\right)\sum_i h|V_i| + \frac{ak}{h}\sum_i h|V_{i-1}| = \|V\|_{\ell^1_h}.\end{aligned}$$

□

Übung 12.13 Zeigen Sie im Setting von Satz 12.12: falls die CFL-Bedingung $|a|k/h \leq 1$ gilt, dann ist $\|E\|_{\ell^\infty \leftarrow \ell^\infty} \leq 1$, d.h. das Verfahren ist in ℓ^∞ stabil. ■

Bemerkung 12.14 vergleiche auch: den Begriff der Monotonie des Verfahrens — siehe das Buch von Asher

Bemerkung 12.15 (physikalische Plausibilität der CFL-Bedingung und des Upwinding) Die Lösungsformel (12.6) zeigt, daß die exakte Lösung bei (x_i, T) durch $u_0(x_i - aT)$ gegeben ist. Damit ist eine notwendige Bedingung an das numerische Verfahren, daß zum Zeitpunkt $t_n = T$ die Startwerte in der Nähe von $x_i - aT$ "gesampled" werden. Betrachtet man das ft/bs-Verfahren, so "sieht" der Punkt (x_i, t_1) die Werte bei (x_i, t_0) und (x_{i-1}, t_1), der Punkt (x_i, t_2) entsprechend die Werte bei (x_{i-2}, t_0), (x_{i-1}, t_0), (x_i, t_0) etc. Bei $t_n = T$ damit die Startwerte bei (x_{i-j}, t_0), $j = 0, \ldots, n$. Falls also das Intervall $[x_{i-n}, x_i]$ *nicht* den Punkt $x_i - aT$ enthält, dann kann das Verfahren nicht funktionieren.

- Im Fall $a < 0$ kann das Verfahren also *nicht* funktionieren.
- Im Fall $a > 0$ muß zusätzlich die Bedingung $nh \geq aT$, d.h. $nh \geq akn$ gelten, was genau die CFL-Bedingung $ka/h \leq 1$ ausdrückt. ■

Bemerkung 12.16 Upwinding kann auch für Systeme formuliert werden. Sei $\mathbf{A}$ eine konstante Matrix und betrachte

$$\mathbf{U}_t + \mathbf{A}\mathbf{U}_x = 0$$

Kann man $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ diagonalisieren, dann würde man in den neuen Variablen $\widetilde{\mathbf{U}} = \mathbf{V}^{-1}\mathbf{U}$ ft/bs für die Komponenten mit $\mathbf{D}_{ii} > 0$ machen und ft/fs für die Komponenten mit $\mathbf{D}_{ii} < 0$. Schreibe

$$\mathbf{D} = \mathbf{D}^+ + \mathbf{D}^-,$$

wobei $\mathbf{D}^+$ die positiven und $\mathbf{D}^-$ die negativen Diagonalelemente von $\mathbf{D}$ hat (formal: $\mathbf{D}^+_{ii} = \max\{\mathbf{D}_{ii}, 0\}$, $\mathbf{D}^-_{ii} = \min\{\mathbf{D}_{ii}, 0\}$). Das Upwind-Verfahren ist dann

$$\widetilde{\mathbf{U}}^{n+1}_j = \widetilde{\mathbf{U}}^n_j - \frac{k}{h}\mathbf{D}^+(\widetilde{\mathbf{U}}^n_j - \widetilde{\mathbf{U}}^n_{j-1}) - \frac{k}{h}\mathbf{D}^-(\widetilde{\mathbf{U}}^n_{j+1} - \widetilde{\mathbf{U}}^n_j);$$

Rücktransformation zur $\mathbf{U}$-Variablen führt mit

$$\mathbf{A}^+ = \mathbf{V}\mathbf{D}^+\mathbf{V}^{-1}, \qquad \mathbf{A}^- = \mathbf{V}\mathbf{D}^-\mathbf{V}^{-1}$$

auf

$$\mathbf{U}^{n+1}_j = \mathbf{U}^n_j - \frac{k}{h}\mathbf{A}^+(\mathbf{U}^n_j - \mathbf{U}^n_{j-1}) - \frac{k}{h}\mathbf{A}^-(\mathbf{U}^n_{j+1} - \mathbf{U}^n_j);$$

12.2 von Neumann-Analyse

Die klassische Stabilitätsanalyse wird nicht in $\|\cdot\|_{\ell_h^1}$ durchgeführt sondern in der Norm

$$\|(V_i)\|_{\ell_h^2} := \left(\sum_i h|V_i|^2\right)^{1/2}.$$

Entsprechend bezeichnen wir mit ℓ_h^2 den Raum der Folgen mit endlicher Norm.

Der Grund ist in erster Linie ein technischer: für lineare Differentialgleichungen mit konstanten Koeffizienten auf regelmäßigen Gittern liefert die Fourieranalyse ein sehr bequemes Werkzeug. Man definiert:

- Die "Fouriertransformation" einer Folge $(v_i)_{i\in\mathbb{Z}}$:

$$\widehat{v}(\xi) := (\mathcal{F}_h(v))(\xi) := h\sum_j e^{-\mathbf{i}\xi x_j} v_j, \qquad \xi \in [-\pi/h, \pi/h]$$

- die L_h^2-norm:

$$\|\widehat{v}\|_{L_h^2}^2 := \int_{-\pi/h}^{\pi/h} |\widehat{v}(\xi)|^2\, d\xi$$

- Die Faltung zweier Folgen $u = (u_i)_i$, $v = (v_i)_i$

$$(u * v)_i := h\sum_j u_{i-j} v_j = h\sum_j u_j v_{i-j}$$

Es gilt:

Satz 12.17 *(i) (Parseval) $\mathcal{F}_h$ ist ein Isomorphismus $\ell_h^2 \to L_h^2$:*

$$\sqrt{2\pi}\|(v_i)_{i\in\mathbb{Z}}\|_{\ell_h^2} = \|\widehat{v}\|_{L_h^2}, \qquad \widehat{v} = \mathcal{F}_h((v_i)_{i\in\mathbb{Z}}).$$

Die Inverse ist gegeben durch

$$v_j = (\mathcal{F}_h^{-1}(\widehat{v}))_j = \frac{1}{2\pi}\int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x_j}\widehat{v}(\xi)\, d\xi$$

(ii) Für $(u_i)_{i\in\mathbb{Z}} \in \ell_h^2$ und $(v_i)_{i\in\mathbb{Z}} \in \ell_h^1$ ist $u * v \in \ell_h^2$ und

$$\widehat{(u * v)}(\xi) = \widehat{u}(\xi)\widehat{v}(\xi)$$

(iii) (Translation) Für $j_0 \in \mathbb{Z}$ ist $\mathcal{F}_h((v_{j+j_0})_{j\in\mathbb{Z}})(\xi) = e^{\mathbf{i}\xi x_{j_0}}\widehat{v}(\xi)$

(iv) (Modulation) Für $\xi_0 \in \mathbb{R}$ ist $\mathcal{F}_h((e^{\mathbf{i}\xi_0 x_j} v_j)_{j\in\mathbb{Z}})(\xi) = \widehat{v}(\xi - \xi_0)$

(v) (Dilation) Für $m \in \mathbb{Z}\setminus\{0\}$ ist $\mathcal{F}_h((v_{mj})_{j\in\mathbb{Z}})(\xi) = \widehat{v}(\xi/m)/|m|$

(vi) (Konjugation) $\mathcal{F}_h((\overline{v_j})_{j\in\mathbb{Z}})(\xi) = \overline{\widehat{v}(-\xi)}$

Beweis: Übung. □

Betrachte ein Einschrittverfahren mit Propagationsoperator E der Form

$$(Ev)_i = \sum_{\ell=-r}^{s} \alpha_\ell v_{i+\ell}$$

für Koeffizienten α_ℓ. Der Operator E ist vom Faltungstyp:

$$(Ev) = a * v, \qquad a_\ell = \frac{1}{h}\alpha_{-\ell}.$$

Damit ist

$$\widehat{(Ev)}(\xi) = \widehat{a}(\xi)\widehat{v}(\xi),$$

wobei $\widehat{a}$ der *Verstärkungsfaktor* genannt wird. Es ist:

Operator	Symbol
Vorwärtsdifferenz D_+: $u_{i+1} - u_i$	$e^{\mathbf{i}\xi} - 1$
Rückwärtsdifferenz D_-: $u_i - u_{i-1}$	$1 - e^{-\mathbf{i}\xi}$
symmetrische Differenz D_0: $u_{i+1} - u_{i-1}$	$2\mathbf{i}\sin\xi$
δ: $u_{i+1/2} - u_{i-1/2}$	$2\mathbf{i}\sin(\xi/2)$
D_+D_-	$-4\sin^2(\xi/2)$

Abbildung 12.5: Symbole einiger Differenzenoperatoren

Übung 12.18

$$\|E\|_{\ell_h^2 \leftarrow \ell_h^2} = \max_{\xi \in [-\pi/h, \pi/h]} |\widehat{a}(\xi)|$$

Damit ist die Stabilitätsanalyse zurückgeführt auf die Berechnung von $\widehat{a}$. Man sagt, daß ein Verfahren die *von-Neumann-Stabilitätsbedingung* erfüllt, falls der zugehörige Verstärkungsfaktor $\widehat{a}$ die folgende Bedingung erfüllt:

$$|\widehat{a}(\xi)| \leq 1 + Ck \qquad \forall \xi \in [-\pi/h, \pi/h], \tag{12.15}$$

wobei $C > 0$ unabhängig von k (und natürlich ξ) ist.

Beispiel 12.19 (Upwind-Verfahren) Das Upwind-Verfahren für die Advektionsgleichung im Fall $a = -1$ (und $g \equiv 0$) ist

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \lambda(v_{j+1}^n - v_j^n), \qquad \lambda = \frac{k}{h}.$$

Dies hat die Form $Ev^n = a * v^n$ mit

$$a_j = \frac{1}{h}\lambda\delta_{-1,j} + \frac{1}{h}(1-\lambda)\delta_{j,0}, \qquad \delta_{n,m} = \text{Kronecker } \delta$$

Die Fouriertransformierte ist

$$\widehat{a}(\xi) = h(e^{-\mathbf{i}\xi x_{-1}}a_{-1} + e^{-\mathbf{i}\xi x_0}a_0) = \lambda e^{\mathbf{i}\xi h} + (1-\lambda)$$

Man sieht, daß im Fall $0 < \lambda \leq 1$ gilt:

$$|\widehat{a}(\xi)| \leq 1 \qquad \forall \xi \in [-\pi/h, \pi/h],$$

d.h. das Verfahren erfüllt die von-Neumann-Bedingung (12.15). ∎

In der Praxis kürzt man die Bestimmung von $g(\xi) := \widehat{a}(\xi)$ mit einer "Rechenregel" erheblich ab: man macht den Ansatz $v_j^n = g^n e^{\mathbf{i}\xi jh}$ und setzt in das Verfahren ein, um eine Formel für $g(\xi)$ zu erhalten.

Beispiel 12.20 (Lax-Wendroff) Für die Advektionsgleichung mit $a = -1$ ist das Verfahren gegeben durch

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \frac{1}{2}\lambda(v_{j+1}^n - v_{j-1}^n) + \frac{1}{2}\lambda^2(v_{j+1}^n - 2v_j^n + v_{j-1}^n), \qquad \lambda = k/h.$$

Einsetzen des Ansatzes $v_j^n = g^n e^{\mathbf{i}\xi jh}$ und Kürzen mit $g^n e^{\mathbf{i}\xi jh}$ liefert

$$g(\xi) = 1 + \frac{1}{2}\lambda(e^{\mathbf{i}\xi h} - e^{-\mathbf{i}\xi h}) + \frac{1}{2}\lambda^2(e^{\mathbf{i}\xi h} - 2 + e^{-\mathbf{i}\xi h}) = 1 + \mathbf{i}\lambda\sin\xi h - 2\lambda^2\sin^2\frac{\xi h}{2}.$$

Eine elementare Rechnung zeigt, daß auch hier für $\lambda \in (0,1]$ die Bedingung

$$\sup_{\xi \in [-\pi/h, \pi/h]} |g(\xi)|^2 = \sup_{\xi \in [-\pi/h, \pi/h]} \left(1 - 4\lambda^2(1-\lambda^2)\sin^4(\xi h/2)\right) \overset{\lambda \in (0,1]}{\leq} 1$$

erfüllt ist, d.h. die von-Neumann-Bedingung (12.15).

Die Form des Verstärkungsfaktors g zeigt auch, daß das Lax-Wendroff-Verfahren *dissipativ*[4] ist: Während $g(\xi) \approx 1$ für $\xi \approx 0$ ist für $|\xi| \approx \pi/h$ sogar $|g(\xi)| \approx 1 - 4\lambda^2(1-\lambda^2) < 1$, falls $\lambda < 1$. M.a.W.: Während die niederfrequenten Term (wegen Konsistenz!) weder verstärkt noch gedämpft werden, werden die hochfrequenten Lösungskomponenten (und damit auch die hochfrequenten Fehleranteile) gedämpft.
■

Die Bedeutung des Faktors Ck in der Bedingung (12.15) ergibt aus dem Wunsch, auch Gleichungen mit Termen niedrigerer Ordnung zu behandeln:

Übung 12.21 Betrachten Sie

$$u_t - u_x + c(x)u = 0$$

mit der Upwinddiskretisierung

$$v_j^{n+1} = (Ev^n)_j = v_j^n + \lambda(v_{j+1}^n - v_j^n) + kc(x_j)v_j, \qquad \lambda = \frac{k}{h}. \tag{12.16}$$

Sei E_0 die Evolution, die zu $c \equiv 0$ gehört. Aus Beispiel 12.19 folgt, daß $\|E\|_{\ell_h^2} \leq 1$. Sei nun c stetig mit $\|c\|_{L^\infty} < \infty$. Überlegen Sie sich, daß dann $\|E\|_{\ell_h^2} \leq \|E_0\|_{\ell_h^2} + k\|c\|_{L^\infty}$ gilt. Schließen Sie, daß damit auch E die von-Neumann-Bedingung erfüllt. ■

Übung 12.22 betrachten Sie die Diskretisierung der Wärmeleitungsgleichung $u_t - u_{xx} = 0$ mit dem expliziten Eulerverfahren in der Zeit (und symmetrischen Differenzen im Ort):

$$u_i^{n+1} - u_i^n = \sigma\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right), \qquad \sigma := \frac{k}{h^2}$$

Zeigen Sie, daß der Verstärkungsfaktor $g(\xi) = 1 - 4\sigma\sin^2(\xi h/2)$. Was können Sie über die Stabilität des Verfahrens in Abhängigkeit von k und h aussagen? ■

Bemerkung 12.23

- die von-Neumann-Analyse kann auch für Systeme (und damit auch für Mehrschrittverfahren wie dem leap frog Verfahren) durchgeführt werden—siehe Übungen.
- Im allg. liefert es nur *notwendige* Bedingungen für Stabilität, aber nicht hinreichend. In der Praxis liefert es aber doch eine recht gute Vorstellung davon, was die CFL-Bedingung ist.
- Für Probleme mit nichtkonstanten Koeffizienten geht man typischerweise so vor, daß man in einem ersten Schritt den Wertebereich der Koeffizienten bestimmt und dann eine von-Neumann-Analyse für die Gleichung mit eingefrorenen Koeffizienten ("freezing of coefficients") durchführt. Im Prinzip kann dies auch für nichtlineare Gleichungen durchgeführt werden. Auf diese Weise erhält man natürlich nicht scharfe Abschätzungen für die CFL-Bedingung, aber oft brauchbare Schätzwerte.

■

12.2.1 Stabilitätsanalyse des Leap Frog Verfahrens

Vorbemerkung: Leap Frog ist ein wichtiger Vertreter von 2-Schritt-Verfahren. Mehrschrittverfahren können als Vektoreinschrittverfahren verstanden und analysiert werden (siehe Übung), aber eine direkte Analyse ist oft einfacher.

Für die Advektionsgleichung ist das Verfahren

$$\frac{u_i^{n+1} - u_i^{n-1}}{2k} + a\frac{u_{i+1}^n - u_{i-1}^n}{2h} = 0 \tag{12.17}$$

Bemerkung 12.24

- Man benötigt 2 Startwerte u_0^0 und u_i^1, $i \in \mathbb{Z}$. Die Werte u_i^1 können mit einem Einschrittverfahren erzeugt werden.

[4] genauer: von der Ordnung 4. Allg. ist ein Verfahren dissipativ von der Ordnung $2r$, falls der Verstärkungsfaktor g die Beziehung $|g(\xi)| \leq (1+Ck)(1-C|\xi h|^{2r})$, für $\xi \in (-\pi/h, \pi/h)$ erfüllt. Siehe die Diskussion in Abschnitt 12.3. LW ist strikt dissipativ, Lax-Friedrichs ist dissipativ aber nicht strikt dissipativ; siehe Diskussion in [11, p. 179].

- Mit dem Vektor
$$U_i^n := \begin{pmatrix} u_i^n \\ u_i^{n-1} \end{pmatrix}$$
kann das Verfahren (12.17) als Einschrittverfahren formuliert und analysiert werden.

■

Wir machen eine Stabilitätsanalyse mittels Fouriertransformation: die Transformierte $\widehat{u}^n(\xi) := \mathcal{F}_h((u_i^n)_i)(\xi)$ erfüllt

$$\widehat{u}^{n+1}(\xi) + 2\mathbf{i}a\lambda\sin(\xi h)\widehat{u}^n(\xi) - \widehat{u}^{n-1}(\xi) = 0, \qquad \lambda = \frac{k}{h}. \tag{12.18}$$

Für feste ξ, h ist dies eine Rekurrenz in n, die aufgelöst werden kann:

Übung 12.25 Seien α, $\beta \in \mathbb{C}$. Seien x_1, x_2 die Lösungen von $0 = x^2 + \alpha x + \beta$. Betrachte alle Folgen $(v_n)_n$ mit
$$v_{n+1} + \alpha v_n + \beta v_{n-1} = 0, \qquad \forall n \geq 1.$$
Dann gilt:

(a) Der Raum diese Folgen ist ein 2-dimensionaler Vektorraum
(b) Falls $x_1 \neq x_2$, dann sind die beiden Folgen $(x_1^n)_n$, $(x_2^n)_n$ eine Basis dieses Raumes.
(c) Falls $x_1 = x_2$, dann sind $(x_1^n)_n$ und $(nx_1^n)_n$ eine Basis dieses Raumes.

Seien nun $g_+(\xi)$ und $g_-(\xi)$ Lösungen von
$$g^2 + 2\mathbf{i}a\lambda\sin(\xi h)g - 1 = 0$$
d.h.
$$g_\pm(\xi h) = -\mathbf{i}a\lambda\sin(\xi h) \pm \sqrt{1 - (a\lambda)^2\sin^2(\xi h)}.$$
Damit können wir $\widehat{u}^n$ schreiben als:
$$\begin{aligned} \widehat{u}^n(\xi) &= \gamma(\xi)(g_+(\xi h))^n + \delta(\xi)(g_-(\xi))^n && \text{falls } g_+(\xi h) \neq g_-(\xi h), \\ \widehat{u}^n(\xi) &= \gamma(\xi)(g_+(\xi h))^n + \delta(\xi)n(g_+(\xi))^n && \text{falls } g_+(\xi h) = g_-(\xi h). \end{aligned}$$
Die Funktionen $\gamma(\xi)$, $\delta(\xi)$ ergeben sich aus den Startwerten $\widehat{u}^0(\xi)$, $\widehat{u}^1(\xi)$. Stabilität der Diskretisierung fordert hier
$$|g_+(\xi h)| \leq 1 \quad \text{und} \quad |g_-(\xi h)| \leq 1 \qquad \forall \xi \in (-\pi/h, \pi/h), \tag{12.19}$$
$$\text{falls } g_+(\xi) = g_-(\xi) \text{ dann muss} |g_+(\xi h)| < 1 \quad \text{sein.} \tag{12.20}$$

Wir beobachten nun:

- $|a\lambda| > 1$ impliziert $|g_-(-\pi/(2h))| > 1$ woraus folgt, daß die CFL-Bedingung $|a\lambda| \leq 1$ gelten muß.
- $|a\lambda| \leq 1$ impliziert $|g_+|^2 = |g_-|^2 = 1 - (a\lambda)^2\sin^2(\xi h) + (a\lambda\sin(\xi h))^2 = 1$, d.h. das Verfahren ist stabil, es sei denn, $g_+(\xi h) = g_-(\xi h)$. In diesem Fall muß $1 = |a\lambda\sin(\xi h)|$, d.h. $|a\lambda| = 1$ und $\xi = \pm\pi/(2h)$. Dann ist $g_+ = g_- = \pm\mathbf{i}$ d.h. (12.20) ist verletzt.

Zusammenfassung: Falls $|a\lambda| < 1$, dann ist Leap Frog stabil. Falls $|a\lambda| = 1$, dann ist es (schwach) instabil.

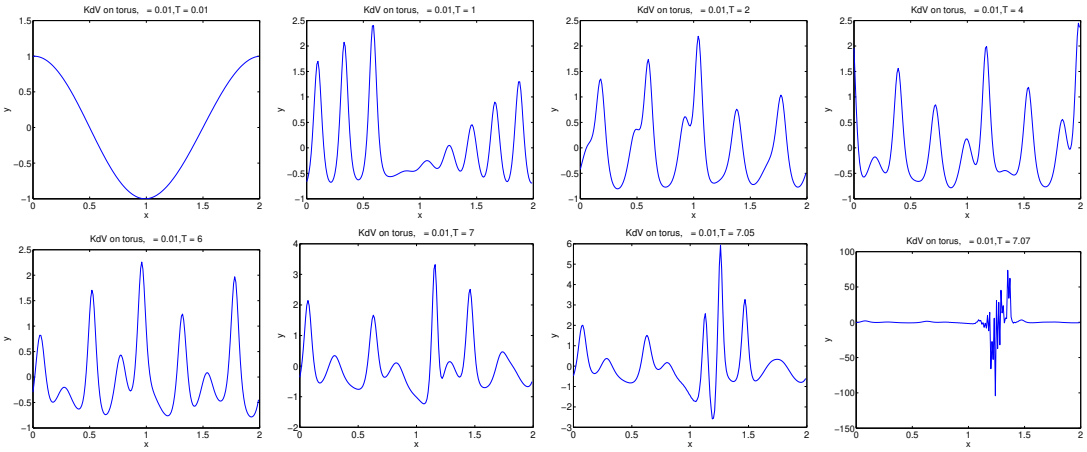
12.3 Dissipative Verfahren

12.3.1 Vorbetrachtung:

Ausgangspunkt:

1. für lineare Gleichungen mit konstanten Koeffizienten liefert die von-Neumann-Analyse ein einfaches Werkzeug zur Stabilitätsanalyse

Abbildung 12.6: das Zabusky-Kruskal-Verfahren für die KdV-Gleichung



2. für *Systeme*

$$\mathbf{U}^{n+1} = E\mathbf{U}^m + k\mathbf{G}^n$$

untersucht man analog die Fouriertransformierte $\widehat{\mathbf{E}}(\xi)$. Der Einfachheit halber untersucht man den Spektralradius $\rho(\widehat{\mathbf{E}})$ für $\xi \in (-\pi/h, \pi/h)$. Man sagt, das Verfahren erfüllt die von-Neumann-Bedingung, falls $\rho(\widehat{\mathbf{E}}(\xi)) \leq 1 + Ck$ für alle $\xi \in (-\pi/h, \pi/h)$.

Probleme: variable Koeffiziente? Nichtlineare Gleichungen?

"Standardvorgehen": von-Neumann-Analyse für "frozen coefficients", d.h. man macht von-Neumann-Analyse für alle (erwarteten) Werte der Koeffizienten. Oft gibt dies gute Hinweise auf Stabilität bzw. Schrittweitenbeschränkungen. Für dissipative Verfahren und gewisse Problemklassen reicht in der Tat die Methode des "frezzing of coefficients". Dies triff insbesondere auf parabolische Probleme zu.

Das folgende Beispiel zeigt, dass die Methode des "freezing of coefficients" nicht immer zuverlässig ist:

Beispiel 12.26 (Zabusky-Kruskal-Verfahren für KdV) Wir betrachten die KdV-Gleichung $u_t + (\alpha u^2 + \nu u_{xx})_x = 0$ mit der Leap frog-artigen Diskretisierung

$$u_j^{n+1} = u_j^{n-1} - \frac{2\alpha\mu}{3}\left(u_{j-1}^n + u_j^n + u_{j+1}^n\right)\left(u_{j+1}^n - u_{j-1}^n\right) - \frac{\nu\mu}{h^2}\left(u_{j+2}^n - 2u_{j+1}^n + 2u_{j-1}^n - u_{j-2}^n\right), \qquad \mu = \frac{k}{h}.$$

Die von Neumann-Analyse liefert nach "freezing of coefficients"

$$k < \frac{h}{4|\nu|/h^2 + 2|\alpha u_{max}|}$$

Wir betrachten das Verfahren mit

- $\nu = 0.022^2, \qquad \alpha = 0.5, \qquad u_0(x) = \cos(\pi x), \qquad u(0,t) = u(2,t)$
- Beobachtung: $|u| \leq 5$
- $\Longrightarrow$ Erwartung: $h = 0.01$ und $k < 0.004$ sind stabil.
- Numerik: $h = 0.01$ und $k = 0.0001$

Die Numerik in Fig. 12.6 zeigt, daß diese konservative Schrittweitenwahl nicht ausreicht! Tatsächlich kommt es bei $T > 21/\pi$ zu einem blow-up, obwohl die exakte Lösung für alle Zeiten existiert. ∎

Das Versagen des Zabusky-Kruskal-Verfahrens [21][5] führt man allgemein auf eine "Instabilität" zurück in dem Sinn, daß hochfrequente Lösungsanteile (d.h. große ξ im Fourierbild) durch die Nichtlinearität verstärkt werden. Ein manchmal propagiertes Mittel ist deshalb, Verfahren mit (etwas) Dissipation zu verwenden.

[5] in diesem Paper wird der Begriff der Solitonen eingeführt

12.3.2 Dissipative Verfahren

Definition 12.27 *Ein Verfahren mit Propagationsoperator* $\mathbf{E}$ *heißt* dissipativ[6] *von der Ordnung* $2r$*, falls*

$$\rho(\widehat{\mathbf{E}}(\xi)) \leq (1 - \delta|\xi h|^{2r})(1 + Ck)$$

für Konstante C*,* $\delta > 0$ *unabhängig von* k*,* h*.*

Beispiel 12.28 Betrachte das Lax-Wendroff-Verfahren für $u_t - u_x = 0$:

$$\begin{aligned} g(\xi) &= 1 + \mathbf{i}\lambda \sin(\xi h) - 2\lambda^2 \sin^2(\xi h/2) \\ |g(\xi)|^2 &= (1 - 2\lambda^2 \sin^2(\xi h/2))^2 + \lambda^2 \sin^2(\xi h) = 1 - 4\lambda^2(1 - \lambda^2)\sin^4(\xi h/2) \end{aligned}$$

Damit ist das Verfahren dissipativ von der Ordnung $2r = 4$, *falls* $\lambda = k/h < 1$ (verwende $\sqrt{1-x} = 1 - 1/2x + O(x^2)$ für kleine x). ∎

Dissipativität tritt relativ natürlich bei Diskretisierungen von parabolischen Gleichungen auf, weil das kontinuierliche Problem ein solche Eigenschaft hat:

Beispiel 12.29 Betrachte die explizite Eulerdiskretisierung der Wärmeleitungsgleichung aus Übung 12.22 mit Verstärkungsfaktor $g(\xi)$ $g(\xi) = 1 - 4\sigma \sin^2(\xi h/2)$. Stabilität ($|g(\xi)| \leq 1$) verlangt also $\sigma \leq 1/2$. Für $\sigma \leq 1/2$ ist das Verfahren dissipativ von der Ordnung 2.

Übung: Für das implizite Eulerverfahren ist das Verstärkungsfaktor $g(\xi) = 1/(1 + 4\sigma \sin^2(\xi h/2))$. Das Verfahren ist also stabil und ebenfalls dissipativ von der Ordnung 2. ∎

Im einfachsten Fall liefert Dissipativität, daß konsistente Verfahren tatsächlich stabil sind:

Satz 12.30 (Kreiss) *Betrachte das strikt hyperbolische System*

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0$$

d.h. die konstante Matrix $\mathbf{A}$ *habe paarweise verschiedene reelle Eigenwerte. Dann ist ein (explizites) Differenzenverfahren stabil, wenn es konsistent und dissipativ von der Ordnung* $2r > 0$ *ist.*

Beweis: siehe [3, Thm. 5.2] und z.B. [11, Thm. 6.5.2]. □

Bemerkung 12.31 Das Zusammenspiel von Konsistenz und Dissipativität bereits bei parabolischen Gleichungen im Abschnitt 11.6.4 gesehen haben bei der Analyse der Funktion F_n: für "kleine $z = \lambda_n k$ nutzten wir die Konsistenz, für große $z = \lambda_n k$ benötigten wir die L-Stabilität des Einschrittverfahrens. Ähnliches liegt auch Satz 12.30 zugrunde: für konsistente Verfahren erwartet man, daß das Stabilitätsverhalten für kleine ξ sich aus dem kontinuierlichen Problem erhalten läßt (cf. Satz 12.32)), während das Stabilitätsverhalten für große ξ eine zusätzliche Eigenschaft ist.

Wie Satz 12.30 zeigt, kann Dissipativität eines Verfahrens eine nützliche Eigenschaft sein. Tatsächlich haben sehr viele Verfahren (auch für hyperbolische Probleme) ein wenig Dissipation—das Lax-Wendroff-Verfahren ist ein Beispiel. Während für parabolische Probleme wie der Wärmeleitungsgleichung Dissipation eine Eigenschaft des kontinuiertlichen Problems ist, ist es bei hyperbolischen Erhaltungsgleichungen meist nicht. Man versucht deshalb meist, die Dissipation des Verfahrens möglichst gering zu halten.

noch ueberarbeiten–streichen Aus der Konsistenz eines Verfahrens und der Wohlgestelltheit des kontinuierlichen Problems kann man $\widehat{\mathbf{E}}(\xi)$ für "kleine " ξ kontrollieren:

Satz 12.32 *Betrachte das lineare Problem* $\mathbf{u}_t = P\mathbf{u}$ *mit einen linearen Ortsdifferentialoperator* P *mit konstanten Koeffizienten. Habe das numerische Verfahren Konsistenzordnung* p *(in Ort und Zeit). Dann gilt:*

$$\|\widehat{\mathbf{E}}(\xi) - e^{\widehat{P}(\mathbf{i}\xi)k}\| \leq C|\xi h|^{p+1},$$

wobei die Matrix $e^{\widehat{P}(\mathbf{i}\xi)k}$ *den exakten Lösungsoperator beschreibt, d.h. die Lösung von* $\mathbf{u}_t = P\mathbf{u}$ *mit Anfangsbedingung* $\mathbf{u}(\cdot, 0) = e^{\mathbf{i}\xi x}\mathbf{u}_0$ *ist* $e^{\widehat{P}(\mathbf{i}\xi)t}\mathbf{u}_0$*, wenn* $\mathbf{u}_0 \in \mathbb{R}^s$ *ist, mit Problemgröße* $s \in \mathbb{N}$*.*

[6] manchmal auch: strikt dissipativ

Beispiel 12.33 Betrachte das Lax-Wendroff-Verfahren für $u_t - u_x = 0$. Der Operator $e^{\widehat{P}(\mathbf{i}\xi)t}$ ergibt sich aus

$$e^{\widehat{P}(\mathbf{i}\xi)t} e^{\mathbf{i}\xi x} \stackrel{!}{=} u(x,t) = e^{\mathbf{i}\xi(x+t)},$$

d.h.

$$e^{\widehat{P}(\mathbf{i}\xi)t} = e^{\mathbf{i}\xi t}.$$

Der Verstärkungsfaktor $g(\xi) = \widehat{E}(\xi)$ von Lax-Wendroff ist $g(\xi) = 1 + \mathbf{i}\lambda\sin(\xi h) - 2\lambda^2\sin^2(\xi h/2)$ Damit

$$g(\xi) - e^{\widehat{P}(\mathbf{i}\xi)k} = 1 + \mathbf{i}\lambda\sin(\xi h) - 2\lambda^2\sin^2(\xi h/2)e^{\mathbf{i}\xi k} = \cdots = O((\xi h)^3),$$

was zur Konsistenzordnung $p = 2$ von Lax-Wendroff paßt. ∎

Beweis:[von Satz 12.32] Sei $\widehat{\mathbf{u}} \in C_0^\infty(-\pi/h, \pi/h)$. Definiere

$$\mathbf{u}(x) = \int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x}\widehat{\mathbf{u}}(\xi)\,d\xi$$

Um die Konsistenz auszunutzen, schreiben wir einen Schritt des Verfahrens:

$$\widehat{\mathbf{E}\mathbf{u}}(\xi) = \widehat{\mathbf{E}}(\xi)\widehat{\mathbf{u}}(\xi)$$
$$\Longrightarrow \mathbf{u}^1(x_j) = \int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x}\widehat{\mathbf{E}}(\xi)\widehat{\mathbf{u}}(\xi)\,d\xi$$

Die exakte Evolution ist

$$\mathbf{u}(x,k) = \int_{-\pi/h}^{\pi/h} e^{\widehat{P}(\mathbf{i}\xi)k}\widehat{\mathbf{u}}(\xi)e^{\mathbf{i}\xi x}\,d\xi$$

Damit ist der Fehler

$$\mathbf{u}(x_j,k) - \mathbf{u}^1(x_j) = \int_{-\pi/h}^{\pi/h} e^{\mathbf{i}\xi x}(e^{\widehat{P}(\mathbf{i}\xi)k} - \widehat{\mathbf{E}}(\xi))\widehat{\mathbf{u}}(\xi)\,d\xi$$

Für $x_j = 0$ folgt damit wegen der Konsistenzordnung p:

$$\left|\int_{-\pi/h}^{\pi/h}(e^{\widehat{P}(\mathbf{i}\xi)} - \widehat{\mathbf{E}}(\xi))\widehat{\mathbf{u}}(\xi)\,d\xi\right| \le Ch^{p+1}\|u^{(p+1)}\|_{L^\infty(\mathbb{R})}$$

Läßt man $\widehat{\mathbf{u}} \to \delta_{\xi'}$ laufen für ein festes $\xi' \in (-\pi/h, \pi/h)$ und nutzt, daß dann $\|u^{(p+1)}\|_{L^\infty(\mathbb{R})} \le C|\xi'|^{p+1}$, so ergibt sich die Behauptung. □

Satz 12.32 ist interessant, weil er zeigt, daß aus der Stabilität des kontinuierlichen Problems (Kontrolle von $e^{\widehat{P}(\mathbf{i}\xi)}$) und Konsistenz des Verfahrens man Kontrolle von $\widehat{\mathbf{E}}(\xi)$ für "kleine" ξ erhalten kann. Für große ξ muß dies aus dem Verfahren herrühren, z.B. durch Dissipativität.

"Herleitung" von Lax-Wendroff

Wie wir es bei parabolischen Gleichungen gemacht haben, könnte man zahlreiche Verfahren erzeugen, indem man zuerst eine Semidiskretisierung im Ort durchführt und dann ein Zeitschrittverfahren darauf anwendet. Das Lax-Wendroff-Verfahren (für $u_t + au_x = 0$) wird anders "hergeleitet": Mit den Differenzenoperatoren D_+, D_-, D_0 (z.B. $D_+u(x) = u(x+h) - u(x)$, $D_0u(x) = u(x+h) - u(x-h)$) ergibt Taylor

$$u(t+k,x) = u + ku_t + \frac{k^2}{2}u_{tt} + O(k^2)$$

Die Gleichung $u_t + au_x = 0$ liefert

$$u_t = -au_x, \qquad u_{tt} = a^2u_{xx}. \tag{12.21}$$

So können Zeitableitungen durch Ortsableitungen ersetzt werden und anschließend durch Differenzenquotienten approximiert werden:

$$u(t+k,x) = u - \frac{k}{2h}aD_0u + \frac{k^2}{2h^2}a^2D_+D_-u + kO(k^2+h^2)$$

Das zeigt, daß die Konsistenzordnung des Verfahrens

$$u_i^{n+1} = u_i^n - \frac{\lambda}{2}\left(u_{i+1}^n - u_{i-1}^n\right) + \frac{\lambda^2}{2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^m\right) = 0, \qquad \lambda = \frac{ka}{h}$$

gerade 2 (genauer: $(2,2)$) ist.

Modifizierte Gleichung

Die Fourieranalyse (d.h. die von-Neumann-Analyse) ist eine Möglichkeit, das Verhalten von Diskretisierungen zu verstehen. Eine andere Möglichkeit, das qualitative Verhalten von Verfahren zu verstehen ist, sie als "bessere" Approximation an eine andere ("modifizierte") Gleichung zu interpretieren und zu stipulieren, daß das qualitative Verhalten dieser kontinuierlichen Gleichung das numerische Verfahren beschreibt. Wir illustrieren das mit drei klassischen Beispielen:

Beispiel 12.34 (modifizierte Gleichung des upwind-Verfahrens) Wir betrachten $u_t + au_x = 0$ mit $a < 0$. Das Verfahren ist

$$u_i^{n+1} = u_i^n - a\lambda(u_{i+1}^n - u_i^n), \qquad \lambda = \frac{k}{h}.$$

Taylor um (t, x) liefert

$$\begin{aligned}
u(t+k,x) &= u + ku_t + \frac{k^2}{2}u_{tt} + \frac{k^3}{6}u_{ttt} + \cdots, \\
u(t,x+h) &= u + hu_x + \frac{h^2}{2}u_{xx} + \frac{h^3}{6}u_{xxx} + \cdots, \\
\tau &= \frac{u(t+k,x)-u}{k} + a\frac{u(t,x+h)-u}{h} = \left[u_t + \frac{k}{2}u_{tt} + \frac{k^2}{6}u_{ttt} + \cdots\right] + a\left[u_x + \frac{h}{2}u_{xx} + \frac{h^2}{6}u_{xxx} + \cdots\right] \\
&= u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} + O(k^2+h^2)
\end{aligned}$$

Falls wir $u_t + au_x = 0$ ausnutzen, dann erhalten wir, daß das Upwindverfahren Konsistenzordnung $(1,1)$ hat. Falls wir jedoch annehmen, daß u die Gleichung

$$u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} = 0 \tag{12.22}$$

erfüllt, dann erhalten wir (formal) Konsistenzordnung $(2,2)$. Man nennt (12.22) die modifizierte Gleichung. Typischerweise formuliert man die modifizierte Gleichung nochmals um, indem man Zeitableitungen durch Ortsableitungen ausdrückt: aus (12.22) ergibt sich (formal) durch Differenzieren nach t und x:

$$u_{tt} + au_{xt} = O(k+h), \qquad u_{xt} + au_{xx} = O(k+h), \tag{12.23}$$

daß wir (12.22) schreiben können (unter Vernachlässigung von Termen $O(h^2+k^2)$)

$$\begin{aligned}
0 = u_t + au_x + \frac{1}{2}ku_{tt} + \frac{1}{2}hau_{xx} &\overset{(12.23)}{=} u_t + au_x + \frac{1}{2}\left[-kau_{xt} + hau_{xx} + kO(k+h)\right] \\
&\overset{(12.23)}{=} u_t + au_x + \frac{1}{2}\left[ka^2u_{xx} + hau_{xx} + kO(k+h)\right] = u_t + au_x + \frac{1}{2}\left[ka^2 + ha\right]u_{xx} + O(k^2+h^2).
\end{aligned}$$

Damit approximiert das obige (!) Upwindverfahren die Gleichung

$$u_t + au_x - \nu u_{xx} = 0, \qquad \nu := -\frac{1}{2}\left[ka^2 + ah\right] \overset{a\le 0}{=} \frac{h}{2k/h}\left[\frac{k}{h}|a| - \frac{k^2}{h^2}a^2\right] \tag{12.24}$$

mit Konsistenzordnung $(2,2)$. Wir bemerken, daß $\nu > 0$, falls die CFL-Bedingung erfüllt ist. Man kann sich also vorstellen, daß das Upwindverfahren zwar die Advektionsgleichung approximiert, aber mit höherer Genauigkeit die Wärmeleitungsgleichung (12.24) mit (kleiner) Diffusion. Da die Wärmeleitungsgleichung Dissipation hat, erwartet man, daß auch das Upwindverfahren dissipativ ist. ∎

Übung 12.35 Zeigen Sie, daß für das Lax-Friedrichs-Schema die modifizierte Gleichung

$$u_t + au_x + \nu u_{xx} = 0, \qquad \nu = \frac{h}{2\lambda}\left(1 - \lambda^2a^2\right), \qquad \lambda = \frac{k}{h}$$

ist. Insbesondere ist die Diffusionskonstante ν für Lax-Friedrichs größer als für Upwind.

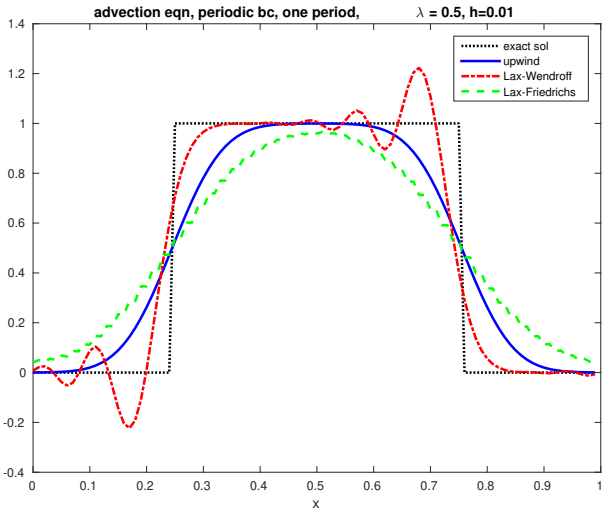


Abbildung 12.7: Upwind, LW, LF für Advektionsgleichung mit unstetigen Anfangsbedingungen

Übung 12.36 Die modifizierte Gleichung für das Lax-Wendroff-Verfahren ist gegeben durch

$$u_t + au_x + \frac{ah^2}{6}\left(1 - \frac{k^2}{h^2}a^2\right)u_{xxx} = 0$$

Beispiel 12.37 Wir betrachten die Advektionsgleichung mit periodischen Randbedingungen:

$$u_t - u_x = 0 \quad \text{auf } (0,1) \times \mathbb{R}^+, \qquad u(0,t) = u(1,t) \quad \forall t > 0. \tag{12.25}$$

Um den Einfluß der Verfahren auf die hochfrequenten Anteil der Anfangsbedingung herauszuarbeiten, ist $u_0(x) = \chi_{[0.25,0.75]}(x)$ (und damit bleibt die exakte Lösung auch eine stückweise konstante Funktion). In Fig. 12.7 wird der Ergebnis verschiedener Verfahren bei $T = 1$ und $k/h = 0.5$ gezeigt. Die modifizierte Gleichung für das Upwindverfahren und Lax-Friedrichs ist eine parabolische Gleichung mit relativ viel Dissipation (Lax-Friedrichs hat sogar noch mehr Dissipation als das Upwindverfahren). Das ist spiegelt sich in Fig. 12.7 wider. Die modifizierte Gleichung von Lax-Wendroff ist eine Gleichung 3. Ordnung, bei der Dispersion eine große Rolle spielt. Das spiegelt sich in den Oszillationen in der numerischen Lösung wider. ■

Bemerkung 12.38 (Dispersion) Die Lösung u von

$$u_t + au_x = 0, \qquad u(0,x) = u_0(x)$$

ist

$$u(t,x) = u_0(x - at) = \frac{1}{\sqrt{2\pi}} \int_\xi \widehat{u}_0(\xi) e^{\mathbf{i}\xi(x-at)}\, d\xi$$

(diese Darstellung erhält man direkt aus der Fourierinversionsformel[7] oder indem man die Differentialgleichung fouriertransformiert und dann löst). Wir interpretieren dies so: Die Fourierkomponente $\widehat{u}_0(\xi)e^{\mathbf{i}\xi x}$ zur Frequenz ξ breitet sich mit Geschwindigkeit a (nach rechts) aus. Die Geschwindigkeit a ist *unabhängig* von ξ. Sei nun $\xi \mapsto a(\xi)$ eine Funktion von ξ. Dann stellt $v(x,t) = \frac{1}{\sqrt{2\pi}} \int_\xi \widehat{u}_0(\xi) e^{\mathbf{i}\xi(x-a(\xi)t)}\, d\xi$ eine Funktion dar mit folgender Eigenschaft: die Fourierkomponente von $v(\cdot,0)$, die zur Frequenz ξ gehört, wandern mit Geschwindigkeit $a(\xi)$. Allgemein spricht man von *Dispersion*, wenn die Fourierkomponenten sich mit ξ-abhängiger Geschwindigkeit ausbreiten.

Wir lösen die Gleichung

$$u_t + au_x + \nu u_{xxx} = 0, \qquad u(0,x) = u_0(x)$$

Fouriertransformation (in x) liefert die ODE

$$\widehat{u}(t,\xi) + a\mathbf{i}\xi\widehat{u}(t,\xi) + \nu(\mathbf{i}\xi)^3\widehat{u}(t,\xi) = 0, \qquad \widehat{u}(0,\xi) = \widehat{u}_0(\xi)$$

[7] Wir verwenden $\widehat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \int_x e^{-\mathbf{i}x\xi}(x)\, dx$ mit Inversionsformel $u(x) = \frac{1}{\sqrt{2\pi}} \int_\xi e^{\mathbf{i}x\xi}\widehat{u}(\xi)\, d\xi$

mit Lösung

$$\widehat{u}(t,\xi) = \exp(-(a\mathbf{i}\xi + \nu(\mathbf{i}\xi)^3)t)\widehat{u}_0(\xi)$$

Rücktransformation liefert

$$u(t,x) = \frac{1}{\sqrt{2\pi}}\int_\xi e^{\mathbf{i}\xi x} e^{-(a\mathbf{i}\xi+\nu(\mathbf{i}\xi)^3)t}\widehat{u}_0(\xi), d\xi \frac{1}{\sqrt{2\pi}}\int_\xi e^{\mathbf{i}\xi\left[x-(a+\nu(\mathbf{i}\xi)^2)t\right]}\widehat{u}_0(\xi), d\xi.$$

Wir erkennen insbesondere, daß Ausbreitung der unterschiedlichen Fourierkomponenten der Anfangswerte von u_0 unterschiedliche Geschwindigkeit haben. Dies ist beim Lax-Wendroff-Verfahren in Fig. 12.7 erkennbar, denn es "erklärt" die Oszillationen. ∎

12.4 Splittingmethoden/fractional step methods

12.4.1 Einführung: der ODE-Fall

Betrachte das ODE-System

$$u' = Au + Bu, \qquad u(0) = u_0, \tag{12.26}$$

wobei A, B Matrizen seien (im Allgemeinen könnten dies auch Operatoren sein). Eine mögliche Herkunft dieser Matrizen könnte eine Ortsdiskretisierung sein. Die Lösungformel ist

$$u(t) = e^{t(A+B)}u_0$$

mit der Matrixexponentialfunktion.[8]

Ziel: eine gute Approximation an $u(k)$, wobei k der Zeitschritt ist.

Spielregel: $u_0 \mapsto e^{k(A+B)}u_0$ ist schwer/teuer zu realisieren, aber die Abbildungen $v \mapsto e^{kA}v$ und $v \mapsto e^{kB}u$ sind "einfach".

Die "klassischen" Verfahren sind:

- Lie-Splitting: $e^{t(A+B)} \approx e^{tB}e^{tA}$. Algorithmisch ist dies für einen Schritt der Länge k:

$$u^{1/2} := e^{kA}u_0, \qquad u^1 := e^{kB}u_{1/2}$$

- Strang-Splitting: $e^{t(A+B)} \approx e^{1/2tA}e^{tB}e^{1/2tA}$. Algorithmisch ist dies für einen Schritt der Länge k:

$$u^{1/3} := e^{k/2A}u_0, \qquad u^{2/3} := e^{kA}u_{1/3}, \qquad u^1 := e^{k/2A}u_{2/3}.$$

  Nota: das Strangsplitting sieht teurer aus als das Lie-Splitting. Tatsächlich ist es fast genauso teuer wie das Liesplitting, wenn man mehrere Schritte hintereinander ausführt. Wegen der Halbgruppeneigenschaft von $t \mapsto e^{tA}$ gilt:

$$e^{k/2A}e^{kB}e^{k/2A} \quad e^{k/2A}e^{kB}e^{k/2A} \quad \cdots \quad e^{k/2A}e^{kB}e^{k/2A} = e^{k/2A}e^{kB}e^{kA}e^{kB}\cdots e^{kB}e^{k/2A}$$

- das SWSS ("symmetrically weighted sequential splitting") $e^{t(A+B)} \approx \frac{1}{2}\left(e^{tA}e^{tB} + e^{tB}e^{tA}\right)$. Algorithmisch ist dies für einen Schritt der Länge k:

$$u^{1/2} := e^{kA}e^{kB}u_0, \qquad \widetilde{u}^{1/2} := e^{kB}e^{kA}u_0, \qquad u^1 := \frac{1}{2}\left(u^{1/2} + \widetilde{u}^{1/2}\right).$$

Weil die Matrizen A, B i.a. nicht kommutieren, d.h. $AB \neq BA$, gilt i.A. *nicht* $e^{t(A+B)} = e^{tA}e^{tB}$—damit sind die obigen Verfahren i.a. wirklich nur (unterschiedliche) Approximationen an die exakte Evolution. Zur Beschreibung des Konsistenzfehlers benötigt man den *Kommutator* zweier Matrizen/Operatoren:

$$[A,B] := AB - BA$$

[8] Es hat sich eingebürgert, auch für (unbeschränkte) Operatoren A das Symbol $e^{tA}u_0$ zu verwenden. Dann ist es lediglich ein Symbol für den Operator $u_0 \mapsto u(t)$, wobei $u(t)$ die Lösung des AWP $u' = Au$, $u(0) = u_0$ ist.

Übung 12.39 Seien A, B diagonalisierbar und $[A, B] = 0$. Dann sind sie simultan diagonalisierbar. Insbesondere gilt

$$e^{t(A+B)} = e^{tA}e^{tB} = e^{tB}e^{tA}$$

■

Lemma 12.40 (Konsistenzfehler bei Splittingverfahren) *Für zwei Matrizen A, B gilt für den Konsistenzfehler der Splittingverfahren:*

(i) (Lie-Splitting)

$$\tau_L = u^1 - e^{k(A+B)}u_0 = \frac{1}{2}k^2[A, B]u_0 + O(k^3)$$

(ii) (Strang-Splitting)

$$\tau_S = u^1 - e^{k(A+B)}u_0 = k^3\left(\frac{1}{12}[B, [B, A]] - \frac{1}{24}[A, [A, B]]\right)u_0 + O(k^4)$$

(iii) (SWSS-Splitting)

$$\tau_{SW} = u^1 - e^{k(A+B)}u_0 = O(k^3)$$

Beweis: Taylor. Taylor. Taylor. □

Bemerkung 12.41 Die vorgestellten Splitting-Verfahren sind maximal 2. Ordnung. Im Prinzip ist es auch möglich, Splitting-Verfahren höherer Ordnung zu erzeugen. Das Lie- und das Strang-Splitting sind jedoch die meistverbreiteten Techniken. ■

12.4.2 Beispiele

Die Idee des Splittings wird oft eingesetzt. Dabei werden folgende Verallgemeinerungen verwendet:

- die Matrizen A, B können auch (Differential-)Operatoren oder deren (Orts-)Diskretisierungen sein
- die Operatoren A, B können auch nichtlinear sein

Bezeichnet man mit E_A und E_B die Evolutionsoperatoren für die Probleme

$$\begin{aligned} u' &= Au, \qquad u(0) = u_0 \\ u' &= Bu, \qquad u(0) = u_0 \end{aligned}$$

so ist ein Schritt (der Länge k) des Lie-Splittings gegeben durch

$$u_1 := E_B(k)E_A(k)u_0$$

und analog ein Schritt des Strang-Splittings $u_1 := E_A(k/2)E_B(k)E_A(k/2)u_0$.

Bemerkung 12.42 In der Praxis kann man die Evolutionen E_A und/oder E_B nicht exakt durchführen. Man behilft sich dann mit geeigneten Approximationen—im einfachsten Fall wäre es z.B. ein expliziter Eulerschritt. ■

Beispiel 12.43 Betrachte die 2D-Advektionsgleichung

$$u_t + au_x + bu_y = 0 \qquad \text{in } \Omega = \mathbb{R}^2, \qquad u(\cdot, 0) = u_0(\cdot) \tag{12.27}$$

Die naheliegende Wahl der Operatoren A und B ist $Au = au_x$ und $Bu = bu_y$. Die exakten Evolutionen sind für eine Funktion $v = v(x, y)$:

$$(E_A v)(x, y, t) = v(x - at, y), \qquad (E_B v)(x, y, t) = v(x, y - bt),$$

Damit ein Schritt des Lie-Splittings $E_B E_A$ gegeben durch

$$v \mapsto v(x - at, y - bt).$$

Dies ist sogar die exakte Lösung, d.h. das Lie-Splitting ist exakt! Da die Operatoren A, B kommutieren, war dies aus dem Matrixfall zu erhoffen. ■

Übung 12.44 Betrachten Sie die Gleichung (12.27) mit $a = 1$ und $b = -1$.

(a) Formulieren Sie ein (klassisches) Upwindverfahren.

(b) Machen Sie analog zum 1D-Fall eine Fourieranalyse, um eine CFL-Bedingung herzuleiten.

(c) Formulieren Sie ein Splittingverfahren, indem Sie die Upwindverfahren (ft/bs bzw. ft/fs) kombinieren. Was ist die CFL-Bedingung des Splittingverfahrens?

(d) Programmieren Sie beide Verfahren für den Fall periodischer Randbedingungen. ∎

Splittingverfahren verwendet man gerne, wenn die beiden Operatoren A und B zu Differentialgleichungen mit verschiedenem Charakter gehören, die entsprechend unterschiedliche Behandlung/Diskretisierung benötigen.

Beispiel 12.45 Die Gleichung

$$u_t = \Delta u + \mathbf{b} \cdot \nabla u$$

wird oft zerlegt mit $Au = \Delta u$ und $Bu = \mathbf{b} \cdot \nabla u$. Die Evolution E_A verlangt dann die (approximative) Lösung einer Wärmeleitungsgleichung und E_B die Lösung der Advektionsgleichung. ∎

Splitting-Verfahren können auch auf der Ebene der Diskretisierung erhebliche Beschleunigungen bringen.

Beispiel 12.46 (Varianten von ADI–alternating directions implicit) Betrachte die 2D Wärmeleitungsgleichung (mit homogenen Dirichlet-Randbedingungen) auf $\Omega = (0,1)^2$: $u_t = u_{xx} + u_{yy}$. Wir betrachten die Semidiskretisierung mit stückweise linearen Ansatzfunktionen im Ort auf einem regelmäßigen Gitter mit $N \times N$ Knoten (d.h. N Knoten in jede Raumrichtung). Der Einsatz eines impliziten Verfahrens (z.B. impliziter Euler) verlangt die Lösung eines 2D-Ortsproblems in jedem Zeitschritt. Kosten: für die typische lexikographische Nummerierung der Unbekannten ist der Speicherbedarf für die LU-Zerlegung $N^2 b$, wobei $b = N$ die Bandbreite der Steifigkeitsmatrix ist. Damit sind die Kosten pro Zeitschritt (ohne Berücksichtigung der Kosten des Aufstellens der LU-Zerlegung) $O(N^3)$.

Ein mögliches Splittingverfahren basiert auf $Au = u_{xx}$ und $Bu = u_{yy}$. Damit ist die (diskrete) Evolution E_A beschrieben durch N *entkoppelte* Gleichungssystem (je von der Größe N), die sogar noch tridiagonal sind. Analog für E_B. Damit sind die Kosten für einen (Lie)-Zeitschritt $2NO(N) = O(N^2)$, was die optimale Komplexität ist. ∎

12.4.3 IMEX

Verwandt mit den oben diskutierten Splittingmethoden mit IMEX-Verfahren ("implicit explicit method"). Da implizite Zeitschrittverfahren insbesondere bei nichtlinearen Problem teuer sind, versucht man, nur Teile des Differentialoperators implizit zu rechnen und nur die "steifen" Anteile implizit.[9]

Betrachte die Evolution

$$\mathbf{u}' = A\mathbf{u} + B\mathbf{u}$$

Um zwei verschiedene Zeitdiskretisierungen unter einen Hut zu bekommen, machen wir gedanklich eine Zerlegung $\mathbf{u} = \phi + \psi$ und betrachten das System

$$\begin{aligned} \phi' &= A\mathbf{u} = A(\phi + \psi) \\ \psi' &= B\mathbf{u} = B(\phi + \psi) \end{aligned}$$

Für diese beiden Gleichungen verwenden wir nun 2 verschiedene Zeitdiskretisierungen (z.B. expliziten und impliziten Euler). Dabei will man das Verfahren schlußendlich so formulieren, daß nur die Funktion $\mathbf{u}$ benötigt wird. Eine typische Konstruktion ist mittels "partitionierten" RK-Verfahren: Seien zwei s-stufige RK-Verfahren mit Butcher-Tableaus

$$\begin{array}{c|c} c^p & A \\ \hline & b^\top \end{array} \qquad \begin{array}{c|c} c^q & \widehat{A} \\ \hline & \widehat{b}^\top \end{array}$$

[9] Verwandt sind die IMEX-Verfahren auch mit den sog. "Rosenbrock-Verfahren" (auch linear-implizite Verfahren genannt) im Kontext der klassischen ODE-Verfahren.

gegeben. Ein Schritt eines s-stufigen *partitionierten RK-Verfahrens* für das System $y' = f(y,z)$, $z' = g(y,z)$ ist definiert durch $y_1 = y_0 + k\sum_j b_j k_j$, $z_1 = z_0 + k\sum_j \widehat{b}_j \widehat{k}_j$ mit

$$\begin{aligned} k_i &= f(y_0 + k\sum_{j=1}^{s} a_{ij}k_j, z_0 + k\sum_{j=1}^{s} \widehat{a}_{ij}\widehat{k}_j), \qquad i = 1,\dots,s \\ \widehat{k}_i &= g(y_0 + k\sum_{j=1}^{s} a_{ij}k_j, z_0 + k\sum_{j=1}^{s} \widehat{a}_{ij}\widehat{k}_j), \qquad i = 1,\dots,s. \end{aligned}$$

Beispiel 12.47 Eine Differenzendiskretisierung im Ort der Advektions-Diffusionsgleichung

$$u_t = (u_{xx} + u_{yy}) - au_x$$

führt auf das ODE-System ($a > 0$, so daß eine "backward space" Diskretisierung die richtige ist)

$$\frac{d}{dt}u_{i,j} = \frac{1}{h^2}\left[D_x^+ D_x^- + D_y^+ D_y^-\right] u_{i,j} - \frac{a}{h} D_x^- u_{i,j}$$

Der Operator A gehört zur (Ortsdiskretisierung) des Laplace and B zum Transportterm. Schreibt man $u = \phi + \psi$ und verwendet für ϕ den impliziten Euler und für ψ den expliziten Euler, so ergibt sich

$$\begin{aligned} \frac{1}{k}\left(\phi_{i,j}^{n+1} - \phi_{i,j}^{n}\right) &= \frac{1}{h^2}\left[D_x^+ D_x^- + D_y^+ D_y^-\right] u_{i,j}^{n+1}, \\ \frac{1}{k}\left(\psi_{i,j}^{n+1} - \psi_{i,j}^{n}\right) &= -\frac{a}{h} D_x^- u_{i,j}^{n}. \end{aligned}$$

Zusammengefaßt also

$$u_{i,j}^{n+1} = u_{ij}^{n} + \frac{k}{h^2}\left[D_x^+ D_x^- + D_y^+ D_y^-\right] u_{i,j}^{n+1} - \frac{ak}{h} D_x^- u_{i,j}^{n}. \tag{12.28}$$

Da der Transportterm explizit behandelt wurde, gibt es die CFL-Bedingung $k|a|/h \le 1$. Das klassische implizite Eulerverfahren hätte auf

$$u_{i,j}^{n+1} = u_{ij}^{n} + \frac{k}{h^2}\left[D_x^+ D_x^- + D_y^+ D_y^-\right] u_{i,j}^{n+1} - \frac{ak}{h} D_x^- u_{i,j}^{n+1} \tag{12.29}$$

geführt (dann ohne CFL-Bedingung). Es ist zu betonen, daß das LGS in (12.28) ein SPD System ist, für das es effiziente Verfahren gibt. Das LGS in (12.29) ist nicht symmetrisch. ∎

Es gibt auch IMEX-Verfahren höherer Ordnung, die auf RK-Verfahren beruhen. Etwas einfacher (und älter) sind IMEX-Verfahren, die auf Mehrschrittverfahren beruhen. Dabei sind insbesondere Verfahren, die auf dem BDF2 (implizit) und Adams-Bashforth (explizit)-Verfahren beruhen, populär.

Übung 12.48 LMM-IMEX für $\mathbf{u}' = \mathbf{f}(t,\mathbf{u}) + \mathbf{g}(t,\mathbf{u})$ sind von der Form

$$\sum_{\ell=0}^{p} \alpha_\ell \mathbf{v}^{n+1-\ell} = k\sum_{\ell=1}^{p} \beta_\ell \mathbf{f}^{n+1-\ell} + k\sum_{\ell=0}^{p} \gamma_\ell \mathbf{g}^{n+1-\ell}.$$

Zeigen Sie: das Verfahren hat Ordnung $\widehat{p}$, wenn gilt:

$$\begin{aligned} &\sum_{\ell=0}^{p} \alpha_\ell = 0, \\ &\frac{1}{i!}\sum_{\ell=1}^{p} \ell^i \alpha_\ell = -\frac{1}{(i-1)!}\sum_{\ell=0}^{p} \ell^{i-1}\beta_\ell = -\frac{1}{(i-1)!}\sum_{\ell=0}^{p} \ell^{i-1}\gamma_\ell, \qquad i = 1,\dots,\widehat{p} \end{aligned}$$

Beispiel 12.49 IMEX-Verfahren oder Varianten werden auch verwendet, um das Lösen von nichtlinearen Gleichungen bei impliziten Verfahren zu umgehen. Im ODE-Fall sind wichtige Beispiele die sog. Rosenbrock-Verfahren ("linear implizite Verfahren"), bei denen eine ODE $y' = f(t,y)$ künstlich als $y' = f(t,y) =: f_1(t,y) + f_2(t,y)$ geschrieben wird. Eine typische Wahl ist $f_1(t,y) = Jy$ mit $J = \partial_y f(t,y)$. Wenn man f_1 implizit behandelt und f_2 explizit, dann müssen nur lineare Gleichungssysteme gelöst werden (Übung: überlegen Sie sich den Fall "impliziter Euler", "expliziter Euler").

Diese Idee kann z.B. bei Diskretisierungen von zeitabhängigen PDEs verwendet werden, indem die nichtlinearen Terme explizit behandelt werden. Z.B. wäre eine implizite Eulerdiskretisierung der inkompressiblen NSE

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{k} + \mathbf{u}^{n+1} \cdot \nabla \mathbf{u}^{n+1} - \nu \Delta \mathbf{u}^{n+1} + \nabla p^{n+1} = \mathbf{f}^{n+1}, \tag{12.30a}$$

$$\operatorname{div} \mathbf{u}^{n+1} = 0 \tag{12.30b}$$

Das System (nach Diskretisierung im Ort) ist nichtlinear. Man kann die Nichtlinearität z.B. so explizit behandeln:

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{k} + \mathbf{u}^{n} \cdot \nabla \mathbf{u}^{n+1} - \nu \Delta \mathbf{u}^{n+1} + \nabla p^{n+1} = \mathbf{f}^{n+1},$$

$$\operatorname{div} \mathbf{u}^{n+1} = 0$$

Dieses Verfahren kann man in verschiedener Art interpretieren:

1. als IMEX-Verfahren mit implizitem und explizitem Eulerverfahren und Zerlegung

$$\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{u}_t + [\mathbf{u}(t_n) \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p] + [(\mathbf{u} - \mathbf{u}(t_n)) \cdot \nabla \mathbf{u}]$$

2. als 1 Schritt eines iterativen Verf. (mit Startwert $\mathbf{u}^n$) zum Lösen des nichtlin. Systems (12.30). ∎

12.5 Raum-Zeit-DG

Wir betrachten eine etwas allgemeinere Form von (12.3):

$$u_t + \mathbf{b}(x,t) \cdot \nabla u + c(x,t) u = g. \tag{12.31}$$

Im Unterschied zu parabolischen Gleichungen hat in (12.31) die t-Variable keine besonders herausgehobene Rolle. Nennt man $x_0 = t$, so können wir auch auch folgendes, allgemeineres Problem betrachten:

$$\mathbf{b}(x) \cdot \nabla u + c(x) u = f \quad \text{auf } \Omega, \tag{12.32}$$

wobei wir jetzt $\Omega \subset \mathbb{R}^{d+1}$ zulassen. Bei dieser Gleichung kann nicht eine Randbedingung auf dem gesamten Rand $\partial\Omega$ gefordert werden. Wir definieren den "Einströmrand", den "Ausströmrand" und den "charakteristischen Rand" durch

$$\Gamma^- := \{x \in \partial\Omega \colon \mathbf{b}(x) \cdot n(x) < 0\}, \tag{12.33}$$

$$\Gamma^+ := \{x \in \partial\Omega \colon \mathbf{b}(x) \cdot n(x) > 0\}, \tag{12.34}$$

$$\Gamma^= := \{x \in \partial\Omega \colon \mathbf{b}(x) \cdot n(x) = 0\}. \tag{12.35}$$

Hier ist $n(x)$ der (äußere) Normalenvektor im Punkt $x \in \partial\Omega$. Tatsächlich können wir für (12.32) nur eine Randbedingung auf Γ^- oder Γ^+ fordern. Wir betrachten deshalb das Randwertproblem

$$\mathbf{b}(x) \cdot \nabla u + c(x) u = g \quad \text{auf } \Omega, \tag{12.36a}$$

$$u = 0 \quad \text{auf } \Gamma^-. \tag{12.36b}$$

Beispiel 12.50 Betrachten Sie das Problem

$$-u' + u = g \quad \text{auf } (0,1),$$

Überlegen Sie sich, daß man nicht sinnvollerweise $u(0) = 0 = u(1)$ fordern kann sondern nur $u(0) = 0$ *oder* $u(1) = 0$. ∎

Unser Ziel ist ein numerisches Verfahren für (12.36). Hierzu sei $\mathcal{T}$ ein Gitter, welches Γ^- auflöst, d.h. jede Randkante e erfüllt entweder $e \subset \Gamma^-$ oder $e \subset \partial\Omega \setminus \Gamma^-$. Für jedes Element K bezeichnet wir mit n_K die (äußere) Normale des Elementes K.

Zur Motivation nehmen wir an, daß die Lösung u von (12.36) hinreichend glatt ist. Sei v eine stückweise glatte Testfunktion, d.h. $v|_K$ ist glatt für jedes K. Betrachtet man ein Element K, so folgt aus (12.36a) durch Multiplikation mit v, Integration über K und partieller Integration:

$$\int_K gv = \int_K (c + \mathbf{b} \cdot \nabla u)v = \int_K u(c - u\nabla \cdot (\mathbf{b}v) + \int_{\partial K} (\mathbf{b} \cdot n_K)uv.$$

Durch Summation über alle Elemente $K \in \mathcal{T}$ ergibt sich:

$$\sum_{K \in \mathcal{T}} \int_K u\,(cv - \nabla \cdot (\mathbf{b}v)) + \int_{\partial K} (\mathbf{b} \cdot n_K)uv = \sum_{K \in \mathcal{T}} \int_K gv.$$

Wir führen den Begriff des Flußes auf dem Rand des Elementes K ein:

$$f_K = (\mathbf{b} \cdot n_K)u.$$

Es wird sich als zweckmäßig erweisen, die gemeinsame Kante/Fläche zwischen zwei Elementen K, K' zu bezeichnen; wir verwenden das Symbol

$$K|K',$$

wobei die Reihenfolge gleichzeitig eine "Orientierung" festlegt. Wir werden im Folgenden kurz "Kante" für diese Schnitte sagen, auch wenn es in höheren Dimensionen eigentlich Hyperfläche sind.

Es wird sich auch als zweckmäßig erweisen, die Nachbarelemente zu definieren:

$$\mathcal{N}(K) := \{K' \in \mathcal{T} \mid \ K \text{ und } K' \text{ teilen sich eine Manigfaltigkeit mit Kodimension } 1\}$$

Bei der Diskretisierung wird man u stückweise glatt ansetzen. Damit liegt es nahe, den Raum

$$\mathcal{S}^{p,0} := \{u \in L^2(\Omega) \colon u|_K \in \mathcal{P}_p \qquad \forall K \in \mathcal{T}\}$$

zu gegebenem $p \in \mathbb{N}_0$ zu betrachten. Für eine numerische Realisierung liegt es dann nahe, die Testfunktionen v ebenfalls aus diesem Raum zu wählen. Hierzu ist zu bemerken, daß bei unstetigem Ansatz für u und unstetigen Testfunktionen v *keine* Kopplung zwischen $u|_K$ und $u|_{K'}$ für benachbarte Elemente K und K' existiert. Diese Kopplung realisieren wir nun dadurch, daß auf der Kante $K|K'$, auf der eigentlich zwei Approximationen (nämlich $u|_K$ und $u|_{K'}$) an die exakte Lösung zur Verfügung stehen die Kopplung über einen "numerischen Fluß" $\widehat{h}_{K|K'}$ zu realisieren, d.h. den Fluß f_K wird durch einen "numerischen Fluß" $\widehat{h}_{K|K'}$ ersetzt. Eine sinnvolle Wahl des numerischen Flußes erscheint z.B.

Begriff der Konsistenz

$$\widehat{h}_{K|K'} = (\mathbf{b} \cdot n_K)\widehat{u}_{K|K'}$$

wobei es für die Wahl von $\widehat{u}$ viele Möglichkeiten gibt; plausibel erscheinen z.B.

- $\widehat{u}_{K|K'} = \frac{1}{2}(u|_K + u|_{K'})|_e$
- $\widehat{u}_{K|K'} = u|_K$
- $\widehat{u}_{K|K'} = u|_{K'}$

Für Randkanten $e \subset \Gamma^-$ wird man sinnvollerweise

$$\widehat{u}|_e = 0 \qquad \forall e \subset \Gamma^-$$

wählen.[10] Im vorliegenden Fall ist also der numerische Fluß $\widehat{h}_{K|K'}$ bereits eindeutig durch die Wahl von $\widehat{u}$ auf den Kanten festgelegt:

$$\widehat{h}_{K|K'} = \mathbf{b} \cdot n_K \widehat{u}|_{K|K'}.$$

[10] diese Wahl ist zwar naheliegend, aber nicht zwingend—weiterhin ist es zwar naheliegend, $\widehat{u}|_e$ nur abhängig von den Funktionswerten in den angrenzenden Elementen zu machen, aber auch das ist nicht zwingend

Es ergibt sich als numerisches Verfahren:

Finde $u \in S^{p,0}(\mathcal{T})$ s.d.

$$B_{DG}^{Trans}(u,v) := \sum_{K\in\mathcal{T}} \int_K u\,(cv + \nabla\cdot(\mathbf{b}v)) + \int_{\partial K} (\mathbf{b}\cdot n_K)\widehat{u}v = l(v) := \int_\Omega vg \qquad \forall v \in S^{p,0}(\mathcal{T}) \tag{12.37}$$

Üblicherweise wird diese Formulierung wieder partiell rückintegriert, und wir erhalten:

Finde $u \in S^{p,0}(\mathcal{T})$ s.d.

$$B_{DG}^{Trans}(u,v) = \sum_{K\in\mathcal{T}} \int_K (cu + \mathbf{b}\cdot\nabla u)v) + \int_{\partial K} (\mathbf{b}\cdot n_K)(\widehat{u} - u)v = l(v) := \int_\Omega vg \qquad \forall v \in S^{p,0}(\mathcal{T}) \tag{12.38}$$

Die Wahl des numerischen Flußes beeinflußt entscheidend die Qualität des numerischen Verfahrens. Wir führen das am folgenden Beispiel vor:

Beispiel 12.51

$$u' + u = g \quad \text{auf } (0,1), \qquad u(0) = 0.$$

Hier ist $g(x) = 1 + x$, so daß die exakte Lsg $u(x) = x$. Sei $x_i = ih$, $i = 0,\ldots,N$ und $K_i = (x_{i-1}, x_i)$. Für die Wahl $p = 0$ (d.h. $S^{0,0}(\mathcal{T})$ besteht aus den stückweise konstanten Funktionen) schreiben wir für $u|_{K_i} = u_i$ und die Testfunktionen bestehen ebenfalls aus stückweise konstanten Funktionen, für die $v|_{K_i} = v_i$ schreiben:

$$\begin{aligned} B_{DG}^{Trans}(u,v) &= \sum_{i=1}^N \int_{K_i} (u' + u)v + (\widehat{u}(x_i)v_i - \widehat{u}(x_{i-1})v(x_i)) \\ &= \sum_{i=1}^N \int_{K_i} uv + (\widehat{u}(x_i) - \widehat{u}(x_{i-1}))v_i \\ &= \sum_{i=1}^N \left[hu_i + (\widehat{u}(x_i) - \widehat{u}(x_{i-1}))\right] v_i \end{aligned}$$

woraus sich als LGS ergibt:

$$\int_{K_i} g\,dx = hu_i + \widehat{u}(x_i) - \widehat{u}(x_{i-1}), \qquad i = 1,\ldots,N.$$

Wir betrachten 2 Wahlen von $\widehat{u}(x_i)$:

- *upwind flux:* $\widehat{u}(x_j) = u_j$ für $1 \le j \le N$ und $\widehat{u}(x_0) = 0$ (und $\widehat{u}(x_N) = u_N$)
- *central flux:* $\widehat{u}(x_j) = \frac{1}{2}(u_j + u_{j+1})$ für $1 \le j \le N-1$ und $\widehat{u}(x_0) = 0$ und $\widehat{u}(x_N) = u_N$.

Wir erkennen in Fig. 12.8, daß die Wahl des *upwind flux* gute Approximationen und sogar die erhoffte Konvergenz $O(h)$ liefert, während die Wahl des *central flux* zu keiner Konvergenz führt. ■

Entscheidend für die Wahl des numerischen Flußes $\widehat{u}$ ist, daß man ihn abhängig vom Vorzeichen von $\mathbf{b}\cdot n_K$ in (12.38) wählt. Wie wir unten sehen werden, führt die richtige Wahl des numerischen Flußes auf ein Verfahren mit guten Stabilitätseigenschaften.

Der *upwind flux* $(\mathbf{b}\cdot n_K)\widehat{u}$ ist definiert durch folgende Wahl von $\widehat{u}$ auf jeder Kante:

- Sei e eine *innere* Kante von $\mathcal{T}$, welche sich K und K' teilen. Für $x \in e$ definieren wir:

$$\begin{aligned} \widehat{u}(x) &= \text{egal} && \text{falls } \mathbf{b}(x)\cdot n_K(x) = \mathbf{b}(x)\cdot n_{K'}(x) = 0 \\ \widehat{u}(x) &= u|_K(x) && \text{falls } \mathbf{b}(x)\cdot n_K(x) > 0 \\ \widehat{u}(x) &= u|_{K'}(x) && \text{falls } \mathbf{b}(x)\cdot n_{K'}(x) > 0. \end{aligned}$$

- Sei e ein Kante auf Γ^-: dann definieren wir $\widehat{u}|_e = 0$

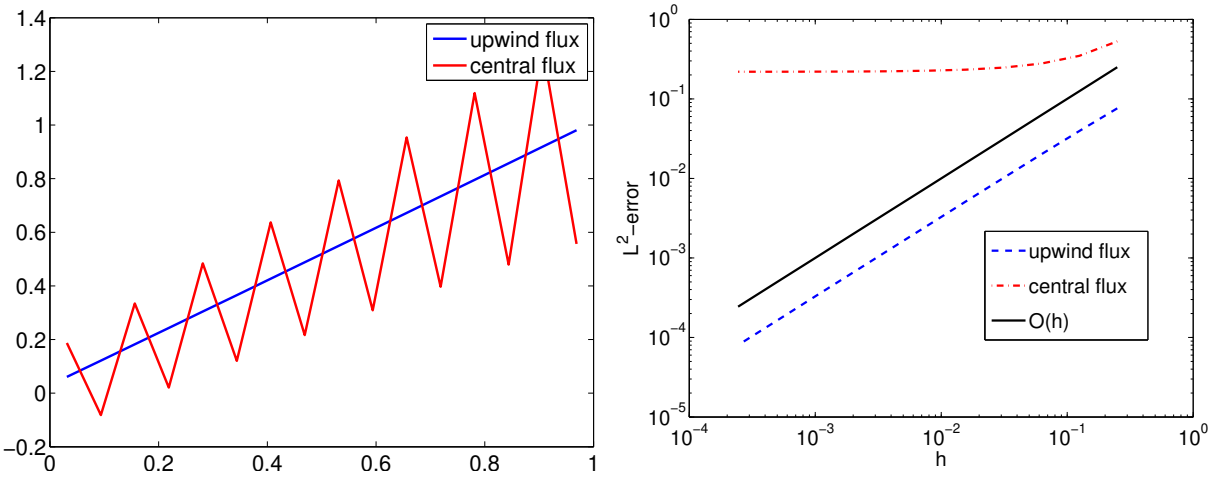


Abbildung 12.8: Links: Lösungsplot. Rechts: Konvergenzverhalten

- Sei e ein Kante auf $\partial\Omega \setminus \Gamma^-$: dann definieren wir $\widehat{u}|_e$ als Limes von u vom angrenzenden Element

Dieses Verfahren hat gut Stabilitätseigenschaften, wie wir nun zeigen. Hierzu benötigen wir den *Sprung* $[\![u]\!]$ auf einer Kanten $e = K|K'$:

$$\begin{aligned} [\![u]\!]|_e &= u|_K n_K + u|_K n_{K'} \qquad \text{falls } e = K|K' \text{ eine innere Kante ist} \\ [\![u]\!]|_e &= u|_K n_K \qquad \text{falls } e \text{ eine Randkante ist mit } e \subset \partial K. \end{aligned}$$

Satz 12.52 *Es gelte*

$$c - \frac{1}{2}\nabla \cdot \mathbf{b} \geq c_0 > 0 \qquad \textit{auf } \overline{\Omega}.$$

Mit dem Sprung $[\![\cdot]\!]$ gilt für die Wahl des "upwind flux" und stückweise glatte Funktionen u:

$$B_{DG}^{Trans}(u,u) \geq \sum_{K\in\mathcal{T}} \|\sqrt{c_0}u\|^2_{L^2(K)} + \sum_{e\in\mathcal{E}} \frac{1}{2}\||\mathbf{b}\cdot n_K|^{1/2}[\![u]\!]\|^2_{L^2(e)},$$

wobei $\mathcal{E}$ die Menge aller Kanten von $\mathcal{T}$ ist. Für Randkanten ist der Sprung einfach definiert als der Wert der Spur.

Beweis: Für glatte Funktionen u ist $u\nabla u = \nabla(\frac{1}{2}u^2)$. Damit gilt für jedes Element $K \in \mathcal{T}$

$$\int_K u(c + \mathbf{b}\cdot\nabla u) = \int_K u^2\left(c - \frac{1}{2}\nabla\cdot\mathbf{b}\right) + \int_{\partial K} \frac{1}{2}\mathbf{b}\cdot n_K u^2$$

Damit

$$B_{DG}^{Trans}(u,u) \quad = \quad \sum_{K\in\mathcal{T}} \int_K u^2 \underbrace{\left(c - \frac{1}{2}\nabla\cdot\mathbf{b}\right)}_{\geq c_0} + \int_{\partial K} \mathbf{b}\cdot n_K \left[\widehat{u}u - \frac{1}{2}u^2\right].$$

Die Summe $\sum_{K\in\mathcal{T}}\int_{\partial K}$ schreiben wir als Summe über Kanten. Dabei:

- Sei e eine *innere* Kante, die sich Elemente K und K' teilen. Sei $x \in e$. Sei oBdA K das Element mit $\mathbf{b}(x)\cdot n_K(x) > 0$ (der Fall $b(x)\cdot n_K(x) = 0$ ist nicht interessant). Dann rechnen wir wegen $n_K = -n_{K'}$ und der Wahl von $\widehat{u}(x)$:

$$\begin{aligned} & \mathbf{b}(x)\cdot n_K(x)\left[\widehat{u}(x)u_K(x) - \frac{1}{2}u_K(x)^2\right] + \mathbf{b}(x)\cdot n_{K'}(x)\left[\widehat{u}(x)u_{K'}(x) - \frac{1}{2}u_{K'}(x)^2\right] \\ = \; & \mathbf{b}(x)\cdot n_K(x)\left[u_K(x)^2 - \frac{1}{2}u_K(x)^2 - u_K(x)u_{K'}(x) + \frac{1}{2}u_{K'}(x)^2\right] \\ = \; & \mathbf{b}(x)\cdot n_K(x)\frac{1}{2}|u_K(x) - u_{K'}(x)|^2. \end{aligned}$$

Diese Rechnung ist auch für den Fall $\mathbf{b}(x) \cdot n_K(x) = 0$ richtig.

- Falls e eine Randkante mit $e \subset \Gamma^-$ ist, dann ist $\widehat{u} = 0$ auf e und somit

$$b \cdot n_K \left[\widehat{u} - \frac{1}{2}u\right] u = -\mathbf{b} \cdot n_K \frac{1}{2}u^2 = \frac{1}{2}|\mathbf{b} \cdot n_K|u^2 = \frac{1}{2}|\mathbf{b} \cdot n_K||[\![u]\!]|^2,$$

  wobei wir im letzten Schritt ausgenutzt haben, wie der Sprung auf Randkanten definiert ist.

- Falls e eine Randkante mit $e \subset \partial\Omega \setminus \Gamma^-$ ist, dann ist $\mathbf{b} \cdot n_K \geq 0$ und $\widehat{u} = u$. Somit

$$b \cdot n_K \left[\widehat{u} - \frac{1}{2}u\right] u = \frac{1}{2}b \cdot n_K u^2 = \frac{1}{2}b \cdot n_K|[\![u]\!]|^2,$$

  wobei wir wieder ausgenutzt haben, wie der Sprung auf Randkanten definiert ist.

Faßt man alle Kantenbeiträge zusammen, dann hat man

$$\sum_{K \in \mathcal{T}} \int_{\partial K} \mathbf{b} \cdot n_K \left(\widehat{u} - \frac{1}{2}u\right) u = \frac{1}{2} \sum_{e \in \mathcal{T}} \||\mathbf{b} \cdot n_K|^{1/2}[\![u]\!]\|^2_{L^2(e)},$$

wobei wir leicht schlampig mit n_K einen Normalenvektor auf e bezeichnen. □

Satz 12.52 zeigt, daß die Bilinearform B koerziv ist. Damit ist insbesondere eindeutige Lösbarkeit des diskreten Verfahrens gegeben.

Die Herleitung der Variationsformulierung zeigt außerdem, daß das Verfahren konsistent ist im folgenden Sinn: Falls u eine Lösung von (9.7) ist und zudem die Regularitätsbedingung $u \in H^1(\Omega)$ gilt, dann ist

$$B_{DG}^{Trans}(u, v) = l(v) \qquad \forall v \in S^{p,0}(\mathcal{T}). \tag{12.39}$$

Damit ergibt sich die Galerkinorthogonalität

$$B_{DG}^{Trans}(u - u_N, v) = 0 \qquad \forall v \in S^{p,0}(\mathcal{T}). \tag{12.40}$$

12.6 DG und Finite Volumenmethoden im Ort—RK in der Zeit

Das numerische Verfahren in Abschnitt 12.5 hat ausgenutzt, daß die Zeitvariable eigentlich keine herausgehobene Rolle hat und deshalb eine Diskretisierung im Raum-Zeit-Zylinder durchgeführt. Wie bei parabolischen Verfahren sind jedoch in der Praxis Zeitschrittverfahren vorherrschend. Wir werden mit k den Zeitschritt bezeichnen.

Die Behandlung von Randbedingungen ist ein eigenes Thema bei hyperbolischen Problemen. Wir behandeln hier den einfachsten Fall eines reinen Anfangswertproblems, d.h. $\Omega = \mathbb{R}^d$. Die Anfangsbedinung u_0 wird mit kompaktem Träger vorausgesetzt.

$$u_t + \nabla \cdot \mathbf{f}(u) = 0 \quad \text{auf } \Omega \times \mathbb{R}^+, \qquad u(\cdot, 0) = u_0 \tag{12.41}$$

Wir gehen von einer Triangulierung $\mathcal{T}$ von Ω aus. Für Testfunktionen $v \in S^{p,0}(\mathcal{T})$ ergibt sich nach partieller Integration (und Vertauschen von Integration und $\frac{d}{dt}$)

$$\sum_K \frac{d}{dt} \int_K uv\, dt - \int_K \nabla v \cdot \mathbf{f}(u) + \int_{\partial K} n_K \cdot \mathbf{f}(u) v = 0.$$

Weil die Testfunktionen unstetig sind, muß ein Kopplung über benachbarte Elemente erfolgen. Dies erfolgt mit dem zu wählenden *numerischen Fluß* $\widehat{h} = \widehat{h}(u, v, n)$. Bezeichnet man mit $\mathcal{N}(K)$ die Nachbarelemente von K, so ergibt sich als numerisches Verfahren: Finde $u \in S^{p,0}(\mathcal{T})$, so daß

$$\sum_K \frac{d}{dt} \int_K uv - \int_K \nabla v \cdot \mathbf{f}(u) + \sum_{K' \in \mathcal{N}(K)} \int_{K|K'} \widehat{h}(u_K, u_{K'}, n_K) v = 0.$$

Hier ist wie oben $K|K'$ eine Kurznotation für den Schnitt $\overline{K} \cap \overline{K'}$. Wir nennen ihn Kante, was aus dem 2D-Fall im Ort motiviert ist.

Definition 12.53 (numerischer Fluß) *Sei* $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$. *Eine Funktion* $\widehat{h} : \mathbb{R} \times \mathbb{R} \times \mathbb{S}^{d-1} \to \mathbb{R}$ *heißt numerischer Fluß, wenn sie lokal lipschitz-stetig stetig ist. Der numerische Fluß heißt*

- konsistent, *wenn* $\widehat{h}(u, u, \mathbf{n}) = \mathbf{f}(u) \cdot \mathbf{n}$ *für alle* u.
- konservativ, *falls* $\widehat{h}(u, v, \mathbf{n}) = -\widehat{h}(v, u, -\mathbf{n})$.
- monoton, *falls* $\widehat{h}$ *monoton wachsend im ersten Argument und monoton fallend im zweiten Argument ist:* $\widehat{h}(\uparrow, \downarrow, \mathbf{n})$

Beispiel 12.54 Der Fall der Advektionsgleichung entspricht $f(u) = \mathbf{b}u$ für konstantes $\mathbf{b}$. Upwinding entspricht der folgenden Wahl des numerisches Flußes auf einer gemeinsamen Kante $K|K'$ von zwei (Orts-)Elementen K, K':

$$\widehat{h}(u_K, u_{K'}, n_K) = \begin{cases} \mathbf{b} \cdot n_K u_K & \text{falls } \mathbf{b} \cdot n_K > 0 \\ \mathbf{b} \cdot n_K u_{K'} & \text{falls } \mathbf{b} \cdot n_K < 0. \end{cases}$$

Man sieht, daß der numerische Fluß konservativ (Übung!) und natürlich auch konsistent ist.

Man kann diese Wahl des Flußes auch ohne Fallunterscheidungen schreiben:

$$\widehat{h}(u_K, u_{K'}, n_K) = \frac{1}{2}\mathbf{b} \cdot n_K \left(u_{K'} + u_K\right) - \frac{1}{2}\left|\mathbf{b} \cdot n_K\right| \left(u_{K'} - u_K\right)$$

■

Bemerkung 12.55 Die Bedingung $\widehat{h}(u, v, n) = -\widehat{h}(v, u, -n)$ drückt eine Erhaltungseigenschaft aus: Für die Testfunktion $v \equiv 1$ ergibt sich:

$$\frac{d}{dt}\int_\Omega u = \sum_K \frac{d}{dt}\int_K u = -\sum_K \sum_{K' \in \mathcal{N}(K)} \int_{K|K'} \widehat{h}(u_K, u_{K'}, n_K).$$

Schreibt man dies als Summe über alle Kanten, so ergibt sich mit der Beobachtung, daß jede Kante e von 2 Elementen K_e, K'_e geteilt wird:

$$\begin{aligned} \frac{d}{dt}\int_\Omega u &= -\sum_K \int_{K|K'} \widehat{h}(u_K, u_{K'}, n_K) = -\sum_e \int_e \widehat{h}(u_{K_e}, u_{K'_e}, n_{K_e}) + \int_e \widehat{h}(u_{K'_e}, u_{K_e}, n_{K'_e}) \\ &= -\sum_e \int_e \widehat{h}(u_{K_e}, u_{K'_e}, n_{K_e}) - \widehat{h}(u_{K_e}, u_{K'_e}, n_{K_e}) = 0; \end{aligned}$$

hier haben wir die Eigenschaft der Konservativität ausgenutzt. ■

Ein vollständiges Verfahren ergibt sich nur die Wahl eines Zeitintegrators. Im einfachsten Fall wird man ein explizites Eulerverfahren wählen. Im Fall des expliziten Eulerverfahrens gilt immer noch die (globale) Erhaltungseigenschaft aus Bemerkung 12.55:

Übung 12.56 Formulieren Sie das explizite Eulerverfahren. Bezeichne mit u^n die numerische Approximation zum Zeitpunkt t_n. Zeigen Sie:

$$\int_\Omega u^{n+1} = \int_\Omega u^n \qquad \forall n \in \mathbb{N}_0.$$

■

strong stability preserving methods

Im Prinzip gibt es viel Auswahl bei den Zeitdiskretisierungen. Meist will man jedoch zusätzliche Eigenschaften erhalten. Bei vielen hyperbolischen Erhaltungsgleichungen gelten gewisse Monotonieeigenschaften auf der kontinuierlichen Ebene, die dann ins Diskrete vererbt werden sollen. Bei skalaren Gleichungen z.B. könnte $\|u(\cdot,t)\|_{L^\infty} \leq \|u_0(\cdot)\|_{L^\infty}$ gelten. Das wesentliche Vorgehen kann man bereits auf dem ODE-Level verstehen.

Betrachte das ODE-System

$$\mathbf{y}' = \mathbf{g}(\mathbf{y}).$$

Wir nehmen an, daß das einfachste RK-Verfahren, das explizite Eulerverfahren, in einer Norm $\|\cdot\|$ die Kontraktivitätsbedingung

$$\|\mathbf{y}^n + k\mathbf{g}(\mathbf{y}^n)\| \leq \|\mathbf{y}^n\|$$

erfüllt. Dann sagen wir, daß ein (anderes) RK-Verfahren die SSP-Eigenschaft[11] hat, falls für es ebenfalls die Eigenschaft $\|\mathbf{y}^{n+1}\| \leq \|\mathbf{y}^n\|$ gilt. Das folgende Lemma 12.58 gibt hinreichende Kriterien an, unter denen ein explizites RK-Verfahren SSP ist. Zuvor eine Umformulierung des klassischen Vorgehens:

Übung 12.57 Sei ein s-stufiges explizites RK-Verfahren beschrieben durch das Butcher-Tableau $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ Dann gilt für das RK-Verfahren angewandt auf $y' = g(y)$:

- Das RK-Verfahren wie folgt definiert mit Hilfe der (Zwischen-)Stufen Y_i, $i = 1, \ldots, s$:

$$\begin{aligned} Y_i &= y_0 + k\sum_{j=1}^{i-1} A_{i,j} g(Y_j), \qquad i = 1, \ldots, s \\ y_1 &= y_0 + k\sum_{j=1}^{s} b_j g(Y_j) \end{aligned}$$

- Erweitert man die Matrix A zu einer Matrix in $\mathbb{R}^{(s+1)\times s}$, indem $A_{s+1,:} := b^\top$ gesetzt wird, dann hat das Verfahren die Form

$$\begin{aligned} Y_1 &= y_0 \\ Y_{i+1} &= y_0 + k\sum_{j=1}^{i} A_{i+1,j} g(Y_j), \qquad i = 1, \ldots, s \\ y_1 &= Y_{s+1} \end{aligned}$$

- Das RK-Verfahren kann in der folgenden Form geschrieben werden mit Koeffizienten $\alpha_{i,j} \geq 0$, die zudem die folgende Bedingung erfüllen: $\alpha_{i,j} = 0$ impliziert $\beta_{i,j} = 0$.

$$\begin{aligned} Y_1 &= y_0 \\ Y_{i+1} &= \sum_{j=1}^{i} \alpha_{i,j} Y_j + k\sum_{j=1}^{i} \beta_{i,j} g(Y_j), \qquad i = 1, \ldots, s \\ y_1 &= Y_{s+1} \end{aligned}$$

  (Die Darstellung ist nicht eindeutig). Zeigen Sie, daß die Konsistenz des Verfahrens die Bedingung $\sum_{j=1}^{i} \alpha_{i,j} = 1$ (für jedes i) erwingt.

■

[11] strong stability preserving

Lemma 12.58 *Sei ein explizites s-stufiges RK-Verfahren gegeben. Möge das explizite Eulerverfahren stabil sein unter der "CFL-Bedingung" $k \leq k_{expEuler}$. Mögen die Koeffizienten $\beta_{i,j}$ in der Darstellung aus Übung 12.57 die Bedingung $\beta_{i,j} \geq 0$ erfüllen. Dann gilt: das RK-Verfahren ist SSP, falls die Schrittweite k die Bedingung*

$$k \leq k_{expEuler} \min_{i,j} \frac{\alpha_{i,j}}{\beta_{i,j}}$$

erfüllt.

Beweis: Die Idee ist, daß das Verfahren eine Konvexkombination aus expliziten Eulerschritten ist. Der Einfachheit nehmen wir an, daß alle $\alpha_{i,j} > 0$. Dann ist

$$Y_{i+1} = \sum_{j=1}^{i} \alpha_{i,j} \left(Y_j + \frac{\beta_{i,j}}{\alpha_{i,j}} kg(Y_j) \right)$$

und damit

$$\|Y_{i+1}\| \leq \sum_{j=1}^{i} \alpha_{i,j} \|Y_j + \frac{\beta_{i,j}}{\alpha_{i,j}} kg(Y_j)\| \leq \sum_{j=1}^{i} \alpha_{i,j} \|Y_j\|,$$

wobei wir im letzten Schritt die Stabilität des expliziten Eulerverfahrens und die stringentere Schrittweitenbedingung verwendet haben. Aus der Konsistenz $\sum_j \alpha_{i,j} = 1$ folgt

$$\|Y_{i+1}\| \leq \max_{j=1,\ldots,i} \|Y_j\|.$$

Induktiv folgt damit $\|Y_{i+1}\| \leq \|Y_1\| = \|y_0\|$ für jedes i und damit $\|y_1\| = \|Y_{s+1}\| \leq \|y_0\|$. □

Übung 12.59

- Zeigen Sie, daß die explizite Trapezregel SSP ist:

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

- Zeigen Sie, daß die explizite Mittelpunktsregel nicht SSP ist:

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

■

Bemerkung 12.60 Wenn einige der $\beta_{i,j}$ negativ sind, dann kann man u.U. immer noch stabile Verfahren erzeugen, wenn geeignet das explizite Eulerverfahren durch das implizite Eulerverfahren ersetzt wird, [10].

■

12.7 klassische finite Volumenverfahren

Historisch älter als das Vorgehen in Abschnitt 12.6 sind die klassischen finiten Volumenverfahren. Bei diesen Verfahren wird, ausgehend von einer Triangulierung $\mathcal{T}$ von Ω, mit stückweise konstanten Funktionen auf den Kontrollvolumina $K \times (t_n, t_{n+1})$, $K \in \mathcal{T}$ gearbeitet. Wir werden im Folgenden den räumlich eindimensionalen Fall auf regelmäßigen Gittern mit Gitterweite h betrachten. Wir bezeichnen die Elemente mit

$$K_i = (x_{i-1/2}, x_{i+1/2}),$$

d.h. wir denken uns x_i als den Mittelpunkt von K_i.

1. konzeptionell lassen sich die Verfahren auch auf unstrukturierte Gitter in mehreren Ortsdimensionen erweitern—entscheidend ist bei diesen Verfahren immer noch die Definition des sog. numerischen Flußes

2. Falls regelmäßige Gitter in höheren Dimensionen verwendet werden, ist ein ganz häufig verwendetes Mittel das der Splittingtechniken. Betrachtet man z.B.

$$u_t + \partial_x(f_1(u)) + \partial_y(f_2(u)) = 0,$$

so kann man wie oben beschrieben eine Splittingtechnik basierend auf $u_t + \partial_x(f_1(u)) = 0$ und $u_t + \partial_y(f_2(u)) = 0$ verwenden.

Bemerkung 12.61 Die Verwendung von regelmäßigen Gittern suggeriert, daß man auch Differenzenverfahren verwenden könnte. Differenzenverfahren interpretieren die Werte als Funktionswerte und die Ordnung/Konvergenz des Verfahrens ergibt sich aus der Glattheit der Lösung. Für viele (auch hyperbolische) Probleme ist das angemessen. Die hier untersuchten Probleme können jedoch nicht glatte und unstetige Lösungen ("Schocks") haben. Finite Volumenverfahren stellen deshalb Verfahren auf, bei denen die Interpretation der Unbekannten die von Elementmittelwerten ist. ∎

Testet man (12.31) mit der charakteristischen Funktion von $K_i \times (t_n, t_{n+1})$, so ergibt sich:

$$\int_{K_i} u(x, t_{n+1})\, dx - \int_{K_i} u(x, t_n)\, dx = \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2}, t))\, dt - \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t))\, dt$$

Mit den Zellmitteln

$$\overline{u}_i^n := \int_{K_i} u(x, t_n)\, dx$$

und den exakten Flüßen

$$\overline{F}_{i+1/2}^n := \frac{1}{k} \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t))\, dt$$

ergibt sich damit

$$\overline{u}_i^{n+1} = \overline{u}_i^n - \frac{h}{k} \left(\overline{F}_{i+1/2}^n - \overline{F}_{i-1/2}^n \right) \tag{12.42}$$

Bemerkung 12.62 (12.42) drückt eine exakte Erhaltung auf dem Kontrollvolumen $K_i \times (t_n, t_{n+1})$ aus. ∎

(12.42) ist noch kein numerisches Verfahren, da die Formel $\overline{F}_{i+1/2}^n$ (eine Mittelung in der Zeit des exakten Flußes im Punkt $x_{i+1/2}$) nicht bekannt ist. Finite Volumenverfahren werden aber immer in der Form (12.42) geschrieben, wobei der exakte Fluß $\overline{F}_{i+1/2}^n$ durch einen numerischen Fluß $F_{i+1/2}^n$ ersetzt wird. Das Verfahren ist dann

$$\overline{u}_i^{n+1} = \overline{u}_i^n - \frac{k}{h} \left(F_{i+1/2}^n - F_{i-1/2}^n \right) \tag{12.43}$$

Typischerweise ist $F_{i+1/2}^n$ von der Form

$$F_{i+1/2}^n = \widehat{h}(\overline{u}_i^n, \overline{u}_{i+1}^n) \tag{12.44}$$

für eine numerische Flußfunktion $\widehat{h}$. Wir nennen Verfahren der Form (12.43) *konservativ*. Falls $\widehat{h}(u, u) = f(u)$, dann heißt das Verfahren *konsistent*.

Übung 12.63 Im vorangegangen Kapitel hatte der numerische Fluß $\widehat{h}$ auch noch das Argument n. Überlegen Sie sich, daß die Sprechweise trotzdem konsistent mit der Notation aus dem vorangegangenen Kapitel ist, da im vorliegenden 1D-Fall n nur die beiden Werte ± 1 annehmen kann. Überlegen Sie sich, daß dann auch die Begriffe "konservativ" und "konsistent" zur vorangegangenen Begriffsbildung passen. ∎

Bemerkung 12.64 Man kann den numerischen Fluß auch als Funktion von mehreren benachbarten Zellmitteln definieren. Die Notation wäre dann $F_{i+1/2}^n = \widehat{h}(\overline{u}_{i-p}^n, \overline{u}_{i-p+1}^n, \ldots, \overline{u}_{i+q}^n)$ für feste p, q. Konsistenz bedeutet dann $\widehat{h}(u, u, \ldots, u) = f(u)$. Solche Schemata zu betrachten macht durchaus Sinn → *reconstruction/higher order methods*. ∎

Beispiel 12.65 Das Lax-Friedrichs-Schema ist

$$\overline{u}_i^{n+1} = \frac{1}{2} \left(\overline{u}_{i-1}^n + \overline{u}_{i+1}^n \right) - \frac{k}{2h} \left(f(\overline{u}_{i+1}^n) - f(\overline{u}_{i-1}^n) \right).$$

Dieses kann in die obige Form ("konservative Form") gebracht werden mit den numerische Flüßen

$$F_{i+1/2}^n := \widehat{h}(\overline{u}_i^n, \overline{u}_{i+1}^n) := \frac{h}{2k} \left(\overline{u}_i^n - \overline{u}_{i+1}^n \right) + \frac{1}{2} \left(f(\overline{u}_i^n) + f(\overline{u}_{i+1}^n) \right)$$

∎

12.7.1 Godunov-Verfahren

Beim Goduonvverfahren wird der numerische Fluß mittels einer (lokalen) Lösung eines sog. Riemannproblems definiert.

Das Riemannproblem

Gegeben u_l und $u_r \in \mathbb{R}$, finde die Lösung u des Anfangswertproblems

$$u_t + (f(u))_x = 0 \quad \text{auf } \mathbb{R} \times \mathbb{R}^+, \qquad u(\cdot, 0) = \begin{cases} u_l & x < 0 \\ u_r & x > 0. \end{cases} \tag{12.45}$$

Die physikalisch relevante Lösung ("Entropielösung") kann im vorliegenden skalaren Fall explizit angegeben werden. Sei (der Einfachheit halber) $f : \mathbb{R} \to \mathbb{R}$ strikt convex. Dann ist $f'' > 0$. Die folgenden Fälle können eintreten:

1. $u_l = u_r$: dann ist die Lösung (natürlich) $u(x,t) = u_l = u_r$ für alle $t > 0$.
2. $u_l > u_r$: dann ist die Lösung ein sog. *Schock*, der sich mit *Schockgeschwindigkeit*
$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l}$$
ausbreitet. Genauer:
$$u(x,t) = \begin{cases} u_l, & \text{falls } x < st \\ u_r, & \text{falls } x > st \end{cases}$$
3. $u_l < u_r$: dann ist die Lösung ein sog. *Verdünnungsfächer*: Weil $f'' > 0$ ist f' invertierbar. Damit läßt sich die Lösung schreiben als:
$$u(x,t) = \begin{cases} u_l, & \text{falls } x \le f'(u_l)t \\ (f')^{-1}(x/t), & \text{falls } f'(u_l)t \le x \le f'(u_r)t \\ u_r, & \text{falls } x > f'(u_r)t. \end{cases}$$

Bemerkung 12.66

- Die Schocklösung z.B. ist unstetig. Damit ist der Lösungsbegriff für die Erhaltungsgleichung $u_t + f(u)_x = 0$ der einer schwachen Lösung. Man kann relativ einfach nachrechnen, daß die oben angegebene Lösung eine schwache Lösung der Erhaltungsgleichung ist.
- Weil schwache Lösungen nicht eindeutig sind, muß eine (schwache) Lösung ausgewählt werden. Die physikalisch relevante Lösung nennt man "Entropielösung", weil das "richtige" Auswahlprinzip bei den hyperbolischen Erhaltungsgleichungen der Gasdynamik dem physikalischen Prinzip der Entropieerhöhung genügen.
- Die Schockgeschwindigkeit s ergibt sich aus der Forderung der Erhaltung von u, wenn man stückweise konstante (schwache) Lösungen sucht.
- Die Form der Lösung des Verdünnungsfächers ergibt sich aus folgender Überlegung: Weil die Differentialgleichung invariant unter der Substitution $(x,t) \mapsto \lambda \cdot (x,t)$ ist, kann man vermuten, daß die Lösung $u = u(x,t)$ von der Form $u(x,t) = v(x/t)$ ist. Mit diesem Ansatz ergibt sich die angegebene Form. ∎

Die Lösung u^R des Riemannproblems ist i.a. unstetig. Dennoch ist es eine Ähnlichkeitslösung

Lemma 12.67 *Sei u^R Lösung von (12.45). Dann gilt: u^R hat die Form einer Ähnlichkeitslösung $u^R(x,t) = v(x/t)$. Für die Funktion v gilt zusätzlich: $\xi \mapsto f(v(\xi))$ ist stetig bei $\xi = 0$. Insbesondere ist damit für jedes feste $t > 0$ die Funktion $x \mapsto f(u^R(x,t)) = f(v(x/t))$ stetig bei $x = 0$. Somit ist der Ausdruck $f(u^R(0,t))$ sinnvoll definiert (und sogar konstant in t).*

Beweis: Die Lösungen u^R des Riemannproblems sind alle Ähnlichkeitslösungen, d.h. von der Form $u^R(x,t) = v(x/t)$. Wir unterscheiden nun 2 Fälle:

1. $\xi \mapsto v(\xi)$ ist stetig bei Null. Dann ist offensichtlich auch $\xi \mapsto f(v(\xi))$ stetig bei Null (f hinreichend glatt wird angenommen).

2. $\xi \mapsto v(\xi)$ ist nicht stetig bei Null. Dieser Fall kann nur im Falle von Schocks mit Schockgeschwindigkeit $s = 0$ eintreten. Die Lösung u^R ist gegeben durch $u^R = u_0$, und die Funktion v hat die Form
$$v(\xi) = \begin{cases} u_l & \text{falls } \xi < 0 \\ u_r & \text{falls } \xi > 0. \end{cases}$$
Die Sprungbedingung liefert dann
$$0 = s = \frac{f(u_r) - f(u_l)}{u_r - u_l},$$
d.h. $f(u_r) = f(u_l)$. Das bedeutet aber gerade, daß $\xi \mapsto f(v(\xi))$ stetig bei $\xi = 0$ ist.

□

stückweise konstante Anfangsdaten

Für stückweise konstante Anfangsdaten läßt sich die Lösung für *kleine Zeiten* direkt als Zusammensetzung von Lösungen von Riemannproblemen darstellen. Dies ist der Tatsache geschuldet, daß die Ausbreitungsgeschwindigkeit endlich ist. Wir illustrieren das Phänomen mit einem Beispiel, bei dem nur Schocks auftreten.

Beispiel 12.68 (Burgers Gleichung) Sei
$$f(u) = \frac{1}{2}u^2$$
und die Anfangsdaten für gegebene $x_1 < x_2$ und $u_1 > u_2 > u_3$:
$$u_0(x) = \begin{cases} u_1 & \text{falls } x < x_1 \\ u_2 & \text{falls } x_1 < x < x_2 \\ u_3 & \text{falls } x > x_2. \end{cases}$$
Wir nehmen an, daß u_1, u_2, u_3 die Bedingungen
$$s_1 := \frac{1}{2}(u_1 + u_2) > s_2 := \frac{1}{2}(u_2 + u_3)$$
erfüllen. Dann ist die Lösung u wie folgt gegeben:

1. für $0 < t < t^* := (x_2 - x_1)/(s_1 - s_2)$ ist
$$u(x,t) = \begin{cases} u_1 & \text{falls } x - x_1 < s_1 t \\ u_2 & \text{falls } x - x_1 > s_1 t \quad \text{und } x - x_2 < s_2 t \\ u_3 & \text{falls } x - x_2 > s_2 t \end{cases}$$

2. für $t > t^*$ definieren wir
$$s_* := \frac{1}{2}(u_1 + u_3), \qquad x_* := x_2 + t^* s_2 = x_1 + t^* s_1$$
$$u(x,t) = \begin{cases} u_1 & \text{falls } x - x_3 < s_3 t \\ u_3 & \text{falls } x - x_3 > s_3 t. \end{cases}$$

■

Das Beispiel zeigt, daß die Schocks bis t^* unabhängig voneinander sind. Bei t^* interagieren die beiden Schocks und vereinigen sich zu einem einzigen, neuen Schock.

Godunovverfahren

Das Godunovverfahren ist das Verfahren (12.43), wobei der Fluß $F^n_{i+1/2}$ durch

$$F^{G,n}_{i+1/2} := \frac{1}{k}\int_{t_n}^{t_{n+1}} f(u^R(x_{i+1/2},t))\,dt \tag{12.46}$$

gegeben ist und die Funktion u^R die Lösung des Riemannproblems

$$u_t + f(u)_x = 0 \quad \text{auf } \mathbb{R}\times(t_n,\infty), \qquad u(x,t_n) = \begin{cases} \overline{u}^n_i & \text{falls } x < x_{i+1/2} \\ \overline{u}^n_{i+1} & \text{falls } x > x_{i+1/2} \end{cases}$$

Wir bemerken, daß Lemma 12.67 besagt, daß der Integrand $f(u^R(x_{i+1/2},t))$ wohldefiniert ist und sogar konstant in t. Das Integral kann damit explizit bestimmt werden. Tatsächlich gilt:

$$F^{G,n}_{i+1/2} = \widehat{h}(\overline{u}^n_i,\overline{u}^n_{i+1}) := \begin{cases} \min_{\overline{u}^n_i\le v\le\overline{u}^n_{i+1}} f(v) & \text{falls } \overline{u}^n_i \le \overline{u}^n_{i+1} \\ \max_{\overline{u}^n_{i+1}\le v\le\overline{u}^n_i} f(v) & \text{falls } \overline{u}^n_i \ge \overline{u}^n_{i+1}. \end{cases} \tag{12.47}$$

Übung 12.69 Zeigen Sie die Formel (12.47) unter der Annahme, daß $f \in C^2(\mathbb{R};\mathbb{R})$ strikt konvex ist. ■

Bemerkung 12.70 Die Formel (12.47) gilt nicht nur für konvexe Flüße f sondern auch für nichtkonvexe. Hierzu hätten wir jedoch den Begriff der Entropielösung auch für nichtkonvexe Flußfunktionen f einführen müßen... ■

Bemerkung 12.71 Eigentlich will man den Fluß $F^n_{i+1/2}$ durch Lösen des Anfangswertproblems $u_t + f(u)_x = 0$ auf $\mathbb{R}$ mit der stückweise konstanten Anfangsbedingung $\overline{u}^n$ (die ja auf $\mathbb{R}$ definiert ist) lösen. Tatsächlich reicht es wegen der endlichen Ausbreitungsgeschwindigkeit der Lösung, lokale Riemannprobleme zu betrachten: erfüllen die Zeitschrittweite k und die Ortsschrittweite h die CFL-Bedingung

$$\frac{k}{h}\max_i |f'(\overline{u}^n_i)| \le \frac{1}{2},$$

dann interagieren die Wellen, die von zwei benachbarten Zellgrenzen ausgehen nicht bis zum nächsten Zeitlevel, so daß sich $F^n_{i+1/2}$ tatsächlich aus rein lokalen Betrachtungen ergibt. ■

Der Godunovfluß kann als Verallgemeinerung des *upwinding* aufgefaßt werden, wie die folgende Übung zeigt.

Übung 12.72 Betrachten Sie die lineare Flußfunktion $f(u) = au$ für ein $a \in \mathbb{R}$. Zeigen Sie, daß der Godunovfluß auf das folgende Verfahren führt:

$$\frac{\overline{u}^{n+1}_i - \overline{u}^n_i}{k} + \frac{a^+}{h}\left(\overline{u}^n_i - \overline{u}^n_{i-1}\right) + \frac{a^-}{h}\left(\overline{u}^n_{i+1} - \overline{u}^n_i\right) = 0, \qquad a^+ := \max\{a,0\}, \qquad a^- := \min\{a,0\} \qquad ■$$

12.7.2 Approximative Riemannlöser

Die Bestimmung der Extrema in (12.47) kann aufwendig sein (insbesondere für nichtkonvexe f). Bei vektorwertigen Problemen kann die Bestimmung der Lösung des Riemannproblem teuer sein. Man ist also an Vereinfachungen interessiert.

Roe-Löser

Man ersetzt die Flußfunktion f im Riemannlöser durch eine lineare Funktion $\widehat{\mathbf{A}}u$, wobei $\widehat{\mathbf{A}} = \widehat{A}(\overline{u}^n_i,\overline{u}^n_{i+1})$. Folgende Bedingungen scheinen plausibel (wir formulieren sie gleich für den Fall von Systemen):

(i) $\widehat{\mathbf{A}}(u_l,u_r)(u_r - u_l) = \mathbf{f}(u_r) - \mathbf{f}(u_l)$

(ii) $\widehat{\mathbf{A}}(u_l,u_r)$ ist reell diagonalisierbar (m.a.W.: die Linearisierung ist wiederum hyperbolisch)

(iii) $\widehat{\mathbf{A}}(u_l, u_r) \to \mathbf{f}'(v)$ falls u_l, $u_r \to v$. (m.a.W.: Konsistenz)

Bemerkung 12.73 Im skalaren Fall ist die Bedingung (i) relativ natürlich: falls u_l und u_r zwei Zustände sind, die durch einen Schock getrennt sind, dann guarantiert sie, daß die obige Linearisierung die *gleiche* Schockgeschwindigkeit hat wie die Lösung des echten Riemannproblems, denn die Schockgeschwindigkeit ist nach der Rankine-Hugoniot-Bedingung gegeben durch

$$s_{ex} = \frac{f(u_r) - f(u_l)}{u_r - u_l}, \qquad s_{approx} = \frac{\widehat{A}u_r - \widehat{A}u_l}{u_r - u_l}$$

■

Im vorliegenden skalaren Fall ergibt sich damit als (approximatives) Riemannproblem zur Bestimmung von $F^n_{i+1/2}$:

$$u_t + \widehat{A}_{i+1/2} u_x = 0 \qquad \text{auf } \mathbb{R} \times (t_n, \infty), \qquad u(x, t_n) = \begin{cases} \overline{u}^n_i & \text{falls } x < x_{i+1/2} \\ \overline{u}^n_{i+1} & \text{falls } x > x_{i+1/2}, \end{cases}$$

wobei $\widehat{A}_{i+1/2}$ explizit gegeben ist als das sog. *Roe-Mittel*[12]

$$\widehat{A}_{i+1/2} = \begin{cases} \frac{f(\overline{u}^n_{i+1}) - f(\overline{u}^n_i)}{\overline{u}^n_{i+1} - \overline{u}^n_i} & \text{falls } \overline{u}^n_{i+1} \neq \overline{u}^n_i \\ f'(\overline{u}^n_i) & \text{falls } \overline{u}^n_i = \overline{u}^n_{i+1} \end{cases} \tag{12.48}$$

Der zugehörige numerische Fluß ist damit gegeben durch

$$F^n_{i+1/2} = F^{Roe}(\overline{u}^n_i, \overline{u}^n_{i+1}) = \begin{cases} f(\overline{u}^n_i) & \text{falls } \widehat{A}_{i+1/2} \geq 0 \\ f(\overline{u}^n_{i+1}) & \text{falls } \widehat{A}_{i+1/2} < 0. \end{cases} \tag{12.49}$$

12.7.3 verbesserte Roe-Löser

Im vorliegenden skalaren Fall ist die Lösung des approximativen Riemannlösers nur eine Schocklösung, die je nach Vorzeichen von $\widehat{A}_{i+1/2}$ nach links oder recht läuft. Falls das exakte Riemannproblem einen Verdünnungsfächer hat, dann wird er nur schlecht von einem Roe-Löser approximiert, denn Information läuft bei Schocks entweder nur nach rechts oder nach links (je nach Vorzeichen von $\widehat{A}_{i+1/2}$) während bei einem Verdünnungsfächer die Information in beide Richungen laufen kann (nämlich dann, wenn $f'(u_l)$ und $f'(u_r)$ unterschiedliches Vorzeichen haben[13]). Diese Problematik kann man dadurch entschärfen, daß man vorgibt, daß die Lösung des Roe-Lösers komplexer ist.

Wir betrachten das Riemannproblem

$$u_t + f(u)_x = 0, \qquad u(x, 0) = \begin{cases} u_l & \text{falls } x < 0 \\ u_r & \text{falls } x > 0. \end{cases}$$

Wir approximieren die exakte Lösung u durch *zwei* Schocks mit Schockgeschwindigkeiten $s^l < 0 < s^r$, die wir noch geeignet wählen werden:

$$U(x, t) = \begin{cases} u_l & \text{falls } x < s^l t \\ u_* & \text{falls } s^l t < x < s^r t \\ u_r & \text{falls } x > s^r t. \end{cases} \tag{12.50}$$

Der Zwischenzustand u_* stellt eine Approximation an den Verdünnungsfächer dar. Er ist bereits durch die Rankine-Hugoniot-Bedingung und die (von uns gewählten) Schockgeschwindigkeiten festgelegt:

$$\begin{aligned} f(u_r) - f(u_*) &= s^r(u_r - u_*) \\ f(u_*) - f(u_l) &= s^l(u_* - u_l). \end{aligned}$$

[12] Phil Roe

[13] man spricht von einer *transonic rarefaction wave*

Auflösen nach $f(u_*)$ liefert

$$f(u_*) = \frac{s^r f(u_l) - s^l f(u_r) + s^r s^l (u_r - u_l)}{s^r - s^l}.$$

Mit $u_l = \overline{u}_i^n$, $u_r = \overline{u}_{i+1}^n$ und $s^l = s_{i+1/2}^l$, $s^r = s_{i+1/2}^r$ ergibt sich dann für den numerischen Fluß $F_{i+1/2}^n$ (beachte: weil $s^l < 0 < s^r$ angenommen ist, ist der gesuchte numerische Fluß damit $F = f(u_*)$!)

$$F_{i+1/2}^n = \widetilde{F}(\overline{u}_i^n, \overline{u}_{i+1}^n) = \frac{s_{i+1/2}^r f(\overline{u}_i^n) - s_{i+1/2}^l f(\overline{u}_{i+1}^n) + s_{i+1/2}^r s_{i+1/2}^l (\overline{u}_{i+1}^n - \overline{u}_i^n)}{s_{i+1/2}^r - s_{i+1/2}^l}. \tag{12.51}$$

Wir betrachten nun den Spezialfall $s_{i+1/2}^l = -s_{i+1/2}^r = -s_{i+1/2}$. Dann vereinfacht sich (12.51) wie folgt:

$$F_{i+1/2}^n = \frac{f(\overline{u}_i^n) + f(\overline{u}_{i+1}^n)}{2} - \frac{s_{i+1/2}}{2} \left(\overline{u}_{i+1}^n - \overline{u}_i^n\right). \tag{12.52}$$

Lax-Friedrichs

Die maximale Geschwindigkeit $s^l = -s^r$, die wir sinnvollerweise zulassen, ist

$$s^r = \frac{h}{k} = -s^l = -\frac{h}{k};$$

dann interagieren die Riemannprobleme von benachbarten Zellgrenzen gerade noch nicht. Mit dieser Wahl von s_r ergibt sich in (12.52)

$$F_{i+1/12}^n = F^{LF}(\overline{u}_i^n, \overline{u}_{i+1}^n) = \frac{f(\overline{u}_i^n) + f(\overline{u}_{i+1}^n)}{2} - \frac{h}{2k} \left(\overline{u}_{i+1}^n - \overline{u}_i^n\right).$$

Dieses Verfahren hatten wir bereits kennengelernt.

Rusanov-Verfahren ("lokales Lax-Friedrichs-Verfahren")

Die Wahl $s^r = \frac{h}{k}$ ist nicht physikalisch motiviert. Plausibler erscheint es, die Schockgeschwindigkeit s^r an das lokalen Verhalten des Flußes f zu koppeln. Diese Betrachtung motiviert die Wahl

$$s_{i+1/2}^l = -s_{i+1/2}^r, \qquad s_{i+1/2}^r = \max\{|f'(\overline{u}_i^n)|, |f'(\overline{u}_{i+1}^n)|\}$$

Das Verfahren ist dann

$$F_{i+1/2}^n = F^{Rus}(\overline{u}_i^n, \overline{u}_{i+1}^n) = \frac{f(\overline{u}_i^n) + f(\overline{u}_{i+1}^n)}{2} - \frac{\max\{|f'(\overline{u}_i^n)|, |f'(\overline{u}_{i+1}^n)|\}}{2} \left(\overline{u}_{i+1}^n - \overline{u}_i^n\right).$$

Engquist-Osher

$$F_{i+1/2}^{EO} = \frac{f(\overline{u}_i^n) + f(\overline{u}_{i+1}^n)}{2} - \frac{1}{2} \int_{\overline{u}_i^n}^{\overline{u}_{i+1}^n} |f'(u)|\, du \tag{12.53}$$

Der Engquist-Osher-Fluß ist ein Beispiel für ein *Fluß-Splitting*-Verfahren—-siehe Bemerkung 12.86.

higher order schemes: Reconstruction, Evolution, Averaging (REA)

Alle bisher vorgestellten Verfahren, insbesondere das Godunovverfahren, sind von der Ordnung 1 in Ort und Zeit. Beschränkt man sich auf Verfahren der Ordnung 1, so kann man viele starke Eigenschaften des kontinuierlichen Problems ins Diskrete vererben wie Monotonieeigenschaften. Obwohl viele dieser Eigenschaften aufgegeben werden müssen, wenn man an Verfahren höherer Ordnung interessiert ist, sind Verfahren höherer Ordnung (sagen wir: Ordnung 2) die praktisch wichtigsten Verfahren.

Sehr viele Algorithmen, die auf finiten Volumenmethoden basieren, gehen wie folgt vor:

1. **Reconstruction:** ausgehend von den Zell*mitteln* $\overline{u}_i^n$ zum Zeitpunkt t_n wird eine stückweise Funktion höherer Ordnung $\widehat{u}(x, t_n)$ auf den Zellen $K_i = (x_{i-1/2}, x_{i+1/2})$ rekonstruiert.

2. **Evolution:** Die rekonstruierte Funktion $\widehat{u}(x,t_n)$ wird mittels eines exakten oder approximativen Lösers zum Zeitpunkt t_{n+1} evolviert.

3. Die durch die Evolution erhaltene Funktion $\widehat{u}(x,t_{n+1})$ wird wieder durch eine stückweise konstante Funktion approximiert: $\overline{u}_i^{n+1} := \frac{1}{|K_i|}\int_{K_i}\widehat{u}(x,t_{n+1})\,dx$.

Frage: wie soll rekonstruiert werden?

Wir betrachten den einfachsten Fall linearer Rekonstruktion $\widehat{u}^n \in S^{1,0}(\mathcal{T})$. Eine sinnvolle Forderung ist

$$\frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} \widehat{u}^n(x)\,dx \overset{!}{=} \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} u^n(x)\,dx = \overline{u}_i^n.$$

Dies fixiert auf Element $K_i = (x_{i-1/2}, x_{i+1/2})$ die Form

$$\widehat{u}^n(x) = \overline{u}_i^n + \sigma_i^n(x - x_i).$$

Naheliegende Wahlen für die Steigung σ_i^n sind:

$$\begin{aligned}
\text{central: } \sigma &= \frac{\overline{u}_{i+1}^n - \overline{u}_{i-1}^n}{2h},\\
\text{forward: } \sigma &= \frac{\overline{u}_{i+1}^n - \overline{u}_i^n}{h},\\
\text{backward: } \sigma &= \frac{\overline{u}_i^n - \overline{u}_{i-1}^n}{h}.
\end{aligned}$$

Um den Einfluß der Wahl zu sehen, betrachten wir

Beispiel 12.74 (REA für Advektionsgleichung $f(u) = au$) Für $a > 0$ ist das Verfahren (Übung!)

$$\begin{aligned}
\overline{u}_i^{n+1} &= a\frac{k}{h}\left(\overline{u}_{i-1}^n + \frac{1}{2}(h - ak)\sigma_{i-1}^n\right) + (1 - a\frac{k}{h})\left(\overline{u}_i^n - \frac{1}{2}ak\sigma_i^n\right)\\
&= \overline{u}_i^n - a\frac{k}{h}\left(\overline{u}_i^n - \overline{u}_{i-1}^n\right) - \frac{1}{2}a\frac{k}{h}(h - ak)(\sigma_i^n - \sigma_{i-1}^n)
\end{aligned}$$

Die Wahl der "forward slope" liefert damit

$$\overline{u}_i^{n+1} = \overline{u}_i^n - a\frac{k}{h}\left(\overline{u}_{i+1}^n - \overline{u}_{i-1}^n\right) - \frac{1}{2}a\frac{k^2}{h^2}\left(\overline{u}_{i+1}^n - 2\overline{u}_i^n + \overline{u}_{i-1}^n\right),$$

was gerade das Lax-Wendroff-Verfahren ist (welches wir auch bereits als formal 2. Ordnung erkannt hatten).

Wir erwähnen, daß die Wahl der "backward slope" auf

$$\overline{u}_i^{n+1} = \overline{u}_i^n - \frac{1}{2}\frac{ak}{h}(3\overline{u}_i^n - 4\overline{u}_{i-1}^n + \overline{u}_{i-2}^n) + \frac{1}{2}a^2\frac{k^2}{h^2}(\overline{u}_i^n - 2\overline{u}_{i-1}^n + \overline{u}_{i-2}^n)$$

führt, was ebenfalls ein klassisches Verfahren ist, das sog. "Beam-Warming-Verfahren". ∎

Die eigentliche Wahl der Steigungen σ_i^n wird typischerweise so motiviert, dass man die Totalvariation nicht vergrößert[14]. Hier ist die Totalvariation einer Funktion u definieren durch[15]

$$|u|_{TV} := \sup\{\sum_i |u(y_{i+1}) - u(y_i)| \mid (y_j)_j \text{ Partition von } \mathbb{R} \text{ mit } y_j \le y_{j+1} \text{ für alle } j\}. \tag{12.54}$$

Die Rekonstruktion $\widehat{u}^n$ der stückweise konstanten Funktion $\overline{u}^n$ soll also erfüllen:

$$|\widehat{u}^n|_{TV} \le |u^n|_{TV}. \tag{12.55}$$

[14] für skalare Probleme ist dies eine Eigenschaft des kontinuierlichen Problems

[15] Es gibt weitere Darstellungen: $|u|_{TV} = \lim_{\varepsilon\to 0}\varepsilon^{-1}\|u(\cdot+\varepsilon) - u(\cdot)\|_{L^1(\mathbb{R})}$ und, falls $u \in W^{1,1}(\mathbb{R})$, auch $\|u'\|_{L^1(\mathbb{R})}$.

Dies erreicht man z.B. mit der Wahl

$$\sigma_i^n := \text{minmod}\left(\frac{\overline{u}_{i+1}^n - \overline{u}_i^n}{h}, \frac{\overline{u}_i^n - \overline{u}_{i-1}^n}{h}\right), \tag{12.56}$$

$$\text{minmod}(a,b) := \frac{1}{2}\left(\text{sign}(a) + \text{sign}(b)\right)\min\{|a|,|b|\} = \begin{cases} \text{sign}(a)\min\{|a|,|b|\} & \text{falls } \text{sign}(a) = \text{sign}(b), \\ 0 & \text{sonst .} \end{cases} \tag{12.57}$$

M.a.W.: man vergleicht "forward slope" und "backward slope". Falls sie das gleiche Vorzeichen haben, dann nimmt man den betragsmäßig kleineren Wert. Ansonsten nimmt man 0.

Übung 12.75 Zeigen Sie, daß der minmod-Limiter TVD ("total variation diminishing") ist, d.h. die Bedingung (12.55) erfüllt. ■

Bemerkung 12.76 Der Minimod-"slope limiter" ist nicht die einzig mögliche Wahl. Weitere klassische Wahlen sind der "superbee"- und der "van Leer"-Limiter. ■

Bemerkung 12.77 Falls die Rekonstruktion TVD ist, dann ist das Verfahren auch TVD, weil der Schritt "average" TVD ist und der Schritt "evolve" ebenfalls (falls die Evolution exakt gemacht wird...). ■

12.7.4 Lineare Systeme

Die Methodologie des Verfahrens (12.43) läßt sich auf Systeme übertragen: Hat man m Zustandvariablen, d.h. $\mathbf{u} \in \mathbb{R}^m$ und entsprechend eine Flußfunktion $f : \mathbb{R}^m \to \mathbb{R}^m$, so ergibt sich das Verfahren analog zu

$$\overline{\mathbf{u}}_i^{n+1} = \overline{\mathbf{u}}_i^n - \frac{k}{h}\left(\mathbf{F}_{i+1/2}^n - \mathbf{F}_{i-1/2}^n\right) \tag{12.58}$$

mit zu bestimmenden Füßen $\mathbf{F}_{i+1/2}^n$. Wir betrachten nun den linearen Fall $\mathbf{f}(\mathbf{u}) = \mathbf{A}\mathbf{u}$ mit einer Matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$. Hyperbolizität verlangt, daß $\mathbf{A}$ reell diagonalisierbar ist.

Bemerkung 12.78 Die Reduktion auf lineare hyperbolische Systeme ist aus zwei Gründen relevant:

- lineare Systeme treten auf (Wellengleichung, Maxwellsche Gleichungen)
- für nichtlineare Probleme führen Ideen wie dem Roe-Löser auf lineare hyperbolische Probleme als Grundbaustein. ■

Für konstante Matrix $\mathbf{A}$ kann das Riemannproblem explizit gelöst werden: Seien $\mathbf{r}_1, \ldots, \mathbf{r}_m \in \mathbb{R}^m$ die (Rechts-)Eigenvektoren von $\mathbf{A}$ mit zugehörigen Eigenwerten λ_p, $p = 1, \ldots, m$. Sei

$$\begin{aligned} \Lambda &:= \text{diag}(\lambda_1, \ldots, \lambda_m), \\ \mathbf{R} &:= (\mathbf{r}_1, \ldots, \mathbf{r}_m) \end{aligned}$$

Dann ist $\mathbf{A} = \mathbf{R}\Lambda\mathbf{R}^{-1}$. Einführen von *charakteristischen Variablen*

$$\mathbf{v} := \mathbf{R}^{-1}\mathbf{u}$$

führt auf ein *entkoppeltes* System: $\mathbf{v}_t + \Lambda\mathbf{v}_x = 0$, oder, ausgeschrieben

$$\mathbf{v}_t^p + \lambda_p \mathbf{v}_x^p = 0, \qquad p = 1, \ldots, m.$$

Die explizite Lösung ist damit mit den Anfangsbedingungen $\mathbf{v}_0 := \mathbf{R}^{-1}\mathbf{u}_0$ in charakteristischen Variablen

$$\mathbf{v}^p(x,t) = \mathbf{v}_0^p(x - \lambda_p t), \qquad p = 1, \ldots, m.$$

Rücktransformation $\mathbf{u} = \mathbf{R}\mathbf{v}$ liefert dann die gesuchte Lösung.

Riemannproblem

Wir betrachten eine stückweise konstante Anfangsbedingung

$$\mathbf{u}_0(x) := \begin{cases} \mathbf{u}_l & \text{falls } x < 0 \\ \mathbf{u}_r & \text{falls } x > 0. \end{cases}$$

In charakteristischen Variablen ist die $p$te Komponente $\mathbf{v}^p$ gegeben durch

$$\mathbf{v}^p(x,t) = \begin{cases} \mathbf{v}_l^p & \text{falls } x < \lambda_p t \\ \mathbf{v}_r^p & \text{falls } x > \lambda_p t, \end{cases}$$

wobei natürlich $\mathbf{v}_l = \mathbf{R}^{-1}\mathbf{u}_l$, $\mathbf{v}_r = \mathbf{R}^{-1}\mathbf{u}_r$. Wie bereits in Lemma 12.67 im skalaren Fall gesehen, ist der Fluß der Riemannlösung auf der Linie $x = 0$ konstant. In charakteristischen Variablen ist er gegeben als

$$\lambda_p v(0,t) = \begin{cases} \lambda_p \mathbf{v}_l^p & \text{falls } \lambda_p \geq 0 \\ \lambda_p \mathbf{v}_r^p & \text{falls } \lambda_p \leq 0. \end{cases}$$

Damit ist

$$\Lambda \mathbf{v}(0,t) = \Lambda^+ \mathbf{v}_l + \Lambda^- \mathbf{v}_r$$

mit $\Lambda^+ = \operatorname{diag}(\max\{\lambda_p, 0\})$ und $\Lambda^- = \operatorname{diag}(\min\{\lambda_p, 0\})$. In den physikalischen Variablen ergibt sich deshalb

$$\begin{aligned} \mathbf{A}\mathbf{u}(0,t) &= \mathbf{R}\Lambda\mathbf{R}^{-1}\mathbf{u}(0,t) \\ &= \mathbf{R}\Lambda\mathbf{v}(0,t) \\ &= \mathbf{R}\Lambda^+\mathbf{v}_l + \mathbf{R}\Lambda^-\mathbf{v}_r \\ &= \mathbf{A}^+\mathbf{u}_l + \mathbf{A}^-\mathbf{u}_r, \end{aligned}$$

wobei

$$\mathbf{A}^+ = \mathbf{R}\Lambda^+\mathbf{R}^{-1}, \qquad \mathbf{A}^- = \mathbf{R}\Lambda^-\mathbf{R}^{-1}.$$

Godunovverfahren

Das Godunovverfahren ist damit gegeben mit der Flußfunktion

$$\mathbf{F}_{i+1/2}^n := \mathbf{F}^G(\overline{\mathbf{u}}_i^n, \overline{\mathbf{u}}_{i+1}^n) := \mathbf{A}^+\overline{\mathbf{u}}_i^n + \mathbf{A}^-\overline{\mathbf{u}}_{i+1}^n.$$

Alternative Darstellungen sind:

$$\begin{aligned} \mathbf{F}_{i+1/2}^n &= \mathbf{A}\overline{\mathbf{u}}_i^n + \mathbf{A}^-\left(\overline{\mathbf{u}}_{i+1}^n - \overline{\mathbf{u}}_i^n\right) \\ &= \mathbf{A}\overline{\mathbf{u}}_{i+1}^n + \mathbf{A}^+\left(\overline{\mathbf{u}}_{i+1}^n - \overline{\mathbf{u}}_i^n\right) \\ &= \frac{1}{2}\mathbf{A}\left(\overline{\mathbf{u}}_i^n + \overline{\mathbf{u}}_{i+1}^n\right) - \frac{1}{2}|\mathbf{A}|\left(\overline{\mathbf{u}}_{i+1}^n - \overline{\mathbf{u}}_i^n\right), \end{aligned}$$

wobei $|\mathbf{A}| = \mathbf{A}^+ - \mathbf{A}^-$.

Lax-Friedrichs und Rusanov

Das Godunovverfahren hat den "Nachteil", daß eine Eigenwert/Eigenvektor-Zerlegung von $\mathbf{A}$ benötigt wird[16]. Man kann das Lax-Friedrichs Verfahren aus dem skalaren Fall wie folgt verallgemeinern:

$$\mathbf{F}^{LF}(\overline{\mathbf{u}}_i^n, \overline{\mathbf{u}}_{i+1}^n) = \frac{1}{2}\mathbf{A}\left(\overline{\mathbf{u}}_i^n + \overline{\mathbf{u}}_{i+1}^n\right) - \frac{h}{2k}\left(\overline{\mathbf{u}}_{i+1}^n - \overline{\mathbf{u}}_i^n\right)$$

Man beachte, daß hier nur $\mathbf{A}$ auftritt und keine Information über die Eigenwerte/Eigenvektoren von $\mathbf{A}$ benötigt werden.

Die Verallgemeinerung des besseren Rusanov-Verfahrens ist

$$\mathbf{F}^{Rus}(\overline{\mathbf{u}}_i^n, \overline{\mathbf{u}}_{i+1}^n) = \frac{1}{2}\mathbf{A}\left(\overline{\mathbf{u}}_i^n + \overline{\mathbf{u}}_{i+1}^n\right) - \frac{\lambda_{max}}{2}\left(\overline{\mathbf{u}}_{i+1}^n - \overline{\mathbf{u}}_i^n\right),$$

wobei λ_{max} der größte Eigenwert von $\mathbf{A}$—in der Praxis wird man sich mit einer Schätzung begnügen.

[16] für konstantes $\mathbf{A}$ ist dies natürlich kein Problem—im nichtlinearen Fall müßte aber für jede Zellgrenze erneut diagonalisiert werden...

12.7.5 Roe-Löser für nichtlineare Probleme

Ziel: Konstruktion einer Matrix $\widehat{\mathbf{A}}$, die die Bedingungen (i)—(iii) einer Roe-Matrix erfüllt.

Im Prinzip kann z.B. die Formel

$$\widehat{\mathbf{A}}(\mathbf{u},\mathbf{v}) := \int_{\tau=0}^{1} \mathbf{f}'(\mathbf{u}+t(\mathbf{v}-\mathbf{u}))\,dt$$

verwendet werden. Allerdings ist man an einfachen Konstruktionen interessiert, bei der sich auch die Eigenwerte und Eigenvektoren leicht ermitteln lassen. Wichtige Beispiele sind von der Form

$$\widehat{\mathbf{A}}(\mathbf{u},\mathbf{v}) = \mathbf{f}'(\widetilde{\mathbf{u}}), \tag{12.59}$$

wobei $\widetilde{\mathbf{u}} = \widetilde{\mathbf{u}}(\mathbf{u},\mathbf{v})$ ein geeigneter Zustand ist. Im folgenden Beispiel der Eulergleichungen ist $\widetilde{\mathbf{u}}$ ein gewichtetes Mittel von $\mathbf{u}$ und $\mathbf{v}$.

Beispiel 12.79 (Roematrix für 1D Euler) Wir betrachten die isothermen 1D-Eulergleichungen[17]

$$\begin{pmatrix} \rho \\ \rho v \end{pmatrix}_t + \begin{pmatrix} \rho v \\ \rho v^2 + a^2\rho \end{pmatrix}_x = 0.$$

Wir schreiben $\mathbf{u} = (\rho, \rho v)^T$. Die Idee besteht darin, eine neue Variable

$$\mathbf{z} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

einzuführen, so daß die Komponenten von $\mathbf{u}$ und $\mathbf{f}(\mathbf{u})$ als quadratische Polynome in den Koeffizienten α, β von $\mathbf{z}$ dargestellt werden. Eine geschickte Wahl ist

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \mathbf{z} := \rho^{-1/2}\mathbf{u} = \begin{pmatrix} \rho^{1/2} \\ \rho^{1/2} v \end{pmatrix}.$$

Dann ist

$$\mathbf{u} = \begin{pmatrix} \alpha^2 \\ \alpha\beta \end{pmatrix}, \qquad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} \alpha\beta \\ a^2\alpha^2 + \beta^2 \end{pmatrix}.$$

Das Besondere an quadratischen Polynomen g ist, daß

$$g(\mathbf{z}_1) - g(\mathbf{z}_2) = \nabla g(\overline{\mathbf{z}}) \cdot (\mathbf{z}_1 - \mathbf{z}_2), \qquad \overline{\mathbf{z}} = \frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2).$$

Es gilt daher:

$$\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{B}(\mathbf{z}_1 - \mathbf{z}_2), \qquad \text{mit } \mathbf{B} = \frac{d}{d\mathbf{z}}\mathbf{u}(\overline{\mathbf{z}}) = \begin{pmatrix} 2\overline{\alpha} & 0 \\ \overline{\beta} & \overline{\alpha} \end{pmatrix},$$

$$\mathbf{f}(\mathbf{u}_1) - \mathbf{f}(\mathbf{u}_2) = \mathbf{C}(\mathbf{z}_1 - \mathbf{z}_2), \qquad \text{mit } \mathbf{C} = \frac{d}{d\mathbf{z}}\mathbf{f}(\mathbf{u}(\overline{\mathbf{z}})) = \begin{pmatrix} \overline{\beta} & \overline{\alpha} \\ 2a^2\overline{\alpha} & 2\overline{\beta} \end{pmatrix}.$$

Damit erfüllt die Matrix $\widehat{\mathbf{A}}(\mathbf{u}_1,\mathbf{u}_2) := \mathbf{C}\mathbf{B}^{-1}$ die Bedingung (i). Weil nach der Kettenregel auch $\mathbf{C} = \mathbf{f}'(\mathbf{u}(\overline{\mathbf{z}}))\mathbf{B}$ gilt, haben wir

$$\widehat{\mathbf{A}}(\mathbf{u}_1,\mathbf{u}_2) = \mathbf{f}'(\mathbf{u}(\overline{\mathbf{z}})).$$

so daß $\widehat{\mathbf{A}}$ die Form (12.59) hat. Die reelle Diagonalisierbarkeit von $\widehat{\mathbf{A}}$, d.h. die Bedingung (ii) vererbt sich aus der von $\mathbf{f}'$. Die Bedingung (iii) ergibt sich aus der Tatsache, daß $\mathbf{u}$ glatt von $\mathbf{z}$ abhängt. Damit ist die Roe-Matrix

$$\widehat{\mathbf{A}}(\mathbf{u}_1,\mathbf{u}_2) = \begin{pmatrix} 0 & 1 \\ a^2 - \widetilde{v}^2 & 2\widetilde{v} \end{pmatrix}, \qquad \widetilde{v} = \frac{\overline{\beta}}{\overline{\alpha}} = \frac{\rho_1^{1/2} v_1 + \rho_2^{1/2} v_2}{\rho_1^{1/2} + \rho_2^{1/2}}.$$

Bemerkung 12.80 Die Technik kann auch für die allgemeinen 1D-Eulergleichungen verwendet werden. Die neue Variable ist $\mathbf{z} = \rho^{1/2}(1, v, H)$ mit der spezifischen Enthalpie H. Siehe Schmeiserskript (oder die meisten Bücher zur numerischen Gasdynamik) für die Formel. ∎

[17] aus $pV = nRT$ für ideale Gase folgt für T =const dann $p \sim \rho$

12.7.6 Monotone Verfahren und der Satz von Lax-Wendroff

Eigenschaften des kontinuierlichen Problems

Wir haben für skalare Probleme für die Entropielösung:

Lemma 12.81 (Kružkov, [12, Thm. 2.14]) *Für u_0, $v_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}) \cap \mathrm{TV}(\mathbb{R})$ erfüllen die (eindeutigen) Entropielösungen u und v der Gleichung $u_t + f(u)_x = 0$:*

(i) (L^∞-Stabilität) $\|u(\cdot,t)\|_{L^\infty} \le \|u_0\|_{L^\infty}$

(ii) (L^1-Kontraktivität) $\|u(\cdot,t) - v(\cdot,t)\|_{L^1} \le \|u_0 - v_0\|_{L^1}$

(iii) (Monotonie) $u_0 \le v_0$ impliziert $u(\cdot,t) \le v(\cdot,t)$

(iv) (Monotonieerhaltung) $u_0(\cdot)$ monoton impliziert $u(\cdot,t)$ monoton

(v) (TVD) $|u(\cdot,t)|_{TV} \le |u_0|_{TV}$.

Beweis: Die Beweise finden sich in [12, Thm. 2.14]. Wir bemerken, daß (ii) die Eigenschaft (v) impliziert. Wesentlich dafür ist die Charakterisierung (vgl. [12, Appendix A])

$$|u|_{TV} = \lim_{\varepsilon \to 0} \varepsilon^{-1} \|u(\cdot + \varepsilon) - u(\cdot)\|_{L^1(\mathbb{R})}$$

Dann folgt aus (ii) mit $v_0(x) := u_0(x + \varepsilon)$

$$\varepsilon^{-1} \|u(\cdot + \varepsilon, t) - u(\cdot, t)\|_{L^1} \le \varepsilon^{-1} \|u_0(\cdot + \varepsilon) - u_0(\cdot)\|_{L^1}$$

Läßt man $\varepsilon \to 0$ folgt die Behautpung. □

der Satz von Lax-Wendroff

Wir schreiben $u^{k,h}$ für die stückweise konstante Funktion mit Zellmitteln $\overline{u}_j^n$.

Satz 12.82 (Lax-Wendroff) *Sei $u^{k,h}$ mittels eines konsistenten, konservativen Verfahrens der Form (12.43) erzeugt mit Anfangsdaten $u_0 \in L^1 \cap L^\infty$. Es sei der numerische Fluß $\widehat{F}$ lipschitzstetig. Es sei, für alle k, h (von Interesse)*

$$\begin{aligned} |u^{k,h}(\cdot,t)|_{TV} &\le M \quad \text{für alle } t > 0, \\ \|u^{k,h}\|_{L^\infty(\mathbb{R}\times\mathbb{R}^+)} &\le M'. \end{aligned}$$

Sei $(k_m, h_m)_m$ eine Nullfolge mit $u^{k_m,h_m} \to u$ in $L^1_{loc}(\mathbb{R} \times \mathbb{R}^+)$. Dann ist die Funktion u eine schwache Lösung von $u_t + (f(u))_x = 0$.

Beweis: z.z.: für alle $\varphi \in C_0^\infty(\mathbb{R}^2)$ gilt

$$\int_{\mathbb{R}\times\mathbb{R}^+} u\varphi_t + f(u)\varphi_x = -\int_{\mathbb{R}} u(\cdot,0)\varphi(\cdot,0). \tag{12.60}$$

Sei $\varphi \in C_0^\infty(\mathbb{R}^2)$. Insbesondere sei $T > 0$ derart, daß $\varphi(x,t) = 0$ für alle $x \in \mathbb{R}$ und $t \ge T$. Die Annahme $\varphi \in C_0^\infty(\mathbb{R}^2)$ stellt sicher, daß im Folgenden alle Summen tatsächlich endliche Summen sind.

Das Verfahren liefert

$$\sum_{n\ge 0}\sum_j \varphi(x_j,t_n)(\overline{u}_j^{n+1} - \overline{u}_j^n) = -\frac{k}{h}\sum_{n\ge 0}\sum_j \varphi(x_j,t_n)\left(F_{j+1/2}^n - F_{j-1/2}^n\right)$$

Durch partielle Summation erhält man

$$-\sum_j \varphi(x_j,0)\overline{u}_j^0 - \sum_j\sum_{n\ge 0}(\varphi(x_j,t_n) - \varphi(x_j,t_{n-1}))\overline{u}_j^n - \frac{k}{h}\sum_{n\ge 0}\sum_j (\varphi(x_{j+1},t_n) - \varphi(x_j,t_n))\, F_{j+1/2}^n = 0.$$

Umformen liefert

$$-h\sum_j \varphi(x_j,0)\overline{u}_j^0 = kh\sum_{n,j}\frac{\varphi(x_j,t_n)-\varphi(x_j,t_{n-1})}{k}\overline{u}_j^n + \frac{\varphi(x_{j+1},t_n)-\varphi(x_j,t_n)}{h}F_{j+1/2}^n$$

Der interessante Term für den Limes k, $h\to 0$ ist der Flußterm. Wir schreiben

$$F_{j+1/2}^n = \widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n) = \widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n) - f(\overline{u}_j^n) + f(\overline{u}_j^n).$$

Wir müssen also zeigen:

$$\lim_{k,h\to 0} kh\sum_{n,j}|\widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n) - f(\overline{u}_j^n)| = 0.$$

Mit der Lipschitzstetigkeit von $\widehat{F}$ und der Konsistenz $\widehat{F}(a,a) = f(a)$ folgt

$$\begin{aligned}\sum_{0\le nk\le T}\sum_j |\widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n) - f(\overline{u}_j^n)| &\le \sum_{0\le nk\le T}\sum_j |\widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n) - \widehat{F}(\overline{u}_j^n,\overline{u}_j^n)| \\ &\le \|\widehat{F}\|_{Lip}\sum_{0\le nk\le T}\sum_j |\overline{u}_{j+1}^n - \overline{u}_j^n| = \sum_{0\le nk\le T}\|\widehat{F}\|_{Lip}|u^{kh}(\cdot,t_n)|_{TV} \le \frac{MT}{k}\|\widehat{F}\|_{Lip}.\end{aligned}$$

Damit folgt

$$kh\sum_{0\le nk\le T}\sum_j |\widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n) - f(\overline{u}_j^n)| \le hMT\|\widehat{F}\|_{Lip} \overset{h\to 0}{\to} 0.$$

□

Bemerkung 12.83 Die Grenzfunktion u in Satz 12.82 ist nicht notwendigerweise eine Entropielösung. Betrachte hierzu die Burgersgleichung mit

$$u_0(x) = \begin{cases} 1 & x>0 \\ -1 & x<0 \end{cases}$$

dessen Entropielösung ein Verdünnungsfächer ist. Der Fluß $\widehat{F}$ ist

$$\widehat{F}(v,w) = \begin{cases} \frac{1}{2}v^2 & v+w\ge 0 \\ \frac{1}{2}w^2 & v+w<0 \end{cases}$$

Mit der Anfangsbedingung $\overline{u}_j^0 = -1$ für $j\le 0$ und $\overline{u}_j^0 = 1$ für $j>0$ folgt $\overline{u}_j^n = \overline{u}_j^0$ für alle n und j. Die Diskretisierung konvergiert also gegen eine schwache Lösung, aber nicht die Entropielösung. ■

Bemerkung 12.83 zeigt, daß man weitere Bedingungen an die Flußfunktion $\widehat{F}$ stellen muß. Weiters zeigt Satz 12.82, daß es sinnvoll ist, die TVD-Eigenschaft des kontinuierlichen Problems ins Diskrete zu vererben. Eine wichtige Klasse von Flußfunktionen, die das leistet, sind *monotone* Flußfunktionen, d.h. $\widehat{F}(\uparrow,\downarrow)$.

Übung 12.84 Zeigen Sie, daß monotone Verfahren die Monotonieeigenschaft von Lemma 12.81, (iii) in dem folgenden Sinn erhalten: die Updateformel $\overline{u}_j^{n+1} = H(\overline{u}_{j-1}^n,\overline{u}_j^n,\overline{u}_{j+1}^n)$ ist monoton wachsend in jedem Argument, falls die Flußfunktion $\widehat{F}$ folgende Eigenschaften hat:

$$\text{Monotonie: } \widehat{F}(\uparrow,\downarrow)$$

$$\text{CFL-Bedingung: } |\partial_1\widehat{F}(a,b)| + |\partial_2\widehat{F}(a,b)| \le \frac{h}{k}.$$

Hinweis: Schreiben Sie $H(a,b,c) = b - \frac{k}{h}(\widehat{F}(b,c) - \widehat{F}(a,b))$, und betrachten Sie $\partial_1 H$, $\partial_3 H$, $\partial_2 H$. *Bemerkung:* vergl. auch Beweis von Lemma 12.88. ■

Monotone Flußfunktionen sind u.a.:

1. die Flußfunktion des Godunovverfahrens
2. Lax-Friedrichs
3. lokaler Lax-Friedrichs Fluß
4. Engquist-Osher

Bemerkung 12.85 (flux splitting) Eine weitere Klasse von monotonen Verfahren, die wir bisher nicht kennengelernt haben, sind sog. *Fluß-Splitting*-Verfahren: Dabei schreibt man $f(u) = f^+(u) + f^-(u)$ für zwei Funktionen f^+, f^- mit $(f^+)' \geq 0$ und $(f^-)' \leq 0$ und setzt

$$\widehat{F}(a,b) := f^+(a) + f^-(b) \tag{12.61}$$

■

Bemerkung 12.86 Das Engquist-Osher-Verfahren kann als Fluß-Splittingverfahren verstanden werden. Falls f konvex ist mit einem Minimum bei ω und kein Maximum hat, dann kann der Engquist-Osher-Fluß F^{EO} geschrieben werden als

$$F^{EO}(a,b) = f(\max\{a,\omega\}) + f(\min\{b,\omega\}) - f(\omega) = f^+(a) + f^-(b)$$

mit

$$f^+(a) := f(\max\{a,\omega\}), \qquad f^-(b) := f(\min\{b,\omega\}).$$

■

Auswahl einer konvergenten Teilfolge

Satz 12.82 nimmt an, daß eine Teilfolge $(u^{k_m,h_m})_m$ in L^1_{loc} konvergiert. Wenn das Verfahren die Totalvariation gleichmäßig beschränkt, dann folgt dies aus dem Auswahl von Helly, einer Variante des Kolmogorovschen Kompaktheitskriterium:

Lemma 12.87 *Sei die Flußfunktion $\widehat{F}$ lipschitzstetig. Sei $u^{k,h}(\cdot,0) \in L^1(\mathbb{R})$ und gelte für die erzeugte Folge $|u^{k_m,h_m}(\cdot,t)|_{TV} \leq M$ für alle $t \geq 0$. Dann gilt:*

(i) für $0 \leq s,t \leq T$:

$$\|u^{k,h}(\cdot,t) - u^{k,h}(\cdot,s)\|_{L^1} \leq CM(|s-t| + k)$$

(ii) Falls zusätzlich $\|u^{k,h}(\cdot,t)\|_{L^\infty} \leq M'$ für alle t, dann kann eine Teilfolge $(u^{k_m,h_m})_m$ ausgewählt werden, die in $L^1_{loc}(\mathbb{R} \times \mathbb{R}^+)$ konvergiert.

Beweis: *Beweis von (i):* Wir schreiben für einen Zeitschritt

$$\begin{aligned}
\|u^{k,h}(\cdot,t_{n+1}) - u^{k,h}(\cdot,t_n)\|_{L^1(\mathbb{R})} &= h\sum_j |\overline{u}_j^{n+1} - \overline{u}_j^n| = k\sum_j |\widehat{F}(\overline{u}_{j-1}^n,\overline{u}_j^n) - \widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n)| \\
&\leq k\sum_j |\widehat{F}(\overline{u}_{j-1}^n,\overline{u}_j^n) - \widehat{F}(\overline{u}_j^n,\overline{u}_j^n)| + |\widehat{F}(\overline{u}_j^n,\overline{u}_j^n) - \widehat{F}(\overline{u}_j^n,\overline{u}_{j+1}^n)| \\
&\leq k\|\widehat{F}\|_{Lip}\sum_j |\overline{u}_{j-1}^n - \overline{u}_j^n| + |\overline{u}_j^n - \overline{u}_{j+1}^n| \\
&\leq 2k\|\widehat{F}\|_{Lip}|u^{k,h}(\cdot,t_n)|_{TV} \leq 2k\|\widehat{F}\|_{Lip}M.
\end{aligned}$$

Damit folgt für t_n und t_m, daß $\|u^{k,h}(\cdot,t_n) - u^{k,h}(\cdot,t_m)\|_{L^1(\mathbb{R})} \leq 2|t_n - t_m|\|\widehat{F}\|_{Lip}M$ und damit die Behauptung.

Beweis von (ii): Der Hellysche Auswahlsatz besagt (vgl. z.B. [12, Cor. A.7]), daß eine Folge $(h_n)_n$ von Funktionen auf dem Intervall $[a,b]$, welche $\|h_n\|_{L^\infty(a,b)} \leq M$ und $|h_n|_{TV} \leq M$ erfüllt, eine Teilfolge hat, die fast überall gegen eine Grenzfunktion h konvergiert, die ebenfalls von beschränkter Variation ist.

Mit dem Hellyschen Auswahlsatz folgt die Aussage dann mit (i): wähle abzählbare, dichte Menge $(t_\ell)_\ell$ von $[0,T]$, wähle Teilfolgen von $u^{k,h}(\cdot,t_\ell)$ mittels Helly, verwende geeignetes Diagonalisierungsargument und die "Stetigkeitsaussage" (i), um die Aussage für alle $t \in [0,T]$ zu bekommen. □

Monotone Verfahren

Monotone Verfahren sind L^∞-stabil und TVD.

Lemma 12.88 (L^∞-Stabilität) *Sei die Flußfunktion $\widehat{F}$ monoton und lipschitzstetig. Dann gilt: aus*

$$u_{min} \leq \overline{u}_j^0 \leq u_{max} \qquad \forall j$$

folgt $u_{min} \leq \overline{u}_j^n \leq u_{max}$ für alle j, n unter der CFL-Bedingung

$$k \leq \frac{h}{2\|\widehat{F}\|_{Lip}}. \tag{12.62}$$

Beweis: Der Beweis erfolgt durch Induktion nach n. Wir schreiben das Verfahren als

$$\overline{u}_j^{n+1} = (1 - b_{j+1/2} - a_{j-1/2})\overline{u}_j^n + b_{j+1/2}\overline{u}_{j+1}^n + a_{j-1/2}\overline{u}_{j-1}, \tag{12.63}$$

$$b_{j+1/2} = \begin{cases} 0 & \text{falls } \overline{u}_{j+1}^n = \overline{u}_j^n \\ \frac{k}{h}\frac{\widehat{F}(\overline{u}_j^n, \overline{u}_{j+1}^n) - \widehat{F}(\overline{u}_j^n, \overline{u}_j^n)}{\overline{u}_j^n - \overline{u}_{j+1}^n} & \text{sonst} \end{cases}$$

$$a_{j-1/2} = \begin{cases} 0 & \text{falls } \overline{u}_{j-1}^n = \overline{u}_j^n \\ \frac{k}{h}\frac{\widehat{F}(\overline{u}_{j-1}^n, \overline{u}_j^n) - \widehat{F}(\overline{u}_j^n, \overline{u}_j^n)}{\overline{u}_{j-1}^n - \overline{u}_j^n} & \text{sonst} \end{cases}$$

Wegen der Monotonie von $\widehat{F}$ und der CFL-Bedingung (12.62)

$$0 \overset{\text{Monotonie}}{\leq} b_{j+1/2} \leq \frac{k}{h}\|\widehat{F}\|_{Lip} \overset{(12.62)}{\leq} \frac{1}{2}, \tag{12.64a}$$

$$0 \overset{\text{Monotonie}}{\leq} a_{j-1/2} \leq \frac{k}{h}\|\widehat{F}\|_{Lip} \overset{(12.62)}{\leq} \frac{1}{2}. \tag{12.64b}$$

Damit ist auch $0 \leq 1 - b_{j+1/2} - a_{j-1/2} \leq 1$. Dies zeigt, daß die Updateformel (12.63) eine Konvexkombination von Werten ist. □

Lemma 12.88 zeigt, daß monotone Schemata die L^∞-Stabilität des kontinuierlichen Problems erben.

Lemma 12.89 (TVD) *Sei die Flußfunktion $\widehat{F}$ monoton und lipschitzstetig. Dann gilt:*

$$\sum_j |\overline{u}_{j+1}^{n+1} - \overline{u}_j^{n+1}| \leq \sum_j |\overline{u}_{j+1}^n - \overline{u}_j^n| \qquad \forall n \in \mathbb{N}_0.$$

Beweis: Wir verwenden die Updateformel (12.63)

$$\begin{aligned} \overline{u}_j^{n+1} &= \overline{u}_j^n + b_{j+1/2}(\overline{u}_{j+1}^n - \overline{u}_j^n) + a_{j-1/2}(\overline{u}_{j-1}^n - \overline{u}_j^n), \\ \overline{u}_{j+1}^{n+1} &= \overline{u}_{j+1}^n + b_{j+3/2}(\overline{u}_{j+2}^n - \overline{u}_{j+1}^n) + a_{j+1/2}(\overline{u}_j^n - \overline{u}_{j+1}^n). \end{aligned}$$

Damit folgt mit der CFL-Bedingung (12.62)

$$\begin{aligned} |\overline{u}_{j+1}^{n+1} - \overline{u}_j^{n+1}| &\leq |\overline{u}_{j+1}^n - \overline{u}_j^n||1 - b_{j+1/2} - a_{j+1/2}| + |b_{j+3/2}||\overline{u}_{j+2}^n - \overline{u}_{j+1}^n| + |a_{j-1/2}||\overline{u}_j^n - \overline{u}_{j-1}^n| \\ &\overset{(12.64)}{=} |\overline{u}_{j+1}^n - \overline{u}_j^n|(1 - b_{j+1/2} - a_{j+1/2}) + b_{j+3/2}|\overline{u}_{j+2}^n - \overline{u}_{j+1}^n| + a_{j-1/2}|\overline{u}_j^n - \overline{u}_{j-1}^n| \end{aligned}$$

Eine Summation über j liefert die Behauptung. □

Lemma 12.89 zeigt, daß monotone Schemata die TVD-Eigenschaft des kontinuierlichen Problems erben.

Eine weitere wichtige Eigenschaft monotoner Schemata ist, daß sie auch die Entropiebedingung erhalten:

Lemma 12.90 *Sei $\widehat{F}$ eine monotone Flußfunktion. Dann gilt unter den Voraussetzungen von Satz 12.82, daß die Grenzfunktion u die Kružkov-Entropiebedingung*

$$\int_{\mathbb{R}\times\mathbb{R}^+} |u-z|\varphi_t + \operatorname{sign}(u-z)(f(u)-f(z))\varphi_x \geq 0 \qquad \forall z \in \mathbb{R}, \quad \varphi \in C_0^\infty(\mathbb{R}\times\mathbb{R}^+), \quad \varphi \geq 0$$

erfüllt.

Beweis: Wir verweisen auf [12, Thm. 3.9]. Entscheidend ist, daß monotone Schemata auch eine diskrete Entropiebedingung erfüllen, die im Limes k, $h \to 0$ die Kružkovbedingung liefert. □

Aus Lemma 12.90 folgt, daß die Grenzfunktion u aus Satz 12.82 für monotone Flußfunktionen tatsächlich die Entropielösung ist; weiters folgt aus den obigen Betrachtungen, daß tatsächlich eine Teilfolge u^{k_m,h_m} existiert, die gegen die Entropielösung konvergiert.

Ein Nachteil monotoner Verfahren ist, daß sie nur 1. Ordnung sein können:

Lemma 12.91 *Monotone Verfahren sind höchstens erster Ordnung.*

Beweis: [12, Thm. 3.10]. □