Kapitel 11

lineare hyperbolische Gleichungen

11.0 Vorbemerkungen

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

$$\partial_t \mathbf{u} + \sum_{j=1}^{d} \partial_{x_j}(\mathbf{f}^j(\mathbf{u})) = 0 \qquad (x,t) \in \Omega \times (0,\infty). \tag{11.1}$$

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

Hyperbolische Erhaltungsgleichungen von der Form (11.1) 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 11.1 (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 (11.1) 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 11.2 Die Gleichung (11.1) 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 11.3 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

$$\partial_t u + \nabla \cdot (\mathbf{F}(u)) = 0. \tag{11.2}$$

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 11.4 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*

$$u_t + \mathbf{b} \cdot \nabla u = 0. \tag{11.3}$$

∎

Bemerkung 11.5 (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 (11.1) noch eine Bedingung der reellen Diagonalisierbarkeit der Linearisierung:

Definition 11.6 (Hyperbolizität einer Erhaltungsgleichung) *(11.1) heißt hyperbolisch, falls die Ableitung $D_{\mathbf{u}}\mathbf{F}(\mathbf{u}, \underline{\omega})$ für jeden Zustandsvektor $\mathbf{u}$ und jede Richtung $\underline{\omega} \in \mathbb{R}^d \setminus \{0\}$ reell diagonalisierbar ist.*

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

Übung 11.7 Sei $d = 1$. Das System hat die Form

$$\partial_t \mathbf{u} + \partial_x(\mathbf{F}(\mathbf{u})) = 0 \tag{11.4}$$

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

Bemerkung 11.8 Das Problem (11.1) muß noch mit Randbedingungen (und Anfangsbedingungen) vervollständigt werden. ∎

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

11.1 klassische Differenzenverfahren am Beispiel der Advektionsgleichung

Eine simultane Diskretisierung in Ort und Zeit wie wir es in Abschnitt 11.4 vorstellen werden erfolgt in der Praxis selten. Fast ausschließlich werden bei (zeitabhängigen) hyperbolischen Problemen Zeitschrittverfahren eingesetzt—meist sogar explizit in der Zeit. Eine der zentralen Fragen bei solchen Verfahren ist die der Stabilität, und der vorliegende Abschnitt ist primär dieser Frage gewidmet. Eine zweite Frage ist, insbesondere bei Differenzenverfahren, die Realisierung von Randbedingungen. Wir werden diese Frage allenfalls kursorisch behandeln. Um die Frage nach Randbedingungen zu umgehen, betrachten wir ein reines Cauchyproblem (d.h. $\Omega = \mathbb{R}^d$– die klassische Alternative ist die Untersuchung von periodischen Randbedindungen). Um die Situation noch einfacher zu gestalten, betrachten wir den räumlichen 1D-Fall. Der einfachste Fall einer linearen hyperbolischen Gleichung ist damit die Advektionsgleichung:

$$u_t + au_x = g \qquad \text{auf } \mathbb{R} \times (0,\infty), \qquad u(x,0) = u_0(x), \tag{11.5}$$

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

$$u(x,t) = u_0(x - at). \tag{11.6}$$

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":

$$\frac{u_i^{n+1} - u_i^n}{k} + a\frac{u_i^n - u_{i-1}^n}{h} = g(x_i, t_n). \tag{11.7}$$

2. "forward time/forward space":

$$\frac{u_i^{n+1} - u_i^n}{k} + a\frac{u_{i+1}^n - u_i^n}{h} = g(x_i, t_n). \tag{11.8}$$

3. "Lax-Friedrichs":

$$\frac{1}{k}\left(u_i^{n+1} - \frac{1}{2}\left(u_{i+1}^n + u_{i-1}^n\right)\right) + \frac{a}{2h}\left(u_{i+1}^n - u_{i-1}^n\right) = g(x_i, t_n). \tag{11.9}$$

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

$$U^{n+1} = EU^n + kG^n, \tag{11.10}$$

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} gegeben durch

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

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

dann ist der Konsistenzfehler

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

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

$$\tau_i^{n+1} = \frac{1}{k}\left(U_{kh,i}^{n+1} - (EU_{kh}^n)_i\right) - (u_t(x_i,t_n) + au_x(x_i,t_n)) \tag{11.11}$$

für jede hinreichend glatte Funktion u.

Übung 11.9 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:

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

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

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

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

$$\sup_{n:0\le nk\le T}\|E^n\|_{\ell_h^1} \le C_T, \tag{11.12}$$

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

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

Für die betrachteten Verfahren ergibt sich somit

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

Man kann sich als Startfehler $\|\varepsilon^0\|_{\ell_h^1}$ also $O(h)$ erlauben, was mit Knotenauswertung oder sogar bei Wahl als Mittelwerte über Elemente der Fall ist. Entscheidend ist die Stabilitätsbedingung (11.12). Praktisch ist sie erfüllt, falls

- $\|E\|_{\ell^1_h} \leq 1$
- oder doch wenigstens $\|E\|_{\ell^1_h} \leq 1 + Ck$ 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]

$$\frac{|ak|}{h} \leq c, \tag{11.13}$$

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

Folie

Übung 11.10 Zeigen Sie, daß für das Lax-Friedrichs-Schema unter der Voraussetzung der CFL-Bedingung $|ak/h| \leq 1$ die folgende Abschätzung gilt:

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

11.1.1 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:

$$\begin{cases} \text{verwende ft/bs (11.7)} & \text{falls } a > 0 \\ \text{verwende ft/fs (11.8)} & \text{falls } a < 0\ . \end{cases} \tag{11.14}$$

Satz 11.11 (Stabilität des Upwind-Verfahrens) *Unter der CFL-Bedingung $|ak/h| \leq 1$ erfüllt der Propagationsoperator E des Upwind-Verfahrens (11.14) die Stabilitätsbedingung $\|E\|_{\ell^1_h} \leq 1$. Für glatte Lösungen ist damit der Fehler $O(k+h)$.*

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

$$\begin{aligned} \|EV\|_{\ell^1_h} &\leq \sum_i h\left[\left|1 - \frac{ka}{h}\right||V_i| + |V_{i-1}|\left|\frac{ka}{h}\right|\right] \\ &\overset{(11.13)}{=} \left(1 - \frac{ak}{h}\right)\sum_i h|V_i| + \frac{ak}{h}\sum_i h|V_{i-1}| = \|V\|_{\ell^1_h}. \end{aligned}$$

□

Bemerkung 11.12 (physikalische Plausibilität der CFL-Bedingung und des Upwinding) Die Lösungsformel (11.6) 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 Intevall $[x_{i-n}, x_i]$ *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. ∎

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

Bemerkung 11.13 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

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

11.2 von Neumann-Analyse

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

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

Entsprechend bezeichnen wir mit ℓ^2_h 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^2_h-norm:

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

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

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

Es gilt:

Satz 11.14 *(i) (Parseval) $\mathcal{F}_h$ ist ein Isomorphismus $\ell^2_h \to L^2_h$:*

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

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 $(u_i)_{i\in\mathbb{Z}} \in \ell^2_h$ *und* $(v_i)_{i\in\mathbb{Z}} \in \ell^1_h$ *ist* $u * v \in \ell^2_h$ *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:

Übung 11.15

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

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

$$|\widehat{a}(\xi)| \le 1 + Ck \qquad \forall \xi \in [-\pi/h, \pi/h], \tag{11.15}$$

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

Beispiel 11.16 (Upwind-Verfahren) Das Upwind-Verfahren für die Advektionsgleichung im Fall $a = -1$ (und $g \equiv 0$) ist

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

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

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

Die Fouriertransformierte ist

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

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

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

d.h. das Verfahren erfüllt die von-Neumann-Bedingung (11.15). ∎

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 11.17 (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

$$\sup_{\xi \in [-\pi/h, \pi/h]} |g(\xi)| \leq 1$$

erfüllt ist, d.h. die von-Neumann-Bedingung (11.15).

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 $|g(\xi)| \approx 1 - 2\lambda^2 < 1$. M.a.W.: Während die niederfrequenten Term (wegen Konsistenz!) weder verstärkt noch gedämpft werden, werden die hochfrequenten Lösungskomponten (und damit auch die hochfrequenten Fehleranteile) gedämpft. ∎

Übung 11.18 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 11.19

- 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 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.

∎

11.3 Splittingmethoden/fractional step methods

11.3.1 Einführung: der ODE-Fall

Betrachte das ODE-System

$$u' = Au + Bu, \qquad u(0) = u_0, \tag{11.16}$$

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.[5]

[4] genauer: von der Ordnung 2 **pruefen!**. 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.

[5] 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.

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 eine 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:

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

- 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 eine 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:

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

Übung 11.20 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 11.21 (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 11.22 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. ∎

11.3.2 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 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 11.23 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 11.24 Betrachte die 2D-Advektionsgleichung

$$u_t + au_x + bu_y = 0 \qquad \text{in } \Omega = \mathbb{R}^2, \qquad u(\cdot, 0) = u_0(\cdot) \tag{11.17}$$

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 11.25 Betrachten Sie die Gleichung (11.17) 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 11.26 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 11.27 (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. ∎

11.3.3 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.[6]

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.

Beispiel 11.28 Eine Differenzendiskretisierung im Ort der Advektions-Diffusionsgleichung

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

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

$$u_{i,j}^{n+1} = u_{ij}^n + \frac{k}{h^2} \left[D_x^+ D_x^- + D_y^+ D_y^- \right] u_{i,j}^{n+1} - \frac{ak}{h} D_x^- u_{i,j}^n. \tag{11.18}$$

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

$$u_{i,j}^{n+1} = u_{ij}^n + \frac{k}{h^2} \left[D_x^+ D_x^- + D_y^+ D_y^- \right] u_{i,j}^{n+1} - \frac{ak}{h} D_x^- u_{i,j}^{n+1} \tag{11.19}$$

geführt (dann ohne CFL-Bedingung). Es ist zu betonen, daß das LGS in (11.18) ein SPD System ist, für das es effiziente Verfahren gibt. Das LGS in (11.19) ist nicht symmetrisch. ∎

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

modified equation?

11.4 Raum-Zeit-DG

Wir betrachten eine etwas allgemeinere Form von (11.3):

$$u_t + \mathbf{b}(x,t) \cdot \nabla u + c(x,t)u = g. \tag{11.20}$$

Im Unterschied zu parabolischen Gleichungen hat in (11.20) die t-Variable keine besonders herausgehobene Rolle. Nennt man $x_0 = t$, so können wir auch auch folgendes, allgemeineres Problem betrachten:

$$\mathbf{b}(x) \cdot \nabla u + c(x)u = f \quad \text{auf } \Omega, \tag{11.21}$$

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

$$\Gamma^- := \{x \in \partial\Omega \colon \mathbf{b}(x) \cdot n(x) < 0\}, \tag{11.22}$$
$$\Gamma^+ := \{x \in \partial\Omega \colon \mathbf{b}(x) \cdot n(x) > 0\}, \tag{11.23}$$
$$\Gamma^= := \{x \in \partial\Omega \colon \mathbf{b}(x) \cdot n(x) = 0\}. \tag{11.24}$$

Hier ist $n(x)$ der (äußere) Normalenvektor im Punkt $x \in \partial\Omega$. Tatsächlich können wir für (11.21) nur eine Randbedingung auf Γ^- oder Γ^+ fordern. Wir betrachten deshalb das Randwertproblem

$$\mathbf{b}(x) \cdot \nabla u + c(x)u = g \quad \text{auf } \Omega, \tag{11.25a}$$
$$u = 0 \quad \text{auf } \Gamma^-. \tag{11.25b}$$

Beispiel 11.29 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 (11.25). 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 (11.25) 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 (11.25a) 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.[7] Im vorliegenden Fall ist also der numerische Fluß $\widehat{h}_{K|K'}$ bereits eindeutig durch die Wahl von $\widehat{u}$ auf den Kanten festgelegt:

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

Es ergibt sich als numerisches Verfahren:

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

$$B_{DG}^{Trans}(u,v) := \sum_{K\in\mathcal{T}} \int_K u\,(cv + \nabla\cdot(\mathbf{b}v)) + \int_{\partial K} (\mathbf{b}\cdot n_K)\widehat{u}v = l(v) := \int_\Omega vg \qquad \forall v \in S^{p,0}(\mathcal{T}) \tag{11.26}$$

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

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

$$B_{DG}^{Trans}(u,v) = \sum_{K\in\mathcal{T}} \int_K (cu + \mathbf{b}\cdot\nabla u)v) + \int_{\partial K} (\mathbf{b}\cdot n_K)(\widehat{u} - u)v = l(v) := \int_\Omega vg \qquad \forall v \in S^{p,0}(\mathcal{T}) \tag{11.27}$$

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

Beispiel 11.30

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

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

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

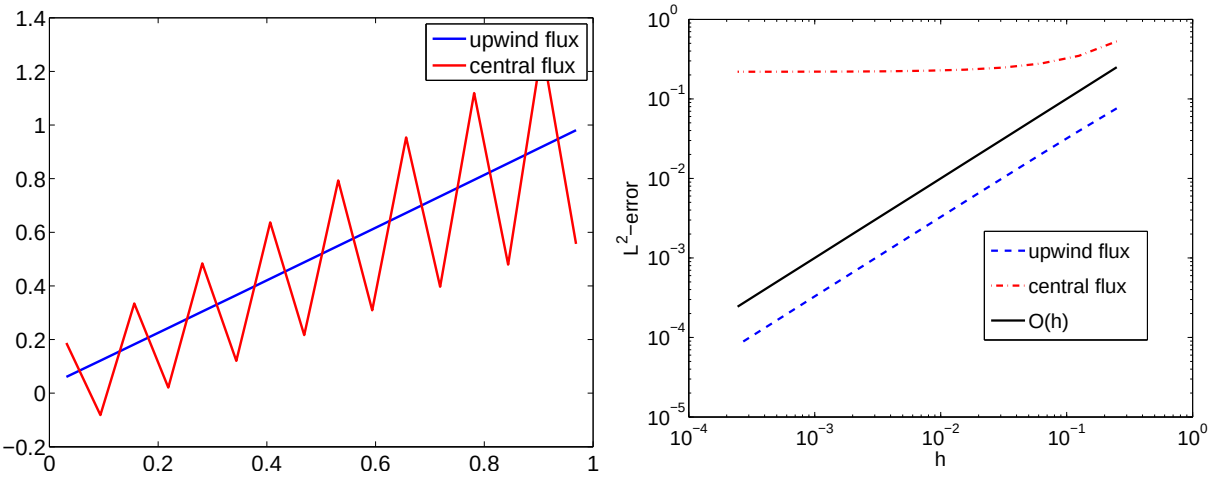

Abbildung 11.1: Links: Lösungsplot. Rechts: Konvergenzverhalten

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. 11.1, 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 (11.27) 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:

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

- Sei e ein Kante auf Γ^-: dann definieren wir $\widehat{u}|_e = 0$
- Sei e ein Kante auf $\partial\Omega \setminus \Gamma^-$: dann definieren wir $\widehat{u}|_e$ als Limes von u vom angrenzenden Element

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

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

Satz 11.31 *Es gelte*

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

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

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

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

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

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

Damit

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

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

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

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

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

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

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

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

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

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

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

Faßt man alle Kantenbeiträge zusammen, dann hat man

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

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

Satz 11.31 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

$$B_{DG}^{Trans}(u,v)=l(v)\qquad\forall v\in S^{p,0}(\mathcal{T}).\tag{11.28}$$

Damit ergibt sich die Galerkinorthogonalität

$$B_{DG}^{Trans}(u-u_N,v)=0\qquad\forall v\in S^{p,0}(\mathcal{T}).\tag{11.29}$$

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

Das numerische Verfahren in Abschnitt 11.4 hat ausgenutzt, daß die Zeitvariable eigentlich keine herausgehobene Rolle hat und deshalb eine Diskretisierung im Raum-Zeit-Zylinder durchgeführt. Wie bei parabolischen Verfahren sind jedoch in der Praxis Zeitschrittverfahren vorherrschend. Wir werden mit k den Zeitschritt bezeichnen.

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

$$u_t+\nabla\cdot\mathbf{f}(u)=0\quad\text{ auf }\Omega\times\mathbb{R}^+,\qquad u(\cdot,0)=u_0\tag{11.30}$$

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

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

Beispiel 11.33 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 numerisch 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 11.34 Die Bedingung $\widehat{h}(u, v, n) = -h(v, u, -n)$ 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 11.34:

Übung 11.35 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[8] hat, falls für es ebenfalls die Eigenschaft $\|\mathbf{y}^{n+1}\| \leq \|\mathbf{y}^n\|$ gilt. Das folgende Lemma 11.37 gibt hinreichende Kriterien an, unter denen ein explizites RK-Verfahren SSP ist. Zuvor eine Umformulierung des klassischen Vorgehens:

[8] strong stability preserving

Übung 11.36 Sei ein s-stufiges explizites RK-Verfahren beschrieben durch das Butcher-Tableau $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ Dann gilt für das RK-Verfahren angewandt auf $y' = g(y)$:

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

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

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

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

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

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

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

■

Lemma 11.37 *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 11.36 die Bedingung $\beta_{i,j} \geq 0$ erfüllen. Dann gilt: das RK-Verfahren ist SSP, falls die Schrittweite k die Bedingung*

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

erfüllt.

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

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

und damit

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

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

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

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

Übung 11.38

- Zeigen Sie, daß die explizite Trapezregel SSP ist:

0	0	0
1	1	0
	1/2	1/2

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

0	0	0
1/2	1/2	0
	0	1

■

Bemerkung 11.39 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, [4]. ■

11.6 klassische finite Volumenverfahren

Historisch älter als das Vorgehen in Abschnitt 11.5 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 11.40 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 (11.20) 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

$$\overline{u}_i^{n+1} = \overline{u}_i^n - \frac{h}{k}\left(\overline{F}_{i+1/2}^n - \overline{F}_{i-1/2}^n\right) \tag{11.31}$$

Bemerkung 11.41 (11.31) drückt eine exakte Erhaltung auf dem Kontrollvolumen $K_i \times (t_n, t_{n+1})$ aus. ∎

(11.31) ist noch kein numerisches Verfahren, da die Formel $\overline{F}^n_{i+1/2}$ (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 (11.31) geschrieben, wobei der exakte Fluß $\overline{F}^n_{i+1/2}$ durch einen numerischen Fluß $F^n_{i+1/2}$ ersetzt wird. Das Verfahren ist dann

$$\overline{u}_i^{n+1} = \overline{u}_i^n - \frac{k}{h}\left(F^n_{i+1/2} - F^n_{i-1/2}\right) \tag{11.32}$$

Typischerweise ist $F^n_{i+1/2}$ von der Form

$$F^n_{i+1/2} = \widehat{h}(\overline{u}^n_i, \overline{u}^n_{i+1}) \tag{11.33}$$

für eine numerische Flußfunktion $\widehat{h}$. Wir nennen Verfahren der Form (11.32) *konservativ*. Falls $\widehat{h}(u,u) = f(u)$, dann heißt das Verfahren *konsistent*.

Übung 11.42 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 11.43 Man kann den numerischen Fluß auch als Funktion von mehreren benachbarten Zellmitteln definieren. Die Notation wäre dann $F^n_{i+1/2} = \widehat{h}(\overline{u}^n_{i-p}, \overline{u}^n_{i-p+1}, \ldots, \overline{u}^n_{i+q})$ 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 11.44 Das Lax-Friedrichs-Schema ist

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

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

$$F^n_{i+1/2} := \widehat{h}(\overline{u}^n_i, \overline{u}^n_{i+1}) := \frac{h}{2k}\left(\overline{u}^n_i - \overline{u}^n_{i+1}\right) + \frac{1}{2}\left(f(\overline{u}^n_i) + f(\overline{u}^n_{i+1})\right)$$

∎

11.6.1 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

$$u_t + (f(u))_x = 0 \quad \text{auf } \mathbb{R}\times\mathbb{R}^+, \qquad u(\cdot,0) = \begin{cases} u_l & x < 0 \\ u_r & x > 0. \end{cases} \tag{11.34}$$

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:

$$u(x,t) = \begin{cases} u_l, & \text{falls } x \leq f'(u_l)t \\ (f')^{-1}(x/t), & \text{falls } f'(u_l)t \leq x \leq f'(u_r)t \\ u_r, & \text{falls } x > f'(u_r)t. \end{cases}$$

Bemerkung 11.45

- 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 11.46 *Sei u^R Lösung von (11.34). 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 11.47 (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 (11.32), wobei der Fluß $F^n_{i+1/2}$ durch

$$F^{G,n}_{i+1/2} := \frac{1}{k}\int_{t_n}^{t_{n+1}} f(u^R(x_{i+1/2}, t))\, dt \tag{11.35}$$

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 11.46 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:

$$F^{G,n}_{i+1/2} = \widehat{h}(\overline{u}^n_i, \overline{u}^n_{i+1}) := \begin{cases} \min_{\overline{u}^n_i \le v \le \overline{u}^n_{i+1}} f(v) & \text{falls } \overline{u}^n_i \le \overline{u}^n_{i+1} \\ \max_{\overline{u}^n_{i+1} \le v \le \overline{u}^n_i} f(v) & \text{falls } \overline{u}^n_i \ge \overline{u}^n_{i+1}. \end{cases} \tag{11.36}$$

Übung 11.48 Zeigen Sie die Formel (11.36) unter der Annahme, daß $f \in C^2(\mathbb{R};\mathbb{R})$ strikt konvex ist. ∎

Bemerkung 11.49 Die Formel (11.36) 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 11.50 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 11.51 Betrachten Sie die lineare Flußfunktion $f(u) = au$ für ein $a \in \mathbb{R}$. Zeigen Sie, daß der Godunovfluß auf das folgende Verfahren führt:

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

11.6.2 Approximative Riemannlöser

Die Bestimmung der Extrema in (11.36) 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 $\widehat{A}u$, wobei $\widehat{A} = \widehat{A}(\overline{u}^n_i, \overline{u}^n_{i+1})$. Folgende Bedinungen scheinen plausibel (wir formulieren sie gleich für den Fall von Systemen):

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

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

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

Bemerkung 11.52 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*

$$\widehat{A}_{i+1/2} = \begin{cases} \frac{f(\overline{u}^n_{i+1}) - f(\overline{u}^n_i)}{\overline{u}^n_{i+1} - \overline{u}^n_i} & \text{falls } \overline{u}^n_{i+1} \neq \overline{u}^n_i \\ f'(\overline{u}^n_i) & \text{falls } \overline{u}^n_i = \overline{u}^n_{i+1} \end{cases} \tag{11.37}$$

Der zugehörige numerische Fluß kann explizit berechnet werden:

Übung 11.53 Zeigen Sie: der numerische Fluß $F^n_{i+1/2}$, der zum obigen approximativen Riemannlöser von Roe gehört, ist gegeben durch

$$F^n_{i+1/2} = F^{Roe}(\overline{u}^n_i, \overline{u}^n_{i+1}) = \begin{cases} f(\overline{u}^n_i) & \text{falls } \widehat{A}_{i+1/2} \geq 0 \\ f(\overline{u}^n_{i+1}) & \text{falls } \widehat{A}_{i+1/2} < 0. \end{cases} \tag{11.38}$$

■

11.6.3 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[9]). 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:

$$U(x,t) = \begin{cases} u_l & \text{falls } x < s^l t \\ u_* & \text{falls } s^l t < x < s^r t \\ u_r & \text{falls } x > s^r t. \end{cases} \tag{11.39}$$

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:

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

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

$$F^n_{i+1/2} = \widetilde{F}(\overline{u}^n_i, \overline{u}^n_{i+1}) = \frac{s^r_{i+1/2} f(\overline{u}^n_i) - s^l_{i+1/2} f(\overline{u}^n_{i+1}) + s^r_{i+1/2} s^l_{i+1/2} (\overline{u}^n_{i+1} - \overline{u}^n_i)}{s^r_{i+1/2} - s^l_{i+1/2}}. \tag{11.40}$$

Wir betrachten nun den Spezialfall $s^l_{i+1/2} = -s^r_{i+1/2} = -s_{i+1/2}$. Dann vereinfacht sich (11.40) wie folgt:

$$F^n_{i+1/2} = \frac{f(\overline{u}^n_i) + f(\overline{u}^n_{i+1})}{2} - \frac{s_{i+1/2}}{2} \left(\overline{u}^n_{i+1} - \overline{u}^n_i \right). \tag{11.41}$$

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

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 (11.41)

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

Dieses Verfahren hatten wir bereits kennengelernt.

Rusanov-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

$$s^l_{i+1/2} = -s^r_{i+1/2}, \qquad s^r_{i+1/2} = \max\{|f'(\overline{u}^n_i)|, |f'(\overline{u}^n_{i+1})|\}$$

Das Verfahren ist dann

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

Enquist-Osher

fill in details here...

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 müssen aufgegeben werden, wenn man an Verfahren höherer Ordnung interessiert ist, sind Verfahren höherer Ordnung (sagen wir: Ordnung 2) das Mittel der Wahl.

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

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

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

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

fill in details here...

11.6.4 Lineare Systeme

Die Methodologie des Verfahrens (11.32) 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

$$\overline{\mathbf{u}}^{n+1}_i = \overline{\mathbf{u}}^n_i - \frac{k}{h}\left(\mathbf{F}^n_{i+1/2} - \mathbf{F}^n_{i-1/2}\right) \tag{11.42}$$

mit zu bestimmenden Füßen $\mathbf{F}^n_{i+1/2}$. 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 11.54 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

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

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 11.46 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

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

Alternative Darstellungen sind:

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

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[10]. Man kann das Lax-Friedrichs Verfahren aus dem skalaren Fall wie folgt verallgemeinern:

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

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

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

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

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