RWTH Aachen

Institut für Geometrie und Praktische Mathematik

Prof. Dr. R. Jeltsch

Diplomarbeit

Im wesentlichen - nicht - oszillierende Verfahren zur Approximation von hyperbolischen Erhaltungssätzen in einer Raumdimension

vorgelegt von

Siegfried Müller

Matrikel - Nr.: 145545

Ich versichere, daß ich diese Diplomarbeit selbständig verfaßt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Wörtliche oder sinngemäße Wiedergaben aus anderen Quellen sind kenntlich gemacht und durch Zitate belegt.

Aachen, den 1.3.1989

Inhaltsverzeichnis

Seite

Einleitung

In dieser Arbeit sollen spezielle Verfahren zur numerischen Lösung von s.g. Erhaltungssätzen näher untersucht werden. Diese — in der Literatur als **E**ssentially-**N**on-**O**scillatory – Verfahren bekannten Schemata — wurden von A. Harten, B. Engquist, S. Osher und S. R. Chakravarthy in den Jahren 1986/87 entwickelt sowie numerisch und analytisch untersucht (s. [ENO1], [ENO2], [ENO3]). Es konnten jedoch bis heute nicht alle Fragen beantwortet werden, so z.B.:

- Erfüllen diese Verfahren eine numerische Entropiebedingung ?
- Sind sie TV–stabil ?

Auf den ersten Punkt soll hier nicht näher eingegangen werden, dafür werden zum zweiten einige interessante Teilaspekte hergeleitet, die vielleicht zur Beantwortung dieser Frage beitragen könnten.

Die vorliegende Arbeit ist nun folgendermaßen aufgebaut:

In Kapitel 1 soll eine in [ENO1] und [ENO2] entwickelte Interpolationstechnik untersucht werden. Schwerpunkt wird hier die Analyse des Oszillationsverhaltens dieser Interpolation sein. Diese wurde z.T. auch schon in [ENO2] durchgeführt, jedoch äußerst knapp und stellenweise fehlerhaft. Darüberhinaus werden einige weitere Eigenschaften hergeleitet, die das numerisch gute Verhalten der ENO–Schemata erklären.

Im zweiten Kapitel sollen dann die eigentlichen Verfahren konstruiert und analysiert werden. Dazu wird zunächst der skalare Fall ausführlich betrachtet und anschließend die Theorie auf Systeme übertragen. Hier wird die Untersuchung von s.g. Rekonstruktionsmethoden, die aus den Mittelwerten einer Funktion diese von gleichmäßig hoher Ordnung approximieren, im Vordergrund stehen. Ferner wird das Ordnungs– sowie das Oszillationsverhalten dieser Verfahren ausführlich analysiert.

Da der durch die Konstruktionsvorschrift der ENO-Verfahren festgelegte numerische Fluß i.a. nur approximativ bestimmbar ist, bleibt dem Benutzer ein beträchtlicher Spielraum für Modifikationen. In Kapitel 3 und 4 werden zwei Möglichkeiten näher vorgestellt, zum einen die Methode von Roe (s. [Ro1]), die auch schon in [ENO3] verwendet wurde, und zum anderen eine von P. LeFloch und P. A. Raviart in den Jahren 1986/87 entwickelte. Letztere wird aus dem Grund näher analysiert, weil sie zu der Hoffnung berechtigt, daß die ENO–Verfahren mit dieser Modifikation sich bei Interaktionsproblemen von Schocks wesentlich besser verhalten. In Kapitel 5 werden deshalb solche Probleme für den skalaren Fall sowie für die Eulergleichung hergeleitet, die als Testprobleme für die numerische Untersuchung dienen.

Da die ENO–Verfahren auch einige Nachteile haben, z.B. den enormen Programmieraufwand und die z.T. langen Laufzeiten, werden im sechsten Kapitel einige Varianten vorgestellt, die diese Probleme umgehen. Dabei gehen natürlich einige gute Eigenschaften verloren, aber nicht alle.

Abschließend werden die in den vorangegangenen Abschnitten hergeleiteten analytischen Ergebnisse an Hand von numerischen Testproblemen überprüft (s. Kap. 7) sowie einige Bemerkungen zur Implementation der ENO–Verfahren gemacht (s. Kap. 8) .

Kapitel 1: ENO–Interpolation

1.1 Konstruktion des ENO–Splines

Grundlage der in Kapitel 2 näher beschriebenen Verfahren ist — wie schon eingangs erwähnt — eine spezielle Interpolationstechnik. Diese soll es ermöglichen, eine "stückweise glatte" Funktion so zu interpolieren, daß die Oszillation dieser Approximation die der zu interpolierenden Funktion nicht "wesentlich" übersteigt. Was man darunter zu verstehen hat, soll im folgenden erklärt werden:

Definition 1.1:

(a) Seien $a, b \in \overline{\mathbb{R}}$, $a < b$ und $k \in \mathbb{N}$.

$v : [a, b] \to \mathbb{R}$ heißt *stückweise* C^k *und Lipschitz–stetig fortsetzbar* : $\Longleftrightarrow$

Es gibt $n \in \mathbb{N}$ und $(\overline{x}_1, \ldots, \overline{x}_n) \in (a, b)^n$ mit:

(1) $a =: \overline{x}_0 < \overline{x}_1 < \ldots < \overline{x}_n < b =: \overline{x}_{n+1}$;

(2) $v_i \equiv v\big|_{(\overline{x}_i, \overline{x}_{i+1})} \in C^k,\ i = 0, \ldots, n$;

(3) v_i ist Lipschitz–stetig fortsetzbar in $\overline{x}_i$ und $\overline{x}_{i+1},\ i = 0, \ldots, n$, mit Lipschitz–Konstante $L_v\ (\ [\ \overline{x}_i, \overline{x}_{i+1}] \ )$

(4) $0 < |\ v_i(\ \overline{x}_i + 0) - v_{i-1}(\ \overline{x}_i - 0)\ | < \infty,\ i = 1, \ldots, n$.

(b) Die $\overline{x}_i,\ i = 1, \ldots, n$ sind die *Sprungstellen (endl. Höhe)* von v.

(c) $SCL(k, [a, b])$ bezeichne die Menge der stückweisen C^k und Lipschitz–stetig fortsetzbaren Funktionen.

Bei den Funktionen aus obiger Definition handelt es sich also insbesondere um solche, deren Definitionsbereich in endlich viele Bereiche zerfällt, in denen sie glatt sind. Dabei ist (3) die wesentliche Bedingung. Dies sieht man an folgendem Beispiel:

$$v(x) \ := \ \left\{ \begin{array}{c} \sqrt{x} + a \ , \ x \geq 0 \\ \sqrt{-x} \ , \ x < 0 \end{array} \right\} \qquad \text{mit } a \in \mathbb{R}\backslash\{0\}.$$

Für v gilt zwar (1) mit $n = 1$ und (2) mit $k \in \mathbb{N}$ und (4) mit $0 < |\ a\ | < \infty$, aber nicht (3), denn:

$$\lim_{x \to 0+} v_1' = +\infty, \qquad \lim_{x \to 0-} v_1' = -\infty.$$

Da im weiteren Verlauf einige Schreibweisen des öfteren verwendet werden, sollen sie vorweg erläutert werden:

(1) $x_{j,h} := j\ h \qquad j \in \mathbb{Z}\ ,\ h > 0;$

(2) Für $v : \mathbb{R} \to \mathbb{R}$, $N \in \mathbb{N}$ gelte:

(a) $I_{j,h} := [x_{j,h}, x_{j+1,h});$

(b) $Q^N_{j,h}(\circ, v) : \mathbb{R} \to \mathbb{R}$ Interpolationspolynom vom Grad N mit $N + 1$ Stützstellen, die $x_{j,h}$ und $x_{j+1,h}$ umfassen, also insb.:

$Q^N_{j,h}(x_{j,h}, v) = v(x_{j,h}) =: v_{j,h}\ ,\ Q^N_{j,h}(x_{j+1,h}, v) = v(x_{j+1,h}) =: v_{j+1,h}$

(c) $Q^{N,\mu}_{j,h}(x, v) : \mathbb{R} \to \mathbb{R}$ Interpolationspolynom zu den Stützstellen $x_{j-N+\mu,h}, \ldots, x_{j+\mu,h},\ 1 \leq \mu \leq N$

(d) $Q^N_h(x, v) := Q^N_{j,h}\big|_{I_{j,h}}(x, v) \qquad$ für $x \in I_{j,h}$.

Falls der Zusammenhang eindeutig ist, so wird der Index h vernachlässigt.
Mit Hilfe dieser Bezeichnungen lassen sich nun die ENO–Polynome und –Splines definieren:

Definition 1.2:

(a) Sei $v : \mathbb{R} \to \mathbb{R}$, $n \in \mathbb{N}$ und $h > 0$. Das *ENO–Polynom* Q_j^N , $j \in \mathbb{Z}$, ist rekursiv definiert durch:

$$Q_j^1(x,v) := v_j + (x - x_j)v[j, j+1] \quad , \; x \in \mathbb{R}$$

$$k_m^{(1)} := j, \; k_M^{(1)} := j+1.$$

Seien $Q_j^{N-1}(\circ, v)$ und $k_m^{(N-1)}$, $k_M^{(N-1)}$ schon definiert. Setze:

$$a^N := v[k_m^{(N-1)}, \ldots, \; k_M^{(N-1)} + 1] \;, \qquad b^N := v[k_m^{(N-1)} - 1, \ldots, \; k_M^{(N-1)}] \;.$$

Definiere $Q_j^N(\circ, v)$ wie folgt:

$$Q_j^N(x,v) := Q_j^{N-1}(x,v) + \overline{m}(a^N, b^N) \prod_{k=k_m^{(N-1)}}^{k_M^{(N-1)}} (x - x_k) \qquad \text{mit}$$

$$\overline{m}(a^N, b^N) := \left\{ \begin{array}{ll} a^N & , |a^N| \; < \; |b^N| \\ b^N & , \text{sonst} \end{array} \right\},$$

$$k_m^{(N)} := \left\{ \begin{array}{c} k_m^{(N-1)}, \; |a^N| < |b^N| \\ k_m^{(N-1)} - 1, \; \text{sonst} \end{array} \right\}, \qquad k_M^{(N)} := \left\{ \begin{array}{c} k_M^{(N-1)} + 1, \; |a^N| < |b^N| \\ k_M^{(N-1)}, \; \text{sonst} \end{array} \right\}.$$

(b) $Q^N(x,v) \equiv Q_h^N(x,v) := Q_j^N(x,v), \; x \in I_j, \; j \in \mathbb{Z}$ heißt *ENO–Spline.*

Da in obiger Definition die "dividierten Differenzen" eine zentrale Rolle spielen und einige ihrer Eigenschaften in der folgenden Analyse benötigt werden, soll an dieser Stelle diesbezüglich ein kleiner Exkurs eingefügt werden:

Definition 1.3:

Sei $v : \mathbb{R} \to \mathbb{R}$. Die *dividierten Differenzen* sind rekursiv definiert durch:

$$v[i] := v(x_i) \equiv v_i, \qquad i \in \mathbb{Z};$$

$$v[i, \ldots, i+k] := \frac{v[i+1, \ldots, i+k] - v[i, \ldots, i+k-1]}{x_{i+k} - x_i}, \qquad k \in \mathbb{N}, \; i \in \mathbb{Z}.$$

In [BS] wurden dazu die folgenden Eigenschaften gezeigt:

Korollar 1.1:

Vor.: $v : \mathbb{R} \to \mathbb{R}, \; i \in \mathbb{Z}, \; k \in \mathbb{N};$

Beh.: a) $v\big|_{(x_i, x_{i+k})} \in C^k \implies \exists \, \xi \in (x_i, x_{i+k}) : \; v[i, \ldots, i+k] = \frac{1}{k!} \, \left(\frac{d}{dx}\right)^k v(\xi);$

b) $v[i, \ldots, i+k] = v[j_0, \ldots, j_k]$ mit $(j_0, \ldots, j_k)$ Permutation von $(i, \ldots, i+k)$.

Die dividierten Differenzen lassen sich auch für beliebige Stützstellen x_i definieren. Da hier aber nur äquidistante verwendet werden, kann man sie auch mittels einer einfachen Summenformel darstellen:

Korollar 1.2:

Vor.: $v : \mathbb{R} \to \mathbb{R},\ k \in \mathbb{Z},\ \nu \in \mathbb{N},\ h > 0;$

Beh.: a) $v[k-\nu, \ldots, k] \;=\; \dfrac{1}{\nu!\, h^\nu} \displaystyle\sum_{j=0}^{\nu-1} \binom{\nu-1}{j} (-1)^j [\, v(x_{k-j}) - v(x_{k-j-1})\,];$

b) $v[k-\nu, \ldots, k] \;=\; \dfrac{1}{\nu!\, h^\nu} \displaystyle\sum_{j=0}^{\nu} \binom{\nu}{j} (-1)^j\ v(x_{k-j});$

c) $v\big|_{(x_{k-\nu}, x_k)} \in C^1 \implies \exists\ \xi_{k-j} \in (x_{k-j-1}, x_{k-j}),\ j = 0, \ldots, \nu-1 :$

$$v[k-\nu, \ldots, k] \;=\; \frac{1}{\nu!\, h^{\nu-1}} \sum_{j=0}^{\nu-1} \binom{\nu-1}{j} (-1)^j\ v'(\xi_{k-j}).$$

Beweis:

Ein ausführlicher Beweis findet sich in [Mü]. Die Grundschritte seien aber hier noch einmal skizziert:

(a) Vollständige Induktion über ν sowie Additionstheorem für die Binomialkoeffizienten;

(b) Folgt aus (a) ebenfalls mit dem Additionstheorem;

(c) Folgt aus (a) mit dem Mittelwertsatz.

■

Aus diesen Darstellungen lassen sich nun folgende Abschätzungen herleiten:

Folgerung 1.1:

Vor.: $a, b \in \mathbb{R},\ a < b,\ v : [a,b] \to \mathbb{R},\ h > 0,\ k \in \mathbb{Z},\ \nu \in \mathbb{N},\ (x_{k-\nu}, x_k) \subset [a,b];$

Beh.: a) $v \in C^0 \implies |\, v[k-\nu, \ldots, k]\, | \;\le\; \dfrac{2}{h^\nu} \max\limits_{x \in [a,b]} \{\, |\, v(x)\, |\, \};$

b) $v \in C^1 \implies |\, v[k-\nu, \ldots, k]\, | \;\le\; \dfrac{1}{h^{\nu-1}} \max\limits_{x \in [a,b]} \{\, |\, v'(x)\, |\, \};$

c) $\overline{x} \in (x_{k-1}, x_k)$ Sprungstelle von v gemäß Def. 1.1 $\implies$

$$|\, v[k-\nu, \ldots, k]\, | \;\le\; \frac{2}{h^\nu} \max_{x \in [a,b]} \{\, |\, v(x)\, |\, \} + \frac{1}{h^{\nu-1}}\, L_v(\, [a, \overline{x}\,]\,);$$

d) v Lipschitz–stetig auf $[a,b]$ mit Lipschitz–Konstante $L \implies$

$$|\, v[k-\nu, \ldots, k]\, | \;\le\; \tfrac{L}{h^{\nu-1}}$$

Näheres zum Beweis findet sich in [Mü].

Anwendung findet die Theorie der dividierten Differenzen in erster Linie bei der Newton–Interpolation (s. [BS], S. 43-49). Zwischen den Newton–Polynomen und den ENO–Polynomen besteht ein direkter Zusammenhang wie man leicht sieht:
Aus Def. 1.2 folgt unmittelbar, daß $Q_j^N(\circ, v)$ die Funktion v in $N+1$ aufeinanderfolgenden Punkten interpoliert, die x_j und x_{j+1} einschließen. Somit gibt es N mögliche Interpolationspolynome vom Grad N unter denen $Q_j^N(\circ, v)$ gemäß obiger Konstruktion zu suchen ist. Aus Kor. 1.1.b folgt nun, daß $Q_j^N(\circ, v)$ identisch ist mit dem Newton–Interpolationspolynom zu den Stützstellen $x_{k_m^{(N)}}, \ldots, x_{k_M^{(N)}}$ mit $N = k_M^{(N)} - k_m^{(N)}$ und $k_M^{(N)} \in \{x_{j+1}, \ldots, x_{j+N}\}$ (s. [BS], S. 43-49).

1.2 ENO–Eigenschaft für den ENO–Spline

Im folgenden soll nun das Oszillationsverhalten des ENO–Splines $Q^N(\circ, v)$ untersucht werden. Aufgrund der Konstruktion gemäß Def. 1.2 hofft man, daß sich die totale Variation von $v : \mathbb{R} \to \mathbb{R}$ durch die Interpolation nicht "wesentlich" erhöht. Was man darunter zu verstehen hat, ist in nachstehender Definition näher erklärt:

Definition 1.4:

Seien $N \in \mathbb{N},\ a, b \in \overline{\mathbb{R}},\ a < b,\ v : [a, b] \to \mathbb{R},\ h > 0$ und $Q_h^N(\circ, v)$ der gemäß Def. 1.2 konstruierte ENO–Spline. Man sagt, daß $Q_h^N(\circ, v)$ die *ENO–Bedingung* erfüllt, falls gilt:

$$\mathrm{TV}\left(Q_h^N(\circ, v)\right) \leq \mathrm{TV}(v) + O\left(h^{N+1}\right).$$

Dabei bezeichnet TV die totale Variation über $[a, b]$ der entsprechenden Funktion.

Ziel ist es nun, zu zeigen, daß der ENO–Spline für Funktionen $v \in SCL(N+1, \mathbb{R})$ diese Eigenschaft besitzt. Dabei ergeben sich beim Beweis die Schwierigkeiten aus solchen Intervallen $I_j,\ j \in \mathbb{Z}$, in denen eine Sprungstelle von v liegt, da man dort die bekannte Fehlerabschätzung für die Polynom–Interpolation (s. [BS], S. 49) nicht mehr anwenden kann. Deshalb soll zunächst das Oszillationsverhalten von $Q^N(\circ, v)$ in solchen Intervallen untersucht werden. Dazu werden vorweg vier Lemmata bewiesen, die aufgrund ihrer Bedeutung und Beweislänge aus dem eigentlichen Beweis ausgelagert sind. Die verwendeten Bezeichnungen werden im wesentlichen analog zu [ENO2] gewählt, um auf Parallelen zu dieser Arbeit leichter verweisen zu können und auf einige dort vorhandene Fehler aufmerksam zu machen.

Lemma 1.1:

Vor.: $N \in \mathbb{N},\ v \in SCL(N+1, \mathbb{R})$;
Beh.: Es gibt Funktionen $H, w : \mathbb{R} \to \mathbb{R}$ mit den folgenden Eigenschaften:
(a) H stückweise konstant,
(b) w Lipschitz–stetig auf $\mathbb{R}$,
(c) $v \equiv w + H$.

Beweis:

Da nach Voraussetzung $v \in SCL(N+1, \mathbb{R})$ ist, gibt es $n \in \mathbb{N}$ und $\overline{x}_1 < \ldots < \overline{x}_n$ derart, daß v in $\overline{x}_i,\ i = 1, \ldots, n$, einen Sprung endlicher Höhe hat. Setze nun

$$(1) \qquad c_i := v(\overline{x}_i + 0) - v(\overline{x}_i - 0) \qquad i = 1, \ldots, n;$$

$$(2) \qquad \overline{x}_0 := -\infty,\ \overline{x}_{n+1} := +\infty;$$

$$(3) \qquad H(x) := \left\{ \begin{array}{ll} \sum_{j=i+1}^{n} c_j & , \overline{x}_i < x \le \overline{x}_{i+1},\ i = 0, \ldots, n-1 \\ & \\ 0 & , \ \overline{x}_n < x \end{array} \right\};$$

$$(4) \qquad w(x) := v(x) - H(x), \qquad x \in \mathbb{R}.$$

Daraus folgen unmittelbar (a) und (c). Zum Nachweis der Lipschitz–Stetigkeit setze

$$(5) \qquad L_w(\,\mathbb{R}\,) := \max_{i=0,\ldots,n} \{\, L_v(\,[\,\overline{x}_i, \overline{x}_{i+1}]\,)\,\}\,.$$

Sei nun $i \in \{1, \ldots, n\}$ beliebig, aber fest gewählt. Nach Voraussetzung gilt :

$$(6) \qquad v\big|_{(\,\overline{x}_{i-1}, \overline{x}_i)} \in C^1.$$

Ferner seien $x, y \in \mathbb{R}$ und o.B.d.A $x < y$. Führe nun folgende Fallunterscheidung durch:

a) $x, y \in (\,\overline{x}_{i-1}, \overline{x}_i)$:

In diesem Fall liegen zwischen x und y keine Sprungstellen von v. Mit (3) und (4) erhält man dann sofort:

$$|w(x) - w(y)| \;=\; |v(x) - v(y)| \;\le\; L_v(\,[\,\overline{x}_{i-1}, \overline{x}_i]\,)\ |x - y| \;\le\; L_w(\,\mathbb{R}\,)\ |x - y|.$$

b) $x \in (\,\overline{x}_i, \overline{x}_{i+1}),\ y \in (\,\overline{x}_j, \overline{x}_{j+1}),\ i < j - 1$:

Zwischen x und y liegt nun mindestens eine Sprungstelle von v. Aus (1),...,(4) folgt wiederum:

$$\begin{aligned} |\ w(x) - w(y)\ | &\overset{(3)}{=} \left|\ v(x) \;-\; \sum_{l=i+1}^{n} c_l \;-\; v(y) \;+\; \sum_{l=j+1}^{n} c_l\ \right| \\ &= \left|\ v(x) \;+\; \sum_{l=i+1}^{j} c_l \;-\; v(y)\ \right| \\ &\overset{(1)}{=} \left|\ v(x) \;-\; \sum_{l=i+1}^{j} [\, v(\overline{x}_l + 0) - v(\overline{x}_{l+1} - 0)\,] \;-\; v(y)\ \right| \end{aligned}$$

$$\leq \; | \, v(x) - v(\overline{x}_{i+1} - 0) \, | \; + \sum_{l=i+1}^{j-1} | \, v(\overline{x}_l + 0) - v(\overline{x}_{l+1} - 0) \, |$$

$$+ \; | \, v(\overline{x}_j + 0) - v(y) \, |$$

$$\leq \; L_v(\, [x, \overline{x}_{i+1}] \,) \; | \, x - \overline{x}_{i+1} \, | \; + \sum_{l=i+1}^{j-1} L_v(\, [\, \overline{x}_l, \overline{x}_{l+1}] \,) \; | \, \overline{x}_l - \overline{x}_{l+1} \, |$$

$$+ \; L_v(\, [\, \overline{x}_j, \overline{x}_{j+1}] \,) \; | \, \overline{x}_j - y \, |$$

$$\overset{(5)}{\leq} L_w(\, \mathbb{R} \,) \left(| \, x - \overline{x}_{i+1} \, | + \sum_{l=i+1}^{j-1} | \, \overline{x}_l - \overline{x}_{l+1} \, | + | \, \overline{x}_j - y \, | \right)$$

$$= \; L_w(\, \mathbb{R} \,) \; | \, x - y \, |.$$

Aus diesen beiden Fällen ergibt sich unmittelbar die Eigenschaft b).

■

Man beachte, daß obiger Beweis mit der Tatsache, daß $v|_{(\,\overline{x}_i, \overline{x}_{i+1})}$ in $\overline{x}_i, \overline{x}_{i+1}$ Lipschitz–stetig fortsetzbar ist, steht und fällt. In [ENO2] wurde diese Voraussetzung nicht verlangt. Daher wurde die im Anschluß an Def. 1.1 definierte Funktion zum Beispiel nicht ausgeschlossen. Zu ihr lassen sich aber — wie man leicht einsieht — keine Funktionen w und H finden, so daß die Eigenschaft b) aus La. 1.1 erfüllt ist. Es können nun einige — später sehr wichtige — Abschätzungen für w und H bewiesen werden. Dabei bezeichnen $Q_j^N(\circ, w)$ bzw. $Q_j^N(\circ, H)$ im folgenden *nicht* die gemäß Def. 1.2 definierten ENO–Polynome zu w bzw. H, sondern die durch das ENO–Polynom $Q_j^N(\circ, v)$ aufgrund der Linearität der Interpolationspolynome festgelegten Interpolierenden, d.h.:

$$Q_j^N(\circ, v) \equiv Q_j^N(\circ, w + H) = Q_j^N(\circ, w) + Q_j^N(\circ, H).$$

Lemma 1.2:

Vor.: $N \in \mathbb{N}, j \in \mathbb{Z}$ und v, w, H gemäß La. 1.1;
Beh.: Es gibt $h_0 > 0, c > 0$, so daß für alle $0 < h \leq h_0$ gilt:

Liegt in $I_{j,h}$ eine Sprungstelle von v, so gilt:

(i) $\quad Q_{j,h}^N(\circ, H)$ ist streng monoton auf $I_{j,h}$;

(ii) $$\min_{x \in I_{j,h}} \left\{ \left| \tfrac{d}{dx} Q_{j,h}^N(x, H) \right| \right\} > \frac{c}{h}.$$

Beweis:

Da H nach Voraussetzung stückweise konstant ist, gibt es $n \in \mathbb{N}$ und $\overline{x}_1 < \ldots < \overline{x}_n$, so daß H nur in $\overline{x}_i, i = 1, \ldots, n$ einen Sprung endlicher Höhe hat. Setze

$$(1) \qquad h_0 := \frac{1}{N+1} \quad \min_{i=1,\ldots,n} \{ \mid \overline{x}_{i+1} - \overline{x}_i \mid \}.$$

Sei nun $0 < h \leq h_0$ beliebig, aber fest gewählt und $j = j(h) \in \mathbb{Z}$, so daß in $I_{j,h}$ (genau) eine Sprungstelle $\overline{x} \in \{\overline{x}_1, \ldots, \overline{x}_n\}$ von H liegt. Dann läßt sich $Q^N_{j,h}(\circ, H)$ wie folgt schreiben:

$$(2) \qquad Q^N_{j,h}(x, H) = \sum_{\nu=k_M^{(N)}}^{N} H[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] \prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} (x - x_l), \qquad x \in \mathbb{R}.$$

Wegen (1) liegen zwischen zwei Sprungstellen mindestens N Intervalle der Länge h_0. Somit ist $H = const$ auf $(\overline{x}, x_{k_M^{(N)}})$ und es gilt für $\nu = 0, \ldots, k_M^{(N)} - j - 1$ mit Kor. 1.1.a:

$$H[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] \;=\; \frac{1}{\nu!} \left(\tfrac{d}{dx}\right)^{\nu} H(\xi_\nu) \;=\; 0,$$
$$\nu = 1, \ldots, k_M^{(N)} - j - 1, \quad \xi_\nu \in (x_{K_M^{(N)}-\nu}, x_{K_M^{(N)}}).$$

Demzufolge läßt sich (2) auch schreiben als:

$$(3) \qquad Q^N_{j,h}(x, H) = H(x_{k_M^{(N)}}) + \sum_{\nu=k_M^{(N)}-j}^{N} H[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] \prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} (x - x_l).$$

Damit kann man nun (i) folgendermaßen zeigen:
Für $N = 1$ ist (i) trivial, da $Q^1_{j,h}(\circ, H)$ eine nicht konstante affine Funktion ist, denn: $\overline{x} \in I_{j,h}$ und H stückweise konstant. Sei nun $N > 1$. Dann gilt wegen (1):

(a) $\qquad H(x_{j-\nu-1}) = H(x_{j-\nu}) = const \qquad$ für $\quad \nu = 0, \ldots, j - k_M^{(N)} + 1;$

(b) $\qquad H(x_{j+\nu+1}) = H(x_{j+\nu}) = const \qquad$ für $\quad \nu = 1, \ldots, k_M^{(N)} - j - 1.$

Also gibt es $(j - k_m^{(N)} + 1) + (k_M^{(N)} - j - 2) = k_M^{(N)} - k_m^{(N)} - 1 = N - 1$ Intervalle, in denen die Funktionswerte von $Q^N_{j,h}(\circ, H)$ in den Intervallgrenzen übereinstimmen. Der Satz von Rolle besagt nun, daß $\frac{d}{dx} Q^N_{j,h}(\circ, H)$ außerhalb von $[x_j, x_{j+1}]$ $N-1$ Nullstellen hat. Da $\frac{d}{dx} Q^N_{j,h}(\circ, H) \in \mathbb{P}_{N-1}$, gilt mit dem Fundamentalsatz der Algebra, daß $\frac{d}{dx} Q^N_{j,h}(\circ, H)$ in $[x_j, x_{j+1}]$ verschieden von Null ist — $\overline{x} \in I_{j,h}$ und H stückweise konstant — also

$$\tfrac{d}{dx} Q^N_{j,h}(x, H) > 0 \quad \text{bzw.} \quad \tfrac{d}{dx} Q^N_{j,h}(x, H) < 0 \quad \text{für} \quad x \in [x_j, x_{j+1}].$$

Somit ist (i) gezeigt.
Sei nun weiter $\tilde{j} = \tilde{j}(h_0) \in \mathbb{Z}$ so gewählt,daß $\overline{x} \in I_{\tilde{j}}$. Setze

$$\nu := k_M^{(N)} - j, \quad \tilde{k}_M^{(N)} := k_M^{(N)} - j + \tilde{j}, \quad \tilde{x}_i := i\, h_0, \; i \in \mathbb{Z}.$$

Zeige nun:

(4)
$$Q^N_{j,h}(x,H) = Q^N_{\tilde{j},h_0}(\tilde{x},H), \qquad x \in [x_j, x_{j+1}]$$
$$\text{mit} \quad \tilde{x} := (\tfrac{x}{h} + \tilde{j} - j)h_0.$$

<u>Dazu:</u>

$Q^N_{j,h}(\circ, H)$ ist gemäß Def. 1.2 ein Polynom vom Grad N, das H in $x_{k_m^{(N)}}, \ldots, x_{k_M^{(N)}}$ interpoliert. Mit obigen Bezeichnungen gilt nun:

(5)
$$H(x_{k_M^{(N)}-l}) = H(\tilde{x}_{\tilde{k}_M^{(N)}-l}), \quad l \in \{0, \ldots, N\},$$

denn:

(a) Für $l \in \{0, \ldots, k_M^{(N)} - j - 1\}$ gilt:
$x_{k_M^{(N)}-l} \geq x_{j+1} > \overline{x}$ und
$\tilde{x}_{\tilde{k}_M^{(N)}-l} = (k_M^{(N)} - j + \tilde{j} - l)h_0 \geq (\tilde{j}+1)h_0 = \tilde{x}_{\tilde{j}+1} > \overline{x}.$

(b) Für $l \in \{k_M^{(N)} - j, \ldots, N\}$ gilt:
$x_{k_M^{(N)}-l} \leq x_j < \overline{x}$ und
$\tilde{x}_{\tilde{k}_M^{(N)}-l} = (k_M^{(N)} - j + \tilde{j} - l)h_0 \leq \tilde{j}\; h_0 = \tilde{x}_{\tilde{j}} < \overline{x}.$

Wegen (1) ist $\overline{x}$ die einzige Sprungstelle von H in $(\tilde{x}_{\tilde{k}_m^{(N)}}, \tilde{x}_{\tilde{k}_M^{(N)}})$ bzw. $(x_{k_m^{(N)}}, x_{k_M^{(N)}})$. Somit gilt (5), da H links und rechts von $\overline{x}$ konstant ist.

Ersetze nun in (3) die dividierten Differenzen durch die Summendarstellung aus Kor. 1.2.b. Damit ergibt sich für $x \in \mathbb{R}$:

(6)
$$Q^N_{j,h}(x,H) = \sum_{\nu=k_M^{(N)}-j}^{N} \sum_{l=0}^{\nu} \binom{\nu}{l}(-1)^l H(x_{k_M^{(N)}-l}) \frac{1}{\nu!\; h^\nu} \prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} (x - x_l)$$
$$+ H(x_{k_M^{(N)}}).$$

Das Produkt aus der Darstellung (6) läßt sich nun wie folgt transformieren:

$$\prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} (x - x_l) = \left(\frac{h}{h_0}\right)^\nu \prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} \left[\left(\tfrac{x}{h} + \tilde{j} - j\right) h_0 - (l + \tilde{j} - j)\, h_0 \right]$$
$$= \left(\frac{h}{h_0}\right)^\nu \prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} (\tilde{x} - \tilde{x}_{l+\tilde{j}-j})$$
$$\text{für} \quad \nu = k_M^{(N)} - j, \ldots, N.$$

Setzt man dies in (6) ein, so erhält man unter Berücksichtigung von (5):

$$Q^N_{j,h}(x,H) = \sum_{\nu=\tilde{k}^{(N)}_M - \tilde{\jmath}}^{N} \sum_{l=0}^{\nu} \binom{\nu}{l}(-1)^l H(\tilde{x}_{\tilde{k}^{(N)}_M - l}) \frac{1}{\nu! h^\nu} \prod_{l=\tilde{k}^{(N)}_M-\nu+1}^{\tilde{k}^{(N)}_M} (\tilde{x}-\tilde{x}_l)$$

$$+ H(\tilde{x}_{\tilde{k}^{(N)}_M}).$$

Durch Anwendung von Kor. 1.2.b erhält man somit:

$$Q^N_{j,h}(x,H) = H(\tilde{x}_{\tilde{k}^{(N)}_M}) + \sum_{\nu=\tilde{k}^{(N)}_M - \tilde{\jmath}}^{N} H[\tilde{k}^{(N)}_M - \nu, \dots, \tilde{k}^{(N)}_M] \prod_{l=\tilde{k}^{(N)}_M-\nu+1}^{\tilde{k}^{(N)}_M} (\tilde{x}-\tilde{x}_l),$$

also gilt:

$$Q^N_{j,h}(x,H) = Q^{N,\nu}_{\tilde{\jmath},h_0}(\tilde{x},H).$$

Aus (4) folgt nun sofort:

$$(7) \qquad \tfrac{d}{dx} Q^N_{j,h}(x,H) = \tfrac{d}{dx} Q^{N,\nu}_{\tilde{\jmath},h_0}\left((\tfrac{x}{h}+\tilde{\jmath}-j)h_0, H\right) = \tfrac{h_0}{h}\ \tfrac{d}{dx} Q^{N,\nu}_{\tilde{\jmath},h_0}(z,H)\Big|_{z=\tilde{x}}$$

Mit Teil (i) der Behauptung, der schon bewiesen wurde, gilt insbesondere:

$$Q^{N,\nu}_{\tilde{\jmath},h_0}(\tilde{x},H) \neq 0 \quad \text{für} \quad \tilde{x} \in [\tilde{x}_{\tilde{\jmath}}, \tilde{x}_{\tilde{\jmath}+1}].$$

Also gibt es $c(\overline{x},h_0) > 0$ mit

$$(8) \qquad \min_{\tilde{x}\in[\tilde{x}_{\tilde{\jmath}},\tilde{x}_{\tilde{\jmath}+1}]} \left\{ \left| \ \tfrac{d}{dx} Q^{N,\nu}_{\tilde{\jmath},h_0}(x,H)\Big|_{x=\tilde{x}} \ \right| \right\} \ > \ c(\overline{x},h_0).$$

Für $x \in [x_j, x_{j+1}]$ gilt ferner:

$$(\tfrac{x}{h}+\tilde{\jmath}-j)h_0 \in [\tilde{x}_{\tilde{\jmath}}, \tilde{x}_{\tilde{\jmath}+1}].$$

Mit (7) und (8) erhält man mit $\tilde{c}(\overline{x},h_0) := h_0\ c(\overline{x},h_0)$:

$$\min_{x\in[x_j,x_{j+1}]} \left\{ \left| \tfrac{d}{dx} Q^N_{j,h}(x,H) \right| \right\} = \frac{h_0}{h} \min_{x\in[\tilde{x}_{\tilde{\jmath}},\tilde{x}_{\tilde{\jmath}+1}]} \left\{ \left| \tfrac{d}{dx} Q^{N,\nu}_{\tilde{\jmath},h_0}(x,H) \right| \right\} > \frac{\tilde{c}(\overline{x},h_0)}{h}.$$

Der Beweis gilt für jede Sprungstelle von H, da $\overline{x}$ beliebig gewählt war.
Sei nun $j = j(i,h) \in \mathbb{Z}$ mit $\overline{x}_i \in I_j$ für $i = 1,\dots,n$. Dann gilt die Behauptung (ii) mit

$$c := \min\{\tilde{c}(\overline{x}_i,h_0) \big| i = 1,\dots,n\}.$$

■

Als nächstes soll nun für die Funktion w aus La. 1.1 eine ähnliche Abschätzung hergeleitet werden:

Lemma 1.3:

Vor.: $N \in \mathbb{N}, j \in \mathbb{Z}$ und v, w, H gemäß La. 1.1;
Beh.: Es gibt $h_0 > 0, c > 0$, so daß für alle $0 < h \leq h_0$ gilt:

Liegt in $I_{j,h}$ genau eine Sprungstelle von v, so gilt:

$$\left| \tfrac{d}{dx} Q^N_{j,h}(\circ, w) \right| < c, \qquad \forall\, x \in I_{j,h}.$$

Beweis:

Seien wieder $n \in \mathbb{N}$ und $\overline{x}_1 < \ldots < \overline{x}_n$ die Sprungstellen von v. Setze

$$h_0 := \frac{1}{N+1} \min_{i=1,\ldots,n} \{ \mid \overline{x}_{i+1} - \overline{x}_i \mid \}.$$

Dann liegt in N aufeinanderfolgenden Intervallen der Länge $\leq h_0$ höchstens eine Unstetigkeit von v. Sei nun $\overline{x}$ eine beliebige Sprungstelle von v und $0 < h \leq h_0$. Wähle $j = j(h) \in \mathbb{Z}$ mit $\overline{x} \in I_{j,h}$. Aus Kor. 1.2 folgt mit La. 1.1 unmittelbar:

$$v[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] = w[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] + H[k_M^{(N)} - \nu, \ldots, k_M^{(N)}], \qquad \nu = 0, \ldots, N.$$

Aus dem Beweis zu La. 1.1 geht ferner hervor, daß

$$H[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] = 0 \qquad \text{für} \quad \nu = 1, \ldots, k_M^{(N)} - j - 1.$$

Aufgrund der Linearität der Polynominterpolation gilt nun für $x \in \mathbb{R}$:

$$\begin{aligned} Q^N_{j,h}(x,w) = w(x_{k_M^{(N)}}) &+ \sum_{\nu=k_M^{(N)}-j}^{N} w[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] \prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} (x - x_l) \\ &+ \sum_{\nu=1}^{k_M^{(N)}-j-1} v[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] \prod_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} (x - x_l). \end{aligned} \tag{1}$$

Daraus ergibt sich unmittelbar

$$\begin{aligned} \tfrac{d}{dx} Q^N_{j,h}(x,w) = &\sum_{\nu=k_M^{(N)}-j}^{N} w[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] \sum_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} \prod_{\substack{s=k_M^{(N)}-\nu+1 \\ s \neq l}}^{k_M^{(N)}} (x - x_s) \ + \\ &\sum_{\nu=1}^{k_M^{(N)}-j-1} v[k_M^{(N)} - \nu, \ldots, k_M^{(N)}] \sum_{l=k_M^{(N)}-\nu+1}^{k_M^{(N)}} \prod_{\substack{s=k_M^{(N)}-\nu+1 \\ s \neq l}}^{k_M^{(N)}} (x - x_s). \end{aligned} \tag{2}$$

Setze nun $a := \overline{x} - Nh_0,\ b := \overline{x} + Nh_0$. Dann gilt insbesondere: $[x_{k_m^{(N)}}, x_{k_M^{(N)}}] \subset [a,b]$. Da w Lipschitz–stetig nach La. 1.1, ist auch $w\big|_{[a,b]}$ Lipschitz–stetig. Sei $L_w = L_w\left([a,b]\right)$ die zugehörige Lipschitz–Konstante. Mit Folg. 1.1.d gilt dann:

$$(3) \qquad w[k_M^{(N)} - \nu, \dots, k_M^{(N)}] \le \frac{L_w}{h^{\nu-1}} \qquad \text{für} \quad \nu = k_M^{(N)} - j, \dots, N.$$

Da v Lipschitz–stetig fortsetzbar in $\overline{x}$, gilt mit Folg. 1.1.b:

$$(4) \qquad v[k_M^{(N)} - \nu, \dots, k_M^{(N)}] \le \frac{L_v\left(\,[\,\overline{x}, b]\,\right)}{h^{\nu-1}} \qquad \text{für} \quad \nu = 1, \dots, k_M^{(N)} - j - 1.$$

Eine weitere Abschätzung liefert für die Produkte aus (2):

$$(5) \qquad x \in I_{j,h} \quad \Longrightarrow \quad \prod_{\substack{s=k_M^{(N)}-\nu+1 \\ s \neq l}}^{k_M^{(N)}} \left|\, (x - x_s)\,\right| \ \le (N\,h)^{\nu-1}.$$

Aus (2) folgt nun mit (3),(4) und (5):

$$\left|\, \tfrac{d}{dx} Q_{j,h}^N(x,w)\,\right| \quad \le \quad c_{k_M^{(N)}} \qquad \text{für} \quad x \in I_{j,h}$$

mit

$$c_{k_M^{(N)}} \quad := \quad \sum_{\nu=k_M^{(N)}-j}^{N} L_w\ \nu\ N^\nu \ + \sum_{\nu=1}^{k_M^{(N)}-j-1} L_v\left(\,[\,\overline{x}, b]\,\right)\ \nu\ N^\nu.$$

Da $k_M^{(N)} \in \{j+1, \dots, j+N\}$ (vgl. Def. 1.2) in Abhängigkeit von h, bilde noch das Maximum über die Konstanten $c_{k_M^{(N)}}$, d.h.:

$$c(\overline{x}) := \max\left\{c_l \;\middle|\; l \in \{j+1, \dots, j+N\}\right\}.$$

Dies ist notwendig, da im allgemeinen gilt:

$$h_1 < h_2 \le h_0 \qquad \Longrightarrow \qquad k_M^{(N)}(h_1) \quad \neq \quad k_M^{(N)}(h_2).$$

Da $c(\overline{x})$ insbesondere von $\overline{x}$ abhängt, setze noch:

$$c := \max_{i=1,\dots,n} \{c(\overline{x}_i)\}.$$

Mit dieser Konstante gilt dann die Behauptung.

■

In den vorhergegangenen Lemmata wurden nur solche Intervalle I_j betrachtet, in denen eine Sprungstelle von v lag. Jetzt soll eine interessante Aussage über solche Intervalle gezeigt werden, in denen *keine* Sprungstelle von v liegt:

Lemma 1.4:

Vor.: $N \in \mathbb{N}, v \in SCL(N+1, \mathbb{R})$;

Beh.: Es gibt $h_0 > 0$, so daß für alle $0 < h \leq h_0$ gilt:

Liegt in $I_{j,h},\ j \in \mathbb{Z}$, keine Sprungstelle von v, so auch nicht im zugehörigen Interpolationsintervall $[x_{k_m^{(N)}}, x_{k_M^{(N)}}]$ von $Q_{j,h}^N(\circ, v)$.

Beweis:

Da nach Voraussetzung $v \in SCL(N+1, \mathbb{R})$ ist, gibt es $n \in \mathbb{N}$ und $\overline{x}_1 < \ldots < \overline{x}_n$, so daß v nur in $\overline{x}_i, i = 1, \ldots, n$ einen Sprung hat. Setze

$$(1) \qquad \tilde{h}_0 := \frac{1}{N+1} \min_{i=1,\ldots,n} \{ \mid \overline{x}_{i+1} - \overline{x}_i \mid \} .$$

Für $h > 0$ bezeichne $\overline{k}_i = \overline{k}_i(h)$ das Intervall mit $\overline{x}_i \in I_{\overline{k}_i},\ i = 1, \ldots, n$. Sei nun $h < \tilde{h}_0$ und $j \in \mathbb{Z}$ beliebig. Liegen zwischen I_j und $I_{\overline{k}_i},\ i = 1, \ldots, n$ mehr als $N-1$ Intervalle der Länge h, dann gilt wegen Def. 1.2 trivialerweise, daß im zugehörigen Interpolationsintervall keine Sprungstelle von v liegt. Ist dies nicht der Fall, so könnte es passieren, daß eine Sprungstelle $\overline{x} \in \{\overline{x}_1, \ldots, \overline{x}_n\}$ mit in das Interpolationsintervall aufgenommen wird. Damit dies nicht geschieht, muß für $k = 1, \ldots, N-1$ gelten (vgl. Def. 1.2):

a) $x_{\overline{k}-N+1} \leq x_j \leq x_{\overline{k}} \implies \mid v[\, \overline{k}-k, \ldots, \overline{k}+1 \,] \mid > \mid v[\, \overline{k}-k-1, \ldots, \overline{k} \,] \mid$;

b) $x_{\overline{k}+1} \leq x_j \leq x_{\overline{k}+N-1} \implies \mid v[\, \overline{k}, \ldots, \overline{k}+k+1 \,] \mid > \mid v[\, \overline{k}+1, \ldots, \overline{k}+k+2 \,] \mid$.

Im folgenden soll nun der Fall a) gezeigt werden, unter der Voraussetzung, daß h klein genug ist:

Da v in $\overline{x}$ einen Sprung endlicher Höhe hat und in einer punktierten Umgebung stetig differenzierbar ist, gibt es $\tilde{c} = \tilde{c}(\overline{x}) > 0$ und $\varepsilon = \varepsilon(\overline{x}) > 0$ mit

$$(1) \qquad \forall\ x \in [\, \overline{x} - \varepsilon, \overline{x}\,),\ y \in [\, \overline{x}, \overline{x} + \varepsilon] : \ \mid v(x) - v(y) \mid > \tilde{c}.$$

Setze $a := \overline{x} - N\,\varepsilon,\ b := \overline{x}$. Sei ferner $0 < h < \min\{\varepsilon(\overline{x}), \tilde{h}_0\}$ und $k \in \{1, \ldots, N-1\}$. Mit Kor. 1.2.a sowie Folg. 1.1 gilt, da $v \in C^1$ auf $(x_{\overline{k}-k}, x_{\overline{k}}) \subset [a, \overline{x}\,]$:

$$(2) \qquad v[\, \overline{k}-k, \ldots, \overline{k}+1 \,] = \frac{\beta}{(k+1)!h^{k+1}} + \frac{\alpha}{h^k}$$

mit

$$\alpha := \sum_{j=1}^{k} \binom{k}{j}(-1)^j \tfrac{1}{(k+1)!}\, v'(\xi_{\overline{k}+1-j}), \qquad \xi_{\overline{k}+1-j} \in (x_{\overline{k}+1-j}, x_{\overline{k}+2-j});$$

$$\beta := v(x_{\overline{k}+1}) - v(x_{\overline{k}}).$$

Da insbesondere $h < \varepsilon$ und $\overline{x} \in [x_{\overline{k}}, x_{\overline{k}+1}] \subset [\overline{x} - \varepsilon, \overline{x} + \varepsilon]$ gilt:

$$(3) \qquad \mid \beta \mid \quad > \quad \tilde{c}.$$

Setze $c := L_v(\,[a, \overline{x}\,]\,)$. Dann gilt weiter:

(4) $$|\,\alpha\,| \quad \leq \quad c.$$

Aus Folg. 1.1.b ergibt sich außerdem:

(5) $$|\, v[\,\overline{k} - k - 1, \ldots, \overline{k}\,]\,| \quad \leq \quad \frac{c}{h^k}.$$

Setze nun $h_0(\overline{x}) := \min\left\{\varepsilon, \tilde{h}_0, \frac{\tilde{c}}{2cN!}\right\}$. Dann läßt sich folgende Aussage zeigen:

(6) $$\forall\; 0 < h \leq h_0(\overline{x}) : \quad |\, v[\,\overline{k} - k, \ldots, \overline{k} + 1\,]\,| > c\, h^{-k}.$$

Dazu:

Sei $h \leq h_0(\overline{x})$. Dann gilt insbesondere für $k = 1, \ldots, N-1$:

(7) $$h \quad \leq \quad \frac{\tilde{c}}{2\,c\,(k+1)!}.$$

Der Beweis wird nun mittels einer Fallunterscheidung geführt:

(a) $\beta > \tilde{c}(\overline{x})$:

Aus (2) folgt unmittelbar:

$$v[\,\overline{k} - k, \ldots, \overline{k} + 1\,] > \frac{\tilde{c}}{(k+1)!\; h^{k+1}} - \frac{c}{h^k} = h^{-k}\;\left[\frac{\tilde{c}}{(k+1)!\; h} - c\right].$$

Mit (7) gilt dann:

$$v[\,\overline{k} - k, \ldots, \overline{k} + 1\,] > h^{-k}\; c > 0$$

und somit auch

$$|\, v[\,\overline{k} - k, \ldots, \overline{k} + 1\,]\,| > h^{-k}\; c > 0.$$

(b) $\beta < -\tilde{c}(\overline{x})$:

Aus (2) folgt wiederum:

$$v[\,\overline{k} - k, \ldots, \overline{k} + 1\,] < -\frac{\tilde{c}}{(k+1)!\; h^{k+1}} + \frac{c}{h^k} = h^{-k}\;\left[c - \frac{\tilde{c}}{(k+1)!\; h}\right].$$

Mit (7) gilt dann ebenfalls:

$$v[\,\overline{k} - k, \ldots, \overline{k} + 1\,] < -h^{-k}\; c < 0$$

und somit

$$|\, v[\,\overline{k} - k, \ldots, \overline{k} + 1\,]\,| > h^{-k}\; c > 0.$$

Aus (5) und (6) folgt somit für $k \in \{1, \ldots, N-1\}$:

$$\forall\; 0 < h \leq h_0(\overline{x}) : \quad |\, v[\,\overline{k} - k, \ldots, \overline{k} + 1\,]\,| > |\, v[\,\overline{k} - k - 1, \ldots, \overline{k}\,]\,|\;.$$

Da h_0 von $\overline{x} \in \{\overline{x}_1, \ldots, \overline{x}_n\}$ abhängt, v aber n Sprungstellen hat, geht man zum Minimum dieser $h_0(\overline{x}_i)$ über . Setze also $h_L := \min\left\{h_0(\overline{x}_i)\;\middle|\; 1 \leq i \leq n\right\}$. Dann gilt a) für beliebige Sprungstelle $\overline{x}_i$ mit $0 < h \leq h_L$.

Zum Beweis von Fall b) geht man analog vor, in dem man a und b sowie h_L durch $\tilde{a} := \overline{x}$ bzw. $\tilde{b} := \overline{x} + \varepsilon N$ und $h_R := \min\left\{\tilde{h}_0(\overline{x}_i)\;\middle|\; 1 \leq i \leq n\right\}$ ersetzt. Dann erhält man die Aussage b) für $0 < h \leq h_R$.

Aus a),b) folgt mit Def. 1.2 für $Q^N_{j,h}(\circ, v)$ die Behauptumg mit $h_0 := \min\{h_L, h_R\}$.

∎

Nachdem alle vorbereitenden Lemmata gezeigt wurden, läßt sich nun der zentrale Satz dieses Kapitels beweisen:

Satz 1.1: (Lokale Oszillation)

Vor.: $N \in \mathbb{N}, j \in \mathbb{Z}$ und $v \in SCL(N+1, \mathbb{R})$;
Beh.: Es gibt $h_0 > 0$, so daß für alle $0 < h \leq h_0$ gilt:

a) Liegt in $I_{j,h}$ keine Sprungstelle von $v \Longrightarrow$

$$\mathrm{TV}\Big(Q^N_{j,h}(\circ, v)\big|_{I_{j,h}}\Big) \quad \leq \quad \mathrm{TV}\,\Big(v\big|_{I_{j,h}}\Big) \;+\; O\left(h^{N+1}\right);$$

b) Liegt in $I_{j,h}$ genau eine Sprungstelle von $v \Longrightarrow$

(i) $Q^N_{j,h}(\circ, v)\big|_{I_{j,h}}$ ist streng monoton,

(ii) $\mathrm{TV}\Big(Q^N_{j,h}(\circ, v)\big|_{I_{j,h}}\Big) \quad \leq \quad \mathrm{TV}\,\Big(v\big|_{I_{j,h}}\Big)$.

Beweis:

Da $v \in SCL(N+1, \mathbb{R})$ ist, gibt es $n \in \mathbb{N}$ und Sprungstellen $\overline{x}_1 < \ldots < \overline{x}_n$ von v. Setze

$$h_1 := \frac{1}{N+1} \quad \min_{i=1,\ldots,n} \{\, |\ \overline{x}_{i+1} - \overline{x}_i\ |\, \}\,.$$

Dann liegt in N aufeinanderfolgenden Intervallen der Länge $\leq h_1$ höchstens eine Sprungstelle.

Aus La. 1.1 folgt ferner, daß es eine Lipschitz–stetige Funktion w sowie eine stückweise konstante Funktion H gibt mit $v \equiv w + H$. Für diese gelten mit La. 1.2–La. 1.4:

(1) $\exists\, h_2 > 0,\ \forall\ 0 < h \leq h_2$:

Liegt in $I_{j,h}$ für ein $j \in \mathbb{Z}$ keine Sprungstelle von v, so auch nicht in dem zugehörigen Interpolationsintervall $[x_{k^{(N)}_m}, x_{k^{(N)}_M}]$ von $Q^N_{j,h}$;

(2) $\exists\, h_3 > 0\,, c(h_3) > 0,\ \forall\ 0 < h \leq h_3$:

Liegt in $I_{j,h}$ für ein $j \in \mathbb{Z}$ eine Sprungstelle von v, so gilt:

$$\left|\ \tfrac{d}{dx} Q^N_{j,h}(\circ, w)\ \right| \quad < \quad c(h_3) \qquad \forall\, x \in I_{j,h};$$

(3) $\exists\, h_4 > 0\,, c(h_4) > 0,\ \forall\ 0 < h \leq h_4$:

Liegt in $I_{j,h}$ für ein $j \in \mathbb{Z}$ eine Sprungstelle von $v \Longrightarrow$

(a) $Q^N_{j,h}(\circ, H)$ ist streng monoton auf $I_{j,h}$;

(b) $\min\, \left\{ \left|\ \tfrac{d}{dx} Q^N_{j,h}(\circ, H)\ \right| \ \middle|\ x \in I_{j,h} \right\} \quad > \quad \frac{c(h_4)}{h}$.

Setze nun $h_0 := \min\, \left\{ h_1, h_2, h_3, h_4, \frac{c(h_4)}{c(h_3)} \right\}$. Dann gelten für $h \in (0, h_0]$ insbesondere (1),(2),(3). Sei also im folgenden $0 < h \leq h_0$ beliebig, aber fest und $j \in \mathbb{Z}$. Betrachte zunächst den Fall, daß in I_j keine Sprungstelle von v liegt:

Zu a:

Da insbesondere $h \leq h_2$, gilt mit (1), daß im Interpolationsintervall $[x_{k_m^{(N)}}, x_{k_M^{(N)}}]$ von $Q_j^N(\circ, v)$ ebenfalls keine Sprungstelle liegt. Nach Voraussetzung ist $v \in SCL(N+1, \mathbb{R})$. Somit gilt:

$$v\big|_{[x_{k_m^{(N)}}, x_{k_M^{(N)}}]} \in C^{N+1}.$$

Aufgrund der Darstellung des Interpolationsfehlers (vgl. [BS], S. 49) folgt also:

$$Q_j^N(x, v) \;=\; v(x) \;+\; O\left(h^{N+1}\right) \qquad \text{für} \quad x \in I_j.$$

Da $Q_j^N(\circ, v) \in \mathbb{P}_N$, gibt es höchstens $N-1$ Extremstellen, denn:

Andernfalls hätte $\frac{d}{dx}Q_j^N(\circ, v)$ mehr als $N-1$ Nullstellen und wegen dem Fundamentalsatz der Algebra müßte $\frac{d}{dx}Q_j^N(\circ, v) \equiv 0$ sein, also $Q_j^N(\circ, v) = const$. Das würde aber bedeuten, daß $Q_j^N(\circ, v)$ kein Extremum besitzen könnte.

Seien nun $\tilde{x}_1 < \cdots < \tilde{x}_m$, $0 \leq m \leq N-1$ die Extremstellen von $Q_j^N(\circ, v)$ *in* (x_j, x_{j+1}). Setze $\tilde{x}_0 := x_j$, $\tilde{x}_{m+1} := x_{j+1}$. Dann gilt:

$$\begin{aligned}
\mathrm{TV}\left(v\big|_{I_j}\right) &:= \sup \left\{ \sum_{i=1}^{k} |\, v(x'_{i+1}) - v(x'_i) \,| \;\Big|\; \{x'_i\}_{i=1}^{k+1} \text{ Zerlegung von } I_j \right\} \\
&\geq \sum_{i=0}^{m} |\, v(\tilde{x}_{i+1}) - v(\tilde{x}_i) \,| \\
&= \sum_{i=0}^{m} |\, Q_j^N(\tilde{x}_{i+1}), v) - Q_j^N(\tilde{x}_i, v) \,| \; + O\left(h^{N+1}\right) \\
&= \mathrm{TV}\left(Q_j^N(\circ, v)\big|_{I_j}\right) \;+\; O\left(h^{N+1}\right).
\end{aligned}$$

Damit ist Teil a) der Behauptung gezeigt. Betrachte nun den Fall, daß in I_j eine Sprungstelle von v liegt.

Zu b:

Da $h \leq h_1$, liegt im Interpolationsintervall $[x_{k_m^{(N)}}, x_{k_M^{(N)}}]$ von $Q_j^N(\circ, v)$ genau eine Sprungstelle und zwar in I_j. Mit (2) und (3) folgt nun wegen der Wahl von h_0 die nachstehende Ungleichungskette für $x \in I_j$:

$$(4) \qquad \left|\tfrac{d}{dx}Q_j^N(x, w)\right| \;<\; c(h_3) \;\leq\; \frac{c(h_4)}{h} \;<\; \min_{y \in I_j} \left\{ \left|\tfrac{d}{dx}Q_j^N(y, H)\right| \right\} \;\leq\; \left|\tfrac{d}{dx}Q_j^N(x, H)\right|.$$

Insbesondere gilt mit (3), daß $\frac{d}{dx}Q_j^N(\circ, H)$ verschieden von Null und $Q_j^N(\circ, H)$ streng monoton auf I_j ist. Damit folgt unmittelbar die strenge Monotonie von $Q_j^N(\circ, v)$ auf I_j, denn:

(α) $\underline{\frac{d}{dx} Q_j^N(\circ, H)\big|_{I_j} \; > \; 0 :}$

Wegen (4) gilt dann für $x \in I_j$: $-\frac{d}{dx} Q_j^N(x,w) \; < \; \frac{d}{dx} Q_j^N(x,H)$, also

$\frac{d}{dx} Q_j^N(x,v) \; = \; \frac{d}{dx} Q_j^N(x,w) \; + \; \frac{d}{dx} Q_j^N(x,H) \; > \; 0.$

(β) $\underline{\frac{d}{dx} Q_j^N(\circ, H)\big|_{I_j} \; < \; 0 :}$

Wegen (4) gilt nun für $x \in I_j$: $\frac{d}{dx} Q_j^N(x,w) \; < \; -\frac{d}{dx} Q_j^N(x,H)$, also

$\frac{d}{dx} Q_j^N(x,v) \; = \; \frac{d}{dx} Q_j^N(x,w) \; + \; \frac{d}{dx} Q_j^N(x,H) \; < \; 0.$

Aus der Monotonie von $Q_j^N(\circ, v)$ folgt mittels der Interpolationsbedingung:

$$\mathrm{TV}\left(Q_j^N(\circ, v)\big|_{I_j} \right) \; \leq \; \left| \, Q_j^N(x_{j+1}, v) - Q_j^N(x_j, v) \, \right| \; = \; \left| \, v_{j+1} - v_j \, \right| \; \leq \; \mathrm{TV}\left(v\big|_{I_j} \right).$$

$\blacksquare$

In [ENO2] wurde die Aussage aus S. 1.1 in einer etwas modifizierteren Form gezeigt. Die dort verwendete Beweistechnik stimmt weitestgehend mit der hier benutzten überein. An einer Stelle gibt es jedoch einen entscheidenden Unterschied:
Es wird im Beweis zu Teil b) behauptet, daß aus

$$\left| \, \tfrac{d}{dx} Q_{j,h}^N(\circ, w) \, \right| \quad \leq \quad c \qquad \text{für} \quad x \in I_{j,h} \quad \underline{\text{und}}$$

$$\mathbf{max} \left\{ \left| \, \tfrac{d}{dx} Q_{j,h}^N(\circ, H) \, \right| \; \Big| \; x \in I_{j,h} \right\} \quad \geq \quad \tfrac{c}{h}, \qquad c > 0, \; c \neq c(h) \qquad \underline{\text{und}}$$

$Q_{j,h}^N(\circ, H)$ monoton auf I_j ,

folgt:

$Q_{j,h}^N(\circ, v)$ monoton auf I_j.

Im Gegensatz zu hier wird die Aussage aus La. 1.2 für das *Maximum* und nicht für das *Minimum* gezeigt. Daß dies nicht ausreicht, zeigt folgende Skizze:

Wie man sofort sieht, gilt:

$$\tfrac{d}{dx} Q_{j,h}^N(\circ, v)\big|_{[x_j, \overline{x})} \; < \; 0 \quad \text{und} \quad \tfrac{d}{dx} Q_{j,h}^N(\circ, v)\big|_{(\overline{x}, x_{j+1}]} \; > \; 0.$$

Daher ist $Q_{j,h}^N(\circ, v)$ *nicht* monoton auf I_j !

Aus S. 1.1 läßt sich nun eine Aussage über die globale Oszillation, d.h. über die totale Variation des ENO–Splines $Q_h^N(\circ, v)$, ableiten:

Folgerung 1.2:

Vor.: $N \in \mathbb{N}$, $a, b \in \mathbb{R}$, $a < b$, $v : \mathbb{R} \to \mathbb{R}$, $v\big|_{[a,b]} \in SCL(N+1, [a,b])$,

$v\big|_{\mathbb{R}\setminus[a,b]} \equiv 0$ oder $v\big|_{\mathbb{R}\setminus[a,b]}$ periodische Fortsetzung von $v\big|_{[a,b]}$;

Beh.: Es gibt $m_0 \in \mathbb{N}$, so daß für $m \geq m_0$ mit $h := \frac{b-a}{m}$ gilt:

$$\mathrm{TV}\left(Q_h^N(\circ, v)\big|_{[a,b]}\right) \;\leq\; \mathrm{TV}\left(v\big|_{[a,b]}\right) \;+\; O\left(h^N\right).$$

Beweis:

Setze $h := \frac{b-a}{m}$ für $m \in \mathbb{N}$. Dann ist $\{x_j\}_{j=0}^m$ eine äquidistante Zerlegung von $[a,b]$. In Abhängigkeit von h bzw. m definiere folgende Indexmengen:

$$\begin{aligned} J_1 &:= \{\, j \mid 0 \leq j \leq m \;;\; \text{in } [x_j, x_{j+1}) \text{ liegt Sprungstelle von } v \,\}\,; \\ J_2 &:= \{\, 0, \ldots, m \,\} \setminus J_1. \end{aligned}$$

S. 1.1 besagt nun, daß es $h_0 > 0$ bzw. $m_0 \in \mathbb{N}$ gibt, so daß für alle $0 < h \leq h_0$ bzw. $m \geq m_0$ gilt:

$$\begin{aligned} j \in J_1 &\implies \mathrm{TV}\left(Q_{j,h}^N(\circ, v)\big|_{I_{j,h}}\right) \;\leq\; \mathrm{TV}\left(v\big|_{I_{j,h}}\right); \\ j \in J_2 &\implies \mathrm{TV}\left(Q_{j,h}^N(\circ, v)\big|_{I_{j,h}}\right) \;\leq\; \mathrm{TV}\left(v\big|_{I_{j,h}}\right) \;+\; O\left(h^{N+1}\right). \end{aligned}$$

Aufgrund der Interpolationsbedingung für $Q_h^N(\circ, v)$ (s. Def. 1.2) gilt:

$$Q_h^N(x_j + 0, v) \;=\; Q_h^N(x_j - 0, v) \;=\; v_j.$$

Daraus folgt:

$$\begin{aligned} \mathrm{TV}\left(Q_h^N(\circ, v)\big|_{[a,b]}\right) &= \sum_{j \in J_1} \mathrm{TV}\left(Q_{j,h}^N(\circ, v)\big|_{I_{j,h}}\right) \;+\; \sum_{j \in J_2} \mathrm{TV}\left(Q_{j,h}^N(\circ, v)\big|_{I_{j,h}}\right) \\ &\leq \sum_{j \in J_1} \mathrm{TV}\left(v\big|_{I_{j,h}}\right) \;+\; \sum_{j \in J_2} \left(\mathrm{TV}\left(v\big|_{I_{j,h}}\right) \;+\; O\left(h^{N+1}\right)\right) \\ &= \mathrm{TV}\left(v\big|_{[a,b]}\right) \;+\; |J_2|\, O\left(h^{N+1}\right). \end{aligned}$$

Da $m = O\left(h^{-1}\right)$, gilt insbesondere: $\mid J_2 \mid \;=\; O\left(h^{-1}\right)$, also auch

$$\mathrm{TV}\left(Q_h^N(\circ, v)\big|_{[a,b]}\right) \;\leq\; \mathrm{TV}\left(v\big|_{[a,b]}\right) \;+\; O\left(h^N\right).$$

■

Somit ist gezeigt, daß der ENO–Spline eine im wesentlichen nicht oszillierende Interpolation einer Funktion $v \in SCL(N+1, \mathbb{R})$ bildet. Dies ist eine sehr wichtige Eigenschaft, die sich bei der Verwendung dieser Interpolationstechnik bei den später beschriebenen ENO–Verfahren positiv auswirkt. Leider können die hier hergeleiteten Aussagen nicht ohne weiteres für die analytische Untersuchung der Verfahren angewendet werden, da die zu interpolierende Funktion dort *stetig* ist, und deren *Ableitung* bis auf endlich viele Stellen nicht existiert, im Gegensatz zu obiger Theorie, wo die zu *interpolierende* Funktion v selbst Sprungstellen hat. Dies geht im Beweis zu La. 1.2 und La. 1.4 wesentlich ein. Es sollen nun zu dieser Problematik einige weitere Eigenschaften der ENO–Polynome hergeleitet werden.

1.3 ENO–Eigenschaft für die Ableitung des ENO–Splines

Zunächst soll die Frage geklärt werden, ob die Ableitungen von $Q^N_{j,h}(\circ, v)$ für eine hinreichend glatte Funktion v auch eine im wesentlichen nicht oszillierende Approximation der entsprechenden Ableitungen von v bilden. Dazu folgendes

Lemma 1.5:

Vor.: $N \in \mathbb{N},\ j \in \mathbb{Z},\ v : \mathbb{R} \to \mathbb{R}\,, v \in \mathbb{C}^{N+1} \quad, h > 0;$

Beh.: Für $\nu = 0, \dots, N,\ x \in (x_{k_m^{(N)}}, x_{k_M^{(N)}})$ gilt:

$$\left(\tfrac{d}{dx}\right)^\nu Q^N_{j,h}(x, v) \;=\; \left(\tfrac{d}{dx}\right)^\nu v(x) \;+\; O\left(h^{N+1-\nu}\right).$$

Für einen Beweis siehe z.B. [IsKe], S.291.

Mit diesem Lemma läßt sich nun obige Frage positiv beantworten:

Satz 1.2: (Oszillation der Ableitung)

Vor.: $N \in \mathbb{N}$, $m \in \mathbb{N}$, $h := \frac{b-a}{m}$ $\ a, b \in \mathbb{R},\ a < b,\ v\big|_{[a,b]} \in C^{N+1}$ mit

$v\big|_{\mathbb{R}\setminus[a,b]} \equiv 0$ oder $v\big|_{\mathbb{R}\setminus[a,b]}$ periodische Fortsetzung von $v\big|_{[a,b]}$;

Beh.: Für $\nu = 0, \dots, N-1$ gilt:

$$\mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu Q^N_h(\circ, v)\big|_{[a,b]}\right) \;\le\; \mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu v\big|_{[a,b]}\right) \;+\; O\left(h^{N-\nu}\right).$$

Beweis:

O.B.d.A. sei $a = 0$. Konstruiere $Q^N_h(\circ, v)$ gemäß Def. 1.2. Für $\nu = 0$ folgt die Behauptung sofort aus S. 1.1.a analog zu Folg. 1.1, da v in $[a, b]$ *keine* Sprungstelle hat. Sei nun $\nu \in \{1, \dots, N-1\}$ und $j \in \mathbb{Z}$. Mit La. 1.5. gilt:

(1) $$\left(\tfrac{d}{dx}\right)^\nu Q^N_h(x, v) \;=\; \left(\tfrac{d}{dx}\right)^\nu v(x) \;+\; O\left(h^{N-\nu+1}\right), \quad x \in (x_{k_m^{(N)}}, x_{k_M^{(N)}}).$$

Da $\left(\frac{d}{dx}\right)^\nu Q^N_h(\circ, v) \in \mathbb{P}_{N-\nu}$, existieren höchstens $N - \nu - 1$ Extremstellen, denn sonst wäre $\left(\frac{d}{dx}\right)^\nu Q^N_h(\circ, v)$ wegen dem Fundamentalsatz der Algebra linear und hätte somit keine Extrema.

Sei nun $k \in \mathbb{N}_0$ die Anzahl der Extrema in (x_j, x_{j+1}) und $\tilde{x}_1 < \ldots < \tilde{x}_k$ die Extremstellen. Setze $\tilde{x}_0 := x_j, \tilde{x}_{k+1} := x_j + 1$. Dann gilt mit (1) :

$$\begin{aligned}
&\mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu v\big|_{I_{j,h}} \right) \\
&= \sup \left\{ \sum_{i=0}^{l} \left| \left(\tfrac{d}{dx}\right)^\nu v(x'_{i+1}) - \left(\tfrac{d}{dx}\right)^\nu v(x'_i) \right| \;\middle|\; \{x'_i\}_{i=0}^{l+1} \text{ Zerl. v. } [x_j, x_{j+1}] \right\} \\
&\geq \sum_{i=0}^{k} \left| \left(\tfrac{d}{dx}\right)^\nu v(\tilde{x}_{i+1}) - \left(\tfrac{d}{dx}\right)^\nu v(\tilde{x}_i) \right| \\
&= \sum_{i=0}^{k} \left| \left(\tfrac{d}{dx}\right)^\nu Q^N_{j,h}(\tilde{x}_{i+1}, v) - \left(\tfrac{d}{dx}\right)^\nu Q^N_{j,h}(\tilde{x}_i, v) \right| + O\left(h^{N+1-\nu}\right) \\
&= \mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu Q^N_{j,h}(\circ, v)\big|_{I_{j,h}} \right) + O\left(h^{N+1-\nu}\right).
\end{aligned}$$

Daraus folgt nun mit (1):

$$\begin{aligned}
&\mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu Q^N_{j,h}(\circ, v)\big|_{[a,b]} \right) \\
&= \sum_{i=0}^{m} \mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu Q^N_{j,h}(\circ, v)\big|_{I_{j,h}} \right) + \\
&\quad \sum_{i=1}^{m-1} \left| \left(\tfrac{d}{dx}\right)^\nu Q^N_{j,h}(x_j + 0, v) - \left(\tfrac{d}{dx}\right)^\nu Q^N_{j-1,h}(x_j - 0, v) \right| \\
&\leq \sum_{i=0}^{m} \left\{ \mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu v(.)\big|_{I_{j,h}} \right) + O\left(h^{N+1-\nu}\right) \right\} + \\
&\quad \sum_{i=1}^{m-1} \left| \left(\tfrac{d}{dx}\right)^\nu v(x_j + 0, v) - \left(\tfrac{d}{dx}\right)^\nu v(x_j - 0, v) + O\left(h^{N+1-\nu}\right) \right| \\
&\leq \mathrm{TV}\left(\left(\tfrac{d}{dx}\right)^\nu v\big|_{[a,b]} \right) + O(m)\; O\left(h^{N+1-\nu}\right).
\end{aligned}$$

Da nach Voraussetzung $m = O\left(h^{-1}\right)$, gilt also die Behauptung.

∎

Bisher wurden nur solche Funktionen betrachtet, die entweder stückweise glatt oder global glatt waren. Wie verhalten sich die ENO–Polynome aber nun bei solchen Funktionen v, die k–mal global differenzierbar sind, deren $(k+1)$–te Ableitung nur bis auf endlich viele Stellen existiert, wobei $0 \leq k < N$ ist? In Analogie zu La. 1.4 soll nun gezeigt werden, daß die ENO–Interpolationstechnik auch solche "versteckten" Unstetigkeiten erkennt.

Lemma 1.6:
Vor.: $N \in \mathbb{N}, 0 \leq k < N, \overline{x} \in \mathbb{R}, v : \mathbb{R} \to \mathbb{R}, v \in C^k$,
$v\big|_{(-\infty,\overline{x})}, v\big|_{(\overline{x},+\infty)} \in C^{k+1}$ und stetig differenzierbar ergänzbar in $\overline{x}$ mit

$$0 \; < \; \left| \left(\tfrac{d}{dx}\right)^{(k+1)} v(\overline{x}-0) \; - \; \left(\tfrac{d}{dx}\right)^{(k+1)} v(\overline{x}+0) \right| \; < \; \infty.$$

Beh.: $\exists\, h_0 > 0,\ \forall\, 0 < h \leq h_0$: Für $j \in \mathbb{Z}$ mit $\overline{x} \in I_{j,h}$ gilt für $k+1 \leq m-1 \leq N$:

$$(i) \quad |\, v[j,\ldots,j+m] \,| \; > \; |\, v[j+1,\ldots,j+m+1] \,| \, ;$$

$$(ii) \quad |\, v[j-m+1,\ldots,j+1] \,| \; > \; |\, v[j-m+1,\ldots,j] \,| \, .$$

Beweis:

Es soll hier nur der Beweis zu (i) geführt werden, da sich (ii) analog zeigen läßt. Aus Kor. 1.2.a folgt mittels vollständiger Induktion über m unter Verwendung des Mittelwertsatzes für $k+2 \leq m \leq N+1$:

$$(1) \qquad v[j,\ldots,j+m] = \beta\, \frac{(m-k-1)!}{m!\; h^{m-k}} \; + \; \alpha\, \frac{1}{m!\; h^{m-1-k}}$$

mit

$$\alpha := \sum_{i=k+1}^{m-1} (-1)^{m-i-1}\; h^{i-k-1}\; v^{(i)}(\xi_{j+1,i+j+1}) \, , \qquad \xi_{i,j} \in (x_i, x_j),\ i < l;$$

$$\beta := \; (-1)^{m-k-1} \; \left[v^{(k)}(\xi_{j+1,k+j+1}) - v^{(k)}(\xi_{j,j+k}) \right] .$$

Analog zeigt man:

$$(2) \qquad \begin{aligned} v[j+1,\ldots,m+j+1] \; = \; & \sum_{i=k+1}^{m-1} (-1)^{(m-1-i)}\; \frac{1}{m!\; h^{m-i}}\; v^{(i)}(\xi_{j+2,j+i+2}) \; + \\ & (-1)^{m-k-1}\; \frac{(m-k-1)!}{m!\; h^{m-k}}\; v^{(k+1)}(\xi_{j+1,j+k+2}). \end{aligned}$$

Sei nun $h > 0$. (Falls mehrere "verborgene" Unstetigkeiten existieren würden, so müßte man analog zum Beweis zu La. 1.4 wieder $\tilde{h}_0$ einführen). Dann gilt:

$$(3) \qquad h^{m-k-1}\; m!\; v[j+1,\ldots,j+m+1] \; \leq \; b$$

mit

$$\begin{aligned} b := \; & \sum_{i=k+1}^{m-1} \max\left\{ \, |v^{(i)}(x)| \; \middle| \; \overline{x} \leq x \leq \overline{x} + N\, h_0 \right\} \; + \\ & (m-k-1)! \max\left\{ \, |v^{(k+1)}(x)| \; \middle| \; \overline{x} \leq x \leq \overline{x} + N\, h_0 \right\} . \end{aligned}$$

$$(4) \qquad |\alpha| \;\le\; a := \sum_{i=k+1}^{m-1} \max\left\{ |v^{(i)}(x)| \;\middle|\; \overline{x} \le x \le \overline{x} + N\,h_0 \right\}$$

Da $v^{(k+1)}$ in $\overline{x}$ einenSprung endlicher Höhe hat, gibt es $\overline{c} > 0$ und $\varepsilon > 0$, so daß gilt:

$$(5) \qquad \forall\, x \in [\,\overline{x} - \varepsilon, \overline{x}\,), y \in [\,\overline{x}, \overline{x} + \varepsilon] : \;|\, v(x) \;-\; v(y)\,| \;>\; \overline{c}.$$

Für $h \le \varepsilon$ gilt dann insbesondere:

$$(6) \qquad |\beta| \;>\; \overline{c},$$

denn: $\overline{x} - \varepsilon \;<\; x_j \;\le\; \overline{x} \;<\; x_{j+1} \;<\; \overline{x} + \varepsilon$. Analog zu La. 1.4 läßt sich nun folgende Fallunterscheidung durchführen. Setze dazu

$$h_0'(\overline{x}) := \min\left\{\varepsilon, \frac{\overline{c}}{a+b}\right\} \quad \left(\text{bzw. } h_0'(\overline{x}) := \min\left\{\varepsilon, \tilde{h}_0, \frac{\overline{c}}{a+b}\right\}\right).$$

Für $0 < h \le h_0'$ gilt dann, da $k + 2 \le m \le N + 1$ für $0 \le k < N$, insbesondere:

$$(7) \qquad \left(\frac{(m-k-1)!\,\overline{c}}{a+b}\right) \;\ge\; h,$$

und somit

(a) $\underline{\beta > \overline{c}}$

Aus (1) folgt unmittelbar: $h^{m-k-1}\; m!\; v[j, \ldots, m+j] \ge -a + \overline{c}\,(m-k-1)!\; h^{-1}$.

Mit (7) gilt dann: $h^{m-k-1}\; m!\; v[j, \ldots, m+j] \;\ge\; b \;> 0$.

Daraus folgt mit (3): $|\, v[j, \ldots, m+j]\,| \;>\; |\, v[j+1, \ldots, m+j+1]\,|$.

(b) $\underline{\beta < -\overline{c}:}$

Aus (1) folgt sofort: $h^{m-k-1}\; m!\; v[j, \ldots, m+j] \le a - \overline{c}\,(m-k-1)!\; h^{-1}$.

Mit (7) gilt dann wiederum: $h^{m-k-1}\; m!\; v[j, \ldots, m+j] \;\le\; -b \;< 0$.

Daraus folgt mit (3): $|\, v[j, \ldots, m+j]\,| \;>\; |\, v[j+1, \ldots, m+j+1]\,|$.

Wegen (a),(b) gilt somit die Behauptung (i). Analog erhält man (ii) für ein h_0''. Dann gilt die Behauptung mit $h_0 := \min\{h_0', h_0''\}$.

■

Obige Aussage gilt auch, falls v mehrere, aber endlich viele, solcher "verborgenen" Unstetigkeiten $\overline{x}_1, \ldots, \overline{x}_n, n \in \mathbb{N}$ hat. Im Beweis gilt dann die Aussage (3) nur für $h \le \tilde{h}_0$.

Aufgrund der Definition von $Q_h^N(\circ, v)$, folgt somit als Anwendung für $k = 0$:

Satz 1.3:

Vor.: $n, N \in \mathbb{N}$, $\overline{x}_1 < \ldots < \overline{x}_n$, $v : \mathbb{R} \to \mathbb{R}$, $v \in C^0$,

$v\big|_{(-\infty,\overline{x}_1)}$, $v\big|_{(\overline{x}_i,\overline{x}_{i+1})}$, $v\big|_{(\overline{x}_n,+\infty)} \in C^1$, $i = 1, \ldots, n-1$ und
stetig differenzierbar ergänzbar in $\overline{x}_i$ mit $0 < |\, v'(\overline{x}_i - 0) - v'(\overline{x}_i + 0)\,| < \infty$;

Beh.: Es gibt $h_0 > 0$, so daß für alle $0 < h \leq h_0$ gilt:

Liegt in $I_{j,h}, j \in \mathbb{Z}$ keine "verborgene" Sprungstelle von v, so auch nicht im zugehörigen Interpolationsintervall $[x_{k_m^{(N)}}, x_{k_M^{(N)}}]$ von $Q_{j,h}^N(\circ, v)$.

Diese Aussage läßt sich wegen Def. 1.2 wie folgt interpretieren:

Liegt $I_{j,h}$ links bzw. rechts einer "verborgenen" Sprungstelle, so wird das zugehörige Interpolationspolynom $Q_{j,h}^N$ nur aus Informationen aufgebaut, die ebenfalls links bzw. rechts dieser Stelle liegen, falls h hinreichend klein gewählt wird. Gerade in dieser Tatsache dürfte das zum Teil sehr gute Verhalten der ENO–Verfahren in Gegenwart von Schocks begründet sein (vgl. dazu auch Kap. 2.3).

Nun aber zu den eigentlichen Verfahren.

Kapitel 2: ENO–Verfahren

Im folgenden sollen nun die eigentlichen Verfahren hergeleitet werden. Dies wird in zwei Schritten geschehen. Zunächst wird der skalare Fall behandelt und dieser anschließend auf Systeme übertragen.

2.1 Skalarer Fall

Um die Herleitung der Verfahren besser motivieren zu können, soll vorweg ein kurzer Einblick in die kontinuierliche Theorie von Erhaltungssätzen gegeben werden.

2.1.1 Grundlagen der kontinuierlichen Theorie

Im folgenden sollen nun skalare Erhaltungssätze in einer Raumdimension betrachtet werden. Diese sind folgendermaßen definiert:

Definition 2.1:

Unter einem *skalaren Erhaltungssatz* versteht man eine partielle Differentialgleichung von folgendem Typ:

$$u_t(x,t) \; + \; (F(u(x,t)))_x \; = \; 0, \quad x \in \mathbb{R}, t > 0;$$

$$u(x,0) \; = \; u_0(x), \quad x \in \mathbb{R}$$

mit $u : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R},\; F \in C^1(\mathbb{R}),\; u_0 : \mathbb{R} \to \mathbb{R}$.

F heißt auch der *Fluß* des Erhaltungssatzes.

Man unterscheidet nun drei verschiedene Arten von Lösungen:

Definition 2.2:

Eine Lösung u eines skalaren Erhaltungssatzes heißt *starke Lösung*, falls gilt:

$$u \in C^1(\mathbb{R} \times \mathbb{R}_+).$$

Diese Lösung läßt sich durch den folgenden Satz näher beschreiben:

Satz 2.1:

Vor.: $F \in C^2(\mathbb{R}),\; u_0 \in C^1(\mathbb{R}),\; -\infty < \mu := \min\left\{F''\left(u_0(x)\right)\; u_0'(x) \;\middle|\; x \in \mathbb{R}\right\}$,

$$\delta \; := \; \left\{ \begin{array}{l} +\infty \;, \mu \geq 0 \\ -\dfrac{1}{\mu} \;, \mu < 0 \end{array} \right\};$$

Beh.: a) $u(x,t) \; = \; u_0\left(\psi^{[-1]}(x)\right)$ mit $x = \psi(y) := y + t\; F'\left(u_0(y)\right)$ bzw.

$u(x,t) \; = \; u_0\left(x - t\; F'\left(u(x,t)\right)\right)$ für $t \in [0,\delta]$

ist die einzige starke Lösung des zugehörigen skalaren Erhaltungssatzes;

b) i) $\frac{d}{dt}x(t) \;=\; F'\left(u(x(t),t)\right)$ hat genau eine Lösung $x : [0,\delta] \to \mathbb{R}$, die s.g. *Charakteristik*;

ii) die Charakteristiken sind Geraden;

iii) $u\left(x(t),t\right) \;=\; const$ entlang den Charakteristiken.

Beweis:

Dazu siehe zum Beispiel [Smo] oder [Str].

■

Diese Charakterisierung wird in Kap. 2.1.6 noch benötigt werden.

Der Nachteil obiger Definition ist, daß die starke Lösung selbst bei glatten Anfangsdaten u_0 im allgemeinen nur endliche Zeit existiert. Deshalb schwächt man die Bedingung der Differenzierbarkeit ab:

Definition 2.3:

Sei $u_0 : \mathbb{R} \to \mathbb{R}$ beschränkt und Lebesque–meßbar. Eine beschränkte und Lebesque–meßbare Funktion $u : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ heißt *schwache Lösung* eines skalaren Erhaltungssatzes $:\Longleftrightarrow$

Für alle $\Phi \in C_0^1(\mathbb{R} \times \mathbb{R}_+)$ gilt:

$$\int_0^{+\infty}\int_{-\infty}^{+\infty} u(x,t)\;\Phi_t(x,t) \;+\; F\left(u(x,t)\right)\;\Phi_x(x,t)\;dx\;dt + \int_{-\infty}^{+\infty} u_0(x)\;\Phi(x,0)dx = 0$$

Die schwache Lösung kann aber auch äquivalent definiert werden durch

Korollar 2.1:

u ist schwache Lösung eines skalaren Erhaltungssatzes, genau dann wenn
für alle $a, b \in \mathbb{R},\; t_1, t_2 \in \mathbb{R}_+$ mit $a < b,\; t_1 < t_2$ gilt:

$$\int_a^b \int_{t_1}^{t_2} u_t(x,t) \;+\; \left(F\left(u(x,t)\right)\right)_x\; dx\; dt \;=\; 0.$$

Siehe dazu auch [HaLL], S. 36.

Man kann sogar zeigen, daß gilt:

Korollar 2.2:

Ist u schwache Lösung eines skalaren Erhaltungssatzes und $u \in C^1(\mathbb{R} \times \mathbb{R}_+)$, dann ist u auch starke Lösung.

Ein Beweis findet sich in [Smo], S. 247.

Ferner gilt:

Korollar 2.3:

Vor.: $M, N \in \mathbb{N},\ \varphi : \mathbb{R}_+ \to \mathbb{R} \times \mathbb{R}_+,\ i = 1, \ldots, N$ glatte Kurven wobei $|\varphi_i'|$ beschränkt ist, und es gibt Gebiete (offen und zusammenhängend) $\Omega_i,\ i = 1, \ldots, M$ mit

$$\mathbb{R} \times \mathbb{R}_+ \backslash \left\{ \bigcup_{i=1}^{N} \varphi(\mathbb{R}_+) \right\} = \bigcup_{i=1}^{M} \Omega_i;$$

u starke Lösung eines Erhaltungssatzes auf Ω_i mit

1) $u|_{\Omega_i}$ stetig differenzierbar fortsetzbar auf $\overline{\Omega}_i$,

2) $\frac{d}{dt}\varphi_i(t)\,[u(\varphi(t)-0,t) - u(\varphi(t)+0,t)] \stackrel{!}{=} [F(u(\varphi(t)-0,t)) - F(u(\varphi(t)+0,t))]$ für $t > 0, i \in \{1, \ldots, N\}$;

Beh.: u ist schwache Lösung des Erhaltungssatzes.

Für einen Beweis siehe man [Daf1] und [Daf2].
Man sieht also, daß die schwache Lösung eine Abschwächung der starken Lösung ist. Sie hat jedoch den Nachteil, daß die Eindeutigkeit verloren geht (s. [Str]). Man kann diese jedoch durch eine zusätzliche Bedingung an die schwache Lösung erzwingen:

Definition 2.4:

Eine schwache Lösung u eines skalaren Erhaltungssatzes heißt *Entropielösung* $:\Longleftrightarrow$

u erfüllt die *Entropiebedingung*, d.h.:

$$\exists\ E = E(M, \mu, A)\ :\ \frac{u(x+a,t) - u(x,t)}{a}\ <\ \frac{E}{t} \quad \forall\ a > 0,\ t > 0\ x \in \mathbb{R} \text{ mit}$$

$M := \|u_0\|_\infty < \infty,\ \mu := \min\{\, F''(u) \mid |u| \leq M \,\},\ A := \max\{\, |F'(u)| \mid |u| \leq M \,\}$
wobei $F \in C^2(\mathbb{R}), u_0 \in L^\infty(\mathbb{R})$.

Daß diese Definition sinnvoll ist, ergibt sich aus dem Satz von Oleinik, der besagt, daß für $F \in C^2(\mathbb{R}),\ F'' > 0$ und $u_0 \in L^\infty(\mathbb{R})$ eine eindeutige Entropielösung des skalaren Erhaltungssatzes existiert (s. [Smo], S. 266). Diese Lösung setzt sich dabei im (x,t)–Raum aus Teillösungen in einzelnen Bereichen zusammen. Diese Bereiche sind durch Schocks, Kontaktunstetigkeiten oder Verdünnungswellen getrennt, in denen die Lösung glatt ist. Näheres dazu kann man zum Beispiel [Smo] oder [Str] entnehmen. In Kap. 2.1.6 wird darauf noch genauer eingegangen.
Im folgenden soll unter der Lösung eines Erhaltungssatzes immer die schwache Lösung verstanden werden.
Von numerischen Verfahren zur Approximation eines skalaren Erhaltungssatzes fordert man nun, daß es im Falle der Konvergenz gegen die Entropielösung konvergiert. Da das Verfahren insbesondere eine Näherung einer schwachen Lösung ist, verlangt man ferner, daß es die äquivalente Formulierung aus Kor. 2.1 simuliert. Diese läßt sich diskret schreiben als

$$u(a,t_2)\ \approx\ u(a,t_1)\ -\ \frac{t_2 - t_1}{b - a}\ [F(u(b,t_2))\ -\ F(u(a,t_2))]\,.$$

Daher betrachtet man in der Regel nur solche Verfahren, die die folgende Form haben:

Definition 2.5:

Ein Differenzenverfahren heißt in *Erhaltungsform*, genau dann wenn es eine Funktion $G = G(y_1, \ldots, y_{2q}),\ q \in \mathbb{N}_0$, gibt, so daß gilt:

$$v_j^{n+1} = v_j^n - \lambda \left[G(v_{j-q+1}^n, \ldots, v_{j+q}^n) - G(v_{j-q}^n, \ldots, v_{j+q-1}^n) \right].$$

Dabei bezeichnet v_i^n die Approximation zu $u(x_i, t_n)$ mit $t_n := n\,\tau,\ x_i := i\,h,\ \lambda := \frac{\tau}{h}$, $\lambda, \tau > 0$.

G heißt auch der *numerische Fluß* des Verfahrens. Er heißt *konsistent*, falls gilt:

$$G(u, \ldots, u) = F(u) \qquad \text{für} \quad u \in \mathbb{R}.$$

Da die Schemata gegen die Entropielösung konvergieren sollen, muß man ein diskretes Analogon zur Entropiebedingung finden. Dazu betrachte folgende — auch für Systeme (s. Kap. 2.2.1) gültige

Definition 2.6:

Eine skalare Funktion $U : \mathbb{R}^m \to \mathbb{R}$ heißt *Entropiefunktion* eines Systems von $m \geq 1$ Erhaltungssätzen, genau dann wenn gilt:

(1) U ist eine konvexe Funktion, d.h. U'' ist positiv definit;
(2) Es existiert eine skalare Funktion $F_E : \mathbb{R} \to \mathbb{R}$ mit $grad\ U\ F'(u) = grad\ F_E(u)$.

Falls U eine solche Entropiefunktion zu einem bestimmten Erhaltungssatz ist, dann gilt insbesondere für alle nichtnegativen $\Phi \in C_0^1(\mathbb{R} \times \mathbb{R}_+)$ (s. [HaLL], S. 36):

$$\int_0^{+\infty} \int_{-\infty}^{+\infty} \left(U(u)\,\Phi_t + F_E(u)\,\Phi_x \right)\Big|_{(x,t)} dx\,dt + \int_{-\infty}^{+\infty} \left(U(u)\,\Phi \right)\Big|_{(x,0)} dx \leq 0$$

Daraus läßt sich nun eine diskrete Form der Entropiebedingung ableiten:

Definition 2.7:

Ein Verfahren erfüllt eine *numerische Entropiebedingung*, genau dann wenn es eine *numerische Entropiefunktion* $G_E = G_E(y_1, \ldots, y_{2q})$ gibt, so daß die Approximation $\{v_j^n\}_{j \in \mathbb{Z}, n \in \mathbb{N}_0}$ folgende Ungleichung erfüllt:

$$U(v_j^{n+1}) - U(v_j^n) + \lambda \left[G_E(v_{j-q+1}^n, \ldots, v_{j+q}^n) - G_E(v_{j-q}^n, \ldots, v_{j+q-1}^n) \right] \leq 0.$$

Dabei ist U eine Entropiefunktion zu dem approximierten Erhaltungssatz.

Leider folgt aus der Tatsache, daß ein Verfahren eine numerische Entropiebedingung erfüllt, noch nicht, daß es gegen die Entropielösung konvergiert. Dazu ist eine weitere Bedingung erforderlich:

Definition 2.8:

(a) Unter der *totalen Variation* eines Verfahrens $\left\{v_j^n\right\}_{j\in\mathbb{Z}}$, $n \in \mathbb{N}_0$ versteht man

$$\mathrm{TV}\left(\left\{v_j^n\right\}\right) := \sum_{j=-\infty}^{+\infty} \mid v_{j+1}^n - v_j^n \mid .$$

(b) Ein Verfahren heißt *TV–stabil*, genau dann wenn für alle Anfangsdaten $u_0 : \mathbb{R} \to \mathbb{R}$ und $T > 0$ eine Konstante $c > 0$ existiert, so daß gilt:

$$\mathrm{TV}\left(\left\{v_j^n\right\}\right) \;\leq\; c\,\mathrm{TV}\left(\left\{v_j^0\right\}\right) \quad \forall\, \tau > 0,\ n \in \mathbb{N},\ n\,\tau \leq T.$$

Wie später (s. Kap. 2.1.6) gezeigt wird, nimmt bei der Entropielösung die totale Variation mit fortschreitender Zeit nicht zu, d.h. sie hängt von der Oszillation der Anfangsdaten ab. Somit ist es sinnvoll obige Bedingung für ein Verfahren zu fordern.

A. Harten zeigte 1974 folgenden

Satz 2.2:

Vor.: Ein Erhaltungssatz besitze eine Entropiefunktion und u_0 habe einen kompakten Träger. Das Verfahren sei in Erhaltungsform und erfülle eine numerische Entropiebedingung und sei TV–stabil.

Beh.: Das Verfahren konvergiert gegen eine Grenzfunktion, und diese ist die Entropielösung.

Zum Beweis siehe [Ha1].

Daher ist es sinnvoll Verfahren zu konstruieren, die TV–stabil sind und eine numerische Entropiebedingung erfüllen.

Seit 1983 werden nun, ausgehend von einer Arbeit von A. Harten (s. [Ha2]), s.g. *TVD–Verfahren* entwickelt und analysiert (s. [Ye1], [Ye2]). Diese sind folgendermaßen definiert:

Definition 2.9:

Ein Verfahren $\left\{v_j^n\right\}_{j\in\mathbb{Z},n\in\mathbb{N}_0}$ heißt **T***otal-* **V***ariation-* **D***iminishing*, falls für alle Anfangsdaten $u_0 : \mathbb{R} \to \mathbb{R}$ und $T > 0$ gilt:

$$\mathrm{TV}\left(\left\{v_j^n\right\}\right) \;\leq\; \mathrm{TV}\left(\left\{v_j^0\right\}\right) \quad \forall\, \tau > 0,\ n \in \mathbb{N},\ n\,\tau \leq T.$$

Aufgrund dieser Eigenschaft sind TVD–Verfahren offensichtlich TV–stabil. Sie haben jedoch den Nachteil, daß die Ordnung (s. Kap. 2.1.5) in s.g. "nonsonic critical points", d.h.: $\frac{\partial}{\partial x}\, u(x,t) = 0,\ \left(\frac{d}{dx}\right)^2 u(x,t) \neq 0,\ F'(u(x,t)) \neq 0$, durch 1 beschränkt ist (s. [HaT]). Daher ist man bestrebt, TVD–ähnliche Verfahren von gleichmäßig hoher Ordnung zu konstruieren. Dazu schwächt man die TVD–Bedingung aus Def. 2.9 ab:

Definition 2.10:

Ein Verfahren $\left\{v_j^n\right\}_{j\in\mathbb{Z},n\in\mathbb{N}_0}$ heißt **E***ssentially-* **N***on-* **O***scillatory*, falls für alle Anfangsdaten $u_0 : \mathbb{R} \to \mathbb{R}$ und $T > 0$ für h hinreichend klein gilt:

$$\mathrm{TV}\left(\left\{v_j^n\right\}\right) \;\le\; \mathrm{TV}\left(\left\{v_j^0\right\}\right) \;+\; O(h) \quad \forall\, \tau > 0,\; n \in \mathbb{N},\; n\,\tau \le T.$$

Wie man leicht einsieht, sind diese Verfahren TV–stabil mit $c = 2$ (vgl. Def. 2.8.b) für h hinreichend klein. Im folgenden sollen nun solche Verfahren konstruiert und analysiert werden.

2.1.2 Konstruktion der ENO–Verfahren

Die hier vorgestellten Verfahren wurden in den Jahren 1986/87 entwickelt und zum Teil analysiert (s. [ENO1],[ENO2],[ENO3]). Im folgenden werden also keine neuartigen Methoden hergeleitet, aber einige weiterführende numerische und analytische Aspekte betrachtet, die zum besseren Verständnis der Verfahren beitragen dürften und vielleicht auch beim Nachweis der TV–Stabilität behilflich sein könnten.

Sei also u Lösung eines skalaren Erhaltungssatzes:

$$u_t(x,t) \;+\; \left(F(u(x,t))\right)_x \;=\; 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+;$$

$$u(x,0) \;=\; u_0(x), \qquad x \in \mathbb{R}.$$

Führe nun folgende Bezeichnungen ein:

$$\overline{u}^n(x) := \frac{1}{h}\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} u(x,t_n)\,dx, \quad t_n := \tau\, n;$$
$$\overline{u}_j^n := \overline{u}^n(x_j), \quad x_j := j\,h,\; j \in \mathbb{Z}.$$

Dann gilt mit Kor. 2.1:

$$\overline{u}_j^{n+1} \;=\; \overline{u}_j^n \;-\; \lambda\;\left[\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}}\right] \quad \text{mit}$$

$$\hat{F}_{j+\frac{1}{2}} := \frac{1}{\tau}\int_{t_n}^{t_{n+1}} F(u(x_{j+\frac{1}{2}},t))\,dt, \quad \lambda := \frac{\tau}{h}.$$

Davon ausgehend läßt sich das ENO–Verfahren folgendermaßen konstruieren:

(2.1.a) Setze $v_j^0 := \overline{u}_j^0 = \frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)\,dx$;

(2.1.b) Sei v_j^n Approximation der Mittelwerte der exakten Lösung $u(x_j,t_n)$, d.h. Näherungen für $\overline{u}_j^n$. Die Verfahrenswerte v_j^{n+1} auf der neuen Zeitstufe berechnen sich dann in drei Schritten wie folgt:

(α) Rekonstruktion:
Approximiere die exakte Lösung $u(\circ,t_n)$ aus den v_j^n, derart, daß für alle $x\in\mathbb{R}$, in deren Umgebung $u(\circ,t_n)$ hinreichend glatt ist, gilt:

$$L^N(x,v_j^n) = u(x,t_n) + O\left(h^{N+1}\right), \quad N\geq 1;$$

(β) Numerischer Fluß:
Approximiere den Fluß $\hat{F}_{j+\frac{1}{2}}$ gegeben durch

$$\hat{F}_{j+\frac{1}{2}} := \frac{1}{\tau}\int_{t_n}^{t_{n+1}} F(v(x_{j+\frac{1}{2}},t))\,dt$$

mit v Lösung zu dem Anfangswertproblem

$$\begin{aligned} v_t(x,t) \;+\; (F(v(x,t)))_x \;&=\; 0 \quad (x,t)\in\mathbb{R}\times(t_n,+\infty);\\ v(x,t_n) \;&=\; L^N(x,v_j^n) \qquad x\in\mathbb{R}; \end{aligned}$$

(γ) Mittelwertbildung:
Berechne die Näherungswerte auf der neuen Zeitstufe durch

$$v_j^{n+1} \;=\; v_j^n \;-\; \lambda\left[\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}}\right].$$

Mit dieser Konstruktionsvorschrift sind die ENO–Verfahren prinzipiell beschrieben. Man benötigt jedoch noch explizite Kenntnis von

(a) der Rekonstruktion $L^N(\circ,v_j^n)$ und

(b) der Approximation vom Fluß $\hat{F}_{j+\frac{1}{2}}$.

Gerade in diesen beiden Punkten liegt aber die Schwierigkeit bei der expliziten Konstruktion der Verfahren. Im folgenden Abschnitt sollen nun zwei Möglichkeiten für $L^N(\circ,v_j^n)$ beschrieben werden und anschließend drei Varianten zur Näherung von $\hat{F}_{j+\frac{1}{2}}$.

2.1.3. Rekonstruktionsmethoden

Im folgenden bezeichne w eine stückweise glatte und Lipschitz–stetig fortsetzbare Funktion gemäß Def. 1.1, und $\overline{w}$ sei als der Mittelwert von w definiert, d.h.:

$$\overline{w}(x) := \frac{1}{h}\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} w(s)\,ds, \qquad h>0.$$

Hierbei entspricht w der unbekannten Lösung $u(\circ,t_n)$ und $\overline{w}$ der bekannten Approximation zu $\overline{u}(\circ,t_n)$.
Ziel ist es nun, eine Funktion $L^N(\circ,\overline{w})$ zu konstruieren, die folgende Eigenschaften besitzt:

(2.2.a) Approximation:

Für alle Punkte x, die eine Umgebung besitzen, in der w glatt ist, gilt:

$$L^N(x,\overline{w}) \;=\; w(x) \;+\; O\left(h^{N+1}\right);$$

(2.2.b) Erhaltung:

$$\overline{L}^N(x_j,\overline{w}) \;:=\; \frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} L^N(x,\overline{w})\,dx \;=\; \overline{w}(x_j);$$

(2.2.c) ENO–Bedingung:

$$\mathrm{TV}\left(L^N(\circ,\overline{w})\right) \;\le\; \mathrm{TV}(w) \;+\; O\left(h^N\right), \qquad N>0.$$

Von diesen drei Eigenschaften lassen sich bei den im folgenden beschriebenen Rekonstruktionsmethoden jedoch nur die ersten beiden nachweisen, während die dritte nur für Spezialfälle gezeigt werden kann. Aufgrund der ENO–Interpolationstechnik, die hier eingeht, ist die Hoffnung berechtigt, daß man sie auch unter allgemeineren Voraussetzungen an w erhält.

2.1.3.1 RP–Methode

Diese Methode beruht auf der Rekonstruktion mittels der Stammfunktion (Reconstruction via Primitive Function) von w und ist folgendermaßen festgelegt:

(2.3.a) $W(x) := \int_a^x w(s)\,ds \quad \left(\text{o.B.d.A. } a=-\frac{h}{2},\ h \text{ vorgegebene Gitterweite}\right);$

(2.3.b) Bestimme $Q^N(\circ,W)$ gemäß Def. 1.2 stückweise auf $I_{j-\frac{1}{2}} := [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$:

$$Q^N(x_{j+\frac{1}{2}},W) \;=\; W(x_{j+\frac{1}{2}}) \;=\; \sum_{k=0}^{j}\overline{w}(x_k)\,h, \quad j\ge 0;$$

(2.3.c) $$L^N(\circ,\overline{w}) \;:=\; \tfrac{d}{dx}\,Q^N(\circ,W).$$

Da das Verfahren nur auf einem Kompaktum arbeitet, stellt die Wahl von a keine Einschränkung dar. Dann lassen sich folgende Eigenschaften zeigen:

Lemma 2.1:

Vor.: $w\in SCL(N+1,\mathbb{R})$, $N\in\mathbb{N}$;

Beh.: a) Ist w in einer Umgebung von $x\in\mathbb{R}$ $(N+1)$–mal differenzierbar, so gilt:

$$L^N(x,\overline{w}) \;=\; w(x)+O\left(h^N\right) \qquad \text{für } h \text{ hinreichend klein;}$$

b) $\overline{L}^N(x_j,\overline{w}) \;=\; \overline{w}(x_j)$.

Beweis:

Zu a): Sei $x > 0$ derart, daß w in einer Umgebung von x $(N+1)$–mal differenzierbar ist. Dann gilt für h klein genug mit La. 1.4:

(1) $\exists\, j = j(h) \in \mathbb{N}_0 :\ x \in I_{j-\frac{1}{2},h}$;

(2) $w|_{I_{j-\frac{1}{2},h}} \in C^{N+1}$;

(3) Sei $[x_{k_m^{(N)}}, x_{k_M^{(N)}}]$ das zu $Q^N_{j-\frac{1}{2},h}(\circ, W)$ gehörende Interpolationsintervall, dann gilt:

$$w|_{(x_{k_m^{(N)}}, x_{k_M^{(N)}})} \in C^{N+1}.$$

Aus La. 1.5 folgt somit die Behauptung.

Zu b): Aus (2.3) folgt aufgrund der Interpolationsbedingung von $Q^N(\circ, W)$:

$$\begin{aligned}\overline{L}^N(x_j, \overline{w}) &\overset{(2.3.c)}{=} \frac{1}{h}\left[\, Q^N(x_{j+\frac{1}{2}}, W) - Q^N(x_{j-\frac{1}{2}}, W)\,\right]\\ &\overset{(2.3.b)}{=} \frac{1}{h}\left[\, W(x_{j+\frac{1}{2}}) - W(x_{j-\frac{1}{2}})\,\right]\\ &\overset{(def.)}{=} \overline{w}(x_j)\end{aligned}$$

■

Die erwünschte Eigenschaft (2.2.c) für die Rekonstruktion läßt sich für stückweise glattes w nicht so ohne weiteres zeigen, nur für $N = 1$ (s. La. 2.3), aber wohl für $w \in C^{N+1}$ (s. La. 2.4). Die Tatsache, daß die Mittelwertbildung TVD ist (s. La. 2.2), legt die Vermutung nahe, daß sie auch für stückweise glattes w gilt.

Lemma 2.2:

Vor.: $a, b \in \mathbb{R},\ a < b,\ w : \mathbb{R} \to \mathbb{R},\ w|_{[a,b]} \in SCL(1, [a,b]),$

$w\big|_{\mathbb{R}\setminus[a,b]} \equiv 0$ oder periodische Fortsetzung von $w|_{[a,b]}$,

$\overline{w}(x) := \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} w(s)\, ds,\ h > 0;$

Beh.: $$\mathrm{TV}\left(\overline{w}|_{[a,b]}\right) \leq \mathrm{TV}\left(w|_{[a,b]}\right).$$

Beweis:

O.B.d.A. habe w genau eine Sprungstelle $\overline{x}$ in (a,b). Sei $\{\tilde{x}_j\}_{j=0}^n,\ n \in \mathbb{N}$, eine Zerlegung von $[a,b]$. Dann gibt es $i_0 \in \{1, \ldots, n-1\}$ mit $\overline{x} \in [\tilde{x}_{i_0} - \frac{h}{2}, \tilde{x}_{i_0} + \frac{h}{2}]$ und es gilt:

(1) $\underline{j \neq i_0 :}$

Daraus folgt: $w|_{(\tilde{x}_j - \frac{h}{2}, \tilde{x}_j + \frac{h}{2})} \in C^1$, also insb.:

$$\overline{w}(\tilde{x}_j) \overset{(MWS)}{=} w(\xi_j) \quad \text{für ein } \xi_j \in \left(\tilde{x}_j - \tfrac{h}{2}, \tilde{x}_j + \tfrac{h}{2}\right).$$

(2) $\underline{j = i_0 :}$

Dann gilt: $w|_{(\tilde{x}_{i_0}-\frac{h}{2},\overline{x})} \in C^1$ und $w|_{(\overline{x},\tilde{x}_{i_0}+\frac{h}{2})} \in C^1$, also insbesondere:

$$\begin{aligned} \overline{w}(x_{i_0}) &= \frac{1}{h}\left[\int_{\tilde{x}_{i_0}-\frac{h}{2}}^{\overline{x}} w(x)\,dx + \int_{\overline{x}}^{\tilde{x}_{i_0}+\frac{h}{2}} w(x)\,dx\right] \\ &\overset{\text{(MWS)}}{=} \frac{1}{h}\left[\left(\overline{x}-\tilde{x}_{i_0}+\frac{h}{2}\right) w(\xi_L) + \left(\tilde{x}_{i_0}+\frac{h}{2}-\overline{x}\right) w(\xi_R)\right] \end{aligned}$$

$$\text{mit } \xi_L \in \left(\tilde{x}_{i_0}-\frac{h}{2},\overline{x}\right),\ \xi_R \in \left(\overline{x},\tilde{x}_{i_0}+\frac{h}{2}\right).$$

Beachte, daß bei mehreren Sprungstellen im gleichen Intervall dementsprechend das Integral analog zu oben in mehrere Teilintervalle aufgespalten werden muß .

Aus (1) und (2) folgt nun:

$$|\,\overline{w}(\tilde{x}_{i_0})-\overline{w}(\tilde{x}_{i_0-1})\,| = \left|\,w(\xi_L)+\frac{1}{h}\left(\tilde{x}_{i_0}+\frac{h}{2}-\overline{x}\right)[w(\xi_R)-w(\xi_L)]-w(\xi_{i_0-1})\,\right|;$$

$$|\,\overline{w}(\tilde{x}_{i_0+1})-\overline{w}(\tilde{x}_{i_0})\,| = \left|\,w(\xi_{i_0+1})+\frac{1}{h}\left(\overline{x}-\tilde{x}_{i_0}+\frac{h}{2}\right)[w(\xi_R)-w(\xi_L)]-w(\xi_R)\,\right|.$$

Damit gilt:

$$\begin{aligned} &\sum_{j=0}^{n-1} |\,\overline{w}(\tilde{x}_{j+1})-\overline{w}(\tilde{x}_j)\,| \\ &= \sum_{j=0}^{i_0-2} |\,\overline{w}(\tilde{x}_{j+1})-\overline{w}(\tilde{x}_j)\,| + \sum_{j=i_0+1}^{n-1} |\,\overline{w}(\tilde{x}_{j+1})-\overline{w}(\tilde{x}_j)\,| + \\ &\quad |\,\overline{w}(\tilde{x}_{i_0})-\overline{w}(\tilde{x}_{i_0-1})\,| + |\,\overline{w}(\tilde{x}_{i_0+1})-\overline{w}(\tilde{x}_{i_0})\,| \\ &\le \sum_{j=0}^{i_0-2} |\,w(\xi_{j+1})-w(\xi_j)\,| + \sum_{j=i_0+1}^{n-1} |\,w(\xi_{j+1})-w(\xi_j)\,| + \\ &\quad |\,w(\xi_L)-w(\xi_{i_0-1})\,| + |\,w(\xi_R)-w(\xi_L)\,| + |\,w(\xi_{i_0+1})-w(\xi_R)\,| \\ &\le \mathrm{TV}\left(w|_{[a,\xi_L]}\right) + \mathrm{TV}\left(w|_{[\xi_L,\xi_R]}\right) + \mathrm{TV}\left(w|_{[\xi_R,b]}\right) \le \mathrm{TV}\left(w|_{[a,b]}\right). \end{aligned}$$

Da die Zerlegung beliebig gewählt war, kann man auf der linken Seite der Ungleichungskette zum Supremum über alle Zerlegungen übergehen und erhält somit die Behauptung.

■

Als Spezialfall für (2.2.c) erhält man

Lemma 2.3:

Vor.: $a, b \in \mathbb{R},\ a < b,\ w : \mathbb{R} \to \mathbb{R},\ w|_{[a,b]} \in SCL(1, [a, b]),$

$w\big|_{\mathbb{R}\setminus[a,b]} \;\equiv\; 0$ oder periodische Fortsetzung von $w|_{[a,b]}$,

$h := \frac{b-a}{n},\ n \in \mathbb{N},\ \overline{w}(x) := \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} w(s)\,ds,\ x_j := a + j\,h,\ j = 0, \ldots, n;$

Beh.: $$\mathrm{TV}\left(L^1(\circ, \overline{w})|_{[a,b]}\right) \;\le\; \mathrm{TV}\left(w|_{[a,b]}\right).$$

Beweis:

Mit Def. 1.2 gilt nun für $x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$:

$$Q^1(x, W) \;=\; W(x_{j-\frac{1}{2}}) \;+\; W[j - \tfrac{1}{2}, j + \tfrac{1}{2}]\,(x - x_{j-\frac{1}{2}}),$$

also mit (2.3)

$$\tfrac{d}{dx}\, Q^1(x, W) \;=\; W[j - \tfrac{1}{2}, j + \tfrac{1}{2}] \;=\; \overline{w}(x_j) \;=\; const.$$

Daraus folgt:

$$\begin{aligned}
\mathrm{TV}\left(L^1(\circ, \overline{w})|_{[a,b]}\right) \;&=\; \sum_{j=1}^{n} \Big|\, L^1(x_{j-\frac{1}{2}} + 0, \overline{w}) - L^1(x_{j-\frac{1}{2}} - 0, \overline{w}) \,\Big| \\
&=\; \sum_{j=1}^{n-1} |\, \overline{w}(x_j) - \overline{w}(x_{j-1}) \,| \le\; \mathrm{TV}\left(\overline{w}|_{[a,b]}\right).
\end{aligned}$$

Mit La. 2.2 folgt somit die Behauptung.

■

Eine weitere Aussage bzgl. Bedingung (2.2.c) liefert:

Lemma 2.4:

Vor.: $a, b \in \mathbb{R},\ a < b,\ w : \mathbb{R} \to \mathbb{R},\ w|_{[a,b]} \in C^{N+1},\ N \ge 2,$

$w\big|_{\mathbb{R}\setminus[a,b]} \;\equiv\; 0$ oder periodische Fortsetzung von $w|_{[a,b]}$;

Beh.: $$\mathrm{TV}\left(L^N(\circ, \overline{w})|_{[a,b]}\right) \;\le\; \mathrm{TV}\left(w|_{[a,b]}\right) \;+\; O\left(h^{N-1}\right).$$

Beweis:

Die Behauptung folgt unmittelbar aus der Definition von $L^N(\circ, \overline{w})$ und S. 1.2.

■

In Kapitel 1 wurde das Oszillationsverhalten von $Q^N(\circ, v)$ ausführlich untersucht. Dabei war $v : \mathbb{R} \to \mathbb{R}$ eine stückweise glatte und Lipschitz–stetig fortsetzbare Funktion mit endlich vielen *Sprüngen*. Die dort hergeleitete Theorie läßt sich wegen folgendem Lemma nicht so ohne weiteres auf $Q^N(\circ, W)$ und somit auf $L^N(\circ, \overline{w})$ übertragen:

Lemma 2.5:

Vor.: $a \in \mathbb{R},\ N \in \mathbb{N},\ w \in SCL(N, \mathbb{R})$;

Beh.: $W(x) := \int_a^x w(s)\, ds$ ist auf $\mathbb{R}$ stetig und weg von den Sprungstellen von w $(N+1)$–mal differenzierbar.

Beweis:

O.B.d.A. sei $\overline{x}$ die einzige Sprungstelle von w. Dann gilt für $x \in \mathbb{R}$:

$$\lim_{h \to 0} [\, W(x+h) - W(x)\,] \;=\; \lim_{h \to 0} \int_x^{x+h} w(s)\, ds \;=\; 0.$$

Daraus folgt die Stetigkeit von W. Da insb. $W'(x) = w(x)$ für $x \neq \overline{x}$, gilt ferner, daß W weg von $\overline{x}$ $(N+1)$–mal differenzierbar ist.

■

Da W auf ganz $\mathbb{R}$ stetig ist, lassen sich S. 1.1 und F. 1.2 nicht auf W anwenden. Wegen S. 1.3 gibt es aber berechtigte Hoffnung, daß sich ähnliche Aussagen in dem hier vorliegenden Fall herleiten lassen.

2.1.3.2 RD–Methode

Im Gegensatz zu der RP–Methode wird w hier nicht mittels einer Stammfunktion rekonstruiert, sondern durch Approximation der ν–ten Ableitungen , $\nu = 0, \ldots, N$ von w. Hierbei steht RD für Reconstruction via Deconvulotion. Eine Motivation für diese Bezeichnung findet sich in [ENO3], S. 23. Die im folgenden beschriebene Methode ist wesentlich aufwendiger zu konstruieren als die RP–Methode. Sie setzt sich aus fünf Schritten zusammen:

(2.4.a) Setze $\overline{w}(x_j) := \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} w(x)\, dx, \qquad h > 0;$

(2.4.b) Bestimme $Q^N(\circ, \overline{w})$ gemäß Def. 1.2.b mit $Q^N(x_j, \overline{w}) = \overline{w}(x_j)$ für $j \in \mathbb{Z}$;

(2.4.c) Setze für $j \in \mathbb{Z},\ r = 0, \ldots, N$:

$$m_{r,j} := \text{minmod} \left[\left(\tfrac{d}{dx}\right)^r Q^N(x_j - 0, \overline{w}), \left(\tfrac{d}{dx}\right)^r Q^N(x_j + 0, \overline{w}) \right] \text{ mit}$$

$$\text{minmod}(x, y) := \left\{ \begin{array}{ll} s \ \min\{|x|, |y|\} & ,\ \text{sgn}\ x = \text{sgn}\ y = s \\ 0 & ,\ \text{sonst} \end{array} \right\};$$

(2.4.d) $A := (a_{i,j})_{i,j=0}^N$ mit $a_{i,j} := \left\{ \begin{array}{ll} \dfrac{1 - (-1)^{j-i+1}}{2^{j-i+1}\, (j-i+1)!} & ,\ i \leq j \\ 0 & ,\ i > j \end{array} \right\}$ und

$$\left(a_{i,j}^{(-1)}\right)_{i,j=0}^N := A^{-1} \text{ gemäß La. 2.6;}$$

(2.4.e) $L^N(x, \overline{w}) := \sum\limits_{r=0}^N \sum\limits_{k=r}^N a_{r,k}^{(-1)}\, m_{k,j}\, \frac{h^{k-r}}{r!}\, (x - x_j)^r, \qquad x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}).$

Aus dieser Konstruktion ist nicht sofort ersichtlich, daß die Bedingungen (2.2.a,b) erfüllt sind. Dazu Bedarf es einer ausführlicheren Untersuchung. Zunächst soll gezeigt werden, daß die Umkehrmatrix aus (2.4.d) existiert.

Lemma 2.6:

Vor.: A gemäß (2.4.d) definiert;
Beh.: A^{-1} existiert und es gilt für $0 \le i,k \le N$:

$$a^{(-1)}_{i,k} = \left\{ \begin{array}{ll} 0 & , \ i > k \\ 1 & , \ i = k \\ - \sum\limits_{l=i+1}^{k} a^{(-1)}_{l,k} \ a_{i,l} & , i = k-1, \ldots, 0 \end{array} \right\}.$$

Beweis:

Da A insbesondere eine obere Dreiecksmatrix ist, gilt:

$$\det A \;=\; \prod_{i=0}^{N} a_{i,i} \;=\; 1,$$

also existiert A^{-1}.

Für die Umkehrmatrix gilt bekanntlich: $A^{-1}A = Id$. Dies läßt sich auch schreiben als $A\ \underline{a}_k^{-1} = \underline{e}_k,\ k = 0, \ldots, N$ wobei a_k^{-1} die k–te Spalte von A^{-1} und $\underline{e}_k$ den k–ten Einheitsvektor des $\mathbb{R}^{N+1}$ bezeichnet. Da A eine obere Dreiecksmatrix ist, erhält man die Lösung dieser linearen Gleichungssysteme unmittelbar durch Rückwärtseinsetzen für $k = 0, \ldots, N$:

$$a^{(-1)}_{N,k} = \ldots = a^{(-1)}_{k+1,k} = 0,\ a^{(-1)}_{k,k} = 1,\ a^{(-1)}_{i,k} = - \sum_{l=i+1}^{k} a^{(-1)}_{l,k}\ a_{i,l},\ i = k-1, \ldots, 0.$$

■

Somit ist $L^N(\circ, \overline{w})$ gemäß (2.4.e) wohldefiniert. Zum Nachweis der Approximationsordnung (2.2.a) nun folgendes

Lemma 2.7:

Vor.: $w \in SCL(N+1, \mathbb{R}),\ N \in \mathbb{N},\ h > 0,$
$w|_{(x-(N+1)h, x+(N+1)h)} \in C^{N+1},\ j \in \mathbb{Z},\ x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}});$
Beh.: a) $m_{r,j} = \left(\frac{d}{dx}\right)^r \overline{w}(x_j) + O\left(h^{N+1-r}\right),\ r = 1, \ldots, N;$

insbesondere: $m_{0,j} = \overline{w}(x_j);$

b) $h^r\ \left(\frac{d}{dx}\right)^r \overline{w}(x) = \sum\limits_{k=0}^{N-r} h^{r+k}\ \left(\frac{d}{dx}\right)^{r+k} w(x)\ a_{r,r+k}\ +\ O\left(h^{N+1}\right), \quad r = 0, \ldots, N;$

c) $h^r\ \left(\frac{d}{dx}\right)^r w(x) = \sum\limits_{k=r}^{N} h^k\ \left(\frac{d}{dx}\right)^k \overline{w}(x)\ a^{(-1)}_{r,k}\ +\ O\left(h^{N+1}\right), \quad r = 0, \ldots, N.$

Beweis:

Zu a): Sei $Q^N(\circ,\overline{w})$ das gemäß Def. 1.2 konstruierte ENO–Polynom. Mit La. 1.5 gilt dann für $\tilde{x} \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$:

$$\left(\tfrac{d}{dx}\right)^r Q^N(\tilde{x},\overline{w}) = \left(\tfrac{d}{dx}\right)^r \overline{w}(\tilde{x}) \;+\; O\left(h^{N+1-r}\right), \quad r = 0,\ldots,N.$$

Aufgrund der Interpolationsbedingung von $Q^N(\circ,\overline{w})$ gilt ferner:

$$m_{0,j} \;=\; Q^N(x_j,\overline{w}) \;=\; \overline{w}(x_j).$$

Wegen (2.4.c) gilt somit die Behauptung a).

Zu b): Sei nun $s \in [-\frac{1}{2},\frac{1}{2}]$. Dann gilt für $r = 0,\ldots,N$:

$$\text{(1)} \qquad \left(\tfrac{d}{dx}\right)^r (w(x+s\,h)) \;=\; \left(\tfrac{d}{dx}\right)^r w(z)\Big|_{z=x+s\,h}.$$

Mittels Taylorentwicklung um x ergibt sich somit:

$$\begin{aligned}
\left(\tfrac{d}{dx}\right)^r \overline{w}(x) \;&=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\tfrac{d}{dx}\right)^r (w(x+s\,h))\; ds \\[2ex]
&\overset{(1)}{=} \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{k=0}^{N-r} \left(\tfrac{d}{dx}\right)^{k+r} w(x)\frac{(s\,h)^k}{k!} \;+\; O\left(h^{N+1-r}\right) ds \\[2ex]
&= h^{-r}\left(\sum_{k=0}^{N-r} \left(\tfrac{d}{dx}\right)^{k+r} w(x)\; h^{k+r}\, a_{0,k} \;+\; O\left(h^{N+1}\right)\right).
\end{aligned}$$

Da $a_{0,k} = a_{r,r+k}$ für $r = 0,\ldots,N$, erhält man b).

Zu c): In Matrixschreibweise läßt sich b) auch darstellen als

$$\left(h^r\;\left(\tfrac{d}{dx}\right)^r \overline{w}(x)\right)^T_{r=0,\ldots,N} \;=\; A\;\left(h^r\;\left(\tfrac{d}{dx}\right)^r w(x)\right)^T_{r=0,\ldots,N} \;+\; O\left(h^{N+1}\right).$$

Durch Multiplikation von A^{-1} folgt c). Beachte dabei, daß A^{-1} unabhängig von h ist und somit komponentenweise gilt:

$$A^{-1}\,O\left(h^{N+1}\right) \;=\; O\left(h^{N+1}\right).$$

■

Mit Hilfe obiger Lemmata lassen sich nun die Bedingungen (2.2.a,b) für die RD–Rekonstruktion nachweisen:

Lemma 2.8:

Vor.: $w \in SCL(N+1, \mathbb{R})$, $N \in \mathbb{N}$, $h > 0$,

$w|_{(x-(N+1)h, x+(N+1)h)} \in C^{N+1}$, $j \in \mathbb{Z}$, $x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$;

Beh.: a) $w(x) \;=\; L^N(x, \overline{w}) \;+\; O\left(h^{N+1}\right)$;

b) $\overline{L}^N(x_j, \overline{w}) \;=\; \overline{w}(x_j)$.

Beweis:

Zu a): Sei $x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$. Wegen der Voraussetzung ist w $N+1$–mal differenzierbar in einer Umgebung von x. Somit gilt:

$$
\begin{aligned}
w(x) \quad &= \quad \sum_{r=0}^{N} \left(\tfrac{d}{dx}\right)^r w(x_j)\, \frac{(x-x_j)^r}{r!} \;+\; O\left(h^{N+1}\right) \\
&\overset{(\mathrm{L.2.7.c})}{=} \quad \sum_{r=0}^{N} h^{-r} \sum_{k=r}^{N} \left(\tfrac{d}{dx}\right)^k \overline{w}(x_j)\; h^k\; a_{r,k}^{(-1)} \frac{(x-x_j)^r}{r!} \;+\; O\left(h^{N+1}\right) \\
&\overset{(\mathrm{L.2.7.a})}{=} \quad \sum_{r=0}^{N} h^{-r} \sum_{k=r}^{N} h^k\; a_{r,k}^{(-1)}\; m_{k,j}\; \frac{(x-x_j)^r}{r!} \;+\; O\left(h^{N+1}\right) \\
&\overset{(2.4.\mathrm{e})}{=} \quad L^N(x, \overline{w}) \;+\; O\left(h^{N+1}\right).
\end{aligned}
$$

Somit ist a) gezeigt.

Zu b): Hier kann $w : \mathbb{R} \to \mathbb{R}$ beliebig zugelassen werden. Mit (2.4.e) bzw.(2.2.b) gilt dann:

$$
\overline{L}^N(x_j) \;=\; \sum_{r=0}^{N} \sum_{k=r}^{N} a_{r,k}^{(-1)}\; m_{k,j}\; \frac{h^{k-r}}{r!}\; \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} (x-x_j)^r \, dx.
$$

Ferner gilt:

$$
\frac{1}{r!\; h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} (x-x_j)^r \, dx \;=\; \frac{h^r}{(r+1)!\; 2^{r+1}} \; \left[1-(-1)^{r+1}\right] \;\overset{(2.4.\mathrm{d})}{=}\; h^r\; a_{0,r}.
$$

Somit erhält man:

$$
\begin{aligned}
\overline{L}^N(x_j, \overline{w}) \quad &= \quad \sum_{r=0}^{N} \sum_{k=r}^{N} a_{r,k}^{(-1)}\; m_{k,j}\; h^k\; a_{0,r} \;\overset{(\mathrm{L.2.6})}{=}\; \sum_{k=0}^{N} m_{k,j}\; h^k \left(\sum_{r=0}^{N} a_{0,r}\; a_{r,k}^{(-1)} \right) \\
&\overset{(\mathrm{L.2.6})}{=} \quad \sum_{k=0}^{N} m_{k,j}\; h^k\; \delta_{0,k} \;=\; m_{0,j} \;\overset{(\mathrm{L.2.7.a})}{=}\; \overline{w}(x_j).
\end{aligned}
$$

■

2.1.3.3 Vergleich der Rekonstruktionsmethoden

Im weiteren Verlauf bezeichne $L^{N,RP}(\circ,\overline{w})$ die RP– und $L^{N,RD}(\circ,\overline{w})$ die RD–Rekonstruktionsmethode. Aufgrund der Bedingungen aus (2.2), die an die obigen Methoden geknüpft wurden, liegt die Vermutung nahe, daß diese sehr ähnlich sind. So folgt mit La. 2.1 sowie La. 2.8 unmittelbar für solche $x \in \mathbb{R}$, in deren Umgebung w glatt ist:

$$L^{N+1,RP}(x,\overline{w}) \;-\; L^{N,RD}(x,\overline{w}) \;=\; O\left(h^{N+1}\right), \quad N \geq 0.$$

Im folgenden soll nun untersucht werden, wann die beiden Methoden identisch sind. Dazu zunächst:

Lemma 2.11:

Beh.: Für $w : \mathbb{R} \to \mathbb{R}$ beliebig gilt: $L^{1,RP}(x,\overline{w}) \;\equiv\; L^{0,RD}(x,\overline{w})$.

Beweis:

Mit (2.4.e) gilt für $x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ zum einen

$$L^{0,RD}(x,\overline{w}) \quad \overset{(2.4.e)}{=} \quad a_{0,0}^{(-1)}\, m_{0,j} \quad \overset{(L.2.6)}{=} \quad m_{0,j} \quad \overset{(L.2.7)}{=} \quad \overline{w}(x_j)$$

und zum anderen (vgl. Beweis zu La. 2.3)

$$L^{0,RD}(x,\overline{w}) \quad = \quad \overline{w}(x_j).$$

Daraus folgt unmittelbar die Behauptung.

■

Da Polynome vom Grad $\leq N$ bekanntlich durch Interpolationspolynome vom Grad N identisch interpoliert werden, liegt folgende Aussage auf der Hand:

Lemma 2.12:

Vor.: $N \in \mathbb{N},\;\; w \in \mathbb{P}_{N-1},\;\; h > 0;$
Beh.: $L^{N,RP}(\circ,\overline{w}) \;\equiv\; L^{N,RD}(\circ,\overline{w}) \;\equiv\; w.$

Beweis:

Da $w \in \mathbb{P}_{N-1}$, gilt offensichtlich: $\overline{w} \in \mathbb{P}_N$ und $W \in \mathbb{P}_N$. Dann gilt:

(1) $$Q^N_{j-\frac{1}{2},h}(\circ,W) \;\equiv\; W \quad \text{und} \quad Q^N_{j,h}(\circ,\overline{w}) \;\equiv\; \overline{w}$$

und somit für die entsprechenden ENO–Splines

$$Q^N_h(\circ,W) \;\equiv\; W \quad \text{und} \quad Q^N_h(\circ,\overline{w}) \;\equiv\; \overline{w}.$$

Für die RP–Rekonstruktion folgt aus (2.3) unmittelbar:

(2) $$L^{N,RP}(x,\overline{w}) \;=\; \tfrac{d}{dx}\, Q^N_h(x,W) \;=\; \tfrac{d}{dx}\, W(x) \;=\; w(x), \quad x \in \mathbb{R}.$$

Für die RD–Rekonstruktion erhält man wegen (1):

$$\left(\tfrac{d}{dx}\right)^r Q_{i,h}^N(\circ,\overline{w}) \;\equiv\; \left(\tfrac{d}{dx}\right)^r \overline{w} \text{ für } r \geq 0,\; j \in \mathbb{Z},\; i = j-1,\; j,$$

also

$$m_{r,j} \;=\; \left(\tfrac{d}{dx}\right)^r w(x_j).$$

Analog zum Beweis zu La. 2.7 gilt:

$$h^r \;\left(\tfrac{d}{dx}\right)^r \overline{w}(x) \;=\; \sum_{k=0}^{N-r} h^{r+k} \;\left(\tfrac{d}{dx}\right)^{r+k} w(x)\, a_{r,r+k} \quad \text{und}$$

$$h^r \;\left(\tfrac{d}{dx}\right)^r w(x) \;=\; \sum_{k=r}^{N} h^{k} \;\left(\tfrac{d}{dx}\right)^{k} \overline{w}(x)\, a_{r,k}^{(-1)}.$$

Hier fallen die $O\left(h^{N+1}\right)$–Terme weg, da $\left(\tfrac{d}{dx}\right)^{N+1} \overline{w} \equiv 0$ ist. Analog zum Beweis zu La. 2.8.a erhält man für $j \in \mathbb{Z}$:

$$L^{N,RD}(\circ,\overline{w}) \;\equiv\; w \quad \text{auf} \quad I_{j-\frac{1}{2}}.$$

Somit gilt:

$$(3) \qquad L^{N,RD}(\circ,\overline{w}) \;\equiv\; w \quad \text{auf } \mathbb{R} \quad .$$

Aus (2) und (3) folgt unmittelbar die Behauptung. ∎

Abschließend sei noch vermerkt, daß aufgrund von La. 2.5 W eine stetige Funktion ist, selbst wenn w eine stückweise glatte und Lipschitz–stetig fortsetzbare Funktion ist. Dies gilt, wie man sich leicht überzeugen kann, auch für $\overline{w}$. In Kapitel 1 wurden die Hauptergebnisse aber — wie schon oben erwähnt — für $Q^N(\circ, v)$ hergeleitet, wobei v eine Funktion mit endlich vielen Sprüngen war. Diese Theorie läßt sich also nicht so ohne weiteres auf $Q^N(\circ, W)$ bzw. $Q^N(\circ,\overline{w})$ übertragen. Wegen La. 1.6 bzw. S. 1.3 werden aber die "verborgenen" Sprungstellen von W bzw. $\overline{w}$ durch die ENO–Interpolationstechnik für h klein genug aufgespürt und somit auch die von w !

2.1.4 Numerischer Fluß

Mittels obiger Rekonstruktionsmethoden läßt sich nun $L^N(\circ, v_j^n)$ aus (2.1.b) bestimmen, in dem man oben $\overline{w}(x_j)$ durch v_j^n ersetzt. Damit ist das Verfahren aber immer noch nicht explizit konstruiert. Als nächstes muß der numerische Fluß $\hat{F}_{j+\frac{1}{2}}$ bestimmt werden, der wie folgt definiert ist:

$$(2.5.a) \qquad \hat{F}_{j+\frac{1}{2}} := \frac{1}{\tau} \int_{t_n}^{t_{n+1}} F(v(x_{j+\frac{1}{2}}, t)\, dt \quad \text{mit } v \text{ Lösung zu}$$

$$(2.5.b) \qquad \begin{aligned} v_t(x,t) \;+\; \left(F(v(x,t))\right)_x \;&=\; 0, \quad (x,t) \in \mathbb{R} \times (t_n, +\infty); \\ v(x,t_n) \;&=\; L^N(x, v_j^n), \qquad x \in \mathbb{R}. \end{aligned}$$

Hier liegt das Problem in der Bestimmung von v. Dies ist im allgemeinen nicht einfacher als die Lösung des Ausgangsproblems zu bestimmen. Man ist daher nur an einer Approximation von hoher Ordnung interessiert, d.h.:

$$(2.6) \qquad \hat{F}_{j+\frac{1}{2}} \stackrel{!}{=} \frac{1}{\tau} \sum_{i=1}^{m} a_i \, F(v(x_{j+\frac{1}{2}}, \tilde{t}_i)) \;+\; O\left(h^{N+1}\right), \; N \in \mathbb{N},$$

wobei mit m die Anzahl der Stützstellen der Quadraturformel, a_i die Gewichte und $\tilde{t}_i$ die Stützstellen in $[t_n, t_{n+1}]$ bezeichnet werden.

Somit ist das Problem auf Bestimmung von *endlich* vielen Funktionswerten reduziert worden. Da $L^N(\circ, v_j^n)$ in $x_{j+\frac{1}{2}}$ im allgemeinen nicht stetig ist, liegt es nahe, die Werte an den Stützstellen durch Lösen von s.g. Riemann–Problemen (vgl. Def. 4.2) zu nähern. Dies soll im folgenden beschrieben werden:

Sei $v|_{(x_{j+\frac{1}{2}}-\varepsilon, x_{j+\frac{1}{2}}+\varepsilon)\times(t_n,t_{n+1})} \in C^{N+1}$ für $\varepsilon > 0$. Durch Taylorentwicklung um t_n erhält man dann:

$$v(x_{j+\frac{1}{2}}, \tilde{t}_i) \;=\; v^N(x_{j+\frac{1}{2}}, \tilde{t}_i) \;+\; O\left(h^{N+1}\right) \quad \text{mit}$$

$$v^N(x_{j+\frac{1}{2}}, \tilde{t}_i) \;:=\; \sum_{i=0}^{N} \left(\tfrac{\partial}{\partial t}\right)^i \, v(x_{j+\frac{1}{2}}, t_n) \, \frac{(\tilde{t}_i - t_n)^i}{i!} \;.$$

Setze nun

$$(2.7) \qquad v_L^N := v^N(x_{j+\frac{1}{2}} - 0, \tilde{t}_i), \quad v_R^N := v^N(x_{j+\frac{1}{2}} + 0, \tilde{t}_i).$$

Damit lassen sich die diskreten Werte aus (2.6) wie folgt approximieren:

$$(2.8) \qquad F(v(x_{j+\frac{1}{2}}, \tilde{t}_i)) \;=\; F\left(V(0, v_L^N, v_R^N)\right) \;+\; O\left(h^{N+1}\right);$$

$$V\left(\frac{x - x_{j+\frac{1}{2}}}{t - \tilde{t}_i}, v_L^N, v_R^N\right) := v(x,t) \quad \text{Lösung des folgenden Riemann–Problems}$$

$$\begin{aligned} v_t(x,t) \;+\; \left(F(v(x,t))\right)_x \;&=\; 0, \quad (x,t) \in \mathbb{R} \times (\tilde{t}_i, +\infty); \\ v(x, \tilde{t}_i) \;&=\; \left\{ \begin{array}{l} v_L^N \;,\; x \leq x_{j+\frac{1}{2}} \\ v_R^N \;,\; x > x_{j+\frac{1}{2}} \end{array} \right\}. \end{aligned}$$

Zur Bestimmung von $F\left(V(0, v_L^N, v_R^N)\right)$ gibt es nun mehrere Möglichkeiten:

(a) Berechne die exakte Lösung des Riemann–Problems (vgl. Kap. 4.2),

(b) Bestimme eine Näherung des Riemann–Problems mittels der Roe–Linearisierung (vgl. Kap. 3).

Die Methode a) wird zum Beispiel auch beim Godunov–Verfahren verwendet. Somit läßt sich (2.8) auch wie folgt schreiben:

$$
(2.9.a)\quad \begin{cases}
v_L^N > v_R^N : F^{God}\left(\tilde{t}_i, v_L^N, v_R^N\right) = \begin{Bmatrix} F(v_L^N),\ 0 < s \\ \\ F(v_R^N),\ 0 > s \end{Bmatrix},\ s = \dfrac{F(v_L^N) - F(v_R^N)}{v_L^N - v_R^N} \\ \\
v_L^N = v_R^N : F^{God}\left(\tilde{t}_i, v_L^N, v_R^N\right) = F(v_L^N) \\ \\
v_L^N < v_R^N : F^{God}\left(\tilde{t}_i, v_L^N, v_R^N\right) = \begin{Bmatrix} F(v_L^N) & , & 0 < F'(v_L^N) \\ F\left((F')^{[-1]}(0)\right), & & F'(v_L^N) < 0 < F'(v_R^N) \\ F(v_R^N) & , & 0 > F'(v_R^N) \end{Bmatrix}
\end{cases}
$$

$$
(2.9.b)\quad \begin{cases}
F^{Roe}\left(\tilde{t}_i, v_L^N, v_R^N\right) = \dfrac{1}{2}\ \left[F(v_L^N) + F(v_R^N) - \left|\lambda(v_L^N, v_R^N)\right|\ (v_R^N - v_L^N)\right] \text{ mit} \\ \\
\lambda(v_L^N, v_R^N) = \begin{Bmatrix} \dfrac{F(v_L^N) - F(v_R^N)}{v_L^N - v_R^N} & , v_L^N \neq v_R^N \\ 0 & , v_L^N = v_R^N \end{Bmatrix}
\end{cases}
$$

Dann gilt:

$$
(2.10.a)\qquad F(v(x_{j+\frac{1}{2}}, \tilde{t}_i)) \ = \ F^{God}\left(\tilde{t}_i, v_L^N, v_R^N\right),
$$

$$
(2.10.b)\qquad F(v(x_{j+\frac{1}{2}}, \tilde{t}_i)) \ = \ F^{Roe}\left(\tilde{t}_i, v_L^N, v_R^N\right) \ + \ O\left(|v_L^N - v_R^N|\right).
$$

Zur Berechnung von v_L^N, v_R^N gemäß (2.7) ist folgendes Lemma nützlich:

Lemma 2.13:

Vor.: u Lösung eines skalaren Erhaltungssatzes, $i \in \mathbb{N}$, $(x,t) \in \mathbb{R} \times \mathbb{R}_+$,
U Umgebung von (x,t) mit $u \in C^{i+1}(U)$;

Beh.: a) $\left(\frac{\partial}{\partial t}\right)^i u(x,t) \ = \ (-1)^i \left(\frac{\partial}{\partial x}\right)^i \left([F'(u(x,t))]^i \ \frac{\partial}{\partial x} u(x,t)\right)$;

b) $\left(\frac{\partial}{\partial t}\right)^i u(x,t) \ = \ (-1)^i \sum\limits_{k=0}^{i-1} \binom{i-1}{k} \left(\frac{\partial}{\partial x}\right)^k \left([F'(u(x,t))]^i\right) \left(\frac{\partial}{\partial x}\right)^{i-k} u(x,t)$.

Beweis:

Teil a) wird in [Ad1] gezeigt und Teil b) folgt aus a) durch $(i-1)$–malige Anwendung der Produktregel.

■

Daraus läßt sich eine explizite Berechnung von (2.7) ableiten:

Folgerung 2.1:

Vor.: $v|_{(x_{j+\frac{1}{2}}-\varepsilon, x_{j+\frac{1}{2}}+\varepsilon)\times(t_n,t_{n+1})} \in C^{N+1},\ \varepsilon > 0, N \in \mathbb{N},\ v$ Lösung zu (2.5.b);

Beh.: a) $v^N(x_{j+\frac{1}{2}} \pm 0, \tilde{t}_i) \ = \ v(x_{j+\frac{1}{2}} \pm 0, \tilde{t}_i) \ + \ O\left(h^{N+1}\right)$;

b) $\left(\frac{\partial}{\partial x}\right)^i \ v(x_{j+\frac{1}{2}} \pm 0, t_n) \ = \ \left(\frac{\partial}{\partial x}\right)^i \ L^N(x_{j+\frac{1}{2}} \pm 0, v_j^n),\ i = 0, \ldots, N$.

In [ENO3] wurde statt der hier verwendeten Taylorentwicklung in t eine s.g. Cauchy–Kowaleski–Prozedur, d.h. Taylorentwicklung in x *und* t, benutzt. Diese hat aber den Nachteil, daß gemischte Ableitungen der Form

$$\frac{\partial^{i+j}}{\partial^i x \ \partial^j t} \ v(x,t)$$

auftreten. Durch die Entwicklung in der einen oder anderen Form kann es ferner zu Ordnungsverlusten kommen, denn:

Da $L^N(\circ, v_j^n)$ in $x_{j+\frac{1}{2}}$ im allgemeinen nicht stetig ist, können bei der Lösung v aus (2.5.b) Diskontinuitätskurven in $(x_{j+\frac{1}{2}}, t_n)$ entspringen. Falls $L^N(\circ, v_j^n)$ nicht stückweise konstant ist, müssen diese nicht notwendig linear verlaufen. Sie kann also z.B. folgenden Verlauf haben:

Das hat zur Folge, daß eine der beiden Taylorentwicklungen aus (2.7) entlang der Zeitachse $\{x_{j+\frac{1}{2}}\} \times [t_n, +\infty)$ über eine Diskontinuität gemacht wird.

Diesen Ordnungsverlust kann man durch eine asymptotische Entwicklung der Lösung v aus (2.5.b) vermeiden. Darunter versteht man folgende Darstellung von v:

(2.11)
$$v(x,t) \ = \ \sum_{j=0}^{N} v^j\left(\tfrac{x}{t}\right) \ t^j \ + \ O\left(t^{N+1}\right).$$

Dabei sind die v^j Funktionen, die nur von $\frac{x}{t}$ abhängen (s. Kap. 4).

Somit lassen sich die diskreten Werte aus (2.6) auch folgendermaßen approximieren:

(2.12)
$$F^{Rav}(\tilde{t}_i) \ = \ F\left(\sum_{j=0}^{N} v^j(0)\,(\tilde{t}_i - t_n)^j\right).$$

Dann gilt, da $\tau = \lambda\ h,\ \lambda = const$:

(2.13)
$$F(v(x_{j+\frac{1}{2}}, \tilde{t}_i) \ = \ F^{Rav}(\tilde{t}_i) \ + \ O\left(h^{N+1}\right).$$

Näheres zu dieser Theorie findet sich in Kapitel 4.

Während man bei F^{God} bzw. F^{Roe} zum erreichen der hohen Ordnung der Approximation von (2.6) voraussetzen mußte, daß v hinreichend glatt ist entlang der Zeitachse $\{x_{j+\frac{1}{2}}\} \times [t_n, +\infty)$ (vgl. (2.10)), gilt die analoge Aussage für F^{Rav} (vgl. (2.13)) *unabhängig* von der Glattheit von v . Ein weiterer Unterschied ergibt sich aus der Praxis, denn:
Bei der Verwendung von (2.10) hat man ähnlich wie beim Godunov–Verfahren für die CFL–Zahl im allgemeinen folgende Schranke:

$$\lambda \ \max\{\, F'(u) \mid |u| \leq \|u_0\|_\infty \} \ < \ \frac{1}{2}.$$

Numerische Beispiele zeigen (vgl. Kap. 7), daß man unter Benutzung von (2.13) mit einer fast doppelt so großen CFL–Zahl arbeiten kann.
Für $\hat{F}_{j+\frac{1}{2}}$ aus (2.5) erhält man nun folgende Approximationen:

$$F^{God}_{j+\frac{1}{2}} \ := \ \frac{1}{\tau} \sum_{i=1}^{m} a_i \ F^{God}(\tilde{t}_i, v_L^N, v_R^N); \tag{2.14.a}$$

$$F^{Roe}_{j+\frac{1}{2}} \ := \ \frac{1}{\tau} \sum_{i=1}^{m} a_i \ F^{Roe}(\tilde{t}_i, v_L^N, v_R^N); \tag{2.14.b}$$

$$F^{Rav}_{j+\frac{1}{2}} \ := \ \frac{1}{\tau} \sum_{i=1}^{m} a_i \ F^{Rav}(\tilde{t}_i). \tag{2.14.c}$$

2.1.5 Ordnungsuntersuchung

In diesem Abschnitt soll nun die Qualität der Verfahren im glatten Bereich der Lösung des zu approximierenden Erhaltungssatzes untersucht werden. Ein Maß dafür ist der folgende Begriff:

Definition 2.11:

Sei u starke Lösung eines Erhaltungssatzes in einer Umgebung von $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ und $v_m^{n+1} := Q\left(v_{m-q}^n, \ldots, v_{m+p}^n\right)$,$q,p \in \mathbb{N}_0$ ein Verfahren. Ferner sei $\lambda = \frac{\tau}{h} = const.$
Man sagt:

a) Das Verfahren hat in (x,t) die *(lokale) Ordnung N*, falls gilt:

$$u(x,t+\tau) \ - \ Q\left(u(x-q\,h,t), \ldots, u(x+p\,h,t)\right) \ = \ O\left(h^{N+1}\right) \quad \text{für } \tau \to 0;$$

b) Das Verfahren hat in (x,t) die *(lokale) Ordnung N bzgl. der Mittelwerte*, falls gilt:

$$\overline{u}(x,t+\tau) \ - \ Q\left(\overline{u}(x-q\,h,t), \ldots, \overline{u}(x+p\,h,t)\right) \ = \ O\left(h^{N+1}\right) \quad \text{für } \tau \to 0$$

$$\text{mit} \quad \overline{u}(x,t) \ := \ \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} u(s,t)\, ds. \tag{2.15}$$

Die Standarddefinition für den Ordnungsbegriff ist a). Da die ENO–Verfahren gemäß Konstruktion die Mittelwerte der exakten Lösung approximieren, ist die Definition b) sinnvoll. Wie die beiden Begriffe zusammenhängen, wird aus folgendem Lemma ersichtlich:

Lemma 2.14:

Vor.: Sei u starke Lösung eines Erhaltungssatzes in einer Umgebung U von $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ mit $u \in C^{N+1}(U)$, $N \geq 2$;

Beh.: Für h hinreichend klein gilt:

$$\overline{u}(x,t) \;=\; u(x,t) \;+\; \sum_{i=1}^{\lceil \frac{N}{2}-1 \rceil} \left(\tfrac{\partial}{\partial x}\right)^{2\,i} \; u(x,t)\, \tfrac{1}{(2\,i+1)!}\; \left(\tfrac{h}{2}\right)^{2\,i} \;+\; O\left(h^N\right).$$

Beweis:

Sei $U(\circ,t)$ Stammfunktion zu $u(\circ,t)$. Aufgrund der Voraussetzung gilt für $h, \varepsilon > 0$ hinreichend klein:

$$u\big|_{(x-\frac{h}{2},x+\frac{h}{2})\times(t-\varepsilon,t+\varepsilon)} \in C^{N+1} \qquad \text{bzw.} \qquad U\big|_{(x-\frac{h}{2},x+\frac{h}{2})\times(t-\varepsilon,t+\varepsilon)} \in C^{N+2}.$$

Mit (2.15) gilt somit:

$$\overline{u}(x,t) \;=\; \tfrac{1}{h}\;\left[\, U(x+\tfrac{h}{2},t) - U(x-\tfrac{h}{2},t) \,\right].$$

Mittels Taylorentwicklung um x erhält man dann weiter:

$$\overline{u}(x,t) \;=\; \tfrac{1}{h}\;\left\{ \sum_{i=0}^{N} \left(\tfrac{\partial}{\partial x}\right)^{i}\; U(x,t)\, \tfrac{1}{i!}\; \left(\tfrac{h}{2}\right)^{i}\; \left[1-(-1)^i\right] \;+\; O\left(h^{N+1}\right) \right\}.$$

Da insbesondere $\frac{d}{dx}U(\circ,t) \;\equiv\; u(\circ,t)$ ergibt sich unter Berücksichtigung der Tatsache, daß die Terme für die geraden Indizes verschwinden, die Behauptung.

■

Daraus folgt unmittelbar:

Folgerung 2.2:

Ein Verfahren approximiert die Mittelwerte der exakten Lösung von 2. Ordnung genau dann, wenn es die exakte Lösung von 2. Ordnung approximiert.

Als nächstes soll nun untersucht werden, von welcher Ordnung die Mittelwerte der exakten Lösung aproximiert werden:

Satz 2.3:

Vor.: $Q\left(v_{j-q}^n, \ldots, v_{j+p}^n\right) \;:=\; v_j^{n+1}$ gemäß (2.1), $j \in \mathbb{Z}$, $p, q, n \in \mathbb{N}_0$, $N \in \mathbb{N}$;
$L^N(\circ, v_j^n)$ erfülle (2.2.a,b);
u starke Lösung eines Erhaltungssatzes in einer Umgebung U von $(x,t) \in \mathbb{R}\times\mathbb{R}_+$ mit $u \in C^{N+1}(U), F \in C^1$;

Beh.: Für $\tau \to 0$ und h hinreichend klein mit $\lambda = \frac{\tau}{h} = const$ gilt :

$$\overline{u}(x,t+\tau) \;-\; Q\left(\overline{u}(x-q\,h,t), \ldots, \overline{u}(x+p\,h,t)\right) \;=\; O\left(h^{N+1}\right).$$

Beweis:

Für τ bzw. h hinreichend klein gilt wegen (2.2.a):

$$L^N(x, \overline{u}(\circ, t)) \;=\; u(x,t) \;+\; O\left(h^{N+1}\right),\; x \in \mathbb{R}.$$

Somit gilt für die Lösung v aus (2.1.b) und $j = -1, 0$:

$$v\left(x + \left(j + \tfrac{1}{2}\right)\; h, \overline{t}\right) \;=\; u\left(x + \left(j + \tfrac{1}{2}\right)\; h, \overline{t}\right) \;+\; O\left(h^{N+1}\right) \text{ für } \overline{t} \in [t, t+\tau]$$

und damit

$$\hat{F}_{j+\frac{1}{2}} \;=\; \frac{1}{\tau} \int_t^{t+\tau} F\left(u\left(x + \left(j + \tfrac{1}{2}\right)\; h, \overline{t}\right)\right) \;+\; O\left(h^{N+1}\right)\; d\overline{t}$$

$$\;=\; \frac{1}{\tau} \int_t^{t+\tau} F\left(u\left(x + \left(j + \tfrac{1}{2}\right)\; h, \overline{t}\right)\right)\; d\overline{t} \;+\; O\left(h^{N+1}\right), \quad j = -1, 0.$$

Also gilt:

$$Q\left(\overline{u}(x - q\,h, t), \ldots, \overline{u}(x + p\,h, t)\right) \;=$$

$$\overline{u}(x,t) \;+\; \lambda\; \left[\; \hat{F}_{+\frac{1}{2}}(u) \;-\; \hat{F}_{-\frac{1}{2}}(u) \right] \;+\; O\left(h^{N+1}\right)$$

$$\text{mit} \quad \hat{F}_{\pm\frac{1}{2}}(u) \;:=\; \frac{1}{\tau} \int_t^{t+\tau} F\left(u\left(x \pm \tfrac{h}{2}, \overline{t}\right)\right)\; d\overline{t}.$$

Da u insbesondere schwache Lösung in einer Umgebung von (x,t), gilt für τ hinreichend klein mit Kor. 2.1 die Behauptung.

■

In obiger Voraussetzung hängen p und q von der Wahl der Rekonstruktionsmethode ab. Für $L^N(\circ, v_j^n) = L^{N,RP}(\circ, v_j^n)$ gilt: $p = q = N$, für $L^N(\circ, v_j^n) = L^{N,RD}(\circ, v_j^n)$ dagegen: $p = q = N + 1$.
Als eine einfache Folgerung erhält man nun aus S. 2.3 bzw. F. 2.2:

Folgerung 2.3:

Die ENO–Verfahren approximieren die Mittelwerte der exakten Lösung von N–ter Ordnung, die exakte Lösung dagegen von 2. Ordnung für $N \geq 2$ sonst von 1. Ordnung.

Beachte, daß sich für $L^N(\circ, v_j^n) = L^{N,RP}(\circ, v_j^n)$ wegen La. 2.1 die Ordnung der Mittelwertapproximation um 1 verringert.
Ferner arbeitet man im allgemeinen nicht mit $\hat{F}_{j+\frac{1}{2}}$, sondern mit einer Approximation mit Fehler $O\left(h^{N+1}\right)$ (vgl. Kap. 2.1.4). Dies ist aber für die Ordnungsuntersuchung unerheblich, denn:

Lemma 2.15:

Vor.: $Q\left(v_{j-q}^n, \ldots, v_{j+p}^n\right) \;:=\; v_j^{n+1}$ gemäß (2.1), $j \in \mathbb{Z},\; p, q, n \in \mathbb{N}_0,\; N \in \mathbb{N}$;

$L^N(\circ, v_j^n)$ erfülle (2.2.a,b); h hinreichend klein;

u starke Lösung eines Erhaltungssatzes in einer Umgebung U von $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ mit $u \in C^{N+1}(U), F \in C^1$;

Beh.: a) $F_{j+\frac{1}{2}}^{God} \;=\; \hat{F}_{j+\frac{1}{2}} \;+\; O\left(h^{N+1}\right)$;

b) $F_{j+\frac{1}{2}}^{Roe} \;=\; \hat{F}_{j+\frac{1}{2}} \;+\; O\left(h^{N+1}\right)$;

c) $F_{j+\frac{1}{2}}^{Rav} \;=\; \hat{F}_{j+\frac{1}{2}} \;+\; O\left(h^{N+1}\right)$.

Beweis:

Die Behauptung folgt unmittelbar aus (2.14) unter Beachtung von (2.10) bzw. (2.12) und (2.6).

■

Abschließend sei noch vermerkt, daß man durch eine Modifikation der ENO–Schemata ein Verfahren erhält, daß die exakte Lösung im glatten Bereich von N–ter Ordnung approximiert:

(2.16.a) Setze $\overline{v}_j^0 := \overline{u}_0(x_j)$;

(2.16.b) Sei $\overline{v}_j^n$ Approximation der Mittelwerte der exakten Lösung $u(x_j, t_n)$, d.h. Näherungen für $\overline{u}_j^n$ und v_j^n Approximation der exakten Lösung $u(x_j, t_n)$. Die Verfahrenswerte v_j^{n+1} auf der neuen Zeitstufe berechnen sich dann wie folgt:

α)

$$\overline{v}_j^{n+1} \;:=\; \overline{v}_j^n \;-\; \lambda\,\left[\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}}\right] \qquad \text{mit}$$

$$\hat{F}_{j+\frac{1}{2}} := \frac{1}{\tau}\int_{t_n}^{t_{n+1}} F(v(x_{j+\frac{1}{2}}, t))\, dt \quad \text{mit } v \text{ Lösung zu}$$

$$v_t(x,t) \;+\; (F(v(x,t)))_x \;=\; 0 \quad (x,t) \in \mathbb{R} \times (t_n, +\infty);$$

$$v(x, t_n) \;=\; L^N(x, \overline{v}_j^n) \qquad x \in \mathbb{R};$$

β)

$$v_j^{n+1} \;:=\; L^N(x_j, \overline{v}_j^{n+1}).$$

Die Modifikation besteht also nur darin, daß man nach der Approximation der Mittelwerte auf der neuen Zeitstufe — analog zum ENO–Verfahren — die Rekonstruktion zu diesen Werten bildet und an der Stelle x_j auswertet.

Dies ist kein wesentlicher Mehraufwand, da $L^N(\circ, \overline{v}_j^{n+1})$ beim ENO–Verfahren sowieso bestimmt werden muß. Der Nachteil ist, daß das modifizierte Verfahren *nicht* in Erhaltungsform ist.

2.1.6 Eigenschaften der ENO–Verfahren

Nachdem die Verfahren nun ausführlich erklärt wurden und auch das Ordnungsverhalten näher betrachtet wurde, sollen jetzt einige weitere Eigenschaften der ENO–Schemata untersucht werden. Im Vordergrund steht dabei die Frage, ob diese nun die geforderte ENO–Bedingung gemäß Def. 2.10 auch wirklich erfüllen. Dazu zeigt man:

(a) Die Entropielösung (vgl. Def. 2.14) ist unter gewissen Voraussetzungen "stückweise glatt";

(b) Die Entropielösung ist TVD, d.h. die Oszillation nimmt mit der Zeit nicht zu;

(c) Es gilt: $\mathrm{TV}\left(v_j^{n+1}\right) \leq \ \mathrm{TV}\left(L^N(\circ, v_j^n)\right)$.

Zunächst soll erklärt werden, was man im folgenden unter "stückweise glatt" zu verstehen hat:

Definition 2.12:

Sei u Lösung eines skalaren Erhaltungssatzes mit konvexem Fluß F. u heiß *stückweise glatt*, falls es eine höchstens abzählbare Familie Γ von s.g. glatten Diskontinuitätskurven in $\mathbb{R} \times \mathbb{R}_+$ gibt, die folgende Eigenschaften (2.17) erfüllen:

(a) jede kompakte Teilmenge von $\mathbb{R} \times \mathbb{R}_+$ wird nur von endlich vielen Kurven aus Γ geschnitten;

(b) die Kurven können endliche oder unendliche Länge haben;

(c) $u \in C^\infty(\mathbb{R} \times \mathbb{R}_+ \backslash \Gamma)$;

(d) eine Diskontinuitätskurve endet nur dann, wenn sie mit einer anderen zusammenstößt und daraus genau eine neue hervorgeht;

(e) die Kurven sind bis auf gemeinsame Endpunkte disjunkt ;

(f) zwei Kurven haben höchstens einen gemeinsamen Punkt;

(g) $\mathbb{R} \times \mathbb{R}_+ \backslash \Gamma$ ist eine offene Teilmenge von $\mathbb{R} \times \mathbb{R}_+$ und u ist entlang den Kurven aus Γ einseitig stetig fortsetzbar.

Wegen Bedingung (2.17.c) werden auch die Ränder von Verdünnungswellen als Diskontinuitätskurven aufgefaßt, denn u ist zwar stetig aber nicht differenzierbar entlang diesen Kurven. Wie aus folgendem Lemma ersichtlich wird, stellt dies kein Widerspruch zu (2.17.e) dar, denn Verdünnungswellen können nur zum Zeitpunkt $t = 0$ entstehen. Somit können zwei Kurven aus Γ also nur in $(x, 0),\ x \in \mathbb{R}$ einen gemeinsamen *Anfangspunkt* haben. Aufgrund der Definition von Γ werden diese Punkte aber nicht mehr als zur Kurve gehörend betrachtet. Dazu nun

Lemma 2.16:

Vor.: $u_0 \in L^\infty(\mathbb{R}),\ F \in C^2(\mathbb{R}), F$ konvex, u Entropielösung von

$$u_t(x,t) \ + \ \left(F(u(x,t))\right)_x \ = \ 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+;$$

$$u(x,0) \ = \ u_0(x), \quad x \in \mathbb{R};$$

Beh.: Beim Zusammenstoß von Schock– bzw. Verdünnungswellen mit genau einer anderen Schock– bzw. Verdünnungswelle entsteht genau ein neuer Schock.

Beweis:

Wie schon in Kapitel 2.1.1 bemerkt, setzt sich die Entropielösung aus Bereichen zusammen, in denen sie glatt ist. Diese werden durch Verdünnungswellen bzw. Diskontinuitätskurven getrennt. Dabei kann es zu folgenden Kollisionsproblemen kommen:

(a) zwei Schockkurven laufen ineinander;
(b) zwei Verdünnungswellen laufen ineinander;
(c) ein Schock läuft von rechs in eine Verdünnungswelle;
(d) ein Schock läuft von links in eine Verdünnungswelle.

Dies sieht wie folgt aus:

Fall a)

Fall b)

Fall c)

Fall d)

Falls eine Diskontinuitätskurve $\varphi(t), 0 \leq t \leq T, T > 0$ in (o.B.d.A.) $\varphi(0) = 0$ entspringt, so setzt sich die Entropielösung im (x,t)–Raum aus den glatten Teillösungen $u_L(\circ,\circ)$ und $u_R(\circ,\circ)$ wie folgt für $0 \leq t \leq T$ zusammen:

$$u(x,t) = \begin{Bmatrix} u_L(x,t) \ , x < s\,t \\ \\ u_R(x,t) \ , x > s\,t \end{Bmatrix} \text{ mit } s = \frac{d}{dt}\,\varphi = \left.\frac{F(u_L) - F(u_R)}{u_L - u_R}\right|_{(\varphi(t),t)}, \ t > 0,$$

und es gilt entlang $\varphi(t): \ u_L(\varphi(t),t) \ > \ u_R(\varphi(t),t), \ t \geq 0.$
Hat dagegen eine Verdünnungswelle ihren Ursprung in (o.B.d.A.) $(0,0)$, d.h. es gibt zwei Geraden $\varphi_L(t), \varphi_R(t), \ 0 \leq t \leq T, t > 0$ mit dem einzigen gemeinsamen Punkt

$\varphi_L(0) = \varphi_R(0) = 0$, so hat die Entropielösung für $0 \leq t \leq T$ folgende Gestalt:

$$u(x,t) = \left\{ \begin{array}{ll} u_L(x,t) & , x < F'(u_L(0-,0))\, t \\ \\ (F')^{[-1]}(\frac{x}{t}) & , F'(u_L(0-,0))\, t < x < F'(u_R(0+,0))\, t \\ \\ u_R(x,t) & , x > F'(u_R(0+,0))\, t \end{array} \right\}$$

Beachte dabei: $u_L(\varphi_{L,R}(t),t) \;=\; const$ und $u_L(\varphi_L(t),t) < u_R(\varphi_R(t),t)$.
Damit lassen sich die Teillösungen aus a),...,d) wie folgt für $0 \leq t \leq T,\ T > 0$ zur Gesamtlösung zusammensetzen:

zu a):

$$u(x,t) = \left\{ \begin{array}{ll} u_L(x,t) & , x - x_L < s_L\, t \\ \\ u_M(x,t) & , x - x_L > s_L\, t \text{ und } x - x_R < s_R\, t \\ \\ u_R(x,t) & , x - x_R > s_R\, t \end{array} \right\} \text{ mit}$$

$$s_L = \left.\frac{F(u_L) - F(u_M)}{u_L - u_M}\right|_{(\varphi_L(t),t)} = F'(\xi_L(t)),\ \xi_L(t) \in \Big(u_M(\varphi_L(t),t), u_L(\varphi_L(t),t)\Big),$$

$$s_R = \left.\frac{F(u_M) - F(u_R)}{u_M - u_R}\right|_{(\varphi_R(t),t)} = F'(\xi_R(t)),\ \xi_R(t) \in \Big(u_R(\varphi_R(t),t), u_M(\varphi_R(t),t)\Big);$$

zu b):

$$u(x,t) = \left\{ \begin{array}{ll} u_L(x,t) & , x - x_L < s_L\, t \\ \\ (F')^{[-1]}(\frac{x-x_L}{t}) & , s_L\, t < x - x_L < s_{M,L}\, t \\ \\ u_M(x,t) & , s_{M,L}\, t < x - x_L \text{ und } x - x_R < s_{M,R}\, t \\ \\ (F')^{[-1]}(\frac{x-x_R}{t}) & , s_{M,R}\, t < x - x_R < s_R\, t \\ \\ u_R(x,t) & , x - x_R > s_R\, t \end{array} \right\} \text{ mit}$$

$$s_L = F'(u_L(x_L - 0,0)),\ s_{M,L} = F'(u_M(x_L + 0,0)),$$

$$s_R = F'(u_R(x_R + 0,0)),\ s_{M,R} = F'(u_M(x_R - 0,0));$$

zu c):

$$u(x,t) = \left\{ \begin{array}{ll} u_L(x,t) & , x - x_L < s_L\, t \\ \\ (F')^{[-1]}(\frac{x-x_L}{t}) & , s_L\, t < x - x_L < s_R\, t \\ \\ u_M(x,t) & , s_R\, t < x - x_L \text{ und } x - x_R < s\, t \\ \\ u_R(x,t) & , x - x_R > s\, t \end{array} \right\} \text{ mit}$$

$$s = \left.\frac{F(u_M) - F(u_R)}{u_M - u_R}\right|_{(\varphi(t),t)} = F'(\xi(t)),\ \ \xi(t) \in \Big((u_R(\varphi(t),t), u_M(\varphi(t),t)\Big),$$

$$s_L = F'(u_L(x_L - 0, 0)),\ s_R = F'(u_M(x_R + 0, 0));$$

zu d):

$$u(x,t) = \left\{\begin{array}{ll} u_L(x,t) & , x - x_L < s\,t \\ u_M(x,t) & , s\,t < x - x_L \text{ und } x - x_R < s_L\,t \\ (F')^{[-1]}(\frac{x-x_R}{t}) & , s_L\,t < x - x_R < s_R\,t \\ u_R(x,t) & , x - x_R > s_R\,t \end{array}\right\} \text{ mit}$$

$$s = \left.\frac{F(u_L) - F(u_M)}{u_L - u_M}\right|_{(\varphi(t),t)} = F'(\xi(t)),\ \ \xi(t) \in \Big(u_M(\varphi(t),t), u_L(\varphi(t),t)\Big),$$

$$s_L = F'(u_M(x_L - 0, 0)),\ s_R = F'(u_R(x_R + 0, 0)).$$

Damit läßt sich nun die Behauptung überprüfen. Dazu betrachtet man die Zustände links und rechts des Kollisionspunktes (x_0, t_0) für die Fälle a),. . .,d). Hier geht wesentlich ein, daß F konvex ist und somit gilt:

$$(1) \qquad x_1 \ < \ x_2 \quad \Longrightarrow \quad F'(x_1) \ < \ F'(x_2).$$

Beachte ferner, daß die Lösung entlang den Rändern φ_L, φ_R einer Verdünnungswelle stetig ist, d.h:

$$(2) \qquad u(\varphi_{L,R}(t) - 0, t) \ = \ u(\varphi_{L,R}(t) + 0, t).$$

Aus der Entropiebedingung (s. Def. 2.4) folgt ferner, daß in (x_0, t_0) genau eine neue Schockkurve entspringt, wenn gilt:

$$(3) \qquad u(x_0 - 0, t_0) \ > \ u(x_0 + 0, t).$$

Diese Bedingung ist also für die obigen Fälle zu überprüfen. Damit die Kurven zusammenlaufen muß nun gelten:

(a) $s_L(t_0) > s_R(t_0) \Longleftrightarrow F'(\xi_L(t_0)) > F'(\xi_R(t_0)) \overset{(1)}{\Longleftrightarrow} \xi_L(t_0) > \xi_R(t_0) \overset{\text{zu a)}}{\Longleftrightarrow} (3)$

(b) $s_{M,L}(t_0) > s_{M,R}(t_0) \overset{(2)}{\Longleftrightarrow} F'(u(x_0 - 0, t_0)) > F'(u(x_0 + 0, 0)) \overset{(1)}{\Longleftrightarrow}$
$u(x_0 - 0, t_0) > u(x_0 + 0, 0) \overset{\text{zu b)}}{\Longleftrightarrow} (3)$

(c) $s_R(t_0) > s(t_0) \overset{(2)}{\Longleftrightarrow} F'(u(x_0 - 0, t_0)) > F'(\xi(t_0)) \overset{(1)}{\Longleftrightarrow} u(x_0 - 0, t_0) > \xi(t_0) \overset{\text{zu c)}}{\Longleftrightarrow} (3)$

(d) $s(t_0) > s_L(t_0) \overset{(2)}{\Longleftrightarrow} F'(\xi(t_0)) > F'(u(x_0 + 0, t_0)) \overset{(1)}{\Longleftrightarrow} \xi(t_0) > u(x_0 + 0, t_0) \overset{\text{zu d)}}{\Longleftrightarrow} (3)$

Damit ist die Behauptung gezeigt.

∎

Um die eigentliche Aussage beweisen zu können, soll vorweg noch ein sehr wichtiger Satz, der die Struktur der Entropielösung charakterisiert, angeführt werden:

Satz 2.4: (Schaeffer, D.G. , 1973)

Vor.: $F \in C^{\infty}(\mathbb{R})$, F konvex, $H(\mathbb{R})$ Hilbertraum;

Beh.: Es gibt $\Omega \subset H(\mathbb{R})$ Menge 1. Kategorie, so daß für $u_0 \in H(\mathbb{R}) \backslash \Omega$ gilt

(1) Die Lösung des Erhaltungssatzes ist stückweise glatt;

(2) Die Anzahl der möglichen Schockkurven ist endlich, d.h.: $\mid \Gamma(u_0) \mid \; < \; \infty$.

Der Beweis findet sich in [Sch].

Die nicht näher beschriebene Menge Ω aus obigem Satz ist sehr klein und von technischer Natur. Sie stellt daher keine wesentliche Einschränkung dar. Wähle im folgenden $H(\mathbb{R}) := L^2(\ [a,b]\ )$, $a, b \in \mathbb{R}$, $a < b$. Dann gilt insbesondere:

$$L^{\infty}(\ [a,b]\ ) \backslash \Omega \subset L^2(\ [a,b]\ ) \backslash \Omega.$$

Dann läßt sich der folgende Satz zeigen, der besagt, daß die totale Variation der Entropielösung mit fortschreitender Zeit *nicht* zunimmt, d.h. sie ist TVD.

Satz 2.5:

Vor.: $a, b \in \mathbb{R}$, $a < b$, $u_0 : \mathbb{R} \to \mathbb{R}$, $u_0\big|_{[a,b]} \in L^{\infty}(\ [a,b]\ ) \backslash \Omega$ und

$u_0\big|_{\mathbb{R} \backslash [a,b]} \equiv 0$ oder periodische Fortsetzung von $u_0\big|_{[a,b]}$, $F \in C^{\infty}(\mathbb{R})$, F konvex,

u Entropielösung des zugehörigen Erhaltungssatzes;

Beh.: Für alle $t > 0$ gilt: $\mathrm{TV}\,(u(\circ, t)) \;\leq\; \mathrm{TV}(u_0)$.

Beweis:

Aufgrund der Voraussetzungen läßt sich über die Lösung u folgendes sagen:

(1) Wegen dem Satz von Oleinik (s. [Smo]) existiert die Entropielösung;

(2) u ist stückweise glatt im Sinne von Def. 2.12 und die Anzahl der auftretenden Diskontinuitätskurven ist endlich (o.B.d.A. $\big|\Gamma(u_0)\big| = N$) (vgl. Satz von Schaeffer).

Sei nun $T > 0$ beliebig, aber fest gewählt. Wegen (2) gibt es eine Zerlegung

$$t_0 := 0 < t_1 < \cdots < t_n < T =: t_{n+1} \quad \text{mit} \quad n \leq N$$

derart, daß zu diesen Zeitpunkten eine neue Schockkurve aus $\Gamma(u_0)$ entsteht, d.h.:
zwei Kurven laufen ineinander bzw. entspringen zum Zeitpunkt $t = 0$.
Die Lösung u läßt sich dann aus n Teillösungen u_i zusammensetzen, d.h.:

$$u(x,t) \;=\; u_i(x,t) \text{ für } t \in [t_i, t_{i+1}), x \in \mathbb{R},\ i = 0, \ldots, n.$$

Es soll nun folgende Ungleichung gezeigt werden:

$$\mathrm{TV}\,(u_{i+1}(\circ, t_{i+1})) \;\leq\; \mathrm{TV}(u_i(\circ, t_i)) \qquad i = 0, \ldots, n.$$

Diese soll hier nur für $i = 0$ nachgewiesen werden, denn die Induktion von i auf $i+1$ verläuft völlig analog.
Ausgangspunkt des Beweises ist die Zerlegung von $\mathbb{R} \times [t_0, t_1]$ in höchstens abzählbar viele Gebiete $G_j \subset \mathbb{R} \times [t_0, t_1]$, in denen u glatt ist, d.h.:

$$\exists \text{ Kurven } \underline{\varphi}_j, \overline{\varphi}_j : G_j = \Big\{ (x,t) \;\Big|\; x \in \mathbb{R},\ t \in [t_0, t_1],\ \underline{\varphi}_j(t) < x < \overline{\varphi}_j(t) \Big\}.$$

Hierbei ist $j \in J_0$ und $J_0 \subset \mathbb{N}$ die durch die obige Zerlegung gegebene Indexmenge, derart, daß die einzelnen Gebiete von links nach rechts durchnumeriert werden.
Man kann nun drei verschiedene "Glattheitsbereiche" unterscheiden:

(a) $D_j := \Big\{ (x,t) \;\Big|\; \underline{\varphi}_j(t) < x < \overline{\varphi}_j(t),\ \underline{\varphi}_j(t_1) < \overline{\varphi}_j(t_1),\ \underline{\varphi}_j(t_0) < \overline{\varphi}_j(t_0), \Big\}$;

(b) $R_j := \Big\{ (x,t) \;\Big|\; \underline{\varphi}_j(t) < x < \overline{\varphi}_j(t),\ \underline{\varphi}_j(t_1) < \overline{\varphi}_j(t_1),\ \underline{\varphi}_j(t_0) = \overline{\varphi}_j(t_0), \Big\}$;

(c) $S_j := \Big\{ (x,t) \;\Big|\; \underline{\varphi}_j(t) < x < \overline{\varphi}_j(t),\ \underline{\varphi}_j(t_1) = \overline{\varphi}_j(t_1),\ \underline{\varphi}_j(t_0) < \overline{\varphi}_j(t_0), \Big\}$.

In diesen Gebieten ist die Lösung wie folgt charakterisiert:

(a) $D_j \,\hat{=}\,$ Glattheitsbereich von u, durch die keine Kurven aus $\Gamma(u_0)$ laufen;
(b) $R_j \,\hat{=}\,$ Verdünnungswelle von u;
(c) $S_j \,\hat{=}\,$ Gebiet, das durch zwei zum Zeitpunkt t_1 zusammenlaufende Kurven aus $\Gamma(u_0)$ eingegrenzt wird.

Zur Veranschaulichung betrachte folgende Skizzen:

Fall a)

Fall b)

Fall c)

In den Gebieten vom Typ (a) sieht die Lösung u wegen S. 2.1 wie folgt aus:

$$\begin{aligned} & u(x,t) = u_0(x-(t-t_0)\,F'(u(x,t))) \text{ bzw.} \\ (3) \qquad & u(x,t) = u_0\left(\Psi^{[-1]}(x)\right) \text{ mit } x=\Psi(y) := y+(t-t_0)\,F'(u(y,t_0)) \\ & \qquad \text{für } (x,t)\in \mathbb{R}\times[t_0,t_1]. \end{aligned}$$

Im Verdünnungsbereich R_j läßt sich u bekanntlich wie folgt darstellen (s. z.B. [Smo]):

$$(4.a) \qquad u(x,t) \;=\; (F')^{[-1]}\left(\tfrac{x}{t}\right) \quad \text{für} \quad (x,t)\in R_j.$$

Insbesondere gilt:

$$(4.b) \qquad \forall\,(x,t)\in\mathbb{R},\ \frac{x}{t} = const \quad \Longrightarrow \quad u(x,t) \;=\; const.$$

Entlang den Kurven $\varphi(t)$ gelten folgende Beziehungen zwischen den Zuständen links und rechts dieser Kurven:
(5.a) ist $\varphi(t)$ eine Diskontinuitätskurve, so gilt:

$$\tfrac{d}{dt}\,\varphi(t)\ \left[u(\varphi(t)-0,t)-u(\varphi(t)+0,t)\right] = \left[F(u(\varphi(t)-0,t))-F(u(\varphi(t)+0,t))\right];$$

(5.b) ist $\varphi(t)$ Rand einer Verdünnungswelle, so gilt:

$$u(\varphi(t)-0,t) \;=\; u(\varphi(t)+0,t).$$

Durch obige Einteilung zerfällt J_0 nun in drei paarweise disjunkte Teilmengen J_0^1, J_0^2, J_0^3, wobei J_0^1 die Gebiete vom Typ (a), J_0^2 die Gebiete vom Typ (b) und J_0^3 die Gebiete vom Typ (c) indiziert. Dann gilt insbesondere:

$$\mathbb{R} \;=\; \bigcup_{j\in J_0^1\cup J_0^2} \left[\,\underline{\varphi}_j(t_1), \overline{\varphi}_j(t_1)\,\right].$$

Somit gilt für die Oszillation von $u(\circ,t_1)$:

$$\begin{aligned} \mathrm{TV}\,(u(\circ,t_1)) \;=\; & \sum_{j\in J_0^1} \mathrm{TV}\left(u(\circ,t_1)|_{[\,\underline{\varphi}_j(t_1),\overline{\varphi}_j(t_1)]}\right) + \\ & \sum_{j\in K_a} |\,u(\overline{\varphi}_j(t_1)-0,t_1) \;-\; u(\overline{\varphi}_j(t_1)+0,t_1)\,| + \\ (6) \qquad & \sum_{j\in K_b} |\,u(\overline{\varphi}_j(t_1)-0,t_1) \;-\; u(\overline{\varphi}_j(t_1)+0,t_1)\,| + \\ & \sum_{j\in K_b} |\,u(\underline{\varphi}_j(t_1)-0,t_1) \;-\; u(\underline{\varphi}_j(t_1)+0,t_1)\,| + \\ & \sum_{j\in K_{c,d,e}} |\,u(\underline{\varphi}_j(t_1)-0,t_1) \;-\; u(\underline{\varphi}_j(t_1)+0,t_1)\,|. \end{aligned}$$

Hierbei sind die Indexmengen $K_a, \dots, K_e$ wie folgt definiert:

$K_a := \{\ j \in J_0^1 \ \mid\ j+1 \in J_0^1 \}$

$\hat{=}$ Menge der Indizes, für die zwei benachbarte Gebiete D_j und D_{j+1} durch eine Diskontinuitätskurve getrennt werden;

$K_b := \{\ j \in J_0^2 \ \mid\ j-1 \in J_0^1, j+1 \in J_0^1 \}$

$\hat{=}$ Menge der Indizes, bei denen eine Verdünnungswelle zwei Bereiche vom Typ a) trennt;

$K_c := \{\ j \in J_0^3 \ \mid (\ j-1 \in J_0^1, j+1 \in J_0^2) \text{ oder } (\ j-1 \in J_0^2, j+1 \in J_0^1) \}$

$\hat{=}$ Menge der Indizes, bei denen eine Diskontinuitätskurve in eine Verdünnungswelle läuft, zum einen von links zum anderen von rechts;

$K_d := \{\ j \in J_0^3 \ \mid j-1 \in J_0^1, j+1 \in J_0^1 \}$

$\hat{=}$ Menge der Indizes, bei denen zwei Diskontinuitätskurven kollidieren;

$K_e := \{\ j \in J_0^3 \ \mid j-1 \in J_0^2, j+1 \in J_0^2 \}$

$\hat{=}$ Menge der Indizes, bei denen zwei Verdünnungswellen ineinander laufen.

Hierbei werden durch die Mengen K_c, K_d, K_e die Interaktionsprobleme analog zum Beweis zu La. 2.16 charakterisiert. Zum besseren Verständnis betrachte folgende Skizzen:

Fall a)

Fall b)

Fall c2)

Fall c1)

Fall d)

Fall e)

Um (6) weiter abschätzen zu können, muß man einen Zusammenhang der Variation von $u(\circ, t_1)$ und $u(\circ, t_0)$ herstellen. Dies geschieht unter Ausnutzung von (3), (4) und (5) sowie der Tatsache, daß F' streng monoton steigend ist. Dazu untersucht man zunächst die Variation in den Bereichen, in denen u glatt ist und anschließend die Sprunghöhe in den Grenzpunkten dieser Gebiete. Sei also

1) $\underline{j \in J_0^1 :}$

Sei $\{\xi_i\}_{i=1}^m$ eine Zerlegung von $[\,\underline{\varphi}_j(t_1), \overline{\varphi}_j(t_1)\,]$. Da u in D_j starke Lösung ist, gibt es wegen (3) $\zeta_1 < \cdots < \zeta_m$, so daß für $i = 1, \ldots, m$ gilt:

(α) $$u_1(\xi_i, t_1) \;=\; u(\zeta_i, t_0) \;=\; u(\Psi^{[-1]}(\xi_i), t_0) \quad \text{mit}$$

$$\xi_i = \Psi(\zeta_i) := \zeta_i + (t - t_0)\, F'(u(\zeta_i, t_0));$$

(β) $$\underline{\varphi}_j(t_0) \;\le\; \zeta_1 \;\le \cdots \le\; \zeta_m \;\le\; \overline{\varphi}_j(t_0).$$

Beachte ferner, daß sich die Charakteristiken in D_j nicht schneiden, da die starke Lösung sonst nicht mehr existieren würde. Somit bildet $\{\zeta_i\}_{i=1}^m$ eine Zerlegung von $X_j^1 := \left[\; \Psi^{[-1]}\,(\underline{\varphi}_j(t_1) + 0), \Psi^{[-1]}\,(\overline{\varphi}_j(t_1) - 0) \right]$ und es gilt folgende Abschätzung:

$$\sum_{i=1}^{m-1} |\, u(\xi_{i+1}, t_1) \;-\; u(\xi_i, t_1)\,| \;=\; \sum_{i=1}^{m-1} |\, u(\zeta_{i+1}, t_0) \;-\; u(\zeta_i, t_0)\,|$$

$$\le\; \mathrm{TV}(u(\circ, t_0)|_{X_j^1}).$$

Da die Zerlegung beliebig gewählt war und die rechte Seite in obiger Ungleichung von ihr unabhängig ist, kann man links zum Supremum übergehen und es gilt somit:

$$\mathrm{TV}(u(\circ, t_1)|_{[\,\underline{\varphi}_j(t_1), \overline{\varphi}_j(t_1)]}) \;\le\; \mathrm{TV}(u(\circ, t_0))|_{X_j^1}).$$

2) $\underline{j \in J_0^2 :}$

Mit F' ist auch $(F')^{[-1]}$ streng monoton wachsend, und somit gilt:

$$\begin{aligned}
\mathrm{TV}(u(\circ,t_1))|_{[\,\underline{\varphi}_j(t_1),\overline{\varphi}_j(t_1)]}) &= |\, u(\,\overline{\varphi}_j(t_1)+0,t_1) \;-\; u(\,\underline{\varphi}_{j+1}(t_1)-0,t_1)\,| \\
&\overset{(5.b)}{=} |\, u(\,\overline{\varphi}_j(t_1)-0,t_1) \;-\; u(\,\underline{\varphi}_{j+1}(t_1)+0,t_1)\,| \\
&\overset{(4.b)}{=} |\, u(\,\overline{\varphi}_j(t_0)-0,t_0) \;-\; u(\,\underline{\varphi}_{j+1}(t_0)+0,t_0)\,|
\end{aligned}$$

Für die Sprungstellen läßt sich nun folgende Fallunterscheidung durchführen:

3) $\underline{j \in K_a :}$

Da u links und rechts von $\overline{\varphi}_j(t) \equiv \underline{\varphi}_{j+1}(t)$ glatt ist, existieren in den Gebieten beiderseits dieser Kurven Charakteristiken gemäß S. 2.1, die in $\overline{\varphi}_j(t_1) = \underline{\varphi}_{j+1}(t_1)$ zusammenlaufen. Damit gilt für $X_j^2 := [\ \Psi^{[-1]}(\overline{\varphi}_j(t_1)-0), \Psi^{[-1]}(\overline{\varphi}_j(t_1)+0)\ ]$ mit $\overline{\varphi}_j(t_1) \pm 0 = \Psi(y \pm 0) := y + (t-t_0)\ F'(u(y \pm 0, t_0))$:

$$|\, u(\overline{\varphi}_j(t_1)-0,t_1) \;-\; u(\underline{\varphi}_{j+1}(t_1)+0,t_1)\,| \overset{(3)}{=}$$

$$|\, u(\Psi^{[-1]}(\overline{\varphi}_j(t_1)-0),t_0) \;-\; u(\Psi^{[-1]}(\underline{\varphi}_{j+1}(t_1)+0),t_0)\,| \;\le \mathrm{TV}(u(\circ,t_0)|_{X_j^2}).$$

4) $\underline{j \in K_b :}$

Da keine Diskontinuitätskurve in die Verdünnungswelle läuft, gilt mit (5.b):

$$u(\underline{\varphi}_j(t_1)-0,t_1) \;-\; u(\underline{\varphi}_j(t_1)+0,t_1) \;=\; 0;$$

$$u(\overline{\varphi}_j(t_1)-0,t_1) \;-\; u(\overline{\varphi}_j(t_1)+0,t_1) \;=\; 0.$$

5) $\underline{j \in K_c :}$

Zeige hier nur den Fall, daß ein Schock von links in die Verdünnungswelle läuft:

Da u in D_{j-1} glatt ist und entlang den Rändern einer Verdünnungswelle stetig, gilt mit $\overline{\varphi}_j(t_1) - 0 = \Psi(y+0) := y + (t-t_0)\ F'(u(y-0,t_0))$ für $X_j^3 := [\ \Psi^{[-1]}(\overline{\varphi}_j(t_1)-0), \overline{\varphi}_j(t_0)]$:

$$|\, u(\overline{\varphi}_j(t_1)-0,t_1) \;-\; u(\overline{\varphi}_j(t_1)+0,t_1)\,| \overset{(3),(4)}{=}$$

$$|\, u(\Psi^{[-1]}(\overline{\varphi}_j(t_1)-0),t_0) \;-\; u(\overline{\varphi}_j(t_1)-0),t_1)\,| \overset{(5.b)}{=}$$

$$|\, u(\Psi^{[-1]}(\overline{\varphi}_j(t_1)-0),t_0) \;-\; u(\overline{\varphi}_j(t_0)-0),t_0)\,| \le \mathrm{TV}(u(\circ,t_0)|_{X_j^3}).$$

Im anderen Fall erhält man für $X_j^3 := [\ \overline{\varphi}_j(t_0), \Psi^{[-1]}(\overline{\varphi}_j(t_1)+0)]$ mit $\overline{\varphi}_j(t_1) + 0 = \Psi(y+0) := y + (t-t_0)\ F'(u(y+0,t_0))$ eine analoge Aussage.

6) $\underline{j \in K_d :}$

Analog zu Fall 5) läßt sich für $X_j^4 := [\,\Psi^{[-1]}(\overline{\varphi}_j(t_1)-0), \Psi^{[-1]}(\overline{\varphi}_j(t_1)+0)]$ mit $\overline{\varphi}_j(t_1) \pm 0 = y + (t-t_0)\, F'(u(y \pm 0, t_0))$ zeigen:

$$| \, u(\overline{\varphi}_j(t_1)-0, t_1) \; - \; u(\overline{\varphi}_j(t_1)+0, t_1) \, | \stackrel{(3)}{=}$$
$$| \, u(\Psi^{[-1]}(\overline{\varphi}_j(t_1)-0), t_0) \; - \; u(\Psi^{[-1]}(\overline{\varphi}_j(t_1)+0), t_0) \, | \leq \mathrm{TV}(u(\circ, t_0)|_{X_j^4}).$$

7) $\underline{j \in K_e :}$

Für $X_j^5 := [\, \underline{\varphi}_j(t_0), \overline{\varphi}_j(t_0)]$ gilt, da u zwischen den beiden Verdünnungswellen glatt ist:

$$| \, u(\underline{\varphi}_j(t_1)-0, t_1) \; - \; u(\underline{\varphi}_j(t_1)+0, t_1) \, | \stackrel{(5.b)}{=}$$
$$| \, u(\underline{\varphi}_j(t_1)+0, t_1-0) \; - \; u(\overline{\varphi}_j(t_1)-0, t_1-0) \, | \stackrel{(4)}{=}$$
$$| \, u(\underline{\varphi}_j(t_1)+0, t_0) \; - \; u(\overline{\varphi}_j(t_1)-0, t_0) \, | \; \leq \mathrm{TV}(u(\circ, t_0)|_{X_j^5}).$$

Aus 1),...,7) folgt nun mit (6):

$$\begin{aligned} \mathrm{TV}\,(u(\circ, t_1)) \;\leq\; & \sum_{j \in J_0^1} \mathrm{TV}\left(u(\circ, t_0)|_{X_j^1}\right) + \\ & \sum_{j \in J_0^2} | \, u(\overline{\varphi}_j(t_0)-0, t_0) \; - \; u(\overline{\varphi}_j(t_0)+0, t_0) \, | \; + \\ & \sum_{j \in K_a} \mathrm{TV}\left(u(\circ, t_0)|_{X_j^2}\right) + \sum_{j \in K_c} \mathrm{TV}\left(u(\circ, t_0)|_{X_j^3}\right) + \\ & \sum_{j \in K_d} \mathrm{TV}\left(u(\circ, t_0)|_{X_j^4}\right) + \sum_{j \in K_e} \mathrm{TV}\left(u(\circ, t_0)|_{X_j^5}\right). \end{aligned} \tag{8}$$

Aufgrund der Festlegung von $\overline{\varphi}_j(t), \underline{\varphi}_j(t)$ sowie den Eigenschaften von Charakteristiken (s. S. 2.1) gilt:

$$\mathbb{R} \;=\; \bigcup_{i=1}^{5} \; \bigcup_{j \in J_0} X_j^i \,.$$

Dabei sind die X_j^i paarweise — bis auf gemeinsame Randpunkte — disjunkt.
Aus (8) folgt nun mit der Definition der totalen Variation:

$$\mathrm{TV}(u(\circ, t_1)) \;\leq\; \mathrm{TV}(u(\circ, t_0)).$$

Die Induktion von i auf $i+1$ läßt sich nun analog behandeln, in dem man alle obigen Indizierungen mit 0 durch i und alle mit 1 durch $i+1$ ersetzt. Dies ist möglich, da wegen La. 2.16 bei der Kollision von Diskontinuitäten genau eine neue entsteht, aber insbesondere keine Verdünnungswellen.
Beachte dabei, daß u wegen dem Satz von Schaeffer nur endlich viele Diskontinuitätskurven hat. Somit folgt die Behauptung mit Induktionsprinzip.

■

Die Aussage aus obigem Satz läßt sich sehr schön aus Abb.15 ablesen. Man sieht, daß die Welle mit fortschreitender Zeit mehr und mehr abflacht.
Mittels dem gerade bewiesenen Satz kann man nun eine wichtige Eigenschaft für die ENO–Verfahren zeigen:

Satz 2.6:

Vor.: $a, b \in \mathbb{R},\ a < b,\ u_0 : \mathbb{R} \to \mathbb{R},\ u_0\big|_{[a,b]} \in L^\infty(\ [a,b]\ )\backslash\Omega$ und
$u_0\big|_{\mathbb{R}\backslash[a,b]} \equiv 0$ oder periodische Fortsetzung von $u_0\big|_{[a,b]}$, $F \in C^\infty(\mathbb{R}),\ F$ konvex,
$L^N(\circ, v_j^n)$ erfülle (2.2.b), $N \in \mathbb{N}$,
$\{v_j^n\}_{j\in\mathbb{Z}, n\in\mathbb{N}_0}$ das durch (2.1) definierte ENO–Verfahren;

Beh.: $\mathrm{TV}\left(\{v_j^{n+1}\}\right) \ \le\ \mathrm{TV}\left(L^N(\circ, v_j^n)\right)$.

Beweis:

Nach Konstruktion der ENO–Verfahren gilt:

(1) $$v_j^{n+1} \ =\ v_j^n \ -\ \lambda\ \left[\ \hat{F}_{j+\frac{1}{2}} \ -\ \hat{F}_{j-\frac{1}{2}}\ \right] \quad \text{mit}$$

(2) $$\hat{F}_{j+\frac{1}{2}} \ :=\ \frac{1}{\tau}\int_{t_n}^{t_{n+1}} F(v(x_{j+\frac{1}{2}}, t))\, dt \quad \text{mit } v \text{ schwache Lösung zu}$$

(3) $$\begin{aligned} v_t(x,t) \ +\ (F(v(x,t)))_x \ &=\ 0, \quad (x,t) \in \mathbb{R}\times(t_n, +\infty);\\ v(x, t_n) \ &=\ L^N(x, v_j^n), \quad x \in \mathbb{R}. \end{aligned}$$

Da u_0 nach Voraussetzung einen kompakten Träger hat oder periodisch ist, so gilt dies auch für $\{v_j^n\}_{j\in\mathbb{Z}}$, $n \in \mathbb{N}_0$ und somit aufgrund der Definition der ENO–Polynome auch für $L^N(\circ, v_j^n)$. Damit sind La. 2.4 und S. 2.5 auch auf v anwendbar. Sei nun $n \in \mathbb{N}$ beliebig aber fest. Dann gilt für $j \in \mathbb{Z}$:

$$\begin{aligned}
(4)\ v_j^{n+1} \ &=\ v_j^n \ -\ \lambda\ \left[\ \hat{F}_{j+\frac{1}{2}} \ -\ \hat{F}_{j-\frac{1}{2}}\ \right]\\
&\overset{\text{K.2.1}}{=} v_j^n \ +\ \frac{1}{h}\left[\ \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x, t_{n+1})\, dx \ -\ \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x, t_n)\, dx\ \right]\\
&\overset{\text{def.}}{=} v_j^n \ +\ \overline{v}(x_j, t_{n+1}) \ -\ \overline{v}(x_j, t_n)\\
&\overset{(3)}{=} v_j^n \ +\ \overline{v}(x_j, t_{n+1}) \ -\ \overline{L}^N(x_j, v_j^n).
\end{aligned}$$

Wegen (2.2.b) gilt: $\overline{L}^N(x_j, v_j^n) \ =\ v_j^n, \quad j \in \mathbb{Z}.$

Daraus folgt:

$$v_{j+1}^{n+1} - v_j^{n+1} = \overline{v}(x_{j+1}, t_{n+1}) - \overline{v}(x_j, t_{n+1}).$$

Da die Mittelwertbildung und die Entropielösung eines skalaren Erhaltungssatzes TVD sind (s. La. 2.4, S. 2.5), gilt:

$$\begin{aligned}
\mathrm{TV}\left(\{v_j^{n+1}\} \right) &:= \sum_{j\in\mathbb{Z}} | v_{j+1}^{n+1} - v_j^{n} | \overset{(4)}{=} \sum_{j\in\mathbb{Z}} | \overline{v}(x_{j+1}, t_{n+1}) - \overline{v}(x_j, t_{n+1}) | \\
&\leq \mathrm{TV}(\overline{v}(\circ, t_{n+1})) \overset{\text{L.2.4.}}{\leq} \mathrm{TV}(v(\circ, t_{n+1})) \overset{\text{S. 2.5.}}{\leq} \mathrm{TV}(v(\circ, t_n)) \\
&\overset{(3)}{=} \mathrm{TV}(L^N(\circ, v_j^n)).
\end{aligned}$$

■

Obige Aussage gilt auch , falls man $\hat{F}_{j+\frac{1}{2}}$ durch eine Approximation ersetzt, die folgende Form (vgl. La. 2.15) hat:

$$\tilde{F}_{j+\frac{1}{2}} = \hat{F}_{j+\frac{1}{2}} + O\left(h^{N+1}\right).$$

Man macht zwar in $x_{j+\frac{1}{2}}$, $j \in \mathbb{Z}$, einen Fehler $O\left(h^{N+1}\right)$, da aber die Anzahl der Intervalle zur Zerlegung von $\mathbb{R}$ von der Schrittweite h abhängt, beträgt diese gerade $O\left(h^{-1}\right)$. Somit ist der globale Fehler $O\left(h^{N}\right)$, d.h.:

$$\mathrm{TV}\left(\{v_j^{n+1}\}\right) \leq \mathrm{TV}\left(\mathrm{Ł}^N(\circ, v_j^n)\right) + O\left(h^N\right).$$

Für gewisse N läßt sich die Ungleichung aus S. 2.6 noch weiter abschätzen:

Folgerung 2.4:

Vor.: $a, b \in \mathbb{R},\ a < b,\ u_0 : \mathbb{R} \to \mathbb{R},\ u_0\big|_{[a,b]} \in L^\infty(\ [a,b]\ )\backslash\Omega$ und

$u_0\big|_{\mathbb{R}\backslash[a,b]} \equiv 0$ oder periodische Fortsetzung von $u_0\big|_{[a,b]}$, $F \in C^\infty(\mathbb{R})$, F konvex,

$L^N(\circ, v_j^n) = L^{1,RP}(\circ, v_j^n)$ oder $L^N(\circ, v_j^n) = L^{0,RD}(\circ, v_j^n)$,

$\left\{v_j^n\right\}_{j\in\mathbb{Z}, n\in\mathbb{N}_0}$ das durch (2.1) definierte ENO–Verfahren;

Beh.: $\mathrm{TV}\left(\{v_j^{n+1}\}\right) \leq \mathrm{TV}\left(\{v_j^n\}\right) \leq \mathrm{TV}\left(u_0\right)$.

Beweis:

Folgt unmittelbar aus S. 2.6, La. 2.11 und La. 2.10 bzw. La. 2.3.

■

Dies besagt, daß die ENO–Verfahren für gewisse $L^N(\circ, v_j^n)$ TV–stabil sind. Das ist nicht weiter erstaunlich, denn für diese sind die ENO–Verfahren identisch mit dem Godunov–Verfahren (s. La. 2.17).

Zuvor soll aber noch das Oszillationsverhalten der Verfahren in solchen Bereichen untersucht werden, in denen die zu approximierende Lösung eines skalaren Erhaltungssatzes hinreichend glatt ist:

Folgerung 2.5:

Vor.: $a, b \in \mathbb{R},\ a < b,\ u_0 : \mathbb{R} \to \mathbb{R},\ u_0\big|_{[a,b]} \in L^\infty(\ [a,b]\ )\backslash\Omega$ und

$u_0\big|_{\mathbb{R}\backslash[a,b]} \equiv 0$ oder periodische Fortsetzung von $u_0\big|_{[a,b]}$, $F \in C^\infty(\mathbb{R})$, F konvex,

$L^N(\circ, v_j^n) = L^{N+1,RP}(\circ, v_j^n)$ oder $L^N(\circ, v_j^n) = L^{N,RD}(\circ, v_j^n),\ N \geq 2,$

$\{v_j^n\}_{j\in\mathbb{Z},n\in\mathbb{N}_0}$ das durch (2.1) definierte ENO–Verfahren,

u Entropielösung des Erhaltungssatzes mit $u \in C^{N+1}(\mathbb{R})$;

Beh.: Für h hinreichend klein gilt:

$$\mathrm{TV}\left(\{v_j^{n+1}\}\right) \ \leq\ \mathrm{TV}\left(\{v_j^n\}\right) \ +\ O\left(h^N\right) \ \leq\ 2\,\mathrm{TV}\left(u_0\right).$$

Beweis:

Mit La. 2.9 bzw. La. 2.4 gilt:

$$\mathrm{TV}\left(L^N(\circ, v_j^n)\right) \ \leq\ \mathrm{TV}\left(\{v_j^n\}\right) \ +\ O\left(h^N\right).$$

Aus S. 2.6 folgt somit:

$$\begin{aligned}\mathrm{TV}\left(\{v_j^{n+1}\}\right) \ &\leq\ \mathrm{TV}\left(\{v_j^n\}\right) \ +\ O\left(h^N\right) \ \leq\ \mathrm{TV}\left(\{v_j^0\}\right) \ +\ n\,O\left(h^N\right) \\ &\leq\ \mathrm{TV}\left(u_0\right) \ +\ n\,O\left(h^N\right).\end{aligned}$$

Da $n = O\left(h^{-1}\right)$ gilt also für h hinreichend klein:

$$\mathrm{TV}\left(\{v_j^{n+1}\}\right) \ \leq\ \mathrm{TV}\left(u_0\right) \ +\ O\left(h^{N-1}\right) \ \leq\ 2\,\mathrm{TV}\left(u_0\right).$$

■

Leider läßt sich die ENO–Eigenschaft gemäß Def. 2.10 für beliebige physikalische Anfangsdaten, d.h. u_0 hat kompakten Träger oder ist periodisch, und $N \geq 2$ so nicht allgemein nachweisen. Dazu wird man das Monotonieverhalten von $L^N(\circ, \overline{w})$ in Gegenwart von "versteckten" Sprüngen näher untersuchen müssen. In [HaOs] wurde gezeigt, daß die ENO–Verfahren mittels $L^1(\circ, \overline{w})$ nicht oszillierend sind in dem Sinne, daß die Anzahl der Extrema von einer Zeitstufe zur nächsten nicht zunimmt. Daraus folgt aber nicht notwendig, daß die totale Variation nicht wächst, da in den Extrema sehr wohl ein Zuwachs möglich ist. An dieser Stelle sei noch vermerkt, daß in [ENO3], S. 19 bzw S. 23 behauptet wird, daß die in Kapitel 2.1.3 vorgestellten Rekonstruktionsmethoden wirklich die ENO–Bedingung (2.2.c) erfüllen. Damit wäre wegen S. 2.6 die TV–Stabilität für die ENO–Verfahren im skalaren Fall gezeigt. Der Beweis soll sich angeblich in einem Artikel befinden, der seit mindestens 1986 "in preperation" ist (vgl. Literatur zu [ENO3]). Dieser ist aber bis heute nicht veröffentlicht und meines Wissens auch noch nicht geschrieben.

Im folgenden sollen noch einige weitere charakterisierende Eigenschaften für die ENO–Verfahren hergeleitet werden:

Lemma 2.17:

Vor .: $\left\{v_j^n\right\}_{j\in\mathbb{Z},n\in\mathbb{N}_0}$ ENO–Verfahren gemäß (2.1);

Beh.: $L^N(\circ, v_j^n) = L^{1,RP}(\circ, v_j^n)$ oder $L^N(\circ, v_j^n) = L^{0,RD}(\circ, v_j^n) \quad \Longrightarrow$

Das ENO–Verfahren ist identisch mit dem Godunov–Verfahren.

Beweis:

Mit der Definition von $L^{1,RP}(\circ, v_j^n)$ bzw. $L^{0,RD}(\circ, v_j^n)$ gilt:

$$L^{1,RP}(x, v_j^n) \;=\; L^{0,RD}(x, v_j^n) \;=\; v_j^n \quad \text{für} \quad x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}).$$

Aufgrund der Konstruktion der ENO–Verfahren gemäß (2.1) folgt die Behauptung.

■

Obige Aussage besagt also, daß die ENO–Verfahren insbesondere eine Verallgemeinerung des Godunov–Verfahren darstellen.

Zur Approximation eines linearen Erhaltungssatzes läßt sich nun noch zeigen:

Lemma 2.18:

Vor.: $u_0 \in L^\infty(\mathbb{R}),\; b, c \in \mathbb{R},\; N \in \mathbb{N},\; F(u) := c\,u \;+\; b,$

$\left\{v_j^n\right\}_{j\in\mathbb{Z},n\in\mathbb{N}_0}$ das gemäß (2.1) definierte ENO–Verfahren, $\lambda := \frac{\tau}{h},\; \tau, h > 0,$

u Lösung des linearen Erhaltungssatzes;

Beh.: a) $L^N(\circ, v_j^n)$ erfülle (2.2.b) und $\lambda = \frac{1}{|c|} \quad \Longrightarrow \quad v_j^n \;=\; \overline{u}(x_j, t_n)\;,\; j \in \mathbb{Z}, n \in \mathbb{N}_0;$

b) $L^N(\circ, v_j^n) = L^{N,RP}(\circ, v_j^n) \quad \Longrightarrow$

$$v_j^{n+1} \;=\; \frac{1}{h}\, \Big[Q^N(x_{j+\frac{1}{2}} - c\,\tau, W) \;-\; Q^N(x_{j-\frac{1}{2}} - c\,\tau, W) \Big]\,.$$

Beweis:

Zu a): Für die exakte Lösung des linearen Erhaltungssatzes gilt (s. S. 2.1):

(1) $$u(x,t) \;=\; u_0(x - c\,t) \quad \text{für} \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+.$$

O.B.d.A. seien $c > 0$ und τ, h so gewählt, daß gilt:

$$\lambda\, c \;=\; 1.$$

Dann ist

(2) $$v(x,t) \;=\; L^N(x - c\,(t - t_n), v_j^n),\; t \geq t_n$$

die Lösung von

$$v_t(x,t) \;+\; (F(v(x,t)))_x \;=\; 0, \quad (x,t) \in \mathbb{R} \times (t_n, +\infty);$$

$$v(x,t_n) \;=\; L^N(x, v_j^n), \quad x \in \mathbb{R}.$$

Somit gilt für den numerischen Fluß gemäß (2.1.b):

$$\hat{F}_{j+\frac{1}{2}} \;=\; \tfrac{1}{\tau} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} L^N(t, v_j^n)\, dt \;+\; b.$$

Daraus folgt mit (2.1.b):

$$v_j^{n+1} \;=\; v_j^n \;-\; \overline{L}^N(x_j, v_j^n) \;+\; \overline{L}^N(x_{j-1}, v_j^n).$$

Wegen (2.2.b) gilt: $\overline{L}^N(x_j, v_j^n) \;=\; v_j^n$ und somit

(3) $$v_j^{n+1} \;=\; v_{j-1}^n, \quad j \in \mathbb{Z}, n \in \mathbb{N}_0.$$

Der eigentliche Beweis wird nun mittels vollständiger Induktion über n geführt:

$n = 0$: $v_j^1 \overset{(3)}{=} v_{j-1}^0 \overset{\text{def.}}{=} \overline{u}(x_{j-1}, t_0) \overset{(1)}{=} \overline{u}(x_j, t_1).$

$n \to n+1$: $v_j^{n+1} \overset{(3)}{=} v_{j-1}^n \overset{\text{I.V.}}{=} \overline{u}(x_{j-1}, t_n) \overset{(1)}{=} \overline{u}(x_j, t_{n+1}).$

Für $c < 0$ läßt sich der Beweis analog durchführen.

Zu b): Mit (2.3.c) gilt:

$$L^N(\circ, v_j^n) \;=\; \tfrac{d}{dx}\, Q^N(\circ, W).$$

Wegen (2) gilt daher für den numerischen Fluß gemäß (2.1.b):

$$\hat{F}_{j+\frac{1}{2}} \;=\; \tfrac{1}{\tau} \left[\, Q^N(x_{j+\frac{1}{2}}, W) \;-\; Q^N(x_{j+\frac{1}{2}} - c\,\tau, W) \,\right].$$

Durch Einsetzen in (2.1.b.γ) erhält man unter Berücksichtigung von (2.3.b) die Behauptung.

■

Des weiteren läßt sich noch zeigen:

Lemma 2.19:

Die ENO–Verfahren gemäß (2.1) sind unter Verwendung der RP– bzw. RD–Rekonstruktion in Erhaltungsform.

Beweis:

Sei $N \in \mathbb{N}$ beliebig, aber fest gewählt.

Wegen (2.1.b) hängt $\hat{F}_{j+\frac{1}{2}}$ von $L^N(\circ, v_j^n)\big|_{[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})}$ und $L^N(\circ, v_j^n)\big|_{[x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}})}$ ab. Es gilt nun aufgrund der Konstruktionsvorschrift (2.3) bzw. (2.4):

$$L_{j+\frac{1}{2}}^{N,RP} \;=\; L_{j+\frac{1}{2}}^{N,RP}(v_{j-N+2}^n, \ldots, v_{j+N}^n) \text{ bzw. } L_{j+\frac{1}{2}}^{N,RD} \;=\; L_{j+\frac{1}{2}}^{N,RD}(v_{j-N+1}^n, \ldots, v_{j+N+1}^n).$$

Daraus folgt unmittelbar:

$$\hat{F}_{j+\frac{1}{2}}^{N,RP} \;=\; \hat{F}_{j+\frac{1}{2}}(v_{j-N+1}^n, \ldots, v_{j+N}^n) \text{ bzw. } \hat{F}_{j+\frac{1}{2}}^{N,RD} \;=\; \hat{F}_{j+\frac{1}{2}}(v_{j-N}^n, \ldots, v_{j+N+1}^n).$$

■

Nachdem die ENO–Verfahren nun für den skalaren Fall ausführlich analysiert wurden, sollen sie jetzt auf Systeme übertragen werden.

2.2 Systeme

2.2.1 Kontinuierliche Theorie

Analog zum skalaren Fall läßt sich ein System von Erhaltungssätzen definieren:

Definition 2.13:

Sei $\underline{F} \in C^1(\mathbb{R}^m)$, $m \in \mathbb{N}$, und $\underline{u}_0 : \mathbb{R} \to \mathbb{R}$ beliebig. Unter einem *System von Erhaltungssätzen* versteht man:

$$\underline{u}_t(x,t) \ + \ (\underline{F}(\underline{u}(x,t)))_x \ = \ 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+;$$

(2.18.a)

$$\underline{u}(x,0) \ = \ \underline{u}_0(x), \quad x \in \mathbb{R},$$

mit $\underline{u} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}^m$.

Dieses heißt *strikt hyperbolisch*, falls die Funktionalmatrix $\underline{F}'$ m reelle Eigenwerte $\lambda_i(\underline{u})$ besitzt mit

(2.18.b) $$\lambda_1(\underline{u}) \ < \cdots < \ \lambda_m(\underline{u}),$$

wobei $\{ \ \underline{l}_k(\underline{u}) \ \}_{k=1}^m$, $\{ \ \underline{r}_k(\underline{u}) \ \}_{k=1}^m$ die zugehörigen Links– bzw. Rechtseigenvektoren sind. $\big($o.B.d.A.: $\langle \ \underline{r}_k(\underline{u}), \underline{l}_j(\underline{u} \ )\rangle \ = \ \delta_{k,j}\big)$.

Analog zum skalaren Fall lassen sich nun starke und schwache Lösung definieren. Bei der Übertragung der Entropielösung gibt es jedoch Schwierigkeiten, denn der Satz von Oleinik (s. [Smo], [Str]), d.h. die eindeutige Existenz der schwachen Lösung, gilt für Systeme nicht mehr. Die Frage der Existenz ist bis heute noch nicht geklärt. Zur Struktur der schwachen Lösung siehe auch Kapitel 4.4.

Wichtig ist hier jedoch nur, daß Kor. 2.1 auch komponentenweise für Systeme gültig ist. Somit ist es sinnvoll Begriffe wie Erhaltungsform, Entropiefunktion, numerische Entropiebedingung usw. auf Systeme zu übertragen.

Im folgenden sollen nun die ENO–Verfahren analog zu (2.1) zur Approximation eines strikt hyperbolischen Systems von Erhaltungssätzen der Form (2.18) hergeleitet werden.

2.2.2 ENO–Verfahren

In Analogie zu Kap. 2.1.2 führe nun für die schwache Lösung von (2.18) folgende Bezeichnungen ein:

$$\underline{u}(x,t) \ := \ (\ u_1(x,t), \ldots, u_m(x,t) \ )^T, \quad u_i : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R};$$

$$\underline{F}(\underline{u}) \;:=\; (\; F_1(\underline{u}), \ldots, \underline{F}_m(\underline{u}) \;), \quad F_i : \mathbb{R}^m \to \mathbb{R};$$

$$\overline{u}^n(x) := \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \underline{u}(s, t_n)\, ds, \quad t_n := \tau\, n, \;\text{(komponentenweise)};$$

$$\overline{u}_j^n := \overline{u}^n(x_j), \quad x_j := j\, h,\; j \in \mathbb{Z}.$$

Damit lassen sich die ENO–Verfahren (vgl. (2.1)) komponentenweise auf Systeme übertragen :

(2.19.a) Setze $\underline{v}_j^0 := (\; v_{1,j}^0, \ldots, v_{m,j}^0 \;)^T$ mit $v_{i,j}^0 := \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_{0,i}(x)\, dx.$

(2.19.b) Sei $\underline{v}_j^n$ Approximation der Mittelwerte der exakten Lösung $\underline{u}(x_j, t_n)$, d.h. Näherungen für $\overline{u}_j^n$. Die Verfahrenswerte $\underline{v}_j^{n+1}$ auf der neuen Zeitstufe berechnen sich dann folgendermaßen:

(α) Rekonstruktion:

Approximiere die exakte Lösung $\underline{u}(\circ, t_n)$ aus den $\underline{v}_j^n$ derart, daß für alle $x \in \mathbb{R}$, in deren Umgebung $\underline{u}(\circ, t_n)$ hinreichend glatt ist, gilt:

$$\underline{L}^N(x, \underline{v}_j^n) = \underline{u}(x, t_n) + O\left(h^{N+1}\right), \quad N \geq 1.$$

(β) Numerischer Fluß:

Approximiere $\hat{\underline{F}}_{j+\frac{1}{2}} := \frac{1}{\tau} \int_{t_n}^{t_{n+1}} \underline{F}(\underline{v}(x_{j+\frac{1}{2}}, t))\, dt$ mit $\underline{v}$ Lösung zu

$$\underline{v}_t(x,t) \;+\; \left(\underline{F}(\underline{v}(x,t))\right)_x \;=\; 0, \quad (x,t) \in \mathbb{R} \times (t_n, +\infty);$$

$$\underline{v}(x, t_n) \;=\; \underline{L}^N(x, \underline{v}_j^n), \qquad x \in \mathbb{R}.$$

(γ) Mittelwertbildung:

Berechne die Näherungswerte auf der neuen Zeitstufe durch

$$\underline{v}_j^{n+1} \;=\; \underline{v}_j^n \;-\; \lambda\, \left[\hat{\underline{F}}_{j+\frac{1}{2}} - \hat{\underline{F}}_{j-\frac{1}{2}}\right].$$

Wie im skalaren Fall liegt die Schwierigkeit in der Bestimmung der Rekonstruktion $\underline{L}^N(\circ, \underline{v}_j^n)$ und in der Approximation des numerischen Flusses $\hat{\underline{F}}_{j+\frac{1}{2}}$. Letzteres ist kein Problem, da man den numerischen Fluß des skalaren Schemas komponentenweise auf Systeme übertragen kann. Dies ist für die Rekonstruktion prinzipiell auch möglich, aber — wie im folgenden gezeigt werden soll — nicht zweckmäßig.

2.2.3. Rekonstruktionsmethoden

Zur Übertragung der Rekonstruktionsmethoden vom skalaren Fall auf Systeme bieten sich nun zwei Möglichkeiten an:

(2.20.a) <u>komponentenweise Rekonstruktion:</u>

$$\underline{L}^N(\circ,\overline{u}^n) \;:=\; \left(\, L^N(\circ,\overline{u}_1^n),\ldots,L^N(\circ,\overline{u}_m^n)\,\right);$$

(2.20.b) <u>charakteristische Rekonstruktion:</u>

$$\underline{L}^N(\circ,\overline{u}^n) \;:=\; \sum_{k=1}^{m} L^N\left(\circ,\overline{w}_j^k\right)\ \underline{r}_k(\overline{u}_j^n),\;\; x\in[x_{j-\frac{1}{2}},x_{j+\frac{1}{2}}) \text{ mit}$$

$$\overline{w}_j^k(x) \;:=\; \langle\, \underline{l}_k(\overline{u}_j^n),\overline{u}^n(x)\,\rangle\ ,x\in\mathbb{R};$$

$\underline{r}_k(\underline{u}),\underline{l}_k(\underline{u})$ die Links– und Rechtseigenvektoren gemäß (2.18).

Während (2.20.a) auf der Hand liegt, beruht (2.20.b) auf der Tatsache, daß die $\{\ \underline{r}_k(\underline{u})\ \}_{k=1}^m$ und $\{\ \underline{l}_k(\underline{u})\ \}_{k=1}^m$ Basen des $\mathbb{R}^m$ sind, und somit gilt aufgrund der Normierung (2.18.b):

(2.21)

$$\underline{u} \;=\; \sum_{k=1}^{m} w_k(\underline{v},\underline{u})\,\underline{r}_k(\underline{v}) \quad \text{mit}$$

$$w_k(\underline{v},\underline{u}) \;:=\; \langle\ l_k(\underline{v}),\underline{u}\ \rangle.$$

Zwischen den beiden Methoden besteht ein entscheidender Unterschied, wie aus folgendem Lemma ersichtlich wird:

Lemma 2.20:

Vor.: $A := \underline{F}'(\underline{u}) = const \in \mathbb{R}^{m\times m},\;\; \underline{u}_L,\underline{u}_R\in\mathbb{R}^m,\;\; \det A\neq 0,$

$$\underline{u}_0(x) \;=\; \left\{\begin{matrix} \underline{u}_L & ,x<0 \\ \underline{u}_R & ,x>0\end{matrix}\right\};$$

Beh.: Für die Lösung von (2.18) gilt:

a)

$$\underline{u}(x,t) \;=\; \left\{\begin{matrix} \underline{u}_L & ,\frac{x}{t}<\lambda_1 \\ \\ \underline{u}^k & ,\lambda_k<\frac{x}{t}<\lambda_{k+1},\;\; k=1,\ldots,m-1 \\ \\ \underline{u}_R & ,\lambda_m<\frac{x}{t}\end{matrix}\right\} \quad \text{mit}$$

$$\underline{u}^k \;=\; \underline{u}_L \;+\; \sum_{i=1}^{k}\langle\ \underline{l}_i,\underline{u}_R-\underline{u}_L\ \rangle\,\underline{r}_i;$$

b)

$$\underline{u}(x,t) \;=\; \sum_{k=1}^{m} w^k(x,t)\,\underline{r}_k \quad \text{mit } w^k(x,t) \;:=\; \left\{\begin{matrix} \langle\ \underline{l}_k,\underline{u}_L\ \rangle & ,\frac{x}{t}<\lambda_k \\ \\ \langle\ \underline{l}_k,\underline{u}_R\ \rangle & ,\frac{x}{t}>\lambda_k\end{matrix}\right\}.$$

Beweis:

Zu a): Da A konstant ist, sind auch die Eigenwerte und die zugehörigen Links– und Rechtseigenvektoren konstant. Somit gilt (vgl. [Smo] bzw. Kap. 4.4):

$$\langle\, grad_{\underline{u}}\, \lambda_k, \underline{r}_k \,\rangle \;\equiv\; 0,$$

d.h.: alle k–Felder sind linear degeneriert (s. Kap. 4.4).
Die Lösung setzt sich deshalb nur aus Kontaktunstetigkeiten zusammen, d.h. sie hat die Struktur aus Behauptung a) (vgl. (4.21)), und für die noch zu bestimmenden Zustände $\underline{u}^k,\ k = 1, \dots, m$ mit $\underline{u}^0 := \underline{u}_L, \underline{u}^0 := \underline{u}_R$ muß gelten (s. [Smo], S. 334):

$$\begin{aligned} \tfrac{d}{d\varepsilon_k}\, \underline{u}_k(\varepsilon_k) \;&=\; \underline{r}_k(\underline{u}_k(\varepsilon_k)), \quad \varepsilon_k > 0 \\ \underline{u}_k(0) \;&=\; \underline{u}^{k-1}. \end{aligned}$$

Die Lösung dieser gewöhnlichen Differentialgleichungen existiert für alle $\varepsilon_k > 0$, denn $\underline{r}_k$ ist konstant, und somit gilt:

(1) $$\underline{u}_k(\varepsilon_k) \;=\; \underline{r}_k\, \varepsilon_k \;+\; \underline{u}_k(0) \;=\; \underline{r}_k\, \varepsilon_k \;+\; \underline{u}^{k-1}, \quad k = 1, \dots, m.$$

Setze nun

(2) $$\underline{u}^k \;:=\; \underline{u}_k(\varepsilon_k), \quad k = 1, \dots, m.$$

Dann gilt insbesondere:

(3) $$\underline{u}_0(\varepsilon_0) \;=\; \underline{u}_L, \quad \underline{u}_m(\varepsilon_m) \;=\; \underline{u}_R.$$

Durch Einsetzen von (2) in (1) erhält man somit:

$$\underline{u}_k(\varepsilon_k) \;=\; \underline{r}_k\, \varepsilon_k \;+\; \underline{u}_{k-1}(\varepsilon_{k-1}).$$

Daraus folgt mit (3) unmittelbar:

$$\underline{u}_R \;=\; \sum_{k=1}^{m} \underline{r}_k\, \varepsilon_k \;+\; \underline{u}_L.$$

Dies läßt sich als lineares Gleichungssystem mit m Unbekannten und m Gleichungen schreiben:

(4) $$\underline{u}_R \;-\; \underline{u}_L \;=\; (\, \underline{r}_1, \dots, \underline{r}_m \,)\; (\varepsilon_1, \dots, \varepsilon_m)^T.$$

Setze:

$$R := (\, \underline{r}_1, \dots, \underline{r}_m \,), L := (\, \underline{l}_1, \dots, \underline{l}_m \,)^T.$$

Dann gilt aufgrund der Normierung (2.18):

(5) $$L\, R \;=\; R\, L \;=\; Id \quad \Longleftrightarrow \quad L \;=\; R^{-1}.$$

Somit gilt für die Lösung von (4):

$$(\varepsilon_1, \dots, \varepsilon_m)^T \;=\; L\, (\underline{u}_R \;-\; \underline{u}_L) \quad \Longleftrightarrow \quad \varepsilon_k \;=\; \langle\, \underline{l}_k, (\underline{u}_R \;-\; \underline{u}_L)\, \rangle, \; k = 1, \dots, m.$$

Durch Rückeinsetzen in (2) erhält man mit (3) die Behauptung a).

Zu b): Setze nun $\Lambda := \text{diag}\,(\lambda_1, \ldots, \lambda_m)$. Dann gilt mit (5): $LAR = \Lambda$. Für $\underline{w} := L\,\underline{u}$ ergibt sich dann:

$$\left\{\begin{array}{lcll} \underline{u}_t(x,t) \;+\; A\,\underline{u}_x(x,t) & = & 0, & (x,t) \in \mathbb{R} \times \mathbb{R}_+ \\ \underline{u}(x,0) \;=\; \underline{u}_0(x), & & & x \in \mathbb{R} \end{array}\right\} \Longleftrightarrow$$

$$\left\{\begin{array}{lcll} \underline{w}_t(x,t) \;+\; \Lambda\,\underline{w}_x(x,t) & = & 0, & (x,t) \in \mathbb{R} \times \mathbb{R}_+ \\ \underline{w}(x,0) \;=\; L\,\underline{u}_0(x), & & & x \in \mathbb{R} \end{array}\right\}.$$

Dies läßt sich als entkoppeltes System wie folgt schreiben:

$$w_t^k(x,t) \;+\; \lambda_k\, w_x^k(x,t) \;=\; 0, \;\; (x,t) \in \mathbb{R} \times \mathbb{R}_+$$

$$w^k(x,0) \;=\; \langle\, \underline{l}_k, \underline{u}_0(x)\,\rangle, \;\; x \in \mathbb{R}, \quad k = 1, \ldots, m.$$

Für die Lösung eines skalaren Erhaltungssatzes gilt bekanntlich:

$$w^k(x,t) \;=\; w^k(x - \lambda_k\, t, 0).$$

Daraus folgt unmittelbar die Behauptung b).

■

Die Lösungen sind zwar identisch, dennoch läßt sich an den beiden Darstellungen der Unterschied zwischen den Rekonstruktionsmethoden (2.20) sehr gut verdeutlichen:

Verwendet man die komponentenweise Methode, so wird die skalare Rekonstruktionstechnik auf die einzelnen Komponenten der Lösung $\underline{u}(\circ, t)$ aus La. 2.20 angewendet. Diese *koppeln* aber alle Kontaktunstetigkeiten. Die charakteristische Methode setzt die skalare Technik dagegen auf die s.g. charakteristischen Variablen $w^k(\circ, t)$, $k = 1, \ldots, m$ aus La. 2.20 an. Dadurch werden die Kontaktunstetigkeiten voneinander getrennt, d.h. jede zu rekonstruierende Größe hat *genau einen* Sprung. Die Kontaktunstetigkeiten werden erst *nach* der Rekonstruktion der einzelnen charakteristischen Variablen wieder mittels (2.21) gekoppelt.

In [ENO3] werden beide Methoden auf die Eulergleichnung mit dem s.g. Sod– bzw. Lax–Testproblem als Anfangsdaten angewendet. In beiden Fällen treten Oszillationen bei der Benutzung der komponentenweisen Rekonstruktion auf. Daher ist es sinnvoll die charakteristische zu verwenden. Diese setzt sich aus drei Schritten zusammen:

(2.22.a) <u>Stückweise Entkoppelung:</u>

Führe mittels (2.21) eine Basistransformation durch:

$$\underline{u} \;=\; \sum_{k=1}^{m} w_k(\overline{u}_j^n, \underline{u})\; \underline{r}_k(\overline{u}_j^n);$$

(2.22.b) <u>Skalare Rekonstruktion:</u>

Wende auf $w_j^k(x) := w_k(\overline{u}_j^n, \underline{u}(x, t_n)), x \in \mathbb{R}$ die Rekonstruktionsmethode aus dem skalaren Fall an, d.h.: bestimme $L^N(\circ, \overline{w}_j^k)$ auf $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ zu den Werten

$$\overline{w}_{j,i}^k := \overline{w}_j^k(x_i) = \langle\, \underline{l}_k(\overline{u}_j^n), \overline{u}_i^n\,\rangle, \;\; i = j - N, \ldots, j + N;$$

(2.22.c) Charakteristische Rekonstruktion:

$$\underline{L}^N(x, \overline{u}_j^n) := \sum_{k=1}^{m} L^N(x, \overline{w}_j^k)\, \underline{r}_k(\overline{u}_j^n).$$

Durch diese Rekonstruktion übertragen sich alle Eigenschaften aus dem skalaren Fall, denn:

Lemma 2.21:

Vor.: $\underline{u} : \mathbb{R} \to \mathbb{R}^m,\ m > 1,\ \{\underline{r}_j\}_{j=1}^m, \{\underline{l}_j\}_{j=1}^m$ Basen des $\mathbb{R}^m$ mit $\langle\, \underline{r}_j, \underline{l}_k \,\rangle \;=\; \delta_{j,k}$;

$\underline{w}(x) := (w^1, \ldots, w^m)$ mit $w^i(x) := \langle\, \underline{l}_i, \underline{u}(x) \,\rangle,\ x \in \mathbb{R}$,

$L^N(\circ, \overline{w}^i)$ skalare Rekonstruktionsmethode, die (2.2) erfüllt,

$\underline{L}^N(\circ, \overline{u})$ gemäß (2.22);

Beh.: $\underline{L}^N(\circ, \overline{u})$ erfüllt ebenfalls (2.2), d.h.:

a) Für alle Punkte x, die eine Umgebung besitzen, in der $\underline{u}$ glatt ist, gilt:
$\underline{L}^N(x, \overline{u}) \;=\; \underline{u}(x) \;+\; O\left(h^{N+1}\right)$;

b) $\overline{\underline{L}}^N(x_j, \overline{u}) := \frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \underline{L}^N(x, \overline{u})\, dx \;=\; \overline{u}(x_j)$;

c) $\mathrm{TV}(\underline{L}^N(\circ, \overline{u})) \;\leq\; \mathrm{TV}(\overline{u}) \;+\; O\left(h^{q+1}\right),\ q > 0.$

Beweis:

Aus (2.2) folgt mit

$$(1) \qquad \underline{u} \;=\; \sum_{k=1}^{m} \langle\, \underline{l}_k, \underline{u} \,\rangle\, \underline{r}_k, \quad \underline{u} \in \mathbb{R}^m$$

die Behauptung, denn:

Zu a):

$$\underline{L}^N(x, \overline{u}) \overset{(2.22.c)}{=} \sum_{k=1}^{m} L(x, \overline{w}^k)\, \underline{r}_k \overset{(2.2.a)}{=} \sum_{k=1}^{m} \left(w^k(x) \;+\; O\left(h^{N+1}\right)\right) \underline{r}_k$$
$$\overset{\text{def.}}{=} \sum_{k=1}^{m} \langle\, \underline{l}_k, \underline{u}(x) \,\rangle\, \underline{r}_k \;+\; O\left(h^{N+1}\right) \overset{(1)}{=} \underline{u}(x) \;+\; O\left(h^{N+1}\right).$$

Zu b):

$$\overline{\underline{L}}^N(x, \overline{u}) \overset{(2.22.c)}{=} \sum_{k=1}^{m} \left(\frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} L(x, \overline{w}^k)\, dx\right) \underline{r}_k \overset{\text{def.}}{=} \sum_{k=1}^{m} \overline{L}(x_j, \overline{w}^k)\, \underline{r}_k$$
$$\overset{(2.2.b)}{=} \sum_{k=1}^{m} \overline{w}^k(x_j)\, \underline{r}_k \overset{\text{def.}}{=} \sum_{k=1}^{m} \langle\, \underline{l}_k, \overline{u}(x_j) \,\rangle\, \underline{r}_k \overset{(1)}{=} \overline{u}(x_j).$$

Zu c): Die totale Variation einer vektorwertigen Funktion läßt sich komponentenweise wie folgt definieren:

$$\mathrm{TV}(\underline{u}) \;:=\; \sum_{k=1}^{m} \mathrm{TV}(u_k) \quad \text{mit } \underline{u} := (u_1, \ldots, u_m)^T.$$

Setze nun

(2)
$$c_l \;:=\; \sum_{i=1}^{m} \sum_{k=1}^{m} l_{k,l}\, r_{k,i}\;;$$

(3)
$$L := (\underline{l}_1, \ldots, \underline{l}_m)^T, R := (\underline{r}_1, \ldots, \underline{r}_m).$$

Da nach Voraussetzung $\langle\, \underline{r}_j\,, \underline{l}_k\,\rangle \;=\; \delta_{j,k}$, gilt insbesondere:

$$L\,R \;=\; R\,L \;=\; Id.$$

Daraus folgt mit (3):

$$\sum_{k=1}^{m} r_{k,i}\; l_{k,j} \;=\; \delta_{i,j}$$

und somit

(4)
$$c_l \;=\; \sum_{i=1}^{m} \delta_{l,i} \;=\; 1.$$

Dann gilt:

$$\begin{aligned}
&\mathrm{TV}(\underline{L}^N(\circ, \overline{u})) \;\overset{\text{def.}}{=} \\
&\sum_{i=1}^{m} \mathrm{TV}(\underline{L}_i^N(\circ, \overline{u})) \;\overset{(2.22.c)}{=}\; \sum_{i=1}^{m} \mathrm{TV}\left(\sum_{k=1}^{m} L^N(\circ, \overline{w}^k)\, r_{k,i} \right) \;\le \\
&\sum_{i=1}^{m} \sum_{k=1}^{m} \mathrm{TV}\left(\, L^N(\circ, \overline{w}^k)\,\right)\; r_{k,i} \;\overset{(2.2.c)}{\le}\; \sum_{i=1}^{m} \sum_{k=1}^{m} \left[\, \mathrm{TV}(w^k) \;+\; O\left(h^{q+1}\right)\,\right]\; r_{k,i} \;\overset{\text{def.}}{=} \\
&\sum_{i=1}^{m} \sum_{k=1}^{m} \mathrm{TV}(\,\langle\, \underline{l}_k, \underline{u}\,\rangle)\, r_{k,i} + O\left(h^{q+1}\right) \;\le\; \sum_{i=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} l_{k,l}\; \mathrm{TV}(u_l)\, r_{k,i} + O\left(h^{q+1}\right) \;\overset{(2)}{=} \\
&\sum_{l=1}^{m} c_l\; \mathrm{TV}(u_l) \;+\; O\left(h^{q+1}\right) \;\overset{(4)}{=}\; \sum_{l=1}^{m} \mathrm{TV}(u_l) \;+\; O\left(h^{q+1}\right) \;\overset{\text{def.}}{=}\; \mathrm{TV}(\underline{u}) \;+\; O\left(h^{q+1}\right).
\end{aligned}$$

■

Verwendet man zur Konstruktion von $\underline{L}^N(\circ, \overline{u}_j^n)$ die skalaren Rekonstruktionsmethoden $L^{N,RP}(\circ, \overline{w}^k)$ bzw. $L^{N,RD}(\circ, \overline{w}^k)$ gemäß Kapitel 2.1.3, dann werden stückweise polynomiale Funktionen $Q^N(\circ, w^k) : \mathbb{R} \to \mathbb{R}$ benötigt mit $Q^N(x_j, w^k) = w^k(x_j)$. Die ENO–Interpolationstechnik aus Kapitel 1 läßt sich nun analog zur charakteristischen Rekonstruktion (2.22) auf Systeme übertragen:

$$(2.23) \qquad \underline{Q}^N(x, \underline{u}) := \sum_{k=1}^{m} Q^N(x, w^k)\, \underline{r}_k(\overline{u}_j^n), \quad x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}).$$

Dann lassen sich alle Eigenschaften aus Kapitel 1 unter Berücksichtigung der Definition von w^k gemäß Voraussetzung zu La. 2.21 analog zum Beweis dieses Lemmas zeigen.

Für die Implementation hat die RD–Methode hier gegenüber der RP– Rekonstruktion einen erheblichen Nachteil:

Zu jedem Intervall $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ wird gemäß (2.22.a) eine Basistransformation vorgenommen, die für verschiedene $j \in \mathbb{Z}$ im allgemeinen nicht identisch ist, da sie von u abhängig ist. Es werden aber $Q_{j-1}^N(\circ, \overline{w}_j^k)$ *und* $Q_j^N(\circ, \overline{w}_j^k)$ (vgl. (2.4)) auf $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ benötigt. Diese hängen jedoch wegen den $\overline{w}_j^k$ von der selben Basistransformation ab (vgl. (2.22)). Im allgemeinen gilt aber:

$$Q_j^N(\circ, \overline{w}_j^k) \;\not\equiv\; Q_j^N(\circ, \overline{w}_{j+1}^k).$$

Somit müssen zu jedem Intervall *zwei* Interpolationspolynome berechnet werden! Bei der RP–Rekonstruktion wird dagegen nur $Q_{j-\frac{1}{2}}^N(\circ, \overline{W}_j^k)$ benötigt, also nur *ein* Polynom. Somit ist die RP–Methode aufgrund der niedrigeren Laufzeiten vorzuziehen. Um jedoch die selbe Ordnung zu erreichen, muß man gegenüber der RD–Rekonstruktion den Grad des ENO–Polynoms um 1 erhöhen (vgl. La. 2.1, La. 2.8 und F. 2.3).

Abschließend sei noch vermerkt, daß sich wegen La. 2.21 die Lemmata 2.1, 2.4, 2.8 und 2.9 aus dem skalaren Fall auf $\underline{L}^N(\circ, \overline{u}_j^n)$ übertragen lassen.

2.2.4 Numerischer Fluß

Nachdem $\underline{L}^N(\circ, v_j^n)$ explizit konstruiert ist, kann man nun den numerischen Fluß

$$(2.24.a) \qquad \hat{\underline{F}}_{j+\frac{1}{2}} := \frac{1}{\tau} \int_{t_n}^{t_{n+1}} \underline{F}(\underline{v}(x_{j+\frac{1}{2}}, t))\, dt \quad \text{mit } v \text{ Lösung zu}$$

$$(2.24.b) \qquad \begin{aligned} \underline{v}_t(x,t) \;+\; (\underline{F}(\underline{v}(x,t)))_x &= 0 \quad (x,t) \in \mathbb{R} \times (t_n, +\infty); \\ \underline{v}(x, t_n) &= \underline{L}^N(x, \underline{v}_j^n), \quad x \in \mathbb{R} \end{aligned}$$

approximieren, in dem man die skalaren Techniken (s. Kap. 2.1.4) komponentenweise auf $\underline{\hat{F}}_{j+\frac{1}{2}}$ anwendet. Dazu ersetzt man zunächst analog zu (2.6) das Integral durch eine Quadraturformel, d.h.:

$$\underline{\hat{F}}_{j+\frac{1}{2}} \stackrel{!}{=} \tfrac{1}{\tau} \sum_{i=1}^{k} a_i \; \underline{F}(\underline{v}(x_{j+\frac{1}{2}}, \tilde{t}_i)) \;+\; O\left(h^{N+1}\right) \tag{2.25}$$

wobei mit k die Anzahl der Stützstellen, a_i die Gewichte und $\tilde{t}_i$ die Stützstellen in $[t_n, t_{n+1}]$ bezeichnet werden.
Für die Approximation von $\underline{v}(x_{j+\frac{1}{2}}, \tilde{t}_i)$ bieten sich nun wieder zwei Möglichkeiten an:

a) Taylorentwicklung und anschließendes Lösen von Riemann–Problemen (vgl. Def. 4.2), entweder exakt mittels einer Godunov–Prozedur oder approximativ mit einer Roe–Prozedur;
b) asymptotische Entwicklung gemäß Raviart.

Bei a) ändert sich gegenüber dem skalaren Fall nur die Taylorentwicklung. Da La. 2.13.a für Systeme im allgemeinen nicht gültig ist (vgl. [Ad1]), verwendet man hier die Cauchy–Kowaleski–Prozedur, d.h.:

$$\underline{v}^N(x, \tilde{t}_i) := \sum_{s=0}^{N} \sum_{l=0}^{s} \frac{\partial^s}{\partial^l x \; \partial^{t-l} t} \, \underline{v}|_{(x_j, t_n)} \frac{(\tilde{t}_i - t_n)^{s-l}}{(s-l)!} \; \frac{(x - x_j)^s}{s!}.$$

In Bereichen, in denen $\underline{v}$ glatt ist, gilt dann wieder:

$$\underline{v}(x, \tilde{t}_i) = \underline{v}^N(x, \tilde{t}_i) \;+\; O\left(h^{N+1}\right).$$

Setze nun in Analogie zu (2.7):

$$\underline{v}_L := \underline{v}^N(x_{j+\frac{1}{2}} - 0, \tilde{t}_i), \;\; \underline{v}_R := \underline{v}^N(x_{j+\frac{1}{2}} + 0, \tilde{t}_i). \tag{2.26}$$

Dann lassen sich die diskreten Stellen aus (2.24) wie folgt approximieren:

$$\underline{F}(\underline{v}(x_{j+\frac{1}{2}}, \tilde{t}_i)) = \underline{F}(\underline{V}(0, \underline{v}_L^N, \underline{v}_R^N)) \;+\; O\left(h^{N+1}\right); \tag{2.27}$$

$$\underline{V}\left(\frac{x - x_{j+\frac{1}{2}}}{t - \tilde{t}_i}, \underline{v}_L^N, \underline{v}_R^N\right) := \underline{v}(x,t) \text{ Lösung des Riemann–Problems}$$

$$\underline{v}_t(x,t) + (\underline{F}(\underline{v}(x,t)))_x = 0, \quad (x,t) \in \mathbb{R} \times (\tilde{t}_i, +\infty);$$

$$\underline{v}(x, \tilde{t}_i) = \left\{ \begin{array}{l} \underline{v}_L^N \;, x < x_{j+\frac{1}{2}} \\ \underline{v}_R^N \;, x > x_{j+\frac{1}{2}} \end{array} \right\}.$$

Im Gegensatz zum skalaren Fall läßt sich die Lösung des Riemann–Problems nicht allgemein angeben, sondern muß in Abhängigkeit von dem zugrundeliegenden System von Erhaltungssätzen jeweils einzeln bestimmt werden. Dies ist im allgemeinen sehr aufwendig, wie man zum Beispiel an der Lösung für die Eulergleichung (s. Kap. 6 bzw. [Smo]) sieht. Es ist daher günstiger mit der Roe–Linearisierung zu arbeiten, die sich für die Eulergleichung verhältnismäßig einfach angeben läßt (s. Kap. 3).

Die Methode von Raviart (vgl. Kap. 4) ist zwar noch um einiges aufwendiger zu programmieren als die exakte Lösung des Riemann–Problems, dafür kann aber zum Teil mit größeren CFL–Zahlen gearbeitet werden. Bei Interaktionsproblemen (s. Kap. 5) wird die Kollision wesentlich besser erfaßt (s. Kap. 7). daher ist der Mehraufwand gerechtfertigt.

Somit ergeben sich wieder (vgl. (2.10), (2.12)) folgende Approximationen:

(2.28.a) $\underline{F}^{God}_{j+\frac{1}{2}}(\tilde{t}_i, \underline{v}^N_L, \underline{v}^N_R) := \underline{F}(\underline{V}(0, \underline{v}^N_L, \underline{v}^N_R))$;

(2.28.b) $\underline{F}^{Roe}_{j+\frac{1}{2}}(\tilde{t}_i, \underline{v}^N_L, \underline{v}^N_R) := \frac{1}{2}\left[\underline{F}(\underline{v}^N_L) + \underline{F}(\underline{v}^N_R) - \sum_{k=1}^{m} |\lambda_k| \langle \underline{l}_k, \underline{v}^N_L - \underline{v}^N_R \rangle \underline{r}_k\right]$

mit λ_k Eigenwert zu $A_{L,R}$ und $\underline{l}_k$ bzw. $\underline{r}_k$ die zugehörigen Links– und Rechtseigenvektoren und $A_{L,R}$ die Roe–Linearisierung gemäß Kap. 3;

(2.28.c) $\underline{F}^{Rav}_{j+\frac{1}{2}}(\tilde{t}_i) := \underline{F}(\tilde{v}(x_{j+\frac{1}{2}}, \tilde{t}_i))$ mit $\tilde{v}(x,t) := \sum_{i=0}^{N} \underline{v}^i \left(\frac{x - x_{j+\frac{1}{2}}}{t - t_n}\right) (t - t_n)^i$

asymptotische Entwicklung von $\underline{v}$ aus (2.24) und $\underline{v}^i$ gemäß Kap. 4.

Durch Einsetzen von (2.28) in (2.24) erhält man nun folgende Näherungen für den numerischen Fluß (2.25):

(2.29.a)
$$\underline{F}^{God}_{j+\frac{1}{2}} := \frac{1}{\tau} \sum_{i=1}^{k} a_i \, \underline{F}^{God}_{j+\frac{1}{2}}(\tilde{t}_i, \underline{v}^N_L, \underline{v}^N_R);$$

(2.29.b)
$$\underline{F}^{Roe}_{j+\frac{1}{2}} := \frac{1}{\tau} \sum_{i=1}^{k} a_i \, \underline{F}^{Roe}_{j+\frac{1}{2}}(\tilde{t}_i, \underline{v}^N_L, \underline{v}^N_R);$$

(2.29.c)
$$\underline{F}^{Rav}_{j+\frac{1}{2}} := \frac{1}{\tau} \sum_{i=1}^{k} a_i \, \underline{F}^{Rav}_{j+\frac{1}{2}}(\tilde{t}_i).$$

Analog zu (2.10) und (2.13) bzw. La. 2.15 gilt nun:

a) $\underline{v} \in C^{N+1}\left((x_{j+\frac{1}{2}} - \varepsilon, x_{j+\frac{1}{2}} + \varepsilon) \times (t_n, t_{n+1})\right), \ \varepsilon > 0$

$\Longrightarrow \ \hat{\underline{F}}_{j+\frac{1}{2}} = \underline{F}^{God}_{j+\frac{1}{2}} \ + \ O\left(h^{N+1}\right) = \underline{F}^{Roe}_{j+\frac{1}{2}} \ + \ O\left(h^{N+1}\right);$

b) $\hat{\underline{F}}_{j+\frac{1}{2}} = \underline{F}^{Rav}_{j+\frac{1}{2}} \ + \ O\left(h^{N+1}\right).$

2.2.5 Eigenschaften

Wegen La. 2.21 lassen sich nun die meisten Eigenschaften aus dem skalaren Fall übertragen. So kann man den Ordnungsbegriff aus Def. 2.11 komponentenweise für Systeme definieren. Damit erhält man unmittelbar aus La. 2.21 und S. 2.3:

Satz 2.7:

Vor.: $\underline{Q}\left(\underline{v}_{j-q}^n, \ldots, \underline{v}_{j+p}^n\right) \;:=\; \underline{v}_j^{n+1}$ gemäß (2.19), $j \in \mathbb{Z},\ p, q, n \in \mathbb{N}_0,\ N \in \mathbb{N}$;
$\underline{L}^N(\circ, \underline{v}_j^n)$ erfülle (2.2.a,b);
$\underline{u}$ starke Lösung eines Erhaltungssatzes in einer Umgebung U von $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ mit $\underline{u} \in C^{N+1}(U), \underline{F} \in C^1(\mathbb{R}^m)$;

Beh.: Für $\tau \to 0$ mit $\lambda = \frac{\tau}{h} = const$ gilt für h hinreichend klein:

$$\overline{u}(x, t+\tau) \;-\; \underline{Q}\left(\overline{u}(x - q\,h, t), \ldots, \overline{u}(x + p\,h, t)\right) \;=\; O\left(h^{N+1}\right).$$

Daraus läßt sich nun wieder (vgl. F. 2.3) ableiten:

Folgerung 2.6:

Die ENO–Verfahren für Systeme approximieren die Mittelwerte der exakten Lösung von N–ter Ordnung, die exakte Lösung dagegen von 2. Ordnung für $N \geq 2$ sonst von 1. Ordnung (komponentenweise).

Hier verringert sich bei Verwendung von $\underline{L}^N(\circ, \underline{v}_j^n) \;=\; \underline{L}^{N,RP}(\circ, \underline{v}_j^n)$ die Ordnung der Mittelwertapproximation wie im skalaren Fall um 1.

Leider läßt sich der Satz 2.6 so nicht übertragen, denn zum einen gilt der Satz von Oleinik (vgl. [Smo]) nicht und bei der Interaktion von zwei Stoßwellen können im Gegensatz zu La. 2.16 auch mehr als eine neue Diskontinuitätskurve entstehen, sogar neue Verdünnugswellen (s. Kap. 5). Zum Satz von Schaeffer (S. 2.4) existiert jedoch ein Analogon für Systeme (s. [Di1]). Dies berechtigt zu der Hoffnung, daß sich S. 2.5 auch auf Systeme übertragen läßt und somit auch S. 2.6.

Keine Probleme ergeben sich dagegen bei La. 2.17, La. 2.18 und La. 2.19:

Lemma 2.22:

Vor .: $\left\{\underline{v}_j^n\right\}_{j \in \mathbb{Z}, n \in \mathbb{N}_0}$ ENO–Verfahren gemäß (2.19);

Beh.: $\underline{L}^N(\circ, \underline{v}_j^n) = \underline{L}^{1,RP}(\circ, \underline{v}_j^n)$ oder $\underline{L}^N(\circ, \underline{v}_j^n) = \underline{L}^{0,RD}(\circ, \underline{v}_j^n) \quad \Longrightarrow$
Das ENO–Verfahren ist identisch mit dem Godunov–Verfahren.

Beweis:

Wegen (2.23) gilt analog zu La. 2.17:

$$\underline{L}^{1,RP}(x, \underline{v}_j^n) \;=\; \underline{L}^{0,RD}(x, \underline{v}_j^n) \;=\; \underline{v}_j^n \quad \text{für} \quad x \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}).$$

Aus der Konstruktion der ENO–Verfahren gemäß (2.19) folgt die Behauptung.

■

Im linearen Fall gilt analog zu La. 2.18:

Lemma 2.23:

Vor.: $\underline{u}_0 \in L^\infty(\mathbb{R})$, $A \in \mathbb{R}^{m\times m}, \underline{b} \in \mathbb{R}^m$, $\det A \neq 0$, $m, N \in \mathbb{N}$, $\underline{F}(\underline{u}) := A\,\underline{u} + \underline{b}$,

$\lambda_1 < \cdots < \lambda_m$ Eigenwerte von A,

$\{\underline{v}_j^n\}_{j\in\mathbb{Z}, n\in\mathbb{N}_0}$ das gemäß (2.19) definierte ENO–Verfahren, $\lambda := \frac{\tau}{h}$, $\tau, h > 0$,

$\underline{u}$ Lösung des linearen Erhaltungssatzes;

Beh.: Falls $\underline{L}^N(\circ, \underline{v}_j^n) = \underline{L}^{N,RP}(\circ, \underline{v}_j^n)$ gilt:

$$\underline{v}_{j,k}^{n+1} \;=\; \frac{1}{h}\,\Big[Q^N(x_{j+\frac{1}{2}} - \lambda_k\,\tau, W^k) \;-\; Q^N(x_{j-\frac{1}{2}} - \lambda_k\,\tau, W^k) \Big], k = 1, \ldots, m.$$

Beweis:

Setze $L := (\,\underline{l}_1, \ldots, \underline{l}_m\,)^T$ wobei die $\underline{l}_k$, $k = 1, \ldots, m$, die Linkseigenvektoren von A zu den Eigenwerten λ_k sind. Dann gilt mit $\underline{w} := L\,\underline{u}$ analog zum Beweis zu La. 2.20.b:

$$\left\{ \begin{array}{ll} \underline{u}_t(x,t) + A\,\underline{u}_x(x,t) = 0, & x \in \mathbb{R}, t > 0 \\ \underline{u}(x,0) = \underline{u}_0(x), & x \in \mathbb{R} \end{array} \right\} \Longleftrightarrow$$

$$\left\{ \begin{array}{lll} w_t^k(x,t) + \lambda_k\, w_x^k(x,t) = 0, & x \in \mathbb{R}, t > 0 & \\ w^k(x,0) = \langle\, \underline{l}_k, \underline{u}_0(x)\,\rangle, & x \in \mathbb{R}, & k = 1, \ldots, m \end{array} \right\}.$$

Dies ist ein System von m unabhängigen skalaren Erhaltungssätzen. Auf jede einzelne Komponente kann man das skalare ENO–Verfahren anwenden. Mit La. 2.18.b erhält man unmittelbar die Behauptung.

■

Des weiteren gilt wegen (2.19), (2.22) und La. 2.19:

Lemma 2.24:

Die ENO–Verfahren für Systeme sind in Erhaltungsform.

Kapitel 3: Riemann–Löser von Roe

Bei der Bestimmung des numerischen Flusses für die ENO–Verfahren (s. Kap. 2.1.4 bzw. Kap. 2.2.4) müssen Riemann–Probleme (s. Def. 4.2) gelöst werden . Dies ist für Systeme in der Regel sehr aufwendig. Man ist daher bestrebt, diesen Aufwand durch Approximationen zu reduzieren. Eine Möglichkeit ist die Linearisierung von Roe (s. [Ro1], [Ro2], [HaLL]), die im folgenden näher erklärt werden soll.

3.1 Herleitung der Roe–Linearisierung

Ausgangspunkt der Betrachtungen ist das Riemann–Problem für ein strikt hyperbolisches System von Erhaltungssätzen:

$$(3.1)\qquad \begin{aligned} \underline{u}_t(x,t) \ + \ (\underline{F}(\underline{u}(x,t)))_x \ &= \ 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+; \\ \underline{u}(x,0) \ &= \begin{Bmatrix} \underline{u}_L & , x < 0 \\ \underline{u}_R & , x > 0 \end{Bmatrix}. \end{aligned}$$

Es soll nun eine Linearisierung $A_{L,R} \in \mathbb{R}^{m\times m}$ von $\underline{F}'(\underline{u})$ bestimmt werden, so daß die Unstetigkeiten der exakten Lösung möglichst genau erfaßt werden.
Da die Lösung entlang einer Schockkurve die Sprungbedingung

$$(3.2)\qquad [\ \underline{F}(\underline{u}_L) - \underline{F}(\underline{u}_R)\ ] = s\ [\ \underline{u}_L - \underline{u}_R\ ] \text{ für ein } s \in \mathbb{R}$$

erfüllen muß, ist es sinnvoll, die folgenden Eigenschaften von $A_{L,R}$ zu fordern:

(3.3.a) $\Lambda : \mathbb{R}\ (\ \underline{u}_L - \underline{u}_R\ ) \to \mathbb{R}\ (\ \underline{F}(\underline{u}_L) - \underline{F}(\underline{u}_R)\ )$ definiert durch $\Lambda(\underline{u}) := A_{L,R}\ \underline{u}$ ist surjektiv;

(3.3.b) $\underline{u}_L = \underline{u}_R = \underline{u} \in \mathbb{R}^m \implies A_{L,R} = \underline{F}'(\underline{u})$;

(3.3.c) $A_{L,R}\ (\ \underline{u}_L - \underline{u}_R\ ) = [\ \underline{F}(\underline{u}_L) - \underline{F}(\underline{u}_R)\ ], \quad \forall\ \underline{u}_L, \underline{u}_R \in \mathbb{R}^m$;

(3.3.d) $A_{L,R}$ hat m verschiedene reelle Eigenwerte.

Die Existenz von $A_{L,R}$ folgt nun aus

Lemma 3.1:

Vor.: $\underline{F} \in C^1(\mathbb{R}^m),\ \underline{u}_L, \underline{u}_R \in \mathbb{R}^m,\ m \in \mathbb{N}$;

Beh.: $A_{L,R} := \int_0^1 \underline{F}'(\underline{u}_L + \tau(\ \underline{u}_R - \underline{u}_L\ ))\ d\tau$ (komponentenweise) erfüllt (3.3).

Beweis:

Setze $\underline{\tilde{u}}(\tau) := \underline{u}_L + \tau(\ \underline{u}_R - \underline{u}_L\ ),\ \tau \in [0,1]$. Dann gilt:

$$\begin{aligned} \underline{F}(\underline{u}_R) - \underline{F}(\underline{u}_L) = \underline{F}(\underline{\tilde{u}}(1)) - \underline{F}(\underline{\tilde{u}}(0)) &= \int_0^1 \tfrac{\partial}{\partial \tau}\ (\underline{F}(\underline{\tilde{u}}(\tau)))\ d\tau \\ &= \int_0^1 \underline{F}'(\underline{\tilde{u}}(\tau))\ \frac{d}{d\tau}\underline{\tilde{u}}(\tau)\ d\tau = A_{L,R}\ (\ \underline{u}_R - \underline{u}_L\ ). \end{aligned}$$

Die Eigenschaften (3.3) ergeben sich nun aus der Definition von $A_{L,R}$ und der Tatsache, daß die Funktionalmatrix von $\underline{F}$ m reelle verschiedene reelle Eigenwerte hat (s. Def. 2.13).

∎

Es soll nun gezeigt werden, daß das linearisierte Problem die Schockkurven der exakten Lösung von (3.1) erfaßt:

Lemma 3.2:

Vor.: $\underline{u}_L, \underline{u}_R \in \mathbb{R}^m$, $m \in \mathbb{N}$, seien so gewählt, daß sie durch genau einen Schock in der Lösung $\underline{u}$ von (3.1) getrennt werden,
$A_{L,R} := A(\underline{u}_L, \underline{u}_R)$ erfülle die Bedingung (3.3); $\underline{v}$ Lösung von

$$\begin{aligned} \underline{v}_t(x,t) \ + \ A_{L,R}\, \underline{v}_x(x,t) &= \ 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+; \\ \underline{v}(x,0) \ &= \begin{Bmatrix} \underline{u}_L & , x < 0 \\ \underline{u}_R & , x > 0 \end{Bmatrix}. \end{aligned}$$

Beh.: Die charakteristische Schockgeschwindigkeit von $\underline{u}$ ist gleich der charakteristischen Geschwindigkeit einer der Kontaktunstetigkeiten der Lösung $\underline{v}$ des linearisierten Problems.

Beweis:

Für die charakteristische Schockgeschwindigkeit $s \in \mathbb{R}$ von $\underline{u}$ gilt:

$$A_{L,R}\,[\,\underline{u}_L - \underline{u}_R\,] \ \overset{(3.3)}{=} \ [\,\underline{F}(\underline{u}_L) - \underline{F}(\underline{u}_R)\,] \ \overset{(3.2)}{=} \ s\,[\,\underline{u}_L - \underline{u}_R\,].$$

Somit ist s ein Eigenwert von $A_{L,R}$. Diese sind aber gerade die charakteristischen Geschwindigkeiten der Kontaktunstetigkeiten von $\underline{v}$ (vgl. La. 2.20.a).

■

3.2 Bestimmung des Roe–Fluß F^{Roe}

Bei numerischen Verfahren ergibt sich nun das Problem, $\underline{F}(\underline{U}(0, \underline{u}_L, \underline{u}_R))$ zu bestimmen, wobei mit $\underline{U}\left(\frac{x}{t}, \underline{u}_L, \underline{u}_R\right)$ die Lösung $\underline{u}(x,t)$ von (3.1) bezeichnet wird. Da das Riemann–Problem aber in der Regel nur sehr aufwendig zu lösen ist, approximiert man dieses durch das linearisierte Problem

$$(3.4) \qquad \begin{aligned} \underline{w}_t(x,t) \ + \ A_{L,R}\, \underline{w}_x(x,t) &= \ 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+; \\ \underline{w}(x,0) \ &= \begin{Bmatrix} \underline{u}_L & , x < 0 \\ \underline{u}_R & , x > 0 \end{Bmatrix}. \end{aligned}$$

Dabei erfülle $A_{L,R}$ die Bedingungen (3.3). Definiere nun den s.g. Roe–Fluß wie folgt:

$$(3.5) \qquad \underline{F}^{Roe}(\underline{u}_L, \underline{u}_R) := \frac{1}{2} \left[\underline{F}(\underline{u}_L) + \underline{F}(\underline{u}_R) - \sum_{i=1}^{m} |\lambda_i|\, \langle\, \underline{l}_i, \underline{u}_R - \underline{u}_L\,\rangle\, \underline{r}_i \right].$$

Dabei sind die $\lambda_i, i = 1\ldots,m$, die Eigenwerte von $A_{L,R}$ und $\underline{l}_i, \underline{r}_i, i = 1,\ldots,m$, die zugehörigen Links– bzw. Rechtseigenvektoren.

Daß diese Definition sinnvoll ist, geht aus dem folgenden hervor:

Satz 3.1:

Vor.: $\underline{U}\left(\frac{x}{t}, \underline{u}_L, \underline{u}_R\right) := u(x,t)$ Lösung von (3.1) und
$\underline{W}\left(\frac{x}{t}, \underline{u}_L, \underline{u}_R\right) := w(x,t)$ Lösung von (3.4),
$\underline{u}_L, \underline{u}_R \in \mathbb{R}^m,\ m \in \mathbb{N},\ \| \underline{u}_L - \underline{u}_R \|$ hinreichend klein;

Beh.: $\underline{F}\left(\underline{U}\left(\frac{x}{t}, \underline{u}_L, \underline{u}_R\right)\right) = \underline{F}^{Roe}(\underline{u}_L, \underline{u}_R) + O(\ \| \underline{u}_L - \underline{u}_R \|\ )$.

Beweis:

Die Entropielösung von (3.1) besteht bekanntlich aus mindestens $m+1$ Zuständen $\underline{u}^i \in \mathbb{R}^m, i = 1, \ldots, m$, sowie Verdünnungsbereichen (vgl. [Smo] bzw. S. 4.1) und es gilt:

$$(1) \qquad \underline{U}\left(\tfrac{x}{t}, \underline{u}_L, \underline{u}_R\right) = \left\{ \begin{array}{ll} \underline{u}^0 = \underline{u}_L & , \frac{x}{t} < \sigma_1^L \\ \underline{u}^i & , \sigma_i^R < \frac{x}{t} < \sigma_{i+1}^L,\ i = 1, \ldots, m-1 \\ \underline{u}^m = \underline{u}_R & , \frac{x}{t} > \sigma_m^R \end{array} \right\} \text{ mit}$$

$$-\infty < \sigma_i^L < \sigma_i^R < \sigma_{i+1}^L < +\infty,\ i = 1, \ldots, m-1;$$

$$(2) \qquad \frac{x}{t} = const \implies \underline{U}\left(\tfrac{x}{t}, \underline{u}_L, \underline{u}_R\right) = const.$$

Beachte, daß im allgemeinen gilt : $\sigma_i^R \neq \sigma_{i+1}^L,\ i = 1, \ldots, m-1$.
Sei nun $h, \tau > 0$ so gewählt, daß gilt:

$$(3) \qquad \lambda := \frac{\tau}{h} < \frac{1}{2} \max\{\ |\sigma_1^L|, |\sigma_m^R|\ \}.$$

Da $\underline{U}$ insbesondere eine schwache Lösung von (3.1) ist, gilt mit Kor. 2.1:

$$(4) \qquad \int_{-\frac{h}{2}}^{0} \underline{u}(x,\tau)\, dx = \int_{-\frac{h}{2}}^{0} \underline{u}(x,0)\, dx + \int_0^\tau \underline{F}(\underline{u}(-\tfrac{h}{2}, t))\, dt - \int_0^\tau \underline{F}(\underline{u}(0,t))\, dt.$$

Wegen (3) gilt insbesondere: $2\,\lambda < |\sigma_1^L|$ und somit:

$$\underline{u}(x,\tau) \overset{(1)}{=} \underline{u}_L = \underline{u}(x,0), \quad x \in [-\tfrac{h}{2}, 0].$$

Setze

$$\underline{F}_{L,R} := \underline{F}(\underline{u}(0,t)) \overset{(2)}{=} \underline{F}(\underline{U}(0, \underline{u}_L, \underline{u}_R)) \text{ für } t > 0.$$

Dann läßt sich (4) auch schreiben als

$$(\alpha) \qquad \underline{F}_{L,R} = -\frac{1}{\tau} \int_{-\frac{h}{2}}^{0} \underline{u}(x,\tau)\, dx + \frac{1}{2\,\lambda}\, \underline{u}_L + \underline{F}(\underline{u}_L).$$

Durch Anwendung von Kor. 2.1 auf $[0, \frac{h}{2}] \times [0, \tau]$ erhält man analog:

$$(\beta) \qquad \underline{F}_{L,R} = \frac{1}{\tau} \int_0^{\frac{h}{2}} \underline{u}(x,\tau)\, dx - \frac{1}{2\,\lambda}\, \underline{u}_R + \underline{F}(\underline{u}_R).$$

Daraus folgt nun unmittelbar:

$$(5)\qquad \begin{aligned} \underline{F}_{L,R} &= \frac{1}{2}\left[\underline{F}(\underline{u}_L) + \frac{1}{2\,\lambda}\,\underline{u}_L + \frac{1}{\tau}\int_0^{\frac{h}{2}} \underline{u}(x,\tau)\,dx\right] + \\ &\ \frac{1}{2}\left[\underline{F}(\underline{u}_R) - \frac{1}{2\,\lambda}\,\underline{u}_R - \frac{1}{\tau}\int_{-\frac{h}{2}}^{0} \underline{u}(x,\tau)\,dx\right]. \end{aligned}$$

Für die exakte Lösung des linearisierten Problems gilt mit La. 2.20:

$$\underline{w}(x,t) \;=\; \underline{u}_L + \sum_{j=1}^{i} w_j\,\underline{r}_j \;\overset{(2.21)}{=}\; \underline{u}_R - \sum_{j=i+1}^{m} w_j\,\underline{r}_j \;=:\; \underline{w}^i$$

$$\text{für } \lambda_i < \frac{x}{t} < \lambda_{i+1}, i = 0,\ldots,m,\ (x,t) \in \mathbb{R}\times\mathbb{R}_+ \text{ mit}$$

$$\lambda_0 := -\infty, \lambda_{m+1} := +\infty,\ \underline{w}^0 := \underline{u}_L,\ \underline{w}^m := \underline{u}_R,\ w_j := \langle\, \underline{l}_j, \underline{u}_R - \underline{u}_L\,\rangle.$$

Ferner gilt komponentenweise:

$$\underline{w}^i = \underline{u}_L + O(\,\|\,\underline{u}_L - \underline{u}_R\,\|\,),\quad i = 1,\ldots,m\ ,$$

$$\underline{u}(x,t) = \underline{u}_L + O(\,\|\,\underline{u}_L - \underline{u}_R\,\|\,),\quad x\in\mathbb{R}, t>0,$$

und somit auch

$$(7)\qquad \underline{w}^i - u(x,t) = O(\,\|\,\underline{u}_L - \underline{u}_R\,\|\,).$$

Ersetze nun in (α) und (β) $\underline{u}(x,\tau)$ durch $\underline{w}(x,\tau)$ und setze

$$(8)\qquad \begin{aligned} \tilde{\underline{F}}_{L,R} &:= \frac{1}{2}\left[\underline{F}(\underline{u}_L) + \frac{1}{2\,\lambda}\,\underline{u}_L + \frac{1}{\tau}\int_0^{\frac{h}{2}} \underline{w}(x,\tau)\,dx\right] + \\ &\ \frac{1}{2}\left[\underline{F}(\underline{u}_R) - \frac{1}{2\,\lambda}\,\underline{u}_R - \frac{1}{\tau}\int_{-\frac{h}{2}}^{0} \underline{w}(x,\tau)\,dx\right]. \end{aligned}$$

Wegen (7) gilt dann:

$$(9)\qquad \tilde{\underline{F}}_{L,R} = \underline{F}_{L,R} + O(\,\|\,\underline{u}_L - \underline{u}_R\,\|\,).$$

Sei nun $\lambda_k < 0 < \lambda_{k+1}$ für ein $k \in \{0,\ldots,m\}$ und

$$\lambda := \frac{\tau}{h} < \frac{1}{2}\ \max\{\ |\sigma_1^L|, |\sigma_m^R|, |\lambda_1|, |\lambda_m|\ \}.$$

Dann gilt mit (6):

$$(10)\qquad \begin{aligned} \int_{-\frac{h}{2}}^{0} \underline{w}(x,t)\,dx &\overset{(6)}{=} \int_{-\frac{h}{2}}^{\lambda_1\,\tau} \underline{u}_L\,dx + \sum_{i=1}^{k-1}\int_{\lambda_i\,\tau}^{\lambda_{i+1}\,\tau} \underline{w}^i\,dx + \int_{\lambda_k\,\tau}^{0} \underline{u}_k\,dx \\ &= \underline{u}_L\,[\lambda_1\,\tau + \tfrac{h}{2}] + \tau\sum_{i=1}^{k-1} \underline{w}^i\,[\lambda_{i+1} - \lambda_i] - \tau\,\lambda_k\,\underline{w}^k. \end{aligned}$$

Für die Summe aus obiger Gleichung gilt nun

$$(11)\qquad \begin{aligned}\sum_{i=1}^{k-1} \underline{w}^i\,[\lambda_{i+1}-\lambda_i] &\overset{(6)}{=} \sum_{i=1}^{k-1}\left[\underline{u}_L + \sum_{j=1}^{i} w_j\,\underline{r}_j\right][\lambda_{i+1}-\lambda_i]\\ &= \underline{u}_L\,[\lambda_k-\lambda_1] + \sum_{i=1}^{k-1}\sum_{j=1}^{i} w_j\,\underline{r}_j\,[\lambda_{i+1}-\lambda_i].\end{aligned}$$

Durch Indexverschiebung erhält man für die Doppelsumme:

$$(12)\qquad \begin{aligned}\sum_{i=1}^{k-1}\sum_{j=1}^{i} w_j\,\underline{r}_j\,[\lambda_{i+1}-\lambda_i] &= \sum_{i=2}^{k}\sum_{j=1}^{i-1} w_j\,\underline{r}_j\,\lambda_i - \sum_{i=1}^{k-1}\sum_{j=1}^{i} w_j\,\underline{r}_j\,\lambda_i\\ &= \sum_{j=1}^{k-1} w_j\,\underline{r}_j\,\lambda_k - w_1\,\underline{r}_1\,\lambda_1 - \sum_{i=2}^{k-2} w_i\,\underline{r}_i\,\lambda_i.\end{aligned}$$

Durch Einsetzen von (11) und (12) in (10) ergibt sich nun:

$$(13)\qquad \begin{aligned}\int_{-\frac{h}{2}}^{0} \underline{w}(x,\tau)\,dx &= \frac{h}{2}\,\underline{u}_L + \tau\,\lambda_k\left[\underline{u}_L - \underline{w}^k + \sum_{j=1}^{k-1} w_j\,\underline{r}_j\right] - \tau\sum_{i=1}^{k-1} w_i\,\underline{r}_i\,\lambda_i\\ &\overset{(6)}{=} \frac{h}{2}\,\underline{u}_L - \tau\sum_{i=1}^{k} w_i\,\underline{r}_i\,\lambda_i.\end{aligned}$$

Für das zweite Integral aus (8) geht man analog vor:

$$(10')\qquad \int_{0}^{\frac{h}{2}} \underline{w}(x,\tau)\,dx \overset{(6)}{=} \lambda_{k+1}\,\tau\,\underline{w}^k + \sum_{i=k+1}^{m-1} \underline{w}^i\,\tau\,(\lambda_{i+1}-\lambda_i) + \underline{u}_R\,[\tfrac{h}{2}-\lambda_m\,\tau].$$

Die Summe läßt sich nun wieder mit (6) umschreiben nach

$$(11')\qquad \sum_{i=k+1}^{m-1} \underline{w}^i\,\tau\,(\lambda_{i+1}-\lambda_i) = \underline{u}_R\,\lambda_m - \lambda_{k+1}\left[\underline{u}_R - \sum_{j=k+2}^{m} w_j\,\underline{r}_j\right] - \sum_{i=k+2}^{m} w_i\,\underline{r}_i\,\lambda_i.$$

Durch Einsetzen von (11′) in (10′) erhält man mit (6)

$$(13')\qquad \int_{0}^{\frac{h}{2}} \underline{w}(x,\tau)\,dx = \frac{h}{2}\,\underline{u}_R - \tau\sum_{i=k+1}^{m} w_i\,\underline{r}_i\,\lambda_i.$$

Aus (13) und (13′) folgt nun für (8):

$$\underline{\tilde{F}}_{L,R} = \frac{1}{2}\left[\underline{F}(\underline{u}_L) + \underline{F}(\underline{u}_R) + \sum_{i=1}^{k} \lambda_i\,w_i\,\underline{r}_i - \sum_{i=k+1}^{m} \lambda_i\,w_i\,\underline{r}_i\right].$$

Da nach Voraussetzung (vgl. auch (3.5)) $\lambda_1 < \cdots < \lambda_k < 0 < \lambda_{k+1} < \cdots < \lambda_m$, gilt insbesondere:

$$\lambda_i = -|\lambda_i| \text{ für } i = 1,\ldots,k \text{ und} \lambda_i = |\lambda_i| \text{ für } i = k+1,\ldots,m.$$

Daraus folgt unmittelbar:

$$\underline{\tilde{F}}_{L,R} = \underline{F}^{Roe}(\underline{u}_L,\underline{u}_R)$$

und wegen (9) die Behauptung.

■

Für den skalaren Fall, d.h. $m = 1$, erhält man insbesondere:

$$F^{Roe}(u_L, u_R) \;=\; \frac{1}{2}\;\{\; F(u_L) + F(u_R) - |\lambda(u_L, u_R)|\,(u_L - u_R)\;\}\;\text{ mit}$$

$$\lambda(u_L, u_R) = \left\{ \begin{array}{ll} \dfrac{F(u_L) - F(u_R)}{u_L - u_R} & , u_L \neq u_R \\ & \\ 0 & , u_L = u_R \end{array} \right\}.$$

3.3 Herleitung von F^{Roe} für die Eulergleichung

Im folgenden soll nun der Roe–Fluß für ein spezielles System von Erhaltungssätzen hergeleitet werden, der s.g. Eulergleichung. Darunter versteht man das für die Gasdynamik sehr wichtige System von strikt hyperbolischen von Erhaltungssätzen, gegeben durch:

$$\text{(3.6.a)} \qquad \underline{F}(\underline{u}) \;:=\; \left(\,\rho\, v, P + \rho\, v^2, v(P+E)\,\right)^T;$$

$$\underline{u} \;:=\; (\rho, m, E)^T \;\text{ mit}$$

$$\rho = \text{Dichte}\;,\, m = \text{Impuls}\;,\, E = \text{Energie}\;,\, P = \text{Druck}$$

sowie den s.g. Zustansgleichungen

$$\text{(3.6.b)} \qquad E = \frac{1}{\gamma - 1}\;P + \frac{1}{2}\;\rho\, v^2,\;\; m = \rho\, v,\;\; c^2 = \gamma\;\frac{P}{\rho},\;\; H = \frac{1}{\rho}(P+E)\;\text{ mit}$$

$$\gamma = \text{Gaskonstante}\;,\;\; H = \text{Enthalpie}\;,\, c = \text{Schallgeschwindigkeit},$$

$$\gamma > 1,\;\; \rho,\;\; P > 0.$$

In diesem System soll nun zunächst die Roe–Linearisierung $A_{L,R}$ bestimmt werden. Die in La. 3.1 angegebene eignet sich dazu nicht, da sie zu aufwendig ist. Man ist also an einer günstigeren Linearisierung bestimmt. Dazu wurde in [HaLL] folgender Satz gezeigt:

Satz 3.2:

Vor.: (3.1) besitze eine Entropiefunktion gemäß Def. 2.6;
Beh.: Es gibt Matrizen $B, P \in \mathbb{R}^{m\times m}$ mit

a) B ist symmetrisch und P positiv definit;
b) $A := B\;P$ erfüllt (3.3).

Beweis:

Siehe dazu [HaLL], S. 42\ 43.

■

Hier bleibt das Problem, B und P geeignet zu konstruieren. Roe geht in [Ro1] einen ähnlichen Weg, der im folgenden für die Eulergleichung beschrieben werden soll.
Dazu sei im folgenden $\underline{u}_L, \underline{u}_R \in D$, $\underline{u}_L \not\equiv \underline{u}_R$ und $D := \{\, \underline{u} \in \mathbb{R}^3 \;\big|\; u_1 > 0, P(\underline{u}) > 0 \,\}$.
Sei ferner $\underline{V}$ eine spezielle Mittelung von $\underline{u}, \underline{v} \in \mathbb{R}^3$ mit

$$\underline{V}(\underline{u}, \underline{v}) \;=\; \underline{V}(\underline{v}, \underline{u}) \qquad \text{und} \qquad \underline{V}(\underline{u}, \underline{u}) \;=\; \underline{u}. \tag{3.7}$$

Es soll nun $A_{L,R}$ so bestimmt werden, daß gilt:

$$A_{L,R} \;=\; A(\underline{V}(\underline{u}_L, \underline{u}_R)).$$

Dazu führt Roe in seiner Arbeit [Ro1] einen s.g. Parameter–Vektor $\underline{w}$ ein. Dies ist eine injektive Abbildung $\underline{w} : D \to \mathbb{R}^3$. Dann bestimmt er $\underline{u} = \underline{u}(\underline{w})$ bzw. $\underline{F} = \underline{F}(\underline{w})$, berechnet $A_{L,R}(\underline{w})$ und erhält durch Rücktransformation $A_{L,R} = A(\underline{V}(\underline{u}_L, \underline{u}_R))$ mit $\underline{V}(\underline{u}_L, \underline{u}_R) := \frac{1}{2}\,[\,\underline{u}_L + \underline{u}_R\,]$. Als Parameter–Vektor wählt Roe

$$\underline{w}(\underline{u}) \;:=\; \sqrt{\rho}\,(1, v, H)^T \;=:\; (w_1, w_2, w_3)^T. \tag{3.8}$$

Setze $\underline{u} = (u_1, u_2, u_3)^T := (\rho, m, E)^T$. Dann ergibt sich, wie man durch Nachrechnen überprüft, für die Transformation:

$$\begin{aligned} &u_1(\underline{w}) = w_1^2,\;\; u_2(\underline{w}) = w_1\, w_2,\;\; u_3(\underline{w}) = \frac{1}{\gamma}\, w_1\, w_3 + \frac{\gamma-1}{2\,\gamma}\, w_2^2; \\ &F_1(\underline{w}) = w_1\, w_2,\;\; F_2(\underline{w}) = \frac{\gamma-1}{\gamma}\, w_1\, w_3 + \frac{\gamma+1}{2\,\gamma}\, w_2^2,\quad F_3(\underline{w}) = w_2\, w_3. \end{aligned} \tag{3.9}$$

Bestimme nun $B \in \mathbb{R}^{3\times 3}$, so daß gilt:

$$\underline{u}_L - \underline{u}_R \;=\; B\,(\underline{w}_L - \underline{w}_R), \qquad \forall\, \underline{u}_L, \underline{u}_R \in D. \tag{3.10}$$

Lemma 3.3:

Vor.: $\underline{u}_L, \underline{u}_R \in D$, $\underline{u}_L \not\equiv \underline{u}_R$, $\overline{w} := \frac{1}{2}\,(\,\underline{w}(\underline{u}_L) + \underline{w}(\underline{u}_R)\,) =: \frac{1}{2}\,(\,\overline{w}_L + \overline{w}_R\,)$;

Beh.: $B := \begin{pmatrix} 2\,\overline{w}_1 & 0 & 0 \\ \overline{w}_2 & \overline{w}_1 & 0 \\ \frac{1}{\gamma}\,\overline{w}_3 & \frac{\gamma-1}{\gamma}\,\overline{w}_2 & \frac{1}{\gamma}\,\overline{w}_1 \end{pmatrix}$ erfüllt (3.10).

Beweis:

B bestimmt sich spaltenweise direkt aus (3.10), in dem man $u_{L,i} - u_{R,i}$ mittels (3.9) durch die Komponenten von $\underline{w}$ ausdrückt. Beachte dabei, daß gilt:

$$w_{L,1}\, w_{L,3} - w_{R,1}\, w_{R,3} \;=\; \overline{w}_1\,(w_{L,3} - w_{R,3}) + \overline{w}_3\,(w_{L,1} - w_{R,1}).$$

■

Als nächstes bestimmt man nun $C \in \mathbb{R}^{3\times 3}$, so daß gilt:

$$\underline{F}(\underline{w}_L) - \underline{F}(\underline{w}_R) \ = \ C\ (\ \underline{w}_L - \underline{w}_R\ ). \tag{3.11}$$

Lemma 3.4:

Vor.: $\underline{u}_L, \underline{u}_R \in D\ \underline{u}_L \not\equiv \underline{u}_R,\ \underline{w}_{L,R} = \underline{w}(\underline{u}_{L,R})$;

Beh.: $C := \begin{pmatrix} \overline{w}_2 & \overline{w}_1 & 0 \\ \frac{\gamma-1}{\gamma}\,\overline{w}_3 & \frac{\gamma+1}{\gamma}\,\overline{w}_2 & \frac{\gamma-1}{\gamma}\,\overline{w}_1 \\ 0 & \overline{w}_3 & \overline{w}_2 \end{pmatrix}$ erfüllt (3.11).

Beweis:

C läßt sich ähnlich wie im Beweis zu La. 3.3 berechnen, in dem man die Matrix mittels (3.9) spaltenweise bestimmt.

■

Damit läßt sich die gesuchte Roe–Linearisierung folgendermaßen definieren:

$$A_{L,R}(\overline{w}) := C\ B^{-1}. \tag{3.12}$$

Daß dies eine geeignete Wahl ist, zeigt:

Lemma 3.5:

Vor.: $\underline{u}_L, \underline{u}_R \in D, \underline{u}_L \not\equiv \underline{u}_R,\ \underline{w}_{L,R} = \underline{w}(\underline{u}_{L,R}),\ A_{L,R}(\overline{w})$ gemäß (3.12),
$\underline{V}(\underline{w}_L, \underline{w}_R) := \overline{w} = \frac{1}{2}\,(\underline{w}_L - \underline{w}_R\ )$;

Beh.: $A_{L,R}$ erfüllt die Bedingung (3.3), und es gilt: $A_{L,R}(\underline{u}_L, \underline{u}_R) = A_{L,R}(\underline{V}(\underline{w}_L, \underline{w}_R))$.

Beweis:

$A_{L,R}(\overline{w})$ ist wohldefiniert, da B aus La. 3.4 invertierbar ist, denn:

$$\det B \ = \ \frac{2}{\gamma}\,\overline{w}_1^3 \overset{(3.8)}{=} \frac{1}{4\,\gamma}\,(\sqrt{\rho_L} + \sqrt{\rho_R})^3 \ > \ 0.$$

Zu überprüfen ist nun die Bedingung (3.3).:

a): Ist klar, da $A_{L,R} = const.$

b): Durch Bestimmung von B^{-1} erhält man:

$$A_{L,R}(\overline{w}) \ = \ \begin{pmatrix} 0 & 1 & 0 \\ \frac{\gamma-3}{2}\left(\frac{\overline{w}_2}{\overline{w}_1}\right)^2 & (3-\gamma)\,\frac{\overline{w}_2}{\overline{w}_1} & \gamma-1 \\ \frac{\gamma-1}{2}\left(\frac{\overline{w}_2}{\overline{w}_1}\right)^3 - \frac{\overline{w}_3\,\overline{w}_2}{\overline{w}_1^2} & \frac{\overline{w}_3}{\overline{w}_1} - (\gamma-1)\left(\frac{\overline{w}_2}{\overline{w}_1}\right)^2 & \gamma\,\frac{\overline{w}_2}{\overline{w}_1} \end{pmatrix}.$$

Sei nun $\underline{u}_L = \underline{u}_R = \underline{u} \in D$. Dann gilt natürlich auch $\underline{w}_L = \underline{w}_R = \underline{w}$ und somit $\overline{w}_L = \overline{w}_R = \underline{w}$. Daraus folgt:

$$A_{L,R}(\overline{w}_L, \overline{w}_R) \ = \ A_{L,R}(\underline{w}, \underline{w}).$$

Durch explizite Berechnung von $\underline{F}'(\underline{u})$ erhält man:

$$A_{L,R}(\underline{u},\underline{u}) \;=\; \underline{F}'(\underline{u}).$$

c): Folgt mit (3.10) und (3.11), denn:

$$\begin{aligned}\underline{F}(\underline{u}_L)-\underline{F}(\underline{u}_R) \;&=\; \underline{F}(\underline{w}(\underline{u}_L))-\underline{F}(\underline{w}(\underline{u}_R)) \;\overset{(3.11)}{=}\; C\,[\,\underline{w}_L-\underline{w}_R\,]\\ &\overset{(3.10)}{=}\; C\,B^{-1}\,[\,\underline{u}_L-\underline{u}_R\,] \;\overset{(3.12)}{=}\; A_{L,R}\,[\,\underline{u}_L-\underline{u}_R\,].\end{aligned}$$

d): Zur Bestimmung der Eigenwerte von $A_{L,R}(\overline{w})$ gehe wie folgt vor:

$$\begin{aligned} A_{L,R}\,(\,\underline{u}_L-\underline{u}_R\,) \;&=\; \lambda\,(\,\underline{u}_L-\underline{u}_R\,)\\ \overset{(3.10)}{\Longleftrightarrow}\quad (A_{L,R}-\lambda\,I)\,B\,(\,\underline{w}_L-\underline{w}_R\,) \;&=\; 0\\ \overset{(3.12)}{\Longleftrightarrow}\quad (C-\lambda\,B)\,(\,\underline{w}_L-\underline{w}_R\,) \;&=\; 0\\ \overset{\underline{w}_L\not\equiv\underline{w}_R}{\Longleftrightarrow}\quad \det(C-\lambda\,B) \;&=\; 0.\end{aligned}$$

Setze

$$\overline{a}^2 := (\gamma-1)\left(\overline{H}-\frac{1}{2}\,\overline{v}^2\right),\; \overline{H} := \frac{\overline{w}_3}{\overline{w}_1},\; \overline{v} := \frac{\overline{w}_2}{\overline{w}_1}.$$

Dann erhält man durch langwieriges Ausrechnen:

$$\det(C-\lambda\,B) \;=\; 0 \quad\Longleftrightarrow\quad (\lambda-\overline{v}\,)\,(\lambda+\overline{v}+\overline{a}\,)\,(\lambda+\overline{v}-\overline{a}\,) \;=\; 0.$$

Somit bestimmen sich die Eigenwerte von $A_{L,R}(\overline{w})$ zu den Eigenvektoren $\underline{u}_L-\underline{u}_R$ als:

(3.13) $$\lambda_1(\overline{w}) \;=\; \overline{v}-\overline{a},\;\; \lambda_2(\overline{w}) \;=\; \overline{v},\;\; \lambda_3(\overline{w}) \;=\; \overline{v}+\overline{a}.$$

Die zugehörigen Rechtseigenvektoren sind:

(3.14) $$\underline{r}_1(\overline{w}) \;=\; \begin{pmatrix}1\\ \overline{v}-\overline{a}\\ \overline{H}-\overline{v}\,\overline{a}\end{pmatrix},\; \underline{r}_2(\overline{w}) \;=\; \begin{pmatrix}1\\ \overline{v}\\ \frac{1}{2}\,\overline{v}^2\end{pmatrix},\; \underline{r}_3(\overline{w}) \;=\; \begin{pmatrix}1\\ \overline{v}+\overline{a}\\ \overline{H}+\overline{v}\,\overline{a}\end{pmatrix}.$$

Diese sind linear unabhängig, da die Eigenwerte paarweise verschieden sind. Dann gibt es $\underline{a}\in\mathbb{R}^3\backslash\{\underline{0}\}$ mit

$$\underline{u}_L-\underline{u}_R \;=\; \sum_{i=1}^{3} a_i\,\underline{r}_i(\overline{w}).$$

Als Lösung dieses Gleichungssystems erhält man

(3.15) $$\begin{aligned} a_1 \;&=\; \rho_L-\rho_R-a_2-a_3;\\ a_2 \;&=\; (\gamma-1)\,\frac{E_R-E_L+(\rho_L-\rho_R)\,(\overline{H}-\overline{v}^2)+(m_L-m_R)\,\overline{v}}{\overline{a}^2};\\ a_3 \;&=\; \frac{1}{2}\left[\,\frac{m_L-m_R-(\rho_L-\rho_R)\,(\overline{v}-\overline{a}\,)}{\overline{a}}-a_2\,\right].\end{aligned}$$

Damit erfüllt $A_{L,R}(\overline{w})$ die Bedingung (3.3).
Es gilt nun ferner:

$$A_{L,R}(\overline{w}) \;\overset{\text{def.}}{=}\; A_{L,R}(\underline{V}(\underline{w}_L,\underline{w}_R)) \;\overset{(3.9)}{=}\; A_{L,R}(\underline{V}(\underline{w}(\underline{u}_L),\underline{w}(\underline{u}_R))) \;=\; A_{L,R}(\underline{u}_L,\underline{u}_R).$$

■

Nach dem nun die Roe–Linearisierung explizit bestimmt ist, kann man auch den Roe–Fluß angeben.

Satz 3.3:

Vor.: $\underline{u}_L, \underline{u}_R \in D, \underline{u}_L \not\equiv \underline{u}_R$, $\underline{w}_{L,R} = \underline{w}(\underline{u}_{L,R})$, $A_{L,R}(\overline{w})$ gemäß (3.12), $\overline{w} = \frac{1}{2}\,(\underline{w}_L + \underline{w}_R\,)$,
$\{\lambda_i(\overline{w})\}_{i=1}^3, \{\underline{r}_i(\overline{w})\}_{i=1}^3, \{a_i\}_{i=1}^3$ gemäß (3.13), (3.14), (3.15);

Beh.: $\underline{F}^{Roe}(\,\underline{u}_L, \underline{u}_R) \;=\; \dfrac{1}{2}\left(\underline{F}(\underline{u}_L) + \underline{F}(\underline{u}_R) - \sum_{i=1}^{3} |\lambda_i(\overline{w})|\; a_i\; \underline{r}_i(\overline{w})\right).$

Beweis:

Gemäß (3.5) gilt:

$$\underline{F}^{Roe}(\,\underline{u}_L, \underline{u}_R) \;=\; \frac{1}{2}\left(\underline{F}(\underline{u}_L) + \underline{F}(\underline{u}_R) - \sum_{i=1}^{3} |\lambda_i(\overline{w})|\;\langle\, \underline{l}_i(\overline{w}), \underline{u}_L - \underline{u}_R\,\rangle\; \underline{r}_i(\overline{w})\right).$$

Da $\{\underline{r}_j(\overline{w})\}_{j=1}^3$ eine Basis des $\mathbb{R}^3$ ist, gibt es $\{\underline{l}_j(\overline{w})\}_{j=1}^3$ mit $\langle\, l_j(\overline{w}), r_i(\overline{w})\,\rangle = \delta_{j,i}$, und es gilt:

$$\underline{u}_L - \underline{u}_R \;=\; \sum_{i=1}^{3} \langle\, \underline{l}_i(\overline{w}), \underline{u}_L - \underline{u}_R\,\rangle\; \underline{r}_i(\overline{w}).$$

Wegen (3.15) gilt andererseits:

$$\underline{u}_L - \underline{u}_R \;=\; \sum_{i=1}^{3} a_i\; \underline{r}_i(\overline{w}).$$

Daraus folgt:

$$a_i \;=\; \langle\, \underline{l}_i(\overline{w}), \underline{u}_L - \underline{u}_R\,\rangle, \quad i = 1, 2, 3.$$

■

Abschließend sei noch vermerkt, daß bei der Verwendung des Roe–Fluß in Gegenwart von Verdünnungswellen unphysikalische Stöße auftreten können. Um dies zu vermeiden, haben Harten und Hyman in [HaH] eine Modifikation vorgeschlagen, die darin besteht, einen s.g. Viskositätsterm zu $\underline{F}^{Roe}(\underline{u}_L, \underline{u}_R)$ hinzu zu addieren.

Kapitel 4: Asymptotische Entwicklung des VRPs

4.1 Einführung

Die im folgenden hergeleitete Theorie beruht im wesentlichen auf der Arbeit von P. A. Raviart und P. LeFloch (s. [LRa]), die eine asymptotische Entwicklung für ein spezielles Anfangswertproblem zu einem strikt hyperbolischen System von Erhaltungssätzen hergeleitet haben. Dazu zunächst:

Definition 4.1:

Unter dem *verallgemeinerten Riemann-Problem* (VRP) versteht man folgendes Anfangswertproblem:

(4.1)
$$\underline{u}_t(x,t) \ + \ (\underline{F}(\underline{u}(x,t)))_x \ = \ 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+;$$

$$\underline{u}(x,0) \ = \ \begin{Bmatrix} \underline{u}_L(x), \ x < 0 \\ \underline{u}_R(x), \ x > 0 \end{Bmatrix};$$

mit $\underline{u} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}^m, \underline{u}_{L,R} : \mathbb{R} \to \mathbb{R}^m, \underline{F} \in C^1(\mathbb{R}^m), m \in \mathbb{N}$.
Für die Eigenwerte $\{\lambda_k(\underline{u})\}_{k=1}^m$ von $F'(\underline{u})$ und die dazu gehörigen Links– bzw. Rechtseigenvektoren $\{ \ \underline{l}_k(\underline{u}) \ \}_{k=1}^m, \{ \ \underline{r}_k(\underline{u}) \ \}_{k=1}^m$ gilt:

(4.2.a)
$$\lambda_1(\underline{u}) \ < \cdots < \ \lambda_m(\underline{u}),$$

(4.2.b)
$$\langle \ \underline{l}_i(\underline{u}), \underline{r}_j(\underline{u}) \ \rangle = \delta_{i,j}, \ 1 \leq i,j \leq m.$$

Wegen (4.2.b) gilt insbesondere:

(4.3)
$$\text{a) } \underline{u} \ = \ \sum_{j=1}^{m} \langle \ \underline{l}_j(\underline{v}), \underline{u}) \ \rangle \ \underline{r}_j(\underline{v}) \quad \forall \ \underline{u}, \underline{v} \in \mathbb{R}^m;$$

$$\text{b) } \forall \ \underline{u} \in \mathbb{R}^m, \ \exists_1 \ \underline{a} \in \mathbb{R}^m \ : \ \underline{u} \ = \ \sum_{j=1}^{m} a_j \ \underline{r}_j(\underline{v}), \ \underline{v} \in \mathbb{R}^m.$$

Diese Tatsache wird im folgenden eine wesentliche Rolle spielen.
Obige Definition ist insbesondere eine Verallgemeinerung von

Definition 4.2:

Sind $\underline{u}_L, \underline{u}_R$ aus (4.1) konstante Funktionen, so versteht man unter (4.1) das *Riemann–Problem* (RP).

Über die Struktur der Lösung des RPs läßt sich zeigen:

Satz 4.1:

Vor.: $\underline{u}_L < \underline{u}_R \in \mathbb{R}^m, \ \| \ \underline{u}_L - \underline{u}_R \ \|$ hinreichend klein;
Beh.: Das durch Def. 4.2 festgelegte RP hat eine eindeutig bestimmte Entropielösung, die aus mindestens $m+1$ konstanten Zuständen besteht, die durch Schocks, Kontaktunstetigkeiten oder Verdünnugswellen getrennt sind. Entlang den Strahlen $\frac{x}{t} = const, \ t > 0$, ist diese Lösung konstant.

Beweis:

Ein ausführlicher Beweis findet sich in [Smo], S. 335,ff.

■

Ähnliches gilt auch für die Lösung $\underline{u}$ des VRPs:

Satz 4.2:

Vor.: $\underline{u}_L, \underline{u}_R : \mathbb{R} \to \mathbb{R}^m$, $\| \underline{u}_L(0) - \underline{u}_R(0) \|$ hinreichend klein;

Beh.: Um den Ursprung $(0,0)$ gibt es eine Umgebung, so daß das durch Def. 4.1 festgelegte VRP dort eine eindeutig bestimmte Entropielösung besitzt, die sich aus $m+1$ Gebieten D zusammensetzt, in denen sie glatt ist. Diese werden durch glatte Kurven $x = \varphi(t)$ oder Verdünnungsbereiche R, die durch glatte Kurven $x = \underline{\varphi}(t), x = \overline{\varphi}(t)$ begrenzt sind, getrennt mit

$$D = \{ (x,t) \in \mathbb{R} \times \mathbb{R}_+ \mid \varphi_L(t) < x < \varphi_R(t), \varphi_L(0) = \varphi_R(0) = 0 \}$$
$$R = \left\{ (x,t) \in \mathbb{R} \times \mathbb{R}_+ \;\middle|\; \underline{\varphi}(t) < x < \overline{\varphi}(t), \underline{\varphi}(0) = \overline{\varphi}(0) = 0 \right\}.$$

Beweis:

Ein Beweis findet sich in einer Arbeit von Li Ta–Tsien und Yu Wen–ci [TW].

■

Die Kurven $x = \varphi(t)$ bzw. $x = \underline{\varphi}(t)$, $x = \overline{\varphi}(t)$ werden im folgenden auch als charakteristischen Kurven bezeichnet. Entlang diesen muß die Lösung des VRPs folgende Bedingungen erfüllen:

Hat $\underline{u}$ entlang der Kurve $x = \varphi(t)$ einen Schock bzw. Kontaktunstetigkeit, dann muß gelten:

$$s \, [\, \underline{u}(\circ,t) \,]\big|_{(\varphi(t))} = [\, \underline{F}(\underline{u}(\circ,t)) \,]\big|_{(\varphi(t))} \quad \text{mit}$$

(4.4.a)

$$[w(\circ)]\big|_{(x)} := w(x+0) - w(x-0), \; s = \tfrac{dx}{dt} = \varphi'(t).$$

Entlang der Ränder $x = \underline{\varphi}(t), x = \overline{\varphi}(t)$ einer Verdünnungswelle muß gelten:

$$[\, \underline{u}(\circ,t) \,]\big|_{(\underline{\varphi}(t))} = 0, \; [\, \underline{u}(\circ,t) \,]\big|_{(\overline{\varphi}(t))} = 0. \tag{4.4.b}$$

Es soll nun $\underline{u}$ asymptotisch von Ordnung $k \geq 1$ entwickelt werden, d.h. bestimme Funktionen $\underline{v}^i(\xi)$, $i = 0, \ldots, k$ mit

$$\underline{u}(x,t) = \sum_{i=0}^{k} \underline{v}^i(\xi)\, t^i + O\left(t^{k+1}\right), \; \xi := \frac{x}{t}, \; (x,t) \in \mathbb{R} \times \mathbb{R}_+. \tag{4.5}$$

Daß dies gerechtfertigt ist, sieht man an folgendem

Lemma 4.1:

Vor.: $\underline{u}$ glatt in D gemäß S. 4.2, $k \in \mathbb{N}$;

Beh.: Es gibt Funktionen $\underline{v}^i : \mathbb{R} \to \mathbb{R}^m, i = 0 \ldots, k$ mit:

i) $\underline{v}^i \in \mathbb{P}_i$, $i = 0, \ldots, k$;

ii) $\underline{u}(x,t) = \sum_{i=0}^{k} \underline{v}^i(\frac{x}{t})\, t^i + O\left(t^{k+1}\right), \qquad \forall\, (x,t) \in D.$

Beweis:
Sei $(x,t) \in D$. Setze $\xi := \frac{x}{t}$. Dann gilt mit Taylorentwicklung um den Ursprung $(0,0)$:

$$\underline{u}(x,t) \;=\; \sum_{i=0}^{k} \frac{1}{i!}\, d^i \underline{u}(0,0) + O\left(x^{k+1}\right) + O\left(t^{k+1}\right) \text{ mit}$$

$$d^i \underline{u}(0,0) \;:=\; \sum_{l=0}^{i} \binom{i}{l} \frac{\partial^i}{\partial^{i-l}x\, \partial^l t}\, \underline{u}(0,0)\, x^{i-l}\, t^l \text{ und}$$

$$\frac{\partial^i}{\partial^{i-l}x\, \partial^l t}\, \underline{u}(0,0) \;:=\; \lim_{\substack{x,t \to 0 \\ |\frac{x}{t}| > \xi_0 \\ t > 0}} \frac{\partial^i}{\partial^{i-l}x\, \partial^l t}\, \underline{u}(x,t),\ \xi_0 \text{ groß genug.}$$

Damit erhält man die Behauptung mit

$$\underline{v}^i(\xi) \;:=\; \frac{1}{i!} \sum_{l=0}^{i} \binom{i}{l} \frac{\partial^i}{\partial^{i-l}x\, \partial^l t}\, \underline{u}(0,0)\, \xi^{i-l},\ i = 0, \ldots, k.$$

Beachte dabei, daß gilt: $x^{k+1} = (\xi\ t)^{k+1} = O(t^{k+1})$.

■

Im folgenden sollen nun $x = \varphi(t)$, $x = \underline{\varphi}(t)$, $x = \overline{\varphi}(t)$ und $\underline{v}^i$ explizit konstruiert werden, so daß gilt:

— $\varphi(t) = \sum\limits_{j=0}^{k} \sigma^j\, t^{j+1} + O(t^{k+2})$ bzw.

$\underline{\varphi}(t) = \sum\limits_{j=0}^{k} \underline{\sigma}^j\, t^{j+1} + O(t^{k+2})$, $\overline{\varphi}(t) = \sum\limits_{j=0}^{k} \overline{\sigma}^j\, t^{j+1} + O(t^{k+2})$;

— $\underline{v}^i$ soll nur von $\underline{v}^j,\ \sigma^j,\ \underline{\sigma}^j,\ \overline{\sigma}^j,\ j = 0, \ldots, i-1$ abhängen;

— in den Bereichen D, in denen $\underline{u}$ glatt ist, soll gelten: $\underline{v}^i(\frac{x}{t}) \in \mathbb{P}_i,\ (x,t) \in D$.

4.2 Transformation des VRPs

Zur Bestimmung von $\underline{v}^i$ in Abhängigkeit von $\xi := \frac{x}{t}, (x,t) \in \mathbb{R} \times \mathbb{R}_+$ ist es sinnvoll, folgende Variablentransformation vorzunehmen:

(4.6)
$$\xi(x,t) \;:=\; \frac{x}{t},\ s(x,t) := t;$$
$$\underline{v}(\xi,s) \;:=\underline{u}(\xi s, s) \;\equiv\; \underline{u}(x,t),\ (x,t) \in \mathbb{R} \times \mathbb{R}_+.$$

Dann gilt für die Ableitungsoperatoren wegen der Kettenregel:

(4.7)
$$\text{a)}\quad \frac{\partial}{\partial x} \;=\; \frac{\partial \xi}{\partial x}\frac{\partial}{\partial \xi} + \frac{\partial s}{\partial x}\frac{\partial}{\partial s} \overset{(4.6)}{=} \frac{1}{s}\frac{\partial}{\partial \xi};$$

$$\text{b)}\quad \frac{\partial}{\partial t} \;=\; \frac{\partial \xi}{\partial t}\frac{\partial}{\partial \xi} + \frac{\partial s}{\partial t}\frac{\partial}{\partial s} \overset{(4.6)}{=} -\frac{\xi}{s}\frac{\partial}{\partial \xi} + \frac{\partial}{\partial s}.$$

Damit erhält man für die Lösung des VRPs mit (4.6), (4.7)

$$\text{(4.8)} \qquad \text{a)} \quad \tfrac{\partial}{\partial t}\, \underline{u}(x,t) \;=\; -\tfrac{\xi}{s}\, \underline{v}_{\xi}(\xi,s) + \underline{v}_s(\xi,s);$$

$$\text{b)} \quad \tfrac{\partial}{\partial x}\, \underline{F}(\underline{u}(x,t)) \;=\; \tfrac{1}{s}\, \tfrac{\partial}{\partial \xi}\, \underline{F}(\underline{v}(\xi,s)).$$

Unter obiger Variablentransformation geht das System (4.1) somit über in

$$\text{(4.9)} \qquad -\xi\; \underline{v}_{\xi}(\xi,s) + s\; \underline{v}_s(\xi,s) \;=\; -\tfrac{\partial}{\partial \xi}\, \underline{F}(\underline{v}(\xi,s)).$$

Daraus folgt durch Einsetzen von (4.5) in die linke Seite von (4.9) mit (4.6)

$$\text{(4.10)} \qquad \begin{aligned} &-\xi\; \underline{v}_{\xi}(\xi,s) + s\; \underline{v}_s(\xi,s) \;= \\ &-\xi\, \tfrac{d}{d\xi}\, \underline{v}^0(\xi) + \sum_{i=1}^{k} s^i\, \Big(i\, \underline{v}^i(\xi) - \xi\, \tfrac{d}{d\xi}\, \underline{v}^i(\xi) \Big) \;+\; O\left(s^{k+1}\right). \end{aligned}$$

Die analoge Entwicklung für die rechte Seite aus (4.9) läßt sich nicht so ohne weiteres hinschreiben. Dazu

Lemma 4.2

Vor.: D gemäß S. 4.2, $\xi, s : D \to \mathbb{R}$ gemäß (4.6), $\underline{u}$ Lösung des VRPs (4.1);

Beh.: a) $\underline{F}(\underline{v}(\xi,s)) \;=\; \underline{F}(\underline{v}^0(\xi)) \;+\; \sum_{i=1}^{k} s^i\, \left\{ \underline{F}'(\underline{v}^0(\xi))\, \underline{v}^i(\xi) + \underline{F}^i(\xi) \right\} \;+\; O\left(s^{k+1}\right)$

mit $\underline{F}^i(\xi) = \underline{F}^i(\xi, \underline{v}^0, \ldots, \underline{v}^{i-1})$;

b) $\underline{v}^i \in \mathbb{P}_i,\; i = 0,\ldots,k-1 \;\Longrightarrow\; \underline{F}^k \in \mathbb{P}_k$.

Beweis:

Zu a): Wegen La. 4.1 besitzt $\underline{u}$ eine Entwicklung der Form (4.5). Mit (4.6) gilt somit für $i = 0, \ldots, k$:

$$\text{(1)} \qquad \left(\tfrac{\partial}{\partial s}\right)^i\, \underline{v}(\xi,s) \;=\; \sum_{l=i}^{k} s^{l-i}\, \frac{l!}{(l-i)\,!}\, \underline{v}^l(\xi) \;+\; O\left(s^{k+1-i}\right);$$

$$\text{(2)} \qquad \left(\tfrac{\partial}{\partial s}\right)^i\, \underline{v}(\xi,s)\, \Big|_{s=0} \;=\; i!\; \underline{v}^i(\xi).$$

Ferner läßt sich durch vollständige Induktion über $i \in \mathbb{N}$ unter Anwendung der Kettenregel zeigen:

$$\text{(3)} \qquad \begin{aligned} \left(\tfrac{\partial}{\partial s}\right)^i\, \underline{F}(\underline{v}(\xi,s)) \;=&\underline{F}'(\underline{v}(\xi,s))\, \left(\tfrac{\partial}{\partial s}\right)^i\, \underline{v}(\xi,s) \,+ \\ & \underline{f}^i\left(\left(\tfrac{\partial}{\partial s}\right)^0 \underline{v}, \ldots, \left(\tfrac{\partial}{\partial s}\right)^{i-1} \underline{v} \right)\Big|_{(\xi,s)} \text{ mit} \end{aligned}$$

$$\underline{f}^1(\underline{v})\Big|_{(\xi,s)} \;\equiv\; 0;$$

$$\underline{f}^{i+1}\left(\left(\tfrac{\partial}{\partial s}\right)^0 \underline{v}, \ldots, \left(\tfrac{\partial}{\partial s}\right)^i \underline{v} \right)\Big|_{(\xi,s)} := \tfrac{\partial}{\partial s}\, \underline{F}'(\underline{v}(\xi,s)) \left(\tfrac{\partial}{\partial s}\right)^i \underline{v}(\xi,s) +$$

$$\tfrac{\partial}{\partial s}\, \underline{f}^i\left(\left(\tfrac{\partial}{\partial s}\right)^0 \underline{v}, \ldots, \left(\tfrac{\partial}{\partial s}\right)^{i-1} \underline{v} \right)\Big|_{(\xi,s)}.$$

Daraus erhält man mit (2)

(4) $$\left(\tfrac{\partial}{\partial s}\right)^i \underline{F}(\underline{v}(\xi,0)) = \underline{F}'(\underline{v}^0(\xi))\, i!\, \underline{v}^i(\xi) + \underline{f}^i\left(\underline{v}^0, \ldots, \underline{v}^{i-1} \right)\Big|_{(\xi,0)}.$$

Durch Taylorentwicklung um $s = 0$ ergibt sich ferner

(5) $$\underline{F}(\underline{v}(\xi,s)) \overset{(4.5)}{=} \sum_{i=0}^{k} \left(\tfrac{\partial}{\partial s}\right)^i \underline{F}(\underline{v}(\xi,s))\Big|_{s=0} \frac{s^i}{i!} + O\left(s^{k+1}\right).$$

Setze

(6) $$\underline{F}^i(\xi) := \frac{1}{i!}\, \underline{f}^i\left(\underline{v}^0(\xi), \ldots, \underline{v}^{i-1}(\xi) \right)\Big|_{(\xi,0)}.$$

Durch Einsetzen von (4) in (5) erhält man mit (6) die Behauptung a).
Zu b): Sei nun $\underline{v}^i \in \mathbb{P}_i,\ i = 0, \ldots, k-1$. Dann ist $\underline{v}^0$ insbesondere konstant und somit gilt komponentenweise

$$\tfrac{\partial}{\partial s}\, \underline{F}'(\underline{v}(\xi,s))\Big|_{s=0} \in \mathbb{P}_1.$$

Mit (3) folgt dann:

$$\underline{f}^i\left(\underline{v}^0, \ldots, \underline{v}^{i-1} \right)\Big|_{(\xi,0)} \in \mathbb{P}_i$$

und wegen (6) gilt somit die Behauptung b).

■

Mit obigem Lemma erhält man nun für die rechte Seite von (4.9):

(4.11) $$\tfrac{d}{d\xi}\, \underline{F}(\underline{v}(\xi,s)) = \tfrac{d}{d\xi}\, \underline{F}(\underline{v}^0(\xi)) + \sum_{i=1}^{k} s^i \left\{ \tfrac{d}{d\xi}\, \underline{F}'(\underline{v}^0(\xi)) + \underline{F}^i(\xi) \right\} + O\left(s^{k+1}\right).$$

Durch Einsetzen von (4.10) und (4.11) in (4.9) ergibt sich mit anschließendem Koeffizientenvergleich bzgl. Potenzen von s:

(4.12) a) $$-\,\xi\, \tfrac{d}{d\xi}\, \underline{v}^0(\xi) + \tfrac{d}{d\xi}\underline{F}(\underline{v}^0(\xi)) = 0;$$

b) $$i\, \underline{v}^i(\xi) - \xi\, \tfrac{d}{d\xi}\, \underline{v}^i(\xi) + \tfrac{d}{d\xi}\left(\underline{F}'(\underline{v}^0(\xi))\, \underline{v}^i(\xi) \right) = \underline{h}^i(\xi) \text{ mit}$$

$$\underline{h}^i(\xi) := -\tfrac{d}{d\xi}\, \underline{F}^i(\xi),\ i = 1, \ldots, k.$$

Wegen La. 4.2.b gilt insbesondere

(4.13) a) $\underline{v}^i \in \mathbb{P}_i,\ i = 0, \ldots, k-1 \quad \Rightarrow \quad \underline{h}^k \in \mathbb{P}_{k-1};$

b) $\underline{h}^1 \equiv 0.$

4.3 Charakterisierung des transformierten VRPs

Entlang den charakteristischen Kurven muß die Lösung $\underline{u}$ des VRPs die Gleichungen (4.4) erfüllen. Im folgenden sollen nun die entsprechenden Bedingungen für das transformierte (4.9) bzw. entwickelte (4.12) System hergeleitet werden. Da die charakteristischen Kurven $x = \varphi(t)$ glatt sind und $\varphi(0) = 0$, existiert die folgende asymptotische Entwicklung

$$\varphi(t) \; = \; \sum_{i=0}^{k} \sigma^i \, t^{i+1} \; + \; O\left(t^{k+2}\right). \tag{4.14}$$

Bestimme davon ausgehend die entsprechende Darstellung für $\underline{u}(x,t)$ und $\underline{F}(\underline{u}(x,t))$ entlang $x = \varphi(t)$. Dazu:

Lemma 4.3:

Vor.: $\underline{u}$ Lösung des VRPs (4.1), $x = \varphi(t)$, $(x,t) \in D$ gemäß S. 4.2;

Beh.: a) $\underline{u}(x,t) \; = \; \underline{v}^0(\sigma^0) + \sum_{i=1}^{k} t^i \left\{ \underline{v}^i(\sigma^0) + \sigma^i \, \frac{d}{d\xi} \, \underline{v}^0(\sigma^0) + \underline{z}^i \right\} \; + \; O\left(t^{k+1}\right)$

mit $\underline{z}^i = \underline{z}^i \left(\sigma^0, \ldots, \sigma^{i-1}, \underline{v}^0(\sigma^0), \ldots, \underline{v}^{i-1}(\sigma^0) \right);$

b) $\underline{F}(\underline{u}(x,t)) \; = \; \underline{F}(\underline{v}^0(\sigma^0)) +$

$$\sum_{i=1}^{k} t^i \left\{ \underline{F}'(\underline{v}^0(\sigma^0)) \left[\underline{v}^i(\sigma^0) + \sigma^i \, \frac{d}{d\xi} \, \underline{v}^0(\sigma^0) \right] + \underline{a}^i \right\} \; + \; O\left(t^{k+1}\right)$$

mit $\underline{a}^i = \underline{a}^i \left(\sigma^0, \ldots, \sigma^{i-1}, \underline{v}^0(\sigma^0), \ldots, \underline{v}^{i-1}(\sigma^0) \right).$

Beweis:

Zu a): Setze $\tilde{\varphi}(t) := \frac{1}{t} \, \varphi(t)$. Dann gilt wegen (4.14)

$$\tilde{\varphi}(t) \; = \; \sum_{l=0}^{k} \sigma^l \, t^l \; + \; O\left(t^{k+1}\right). \tag{1}$$

Mit (4.5) und (4.6) gilt nun

$$\underline{u}(x,t) \;\overset{(4.5)}{=}\; \underline{v}(\tilde{\varphi}(t), t) \;\overset{(4.6)}{=}\; \sum_{l=0}^{k} \underline{v}^l(\tilde{\varphi}(t)) \, t^l \; + \; O\left(t^{k+1}\right). \tag{2}$$

Aus (1) folgt nun unmittelbar für $0 \leq s \leq k$:

$$(3) \qquad \left(\tfrac{d}{dt}\right)^s \; \tilde{\varphi}(t) \;=\; \sum_{l=s}^{k} \frac{l!}{(l-s)!}\, t^{l-s}\, \sigma^l \;+\; O\left(t^{k+1-s}\right);$$

$$(4) \qquad \left(\tfrac{d}{dt}\right)^s \; \tilde{\varphi}(0) \;=\; s!\, \sigma^s.$$

Durch vollständige Induktion über $i \in \mathbb{N}$ läßt sich ferner zeigen:

$$(5) \qquad \begin{aligned} &\left(\tfrac{d}{dt}\right)^i \; \underline{v}^j(\tilde{\varphi}(t)) \;=\; \tfrac{d}{d\xi}\, \underline{v}^j(\tilde{\varphi}(t)) \; \left(\tfrac{d}{dt}\right)^i \; \tilde{\varphi}(t) \;+\; \underline{z}_j^i(t), \quad j = 0, \ldots, k \\ &\text{mit } \underline{z}_j^0 \;\equiv\; 0,\; \underline{z}_j^{i+1}(t) := \left(\frac{d}{d\xi}\right)^2 \; \underline{v}^j(\tilde{\varphi}(t))\, \tilde{\varphi}'(t) \left(\tfrac{d}{dt}\right)^i \; \tilde{\varphi}(t) \;+\; \tfrac{d}{dt}\, \underline{z}_j^i(t). \end{aligned}$$

Daraus folgt mit (4) sofort:

$$(6) \qquad \begin{aligned} &\left(\tfrac{d}{dt}\right)^i \; \underline{v}^j(\tilde{\varphi}(0)) \;=\; \tfrac{d}{d\xi}\, \underline{v}^j(\sigma^0)\; i!\, \sigma^i \;+\; \underline{z}_j^i(0), \quad j = 0, \ldots, k \\ &\text{mit } \underline{z}_j^i(0) \;=\; \underline{z}_j^i\left(\; \underline{v}^j(\sigma^0), \sigma^0, \ldots, \sigma^{i-1} \;\right) \end{aligned}$$

Durch Taylorentwicklung von $\underline{v}^i \circ \tilde{\varphi}(\circ)$ um $t = 0$ erhält man mit (2)

$$(7) \qquad \begin{aligned} \underline{u}(x,t) \;=\; &\sum_{i=0}^{k} \left(\tfrac{d}{dt}\right)^i \; \underline{v}^0(\tilde{\varphi}(t))\Big|_{t=0} \; \frac{t^i}{i!} \;+ \\ &\sum_{i=1}^{k} t^i \sum_{l=0}^{k-i} \left(\tfrac{d}{dt}\right)^l \; \underline{v}^i(\tilde{\varphi}(t))\Big|_{t=0} \; \frac{t^l}{l!} \;+\; O\left(t^{k+1}\right). \end{aligned}$$

Mit (2) und (6) folgt daraus

$$(8) \qquad \begin{aligned} \underline{u}(x,t) \;=\; &\underline{v}^0(\sigma^0) \;+\; \sum_{i=1}^{k} t^i \left\{ \tfrac{d}{d\xi}\, \underline{v}^0(\sigma^0)\, \sigma^i \;+\; \frac{1}{i!}\, \underline{z}_0^i(0) \;+\; \underline{v}^i(\sigma^0) \right\} + \\ &\sum_{i=1}^{k-1} t^i \sum_{l=1}^{k-i} t^l \left\{ \tfrac{d}{d\xi}\, \underline{v}^i(\sigma^0)\, \sigma^l \;+\; \frac{1}{l!}\, \underline{z}_i^l(0) \right\} \;+\; O\left(t^{k+1}\right). \end{aligned}$$

Durch Umsummieren der Doppelsumme aus (8) in Potenzen von t erhält man dann

$$(9) \qquad \begin{aligned} \underline{u}(x,t) \;=\; &\underline{v}^0(\sigma^0) \;+\; \sum_{i=1}^{k} t^i \left\{ \tfrac{d}{d\xi}\, \underline{v}^0(\sigma^0)\, \sigma^i \;+\; \frac{1}{i!}\, \underline{z}_0^i(0) \;+\; \underline{v}^i(\sigma^0) \right\} + \\ &\sum_{i=1}^{k-1} t^{i+1} \underline{z}^{i+1} \;+\; O\left(t^{k+1}\right). \end{aligned}$$

$$\text{mit } \underline{\tilde{z}}^{i+1} \ := \ \sum_{l=1}^{i} \left\{ \frac{1}{l!} \, \underline{z}_{i+1-l}^{l}(0) \ + \ \tfrac{d}{d\xi} \, \underline{v}^{i+1-l}(\sigma^0) \, \sigma^l \right\}$$

$$= \ \underline{\tilde{z}}^{i+1} \left(\sigma^0, \ldots, \sigma^i, \underline{v}^1(\sigma^0), \ldots, \underline{v}^i(\sigma^0) \right).$$

Fasse nun in (9) die t–Potenzen zusammen. Dann erhält man die Behauptung a) mit

$$\underline{z}^1 \ := \ \underline{z}_0^1(0) \ = \ 0, \ \underline{z}^i \ := \ \frac{1}{i!} \, \underline{z}_0^i(0) + \underline{\tilde{z}}^i, \ i = 2, \ldots, k.$$

Zu b): Setze hierzu $\underline{w}(t) := \underline{u}(\varphi(t), t)$. Dann gilt mit a)

$$\text{(10)} \qquad \underline{w}(t) \ = \ \underline{v}^0(\sigma^0) + \sum_{i=1}^{k} t^i \ \left\{ \underline{v}^i(\sigma^0) + \sigma^i \ \tfrac{d}{d\xi} \, \underline{v}^0(\sigma^0) + \underline{z}^i \right\} \ + \ O\left(t^{k+1}\right).$$

Daraus folgt unmittelbar

$$\text{(11)} \quad \left(\tfrac{d}{dt}\right)^s \ \underline{w}(t) \ = \ \sum_{i=s}^{k} \frac{i!}{(i-s)\,!} \ t^{i-s} \ \left\{ \underline{v}^i(\sigma^0) + \sigma^i \ \tfrac{d}{d\xi} \, \underline{v}^0(\sigma^0) + \underline{z}^i \right\} \ + \ O\left(t^{k+1-s}\right);$$

$$\text{(12)} \quad \begin{aligned} &\left(\tfrac{d}{dt}\right)^s \ \underline{w}(0) \ = \ s! \ \left\{ \underline{v}^s(\sigma^0) + \sigma^s \ \tfrac{d}{d\xi} \, \underline{v}^0(\sigma^0) + \underline{z}^s \right\}, \ 1 \le s \le k, \\ &\text{insbesondere: } \underline{w}(0) \ = \ \underline{v}^0(\sigma^0); \end{aligned}$$

Durch vollständige Induktion über $l \in \mathbb{N}$ läßt sich nun zeigen

$$\text{(13)} \qquad \left(\tfrac{d}{dt}\right)^l \ \underline{F}(\underline{w}(t)) \ = \ \underline{F}'(\underline{w}(t)) \ \left(\tfrac{d}{dt}\right)^l \ \underline{w}(t) \ + \ \underline{b}^l(t) \text{ mit}$$

$$\begin{aligned} \underline{b}^1(t) \ \equiv \ 0, \ \underline{b}^{l+1}(t) \ &= \ \tfrac{d}{dt} \ \underline{F}'(\underline{w}(t)) \ \left(\tfrac{d}{dt}\right)^l \ \underline{w}(t) \ + \ \tfrac{d}{dt} \, \underline{b}^l(t) \\ &= \ \underline{b}^{l+1} \left(\left(\tfrac{d}{dt}\right)^0 \ \underline{w}(t), \ldots, \left(\tfrac{d}{dt}\right)^l \ \underline{w}(t) \right). \end{aligned}$$

Aus (12) und (13) folgt nun

$$\text{(14)} \qquad \left(\tfrac{d}{dt}\right)^l \ \underline{F}(\underline{w}(t))\Big|_{t=0} \ = \ \underline{F}'(\underline{v}^0(\sigma^0)) \ l! \ \left(\underline{v}^l(\sigma^0) + \sigma^l \ \tfrac{d}{d\xi} \, \underline{v}^0(\sigma^0) \right) \ + \ \underline{c}^l$$

$$\begin{aligned} \text{mit } \underline{c}^l \ &:= \ \underline{b}^l(0) + \underline{F}'(\underline{v}^0(\sigma^0)) \ l! \ \underline{z}^l \\ &= \ \underline{c}^l \left(\sigma^0, \ldots, \sigma^{l-1}, \underline{v}^0(\sigma^0), \ldots, \underline{v}^{l-1}(\sigma^0) \right). \end{aligned}$$

Entwickle nun $\underline{F} \circ \underline{w}(\circ)$ um $t = 0$. Daraus ergibt sich

$$\text{(15)} \qquad \underline{F}(\underline{u}(x,t)) \ = \ \sum_{l=0}^{k} \left(\tfrac{d}{dt}\right)^l \ \underline{F}(\underline{w}(t))\Big|_{t=0} \ \frac{t^l}{l!} \ + \ O\left(t^{k+1}\right).$$

Durch Einsetzen von (14) in (15) erhält man die Behauptung b) mit

$$\underline{a}^l \ := \ \frac{1}{l!} \, \underline{c}^l, \ l = 1, \ldots, k.$$

$\blacksquare$

Mittels obigem Lemma läßt sich nun das Verhalten von $\underline{v}^k$ entlang den charakteristischen Kurven charakterisieren. Dabei muß man unterscheiden, ob die Lösung $\underline{u}$ des VRPs entlang diesen Kurven unstetig oder stetig ist (vgl. (4.4)), d.h. sie ist entweder eine Schockkurve bzw. Kontaktunstetigkeit oder der Rand einer Verdünnungswelle und erfüllt entweder (4.4.a) oder (4.4.b).

Fall 1: $\underline{u}$ stetig entlang $x = \varphi(t)$:

Setze in (4.4.b) die Darstellung für $\underline{u}$ aus La. 4.3.a ein. Durch einen Koeffizientenvergleich bzgl. Potenzen von t erhält man

(4.15) a) $$[\,\underline{v}^0\,]\Big|_{(\sigma^0)} = 0;$$

b) $$\left[\,\underline{v}^i + \sigma^i \tfrac{d}{d\xi}\,\underline{v}^0 + \underline{z}^i\,\right]\Big|_{(\sigma^0)} = 0, \quad i = 1, \ldots, k.$$

Fall 2: $\underline{u}$ unstetig entlang $x = \varphi(t)$:

Wegen (4.4.a) muß gelten: $\varphi'(t)\ [\,\underline{u}(\circ, t)\,]\Big|_{(\varphi(t))} = [\,\underline{F}(\underline{u}(\circ, t))\,]\Big|_{(\varphi(t))}$.

Setze nun in die rechte Seite die Darstellung aus La. 4.3.b für $\underline{F}(\underline{u}(\varphi(t), t))$ ein und auf der linken Seite die Darstellung für $\underline{u}(\varphi(t), t)$ aus La. 4.3.a sowie die Entwicklung von $\varphi(t)$ aus (4.14). Durch Koeffizientenvergleich bzgl. Potenzen von t erhält man dann

(4.16.a) $$\sigma^0\ [\,\underline{v}^0\,]\Big|_{(\sigma^0)} = [\,\underline{F}(\underline{v}^0)\,]\Big|_{(\sigma^0)};$$

(4.16.b) $$[\,\{\underline{F}'(\underline{v}^0) - \sigma^0\}\,\underline{v}^i\,]\Big|_{(\sigma^0)} + \sigma^i\left[\,\{\underline{F}'(\underline{v}^0) - \sigma^0\}\,\tfrac{d}{d\xi}\,\underline{v}^0 - (i+1)\,\underline{v}^0\,\right]\Big|_{(\sigma^0)}$$
$$= -[\,\underline{w}^i\,]\Big|_{(\sigma^0)}, \quad i = 1, \ldots, k$$

mit
$$\underline{w}^i(\sigma^0) := \underline{a}^i - \underline{b}^i;$$
$$\underline{b}^1 := \sigma^0\,\underline{z}^1,\ \underline{b}^l := \underline{c}^l + \sigma^0\,\underline{z}^l,\ l \geq 2;$$
$$\underline{c}^l := \sum_{i=1}^{l-1}\left\{\,\underline{v}^i(\sigma^0) + \sigma^i\,\tfrac{d}{d\xi}\,\underline{v}^0(\sigma^0) + \underline{z}^i\,\right\}\,(l-i+1)\,\sigma^{l-i};$$
$\underline{a}^i,\ \underline{z}^i$ gemäß La. 4.3.

Als nächstes soll nun untersucht werden, wie die Anfangsbedingungen aus (4.1) in die $\underline{v}^k$ eingehen. Dazu betrachte solche Punkte $(x, t) \in \mathbb{R} \times \mathbb{R}_+$, die nahe an der x–Achse liegen, d.h. $|\frac{x}{t}|$ hinreichend groß . Dazu:

Lemma 4.4:

Vor.: $\underline{u}_{L,R}$ gemäß (4.1) seien in einer Umgebung von $x = 0$ k–mal differenzierbar, $k \in \mathbb{N}_0$;

Beh.: a) $\left(\frac{d}{d\xi}\right)^k\ \underline{v}^k(\xi) = \left(\frac{d}{d\xi}\right)^k\ \underline{u}_L(0)$ für $\xi < \xi_- < 0$, ξ_- klein genug;

b) $\left(\frac{d}{d\xi}\right)^k\ \underline{v}^k(\xi) = \left(\frac{d}{d\xi}\right)^k\ \underline{u}_R(0)$ für $\xi > \xi_+ > 0$, ξ_+ groß genug.

Beweis:

Durch Taylorentwicklung von $\underline{v}^i(\xi),\ i = 0,\dots,k,$ um $\xi = 0$ erhält man

$$(1) \qquad \underline{v}^i(\xi) \ = \ \sum_{j=0}^{i} \left(\tfrac{d}{d\xi}\right)^j \ \underline{v}^i(0) \ \frac{1}{j!} \ \xi^j \ + \ O\left(\xi^i\right).$$

Setze diese Darstellung in (4.5) ein. Dann gilt für $\xi = \frac{x}{t} > \xi_+$:

$$(2) \qquad \underline{u}(x,t) \ = \ \sum_{i=0}^{k} \sum_{j=0}^{i} \left(\tfrac{d}{d\xi}\right)^j \ \underline{v}^i(0) \ \frac{1}{j!} \ x^j \ t^{i-j} \ + \ O\left(t^{k+1}\right).$$

Daraus folgt für $t \to 0+$:

$$(3) \qquad \lim_{\substack{\frac{x}{t} > \xi_+ \\ t \to 0+}} \underline{u}(x,t) \ = \ \sum_{i=0}^{k} \frac{1}{i!} \left(\tfrac{d}{d\xi}\right)^i \ \underline{v}^i(0) \ x^i.$$

Für $\underline{u}_R$ erhält man durch Taylorentwicklung um $x = 0$:

$$(4) \qquad \underline{u}_R(x) \ = \ \sum_{i=0}^{k} \left(\tfrac{d}{dx}\right)^i \ \underline{u}_R(0) \ \frac{1}{i!} \ x^i \ + \ O\left(x^{k+1}\right).$$

Aus (3) und (4) folgt die Behauptung a) durch Koeffizientenvergleich bzgl. Potenzen von t.
Analog läßt sich b) zeigen.

■

4.4 Konstruktion von $\underline{v}^k$

Im folgenden soll mittels der Charakterisierung $\underline{v}^k$ durch die Sprungbedingungen (4.15) und (4.16) und den Randbedingungen aus La. 4.4 konstuiert werden. Für $\underline{v}^0$ ist dies kein Problem, denn:

Lemma 4.5:

Vor.: $\| \underline{u}_L(0) - \underline{u}_R(0) \|$ hinreichend klein,
$\underline{V}(\circ, \underline{u}_L(0), \underline{u}_R(0))$ Lösung des RPs gemäß Def. 4.2;

Beh.: $\underline{v}^0(\frac{x}{t}) \ = \ \underline{V}(\circ, \underline{u}_L(0), \underline{u}_R(0)),\ (x,t) \in \mathbb{R} \times \mathbb{R}_+.$

Beweis:

Die Bedingungen (4.12.a), (4.15.a), (4.16.a) und La. 4.4 mit $k = 0$ charakterisieren $\underline{v}^0$. Wegen S. 4.1 erfüllt aber die Lösung $\underline{V}(\circ, \underline{u}_L(0), \underline{u}_R(0))$ des RPs ebenfalls diese Bedingungen. Daher müssen beide Funktionen identisch sein, da die Entropielösung des RPs eindeutig ist.

■

Da $\underline{v}^0$ nun bekannt ist, können die $\underline{v}^k,\ k \geq 1$, konstruiert werden. Diese sollen folgende Form haben:

$$(4.17) \qquad \underline{v}^k(\xi) \stackrel{!}{=} \sum_{j=1}^{k} \alpha_j^k(\xi)\, \underline{r}_j(\underline{v}^0(\xi)) \text{ mit}$$

$$\alpha_j^k : \mathbb{R} \to \mathbb{R},\ \alpha_j^k \in \mathbb{P}_k,\ \underline{r}_j \text{ gemäß (4.2).}$$

Dabei sollen die α_j^k nur von $\underline{v}^0, \ldots, \underline{v}^{k-1}$ sowie den charakteristischen Geschwindigkeiten $\sigma_i^0, \ldots, \sigma_i^{k-1}$ und $\underline{\sigma}_i^0, \ldots, \underline{\sigma}_i^{k-1}$ bzw. $\overline{\sigma}_i^0, \ldots, \overline{\sigma}_i^{k-1}$ (vgl. (4.14)) der charakteristischen Kurven $x = \varphi_i(t), x = \underline{\varphi}_i(t), x = \overline{\varphi}_i(t),\ 1 \leq i \leq m$ abhängen. Darin geht die Struktur der Lösung des RPs wesentlich ein. Sie soll deshalb zunächst angegeben werden ohne auf die näheren analytischen Aspekte einzugehen. Dazu siehe zum Beispiel in einer Arbeit von J. Smoller [Smo], S. 320,ff.

Die Grundstruktur der Lösung des RPs setzt sich aus s.g. i–Wellen, $i = 1, \ldots, m$, zusammen. Man unterscheidet dabei zwischen

(A) Verdünnungswellen;

(B) Schockwellen;

(C) Kontaktunstetigkeiten.

Der Typ der i–Welle bestimmt sich wie folgt:

$$(4.18) \qquad \text{a)} \quad \langle\, \text{grad}_{\underline{u}}\ \lambda_i(\underline{u}), \underline{r}_i(\underline{u})\, \rangle \stackrel{\text{o.B.d.A.}}{\equiv} 1,\ \underline{u} \in \mathbb{R}^m \quad \Longleftrightarrow$$
$$\text{die } i\text{–Welle ist vom Typ A oder B;}$$

$$\text{b)} \quad \langle\, \text{grad}_{\underline{u}}\ \lambda_i(\underline{u}), \underline{r}_i(\underline{u})\, \rangle \equiv 0,\ \underline{u} \in \mathbb{R}^m \quad \Longleftrightarrow$$
$$\text{die } i\text{–Welle ist vom Typ C;}$$

Hierbei sind die $\{\lambda_i(\underline{u})\}, \{\underline{r}_i(\underline{u})\}$ gemäß Def. 4.2 festgelegt. Wegen S. 4.1 und La. 4.5 gibt es nun $m+1$ Zustände $\underline{v}_i^0,\ 0 \leq i \leq m$ mit

$$(4.19) \qquad \underline{v}^0(\xi) = \left\{ \begin{array}{ll} \underline{v}_0^0 = \underline{u}_L(0) & , \xi < \underline{\sigma}_1^0 \\ \underline{v}_i^0 & , \overline{\sigma}_i^0 < \xi < \underline{\sigma}_{i+1}^0,\ 1 \leq i \leq m-1 \\ \underline{v}_m^0 = \underline{u}_R(0) & , \overline{\sigma}_m^0 < \xi \end{array} \right\}.$$

Dann gilt für die Sprungbedingungen und den Verdünnungsbereich:

Fall 1: i–Welle ist vom Typ A:

$$(4.20) \qquad \text{a)} \quad \underline{\sigma}_i^0 \stackrel{!}{=} \lambda_{i-1}(\underline{v}_i^0 \stackrel{(4.2.a)}{<} \lambda_i(\underline{v}_i^0) \stackrel{!}{=} \overline{\sigma}_i^0;$$

$$\text{b)} \quad \tfrac{d}{d\xi}\, \underline{v}^0(\xi) \stackrel{!}{=} \underline{r}_i(\underline{v}^0(\xi)),\ \underline{\sigma}_i^0 < \xi < \overline{\sigma}_i^0;$$

$$\lambda_i(\underline{v}^0(\xi)) \stackrel{!}{=} \underline{\xi}.$$

Fall 2: i–Welle ist vom Typ B oder C:

(4.21) a) $\sigma_i^0 := \underline{\sigma}_i^0 \overset{!}{=} \overline{\sigma}_i^0;$

b) Typ B: $\lambda_{i-1}(\underline{v}_{i-1}^0) < \sigma_i^0 < \lambda_i(\underline{v}_{i-1}^0);$

$$\lambda_i(\underline{v}_i^0) < \sigma_i^0 < \lambda_{i+1}(\underline{v}_i^0);$$

c) Typ C: $\lambda_i(\underline{v}_i^0) = \sigma_i^0 = \lambda_{i-1}(\underline{v}_{i-1}^0).$

An Hand dieser Struktur werden nun die $\underline{v}^k, \; k \geq 1,$ konstruiert, in dem man die Funktionen α_j^k stückweise bestimmt für:

a) $\xi \in D_i := (\, \overline{\sigma}_i^0, \underline{\sigma}_{i+1}^0 \,), \; i = 0, \ldots, m, \; \overline{\sigma}_0^0 := -\infty, \; \underline{\sigma}_{m+1}^0 := \infty;$
b) $\xi \in R_i := (\, \underline{\sigma}_i^0, \overline{\sigma}_i^0 \,)$ falls i–Welle vom Typ A.

Mit (4.19) gilt dann insbesondere:

(4.22) a) $\underline{v}^0(\xi) = \underline{v}_i^0 = const,$ bzw. $\frac{d}{d\xi}\, \underline{v}^0(\xi) = 0, \; \overline{\sigma}_i^0 < \xi < \underline{\sigma}_{i+1}^0;$

b) $\frac{d}{d\xi}\, \underline{v}^0(\xi) \overset{(4.20.b)}{\not\equiv} 0, \; \underline{\sigma}_i^0 < \xi < \overline{\sigma}_i^0.$

Für die Bereiche in denen $\underline{v}^0$ konstant ist, erhält man nun

Lemma 4.6:

Vor.: $\xi \in D_i, \; 0 \leq i \leq m, \; \underline{v}^j\big|_{D_i} \in \mathbb{P}_j, \; 0 \leq j \leq k-1, k \in \mathbb{N};$

Beh.: Die Lösung von (4.12) ist auf D_i gegeben durch:

$$\begin{aligned} \underline{v}^k(\xi) \;&=\; \sum_{j=1}^{m} \alpha_j^k(\xi)\, \underline{r}_j(\underline{v}_i^0) \text{ mit} \\ \alpha_j^k(\xi) \;&=\; \big(\xi - \lambda_j(\underline{v}_i^0)\big)^k \; a_{j,i}^k \;+\; p_{j,i}^k(\xi), \; p_{j,i}^k \in \mathbb{P}_{k-1}, \; a_{j,i}^k \in \mathbb{R} \text{ beliebig.} \end{aligned} \tag{4.23}$$

Beweis:

Für $\xi \in D_i$ gilt mit (4.22.a) und (4.12)

(1) $$\big(\, \underline{F}'(\underline{v}_i^0) - \xi \,\big) \; \frac{d}{d\xi}\, \underline{v}^k(\xi) \;+\; k\, \underline{v}^k(\xi) \;=\; \underline{h}^k(\xi).$$

Da $\underline{v}^j \in \mathbb{P}_j$ für $j = 0, \ldots, k-1$, gilt mit (4.13):

$$\underline{h}^k \in \mathbb{P}_{k-1},$$

und somit:

(2) $$\exists_1 \; (\, \underline{b}_0, \ldots, \underline{b}_{k-1} \,) \in (\mathbb{R}^m)^k : \; \underline{h}^k(\xi) = \sum_{l=0}^{k-1} \underline{b}_l\, \xi^l.$$

Setze nun (2) und (4.17) in (1) ein. Dann erhält man mit (4.3):

$$(3) \qquad (\ \lambda_j(\underline{v}_i^0) - \xi\ )\ \tfrac{d}{d\xi}\ \alpha_j^k(\xi)\ +\ k\ \alpha_j^k(\xi)\ =\ \sum_{l=0}^{k-1} \langle\ \underline{l}_j(\underline{v}_i^0), \underline{b}_l\ \rangle\ \xi^l,\ j = 1, \ldots, m.$$

Dies ist ein inhomogenes System von Differentialgleichungen 1. Ordnung. Die Lösungen bestimmen sich durch

$$\alpha_j^k(\xi)\ =\ \alpha_{part}^{k,j}(\xi)\ +\ \alpha_{hom}^{k,j}(\xi),$$

wobei $\alpha_{part}^{k,j}$ eine spezielle partielle Lösung von (3) ist und $\alpha_{hom}^{k,j}$ alle Lösungen des zu (3) gehörigen homogenen Systems durchläuft (vgl. [Wa]). Dazu:

a) Partielle Lösung:

Setze

$$p_{j,i}^k(\xi) := \sum_{l=0}^{k-1} a_l^{j,i}\ \xi^l \text{ mit } \underline{a}^{j,i} := (a_0^{j,i}, \ldots, a_{k-1}^{j,i}) \in \mathbb{R}^k$$

und ersetze in (3) α_j^k durch $p_{j,i}^k$. Dann erhält man durch Koeffizientenvergleich bzgl. Potenzen von ξ:

$$(4) \qquad \begin{aligned} a_{k-1}^{j,i} &= \langle\ \underline{l}_j, \underline{b}_{k-1}\ \rangle; \\ a_l^{j,i} &= \frac{1}{k-l}\ \Big[\ \langle\ \underline{l}_j, \underline{b}_l\ \rangle - \lambda_j(\underline{v}_i^0)\ a_{l+1}^{j,i}\ (l+1)\ \Big],\ l = k-2, \ldots, 1. \end{aligned}$$

Damit ist eine partielle Lösung bestimmt mit $\alpha_{part}^{k,j} := p_{j,i}^k$.

b) Raum der Lösungen des homogenen Systems:

Mittels Seperation der Veränderlichen (vgl. [Wa]) erhält man:

$$(5) \qquad \alpha_{hom}^{k,j}(\xi)\ =\ (\ \xi - \lambda_j(\underline{v}_i^0)\ )^k\ a_{j,i}^k \text{ mit } a_{j,i}^k \in \mathbb{R} \text{ beliebig.}$$

Daraus folgt aufgrund der Vorüberlegungen die Behauptung.

■

Für den Verdünnungsbereich von $\underline{v}^0$ gilt desweiteren:

Lemma 4.7:

Vor.: i–Welle vom Typ A, $i \in \{1, \ldots, m\}$, $m > 1$, $\xi \in R_i$,
$\alpha_j^k(\xi)$, $j = 1, \ldots, m$, gemäß (4.17),
$\underline{\alpha}(\xi) := (\ \alpha_1^k(\xi), \ldots, \alpha_{i-1}^k(\xi), \alpha_{i+1}^k(\xi), \ldots, \alpha_m^k(\xi)\ );$

Beh.: a) Es gibt eindeutig bestimmte $A(\xi) \in \mathbb{R}^{m-1} \times \mathbb{R}^{m-1}$, $\underline{f}(\xi) \in \mathbb{R}^{m-1}$, so daß (4.12.b) äquivalent ist zu:

$i)$ $\frac{d}{d\xi}\ \underline{\alpha}(\xi)\ +\ A(\xi)\ \underline{\alpha}(\xi)\ =\ \underline{f}(\xi),$

$ii)$ $\alpha_i^k(\xi)$ läßt sich als Linearkombination der Komponenten von $\underline{\alpha}(\xi)$ darstellen;

b) $\underline{\alpha}(\xi)\ =\ H(\xi)\ \alpha(\underline{\sigma}_i^0 + 0)\ +\ \int_{\underline{\sigma}_i^0}^{\xi} H(\xi)\ H^{-1}(\tau)\ \underline{f}(\tau)\ d\tau$

ist die eindeutig bestimmte Lösung von i) wobei die $\alpha(\underline{\sigma}_i^0 + 0)$ beliebig gegebene Anfangswerte sind und H das Hauptsystem der homogenen Differentialgleichung von i) mit $H(\underline{\sigma}_i^0) = Id$.

Beweis:

Zu a): Zunächst soll die Gleichung (4.12.b) so umgeschrieben werden, daß sie als Linearkombination der Rechtseigenvektoren $\{\underline{r}_j(\underline{v}^0(\xi)\}_{j=1}^m$ gegeben ist. Dazu führe mit (4.3) folgende Transformationen durch:

$$(1) \qquad \underline{h}^k(\xi) \;=\; \sum_{j=1}^m \gamma_j^k(\xi)\,\underline{r}_j(\underline{v}^0(\xi)) \text{ mit } \gamma_j^k(\xi) := \langle\, \underline{l}_j(\underline{v}^0(\xi)), \underline{h}^k(\xi)\,\rangle;$$

$$(2) \qquad \underline{r}_j'(\underline{v}^0(\xi))\,\underline{r}_i(\underline{v}^0(\xi)) \;=\; \sum_{l=1}^m \omega_{i,j}^l(\xi)\,\underline{r}_l(\underline{v}^0(\xi)) \text{ mit}$$

$$\omega_{i,j}^l(\xi) := \langle\, \underline{l}_l(\underline{v}^0(\xi)), \underline{r}_j'(\underline{v}^0(\xi))\,\underline{r}_i(\underline{v}^0(\xi))\,\rangle.$$

Da $\{\underline{r}_j(\underline{u})\}_{j=1}^m$ die Rechsteigenvektoren von $\underline{F}'(\underline{u})$ sind, gilt mit (4.2):

$$(3) \qquad \underline{F}'(\underline{u})\,\underline{r}_j(\underline{u}) \;=\; \lambda_j(\underline{u})\,\underline{r}_j(\underline{u}).$$

Setze nun (1) und (4.17) in (4.12.b) ein. Durch Anwendung der Produktregel erhält man dann unter Berücksichtigung von (2), (3) und (4.20.a,b) und abschließendem Koeffizientenvergleich bzgl. der Basisvektoren $\{\underline{r}_j(\underline{u})\}_{j=1}^m$ (vgl. (4.3.b)) folgende Gleichungen:

$$(4) \qquad [\,\lambda_j(\underline{v}^0(\xi)) - \xi\,]\;\tfrac{d}{d\xi}\,\alpha_j^k(\xi) \;+\; [\,k + \langle\, grad_{\underline{u}}\,\lambda_j(\underline{v}^0(\xi)), \underline{r}_i(\underline{v}^0(\xi))\,\rangle\,]\;\alpha_j^k(\xi) \;=$$

$$\gamma_j^k(\xi) \;-\; \sum_{\substack{l=1\\ l\neq i}}^m [\,\lambda_l(\underline{v}^0(\xi)) - \xi\,]\;\omega_{i,l}^j\;\alpha_l^k(\xi), \quad j = 1,\ldots,m.$$

Wegen (4.20.b) gilt nun mit (4.2.a):

$$\lambda_1(\underline{v}^0(\xi)) < \ldots < \lambda_i(\underline{v}^0(\xi)) = \xi < \ldots < \lambda_m(\underline{v}^0(\xi)) \quad \text{für } \xi \in R_i.$$

Daraus folgt mit (4):

$$(4.24.a) \qquad \tfrac{d}{d\xi}\,\alpha_j^k(\xi) \;+\; \sum_{\substack{l=1\\ l\neq i}}^m a_{j,l}\,\alpha_j^k(\xi) \;=\; f_j(\xi), \quad j \neq i$$

$$\text{mit } a_{j,j} := \frac{k + \langle\, grad_{\underline{u}}\,\lambda_j(\underline{v}^0(\xi)), \underline{r}_i(\underline{v}^0(\xi))\,\rangle}{\lambda_j(\underline{v}^0(\xi)) - \xi} \;+\; \omega_{i,j}^j,$$

$$a_{j,l} := \frac{\lambda_l(\underline{v}^0(\xi)) - \xi}{\lambda_j(\underline{v}^0(\xi)) - \xi}\,\omega_{i,l}^j \quad l \neq i, j\ ,$$

$$f_j(\xi) := \frac{\gamma_j^k(\xi)}{\lambda_j(\underline{v}^0(\xi)) - \xi};$$

(4.24.b) $$\alpha_i^k(\xi) \;=\; \frac{1}{k+1}\left[\gamma_i^k(\xi) - \sum_{\substack{l=1\\ l\neq i}}^{m} \big(\, \lambda_l(\underline{v}^0(\xi)) - \xi \,\big)\; \omega_{i,l}^i\, \alpha_l^k(\xi)\right].$$

Damit gilt die Behauptung a).

Zu b): Seien nun $\alpha_k^j(\underline{\sigma}_i^0+0)$, $j\neq i$, vorgegeben. Da $\underline{l}_j(\underline{v}^0(\circ)), \underline{h}^k(\underline{v}^0(\circ)), \lambda(\underline{v}^0(\circ))$ stetig sind, ist auch $\underline{f}(\circ)$ stetig. Somit besitzt (4.24.a) als inhomogenes System von Differentialgleichungen 1. Ordnung mit nicht konstanten Koeffizienten eine eindeutig existierende Lösung, die die Form aus Behauptung b) hat (vgl. [Wa]). Dabei versteht man unter einem Hauptsystem der homogenen Differentialgleichung, die Matrix, deren Zeilen aus linear unabhängigen Lösungen der homogenen Differentialgleichung besteht.

■

Wie aus La. 4.6 und La. 4.7 hervorgeht, sind die $\alpha_j^k(\circ)$ in dem Verdünnungsbereich R_i von $\underline{v}^0$ bekannt, falls man die Anfangswerte kennt, und sonst sind sie bis auf einen skalaren Parameter bestimmt. Um diese zu ermitteln, muß man die $\{a_{i,j}^k\}_{j=1}^m$ mit den $\{a_{i-1,j}^k\}_{j=1}^m$ in Verbindung bringen. Dazu untersucht man das Verhalten von $\underline{v}^k$ entlang den charakteristischen Kurven. Man unterscheidet dabei wieder zwischen den beiden Fällen, daß $\underline{v}^0$ entlang $x=\varphi(t)$ stetig oder unstetig ist. Dazu zunächst

Lemma 4.8:

Vor.: $x=\varphi(t) \overset{!}{=} \sum\limits_{l=0}^{k} \sigma_i^l\, t^{l+1} \;+\; O\left(t^{k+2}\right),\ k\in\mathbb{N},\ t>0,$

$\beta_{j,l} := \langle\, \underline{l}_l(\underline{v}_i^0), \underline{r}_j(\underline{v}_{i-1}^0)\,\rangle,\ \gamma_j := -\ \langle l_j(\underline{v}_i^0), [\underline{v}^0]|_{(\sigma_i^0)}\,\rangle,$

$q_{i,j}^k := -\langle\, \underline{l}_j(\underline{v}_i^0), [\underline{w}^k]|_{(\sigma_i^0)}\,\rangle,\ 1\leq i,j,l\leq m,\ \underline{w}^k$ gemäß (4.16)

Beh.: a) Ist die i–Welle von $\underline{v}^0$ vom Typ B, dann gilt:

$$\alpha_j^k(\sigma_i^0+0) = \sum_{l=1}^{m}\left[\frac{\lambda_l(\underline{v}_{i-1}^0)-\sigma_i^0}{\lambda_j(\underline{v}_i^0)-\sigma_i^0}\right]\alpha_l^k(\sigma_i^0-0)\,\beta_{l,j} \;+\; \frac{q_{i,j}^k \;-\; \sigma_i^k\,(k+1)\,\gamma_j}{\lambda_j(\underline{v}_i^0)-\sigma_i^0};$$

b) Ist die i–Welle von $\underline{v}^0$ vom Typ C, so gilt:

$$\sigma_i^k = \frac{1}{(k+1)\,\gamma_i}\left(\sum_{\substack{l=1\\ l\neq i}}^{m}\left[\,\lambda_l(\underline{v}_{i-1}^0)-\lambda_i(\underline{v}_{i-1}^0)\,\right]\,\alpha_l^k(\sigma_i^0-0)\,\beta_{l,i} \;+\; q_{i,i}^k\right),$$

$$\alpha_j^k(\sigma_i^0+0) = \sum_{\substack{l=1\\ l\neq i}}^{m}\left[\frac{\lambda_l(\underline{v}_{i-1}^0)-\lambda_i(\underline{v}_{i-1}^0)}{\lambda_j(\underline{v}_i^0)-\lambda_i(\underline{v}_i^0)}\right]\alpha_l^k(\sigma_i^0-0)\,\beta_{l,j}$$

$$+\;\frac{q_{i,j}^k-\sigma_i^k\,(k+1)\,\gamma_j}{\lambda_j(\underline{v}_i^0)-\lambda_i(\underline{v}_i^0)},\ j\neq i.$$

Beweis:

Ist die i–Welle von $\underline{v}^0$ vom Typ B oder C, so sind die Zustände links und rechts der charakteristischen Kurve $x = \sigma_i^0\, t$ konstant (vgl. (4.19)). Mit (4.19) gilt somit:

(1) $$\underline{v}^0(\sigma_i^0 + 0) \;=\; \underline{v}_i^0,\; \underline{v}^0(\sigma_i^0 - 0) \;=\; \underline{v}_{i-1}^0;$$

(2) $$\tfrac{d}{d\xi}\, \underline{v}^0(\sigma_i^0 \pm 0) \;=\; 0.$$

Aufgrund der Bezeichnungen aus der Voraussetzung gelten mit (4.3) ferner:

(3)
$$\text{a)} \qquad \underline{r}_j(\underline{v}_{i-1}^0) \;=\; \sum_{l=1}^{m} \beta_{j,l}\, \underline{r}_l(\underline{v}_i^0), \quad j = 1,\ldots,m;$$
$$\text{b)} \qquad -\,[\underline{w}^k]\Big|_{(\sigma_i^0)} \;=\; \sum_{l=1}^{m} q_{i,l}^k\, \underline{r}_l(\underline{v}_i^0);$$
$$\text{c)} \qquad -\,[\underline{v}^0]\Big|_{(\sigma_i^0)} \;=\; \sum_{l=1}^{m} \gamma_l\, \underline{r}_l(\underline{v}_i^0);$$

Da $\lambda_i(\underline{u})$ ein Eigenwert von $\underline{F}'(\underline{u})$ ist (vgl. (4.2)), gilt insbesondere auch:

$$\big(\underline{F}'(\underline{u})\big)^k\; \underline{r}_i(\underline{u}) \;=\; \big(\underline{F}'(\underline{u})\big)^{k-1}\; \underline{F}'(\underline{u})\, \underline{r}_i(\underline{u}) \;=\; \big(\underline{F}'(\underline{u})\,\big)^{k-1}\; \underline{r}_i(\underline{u})\, \lambda_i(\underline{u}).$$

Daraus folgt unmittelbar:

(4) $$(\lambda_i(\underline{u}))^k \text{ ist ein Eigenwert von } \big(\,\underline{F}'(\underline{u})\big)^k .$$

Unter Berücksichtigung von (1),…,(4) erhält man mit (4.16.b) durch anschließenden Koeffizientenvergleich bzgl. der Basisvektoren $\{\underline{r}_j(\underline{v}_i^0)\}_{j=1}^m$ wegen (4.3.b) folgendes Gleichungssystem:

(5) $$\sum_{l=1}^{m} \big[\; \lambda_l(\underline{v}_{i-1}^0) - \sigma_i^0 \;\big]\;\; \alpha_l^k(\sigma_i^0 - 0)\, \beta_{l,j} \;-\; \sigma_i^k\; (k+1)\, \gamma_j \;+\; q_{i,j}^k \;\stackrel{!}{=}$$
$$\big[\; \lambda_j(\underline{v}_i^0) - \sigma_i^0 \;\big]\;\; \alpha_j^k(\sigma_i^0 + 0),\;\; j = 1,\ldots,m.$$

Fall 1: i–Welle ist vom Typ B:

Aus (4.21.b) folgt mit (4.2.a):

$$\lambda_i(\underline{v}_i^0) < \sigma_i^0 < \lambda_i(\underline{v}_{i-1}^0)$$

und somit
$$\lambda_j(\underline{v}_i^0) \neq \sigma_i^0,\;\; \lambda_j(\underline{v}_{i-1}^0) \neq \sigma_i^0, \quad j = 1,\ldots,m.$$

Wegen (5) folgt nun unmittelbar die Behauptung a).

Fall 2: i–Welle ist vom Typ C:

Aufgrund von (4.21.a) und (4.2.a) gilt:

$$\lambda_i(\underline{v}_i^0) = \sigma_i^0 = \lambda_i(\underline{v}_{i-1}^0) \text{ und } \lambda_j(\underline{v}_i^0) \neq \sigma_i^0,\ j \neq i.$$

Ferner gilt: $\gamma_i \neq 0$, denn:

Da $x = \sigma_i^0\, t$ eine Diskontinuitätskurve von $\underline{v}^0$ ist (vgl. (4.19)), gilt insbesondere:

$$(6) \qquad [\underline{v}^0]\Big|_{(\sigma_i^0)} \neq \underline{0}.$$

Falls $\gamma_i = 0$ wäre, so würde wegen (3c) gelten:

$$-\ [\underline{v}^0]\Big|_{(\sigma_i^0)} = \sum_{\substack{s=1\\ s\neq i}}^{m} \gamma_s\, \underline{r}_s(\underline{v}_i^0).$$

Eine Multiplikation von $\underline{l}_j(\underline{v}_i^0)$ von links an obige Gleichung würde ergeben:

$$\gamma_j \overset{(\text{def.})}{=} -\langle\, \underline{l}_j(\underline{v}_i^0), [\underline{v}^0]|_{(\sigma_i^0)}\,\rangle = \sum_{\substack{s=1\\ s\neq i}}^{m} \gamma_s\, \langle\, \underline{l}_j(\underline{v}_i^0), \underline{r}_s(\underline{v}_i^0)\,\rangle \overset{(4.2.\text{b})}{=} 0.$$

Wegen (3c) würde somit gelten:

$$[\underline{v}^0]\Big|_{(\sigma_i^0)} = \underline{0}.$$

Dies wäre ein Widerspruch zu (6).

Aus obigem folgt mit (5) unmittelbar die Behauptung b).

■

An dieser Stelle sei noch vermerkt, daß die $\{q_{i,j}^k\}$ aus obigem Lemma nur von $\underline{v}^l(\sigma_i^0 \pm 0)$ und σ_i^l für $l = 0, \dots, k-1$ abhängen!

Eine ähnliche Aussage läßt sich nun über das Verhalten von $\underline{v}^k$ entlang dem Rand einer Verdünnungswelle von $\underline{v}^0$ machen:

Lemma 4.9:

Vor.: $x = \underline{\varphi}_i(t) \overset{!}{=} \sum\limits_{l=0}^{k} \underline{\sigma}_i^l\, t^{l+1} + O\left(t^{k+2}\right),\ x = \overline{\varphi}_i(t) \overset{!}{=} \sum\limits_{l=0}^{k} \overline{\sigma}_i^l\, t^{l+1} + O\left(t^{k+2}\right),$

$k \in \mathbb{N},\ t > 0,\ i$–Welle vom Typ A, $i \in \{1, \dots, m\}$;

$\underline{q}_{i,j}^k := \langle\, \underline{l}_j(\underline{v}_{i-1}^0), [\underline{z}^k]|_{(\underline{\sigma}_i^0)}\,\rangle, \overline{q}_{i,j}^k := \langle\, \overline{l}_j(\underline{v}_i^0), [\underline{z}^k]|_{(\overline{\sigma}_i^0)}\,\rangle, \underline{z}^k$ gemäß (4.3)

Beh.: a) $\alpha_j^k(\underline{\sigma}_i^0 + 0) = \alpha_j^k(\underline{\sigma}_i^0 - 0) + \underline{q}_{i,j}^k, \qquad j \neq i;$

$\alpha_i^k(\underline{\sigma}_i^0 + 0) = \alpha_i^k(\underline{\sigma}_i^0 - 0) + \underline{q}_{i,i}^k - \underline{\sigma}_i^k;$

b) $\alpha_j^k(\overline{\sigma}_i^0 + 0) = \alpha_j^k(\overline{\sigma}_i^0 - 0) - \overline{q}_{i,j}^k, \qquad j \neq i;$

$\alpha_i^k(\overline{\sigma}_i^0 + 0) = \alpha_i^k(\overline{\sigma}_i^0 - 0) - \overline{q}_{i,i}^k - \overline{\sigma}_i^k.$

Beweis:

Wegen (4.22) gilt mit (4.16.a) und (4.20):

(1) $\underline{v}^0(\underline{\sigma}_i^0 - 0) \;=\; \underline{v}^0(\underline{\sigma}_i^0 + 0) \;=\; \underline{v}_{i-1}^0;$

$\frac{d}{d\xi}\,\underline{v}^0(\underline{\sigma}_i^0 - 0) \;=\; 0,\; \frac{d}{d\xi}\,\underline{v}^0(\underline{\sigma}_i^0 + 0) \;=\; \underline{r}_i(\underline{v}_{i-1}^0);$

(2) $\underline{v}^0(\overline{\sigma}_i^0 - 0) \;=\; \underline{v}^0(\overline{\sigma}_i^0 + 0) \;=\; \underline{v}_i^0;$

$\frac{d}{d\xi}\,\underline{v}^0(\overline{\sigma}_i^0 + 0) \;=\; 0,\; \frac{d}{d\xi}\,\underline{v}^0(\overline{\sigma}_i^0 - 0) \;=\; \underline{r}_i(\underline{v}_i^0);$

Aus (4.2.a) folgt ferner:

(3) $\underline{F}'(\underline{v}^0)\,\underline{r}_j(\underline{v}^0) \;=\; \underline{r}_j(\underline{v}^0)\,\lambda_j(\underline{v}^0), \quad j = 1,\ldots,m$

Somit folgt durch Einsetzen von (4.17) in (4.16.b) für $\sigma^0 = \underline{\sigma}_i^0$ bzw. $\sigma^0 = \overline{\sigma}_i^0$ unter Berücksichtigung von (1) bzw. (2) und (3) durch abschließenden Koeffizientenvergleich bzgl. der Basen $\{\underline{r}_j(\underline{v}_{i-1}^0)\}_{j=1}^m$ bzw. $\{\underline{r}_j(\underline{v}_i^0)\}_{j=1}^m$ des $\mathbb{R}^3$:

(a) $\alpha_j^k(\underline{\sigma}_i^0 + 0) - \alpha_j^k(\underline{\sigma}_i^0 - 0) + \underline{q}_{i,j}^k + \delta_{i,j}\underline{\sigma}_i^k \;=\; 0;$

(b) $\alpha_j^k(\overline{\sigma}_i^0 + 0) - \alpha_j^k(\overline{\sigma}_i^0 - 0) + \overline{q}_{i,j}^k + \delta_{i,j}\overline{\sigma}_i^k \;=\; 0.$

Dabei wird mit $\delta_{i,j}$ das Kronecker–Symbol bezeichnet. Aus (a) und (b) folgt nun unmittelbar die Behauptung.

■

Hier sei wiederum vermerkt, daß die $\{\underline{q}_{i,j}^k\}$ bzw. $\{\overline{q}_{i,j}^k\}$ aus obigem Lemma nur von $v^l(\underline{\sigma}_i^0 \pm 0)$ und $\underline{\sigma}_i^l$ bzw. $\underline{v}^l(\overline{\sigma}_i^0 \pm 0)$ und $\overline{\sigma}_i^l$ für $l = 0,\ldots,k-1$ abhängen!
Nachdem nun zum einen die $\alpha_j^k(\circ)$ aus (4.17) *zwischen* (vgl. La. 4.6, La. 4.7) und zum anderen *entlang* (vgl. La. 4.8, La. 4.9) den charakteristischen Kurven bis auf bestimmte Parameter bestimmt sind, sollen nun diese Werte ermittelt werden. Dabei macht man sich zu nutze, daß die $\alpha_j^k(\circ)$ für $\xi \in D_0$ und $\xi \in D_m$ wegen der Randbedingungen aus La. 4.4 bekannt sind. Somit sind nur die $\{a_{i,j}^k\}$ für $i = 1,\ldots,m-1,\; j = 1,\ldots,m$ aus (4.23) zu bestimmen. Die σ_i^k bzw. $\underline{\sigma}_i^k$ und $\overline{\sigma}_i^k$ lassen sich anschließend mittels La. 4.8.b bzw. La. 4.9.b ermitteln. Dazu verbindet man den Zustand in D_m mit dem aus D_0 von rechts nach links durch Ausnutzung der Beziehungen an den Rändern der charakteristischen Kurven (vgl. La. 4.8, La. 4.9) und erhält mit den Randbedingungen aus La. 4.4 ein lineares Gleichungssystem (LGS). Leider läßt sich dieses hier nicht so allgemein explizit angeben, da es von den i–Wellen von $\underline{v}^0$ abhängt und diese sind von System zu System und je nach Anfangsdaten unterschiedlich. Im folgenden soll deshalb nur der Weg beschrieben werden, wie man das LGS erstellt. Da dieser in [LRa] nicht detailliert hergeleitet wurde, soll dies hier geschehen. Dazu:

Lemma 4.10.a:

Vor.: i–Welle von $\underline{v}^0$ sei vom Typ B, $\xi \in D_i,\; i \in \{1,\ldots,m\},\; p_{j,i}^k(\xi)$ gemäß La. 4.6;

Beh.: Für $j = 1,\ldots,m$ gibt es eindeutig bestimmte $c_j^i, K_j^i \in \mathbb{R}, \{K_{j,l}^i\}_{l=1}^m \in \mathbb{R}^m$, so daß gilt:

i) $\alpha_j^k(\xi) \;=\; [\,\xi - \lambda_j(\underline{v}_i^0)\,]^k \;\left(\sum\limits_{l=1}^m K_{j,l}^i\,\alpha_l^k(\sigma_i^0 - 0) + c_j^i + \sigma_i^k\,K_j^i\right) \;+\; p_{j,i}^k(\xi);$

ii) $\left(\frac{d}{d\xi}\right)^k\;\alpha_j^k(\xi) \;=\; k!\;\left(\sum\limits_{l=1}^m K_{j,l}^i\,\alpha_l^k(\sigma_i^0 - 0) + c_j^i + \sigma_i^k\,K_j^i\right).$

Beweis:

Wegen (4.21) gilt : $\sigma_i^0 := \overline{\sigma}_i^0 \overset{!}{=} \underline{\sigma}_i^0$, und mit La. 4.6 folgt:

$$(1) \qquad \alpha_j^k(\sigma_i^0 + 0) \;=\; \left(\, \sigma_i^0 - \lambda_j(\underline{v}_i^0)\,\right)^k \; a_{j,i}^k \;+\; p_{j,i}^k(\sigma_i^0 + 0), \quad j = 1, \ldots, m.$$

Andererseits gilt mit La. 4.8.a und den dort verwendeten Bezeichnungen:

$$(2) \qquad \begin{gathered} \alpha_j^k(\sigma_i^0 + 0) \;=\; \sum_{l=1}^{m} \tilde{K}_{j,l}^i \, \alpha_l^k(\sigma_i^0 - 0) \;+\; \tilde{c}_j^i \;+\; \sigma_i^k \, \tilde{K}_j^i \text{ mit} \\ \tilde{K}_{j,l}^i \;:=\; \frac{\lambda_l(\underline{v}_{i-1}^0) - \sigma_i^0}{\lambda_j(\underline{v}_{i-1}^0) - \sigma_i^0} \, \beta_{l,j}, \;\; \tilde{c}_j^i \;:=\; \frac{q_{i,j}^k}{\lambda_j(\underline{v}_{i-1}^0) - \sigma_i^0}, \;\; \tilde{K}_j^i \;:=\; -\frac{k+1}{\lambda_j(\underline{v}_{i-1}^0) - \sigma_i^0} \, \gamma_j. \end{gathered}$$

Durch Gleichsetzen von (1) und (2) erhält man durch Auflösen nach $a_{j,i}^k$ und anschließendem Einsetzen in (4.23) die Behauptung mit

$$K_{j,l}^i \;:=\; \frac{\tilde{K}_{j,l}^i}{\left(\, \sigma_i^0 - \lambda_j(\underline{v}_i^0)\,\right)^k}, \qquad c_j^i \;:=\; \frac{\tilde{c}_j^i - p_{j,i}^k(\sigma_i^0 + 0)}{\left(\, \sigma_i^0 - \lambda_j(\underline{v}_i^0)\,\right)^k}.$$

Die Darstellung für die k–te Ableitung ergibt sich damit unmittelbar aus der Tatsache, daß gilt (vgl. (4.23)): $p_{j,i}^k \in \mathbb{P}_{k-1}$.

■

Lemma 4.10.b:

Vor.: i–Welle von $\underline{v}^0$ sei vom Typ C, $\xi \in D_i$, $i \in \{1, \ldots, m\}$, $p_{j,i}^k(\xi)$ gemäß La. 4.6;

Beh.: Für $j = 1, \ldots, m$ gibt es eindeutig bestimmte $\hat{c}_j^i \in \mathbb{R}$, $\{\hat{K}_{j,l}^i\}_{l=1,\neq i}^m \in \mathbb{R}^{m-1}$, so daß gilt:

i) $$\alpha_j^k(\xi) \;=\; \left[\, \xi - \lambda_j(\underline{v}_i^0)\,\right]^k \left\{ \sum_{\substack{l=1 \\ l \neq i}}^{m} \hat{K}_{j,l}^i \, \alpha_l^k(\sigma_i^0 - 0) \;+\; \hat{c}_j^i \right\} \;+\; p_{j,i}^k(\xi), \;\; j \neq i;$$

$$\alpha_i^k(\xi) \;=\; \left[\, \xi - \lambda_i(\underline{v}_i^0)\,\right]^k \; a_{i,i}^k \;+\; p_{i,i}^k(\xi);$$

ii) $$\left(\tfrac{d}{d\xi}\right)^k \; \alpha_j^k(\xi) \;=\; k! \left\{ \sum_{\substack{l=1 \\ l \neq i}}^{m} \hat{K}_{j,l}^i \, \alpha_l^k(\sigma_i^0 - 0) + \hat{c}_j^i \right\}, \;\; j \neq i;$$

$$\left(\tfrac{d}{d\xi}\right)^k \; \alpha_i^k(\xi) \;=\; k! \; a_{i,i}^k.$$

Beweis:

Wegen (4.21) gilt : $\sigma_i^0 := \overline{\sigma}_i^0 \overset{!}{=} \underline{\sigma}_i^0 \;=\; \lambda_i(\underline{v}_i^0) = \lambda_i(\underline{v}_{i-1}^0)$. Mit La. 4.6 folgt somit:

$$(1) \qquad \alpha_j^k(\sigma_i^0 + 0) \;=\; \left(\, \lambda_i(\underline{v}_i^0) - \lambda_j(\underline{v}_i^0)\,\right)^k \; a_{j,i}^k \;+\; p_{j,i}^k(\sigma_i^0 + 0), \quad j \neq i.$$

Andererseits gilt mit La. 4.8.b und den dort verwendeten Bezeichnungen durch Einsetzen von σ_i^0 in $\alpha_j^k(\sigma_i^0+0)$:

$$\alpha_j^k(\sigma_i^0+0) \;=\; \sum_{\substack{l=1\\ l\neq i}}^{m} \tilde{K}_{j,l}^i\, \alpha_l^k(\sigma_i^0-0) \;+\; \tilde{c}_j^i,\; j\neq i, \qquad \text{mit}$$

(2)

$$\tilde{K}_{j,l}^i \;:=\; \frac{\lambda_l(\underline{v}_{i-1}^0)-\lambda_i(\underline{v}_{i-1}^0)}{\lambda_j(\underline{v}_i^0)-\lambda_i(\underline{v}_i^0)}\left[\beta_{l,j}-\beta_{l,i}\,\frac{\gamma_j}{\gamma_i}\right],\; \tilde{c}_j^i \;:=\; \frac{q_{i,j}^k \;+\; q_{i,i}^k\,\frac{\gamma_j}{\gamma_i}}{\lambda_j(\underline{v}_i^0)-\lambda_i(\underline{v}_i^0)}.$$

Setze nun (1) und (2) gleich und löse nach $a_{j,i}^k$ auf. Den so bestimmten Parameter $a_{j,i}^k$ setze in (4.23) ein. Dann erhält man die Behauptung mit

$$\hat{K}_{j,l}^i \;:=\; \frac{\tilde{K}_{j,l}^i}{\big(\,\lambda_i(\underline{v}_i^0)-\lambda_j(\underline{v}_i^0)\,\big)^k}\;;\quad \hat{c}_j^i \;:=\; \frac{\tilde{c}_j^i - p_{j,i}^k(\sigma_i^0+0)}{\big(\,\lambda_i(\underline{v}_i^0)-\lambda_j(\underline{v}_i^0)\,\big)^k}.$$

Da die $p_{j,i}^k$ Polynome vom Grad $k-1$ sind (vgl. (4.23)), folgt nun unmittelbar mit obigem die Darstellung für die k–te Ableitung.

■

Lemma 4.10.c:

Vor.: i–Welle von $\underline{v}^0$ sei vom Typ A, $\xi\in D_i,\; i\in\{1,\ldots,m\},\; p_{j,i}^k(\xi)$ gemäß La. 4.6;

Beh.: Für $j=1,\ldots,m,\; j\neq i$, gibt es eindeutig bestimmte $\tilde{c}_j^i, \tilde{K}_j^i\in\mathbb{R}$, so daß gilt:

i) $\alpha_j^k(\xi) \;=\; \big[\,\xi-\lambda_j(\underline{v}_i^0)\,\big]^k\,\Big\{\,\tilde{K}_j^i\,\alpha_j^k(\overline{\sigma}_i^0-0) \;+\; \tilde{c}_j^i\Big\} \;+\; p_{j,i}^k(\xi),\; j\neq i;$

$\alpha_i^k(\xi) \;=\; \big[\,\xi-\lambda_i(\underline{v}_i^0)\,\big]^k\; a_{i,i}^k \;+\; p_{i,i}^k(\xi);$

ii) $\left(\frac{d}{d\xi}\right)^k\,\alpha_j^k(\xi) \;=\; k!\,\Big\{\,\tilde{K}_j^i\,\alpha_j^k(\overline{\sigma}_i^0-0)+\tilde{c}_j^i\Big\},\; j\neq i;$

$\left(\frac{d}{d\xi}\right)^k\,\alpha_i^k(\xi) \;=\; k!\,a_{i,i}^k.$

Beweis:

Mit (4.20.a) gilt : $\underline{\sigma}_i^0=\lambda_{i-1}(\underline{v}_i^0)$ und $\overline{\sigma}_i^0=\lambda_i(\underline{v}_i^0)$. Aus La. 4.6 folgt damit:

(1) $\qquad \alpha_j^k(\overline{\sigma}_i^0+0) \;=\; \big(\,\lambda_i(\underline{v}_i^0)-\lambda_j(\underline{v}_i^0)\,\big)^k\; a_{j,i}^k \;+\; p_{j,i}^k(\overline{\sigma}_i^0+0), \quad j=1\ldots,m.$

Andererseits gilt wegen La. 4.9.b:

$$\alpha_j^k(\overline{\sigma}_i^0+0) \;=\; \alpha_j^k(\overline{\sigma}_i^0-0) \;-\; \overline{q}_{i,j}^k,\quad j\neq i.$$

Gleichsetzen von (1) und (2) liefert durch Auflösen nach $a_{j,i}^k$ und Einsetzen in (4.23) die Behauptung mit

$$\tilde{K}_j^i \;:=\; \frac{1}{\big(\,\lambda_i(\underline{v}_i^0)-\lambda_j(\underline{v}_i^0)\,\big)^k},\quad \tilde{c}_j^i \;:=\; -\frac{p_{j,i}^k(\sigma_i^0+0) \;+\; \overline{q}_{i,j}^k}{\big(\,\lambda_i(\underline{v}_i^0)-\lambda_j(\underline{v}_i^0)\,\big)^k}.$$

Die Darstellung für die k–te Ableitung folgt daraus unter Berücksichtigung der Tatsache (vgl. (4.23)), daß gilt: $p_{j,i}^k\in\mathbb{P}_{k-1}$.

■

Falls die i–Welle von $\underline{v}^0$ vom Typ A ist, so gilt wegen (4.20.b): $\underline{\sigma}_i^0 < \overline{\sigma}_i^0$.
Somit ist $\alpha_j^k(\overline{\sigma}_i^0 - 0)$ vom Verdünnungsbereich von $\underline{v}^0$ abhängig. In La. 4.10.a,b wurden die α_j^k jedoch so bestimmt, daß sie von dem Intervall $D_{i-1} := (\overline{\sigma}_{i-1}^0, \underline{\sigma}_i^0)$ abhängen, in dem $\underline{v}^0$ konstant ist (vgl. (4.19)). Es ist daher sinnvoll, im Falle einer i–Verdünnungswelle die $\alpha_j^k|_{R_i}$ bzw. $\alpha_j^k|_{D_i}$ in Abhängigkeit von D_{i-1} zu bestimmen. Dazu:

Lemma 4.10.d:

Vor.: i–Welle von $\underline{v}^0$ sei vom Typ A, $\xi \in R_i$, $i \in \{1, \ldots, m\}$, $\omega_{i,l}^i(\xi)$ gemäß (4.24);

Beh.: Für $j = 1, \ldots, m$ gibt es eindeutig bestimmte $\acute{c}_j^i, \acute{K}_{j,l}^i : R_i \to \mathbb{R}$, so daß gilt:

$$\alpha_j^k(\xi) \;=\; \sum_{\substack{l=1\\ l\neq i}}^{m} \acute{K}_{j,l}^i(\xi)\, \alpha_l^k(\underline{\sigma}_i^0 - 0) + \acute{c}_j^i(\xi).$$

Beweis:

Für $\xi \in R_i$ gilt wegen La. 4.7:

$$(1) \qquad \underline{\alpha}(\xi) \;:=\; H(\xi)\, \underline{\alpha}(\underline{\sigma}_i^0 + 0) \;+\; \underline{K}(\xi) \text{ mit } \underline{K}(\xi) \;:=\; \int_{\underline{\sigma}_i^0}^{\xi} H(\xi)\, H^{-1}(\tau)\, \underline{f}(\tau)\, d\tau.$$

Dabei ist $\underline{\alpha}$ gemäß den Voraussetzungen zu La. 4.7 definiert und H das Hauptsystem der homogenen Differentialgleichung aus (4.24.a) mit $H(\underline{\sigma}_i^0) \;=\; Id$. Somit ist $H = (h_{j,l})$ insbesondere unabhängig von $\alpha(\underline{\sigma}_i^0 + 0)$.
Komponentenweise läßt sich (1) auch schreiben als:

$$(2) \qquad \alpha_j^k(\xi) \;=\; \sum_{\substack{l=1\\ l\neq i}}^{m} h_{j,l}(\xi)\, \alpha_l^k(\underline{\sigma}_i^0 + 0) \;+\; K_j(\xi), \quad j \neq i.$$

Durch Einsetzen von La. 4.9.a in (2) erhält man dann für $j \neq i$

$$(3) \qquad \begin{aligned} \alpha_j^k(\xi) \;&=\; \sum_{\substack{l=1\\ l\neq i}}^{m} \acute{K}_{j,l}^i(\xi)\, \alpha_l^k(\underline{\sigma}_i^0 - 0) + \acute{c}_j^i(\xi) \quad \text{mit} \\ \acute{K}_{j,l}^i(\xi) \;&:=\; h_{j,l}(\xi),\;\; \acute{c}_j^i(\xi) \;:=\; K_j(\xi) \;+\; \sum_{\substack{l=1\\ l\neq i}}^{m} h_{j,l}(\xi)\, \underline{q}_{i,j}^k. \end{aligned}$$

Setze (3) in (4.24.b) ein. Dann folgt die Behauptung für $j = i$ mit

$$\acute{K}_{i,l}^i(\xi) \;:=\; -\frac{1}{k+1} \sum_{\substack{s=1\\ s\neq i}}^{m} \left[\, \lambda_s(\underline{v}^0(\xi)) - \xi \,\right]\, \omega_{i,s}^i\, \acute{K}_{s,l}^i(\xi),$$

$$\acute{c}_i^i(\xi) \;:=\; \frac{1}{k+1} \left(\gamma_i^k(\xi) \;-\; \sum_{\substack{l=1\\ l\neq i}}^{m} \left[\, \lambda_l(\underline{v}^0(\xi)) - \xi \,\right]\, \omega_{i,l}^i\, \acute{c}_l^i(\xi) \right).$$

■

Nachdem nun je zwei benachbarte Gebiete, in denen $\underline{v}^0$ konstant ist, d.h. D_{i-1} und D_i für $i = 1, \ldots, m$, in Verbindung gebracht worden sind, lassen sich nun die $\alpha_j^k(\xi)$ für $\xi > \overline{\sigma}_m^0$ wie folgt darstellen:

Lemma 4.11:

Vor.: $\xi \in D_m,\ i \in \{1, \ldots, m\}$;

Beh.: Es gibt für $j = 1, \ldots, m$ eindeutig bestimmte, nur vom Typ der m–Welle von $\underline{v}^0$ abhängige Konstanten $\overline{K}_{j,l}, \overline{K}_{j,l}^i, \overline{c}_j^i \in \mathbb{R},\ l \neq i$, so daß für $(a_i, \ldots, a_m) \in \mathbb{R}^{m-i}$ beliebig gilt:

a) Ist die m–Welle von $\underline{v}^0$ vom Typ A oder C, dann gilt:

$$\frac{1}{k!}\left(\frac{d}{d\xi}\right)^k \alpha_j^k(\xi) \;=\; \sum_{l=i}^{m-1} a_l\, \overline{K}_{l,j} \;+\; \sum_{\substack{l=1\\ l\neq i}}^{m} \overline{K}_{j,l}^i\, \alpha_l^k(\underline{\sigma}_i^0 - 0) \;+\; \overline{c}_j^i,\ j \neq m;$$

$$\frac{1}{k!}\left(\frac{d}{d\xi}\right)^k \alpha_m^k(\xi) \;=\; a_m;$$

b) Ist die m–Welle von $\underline{v}^0$ vom Typ B, dann gilt:

$$\frac{1}{k!}\left(\frac{d}{d\xi}\right)^k \alpha_j^k(\xi) \;=\; \sum_{l=i}^{m} a_l\, \overline{K}_{l,j} \;+\; \sum_{l=1}^{m} \overline{K}_{j,l}^i\, \alpha_l^k(\underline{\sigma}_i^0 - 0) \;+\; \overline{c}_j^i.$$

Beweis:

Die Behauptung wird nun mittels vollständiger Induktion über i geführt. Sei dazu zunächst $i = m$. Dann gilt:

Fall 1: m–Welle von $\underline{v}^0$ ist vom Typ A:

Setze in La. 4.10.d $\xi := \overline{\sigma}_m^0 - 0$. Diese Darstellung von $\alpha_j^k(\overline{\sigma}_m^0 - 0)$ setze in La. 4.10.c ein. Dann erhält man die Behauptung mit

$$\overline{K}_{j,l}^m := \tilde{K}_j^m\, \acute{K}_{j,l}^m(\overline{\sigma}_m^0 - 0),\ \overline{c}_j^m := \tilde{K}_j^m\, \acute{c}_j^m(\overline{\sigma}_m^0 - 0) + \tilde{c}_j^m.$$

Dabei entspricht a_m der unbestimmten Größe $a_{m,m}^k$.

Fall 2: m–Welle von $\underline{v}^0$ ist vom Typ B:

Aus La. 4.10.a folgt die Behauptung mit $\overline{K}_{j,l}^m := K_{j,l}^m,\ \overline{K}_{m,j} := K_j^m,\ \overline{c}_j^m := c_j^m$.

Dabei entspricht a_m dem unbestimmten Parameter σ_m^k.

Fall 3: m–Welle von $\underline{v}^0$ ist vom Typ C:

Aus La. 4.10.b folgt die Behauptung wobei a_m der unbestimmten Größe $a_{m,m}^k$ entspricht und für die Koeffizienten gilt: $\overline{K}_{j,l}^i := \hat{K}_{j,l}^i,\ \overline{c}_j^m := \hat{c}_j^m$.

Sei nun die Behauptung für ein $i \in \{2, \ldots, m\}$ erfüllt. Dann gilt für $i-1$:

Fall 1: $(i-1)$–Welle von $\underline{v}^0$ ist vom Typ A:

Berechne in La. 4.10.d $\alpha_j^k(\xi)$ für $\xi = \overline{\sigma}_{i-1}^0 - 0$. Setze diese Darstellung von $\alpha_j^k(\overline{\sigma}_{i-1}^0 - 0)$ in La. 4.10.c ein. Mit der Induktionsvoraussetzung erhält man dann die Behauptung für $i-1$ mit

$$\overline{K}_{i-1,j}^{i-1} \;:=\; \left[\,\underline{\sigma}_i^0 - \lambda_{i-1}(\underline{v}_{i-1}^0)\,\right]^k\; \overline{K}_{j,i-1}^i;$$

$$\overline{K}_{j,l}^{i-1} \;:=\; \sum_{\substack{s=1\\ s\neq i,i-1}}^{m} \left[\,\underline{\sigma}_i^0 - \lambda_s(\underline{v}_{i-1}^0)\,\right]^k\; \acute{K}_{s,l}^{i-1}(\overline{\sigma}_{i-1}^0 - 0)\, \tilde{K}_s^{i-1}\, \overline{K}_{j,s}^i,\ l \neq i-1;$$

$$\overline{c}_j^{i-1} := \sum_{\substack{l=1\\l\neq i,i-1}}^{m} \overline{K}^i_{j,l} \left\{ \left[\underline{\sigma}_i^0 - \lambda_l(\underline{v}_{i-1}^0)\right]^k \left[\tilde{K}_l^{i-1} \acute{c}_l^{i-1}(\overline{\sigma}_{i-1}^0 - 0) + \tilde{c}_l^{i-1} \right] + p_{l,i-1}^k(\underline{\sigma}_i^0 - 0)\right\}$$

$$+ \overline{K}^i_{j,i-1}\ p^k_{i-1,i-1}(\underline{\sigma}_i^0 - 0)\ +\ \overline{c}_j^i.$$

Hier entspricht a_{i-1} dem Parameter $a^k_{i-1,i-1}$.

Fall 2: $(i-1)$–Welle von $\underline{v}^0$ ist vom Typ B:

Aus La. 4.10.a folgt mit der Induktionsvoraussetzung die Behauptung für $i-1$ mit

$$\overline{K}_{i-1,j}\ :=\ \sum_{l=1}^{m} \left[\ \underline{\sigma}_i^0 - \lambda_l(\underline{v}_{i-1}^0)\ \right]^k\ \overline{K}^i_{j,l}\ K_l^{i-1};$$

$$\overline{K}^{i-1}_{j,l}\ :=\ \sum_{s=1}^{m} \left[\ \underline{\sigma}_i^0 - \lambda_s(\underline{v}_{i-1}^0)\ \right]^k\ \overline{K}^i_{j,s}\ K^{i-1}_{s,l};$$

$$\overline{c}_j^{i-1}\ :=\ \sum_{l=1}^{m} \left\{\ \left[\ \underline{\sigma}_i^0 - \lambda_l(\underline{v}_{i-1}^0)\ \right]^k\ \overline{K}_{j,l}\ c_l^{i-1}\ +\ p^k_{l,i-1}(\underline{\sigma}_i^0 - 0)\ \right\}\ +\ \overline{c}_j^i.$$

Hier entspricht a_{i-1} dem unbestimmten Parameter σ^k_{i-1}.

Fall 3: $(i-1)$–Welle von $\underline{v}^0$ ist vom Typ C:

Die Behauptung für $i-1$ folgt mittels der Induktionsvoraussetzung unmittelbar aus La. 4.10.b mit

$$\overline{K}^{i-1}_{i-1,j}\ :=\ \left[\ \underline{\sigma}_i^0 - \lambda_{i-1}(\underline{v}_{i-1}^0)\ \right]^k\ \overline{K}^i_{j,i-1};$$

$$\overline{K}^{i-1}_{j,l}\ :=\ \sum_{\substack{s=1\\s\neq i,i-1}}^{m} \left[\ \underline{\sigma}_i^0 - \lambda_s(\underline{v}_{i-1}^0)\ \right]^k\ \overline{K}^i_{j,s}\ \hat{K}^{i-1}_{s,l},\ l \neq i-1;$$

$$\overline{c}_j^{i-1}\ :=\ \sum_{\substack{l=1\\l\neq i,i-1}}^{m} \overline{K}^i_{j,l}\ \left\{\ \left[\ \underline{\sigma}_i^0 - \lambda_l(\underline{v}_{i-1}^0)\ \right]^k\ \ \hat{c}_l^{i-1}\ +\ p^k_{l,i-1}(\underline{\sigma}_i^0 - 0)\ \right\}$$

$$+ \overline{K}^i_{j,i-1}\ p^k_{i-1,i-1}(\underline{\sigma}_i^0 - 0)\ +\ \overline{c}_j^i.$$

Dabei entspricht a_{i-1} der freien Größe $a^k_{i-1,i-1}$.

Mit dem Induktionsprinzip folgt aus obigen Fällen die Behauptung.

■

Mit obigem Lemma läßt sich nun das gesuchte LGS mittels der Randbedingungen aus La. 4.4 aufstellen:

Folgerung 4.1:

Vor.: $r_j := \langle\, \underline{l}_j(\underline{v}_m^0), \left(\frac{d}{dx}\right)^k\ \underline{u}_R(0)\,\rangle,\ l_j := \langle\, \underline{l}_j(\underline{v}_0^0), \left(\frac{d}{dx}\right)^k\ \underline{u}_L(0)\,\rangle,\ j = 1,\ldots,m$

$\overline{K}_{l,j},\ \overline{K}^1_{j,l}$ gemäß La. 4.11;

Beh.: a) Ist die m–Welle von $\underline{v}^0$ vom Typ A oder C, dann gilt:

$$a_m \overset{!}{=} l_m, \qquad \sum_{l=1}^{m-1} \overline{K}_{l,j}\ a_l \overset{!}{=} b_j,\ j \neq i,\ \text{mit}$$

$$b_j\ :=\ \tfrac{1}{k!}\, r_j\ -\ \sum_{l=2}^{m} \overline{K}^1_{j,l}\ \Big\{ \left[\, \underline{\sigma}_1^0 - \lambda_l(\underline{v}_0^0)\,\right]^k\ \tfrac{1}{k!}\ l_l\ +\ p^k_{j,0}(\underline{\sigma}_1^0 - 0) \Big\}\ -\ \overline{c}^1_j.$$

b) Ist die m–Welle von $\underline{v}^0$ vom Typ B, dann gilt:

$$\sum_{l=1}^{m-1} \overline{K}_{l,j}\ a_l \overset{!}{=} b_j,\ j = 1,\ldots,m\ \text{mit}$$

$$b_j\ :=\ \tfrac{1}{k!}\, r_j\ -\ \sum_{l=1}^{m} \overline{K}^1_{j,l}\ \Big\{ \left[\, \underline{\sigma}_1^0 - \lambda_l(\underline{v}_0^0)\,\right]^k\ \tfrac{1}{k!}\ l_l\ +\ p^k_{j,0}(\underline{\sigma}_1^0 - 0) \Big\}\ -\ \overline{c}^1_j.$$

Beweis:

Wegen La. 4.6 gilt für $\xi \in D_i,\ i = 0,\ldots,m$:

(1)
$$\left(\frac{d}{d\xi}\right)^k\ \underline{v}^k(\xi)\ \overset{!}{=}\ \sum_{j=1}^{m} k!\ a^k_{i,j}\ \underline{r}_j(\underline{v}_i^0).$$

Andererseits muß aufgrund der Randbedingungen aus La. 4.4 gelten:

(2)
$$\begin{aligned}
\left(\frac{d}{d\xi}\right)^k\ \underline{v}^k(\xi)\ &\overset{!}{=}\ \left(\frac{d}{d\xi}\right)^k\ \underline{u}_L(0)\ \overset{(4.3.a)}{=}\ \sum_{j=1}^{m} l_j\ \underline{r}_j(\underline{v}_0^0), \quad \xi < \underline{\sigma}_1^0;\\
\left(\frac{d}{d\xi}\right)^k\ \underline{v}^k(\xi)\ &\overset{!}{=}\ \left(\frac{d}{d\xi}\right)^k\ \underline{u}_R(0)\ \overset{(4.3.a)}{=}\ \sum_{j=1}^{m} r_j\ \underline{r}_j(\underline{v}_m^0), \quad \xi < \underline{\sigma}_m^0.
\end{aligned}$$

Durch Koeffizientenvergleich bzgl. der Basisvektoren $\underline{r}_j(\underline{v}_0^0)$ bzw. $\underline{r}_j(\underline{v}_m^0)$ des $\mathbb{R}^m$ erhält man wegen (1) und (2):

(4.25)
$$\text{a)}\qquad a^k_{j,0}\ \overset{!}{=}\ \frac{1}{k!}\, l_j;$$
$$\text{b)}\qquad a^k_{j,m}\ \overset{!}{=}\ \frac{1}{k!}\, r_j.$$

Aus (4.25.a) folgt mit (4.23):

(3)
$$\alpha^k_j(\underline{\sigma}_1^0 - 0)\ =\ \left[\, \underline{\sigma}_1^0 - \lambda_j(\underline{v}_0^0)\,\right]^k\ \frac{1}{k!}\ l_j\ +\ p^k_{j,0}(\underline{\sigma}_1^0 - 0).$$

Durch Einsetzen von (4.25.b) in La. 4.11 mit $i = 1$ erhält man somit die Behauptung.

■

Aus obigem sieht man, daß, falls die m–Welle von $\underline{v}^0$ vom Typ A oder C ist, so ist ein $(m-1)\times(m-1)$ sonst ein $m \times m$ LGS zu lösen. Diese sind gemäß [LRa] eindeutig lösbar. Um das zu zeigen, haben P.A. Raviart und P. LeFloch aber das LGS nicht exakt bestimmt, sondern dazu die Eigenwerte $\lambda_i(\underline{u}), i = 1,\ldots,m$ approximiert und so ein genähertes LGS erstellt. Von diesem konnten sie dann die eindeutige Existenz der Lösung zeigen. Wegen der Stetigkeit von Matrizen in den einzelnen Komponenten gilt dies auch für das oben erstellte LGS, falls $\parallel \left(\frac{d}{dx}\right)^k \ \underline{u}_L(0) - \left(\frac{d}{dx}\right)^k \ \underline{u}_R(0) \parallel$ hinreichend klein ist.

Ist das entsprechende Gleichungssystem gelöst, so lassen sich die noch freien Parameter mittels La. 4.6–La. 4.9 durch sukzessives ermitteln der fehlenden Größen von $\alpha_j^k(\xi)$ für $\xi \in D_i, \ i = 1,\ldots,m$ bestimmen.

In der Praxis ist man in der Regel nur an $\underline{v}^k(0)$ bzw. $\alpha_j^k(0)$ interessiert (vgl. Kap. 2.1.4 bzw. Kap. 2.2.4). Dazu braucht man aber nicht immer das LGS aufzustellen, denn für $\underline{\sigma}_1^0 > 0$ oder $\overline{\sigma}_m^0 < 0$ ist $\underline{v}^k(0)$ wegen (4.25) unmittelbar berechenbar. Falls für ein $i_0 \in \{1,\ldots,m\}$ gilt: $\overline{\sigma}_{i_0}^0 < 0 < \underline{\sigma}_{i_0+1}^0$ bzw. $\underline{\sigma}_{i_0}^0 < 0 < \underline{\sigma}_{i_0}^0$, so muß zwar das LGS gelöst werden, aber die $\alpha_j^k(\xi), \ \xi \in D_i$ bzw. $\xi \in R_i$ nur für $i = 0,\ldots,m$ bestimmt werden.
Angewendet wurde diese Methode in [ScR] auf das s.g. van Leer–Verfahren für die Eulergleichung (vgl. (3.6)). Dabei wurde $k = 1$ verwendet. Für $k > 1$ ist die Bestimmung von $\underline{v}^k$ in der Regel zu aufwendig. Dies wird schon aus dem linearen Fall für Systeme deutlich:

Korollar 4.1:

Vor.: $A := \underline{F}'(\underline{u}) = const, \ m, k > 1$ mit (4.1), (4.2),
$\lambda_i, \ \underline{r}_i, \ \underline{l}_i$ die zu A gehörenden Eigenwerte und Rechts- bzw. Linkseigenvektoren,
$a_{i,j}^k, \ 1 \le j \le m, \ 0 \le i \le m$ gemäß (4.23);
$l_j^k \ := \ \langle \ \underline{l}_j, \left(\frac{d}{dx}\right)^k \ \underline{u}_L(0) \ \rangle, \ r_j^k \ := \ \langle \ \underline{l}_j, \left(\frac{d}{dx}\right)^k \ \underline{u}_R(0) \ \rangle;$

Beh.: Das zugehörige LGS zur Bestimmung von $\underline{v}^k$ hat folgende Gestalt:

$$\sum_{s=0}^{m-2} \overline{K}_{s-1,j} \ a_{m-s-1,m-s-1}^k \ \stackrel{!}{=} \ \frac{1}{k!} \left\{ r_j^k \ - \ \sum_{l=2}^{m} \overline{K}_{j,l}^1 \ l_l^k \right\}, \ j = 1,\ldots,m-1;$$

$$a_{m,m}^k \ \stackrel{!}{=} \ r_m^k \qquad \text{mit}$$

$$\overline{K}_{s-1,j} \ := \ \sum_{\substack{l_0,\ldots,l_s=1 \\ l_\nu \neq m-\nu, m-\nu-1 \\ \nu=0,\ldots,s}}^{m} \left[\frac{\lambda_{m-s-2} - \lambda_{m-1-s}}{\lambda_j - \lambda_m} \right] \prod_{n=0}^{s} \left[\frac{\lambda_{l_n} - \lambda_{m-n}}{\lambda_{l_n} - \lambda_{m-1-n}} \right]^{k+1};$$

$$\overline{K}^1_{j,l} := \sum_{\substack{l_0,\ldots,l_{m-2}=1 \\ l_\nu \neq m-\nu, m-\nu-1 \\ \nu=0,\ldots,m-2}}^{m} \left[\frac{\lambda_l - \lambda_1}{\lambda_j - \lambda_m} \right]^{k+1} \prod_{n=0}^{m-2} \left[\frac{\lambda_{l_n} - \lambda_{m-n}}{\lambda_{l_n} - \lambda_{m-1-n}} \right]^{k+1};$$

$$\overline{K}_{-1,j} := \left[\frac{\lambda_{m-1} - \lambda_m}{\lambda_j - \lambda_m} \right]^{k+1}.$$

Auf einen Beweis soll hier aufgrund des enormen Rechenaufwands verzichtet werden. Dies ist ein noch sehr einfacher Fall, denn die i–Wellen von $\underline{v}^0$ sind alle vom Typ C, da die Eigenwerte von $\underline{F}'(\underline{u})$ konstant sind (vgl. La. 2.20 bzw. (4.18.b)). Somit erhält man genau ein LGS. Bei der Eulergleichung ist dagegen die 1–Welle bzw. 3–Welle vom Typ A oder B, je nach Wahl der Anfangsdaten, und die 2–Welle stets vom Typ C (vgl. S. 5.1). Man muß somit also *vier* mögliche LGS in Abhängigkeit von den Anfangsdaten für das RP bestimmen. Im Extremfall können dies bei m Gleichungen des Systems m^m LGS sein. Dies macht die Anwendung der obigen Methode für die Numerik in manchen Fällen für Systeme ($m > 1$) unbrauchbar. Im skalaren Fall ($m = 1$) ist dies anders, da man dort *ohne* Lösen eines LGS die v^k, $k \geq 1$ explizit bestimmen kann. Dies soll im folgenden getan werden.

4.5 Anwendung auf den skalaren Fall

Im folgenden soll nun die asymptotische Entwicklung (4.5) für einen skalaren Erhaltungssatz hergeleitet werden, d.h. das Anfangswertproblem hat die Form (4.1) mit $m = 1$. Für den Fluß gelte zu dem

$$F \in C^2(\mathbb{R}), \quad F'' > 0. \tag{4.26}$$

Da in Kapitel 4.1–4.3 und in La. 4.5 sämtliche Aussagen für $m \geq 1$ hergeleitet wurden, gelten diese insbesondere auch für den skalaren Fall. Auf die Bedingungen entlang den charakteristischen Kurven kann verzichtet werden, da diese nur dazu dienten, zwei Zustände, die durch eben solche Kurven getrennt wurden, zu verbinden und somit ein LGS zu erstellen. Dies ist hier nicht nötig, da man $v^k(\xi)$ für $\xi \in D_0$ bzw. $\xi \in D_1$ wegen La. 4.5 unmittelbar aus den Randbedingungen aus La. 4.4 bestimmen kann. Um dies explizit durchführen zu können, muß aber zunächst $h^k(\xi)$ (vgl. (4.12)) und somit $F^k(\xi)$ (vgl. La. 4.2)) bestimmt werden. Dazu

Korollar 4.2:

Vor.: $v : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ gemäß (4.6) mit $v(\xi, s) \;=\; \sum\limits_{i=0}^{k} v^i(\xi)\; s^i \;+\; O\left(s^{k+1}\right),$

$$f^1(\xi, s) \;\equiv\; 0,\;\; f^{i+1}(\xi, s) \;:=\; \tfrac{\partial}{\partial s}\, F'(v(\xi, s))\, \left(\tfrac{\partial}{\partial s}\right)^i\, v(\xi, s) \;+\; \tfrac{\partial}{\partial s}\, f^i(\xi, s),\;\; i > 1,$$

$(\xi, s) \in \mathbb{R} \times \mathbb{R}_+;$

Beh.: $f^i(\xi, s) \;=\; \sum\limits_{j=1}^{i-1} \binom{i-1}{j}\, \left(\tfrac{\partial}{\partial s}\right)^j\, F'(v(\xi, s))\, \left(\tfrac{\partial}{\partial s}\right)^{i-j}\, v(\xi, s),\;\; i > 2.$

Beweis:

Durch vollständige Induktion über $i \in \mathbb{N}$ folgt unter Berücksichtigung der Definition von $f^{i+1}(\xi, s)$ sowie des Additionstheorems für Binomialkoeffizienten und Anwendung der Produktregel auf $\frac{\partial}{\partial s}\, f^i(\xi, s)$ die Behauptung.

■

Diese Darstellung für die $f^i(\xi, s)$ läßt sich noch weiter transformieren:

Korollar 4.3:

Vor.: $v : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ gemäß (4.6) mit $v(\xi, s) \;=\; \sum\limits_{i=0}^{k} v^i(\xi)\; s^i \;+\; O\left(s^{k+1}\right),$

$$f^1(\xi, s) \;\equiv\; 0,\;\; f^{i+1}(\xi, s) \;:=\; \tfrac{\partial}{\partial s}\, F'(v(\xi, s))\, \left(\tfrac{\partial}{\partial s}\right)^i\, v(\xi, s) \;+\; \tfrac{\partial}{\partial s}\, f^i(\xi, s),\;\; i > 1,$$

$(\xi, s) \in \mathbb{R} \times \mathbb{R}_+,\;\; j_0 \geq 3;$

Beh.: Für $l \in \{1, \ldots, j_0 - 2\}$ gilt:

$$f^{j_0}(\xi, s) \;=$$

$$\sum_{j_1=l+1}^{j_0-1} \cdots \sum_{j_{l+1}=1}^{j_l-1} \left(\tfrac{\partial}{\partial s}\right)^{j_{l+1}}\, F^{(l+1)}(v)\, \prod_{m=0}^{l} \left[\, \binom{j_m - 1}{j_{m+1}}\, \left(\tfrac{\partial}{\partial s}\right)^{j_m - j_{m+1}}\, v \,\right]\Bigg|_{(\xi,s)} +$$

$$\sum_{k=1}^{l} \sum_{j_1=k}^{j_0-1} \cdots \sum_{j_k=1}^{j_{k-1}-1} F^{(k+1)}(v)\, \left(\tfrac{\partial}{\partial s}\right)^{j_k}\, v\, \prod_{m=0}^{k} \left[\, \binom{j_m - 1}{j_{m+1}}\, \left(\tfrac{\partial}{\partial s}\right)^{j_m - j_{m+1}}\, v \,\right]\Bigg|_{(\xi,s)}.$$

Beweis:

Mittels vollständiger Induktion über l folgt die Behauptung. Beim Induktionsschritt beachte man dabei:

$$\left(\tfrac{\partial}{\partial s}\right)^{j_{l+1}}\, F^{(l+1)}(v(\xi, s)) \;=\; \left(\tfrac{\partial}{\partial s}\right)^{j_{l+1}-1}\, \left(F^{(l+2)}(v)\, \tfrac{\partial}{\partial s}\, v\right)\Bigg|_{(\xi,s)} \;=$$

$$\sum_{j_{l+2}=1}^{j_{l+1}-1} \binom{j_{l+1}-1}{j_{l+2}}\, \left(\tfrac{\partial}{\partial s}\right)^{j_{l+2}}\, F^{(l+2)}(v)\, \left(\tfrac{\partial}{\partial s}\right)^{j_{l+1}-j_{l+2}}\, v \;+\; F^{(l+2)}(v)\, \left(\tfrac{\partial}{\partial s}\right)^{j_{l+1}}\, v\Bigg|_{(\xi,s)}.$$

Für die Indizes der einzelnen Summanden gilt ferner:

$$j_i \in \{l + 2 - i, \ldots, j_{i-1} - 1\} \;=\; \{l + 2 - i, \ldots, j_0 - i\},\;\; i = 1, \ldots, l + 1.$$

■

Damit erhält man nun folgende Darstellung für die $f^i(\circ, 0)$:

Folgerung 4.2:

Vor.: $v : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ gemäß (4.6) mit $v(\xi, s) \;=\; \sum_{i=0}^{k} v^i(\xi)\, s^i \;+\; O\left(s^{k+1}\right),$

$f^1(\xi, s) \;\equiv\; 0,\;\; f^{i+1}(\xi, s) \;:=\; \frac{\partial}{\partial s}\, F'(v(\xi, s))\, \left(\frac{\partial}{\partial s}\right)^i\, v(\xi, s) \;+\; \frac{\partial}{\partial s}\, f^i(\xi, s),\;\; i > 1,$

$(\xi, s) \in \mathbb{R} \times \mathbb{R}_+,\;\; j_0 \geq 3;$

Beh.: $f^{j_0}(\xi, 0) \;=$

$$\sum_{k=1}^{j_0-2} \sum_{j_1=k}^{j_0-1} \cdots \sum_{j_k=1}^{j_{k-1}-1} j_0!\; \alpha_k\; F^{(k+1)}(v^0)\; v^{j_k} \prod_{m=0}^{k-1} v^{j_m - j_{m+1}}\Big|_{(\xi)} \;+\; F^{(j_0)}(v^0)\; \left[v^1\right]^{j_0}\Big|_{(\xi)}$$

mit $\alpha_1 \;:=\; \frac{j_0 - j_1}{j_0},\;\; \alpha_k \;:=\; \frac{j_k - j_{k+1}}{j_0} \prod_{m=0}^{k-2} \frac{j_m - j_{m+1}}{j_{m+1}},\;\; k \geq 2.$

Beweis:

Wegen (4.5) gilt (vgl. Beweis zu La. 4.2):

$$(1) \qquad \left(\frac{\partial}{\partial s}\right)^i\, v(\xi, s)\Big|_{s=0} \;=\; i!\; v^i(\xi), \qquad i = 0, \ldots, k.$$

Setze nun in Kor. 4.3 $l := j_0 - 2$. Dann besteht die erste Summe gerade aus einem einzigen Summanden, denn es gilt für die Indizes:

$$j_{i+1} \in \{l + 1 - i, \ldots, j_i - 1\} \text{ für } i = 0, \ldots, l,$$

und somit auch:

$$j_i = j_0 - i.$$

Aus der so gewonnenen Darstellung für $f^i(\xi, s)$ erhält man für $s = 0$ mit (1) die Behauptung.

∎

Mit obiger Dartstellung für die $f^i(\xi, 0)$ lassen sich nun die $h^i(\xi)$ aus (4.12) explizit angeben:

Lemma 4.12:

Vor.: $v : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ gemäß (4.6) mit $v(\xi, s) \;=\; \sum_{i=0}^{k} v^i(\xi)\, s^i \;+\; O\left(s^{k+1}\right),$

$f^1(\xi, s) \;\equiv\; 0,\;\; f^{i+1}(\xi, s) \;:=\; \frac{\partial}{\partial s}\, F'(v(\xi, s))\, \left(\frac{\partial}{\partial s}\right)^i\, v(\xi, s) \;+\; \frac{\partial}{\partial s}\, f^i(\xi, s),\;\; i > 1,$

$(\xi, s) \in \mathbb{R} \times \mathbb{R}_+,\;\; \alpha_k$ gemäß F. 4.2;

Beh.: a) $h^1 \;\equiv\; 0;$

b) $h^2 \;\equiv\; -\frac{1}{2}\, F^{(3)}(v^0)\, \frac{d}{d\xi}\, v^0\, \left[v^1\right]^2 \;-\; F^{(2)}(v^0)\, v^1\, \frac{d}{d\xi} v^1;$

c) Für $j_0 \geq 3$ gilt:

$$h^{j_0} \equiv -\tfrac{1}{j_0!}\, F^{(j_0+1)}(v^0)\, \tfrac{d}{d\xi}\, v^0\, \left[v^1\right]^{j_0} - \tfrac{1}{(j_0-1)!}\, F^{(j_0)}(v^0)\, \left[v^1\right]^{j_0-1}\, \tfrac{d}{d\xi}\, v^1 - \sum_{k=1}^{j_0-2} \sum_{j_1=k}^{j_0-1} \cdots \sum_{j_k=1}^{j_{k-1}-1} \alpha_k\, (a_k + b_k) \prod_{m=0}^{k-1} v^{j_m - j_{m+1}}$$

$$\text{mit} \quad a_k := F^{(k+2)}(v^0)\, \tfrac{d}{d\xi}\, v^0\, v^{j_k},$$

$$b_k := F^{(k+1)}(v^0) \left\{ \tfrac{d}{d\xi}\, v^{j_k} + v^{j_k} \sum_{l=0}^{k-1} \tfrac{d}{d\xi}\, v^{j_l - j_{l+1}}\, \tfrac{1}{v^{j_l - j_{l+1}}} \right\}.$$

Beweis:

Wegen (4.12) gilt mit dem Beweis zu La. 4.2:

$$h^i(\xi) := -\frac{1}{i!}\, \tfrac{d}{d\xi}\, f^i(\xi, 0). \tag{1}$$

Aus (4.13) folgt unmittelbar die Behauptung a). Für $j_0 = 2$ läßt sich h^{j_0} leicht berechnen. Zum Beweis von c) setze in (1) die Darstellung von $f^i(\xi, 0)$ aus F. 4.2 ein. Durch Ableiten erhält man dann die Behauptung.

∎

Nachdem man die rechte Seite aus (4.13) explizit bestimmt hat, sollen nun die $v^k,\ k \geq 0$, ermittelt werden. Dabei ist v^0 wegen La. 4.5 durch die Lösung $V(\circ, u_L(0), u_R(0))$ des RPs gemäß Def. 4.2 festgelegt. Diese existiert immer, da F als konvex vorausgesetzt wurde (vgl. (4.26)), also ist sie insbesondere unabhängig von $\mid u_L(0) - u_R(0) \mid$. Damit hat v^0 folgende Gestalt:

Lemma 4.13:

Beh.: a) $u_L(0) > u_R(0) \implies v^0(\xi) = \left\{ \begin{matrix} u_L(0),\ \xi < \sigma^0 \\ u_R(0),\ \xi > \sigma^0 \end{matrix} \right\},\ \sigma^0 := \dfrac{F(u_R(0)) - F(u_L(0))}{u_R(0) - u_L(0)};$

b) $u_L(0) = u_R(0) \implies v^0(\xi) = u_L(0) = u_R(0),\ \sigma^0 := F'(u_L(0));$

c) $u_L(0) < u_R(0) \implies v^0(\xi) = \left\{ \begin{matrix} u_L(0), & \xi < \underline{\sigma}^0 \\ (F')^{[-1]}(\xi), & \underline{\sigma}^0 < \xi < \overline{\sigma}^0 \\ u_R(0), & \overline{\sigma}^0 < \xi \end{matrix} \right\}$ mit

$$\underline{\sigma}^0 := F'(u_L(0)),\ \overline{\sigma}^0 := F,(u_R(0)).$$

Für einen Beweis der Lösung des RPs im skalaren Fall sehe man in [Smo], Kap. 16.D. Der Fall b) ist zwar trivial, aber für $k \geq 1$ gilt im allgemeinen $v^k \not\equiv const$, da in der Regel gilt: $\mid \frac{d}{dx}\, u_L(0) - \frac{d}{dx}\, u_R(0) \mid \neq 0$. Deshalb muß dieser Fall berücksichtigt werden. Führe nun folgende Bezeichnungen ein:

$$D_L := \left\{ \begin{matrix} (-\infty, \sigma^0),\ u_L(0) \geq u_R(0) \\ (-\infty, \underline{\sigma}^0),\ u_L(0) < u_R(0) \end{matrix} \right\},\ D_R := \left\{ \begin{matrix} (\sigma^0, +\infty),\ u_L(0) \geq u_R(0) \\ (\overline{\sigma}^0, +\infty),\ u_L(0) < u_R(0) \end{matrix} \right\},$$

$$R := (\underline{\sigma}^0, \overline{\sigma}^0).$$

Dann lassen sich die $v^k, k \geq 1$, analog zu Kap. 4.4 berechnen, in dem man die Lösung zum einen auf D_L bzw. D_R und zum anderen auf R bestimmt. Dazu:

Korollar 4.4:

Vor.: $\xi \in D_L$ oder $\xi \in D_R$, $k \geq 1$, $h^k(\xi) \stackrel{!}{=} \sum_{j=0}^{k-1} b_j^k \, \xi^j$ gemäß (4.13.a);

Beh.: Die Lösung von (4.12.b) ist gegeben durch

$$v^k(\xi) \;=\; [\,\xi - F'(v^0(\xi))\,]^k \; c_{L,R}^k \;+\; p_{L,R}^k(\xi)$$

$$\text{mit } p_{L,R}^k(\xi) := \sum_{j=0}^{k-1} a_{k,j}^{L,R} \, \xi^j, \;\; c_{L,R}^k := \frac{1}{k!} \, \left(\tfrac{d}{dx}\right)^k \; u_{L,R}(0),$$

$$a_{k,j}^{L,R} := \left\{ \begin{array}{ll} b_{k-1}^k & , j = k-1 \\ \frac{1}{k-j} \; \left[\, b_j^k - a_{k,j+1}^{L,R} \; (j+1) \; F'(u_{L,R}(0)) \,\right] & , j = k-2, \ldots, 0 \end{array} \right\}.$$

Beweis:

Sei $\xi \in D_L$ bzw. $\xi \in D_R$. Dann ist $v^0(\xi) \;=\; u_{L,R}(0) \;=\; const.$ Somit gilt wegen (4.12)

$$(1) \qquad [\, F'(u_{L,R}(0)) - \xi \,] \; \tfrac{d}{d\xi} \, v^k(\xi) \;+\; k \; v^k(\xi) \;=\; h^k(\xi),$$

$$(2) \qquad v^i \in \mathbb{P}_{k-1} \text{ für } i = 0, \ldots, k-1.$$

Damit ist $h^k \in \mathbb{P}_{k-1}$ aufgrund von (4.13), und somit existiert die Darstellung aus der Voraussetzung für h^k. Man kann nun leicht zeigen, daß $F'(u_{L,R}(0)) \neq \xi$ gilt, in dem man die Intervallgrenzen und die Definition der charakteristischen Geschwindigkeiten σ^0, $\underline{\sigma}^0$, $\overline{\sigma}^0$ aus La. 4.13 beachtet. Somit hat (1) die gleiche Struktur wie die Differentialgleichung (3) aus dem Beweis zu La. 4.6. Man kann nun völlig analog dazu die Lösung von (1) bestimmen als:

$$(3) \qquad v^k(\xi) \;=\; \left[\, \xi - F'(v^0(\xi)) \,\right]^k \; c_{L,R}^k \;+\; p_{L,R}^k.$$

Dabei sind die $c_{L,R}^k \in \mathbb{R}$ beliebig. Damit gilt sofort:

$$(4) \qquad \left(\tfrac{d}{d\xi}\right)^k \; v^k(\xi) \;\stackrel{!}{=}\; k! \; c_{L,R}^k.$$

Mit La. 4.4 folgt dann:

$$c_{L,R}^k \;\stackrel{!}{=}\; \frac{1}{k!} \; \left(\tfrac{d}{d\xi}\right)^k \; u_{L,R}(0).$$

∎

Aus obigem Lemma folgt unter Anwendung der Binomischen Formel unmittelbar folgende polynomiale Darstellung für v^k in D_L bzw. D_R:

Folgerung 4.3:

Vor.: $\xi \in D_L$ bzw. $\xi \in D_R$, $k \geq 1$, Bezeichnungen gemäß Kor. 4.4;

Beh.: $v^k(\xi) \; = \; \sum\limits_{j=0}^{k} v_j^{L,R} \, \xi^j$ mit

$$v_j^{L,R} \;:=\; \left\{ \begin{array}{ll} \binom{i}{j} \; [\, -F'(u_{L,R}(0))\,]^{k-j} \;\; c_{L,R}^k \; + \; a_{k,j}^{L,R} & , j = 0, \ldots, k-1 \\ c_{L,R}^k & , j = k \end{array} \right\}.$$

Im Gegensatz zu den Systemen, d.h. $m > 1$, ist der Verdünnungsbereich R von v^0 hier für v^k, $k \geq 1$, trivial, denn:

Lemma 4.14:

Beh.: Für $k \geq 1$ und $\xi \in R$ gilt: $\quad v^k(\xi) \; = \; 0.$

Beweis:

Sei $\xi \in R$. Aus (4.12.b) folgt insbesondere

$$(1) \qquad \left[\, F'(v^0(\xi)) - \xi \,\right] \; \tfrac{d}{d\xi} \, v^k(\xi) \; + \; \left[\, k + F''(v^0(\xi)) \, \tfrac{d}{d\xi} \, v^0(\xi) \,\right] \; v^k(\xi) \; \stackrel{!}{=} \; h^k(\xi).$$

Des weiteren gilt mit La. 4.13.c:

$$(2) \qquad F'(v^0(\xi)) \; = \; \xi,$$

$$(3) \qquad \tfrac{d}{d\xi} \, v^0(\xi) \; \stackrel{(2)}{=} \; \tfrac{d}{d\xi} \, (F')^{[-1]} \, (\xi) \; = \; \frac{1}{F''\Big((F')^{[-1]} \, (\xi) \Big)} \; \stackrel{(2)}{=} \; \frac{1}{F''(v^0(\xi))}.$$

Setze nun (2) und (3) in (1) ein. Dann erhält man

$$(4) \qquad v^k(\xi) \; = \; \frac{1}{k+1} \, h^k(\xi).$$

Wegen La. 4.12.a folgt daraus unmittelbar die Behauptung für $k = 1$. Seien nun $k > 1$ und $v^i \equiv 0$ für $i = 1, \ldots, k-1$. Mit La. 4.13.b,c erhält man dann sofort:

$$h^k(\xi) \; = \; 0 \text{ für } \xi \in R$$

und wegen (4) somit auch

$$v^k(\xi) \; = \; 0 \text{ für } \xi \in R.$$

Mit dem Induktionsprinzip folgt nun die Behauptung.

∎

Mit F. 4.3 und La. 4.14 läßt sich somit v^k explizit angeben:

Lemma 4.15:

Vor.: $k \geq 1$, Bezeichnungen gemäß Kor. 4.4;

$$\text{Beh.: } v^k(\xi) \;=\; \left\{ \begin{array}{ll} [\,\xi - F'(u_L(0))\,]^k \; \frac{1}{k!} \; \left(\frac{d}{d\xi}\right)^k \; u_L(0) \;+\; p_L^k(\xi) & ,\xi \in D_L \\ 0 & ,\xi \in R \\ [\,\xi - F'(u_R(0))\,]^k \; \frac{1}{k!} \; \left(\frac{d}{d\xi}\right)^k \; u_R(0) \;+\; p_R^k(\xi) & ,\xi \in D_R \end{array} \right\}.$$

Da man in Kap. 2.1.4 nur an der Darstellung für v^k an der Stelle $\xi = 0$ interessiert ist, zusätzlich noch

Folgerung 4.4:

Vor.: $k \geq 1$, Bezeichnungen gemäß Kor. 4.4;

$$\text{Beh.: } v^k(0) \;=\; \left\{ \begin{array}{ll} (-1)^k \; F'(u_L(0)) \; \frac{1}{k!} \; \left(\frac{d}{d\xi}\right)^k \; u_L(0) \;+\; p_L^k(0) & ,0 \in D_L \\ 0 & ,0 \in R \\ (-1)^k \; F'(u_R(0)) \; \frac{1}{k!} \; \left(\frac{d}{d\xi}\right)^k \; u_R(0) \;+\; p_R^k(0) & ,0 \in D_R \end{array} \right\}.$$

Im skalaren Fall liegt also die Schwierigkeit nicht im Verdünnungsbereich R des RPs, sonderen einzig und allein in der Bestimmung der $h^k(\xi)$ für $\xi \in D_L$ oder $\xi \in D_R$. Daß dies sehr aufwendig sein kann, soll an dem folgenden Beispiel demonstriert werden:

Lemma 4.16:

Vor.: $F(u) \;:=\; \frac{1}{2}\,u^2,\; k \geq 1,\; \xi \in D_L$ oder $\xi \in D_R$;

Beh.: a) $h^k(\xi) \;=\; -\sum\limits_{j=1}^{k-1} \frac{d}{d\xi}\, v^j(\xi)\, v^{k-j}(\xi);$

$$\text{b) } v^j(\xi) \;=\; \sum_{i=0}^{j} d_i^j\, \xi^j,\; j = 0, \ldots, k-1 \;\Longrightarrow\; h^k(\xi) \;=\; \sum_{i=0}^{k-1} \left(-\sum_{j=1}^{k-1} c_i^j \right) \xi^k$$

$$\text{mit } c_i^j \;:=\; \left\{ \begin{array}{ll} \sum\limits_{l=\max\{0,i-k+j\}}^{\min\{i,j-1\}} d_{l+1}^j\, (l+1)\, d_{i-l}^{k-j} & ,j > \lfloor\frac{k+1}{2}\rfloor \;+\; 1 \\ \sum\limits_{l=\max\{0,i-j+1\}}^{\min\{i,k-j\}} d_l^{k-j}\, (i-l+1)\, d_{i+1-l}^j & ,j \leq \lfloor\frac{k+1}{2}\rfloor \end{array} \right\}.$$

Beweis:

Aus La. 4.12 folgt a) und daraus die Behauptung b) durch Umsortieren der Potenzen von ξ.

■

Da die Bestimmung von $h^k(\xi)$ als Polynom in ξ schon im skalaren Fall mühsam ist, dürfte dies bei Systemen für $k > 1$ noch um einiges aufwendiger sein.

Kapitel 5: Schock–Interaktion

In Kapitel 1 wurde das Oszillationsverhalten der ENO–Splines $Q^N(\circ, v)$ ausführlich untersucht. Es wurde unter anderem gezeigt, daß diese die ENO–Bedingung gemäß Def. 1.4 erfüllen, falls für die zu interpolierende Funktion gilt: $v \in SCL(N+1, R)$ (vgl. Kap. 1). Dabei wurde allerdings vorausgesetzt, daß in einem Intervall $[x_j, x_{j+1})$ höchstens eine Sprungstelle von v liegt. Es stellt sich nun die Frage, wie sich das ENO–Polynom verhält, wenn zwei oder mehr Unstetigkeitsstellen in *einem* Intervall liegen. Dies ist deshalb wichtig, weil bei den ENO–Verfahren dieser Fall eintreten kann, wenn die exakte Lösung aus zwei Schocks besteht, die zu einem bestimmten Zeitpunkt aufeinandertreffen. In diesem Zusammenhang stellt sich ferner die Frage, in wie weit die Wahl des numerischen Flusses (vgl. Kap. 2.1.4 bzw. Kap. 2.2.4) die Qualität der Approximation mittels der ENO–Verfahren beeinflußt. Es sollen dazu im folgenden Interaktionsprobleme konstruiert werden, zum einen für die Burgersgleichung im skalaren Fall und zum anderen für die Eulergleichung für Systeme.

5.1 Skalarer Fall

Hier soll die Entropielösung des Anfangswertproblems

(5.1) a) $$u_t(x,t) \;+\; u(x,t)\, u_x(x,t) \;=\; 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+;$$

b) $$u(x,0) \;=\; \left\{ \begin{array}{c l} 3 & , x \le 0.1 \\ -2 & , 0.1 < x < 0.5 \\ 5 & , 0.5 < x < 0.9 \\ -5 & , 0.9 < x \end{array} \right\}$$

sowie die charakteristischen Kurven bestimmt werden. Dazu geht man analog zum Beweis zu S. 2.5 vor, in dem man die Bereiche, in denen die Lösung von (5.1) glatt ist, sowie die Diskontinuitätskurven bzw. Verdünnungswellen, die diese voneinander trennen, ermittelt. Falls zwei charakteristische Kurven ineinanderlaufen, ist ein neues Riemann–Problem (RP) zu lösen. Dieser Fall kann wegen dem Satz von Schaeffer (S. 2.4) nur endlich oft eintreten. Beachte dabei, daß entlang einer Unstetigkeitskurve $x = \varphi(t)$ die Sprungbedingung erfüllt sein muß ,d.h. es muß gelten:

(5.2) $$\begin{aligned} s\;[\,u(\varphi(t)-0,t) - u(\varphi(t)+0,t)\,] \;&\overset{!}{=}\; [\,F(u(\varphi(t)-0,t)) - F(u(\varphi(t)+0,t))\,]\,, \\ s \;&\overset{!}{=}\; \tfrac{d}{dt}\varphi. \end{aligned}$$

Wegen Kor. 2.3 lassen sich die Teillösungen aufgrund von (5.2) zu einer schwachen Lösung zusammensetzen. Diese ist insbesondere Entropielösung, falls die Entropiebedingung erfüllt ist.

Da $u(\circ, 0)$ stückweise konstant ist, sind also zunächst drei gewöhnliche RPs zu lösen. Wegen La. 4.13 haben deren Lösungen die folgende Gestalt:

$$V(\tfrac{x}{t}, 3, -2) \;=\; \left\{ \begin{array}{c l} 3 & , x < s_1(t) \\ -2 & , x > s_1(t) \end{array} \right\}, \; V(\tfrac{x}{t}, 5, -5) \;=\; \left\{ \begin{array}{c l} 5 & , x < s_4(t) \\ -5 & , x > s_4(t) \end{array} \right\},$$

$$V(\tfrac{x}{t}, -2, 5) \;=\; \left\{ \begin{array}{c l} -2 & , x < s_2(t) \\ \frac{x-0.5}{t} & , s_2(t) < x < s_3(t) \\ 5 & , x > s_3(t) \end{array} \right\} \;\; \text{mit}$$

(5.3) $s_1(t) = 0.5\,t + 0.1,\ s_2(t) = -2\,t + 0.5,\ s_3(t) = 5\,t + 0.5,\ s_4(t) = 0.9.$

Überprüfe nun, ob die charkteristischen Kurven ineinanderlaufen:

(5.4)
a) $s_1(t) = s_2(t) \iff t_2 = 0.16\ (x_2 := s_1(t_2) = 0.18);$
b) $s_3(t) = s_4(t) \iff t_1 = 0.08\ (x_1 := s_3(t_1) = 0.09);$
c) $s_2(t) < s_3(t)$ für $t \geq 0.$

Somit ergibt sich die erste Teillösung durch Zusammensetzen der einzelnen Lösungen der RPs gemäß (5.3)

$$u_0(x,t) = \left\{\begin{array}{cl} 3 & , x < s_1(t) \\ -2 & , s_1(t) < x < s_2(t) \\ \frac{x-0.5}{t} & , s_2(t) < x < s_3(t) \\ 5 & , s_3(t) < x < s_4(t) \\ -5 & , s_4(t) < x \end{array}\right\} \text{ für } (x,t) \in \mathbb{R} \times [0, t_1). \tag{5.5}$$

Zum Zeitpunkt t_1 hat man nun in (x_1, t_1) ein neues RP zu lösen, dessen Lösung sich aus La. 4.13 wie folgt bestimmt:

$$V(\tfrac{x}{t}, \tfrac{x-0.5}{t}, -5) = \left\{\begin{array}{cl} \frac{x-0.5}{t} & , x < s_5(t) \\ -5 & , s_5(t) < x \end{array}\right\}.$$

Dabei bestimmt sich $s_5(t)$ folgendermaßen:
Wegen (5.4.c) gilt

$$x - 0.5 > -5\,t \text{ für } (x,t) \in \mathbb{R} \times \mathbb{R}_+.$$

Somit handelt es sich bei $s_5(t)$ um eine Schockkurve, entlang der — wegen (5.2) — gelten muß

$$\frac{d}{dt}\, s_5(t) \stackrel{!}{=} 0.5 \left\{ \frac{s_5(t) - 0.5}{t} - 5 \right\},\ t > t_1;$$

$$s_5(t_1) \stackrel{!}{=} x_1.$$

Dies ist eine nichtlineare Differentialgleichung 1. Ordnung, deren Lösung sich gemäß [Wa] eindeutig bestimmt als

$$s_5(t) = -5\,t + \sqrt{8\,t} + 0.5,\ t \geq t_1. \tag{5.6}$$

Überprüfe nun, ob diese Kurve in eine der übrigen charakteristischen Kurven s_1, s_2 hineinläuft. Zeige dazu:

$$s_5(t) > s_2(t) > s_1(t) \text{ für } t \in [t_1, t_2). \tag{5.7}$$

Somit erhält man die zweite Teillösung:

$$u_1(x,t) = \left\{\begin{array}{cl} 3 & , x < s_1(t) \\ -2 & , s_1(t) < x < s_2(t) \\ \frac{x-0.5}{t} & , s_2(t) < x < s_5(t) \\ -5 & , s_5(t) < x \end{array}\right\} \text{ für } (x,t) \in \mathbb{R} \times [t_1, t_2). \tag{5.8}$$

Wegen (5.4.a) und (5.7) hat man zum Zeitpunkt t_2 in (x_2, t_2) ein weiteres RP zu lösen, dessen Lösung sich mit La. 4.13 bestimmt als

$$V(\tfrac{x}{t}, 3, \tfrac{x-0.5}{t}) \;=\; \left\{ \begin{array}{ll} 3 & , x < s_6(t) \\ \frac{x-0.5}{t} & , s_6(t) < x \end{array} \right\}.$$

Wegen (5.7) gilt

$$3 \;>\; \frac{x-0.5}{t} \text{ für } t > t_2.$$

Somit handelt es sich bei s_6 wieder um einen Schock, für den aufgrund von (5.2) gelten muß:

$$\begin{aligned} \tfrac{d}{dt}\, s_6(t) \;&\overset{!}{=}\; 0.5\left(\tfrac{s_6(t)-0.5}{t} \;+\; 3 \right),\; t > t_2; \\ s_6(t_2) \;&\overset{!}{=}\; x_2. \end{aligned}$$

Damit erhält man für die neue charakteristische Kurve:

$$s_6(t) \;=\; 3\,t \;-\; 2\sqrt{t} \;+\; 0.5,\; t \geq t_2. \tag{5.9}$$

Diese kollidiert mit der anderen charakteristischen Kurve s_5, denn:

$$s_6(t) \;=\; s_5(t) \iff t_3 \;=\; \left(\frac{\sqrt{2} \;+\; 1}{4} \right)^2,\; (\; x_3 := s_5(t_3)\;). \tag{5.10}$$

Die dritte Teillösung setzt sich also wie folgt zusammen:

$$u_2(x,t) \;=\; \left\{ \begin{array}{ll} 3 & , x < s_6(t) \\ \frac{x-0.5}{t} & , s_6(t) < x < s_5(t) \\ -5 & , s_5(t) < x \end{array} \right\} \text{ für } (x,t) \in \mathbb{R} \times [t_2, t_3). \tag{5.11}$$

Zum Zeitpunkt t_3 ist noch ein weiteres RP in (x_3, t_3) zu lösen. Da aber für $t = t_3$ nur noch zwei konstante Zustände existieren, ist dies die letzte Teillösung, die noch zu bestimmen ist. Somit gilt mittels La. 4.13

$$\begin{aligned} u_3(x,t) \;=\; V(\tfrac{x}{t}, 3, -5) \;&=\; \left\{ \begin{array}{ll} 3 & , x < s_7(t) \\ -5 & , s_7(t) < x \end{array} \right\},\; t \geq t_3 \text{ mit} \\ s_7(t) \;&\overset{(5.2)}{=}\; -t \;+\; 0.75. \end{aligned} \tag{5.12}$$

Die gesuchte Entropieösung sieht also wie folgt aus:

Lemma 5.1:

Vor.: $u_i(\circ,\circ), i = 0, \ldots, 3$ gemäß (5.5)–(5.12), $t_0 := 0,\; t_4 := +\infty$;
Beh.: Die (eindeutige) Entropielösung von (5.1) ist gegeben durch:
$u(x,t) \;=\; u_i(x,t),\; (x,t) \in \mathbb{R} \times [t_i, t_{i+1}),\; i = 0, \ldots, 3.$

Beweis:

Da sich die Teillösungen aus den Entropielösungen von endlich vielen RPen zusammensetzen, ist $u(x,t)$ die gesuchte Entropielösung. Wegen dem Satz von Oleinik (s. [Smo]) ist diese eindeutig bestimmt.

■

Die Lösungsstruktur, insbesondere der Verlauf der charakteristischen Kurven, läßt sich sehr gut aus der 2 D–Projektion in Abb. 16 entnehmen.
Es sei noch vermerkt, daß obiges Beispiel insbesondere die Gültigkeit der Aussagen aus La. 2.16 sowie S. 2.4 und S. 2.5 bestätigt.

5.2 Systeme

Hier sollen nun ganz allgemein Interaktionsprobleme für die Eulergleichung hergeleitet werden. Dazu sei folgendes Anfangswertproblem gegeben:

(5.13) a) $\quad \underline{u}_t(x,t) \;+\; (\underline{F}(\underline{u}(x,t)))_x \;=\; 0,\ (x,t) \in \mathbb{R} \times \mathbb{R}_+,\ \underline{F},\ \underline{u}$ gemäß (3.6);

b) $$\underline{u}(x,0) \;=\; \left\{ \begin{array}{ll} \underline{u}_L & , x < x_0 \\ \underline{u}_M & , x_0 < x < x_1 \\ \underline{u}_R & , x_1 < x \end{array} \right\},\ \underline{u}_L, \underline{u}_M, \underline{u}_R \in \mathbb{R}^3.$$

Ziel ist es nun, $\underline{u}_l$, $\underline{u}_M$, $\underline{u}_R$ so zu bestimmen, daß die Lösung von (5.13) aus genau zwei 3–Schocks (vgl. (4.18), (4.20)) besteht, die ineinanderlaufen. Dazu sollen zunächst notwendige und hinreichende Bedingungen hergeleitet werden, wann die Lösung des RPs

(5.14) a) $\quad \underline{u}_t(x,t) \;+\; (\underline{F}(\underline{u}(x,t)))_x \;=\; 0,\ (x,t) \in \mathbb{R} \times \mathbb{R}_+;$

b) $$\underline{u}(x,0) \;=\; \left\{ \begin{array}{ll} \underline{w}_L & , x < 0 \\ \underline{w}_R & , x > 0 \end{array} \right\},\ \underline{w}_L, \underline{w}_R \in \mathbb{R}^3$$

aus genau einem 3–Schock besteht. Im folgenden wird dabei nicht in $\underline{u} = (\rho, m, E)$ sondern in $\underline{w} = (\rho, P, v)$ (vgl. (3.6)) gerechnet. Mittels der Zustandsgleichungen lassen sich diese ineinander überführen. Führe nun folgende Bezeichnungen ein (vgl. [Smo]):

(5.15) a) $$\tau := \frac{\gamma - 1}{2\,\gamma},\ \beta := \frac{\gamma + 1}{\gamma - 1},\ \gamma > 1;$$

b) $$A := \frac{\rho_R}{\rho_L},\ B := \frac{P_R}{P_L},\ C := \frac{v_R - v_L}{c_L};$$

c) $$T_x^{(1)}\,\underline{w} := \left(f_1(x)\,w_1, e^{-x}\,w_2, w_3 + \left(\frac{\gamma\,w_2}{w_1} \right)^{\frac{1}{2}} h_1(x) \right)^T;$$

$$T_x^{(2)}\,\underline{w} := (\; e^x\,w_1, w_2, w_3\;)^T;$$

$$T_x^{(3)}\ \underline{w} := \left(\frac{w_1}{f_1(x)}, e^x\ w_2, w_3 + \left(\frac{\gamma\ w_2}{w_1}\ f_1(x)\ e^x \right)^{\frac{1}{2}}\ h_1(x) \right)^T;$$

$$h_1(x) := \left\{ \begin{array}{ll} \frac{2}{\gamma-1}\ (\ 1 - e^{-\tau\ x}\ ) & , x \geq 0 \\ \frac{2\ \sqrt{\tau}\ (\ 1-e^{-x}\ )}{(\ \gamma-1\ )\ \sqrt{1+\beta\ e^{-x}}} & , x \leq 0 \end{array} \right\};$$

$$f_1(x) := \left\{ \begin{array}{ll} e^{\frac{-x}{\gamma}} & , x \geq 0 \\ \frac{\beta+e^x}{1+\beta\ e^x} & , x \leq 0 \end{array} \right\};$$

d) $\lambda_1(\underline{w}) = v - c,\ \lambda_2(\underline{w}) = v,\ \lambda_3(\underline{w}) = v + c$ Eigenwerte von $\underline{F}'(\underline{u})$.

Diese sind mit Absicht an diejenigen aus [Smo] angelehnt, da einige Aussagen bzgl. der Struktur der Lösung des RPs (5.14) im folgenden noch des öfteren benötigt werden. Zunächst aber zur Entropielösung des RPs:

Satz 5.1:

Vor.: $\underline{w}_L,\ \underline{w}_R \in \mathbb{R}^3 \setminus \{(x,y,z)\,|\,x,y>0,\ z \in \mathbb{R}\},\ \xi := \frac{x}{t},\ (x,t) \in \mathbb{R} \times \mathbb{R}_+;$

Beh.: 1) Die Lösung des RPs (5.14) existiert eindeutig, genau dann wenn gilt

$$v_R - v_L\ <\ \frac{2}{\gamma-1}\ (\ c_L + c_R\ );$$

2) Die 1–Welle ist vom Typ A (vgl. (4.18)), genau dann wenn gilt

$$\sqrt{\frac{B}{A}}\ h_1(\ln\ B)\ <\ C\ <\ \frac{2}{\gamma-1}\ \left(1 + \sqrt{\frac{B}{A}} \right),$$

sonst vom Typ B;

3) Die 1–Welle ist vom Typ A (vgl. (4.18)), genau dann wenn gilt

$$h_1(-\ln\ B)\ <\ C\ <\ \frac{2}{\gamma-1}\ \left(1 + \sqrt{\frac{B}{A}} \right),$$

sonst vom Typ B;

4) Für die Zwischenzustände (vgl. (4.19)) gilt:

$$\underline{w}_1 = T_{x_1}^{(1)}\ \underline{w}_L,\ \underline{w}_2 = T_{x_2}^{(2)}\ \underline{w}_1,\ \underline{w}_3 = T_{x_3}^{(3)}\ \underline{w}_2$$

wobei (x_1, x_2, x_3) gegeben ist durch die Lösung des nichtlinearen Gleichungssystems:

$$\text{a)}\qquad h_1(x_1)\ +\ \sqrt{\frac{B}{A}}\,h_1(x_1\ +\ \ln\ B)\ =\ C,$$
$$\text{b)}\qquad x_3\ =\ \ln\ B\ +\ x_1,$$
$$\text{c)}\qquad x_2\ =\ \ln\left(A\,\frac{f_1(x_3)}{f_1(x_1)}\right);$$

5) Falls die 1,(3)–Welle vom Typ A ist, gilt für $\lambda_{1,(3)}(\underline{w}_{L,(2)}) < \xi < \lambda_{1,(3)}(\underline{w}_{1,(R)})$:

$$\rho(x,t)\ =\ \rho_{R,(L)}\left(\frac{v_{R,(L)}-\xi}{\beta\ c_{R,(L)}}-\frac{2}{\gamma+1}\right)^{\dfrac{1}{\tau\ \gamma}},$$

$$P(x,t)\ =\ P_{R,(L)}\left(\frac{v_{R,(L)}-\xi}{\beta\ c_{R,(L)}}-\frac{2}{\gamma+1}\right)^{\dfrac{1}{\tau}},$$

$$v(x,t)\ =\ c_{R,(L)}\left(\frac{v_{R,(L)}-\xi}{\beta\ c_{R,(L)}}-\frac{2}{\gamma+1}\right)\ +\ \xi;$$

6) Ist die 1,(3)–Welle vom Typ B, dann gilt für die charakteristische Geschwindigkeit:

$$s_1\ =\ \left\{\begin{array}{ll}\dfrac{\rho_1\ v_1\ -\ \rho_L\ v_L}{\rho_1\ -\ \rho_L} & ,\rho_L\neq\rho_1\\ \lambda_1(\underline{w}_L) & ,\rho_L=\rho_1\end{array}\right\},\ s_3\ =\ \left\{\begin{array}{ll}\dfrac{\rho_R\ v_R\ -\ \rho_2\ v_2}{\rho_R\ -\ \rho_2} & ,\rho_R\neq\rho_2\\ \lambda_3(\underline{w}_R) & ,\rho_R=\rho_2\end{array}\right\};$$

7) Die 2–Welle ist vom Typ C mit charakteristischer Geschwindigkeit

$$s_2 := \lambda_2(\underline{w}_1) = \lambda_2(\underline{w}_2).$$

Der Beweis findet sich in [Smo], S. 346,ff.
In vier Schritten lassen sich nun die gewünschten Kriterien herleiten. Zunächst:

Lemma 5.2:

Vor.: $\underline{w}_L,\ \underline{w}_R \in \mathbb{R}^3\setminus\{(x,y,z)\,|\,x,y>0,\ z\in\mathbb{R}\}\,,\ \xi := \frac{x}{t},\ (x,t)\in\mathbb{R}\times\mathbb{R}_+;$

Beh.: Die Lösung des RPs (5.14) besteht aus einem einzigen 3–Schock, genau dann wenn gilt:

$$1)\qquad x_1\ =\ x_2\ =\ 0;$$

$$2)\qquad (x_1,x_2,x_3)\text{ ist Lösung von (S.5.2.4);}$$

$$3)\qquad C\ <\ \frac{2}{\gamma-1}\left(1+\sqrt{\frac{B}{A}}\right);$$

$$4)\qquad C\ <\ \min\left\{\sqrt{\frac{B}{A}}\,h_1(\ln\ B),h_1(-\ln\ B)\right\}.$$

Beweis:

Wie man leicht sieht, ist (L.5.2.3) äquivalent zu (S.5.1.1) (vgl. (5.15.b)), d.h. die Lösung des RPs existiert, also auch (x_1, x_2, x_3) gemäß (S.5.1.4). Wegen (L.5.2.4) kann die 1– bzw. 3–Welle nicht vom Typ A sein, somit besteht die Lösung also aus einem 1–Schock und einem 3–Schock sowie einer 2–Kontaktunstetigkeit. Damit die Zustände links und rechts der 1– bzw. 2–Welle identisch sind, muß wegen (S.5.1.4) gelten:

$$\underline{w}_L \overset{!}{=} T^{(1)}_{x_1}\, \underline{w}_L \text{ und } T^{(2)}_{x_2}\, T^{(1)}_{x_1}\, \underline{w}_L \overset{!}{=} T^{(1)}_{x_2}\, \underline{w}_L = \underline{w}_L.$$

Dies ist aufgrund von (5.15.c) genau dann der Fall, wenn gilt: $x_1 = x_2 = 0$.

■

Daraus lassen sich nun weitere äquivalente Kriterien herleiten:

Lemma 5.3:

Vor.: $\underline{w}_L,\ \underline{w}_R \in \mathbb{R}^3 \setminus \{(x,y,z)\,|\,x,y>0,\ z\in\mathbb{R}\}\,,\ \xi := \frac{x}{t},\ (x,t)\in\mathbb{R}\times\mathbb{R}_+,$
Bedingung (L.5.2.3) sei erfüllt;

Beh.: Die Bedingungen (L.5.2.1) und (L.5.2.2) sind äquivalent zu:

$$1)\qquad C = \sqrt{\frac{B}{A}}\, h_1(\ln\ B);$$

$$2)\qquad A = \frac{1}{f_1(\ln\ B)}.$$

Beweis:

Wegen (L.5.2.3) existiert die Lösung des RPs. Somit ist (x_1, x_2, x_3) eindeutig bestimmt. Dann sind die Bedingungen (L.5.2.1) und (L.5.2.2) äquivalent mit

$$h_1(0) + \sqrt{\frac{B}{A}}\, h_1(\ln\ B) = C,\ x_3 = \ln\ B,\ 0 = \ln\left(A\, \frac{f_1(\ln\ B)}{f_1(0)} \right).$$

Diese Gleichungen lassen sich aber zu den Bedingungen (L.5.3.1) und (L.5.3.2) umschreiben.

■

Des weiteren gilt:

Lemma 5.4:

Vor.: $\underline{w}_L,\ \underline{w}_R \in \mathbb{R}^3 \setminus \{(x,y,z)\,|\,x,y>0,\ z\in\mathbb{R}\}\,,\ \xi := \frac{x}{t},\ (x,t)\in\mathbb{R}\times\mathbb{R}_+,$
Bedingung (L.5.3.1) sei erfüllt;

Beh.: Die Bedingung (L.5.2.4) ist genau dann erfüllt, wenn gilt: $0 < B \leq 1$.

Beweis:

Wegen (5.15.a) gilt: $\gamma > 0$. Daraus folgt:

$$(1) \qquad \tau \geq 0;$$

$$(2) \qquad B > 1 \iff h_1(-\ln\ B) \leq 0 \leq h_1(\ln\ B).$$

Mit (1) läßt sich dann auch zeigen:

$$(3) \qquad 0 < B \leq 1 \iff h_1(\ln\ B) \leq 0 \leq h_1(-\ln\ B).$$

Somit gilt wegen (L.5.2.4):

$$(4) \qquad B > 1 \iff C < h_1(-\ln\ B);$$

$$(5) \qquad 0 < B \leq 1 \iff C < \sqrt{\frac{B}{A}}\, h_1(\ln\ B).$$

Beachte dabei, daß h_1 streng monoton steigend ist (s. [Smo], S. 355).
Mit (L.5.3.1) gilt nun:

$$B > 1 \implies C \overset{(4)}{<} h_1(-\ln\ B) \overset{(2)}{<} \sqrt{\frac{B}{A}}\, h_1(\ln\ B) \overset{(L.5.3.1)}{=} C.$$

Dies ist ein Widerspruch. Somit kann unter diesen Bedingungen nicht $B > 1$ gelten.
Im anderen Fall gilt:

$$(7) \qquad 0 < B \leq 1 \overset{(5)}{\iff} \sqrt{\frac{B}{A}}\, h_1(\ln\ B) \overset{(3)}{\leq} h_1(-\ln\ B).$$

Daraus folgt die Behauptung.

■

Aus La. 5.2, La. 5.3 und La. 5.4 erhält man nun die gewünschten Kriterien:

Satz 5.2:

Vor.: $\underline{w}_L,\ \underline{w}_R \in \mathbb{R}^3 \setminus \{(x,y,z)\,|\,x,y>0,\ z\in\mathbb{R}\}$, $\xi := \frac{x}{t}$, $(x,t) \in \mathbb{R}\times\mathbb{R}_+$;

Beh.: Die Lösung des RPs besteht aus einem einzigen 3–Schock, genau dann wenn gilt:

1) $\quad 0 < B \leq 1;$

2) $\quad C = \dfrac{2\sqrt{\tau}\ (1-B^{-1})}{(\gamma-1)\sqrt{1+\beta\, B^{-1}}}\sqrt{\dfrac{B}{A}};$

3) $\quad A = \dfrac{1+\beta\, B}{\beta + B}.$

Beweis:
Wegen La. 5.3 und La. 5.4 gilt unter der Voraussetzung von (L.5.2.3):

$$(L.5.2.1),\ (L.5.2.4),\ (L.5.2.4) \overset{(5.15.c)}{\Longleftrightarrow} (S.5.2.1),\ (S.5.2.2),\ (S.5.2.3).$$

Mit [Smo], S. 355 gilt insbesondere:

$$(1) \qquad h_1(\mathbb{R}) \;=\; \left(-\infty, \frac{2}{\gamma - 1}\right).$$

Daraus folgt mit (L.5.3.1) bzw. (S.5.2.2) die Bedingung (L.5.2.3), denn:

$$c \overset{(L.5.3.1)}{=} \sqrt{\frac{B}{A}}\, h_1(\ln\ B) \overset{(1)}{=} \sqrt{\frac{B}{A}}\,\frac{2}{\gamma - 1} \;<\; \frac{2}{\gamma - 1}\left(\sqrt{\frac{B}{A}} + 1\right).$$

Somit gilt also die Behauptung.

■

Aus den Bedingungen von S. 5.2 folgt insbesondere mit (5.15.b), daß die Zustände links vom 3–Schock komponentenweise größer sein müssen als die rechts davon. Dies gilt wegen dem Satz von Oleinik (s. [Smo]) ebenso für die Entropielösung im skalaren Fall. Aufgrund von S. 5.2 und (S.5.1.7) gilt nun

Folgerung 5.1:

Vor.: $\underline{v}_L,\ \underline{v}_M,\ \underline{v}_R \in \mathbb{R}^3 \setminus \{(x,y,z)\,|\,x,y > 0,\ z \in \mathbb{R}\}$,
$(\underline{v}_L, \underline{v}_M)$ und $(\underline{v}_M, \underline{v}_R)$ erfüllen die Bedingungen (S.5.2.1–3);

Beh.: a) Die Entropielösung zu (5.13) besteht aus genau zwei 3–Schocks mit folgenden charakteristischen Geschwindigkeiten:

$$s_{3,L} = \begin{Bmatrix} \dfrac{\rho_M v_M - \rho_L v_L}{\rho_M - \rho_L} & ,\rho_L \neq \rho_M \\ \lambda_3(\underline{v}_M) & ,\rho_L = \rho_M \end{Bmatrix},\ s_{3,R} = \begin{Bmatrix} \dfrac{\rho_R v_R - \rho_M v_M}{\rho_R - \rho_M} & ,\rho_R \neq \rho_M \\ \lambda_3(\underline{v}_R) & ,\rho_R = \rho_M \end{Bmatrix};$$

b) Die beiden 3–Schocks laufen ineinander, genau dann wenn gilt: $s_{3,L} > s_{3,R}$.
Der Interaktionspunkt ist:

$$(x_0, t_0) \;=\; \left(x_1 \;+\; \frac{x_2 - x_1}{s_{3,L} - s_{3,R}}\, s_{3,L},\ \frac{x_2 - x_1}{s_{3,L} - s_{3,R}}\right).$$

Mittels obiger Folgerung lassen sich nun $\underline{v}_L, \underline{v}_M, \underline{v}_R$ so bestimmen, daß sie das geforderte leisten. Eine ganze Familie von Testwerten ist nun angegeben in

Lemma 5.5:

Vor.: $0 < k < 1,\ P_L, \rho_L > 0,\ v_L \in \mathbb{R},\ \gamma = 1.4,$
$\rho_M = \frac{1+6k}{6+k}\,\rho_L,\ \rho_R = \frac{1+6k}{6+k}\,\rho_M, P_M = k\,P_L,\ P_R = k\,P_M,$
$v_M = \sqrt{5}\,\frac{k-1}{\sqrt{1+6k}}\,\sqrt{\frac{P_L}{\rho_L}} + v_L,\ v_R = \sqrt{5}\,\frac{k-1}{\sqrt{1+6k}}\,\sqrt{\frac{P_M}{\rho_M}} + v_M;$

Beh.: Die Lösung zu (5.13) mit obigen Anfangsdaten besteht aus genau zwei 3–Schocks, die kollidieren. Die zugehörigen charakteristischen Geschwindigkeiten sind:

$$s_{3,L} \;=\; \frac{v_L\,(6+k) \;-\; v_M\,(1+6\,k)}{5 - 5\,k},\quad s_{3,R} \;=\; \frac{v_M\,(6+k) \;-\; v_R\,(1+6\,k)}{5 - 5\,k}.$$

Beweis:

Die Behauptung folgt aus F. 5.1, in dem man die Bedingungen aus S. 5.2 für $(\underline{v}_L, \underline{v}_M)$ bzw. $(\underline{v}_M, \underline{v}_R)$ verifiziert.

■

Kapitel 6: Modifikationen der ENO–Verfahren

Die in Kapitel 2 vorgestellten Verfahren haben zwei entscheidende Nachteile, die zum einen von numerischer und zum anderen von analytischer Natur sind:

(6.1) die Bestimmung des numerischen Flusses gemäß (2.1) bzw. (2.9) ist sehr aufwendig;

(6.2) die Rekonstruktion $L^N(\circ, v_j^n)$ (vgl. (2.3), (2.4)) bzw. $\underline{L}^N(\circ, \underline{v}_j^n)$ (vgl. (2.22)) ist in $x_{j\pm\frac{1}{2}}$ im allgemeinen *nicht* stetig bzw. stetig differenzierbar.

Dadurch hat man bei der Verwendung der ENO–Verfahren lange Laufzeiten und bei der Analyse des Oszillationsverhaltens der Rekonstruktionsmethoden Probleme. Um diese zu umgehen, lassen sich Modifikationen an den Verfahren vornehmen. Dazu gibt es mehrere Möglichkeiten:

zu (6.1) — Semidiskrete Verfahren (s. [ENO2], [Ad2]);
— Fluß–Interpolation (s. [OsS1], [OsS2]);

zu (6.2) — Hermite–Interpolation (s. [Wan])

Diese Varianten sollen nun im folgenden kurz beschrieben werden, ohne sie jedoch genauer zu analysieren.

6.1 Modifikation des numerischen Flusses

6.1.1 Semidiskrete Verfahren

Sei u (schwache) Lösung eines Erhaltungssatzes gemäß Def. 2.1 bzw. Def. 2.13. Dann gilt für $(x,t) \in \mathbb{R} \times \mathbb{R}_+,\ \tau, h > 0$ wegen Kor. 2.1 (vgl. Kap. 2.1.2 bzw. Kap. 2.2.2):

$$(6.3) \qquad \overline{u}(x, t+\tau) \;=\; \overline{u}(x,t) \;-\; \frac{\tau}{h}\left[\,\hat{F}_{+\frac{1}{2}} \;-\; \hat{F}_{-\frac{1}{2}}\,\right] \qquad \text{mit}$$

$$\hat{F}_{\pm\frac{1}{2}} \;:=\; \frac{1}{\tau}\int_t^{t+\tau} F\left(u(x\pm\tfrac{h}{2},\tau)\right)\,d\tau,\ \overline{u}(x,t) \;:=\; \frac{1}{h}\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} u(s,t)\,ds.$$

Dies läßt sich auch äquivalent schreiben als

$$(6.4) \qquad \frac{\overline{u}(x,t+\tau) \;-\; \overline{u}(x,t)}{\tau} \;=\; -\frac{\hat{F}_{+\frac{1}{2}} \;-\; \hat{F}_{-\frac{1}{2}}}{h}.$$

Geht man in (6.4) auf beiden Seiten zum Grenzwert für $\tau \to 0+$, so gilt:

$$(6.5) \qquad \tfrac{\partial}{\partial t}\,\overline{u}(x,t) \;=\; -\frac{F(u(x+\frac{h}{2},t+0)) \;-\; F(u(x-\frac{h}{2},t+0))}{h}.$$

Die linke Seite läßt sich wie folgt durch einen Differenzenquotienten approximieren:

$$(6.6) \qquad \tfrac{\partial}{\partial t}\,\overline{u}(x,t) \;=\; \frac{\overline{u}(x,t+\tau) \;-\; \overline{u}(x,t)}{\tau} \;+\; O(\tau).$$

Führe nun folgende Bezeichnungen ein:

a) v_j^n Approximation von $\overline{u}(x_j, t_n)$;

b) $v,\ \tilde{v}$ Lösungen von

$$\left\{\begin{array}{ll} v_t(x,t) + (F(v(x,t)))_x = 0, \; x \in \mathbb{R},\ t > t_n \\ v(x,t_n) = L^N(x, v_j^n), & x \in \mathbb{R} \end{array}\right\} \text{ bzw.}$$

$$\left\{\begin{array}{ll} \tilde{v}_t(x,t) + (F(\tilde{v}(x,t)))_x = 0, & x \in \mathbb{R},\ t > t_n \\ \tilde{v}(x,t_n) = \left\{\begin{array}{l} L^N(x_{j+\frac{1}{2}} - 0, v_j^n),\ x < x_{j+\frac{1}{2}} \\ L^N(x_{j+\frac{1}{2}} + 0, v_j^n),\ x > x_{j+\frac{1}{2}} \end{array}\right\}, & x \in \mathbb{R} \end{array}\right\}.$$

Dann gilt für den Bereich, in dem u glatt ist wegen (2.2.a) und La. 2.15 bzw. La. 2.21:

$$\begin{aligned} F(u(x_{j+\frac{1}{2}}, t_n + 0)) &= F(v(x_{j+\frac{1}{2}}, t_n + 0)) \; + \; O\left(h^N\right) \\ &= F(\tilde{v}(x_{j+\frac{1}{2}}, t_n + 0)) \; + \; O\left(h^N\right) \\ &= F^{God}\left(L^N(x_{j+\frac{1}{2}} - 0, v_j^n), L^N(x_{j+\frac{1}{2}} + 0, v_j^n)\right) \; + \; O\left(h^N\right). \end{aligned}$$

Mit (6.5), (6,6) und (6.7) erhält man nun folgendes Verfahren

$$v_j^0 := \overline{u}(x_j, 0),\ v_j^{n+1} = v_j^n - \lambda \left(F_{j+\frac{1}{2}}^{God} - F_{j-\frac{1}{2}}^{God}\right) \text{ mit} \tag{6.8}$$

$$F_{j-\frac{1}{2}}^{God} = F^{God}\left(L^N(x_{j+\frac{1}{2}} - 0, v_j^n), L^N(x_{j+\frac{1}{2}} + 0, v_j^n)\right),\ \lambda := \frac{\tau}{h}.$$

Diese Modifikation hat den Vorteil, daß die Bestimmung des numerischen Flusses durch die Einführung eines Godunov–Fluß vereinfacht wird, in dem man die Diskretisierung in t–Richtung fallen läßt. Für τ hinreichend klein hat dieses Verfahren dann ähnliche Eigenschaften wie das ursprüngliche. Der Nachteil liegt nun zum einen im Ordnungsverlust, den man wegen (6.6) in Kauf nehmen muß, und zum anderen kann das ursprüngliche Verfahren nicht zurückgewonnen werden, da die Informationen in t–Richtung fehlen.

Eine Variante dieser Modifikation für den *skalaren* Fall besteht nun darin, daß man an Stelle des Godunov–Fluß den Fluß eines s.g. 3–Punkte–E–Schemata (s. [Os]) verwendet.

Definition 6.1:

Ein $2q+1$ Punkte–Verfahren in Erhaltungsform mit konsistentem Fluß F^E, d.h. $F^E(u,\ldots,u) = F(u)$, heißt ein *E-Schema*, genau dann wenn gilt:

$$\left(F_{j+\frac{1}{2}}^E - F(u)\right) \text{ sign}\,(v_{j+1}^n - v_j^n) \leq 0,\ \forall\, j \in \mathbb{Z},\ u \in \left[\,\min\{v_{j+1}^n, v_j^n\}, \max\{v_{j+1}^n, v_j^n\}\,\right].$$

Von diesen Verfahren lassen sich folgende Eigenschaften zeigen (s. [Ta], [Le]):

Lemma 6.1:

Vor.: Sei ein E–Schema gemäß Def. 6.1 gegeben;

Beh.: a) Das E–Schema ist von 1. Ordnung;

b) Ein 5–Punkte–E–Schema ist TVD und erfüllt eine numerische Entropiebedingung, d.h. es konvergiert gegen die Entropielösung;

c) Ein 3–Punkte Verfahren in Erhaltungsform, das monoton ist, ist ein E–Schema.

6.1.2 Fluß–Interpolation

Bei den ENO–Verfahren werden die *Zellmittel* $\overline{u}(\circ,t)$ der exakten Lösung $u(\circ,t)$ interpoliert. Dazu muß zum einen $u(\circ,t)$ aus den Näherungen der Mittelwerte rekonstruiert und zum anderen der Fluß $F(u)$ approximiert werden. Dies ist sehr aufwendig. Eine mögliche Vereinfachung besteht nun darin, den *Fluß* durch eine geeignete Interpolationsmethode zu approximieren. Es soll also nun ein Verfahren in Erhaltungsform mit konsistentem numerischen Fluß konstruiert werden, d.h.:

$$
(6.9) \qquad \begin{aligned} v_j^{n+1} &= v_j^n - \lambda \left(\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}} \right), \; v_j^0 := u(x_j,0), \; \lambda := \frac{\tau}{h}, \; \tau, h > 0 \\ \hat{F}_{j+\frac{1}{2}} &= \hat{F}(v_{j-l}^n, \ldots, v_{j+k}^n), \; \hat{F}(u,\ldots,u) = F(u). \end{aligned}
$$

Vom numerischen Fluß verlangt man nun:

a) $\exists\, a_2, \ldots, a_{2N} \in \mathbb{R} : \hat{F}_{j+\frac{1}{2}} \overset{!}{=} F(u(x_{j+\frac{1}{2}},t_n)) + \sum_{k=1}^{N} \left(\frac{\partial}{\partial x}\right)^{2k} F(u(x_{j+\frac{1}{2}},t_n))\, a_{2k}\, h^{2k};$

b) $\frac{1}{h} \left[\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}} \right] \overset{!}{=} \frac{\partial}{\partial x} F(u(x_j,t_n)) + O\left(h^{2N+1} \right).$

Die Koeffizienten a_{2k} bestimmen sich dabei aus

Lemma 6.2:

Vor.: $\hat{F}_{j\pm\frac{1}{2}}$ erfülle (6.10), $N > 2,\; k \in \{1,\ldots,N-1\}$;

Beh.: $a_2 = -\frac{1}{24};\; a_{2(k+1)} = -\frac{1}{4^{k+1}} \left[\frac{1}{(2k+3)\,!} + \sum_{i=1}^{k} a_{2i} \frac{4^i}{(\,2(i-1)+3\,)\,!} \right].$

Beweis:

Entwickle $\left(\frac{\partial}{\partial x}\right)^{2k} F(u(x_{j+\frac{1}{2}},t_n))$ um x_j bis zur Ordnung $2(N-k)+1$ für $k = 1,\ldots,N$. Setze diese Taylorentwicklung in (6.10.a) ein. Dann erhält man folgende Gleichung (beachte: $u_j^n := u(x_j,t_n)$):

$$
\frac{1}{h} \left(\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}} \right)
$$

$$
(1) \qquad \overset{!}{=} \frac{\partial}{\partial x} F(u_j^n) + \frac{1}{h} \sum_{k=1}^{N-1} \left(\frac{\partial}{\partial x}\right)^{2k+1} F(u_j^n) \left(\frac{h}{2}\right)^{2k+1} \frac{2}{(2k+1)\,!} +
$$

$$
\frac{1}{h} \sum_{k=1}^{N-1} \left(\frac{\partial}{\partial x}\right)^{2k+1} F(u_j^n) \sum_{i=1}^{k} h^{2i}\, a_{2i} \frac{2}{(2(k-i)+1)\,!} \left(\frac{h}{2}\right)^{2(k-i)+1} + O\left(h^{2N+1} \right).
$$

Wegen (6.10.b) muß andererseits gelten:

$$
(2) \qquad \frac{1}{h} \left(\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}} \right) \overset{!}{=} \frac{\partial}{\partial x} F(u_j^n) + O\left(h^{2N+1} \right).
$$

Durch Gleichsetzen von (1) und (2) erhält man durch Koeffizientenvergleich die Behauptung.

∎

Setze nun

$$(6.11)\qquad F^+(u) := \frac{1}{2}\ (\ F(u) + a\ u\ ),\ \ F^-(u) := \frac{1}{2}\ (\ F(u) - a\ u\ )\ \text{ mit}$$
$$a \geq \max\left\{|F'(u)|\ \big|\ |u| \leq \|u_0\|_\infty\right\}.$$

Dann gilt offensichtlich:

$$(6.12)\qquad \tfrac{\partial}{\partial u}\,F^+\ \geq\ 0,\ \ \tfrac{\partial}{\partial u}\,F^-\ \leq\ 0,\ \ F\ \equiv\ F^+\ +\ F^-.$$

Es sollen nun Interpolationspolynome $P_j^{\pm} \in \mathbb{P}_{2N}$ zu $F^{\pm}$ konstruiert werden mit

$$(6.13)\qquad P_j^{\pm}(x)\ \stackrel{!}{=}\ F^{\pm}(u(x,t_n))\ +\ O\left(\,h^{2N+1}\right),\ \ x \in I_{j-\frac{1}{2}}.$$

Dazu:

Definition 6.2:

Sei $h > 0, N \in \mathbb{N}_0$. Das *Fluß–Interpolationspolynom* $Q_{\pm}^N(\circ, F(u))$ ist für $x \in I_{j-\frac{1}{2}}$ rekursiv definiert. Für $N = 0$ setze

$$\begin{aligned}
&\text{a) } Q_+^0(x,F(u))\ =\ F^+(u_j^n),\ k_m^{(0)}\ =\ k_M^{(0)}\ :=\ j;\\
&\text{b) } Q_-^0(x,F(u))\ =\ F^-(u_{j+1}^n),\ k_m^{(0)}\ =\ k_M^{(0)}\ :=\ j+1.
\end{aligned}$$

Seien $Q_{\pm}^{N-1}(\circ, F(u))$ sowie $k_m^{(N-1)} <\ k_M^{(N-1)}$ schon definiert. Setze

$$a_{\pm}^{(N)}\ :=\ F^{\pm}[u_{k_m^{(N-1)}}, \ldots, u_{k_M^{(N-1)}+1}],\qquad b_{\pm}^{(N)}\ :=\ F^{\pm}[u_{k_m^{(N-1)}-1}, \ldots, u_{k_M^{(N-1)}}].$$

Definiere $Q_{\pm}^N(\circ, F(u))$ wie folgt:

$$Q_{\pm}^N(x,F(u))\ :=\ Q_{\pm}^{N-1}(x,F(u))\ +\ \overline{m}\left(\,a_{\pm}^{(N)}, b_{\pm}^{(N)}\,\right)\prod_{k=k_m^{(N-1)}}^{k_M^{(N-1)}}(x - x_k)\ \text{mit}$$

$$k_m^{(N)}\ :=\ \left\{\begin{array}{ll} k_m^{(N-1)} - 1 & , |a_{\pm}^{(N)}| \geq |b_{\pm}^{(N)}| \\ k_m^{(N-1)} & , \text{sonst} \end{array}\right\},\ k_M^{(N)}\ :=\ \left\{\begin{array}{ll} k_M^{(N-1)} & , |a_{\pm}^{(N)}| \geq |b_{\pm}^{(N)}| \\ k_M^{(N-1)} + 1 & , \text{sonst} \end{array}\right\},$$

$$\overline{m}\left(\,a_{\pm}^{(N)}, b_{\pm}^{(N)}\,\right)\ :=\ \left\{\begin{array}{ll} b_{\pm}^{(N)} & , |a_{\pm}^{(N)}| \geq |b_{\pm}^{(N)}| \\ a_{\pm}^{(N)} & , \text{sonst} \end{array}\right\}.$$

Setze nun

(6.14)
$$\text{a) } P_j^{\pm}(x) \;:=\; Q_{\pm}^{2N}(x, F(u)) \text{ für } x \in I_{j-\frac{1}{2}};$$
$$\text{b) } F_{j+\frac{1}{2}}^{\pm} \;:=\; P_j^{\pm}(x_{j+\frac{1}{2}}) \;+\; \sum_{k=1}^{m} a_{2k} \left(\tfrac{d}{dx}\right)^{2k} \; P_j^{\pm}(x_{j+\frac{1}{2}});$$
$$\text{c) } \hat{F}_{j+\frac{1}{2}} \;:=\; \hat{F}_{j+\frac{1}{2}}^{+} \;+\; \hat{F}_{j+\frac{1}{2}}^{-}.$$

Damit ist das Verfahren (6.9) explizit bestimmt. Beachte dabei, daß aufgrund der Fehlerdarstellung für Interpolationspolynome (s. [BS], S. 49,ff) in den Bereichen, in denen $F^{\pm}$ glatt ist, mit (6.14) auch (6.10) erfüllt ist.

Der Vorteil des so konstruierten Verfahrens liegt zum einen darin, daß $\hat{F}_{j+\frac{1}{2}}$ beliebig oft differenzierbar ist und zum anderen der Fluß F in x–Richtung von hoher Ordnung durch den numerischen Fluß approximiert wird. Da aber die Informationen in t–Richtung verloren gehen, ist das Verfahren im allgemeinen nicht von hoher Ordnung gemäß Def. 2.11.

Abschließend sei noch vermerkt, daß man für $N = 1$ das s.g. Lax–Friedrichs–Verfahren erhält. Obige Verfahren können somit als Verallgemeinerung dieses Schemas aufgefaßt werden (vgl. auch La. 2.17.a).

6.2 Hermite–Rekonstruktion

Die Rekonstruktionen $L^N(\circ, \overline{w})$ gemäß der RP– bzw. RD–Methode (vgl. Kap. 2.1.3) haben den Nachteil, daß sie in $x_{j+\frac{1}{2}}$ im allgemeinen nicht stetig sind. Dabei wurde $w \in SCL(N+1, \mathbb{R})$ vorausgesetzt und mit $\overline{w}$ der zugehörige Mittelwert bezeichnet. Es soll nun ein Rekonstruktionspolynom $L^{2N-2}(\circ, \overline{w})$ konstruiert werden mit

(6.15.a) Approximation:

Für alle Punkte x, die eine Umgebung besitzen, in der w glatt ist, gilt:

$$L^{2N-2}(x, \overline{w}) \;=\; w(x) \;+\; O\left(h^{N+1}\right);$$

(6.15.b) Erhaltung:

$$\overline{L}^{2N-2}(x_j, \overline{w}) \;:=\; \frac{1}{h} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} L^{2N-2}(x, \overline{w})\, dx \;=\; \overline{w}(x_j);$$

(6.15.c) ENO–Bedingung:

$$\mathrm{TV}\left(L^{2N-2}(\circ, \overline{w})\right) \;\leq\; \mathrm{TV}(w) \;+\; O\left(h^{q+1}\right), \qquad q > 0;$$

(6.15.d) Stetigkeit:

$$L^{2N-2}(x_{j+\frac{1}{2}} - 0, \overline{w}) \;=\; L^{2N-2}(x_{j+\frac{1}{2}} + 0, \overline{w}).$$

Kapitel 7: Numerische Untersuchung

Konstruiere zunächst ein Hermite–Interpolationspolynom. Dazu sei $w \in SCL(N+1, \mathbb{R})$ und W eine Stammfunktion zu w. Gehe wie folgt vor:

(6.16.a) Konstruiere $Q^N(\circ, W)$ gemäß Kap. 1;

(6.16.b) $H^{2N-1}(\circ, W)$ sei bestimmt durch

1) $H^{2N-1}(x_{j-\frac{1}{2}}, W) \stackrel{!}{=} W_{j-\frac{1}{2}}, \quad j \in \mathbb{Z};$

2) $$\frac{d}{dx} H^{2N-1}(x_{j-\frac{1}{2}}, W) \stackrel{!}{=} \operatorname{minmod}\Big(\frac{d}{dx} Q^N(x_{j-\frac{1}{2}} + 0, W), \frac{d}{dx} Q^N(x_{j-\frac{1}{2}} - 0, W) \Big) := m_{j-\frac{1}{2}};$$

3) $$H^{2N-1}(x, W) \stackrel{!}{=} \sum_{j=k_m^{(N)}}^{k_M^{(N)}} \Big\{ W_{j-\frac{1}{2}} + (x - x_{j-\frac{1}{2}}) \Big[m_{j-\frac{1}{2}} - 2\, W_{j-\frac{1}{2}}\; l_{j-\frac{1}{2}}(x_{j-\frac{1}{2}}) \Big] \Big\}\; l^2_{j-\frac{1}{2}}(x)$$

$$l_{j-\frac{1}{2}}(x) := \prod_{\substack{i=k_m^{(N)} \\ i \neq j}}^{k_M^{(N)}} \frac{x - x_{i-\frac{1}{2}}}{x_{j-\frac{1}{2}} - x_{i-\frac{1}{2}}}, \qquad x \in I_{j-\frac{1}{2}}.$$

Definiere nun die s.g. Hermite–Rekonstruktion durch

$$L^{2N-2}(\circ, \overline{w}) := \frac{d}{dx} H^{2N-1}(\circ, W).$$

Dann ist (6.15.d) wegen (6.16.b) trivialerweise erfüllt. Die übrigen Eigenschaften lassen sich analog zu Kap. 2.1.3 zeigen.

Der Nachteil dieser Rekonstruktionsmethode im Gegensatz zu $L^{N-1,RP}(\circ, \overline{w})$ ist der, daß doppelt so viele Punkte pro Intervall benötigt werden, um die gleiche Ordnung zu erzielen. Dafür hat man aber eine glattere Rekonstruktion von w.

Abschließend sei noch vermerkt, daß Harten in [Ha3] eine weitere Modifikation der ENO–Schemata hergeleitet hat. Diese besteht darin, daß er zum numerischen Fluß der ENO–Verfahren einen Korrekturterm hinzuaddiert — ähnlich wie bei seinen 5–Punkte–TVD–Verfahren (s. [Ha2]). Hier soll auf diese Methode aber nicht näher eingegangen werden, da sie nur 2. Ordnung und nicht von beliebiger konstruiert ist.

Im folgenden sollen nun die analytischen Ergebnisse an Hand von numerischen Testbeispielen überprüft werden. Dabei stehen folgende Aspekte im Vordergrund:

a) numerische Ordnungsuntersuchung;

b) Verhalten der ENO–Verfahren in s.g. noncritical sonic points (vgl. Kap. 2.1.1);

c) Vergleich der numerischen Flüsse: F^{God} — F^{Roe} — F^{Rav};

d) Schock–Interaktion.

7.1. Numerische Ordnungsuntersuchung

Dazu soll zunächst einmal definiert werden, was man unter der numerischen Ordnung eines Verfahrens zu verstehen hat:

Definition 7.1:

Sei $\{v_{h_j}\}_{j\in\mathbb{N}}$ eine Folge von Approximationen zu $u(x,t)$, $(x,t)\in\mathbb{R}\times\mathbb{R}_+$ beliebig, aber fest zu den Schrittweiten h_j mit $\lambda := \frac{\tau_j}{h_j} = const$ wobei $\{h_j\}_{j\in\mathbb{Z}}$ eine streng monoton fallende Nullfolge ist. Das Verfahren hat (lokal) die *numerische Ordnung* $p \geq 1$, genau dann wenn gilt:

$$\exists\, C > 0 : \limsup_{j\to+\infty} \frac{e_{j+1}}{h_j^p} \;=\; C \qquad \text{mit} \qquad e_{j+1} \;:=\; |\, v_{h_j} \;-\; u(x,t)\,|.$$

Falls $p = 1$, so muß gelten: $C < 1$.

Diese Definition läßt sich nun zur Überprüfung der numerischen Ordnung der ENO–Verfahren wie folgt anwenden:

Lemma 7.1:

Vor.: $h_0 > 0$, $h_j := h_0\, 2^{-j}$, $\{v_{h_j}\}_{j\in\mathbb{N}}$, $\{e_j\}_{j\in\mathbb{N}}$ gemäß Def. 7.1;
Beh.: Falls das Verfahren numerisch die Ordnung $p \geq 1$ hat, so gilt: für $i \in \mathbb{N}$ hinreichend groß

$$p \;\approx\; \log_2\left(\frac{e_i}{e_{i+1}}\right).$$

Beweis:

Wegen Def. 7.1 existiert $C > 0$ so daß für $i \in \mathbb{N}$ hinreichend groß gilt:

$$e_{i+1} \;=\; C\, h_{i+1}^p \qquad \text{und} \qquad e_i \;=\; C\, h_i^p,$$

also auch

$$e_{i+1} \;\approx\; e_i\, 2^{-p}.$$

Daraus folgt die Behauptung.

■

Als Testbeispiel diente

$$\begin{aligned} u_t(x,t) \;+\; u_x(x,t) \;&=\; 0, \quad (x,t)\in\mathbb{R}\times\mathbb{R}_+; \\ u(x,0) \;&=\; \sin(\pi\, x), \quad x\in\mathbb{R}. \end{aligned} \tag{7.1}$$

Untersucht wurde der globale Fehler auf dem Intervall $[0,2]$ in der L_1– bzw. in der L_∞–Norm zum Zeitpunkt $t = 1$ für $\lambda = 1$ zum einen bzgl. der exakten Lösung u von (7.1) und zum anderen bzgl. der exakten Mittelwerte $\overline{u}$, d.h.:

$$\begin{aligned} E_1 \;&:=\; \sum_{j=0}^{m} |\, v_j^n \;-\; u(x_j,1)\,|,\ E_\infty \;:=\; \max\left\{|\, v_j^n \;-\; u(x_j,1)\,| \;\middle|\; j = 0,\ldots,m\right\}; \\ \overline{E}_1 \;&:=\; \sum_{j=0}^{m} |\, v_j^n \;-\; \overline{u}(x_j,1)\,|,\ \overline{E}_\infty \;:=\; \max\left\{|\, v_j^n \;-\; \overline{u}(x_j,1)\,| \;\middle|\; j = 0,\ldots,m\right\} \end{aligned}$$

wobei m die Anzahl der Zerlegung des Intervalls $[0,2]$ bezeichnet und n die Anzahl der Zeitschritte in $[0,1]$. Dabei wurden $L^{N,RD}(\circ, v_j^n)$ und $L^{N,RP}(\circ, v_j^n)$ für $N = 1,2,3$ verwendet und $\hat{F}_{j+\frac{1}{2}} := F^{God}_{j+\frac{1}{2}}$.
Aus Tabelle 1,2 läßt sich nun folgendes ablesen:

1) mit wachsendem N nimmt die numerische Ordnung der Mittelwertbildung entsprechend zu (vgl. S. 2.3);
2) die numerische Ordnung der exakten Lösung wird für $N \geq 2$ nicht größer als 2 (vgl. F. 2.3);
3) die numerische Konvergenzordnung ist in der L_1–Norm im allgemeinen etwas grösser als in der L_∞–Norm.

7.2 Noncritical Sonic Points

Unter solchen Punkten $(x,t) \in \mathbb{R} \times \mathbb{R}_+$ versteht man diejenigen, in denen für die Lösung u eines Erhaltungssatzes gilt:

$$u_x(x,t) \;=\; 0, \;\left(\tfrac{d}{dx}\right)^2 \; u(x,t) \;\neq\; 0, \;F'(u(x,t)) \;\neq\; 0.$$

Wie in [HaT] gezeigt wird, haben die TVD–Verfahren in solchen Punkten einen Ordnungsverlust, d.h. sie können dort nur von 1. Ordnung sein. Ziel der ENO–Konstruktion war es unter anderem, diese Ordnungsminderung zu vermeiden. Als Testbeispiel wurde diesmal die Burgergleichung verwendet:

$$\begin{aligned} u_t(x,t) \;+\; u(x,t)\, u_x(x,t) \;&=\; 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+; \\ u(x,0) \;&=\; \sin(\pi\, x), \quad x \in \mathbb{R}. \end{aligned} \tag{7.1}$$

Die Plots in Abb. 1 wurden zu folgenden Parametern erstellt:

$$t \;=\; 0.3, \; x \in [0,2], \; \lambda \;=\; 0.25, \; h \;=\; 0.1.$$

Dabei wurden die folgenden Verfahren verwendet:

a) ENO–Verfahren mit $L^{3,RP}(\circ, v_j^n)$ und $\hat{F}_{j+\frac{1}{2}} = F^{God}_{j+\frac{1}{2}}$;
b) ENO–Verfahren mit $L^{2,RD}(\circ, v_j^n)$ und $\hat{F}_{j+\frac{1}{2}} = F^{God}_{j+\frac{1}{2}}$;
c) TVD–Verfahren von Harten (s. [Ha2]) mit numerischer Viskosität $Q(x) := \dfrac{x^2+1}{2}$.

Aus diesen Abbildungen geht hervor, daß der noncritical sonic point $(0.8, 0.3)$ von den ENO–Verfahren nahezu exakt erfaßt wird, während das TVD–Verfahren die Approximation dort erheblich dämpft.

7.3. Vergleich der numerischen Flüsse

In Kap. 2.1.4 bzw. Kap. 2.2.4 sind drei verschiedene Möglichkeiten der Konstruktion von $\hat{F}_{j+\frac{1}{2}}$ angegeben. Dabei beruhen $F^{God}_{j+\frac{1}{2}}$ und $F^{Roe}_{j+\frac{1}{2}}$ auf Lösen von Riemann–Problemen, während $F^{Rav}_{j+\frac{1}{2}}$ auf einer asymptotischen Entwicklung der Lösung eines verallgemeinerten Riemann–Problems aufgebaut ist. Im folgenden sollen nun die nachstehenden Fragen untersucht werden:

1) Sind Unterschiede bei der Anwendung der ENO–Verfahren zu den gleichen Testdaten bei Verwendung unterschiedlicher Rekonstruktionsmethoden bzw. verschiedener numerischer Flüsse zu erkennen?
2) Wie verhalten sich die ENO–Verfahren bei Verwendung von $F^{Rav}_{j+\frac{1}{2}}$, $F^{Roe}_{j+\frac{1}{2}}$, $F^{God}_{j+\frac{1}{2}}$ bei "höheren" CFL–Zahlen?

Als Testbeispiel wurde hier das s.g. Sod–Problem gewählt:

$$\underline{u}_t(x,t) \;+\; (\underline{F}(\underline{u}(x,t)))_x \;=\; 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+,\ \underline{F},\ \underline{u} \text{ gemäß (3.6);}$$

$$\underline{u}(x,0) \;=\; \begin{Bmatrix} \underline{u}_L,\ x < 0 \\ \underline{u}_R,\ x > 0 \end{Bmatrix} \text{ mit } \begin{Bmatrix} (\rho_L, P_L, v_L)^T = \ (1,1,0)^T \\ (\rho_R, P_R, v_R)^T = \ (0.1, 0.125, 0)^T \end{Bmatrix}. \tag{7.3}$$

Die Plots in Abb. 2–7 wurden für die Dichte (Abb. 2, 3), den Druck (Abb. 4, 5) und den Impuls (Abb. 6, 7) zu den folgenden Parametern erstellt:

$$t \;=\; 0.2,\ x \in [-0.5, 0.5],\ \lambda \;=\; 0.2,\ h \;=\; 0.01.$$

Aus ihnen ist kaum ein Unterschied auszumachen, weder in der Wahl des numerischen Flusses noch in der Wahl der Rekonstruktionsmethode. Zum Vergleich ist das Godunov–Verfahren zu den selben Parametern angegeben (s. Abb. 8).

Für die Plots in Abb. 9 (Dichte), Abb. 10 (Druck), Abb. 11 (Impuls) und Abb. 12 (Energie) wurden folgende Parameter verwendet:

$$x \in [-0.7, 0.7],\ \lambda \;=\; 0.45,\ h \;=\; 0.02,\ L^N(\circ, v_j^n) \;=\; L^{Rd,1}(\circ, v_j^n).$$

Mit dem numerischen Fluß $F^{Rav}_{j+\frac{1}{2}}$ erhält man ähnlich gute Ergebnisse wie bei der Verwendung einer nicht einmal halb so großen CFL–Zahl (vgl. Abb. 2a, 4a, 6a). Bei $F^{God}_{j+\frac{1}{2}}$ kommt es zu starken Oszillationen ebenso mit $F^{Roe}_{j+\frac{1}{2}}$. In letzterem Fall führen diese sehr rasch zum Abbruch des Verfahrens.

Durch die Verwendung von $F^{Rav}_{j+\frac{1}{2}}$ kann man gegenüber $F^{Roe}_{j+\frac{1}{2}}$ und $F^{God}_{j+\frac{1}{2}}$ also mit einer doppelt so großen Schrittweite arbeiten. Dies rechtfertigt den zum Teil erheblichen Mehraufwand bei der Berechnung und der Implementation der Raviart–Prozedur.

7.4. Schock–Interaktion

In Kapitel 1 wurde gezeigt, daß die ENO–Interpolation, falls zwischen zwei Sprüngen der zu interpolierenden Funktion $v \in SCL(N+1, \mathbb{R})$ genügend viele Zwischenintervalle liegen, im wesentlichen nichtoszillierend ist. Im folgenden soll nun untersucht werden, wie sich die ENO–Verfahren verhalten, wenn zwei Schocks ineinanderlaufen. Für den skalaren Fall (vgl. Abb. 13) wurde dabei das Testproblem (5.1) mit folgenden Parametern verwendet:

$$x \in [0,1],\ \lambda \ = \ 0.18,\ h \ = \ 0.025,\ L^N(\circ, v_j^n) \ = \ L^{3,RD}(\circ, v_j^n),\ \hat{F}_{j+\frac{1}{2}} \ = \ F^{God}_{j+\frac{1}{2}}.$$

Es sind jeweils die Zustände unmittelbar vor und nach einer Interaktion abgebildet. Es zeigt sich, daß die ENO–Verfahren im skalaren Fall auch bei Schockinteraktionen sehr gut arbeiten, falls man den Grad der verwendeten Interpolation nicht zu klein wählt. Die Verwendung von $F^{Rav}_{j+\frac{1}{2}}$ brachte keine hier keine Verbesserung.

Als Testbeispiel für Systeme wurde (5.13) mit den folgenden Anfangsdaten gewählt:

$$\begin{aligned} \rho_L &= 1 \qquad\qquad\ , & P_L &= 0.1, & v_L &= \quad 0.01; \\ \rho_M &= 0.262295, & P_M &= 0.1, & v_M &= -1.580990; \\ \rho_R &= 0.068798, & P_M &= 0.1, & v_M &= -2.563354; \end{aligned} \tag{7.4}$$

$$x_0 \ = \ 0.4,\ x_1 \ = \ 0.6.$$

Diese Anfangsdaten sind wegen La. 5.5 mit $k = 0.1$ gerade so konstruiert, daß die Lösung aus genau zwei 3–Schocks besteht, die in $(x,t) = (0.4637, 0.11606)$ kollidieren und dort eine 1–Verdünnungswelle, eine 2–Kontaktunstetigkeit und einen 3–Schock produzieren. Das ENO–Verfahren wurde auf diese Daten zum einen mit $F^{Rav}_{j+\frac{1}{2}}$ und zum anderen mit $F^{Roe}_{j+\frac{1}{2}}$ als numerischen Fluß angesetzt. Dabei wurden folgende Parameter verwendet:

$$x \in [0,1],\ \lambda \ = \ 0.1,\ h \ = \ 0.02,\ L^N(\circ, v_j^n) \ = \ L^{1,RD}(\circ, v_j^n).$$

Aus den Plots für den Impuls (s. Abb. 14) geht hervor, daß beide Methoden die beiden 3–Schocks gut erfassen. Bei der Kollision haben beide Verfahren Probleme, aber dasjenige mit $F^{Rav}_{j+\frac{1}{2}}$ als numerischen Fluß "fängt" sich wesentlich schneller als das mit $F^{Roe}_{j+\frac{1}{2}}$, d.h. bei Interaktionsproblemen arbeitet das ENO–Verfahren mit dem Raviart–Fluß besser. In diesem Fall ist der Mehraufwand also wiederum gerechtfertigt.

Kapitel 8: Implementation

In diesem abschließenden Kapitel sollen nun noch einige Bemerkungen zur effizienten Programmierung der ENO–Verfahren gemacht werden.

8.1. Algorithmus zur Transformation eines Newton–Polynoms

Bei der Implementation der Rekonstruktionsmethoden gemäß Kap. 2.1.3 wird man vor das Problem gestellt, die Ableitungen des ENO–Splines $Q^N(\circ, W)$ bzw. $Q^N(\circ, \overline{w})$ an bestimmten Stellen zu berechnen. Da diese aber aufgrund der Def. 1.2 in der Newton–Darstellung für Interpolationspolynome gegeben sind, bietet es sich an, dieses in ein nach Potenzen geordnetes Polynom zu transformieren. Dazu:

Satz 8.1:

Vor.: $n \in \mathbb{N}_0,\ p_n(x) := \prod_{i=0}^{n} (x - x_i),\ (x_0, \ldots, x_n) \in \mathbb{R}^{n+1},\ x \in \mathbb{R};$

Beh.: $p(x) = x^{n+1} + \sum_{k=0}^{n} a_k^{(n)}\ x^{n-i}$ mit

$$a_k^{n+1} := (-1)^{k+1} \sum_{\substack{i_0, \ldots, i_k = 0 \\ i_0 < \cdots < i_k}}^{n} x_{i_0} \cdots x_{i_k} = (-1)^{k+1} \sum_{i_0=0}^{n} \sum_{i_1=i_0+1}^{n} \cdots \sum_{i_k=i_{k-1}+1}^{n} x_{i_0} \cdots x_{i_k}.$$

Beweis:

Mittels vollständiger Induktion über $n \in \mathbb{N}_0$ folgt die Behauptung, denn:
Für $n = 0$ ist die Aussage klar. Für $n + 1$ gilt dann mit Induktionsvoraussetzung (IV.):

$$\begin{aligned} p_{n+1}(x) &= x^{n+2} + \left(a_0^{(n+1)} - x_{n+1} \right) - a_n^{(n+1)}\ x_{n+1} \\ &\quad + \sum_{k=1}^{n} \left(a_k^{(n+1)} - a_{k-1}^{(n+1)}\ x_{n+1} \right) x^{n-k+1}; \end{aligned} \tag{1}$$

$$-a_n^{(n+1)}\ x_{n+1} \overset{\text{IV.}}{=} (-1)^{n+2}\ x_0 \cdots x_n\ x_{n+1} \overset{\text{def.}}{=} a_{n+2}^{(n+2)}; \tag{2}$$

$$\begin{aligned} &a_k^{(n+1)} - a_{k-1}^{(n+1)}\ x_{n+1} \overset{\text{IV.}}{=} \\ &(-1)^{k+1} \sum_{i_0=0}^{n} \cdots \sum_{i_{k-1}=i_{k-2}+1}^{n} x_{i_0} \cdots x_{i_{k-1}} \left(\sum_{i_k=i_{k-1}+1}^{n} x_{i_k} + x_{n+1} \right) \overset{\text{def.}}{=} a_k^{(n+2)}; \end{aligned} \tag{3}$$

$$a_0^{(n+1)} - x_{n+1} \overset{\text{IV.}}{=} - \sum_{i_0=0}^{n+1} x_{i_0} \overset{\text{def.}}{=} a_0^{(n+2)}. \tag{4}$$

Daraus folgt unmittelbar die Behauptung für $n + 1$.

∎

Aus obigem Satz läßt sich nun ein Algorithmus ableiten:

Folgerung 8.1:

Vor.: $N \in \mathbb{N},\ (n_0, \ldots, n_N) \in \mathbb{R}^{N+1},\ m \in \mathbb{Z},\ c \in \mathbb{R},\ x_m < \cdots < x_{m+N-1},$

$\tilde{x}_l := x_{m+l} - c,\ l = 0, \ldots, N-1,\ x \in \mathbb{R},\ Q^N(x) := \sum\limits_{i=0}^{N} n_i \prod\limits_{l=0}^{i-1} (x - x_{m+l}),$

$a_{-1}^{(k)} := 1,\ k = -1, \ldots, N-1,\ a_i^{(k)},\ k = 0, \ldots, N,\ i = 0, \ldots, k-1$ gemäß S. 8.1;

Beh.: a) $Q^N(x) = \sum\limits_{i=0}^{N} \left(\sum\limits_{l=0}^{N-i} n_{l+i}\, a_{l-1}^{(l-1+i)} \right) (x-c)^i;$

b) i) $a_l^{(l)} = -a_{l-1}^{(l-1)}\, \tilde{x}_l,\ l = 0, \ldots, N-1;$

ii) $a_l^{(k)} = a_l^{(k-1)} - \tilde{x}_{k-1}\, a_{l-1}^{(k-1)},\ l = 0, \ldots, N-1,\ k = l+1, \ldots, N-1.$

Beweis:

Zu a): Setze $\tau := x - c$. Dann gilt:

$$Q^N(x) = \sum_{i=0}^{N} n_i \prod_{l=0}^{i-1} (\tau - \tilde{x}_l).$$

Mit S. 8.1 folgt unmittelbar die Behauptung.

Zu b): Folgt aus dem Beweis zu S. 8.1 aus dem Induktionsschritt (vgl. Beweis zu S. 8.1, (2), (3)).

■

Der durch F. 8.1.b gegebene Algorithmus zur Bestimmung der Koeffizienten $a_l^{(k)}$ aus S. 8.1 läßt sich nun bei der Implementation der Rekonstruktionsmethoden folgendermaßen anwenden:

Bei der Bestimmung von $L^{N,RP}(\circ, \overline{w})$ muß $Q_{j-\frac{1}{2}}^N(\circ, W)$ in der Form aus der Voraussetzung zu F. 8.1 bestimmt werden mit

$$m := k_m^{(N)} - \tfrac{1}{2},\ x_l := a + (l - \tfrac{1}{2})\, h,\ h > 0,\ a \in \mathbb{R}.$$

Transformiere dieses in die Form von F. 8.1.a mit

$$c := (k_m^{(N)} - \tfrac{1}{2})\, h,\ \tilde{x}_l := (l - \tfrac{1}{2})\, h.$$

Wegen der speziellen Wahl von c ist $\tilde{x}_l$ unabhängig von j, d.h. die Koeffizienten $a_l^{(k)}$ aus S. 8.1 sind für alle $j \in \mathbb{Z}$ die selben und brauchen somit nur *einmal* berechnet zu werden und zwar *vor* dem eigentlichen Verfahren.

Bei der RD–Methode ist die Polynome $Q_j^N(\circ, \overline{w}),\ j \in \mathbb{Z}$, zu berechnen. Diese transformiert man analog zu oben mit

$$m := k_m^{(N)},\ x_l := a + l\, h,\ c := m\, h, \tilde{x}_l := x_l,\ h > 0,\ a \in \mathbb{R}.$$

Dann sind die $a_l^{(k)}$ wiederum unabhängig von $j \in \mathbb{Z}$ bestimmbar.

8.2. Randbedingungen

Der Nachteil eines $2q+1$–Punkte–Verfahrens in Erhaltungsform ist der, daß von einem Zeitschritt auf den nächsten links und rechts des Kompaktums $[a,b]$, auf dem die Lösung eines Erhaltungssatzes approximiert werden soll, jeweils q Punkte verloren gehen. Dies sind bei n zu berechnenden Zeitschritten $2nq$ Punkte, d.h. die Zahl der verlorenen Punkte ist proportional zu der Anzahl der Zeitschritte. Um dies zu vermeiden, ist es sinnvoll mit Randbedingungen zu arbeiten, d.h.:

1) Berechne $v_{-q}^0, \ldots, v_{-1}^0$ bzw. $v_{m+1}^0, \ldots, v_{m+q}^0$ aus den Anfangsdaten $u_0(\circ)$;
   $x_0 := a,\ x_m := b,\ a < b,\ h := \frac{b-a}{m},\ m \in \mathbb{N}$;
2) a) Berechne $v_0^{n+1}, \ldots, v_m^{n+1}$ gemäß dem gegebenen Schema aus $v_{-q}^n, \ldots, v_{m+q}^n$;
   b) Bestimme $v_{-q}^{n+1}, \ldots, v_{-1}^{n+1}$ bzw. $v_{m+1}^{n+1}, \ldots, v_{m+q}^0$ aus $v_0^{n+1}, \ldots, v_m^{n+1}$.

Dazu seien zwei Beispiele angegeben:

(a) **Stückweise konstante Anfangsdaten:**

$$u_0(x) := \begin{Bmatrix} u_L,\ x < 0 \\ u_R,\ x > 0 \end{Bmatrix},\ u_L, u_R \in \mathbb{R};$$

$$v_{-q}^{n+1} = \cdots = v_{-1}^{n+1} := v_0^{n+1};\ v_{m+1}^{n+1} = \cdots = v_{m+q}^{n+1} := v_m^{n+1}.$$

Dies ist nur solange zulässig, wie die charakteristische Kurven nicht aus dem Kompaktum $[a,b] \times [0,T],\ T > 0$, hinauslaufen.

(b) **Periodische Anfangsdaten:**

$$u_0(x) \overset{!}{=} u_0(x + (b-a));$$

$$v_j^{n+1} := \begin{Bmatrix} v_{m+j \bmod m}^{n+1} & , j = -q, \ldots, -1 \\ \\ v_{j \bmod m}^{n+1} & , j = m+1, \ldots, m+q \end{Bmatrix}.$$

8.3. Laufzeiten

Zum Abschluß dieser Arbeit soll nun noch etwas über das Laufzeitverhalten der ENO–Verfahren gesagt werden. Als Testproblem wurde das s.g. Sodproblem (vgl. (7.3)) zu den Parametern

$$a = -0.5,\ b = 0.5,\ \lambda = 0.2,\ h = 0.025$$

gewählt.

Um die gleiche Ordnung zu erzielen (s. Kap. 2.1.5), muß bei Verwendung der RP–Methode der Grad des ENO–Splines um eins höher gewählt werden als bei der RD–Rekonstruktion. Deshalb wurden jeweils die Rekonstruktionsmethoden $L^{1,RD}(\circ, v_j^n)$ bzw. $L^{2,RP}(\circ, v_j^n)$ mit den numerischen Flüssen $F_{j+\frac{1}{2}}^{Roe}$, $F_{j+\frac{1}{2}}^{Rav}$ und $F_{j+\frac{1}{2}}^{God}$ verwendet. In Tab. 3 und Tab. 4 werden die pro Zeitschritt benötigten Laufzeiten in Sekunden angegeben. Es läßt sich folgendes ablesen:

1) bei $F^{Roe}_{j+\frac{1}{2}}$ sind die Laufzeiten pro Zeitschritt konstant;
2) bei $F^{God}_{j+\frac{1}{2}}$ nimmt die Laufzeit von Zeitschritt zu Zeitschritt geringfügig zu;
3) bei $F^{Rav}_{j+\frac{1}{2}}$ steigt die Laufzeit von einem Zeitschritt zu dem nächsten erheblich;
4) bei Verwendung von $L^{1,RD}(\circ, v^n_j)$ sind die Laufzeiten geringer als bei $L^{2,RP}(\circ, v^n_j)$.

Anhang

	m	N=1		N=2		N=3	
Rek.		Fehler	P(mw)	Fehler	P(mw)	Fehler	P(mw)
RP	10	$3.846 \cdot 10^{-1}$		$2.090 \cdot 10^{-1}$		$5.275 \cdot 10^{-2}$	
			0.789		1.049		2.586
	20	$2.222 \cdot 10^{-1}$		$1.010 \cdot 10^{-1}$		$8.784 \cdot 10^{-3}$	
			0.927		1.249		2.447
	40	$1.170 \cdot 10^{-1}$		$4.251 \cdot 10^{-2}$		$1.610 \cdot 10^{-3}$	
			0.963		1.279		2.290
	80	$6.007 \cdot 10^{-2}$		$1.751 \cdot 10^{-2}$		$3.294 \cdot 10^{-4}$	
			0.981		1.295		2.169
	160	$3.043 \cdot 10^{-2}$		$7.140 \cdot 10^{-3}$		$7.325 \cdot 10^{-5}$	
RD	10	$2.001 \cdot 10^{-1}$		$4.811 \cdot 10^{-2}$		$3.107 \cdot 10^{-2}$	
			0.986		2.483		2.251
	20	$1.010 \cdot 10^{-1}$		$8.607 \cdot 10^{-3}$		$6.528 \cdot 10^{-3}$	
			1.249		2.431		2.325
	40	$4.251 \cdot 10^{-2}$		$1.596 \cdot 10^{-3}$		$1.302 \cdot 10^{-3}$	
			1.279		2.282		2.178
	80	$1.751 \cdot 10^{-2}$		$3.282 \cdot 10^{-4}$		$2.878 \cdot 10^{-4}$	
			1.295		2.166		2.088
	160	$7.140 \cdot 10^{-3}$		$7.317 \cdot 10^{-5}$		$6.769 \cdot 10^{-5}$	

Tab. 1.a: Numerische Konvergenzordnung bzgl. der exakten Lösung in der L_∞–Norm

	m	N=1		N=2		N=3	
Rek.		Fehler	P(mw)	Fehler	P(mw)	Fehler	P(mw)
RP	10	$3.691 \cdot 10^{-1}$		$1.935 \cdot 10^{-1}$		$3.719 \cdot 10^{-2}$	
			0.756		0.997		2.991
	20	$2.185 \cdot 10^{-1}$		$9.693 \cdot 10^{-2}$		$4.677 \cdot 10^{-3}$	
			0.913		1.224		3.004
	40	$1.160 \cdot 10^{-1}$		$4.148 \cdot 10^{-2}$		$5.829 \cdot 10^{-4}$	
			0.956		1.265		3.009
	80	$5.982 \cdot 10^{-2}$		$1.726 \cdot 10^{-2}$		$7.239 \cdot 10^{-5}$	
			0.978		1.286		3.008
	160	$3.037 \cdot 10^{-2}$		$7.076 \cdot 10^{-3}$		$9.001 \cdot 10^{-6}$	
RD	10	$1.845 \cdot 10^{-1}$		$3.254 \cdot 10^{-2}$		$1.550 \cdot 10^{-2}$	
			0.929		2.854		2.679
	20	$9.693 \cdot 10^{-2}$		$4.500 \cdot 10^{-3}$		$2.421 \cdot 10^{-3}$	
			1.224		2.984		3.138
	40	$4.148 \cdot 10^{-2}$		$5.688 \cdot 10^{-4}$		$2.750 \cdot 10^{-4}$	
			1.265		2.996		3.157
	80	$1.726 \cdot 10^{-2}$		$7.128 \cdot 10^{-5}$		$3.084 \cdot 10^{-5}$	
			1.286		2.999		3.164
	160	$7.076 \cdot 10^{-3}$		$8.916 \cdot 10^{-6}$		$3.442 \cdot 10^{-6}$	

Tab. 1.b: Numerische Konvergenzordnung bzgl. der Mittelwerte in der L_∞–Norm

	m	N=1		N=2		N=3	
Rek.		Fehler	P(ex)	Fehler	P(ex)	Fehler	P(ex)
RP	10	$4.979 \cdot 10^{-1}$		$1.925 \cdot 10^{-1}$		$6.249 \cdot 10^{-1}$	
			0.825		1.281		2.524
	20	$2.811 \cdot 10^{-1}$		$7.923 \cdot 10^{-2}$		$1.086 \cdot 10^{-2}$	
			0.981		1.703		2.421
	40	$1.487 \cdot 10^{-1}$		$2.434 \cdot 10^{-2}$		$2.028 \cdot 10^{-3}$	
			0.960		1.821		2.280
	80	$7.645 \cdot 10^{-2}$		$6.890 \cdot 10^{-3}$		$4.177 \cdot 10^{-4}$	
			0.980		1.881		2.165
	160	$3.874 \cdot 10^{-2}$		$1.870 \cdot 10^{-3}$		$9.315 \cdot 10^{-5}$	
RD	10	$1.874 \cdot 10^{-1}$		$6.227 \cdot 10^{-2}$		$3.093 \cdot 10^{-2}$	
			1.242		2.518		2.510
	20	$7.923 \cdot 10^{-2}$		$1.086 \cdot 10^{-2}$		$5.431 \cdot 10^{-3}$	
			1.703		2.422		2.046
	40	$2.432 \cdot 10^{-2}$		$2.028 \cdot 10^{-3}$		$1.315 \cdot 10^{-3}$	
			1.820		2.280		2.006
	80	$6.890 \cdot 10^{-3}$		$4.177 \cdot 10^{-4}$		$3.273 \cdot 10^{-4}$	
			1.881		2.165		2.001
	160	$1.870 \cdot 10^{-3}$		$9.315 \cdot 10^{-5}$		$8.181 \cdot 10^{-5}$	

Tab. 2.a: Numerische Konvergenzordnung bzgl. der exakten Lösung in der L_1–Norm

	m	N=1		N=2		N=3	
Rek.		Fehler	P(mw)	Fehler	P(mw)	Fehler	P(mw)
RP	10	$4.777 \cdot 10^{-1}$		$1.723 \cdot 10^{-1}$		$4.234 \cdot 10^{-2}$	
			0.792		1.163		2.898
	20	$2.759 \cdot 10^{-1}$		$7.699 \cdot 10^{-2}$		$5.680 \cdot 10^{-3}$	
			0.904		1.643		2.974
	40	$1.474 \cdot 10^{-1}$		$2.464 \cdot 10^{-2}$		$7.227 \cdot 10^{-4}$	
			0.954		1.808		2.994
	80	$7.612 \cdot 10^{-2}$		$7.039 \cdot 10^{-3}$		$9.071 \cdot 10^{-5}$	
			0.977		1.884		2.999
	160	$3.868 \cdot 10^{-2}$		$1.907 \cdot 10^{-3}$		$1.135 \cdot 10^{-5}$	
RD	10	$1.672 \cdot 10^{-1}$		$4.212 \cdot 10^{-2}$		$1.402 \cdot 10^{-2}$	
			1.119		2.890		3.020
	20	$7.699 \cdot 10^{-2}$		$5.682 \cdot 10^{-3}$		$1.729 \cdot 10^{-3}$	
			1.644		2.975		3.848
	40	$2.463 \cdot 10^{-2}$		$7.227 \cdot 10^{-4}$		$1.200 \cdot 10^{-4}$	
			1.807		2.994		3.799
	80	$7.039 \cdot 10^{-3}$		$9.071 \cdot 10^{-5}$		$8.629 \cdot 10^{-6}$	
			1.884		2.999		3.886
	160	$1.907 \cdot 10^{-3}$		$1.135 \cdot 10^{-5}$		$5.840 \cdot 10^{-7}$	

Tab. 2.b: Numerische Konvergenzordnung bzgl. der Mittelwerte in der L_1–Norm

n	F^{God}	F^{Roe}	F^{Rav}
1	09 : 11	09 : 65	08 : 23
2	09 : 28	09 : 72	08 : 77
3	09 : 34	09 : 77	09 : 23
4	09 : 39	09 : 78	09 : 33
5	09 : 44	09 : 76	09 : 67
6	09 : 45	09 : 77	09 : 60
7	09 : 49	09 : 78	10 : 09
8	09 : 49	09 : 71	09 : 94
9	09 : 60	09 : 72	09 : 93
10	09 : 55	09 : 76	10 : 38
11	09 : 61	09 : 73	10 : 76
12	09 : 61	09 : 77	10 : 70
13	09 : 66	09 : 72	10 : 99
14	09 : 72	09 : 78	10 : 82
15	09 : 67	09 : 77	10 : 81
16	09 : 71	09 : 71	10 : 82
17	09 : 72	09 : 71	10 : 93
18	09 : 72	09 : 72	11 : 42
19	09 : 78	09 : 72	11 : 70
20	09 : 78	09 : 77	11 : 59
21	09 : 76	09 : 72	11 : 91
22	09 : 78	09 : 72	11 : 85
23	09 : 78	09 : 77	11 : 81
24	09 : 82	09 : 72	12 : 19
25	09 : 82	09 : 72	12 : 19
26	09 : 83	09 : 77	12 : 41
27	09 : 89	09 : 72	12 : 40
28	09 : 82	09 : 72	12 : 41
29	09 : 83	09 : 72	12 : 57
30	09 : 87	09 : 72	12 : 58
31	09 : 88	09 : 72	12 : 74
32	09 : 89	09 : 72	12 : 73
33	09 : 87	09 : 71	13 : 23
34	09 : 89	09 : 72	13 : 29
35	09 : 93	09 : 71	12 : 91
36	09 : 94	09 : 71	13 : 23
37	09 : 94	09 : 72	13 : 22
38	10 : 00	09 : 67	13 : 68
39	09 : 94	09 : 67	13 : 18
40	09 : 99	09 : 71	13 : 40

Tab. 3: Laufzeiten in *sec* für ENO–Verf. mit $L^{1,RD}(\circ, v_j^n)$

n	F^{God}	F^{Roe}	F^{Rav}
1	$09:83$	$10:27$	$08:62$
2	$09:93$	$10:48$	$09:17$
3	$10:05$	$10:44$	$09:54$
4	$10:05$	$10:48$	$09:73$
5	$10:15$	$10:49$	$10:05$
6	$10:16$	$10:42$	$09:98$
7	$10:22$	$10:48$	$10:33$
8	$10:22$	$10:50$	$10:27$
9	$10:26$	$10:42$	$10:43$
10	$10:32$	$10:44$	$10:66$
11	$10:33$	$10:49$	$10:60$
12	$10:38$	$10:42$	$10:88$
13	$10:38$	$10:43$	$11:31$
14	$10:37$	$10:44$	$11:32$
15	$10:44$	$10:44$	$11:14$
16	$10:43$	$10:42$	$11:31$
17	$10:44$	$10:43$	$11:69$
18	$10:42$	$10:44$	$11:43$
19	$10:44$	$10:43$	$11:47$
20	$10:49$	$10:44$	$11:80$
21	$10:49$	$10:49$	$11:87$
22	$10:54$	$10:44$	$11:74$
23	$10:53$	$10:43$	$12:42$
24	$10:54$	$10:44$	$12:51$
25	$10:53$	$10:43$	$12:57$
26	$10:60$	$10:50$	$12:56$
27	$10:59$	$10:43$	$12:91$
28	$10:60$	$10:38$	$12:63$
29	$10:64$	$10:38$	$12:86$
30	$10:65$	$10:44$	$12:62$
31	$10:64$	$10:43$	$13:28$
32	$10:65$	$10:39$	$13:23$
33	$10:64$	$10:37$	$13:24$
34	$10:70$	$10:44$	$13:12$
35	$10:66$	$10:43$	$13:18$
36	$10:71$	$10:44$	$13:29$
37	$10:70$	$10:38$	$12:96$
38	$10:77$	$10:38$	$14:07$
39	$10:71$	$10:43$	$13:95$
40	$10:81$	$10:38$	$13:88$

Tab. 4: Laufzeiten in sec für ENO–Verf. mit $L^{2,RP}(\circ, v_j^n)$

Abb. 1.a: $L^{3,RP}(\circ, v_j^n),\ F^{God}$

Abb. 1.b: $L^{2,RD}(\circ, v_j^n),\ F^{God}$

Abb. 1.c: TVD–Verfahren von Harten

Dichte

Abb. 2.a: $L^{1,RD}(\circ, v_j^n)$, F^{Rav}

Abb. 3.a: $L^{2,RP}(\circ, v_j^n)$, F^{Rav}

Abb. 2.b: $L^{1,RD}(\circ, v_j^n)$, F^{Roe}

Abb. 3.b: $L^{2,RP}(\circ, v_j^n)$, F^{Roe}

Abb. 2.c: $L^{1,RD}(\circ, v_j^n)$, F^{God}

Abb. 3.c: $L^{2,RP}(\circ, v_j^n)$, F^{God}

Druck

Abb. 4.a: $L^{1,RD}(\circ, v_j^n),\ F^{Rav}$

Abb. 5.a: $L^{2,RP}(\circ, v_j^n),\ F^{Rav}$

Abb. 4.b: $L^{1,RD}(\circ, v_j^n),\ F^{Roe}$

Abb. 5.b: $L^{2,RP}(\circ, v_j^n),\ F^{Roe}$

Abb. 4.c: $L^{1,RD}(\circ, v_j^n),\ F^{God}$

Abb. 5.c: $L^{2,RP}(\circ, v_j^n),\ F^{God}$

Impuls

Abb. 6.a: $L^{1,RD}(\circ, v_j^n)$, F^{Rav}

Abb. 7.a: $L^{2,RP}(\circ, v_j^n)$, F^{Rav}

Abb. 6.b: $L^{1,RD}(\circ, v_j^n)$, F^{Roe}

Abb. 7.b: $L^{2,RP}(\circ, v_j^n)$, F^{Roe}

Abb. 6.c: $L^{1,RD}(\circ, v_j^n)$, F^{God}

Abb. 7.c: $L^{2,RP}(\circ, v_j^n)$, F^{God}

Abb. 8.a: Godunov–Verfahren, Dichte

Abb. 8.b: Godunov–Verfahren, Druck

Abb. 8.c: Godunov–Verfahren, Impuls

Dichte

Druck

Abb. 9.a: $F^{Rav},\ t = 0.36$

Abb. 10.a: $F^{Rav},\ t = 0.36$

Abb. 9.b: $F^{God},\ t = 0.36$

Abb. 10.b: $F^{God},\ t = 0.36$

Abb. 9.c: $F^{Roe},\ t = 0.036$

Abb. 10.c: $F^{Roe},\ t = 0.036$

Impuls

Energie

Abb. 11.a: $F^{Rav},\ t = 0.36$

Abb. 12.a: $F^{Rav},\ t = 0.36$

Abb. 11.b: $F^{God},\ t = 0.36$

Abb. 12.b: $F^{God},\ t = 0.36$

Abb. 11.c: $F^{Roe},\ t = 0.036$

Abb. 12.c: $F^{Roe},\ t = 0.036$

Abb. 13.a: $n = 15$

Abb. 13.b: $n = 20$

Abb. 13.c: $n = 30$

Abb. 13.d: $n = 35$

Abb. 13.e: $n = 75$

Abb. 13.f: $n = 85$

Impuls

Abb. 14.a: F^{Rav}, $n = 2$

Abb. 14.b: F^{Roe}, $n = 2$

Abb. 14.c: F^{Rav}, $n = 40$

Abb. 14.d: F^{Roe}, $n = 40$

Abb. 14.e: F^{Rav}, $n = 60$

Abb. 14.f: F^{Roe}, $n = 60$

Impuls

Abb. 14.g: F^{Rav}, $n = 70$

Abb. 14.h: F^{Roe}, $n = 70$

Abb. 14.i: F^{Rav}, $n = 100$

Abb. 14.j: F^{Roe}, $n = 100$

Abb. 14.k: F^{Rav}, $n = 140$

Abb. 14.l: F^{Roe}, $n = 140$

Abb. 15: Testproblem (7.2) zu den Parametern
$x \in [0,2],\ t \in [0,2],\ h = 0.05,\ \tau = 0.0125$

Abb. 16: Testproblem (5.1) zu den Parametern
$x \in [0, 1],\ t \in [0, 0.45],\ h = 0.025,\ \tau = 0.0045$

Literatur

[Ad1] Ademes, R.

Differenzenverfahren höherer Ordnung zur Lösung eines skalaren hyperbolischen Erhaltungssatzes

Diplomarbeit, RWTH Aachen, 1988;

[Ad2] Ademes, R.

Numerik partieller Differentialgleichungen

Seminararbeit, Inst. für Geometr. und prakt. Math., RWTH Aachen, 1987;

[BLRa] Bourgeade, A., LeFloch, P., Raviart, P. A.

Approximate solution of the generalized Riemann Problem and Applications

in: Actes du Congres Hyperbolique of Saint–Etienne, 1986;

[BS] Bulirsch, R., Stoer, J.

Introduction to Numerical Analysis

Springer Verlag, korrig. 2. Auflage, 1983;

[Daf1] Dafernos, C.

Structure of solutions of the Riemann–Problem for Hyperbolic Systems of Conservation Laws

Arch. Rat. Mech. Anal., 53, S. 203–217, 1974;

[Daf2] Dafernos, C.

Generalized characteristics and the structure of solutions of the Hyperbolic Conservation Laws

Ind. Univ. Math. J., 26, S. 1097–1119, 1977;

[Di1] DiPerna, R.

Singularities of Solutions of nonlinear hyperbolic systems of Conservation Laws

Arch. Rat. Mech. Anal., 60, S. 75–100, 1975;

[Di2] DiPerna, R.

Uniqueness of Solutions to Hyperbolic Conservation Laws

Ind. U. Math. J., 28, S. 244–257, 1979;

[Ei] Einfeldt, B.

Zur Numerik stoßauflösender Verfahren

Dissertation, RWTH Aachen, 1988;

[ENO1] Chakravarthy, S. R., Harten, A., Osher, S.:

Essentially Non–Oscillatory Shock–Capturing Schemes of Arbitrarly–High Accuracy

AIAA 24th Aerospace Sciences Meeting, AIAA–86–0339, 1986;

[ENO2] Chakravarthy, S. R., Engquist, B., Harten, A., Osher, S.:
An Overview of uniformly high order accurate Essentially Non–Oscillatory Schemes, II
NASA Langley Report, 1986;

[ENO3] Chakravarthy, S. R., Engquist, B., Harten, A., Osher, S.:
Uniformly high order accurate Essentially Non–Oscillatory Schemes, III
ICASE Report No. 86–22, 1986;

[Ha1] Harten, A.
On a class of High Resolution Total–Variation–Stable Finite–Difference Schemes
Siam J. Num. Anal., 21, S. 1–23, 1984;

[Ha2] Harten, A.
High Resolution Schemes for Hyperbolic Conservation Laws
J. Comp. Phys., 49, S. 357–393, 1983;

[Ha3] Harten, A.
ENO Schemes with Subcell Resolution
ICASE Report No. 87–56, Langley Research Center, 1987;

[HaH] Harten, A., Hyman, J. M.
A self– adjusting grid for the computation of Weak Solutions of Hyperbolic Conservation Laws
Report LA 9105, Center for Nonlinear Studies, Theoretical Division, Los Alamos National Lab., 1981;

[HaLL] Harten, A., Lax, P. D.,van Leer, B.
On Upstream Differencing and Godunov–Type Schemes for Hyperbolic Conservation Laws
SIAM Review, 25, S. 35–61, 1983;

[HaOs] Harten, O., Osher, S.
Uniformly High Order Accurate Non Oscillatory Schemes, I
MRC Technical Summary Report # 2823, 1985;

[HaT] Harten, A., Tadmor, E.
On the convergence of Difference Approximation to scalar conservation Laws
NASA Contractor Report 172614

[IsKe] Isaacson, E., Keller, H.B.
Analysis of Numerical Methods
John Wiley & Sons, Inc., 1966

[Le] van Leer, B.

Towards the ultimate conservative difference schemes. A new approach to numerical convection
J. Comp. Phys., 23, S. 276–298, 1977;

[LRa] LeFloch, P., Raviart, P. A.

An asymptotic expansion for the solution of the generalized Riemann Problem
Ecole Polytechnique, Centre de Mathematiques Appliques, No. 160, 1987;

[Mü] Müller, S.

Numerik Partieller Differentialgleichungen
Seminar, Inst. für Geometr. und prakt. Math., RWTH Aachen, 1987

[Os] Osher, S.

Riemann Solvers, the entropy Condition and Difference Approximations
SINUM, 21, S. 217–235, 1984;

[OsS1] Osher, S., Shu, C.-W.

Efficient Implementation of Essentially Non–Oscillatory Shock Capturing Schemes
UCLA, CAM Report 87–11, 1987;

[OsS2] Osher, S., Shu, C.-W.

Efficient Implementation of Essentially Non Oscillatory Shock Capturing Schemes, II
ICASE Report No. NAS1–18107, 1988;

[Ro1] Roe, P. L.

Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes
J. Comp. Phys., 43, S. 357–372, 1981;

[Ro2] Roe, P. L.

Upwind Differencing Schemes for Hyperbolic Conservation Laws with source terms
in: Proceedings, 7. Int. Conference Num. Math. Fluid Dynamics,
Springer Verlag, 1981;

[Sch] Schaeffer, D. G.

A Regularity Theorem for Conservation Laws
Adv. Math., 11, S. 368–386, 1973:

[Schn] Schneiders, Robert

Hyperbolische Differentialgleichungen mit Quellterm
Diplomarbeit, RWTH Aachen, 1988;

[Smo] Smoller, J.
Shock Waves and Reaction–Diffusion Equations
Springer Verlag, 1983;

[Str] Strack, K.-G.
Zur numerischen Behandlung nichtlinearer Erhaltungssätze
Dissertation, RWTH Aachen, 1984;

[Ta] Tadmor, E.
Numerical Viscosity and the Entropy Condition for Conservative Schemes
NASA Langley, Math. Comp., 43, S. 369–381, 1984;

[TW] Tatsien, L., Wenci, Y.
Boundary value problem for quasilinear hyperbolic systems
Duke University Mathematics Series, 1985;

[Wa] Walter, W.
Gewöhnliche Differentialgleichungen
Springer Verlag, 2. Auflage, 1976;

[Wan] Wang, Z.-D.
ENO–Interpolation and Uniformly High Order Accurate ENO–Schemes
in: Proceedings, 2. Int. Conference of Hyperbolic Fluid Dynamics,
RWTH Aachen, 1988;

[Ye1] Yee, H. C.
Construction of Explicit and Implicit Symmetric TVD–Schemes and their Applications
J. Comp. Phys., 68, S. 151–179, 1987;

[Ye2] Yee, H. C.
Upwind and symmetric Shock Capturing Schemes
NASA Technical Memorandum 89464, 1987;