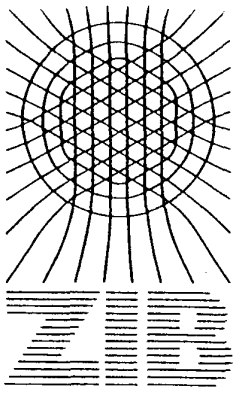
Konrad-Zuse-Zentrum für Informationstechnik Berlin



Artur Rudolf Walter

Ein Finite-Elemente-Verfahren zur numerischen Lösung von Erhaltungsgleichungen

Technical Report TR 89-3 (Februar 1989)

Abstract

A new approach for the discretisation of hyperbolic conservation laws via a finite element method is developed and analysed. Appropriate forms of the Euler's equation of gas dynamic are considered to employ the algorithm in a reasonable way for this system of nonlinear equations. Both mathematical and physical stability results are obtained. A main part of the paper is devoted to the convergence proof with energy methods under strong regularity of the solution of a scalar nonlinear conservation law. Some hints on the implementation and numerical results for the calculation of transonic gasflow through a Laval nozzle are given. The necessary amount of numerical work is compared to an established finite difference method and the efficiency of the algorithm is shown. A survey on recent literature about finite element methods for hyperbolic problem is included.

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Fachbereich Angewandte Mathematik und Informatik der Universität des Saarlandes in Saarbrücken und als freier Mitarbeiter des Siemens Forschungszentrums Abteilung Plasmaphysik in Erlangen.

Mein besonderer Dank gilt Herrn Prof. Dr. R. Rannacher, der mir die Gelegenheit gab, das von der Firma Siemens vorgeschlagene Thema an seinem Lehrstuhl zu bearbeiten. Durch seine rege Anteilnahme und durch zahlreiche Diskussionen hat er mich bei der Durchführung der Arbeit sehr unterstützt.

Außerdem möchte ich Herrn Dr. H. Blum für seine Hilfestellung insbesondere in programmiertechnischen Fragen und der Firma Siemens für die finanzielle Förderung der Arbeit danken.

Inhaltsverzeichnis

Einleitung

Die Berechnung der transsonischen Gasströmung durch eine Laval-Düse stellt ein schwieriges, praxisrelevantes Problem dar. Die Aufgabe, eine solche Strömung zu berechnen, wurde von der Abteilung für Plasmaphysik der Firma Siemens gestellt. Dort werden solche Strömungen in Hochleistungsschaltern untersucht. Das Öffnen der Kontakte in diesen Schaltern verursacht einen Lichtbogen, der den Schalter zerstören kann. Durch eine Druckdifferenz in zwei Behältern wird Schwefelhexafluorid mit Hilfe einer Laval-Düse auf Überschallgeschwindigkeit beschleunigt, um den Lichtbogen zu löschen. Da die Laval-Düse in den Schalter eingebaut ist, ändert sich während des Schaltvorganges die Gestalt des Lösungsgebietes. Deshalb ist die Zielsetzung dieser Arbeit, ein Verfahren zur numerischen Berechnung einer transsonischen Düsenströmung auf unregelmäßigen Gebieten zu entwickeln und zu analysieren.

Die instationäre, reibungsfreie und kompressible Strömung eines Fluids wird durch die Euler-Gleichungen der Gasdynamik beschrieben, die durch

$$\frac{\partial}{\partial t} u + \operatorname{div} f(u) = 0$$

gegeben sind. Dabei ist u ein Vektor, der im $\mathbb{R}^3$ fünf Komponenten hat. Der Vektor u enthält die physikalischen Größen Dichte, Impuls und Energie. Die Euler-Gleichungen der Gasdynamik gehören zur Klasse der hyperbolischen Erhaltungsgleichungen. Diese Gleichungen sind schon seit langem, insbesondere auch wegen ihrer Bedeutung in den Ingenieurwissenschaften Gegenstand intensiver mathematischer Forschung. Das durch diese Gleichungen modellierte Strömungsfeld enthält im allgemeinen Über- und Unterschallgebiete, die durch Schocks oder Kontaktunstetigkeiten getrennt sind. Entsprechend entwickeln die nichtlinearen Euler-Gleichungen aus beliebig regulären Anfangsdaten unstetige Lösungen. Wegen der komplexen Struktur dieser Lösungen werden sie vorwiegend mit numerisch-

schen Verfahren bestimmt. Dabei finden in neuerer Zeit hauptsächlich solche Methoden Verwendung, die es ermöglichen, die Lösung über einen möglichst großen Geschwindigkeitsbereich zu berechnen. Die numerische Behandlung hyperbolischer Probleme ist durch die Anwendung von Differenzenverfahren geprägt. Einen Überblick über die Grundlagen und die hauptsächlich verwendeten Algorithmen geben die Lehrbücher / 1/ bis / 4/. Da diese Verfahren schon längere Zeit entwickelt und erprobt werden, sind sie schon so robust, daß sie für den industriellen Einsatz verwendet werden können.
Ein Nachteil der Differenzenverfahren ist, daß sie auf kartesischen Gittern formuliert sind. Will man diese Verfahren auf Gebieten mit gekrümmten Rändern verwenden, so muß das kartesische Gitter im "Rechenraum" auf den "physikalischen Raum" abgebildet werden. Dies kann bei der Diskretisierung von kompliziert geformten zwei- und dreidimensionalen Gebieten Schwierigkeiten bereiten. Deshalb bieten sich Verfahren an, die auch auf unstrukturierten Gittern durchführbar sind. Solche ungleichmäßigen Zerlegungen können effizienter erzeugt werden. Diese Flexibilität bietet auch die Möglichkeit, moderne adaptive Verfahren zur Auflösung von Unstetigkeiten und starken Gradienten in der Lösung zu verwenden.
Ein solches Verfahren, das auch auf unstrukturierten Netzen durchführbar ist, ist die Methode der finiten Elemente, die im folgenden mit FEM bezeichnet wird. Nachdem sich dieses Verfahren im Bereich der elliptischen und parabolischen Differentialgleichungen etabliert hatte, began man, es auch auf Strömungsmechanische Probleme anzuwenden. Zunächst wurde es zur Berechnung von Potentialströmungen herangezogen, da diese sich wie viele Probleme der Strukturmechanik durch elliptische Gleichungen beschreiben lassen. Als nächstes wurden verschiedene Konzepte untersucht, das FEM-Verfahren für die Behandlung hyperbolischer Probleme zu nutzen. Die wichtigsten Varianten sind wohl das Taylor-Galerkin-Verfahren von Donea /67/, das dem Lax-Wendroff-Verfahren im FEM-Kontext entspricht, das darauf aufbauende "flux-corrected-tranport"-Verfahren von Löhner /70/ und die Stromliniendiffusionsmethode von Hughes und Brooks /36/. Dieses

zuletzt erwähnte Verfahren wurde als Dämpfungsstrategie entwickelt, um die bei einer zentralen Diskretisierung nicht selbstadjungierter Operatoren typischen Oszillationen der Approximierenden zu unterdrücken ohne die Konsistenzordnung zu beeinträchtigen. Es wurde zunächst für stationäre skalare hyperbolische Probleme formuliert und von Johnson /29/ auf instationäre Gleichungen übertragen. Johnson und seine Mitarbeiter, /29/ bis /33/ , haben die von ihnen gewählte Formulierung der Stromliniendiffusionsmethode analysiert. Eine numerische Verifikation dieses Verfahrens an "harten" Problemen der Strömungsmechanik steht noch aus. Alle diese Verfahren sind allerdings noch im Entwicklungsstadium und haben deshalb noch nicht diese Verbreitung und die Ausgereiftheit wie die Differenzenverfahren. Ein entscheidender Vorteil der FEM ist aber, daß sie einen systematischen Zugang sowohl zur Konstruktion als auch zur Analyse numerischer Verfahren für hyperbolische Gleichungen ermöglicht. Bei der Herleitung von Differenzenverfahren geht man meist von eindimensionalen skalaren Erhaltungsgleichungen aus und formuliert hierfür unter der Annahme einer glatten Lösung ein numerisches Verfahren. Die Übertragung der Lösungsstrategie auf Systeme und mehrdimensionale Probleme geschieht dann ad hoc. Für viele Differenzenverfahren, wie z.B. bei der Methode der angenäherten Faktorisierung /72/ ist es bis jetzt noch unklar, wie sie im $\mathbb{R}^3$ zu formulieren ist. Die FEM-Algorithmen gehen immer von der schwachen Formulierung des Problems aus. Damit ist es möglich, sie direkt für Systeme in mehreren Raumdimensionen zu formulieren. Die Analyse der so gewonnenen Algorithmen kann deshalb auch für solche Lösungen durchgeführt werden, die die zu erwartende geringe Regularität besitzen.

Die FEM kann auch dazu dienen, bereits bestehende Differenzenverfahren unter schwächeren Voraussetzungen an die Lösung zu analysieren. So lassen sich z. B. die "flux-limiter"-Verfahren als Petrov-Galerkin-Verfahren interpretieren und können so mit den in dieser Arbeit verwendeten Mitteln untersucht werden. Es gibt mehrere Möglichkeiten, die Stromliniendiffusionsmethode auf instationäre Probleme zu übertragen. Da das Ziel

die Berechnung einer Düsenströmung mit den am Lehrstuhl vorhandenen Mitteln war, wurde das Verfahren auf eine andere Art zur Diskretisierung der Euler-Gleichungen verwandt, als es von Johnson /30/ vorgeschlagen wird. Wegen der anwendungsorientierten Ausrichtung folgt das hier entwickelte Verfahren dem im ingenieurwissenschaftlichen Bereich üblichen und auch von Hughes /36/, /41/ verwendeten Vorgehen, die Orts- und Zeitdiskretisierung getrennt durchzuführen. Dieser Zugang ist auch sehr viel weniger aufwendig, als das von Johnson vorgeschlagene Galerkin-Verfahren, das in der Zeit unstetige Ansatzfunktionen benutzt.

Ein Nachteil der FEM ist der im Vergleich zu Differenzenverfahren höhere Programmier- und Speicheraufwand, da zur Charakterisierung eines unstrukturierten Gitters mehr Größen erforderlich sind, als bei regelmäßigen Netzen. Die Netzverwaltung, die Auswertung der Kubaturformeln und die kompliziertere Struktur der entstehenden Gleichungssysteme bedingen einen höheren Speicherplatzbedarf und längere Rechenzeiten.

Zur Berechnung transsonischer Strömungen, wie sie hier durchgeführt werden, bieten sich implizite Verfahren an. Bei diesen Verfahren ist die Beschränkung der Zeitschrittweite nicht so restriktiv wie bei expliziten Verfahren. Stellt sich eine stationäre Lösung ein, so kann das Zeitschrittverfahren als eine Variante des Newton-Raphson-Verfahrens zur Linearisierung der Euler-Gleichungen interpretiert werden. In diesem Fall ist es dann möglich, die stationäre Lösung nach weniger Zeitschritten als mit einem expliziten Zeitschrittverfahren zu erhalten. Ein implizites Verfahren verlangt zur Durchführung eines Zeitschrittes die Invertierung eines großen linearen Gleichungssystems. Auf kartesischen Gittern wurden für Differenzenverfahren effiziente "splitting"-Algorithmen oder ADI-Verfahren entwickelt. Sie nutzen die regelmäßige Struktur der block-pentadiagonalen Gleichungssysteme aus, die auf unstrukturierten Gittern nicht mehr vorhanden ist. Deshalb ist ein Hauptziel dieser Arbeit, den Lösungsalgorithmus für die Euler-Gleichungen so zu formulieren, daß die entstehenden linearen Gleichungssysteme iterativ lösbar sind.

Die vorliegende Arbeit ist in die folgenden Abschnitte gegliedert:
Im ersten Kapitel werden die Euler-Gleichungen der Gasdynamik und ihre verschiedenen Formulierungen besprochen, wobei besonders auf die symmetrische Form dieses hyperbolischen Systems eingegangen wird. Es werden die Bedingungen angegeben, unter welchen das System der Euler-Gleichungen symmetrisierbar ist. Nach der Beschreibung des hier betrachteten gasdynamischen Anfangs- Randwertproblems, der Strömung eines idealen Gases durch eine Laval-Düse, wird untersucht, in wie weit die zur Symmetrisierung notwendigen Annahmen hierfür erfüllt sind. Mit Hilfe der Theorie von Kreiss /62/ , die Aussagen über die Wohlgestelltheit von Anfangs- Randwertproblemen bei hyperbolischen Systemen macht, werden dann die Randbedingungen, die in diesem Fall an das System der Euler-Gleichungen zu stellen sind diskutiert.

Das zweite Kapitel enthält die Beschreibung der in dieser Arbeit entwickelten Variante der Stromliniendiffusionsmethode und deren Analyse. An Hand einer linearen Modellgleichung, der Advektionsgleichung, werden die prinzipiellen Schwierigkeiten bei der Standarddiskretisierung nicht selbstadjungierter Differentialoperatoren aufgezeigt. Die klassischen Stabilisierungsmethoden sind mit Konvergenz- oder Konsistenzordnungsverlust verbunden. Es zeigt sich, daß das hier verwendete FEM-Verfahren konsistent ist, und seine Konvergenz nur eine halbe Ordnung schlechter ist als die der Bestapproximation. Unter einer zusätzlichen Vorzeichenbedingung an die Koeffizienten der Advektionsgleichung ergibt sich, daß der Diskretisierungsfehler höchstens linear in der Zeit wächst.
Als nächstes wird das Verfahren auf die eindimensionale Burger-Gleichung angewandt. Im Fall einer glatten Lösung können die gleichen Aussagen, wie für das lineare Modellproblem bewiesen werden. Die Annahme einer glatten Lösung ist im allgemeinen natürlich unrealistisch. Es wird gezeigt, daß ohne diese Voraussetzung das Verfahren immer noch gegen eine schwache Lösung der Burger-Gleichung konvergiert, die die Entropiebedingung erfüllt. Bei den Beweisen wird die von DiPerna und Tartar eingeführte Methode der "kompensierten

schwache Lösung der Burger-Gleichung konvergiert, die die Entropiebedingung erfüllt. Bei den Beweisen wird die von DiPerna und Tartar eingeführte Methode der "kompensierten Kompaktheit" verwendet.

Im dritten Kapitel wird das Stromliniendiffusionsverfahren für das System der Euler-Gleichungen formuliert. Dabei wird deutlich, daß man auf die Symmetrisierung des hyperbolischen Systems nicht verzichten kann. Nur in dieser Form läßt sich das Verfahren so auf das System der Euler-Gleichungen anwenden, daß eine physikalische Interpretation der entstehenden Gleichungen möglich ist.

Das vierte Kapitel enthält die Beschreibung der Implementierung des entwickelten FEM-Verfahrens. Dabei wurde aus den schon oben aufgeführten Gründen der numerischen Effizienz größte Bedeutung beigemessen. Allerdings wurden dabei nur solche Aspekte berücksichtigt, die die Durchführung des Verfahrens betreffen. Maschinenabhängige Optimierungsmöglichkeiten wie z.B. die Vektorisierbarkeit des Codes blieben außer Betracht. Die bei der Durchführung eines FEM-Verfahrens immer auftretenden Aufgaben der Netzverwaltung, der Bestimmung der Zeigerstruktur und der Auswertung der Ansatzfunktionen wurden mit Hilfe des Programmpakets SAFE2 /76/ gelöst.

Im fünften Kapitel wird die praktische Realisierbarkeit der Stromliniendiffusionsmethode gezeigt. Die numerische Dissipation des Verfahrens wird am "rotating cone" Problem veranschaulicht. Dabei handelt es sich um eine zweidimensionale Advektionsgleichung, deren Lösung in einem abgeschlossenen Gebiet bleibt. Dadurch kann sie über längere Zeitintervalle berechnet werden, ohne das Wechselwirkungen mit dem Rand des Lösungsgebietes auftreten. Zu diesem Problem gibt es viele Referenzen, die den Vergleich des FEM-Verfahrens mit Differenzenverfahren ermöglichen.

Die Berechnung der transsonischen Gasströmung wurde an einer Laval-Düse mit einem Ein- und Ausströmbereich durchgeführt. Die Düsenkontur war durch eine Parabel gegeben. An dieser Konfiguration wurden verschiedene Randbedingungen erprobt und die Resultate mit der quasieindimensionalen Näherung verglichen. In einem durch die Stabilität und die Durchführbarkeit des Verfahrens begrenzten Bereich wurden die Parameter der Stromliniendiffusionsmethode variiert. Sämtliche Berechnungen wurden auf einem SIEMENS-PC MX-2 mit einem Speicherausbau von 3 MB durchgeführt. Wegen dieser beschränkten Computerkapazität wurden die zweidimensionalen Euler-Gleichungen berechnet. Das Verfahren ist aber ohne wesentliche Änderungen auch für dreidimensionale Strömungen verwendbar.

Abschließend wird eine zusammenfassende Bewertung der Resultate und ein Ausblick auf zukünftige Anwendungs- und Entwicklungsmöglichkeiten der verwendeten Methode gegeben.

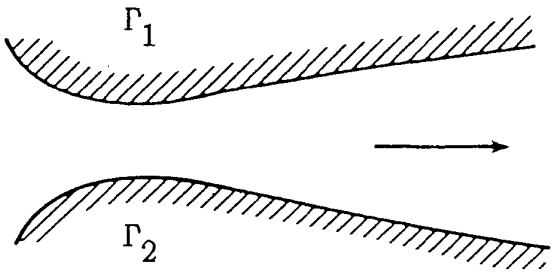
Das angefügte Literaturverzeichnis enthält mehr Literaturstellen als in der Arbeit verwendet wurden. Es soll dem interessierten Leser einen Überblick über FEM-Verfahren bei hyperbolischen Problemen bieten.

1. Die Euler-Gleichungen der Gasdynamik

In diesem Kapitel werden die Euler-Gleichungen der Gasdynamik formuliert und insbesondere die verschiedenen Darstellungsformen dieses Systems von Erhaltungsgleichungen diskutiert. Für eine ausführliche Herleitung dieser Gleichungen sei auf die umfangreiche Lehrbuchliteratur verwiesen. Darunter insbesondere auf Chorin, Marsden /18/ und Courant, Friedrichs /19/, die mehr mathematisch orientiert sind und für eine ingenieurwissenschaftliche Darstellung auf Zierep /16/ .

Im ersten Abschnitt werden die Bewegungsgleichungen eines Fluids angegeben und als nächstes die notwendigen thermodynamischen Hilfsmittel bereitgestellt. Dann wird die in vielen Aspekten noch offene Frage diskutiert, für welche der physikalisch sinnvollen Anfangs- und Randbedingungen sich eine beschränkte Lösung der Euler-Gleichungen einstellt. Wie in der Arbeit von Schochet /77/ gezeigt wird, ist die Frage nach der Wohlgestelltheit des Problems und der Existenz von Lösungen eng verbunden mit der Frage, ob man das System der Euler-Gleichungen als symmetrisches hyperbolisches System schreiben kann. Im Fall eines symmetrischen hyperbolischen Systems kann man, wie es auch in dieser Arbeit zur Analyse der Randbedingungen getan wird, die von Friedrichs /61/ entwickelten Hilfsmittel benutzen. Es wird gezeigt, daß diese Symmetrisierung unter bestimmten Bedingungen an die Lösung möglich ist. Als letztes werden die Euler-Gleichungen "quasieindimensional" formuliert, d.h. man betrachtet eine zweidimensionale Gasströmung in einem achsensymmetrischen Gebiet, wie es z.B. ein Rohr oder eine Laval-Düse ist. Aus dieser Formulierung können dann leichter Schlüsse über das Verhalten der Unstetigkeiten, die in diesem Fall durch den Wechsel von Überschall- zu Unterschallströmung gegeben sind, gezogen werden.

Diese Ergebnisse und die der folgenden Abschnitte werden auf das hier gestellte *gasdynamische Problem* einer Innenraumströmung angewendet.



Dabei sollen wie üblich gewisse Einfluß- und Ausflußgrößen und die physikalischen Eigenschaften der festen Wände Γ_1 und Γ_2 vorgegeben sein.

1.1 Die Bewegungsgleichungen

Bei der Aufstellung der Bewegungsgleichungen wird vorausgesetzt, daß

a.) die Masse des Fluids erhalten bleibt,

b.) das zweite Newtonsche Gesetz, $F = m \cdot a$, gilt und

c.) die Energie im System erhalten bleibt.

Außerdem soll die *Kontinuumshypothese* erfüllt sein, d.h. für jeden Ortspunkt $x \in \Omega \subset \mathbb{R}^2$ soll zu jedem Zeitpunkt $t \in \mathbb{R}_+$ eine *Massendichte* $\rho(x,t)$ definiert sein. Ist $u(x,t)$ das *Geschwindigkeitsfeld* des Fluids dann bedeutet die Forderung der Massenerhaltung, daß für alle genügend glatten Teilgebiete D von Ω gilt

$$\frac{d}{dt} \int_D \rho(x,t)\, dx = - \int_{\partial D} \rho(x,t)\, u(x,t) \cdot n\, ds \ .$$

Dabei ist n die äußere Normale an den Rand ∂D von D. Mit dem Gaußschen Integralsatz folgt

(1.1.1) $$\int_D \frac{\partial}{\partial t} \rho(x,t) + \operatorname{div}(\rho u)\, dx = 0 \ .$$

Da dies für alle glatten Teilgebiete D gilt, erhält man daraus die differentielle Form der *Massenerhaltungsgleichung*

(1.1.2) $$\frac{\partial}{\partial t} \rho(x,t) + \operatorname{div}(\rho u) = 0 \ .$$

Diese Gleichung wird auch oft als *Kontinuitätsgleichung* bezeichnet. Werden diese Gleichungen im klassischen Sinne verstanden, so setzt die Form (1.1.1) weniger Regularität von u und ρ voraus als (1.1.2). Deshalb geht man bei der Diskussion des Verhaltens dieser Größen an Unstetigkeiten, z. B. bei der Ableitung der Rankine-Hugoniot-Gleichungen von der Formulierung (1.1.1) aus.

Die Formulierung der Impulsgleichung soll hier für ein *ideales Fluid* angegeben werden. Ein ideales Fluid hat die Eigenschaft, daß für jeden Bewegungszustand des Fluids eine *Druckfunktion* $p(x,t)$ existiert, mit der sich die Oberflächenkraft schreiben läßt als

$$F = \int_{\partial\Omega} p(x,t)\, ds \ ,$$

wobei ds das Einheitsoberflächenelement ist. Damit ist insbesondere keine tangentiale Kraftübertragung möglich. Dies schließt eine Haftung des Fluids an festen Wänden oder Reibungskräfte durch viskose Effekte aus. (Diese Kräfte werden in den Navier-Stokes-Gleichungen berücksichtigt.) Mit dieser Voraussetzung lautet die *Impulsgleichung*

(1.1.3) $$\rho\,(u_{,t} + u\cdot\nabla u) = -\nabla p + \rho b \ ,$$

wobei b die Volumenkraft pro Einheitsmasse ist. Die verkürzende Schreibweise $u_{,t}$ soll hier und im weiteren die Ableitung nach der tiefgestellten Größe bedeuten. Sie wird neben der ausführlichen Schreibweise $\frac{\partial}{\partial t}$ benutzt. Mit Hilfe der Kontinuitätsgleichung (1.1.2) kann man (1.1.3) auch schreiben als

(1.1.4) $$(\rho u)_{,t} + u \operatorname{div}(\rho u) + \rho(u\cdot\nabla)\, u = -\nabla p + \rho b \ .$$

Die integrale Form der Impulserhaltung ist

$$\frac{\partial}{\partial t}\int_D \rho u \, dx = -\int_{\partial D} (pn + \rho u(u\cdot n)) \, ds + \int_D \rho b \, dx \quad . \tag{1.1.5}$$

Bei der Beschreibung des gestellten gasdynamischen Problems, der Gasströmung durch eine Laval-Düse, werden die Volumenkräfte, wie z.B. die Gravitation vernachlässigt.
Die *Gesamtenergiedichte* läßt sich als Summe der *inneren Energiedichte* und der *kinetischen Energiedichte* schreiben,

$$e = i + \frac{1}{2}|u|^2 \quad , \tag{1.1.6}$$

wobei $|\,.\,|$ die euklidische Vektornorm ist. Mit dem Postulat, daß dem Fluid nur durch Oberflächenkräfte Energie zugeführt werden soll, lautet die differentielle Form der *Energiegleichung*

$$(\rho e)_{,t} + \nabla\cdot(\rho e u) = -\nabla\cdot(pu) \quad , \tag{1.1.7}$$

da auch kein Wärmeübergang an $\partial\Omega$ betrachtet werden soll. Die integrale Form dieser Gleichung ist der erste Hauptsatz der Thermodynamik,

$$\int_D \frac{\partial}{\partial t} e \, dx + \int_{\partial D} e\, u\cdot n \, ds + \int_{\partial D} p\, u\cdot n \, ds = 0 \quad . \tag{1.1.8}$$

1.2 Thermodynamische Begriffe

Die hier zusammengestellten Begriffe der Thermodynamik sind hauptsächlich den Büchern von Anderson /17/, Zierep /16/ und Courant, Friedrichs /19/ entnommen, auf die wieder für eine ausführlichere Herleitung verwiesen wird. Es werden nur die thermodynamischen Geschichtspunkte aufgeführt, die für die spätere Symmetrisierung des Systems relevant sind.

Ein thermodynamischer Zustand eines Fluids ist beschrieben durch

den *Druck* p ,

die *Dichte* ρ ,

die *Temperatur* T ,

das *spezifische Volumen* τ , das durch $\rho\tau = 1$ definiert ist,

die *spezifische Entropie* S ,

die *spezifische Energie* e und

die innere Energie oder *spezifische Enthalpie* i .

Von diesen Parametern sind nur zwei unabhängig. Im weiteren werden alle Parameter als Funktionen von τ und S aufgefaßt. Unter der schon oben gemachten Annahme, daß dem Fluid nur Energie durch Temperaturerhöhung oder durch Druckerhöhung zugeführt werden kann, gilt für eine infinitesimale Energieänderung

$$de = TdS - pd\tau \quad . \tag{1.2.1}$$

Ist bekannt, wie die Energie e von τ und S abhängt, so kann man mit (1.2.1) p und T bestimmen :

$$p = -e_{,\tau} \quad \text{und} \quad T = e_{,S} \quad , \tag{1.2.2}$$

oder allgemein

$$p = f(\rho,S) = g(\tau,S) \quad . \tag{1.2.3}$$

Diese Gleichung nennt man *Zustandsgleichung* des Fluids.

Aus physikalischen Gründen fordert man, daß bei konstanter Entropie der Druck bei Volumenverkleinerung steigt, also

$$f_{,\rho}(\rho,S) > 0 \quad \text{oder} \quad g_{,\tau}(\tau,S) < 0, \tag{1.2.4}$$

außer in dem Fall $\rho = 0$ für den man $f_{,\rho} = 0$ fordert. Wegen (1.2.4) kann man jetzt eine positive Größe c , die Schallgeschwindigkeit definieren durch

$$c^2 = \frac{\partial}{\partial\rho} p = f_{,\rho}(\rho,S) \quad . \tag{1.2.5}$$

Unter der Annahme, daß das Fluid ein ideales Gas ist, lautet die Zustandsgleichung

$$p\tau = RT \quad , \tag{1.2.6}$$

wobei R die *universelle Gaskonstante* ist.

Gilt außerdem

(1.2.7) $$e = c_v T$$

so bezeichnet man das Gas als *polytrop*. Dabei heißt c_v *spezifische Wärmekapazität* bei konstantem Volumen. Man kann zeigen, daß die Forderung (1.2.7) dazu äquivalent ist, daß die Zustandsgleichung die Form

(1.2.8) $$p = A\rho^{\gamma}$$

hat; siehe z.B. Courant, Friedrichs /19/ pp. 8-10. Die Konstante γ heißt *Adiabatenkonstante*. Der Koeffizient A hängt über die Formel

$$A = (\gamma - 1)\,(\exp(S - S_0) - \exp(c_v))$$

von der Entropie S ab, wobei S_0 eine beliebige Bezugsentropie ist. Löst man (1.2.6) nach T auf und setzt das Ergebnis in (1.2.7) ein, so erhält man für die Energie

(1.2.9) $$e = \frac{p\tau}{\gamma - 1} \,.$$

Dabei wurde die Beziehung

$$c_v = \frac{R}{\gamma - 1}$$

ausgenutzt.

Um den thermodynamischen Zustand eines Fluids als Funktion von p und τ auszudrücken, benötigt man noch eine Bestimmungsgleichung für die Entropie. Aus Gleichung (1.2.8) erhält man

(1.2.10) $$S - S_0 = c_v \log\left(\frac{p\tau^{\gamma}}{\gamma - 1}\right) .$$

Im weiteren wird S_0 zu null gesetzt und die Entropie dimensionslos gemacht,

$$S := \frac{S}{c_v}$$

Der zweite Hauptsatz der Thermodynamik lautet mit diesen Voraussetzungen

(1.2.11) $$\frac{d}{dt}\int_D \rho S \, dx = 0 \,,$$

siehe z.B. Zierep /15/ p. 8 ff. Da (1.2.11) für alle glatten Teilgebiete D gelten soll, ergibt sich wie oben die differentielle Form zu

$$(\rho S)_{,t} + u \cdot \nabla(\rho S) = 0 \tag{1.2.12}$$

oder unter Verwendung der Kontinuitätsgleichung

$$S_{,t} + u \cdot \nabla S = 0 \ . \tag{1.2.13}$$

In der Thermodynamik wird (1.2.11) verwendet, um zulässige thermodynamische Prozesse zu kennzeichnen. Für sie soll gelten, daß

$$\frac{d}{dt} \int_D \rho S \, dx \geq 0 \tag{1.2.14}$$

beziehungsweise

$$(\rho S)_{,t} + u \cdot \nabla(\rho S) \geq 0 \tag{1.2.15}$$

gilt. Das ist das physikalische Analogon zur Entropiebedingung (2.6.8). Unter der Annahme, daß die Zustandsänderungen adiabatisch verlaufen, sind sie reversibel. Daraus wurde dann gefolgert, daß die substantielle Ableitung von S , d. h. (1.2.13) verschwindet. Dies bedeutet, daß die Entropie entlang Stromlinien konstant ist. Senkrecht zu den Stromlinien kann sie sich aber ändern. Damit verläuft die Strömung isentropisch. Treten Schocks in der Strömung auf, so sind diese mit Entropieerhöhung verbunden, und es gilt nur noch die Relation (1.2.15).

Für das System der Euler-Gleichungen gibt es eine Vielzahl von Darstellungsmöglichkeiten, je nachdem, welche Variablen benutzt werden. Hier soll von den sogenannten "konservativen Variablen" ausgegangen werden. Die Integrale dieser Größen bleiben wegen dem Massen-, Impuls-, und Energieerhaltungssatz unverändert. Sie lauten im $\mathbb{R}^2$

$$u^T = (\rho, \rho u, \rho v, E) := (u_1, u_2, u_3, u_4) \ ,$$

d.h. der Vektor U besteht aus der Dichte , der Impulsdichte und der Gesamtenergiedichte. Dabei sind u und v die x- beziehungsweise die y - Komponente des Geschwindigkeits-

vektors. Der Exponent T soll die *Transposition* eines Vektors oder einer Matrix bedeuten. Damit lautet das *hyperbolische System der gasdynamischen Gleichungen*

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f^x(u) + \frac{\partial}{\partial y} f^y(u) = 0 \quad , \tag{1.2.17}$$

mit den *Flüssen*

$$f^x(u)^T = u_1^{-2} \{u_1^2 u_2,\ (\gamma - 1)u_1(u_1 u_4 - \tfrac{1}{2} u_3^2) + \tfrac{1}{2}(3 - \gamma)u_1 u_2^2,$$

$$u_1 u_2 u_3,\ u_2(\gamma u_1 u_4 - \tfrac{1}{2}(\gamma - 1)(u_2^2 + u_3^2)) \}$$

und

$$f^y(u)^T = u_1^{-2} \{u_1^2 u_3,\ u_1 u_2 u_3,\ (\gamma - 1)u_1(u_1 u_4 - \tfrac{1}{2} u_2^2) + \tfrac{1}{2}(3 - \gamma)u_1 u_3^2,$$

$$u_3(\gamma u_1 u_4 - \tfrac{1}{2}(\gamma - 1)(u_2^2 + u_3^2)) \} \ .$$

Andere Schreibweisen, z. B. in den "primitiven Variablen " ρ, u, v und p und deren wechselseitige Umformungen findet man in Warming, Beam, Hyett /42/ oder in Peyret, Taylor / 3/.

1.3 Symmetrisierung der gasdynamischen Gleichungen

In den vorherigen Abschnitten wurde das System der Euler-Gleichungen für ein ideales polytropes Gas eingeführt. Dieses System gehört in die Klasse der hyperbolischen Erhaltungsgleichungen, was z.B. in Warming, Beam, Hyett /42/ gezeigt wird. Hier sollen jetzt allgemeine Systeme hyperbolischer Erhaltungsgleichungen der Form

$$u_{,t} + \operatorname{div} f(u) = 0 \tag{1.3.1}$$

betrachtet werden. Die dabei gewonnenen Ergebnisse finden dann wieder Anwendung auf die Euler-Gleichungen.

Im folgenden sei u ein m-Vektor, $f^i(u)$ eine vektorwertige Funktion mit m Komponen-

ten, $x \in \mathbb{R}^n$ und $f = (f^1,...,f^n)$. Die Gleichung (1.3.1) lautet in der *quasilinearen Form* :

$$u_{,t} + \sum_{i=1}^{n} A^i(u)\, u_{,x_i} = 0 \quad , \tag{1.3.2}$$

mit

$$A^i(u) := f^i_{,u} \quad .$$

Grundlegend für die Diskussion von Erhaltungsgleichungen sind die in den folgenden beiden Definitionen eingeführten Begriffe.

Definition 1 :

Ein skalare Funktion $U(u)$ heißt eine *Entropiefunktion* für das System (1.3.2), wenn

a.) Die Funktion U ist eine konvexe Funktion von u ist, und

b.) die Funktion U erfüllt

$$U_{,u} f^i_{,u} = F^i_{,u} \quad , \text{ für } i = 1,...,n \quad ,$$

wobei $F^i(u)$ eine skalare Funktion ist, die *Entropiefluß* in Richtung x_i heißt.

■

Definition 2 :

Ein System von Gleichungen der Form

$$B^0 v_{,t} + \sum_{i=1}^{n} B^i v_{,x_i} = 0 \tag{1.3.3}$$

heißt *symmetrisch hyperbolisch*, wenn alle Matrizen B^i symmetrisch sind und falls die Matrix B^0 positiv definit ist.

■

Für symmetrische hyperbolische Systeme existiert eine umfangreiche Lösungstheorie von Lax /24/ , die viele Ergebnisse liefert, die für den unsymmetrischen Fall nicht gelten oder noch nicht bekannt sind. Deshalb wird für eine Diskussion der Euler-Gleichungen das System um einen festen Zustand z_0 in eine Taylorreihe entwickelt und die Flüße bezüglich z_0 linearisiert. Damit ist dann eine im Zustandsraum lokale Symmetrisierung der Euler--Gleichungen möglich und es können lokale Aussagen bzgl. der Wohlgestelltheit des Problems gemacht werden. Diese Technik wird in Kapitel 1.4 verwendet. Die Theorie von Lax für symmetrische hyperbolische Systeme und die Linearisierung der Euler-Gleichungen werden in Majda /21/ ausführlich besprochen. Inhalt dieses Abschnittes ist es, neuere Ergebnisse zur nichtlinearen globalen Symmetrisierung aufzuzeigen und sie für die Behandlung der kontinuierlichen und der diskretisierten Euler-Gleichungen zu nutzen.

Die Symmetrisierung von (1.3.2) entspricht einer Variablentransformation. Führt man die neue Varaible v ein, d.h. $v = v(u)$ dann ist

$$u(v)_{,t} + \sum_{i=1}^{n} f^i(u(v)))_{,x_i} = u_{,v} v_{,t} + \sum_{i=1}^{n} f^i_{,v} v_{,x_i} = 0 \quad . \tag{1.3.4}$$

Damit hat (1.3.4) die Gestalt (1.3.3) mit

$$B^0 = u_{,v} \quad \text{und} \quad B^i = f^i_{,v} \quad .$$

Aus der Symmetrie der Matrizen $u_{,v}$ und $f^i_{,v}$ folgt, daß sie sich als Gradienten von skalaren Funktionen $q(v)$ und $r^i(v)$ darstellen lassen, so daß

$$q_{,v} = u^T \quad \text{und} \quad r^i_{,v} = (f^i)^T \tag{1.3.5}$$

ist. Da $u_{,v}$ positiv definit sein soll, ist $q(v)$ eine konvexe Funktion von v . Daraus folgt, daß die Abbildung $v \to q_{,v} = u^T$ ein Isomorphismus ist. Damit ist sie umkehrbar und u eine Funktion von v .

Nicht jedes hyperbolische System hat eine Entropie. Beispiele hierfür werden in Smoller

/20/ mit den sogenannten p-Systemen gegeben. Das folgende klassische Resultat geht auf Godunov /83/ zurück.

Satz 1 (Harten, Lax /46/)Kann das hyperbolische System (1.3.2) durch eine Variablentransformation symmetrisiert werden, dann hat (1.3.2) eine Entropiefunktion U(u) , die durch

(1.3.6) $$U(u) = u^T v - q(v)$$

gegeben ist. Der zugehörige Entropiefluß $F^i(u)$ ist

(1.3.7) $$F^i(u) = (f^i)^T v - r^i(v) \ .$$

■

Wichtiger ist natürlich die Umkehrung dieses Satzes, die von Mock /84/ gegeben wurde.

Satz 2 (Harten, Lax /46/)

Ist U(u) eine Entropiefunktion von (1.3.2), dann symmetrisiert

(1.3.8) $$v^T = U_{\cdot,u}$$

das System (1.3.2).

■

Diese Ergebnisse werden jetzt auf die Euler-Gleichung (1.2.17) angewendet.

Die Jakobimatrizen der Flußfunktionen der Euler-Gleichungen (1.2.17) lauten für $A^x := - u_1^3 f^x_{,u}$:

$a_{11} = 0$ $\qquad$ $a_{12} = - u_1^3$

$a_{13} = 0$ $\qquad$ $a_{14} = 0$

$a_{21} = (\frac{3-\gamma}{2} u_2^2 + \frac{1-\gamma}{2} u_3^2)u_1$ $\qquad$ $a_{22} = (\gamma - 3)u_1^2 u_2$

$a_{23} = (\gamma - 1)u_1^2 u_3$ $\qquad$ $a_{24} = (1 - \gamma)u_1^2$

$a_{31} = u_1 u_2 u_3$ $\qquad$ $a_{32} = - u_1^2 u_3$

$a_{33} = - u_1^2 u_2$ $\qquad$ $a_{34} = 0$

$a_{41} = (\gamma u_1 u_4 + (1 - \gamma)(u_2^2 + u_3^2))u_2$ $\qquad$ $a_{42} = (- \gamma u_1 u_4 + \frac{\gamma - 1}{2} (3 u_2^2 + u_3^2))u_1$

$a_{43} = (\gamma - 1)u_1 u_2 u_3$ $\qquad$ $a_{44} = - \gamma u_1^2 u_2$

und für $A^y := -u_1^3 f^y_{,u}$:

$a_{11} = 0$ $\qquad$ $a_{12} = 0$

$a_{13} = - u_1^3$ $\qquad$ $a_{14} = 0$

$a_{21} = u_1 u_2 u_3$ $\qquad$ $a_{22} = - u_1^2 u_3$

$a_{23} = - u_1^2 u_2$ $\qquad$ $a_{24} = 0$

$a_{31} = u_1(3 - \gamma) u_3^2 + \frac{1-\gamma}{2} u_2^2$. $\qquad$ $a_{32} = (\gamma - 1) u_1^2 u_2$

$a_{33} = (\gamma - 3)u_1^2 u_3$ $\qquad$ $a_{34} = (1 - \gamma) u_1^3$

$$a_{41} = (\gamma u_1 u_4 + (1 - \gamma)(u_2^2 + u_3^2))u_3 \qquad a_{42} = (\gamma - 1)\, u_1 u_2 u_3$$

$$a_{43} = (-\gamma u_1 u_4 + \frac{\gamma - 1}{2}(3u_3^2 + u_2^2))u_1 \qquad a_{44} = -\gamma u_1^2 u_3 \ .$$

Die Jakobimatrizen haben die Eigenwerte

$$e_1^x = u_1^{-1}(u_2 - (\gamma(\gamma - 1)(u_1 u_4 - \tfrac{1}{2}(u_2^2 + u_3^2)))^{\frac{1}{2}})$$

$$e_2^x = e_3^x = u_1^{-1} u_2$$

$$e_4^x = u_1^{-1}(u_2 + (\gamma(\gamma - 1)(u_1 u_4 - \tfrac{1}{2}(u_2^2 + u_3^2)))^{\frac{1}{2}})$$

und

$$e_1^y = u_1^{-1}(u_3 - (\gamma(\gamma - 1)(u_1 u_4 - \tfrac{1}{2}(u_2^2 + u_3^2)))^{\frac{1}{2}})$$

$$e_2^y = e_3^y = u_1^{-1} u_3$$

$$e_4^y = u_1^{-1}(u_3 + (\gamma(\gamma - 1)(u_1 u_4 - \tfrac{1}{2}(u_2^2 + u_3^2)))^{\frac{1}{2}}) \ .$$

Die Gleichung (1.2.10) für die Entropie lautet in der hier verwendeten Schreibweise :

(1.3.9) $$S = \log(u_1^{-\gamma - 1}) - \log(\gamma - 1) + \log(u_1 u_4 - \tfrac{1}{2}(u_2^2 + u_3^2)) \ .$$

Der Druck p schreibt sich mit (1.2.9) als

(1.3.10) $$p = (\gamma - 1)\, u_1^{-1}(u_4 u_1 - \tfrac{1}{2}(u_2^2 + u_3^2)) \ .$$

Da der Fluß als isentropisch angenommen wurde, ist die substantielle Ableitung $\frac{D}{Dt}$ der Entropie null und es gilt (1.2.13)

$$u_1 S_{,t} + u_2 S_{,x} + u_3 S_{,y} = 0 \ .$$

Bemerkung :

Die differentielle Form der Euler-Gleichungen gilt im klassischen Sinne nur in den Gebieten, in denen die Strömung keine Unstetigkeiten enthält. Über diese Unstetigkeiten hinweg wird die Lösung mit Hilfe der Rankine-Hugoniot-Gleichungen fortgesetzt. Da bei der Herleitung der symmetrischen Form der Euler-Gleichungen alle Differentiale im klassischen Sinne zu verstehen sind, gilt die Gleichung der Entropieerhaltung (1.2.13) nur für solche Strömungen, die keine Entropiequellen enthalten. Im Falle der hier betrachteten Düsenströmung ist sie deshalb nur vor und hinter der Schallinie erfüllt. In Kapitel 1.5 wird mit Hilfe der eindimensionalen Euler-Gleichungen eine Begründung dafür gegeben, warum man für die Berechnung einer Strömung mit "kleinen" Sprüngen in den Komponenten des Lösungsvektors u trotzdem die Gültigkeit der Gleichung (1.2.13) annehmen kann.
Mögliche Umformungen hyperbolischer Systeme im Rahmen einer schwachen Lösungstheorie in $L^\infty \cap BV(\mathbb{R} \times \mathbb{R}^+)$, d.h. die Gleichung (1.3.3) soll im Raum der lokal endlichen Maße erfüllt sein, werden in der Arbeit von Floch /81/ gegeben. Aber auch dort werden, wie bei der Methode der kompensierten Kompaktheit, Techniken verwendet, die im wesentlichen auf das Eindimensionale beschränkt sind.

■

Mit den obigen Voraussetzungen gilt also für jede differenzierbare Funktion $h(S)$

$$u_1 h'(s)\frac{D}{Dt} S = 0 = u_1 h(S)_{,t} + u_2 h(S)_{,x} + u_3 h(S)_{,y} \quad , \tag{1.3.11}$$

wobei h' die Ableitung von h nach seinem Argument bedeutet.
Wird die Kontinuitätsgleichung (1.1.2) mit $-h(S)$ multipliziert und das Ergebnis von (1.3.11) subtrahiert, so erhält man

$$(-u_1 h(S))_{,t} + (-u_2 h(S))_{,x} + (-u_3 h(S))_{,y} = 0 \quad .$$

Mit der Identifikation

$$U(u) = -u_1 h(S),$$

$$F^x(u) = -u_2 h(S) \quad \text{und}$$

$$F^y(u) = -u_3 h(S)$$

ergibt sich damit die Entropiegleichung aus Definition 1 für (1.2.17). Mit der Bestimmungsgleichung (1.3.8) für die Variable v, für die das System (1.2.17) symmetrisch ist, kann man diese jetzt explizit angeben :

$$(1.3.12) \quad v^T = -(\gamma-1)\frac{h'(S)}{p}\{u_4 + \frac{p}{(\gamma-1)}(\frac{h(S)}{h'(S)} - \gamma - 1),\ -u_2,\ -u_3,\ u_1\} \ .$$

Die Jakobimatrix der Transformation ist

$$(1.3.13) \qquad v_{,u} = \left\{\frac{\gamma-1}{p}\right\}^2 u_1 h'(S) \cdot D \ ,$$

wobei die Matrix D gegeben ist durch

$$d_{11} = \tfrac{1}{4}q^4 + \frac{c^4}{\gamma} - K(\tfrac{1}{2}q^2 - c^2)^2 \qquad d_{12} = -q_1(\tfrac{1}{2}q^2(1-K) + Kc^2$$

$$d_{13} = -q_2(\tfrac{1}{2}q^2(1-K) + Kc^2) \qquad d_{14} = \tfrac{1}{2}q^2(1-K) - c^2(\tfrac{1}{\gamma} - K)$$

$$d_{22} = q_1^2(1-K) + \frac{c^2}{\gamma} \qquad d_{23} = q_1 q_2(1-K)$$

$$d_{24} = -q_1(1-K) \qquad d_{33} = q_2^2(1-K) + \frac{c^2}{\gamma}$$

$$d_{34} = -q_2(1-K) \qquad d_{44} = 1-K$$

und $D = D^T$. Dabei bedeutet $c = \frac{\gamma}{\gamma-1}\frac{p}{u_1}$ wieder die Schallgeschwindigkeit, q_1 beziehungsweise q_2 sind die x- bzw. die y-Komponenten der Geschwindigkeit und q ist deren Betrag. K ist eine Funktion von S mit $K := \frac{h''(S)}{h'(S)}$.

In der Definition symmetrischer hyperbolischer Systeme wird gefordert, daß diese Jakobimatrix der Variablentransformation positiv definit sein soll. Daraus ergibt sich eine Einschränkung an die als beliebig angenommene Funktion h . Sie wird durch den folgenden Satz gegeben. Seine Aussage läßt sich leicht verifizieren, indem man sämtliche Unterdeterminanten von D berechnet.

Satz 3

Die symmetrische Matrix D ist genau dann positiv definit, wenn

$$K < \frac{1}{\gamma}$$

ist.

■

In dem Artikel von Harten /45/ wird

(1.3.14) $$h(S) = \alpha \exp(\frac{S}{\beta + \gamma})$$

gesetzt, wobei α, β und γ zu wählende Konstanten sind. Der Vorteil dieser Entropiefunktion ist, daß die entsprechenden Flüße homogene Funktionen vom Grad r ihrer Argumente sind

$$f(\alpha v) = \alpha^r f(v) \quad \forall\, \alpha \in \mathbb{R} \ .$$

Dies ist eine wesentliche Eigenschaft, die bei der Konstruktion von "flux-splitting" Verfahren zur Diskretisierung von (1.2.17) ausgenutzt wird, siehe z.B. Steger, Warming /73/ . Diese Verfahren gehen aber von der konservativen Formulierung der Euler-Gleichungen aus. In diesen Variablen sind die Flüße auch homogene Funktionen.

Für die Anwendung hier wurde im Gegensatz dazu die "physikalische Entropie"

(1.3.15) $$h(S) = S$$

gewählt. Damit ist die Voraussetzung von Satz 3 trivialerweise erfüllt. Allerdings sind die resultierenden Flüße keine homogenen Funktionen mehr. Diese Wahl läßt aber, wie sich

zeigen wird, im Gegensatz zu (1.3.14) bei der Frage nach der Wohlgestelltheit des Problems eine physikalische Deutung zu. Für die Diskretisierung von (1.2.17) wurden sowohl (1.3.14) als auch (1.3.15) verwendet. Die numerischen Ergebnisse haben sich von der Wahl der Funktion h als weitgehend unabhängig erwiesen.

Für die hier getroffene spezielle Wahl von h läßt sich die Variable v mit Hilfe von (1.3.12) angeben.

$$v^T = \frac{1}{\rho i} \{-u_4 + \rho i(\gamma + 1 - S),\ u_2,\ u_3,\ -u_1\} \quad . \tag{1.3.16}$$

Dabei kann die innere Energie ρi aus (1.1.6) ,

$$\rho i = u_4 - \frac{1}{2u_1}(u_2^2 + u_3^2) \tag{1.3.17}$$

und die Entropie S aus (1.2.10) ,

$$S = \ln((\gamma - 1)\rho i) + \gamma \ln(u_1) \quad , \tag{1.3.18}$$

berechnet werden. Die Umkehrabbildung ist gegeben durch

$$u^T = \rho i \{-v_4,\ v_2,\ v_3,\ 1 - \frac{1}{2v_4}(v_2^2 + v_3^2)\} \quad , \tag{1.3.19}$$

wobei

$$\rho i = (\frac{1}{-\beta\ v_4^{\gamma}})^{\gamma} \exp(-S\ \beta) \qquad \text{mit } \beta := \frac{1}{\gamma - 1} \tag{1.3.20}$$

und

$$S = \gamma - v_1 + \frac{1}{2v_4}(v_2^2 + v_3^2) \tag{1.3.21}$$

ist. Daraus sieht man, daß bis auf Skalierung durch die Symmetrisierung der Euler-Gleichungen (1.2.17) nur die Dichte und die Energie vertauscht werden. Dadurch gestaltet sich die Umformung und die Berücksichtigung der Randbedingungen besonders einfach.

Die Koeffizientenmatrizen der quasilinearen Form sollen nun in der neuen Variablen v mit der Setzung

$$\alpha := \frac{1}{2v_4}(v_2^2 + v_3^2) \quad \text{und} \quad \beta := \gamma - 1$$

angegeben werden.

Die Matrix $\frac{\beta v_4}{\rho i} B^0$ lautet

$$b_{11} = -v_4^2 \qquad b_{12} = v_2 v_4$$

$$b_{13} = v_3 v_4 \qquad b_{14} = v_4(1 - \alpha)$$

$$b_{22} = \beta v_4 - v_2^2 \qquad b_{23} = -v_2 v_3$$

$$b_{24} = v_2(\alpha - \gamma) \qquad b_{33} = \beta(v_4 - v_3^2)$$

$$b_{34} = v_3(\alpha - \gamma) \qquad b_{44} = -\alpha(\alpha - 2\gamma) - \gamma \ .$$

Die Matrix $-(\rho i v_4)\,(B^0)^{-1}$ ist

$$b_{11} = \alpha^2 + \gamma \qquad b_{12} = \alpha v_2$$

$$b_{13} = \alpha v_3 \qquad b_{14} = v_4(1 + \alpha)$$

$$b_{22} = v_2^2 - v_4 \qquad b_{23} = v_2 v_3$$

$$b_{24} = v_2 v_4 \qquad b_{33} = v_3^2 - v_4$$

$$b_{34} = v_3 v_4 \qquad b_{44} = v_4^2 \ .$$

Diese Matrix wird später zur Durchführung des numerischen Lösungsverfahrens benötigt.

Die Matrix $\frac{\beta v_4^2}{\rho i} B^1$ ist gegeben durch

$$b_{11} = v_2 v_4^2 \qquad b_{12} = \beta v_4^2 - v_2^2 v_4$$

$$b_{13} = - v_2 v_3 v_4 \qquad b_{14} = (\alpha - \gamma) v_2 v_4$$

$$b_{22} = - (3\beta v_4 - v_2^2) v_2 \qquad b_{23} = - (\beta v_4 - v_2^2) v_3$$

$$b_{24} = (\beta v_4 - v_2^2)(\alpha - \gamma) + \beta v_2^2 \qquad b_{33} = - (\beta v_4 - v_3^2) v_2$$

$$b_{34} = - (\alpha - 2\gamma + 1) v_2 v_3 \qquad b_{44} = (\alpha(\alpha - 4\gamma + 2) + \gamma(2\gamma - 1)) v_2 \ .$$

Und schließlich ergibt sich für die Matrix $\frac{\beta v_4^2}{\rho i} B^2$

$$b_{11} = v_3 v_4^2 \qquad b_{12} = - v_2 v_3 v_4$$

$$b_{13} = (\beta v_4 - v_3^2) v_4 \qquad b_{14} = (\alpha - \gamma) v_3 v_4$$

$$b_{22} = - (\beta v_4 - v_2^2) v_3 \qquad b_{23} = - (\beta v_4 - v_3^2) v_2$$

$$b_{24} = - (\alpha - 2\gamma + 1) v_2 v_3 \qquad b_{33} = - (3\beta v_4 - v_3^2) v_3$$

$$b_{34} = (\beta v_4 - v_3^2)(\alpha - \gamma) + \beta v_3^2 \qquad b_{44} = (\alpha(\alpha - 4\gamma + 2) + \gamma(2\gamma - 1)) v_3 \ .$$

1.4 Anfangs- und Randbedingungen

Um das gestellte gasdynamische Problem vollständig zu beschreiben, benötigt man neben den Euler-Gleichungen (1.2.17) noch Anfangs- und Randbedingungen. Dabei erhebt sich die Frage, welche Bedingungen gefordert werden dürfen, d.h. für welche Anfangs- und Randbedingungen ist das Problem, wenigstens für glatte Lösungen, wohlgestellt im Sinne von Hadamard.

Zur Verdeutlichung der Problematik bei hyperbolischen Systemen wird zunächst ein symmetrisches hyperbolisches System in einer Raumdimension betrachtet.

(1.4.1) $$u_{,t} = A\, u_{,x} \quad , \; 0 \le x \le 1 \quad , t \ge 0 \; .$$

Ohne Beschränkung der Allgemeinheit kann man annehmen, daß die $n \times n$ - Matrix A eine Diagonalmatrix ist,

(1.4.2) $$A = \mathrm{diag}\{\Lambda_1, O, -\Lambda_3\}$$

wobei $\forall\, \lambda \in \Lambda : \lambda > 0$ und $\forall\, \lambda \in O : \lambda = 0$ sein soll. Der Rang der Teilmatrizen ist gleich der Anzahl der Eigenwerte mit dem entsprechenden Vorzeichen. Mit Hilfe von (1.4.2) zerfällt (1.4.1) in die drei Gruppen von jeweils skalaren Gleichungen

$$u^1_{,t} = \Lambda_1 u^1_{,x} \quad , \; u^2_{,t} = 0 \;\text{ und }\; u^3_{,t} = -\Lambda_3 u^3_{,x} \; .$$

Die Lösungskomponenten von u sind konstant entlang der zugehörigen Charakteristiken. Die Charakteristik, die zu u^1 gehört läuft nach links, also darf man am rechten Rand, $x = 1$, Randbedingungen stellen, z.B.

$$u^1(1,t) = g^1(t) \; ,$$

und entsprechend für u^3

$$u^3(0,t) = g^3(t) \; .$$

Zu den vorgegebenen Randbedingungen kommen noch die von den anderen Charakteristiken gelieferten Werte

$$u^1(1,t) = S^1 \begin{Bmatrix} u^2(1,t) \\ u^3(1,t) \end{Bmatrix} + g^1(t)$$

und

$$u^3(0,t) = S^3 \begin{Bmatrix} u^1(0,t) \\ u^2(0,t) \end{Bmatrix} + g^3(t) \; .$$

Die konstanten rechteckigen Matrizen S^1 und S^2 beschreiben die Reflexion von u an den Rändern des Gebiets. Allgemein kann man also Randbedingungen der Form

(1.4.3) $$R_0 u(0,t) = 0 \quad \text{und} \quad R_1 u(1,t) = 0$$

stellen, wobei die R_i rechteckige Matrizen sind. Die Anzahl der linear unabhängigen Randbedingungen am Punkt $x = 0$ ist gleich der Anzahl der negativen Eigenwerte von $A(0,t)$ und am Punkt $x = 1$ gleich der Anzahl der positiven Eigenwerte von $A(1,t)$.
Nun soll die Energie, d.h. in diesem Fall die L^2- Norm der Lösung u von (1.4.1) abgeschätzt werden. Im folgenden bezeichne

$$(u,v) := \int_0^1 uv \, dx \quad \text{und} \quad \| u \|^2 := (u,u) \ .$$

Für die Lösung u von (1.4.1) gilt

(1.4.4) $$\frac{d}{dt} \|u\|^2 = \frac{\partial}{\partial t} (u,u) = (\frac{\partial}{\partial t} u, u) + (u, \frac{\partial}{\partial t} u) = (A\frac{\partial}{\partial x} u, u) + (u, A\frac{\partial}{\partial x} u)$$

$$= u^T A u \Big|_{x=0}^{x=1} + (u, A_{,x} u) \ .$$

Für eine Abschätzung der Energie von u nach oben muß gefordert werden, daß

(1.4.5) $$-y^T A(0,t)\, y \le 0 \quad \forall y : R_0 y = 0$$

und

(1.4.6) $$y^T A(1,t)\, y \le 0 \quad \forall y : R_1 y = 0 \ .$$

Damit erhält man aus (1.4.4)

$$\frac{\partial}{\partial t} \|u\|^2 \le c \|u\|^2 \quad , \text{ mit } c = \|A_{,x}\|$$

und mit dem Lemma von Gronwall

(1.4.7) $$\|u(x,t)\| \le e^{ct} \|u(x,0)\| \ .$$

Die Randbedingungen, die außerdem (1.4.5) und (1.4.6) erfüllen sind natürlich nur eine Teilmenge aller erlaubten Randbedingungen. Sie ermöglichen aber eine Energieabschätzung der Form (1.4.7).
Diese Ergebnisse sollen nun auf das System der Euler-Gleichungen angewendet werden. Dabei sei vorausgesetzt, daß die Lösung für ein $t > 0$ so glatt ist, daß Gleichung (1.2.13) gilt, damit das System gemäß Kapitel 1.3 symmetrisiert werden kann.

Für alle Komponenten des Lösungsvektors v von (1.2.17) müssen Anfangsbedingungen vorgeschrieben werden. Es werden dafür die Werte der ungestörten Strömung im Unendlichen für alle Ortspunkte vorgegeben. Die Störung der Strömung durch die Berandung des Lösungsgebietes stellt sich dann im zeitlichen Verlauf der Rechnung ein.

Bei der Stabilitätsanalyse des symmetrischen hyperbolischen Systems (1.3.3) wird, wie in der Arbeit von Friedrichs /61/ , von einer glatten Lösung $\tilde{v}$ ausgegangen, und das System um diese glatte Lösung linearisiert:

$$d(x,t) := v(x,t) + \tilde{v}(x,t) \tag{1.4.8}$$

Damit erhält man für die Differenz d das lineare symmetrische hyperbolische System

$$B^0(\tilde{v})\, d_{,t} + \sum_{i=1}^{2} B^i(\tilde{v}) d_{,x_i} = 0 \ . \tag{1.4.9}$$

Die positiv definite Matrix B^0 definiert eine Energie durch

$$E(t) := (B^0(\tilde{v})v, v) \ .$$

Mit der Gleichung (1.4.9) erhält man jetzt wie oben

$$\frac{d}{dt} E(t) = (\{ \sum_{i=1}^{2} B^i(\tilde{v})_{,x_i} + B^0_{,t} \}\, d, d) - \int_{\partial\Omega} \{ d^T \sum_{i=1}^{2} B^i(\tilde{v})\, n^T \cdot e_i\, d \}\, ds \ . \tag{1.4.10}$$

Betrachtet man das Cauchy-Problem für Anfangswerte v_0 mit kompakten Träger, so verschwindet der zweite Summand in (1.4.10) . Da B^0 positiv definit und eine stetige Funktion seiner Argumente ist, gilt

$$cI \le B^0(\tilde{v}) \le c^{-1}I \quad \forall\, v \text{ aus einer Umgebung von } \tilde{v} \text{ im Phasenraum} \ .$$

Damit erhält man wieder wie oben mit dem Lemma von Gronwall

$$\|v(x,t)\| \le c^{-1} e^{\alpha t} \|v_0\| \quad \text{mit} \quad \alpha = \frac{1}{2} c^{-1} \|B^0_{,t} + \sum_{i=1}^{2} B^i(\tilde{v})_{,x_i}\| \ .$$

Betrachtet man jetzt das gegebene Anfangs- Randwertproblem, so muß man um Stabilität zu garantieren zeigen, daß das Randintegral in (1.4.10)

$$\int_{\partial\Omega} d^T \sum_{i=1}^{2} B^i(\tilde{v})\, n^T \cdot e_i \, d \, ds \geq 0$$

erfüllt.

Für die festen Wände Γ_1 und Γ_2 des vorgegebenen Stömungsgebiets nimmt man an, daß sie für das Fluid undurchdringbar sind, d.h. für den Geschwindigkeitsvektor u und die Normale n an die Wand gilt

(1.4.11) $$u \cdot n = 0$$

oder falls die Berandung durch eine Funktion $f(x,y,t)$ gegeben ist

$$u \cdot \nabla f(x,y,t) = 0 \quad .$$

Dies entspricht einer Bedingung an v und ist damit konsistent mit der Anzahl der zum negativen Eigenwert gehörigen Charakteristik. Diese Charakteristik läuft in das Gebiet Ω hinein. Für die Randbedingung (1.4.11) ist

(1.4.12) $$d^T \sum_{i=1}^{2} B^i(\tilde{v})\, n^T \cdot e_i \, d = 0$$

und man hat damit Stabilität gezeigt.

Im Fall eines offenen Randes, das sind die Einfluß- und Ausflußränder, kann man nur für den supersonischen Bereich eine Aussage machen. Für einen supersonischen Einflußrand zeigen alle Charakteristiken ins Innere des Gebietes und man darf alle Werte von v vorgeben und (1.4.12) ist erfüllt. Entsprechend darf man für einen supersonischen Ausflußrand keine Komponente von v vorgeben. Damit ist

$$d^T \sum_{i=1}^{2} B^i(\tilde{v})\, n^T \cdot e_i \, d \geq 0$$

da es sich um eine quadratische Form in d handelt. Für subsonische Ränder ist bis jetzt nicht klar, welche physikalisch sinnvollen Randbedingungen gestellt werden dürfen und ob eine Stabilitätsaussage dafür möglich ist, siehe z.B. Roe /63/ oder Schochet /77/ . Das gleiche gilt auch für zusätzlich an den festen Wänden gestellte Randbedingungen wie z.B. die

häufig verwandte Vorgabe des Druckgradienten, siehe z.B. Henke /72/ .

Einen anderen Stabilitätsaspekt sieht man, wenn man das symmetrisierte gasdynamische Problem variationell formuliert. Dabei ist dann eine Funktion $v \in H^1((0;T], H^1(\Omega))$ gesucht, die

$$\int_\Omega \varphi^T (B^0 v_{,t} + \sum_{i=1}^{2} B^i v_{,x_i}) \, dx = 0 \qquad \forall \varphi \in H^1(\Omega) \tag{1.4.13}$$

erfüllt. Wählt man jetzt $\varphi = v$ und verwendet die Definition von v und B^i so kann man (1.4.13) umformen.

$$v^T B^0 v = U_{,u} u_{,v} v_{,t} = U_{,t} = (-u_1 h(S))_{,t}$$

und

$$v^T \sum_{i=1}^{2} B^i v_{,x_i} = \sum_{i=1}^{2} (-u_i h(S))_{,x_i} \, .$$

Damit hat man aus (1.4.13) , da $h(S) = S$ gewählt war

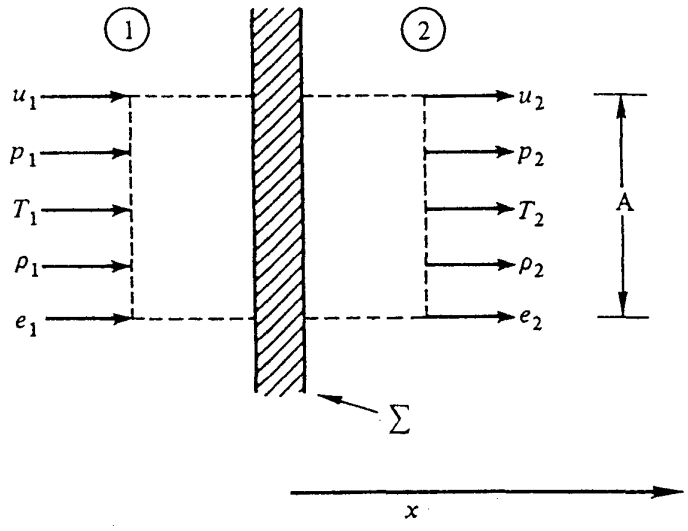
$$\int_\Omega \frac{D}{Dt} (\rho S) \, dx = 0 = \frac{d}{dt} \int_\Omega (\rho S) \, dx \, .$$

Das bedeudet, daß die Entropie des Systems erhalten bleibt, wie es bei der Herleitung der symmetrischen Form der Euler-Gleichungen gefordert wurde.

Die Methode der finiten Elemente geht bei der Diskretisierung von (1.3.3) von der variationellen Formulierung (1.4.13) aus. Verwendet man also ein standard Galerkin-Verfahren, so bleibt die Entropie der FEM-Lösung erhalten. Um Unstetigkeiten in der Lösung, die eine Erhöhung der Entropie bedeuten zu approximieren, muß also ein abgeändertes Galerkin-Verfahren verwendet werden. Ein solches Verfahren wird im zweiten Abschnitt entwickelt.

1.5 Die eindimensionalen Euler-Gleichungen

Wie schon in Kapitel 1.1 erwähnt wurde, ist es bei der Analyse von unstetigen Lösungen der Euler-Gleichungen sinnvoll, von deren integraler Formulierung auszugehen, da diese auch für den Fall einer Strömung mit Schock gültig sind. Es soll hier eine stationäre eindimensionale Strömung über eine Unstetigkeit hinweg untersucht werden. Diese Situation liegt z.B. auf der Staustromlinie einer Kugelanströmung vor. Dabei wird ausgenutzt, daß der Schock immer senkrecht zu den Stromlinien verläuft. Andernfalls würden bei der Betrachtung der Strömungsgrößen ihre Werte von dem Winkel zwischen den Stromlinien und der Unstetigkeit abhängen. Eine Strömung mit senkrechtem Stoß soll durch die folgende Graphik dargestellt werden.



Beim Durchgang durch die Unstetigkeit Σ ändern sich die Strömungsgrößen nur in x-Richtung. Man kann leicht aus den Ergebnissen von Kapitel 1.1 herleiten, daß die Euler-Gleichungen die folgende Form haben :

(1.5.1)
$$\rho_1 u_1 = \rho_2 u_2$$

$$p_1 + \rho_1 u_1^2 = p_2 + \rho_2 u_2^2$$

$$i_1 + \frac{1}{2} u_1^2 = i_2 + \frac{1}{2} u_2^2 \quad .$$

Im weiteren sei der Zustand 1 des Fluids vor dem Stoß gegeben. Aus (1.5.1) kann man dann den Zustand 2 nach dem Passieren der Unstetigkeit berechnen. Dabei soll angenommen werden, daß eine Unstetigkeit vorhanden ist, also die triviale Lösung von (1.5.1) nicht auftritt. Benutzt man für die Energie die Formel (1.2.9), dann erhält man aus (1.5.1)

(1.5.2)
$$\frac{u_2}{u_1} = \frac{\rho_1}{\rho_2} = 1 - \frac{2\gamma}{\gamma + 1} \left(1 - \frac{\gamma p_1}{\rho_1 u_1^2}\right)$$

$$\frac{p_2}{p_1} = 1 + \frac{2\gamma}{\gamma + 1} \left(\frac{u_1^2 \rho_1}{\gamma p_1} - 1\right) \quad .$$

Die Definition der Schockstärke kann auf mehrere Arten erfolgen, siehe Courant, Friedrichs /19/, so z.B. durch die Größen

$$\frac{p_2 - p_1}{p_1} \quad , \quad \frac{\rho_2 - \rho_1}{\rho_1} \quad \text{oder} \quad M^2 - 1 \quad .$$

Dabei ist M die Machzahl, die durch

$$M^2 := \frac{u^2}{c^2} = \frac{u^2 \rho}{\gamma p}$$

definiert ist. Sie gibt an, ob die Strömung subsonisch, $M < 1$ oder supersonisch, $M > 1$ ist. Damit schreibt sich (1.5.2) als

(1.5.3)
$$\frac{u_2}{u_1} = \frac{\rho_1}{\rho_2} = \frac{1}{M^2} \left(1 + \frac{\gamma - 1}{\gamma + 1} (M^2 - 1)\right)$$

$$\frac{p_2}{p_1} = 1 + \frac{2\gamma}{\gamma + 1} (M^2 - 1) \quad ,$$

wobei M die Machzahl vor der Unstetigkeit ist. Aus Gleichung (1.2.10) erhält man für die Änderung der Entropie über den Schock hinweg

(1.5.4) $$S_2 - S_1 = c_v \left(\ln(p_2\rho_1^{\gamma}) - \ln(p_1\rho_2) \right) =$$

$$\ln \left\{ \left(1 + \frac{2\gamma}{\gamma + 1} (M^2 - 1) \right) \left(1 - \frac{2}{\gamma + 1} \left(1 - \frac{1}{M^2}\right)^{\gamma} \right) \right\} .$$

Um die Entropieänderung mit der Änderung der anderen Strömungsgrößen vergleichen zu können, entwickelt man (1.5.4) um die Schallgrenze $M = 1$, die die Unstetigkeit darstellt. Mit

$$\delta := M^2 - 1$$

erhält man für die normierte Entropie S

(1.5.5) $$S_2 - S_1 = \{ \ln(1 + \frac{2\gamma\delta}{\gamma + 1}) + \gamma \ln(1 + \delta \frac{\gamma - 1}{\gamma + 1}) - \gamma \ln(1 + \delta) \}$$

$$= \frac{2\gamma(\gamma - 1)}{3(\gamma + 1)^2} \delta^3 + O(\delta^4) .$$

Für Luft als ideales Gas ist $\gamma = 1.4$ und (1.5.5) hat den Wert

$$S_2 - S_1 \approx 0.065(M^2 - 1)^3 ,$$

oder durch den Drucksprung ausgedrückt

$$S_2 - S_1 = \frac{(\gamma + 1)(\gamma - 1)}{12\gamma^2} \left\{ \frac{p_2}{p_1} - 1 \right\}^3 + O\left\{ \frac{p_2}{p_1} - 1 \right\}^4 \cdot$$

$$= 0.041 \left\{ \frac{p_2}{p_1} - 1 \right\}^3 .$$

Dies besagt, daß die Entropieänderung erst mit der dritten Potenz der Schockstärke erfolgt. Da in dem in dieser Arbeit vorgegebenen Fall der Düsenströmung nur solche schallnahen Strömungen, d.h. $|M^2 - 1| << 1$ untersuchet werden sollen, kann man mit guter Näherung von der Gültigkeit der Gleichung (1.2.13) ausgehen. Sie besagt, daß die Entropie entlang Stromlinien, aber nicht unbedingt im gesamten Strömungsfeld, konstant ist und gilt damit auch für einen gekrümmten Stoß, der bei der Düsenströmung auftritt. Sollen "stärkere" Unstetigkeiten untersucht werden, ist nicht sicher, ob man noch von der symmetrischen Form der Euler-Gleichungen ausgehen kann, wie es z.B. von Johnson /32/ und Hughes /41/ getan wird. Diese Autoren gehen zur Herleitung und bei der Analyse ihres

Verfahrens ohne weitere Begründung von der symmetrischen Form der Euler-Gleichungen aus. In den zitierten Arbeiten werden auch keine Aussagen dazu gemacht, in welchem Geschwindigkeitsbereich die von ihnen verwendeten Euler-Gleichungen gültig sein sollen. Es sind dem Verfasser bis jetzt auch keine Veröffentlichungen bekannt, in denen das von Johnson und Hughes vorgeschlagene Verfahren für Strömungen mit starken Schocks erprobt wurde. Selbst in der Arbeit von Hughes, Franca und Mallet /41/ wird ein anderes Verfahren verwendet.

Eine weitere Möglichkeit die Euler-Gleichungen zu vereinfachen besteht darin, die sogenannte "quasieindimensionale Formulierung" zu verwenden. Bei Strömungen durch achsensymmetrische zweidimensionale Gebiete, wie z.B. eine Laval-Düse kann man statt der x- und y-Koordinate den Gebietsquerschnitt A als unabhängige Koordinate wählen. Dadurch erhält man ein hyperbolisches System in einer "Raumdimension". Die Herleitung wird in Anderson /17/ und exakte Lösungen für die Laval-Düse werden in Zierep /15/ gegeben. Nach Zierep /15/, pp.334 stimmt diese Näherung mit der Lösung der zweidimensionalen Euler-Gleichung für schallnahe Strömungen gut überein. Deshalb wurde diese Näherung auch zur Überprüfung der erzielten numerischen Ergebnisse in Kapitel 5 verwandt.

2. Die Stromliniendiffusionsmethode

Zur Diskretisierung hyperbolischer Gleichungen werden, im Gegensatz zur Diskretisierung elliptischer Gleichungen, hauptsächlich Differenzenverfahren verwandt. Deshalb ist die Literatur numerischer Verfahren für diesen Gleichungstyp auch im wesentlichen auf Differenzenverfahren beschränkt. Um den Zusammenhang der Differenzenverfahren mit der Methode der finiten Elemente zu diskutieren und die numerischen Schwierigkeiten zu verdeutlichen, wird das Beispiel der einfachsten Advektionsgleichung

(2.1) $$u_{,t} = cu_{,x} \quad x \in \Omega \subset\subset \mathbb{R},\ t \in (0, T]$$

(2.2) $$u(x,0) = u_0(x) \quad x \in \Omega$$

mit den Charakteristiken

$$x - ct = \text{const.}$$

herangezogen. Der Anfangswert u_0 habe kompakten Träger in Ω. Das Lösungsgebiet Ω oder der Endzeitpunkt T seien so gewählt, daß für alle $t \leq T$ $\text{supp}(u) \subset\subset \Omega$ ist. Man kann dann $u(x,t) = 0$ für alle $x \in \partial\Omega$, $t \in [0, T]$ fordern. Die verwendeten Begriffe aus der Theorie der Differenzenverfahren und der finiten Elemente werden im allgemeinen nicht explizit eingeführt. Dafür wird auf die Lehrbücher Richtmeyer, Morton / 2/ oder Sod / 1/ für Differenzenverfahren und Strang, Fix / 9/ oder Ciarlet / 6/ für FEM-Verfahren verwiesen.

Sei k die Zeitschrittweite, h die Ortsschrittweite, $\lambda := k/h$ und wie üblich $u_i^n := u(ih,nk)$. Ein einfaches Differenzenverfahren für (2.1) hat die Gestalt

(2.3) $$u_i^{n+1} = u_i^n + \frac{1}{2}\lambda c(u_{i+1}^n - u_{i-1}^n) \ .$$

Das Verfahren hat die Konsistenzordnung $O(k) + O(h^2)$. Um die Stabilität des Verfahrens zu analysieren berechnet man sein Symbol

$$\rho(\xi) = -c\lambda e^{i\xi} + 1 + ce^{-i\xi} \ ,$$

wobei i die imaginäre Einheit ist. Die von Neuman Bedingung fordert für die Stabilität des Verfahrens, daß $|\rho(\xi)| \leq 1$ ist. Für das Verfahren (2.3) ist

$$|\rho(\xi)| = \rho(\xi)\overline{\rho(\xi)} = 1 + c^2\lambda^2(\sin\xi)^2 > 1 \ .$$

Also ist das Verfahren instabil.

Die Differenzenverfahren bieten die Möglichkeit, eine einseitige Differenzenapproximation zu bilden. Dies ist bei den FEM-Verfahren nicht direkt möglich, da die Ansatzfunktionen im allgemeinen ihren Träger auf mehreren Elementen haben. Das durch "upwinding" verbesserte Verfahren lautet

$$u_i^{n+1} = u_i^n + \lambda c(u_i^n - u_{i-1}^n) \tag{2.4}$$

und hat das Symbol $\rho(\xi) = 1 + c\lambda - c\lambda e^{i\xi}$. Das Verfahren hat die Konsistenzordnung $O(h) + O(k)$. Die Norm des Symbols ist für $-c\lambda \leq 1$

$$|\rho(\xi)| \leq |1 + c\lambda| + |-c\lambda e^{i\xi}| = 1 \ .$$

Damit ist das Verfahren für $-c\lambda \leq 1$ stabil, sonst aber instabil. Für variables c ist diese Verbesserung also aufwendig und mit einem Ordnungsverlust verbunden. Eine schöne Erklärung für die Instabilität von (2.3) wird von Sod / 1/ gegeben. Er zeigt, daß die Werte von u an den Gitterpunkten mit ungerader Nummer unabhängig von den Werten von u an den Gitterpunkten mit gerader Nummer transportiert werden. Abhilfe gegen diese Entkopplung schafft das Verfahren von Lax und Friedrich, das stabil ist aber nur die Konsistenzordnung $O(h) + O(k)$ hat, also auch mit einem Ordnungsverlust verbunden ist.

Bei der Durchführung der Methode der finiten Elemente geht man von der schwachen Formulierung von (2.1) aus. Sie lautet : Finde ein $u \in H^1([0; T) ; H^1(\mathbb{R}))$, das

$$(u_{,t} , v) = (u_{,x} , v) \qquad \forall\, v \in V_h \subset H^1(\Omega) \tag{2.5}$$

erfüllt. Dabei sei V_h der Raum der auf dem Gitter $\{0 \pm hj, j \leq J \in \mathbb{N}\}$ stückweise linearen Funktionen. Ist $\{\varphi_i\}$ eine Basis von V_h , so erhält man mit der Darstellung

$$u(x,t) = \sum_{i=1}^{N} \alpha_i(t)\varphi_i(x)$$

aus (2.5) ein lineares Gleichungssystem der Form

$$\sum_{i=1}^{N} \frac{d}{dt}\,\alpha_i(t)(\varphi_i,\varphi_k) - \sum_{i=1}^{N} \alpha_i(t)(\frac{d}{dx}\,(c\varphi_i),\varphi_k) = 0 \qquad k = 1,\ldots,N \ .$$

Nimmt man c als konstant an, so können die auftretenden Integrale exakt ausgewertet werden. Das entstehende System gewöhnlicher Differentialgleichungen wird mit dem expliziten Eulerverfahren diskretisiert. Dann hat die i-te Zeile des entsprechenden linearen Gleichungssystems die folgende Gestalt :

$$(2.6)\quad \frac{1}{6}\,(u_{i-1}^{n+1} + 4u_i^{n+1} + u_{i+1}^{n+1}) = \frac{1}{6}\,(u_{i-1}^{n} + 4u_i^{n} + u_{i+1}^{n}) + \frac{1}{2}\,c\lambda(u_{i+1}^{n} - u_{i-1}^{n}) \ .$$

Wird die Massematrix $\frac{1}{6}\,(u_{i-1}^{n} + 4u_i^{n} + u_{i+1}^{n})$ "gelumpt", das heißt die Integrale werden mit der Trapezregel ausgewertet, siehe Thomee / 8/, so erhält man wieder das Verfahren (2.3), von dem man bereits weiß, daß es instabil ist. Aus der Äquivalenz der Verfahren sieht man auch, daß dieser Prozess keinen Einfluß auf die Konvergenzordnung hat. Eine hier vorgeschlagene Möglichkeit, das Verfahren (2.6) zu stabilisieren ist, das "Lumpen" nur auf der linken Seite auszuführen. Dadurch erhält man das Verfahren

$$(2.7)\qquad u_i^{n+1} = \frac{1}{6}\,(u_{i-1}^{n} + 4u_i^{n} + u_{i+1}^{n}) + \frac{1}{2}\,c\lambda(u_{i+1}^{n} - u_{i-1}^{n})$$

mit dem Symbol

$$\rho(\xi) = \frac{1}{6}\,(e^{i\xi} + 4 + e^{-i\xi}) + \frac{1}{2}\,c\lambda(e^{i\xi} - e^{-i\xi}) \ .$$

Dieses Symbol hat die Norm

$$|\rho(\xi)| \le 1 - (\frac{1}{9} - c^2\lambda^2)\,(\sin\xi)^2 \ .$$

Sie ist ≤ 1 für $\frac{1}{9} > c^2\lambda^2$. Damit sieht man auch, daß das "Lumpen" die Konvergenzordnung erhält aber die Stabilität beeinträchtigt.

An diesem einfachen Beispiel wurde gezeigt, daß die Problematik bei der Lösung von hyperbolischen Gleichungen in der Instabilität der Standard-Verfahren oder dem Verlust an Konsistenzordnung liegt. Das gilt auch für andere klassische Dämpfungsverfahren

wie zum Beispiel die "künstliche Viskosität", siehe Roache / 4/, das die Konsistenzordnung auf eins beschränkt.

In diesem Kapitel soll an Hand einer hyperbolischen Modellgleichung ein FEM-Ansatz motiviert und seine Herleitung gegeben werden, der die obengenannten Nachteile nicht hat. Dabei soll zunächst der Fall einer linearen skalaren Modellgleichung betrachtet werden. Ausgehend von der stationären Modellgleichung werden dann die verschiedenen Möglichkeiten diskutiert, das instationäre Problem zu behandeln. Das hier gewählte Verfahren zur Berechnung des instationären Modellproblems wird zunächst für die lineare Gleichung analysiert. Für die Übertragung des Verfahrens auf eine Erhaltungsgleichung wird ein spezieller Vertreter, nämlich die Burger-Gleichung ausgewählt. Sie gilt als Modell für die Euler-Gleichungen der Gasdynamik und wird deshalb zur Verifikation von theoretischen Ergebnissen und zur Erprobung von numerischen Algorithmen für die Euler-Gleichungen häufig in der Literatur verwendet. Außerdem kann man für viele Situationen exakte Lösungen angeben, siehe Fletcher /49/ .

Bei der Analyse des Verfahrens für diese Gleichung wird in einem ersten Schritt der Fall einer glatten Lösung mit herkömmlichen Energie-Methoden untersucht. Da dieser Fall für die Praxis, in der die Lösung im allgemeinen unstetig ist, nicht so relevant ist, wird in einem zweiten Schritt der Fall einer beschränkten Lösung mit der Methode der "kompensierten Kompaktheit ", siehe z.B. DiPerna /51/, Tartar /57/, Dacarogna /60/ analysiert. Diese Methode wurde von Tartar für Fragestellungen aus der Kontrolltheorie entwickelt. Danach wurde sie von DiPerna zur Analyse von hyperbolischen Systemen in einer Raumdimension verwendet. Ein entscheidenter Nachteil der Methode ist, daß ihre Ergebnisse bis jetzt nur in einer Raumdimension gelten. Zur Herleitung der Methode der kompensierten Kompaktheit werden hauptsächlich Eigenschaften der Schwach-*-Topologie von $L^{\infty}(\mathbb{R})$ und der Begriff des Youngschen Maßes verwendet. Diese Resultate dienen dann mit Hilfe des Lemmas von Murat zum Existenzbeweis einer Lösung des hyperbolischen Systems von Erhaltungsgleichungen. Eine Einführung in die Methode wird in Dacarogna /60/ geben. Das Lemma von Murat wird in Tartar /58/ bewiesen.

Die Anwendung dieser Methode beschränkt sich bis jetzt fast ausschließlich auf die kontinuierlichen Gleichungen. Anwendungen für einen Konvergenzbeweis eines numerischen Verfahrens wurden zunächst von DiPerna /56/ und darauf von Johnson, Szepessy /31/ publiziert. Für eine weiterführende Diskussion der Methode der kompensierten Kompaktheit und der in den Sätzen von Kapitel 2.6 verwendeten Ergebnisse wird auf die Arbeiten von DiPerna und Tartar im Literaturverzeichnis verwiesen.

2.1. Das Modellproblem

Für ein beschränktes Gebiet $\Omega \subset \mathbb{R}^n$ und einen Vektor $\beta \in C(\Omega)$ mit $|\beta| \approx 1$ definiert man für den Rand Γ von Ω den

Einflußrand $\quad \Gamma_- := \{ x \in \Gamma : n(x) \cdot \beta(x) < 0 \}$

und den

Ausflußrand $\quad \Gamma_+ := \{ x \in \Gamma : n(x) \cdot \beta(x) \geq 0 \}$,

wobei $n(x)$ der äußere Normalenvektor an Γ ist. Gesucht ist jetzt eine Lösung $u \in H^1(\Omega)$ der *Randwertaufgabe*

$$\beta \cdot \nabla u + \sigma u = f \quad \text{in } \Omega \,, \tag{2.1.1}$$

$$u = g \quad \text{auf } \Gamma_- \,,$$

mit $f \in L^2(\Omega)$ und $g \in H^{\frac{1}{2}}(\Gamma_-)$.

Es dürfen nur Randbedingungen auf Γ_- oder die durch den Verlauf der Charakteristiken bestimmten Randbedingungen auf Γ_+ vorgegeben werden. Mit diesen Vorgaben ist die Randwertaufgabe (2.1.1) eindeutig lösbar. Für eine ausführliche Analyse von Gleichungen des obigen Typs wird auf Whitham /25/ verwiesen.

Bei instationären hyperbolischen Gleichungen sind der Zeit- und der Ortsdifferentialoperator von gleicher Ordnung. Deshalb kann man die zu (2.1.1) analoge instationäre *Anfangs–Randwertaufgabe*

(2.1.2)
$$u_{,t} + \beta \cdot \nabla u + \sigma u = f \quad \text{in } \Omega \times I \,,$$
$$u = g \quad \text{auf } \Gamma_- \,,$$
$$u(x,0) = u_0(x) \quad \text{in } \Omega \,,$$

mit $I = (0;T]$, $T > 0$, auch schreiben als

(2.1.3)
$$\tilde{\beta} \cdot \tilde{\nabla} u + \sigma u = f \quad \text{in } \Omega \times I \,,$$
$$u = g \quad \text{auf } \Gamma_- \,,$$
$$u(x,0) = u_0(x) \quad \text{in } \Omega \,,$$

wobei $\tilde{\nabla} := (\frac{\partial}{\partial t} , \nabla)$, $\tilde{\beta} := (1 , \beta)$ und $n := (n_t , n_x)$ bedeutet. Der Einflußrand Γ_- ist jetzt entsprechend im Orts-Zeit-Raum definiert. Mit dieser Notation kann man die Theorie für die stationäre Gleichung auf die Gleichung im Orts-Zeit-Raum übertragen. Dieser Weg wird von Johnson /29/ und Nävert /34/ beschritten.

Hingegen ist es im ingenieurwissenschaftlichen Bereich üblich, bei der Diskretisierung von Gleichungen des Typs (2.1.2) die Orts- und die Zeitableitungen getrennt zu behandeln, wie es zum Beispiel bei der Einführung des Stromliniendiffusionsverfahrens durch Hughes /36/ getan wird. Dieses Vorgehen ermöglicht auch eine getrennte Analyse der Orts- und Zeitdiskretisierungen und eine Kopplung verschiedener Verfahren für die jeweilige Diskretisierung. Insbesondere für die numerische Lösung der Euler-Gleichungen für reale Gase sollte man sich diese Möglichkeit offen halten. Bei diesen Strömungen treten Effekte auf, die eine hohe Auflösung der Ortsvariablen erfordern, wie z.B. lokal abgelöste Überschallgebiete, aber auch Effekte, die eine zeitgenaue Rechnung erfordern, wie z.B. nichtkonvexe Zustandsgleichungen oder die Berücksichtigung chemischer Reaktionen.

Aus diesen Gründen wurde in der vorliegenden Arbeit ein Verfahren gewählt, bei dessen Herleitung die Orts- und die Zeitdiskretisierung getrennt durchgeführt werden.

2.2. Herleitung und Beschreibung des Verfahrens

Die numerische Berechnung der Lösung von (2.1.1) mit der Methode der finiten Elemente, oder was auf regelmäßigen Gittern äquivalent ist, durch eine Diskretisierung mit zentralen Differenzenquotienten ist problematisch, da die Näherungslösungen starke Oszillationen aufweisen. Um dies zu verhindern muß das Verfahren gedämpft und damit stabilisiert werden. Hier soll jetzt eine Möglichkeit zur Stabilisierung eines Standard-Verfahrens vorgestellt werden, die im Gegensatz zu den obengenannten die Konsistenzordnung mit dem Randwertproblem erhält.

Im weiteren wird die folgende Notation benutzt :

$$(v,w) := \int_\Omega vw \, dx \quad , \quad \|v\| := \|v\|_{L^2(\Omega)} \quad , \|v\|_s := \|v\|_{H^s(\Omega)} \quad ,$$

$$\langle v,w\rangle_\pm := \int_{\Gamma_\pm} v \cdot n \cdot \beta \, ds \quad , \quad |v|^2 := \int_\Gamma v^2 \, |n \cdot \beta| \, ds \ ,$$

wobei $L^2(\Omega)$ der Hilbertraum der Lebesgue-integrierbaren Funktionen und $H^s(\Omega)$ der Sobolewraum der Funktionen mit verallgemeinerten s-ten Ableitungen ist. Weitere Bezeichnungen werden an der Stelle ihres ersten Auftretens vereinbart. In dieser Notation schreibt sich die *Greensche Formel* als

$$(\beta \cdot \nabla v \, , \, w) = \langle v \, , \, w\rangle - (v \, , \, \beta \cdot \nabla w) - (v \, , \, w \, \mathrm{div} \, \beta) \ .$$

Sei $\{T_h\}$ eine Familie von quasiregulären Triangulierungen $T_h = \{K\}$ des Polygongebietes Ω , die die Minimalwinkelbedingung erfüllen. Der Parameter h ist der größte Durchmesser aller Polygone $K \in T_h$. Für ein $k \in \mathbb{N}$ definiert man den *Finiten–Elemente–Raum*

$$V_h := \{ \, v \in H^1(\Omega) \, : \, v|_K \in \mathbb{P}_k(K) \quad \forall \, K \in T_h \, \} \ ,$$

wobei $\mathbb{P}_k(K)$ die Menge der Polynome auf K vom Grad k ist, d.h. V_h ist der Raum der stetigen und stückweise vom Grad k polynomialen Funktionen. Aus der Approximations-

theorie, siehe z.B. Ciarlet / 6/, weiß man, daß für Funktionen $u \in H^{k+1}(\Omega)$ eine Interpolierende $I_h u \in V_h$ existiert, so daß

$$\|u - I_h u\| + h\|u - I_h\|_1 \leq ch^{k+1}\|u\|_{k+1} \tag{2.2.1}$$

$$|u - I_h u| \leq ch^{k+\frac{1}{2}}\|u\|_{k+1}$$

gilt. Das gleich ist auch für die L^2-Projektion $\pi u \in V_h$ von u richtig.

Zunächst wird das normale Galerkin-Verfahren für Gleichung (2.1.1) vorgestellt und gezeigt, worin die Verbesserung der Stromliniendiffusionsmethode besteht. Im Standard-Galerkin-Verfahren für Gleichung (2.1.1) mit $\sigma \equiv 1$ sucht man eine Funktion $u_h \in V_h$ die

$$(\beta\cdot\nabla u_h + u_h , \varphi) - \langle u_h , \varphi\rangle_- = (f , \varphi) - \langle g , \varphi\rangle_- \qquad \forall \varphi \in V_h \tag{2.2.2}$$

erfüllt. Das läßt sich mit Hilfe der durch (2.2.2) definierten Bilinearform bzw. Linearform kürzer schreiben als

$$b(u,\varphi) = l(\varphi) \qquad \forall\, \varphi \in V_h \ . \tag{2.2.3}$$

Für ein $u \in H^1(\Omega)$ gilt mit der Greenschen Formel

$$b(u,u) = \|u\|^2 + \tfrac{1}{2}\,|u|^2 \ . \tag{2.2.4}$$

Aus der Koerzitivitätsgleichung (2.2.4) folgt mit dem Satz von Lax-Milgram, z.B. in Ciarlet / 6/ , die eindeutige Lösbarkeit von (2.2.3).

Der *Approximationsfehler* e wird gesplittet zu

$$e := u - u_h = u - I_h u + I_h u - u_h =: \rho + \theta \ . \tag{2.2.5}$$

Mit der Orthogonalitätsrelation für den Fehler e erhält man in der durch $b(.,.)$ induzierten Norm

$$\|\theta\|^2 + \tfrac{1}{2}\,|\theta|^2 = b(\rho,\theta) \leq \|\beta\cdot\nabla\rho\|^2 + \|\rho\|^2 + \tfrac{1}{2}\,\|\theta\|^2 + |\rho|^2 + \tfrac{1}{4}\,|\theta|^2 \ .$$

Der erste Term auf der rechten Seite wird abgeschätzt zu

$$\|\beta\cdot\nabla\rho\| \leq ch^r\,\|u\|_{r+1} \ , \tag{2.2.6}$$

was den Verlust einer vollen h-Potenz im Vergleich zur Bestapproximation (2.2.1) bedeutet. Die Dreiecksungleichung liefert jetzt für den Gesamtfehler e

$$\|e\| + |e| \leq ch^r \|u\|_{r+1} \quad . \tag{2.2.7}$$

Da die gesuchte Lösung $u \in H^1(\Omega)$ ist, ist im allgemeinen keine Konvergenzordnung zu erwarten.

Die Herleitung der Stromliniendiffusionsmethode soll hauptsächlich motivierenden Charakter haben. Deshalb wird angenommen, daß die Funktion u genügend Regularität besitzt, um alle auftretenden Operationen durchführen zu können. Als erstes differenziert man die Gleichnung (2.1.1) in β-Richtung, also in Stromlinienrichtung, und multipliziert das Ergebnis mit $\delta = \epsilon h$, $h \geq 0$, $\epsilon > 0$ beliebig aber von h unabhängig.

$$\delta\beta\cdot\nabla(\beta\cdot\nabla u + u) = \delta\beta\cdot\nabla f \quad . \tag{2.2.8}$$

Anschließend subtrahiert man (2.1.1) und (2.2.8)

$$\begin{aligned} -\delta\beta\cdot\nabla(\beta\cdot\nabla u + u) + \beta\cdot\nabla u + u &= f - \delta\beta\cdot\nabla f \qquad \text{in } \Omega \ , \\ -\delta(\beta\cdot\nabla u + u) + u &= g - \delta f \qquad \text{auf } \Gamma_- \ , \\ \delta(\beta\cdot\nabla u + u) &= \delta f \qquad \text{auf } \Gamma_+ \ . \end{aligned} \tag{2.2.9}$$

Auf (2.2.9) wendet man das Standard-Galerkin-Verfahren an und integriert partiell.

$$\begin{aligned} &(\beta\cdot\nabla u_h + u_h \ , \ \varphi + \delta\beta\cdot\nabla\varphi) - (1+\delta)\langle u_h \ , \ \varphi\rangle_- = \\ &(f \ , \ \varphi + \delta\beta\cdot\nabla\varphi) - (1+\delta)\langle g \ , \ \varphi\rangle_- \qquad \forall \varphi \in V_h \ , \end{aligned} \tag{2.2.10}$$

oder in kürzerer Schreibweise mit den durch (2.2.10) definierten Linearformen

$$B(u,\varphi) = L(\varphi) \qquad \forall \varphi \in V_h \ . \tag{2.2.11}$$

Für ein $u \in H^1(\Omega)$ ist wie oben

$$B(u,u) = \frac{1+\delta}{2}|u|^2 + \|u\|^2 + \delta\|\beta\cdot\nabla u\|^2 =: \|u\|_\beta^2 \ . \tag{2.2.12}$$

Damit wird ein h als Faktor vor dem in der Abschätzung (2.2.6) problematischen Term $\|\beta\cdot\nabla\rho\|$ erzeugt. Der Fehler e läßt sich in der durch B(u,u) induzierten Norm kontrollieren.

$$(2.2.13)\quad \|e\|_\beta^2 = B(e,\rho) \le \frac{h}{2}\|\beta\cdot\nabla e\|^2 + h^{-1}\|\rho\|^2 + \frac{h}{4}\|\beta\cdot\nabla e\|^2 + h\|\beta\cdot\nabla\rho\|^2 +$$

$$\frac{1}{2}\|e\|^2 + \|\rho\|^2 + h^2\|\beta\cdot\nabla\rho\|^2 + \frac{1+h}{4}|e|^2 + (1+h)|\rho|^2$$

$$\le ch^{2r+1}\|u\|_{r+1}^2 .$$

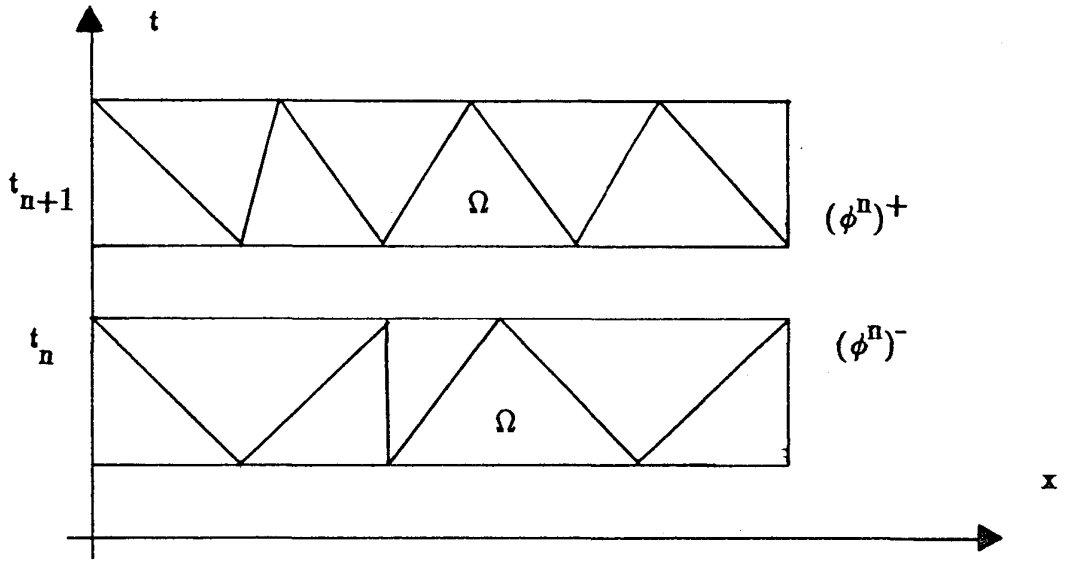
Damit gewinnt man eine halbe h-Potenz in der Konvergenzordnung im Gegensatz zu (2.2.7) und kann den Fehler auch für Lösungen $u \in H^1(\Omega)$ kontrollieren. Ein Nachteil dieser Analyse ist, daß auch für beliebig glatte Lösungen keine optimale Konvergenzordnung erzielt werden kann.

Eine andere Möglichkeit die Stromliniendiffusionsmethode zu behandeln besteht darin, sie im Rahmen von sogenannten Petrov-Galerkin-Verfahren einzuführen, wie es z.B. in Morton /65/ getan wird. Dieses Vorgehen ist dadurch charakterisiert, daß Ansatz- und Testraum verschieden gewählt werden und man auf eine Herleitung wie oben verzichten kann.
Eine ausführliche Diskussion der Stromliniendiffusionsmethode für den Fall $\sigma - \frac{1}{2}\operatorname{div}\beta \ge c > 0$ wird in Nävert /34/ dargestellt.

Um das Verfahren auf eine instationäre Gleichung anzuwenden, definiert sich Johnson, wie oben erwähnt, einen Operator $\tilde{\nabla}$ und ein $\tilde{\beta}$ und überträgt das Verfahren für die stationäre Gleichung. Benutzt man den Ansatzraum V_h jetzt im Orts-Zeit-Raum, so sind dort alle Punkte miteinander gekoppelt. Bei der Durchführung des Verfahrens müßte also ein Gleichungssystem für alle Orts-Zeit-Punkte simultan gelöst werden. Dieses Problem umgeht Johnson durch Einführung eines neuen Finite-Elemente-Raums

$$W_{hk} = \{\, u : I \to V_h : u_{|I_n} \in P_l(I_n)\,,\ n = 1,\ldots,N \,\} .$$

Dieser Ansatzraum enthält Funktionen, die auf jedem Zeitintervall $I_n := t_n - t_{n-1}$ polynomial vom Grad l und über die Intervallgrenzen hinweg unstetig sind.



Dadurch werden die Zeitebenen entkoppelt und man erhält ein Verfahren, mit dem man in der Zeit voranschreiten kann. In Analogie zur Stromliniendämpfung im Ort wird jetzt in Stromlinienrichtung im Orts-Zeit-Raum gedämpft.

$$(2.2.14) \qquad (u_{h,t} + \beta\cdot\nabla u_h \,,\, \varphi + \delta(\varphi_{,t} + \beta\cdot\nabla\varphi)) - (1+\delta)\langle u_h \,,\, \varphi\rangle_- =$$

$$(f \,,\, \varphi + \delta(\varphi_{,t} + \beta\cdot\nabla\varphi)) - (1+\delta)\langle g \,,\, \varphi\rangle_- \quad \forall\, \varphi \in W_{hk} \;.$$

Die Gebiets- und Randintegrale in (2.2.14) beziehen sich jetzt auf den Orts-Zeit-Raum. Wählt man den Polynomgrad $l = 0$, so ist das Verfahren, für homogene rechte Seite in (2.1.2) äquivalent zum impliziten Eulerverfahren. Für $l \geq 1$ ergeben sich implizite Runge-Kutta-Formeln, wie sie z.B. von Lesaint, Raviart /38/ und Johnson /12/ zur Diskretisierung der Neutronen-Transport-Gleichung verwendet wurden.

Im Vergleich der hier vorgestellten Verfahren ist es wichtig, den jeweiligen numerischen Aufwand darzustellen. Dazu wird das Galerkin-Verfahren mit Test- und Ansatzfunktionen aus W_{hk} auf eine Gleichung vom Typ (2.1.2) angewendet.

$$\int_{I_n} ((\frac{\partial}{\partial t} u_h , \varphi) + b(u_h , \varphi)) \, dt + (u_{h+}^{n-1} , \varphi_+^{n-1}) =$$

$$\int_{I_n} (f , \varphi) \, dt + (u_{h-}^{n-1} , \varphi_+^{n-1}) \quad \forall \varphi \in \mathbb{P}_1(I_n) \ ,$$

wobei $u_\pm(x,t) := \lim_{s \to \pm 0} v(x + s\beta , t + s) \qquad \forall (x,t) \in \Omega \times \{t_{n-1}, t_n\}$.

Nun wird für u^n und φ die Basisdarstellung eingesetzt und die Zeitintegration ausgeführt. Das entstehende Blockgleichungssystem hat die folgende Gestalt:

$$(2.2.15) \quad (M + k_n \cdot A)U_0 + (M + \tfrac{1}{2} k_n \cdot A)U_1 = F_1 + (U_-)^{n-1}$$

$$\tfrac{1}{2} k_n \cdot AU_0 + (\tfrac{1}{2} M + \tfrac{1}{3} k_n \cdot A)U_1 = F_2 \ ,$$

wobei $k_n := t_n - t_{n-1}$ und $U^n := U_0 + U_1$. M und A sind $N \times N$ - Matrizen, wenn N die Dimension des Raumes V_h ist.

Dies zeigt, daß das Verfahren, insbesondere im nichtlinearen Fall, bei dem $M = M(U)$ und $K = K(U)$, im Vergleich zu herkömmlichen Einschrittverfahren mehr als viermal so aufwendig ist. Es ist im Fall einer skalaren Gleichung ein 4×4-Blockgleichungssystem zu lösen. Außerdem kommt für eine nichtlineare Gleichung der sehr rechenzeitintensive Matrizenaufbau pro Zeitschritt hinzu. Dies ist insbesondere wichtig im Hinblick auf die Anwendung des Verfahrens auf die im Kapitel 1 besprochenen nichtlinearen Systeme, da hier der zusätzliche Aufwand nocheinmal um einen Faktor vier größer ist. Das in Gleichung (2.2.15) verwendete Galerkin-Verfahren mit unstetigen Ansatzfunktionen in der Zeit wurde ausführlich von Návert /34/ im Fall einer linearen Gleichung und für nichtlineare skalare Probleme von Johnson und Szepessy /31/ analysiert.

Aus den oben genannten Gründen wird hier eine andere Möglichkeit zur Behandlung des instationären Problems gewählt. Im weiteren soll die Anfangs-Randwertaufgabe (2.1.2) mit $\sigma \equiv 0$, $f \equiv 0$, $g \equiv 0$ und $\beta = \beta(x)$ betrachtet werden. Außerdem habe wieder

$u_0(x)$ kompakten Träger in Ω . Dies stellt keine Einschränkung an das Verfahren oder die Analyse dar, vereinfacht aber die Schreibarbeit.

Wie im stationären Fall wird die Gleichung (2.1.2) nur in Ortsstromlinienrichtung gedämpft.

$$(2.2.16) \qquad (\frac{\partial}{\partial t} u_h + \beta \cdot \nabla u_h \, , \, \varphi + \delta\beta \cdot \nabla \varphi) = 0 \qquad \forall \varphi \in V_h \, .$$

Zur Zeitdiskretisierung wird das Crank-Nicolson-Verfahren verwendet.

$$(2.2.17) \qquad (\frac{1}{k}(u_h^n - u_h^{n-1}) + \beta \cdot \nabla (\frac{1}{2}(u_h^n + u_h^{n-1}) \, , \, \varphi + \delta\beta \cdot \nabla \varphi) = 0 \qquad \forall \varphi \in V_h \, .$$

Dies stellt,wie leicht zu sehen ist, eine erhebliche Vereinfachung im Vergleich zu (2.2.15) dar. Die folgende Analyse wird zeigen, daß die für (2.2.15) gewonnen Ergebnisse auch für das vereinfachte Verfahren (2.2.17) gelten.

2.3 Analyse des Verfahrens für die lineare Modellgleichung

Das oben eingeführte Verfahren wird auf die folgende Advektionsgleichung angewendet:

$$(2.3.1) \qquad \begin{aligned} u_{,t} + \beta \cdot \nabla u &= 0 \quad \text{in } \Omega \times (0,T] \, , \\ u &= 0 \quad \text{auf } \Gamma \, , \\ u(x,0) &= u_0(x) \quad \text{in } \Omega \, , \end{aligned}$$

wobei $u \in H^1((0;T] \, , H^1(\Omega))$ ist, und das beschränkte Gebiet Ω oder Γ sollen so gewählt werden, daß die Verträglichkeitsbedingung an die Randwerte erfüllt ist.

Die Divergenz von β soll beschränkt sein

$$(2.3.2) \qquad \alpha := \max_{x \in \Omega} | \operatorname{div} \beta | < \infty \, ,$$

muß aber keiner Vorzeichenbedingung genügen. Die variationelle Formulierung von (2.3.1) ist

(2.3.3) $$(u_{,t} + \beta \cdot \nabla u, \varphi + \delta\, \beta \cdot \nabla \varphi) = 0 \quad \forall\, \varphi \in V \ ,$$

$$u(x,0) = u_0(x) \ ,$$

wobei $V = H_0^1(\Omega)$ und $\delta = \epsilon h \quad \epsilon > 0 \ , \quad h \geq 0$ beliebig . Definiert man in der üblichen Weise einen Finite-Elemente-Raum

$$V_h := \{ v \in V \ : \ v|_K \in \mathbb{P}_k(K), \ \forall\, K \in T_h \}$$

und setzt ein $U \in H^1((0,T] ; V_h)$ in (2.3.3) ein, so erhält man die *Semidiskretisierung*

(2.3.4) $$(U_{,t} + \beta \cdot \nabla U \ , \varphi + \delta\beta \cdot \nabla \varphi) = 0 \quad \forall\, \varphi \in V_h \ ,$$

$$U(x,0) = \pi u_0 \ .$$

Führt man die Zeitdiskretisierung mit dem Crank-Nicolson-Verfahren durch, so lautet das *volldiskrete Verfahren*

(2.3.5) $$(d_t U^n + \beta \cdot \nabla \bar{U}^n \ , \varphi + \delta\beta \cdot \nabla \varphi) = 0 \quad \forall\, \varphi \in V_h$$

$$U^0 = \pi u_0$$

In (2.3.5) bedeutet

$$d_t U^n := \frac{1}{k}(U^n - U^{n-1}) \ , \text{ und } \bar{U}^n := \frac{1}{2}(U^n + U^{n-1})$$

mit $k := t_n - t_{n-1}$.

Als erstes wird die Stabilität der Semidiskretisierung untersucht.

Satz 1

Für die Lösung U von (2.3.4) gilt unter der Voraussetzung $1 > \alpha\delta$ mit einer Konstanten $c > 0$

(2.3.6) $$\| U(t) \| + \delta \| \beta \cdot \nabla U(t) \| \leq e^{c\alpha t} (\| U^0 \| + \delta \| \beta \cdot \nabla U^0 \|) \ .$$

Beweis :

Wählt man in (2.3.4) $\varphi := U$ und $\varphi := U_{,t}$. Dann gilt

$$(2.3.7) \qquad \frac{1}{2}\frac{d}{dt}\| U \|^2 + (\beta \cdot \nabla U , U) + \delta (U_{,t} , \beta \cdot \nabla U) + \delta \| \beta \cdot \nabla U \|^2 = 0$$

und

$$(2.3.8) \qquad \delta \| U_{,t} \|^2 + \delta (\beta \cdot \nabla U , U_{,t}) + \delta^2 (U_{,t} , \beta \cdot \nabla U_{,t})$$

$$+ \frac{1}{2}\delta^2 \frac{d}{dt}\| \beta \cdot \nabla U \|^2 = 0 \ .$$

Die Gleichung (2.3.8) wurde außerdem noch mit δ multipliziert. Da

$$\| U_{,t} + \beta \cdot \nabla U \|^2 = \| U_{,t} \|^2 + 2(\beta \cdot \nabla U , U_{,t}) + \| \beta \cdot \nabla U \|^2 \ ,$$

ergibt die Addition von (2.3.7) und (2.3.8)

$$\frac{1}{2}\frac{d}{dt}\{ \| U \|^2 + \delta^2 \| \beta \cdot \nabla U \|^2 \} + \delta \| U_{,t} + \beta \cdot \nabla U \|^2 = - (\beta \cdot \nabla U , U)$$

$$- \delta^2 (U_{,t} , \beta \cdot \nabla U_{,t}) \ .$$

Um die rechte Seite der letzten Gleichung umzuformen, wird partiell integriert.

$$(2.3.9) \qquad \frac{d}{dt}\{ \| U \|^2 + \delta^2 \| \beta \cdot \nabla U \|^2 \} + 2\delta \| U_{,t} + \beta \cdot \nabla U \|^2 \leq$$

$$\alpha \| U \|^2 + \alpha\delta^2 \| U_{,t} \|^2 \ .$$

In dieser Ungleichung stört bei der Anwendung des Lemmas von Gronwall noch der Term $\| U_{,t} \|$ auf der rechten Seite. Eine Schranke dafür verschafft man sich aus Gleichung (2.3.8). Nach partieller Integration und Anwendung der Cauchy-Schwarzschen und der Youngschen Ungleichung ist (2.3.8)

$$\| U_{,t} \|^2 + \frac{1}{2}\delta \frac{d}{dt}\| \beta \cdot \nabla U \|^2 = - (\beta \cdot \nabla U , U_{,t}) + \frac{1}{2}\delta (U_{,t} , U_{,t} \operatorname{div} \beta)$$

$$\leq \frac{1}{2}\| \beta \cdot \nabla U \|^2 + \frac{1}{2}\| U_{,t} \|^2 + \frac{1}{2}\delta\alpha \| U_{,t} \|^2 \ .$$

Da nach Voraussetzung $1 > \alpha\delta$ sein soll, folgt

(2.3.10) $$\| U_{,t} \|^2 \leq \frac{-\delta}{(1-\alpha\delta)} \left(\frac{d}{dt} \| \beta \cdot \nabla U \|^2 + \| \beta \cdot \nabla U \|^2\right).$$

Zusammen mit (2.3.10) erhält man aus (2.3.9)

(2.3.11) $$\frac{d}{dt} \{ \| U \|^2 + \delta^2 \| \beta \cdot \nabla U \|^2 \} + 2\delta \| U_{,t} + \beta \cdot \nabla U \|^2$$

$$\leq \alpha \| U \|^2 + \alpha\delta^2 \| \beta \cdot \nabla U \|^2 - \frac{\alpha\delta^3}{(1-\alpha\delta)} \frac{d}{dt} \| \beta \cdot \nabla U \|^2 \ .$$

Dies zeigt insbesondere, daß die δ-Terme auch in der richtigen Größenordnung zueinander sind. Nach der Zeitintegration ergibt sich mit einer Konstanten $c > 0$

$$\| U(t) \|^2 + \delta^2 \| \beta \cdot \nabla U(t) \|^2 + \delta \int_0^t \| U_{,s} + \beta \cdot \nabla U \|^2 \, ds \leq$$

$$\| U(0) \|^2 + \delta^2 \| \beta \cdot \nabla U(0) \|^2 + c\alpha \int_0^t \| U \|^2 + \delta^2 \| \beta \cdot \nabla U \|^2 \, ds \ .$$

Darauf wendet man das Lemma von Gronwall an.

(2.3.12) $$\| U(t) \|^2 + \delta^2 \| \beta \cdot \nabla U(t) \|^2 + \delta \int_0^t \| U_{,s} + \beta \cdot \nabla U \|^2 \, ds \leq$$

$$e^{c\alpha t} (\| U(0) \|^2 + \delta^2 \| \beta \cdot \nabla U(0) \|^2) \ .$$

Daraus folgt die Aussage des Satzes.

■

Die Untersuchung der Konvergenz soll sich hier auf den Fall stückweise linearer Ansatzfunktionen beschränken. Die Aussagen der Sätze gelten entsprechend modifiziert auch für Ansatzfunktionen mit höherem Polynomgrad. Bei der Realisation der Algorithmen werden die Ansatzfunktionen als stückweise bilinear gewählt.

Für eine Funktion $u \in L^2(\Omega)$ definiert man durch

$$(u,\varphi) = (\pi u,\varphi) \quad \forall \ \varphi \in V_h$$

die L^2 - Projektion von u auf V_h . Der Fehler e wird wie üblich gesplittet zu

$$e := u - U = u - \pi u + \pi u - U =: \rho + \tau \ .$$

Nach Einsetzen der exakten Lösung u in die semidiskrete Gleichung (2.3.4) gilt für den Fehler e die Gleichung

$$(2.3.13) \qquad (e_{,t} , \varphi) + (\beta \cdot \nabla e , \varphi) + \delta (e_{,t} , \beta \cdot \nabla \varphi) + \delta (\beta \cdot \nabla e , \beta \cdot \nabla \varphi) = 0$$

$$\forall \ \varphi \in V_h \ .$$

Eine Aussage über die Konvergenz der Semidiskretisierung macht der folgende

Satz 2

Unter der Voraussetzung $1 > \alpha\delta$ gilt für $t \in (0,T]$

$$\|e(t)\| + \delta \|\beta \cdot \nabla e(t)\| \ \leq \ e^{c\alpha t} (\|e(0)\| + \delta \|\beta \cdot \nabla e(0)\|$$

$$+ h^{r+\frac{1}{2}} \{\int_0^t (\|u\|_{r+1}^2 + \|u_{,s}\|_1^2) \, ds\}^{\frac{1}{2}} \ , \ r = 0,1 \ .$$

Beweis :

Man wählt in (2.3.13) $\varphi := e - \rho$.

$$(2.3.14) \qquad \frac{1}{2} \frac{d}{dt} \|e\|^2 + (\beta \cdot \nabla e , e) + \delta (e_{,t} , \beta \cdot \nabla e) + \delta \|\beta \cdot \nabla e\|^2 =$$

$$(e_{,t} , \rho) + (\beta \cdot \nabla r , \rho) + \delta (e_{,t} , \beta \cdot \nabla \rho) + \delta (\beta \cdot \nabla e , \beta \cdot \nabla \rho) \ ,$$

oder mit der Greenschen Formel

$$\frac{d}{dt} \|e\|^2 + 2\delta (e_{,t} , \beta \cdot \nabla e) + 2\delta \|\beta \cdot \nabla e\|^2 =$$

$$(e , e \operatorname{div} \beta) + 2(e_{,t} , \rho) + 2 (\beta \cdot \nabla e , \rho) + 2\delta (e_{,t} , \beta \cdot \nabla \rho) + 2\delta (\beta \cdot \nabla e , \beta \cdot \nabla \rho) \ .$$

Jetzt setzt man in (2.3.13) $\varphi := e_t - \rho_t$.

$$(2.3.15) \qquad 2 \|e_{,t}\|^2 + 2 (\beta \cdot \nabla e , e_{,t}) + \delta \frac{d}{dt} \|\beta \cdot \nabla e\|^2 = \delta (e_{,t} , e_{,t} \operatorname{div} \beta) +$$

$$2 (e_{,t} , \rho_{,t}) + 2 (\beta \cdot \nabla e , \rho_{,t}) + 2\delta (e_{,t} , \beta \cdot \nabla \rho_{,t}) + 2\delta (\beta \cdot \nabla e , \beta \cdot \nabla \rho_{,t}) \ .$$

Aus (2.3.15) gewinnt man eine Abschätzung für $\| e_{,t} \|$. Es ist

$$2\,\|e_{,t}\|^2 \le (\tfrac{3}{4} + \delta\alpha)\|e_{,t}\|^2 - \delta\frac{d}{dt}\|\beta\cdot\nabla e\|^2 + 4\,\|\beta\cdot\nabla e\|^2 + 3\,\|\rho_{,t}\|^2 + 3\delta^2\,\|\beta\cdot\nabla\rho_{,t}\|^2 \ ,$$

und mit der Voraussetzung $1 > \alpha\delta$ folgt

$$\|e_{,t}\|^2 \le \frac{1}{1-\delta}\,(-\delta\frac{d}{dt}\|\beta\cdot\nabla e\|^2 + 4\,\|\beta\cdot\nabla e\|^2 + 3\,\|\rho_{,t}\|^2 + 3\delta^2\,\|\beta\cdot\nabla\rho_{,t}\|^2 \ . \tag{2.3.16}$$

Die Gleichung (2.3.15) multipliziert man mit δ und addiert das Ergebnis zu (2.3.14).

$$\frac{d}{dt}\{\|e\|^2 + \delta^2\,\|\beta\cdot\nabla e\|^2\} + 2\delta\,\|e_{,t} + \beta\cdot\nabla e\|^2 \le$$

$$\alpha\,\|e\|^2 + \alpha\delta^2\,\|e_{,t}\|^2 + \delta\,\|e_{,t} + \beta\cdot\nabla e\|^2 + \frac{4}{\delta}\,\|\rho\|^2 + 4\delta\,\|\beta\cdot\nabla\rho\|^2$$

$$+ 4\delta\,\|\rho_{,t}\|^2 + 4\delta^3\,\|\beta\cdot\nabla\rho_{,t}\|^2 \ .$$

Hier setzt man das Ergebnis (2.3.16) ein.

$$\frac{d}{dt}\{\|e\|^2 + \delta^2\,\|\beta\cdot\nabla e\|^2\} + 2\delta\,\|e_{,t} + \beta\cdot\nabla e\|^2 \le$$

$$c\,\{\alpha\,\|e\|^2 + \delta^2\,\|\beta\cdot\nabla e\|^2 + h^{2r+1}\,\|u\|^2_{r+1} + \|u_{,t}\|^2\} \quad ,r = 0,1 \ .$$

Dabei wurde die Fehlerabschätzung (2.2.1) für die L^2-Projektion verwendet, die auch für die nach der Zeit differenzierte Funktion u gilt. Eine Zeitintegration der letzten Gleichung

$$\|e(t)\|^2 + \delta^2\,\|\beta\cdot\nabla e(t)\|^2 + \delta\int_0^t \|e_s + \beta\cdot\nabla e\|^2\,ds \le$$

$$c\alpha\int_0^t (\|e\|^2 + \delta^2\,\|\beta\cdot\nabla e\|^2)\,ds + h^{2r+1}\int_0^t (\|u\|^2_{r+1} + \|u_{,s}\|^2)\,ds \ ,$$

und das Lemma von Gronwall liefern

$$\|e(t)\|^2 + \delta^2\,\|\beta\cdot\nabla e(t)\|^2 + \delta\int_0^t \|e_{,s} + \beta\cdot\nabla e\|^2\,ds \le$$

$$e^{c\alpha t}(\|e(0)\|^2 + \delta^2\,\|\beta\cdot\nabla e(0)\|^2 + h^{2r+1}\int_0^t (\|u\|^2_{r+1} + \|u_{,s}\|^2_1)\,ds \ .$$

Das ist die Aussage des Satzes.

■

Nach der Durchführung der Semidiskretisierung erhält man ein System von gewöhnlichen Differentialgleichungen, siehe z.B. Thomee / 8/. Zur Zeitdiskretisierung dieses Systems wird hier das Crank-Nicolson-Verfahren verwendet. Es ist eine lineare Einschrittmethode und entspricht der Zeitintegration mit der Trapezregel. Dieses Verfahren hat die Konvergenzordnung 2 und die kleinste Fehlerkonstante aller A-stabilen linearen Mehrschrittverfahren.

Der nächste Satz zeigt, daß das volldiskrete Verfahren die gleiche Stabilitätseigenschaft hat, wie die Semidiskretisierung. In den weiteren Aussagen werden die folgenden Bezeichnungen verwendet :

$u^n := u(x,nk)$, das heißt die exakte Lösung ausgewertet zum Zeitpunkt t_n ,

$U^n :=$ volldiskrete Lösung zum Zeitpunkt t_n ,

$e^n := u^n - U^n = u^n - \pi u^n + \pi u^n - U^n := \rho^n + \theta^n$.

Satz 3

Ist δ hinreichend klein, so gilt für die Lösung U^n von (2.3.5)

$$\|U^n\| + \delta \|\beta\cdot\nabla U^n\| \le e^{c\alpha t} (\|U^0\| + \delta \|\beta\cdot\nabla U^0\|) \ . \tag{2.3.17}$$

Beweis :

In (2.3.5) wählt man $\varphi := \bar{U}^n$.

$$d_t\|U^n\|^2 + \delta (d_t U^n , \beta\cdot\nabla \bar{U}^n) + \delta \|\beta\cdot\nabla \bar{U}^n\|^2 = (\bar{U}^n , \bar{U}^n \operatorname{div} \beta) \tag{2.3.18}$$

und $\varphi := d_t U^n$

(2.3.19) $$\|d_t U^n\|^2 + (\beta\cdot\nabla \bar{U}^n , d_t U^n) + \frac{1}{2}\,\delta\, d_t \|\beta\cdot\nabla U^n\|^2 =$$

$$\delta\,(d_t U^n , d_t U^n \operatorname{div}\beta) \ .$$

Die letzte Gleichung liefert außerdem eine Abschätzung für $\| d_t U^n \|$.

$$\|d_t U^n\|^2 \le \alpha\delta \|d_t U^n\|^2 + \frac{1}{4}\|d_t U^n\|^2 + \|\beta\cdot\nabla \bar{U}^n\|^2 - \frac{1}{2}\,\delta\, d_t \|\beta\cdot\nabla U^n\|^2 \ ,$$

und mit der Voraussetzung an δ

(2.3.20) $$\|d_t U^n\|^2 \le c\,\{\, \|\beta\cdot\nabla \bar{U}^n\|^2 - \frac{1}{2}\,\delta\, d_t \|\beta\cdot\nabla U^n\|^2 \,\} \ .$$

Man addiert wieder (2.3.18) und das δ-fache von (2.3.19)

(2.3.21) $$d_t\{\|U^n\|^2 + \frac{1}{2}\,\delta^2 \|\beta\cdot\nabla U^n\|^2\} + \delta \|d_t U^n + \beta\cdot\nabla \bar{U}^n\|^2 \le$$

$$\alpha \|\bar{U}^n\|^2 + \alpha\delta^2 \|d_t U^n\|^2 \ ,$$

und mit (2.3.20)

$$\le c\{\alpha \|\bar{U}^n\|^2 + \delta^2 \|\beta\cdot\nabla \bar{U}^n\|^2\} \ .$$

In Analogie zur Zeitintegration im letzten Beweis wird (2.3.21) summiert für i = 1,..,n

$$\|U^n\|^2 + \delta^2 \|\beta\cdot\nabla U^n\|^2 + k\delta \sum_{i=1}^{n} (\|d_t U^i + \beta\cdot\nabla \bar{U}^i\|^2) \le$$

$$\|U^0\|^2 + \delta^2 \|\beta\cdot\nabla U^0\|^2 + kc\alpha \sum_{i=1}^{n} (\|U^i\|^2 + \delta^2 \|\beta\cdot\nabla \bar{U}^i\|^2) \ .$$

Mit dem diskreten Gronwallschen Lemma erhält man die Aussage des Satzes.

$$\|U^n\|^2 + \delta^2 \|\beta\cdot\nabla U^n\|^2 + k\delta \sum_{i=1}^{n} (\|d_t U^i + \beta\cdot\nabla \bar{U}^i\|^2) \leq$$

$$e^{c\alpha t}(\|U^0\|^2 + \delta^2 \|\beta\cdot\nabla U^0\|^2) \ .$$

■

Um den Abschneidefehler für die Zeitdiskretisierung anzugeben, wird eine Formel verwendet, die ein Spezialfall der Euler-MacLaurinschen Summenformel ist.

Wird eine Funktion $f \in C^2(a;b)$ mit der Trapezregel numerisch integriert so gilt :

$$\int_a^b f(x)dx = \frac{h}{2}(f(a) + f(b)) + \frac{h^2}{2}\int_a^b f''(x)s(x)dx \ ,$$

wobei $h := b - a$ und $s(x) = \frac{(x-a)}{h} \cdot \frac{(x-b)}{h}$ ist.

Unter Verwendung von (2.3.5) erhält man für den Fehler e^n :

(2.3.22) $$(d_t e^n , \varphi) + (\beta\cdot\nabla \bar{e}^n , \varphi) + \delta (d_t e^n, \beta\cdot\nabla \varphi) + \delta (\beta\cdot\nabla \bar{e}^n , \beta\cdot\nabla \varphi) =$$

$$(r^n , \varphi) \qquad \forall \varphi \in V_h \ ,$$

wobei

$$(r^n , \varphi) :\equiv (d_t u^n - \bar{u}^n_{,t} , \varphi) + \delta (d_t u^n - \bar{u}^n_{,t} , \beta\cdot\nabla \varphi) \ .$$

Eine Abschätzung für den *Abschneidefehler* r^n gibt das folgende

Lemma 1

Ist $u \in H^3((0;T] ; L^2(\Omega))$, so gilt für den Abschneidefehler

$$\|r^n\| \leq c \cdot k \int_{t_{n-1}}^{t_n} \|u_{,sss}\| \, ds \ .$$

Ist u darüberhinaus aus $W^{3,\infty}((0;T] ; L^2(\Omega))$ dann gilt

(2.3.22) $$\|r^n\| \leq c \cdot k^2 \, \|u_{,ttt}\|_{L^\infty((0;T] ; L^2(\Omega))} \ .$$

Beweis :

Aus der Definition von (r^n,φ) folgt

$$(r^n,\varphi) = \frac{1}{k} \int_{t_{n-1}}^{t_n} (u_{,s},\varphi) + \delta \, (u_{,s} \, , \beta \cdot \nabla \, \varphi) \, ds - (\bar{u}^n_{,t} \, , \, \varphi) - \delta \, (\bar{u}^n_{,t} \, , \beta \cdot \nabla \, \varphi) \ .$$

Dieses Integral wird mit der Trapezregel ausgewertet. Mit der obigen Überlegung ergibt sich

$$(r^n,\varphi) = k \int_{t_{n-1}}^{t_n} \{ \, (u_{,sss} \, , \, \varphi) + \delta \, (u_{,sss} \, , \beta \cdot \nabla \, \varphi) \, \} \, s(t) \, ds \ .$$

Mit der Wahl $\varphi := r^n$ erhält man

$$\|r^n\|^2 \leq k \int_{t_{n-1}}^{t_n} \|u_{,sss}\| \cdot \|r^n\| + \delta \, \|u_{,sss}\| \cdot \|\beta \cdot \nabla \, r^n\| \, s(t) \, ds \ .$$

Da $r^n \in H^1((0;t] ; V_h(\Omega))$ gilt für r^n eine inverse Beziehung,

$$\|r^n\|_1 \leq c h^{-1} \|r^n\| \ ,$$

und mit $\delta = \epsilon h$ ergibt sich

$$\|r^n\| = ck \int_{t_{n-1}}^{t_n} \|u_{,sss}\| dt \leq c \cdot k^2 \, \|u_{,ttt}\|_{L^\infty((0;T] ; L^2(\Omega))}$$

■

Mit Hilfe dieses Lemmas kann man die Konvergenzordnung des volldiskreten Verfahrens angeben.

Satz 4

Ist die Lösung u von (2.3.1) aus $H^3((0;T] ; L^2(\Omega)) \cap H^1((0;T] ; H^1(\Omega)) \cap L^2((0;T] ; H^{r+1}(\Omega))$, $r = 0,1$, und ist $1 > \alpha\delta$ so gilt mit einer von div β und t_n exponentiell abhängigen Konstanten c

$$\|e^n\| + \delta \|\beta\cdot\nabla e^n\| \le c\{ \|e^0\| + \delta \|\beta\cdot\nabla e^0\| +$$

$$k \left(\int_0^{t_n} \|u_{,sss}\|^2 \, ds\right)^{\frac{1}{2}} + h^{r+\frac{1}{2}} \left(\int_0^{t_n} \{\|u\|_{r+1}^2 + \|u_{,s}\|_1^2\} \, ds\right)^{\frac{1}{2}} .$$

Beweis :

In der Fehleridentität (2.3.13) wählt man $\varphi := \bar{e}^n - \bar{\rho}^n$,

$$(2.3.23) \quad \tfrac{1}{2} d_t\|e^n\|^2 + \delta (d_t e^n , \beta\cdot\nabla \bar{e}^n) + \delta \|\beta\cdot\nabla \bar{e}^n \|^2 = (\bar{e}^n , \bar{e}^n \operatorname{div} \beta) +$$

$$(r^n , \bar{e}^n - \bar{\rho}^n) + (d_t e^n + \beta\cdot\nabla \bar{e}^n , \bar{\rho}^n) + \delta (d_t e^n + \beta\cdot\nabla \bar{e}^n , \beta\cdot\nabla \bar{\rho}^n) ,$$

und $\varphi := d_t e^n - d_t \rho^n$

$$(2.3.24) \quad \|d_t e^n\|^2 + (\beta\cdot\nabla \bar{e}^n , d_t e^n) + \tfrac{1}{2} \delta d_t\|\beta\cdot\nabla e^n\|^2 = \delta (d_t e^n , d_t e^n \operatorname{div} \beta) +$$

$$(r^n , d_t e^n - d_t \rho^n) + (d_t e^n + \beta\cdot\nabla \bar{e}^n , d_t \rho^n + \delta\beta\cdot\nabla d_t \rho^n) .$$

Aus der letzten Identität verschafft man sich eine Abschätzung für $\| d_t e^n \|$.

$$\|d_t e^n\|^2 \le \tfrac{1}{2} \delta d_t\|\beta\cdot\nabla e^n\|^2 + \|\beta\cdot\nabla \bar{e}^n \|^2 + \tfrac{1}{4} \|d_t e^n\|^2 + \alpha\delta \|d_t e^n\|^2 + 2 \|r^n\|^2 +$$

$$\tfrac{1}{4} \|d_t e^n\|^2 + \tfrac{1}{4} \|d_t \rho^n\|^2 + \tfrac{1}{4} \|d_t e^n\|^2 + 2 \| d_t \rho^n + \delta\beta\cdot\nabla d_t \rho^n\|^2 + \tfrac{1}{4} \|\beta\cdot\nabla \bar{e}^n \|^2 .$$

Da $1 - \alpha\delta > 0$ vorausgesetzt wurde, folgt

(2.3.25) $$\|d_t e^n\|^2 \le c\{ - \delta d_t \|\beta\cdot\nabla e^n\|^2 + \|\beta\cdot\nabla \bar{e}^n\|^2 + \|d_t\rho^n\|^2 +$$

$$\delta \|\beta\cdot\nabla d_t\rho^n\|^2 + \|r^n\|^2\} ,$$

mit $c \approx \frac{1}{1-\alpha\delta}$.

Die Multiplikation von (2.3.24) mit δ und die Addition des Ergebnisses zu (2.3.23) liefert

(2.3.26) $$\tfrac{1}{2} d_t\{\|e^n\|^2 + \delta^2 \|\beta\cdot\nabla e^n\|^2\} + \delta \|d_t e^n + \beta\cdot\nabla \bar{e}^n\|^2 \le (\alpha + \tfrac{1}{2}) \|e^n\|^2 +$$

$$2\|r^n\|^2 + (\tfrac{1}{2} + \tfrac{1}{\delta}) \|\bar{\rho}^n\|^2 + \tfrac{3\delta}{4} \|d_t e^n + \beta\cdot\nabla \bar{e}^n\|^2 + (\alpha\delta^2 + \delta^2) \|d_t e^n\|^2 +$$

$$(\delta^2 + 2\delta) \|d_t\rho^n\|^2 + \delta \|\beta\cdot\nabla \bar{\rho}^n\|^2 + 2\delta^2 \|\beta\cdot\nabla d_t\rho^n\|^2 .$$

Hier benutzt man das Ergebniss (2.3.25) und erhält

$$d_t\{\|e^n\|^2 + \delta^2 \|\beta\cdot\nabla e^n\|^2\} + \delta \|d_t e^n + \beta\cdot\nabla \bar{e}^n\|^2 \le c\{\|\bar{e}^n\|^2 + \delta^2 \|\beta\cdot\nabla \bar{e}^n\|^2 +$$

$$\|r^n\|^2 + \tfrac{1}{\delta} \|\bar{r}^n\|^2 + \delta \|d_t\rho^n\|^2 + \delta \|\beta\cdot\nabla \bar{\rho}^n\|^2 + \delta^2 \|\beta\cdot\nabla d_t\rho^n\|^2\} .$$

Diese Ungleichung wird mit k multipliziert und von $i = 1$ bis n summiert.

$$\|e^n\|^2 + \delta^2 \|\beta\cdot\nabla e^n\|^2\} + \delta k \sum_{i=1}^{n} \|d_t e^i + \beta\cdot\nabla \bar{e}^i\|^2 \le c\{\|e^0\|^2 + \delta^2 \|\beta\cdot\nabla e^0\|^2 +$$

$$ck \sum_{i=1}^{n} (\|\bar{e}^i\|^2 + \delta^2 \|\beta\cdot\nabla \bar{e}^i\|^2) + k^2 \int_0^{t_n} \|u_{,sss}\|^2 ds +$$

$$h^{2r+1} \int_0^{t_n} \{\|u\|_{r+1}^2 + \|u_{,s}\|_1^2\} ds .$$

Mit dem diskreten Lemma von Gronwall kann man nun die Aussage des Satzes folgern.

$$\|e^n\|^2 + \delta^2 \|\beta\cdot\nabla e^n\|^2\} + \delta k \sum_{i=1}^{n} \|d_t e^i + \beta\cdot\nabla \bar{e}^i\|^2 \le e^{ct}\, c\, \{\|e^0\|^2 + \delta^2 \|\beta\cdot\nabla e^0\|^2 +$$

$$k^2 \left(\int_0^{t_n} \|u_{,sss}\|^2 ds + h^{2r+1} \int_0^{t_n} \{\|u\|_{r+1}^2 + \|u_{,s}\|_1^2\}\, ds \right) .$$

■

Lemma 1 zeigt, daß man die Aussage von Satz 4 noch verbessern kann. Ist nämlich $\| u_{,ttt} \|$ beschränkt in der Zeit, so erhält man eine Zeitdiskretisierung von 2. Ordnung.

Bei der algorithmischen Durchführung des Verfahrens wählt man k in der Größenordnung von h . Unter dieser Voraussetzung kann man den in (2.3.26) störenden Term $\delta^2 \| d_t e^n \|^2$ umformen und benötigt nicht die Abschätzung (2.3.25) .

$$(1+\alpha)\delta^2 \|d_t e^n\|^2 = (1+\alpha)\, \epsilon\, h^2 k^{-2} \|e^n - e^{n-1}\|^2 \le (1+\alpha)\, \epsilon\, c\, \|e^n - e^{n-1}\|^2 .$$

Das Ergebnis dieser Umformung kann man jetzt mit dem Term $\alpha \| \bar{e}^n \|$ zusammenfassen. Dies bedeutet, daß bei der Realisation des Verfahrens keine Restriktion an δ gefordert werden muß.

2.4. Analyse des Verfahrens für den Fall div $\beta < 0$

Wird das *Modellproblem*

$$u_{,t} + \beta\cdot\nabla u = 0 \quad \text{in } \Omega \times (0;T] ,$$
$$u(x,0) = u_0 \quad \text{in } \Omega \text{ für } t = 0 ,$$

variationell formuliert

$$(u_{,t} + \beta\cdot\nabla u\, ,\, v) = 0 \quad \text{in } \Omega \times (0;T] ,$$
$$u(x,0) = u_0 \quad \text{in } \Omega \text{ für } t = 0, \quad \forall\, v \in H^1(\Omega) ,$$

so erhält man, da keine Vorzeichenbedingung an div β gestellt wurde, als zweiten Summan-

den den indefiniten Term $(\beta\cdot\nabla u , v)$. Im weiteren wird die Voraussetzung

(2.4.1) $$\operatorname{div}\beta \leq 0$$

gemacht, unter der auch Nävert /34/ sein Verfahren analysiert. Die Greensche Formel liefert dann für den fraglichen Term

(2.4.2) $$(\beta\cdot\nabla u , u) = -\frac{1}{2}(u , u \operatorname{div}\beta) + \frac{1}{2}\langle u , u\rangle \geq 0 \quad \forall u \in H^1(\Omega) .$$

D.h. unter der getroffenen Voraussetzung ist $(\beta\cdot\nabla u , u)$ ein definiter Term.

Die weitere Analyse des Verfahrens für diesen Spezialfall verläuft weitgehend analog zur bisherigen mit dem Unterschied, daß ein verallgemeinertes Gronwallsches Lemma benutzt wird. Das wichtigste Ergebnis dieser Analyse ist, daß unter der obigen Voraussetzung (2.4.1) die Konstante c in den Aussagen der vorausgegangenen Sätze nicht mehr exponentiell mit der Länge des Zeitintervalls wächst, sondern nur noch linear.

Als erstes wird die Stabilität der *Semidiskretisierung* betrachtet. Die Semidiskretisierung lautet :

Finde ein $U \in H^1((0;T] ; V_h)$, das

(2.4.3) $$(U_{,t} + \beta\cdot\nabla U , \varphi + \delta\beta\cdot\nabla \varphi) = 0 \quad \forall \varphi \in V_h$$

erfüllt.

Satz 1

Für die Lösung U der Semidiskretisierung (2.4.3) gilt $\forall t \in (0;T]$

$$\|U(t)\| + \delta \|\beta\cdot\nabla U\| \leq \|U(0)\| + \delta \|\beta\cdot\nabla U(0)\| .$$

Beweis:

Man benutzt (2.4.2) und setzt in der semidiskreten Gleichung (2.4.3) $\varphi := U$,

(2.4.4) $$\frac{1}{2}\frac{d}{dt}\|U\|^2 + \delta (U_{,t} , \beta\cdot\nabla U) + \delta \|\beta\cdot\nabla U\|^2 \leq 0$$

und $\varphi := U_{,t}$

(2.4.5) $$\|U_{,t}\|^2 + (\beta\cdot\nabla U , U_{,t}) + \tfrac{1}{2}\,\delta\,\frac{d}{dt}\,\|\beta\cdot\nabla U\|^2 \le 0$$

Nach der Multiplikation von (2.4.5) mit δ addiert man diese Gleichung zu (2.4.4) und erhält

$$\frac{d}{dt}\,(\|U\|^2 + \delta^2\,\|\beta\cdot\nabla U\|^2) + \delta\,\|U_{,t} + \beta\cdot\nabla U\|^2 \le 0 \ .$$

Die zeitliche Integration dieser Gleichung liefert

(2.4.6) $$\|U(t)\|^2 + \delta^2\,\|\beta\cdot\nabla U(t)\|^2 + \delta\int_0^t \|U_{,s} + \beta\cdot\nabla U\|^2 \le$$

$$\|U(0)\|^2 + \delta^2\,\|\beta\cdot\nabla U(0)\|^2 \ ,$$

und damit die Aussage des Satzes.

■

Wie in Kapitel 2.3 erhält man eine Gleichung für den Fehler e :

(2.4.7) $$(e_{,t} + \beta\cdot\nabla e\,,\ \varphi + \delta\beta\cdot\nabla\varphi) = 0 \qquad \forall\,\varphi\in V_h \ ,$$

und kann damit die Konvergenz der Semidiskretisierung untersuchen.

Satz 2

Ist u die Lösung des Modellproblems, so gilt für den Fehler $e := u - U$

$$\|e(t)\| + \delta\,\|\beta\cdot\nabla e(t)\| \le \|e(0)\| + \delta\,\|\beta\cdot\nabla e(0)\| + ch^{r+\frac{1}{2}}\int_0^t \{\|u\|^2_{r+1} + \|u_{,s}\|^2_1\}\,ds$$

für $r = 0$ oder 1 und $\forall\, t\in(0;T]$.

Beweis:

Man wählt in der Fehleridentität (2.4.7) $\varphi := e - \rho$

$$(2.4.8)\quad \frac{d}{dt}\|e\|^2 + 2\delta\,(e_{,t}\,,\,\beta\cdot\nabla\, e) + 2\delta\,\|\beta\cdot\nabla\, e\|^2 \;\le\; 2\,(e_{,t} + \beta\cdot\nabla\, e\,,\,\rho + \delta\beta\cdot\nabla\,\rho)$$

$$\le \tfrac{1}{2}\,\delta\,\|e_{,t} + \beta\cdot\nabla\, e\|^2 + \tfrac{4}{\delta}\,\|\rho\|^2 + 4\delta\,\|\beta\cdot\nabla\,\rho\|^2 \quad ,$$

und $\varphi := e_{,t} - \rho_{,t}$

$$(2.4.9)\quad 2\,\|e_{,t}\|^2 + 2\,(\beta\cdot\nabla\, e\,,\,e_{,t}) + \delta\,\frac{d}{dt}\,\|\beta\cdot\nabla\, e\|^2 \;\le\; 2\,(e_{,t} + \beta\cdot\nabla\, e\,,\,\rho_{,t} + \delta\beta\cdot\nabla\,\rho_{,t})$$

$$\le \tfrac{1}{2}\,\|e_{,t} + \beta\cdot\nabla\,\|^2 + 4\,\|\rho_{,t}\|^2 + 4\delta^3\,\|\beta\cdot\nabla\,\rho_{,t}\|^2 \quad .$$

Die Addition von (2.4.8) mit dem δ-fachen von (2.4.9) ergibt

$$\frac{d}{dt}\,\{\|e\|^2 + \delta^2\,\|\beta\cdot\nabla\, e\|^2\} + \delta\,\|e_t + \beta\cdot\nabla\, e\|^2 \;\le\; 4\,\{\tfrac{1}{\delta}\,\|\rho\|^2 + \delta\,\|\beta\cdot\nabla\,\rho\|^2 +$$

$$\delta\,\|\rho_t\|^2 + \delta^3\,\|\beta\cdot\nabla\,\rho_t\|^2\} \quad .$$

Jetzt integriert man in der Zeit und erhält mit den bekannten Abschätzungen für $\|\,\rho\,\|$

$$(2.4.10)\qquad \|e(t)\|^2 + \delta^2\,\|\beta\cdot\nabla\, e(t)\|^2 + \delta\int_0^t \|e_{,s} + \beta\cdot\nabla\, e\|^2\,ds \;\le\; \|e(0)\|^2$$

$$+ \delta^2\,\|\beta\cdot\nabla\, e(0)\|^2 + ch^{2r+1}\int_0^t \{\|u\|_{r+1} + \|u_{,s}\|_1\}\,ds \quad .$$

Damit ist der Satz bewiesen.

■

Der nächste Satz zeigt, daß das volldiskrete Verfahren die selben Stabilitätseigenschaften hat, wie die Semidiskretisierung. Das *volldiskrete Verfahren* lautet :

(2.4.11) $$(d_t U^n + \beta\cdot\nabla \bar{U}^n\,,\, \varphi + \delta\beta\cdot\nabla\varphi) = 0 \quad \forall\,\varphi \in V_h\ ,$$

$$U^0 = \pi u_0$$

Satz 3

Für die Lösung U^n, $n \in \{1,\ldots,N\}$ von (2.4.11) gilt die Abschätzung

$$\|U^n\| + \delta\,\|\beta\cdot\nabla U^n\| \;\le\; \|U^0\| + \delta\,\|\beta\cdot\nabla U^0\|\ .$$

Beweis:

In (2.4.11) wählt man $\varphi := \bar{U}^n$

(2.4.12) $$d_t\|U^n\|^2 + \delta\,(d_t U^n\,,\,\beta\cdot\nabla\bar{U}^n) + \delta\,\|\beta\cdot\nabla\bar{U}^n\|^2 \;\le\; 0\ ,$$

und $\varphi := d_t U^n$

(2.4.13) $$\|d_t U^n\|^2 + (\beta\cdot\nabla\bar{U}^n\,,\,d_t\,U^n) + \tfrac{1}{2}\,\delta\,\frac{d}{dt}\,\|\beta\cdot\nabla U^n\|^2 \;\le\; 0\ .$$

Die Addition von (2.4.12) mit dem δ-fachen von (2.4.13) liefert

$$d_t\{\|U^n\|^2 + \delta^2\,\|\beta\cdot\nabla U^n\|^2\} + \delta\,\|d_t U^n + \beta\cdot\nabla\bar{U}^n\|^2 \;\le\; 0\ .$$

Jetzt summiert man diese Gleichung von $i = 1$ bis n und erhält damit

(2.4.14) $$\|U^n\|^2 + \delta^2\,\|\beta\cdot\nabla U^n\|^2 + \delta k \sum_{i\,=\,1}^{n} \|d_{,t}U^i + \beta\cdot\nabla\bar{U}^i\|^2 \;\le$$

$$\|U^0\|^2 + \delta^2\,\|\beta\cdot\nabla U^0\|^2\ .$$

Daraus ergibt sich die Aussage des Satzes.

■

Als letztes wird die Konvergenzordnung des volldiskreten Verfahrens angegeben. Wie bereits erwähnt wurde, liegt der Unterschied zur entsprechenden Analyse im vorherigen Kapitel darin, daß jetzt nur noch ein lineares Anwachsen des Fehlers in der Zeit stattfinden kann. Außerdem werden keine Bedingungen an δ gestellt.

Satz 4

Ist u die Lösung des Modellproblems und ist U^n durch das volldiskrete Verfahren (2.4.11) gegeben, dann gilt für hinreichend kleines k , $t = nk$

$$\|e^n\| + \delta \|\beta\cdot\nabla e^n\| + \delta k \sum_{i=1}^{n} \|d_t e^i + \bar{e}^i\| \le ct \{\|e^0\| + \delta \|\beta\cdot\nabla e^0\| +$$

$$\int_0^t k \|u_{,sss}\|^2 + h^{r+\frac{1}{2}} (\|u\|^2_{r+1} + \|u_{,s}\|^2_1)\ ds \} \ .$$

Beweis:

In der Gleichung (2.4.7) für den Fehler e wählt man als Testfunktion $\varphi := \bar{e}^n - \bar{\rho}^n$

$$(2.4.15) \qquad \tfrac{1}{2} d_t \|e^n\|^2 + \delta (d_t e^n , \beta\cdot\nabla \bar{e}^n) + \delta \|\beta\cdot\nabla \bar{e}^n\|^2 \le (r^n , \bar{e}^n - \bar{\rho}^n) +$$

$$\tfrac{\delta}{4} \|d_t e^n + \beta\cdot\nabla \bar{e}^n\|^2 + \tfrac{2}{\delta} \|\bar{\rho}^n\|^2 + 2\delta \|\beta\cdot\nabla \bar{\rho}^n\|^2 \ ,$$

und $\varphi := d_t e - d_t \rho^n$

$$(2.4.16) \qquad \|d_t e^n\|^2 + (\beta\cdot\nabla \bar{e}^n , d_t e^n) + \tfrac{1}{2} \delta d_t \|\beta\cdot\nabla e^n\|^2 \le (r^n , d_t e^n - d_t \rho^n) +$$

$$\tfrac{1}{4} \|d_t e^n + \beta\cdot\nabla \bar{e}^n\|^2 + 2 \|d_t \rho^n\|^2 + 2\delta^2 \|\beta\cdot\nabla d_t \rho^n\|^2 \ .$$

Aus der letzten Gleichung erhält man eine Abschätzung für $\| d_t e^n \|$ ohne Bedingung an δ .

(2.4.17) $$\|d_t e^n\|^2 \le -\delta\, d_t \|\beta\cdot\nabla e^n\|^2 + c\,\{\|\beta\cdot\nabla \bar{e}^n\|^2 + \|d_t\rho^n\|^2 +$$

$$\delta^2 \|\beta\cdot\nabla\, d_t\rho^n\|^2 + \|r^n\|^2\}\ .$$

Nun multipliziert man Gleichung (2.4.16) mit δ und addiert sie zu (2.4.15).

$$d_t\{\|e^n\|^2 + \delta^2 \|\beta\cdot\nabla e^{n2}\} + \delta\,\|d_t e^n + \beta\cdot\nabla \bar{e}^n\|^2 \le 4\,\{\tfrac{1}{\delta}\|\bar{\rho}^n\|^2 +$$

$$\delta\,\|\beta\cdot\nabla\bar{\rho}^n\|^2 + \delta\,\|d_t\rho^n\|^2 + \delta^3\,\|\beta\cdot\nabla\, d_t\rho^n\|^2\} +$$

$$2\,\|r^n\|\cdot\{\|\bar{e}^n\| + \|\bar{\rho}^n\| + \delta\,\|d_t e^n\| + \delta\,\|d_t\rho^n\|\}\ .$$

Mit Hilfe von (2.4.17) kann man die rechte Seite weiter abschätzen.

$$\le 4\,\{\tfrac{1}{\delta}\|\bar{\rho}^n\|^2 + \delta\,\|\beta\cdot\nabla\bar{\rho}^n\|^2 + \delta\,\|d_t\rho^n\|^2 + \delta^3\,\|\beta\cdot\nabla\, d_t\rho^n\|^2\} +$$

$$\|r^n\|\cdot\{\|\bar{e}^n\| + \delta c\,\|\beta\cdot\nabla\bar{e}^n\|\} + 4\,\|r^n\|^2 +$$

$$\|\rho^n\|^2 - \delta\, d_t\|\beta\cdot\nabla e^n\|^2 + c\delta\,\|d_t\rho^n\|^2 + c\delta^2\,\|\beta\cdot\nabla\, d_t\rho^n\|^2\ .$$

Jetzt multipliziert man die Ungleichung mit k und summiert auf beiden Seiten von $i = 1$ bis n.

$$\|e^n\|^2 + \delta^2\,\|\beta\cdot\nabla e^n\|^2 + \delta k \sum_{i=1}^{n} \|d_t e^i + \beta\cdot\nabla \bar{e}^i\|^2 \le \|e^0\|^2 + \delta^2\,\|\beta\cdot\nabla e^0\|^2 +$$

$$ck \sum_{i=1}^{n} \|r^i\|\{\|\bar{e}^i\| + \delta\,\|\beta\cdot\nabla\bar{e}^i\|\} +$$

$$ck^2 \int_0^t \|u_{,sss}\|^2\, ds + h^{2r+1}\,(\|u\|_{r+1}^2 + \|u_{,s}\|_1^2)\, ds\ .$$

Für hinreichend kleines k kann man nun mit dem diskreten verallgemeinerten Gronwallschen Lemma folgern, daß

$$(2.4.18)\qquad \|e^n\|^2 + \delta^2 \|\beta\cdot\nabla e^n\|^2 + \delta k \sum_{i=1}^{n} \|d_t e^i + \beta\cdot\nabla \bar{e}^i\|^2 \le ct^2 (\|e^0\|^2 +$$

$$\delta^2 \|\beta\cdot\nabla e^0\|^2 + c \int_0^t k^2 \|u_{,sss}\|^2 ds + h^{2r+1} (\|u\|^2_{r+1} + \|u_{,s}\|^2_1) ds$$

mit $r = 0$ oder 1 .

Damit ist die Aussage des Satzes bewiesen.

■

2.5 Analyse des Verfahrens für eine Erhaltungsgleichung

In diesem Kapitel wird das entwickelte Verfahren auf eine nichtlineare hyperbolische Erhaltungsgleichung angewendet und analysiert. Dazu wählt man als speziellen Vertreter die eindimensionale Burger-Gleichung, d.h. man sucht eine Funktion $u(x,t) : \Omega \times I \to \mathbb{R}$, mit $I \subset\subset \mathbb{R}_+$ und $\Omega \subset\subset \mathbb{R}$, die die Gleichung

$$(2.5.1)\qquad \begin{aligned} u_{,t} + uu_{,x} &= 0 \quad \forall x \in \Omega , t \in I \\ u(x,0) &= u_0 \quad \forall x \in \Omega \end{aligned}$$

erfüllt. Der Einfachheit halber soll wieder angenommen werden, daß u_0 kompakten Träger auf $\Omega' \subset\subset \Omega$ hat. Daraus folgt, daß auch u für alle $t \in I$ kompakten Träger in dem beschränkten Gebiet Ω hat. Die Gleichung (2.5.1) ist in nicht-konservativer Schreibweise gegeben und lautet in konservativer Form

$$(2.5.2)\qquad \begin{aligned} u_{,t} + (\tfrac{1}{2} u^2)_{,x} &= 0 \quad \forall x \in \Omega , t > 0 , t \in I , \\ u(x,0) &= u_0 \quad \forall x \in \Omega \end{aligned}$$

Bei der numerischen Lösung einer Erhaltungsgleichung ist es von Bedeutung, welche Formulierung man wählt. Z.B. gilt für die Diskretisierung von Gleichungen der Form (2.5.2) die Theorie von Lax /82/ , die auf Diskretisierungen von (2.5.1) nicht anwendbar ist.

Eine ausführliche Analyse der Burger-Gleichung wird in Springer Lecture Notes 1047 /23/ gegeben.

Das *volldiskrete Verfahren* erhält man wieder durch einen FEM-Ansatz in der Ortsvariablen und einer Zeitdiskretisierung mit dem Crank-Nicolson-Verfahren. Die hier und später für das hyperbolische System verwendete Form des Crank-Nicolson-Verfahrens wird in der Literatur oft als "semiimplizite Mittelpunktsregel" bezeichnet. Das Verfahren lautet in der üblichen Notation :

Finde ein $U^n \in V_h \quad n \geq 1$, das

(2.5.3) $$(d_t U^n + \bar{U}^n \bar{U}^n{}_{,x} , \varphi + \delta \bar{U}^n \varphi_{,x}) = 0 \ ,$$

$$U^0(x) = \pi u_0 \quad \forall\, x \in \Omega, \forall\, \varphi \in V_h$$

erfüllt. Da es sich um ein vollimplizites Zeitschrittverfahren handelt, muß zur Durchführung eines Zeitschrittes ein nichtlineares Gleichungssystem gelöst werden. Mit Hilfe einer Variante des Brouwerschen Fixpunktsatzes wird gezeigt, daß dies immer möglich ist.

Lemma 1 (Temam /10/ pp.166 ff)

Sei X ein Hilbertraum mit dim X $\equiv$ n , n $\in \mathbb{N}$ versehen mit dem Skalaprodukt [.,.] und der dadurch erzeugten Norm $\| \cdot \|$. Erfüllt die stetige Selbstabbildung P

$$[P(\xi) , \xi] > 0 \qquad \forall \| \xi \| = K > 0 \ ,$$

dann gibt es ein $\xi \in X$ mit $\| \xi \| \leq K$, so daß

$$P(\xi) = 0$$

ist.

■

Damit kann man den folgenden Satz beweisen, der die Durchführbarkeit des Verfahrens sichert.

Satz 1

Für jedes n mit $0 \le n \le N$ hat (2.5.3) mindestens eine Lösung.

Beweis:

Mit der neuen Variablen $V = \bar{U}^n$ hat (2.5.3) die Gestalt

$$(V , \varphi) + \frac{k}{2}(VV_{,x} , \varphi) + \delta (V , V\varphi_{,x}) + \frac{k\delta}{2} (VV_{,x} , V\varphi_{,x}) -$$

$$(U^n , \varphi) - \delta (U^n , V\varphi_{,x}) = 0 \ .$$

Diese Gleichung dient zur Definition des Funktionals

$$P : V_h \to V_h^* \ ,$$

wobei V_h^* wie üblich den Dualraum von V_h bezeichnet. Für die duale Paarung erhält man mit dem L^2-Skalarprodukt

$$(P(V) , \varphi) := (V , \varphi) + \frac{k}{2} (VV_{,x} , \varphi) + \delta (V , V\varphi_{,x}) + \frac{k\delta}{2} (VV_{,x} , V\varphi_{,x}) -$$

$$(U^n , \varphi) - \delta (U^n , V\varphi_{,x}) \ .$$

In dieser Definitionsgleichung wählt man $\varphi := V$.

$$(P(V) , V) = \|V\|^2 + \frac{k\delta}{2} \|VV_{,x}\|^2 - (U^n , V) - \delta (U^n , VV_{,x})$$

$$\ge \frac{3}{4} \|V\|^2 + (\frac{k\delta}{2} - \frac{k\delta}{3}) \|VV_{,x}\|^2 - (\frac{3\delta}{4k} + 1) \|U^n\|^2 \ .$$

Da $\delta = O(h) = O(k)$ folgt

$$(P(V) , V) \ge c \{\|V\|^2 + \|VV_{,x}\|^2 - \|U^n\|^2\} \ge 0 \ ,$$

und damit

$$(P(V) , V) > 0 \quad \text{für } \|V\|^2 = K \text{ hinreichend groß} \ .$$

Also existiert nach dem vorausgegangenen Lemma mindestens eine Lösung.

■

Als generelle Voraussetzung an die Lösung U der Semidiskretisierung soll

$$\|U\|_{L^\infty((0;T]\,;\,L^\infty(\Omega))} + \sqrt{\delta}\,\|U_{,x}\|_{L^\infty((0;T]\,;\,L^\infty(\Omega))} \le c \tag{2.5.4a}$$

erfüllt sein. Das gleiche soll auch für die Lösung U^n des volldiskreten Problems gelten :

$$\sup_{n\le N} \{\|U^n\|_{L^\infty(\Omega))} + \sqrt{\delta}\,\|U^n_{,x}\|_{L^\infty(\Omega))}\} \le c\ . \tag{2.5.4b}$$

Diese Schranken sind für die Stabilitätsbeweise der Semidiskretisierung und des volldiskreten Verfahrens nötig. Die Aussagen dieser Stabilitätssätze zeigen darüber hinaus, daß die Lösungen der Diskretisierungen durch die Anfangsdaten beschränkt sind. Es ist eine noch offene Frage, wie man die Beschränktheit (2.5.4) der U^n , die auch im numerischen Experiment zu beobachten ist, für das hier entwickelte Verfahren zeigen kann. Beweise für die Beschränktheit der Folge der Approximierenden finden sich in der Literatur für solche Verfahren, die einen FEM-Ansatz für (2.5.1) mit einer "least-squares"- Methode koppeln, z.B. Hughes /37/. Da damit eine Symmetrisierung des Problems erreicht wird, sind diese Beweistechniken auf das hier entwickelte Verfahren nicht übertragbar.

Die *Semidiskretisierung* von (2.5.1) lautet : Gesucht ist ein $U \in H((0;T]\,;\,V_h)$, das

$$(U_{,t} + UU_{,x}\,,\,\varphi + \delta U\varphi_{,x}) = 0 \qquad \forall\, x \in \Omega,\quad t > 0\ , \tag{2.5.5}$$

$$U(x,0) = \pi u_0 \qquad \forall\, x \in \Omega,\quad \forall\, \varphi \in V_h$$

Der nächste Satz garantiert die Stabilität der Semidiskretisierung. Er zeigt, daß die Norm der Lösung U der Semidiskretisierung durch die Norm der Anfangsdaten beschränkt ist.

Satz 2

Unter der Voraussetzung (2.5.4a) gilt für die Lösung U von (2.5.5) für hinreichend kleinen Dämpfungsparameter δ

$$\|U(t)\| + \delta\,\|UU_{,x}\| + \sqrt{\delta}\,\|\|U_{,t} + UU_{,x}\|\| \le c\,\{\|U^0\| + \delta\,\|U^0 U^0_{,x}\|\} \tag{2.5.6}$$

$$\forall\, t \in (0;T]\ .$$

Bemerkung

Hier und im weiteren wird die Notaion

$$|||v|||_r := (\int_0^T \|v\|_r^2 \, dt)^{\frac{1}{2}}$$

verwendet. In diesem Abschnitt wird das Lemma von Gronwall wieder in seiner üblichen Form angewandt. Deshalb hängen die auftretenden Konstanten c im allgemeinen exponentiell vom Endzeitpunkt T ab.

■

Beweis von Satz 2

Da u_0 und damit U für endliche T kompakten Träger haben gilt

$$\int_{\mathbb{R}} UUU_{,x} \, dx = \frac{1}{3} \int_{\mathbb{R}} (U)^3_{,x} \, dx = 0 \ .$$

Damit ergibt sich mit $\varphi := U$ aus (2.5.5)

$$\frac{1}{2} \frac{d}{dt} \|U\|^2 + \delta (U_{,t} , UU_{,x}) + \delta \|UU_{,x}\|^2 = 0 \ . \tag{2.5.6}$$

Wählt man in (2.5.5) $\varphi := U_{,t}$ und beachtet, daß

$$\frac{d}{dt} \int_{\mathbb{R}} (uv)^2 \, dx = 2 \int_{\mathbb{R}} (uv)(u_{,t}v + uv_{,t}) \, dx \ ,$$

so erhält man

$$\|U_{,t}\|^2 + (UU_{,x} , U_{,t}) + \frac{1}{2} \delta \frac{d}{dt} \|UU_{,x}\|^2 = \frac{1}{2} \delta (U_{,t} , U_{,x}U_{,t}) + \tag{2.5.7}$$

$$\frac{1}{2} \delta (UU_{,x} , U_{,x}U_{,t}) \ .$$

Aus dieser Gleichung gewinnt man eine Abschätzung für $\| U_{,t} \|$:

$$\|U_{,t}\|^2 \leq -\frac{1}{2} \delta \frac{d}{dt} \|UU_{,x}\|^2 + 3 \|UU_{,x}\|^2 + \frac{1}{8} \|U_{,t}\|^2 + \tag{2.5.8}$$

$$\frac{1}{2} \delta \|U_{,x}\|_\infty \|U_{,t}\|^2 + \frac{1}{8} \delta^2 \|U_{,x}\|_\infty^2 \|U_{,t}\|^2 \ .$$

Für U gilt die *inverse Beziehung*

$$\|U_x\|_\infty \leq \frac{2}{h} \|U\|_\infty \ .$$

Damit ist

$$\delta \|U_{,x}\|_\infty = \epsilon h \|U_{,x}\|_\infty \leq \epsilon \|U\|_\infty \ ,$$

und mit der Voraussetzung (2.5.4a) gilt für hinreichend kleines ϵ

$$1 - \epsilon \|U\|_\infty \geq c > 0 \ .$$

Dies verwendet man in (2.5.8)

$$\|U_{,t}\|^2 \leq c \{ - \delta \frac{d}{dt} \|UU_{,x}\|^2 + \|UU_{,x}\|^2 \} \ . \tag{2.5.9}$$

Jetzt wird (2.5.7) mit δ multipliziert und zu (2.5.6) addiert.

$$\frac{1}{2} \frac{d}{dt} \{\|U\|^2 + \delta^2 \|UU_{,x}\|^2\} + \delta \|U_{,t} + UU_{,x}\|^2 \leq \frac{1}{2} \delta^2 \|U_{,x}\|_\infty \|U_{,t}\|^2 +$$

$$\delta^2 \|UU_{,x}\|^2 + \frac{1}{4} \delta^2 \|U_{,x} U_{,t}\|^2 \ .$$

Mit der Voraussetzung (2.5.4) ist

$$\frac{1}{4} \delta^2 \|U_{,x} U_{,t}\|^2 \leq \delta^2 \|U_{,x}\|_\infty^2 \|U_{,t}\|^2 \leq \delta \|U_{,t}\|^2 \ ,$$

und man kann in der letzten Ungleichung die rechte Seite weiter abschätzen zu

$$\leq c \{\delta \|U_{,t}\|^2 + \delta^2 \|UU_{,x}\|^2\} \ .$$

Den Term $\| U_{,t} \|^2$ auf der rechten Seite schätzt man mit (2.5.9) ab und erhält

$$\frac{d}{dt} \{\|U\|^2 + \delta^2 \|UU_{,x}\|^2\} + \delta \|U_{,t} + UU_{,x}\|^2 \leq c \{ \delta^2 \|UU_{,x}\|^2 + \delta \|UU_{,x}\|^2\} \ ,$$

was mit der Voraussetzung (2.5.4) schließlich

$$\frac{d}{dt} \{\|U\|^2 + \delta^2 \|UU_{,x}\|^2\} + \delta \|U_{,t} + UU_{,x}\|^2 \leq c \{ \|U\|^2 + \delta^2 \|UU_{,x}\|^2\}$$

ergibt. Diese Gleichung integriert man von 0 bis t und erhält daraus mit dem Lemma von Gronwall

$$\|U(t)\|^2 + \delta^2 \|U(t)U(t)_{,x}\|^2 + \delta \int_0^t \|U_{,s} + UU_{,x}\|^2 \, dt \leq$$

$$c \{\|U(0)\|^2 + \delta^2 \|U(0)U(0)_{,x}\|^2\} \ .$$

Damit ist der Satz bewiesen.

■

Als nächstes wird die Stabilität und die Konvergenz mit klassischen Energie-Methoden, wie sie z.B auch in Thomee / 8/ verwendet werden, bewiesen. Dazu muß man voraussetzen, daß die Lösung u von (2.5.1) hinreichend regulär ist. Dann ist es möglich, eine Konvergenzordnung für das Verfahren anzugeben. Sei also

$$u_{,t} \in L^2((0;T] \, ; H^2(\Omega)) \qquad \text{und} \qquad u_0 \in H^2(\Omega)$$

die Lösung von (2.5.1). Sie ist damit insbesondere eindeutig. Für den Fehler e hat man aus

$$(u_{,t} + uu_{,x} \, , \varphi + \delta U\varphi_{,x}) = 0 \quad \forall \, \varphi \in V_h \ ,$$

und

$$(U_{,t} + UU_{,x} \, , \varphi + \delta U\varphi_{,x}) = 0 \quad \forall \, \varphi \in V_h \ ,$$

daß

$$(e_{,t} + Ue_{,x} \, , \varphi + \delta U\varphi_{,x}) = (-eu_{,x} \, , \varphi + \delta U\varphi_{,x}) \quad \forall \, \varphi \in V_h \ , \tag{2.5.10}$$

ist. Damit kann man die Konvergenz der Semidiskretisierung angeben. Als Voraussetzung für die nachfolgenden Sätze muß man zusätzlich zur Voraussetzung (2.5.4) fordern, daß für die Lösung U der Semidiskretisierung

$$\|U_{,x}\|_{L^\infty((0;T] \, ; L^\infty(\mathbb{R}))} \leq c$$

ist. Die in den nachfolgenden Sätzen auftretenden Konstanten hängen im allgemeinen von $\| u \|_2$ ab.

Satz 3

Unter den obigen Voraussetzungen gilt für den Fehler e der Semidiskretisierung

$$\|e(t)\| + \delta \|Ue(t)_{,x}\| \leq c \{\|e(0)\| + \delta \|e(0)_{,x}\| + h^{r+\frac{1}{2}} (|||u|||_{r+1} + |||u_{,t}|||_1) .$$

Beweis:

In (2.5.10) wählt man $\varphi := e - \rho$.

$$(e_{,t} + Ue_{,x} , e + \delta Ue_{,x}) = (e_{,t} + Ue_{,x} , \rho + \delta U\rho_{,x}) -$$

$$- (eu_{,x} , e + \delta Ue_{,x}) + (eu_{,x} , \rho + \delta U\rho_{,x}) .$$

Um im weiteren keine zusätzlichen Voraussetzungen an $U_{,x}$ stellen zu müssen, kann man ausnutzen, daß die folgende Identität gilt :

$$(U_{,x}e , e) = (u_{,x}e , e) - (ee_{,x} , e) = (u_{,x}e , e) .$$

Damit erhält man aus der letzten Gleichung

$$\frac{1}{2}\frac{d}{dt}\|e\|^2 + \delta (e_{,t} , Ue_{,x}) + \delta \|Ue_{,x}\|^2 \leq \frac{1}{2}\|u_{,x}\|_\infty \|e\|^2 + \frac{\delta}{4}\|e_{,t} + Ue_{,x}\|^2 +$$

$$(\frac{1}{\delta} + \frac{1}{2}) \|\rho + \delta U\rho_{,x}\|^2 + (\|u_{,x}\|_\infty + \|u_{,x}\|_\infty^2) \|e\|^2 + \frac{1}{2}\delta^2 \|Ue_{,x}\|^2 .$$

Da die Funktion $u_{,x} \in H^1(\Omega)$ ist, ist sie nach dem Lemma von Sobolew stetig und damit auf ihrem kompakten Träger beschränkt. Also ist

$$(2.5.11) \qquad \frac{1}{2}\frac{d}{dt}\|e\|^2 + \delta (e_{,t} , Ue_{,x}) + \delta \|Ue_{,x}\|^2 \leq \frac{\delta}{4}\|e_{,t} + Ue_{,x}\|^2 +$$

$$c \{\|e\|^2 + \frac{1}{\delta}\|\rho + \delta U\rho_{,x}\|^2 + \delta^2 \|Ue_{,x}\|^2\} .$$

Jetzt setzt man in (2.5.10) $\varphi := e_{,t} - \rho_{,t}$.

$$(e_{,t} + Ue_{,x} , e_{,t} + \delta Ue_{,tx}) = (e_{,t} + Ue_{,x} , \rho_{,t} + \delta U\rho_{,tx}) - (eu_{,x} , e_{,t} + \delta Ue_{,tx}) +$$

$$(eu_{,x} , \rho_{,t} + \delta U\rho_{,tx}) .$$

Diese Gleichung ergibt nach partieller Integration

(2.5.12) $$\|e_{,t}\|^2 + (Ue_{,x}\,,\, e_{,t}) + \frac{1}{2}\frac{d}{dt}\|Ue_{,x}\|^2 \le \frac{1}{2}\delta\,(U_{,x}e_{,t}\,,\, e_{,t}) +$$

$$\frac{1}{2}\delta\,(Ue_{,x}\,,\, U_{,t}e_{,x}) + \frac{1}{4}\|e_{,t} + Ue_{,x}\|^2 + 3\,\|\rho_{,t} + \delta\rho_{,tx}\|^2 + \frac{\delta}{8}\|e_{,t}\|^2 +$$

$$\frac{1}{2\delta}\|u_{,x}\|_\infty\,\|e\|^2 + \delta\,(e_{,x}Uu_{,x}\;e_{,t}) + \delta\,(eu_{,xx}U\,,\, e_{,t}) + \delta\,(eu_{,x}U_{,x}\,,\, e_{,t})\ .$$

An dieser Stelle ist es jetzt notwendig

$$\|U_{,x}\|_{L^\infty((0;T]\,;\,L^\infty(\mathbb{R}))} \le c$$

zu fordern. Als Lösung eines Systems gewöhnlicher Differentialgleichungen zu Lipschitz-stetiger rechter Seite ist mit der Voraussetzung (2.5.4a)

$$\|U_{,t}\|_{L^\infty((0;T]\,;\,L^\infty(\mathbb{R}))} \le c\ .$$

Also erhält man aus (2.5.12)

(2.5.13) $$\|e_{,t}\|^2 + (Ue_{,x}\,,\, e_{,t}) + \frac{1}{2}\frac{d}{dt}\|Ue_{,x}\|^2 \le \frac{1}{2}\delta\,\|e_{,t}\|^2 + c\delta^2\,\|Ue_{,x}\|^2 +$$

$$\frac{1}{4}\|e_{,t} + Ue_{,x}\|^2 + 3\,\|\rho_{,t} + \delta\rho_{,tx}\|^2 + (\frac{\delta}{8} + \frac{3\delta}{2})\,\|e_{,t}\|^2 + (c + c\delta)\,\|e\|^2\ .$$

Die Gleichung (2.5.13) liefert außerdem wieder eine Abschätzung für $\|\,e_{,t}\,\|^2$. Für δ hinreichend klein gilt :

(2.5.14) $$\|e_{,t}\|^2 \le c\{-\delta\frac{d}{dt}\|Ue_{,x}\|^2 + \|Ue_{,x}\|^2 + \|e\|^2 + \|\rho_{,t} + \delta\rho_{,tx}\|^2\} +$$

$$\frac{1}{4}\|e_{,t} + Ue_{,x}\|^2\ .$$

Nachdem man die Gleichung (2.5.13) mit δ multipliziert hat, addiert man sie zu (2.5.11)

$$\frac{1}{2}\frac{d}{dt}\{\|e\|^2 + \delta^2\,\|Ue_{,x}\|^2\} + \delta\,\|e_{,t} + Ue_{,x}\|^2 \le \frac{\delta}{2}\|e_{,t} + Ue_{,x}\|^2 +$$

$$c\,\{\delta^2\,\|Ue_{,x}\|^2 + \frac{1}{\delta}\|\rho + \delta U\rho_{,x}\|^2 + \delta\,\|\rho_{,t} + \delta U\rho_{,tx}\|^2 + \|e\|^2 + \delta^2\,\|e_{,t}\|^2\}\ .$$

Verwendet man auf der rechten Seite (2.5.14) , so ergibt sich

(2.5.15) $$\frac{d}{dt}\{\|e\|^2 + \delta^2 \|Ue_{,x}\|^2\} + \delta \|e_{,t} + Ue_{,x}\|^2 \le c\{\delta^2 \|Ue_{,x}\|^2 +$$

$$\frac{1}{\delta}\|\rho + \delta U\rho_{,x}\|^2 + \delta \|\rho_{,t} + \delta U\rho_{,tx}\|^2 + \|e\|^2\} .$$

Eine Zeitintegration von 0 bis t und die übliche Abschätzung für den Interpolationsfehler liefert

$$\|e(t)\|^2 + \delta^2 \|U(t)e(t)_{,x}\|^2 \le \|e(0)\|^2 + \delta^2 \|U(0)e(0)_{,x}\|^2 +$$

$$c\int_0^t \|e(s)\|^2 + \delta^2 \|U(s)e(s)_{,x}\|^2 ds + ch^{2r+1}\int_0^t \|u\|_{r+1}^2 + \|u_{,s}\|_1^2 ds .$$

Darauf wendet man das Gronwallsche Lemma an.

$$\|e(t)\|^2 + \delta^2 \|U(t)e(t)_{,x}\|^2 \le \exp(ct)\{\|e(0)\|^2 + \delta^2 \|U(0)e(0)_x\|^2 +$$

$$h^{2r+1}\int_0^t \|u\|_{r+1}^2 + \|u_{,s}\|_1^2 ds\} .$$

Damit ist der Satz bewiesen.

■

Als nächstes wird das volldiskrete Problem studiert und zunächst die Stabilität des Verfahrens gezeigt. Hier und im folgenden Kapitel 2.6 werden die Integrale, wie bei reinen Cauchy-Problemen üblich als Integrale über $\mathbb{R}$ geschrieben. Man beachte aber, daß sie sich nur über das Kompaktum Ω oder dessen Rand erstrecken.

Satz 4

Für die Lösung U des volldiskreten Verfahrens (2.5.3) gilt

$$\|U^n\|^2 + \delta \|U^n U^n_{,x}\|^2 \le c\{\|U^0\|^2 + \delta \|U^0 U^0_{,x}\|^2\} \le c\|u^0\|^2 .$$

Beweis:

In der Gleichung (2.5.3) wählt man $\varphi := \bar{U}^n$ und beachtet, daß wegen der homogenen Randbedingung

$$(\bar{U}^n \bar{U}^n_{,x}, \bar{U}^n) = \frac{1}{3}\int_{\mathbb{R}} (\bar{U}^n)^3_{,x}\, dx = 0$$

gilt.

$$(2.5.16) \qquad (d_t\|U^n\|^2 + \delta\,(d_t U^n, \bar{U}^n \bar{U}^n_{,x}) + \delta\,\|\bar{U}^n \bar{U}^n_{,x}\|^2 = 0 \ .$$

Jetzt wählt man in (2.5.3) $\varphi := d_t U^n$.

$$\|d_t U^n\|^2 + (\bar{U}^n \bar{U}^n_{,x}, d_t U^n) + \delta\,(\bar{U}^n \bar{U}^n_{,x}, \bar{U}^n d_t U^n_{,x}) = \frac{1}{2}\,\delta\,(d_t U^n, \bar{U}^n_{,x} d_t U^n) \ .$$

Der letzte Summand auf der linken Seite soll so umgeformt werden, daß sich die Zeitableitung einer Norm ergibt.

$$(\bar{U}^n \bar{U}^n_{,x}, \bar{U}^n d_t U^n_{,x}) = \frac{1}{2k}\{\|\bar{U}^n U^n_x\|^2 - \|\bar{U}^n U^{n-1}_{,x}\|^2\}$$

$$\geq \frac{1}{2k}\{\|U^n U^n_{,x}\|^2 - \|U^{n-1} U^{n-1}_{,x}\|^2 - \|U^n U^{n-1}_{,x}\|^2 - \|U^{n-1} U^n_{,x}\|^2\} \ .$$

Da $\|U\|_{L^\infty((0;T]\,;\,L^\infty(\mathbb{R}))}$ als beschränkt vorausgesetzt war und $h \approx k$ ist erhält man

$$(2.5.17) \qquad \|d_t U^n\|^2 + (\bar{U}^n \bar{U}^n_{,x}, d_t U^n) + \frac{1}{2}\,\delta\, d_t\|U^n U^n_{,x}\|^2 \leq c\,\|\bar{U}^n \bar{U}^n_{,x}\|^2 +$$

$$\frac{1}{2}\,\delta\,(d_t U^n, \bar{U}^n_{,x} d_t U^n) \ .$$

Um im weiteren keine implizite Abhängigkeit des Dämpfungsparameters δ von U^n zu erhalten, wird der zweite Summand auf der rechten Seite von (2.5.17) umgeformt.

$$(d_t U^n, \bar{U}^n_{,x} d_t U^n) = (d_t U^n, \{\frac{k}{2} d_t U^n_{,x} + U^n_{,x}\}\, d_t U^n)$$

$$= \frac{k}{2}(d_t U^n d_t U^n_{,x}, d_t U^n) + (d_t U^n U^n_{,x}, d_t U^n) = (d_t U^n U^n_{,x}, d_t U^n) \ .$$

Aus der Voraussetzung (2.5.6) folgt mit der inversen Beziehung

$$\|U^n_{,x}\|_\infty \le \frac{2}{h} \|U^n\|_\infty ,$$

daß

$$\delta\|U^n_{,x}\|_\infty = \epsilon h \|U^n_{,x}\|_\infty \le \epsilon \|U^n_{,x}\|_\infty .$$

Damit gilt für hinreichend kleines ϵ

$$1 - \epsilon \|U^n\|_\infty \ge c > 0 .$$

Aus (2.5.17) erhält man damit

$$\|d_t U^n\|^2 \le c\{-\delta\, d_t \|U^n U^n_{,x}\|^2 + \|\bar{U}^n \bar{U}^n_{,x}\|^2\} , \tag{2.5.18}$$

wobei ϵ aus $\| U^{n-1} \|_\infty$ bestimmt werden kann. Die Addition von (2.5.16) zu dem δ-fachen von (2.5.17) liefert mit (2.5.18)

$$d_t\{\|U^n\|^2 + \delta^2 \|U^n U^n_{,x}\|^2\} + \delta \|d_t U^n + \bar{U}^n \bar{U}^n_{,x}\|^2 \le c \{\|U^n\|^2 + \delta^2 \|\bar{U}^n \bar{U}^n_{,x}\|^2\} .$$

Diese Gleichung summiert man über n und wendet das diskrete Gronwallsche Lemma an.

$$\|U^n\|^2 + k\delta \sum_{i=1}^{n} \|d_t U^i + \bar{U}^i \bar{U}^i_{,x}\|^2 + \delta^2 \|U^n U^n_{,x}\|^2 \le$$

$$c \{\|U^0\|^2 + \delta^2 \|U^0 U^0_{,x}\|^2\} \le c\|U^0\|^2 .$$

Damit ist die Aussage des Satzes bewiesen.

■

Um die Konvergenzordnung des volldiskreten Verfahrens zu untersuchen leitet man wie im linearen Fall eine Gleichung für den Fehler e her und erhält

$$(d_t e^n + \bar{U}^n \bar{e}^n_{,x} , \varphi + \delta\bar{U}^n \varphi_{,x}) = (-\bar{e}^n \bar{u}^n_{,x} , \varphi + \delta\bar{U}^n \varphi_{,x}) + (r^n , \varphi) , \tag{2.5.19}$$

wobei der Abschneidefehler r^n jetzt die folgende Gestalt hat :

$$(r^n , \varphi) := (d_t u^n - \bar{u}^n_{,x} , \varphi + \delta\bar{U}^n \varphi_{,x}) \tag{2.5.20}$$

Als erstes soll eine Abschätzung für den Abschneidefehler r^n angegeben werden.

Lemma 2

Ist $u \in H^3((0;T] ; L^2(\Omega))$ und erfüllt die Lösung U^n der volldiskreten Gleichung (2.5.4b), so gilt für den Abschneidefehler

$$\|r^n\| \leq c \cdot k \int_{t_{n-1}}^{t_n} \|u_{,sss}\| \, ds \ .$$

Ist u darüberhinaus aus $W^{3,\infty}((0;T] ; L^2(\Omega))$ dann gilt

$$\|r^n\| \leq c \cdot k^2 \, \|u_{,ttt}\|_{L^\infty((0;T] ; L^2(\Omega))} \ . \tag{2.5.21}$$

Beweis:

Aus der Definition (2.5.20) von (r^n,φ) folgt

$$(r^n,\varphi) = \frac{1}{k} \int_{t_{n-1}}^{t_n} (u_{,s},\varphi) + \delta\,(u_{,s},\bar{U}^n \varphi_{,x}) ds - (\bar{u}_t ,\varphi) - \delta\,(\bar{u}_t^n ,\bar{U}^n \varphi_{,x}) \ .$$

Wertet man das Integral mit der Trapezregel aus und verwendet die entsprechende Bemerkung aus Kapitel 2.3. dann gilt :

$$(r^n,\varphi) = k \int_{t_{n-1}}^{t_n} \{ (u_{,sss} , \varphi) + \delta\,(u_{,sss} , U^n \varphi_{,x}) \} \, s(t) \, ds \ .$$

Dabei wurde ausgenutzt, daß U^n eine auf dem Intervall $[t_{n-1} , t_n)$ konstante Funktion ist, und damit exakt integriert wird. Wählt man $\varphi := r^n$, so erhält man daraus

$$\|r^n\|^2 \leq k \int_{t_{n-1}}^{t_n} \{ \|u_{,ttt}\| \cdot \|r^n\| + c\delta \, \|u_{,ttt}\| \cdot \|r^n_{,x}\| \} \, dt \ .$$

Da $r^n \in H^1((0;t] ; V_h(\Omega))$ gilt für r^n eine inverse Beziehung

$$\|r^n\|_1 \leq c h^{-1} \|r^n\| \ ,$$

und mit $\delta = \epsilon h$ ergibt sich

$$\|r^n\| = ck \int_{t_{n-1}}^{t_n} \|u_{,sss}\| \, dt \;\le\; c \cdot k^2 \, \|u_{,ttt}\|_{L^\infty((0;T]\,;\,L^2(\Omega))} \,.$$

■

Als letzter Satz in diesem Abschnitt soll nun gezeigt werden, daß man für das volldiskrete Verfahren die gleiche Konvergenzordnung wie im linearen Fall erhält.

Satz 5

Unter den obigen Voraussetzungen gilt für den Fehler $e := u - U$

$$\|e^n\| + \delta \, \|U^n e^n\| \;\le\; c \, \{\|e^0\| + \delta \, \|U^0 e^0\| + k \, |||u_{,ttt}||| +$$

$$h^{r+\frac{1}{2}} \, (\, |||u|||_{r+1} + |||u_{,t}|||_1 \,)$$

$$\le\; c \, \{\|e^0\| + \delta \, \|U^0 e^0\| + k^2 \, |||u_{,ttt}|||_\infty + h^{r+\frac{1}{2}} \, (\, |||u|||_{r+1} + |||u_{,t}|||_1 \,) \,.$$

Beweis:

In der Fehleridentität (2.5.19) wählt man als Testfunktion $\varphi := \bar{e}^n - \bar{\rho}^n$.

$$(d_t e^n + \bar{U}^n \bar{e}^n_{,x} \,,\, \bar{e}^n + \delta \bar{U}^n \bar{e}^n_{,x}) = (d_t e^n + \bar{U}^n \bar{e}^n_{,x} \,,\, \bar{\rho}^n + \delta \bar{U}^n \bar{\rho}^n_{,x}) -$$

$$(\bar{e}^n \bar{u}^n_{,x} \,,\, \bar{e}^n + \delta \bar{U}^n \bar{e}^n_{,x}) + (\bar{e}^n \bar{u}^n_{,x} \,,\, \bar{\rho}^n + \delta \bar{U}^n \bar{\rho}^n_{,x}) + (r^n \,,\, \bar{e}^n) + (r^n \,,\, \bar{\rho}^n) \,,$$

und erhält daraus

$$(2.5.22)\quad \tfrac{1}{2}\, d_t\|e^n\|^2 + \delta\,(d_t e^n\,,\, \bar{U}^n\, \bar{e}^n_{,x}) + \delta\,\|\bar{U}^n\, \bar{e}^n_{,x}\|^2 \le \tfrac{1}{2}\,(\bar{U}^n_{,x}\, \bar{e}^n\,,\, \bar{e}^n) +$$

$$\tfrac{\delta}{4}\,\|d_t e^n + \bar{U}^n\, \bar{e}^n_{,x}\|^2 + (\tfrac{1}{\delta} + 1)\,\|d_t \rho^n + \bar{U}^n\, \bar{\rho}^n_{,x}\|^2 +$$

$$(\|\bar{u}^n_{,x}\|_\infty + \|\bar{u}^n_{,x}\|^2_\infty + \tfrac{1}{2})\,\|e^n\|^2 + \delta^2\,\|\bar{U}^n\, \bar{e}^n_{,x}\|^2 + \|r^n\|^2 + \tfrac{1}{2}\,\|\bar{\rho}^n\|^2\ .$$

Wie schon im Beweis des letzten Satzes gezeigt wurde gilt

$$(\bar{U}^n_{,x}\, \bar{e}^n\,,\, \bar{e}^n) = (\bar{u}^n_{,x}\, \bar{e}^n\,,\, \bar{e}^n) \le \|\bar{u}^n_{,x}\|_\infty \cdot \|\bar{e}^n\|^2\ .$$

Als nächstes testet man in (2.5.19) mit $\varphi := d_t e^n - d_t \rho^n$.

$$(d_t e^n + \bar{U}^n\, \bar{e}^n_{,x}\,,\, d_t \bar{e}^n + \delta \bar{U}^n\, d_t \bar{e}^n_{,x}) = (d_t e^n + \bar{U}^n\, \bar{e}^n_{,x}\,,\, d_t \bar{\rho}^n + \delta \bar{U}^n\, d_t \bar{\rho}^n_{,x}) -$$

$$(\bar{e}^n \bar{u}^n_{,x}\,,\, d_t \bar{e}^n + \delta \bar{U}^n\, d_t \bar{e}^n_{,x}) + (\bar{e}^n \bar{u}^n_{,x}\,,\, d_t \bar{\rho}^n + \delta \bar{U}^n\, d_t \bar{\rho}^n_{,x}) +$$

$$(r^n\,,\, d_t \bar{e}^n) + (r^n\,,\, d_t \bar{\rho}^n)\ .$$

In dieser Gleichung ist zu beachten, daß

$$(\bar{U}^n\, \bar{e}^n_{,x}\,,\, \delta \bar{U}^n\, d_t e^n_{,x}) = \tfrac{1}{2k}\,\{\|\bar{U}^n\, e^n_{,x}\|^2 - \|\bar{U}^n\, e^{n-1}_{,x}\|^2\}$$

und

$$(\bar{u}^n_{,x}\, \bar{e}^n\,,\, \delta \bar{U}^n\, d_t e^n_{,x}) = -\,\delta\,(\bar{e}^n_{,x}\, \bar{u}^n_{,x}\, \bar{U}^n + \bar{e}^n\, \bar{u}^n_{,xx} \bar{U}^n + \bar{e}^n\, \bar{u}^n_{,x} \bar{U}^n_{,x}\,,\, d_t e^n)$$

ist. Mit (2.5.4) erhält man die Abschätzung

$$(2.5.23)\quad \|d_t e^n\|^2 + (\bar{U}^n\, \bar{e}^n\,,\, d_t e^n) + c\delta\, d_t \|U^n e^n_{,x}\|^2 \le \tfrac{1}{2}\,\delta\,(\bar{U}^n_{,x}\, d_t e^n\,,\, d_t e^n) +$$

$$\tfrac{1}{4}\,\|d_t e^n + \bar{U}^n\, \bar{e}^n_{,x}\|^2 + 3\,\|d_t \rho^n + \bar{U}^n\, \bar{\rho}^n_{,x}\|^2 + \|d_t \rho^n\|^2 + (\tfrac{\delta}{2} + c\delta)\,\|d_t e^n\|^2 +$$

$$(\tfrac{1}{2} + \tfrac{2}{\delta}\,\|\bar{u}^n_{,x}\|^2_\infty + \tfrac{1}{2}\,\|\bar{u}^n_{,x}\|^2_\infty + c + c\delta + c\delta\,\|\bar{u}^n_{,x}\|^2_\infty)\,\|\bar{e}^n\|^2 +$$

$$\delta\,\|\bar{U}^n\, \bar{e}^n\|^2 + (2 + \tfrac{4}{\delta})\,\|r^n\|^2\ .$$

Aus dieser Ungleichung kann man jetzt für δ hinreichend klein eine Abschätzung für $\|\, d_t e^n\,\|^2$ gewinnen.

$$(2.5.24)\quad \|d_t e^n\|^2 \le c\,\{\|\bar{U}^n \bar{e}^n_{,x}\|^2 - \delta\, d_t\|U^n e^n_{,x}\|^2 + \|d_t\rho^n + \delta\bar{U}^n\, d_t\rho^n_{,x}\|^2 +$$

$$\|\bar{e}^n\|^2 + \tfrac{1}{\delta}\|r^n\|^2 + \|d_t\rho^n\|^2\}\ .$$

Nachdem man (2.5.23) mit δ multipliziert hat addiert man sie zu (2.5.22).

$$d_t\{\|e^n\|^2 + \delta^2\,\|U^n e^n_{,x}\|^2\} + \delta\,\|d_t e^n + \bar{U}^n\,\bar{e}^n_{,x}\|^2 \le c\,\{\|\bar{e}^n\|^2 + \delta^2\,\|\bar{U}^n\,\bar{e}^n_{,x}\|^2 +$$

$$\delta^2\,\|d_t e^n\|^2 + \tfrac{1}{\delta}\,\|\rho^n + \bar{U}^n\,\bar{\rho}^n_{,x}\|^2 + \delta\,\|d_t\rho^n + \bar{U}^n\, d_t\rho^n_{,x}\|^2 + \|r^n\|^2 + \delta\|d_t\rho^n\|^2\}\ .$$

Daraus erhält man dann schließlich mit (2.5.24)

$$d_t\{\|e^n\|^2 + \delta^2\,\|U^n e^n_{,x}\|^2\} + \delta\,\|d_t e^n + \bar{U}^n\,\bar{e}^n_{,x}\|^2 \le c\,\{\|\bar{e}^n\|^2 + \delta^2\,\|\bar{U}^n\,\bar{e}^n_{,x}\|^2 +$$

$$\tfrac{1}{\delta}\,\|\rho^n\|^2 + \delta\,\|\,\bar{U}^n\,\bar{\rho}^n_{,x}\|^2 + \delta\,\|d_t\rho^n\|^2 + \delta^2\,\|\bar{U}^n\, d_t\rho^n_{,x}\|^2 + \|r^n\|^2\}\ .$$

Diese Ungleichung wird mit k multipliziert und von i = 1 bis n summiert.

$$\|e^n\|^2 + \delta^2\,\|U^n e^n_{,x}\|^2 \le \|e^0\|^2 + \delta^2\,\|U^0 e^0_{,x}\|^2 + ck\sum_{i=1}^{n}\{\|\bar{e}^i\|^2 + \delta^2\,\|\bar{U}^i\,\bar{e}^i_{,x}\|^2\} +$$

$$k^2\|\|u_{,ttt}\|\|^2 + h^{2r+1}(\,\|\|u\|\|^2_{r+1} + \|\|u_{,t}\|\|^2_1)\ .$$

Daraus erhält man mit dem diskreten Gronwallschen Lemma die Aussage des Satzes.

$$\|e^n\|^2 + \delta^2\,\|U^n e^n_{,x}\|^2 \le e^{ct}\,\{\,\|e^0\|^2 + \delta^2\,\|U^0 e^0_{,x}\|^2 + k^2\,\|\|u_{,ttt}\|\|^2 +$$

$$h^{2r+1}\,(\,\|\|u\|\|^2_{r+1} + \|\|u_{,t}\|\|^2_1)\ .$$

■

In der durchgeführten Analyse wurden die Orts- und die Zeitdiskretisierung getrennt betrachtet. Dadurch war es möglich, die Stellen zu sehen, an denen die verschiedenen Beschränktheitsvoraussetzungen benötigt wurden. Bei dem von Johnson gewählten Zugang wird nur das volldiskrete Problem betrachtet, da er einen Finiten-Elemente-Raum im

Orts-Zeit-Raum benutzt. Da auch er Voraussetzungen der Art (2.5.4) benötigt, ist es bei seinem Zugang schwer zu entscheiden, ob man sich etwa durch eine andere Wahl der Zeitdiskretisierung von diesen Bedingungen befreien könnte.

2.6. Konvergenz gegen schwache - und Entropielösung

In diesem Abschnitt wird das Verfahren wieder auf die eindimensionale Burger-Gleichung, die als einfachstes nichtlineares Modell der Euler-Gleichungen gilt, angewandt und analysiert. Im Gegensatz zum vorausgegangenen Abschnitt soll jetzt auf die, wie in einem einführenden Teil gezeigt wird, im allgemeinen nicht erfüllten Regularitätsannahmen an die Lösung u der kontinuierlichen Gleichung (2.5.1) verzichtet werden. Zuerst werden kurz einige Begriffe aus der Theorie der Erhaltungsgleichungen eingeführt. Dann wird gezeigt, daß falls die Folge der durch das Verfahren gegebenen Approximierenden konvergiert, der Grenzwert eine Entropie-Lösung ist. Das bedeutet, daß mit dem hier entwickelten FEM-Verfahren keine unphysikalischen Lösungen approximiert werden können. Wird die Burger-Gleichung mit dem Standard-Galerkin-Verfahren approximiert, so können sich unphysikalische Lösungen einstellen. Da in diesem Kapitel das Stabilitätsresultat von Satz 4 aus Kapitel 2.5 verwendet wird, muß wieder die Beschränktheit der Lösungen gemäß (2.5.4) gefordert werden. Es wir gezeigt, daß die Beschränktheit der Folge der Approximierenden die Konvergenz einer Teilfolge in der Supremumsnorm impliziert. Dieses Resultat erhält man mit der Methode der kompensierten Kompaktheit. In den meisten Konvergenzbeweisen für Differenzenverfahren wird der Auswahlsatz von Helly verwendet. Er verlangt eine a priori Schranke für u in $W^{1,1}(\Omega\times I)$. Solche Abschätzungen sind insbesondere für hyperbolische Systeme von Erhaltungsgleichungen sehr schwer zu zeigen. Im Gegensatz dazu, ist bei der hier verwendeten Beweistechnik keine Kontrolle der Ableitung von u notwendig.

Zur Definition der im folgenden verwendeten Begriffe betrachtet man das Cauchy-Problem : Finde eine Funktion $u(x,t) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, die die Gleichung

(2.6.1) $$u_{,t} + f(u)_{,x} = 0 \quad \forall\, x \in \mathbb{R}\,,\ t > 0\,,\ t \in I\ ,$$

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

erfüllt. Der Einfachheit halber soll wieder angenommen werden, daß u_0 kompakten Träger in $\mathbb{R}$ hat. Damit hat u für endliche Zeiten kompakten Träger in $\mathbb{R}$. Das Gebiet $\Omega \subset\subset \mathbb{R}$ auf dem die Lösung gesucht ist, hängt von der Wahl des Endzeitpunkts T ab. Deshalb werden im folgenden alle Ortsintegrale über $\mathbb{R}$ integriert. Sie erstrecken sich aber nur über den Träger von u. Damit erhält man insbesondere einen endlichdimensionalen Ansatzraum für das FEM-Verfahren.

Setzt man $a(u) := \frac{d}{du} f(u)$, so können die durch Gleichung (2.6.1) definierten Charakteristiken C geschrieben werden als

(2.6.2) $$C : \frac{dx}{dt} = a(u)\ .$$

Wie im linearen Fall, ist das Verhalten von u entlang der durch t parametrisierten Geraden C durch

(2.6.3) $$\frac{du}{dt} = 0\ ,$$

bestimmt, das heißt u ist konstant entlang jeder Charakteristik. Damit erhält man mit (2.6.2), daß die Charakteristiken Geraden sind. Die Charakteristik, die durch $(s, 0)$ geht erfüllt mit (2.6.2) die Gleichung

(2.6.4) $$x = s + ta(u_0(s))\ .$$

Auf dieser Charakteristik gilt

(2.6.5) $$u = u_0(s)\ .$$

Die Gleichungen (2.6.4) und (2.6.5) ergeben also eine implizite Lösungsdarstellung der Form

(2.6.6) $$u(x,t) = u_0(x - ta(u(x,t)))$$

auf C. Ist u_0 differenzierbar, so existiert für genügend kleine t nach dem Satz über implizite Funktionen eine nach x und t differenzierbare Lösung u von (2.6.6). Seien jetzt

$s_1 < s_2$ zwei Punkte auf der x-Achse. Die Steigung der charakteristischen Geraden ist mit (2.6.4) gegeben durch

$$\frac{dt}{dx} = \frac{1}{a(u_0(s))} \quad .$$

Ist also $a(u_0(s_1)) > a(u_0(s_2))$, so treten Schnittpunkte der Charakteristiken für $t > 0$ auf. Damit ist die Lösung nicht mehr eindeutig bestimmt und es kommt zu Unstetigkeiten. Das Auftreten dieser Unstetigkeiten ist also unabhängig von der Regularität der Anfangsdaten oder der Koeffizienten der Differentialgleichung. Um auch für große Zeiten t Aussagen machen zu können, führt man den Begriff der schwachen Lösung ein.

Definition

Eine Funktion $u \in L_\infty (\mathbb{R} \times I)$ heißt *schwache Lösung* von (2.5.1) , wenn

$$\int_0^T \int_{\mathbb{R}} (u\varphi_{,t} + f(u)\varphi_{,x})\, dxdt + \int_{\mathbb{R}} u_0\, \varphi(x,0)\, dx = 0 \tag{2.6.7}$$

$$\forall\, \varphi \in C_0^\infty (\mathbb{R} \times [0;T]) \quad .$$

■

Die Definitionsgleichung (2.6.7) entsteht aus (2.6.1) durch Multiplikation mit φ und anschließendem partiellen integrieren. Dadurch wird wie üblich die Regularitätsforderung an die Lösung u reduziert. Mit dem Fundamentallemma der Variationsrechnung folgt, daß jede reguläre schwache Lösung auch klassische Lösung ist.

Ein Nachteil des schwachen Lösungsbegriffes ist, daß die Lösung im allgemeinen nicht mehr eindeutig ist. Um eine bestimmte Klasse von Lösungen zu beschreiben wurden verschiedene, zum Teil äquivalente Zusatzbedingungen an die Lösung von (2.6.7) angegeben. Die beiden wichtigsten sind wohl die Entropiebedingung von Hopf und Krushkov und die Bedingung von Hopf, daß sich die Lösung u von (2.6.1) als Grenzwert einer parabolisch regularisierten Gleichung darstellen läßt. Das heißt, daß man für alle $\epsilon > 0$ eine

Funktion $u^\epsilon(x,t) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ sucht, die die Gleichung

$$u^\epsilon_{,t} + f(u^\epsilon)_{,x} - \epsilon u_{,xx} = 0 \qquad \forall\, x \in \mathbb{R},\ t \in I$$

erfüllt, und fordert, daß

$$\lim_{\epsilon \to 0} u^\epsilon = u \qquad \forall\, x \in \mathbb{R},\ t \in I \subset\subset \mathbb{R}^+$$

gilt.

Eine Zusammenstellung der verschiedenen Bedingungen und eine Diskussion ihrer Abhängigkeiten findet man in Bustos /22/. Hier wird, auch in Hinblick auf eine spätere physikalische Interpretation, die Entropiebedingung, wie sie von Kruskov eingeführt wurde gewählt.

Definition

Eine schwache Lösung u von (2.5.1) heißt *Entropie–Lösung*, wenn u die *Entropieungleichung*

$$\int_0^T \int_{\mathbb{R}} (\eta(u)\, \varphi_{,t} + \sigma(u)\, \varphi_{,x})\, dxdt \geq 0 \qquad \forall\, \varphi \in C_0^\infty(\mathbb{R} \times [0;T]) \tag{2.6.8}$$

erfüllt für jede konvexe Funktion $\eta : \mathbb{R} \mapsto \mathbb{R}$; η heißt die *Entropiefunktion* von u . Der zugehörige *Entropiefluß* σ ist definfiert durch

$$\frac{\partial}{\partial u}\, \sigma(U) = \frac{\partial}{\partial u}\, f(u)\, \frac{\partial}{\partial u}\, \eta(u) \ .$$

■

Die Lösungen u von (2.6.1) , die außerdem (2.6.8) erfüllen, heißen auch oft *physikalisch relevante Lösungen.* Für klassische, d.h. glatte Lösungen gilt in (2.6.8) die Gleichheit. Im allgemeinen sind auch Entropielösungen ohne weitere Zusatzbedingungen nicht eindeutig bestimmt. Setzt man z.B. bei der Burger-Gleichung außerdem voraus, daß die Entropielösung von beschränkter Totalvariation ist, so ist diese Lösung eindeutig.

Im weiteren wird gezeigt, daß das Stromliniendiffusionsverfahren eine Lösung u approximiert, die (2.6.7) und (2.6.8) erfüllt.

Die Analyse des Verfahrens wird wieder wie im vorherigen Kapitel an Hand der eindimensionalen Burger-Gleichung durchgeführt. Sie lautete : Finde eine Funktion $u(x,t) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, die die Gleichung

(2.6.9)
$$u_{,t} + uu_{,x} = 0 \qquad \forall x \in \mathbb{R}, \quad t > 0, \quad t \in I ,$$
$$u(x,0) = u_0 \qquad \forall x \in \mathbb{R}$$

erfüllt. Die beweistechnischen Vorteile, die sich aus dieser Wahl ergeben, werden an den entsprechenden Stellen aufgezeigt.

Das volldiskrete Verfahren erhält man wieder durch einen FEM-Ansatz in der Ortsvariablen und einer Zeitdiskretisierung mit dem Crank-Nicolson-Verfahren. Das Verfahren wurde im vorherigen Kapitel eingeführt und lautet :

Finde $U^n \in V_h$ $n \geq 1$, das

(2.6.10)
$$(d_t U^n + \bar{U}^n \bar{U}^n{}_{,x} , \varphi + \delta \bar{U}^n \varphi_{,x}) = 0 \qquad \forall x \in \Omega,$$
$$U^0(x) = \pi u_0 \qquad \forall x \in \Omega, \quad \forall \varphi \in V_h$$

erfüllt. Als generelle Voraussetzung an U^n soll wieder

(2.6.11)
$$\sup_{n \leq N} \{ \|U\|_{L^\infty(\mathbb{R}))} + \sqrt{\delta} \, \|U_{,x}\|_{L^\infty(\mathbb{R}))} \} \leq c$$

erfüllt sein.

Unter dieser Voraussetzung gilt ,wie im letzten Kapitel gezeigt wurde, für die Stabilität des volldiskreten Verfahrens

(2.6.12)
$$\|U^n\|^2 + \delta \, \|U^n U^n{}_{,x}\|^2 + \delta k \sum_{i=1}^{n} \|d_t U^i + U^i U^i{}_{,x}\|^2 \leq c \, \|u^0\|^2 .$$

Für die weitere Analyse ist es besonders wichtig, daß in dieser Aussage auch der Term

$$\delta k \sum_{i=1}^{n} \| d_t U^i + U^i U^i_{,x} \|^2$$

abgeschätzt wird, was in den vorherigen Kapiteln vernachlässigt wurde.

Zunächst werden noch einige Bezeichnungen eingeführt. Für eine Funktion $\phi(x,t)$ bezeichne $\pi\phi$ die L^2-Projektion auf V_h zu einem festen Zeitpunkt t. Um mit einer *Gitterfunktion* $\{G(x,k); G(x;2k);.....;G(x,Nk)\}$ besser arbeiten zu können, definiert man deren stückweise konstante *Fortsetzung* auf $\mathbb{R} \times [0, T + k]$ durch

$$\tilde{G}(x,t) := \begin{cases} G(x,0) & , \quad t = 0 , \quad x \in \mathbb{R} \\ G(x,nk) & , \ (n-1)k < t \leq nk , \quad x \in \mathbb{R} \\ 0 & , \quad T > t \leq T + k, \quad x \in \mathbb{R} \end{cases} .$$

Im weiteren soll immer jede Gitterfunktion G mit ihrer stückweise konstanten Fortsetzung identifiziert werden. Eine *schwache Lösung* $u \in L_\infty(\mathbb{R} \times I)$ *der Burger–Gleichung* muß

$$(2.6.13) \qquad \int_0^T \int_{\mathbb{R}} (u\varphi_{,t} + (\tfrac{1}{2} u^2)\varphi_{,x}) \, dxdt + \int_{\mathbb{R}} u_0 \, \varphi(x,0) \, dx = 0$$

$$\forall \, \varphi \in C_0^\infty(\mathbb{R} \times [0;T])$$

erfüllen.

Als erstes wird gezeigt, daß das Verfahren eine Folge von Approximationen der Lösung der Burger-Gleichung liefert, die falls sie konvergiert, gegen eine schwache Lösung konvergiert.

Satz 1

Konvergiert die Folge der Lösungen $\{U_{h,k}\}$ in $L^\infty((0;T] ; L^\infty(\mathbb{R}))$ gegen eine Lösung u von (2.6.9), dann ist u eine schwache Lösung von (2.6.9).

Bemerkung:

Aus der Konvergenz der Folge $\{U_{h,k}\}$ folgt insbesondere, daß alle $U_{h,k}$ beschränkt sind, d. h.

$$\|U\|_{L^\infty((0;T]\,;\,L^\infty(\mathbb{R}))} \leq c \ .$$

Da aber die Stabilität (2.6.12) des volldiskreten Verfahrens verwendet werden wird, muß außerdem

$$\sqrt{\delta}\,\|U_{,x}\|_{L^\infty((0;T]\,;\,L^\infty(\mathbb{R}))} \leq c$$

gefordert werden.

Beweis von Satz 1 :

Aus der Definition des volldiskreten Verfahrens (2.6.10) erhält man mit $\phi \in C_0^\infty(\overline{\mathbb{R}\times I})$ und der Testfunktion $\varphi := \pi\phi$

$$(2.6.14)\qquad \int_{\mathbb{R}} k \sum_{n=1}^{N} (d_t U^n)\,\phi(x,nk)\,dx + \int_{\mathbb{R}} k \sum_{n=1}^{N} (\bar{U}^n \bar{U}^n_{,x})\,\phi(x,nk) =$$

$$\int_{\mathbb{R}} k \sum_{n=1}^{N} (d_t U^n + \bar{U}^n \bar{U}^n_{,x})\,(\phi(x,nk) - \pi\phi(x,nk))\,dx -$$

$$\delta \int_{\mathbb{R}} k \sum_{n=1}^{N} ((d_t U^n + \bar{U}^n \bar{U}^n_{,x})\,\bar{U}^n\,(\pi\phi(x,nk))_x\,dx \ .$$

Der erste Term auf der linken Seite von (2.6.14) wird mittels partieller Summation umgeformt.

$$\sum_{n=1}^{N} \frac{1}{k}\,(U(x)^n - U(x)^{n-1})\,\phi(x,nk) = \sum_{n=1}^{N-1} -\frac{1}{k}\,(\phi(x,(n+1)k) - \phi(x,nk))\,U(x)^n -$$

$$\frac{1}{k}\,\phi(x,k)U^0(x) + \frac{1}{k}\,\phi(x,Nk)U^N(x) \ .$$

Da nach Voraussetzung ϕ kompakten Träger hat, ist

$$\phi(x,Nk) = 0 \qquad \forall\, x \in \mathbb{R} \ ,$$

und die linke Seite von (2.6.14) hat jetzt die Gestalt

$$(2.6.15) \qquad \int_{\mathbb{R}} k \sum_{n=1}^{N-1} -\frac{1}{k} \left(\phi(x,(n+1)k) - \phi(x,nk)\right) U(x)^n \, dx - \int_{\mathbb{R}} \phi(x,k)\, U^0(x) -$$

$$\int_{\mathbb{R}} k \sum_{n=1}^{N} \frac{1}{2} (U^2)^n \, \phi_{,x}(x,nk) \, dx = \text{rechte Seite von (2.6.14).}$$

Da die Gitterfunktionen als stückweise konstante Funktionen aufgefaßt werden sollen, lassen sich die Summen in der Gleichung (2.6.15) als Integrale schreiben

$$(2.6.16) \qquad \int_{\mathbb{R}} \int_0^T -\frac{1}{k} \left(\phi(x,t+k) - \phi(x,t)\right) U(x)^n \, dtdx - \int_{\mathbb{R}} \phi(x,k)\, u_0 \, dx -$$

$$\int_{\mathbb{R}} \int_0^T \frac{1}{2} U^2 \, \phi_{,x}(x,t) \, dtdx = \text{rechte Seite von (2.6.14).}$$

Die rechte Seite schätzt man nach oben ab durch

$$(2.6.17) \qquad \text{linke Seite von (2.6.14)} \le \sum_{n=1}^{N} k \, \{ \, \| d_t U^n + \bar{U}^n \bar{U}^n_{,x} \| \, ch \, \|\phi\|_1 +$$

$$\delta \, \| d_t U^n + \bar{U}^n \bar{U}^n_{,x} \| \cdot |||U||| \cdot \|\pi\phi_{,x}\| \ .$$

Mit der Stabilitätsaussage (2.6.12) erhält man weiter

$$\le c\sqrt{h} \, \|\phi\|_1 \ .$$

Läßt man jetzt h und k gegen Null gehen, so ergibt sich aus (2.6.16) und (2.6.17)

$$\left| \int_{\mathbb{R}} \int_0^T u\phi_{,t} - \frac{1}{2} u^2 \, \phi_{,x} \, dtdx \right| \le 0$$

und damit

$$\int_{\mathbb{R}} \int_0^T u\phi_{,t} - \frac{1}{2} u^2 \phi_{,x} \, dtdx = 0 \ ,$$

was die Aussage des Satzes ist.

■

Bemerkung:

Für die Aussage des Satzes ist die Beschränktheitsforderung $\|U\|_{L^\infty((0;T] \, ; \, L^\infty(\mathbb{R}))} \leq c$ nicht notwendig. In der Abschätzung (2.6.17) könnte man bei der Anwendung der Hölderschen Ungleichung eine schwächere Norm für U verwenden, z.B.

$$\delta \int_{\mathbb{R}} k \sum_{n=1}^{N} ((d_t U^n + \bar{U}^n \bar{U}^n_{,x}) \bar{U}^n (\pi\phi(x,nk))_{,x} \, dx \leq$$

$$c \sqrt{h} \, \|U\|_{L^2((0;T] \, ; \, L^2(\mathbb{R}))} \, \|\phi\|_{L^\infty((0;T] \, ; \, W^{1,\infty}(\mathbb{R}))} \ ,$$

da die L^2-Projektion in jeder Norm stabil ist. Die Regularitätsforderung an U ist also dadurch bedingt, daß die Stabilitätsaussage (2.6.12) verwendet wird.

Um die Konvergenz gegen eine Entropielösung zu zeigen, benötigt man das folgende einfache technische Hilfsmittel, das in der Literatur oft als Superapproximationseigenschaft bezeichnet wird.

Lemma 1 (siehe z.B. Johnson, Szeppessy /31/)
Ist $u \in H^1(\Omega)$ und ist $\varphi \in V_h$ dann gilt

$$\|u\varphi - \pi(u\varphi)\|_r \leq c h^{1-r} \|\varphi\|_{L^\infty(\Omega)} \|u\|_1 \ , \quad s = 0,1 \ .$$

■

Da die Burger-Gleichung eine strikt konvexe Flußfunktion hat, $f(u) = \frac{1}{2} u^2$, genügt es, die Entropiebedingung für

(2.6.18) a) die konvexe Entropie $\eta(u) = u$ und den

zugehörigen Entropiefluß $\sigma(u) = f(u) = \frac{1}{2} u^2$

und für

b) die strikt konvexe Entropie $\eta(u) = \frac{1}{2} u^2$ und den

zugehörigen Entropiefluß $\sigma(u) = \frac{1}{3} u^3$

zu zeigen. Einen Beweis für diese Vereinfachung der Entropiebedingung findet man z.B. in Tartar /57/. Die Bedingung a) wurde in Satz 1 gezeigt. Im nächsten Satz wird die Bedingung b) nachgewiesen. Damit hat man dann gezeigt, daß das Verfahren eine Entropielösung liefert.

Satz 2

Konvergiert die Folge der Lösungen $\{U_{h,k}\}$ in $L^\infty((0;T] ; L^\infty(\mathbb{R}))$ gegen eine Lösung u von (2.6.9), dann erfüllt u die Entropiebedingung (2.6.18) b) .

Beweis:

Der Beweis verläuft im wesentlichen analog zum Beweis von Satz 1. Mit der Testfunktion $\varphi := \pi(\bar{U}^n \phi)$, wobei wieder $\phi \in C_0^\infty (\overline{\mathbb{R} \times I})$ ist, erhält man aus der Definition des volldiskreten Verfahrens (2.6.10)

$$(2.6.19) \quad \int_{\mathbb{R}} k \sum_{n=1}^{N} (d_t U^n)\ \bar{U}^n \phi(x,nk)\, dx + \int_{\mathbb{R}} k \sum_{n=1}^{N} (\bar{U}^n \bar{U}^n_{,x})\ \bar{U}^n \phi(x,nk) =$$

$$\int_{\mathbb{R}} k \sum_{n=1}^{N} (d_t U^n + \bar{U}^n \bar{U}^n_{,x}) (\bar{U}^n \phi(x,nk) - \pi(\bar{U}^n \phi(x,nk)))\, dx +$$

$$\delta \int_{\mathbb{R}} k \sum_{n=1}^{N} ((d_t U^n + \bar{U}^n \bar{U}^n_{,x}) \bar{U}^n ((\bar{U}^n \phi(x,nk))_{,x} - \pi(\bar{U}^n \phi(x,nk))_{,x}) \, dx -$$

$$- \delta \int_{\mathbb{R}} k \sum_{n=1}^{N} ((d_t U^n + \bar{U}^n \bar{U}^n_{,x}) (\bar{U}^n)^2 \phi(x,nk)_{,x} \, dx \quad .$$

Der erste Summand auf der linken Seite von (2.6.19) wird mittels partieller Summation umgeformt.

$$\sum_{n=1}^{N} \frac{1}{k} (U(x)^n - U(x)^{n-1}) \frac{1}{2} (U(x)^n + U(x)^{n-1}) \phi(x,nk) =$$

$$\frac{1}{2} \sum_{n=1}^{N} \frac{1}{k} ((U(x)^2)^n - (U(x)^2)^{n-1}) \phi(x,nk) =$$

$$\frac{1}{2} \sum_{n=1}^{N-1} -\frac{1}{k} (\phi(x,(n+1)k) - \phi(x,nk)) (U(x)^2)^n - \frac{1}{k} \phi(x,k) u_0^2(x) + \frac{1}{k} \phi(x,Nk) U^N(x)^2 \quad .$$

Da nach Voraussetzung ϕ kompakten Träger hat ist

$$\phi(x,Nk) = 0 \qquad \forall x \in \mathbb{R}$$

und die linke Seite von (2.6.19) hat jetzt die Gestalt

$$\int_{\mathbb{R}} k \sum_{n=1}^{N-1} -\frac{1}{k} (\phi(x,(n+1)k) - \phi(x,nk)) (U(x)^2)^n \, dx -$$

$$\int_{\mathbb{R}} \phi(x,k) u_0^2(x) - \int_{\mathbb{R}} k \sum_{n=1}^{N} \frac{1}{3} (\bar{U}^n)^3 \phi_{,x}(x,nk) \, dx = \text{rechte Seite von (2.6.19).}$$

Da die Gitterfunktionen wieder als stückweise konstante Funktionen aufgefaßt werden sollen, lassen sich die Summen in der letzten Gleichung als Integrale schreiben.

$$\int_{\mathbb{R}} \int_0^T -\frac{1}{k} (\phi(x,t+k) - \phi(x,t))\, U(x)^2 \, dtdx - \int_{\mathbb{R}} \phi(x,k)\, u_0 \, dx -$$

$$\int_{\mathbb{R}} \int_0^T \frac{1}{3} U^3 \, \phi_{,x}(x,t) \, dtdx = \text{rechte Seite von (2.6.19).}$$

Die rechte Seite schätzt man nach oben ab durch

$$\text{rechte Seite von (2.6.19)} \leq \sum_{n=1}^{N} k \{ \| d_t U^n + \bar{U}^n \bar{U}^n_{,x} \| (\| \bar{U}^n \phi - \pi(\bar{U}^n \phi) \| +$$

$$\delta \| \| U \| \| \cdot \| (\bar{U}^n \phi)_{,x} - (\pi(\bar{U}^n \phi))_{,x} \| + \delta \| \| (\bar{U}^n)^2 \| \| \cdot \| \phi_{,x} \|) \} \ .$$

Mit der Stabilitätsaussage (2.6.12), dem Lemma 1 und der Voraussetzung (2.6.11) erhält man weiter

$$\leq c \sqrt{h} \, \| \phi \|_1 \ .$$

Läßt man jetzt h und k gegen Null gehen so ergibt sich insgesamt

$$-\int_{\mathbb{R}} \int_0^T \frac{1}{2} u^2 \, \phi_{,t} - \frac{1}{3} u^3 \, \phi_{,x} \, dtdx \leq 0$$

und damit die Aussage des Satzes.

■

Es bleibt noch zu zeigen, daß das Verfahren eine konvergente Folge von Näherungen liefert. Dazu wird ein Satz von Murat und Tartar verwendet und mit der Methode der kompensierten Kompaktheit gezeigt, daß die Folge $\{U_{h,k}\}$ eine konvergente Teilfolge enthält, die mit Satz 2 dann Entropielösung von (2.6.9) ist. Da hier die Ergebnisse aus den Arbeiten von Tartar und DiPerna verwendet werden, soll zwar auf eine Herleitung, die außerdem sehr langwierig ist, verzichtet werden, aber stattdessen ein motivierendes Beispiel gegeben werden.

Gegeben sei eine Folge von Funktionen u_n die schwach gegen eine Grenzfunktion u konvergiert. Ist f eine affin lineare Funktion, dann konvergiert $f(u_n)$ schwach gegen f(u) .

Im allgemeinen ist das für nichtlineare Funktionen f falsch, wie das folgende Beispiel zeigt. Ist $u_n(x) := \sin(nx)$ und $f(u) := u^2$ dann ist

$$\int_0^{2\pi} \sin(nx)\phi(x)dx = 0 \qquad \forall\, n \in \mathbb{N}$$

aber es ist für $\phi \equiv 1$

$$\int_0^{2\pi} \sin^2(nx)dx = \pi \qquad \forall\, n \in \mathbb{N} \ .$$

Eine Möglichkeit auch im nichtlinearen Fall eine Konvergenzaussage zu machen gibt der folgende Satz von Murat und Tartar

Satz 3 (z.B. in Dacarogna /60/)

Sei $\Omega \subset \mathbb{R}^2$ eine beschränkte offene Menge und $f : \mathbb{R} \to \mathbb{R}$ einmal stetig differenzierbar. Die Funktionenfolge $\{u^\epsilon\}$ konvergiere gegen die Funktion u in $L^\infty(\Omega)$ in der Schwach-*-Topologie und es gelte, daß für jede Entropie und den zugehörigen Entropiefluß

$$\frac{\partial}{\partial t}\eta(u) + \frac{\partial}{\partial x}\sigma(u) \in A + B \tag{2.6.20}$$

ist, wobei

A := kompakte Teilmenge von $H^{-1}(\Omega)$, und

B := beschränkte Teilmenge der beschränkten Maße auf Ω

ist.

Dann konvergiert $f(u^\epsilon)$ gegen f(u) in der Schwach-*-Topologie von $L^\infty(\Omega)$.

■

Damit kann man jetzt die Konvergenz der Folge $\{U_{h,k}\}$ gegen eine Lösung von (2.6.1) zeigen.

Satz 4

Erfüllt die Folge $\{U_{h,k}\}$ die Voraussetzung (2.6.10), dann besitzt $\{U_{h,k}\}$ eine Teilfolge, die gegen eine Entropielösung von (2.6.1) konvergiert.

Beweis:

Aus der Voraussetzung (2.6.10) folgt, daß eine Teilfolge der Folge $\{U_{h,k}\}$ gegen eine Funktion U in der Schwach-*-Topologie von $L^\infty(\Omega)$ konvergiert. Die Teilfolge soll mit der Folge $\{U_{h,k}\}$ identifiziert werden. Wie schon vorher erwähnt, genügt es im Fall einer strikt konvexen Flußfunktion $f(u) = \frac{1}{2}u^2$ die Bedingung (2.6.19) nur für die Entropien

$$\eta(u) = u, \qquad \sigma(u) = \tfrac{1}{2}u^2 \tag{2.6.21}$$

und

$$\eta(u) = \tfrac{1}{2}u^2, \qquad \sigma(u) = \tfrac{1}{3}u^3 \tag{2.6.22}$$

zu verifizieren. Zunächst wird gezeigt, daß (2.6.20) für die Wahl (2.6.21) erfüllt ist. Dazu sei $\phi \in C_0^\infty(\mathbb{R} \times I)$. Als Testfunktion wählt man $\varphi := \pi\phi$ und erhält damit wie in Satz 1

$$\int_{\mathbb{R}} \int_0^T -\tfrac{1}{k}\,(\phi(x,t+k) - \phi(x,t))\, U(x)^n \,dtdx - \int_{\mathbb{R}} \int_0^T \tfrac{1}{2}\,U^2\,\phi_x(x,t)\,dtdx =$$

$$= \text{rechte Seite von (2.6.14).}$$

Da ϕ seinen Träger in $\mathbb{R} \times I$ hat, treten beim partiellen summieren bzw. integrieren keine Randterme auf. Die rechte Seite schätzt man nach oben ab durch

$$\text{linke Seite von (2.6.14)} \le \sum_{n=1}^{N} k\,\{\, \|d_t U^n + \bar{U}^n \bar{U}^n_{,x}\|\, \mathrm{ch}\, \|\phi\|_1 +$$

$$\delta\, \|d_t U^n + \bar{U}^n \bar{U}^n_{,x}\| \cdot |||U||| \cdot \|\pi\phi_{,x}\| \,.$$

Mit der Stabilitätsaussage (2.6.12) erhält man weiter

$$\le c\sqrt{h}\, \|\phi\|_1 \,.$$

Läßt man jetzt h und k gegen Null gehen so ergibt sich aus (2.6.16) und (2.6.17)

$$\left| \int_{\mathbb{R}} \int_0^T u\, \phi_{,t} - \tfrac{1}{2} u^2\, \phi_{,x}\, dtdx \right| \leq c \sqrt{h}\, \|\phi\|_1$$

und damit

$$\frac{\partial}{\partial t}\, \eta(u) + \frac{\partial}{\partial x}\, \sigma(u) \in A \quad ,$$

womit (2.6.20) im Fall (2.6.21) gezeigt ist.

Um den Fall (2.6.22) zu behandeln, wählt man wie in Satz 2 $\phi \in C_0^\infty(\mathbb{R} \times I)$ und als Testfunktion wieder $\varphi := \pi(\bar{U}^n)\phi$ und erhält damit

$$\int_{\mathbb{R}} \int_0^T -\frac{1}{k} (\phi(x,t+k) - \phi(x,t))\, U(x)^2\, dtdx - \int_{\mathbb{R}} \int_0^T \frac{1}{3} U^3\, \phi_{,x}(x,t)\, dtdx =$$

$$= \text{rechte Seite von (2.6.19).}$$

Die rechte Seite schätzt man nach oben ab durch

$$\text{linke Seite von (2.6.19)} \leq \sum_{n=1}^{N} k \{ \|d_t U^n + \bar{U}^n \bar{U}^n_{,x}\| (\|\bar{U}^n \phi - \pi(\bar{U}^n \phi)\| +$$

$$\delta \|\|U\|\| \cdot \|(\bar{U}^n \phi)_{,x} - (\pi(\bar{U}^n \phi))_{,x}\| + \delta \|\|(\bar{U}^n)^2\|\| \cdot \|\phi_{,x}\|) \} \quad .$$

Mit der Stabilitätsaussage (2.6.12), dem Lemma 1 und der Voraussetzung (2.6.10) erhält man weiter

$$\leq c \sqrt{h}\, \|\phi\|_1 \quad .$$

Läßt man jetzt h und k gegen Null gehen so ergibt sich daraus

$$\left| \int_{\mathbb{R}} \int_0^T \frac{1}{2} u^2\, \phi_{,t} + \frac{1}{3} u^3\, \phi_{,x}\, dtdx \right| \leq c \|\phi\|_{L^\infty((0;\, T]\, ;\, L^\infty(\mathbb{R}))} + c\sqrt{h}\, \|\phi\|_1 \quad ,$$

womit (2.6.20) im Fall (2.6.22) und damit die Aussage des Satzes gezeigt ist.

■

Die wesentlichsten Voraussetzungen zur Herleitung der obigen Sätze waren, daß die Flußfunktion f in (2.6.1) strikt konvex ist, und daß die Erhaltungsgleichung im $\mathbb{R}_+ \times \mathbb{R}$ diskutiert wurde. Die Ausdehnung der Resultate auf den $\mathbb{R}_+ \times \mathbb{R}^n$ erfordert einen anderen Zugang und ist keine analoge Übertragung der Begriffe. Im Falle einer allgemeinen Flußfunktion f müßte (2.6.19) für alle konvexen Entropien η gezeigt werden, was ein schwierigeres Problem wäre.

Anhang : Das Lemma von Gronwall

In diesem Anhang soll das in den vorausgegangenen Kapiteln häufig verwendete Lemma von Gronwall in einer etwas allgemeineren Form bewiesen werden. Diese Verallgemeinerung hat in Kapitel 2.4 die verbesserten Fehlerabschätzugen erlaubt.

Verallgemeinertes Lemma von Gronwall

Sei $I = [t_0, \infty) \subset \mathbb{R}$ und $w(t) \in C(I)$, $a(T), c(t) \in L^1(I)$, $w(t), a(t), c(t),$ und $b(t) > 0$ $\forall\, t \in I$. Für die stetige isotone Funktion f gelte $f(u) > 0 \quad \forall\, u > 0$. Aus der Ungleichung

$$(1) \qquad w(t) + \int_{t_0}^{t} c(s)ds \leq \int_{t_0}^{t} a(s)\, f(w(s))\, ds + b(t) , \quad \forall\, t \in I$$

folgt dann

$$(2) \qquad w(t) + \int_{t_0}^{t} c(s)ds \leq F^{-1}\left(\int_{t_0}^{t} a(s)ds + F(b(t))\right) \quad \forall\, t \in I' ,$$

wobei

$$(3) \qquad F(u) := \int_{u_0}^{u} \frac{1}{f(x)}\, dx \quad , \quad u_0 > 0, \quad u \geq 0$$

und $F^{-1}(F(u)) = u \quad \forall\, u$ sein soll.

Bemerkung:

a) Die Umkehrfunktion F^{-1} existiert und ist isoton, da F strikt isoton ist.

b) In (2) ist das Teilintervall $I' \subset I$ so zu wählen, daß $\int_{t_0}^{t} a(s)ds + F(b(t))$ im Definitionsbereich von F^{-1} liegt.

c) Die Behauptung des Satzes gilt unabhängig von der Wahl von u_0 , denn sei

$$\tilde{F}(u) := \int_{\tilde{u}_0}^{u} \frac{1}{f(x)}\, dx = F(u) - \delta$$

mit $$\delta := \int_{u_0}^{\tilde{u}_0} \frac{1}{f(x)}\, dx \quad .$$

Dann ist $\tilde{F}^{-1}(u) = F^{-1}(u + \delta)$ und somit

$$\tilde{F}^{-1}(A + \tilde{F}(b)) = \tilde{F}^{-1}(A + F(b) - \delta) = F^{-1}(A + F(b)) \, .$$

Beweis des Lemmas:

Es sei $v(t)$ definiert durch

$$v(t) := \int_{t_0}^{t} a(s) f(w(s)) ds \quad .$$

Mit dieser Setzung folgt aus (1) für $w(t) \leq v(t)$

(4) $$f(w) \leq f(v), \quad \text{da } f \text{ isoton ist.}$$

Wäre $f(v) = 0$, so wäre auch $f(w) = 0$ für alle $w \leq v$ und die Behauptung des Lemmas folgt für $F = id$. Deshalb sei im folgenden o.B.d.A. $f(v) > 0$ vorausgesetzt. Dann gilt mit (4)

$$\frac{1}{f(v)} \cdot f(w) \cdot a(s) \leq a(s)$$

und mit der Definition von v

$$\frac{1}{f(v)} \cdot \frac{d}{dt} v(t) \leq a(s) \quad .$$

Daraus erhält man

$$\frac{d}{dt} F(v) \leq a(s) \quad .$$

Die letzte Gleichung integriert man bezüglich t .

$$F(v(t)) \leq \int_{t_0}^{t} a(s)ds + F(b(t)) \quad .$$

Darauf wendet man die isotone Funktion F^{-1} an.

$$v(t) \leq F^{-1}(\int_{t_0}^{t} a(s)ds + F(b(t)) \quad .$$

Daraus ergibt sich mit (4) die Behauptung des Lemmas.

■

Im allgemeinen wird im Lemma von Gronwall $f(u) \equiv u$ gesetzt. Mit dieser Setzung wird $F(u) = \ln(u)$ und $f^{-1}(u) = \exp(u)$. Die Behauptung des Lemmas lautet in diesem Spezialfall

$$w(t) + \int_{t_0}^{t} c(s)ds \leq \exp\left(\int_{t_0}^{t} a(s)ds\right) \cdot b(t) \quad .$$

Diese spezielle Form wird in der Literatur und auch in dieser Arbeit als "Lemma von Gronwall" bezeichnet.

Wählt man $f(u) = \sqrt{u}$, so ist $F(u) = 2\sqrt{u}$ und $F^{-1}(u) = \frac{1}{4}u^2$. Die entsprechende Form der Behauptung (2) lautet jetzt

$$w(t) + \int_{t_0}^{t} c(s)ds \leq \frac{1}{4}\left(\int_{t_0}^{t} a(s)ds + 2\sqrt{b(t)}\right)^2 \quad .$$

Diese Form wird hier als "verallgemeinertes Lemma von Gronwall" bezeichnet. Sie garantierte im Kapitel 2.4, daß der Fehler höchtens linear in der Zeit wächst.

Um die Konvergenz des volldiskreten Verfahrens zu zeigen, benötigte man eine diskrete Version des Gronwallschen Lemmas, das als nächtes bewiesen werden soll. Da für die Verbesserung der Fehlerabschätzung nur der Fall $f(u) = \sqrt{u}$ benötigt wird, soll im folgenden vorausgesetzt werden, daß

$$0 \leq f(u) \leq c_f \qquad \forall\, 0 \leq u \leq \tilde{u}$$

erfüllt ist.

Diskretes verallgemeinertes Lemma von Gronwall :

Sei $0 < t_0 < t_1 < \ldots < t_{n+1}$ und $\{a_i\}$, $\{b_i\}$, $\{c_i\}$ und $\{w_i\}$ Folgen in $\mathbb{R}_+$ mit $a_n < 1/c$. Die Funktion f sei stetig und isoton. Dann folgt aus der Ungleichung

$$w_n + \sum_{i=0}^{n} c_i \leq \sum_{i=0}^{n} a_i f(w_i) + b_n , \qquad \forall\, n \geq 0 \ , \tag{5}$$

daß

$$w_n + \sum_{i=0}^{n} \tilde{c}_i \leq F^{-1}\Big(\sum_{i=0}^{n-1} \tilde{a}_i f(w_i) + F(\tilde{b}_n)\Big) , \qquad \forall\, n \geq 0 \tag{6}$$

erfüllt ist, wobei F und F^{-1} wie im Lemma von Gronwall definiert sind. Außerdem ist

$$\tilde{a}_i := (1 - a_n c_f)^{-1} a_i \quad .$$

Entsprechend sind $\tilde{b}_i$ und $\tilde{c}_i$ definiert.

Beweis

Aus der Voraussetzung (5) folgt

$$(1 - a_n f)(w_n) + \sum_{i=0}^{n} c_i \leq \sum_{i=0}^{n-1} a_i f(w_i) + b_n$$

und damit

$$(1 - a_n c_f)\, w_n + \sum_{i=0}^{n} c_i \leq \sum_{i=0}^{n-1} a_i f(w_i) + b_n \quad .$$

Also hat man insgesamt

$$w_n + \sum_{i=0}^{n} \tilde{c}_i \leq \sum_{i=0}^{n-1} \tilde{a}_i f(w_i) + \tilde{b}_n \quad .$$

Jetzt definiert man die folgenden Treppenfunktionen als Fortsetzungen der Gitterfunktionen

$$z(t) := z_i \qquad \forall\, t \in [t_i, t_{i+1}) \quad ,$$

wobei $z \in \{w, c, a, b\}$ ist. Für diese Funktionen gilt nach obiger Überlegung

$$w(t_n) + \int_{t_0}^{t_n} \frac{1}{k_i} c(s)ds \leq \int_{t_0}^{t_n} \frac{1}{k_i} a(s)\, f(w(s))\, ds + b(t_n) \quad ,$$

mit $k_i := t_{i+1} - t_i$. Daraus folgt mit dem (kontinuierlichen) verallgemeinerten Lemma von Gronwall

$$w_n + \sum_{i=0}^{n} \tilde{c}_i \leq F^{-1}\Big(\sum_{i=0}^{n-1} \tilde{a}_i\, f(w_i) + F(\tilde{b}_n) \Big) \quad .$$

■

3. Stromliniendiffusion für Systeme von Erhaltungsgleichungen

In diesem Kapitels wird unter zusätzlichen Forderungen an das Verfahren, der in Kapitel 2 für eine skalare Gleichung entwickelte FEM-Algorithmus auf ein hyperbolisches System von Erhaltungsgleichungen in mehreren Raumdimensionen verallgemeinert. Das Hauptproblem liegt in der Definition einer "Stromlinienrichtung" für ein System von Gleichungen. Es zeigt sich, daß das Verfahren nur sinnvoll für *symmetrische* hyperbolische Systeme definiert werden kann. Das hier aufgezeigte Vorgehen ist nicht auf zwei Raumdimensionen beschränkt. Es wird aber im Hinblick auf seine hier vorgesehene numerische Anwendung für diesen Fall formuliert. Dabei soll die folgende naheliegende Forderung erfüllt sein :

(3.1) Für den Fall, daß das System entkoppelt ist, soll für jede skalare zweidimensionale Gleichung das Verfahren aus Kapitel 2 reproduziert werden.

Diese Forderung wird z.B. vom Lax-Wendroff- oder seinem FEM-Analogon dem Taylor-Galerkin-Verfahren nicht erfüllt. Dieses Verfahren lautet für ein hyperbolisches System in einer Raumdimesion

$$u_{,t} + Au_{,x} = 0$$

mit den üblichen Bezeichnungen

$$(u^{n+1} - u^n, \varphi) = -\Delta t\,(Au^n_{,x}, \varphi) + \tfrac{1}{2}\Delta t^2\,(A^2 u_{,x}, \varphi_{,x}) \quad \forall\, \varphi \in V_h \ .$$

Dabei werden alle Komponenten von u mit dem selben Faktor $1/2\,\Delta t$ gedämpft. Dadurch wird das unterschiedliche Verhalten der Komponenten von u nicht berücksichtigt. Außerdem hat sich bei der Analyse in Kapitel 2 gezeigt, das die Dämpfung auch von der lokalen Schrittweite h abhängen sollte. Daß diese Dämpfungsstrategie nicht ausreichend ist, zeigt sich auch bei Rechnungen, die in Lötzerich /71/ durchgeführt werden. Dort muß zur Dämpfung von Oszillationen der Lösung noch zusätzlich ein sogenannter Lapidus-Filter verwendet werden.

In diesem Abschnitt wird wieder das System der Euler-Gleichungen aus Kapitel 1.3 be-

trachtet. Es lautet in quasilinearer Form

$$u_{,t} + \sum_{i=1}^{2} A^i u_{,x_i} = 0 \tag{3.2}$$

mit dem dort angegebenen Vektor u und den Matrizen A^i . Formuliert man diese Gleichung variationell,

$$\int_\Omega \varphi^T (u_{,t} + \sum_{i=1}^{2} A^i u_{,x_i}) \, dx = 0 \ , \ \forall \varphi \in H^1(\Omega) \ , \tag{3.3}$$

so ergibt sich bei der Setzung $\varphi := u$ für den ersten Term unter Verwendung der Definition der Matrix A^1

$$\tfrac{1}{2} \tfrac{d}{dt} \int_\Omega \rho^2 (1 + u^2 + e^2) \, dx \ , \tag{3.4}$$

was physikalisch unsinnige Größen sind. Dies ist mit ein Grund, von der symmetrischen Formulierung von (3.2) auszugehen, die in Kapitel 1.3 angegeben ist. Sie lautet

$$B^0 v_{,t} + \sum_{i=1}^{2} B^i v_{,x_i} = 0 \tag{3.5}$$

mit dem in Kapitel 1.3 angegebenen Vektor v und den Matrzen B^i . Bei der variationellen Formulierung von (3.5) ergibt sich, wie in Kapitel 1.4 gezeigt wurde der zweite Hauptsatz der Thermodynamik, also eine physikalisch sinnvolle Aussage. Das bietet außerdem die Möglichkeit, ein Stabilitätskonzept, das der Entropieerhaltung, für das System (3.5) zu formulieren. Ein weiterer Grund von der Darstellung (3.5) auszugehen ist, daß in diesem Fall die Forderung (3.1) leichter zu erfüllen ist, siehe Hughes /39/.

Im folgenden wird das Cauchy-Problem mit der Anfangsbedingung

$$u(x,0) = u_0(x) \tag{3.6}$$

mit $u_0(x) \in [L_2(\mathbb{R}^2)]^4$ und kompakten Träger in $\Omega \subset \mathbb{R}^2$ betrachtet. Die Stromliniendiffusionsmethode für das System (3.5) lautet : Finde ein $v \in [H^1((0;T] , H^1(\Omega)]^4$, das

$$(3.7) \qquad \int_\Omega (B^0(v)v_{,t} + \sum_{i=1}^{2} B^i(v)v_{,x_i}) \cdot (\varphi + \delta \sum_{i=1}^{2} B^i(v)\varphi_{,x_i}) \, dx = 0$$

$$\forall \varphi \in [H^1(\Omega)]^4$$

erfüllt. Durch eine langwierige Eigenvektor- und Eigenwertberechnung für ein System der Gestalt (3.5) hat Hughes /39/ gezeigt, wie man im Fall konstanter Matrizen B^i durch eine geeignete Wahl des Dämpfungsparameters δ erreichen kann, daß die Forderung (3.1) erfüllt ist. Die Ergebnisse dieser Arbeit von Hughes werden in kurzer Form dargestellt.
Wie schon oben gezeigt wurde, sollte der Dämpfungsparameter $\delta = \delta(x,t)$ für Systeme eine Matrix sein. Die einfachste Wahl von δ als $\delta := chI$ erfüllt (3.1) nicht. Deshalb betrachtet man zunächst ein symmetrisches hyperbolisches System in einer Raumdimension

$$(3.8) \qquad B^0 v_{,t} + B v_{,x} = 0 \ .$$

Man definiert eine Transformationsmatrix T durch

$$(3.9) \qquad S := (B^0)^{-\frac{1}{2}} E \ ,$$

wobei E die orthogonale Matrix ist, die aus Eigenvektoren von

$$\tilde{B} := (B^0)^{-\frac{1}{2}} B \, (B^0)^{-\frac{1}{2}}$$

besteht. Die zugehörigen Eigenwerte seien λ_i . Dann gilt

$$(3.10) \qquad S^T B^0 S = E^T (B^0)^{-\frac{1}{2}} B^0 (B^0)^{-\frac{1}{2}} E = I$$

und

$$S^T B S = E^T (B^0)^{-\frac{1}{2}} B \, (B^0)^{-\frac{1}{2}} E = \Lambda = \mathrm{diag}(\lambda_i) \ .$$

Führt man die neue Variable $w := Sv^{-1}$ ein, dann ist mit (3.10)

$$(3.11) \qquad (B^0 v_{,t} + Bv_{,x}) \cdot (\varphi + \delta B \varphi_{,x}) = (B^0 Sw_{,t} + B Sw_{,x}) \cdot (S\phi + \delta B S \phi_{,x})$$

$$= (w_{,t} + \Lambda w_{,x}) \cdot (\phi + S^{-1} \delta S^{-T} \Lambda \phi_{,x}) \ .$$

Das System $w_{,t} + \Lambda w_{,x} = 0$ ist entkoppelt. Um die Forderung (3.1) zu realisieren muß

$$S^{-1} \delta S^{-T} = ch \, \Lambda^{-1}$$

sein. Daraus folgt

$$(3.12) \qquad \delta = chS^{-1} S^T = ch \, (B^0)^{-\frac{1}{2}} \tilde{B}^{-1} (B^0)^{-\frac{1}{2}} = ch \, B^{-1} \ .$$

Bei einem System in mehreren Raumdimensionen wie es (3.5) ist, ist es im allgemeinen nicht möglich, alle Matrizen B^i mit derselben Transformationsmatrix zu diagonalisieren. Als Verallgemeinerung von (3.12) verwendet Hughes die Formel

$$(3.13) \qquad \delta = ch(B^0)^{-\frac{1}{2}} \Big(\sum_{i=1}^{2} (B^i(v))^2 \Big)^{-\frac{1}{2}} (B^0)^{-\frac{1}{2}} \ .$$

Mit dieser Wahl ist es möglich, für den Fall eines stationären Systems mit konstanten Koeffizientenmatrizen B^i Stabilität der Stromliniendiffusionsmethode zu zeigen, siehe Hughes /40/.

Im Rahmen der Herleitung der Stromliniendiffusionsmethode in Kapitel 2.2 war der Dämpfungsparameter δ frei wählbar. Man kann für eine skalare lineare eindimensionale Gleichung zeigen, daß es einen optimalen Wert für den Dämpfungsparameter gibt, in dem Sinne, daß sich an den Knoten des Netzes die exakte Lösung einstellt. Dies ist auch für die hier gegebene Herleitung möglich, wenn man sich auf diesen Spezialfall beschränkt. Es scheint aber unmöglich, solch ein optimales δ für Systeme der Form (3.5) zu finden.

Zur Zeitdiskretisierung wird hier, wie in Kapitel 2 beschrieben, das Crank-Nicolson-Verfahren verwendet. Mit den dort eingeführten Bezeichnungen lautet das *volldiskrete Verfahren* :

Finde ein $v \in [V_h]^4$, das für jeden Zeitpunkt t_n

$$(3.13)\quad \int_\Omega (B^0(\bar{v}^n)\, d_t v^n + \sum_{i=1}^{2} B^i(\bar{v}^n)\bar{v}^n_{,x_i}) \cdot (\varphi + \delta \sum_{i=1}^{2} B^i(\bar{v}^n)\varphi_{,x_i})\, dx = 0$$

$$\forall\, \varphi \in [V_h]^4$$

erfüllt.

Die Auswertung des Dämpfungsparameters ist kompliziert, da die Wurzel aus einer Matrix gezogen werden muß, die von der aktuellen Lösung abhängt. Die hier verwendete Dämpfungsmethode ist eine *globale* Dämpfung, und damit sehr aufwendig. Die meisten Differenzenverfahren verwenden *lokale* Methoden. Dabei wird eine Größe, in der Regel der Druck als Indikator für eine Unstetigkeit im Strömungsfeld verwendet. Diese Strategie ist zwar sparsamer, ermöglicht aber, da sie auch oft empirische Parameter enthält, kein systematisches Vorgehen bei allgemeinen hyperbolischen Systemen, wie es die hier verwendete Methode erlaubt.

4. Implementierung der Stromliniendiffusionsmethode

Wird das hier gestellte strömungsmechanische Problem mit dem in Kapitel 2 und 3 beschriebenen Galerkin-Verfahren diskretisiert, so sind im wesentlichen vier Schritte durchzuführen:

1.) Zerlegung des Grundgebietes Ω in Vierecke,

2.) Aufbau der Systemmatrizen,

3.) Durchführung eines Zeitschrittes und

4.) Lösen der entstehenden linearen Gleichungssysteme.

Die standardmäßig bei einem FEM-Verfahren auftretenden programmtechnischen Teilprobleme wurden mit Hilfe des am Lehrstuhl für Angewandte Mathematik der Universität Saarbrücken entwickelten Programmpaketes SAFE2 /76/ gelöst, das eine Sammlung der dafür erforderlichen Unterprogramme ist. Damit wurde die Netzgenerierung, die Erzeugung der Zeigerstruktur zum kompakten Abspeichern einer schwach besetzten Matrix und der Aufbau der Systemmatrizen durchgeführt. Die vier Schritte zur Implementierung des Verfahrens werden im folgenden diskutiert.

Gebietszerlegung

Das Grundgebiet Ω wurde ausgehend von einer Grobeinteilung in Vierecke zerlegt. Dabei wurde das Netz an der zu erwartenden Unstetigkeit und im Inneren der Laval-Düse verfeinert. Die feinste Zerlegung hatte 513 Knoten und 448 Elemente. Die Düsenkontur wurde im konvergent-divergenten Teil durch eine Parabel vorgegeben. Dadurch kann man durch zwei Parameter die Düsengestalt für numerische Tests leicht verändern. Außerdem ergibt sich hierdurch eine einfache Darstellung der quasieindimensionalen Lösung, siehe Zierep /15/. Das Netz wurde in einem Vorlauf zur eigentlichen Strömungsberechnung erzeugt, da keine adaptiven Verfahren zur Verfügung standen. Deshalb ist das Netz nach der Verfeinerung noch relativ gleichmäßig und die volle Flexibilität des FEM-Ansatzes ist noch nicht ausgenutzt.

Aufbau der Systemmatrizen

Während der Durchführung des Verfahrens wird ungefähr 75% der Rechenzeit für die Aufstellung der Systemmatrizen verbraucht. Deshalb wird dieser Gesichtspunkt etwas ausführlicher beschrieben, um zu zeigen, daß hier ein erheblicher Rechenzeitgewinn möglich ist. Wie in Kapitel 2 gezeigt wurde, sind zum Aufstellen der Systemmatrizen Integrale der Form

$$M_{ii} = \int_{\Omega} w(x)u(x)\varphi_i(x)\,dx \tag{4.1}$$

auszuwerten. Dabei ist M_{ii} das i-te Diagonalelement der Systemmatrix M und $\varphi_i(x)$ die zum Knoten i gehörige Basisfunktion. Die Funktion $w(x)$ soll die Nichtlinearität repräsentieren und $u(x)$ ist die gesuchte Lösung. Beide Funktionen werden als Linearkombinationen der Basisfunktionen dargestellt,

$$f(x) = \sum_{n=1}^{N} f(x_n)\varphi_n(x) \quad . \tag{4.2}$$

Die hier verwendeten Ansatzfunktionen sind stückweise bilineare Funktionen auf den Vierecken der Gebietzerlegung. Deshalb kann man die nachfolgenden Überlegungen zur effizienten numerischen Integration im Eindimensionalen durchführen. Die Übertragung auf das Zweidimensionale geht dann analog zu anderen Tensorproduktformeln. In der folgenden Darstellung wird die Berechnung des Diagonalelements (4.1) betrachtet, da bei der Berechnung der Subdiagonalelemente der numerische Aufwand halbiert ist.

Zerlegt man das Gebiet $\Omega \subset\subset \mathbb{R}$ gemäß

$$T = \{x_1,\ldots,x_N\} \text{ mit } x_i < x_{i+1}$$

dann ist, da $\varphi_i(x)$ seinen Träger auf (x_{i-1}, x_{i+1}) hat die Berechnung (4.1) gegeben durch

$$\begin{aligned} M_{11} &= R_1 \\ M_{ii} &= R_i + S_{i-1} \quad \text{für } i = 2,\ldots,N-1 \\ M_{NN} &= S_{N-1} \quad , \end{aligned} \tag{4.3}$$

mit

$$R_i := \int_{x_i}^{x_{i+1}} w(x)u(x)\varphi_i(x)\,dx \quad \text{für } i = 2,...,N-1 \tag{4.4}$$

und

$$S_i := \int_{x_i}^{x_{i+1}} w(x)u(x)\varphi_{i+1}(x)\,dx \quad \text{für } i = 1,...,N-1\ . \tag{4.5}$$

Da die Integrale in (4.4) und (4.5) kubische Polynome sind, werden sie in der Regel mit der 2-Punkt Gaußformel integriert, siehe z.B. Dhatt-Touzot /75/ oder die Anwendungen in Hughes /36/ und Hansbo /35/. Diese Formel lautet

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = \tfrac{1}{2} h_i\,(f(z_i^1) + f(z_i^2))\ ,\ h_i := x_{i+1} - x_i\ , \tag{4.6}$$

mit den Stützstellen

$$z_i^1 := \tfrac{1}{2}(x_i + x_{i+1}) + \frac{1}{2\sqrt{3}} h_i \quad \text{und} \quad z_i^2 := \tfrac{1}{2}(x_i + x_{i+1}) - \frac{1}{2\sqrt{3}} h_i\ . \tag{4.7}$$

Setzt man dies in (4.6) und (4.5) ein, so erhält man

$$R_i = \tfrac{1}{2} h_i\,(w(z_i^1)u(z_i^1)\varphi_i(z_i^1) + w(z_i^2)u(z_i^2)\varphi_i(z_i^2)) \tag{4.8}$$

und

$$S_i = \tfrac{1}{2} h_i\,(w(z_i^1)u(z_i^1)\varphi_{i+1}(z_i^1) + w(z_i^2)u(z_i^2)\varphi_{i+1}(z_i^2))\ . \tag{4.9}$$

Im allgemeinen wird die Lösung u nur an den Knoten abgespeichert. Deshalb ist die Berechnung von (4.8) und (4.9) aufwendig, da die Werte von u an den Stützstellen durch Interpolation gewonnen werden müssen. Zusammenfassend erhält man den Algorithmus "Gauß" :

Für $i = 1,...,N-1$ ist

$$U_i^{1,2} := u_i + [\varphi_{i+1}(z_i^1) \,/\, \varphi_i(z_i^1)]\, u_{i+1}$$

$$V_i^{1,2} := w_i + [\varphi_{i+1}(z_i^1) \,/\, \varphi_i(z_i^1)]\, w_{i+1}$$

$$G_i^{1,2} := U_i^{1,2}\, V_i^{1,2}$$

$$R_i = [\tfrac{1}{2} h_i\, (\varphi_i(z_i^1))^3\,]\, G_i^1 + [\tfrac{1}{2} h_i\, (\varphi_i(z_i^2))^3\,]\, G_i^2$$

$$S_i = [\tfrac{1}{2} h_i\, (\varphi_i(z_i^1))^2\, \varphi_{i+1}(z_i^1)]\, G_i^1 + [\tfrac{1}{2} h_i\, (\varphi_i(z_i^2))^2\, \varphi_{i+1}(z_i^2)]\, G_i^2 \ .$$

Für $i = 1,...,N$ ist jetzt M_{ii} durch (4.3), (4.4) und (4.5) gegeben. Die Terme in eckigen Klammern hängen nicht von der aktuellen Lösung ab und können in einem Vorlauf berechnet werden. Hier ist allerdings wieder zwischen Rechenzeitgewinn und Speicherplatzbedarf abzuwägen, wie es auch bei der Verwendung iterativer Löser zu tun ist. Für den Vergleich der Algorithmen soll davon ausgegangen werden, daß die Größen in eckigen Klammern bereits berechnet sind. Zählt man die Operationen im Algorithmus "Gauß" , so sieht man, daß er insgesamt 10 Multiplikationen und 7 Additionen erfordert.

Um das aufwendige Interpolieren der Funktionswerte zu vermeiden, sucht man eine Formel, deren Stützstellen günstiger liegen. Die Simpson-Regel, die auch kubische Polynome exakt integriert, erfüllt diese Forderung. Sie lautet

$$\int_{x_i}^{x_{i+1}} f(x)\, dx = \frac{1}{6} h_i\, (f(x_i) + 4f(x_{i+\frac{1}{2}}) + f(x_i)) \ . \tag{4.10}$$

Berechnet man damit (4.4) und (4.5) , so erhält man

$$R_i = \frac{1}{6} h_i\, (w_i u_i + 2w_{i+\frac{1}{2}}\, u_{i+\frac{1}{2}}) \tag{4.11}$$

und

$$S_i = \frac{1}{6} h_i\, (2w_{i+\frac{1}{2}}\, u_{i+\frac{1}{2}} + w_{i+1} u_{i+1}) \ . \tag{4.12}$$

Diese weitere Vereinfachung kommt daher, daß die Basisfunktionen an den entsprechenden Kubaturpunkten verschwinden. Damit erhält man den

Algorithmus "Simpson" :

Für $i = 1,...,N - 1$ ist

$$A_{i+\frac{1}{2}} := u_i + u_{i+1}$$

$$B_{i+\frac{1}{2}} := w_i + w_{i+1}$$

$$C_{i+\frac{1}{2}} := [\tfrac{1}{12} h_i]\, A_{i+\frac{1}{2}}\, B_{i+\frac{1}{2}} \ ,$$

für $i = 1,...,N$ ist $D_i = w_i u_i$,

für $i = 1$ ist $M_{11} = \frac{1}{6} h_1 D_1 + C_{3/2}$,

für $i = 2,...,N - 1$ ist $M_{ii} = [\frac{1}{6}(h_{i-1} + h_i)]\, D_i + C_{i-\frac{1}{2}} + C_{i+\frac{1}{2}}$

und für $i = N$ ist $M_{NN} = \frac{1}{6} h_N D_N + C_{N-\frac{1}{2}}$.

Nimmt man wieder an, daß die Terme in eckigen Klammern bereits berechnet sind, so muß man also ingesamt 4 Multiplikationen und 4 Additionen ausführen. Der Algorithmus "Gauß" ist damit doppelt so aufwendig wie der Algorithmus "Simpson", der implementiert wurde. Bei dem hier berechneten zeitabhängigen Problem, müssen die Systemmatrizen häufig neu aufgebaut werden. Deshalb hat sich diese Verbesserung durch eine Einsparung von ungefähr 35% an Rechenzeit bemerkbar gemacht.

Zu Testzwecken wurde die Ordnung der Kubaturformel weiter verringert und die Integrale der Form (4.1) mit der Trapezregel ausgewertet. Dies führte aber, insbesondere im Fall einer transsonischen Strömung zu einem instabilen Verfahren.

Die Aufstellung der Systemmatrizen verlangt außerdem die Berechnung der in Kapitel 3 angegebenen Dämpfungsmatrix δ. Dazu muß die Wurzel aus einer positiv definiten Matrix gezogen werden. Dies ist prinzipiell mit dem Satz von Cayley-Hamilton exakt möglich, wie es von Hoger, Carlson /74/ angegeben wurde. Die Berechnungen hierzu sind allerdings, da sie das Lösen quadratischer Gleichungen erfordern aufwendig. Deshalb geht man von der Darstellung

$$A = \sum_{i=1}^{4} \lambda_i \phi_i \phi_i^T \tag{4.13}$$

einer positiv definiten Matrix mit Hilfe der Eigenwerte und des dyadischen Produkts der zugehörigen Eigenvektoren aus. Die Eigenwerte und die Eigenvektoren der 4 × 4-Matrix können effizient mit dem QR-Algorithmus wie er in Golub, Van Loan /13/ angegeben ist berechnet werden. Damit gilt

$$A^{\frac{1}{2}} = \sum_{i=1}^{4} \lambda_i^{\frac{1}{2}} \phi_i \phi_i^T \quad .$$

Die Auswertung der nichtlinearen Koeffizientenfunktionen zur Berechnung der Matrixeinträge wurde an den Elementschwerpunkten vorgenommen. Dazu bildet man das arithmetische Mittel aus den vier Eckknotenwerten von u .

Da die Systemmatrizen dünn besetzt sind, werden nur die Matrixelemente ungleich null und die zugehörigen Zeigervektoren abgespeichert. Das ist möglich, da zur Lösung der entstehenden Gleichungssysteme ein iteratives Lösungverfahren verwendet wird, und es deshalb nicht zu einem Auffüllen der Bänder wie bei einem direkten Löser kommt. Eine wesentliche Eigenschaft des hier entwickelten Verfahrens ist es, daß die entstehenden linearen Gleichungssysteme iterativ gelöst werden können. Dadurch wurde es eigentlich erst mit den zur Verfügung stehenden Geräten durchführbar.

Durchführung des Zeitschrittverfahrens.

Schreibt man das volldiskrete Verfahren (3.13) mit den in Kapitel 2 eingeführten Bezeichnungen, so lautet es

$$M(\bar{v}^n)(\tfrac{1}{k}(v^n - v^{n-1}) + K(\bar{v}^n)(\bar{v}^n) = 0 \quad , \tag{4.14}$$

wobei die Matrizen M und K durch (3.13) gegeben sind. Zur Lösung des nichtlinearen Gleichungssystems (4.14) wurde in jedem Zeitschritt eine Fixpunktiteration verwandt, wie sie auch von Hansbo /35/ vorgeschlagen wird. Nach 3 - 5 Iterationen wurde sie mit einem Residuum von 10^{-8} in der Maximumnorm abgebrochen. Die Durchführung dieses Ver

fahrens verlangt, daß in jedem Iterationsschritt die Systemmatrizen neu aufgebaut und berechnet werden müssen. Es liefert eine qualitativ gute Übereinstimmung mit der quasieindimensionalen Lösung nach Zierep /15/, ist aber sehr rechenzeitintensiv und für feinere Gebietszerlegungen zu aufwendig. Deshalb wurde es nach einigen numerischen Tests verworfen.

Als nächstes wurde eine Prediktor-Korrektor-Version des Crank-Nicolson-Verfahrens implementiert, wie sie von Douglas, Dupont /14/ angegeben wird. Sie lautet für (4.14)

$$M(v^{n-1})(\tfrac{1}{k}(w - v^n)) + K(v^{n-1})(\tfrac{1}{2}(w + v^n)) = 0 \tag{4.15}$$

$$M(\tfrac{1}{2}(w + v^{n-1}))(\tfrac{1}{k}(v^n - v^{n-1})) + K(\tfrac{1}{2}(w + v^{n-1}))(\tfrac{1}{2}(v^n + v^{n-1})) = 0 \ .$$

Das Verfahren (4.15) entspricht dem Verfahren (4.14), wobei zur Lösung des nichtlinearen Gleichungssystems ein Schritt des Newton-Verfahrens durchgeführt wird. Die Ergebnisse stimmen gut mit denen von (4.14) überein, mit dem Vorteil, daß pro Zeitschritt nur noch ein zusätzlicher Matrixaufbau notwendig ist. Allerdings müssen hierbei zwei Lösungsvektoren gespeichert werden. Mit der letzten Vereinfachung kann man pro Zeitschritt noch einen Matrixaufbau einsparen, indem man die Lösung zum Auswertungszeitpunkt $t_{n-1/2}$ aus der Lösung zu zwei vorherigen Zeitpunkten extrapoliert,

$$\bar{v}^n = \tfrac{3}{2} v^{n-2} - \tfrac{1}{2} v^{n-1} + O(k^2) \tag{4.16}$$

und in (4.14) einsetzt. Diese Variante wird z. B. auch in Thomee / 8/ betrachtet. Gestartet wird der Algorithmus mit dem Verfahren (4.15). Ein weiterer Vorteil dieser Methode ist, daß man durch ein günstiges Anordnen der Matrix-Vektor-Multiplikationen keinen zusätzlichen Lösungsvektor speichern muß. Auch mit diesem entgültig implementierten Verfahren konnte eine gute Übereinstimmung der numerischen Ergebnisse mit der quasieindimensionalen Lösung festgestellt werden, wie in Kapitel 5 dargestellt ist. Analog zu den in den Arbeiten von Douglas, Dupont und Thomee angegebenen Beweisen kann man zeigen, daß die Verfahren (4.15) und (4.14) mit der Extrapolation (4.16) unter genügender Regularitätsvoraussetzung an die Lösung die gleichen Stabilitäts- und Konvergenzeigenschaften haben

wie die, die in der vorliegenden Arbeit für das Verfahren (4.14) bewiesen wurden.
Der Aufwand des implementierten Zeitschrittverfahrens soll jetzt im Vergleich mit dem häufig verwendeten Verfahren von Beam und Warming /80/ diskutiert werden. Dazu werden wieder die Euler-Gleichungen aus Kapitel 1 betrachtet. Damit die weitere Darstellung übersichtlicher ist, wurde die Notation geändert:

$$u_{,t} + f(u)_{,x} + g(u)_{,y} = 0 \quad , \tag{4.17}$$

mit dem durch (1.2.17) gegeben Vektor u und den entsprechenden Flüssen f und g .
Das Verfahren von Beam und Warming geht von der konservativen Formulierung (4.17) aus und basiert auf einer Taylorentwicklung für u in der Zeit. Im einfachsten Fall erhält man ein implizites Eulerverfahren für die Änderung der Lösung

$$\Delta u^n = \Delta t \frac{\partial}{\partial t} \Delta u^n + \Delta t \frac{\partial}{\partial t} u^n + O((\Delta t)^2) \quad , \tag{4.18}$$

wobei

$$\Delta u := u^{n+1} - u^n$$

ist. Setzt man (4.17) in (4.18) ein, so ergibt sich unter Vernachlässigung des Abschneidefehlers

$$\Delta u^n + \Delta t(\frac{\partial}{\partial x} \Delta f^n + \frac{\partial}{\partial y} \Delta g^n) = -\Delta t(\frac{\partial}{\partial x} f^n + \frac{\partial}{\partial y} g^n) \quad . \tag{4.19}$$

Dadurch erhält man ein implizites Zeitschrittverfahren, das wegen der Terme Δf^n und Δg^n auf eine nichtlineare Gleichung führt. Die Linearisierung nach Beam und Warming benutzt eine Entwicklung der Flüsse in der Zeit. Dazu betrachtet man die Jakobimatrizen

$$J_f := f_{,u} \quad \text{und} \quad J_g := g_{,u} \quad .$$

Damit folgt

$$\Delta f^n = J_f \Delta u^n + O((\Delta t)^2) \quad \text{und} \quad \Delta g^n = J_g \Delta u^n + O((\Delta t)^2) \quad . \tag{4.20}$$

Die Gleichung (4.20) wird in (4.19) eingesetzt. Das Ergebnis ist

$$(I + \Delta t(\frac{\partial}{\partial x} J_f + \frac{\partial}{\partial y} J_g))\Delta u^n = -\Delta t(\frac{\partial}{\partial x} f^n + \frac{\partial}{\partial y} g^n) \quad , \tag{4.21}$$

oder mit den durch (4.21) gegebenen Gesamtmatrizen und Vektoren

$$A^n \Delta u^n = -\Delta t E^n \quad . \tag{4.22}$$

Die Gestalt der Matrix A wird durch die Knotennummerierung bestimmt. Hier soll

$$\Delta u := (\Delta u_1, ..., \Delta u_k)^T \text{ mit } \Delta u_i := (\rho_i, (\rho u)_i, (\rho v)_i, e_i)^T \quad ,$$

wobei $1 \leq i \leq k$ und k die Anzahl der Gitterpunkte ist, gewählt werden. Dann ist A *auf einem regelmäßigen Gitter* eine block-pentadiagonale Matrix. Dabei besteht jeder Block aus einer 4×4-Matrix. Verwendet man ADI-Techniken, wie sie auch von Beam und Warming vorgeschlagen werden, dann wird A in zwei block-tridiagonale Matrizen faktorisiert. Dieses System kann dann mit einem, die spezielle Struktur berücksichtigenden Gauß-Algorithmus gelöst werden. Dieses Vorgehen setzt aber ein regelmäßiges Gitter voraus. Andernfalls kann kein direkter Löser verwendet werden, ohne die Besetzungsstruktur der Matrix A während der Lösung des Systems zu ändern. Aus (4.2.1) sieht man auch, daß der Zeitschritt Δt "klein" sein muß, da für große Zeitschritte die Diagonaldominanz der Matrix A nicht mehr garantiert ist. In diesem Fall muß der Gauß-Algorithmus mit Pivotierung durchgeführt werden. Das ist aber nicht möglich, da durch eventuelle Zeilenvertauschungen die regelmäßige Struktur der Matrix A zerstört würde. Der Algorithmus von Beam und Warming ist also nicht direkt, so wie das hier entwickelte FEM-Verfahren auf unregelmäßigen Gittern durchführbar.

Häufig werden in der Matrix A die Superdiagonal- oder sogar alle Ausserdiagonalelemente vernachlässigt. Dadurch werden die Gleichungen teilweise oder völlig entkoppelt. Ist man nur an einer stationären Lösung interessiert, so dient das Zeitschrittverfahren zur Linearisierung der Euler-Gleichungen. In diesem Fall ist das Entkoppeln gerechtfertigt, und natürlich auch für den FEM-Algorithmus möglich. Damit wird der Aufwand zum Aufbau der Matrizen und zum Lösen des linearen Gleichungssystems stark reduziert. Wie aber schon in der Arbeit von Beam und Warming gezeigt wird, ist mit dieser Entkopplung ein erheblicher Stabilitätsverlust des Verfahrens verbunden.

Zusammenfassend kann man sagen, daß das Zeitschrittverfahren im FEM-Algorithmus

nicht aufwendiger ist, als bei dem häufig verwendeten Verfahren von Beam und Warming. Ein großer Vorteil des FEM-Verfahrens ist, daß auch für dreidimensionale Probleme auf unstrukturierten Gittern die entstehenden linearen Gleichungssysteme iterativ lösbar sind. Im Rahmen der "splitting"-Verfahren, zu denen auch das Verfahren von Beam und Warming gehört, ist die Übertragung der Lösungsalgorithmen auf dreidimensionale Probleme noch eine offene Frage.
Der wesentliche Unterschied im Aufwand der betrachteten Verfahren liegt also in dem für FEM-Verfahren komplizierteren Aufbau der Systemmatrizen.

Lösen der Gleichungssysteme.
Die Koeffizientenmatrix B^0 ist positiv definit und somit auch die zugehörige Systemmatrix M aus (4.14). Die Matrix K, die im wesentlichen den Differenzenquotienten der ersten Ableitung enthält hat damit die Kondition $O(h^{-1})$. Deshalb hat das Gleichungssystem aus (4.14) für Δt und δ hinreichend klein die Kondition $O(1)$ und man kann es mit dem Verfahren der sukzessiven Überrelaxation lösen, siehe Golub, Van Loan /13/. Bei den numerischen Experimenten konnte man beobachten, daß sich ein Sortieren der Knoten in Strömungsrichtung konvergenzbeschleunigend auswirkt. In der Regel mußte man 5 bis 10 Iterationen im SOR-Verfahren zur Lösung des linearen Gleichungssytems ausführen um ein Residuum von 10^{-8} in der Maximumnorm zu erhalten. Dabei wurde der Relaxationsparameter je nach Strömungsgeschwindigkeit in einem Bereich zwischen 0.8 und 1.15 gewählt.

Als Anfangswerte wurden die Freistromwerte im Unendlichen im Gebiet Ω vorgegeben. An den festen Wänden wurden für die Komponenten v_2 und v_3 des Lösungsvektors v gefordert, daß

$$v_2 \cdot n_x + v_3 \cdot n_y = 0$$

ist. Das entspricht der transformierten Randbedingung (1.4.11). Am Einströmrand wurde für den Fall der Unterschallanströmung v_1, v_2, v_3 und für den Fall der Überschallanströmung v_1, v_2, v_3, v_4 vorgegeben. Am Ausströmrand und am Einströmrand wurden für die übrigen Komponenten keine Randbedingungen gestellt.

5. Numerische Anwendungen

Mit den in diesem Kapitel vorgestellten numerischen Anwendungen soll die Realisierbarkeit des in dieser Arbeit entwickelten FEM-Verfahrens an dem "harten" Problem einer instationären transsonischen Gasströmung in einer Düse demonstriert werden. Die Euler-Gleichungen sind ein nichtlineares System von partiellen Differentialgleichungen. Sie sind zur prinzipiellen Verifikation numerischer Verfahren weniger geeignet, da zu viele Schwierigkeiten gleichzeitig auftreten. Deshalb soll zunächst die schon in den vorherigen Kapiteln benutzte lineare Modellgleichung für den Transportterm der Euler-Gleichungen betrachtet werden.

Für die zweidimensionale Advektionsgleichung

$$u_{,t} + \beta \cdot \nabla u = 0 \qquad \text{in } \Omega \times (0; T]$$
$$u(x,y,0) = u_0(x,y) \quad \text{in } \Omega ,$$

auf dem Gebiet $\Omega = [-1; 1] \times [-1; 1]$ wird das in der Literatur der Differenzenverfahren häufig zitierte "rotating cone problem" gerechnet. In diesem Fall setzt man für den Vektor β der charakteristischen Richtung

$$\beta = \alpha\,(y, -x)^T .$$

Auf den Randstücken

$$\Gamma_- := \{\, (x;y) \in \partial\Omega : n(x,y) \cdot \beta(x,y) < 0 \,\}$$

wird der Randwert $u(x,y) = 0 \;\forall\, x,y \in \Gamma_-$ vorgeben. Die exakte Lösung wird durch eine Rotation des Anfangswertes u_0 um den Ursprung mit der Winkelgeschwindigkeit α beschrieben. Dieser Testfall hat den Vorteil, daß falls $\mathrm{supp}(u_0) \subset\subset \Omega$, keine Wechselwirkung der Lösung mit dem Rand von Ω stattfindet. Deshalb kann man auch auf dem gesamten Rand $\partial\Omega$ den Wert 0 vorgeben. Als Anfangsbedingung wählt man einen Kegel. Damit ist die Höhenreduktion des Kegels eine Veranschaulichung für die Größe der numerischen Dissipation des Verfahrens.

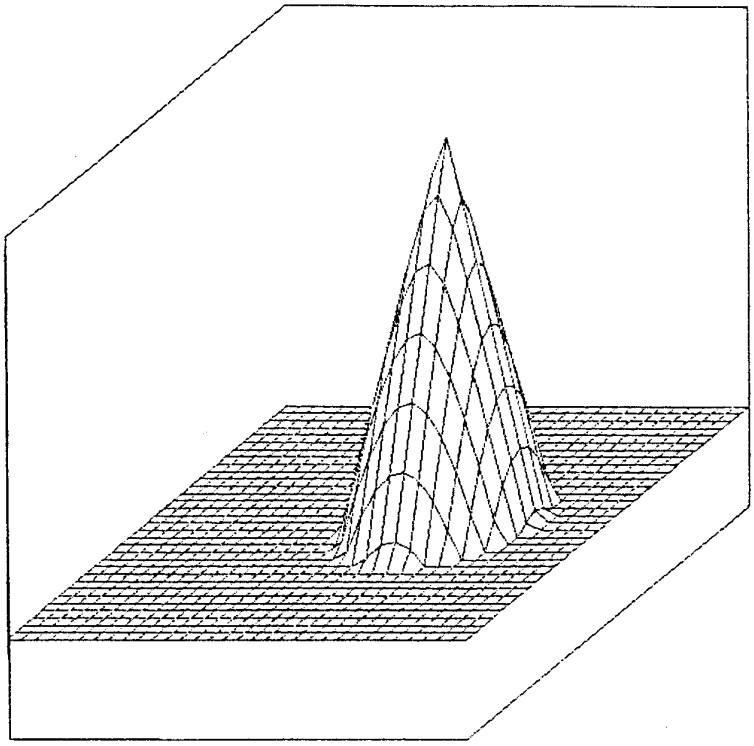


Abbildung 1

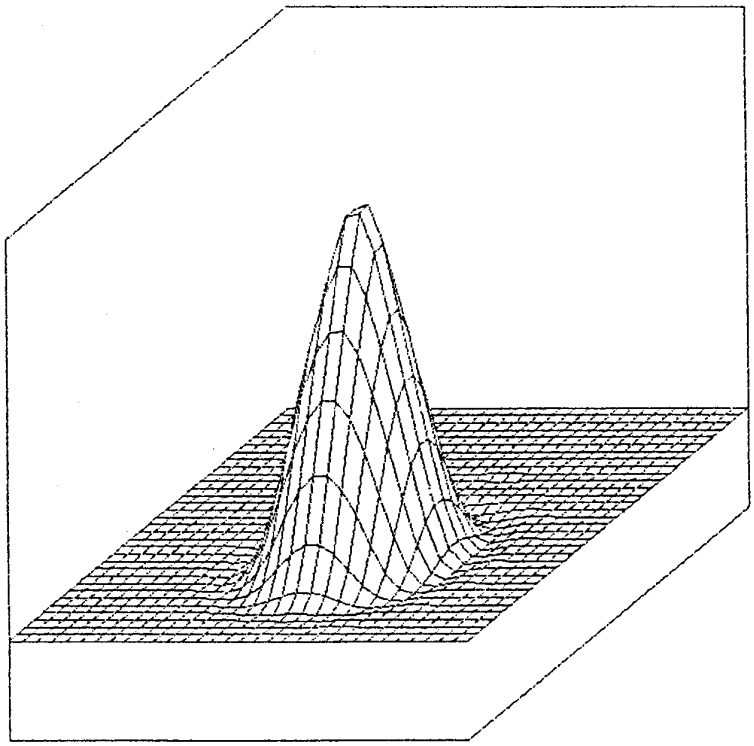


Abbildung 2

In Abbildung 1 ist der Anfangswert u_0 und in Abbildung 2 ist u nach einer Umdrehung dargestellt. Auf dem feinsten Gitter mit einer Schrittweite von 1 / 64 hat der Kegel nach einer Umdrehung noch 92% seiner Ausgangshöhe. Durch mehrmalige Halbierung der Ausgangsschrittweite konnte eine quadratische Konvergenz des L^2-Fehlers des Verfahrens beobachtet werden, wobei durch eine Verschiebung der berechneten Lösung der Phasenfehler ausgeglichen wurde. Ohne diese Korrektur ergibt sich eine Konvergenzordnung von 1.91 . Das Konvergenzverhalten des Verfahrens ist in Abbildung 3 dargestellt.

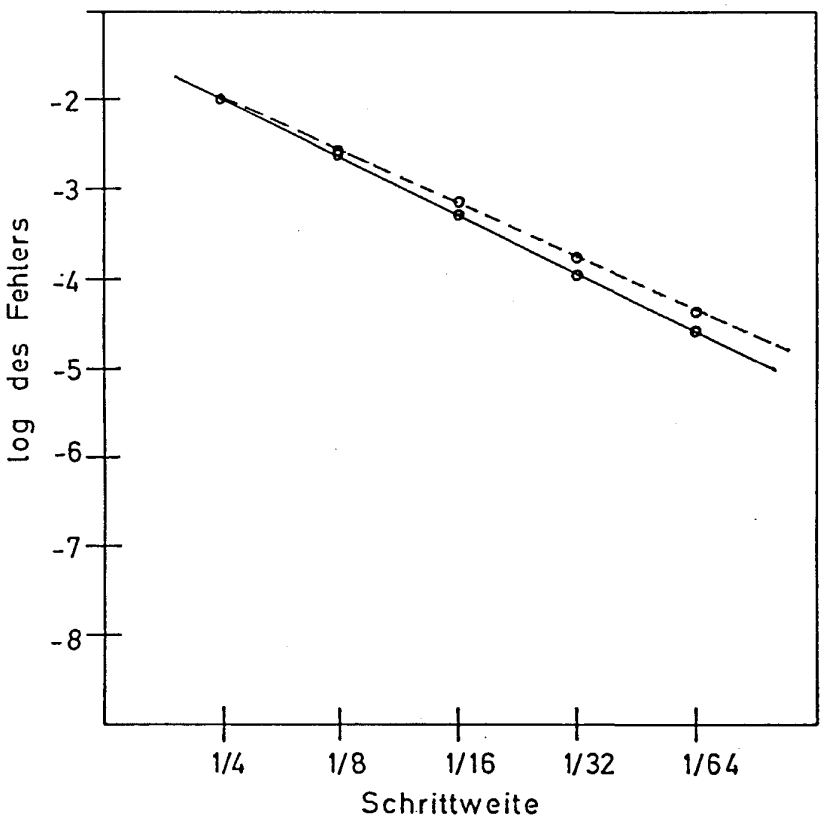


Abbildung 3

Um den Fall einer unstetigen Anfangsvorgabe für u zu testen, wurde ein Plateau,

$$u_0(x,y) = 1 \quad ,\forall\, x,y \;:\; 0 < x < 0.3 \;\text{ und }\; |y| < 0.2 \;,$$

gewählt. Hier stellte sich nach einer Umdrehung eine Höhe von 0.86 ein. Die Konvergenzordnung betrug mit Phasenfehlerausgleich 0.49 und ohne Phasenfehlerausgleich 0.46 .
Das Ergebnis ist in Abbildung 4 dargestellt. Ähnliche Resultate wurden auch von Nävert /34/ für die Johnsonsche Variante der Stromliniendiffusion dokumentiert.

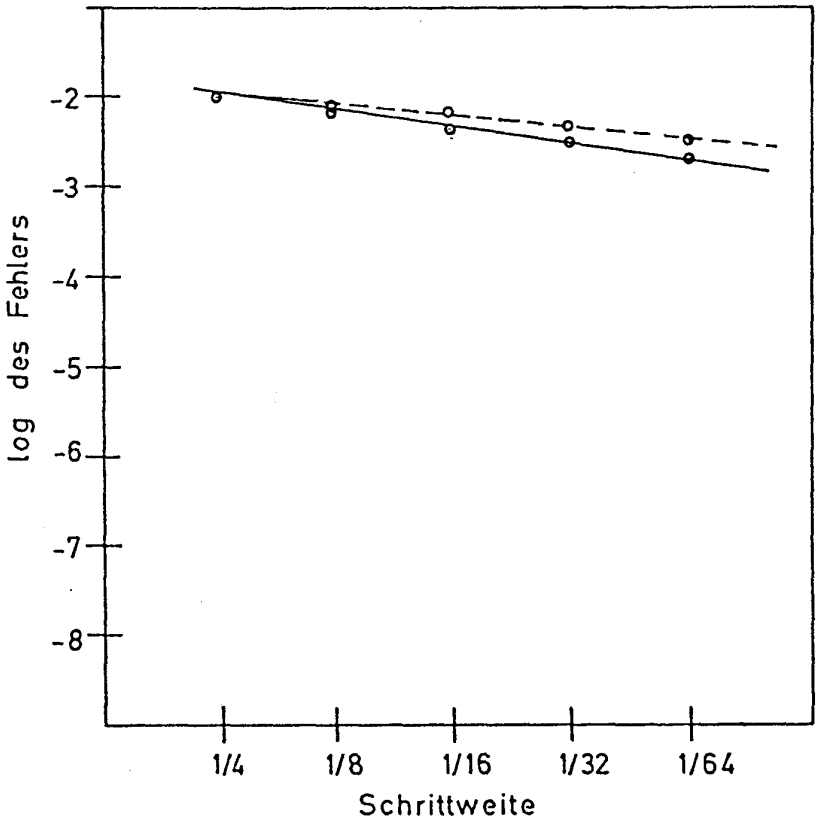


Abbildung 4

Bei vielen Differenzenverfahren bringt die Berücksichtigung der Randbedingungen Stabilitätsprobleme mit sich. Dem Studium numerischer Randbedingungen, das sind Randbedingungen, die vom Verfahren, nicht aber vom kontinuierlichen Problem verlangt werden, ist im Bereich der Differenzenverfahren eine umfangreiche Literatur gewidmet, siehe z. B. Sod / 1/ und die darin angegebenen Zitate. Um zu testen, wie sich das Verfahren bei Wechselwirkung der Lösung mit dem Rand des Gebietes Ω verhält, wurde

$$\beta = (-1, 1)^T$$

und damit der Einströmrand zu

$$\Gamma = \{ (x;y) : x = 1 , y \in (-1;1) \}$$

gewählt.

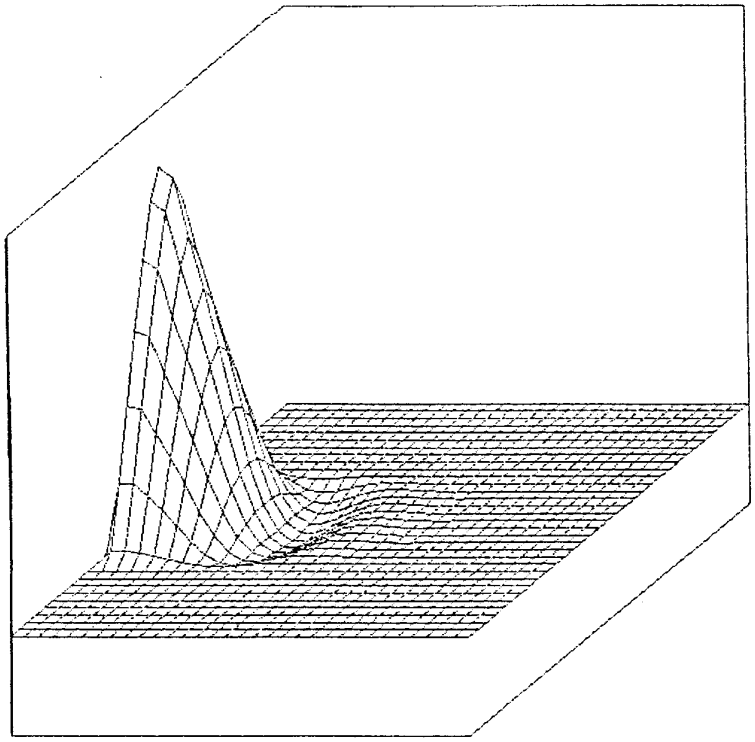


Abbildung 5

Die Abbildung 5 zeigt, daß sich keine Stabilitätsprobleme durch Reflexion ergeben. Für große Zeiten verschwindet der Anfangswert u_0 völlig aus dem Gebiet Ω .

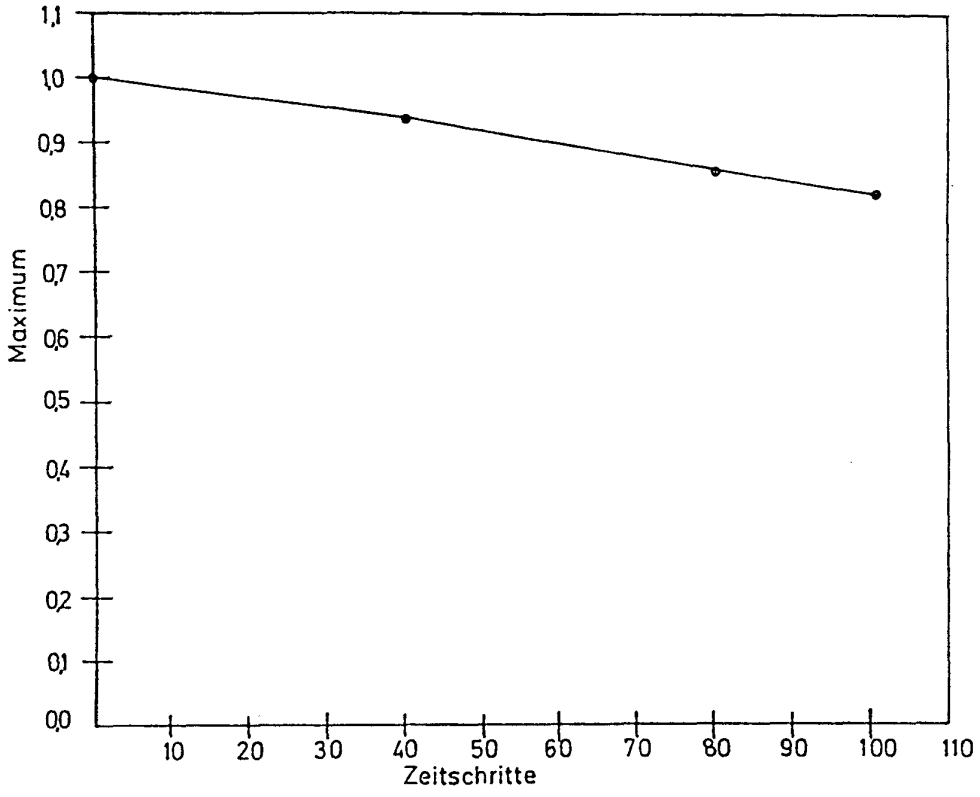
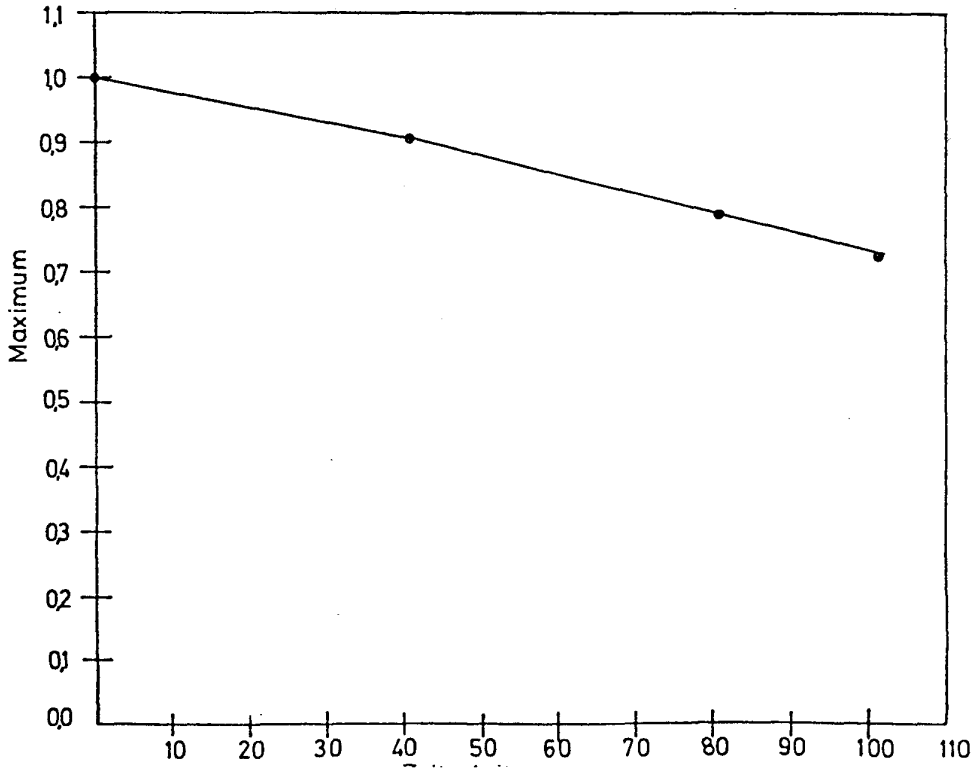


Abbildung 6

Als nächstes wurde der Einfluß des Dämpfungsparameters untersucht. Die Abbildung 6 zeigt die Abnahme der Kegelhöhe im Verlauf von zwei Rotationen für den Dämpfungsparameter $\delta = h$.



Die Abbildung 7 zeigt die Abnahme der Kegelhöhe für die gleiche Rechnung. Hierbei wurde die Dämpfung $\delta = 2h$ gewählt. Bei der Herleitung des Verfahrens war der Dämpfungsparameter δ beliebig . Man sieht aber, daß die Wahl des Dämpfungsparameters die Dissipativität des Verfahrens beeinflußt.

Das "rotating cone problem" zeigt auch ein für viele strömungsmechanische Situationen typisches Erscheinungsbild der Lösung. Eine fast gleichförmige Lösung, hier $u = 0$ ist an nur wenigen Punkten gestört. Diese Störung wird durch den Kegel dargestellt. Deshalb arbeiten momentan viele Autoren, z. B. Johnson /12/ und vor allem Löhner /70/ an adaptiven oder "domain splitting"-Verfahren, die diesen Sachverhalt berücksichtigen. Solche Zugänge sind natürlich wegen ihrer erforderlichen Gitterflexibilität im FEM-Kontext sehr viel besser als mit Differenzenverfahren zu realisieren.
Die bekanntesten Differenzenverfahren wurden in einer Arbeit von Munz, Schmidt /78/ für die oben dargestellten Fälle einer stetigen (Kegel) und einer unstetigen (Plateau) Anfangsvorgabe numerisch untersucht. Ein Vergleich dieser Resultate mit den hier erzielten zeigt, daß das FEM-Verfahren gleichwertige Ergebnisse liefert.

Nach diesen Testrechnungen wurden zur tatsächlichen Prüfung des Verfahrens die zweidimensionalen Euler-Gleichungen aus Kapitel 1 berechnet. Dazu wurden zunächst alle physikalischen Größen dimensionslos gemacht:

$$x := \frac{\tilde{x}}{\tilde{L}} \qquad y := \frac{\tilde{y}}{\tilde{L}} \qquad t := \frac{\tilde{t}\ \tilde{c}_0}{\tilde{L}} \qquad u := \frac{\tilde{u}}{\tilde{c}_0}$$

$$v := \frac{\tilde{v}}{\tilde{c}_0} \qquad \rho := \frac{\tilde{\rho}}{\tilde{\rho}_0} \qquad p := \frac{\tilde{p}}{\tilde{\rho}_0\,\tilde{c}_0^{\,2}} \qquad e := \frac{\tilde{e}}{\tilde{\rho}_0\,\tilde{c}_0^{\,2}} \ ,$$

wobei L eine charakteristische Länge ist. Die mit " 0 " indizierten Größen sind die zum Zeitpunkt $t = 0$ angegebenen. Das Zeichen " ~ " soll bedeuten, daß diese Größen dimen-

sionsbehaftet sind. Als Anfangsbedingung erhält man damit

$$v_0 = 0 \qquad u_0 = Mc_0 \qquad c_0 = 1 \qquad \rho_0 = 1$$

$$e_0 = 1 / (\gamma(\gamma-1)) + 1/2\, u_0^2 \qquad p_0 = 1 / \gamma \ .$$

Die Jakobimatrizen der Flüsse haben nach Kapitel 1.3 damit die Eigenwerte

$$e_1^x = u + c \quad e_2^x = e_3^x = u \quad e_4^x = u - c$$

und

$$e_1^y = v + c \quad e_2^y = e_3^y = v \quad e_4^y = v - c \ .$$

Also können für die transformierten Variablen im Fall der subsonischen Einströmung die Impulskomponenten $(\rho u)_0$ und $(\rho v)_0$ und die spezifische Energie e_0 am Einströmrand vorgegeben werden. Die Einströmmachzahl M war in allen Fällen 0.6 .

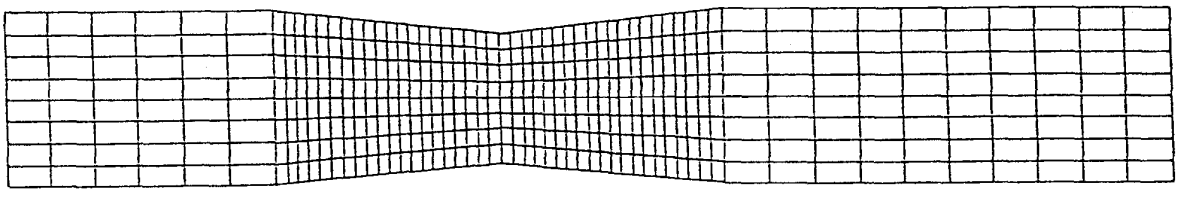


Abbildung 8

Die Abbildung 8 zeigt das verwendete Netz. Der Düsenquerschnitt ist in Abhängigkeit der Länge x gegeben durch

$$A(x) = \begin{cases} 1 & \text{für } 0 \le x \le 3 \\ 1 - \frac{0.2}{24.5}(x-3)(2.3 - 0.2x) & \text{für } 3 \le x \le 10 \\ 1 & \text{für } 10 \le x \le 16 \end{cases} .$$

Wie die Darstellung des Gitters zeigt, wurde keine Ausrichtung an den zu erwartenden Lösungsverlauf vorgenommen. Die Düsenströmung war auch ein Testfall für den GAMM-Workshop 1980 in Stockholm /79/ . Ein Vergleich mit den dort durchgeführten

Rechnungen zeigt, daß hier für die sich einstellende komplizierte Lösung ein relativ grobes Netz verwendet wird. Dies war durch den zur Verfügung stehenden Speicherplatz bedingt. Mit dieser Konfiguration und den obengenannten Anfangsbedingungen wurden der Einfluß der Ausströmrandbedingungen und die Genauigkeit der erzielten Ergebnisse getestet. Wie in Zierep /15/ gezeigt wird, hängt die zu erwartende Lösung bei festgehaltener Gestalt der Düse im wesentlichen von der Ausströmbedingung ab. An der festen Wand wurde, wie schon in Kapitel 4 beschrieben, die Bedingung

$$u\, n_x + v\, n_y = 0$$

verwendet. Für die Kontinuitätsgleichung und die Energiegleichung werden die entsprechenden Randintegrale zu Null gesetzt. Diese Bedingungen werden jedoch durch die natürlichen Randbedingungen automatisch erfüllt.

Zunächst wurden am Ausströmrand keine Randbedingungen gestellt. Es ist bemerkenswert, daß das Verfahren trotzdem stabil blieb und sich qualitativ gute Werte ergaben. Bei einem Differenzenverfahren wär auf jeden Fall eine Extrapolation der Randwerte nötig gewesen. Es stellte sich eine transsonische Strömung mit supersonischer Ausströmung ein. Zum Vergleich mit den quasieindimensionalen Lösungen nach Zierep /15/ wurden jeweils die berechneten Werte auf der Symmetriachse der Düse verwandt.

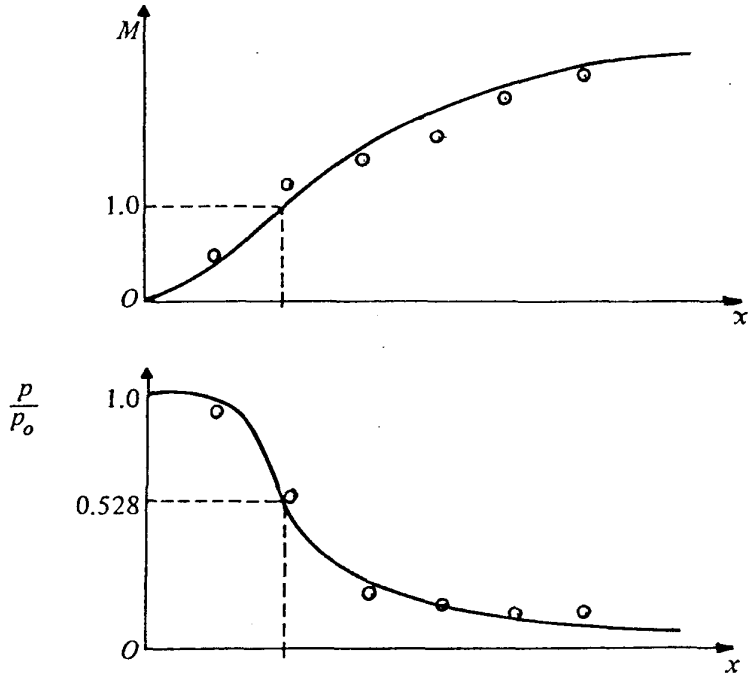


Abbildung 9

In Abbildung 9 sind die durch die quasieindimensionale Näherung gegebenen Werte dargestellt. Beim Vergleich dieser Werte ist natürlich zu beachten, daß die quasieindimensionale Rechnung selbst eine Näherung darstellt. Die berechneten Werte sind mit einer " 0 " markiert. Eine Untersuchung zur Konvergenz konnte nicht gemacht werden, da keine exakte Lösung zur Verfügung steht. Gröbere Gitter lieferten wenig brauchbare Resultate und feinere Gitter konnten mit den vorhandenen Mitteln nicht gerechnet werden. Das Verfahren wird für den Fall einer supersonischen Ausströmung nicht voll gefordert, da sich eine glatte Lösung einstellt. Die Symmetrie der Lösung konnte reproduziert werden.

Als nächstes wurde am Ausströmrand die Dicht zu ρ_0 vorgeschrieben. Dadurch stellte sich eine transsonische Strömung mit subsonischer Ausströmung ein. In diesem Fall tritt eine Diskontinuität an der Schallinie in der Düse auf. Beim Vergleich der Lösungen kommt hier nur noch die Symmetrieachse der Düse in Frage. Die Herleitung der quasieindimensionalen Gleichungen setzt voraus, daß der Stoß senkrecht zur Stromlinie ist. Da hier ein gekrümmter Stoß vorliegt, und deshalb alle Strömungsgrößen nach dem Stoß vom Stoßwinkel abhängen, ist diese Annahme nur noch auf der Symmetrieachse erfüllt.

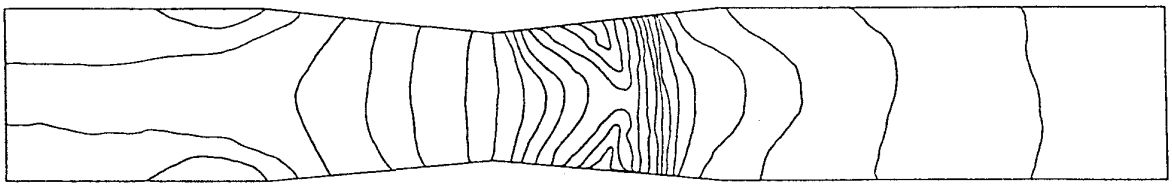


Abbildung 10

Die Darstellung der Niveaulinien für die Dichte in Abbildung 10 zeigt einen steilen Gradienten kurz nach der Stelle mit dem kleinsten Düsenquerschnitt A . Deutlich ist der gekrümmte Verlauf des Stoßes zu erkennen. Obwohl es sich um ein achsensymmetrisches Pro-

blem handelt, wurde der gesamte Düsenquerschnitt diskretisiert. Es sollte untersucht werden, ob das Verfahren die Symmetrie der Lösung reproduziert. Man erkennt aus der obigen Abbildung, daß sich eine nahezu symmetrische Strömung einstellt. Um eine symmetrische Lösung zu erhalten wurden 600 Zeitschritte durchgeführt. Im Verlauf der Rechnung konnte man ein Oszillieren der Lösung um die Symmetrieachse beobachten. Es stellte sich jedoch keine stationäre symmetrische Lösung ein. Allerdings ist zu bemerken, daß die Störung der Symmetrie in einem Bereich von ein bis drei Prozent lag. Eine Ursache für dieses Verhalten ist wohl darin zu sehen, daß der Abschneidefehler durch die Nichtlinearität unsymmetrisch ausgebreitet wird. Auch hat das Verfahren, wie jede Diskretisierung, eine gewisse numerische Viskosität und es erhebt sich die Frage ob die berechnete diskrete Lösung überhaupt stationär ist. Eine exakte Lösung ist insbesondere für den Fall einer viskosen Strömung nicht bekannt. Dieses Beispiel macht auch deutlich, daß der Frage nach der Dissipativität des Verfahrens größte Aufmerksamkeit gewidmet werden muß. Das Ziel jedes Lösungsverfahrens strömungsmechanischer Probleme sollte es sein, die kompressiblen Navier-Stokes-Gleichungen berechnen zu können. Diese Gleichungen reagieren insbesondere für hohe Strömungsgeschwindigkeiten in Ablösegebieten sehr empfindlich auf Änderungen der Viskosität des strömenden Mediums. Deshalb sollte das Verfahren möglichst wenig "numerische" Viskosität besitzen. Dies wurde für das FEM-Verfahren im linearen Fall mit dem "rotating cone problem" dokumentiert. Andererseits sollen Unstetigkeiten und starke Oszillationen gedämpft werden. Daß das Verfahren auch dazu in der Lage ist, zeigen die Lösungen zur Düsenströmung.

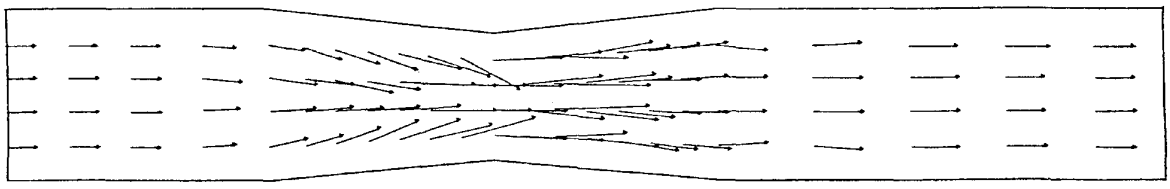


Abbildung 11

Aus der Darstellung der Geschwindigkeitsvektoren in Abbildung 11 sieht man, daß die größten Geschwindigkeiten an der engsten Stelle der Düse auftreten. Die Störung der laminaren Einströmung durch die Düse ist am Ausströmrand des Gebietes schon fast wieder abgeklungen. Die Herleitung der symmetrischen Form der Euler-Gleichungen in Kapitel 1.4 konnte nur unter der Annahme der Isentropie der Strömung geschehen. In Kapitel 1.5 wurde dann diskutiert, in wieweit diese Annahme im Fall der Düsenströmung erfüllt ist.

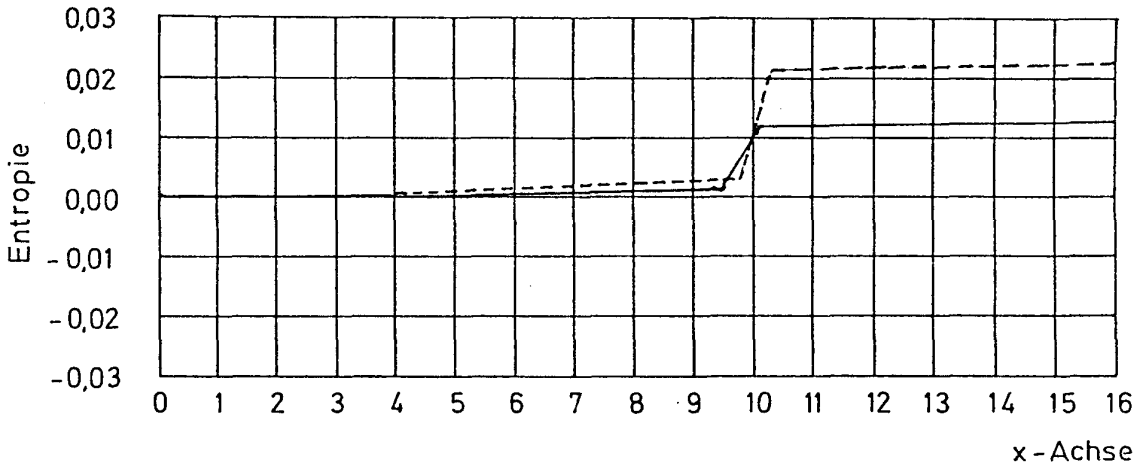


Abbildung 12

In der Abbildung 12 ist der Entropieverlauf in der Mitte der Düse (durchgezogene Linie) und für 1/4 des Düsenquerschnitts (gestrichelte Linie) dargestellt. Es zeigt sich, daß im konvergenten Teil ($x = 4$) und für die Schallinie ($x \approx 10$) Entropieschwankungen auftreten. Sie bewegen sich aber in einer Größenordnung, die die Annahme der Isentropie rechtfertigen. Die Darstellung zeigt auch das in den Kapitel 1.3 und 1.4 bewiesene Verhalten des FEM-Verfahrens, die Entropie zu erhalten. Ein einmal erreichtes Entropieniveau wird nicht wieder unterschritten. Dies entspräche einer unphysikalischen Lösung und würde den zweiten Hauptsatz der Thermodynamik verletzen. Die gleiche Rechnung wurde auch mit der quasilinearen Formulierung von (1.2.17) und ohne Stromliniendiffusion durchgeführt.

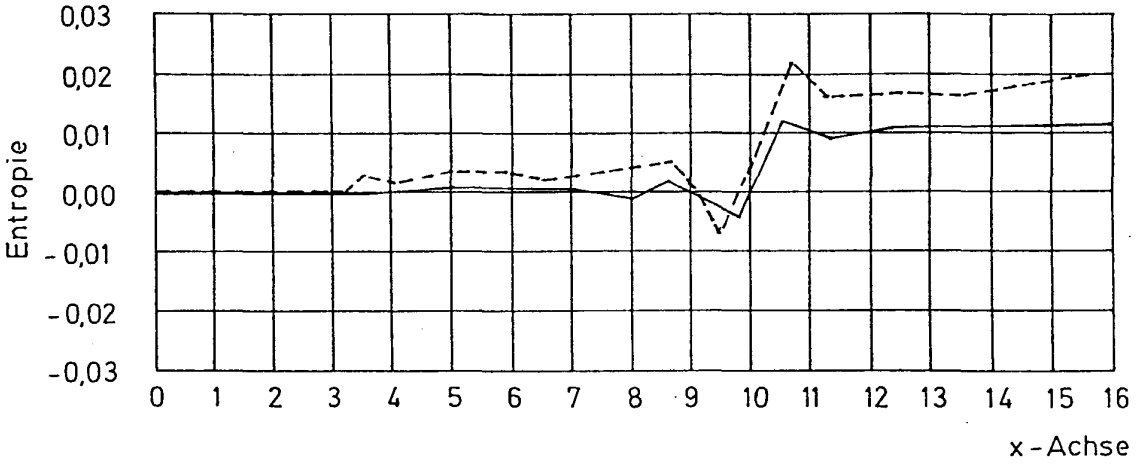


Abbildung 13

Die Abbildung 13 zeigt die entsprechenden Entropieverläufe. Deutlich sind die Oszillationen in der Entropie und sogar das auftreten negativer Entropien zu sehen. Dies bedeutet, daß eine unphysikalische Lösung approximiert wird. Allerding konnten für diese Rechnung nur wenige Zeitschritte durchgeführt werden, da das Verfahren instabil war.

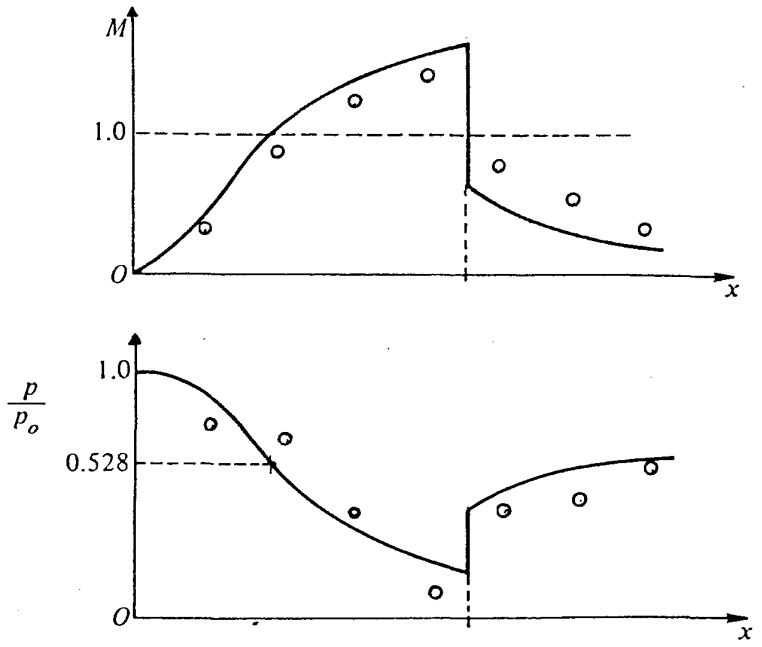


Abbildung 14

Ein Vergleich der berechneten Werte mit der quasieindimensionalen Theorie ist in Abbildung 14 dargestellt. Daraus wird insbesondere deutlich, daß das Verfahren noch keine so genaue Auflösung der Diskontinuität wie die Differenzenverfahren erreicht hat. Eine Lokalisierung der Unstetigkeit auf ein bis zwei Gitterpunkte, wie sie von den "flux-splitting" - Verfahren geliefert wird, war nicht möglich. Der Stoß war über ein bis drei Elemente verschmiert. Allerdings wurde auch kein an den Verlauf der Lösung angeglichenes Netz verwendet.

In allen Berechnungsbeispielen wurde der Dämpfungsparameter δ in einem Bereich zwischen 1.5 und 0.5 mal der lokalen Ortsschrittweite variiert. Dabei konnten keine signifikanten Änderungen in der Lösung festgestellt werden. Beim Verlassen dieses Bereiches war allerdings ein Verschlechtern der Lösung zu beobachten. Enthielt die Lösung Unstetigkeiten, so kam es bereits ab einem Dämpfungswert von 0.8 mal der lokalen Ortsschrittweite zu Oszillationen in der Lösung.
Es war nicht möglich, andere Werte für δ zu testen, da hierfür auch das SOR-Verfahren nicht mehr konvergierte. Ebenso wurde der Zeitschritt Δt gleich einer mittleren Ortsschrittweite gewählt. Für große Zeitschritte wurde das Verfahren instabil, was sicherlich auch an der zur Linearisierung verwendeten Extrapolation liegt. Wurde der Zeitschritt verdoppelt, so blieb das Verfahren mit einem Relaxationsparameter von 0.6 im SOR-Verfahren, d. h. starker Unterrelaxation, stabil. Allerding nahm die Anzahl der notwendigen SOR-Schritte stark zu. Da das SOR-Verfahren im wesentlichen nur Matrix-Vektor-Multiplikationen erfordert, ist diese Aufwandserhöhung nicht entscheidend. Wurde der Zeitschritt verdreifacht, so war das Verfahren für keine Wahl des Dämpfungs- und des Relaxationsparameters stabil.

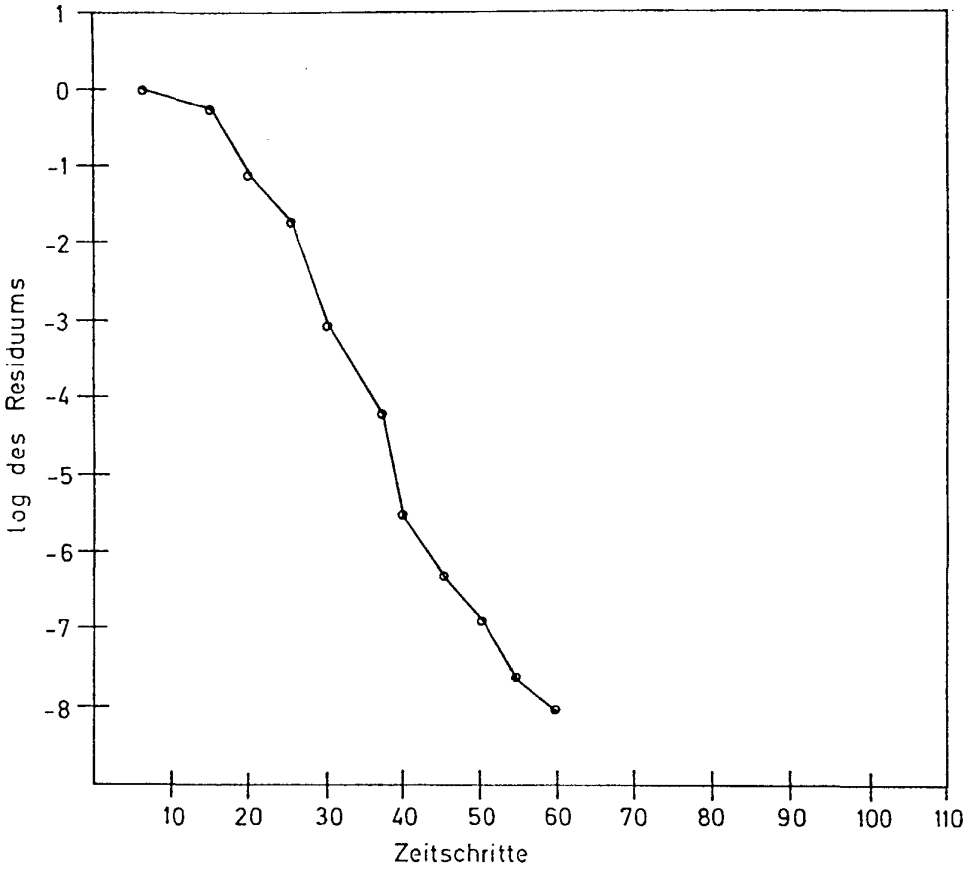


Abbildung 15

Die Abbildung 15 zeigt die Konvergenzgeschichte der Düsenrechnung für den Fall einer supersonischen Ausströmung. Der Dämpfungsparameter δ wurde gleich der lokalen Ortsschrittweite und der Zeitschritt Δt wurde gleich einer mittleren Ortsschrittweite gewählt. Es ist jeweils das Residuum nach der ersten SOR-Iteration dargestellt. Für die Strömung ohne Schock konnte mit leichter Überrelaxation, der Parameterwert war 1.15, nach maximal 6 Iterationen eine Residuum von 10^{-8} in der Maximumnorm erreicht werden. Die numerischen Experimente haben gezeigt, daß für die Stabilität des Verfahrens auch ein Residuum von 10^{-4} genügt. Für größere Residuen wird das Verfahren allerdings instabil. Ist man nur an einer stationären Lösung interessiert und betrachtet das Verfahren als Pseudozeitschrittverfahren, so kann man auch das Residuum von 10^{-4} als Abbruchkriterium ver-

wenden. Nach etwa 60 Zeitschritten führte das SOR-Verfahren nur noch jeweils eine Iteration aus. Die dann erreichte Approximation wurde als stationäre Lösung betrachtet.

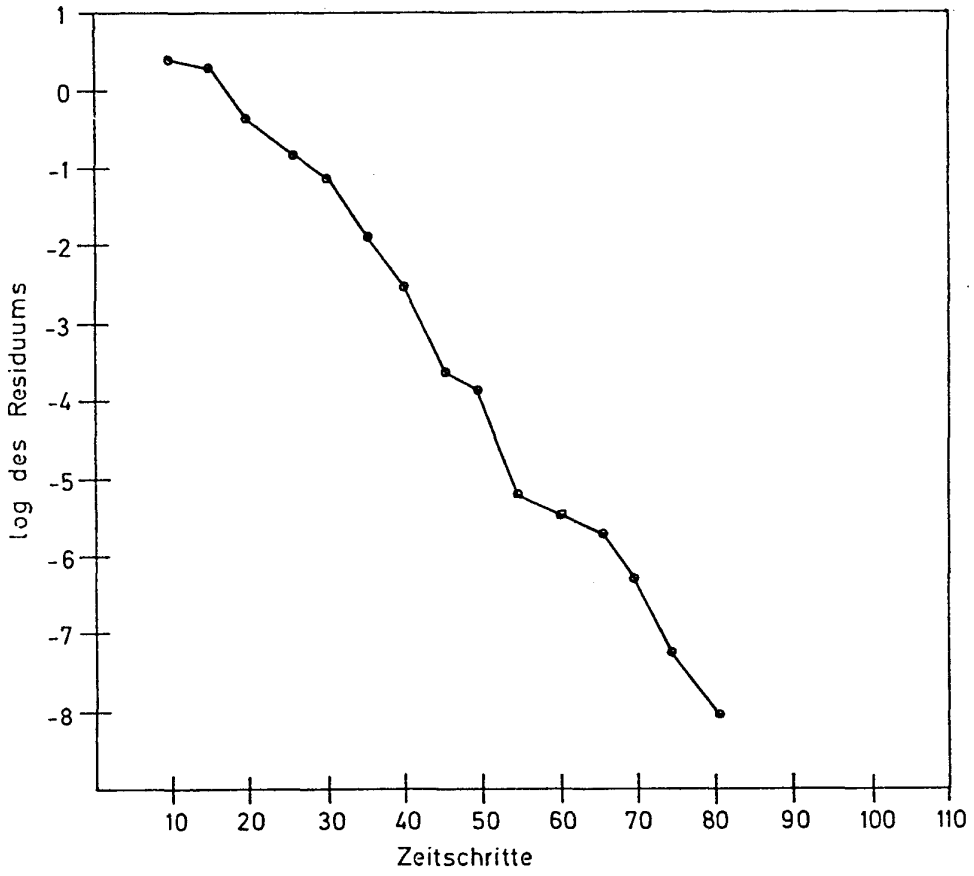


Abbildung 16

Die Darstellung in Abbildung 16 zeigt die Konvergenzgeschichte im Fall einer subsonischen Ausströmung. Dieses Strömungsfeld enthält einen gekrümmten Stoß, der die Schallinie darstellt. Der Dämpfungsparameter und die Zeitschrittweite wurden wie im vorherigen Beispiel gewählt. Der Relaxationsparameter im SOR-Verfahren mußte auf 0.8 gesetzt werden. Nach ungefähr 30 Iterationen konnte er aber auf 1.0 erhöht werden. Die Anzahl der SOR-Iterationen um ein Residuum von 10^{-8} zu erreichen betrug maximal 10 . Um für diesen Fall eine nach dem hier gewählten Kriterium stationäre Lösung zu erreichen, mußten etwa 80 Zeitschritte durchgeführt werden.

Das Verfahren benötigte in der hier implemetierten Version auf einem SIEMENS PC MX-2 für die Düsenberechnung auf dem feinsten Gitter mit 513 Knoten einen Speicherplatz von 3 Megabyte. Die Rechenzeit betrug für den Fall einer Strömung mit Schock ungefähr 95 Minuten.

Der Aufwand zur Durchführung des hier entwickelten Verfahrens verteilt sich etwa wie folgt :

	Rechenzeit	Speicherplatz
Gittererzeugung	5%	20%
Aufbau der Systemmatrizen	80%	70%
Durchführen eines Zeitschritts	5%	5% (Hilfsvektor)
Lösen des Gleichungssystems	10%	5%

In den vorherigen Kapiteln wurde gezeigt, daß das hier verwendete Verfahren etwa um einen Faktor 16 sparsamer ist, als die von Johnson vorgeschlagene Version der Stromliniendiffusion für instationäre Probleme. Diese Einsparung kommt hauptsächlich durch die geringere Anzahl der Systemmatrizen des hier entwickelten Verfahrens zustande. Deshalb bezieht sich dieser Faktor sowohl auf den Speicherplatz, als auch auf die Rechenzeit. Damit ist sein Verfahren für technisch interessante Aufgaben praktisch nicht durchführbar. Dies mag auch ein Grund dafür sein, daß Hughes /41/ für seine Berechnungen zur Euler-Gleichung das explizite Eulerverfahren benutzt und Hansbo /35/ sich auf die inkompressible Navier-Stokes-Gleichung in der Stromfunktionsformulierung beschränkt.

Aus der Aufstellung des numerischen Aufwandes wird deutlich, daß er größtenteils in der Verwaltung des Gitters und als Folge davon, im Aufbau der Systemmatrizen liegt. Bei einem Vergleich mit Differenzenverfahren muß aber beachtet werden, daß dies eigentlich kein Unterschied ist, da eine unregelmäßige Netzstruktur eben mehr Informationen, z. B. über Zusammenhangskomponenten oder Nachbarschaftsbeziehungen der Elemente enthält. Die entstehende Systemmatrix, die durch ihre Besetzungsstruktur den geometrischen Zusammenhang zwischen den Knoten widerspiegelt, muß dann natürlich entsprechend kompliziер-

ter sein. Diese Netzinformationen wären im wesentlichen auch für ein Differenzenverfahren nötig. Sie müssen aber dafür nicht explizit berechnet und gespeichert werden, da sie auf regelmäßigen Gittern in offensichtlicher Weise gegeben sind. Da ein iterativer Löser verwendet wird, der unter der Voraussetzung einer guten Startlösung nur wenige Iterationen benötigt, könnte man den Speicherplatzbedarf reduzieren und die Matrixelemente bei Bedarf jeweils neu berechnen. Dies wird häufig beim Verfahren von Beam und Warming getan, da hier beim Lösen auf jedes Matrixelement nur zweimal zugegriffen wird.
Werden die hier durchgeführten Rechnungen mit denen des GAMM-Workshops /79/ verglichen, so sieht man, daß ein sehr grobes Netz verwendet wurde. Umso bemerkenswerter ist doch die gute Übereinstimmung der berechneten Werte mit den quasieindimensionalen Näherungen.

6. Zusammenfasssung und Ausblick

In der vorliegenden Arbeit wird ein FEM-Verfahren zur Diskretisierung der Euler-Gleichungen der Gasdynamik für eine reibungsfreie, kompressible transsonische Düsenströmung entwickelt und analysiert.

Dazu wird eine Herleitung dieser Gleichungen und der damit zusammenhängenden physikalischen Begriffe gegeben. Unter der Annahme einer isentropen Strömung können diese Gleichungen, wenn alle Differentiationen im klassischen Sinn verstanden werden, in ein symmetrisches hyperbolisches System von Erhaltungsgleichungen überführt werden. Die Frage, ob dies auch in der schwachen Formulierung der Euler-Gleichungen und damit für allgemeine Strömungen gilt, ist noch offen.

Die Zulässigkeit der verschiedenen für das Innenraumproblem notwendigen Randbedingungen wird mit Hilfe der von Lax entwickelten Theorie für symmetrische hyperbolische Systeme diskutiert. Für die Randbedingungen an der festen Wand konnte Stabilität gezeigt werden.

Im nächsten Abschnitt der Arbeit wird das Verfahren der Stromliniendiffusion eingeführt. Dies geschieht zunächst für eine skalare Modellgleichung und wird dann auf das symmetrische hyperbolische System der Euler-Gleichungen übertragen. Es zeigt sich, daß mit dem zweiten Hauptsatz der Thermodynamik die Stabilität des Verfahrens für die Euler-Gleichungen garantiert ist.

Die mathematische Analyse des Verfahrens wird an skalaren Modellgleichungen durchgeführt. Dies stellt keine Einschränkung dar, da man das Vorgehen auf symmetrische hyperbolische Systeme verallgemeinern kann. Im Gegensatz zu vielen anderen Zugängen kann die Analyse auch für die nichtlineare hyperbolische Modellgleichung durchgeführt werden. Das wichtigste Hilfsmittel hierzu ist die von Tartar und DiPerna entwickelte Methode der kompensierten Kompaktheit.

Der letzte Teil der Arbeit beschäftigt sich mit der effizienten Implementierung des hier entwickelten Verfahrens. Besonderes Augenmerk gilt dabei dem Aufbau der Systemmatrizen,

da dies der rechenzeit- und speicherplatzintensivste Teil des FEM-Verfahrens ist. Dabei wird gezeigt, wie eine Halbierung des Aufwandes erreicht werden kann. Das aus numerischer Sicht wichtigste Ergebnis der vorherigen Überlegungen ist, daß nach der Symmetrisierung die Koeffizientenmatrix des Systems der Euler-Gleichungen positiv definit ist. Die im Crank-Nicolson-Verfahren zu invertierende Matrix hat die Kondition O(1), und das lineare Gleichungssystem kann mit dem SOR-Verfahren gelöst werden. Am Beispiel eines weitverbreiteten impliziten Differenzenverfahrens und der von Johnson /31/ vorgeschlagenen Variante der Stromliniendiffusion wird der numerische Aufwand mit dem hier entwickelten FEM-Verfahren verglichen.
Die mit dem implementierten Programm durchgeführten Rechnungen bestätigen die theoretischen Ergebnisse, und zeigen, daß das Verfahren prinzipiell zur Diskretisierung hyperbolischer Probleme geeignet ist.

Eine Gegenüberstellung der hier erzielten numerischen Ergebnisse mit Veröffentlichungen aus dem Bereich der Differenzenverfahren macht aber auch deutlich, daß das Verfahren insbesondere für die Euler-Gleichungen noch verbessert werden muß. Dabei ist aber auch zu sehen, daß das FEM-Verfahren einen Lösungszugang für "allgemeine" Situationen liefert. Viele Differenzenverfahren, wie z. B. die "shock-fitting"-Verfahren sind auf eine spezielle strömungsmechanische Situation zugeschnitten und liefern dort gute Ergebnisse. Die Übertragung auf andere Situationen ist im allgemeinen aufwendig.
Um die Genauigkeit der Stromliniendiffusionsmethode zu erhöhen hat Hughes /37/ das Verfahren für stationäre hyperbolische Probleme mit einem "least-squares"-Ansatz gekoppelt. Dadurch erhält man im Diskreten eine positiv definite Systemmatrix. Diese Erweiterung wurde von Johnson und Szepessy /32/ mit ihrer Übertragung auf instationäre Probleme analysiert. Ein Vorteil dieses Zuganges ist, daß im Gegensatz zu dem in dieser Arbeit geführten Beweis bei der Analyse der nichtlinearen Modellgleichung keine Annahmen an die Folge der Approximationen nötig ist. Ein entscheidender Nachteil ist der noch größere

numerische Aufwand dieses Verfahrens.

Momentan liegt der Vorteil dieser FEM-Zugänge noch in ihrer Allgemeingültigkeit und in der Möglichkeit, einer mathematischen Analyse der Verfahren unter realistischen Voraussetzungen an die Lösung. Die dazu entwickelten Techniken sind auch auf bestimnte Differenzenverfahren anwendbar, wenn man sie im FEM-Kontext sieht. Ein Problem bei der Analyse der Stromliniendiffusionsmethode liegt darin, daß man die im numerischen Experiment beobachtete quadratische Konvergenz für glatte Lösungen nicht zeigen kann. Da die Theorie aus Kapitel 2 nur eine Konvergenzordnung von 1.5 liefert, müssen auch andere Analysetechniken angewandt werden.

Das Ziel jeder Verfahrensentwicklung für strömungsmechanische Aufgabenstellungen sollte es sein, die kompressiblen Navier-Stokes-Gleichungen auf unregelmäßigen Gebieten des dreidimensionalen Raumes zu lösen. Diese Gleichungen bilden ein unvollständiges parabolisches System. Für hohe Strömungsgeschwindigkeiten findet beim Verlassen der Grenzschicht faktisch ein Typenwechsel von parabolischen zu hyperbolischen Gleichungen statt, da die Viskosität vernachlässigbar wird. Bereits von Johnson /30/ wurde eine Möglichkeit vorgeschlagen, wie viskose Terme im Rahmen der Stromliniendiffusionsmethode zu berücksichtigen sind. Seine Vorgehensweise ist aber unbefriedigend und weitere Untersuchungen sollen zeigen, wie dies in systematischer Weise geschehen kann.

Insgesamt kann man sagen, daß die FEM-Verfahren die Möglichkeit bieten, auf mathematisch fundierte Weise Diskretisierungsmethoden für hyperbolische Gleichungen zu entwikkeln und die gewonnenen Algorithmen zu analysieren. Diese Verfahren können dann dem speziellen strömungsmechanischen Problem angepaßt werden. Die für FEM-Verfahren anwendbaren Analysemethoden sind insbesondere für nichtlineare Gleichungen denen der Differenzenverfahren überlegen. In der numerischen Praxis haben allerdings die älteren Differenzenverfahren einen Erfahrungs- und Entwicklungsvorsprung, den die FEM-Verfahren

noch aufholen müssen. Nach den hier gesammelten theoretischen und numerischen Erfahrungen sollten sie dazu aber in der Lage sein.

7. Literaturverzeichnis

Numerische Lehrbücher :

/ 1/ G. A. Sod

Numerical Methods in Fluid Dynamics , Cambridge University Press 1986.

/ 2/ R. D. Richtmyer, K. W. Morton

Difference Methods for Initial Value Problems , Interscience Publishers 1967.

/ 3/ R. Peyret, T. D. Taylor

Computational Methods for Fluid Flow , Springer Verlag 1983.

/ 4/ P. J. Roache

Computational Fluid Dynamics , Hermosa Publishers Albuquerque 1972.

/ 5/ A. J. Baker

Finite Element Computational Fluid Mechanics , McGraw-Hill 1983.

/ 4/ P. G. Ciarlet

The Finite Element Method for Elliptic Problems , North-Holland 1979.

/ 7/ T. J. Chung

Finite Elemente in der Strömungsmechanik , Hanser Verlag 1982.

/ 8/ V. Thomee

Galerkin Finite Element Methods for Parabolic Problems ,

Springer Lecture Notes in Mathematics No. 1054 , 1984.

/ 9/ G. Strang, G. J. Fix

An Analysis of the Finite Element Method , Prentice-Hall 1973.

/10/ R. Temam

Navier–Stokes Equations, Theory and Numerical Analysis, North-Holland 1979.

/11/ R. Vichnevetsky, L. B. Bowles

Fourier Analysis of Numerical Approximations of Hyperbolic Equations,

S I A M Philadelphia 1982

/12/ C. Johnson

Numerical Solution of Partial Differential Equations by the Finite Element Method,

Cambridge University Press 1987.

/13/ G. H. Golub, Ch. Van Loan

Matrix Computations

North Oxford Academic 1983.

/14/ J. Douglas, T. Dupont

Galerkin Methods for Parabolic Equations,

SIAM J. Numer. Anal. Vol. 7, pp. 575-676, 1970.

Gasdynamische Lehrbücher :

/15/ J. Zierep

Theoretische Gasdynamik , Braun Verlag Karlsruhe 1976.

/16/ J. Zierep

Grundzüge der Strömungslehre , Braun Verlag Karlsruhe 1979.

/17/ J. D. Anderson, Jr.

Modern Compressible Flow , McGraw-Hill 1982.

/18/ A. J. Chorin, J. E. Marsden

A Mathematical Introduction to Fluid Mechanics , Springer Verlag 1984.

/19/ R. Courant, K. O. Friedrichs

Supersonic Flow and Shock Waves , Applied Mathematical Sciences 21 ,

Springer Verlag 1976.

Theorie hyperbolischer Gleichungen :

/20/ J. Smoller

Shock Waves and Reaction–Diffusion Equations , Springer Verlag 1983.

/21/ A. Majda

Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables,

Applied Mathematical Sciences Volume 53 , Springer Verlag 1984.

Viele Zitate zur Theorie hyperbolischer Gleichungen enthält

/22/ M. C. Bustos

On the Existence and Determination of Discontinuous Solutions to Hyperbolic Conservation Laws in the Theory of Sedimentation,

Dissertation Darmstadt 1984.

/23/ H. B. da Veiga

Fluid Dynamics,

Lecture Notes in Mathematics No. 1047 , Springer Verlag 1984.

/24/ P. D. Lax

Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves , S I A M Philadelphia 1973.

/25/ G. B. Witham

Linear and Nonlinear Waves, John Wiley & Sons 1974.

/26/ A. Jeffrey

Quasilinear Hyperbolic Systems and Waves, Pitman Publishing 1976.

/27/ C. Carasso, P.-A. Raviart, D. Serre

Nonlinear Hyperbolic Problems,

Lecture Notes in Mathematics No. 1270 ; Springer Verlag 1986.

Methode der Stromliniendiffusion :

/28/ C. Johnson

Streamline Diffusion Methods for Problems in Fluid Mechanics,

in R. H. Gallagher, *Finite Elements in Fluids,* Vol. 6 pp. 251-261, Wiley 1986.

/29/ C. Johnson, U. Nävert, J. Pitkäranta

Finte Element Methods for Linear Hyperbolic Problems,

Comp. Meth. in Appl. Mech. Eng. 45, pp.285-312, 1984.

/30/ C. Johnson, J. Saranen

Streamline Diffusion Methods for Incompressible Euler and Navier–Stokes Equations, Math. Comp. 47, pp. 1-18, 1986.

/31/ C. Johnson, A. Szepessy

On the Convergence of a Finite Element Method for a Nonlinear Conservation Law,

Math. Comp. 49, pp. 427-444, 1987.

/32/ C. Johnson, A. Szepessy, P. Hansbo

On the Convergence of Shockcapturing Streamline Diffusion Finite Element Methods for Hyperbolic Conservation Laws,

Technical Report 1987-21, Mathematics Department, Chalmers University of Technology, Göteborg Sweden , 1987.

/33/ C. Johnson, A. Szepessy

A Shock–Capturing Streamline Diffusion Finite Element Method for a Nonlinear Hyperbolic Conservation Law,

Technical Report 1986-09, Mathematics Department, Chalmers University of Technology, Göteborg Sweden , 1986.

/34/ U.Nävert

A Finite Element Method for Convection–Diffusion Problems,

Dissertation, Department of Computer Science, Chalmers University of Technology, Göteborg Sweden , 1982.

/35/ P. Hansbo

Finite Element Procedures for Conduction and Convection Problems,

Publication 86:7, Department of Structural Mechanics, Chalmers University of Technology, Göteborg Sweden , 1986.

/36/ T. J. R. Hughes, A. N. Brooks

A Theoretical Framework for Petrov–Galerkin Methods with Discontinuous Weighting Functions : Application to the Streamline Upwind Procedure,

in R. H. Gallagher, *Finite Elements in Fluids,* Vol. 6 pp. 47-65, Wiley 1982.

/37/ T. J. R. Hughes, M. Mallet, M. Mizukami

A New Finite Element Formulation for Computational Fluid Dynamics : II. Beyond SUPG,

Comp. Meth. in Appl. Mech. and Eng. 54, pp.341-355, 1986.

/38/ P. Lesaint, P.-A. Raviart

On a Finite Element Method for Solving the Neutron Transport Problem,

in C. de Boor, *Mathematical Aspects of Finite Element in Partial Differential Equations,* pp. 89-123, Academic Press, 1974.

/39/ T. J. R. Hughes, M. Mallet

A New Finite Element Formulation for Computational Fluid Dynamics : III. The Generalized Streamline Operator for Multidimensional Advection–Diffusive Systems,

Comput. Meth. in Appl. Mech. and Eng. 58, pp.305-328, 1986.

/40/ T. J. R. Hughes, L. P. Franca, M. Mallet

A New Finite Element Formulation for Computational Fluid Dynamics : VI. Convergence Analysis of the Generalized SUPG Formulation for Linear Time–Dependent Multidimensional Advection–Diffusive Systems,

Comput. Meth. in Appl. Mech. and Eng. 60, 1987.

/41/ T. J. R. Hughes, L. P. Franca, M. Mallet

Entropy–Stable Finite Element Methods for Compressible Fluids : Application to High Mach Number Flow with Shocks,

in P. Bergan et al. , *Finite Element Methods for Nonlinear Problems*, pp. 761-773, Springer Verlag 1986.

Symmetrisieren hyperbolischer Systeme :

/42/ R. F. Warming, R. M. Beam, B. J. Hyett

Diagonalization and Simultaneous Symmetrization of the Gas–Dynamic Matrices,

Math. Comp. 29, pp. 1037-1045, 1975.

/43/ E. Tadmor

Skew–Selfadjoint Form for Systems of Conservation Laws,

J. of Math. Anal. and Appl. 103, pp. 428-442, 1984.

/44/ E. Tadmor

Entropy Functions for Symmetric Systems of Conservation Laws,

J. of Math. Anal. and Appl. 122, pp. 355-359, 1987.

/45/ A. Harten

On the Symmetric Form of Systems of Conservation Laws with Entropy,

J. of Comp. Physics 49, pp. 151-164, 1983.

/46/ A. Harten, P. D. Lax

A Random Choice Finite Difference Scheme for Hyperbolic Conservation Laws,

SIAM J. Num. Anal. 18, pp. 289 -301, 1981.

Finite-Element-Verfahren für symmetrische hyperbolische Systeme :

/47/ M. S. Mock

Explicit Finite Element Schemes for First Order Symmetric Hyperbolic Systems,

Num. Math. 26, pp. 367-378, 1976.

/48/ P. Lesaint

Finite Element Methods for Symmetric Hyperbolic Equations,

Num. Math. 21, pp. 244-255, 1976.

Burger-Gleichung :

/49/ C. A. J. Fletcher

Burger's Equation : A Modell for All Reasons,

in J. Noye, *Numerical Solution of Partial Differential Equations*,

North-Holland 1982.

Theorie der kompensierten Kompaktheit :

/50/ R. J. Diperna

Uniqueness of Solutions to Hyperbolic Conservation Laws,

Indiana Uni. Math. Journal 28, pp. 137-188, 1979.

/51/ R. J. Diperna

Measure–Valued Solutions of Conservation Laws,

Arch. Rat. Mech. Anal. , pp. 223-270, 1985.

/52/ R. J. Diperna

Compensated Compactness and General Systems of Conservation Laws,

Transactions of the AMS 292, pp. 383-420, 1985.

/53/ R. J. Diperna

Convergence of the Viscosity Method for Isentropic Gas Dynamics,

Comm. Math. Physics 91, pp. 1-30, 1983.

/54/ R. J. Diperna

Conservation Laws and the Weak Topology,

Lecture Notes in Num. Appl. Anal. 5, pp. 1-15, 1982.

/55/ R. J. Diperna

Nonlinear Partial Differential Equations and the Weak Topology,

AMS Lectures in Appl. Math. 23, pp. 267-281, 1986.

/56/ R. J. Diperna

Convergence of Approximate Solutions to Conservation Laws,

Arch. Rat. Mech. Anal. 82, pp. 27-70, 1983.

/57/ L. Tartar

The Compensated Compactness Method Applied to Systems of Conservation Laws, in . M. Ball , *Systems of Nonlinear Partial Differential Equations,* pp. 263-285, NATO ASI Series 1983.

/58/ L. Tartar

Compensated Compactness and Application to Partial Differential Equations, in R. J. Knops , Research Notes in Mathematics, Nonlinear Analysis and Mechanics : Heriot-Watts Symposium, Vol. 4, Pitman Press 1979.

/59/ L. Tartar

Oscillation in Nonlinear Partial Differential Equations : Compensated Compactness and Homogenization,

AMS Lectures in Appl. Math. 23, pp. 243-266, 1986.

/60/ B. Dacorogna

Weak Continuity and Weak LowerSemicontinuity of Non−Linear Finctionals,

Lecture Notes in Mathematics No. 922 , Springer Verlag 1982.

Stabilitätstheorie :

/61/ K. O. Friedrichs

Symmetric Hyperbolic Linear Differential Equations,

Comm. Pure and Appl. Math 7, pp. 345-392, 1954.

/62/ H.-O. Kreiss

Initial Boundary Value Problems for Hyperbolic Systems,

Comm. Pure and Appl. Math 23, pp. 277-298, 1970.

/63/ P. L. Roe

Remote Boundary Conditions for Unsteady Multidimensional Aerodynamic Computations.

ICASE Report No. 86-75, 1986.

Petrov-Galerkin Verfahren :

/64/ J. W. Barrett, K. W. Morton

Optimal Petrov–Galerkin Methods through Approximate Symmetrization,

IMA J. of Num. Anal. 1, pp. 434-468, 1981.

/65/ K. W. Morton, A. K. Parrott

Generalised Galerkin Methods for First–Order Hyperbolic Equations,

J. of Comp. Physics 36, pp. 249-270, 1980.

/66/ D. B. Duncan, D. F. Griffiths

The Study of a Petrov–Galerkin Method for First–Order Hyperbolic Equations,

Comp. Meth. in Appl. Mech. and Eng. 45, pp. 147-166, 1984.

Taylor-Galerkin Verfahren :

/67/ J. Donea

A Taylor–Galerkin Method for Convective Transport Problems,

Int. J. Num. Eng. 20,pp. 101-119, 1984.

/68/ J. Donea, S. Giuliani

A Simple Method to Generate High Order Accurate Convection Operators for Explicit Schemes Based on Linear Finite Elements,

Int. J. Num. Meth. in Eng. 1, pp. 63-79, 1981.

/69/ R. Löhner, K. Morgan, O. C. Zienkiewicz

The Solution of Non–Linear Hyperbolic Equation Systems by the Finite Element Method, Int. J. for Num. Meth. in Fluids 4, pp. 1043-1063, 1984.

/70/ R. Löhner

Finite Element Flux–Corrected Transport (FEM - FCT) for the Euler and Navier–Stokes Equations,

Int. J. for Num. Meth. in Fluids 7, pp. 1093-1109, 1987.

/71/ M. Lötzerich

Beitrag zur Strömungsberechnung in Turbomaschinen mit Hilfe der Methode der finiten Elemente,

Dissertation an der RWTH Aachen 1987.

Sonstige Referenzen :

/72/ H. H. Henke

Lösung der Euler–Gleichungen mit der Methode der angenäherten Faktorisierung,

Dissertation an der RWTH Aachen 1986.

/73/ J. L. Steger, R. F. Warming

Flux Vector Splitting of the Inviscid Gasdynamic Equations with Application to Finite–Difference Methods,

J. of Comp. Physics 40, pp. 263-293, 1981.

/74/ A. Hoger, D. E. Carlson

Determination of the Stretch and Rotation in the Polar Decomposition of the Deformation Gradient,

Quart. Appl. Math. 42 (1), pp. 113-117, 1984.

/75/ Dhatt, Touzot

The Finite Element Method Displayed, Wiley & Sons 1984.

/76/ H. Blum, J. Harig, S. Müller, A. Walter

SAFE 2 , A Finite Element Package,

Universität des Saarlandes, Saarbrücken 1987.

/77/ S. Schochet

The Compressible Euler Equations in a Bounded Domain : Existence of Solutions and the Incompressible Limit,

Commun. Math. Phys. , pp. 49-75, 1986.

/78/ C.-D. Munz, L. Schmidt

Näherungsverfahren höherer Ordnung zur Approximation von Stoßwellen

Bericht Nr. 28, Fakultät für Mathematik, Universität Karlsruhe,

1985

/79/ A. Rizzi, H. Viviand

Numerical methods for the computation of inviscid transonic flows with shock waves (A GAMM workshop),

Notes on Numerical Fluid Mechanics Vol. 3, 1981, Vieweg Verlag Braunschweig.

/80/ R. M. Beam, R. F. Warming

An Implicit Finite–Difference Algorithm for Hyperbolic Systems in Conservation Law Form

J. of Comp. Physics 22, pp. 87-110, 1976

/81/ P. Le Floch

Entropy Weak Solutions to Nonlinear Hyperbolic Systems under Nonconservative Form

Commun. in Part. Diff. Equa. 13, pp. 669-727, 1988

/82/ P. D. Lax

Weak Solutions of Nonlinear Hyperbolic Equations and their Numerical Computation

Commun. Pure and Appl. Math. 7, pp. 159-193, 1954

/83/ S. K. Godunov

An Interesting Class of Quasilinear Systems

Dokl. Acad. Nauk. SSSR, 139, pp. 521-523, 1961

/84/ M. S. Mock

Systems of Conservation Laws of Mixed Type

J. Diff. Equa. , 37, pp. 70-88, 1980

Veröffentlichungen des Konrad-Zuse-Zentrums für Informationstechnik Berlin
Preprints Dezember 1988

SC 86-1. P. Deuflhard; U. Nowak. *Efficient Numerical Simulation and Identification of Large Chemical Reaction Systems.* (vergriffen)

SC 86-2. H. Melenk; W. Neun. *Portable Standard LISP for CRAY X-MP Computers.*

SC 87-1. J. Anderson; W. Galway; R. Kessler; H. Melenk; W. Neun. *The Implementation and Optimization of Portable Standard LISP for the CRAY.*

SC 87-2. Randolph E. Bank; Todd F. Dupont; Harry Yserentant. *The Hierarchical Basis Multigrid Method.* (vergriffen)

SC 87-3. Peter Deuflhard. *Uniqueness Theorems for Stiff ODE Initial Value Problems.*

SC 87-4. Rainer Buhtz. *CGM-Concepts and their Realization.*

SC 87-5. P. Deuflhard. *A Note on Extrapolation Methods for Second Order ODE Systems.*

SC 87-6. Harry Yserentant. *Preconditioning Indefinite Discretization Matrices.*

SC 88-1. Winfried Neun; Herbert Melenk. *Implementation of the LISP-Arbitrary Precision Arithmetic for a Vector Processor.*

SC 88-2. H. Melenk; H.M. Möller; W. Neun. *On Gröbner Bases Computation on a Supercomputer Using REDUCE.* (vergriffen)

SC 88-3. J. C. Alexander; B. Fiedler. *Global Decoupling of Coupled Symmetric Oscillators.*

SC 88-4. Herbert Melenk; Winfried Neun. *Parallel Polynomial Operations in the Buchberger Algorithm.*

SC 88-5. P. Deuflhard; P. Leinen; H. Yserentant. *Concepts of an Adaptive Hierarchical Finite Element Code.*

SC 88-6. P. Deuflhard; M. Wulkow. *Computational Treatment of Polyreaction Kinetics by Orthogonal Polynomials of a Discrete Variable.* (vergriffen)

SC 88-7. H. Melenk; H. M. Möller; W. Neun. *Symbolic Solution of Large Stationary Chemical Kinetics Problems.*

SC 88-8. Ronald H. W. Hoppe; Ralf Kornhuber. *Multi-Grid Solution of Two Coupled Stefan Equations Arising in Induction Heatiing of Large Steel Slabs.*

SC 88-9. Ralf Kornhuber; Rainer Roitzsch. *Adaptive Finite-Element-Methoden für konvektionsdominierte Randwertprobleme bei partiellen Differentialgleichungen.*

SC 88-10. C. -N. Chow; B. Deng; B. Fiedler. *Homoclinic Bifurcation at Resonant Eigenvalues.*

Veröffentlichungen des Konrad-Zuse-Zentrums für Informationstechnik Berlin
Technical Reports Dezember 1988

TR 86-1. H.J. Schuster. *Tätigkeitsbericht 1985.* (vergriffen)

TR 87-1. Hubert Busch; Uwe Pöhle; Wolfgang Stech. *CRAY-Handbuch. - Einführung in die Benutzung der CRAY..*

TR 87-2. Herbert Melenk; Winfried Neun. *Portable Standard LISP Implementation for CRAY X-MP Computers. Release of PSL 3.4 for COS.*

TR 87-3. Herbert Melenk; Winfried Neun. *Portable Common LISP Subset Implementation for CRAY X-MP Computers.*

TR 87-4. Herbert Melenk; Winfried Neun. *REDUCE Installation Guide for CRAY 1 / X-MP Systems Running COS Version 3.2.*

TR 87-5. Herbert Melenk; Winfried Neun. *REDUCE Users Guide for the CRAY 1 / X-MP Series Running COS. Version 3.2.*

TR 87-6. Rainer Buhtz; Jens Langendorf; Olaf Paetsch; Danuta Anna Buhtz. *ZUGRIFF - Eine vereinheitlichte Datenspezifikation für graphische Darstellungen und ihre graphische Aufbereitung.*

TR 87-7. J. Langendorf; O. Paetsch. *GRAZIL (Graphical ZIB Language).*

TR 88-1. Rainer Buhtz; Danuta Anna Buhtz. *TDLG 3.1 - Ein interaktives Programm zur Darstellung dreidimensionaler Modelle auf Rastergraphikgeräten.*

TR 88-2. Herbert Melenk; Winfried Neun. *REDUCE User's Guide for the CRAY 1 / CRAY X-MP Series Running UNICOS. Version 3.3.*

TR 88-3. Herbert Melenk; Winfried Neun. *REDUCE Installation Guide for CRAY 1 / X-MP Systems Running UNICOS. Version 3.3.*

TR 88-4. Danuta Anna Buhtz; Jens Langendorf; OLaf Paetsch. *GRAZIL-3D. Ein graphisches Anwendungsprogramm zur Darstellung von Kurven- und Funktionsverläufen im räumlichen Koordinatensystem.*

TR 88-5. Gerhard Maierhöfer; Georg Skorobohatyj. *Ein paralleler, adaptiver Algorithmus zur numerischen Integration ; seine Implementierung für SUPRENUM-artige Architekturen mit SUSI.*

TR 89-1. *CRAY-HANDBUCH. Einführung in die Benutzung der CRAY X-MP unter UNICOS.*