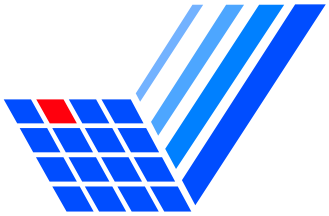
Hochauflösende FEM-FCT-Verfahren zur Diskretisierung von konvektionsdominanten Transportproblemen mit Anwendung auf die kompressiblen Eulergleichungen

– Diplomarbeit –

dem Fachbereich Mathematik
der Universität Dortmund vorgelegt von

Matthias Möller



Universität Dortmund,
Institut für Angewandte Mathematik und Numerik

Hiermit versichere ich, daß ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfaßt habe.

Dortmund, den

Einleitung

Auch in der Zeit von moderner CFD-Software stehen Forschung und Industrie bei der numerischen Simulation von kompressiblen sowie inkompressiblen Strömungen vor der bisher noch nicht zufriedenstellend gelösten Herausforderung, die in vielen Anwendungen auftretenden konvektionsdominanten Transportprobleme in adequater Weise zu behandeln. Dem Wunsch von hoher Genauigkeitsordnung und präziser Auflösung von Unstetigkeiten steht die physikalisch begründete Notwendigkeit gegenüber, keine positivitätsverletzenden Oszillationen entstehen zu lassen. Während klassische Upwind-Diskretisierungen in erschreckendem Maße diffusiv – dafür jedoch positiv – sind, macht das Godunovsche Theorem alle Hoffnung auf eine lineare, die Monotonie erhaltende Methode von höherer als erster Ordnung zunichte. Es bleibt der einzige Ausweg, nichtlineare künstliche Diffusion in Abhängigkeit von der lokalen Glattheit zu addieren, um eine gute Auflösung von Unstetigkeiten zu erzielen ohne jedoch wichtige Eigenschaften der exakten Lösung wie Positivität oder Monotonie zu opfern. Die direkte Implementierung dieser Idee führt auf die traditionellen *shock-capturing*-Verfahren, wobei man bei der Wahl des eingeführten freien Parameters, der den Anteil an künstlicher Viskosität bestimmt, stets zwischen Genauigkeit und Positivität schwankt und keines in der Regel von beiden garantieren kann.

Das von Boris und Book um 1970 entwickelte Konzept von *flux-corrected-transport* [4] bildet die Grundlage einer ganzen Generation von nichtlinearen Diskretisierungstechniken und beeinflußt die Erforschung von modernen hochauflösenden Verfahren, die auf dem Einsatz von *flux/slope* Limitern basieren. Allen gemeinsam ist das Prinzip, in Bereichen mit glattem Lösungsverlauf eine Diskretisierung hoher Ordnung anzuwenden und diese in der Nähe von Unstetigkeitsstellen schrittweise mit einer Methode niedriger Ordnung zu überblenden. In den 30 Jahren von FCT hat sich der ursprünglich eindimensionale Finite Differenzen Algorithmus von Boris und Book – SHASTA – durch die Einführung des für weitgehend beliebige Kombinationen aus Diskretisierungen hoher und niedriger Ordnung anwendbaren Zalesak-Limiters [79] zu einem mehrdimensionalen Verfahren weiterentwickelt. Den Sprung in die Welt der Finiten Elemente schaffte

das FCT-Konzept durch den Beitrag von Parrott und Christie [58], doch erst die beeindruckenden Leistungen von Löhner *et al.* [51], [52] machten FEM-FCT zu einer ausgereiften Methode für unstrukturierte Gitter. Die klassische Formulierung baut auf Zalesaks Limitingstrategie auf, wobei die dortigen (FD-)Flüsse durch antidiffusive Elementbeiträge ersetzt werden, so daß der Massenaustausch nicht nur entlang einer Kante des verwendeten Gitters sondern zwischen allen Knoten eines Elements stattfindet. Die entstehende engere Kopplung der Knotenwerte untereinander macht die Anwendung einer seit Boris und Book bekannten Prelimitingtechnik zur Verbesserung der Eckenauflösung unmöglich. Auch kann die Wahl eines zu kleinen Koeffizienten für die konstante Massendiffusion zur Entstehung kleiner Oszillationen führen. Nicht zuletzt ist die ursprüngliche Formulierung nur auf explizite Zeitdiskretisierungen anwendbar, welche einer restriktiven CFL-artigen Bedingung unterliegen, so daß FEM-FCT für die Behandlung von stationären Strömungen aus Effizienzgründen nicht ganz so attraktiv erscheint. Trotz dieser Einschränkungen haben Löhner *et al.* [50], [49] spektakuläre Simulationsergebnisse für eine Reihe von äußerst komplexen Problemstellungen erzielt.

Im Kontext von Finiten Differenzen oder Finiten Volumen existiert eine Reihe von robusten Verfahren, die auf *flux/slope* Limitern basieren. Teilweise sind sie jedoch auf eindimensionale Probleme oder Cartesische Gitter beschränkt, so daß ihre Übertragung auf Finite Elemente, und damit auf unstrukturierte Gitter, eine große Herausforderung darstellt. Die von Peraire *et al.* [60] vorgestellte kantenbasierte Datenstruktur ermöglicht nicht nur die Übertragung der eindimensionalen Theorie auf unstrukturierte Gitter, sondern bringt auch eine Verbesserung der Laufzeit und eine Verringerung der Speicherplatzanforderung mit sich, da sie die Kosten der indirekten Adressierung reduziert. Viele populäre Verfahren, die auf *flux difference* und *flux vector splitting* basieren, können dadurch auf Finite Elemente verallgemeinert werden. Die Ortsgenauigkeit läßt sich durch die Rekonstruktion eines eindimensionalen *stencils* für jede Kante durch Einfügen sogenannter *Dummyknoten* weiter verbessern, was die Herleitung einer Reihe von TVD-artigen Verfahren [53] ermöglichte. Leider ist die der kantenbasierten Datenstruktur [60] zugrunde liegende konservative Zerlegung der Galerkinintegrale in Flüsse nur für Simplexelemente mit linearen Basisfunktionen möglich, so daß die Verwendung von multilinearen Finiten Elementen oder solchen von höherer Ordnung ausgeschlossen ist. Gleichzeitig bleibt den auf elementorientierten Routinen basierenden Codes der Einsatz moderner flussbasierter Verfahren verschlossen.

Beide Verfahrensrichtungen scheinen in ihrer Weiterentwicklung zu stagnieren. Den Weg in die Welt der Finiten Elemente haben sie sich mit der Einführung einer sehr speziellen Datenstruktur oder von unhandlichen antidiffusiven Elementbeiträgen erkauft, was sie zu – durchaus sehr wertvollen – *special purpose* Werkzeugen macht. Um eine in sich geschlossene Theorie entwickeln zu können, verlangen die folgenden Fragen nach einer Antwort:

- Wie läßt sich eine FE-Methode in eine konservative Form überführen?
- Wie läßt sich Upwinding im Kontext von Finiten Elementen durchführen?
- Wie lassen sich nichtlineare Quellterme diskretisieren?
- Wie lassen sich nichtlineare Probleme mit impliziten Methoden behandeln?
- Wie läßt sich die Explizitheit und Zeitschrittabhängigkeit des Zalesak-Limiters aufheben?
- Wie läßt sich die skalare Theorie auf hyperbolische Gleichungssysteme wie die kompressiblen Eulergleichungen verallgemeinern?

Aufbau

In einer Serie von Veröffentlichungen [37], [38], [39], [40] konnte die klassische FEM-FCT Formulierung auf implizite Zeitdiskretisierungen erweitert und ein theoretisches Fundament zur Konstruktion von positivitätserhaltenden Methoden geschaffen werden. In den beiden letztgenannten Referenzen haben wir am Beispiel der kompressiblen Eulergleichungen demonstriert, wie sich die skalare Theorie auf hyperbolische Gleichungssysteme verallgemeinern läßt. Dieser Übertragungsprozess stützt sich auf mathematisch fundierte Grundlagen und wird in [40] fortgeführt. Mit der vorliegenden Arbeit möchten wir diese Entwicklung – angefangen bei der skalaren Theorie bis hin zu den aktuellen Forschungsergebnissen für hyperbolische Systeme – in einer zusammenhängenden Form präsentieren.

Im **ersten Kapitel** werden wir mit der Analyse des klassischen FEM-FCT Algorithmus eine Basis zur Weiterentwicklung des FCT-Paradigmas schaffen und das Prinzip des verwendeten Zalesak-Limiters beleuchten. Im Anschluß werden wir demonstrieren, daß die Galerkin Methode eine konservative Zerlegung in antisymmetrische Flüsse zwischen einzelnen Knoten erlaubt. Diese Flüsse lassen sich nur für P_1-Elemente mir den 'physikalischen' Kanten des Gitters assoziieren und stellen im allgemeinen 'virtuelle' Verbindungen zwischen Knoten, deren Basisfunktionen sich überlappen, dar. In [41] wird aus der Galerkin Flußzerlegung eine Alternative zu der auf Simplexelemente beschränkten kantenbasierten Datenstruktur von Peraire *et al.* [60] hergeleitet. Mit ihrer Hilfe läßt sich die auf einer kantenweisen (Re-)Konstruktion basierende FEM-TVD Methodik von Lyra *et al.* [53] auf allgemeine Finite Elemente Diskretisierungen erweitern. Gleichzeitig wird in [41] ein neuer mehrdimensionaler TVD-Zugang vorgestellt. Für FEM-FCT ist der generelle Übergang zu einer efizienten kantenbasierten Datenstruktur möglich. Es reicht jedoch aus, nur die diffusiven Terme als Flüsse darzustellen, so daß sich FCT einfach in bestehende Codes integrieren läßt.

Einen wichtigen Abschnitt des ersten Kapitels wird die Herleitung von notwendigen Positivitätskriterien und die Definition von diskreten Diffusionsoperatoren ausmachen. Mit ihrer Hilfe werden wir eine positivitätserhaltende Methode niedriger Ordnung für skalare Konvektions-Diffusions-Reaktionsgleichungen konstruieren, welche auf mathematisch fundierten Kriterien beruht. Dieses diskrete Upwinding für Finite Elemente wird durch die Elimination von negativen Nebendiagonaleinträgen aus dem diskreten Transportoperator hoher Ordnung bewerkstelligt, wobei diese Modifikation die M-Matrixeigenschaft des resultierenden Operators garantiert und auf eine *local extremum diminishing* Diskretisierung führt. Im Unterschied zu anderen Upwind-artigen Verfahren arbeitet *discrete Upwinding* unabhängig von der zugrunde liegenden Gittertopologie und den verwendeten Elementtypen ausschließlich durch Modifikation des Transportoperators auf diskreter Ebene. Weiterhin werden wir eine Form der Quelltermlinearisierung vorstellen, welche mit den Positivitätskriterien konsistent ist.

Um in Regionen mit glatten Lösungsprofilen die gute Ortsgenauigkeit der Diskretisierung hoher Ordnung zurückzugewinnen, wird die eingefügte künstliche Dissipation durch die Anwendung von nichtlinearer Antidiffusion entfernt. Dazu werden wir eine flußbasierte FEM-FCT Formulierung vorstellen, die den Anteil an zu addierender Antidiffusion mit Hilfe des Zalesak-Limiters bestimmt, für welchen wir einen mathematischen Positivitätsbeweis führen werden. Zum Verständnis der mitunter am Rand auftretenden Oszillationen, die sich ins Innere des Rechengebiets fortpflanzen, werden wir ein sogenanntes Hebelmodell einführen und einen leicht zu implementierenden Postlimitingschritt zu ihrer Elimination vorstellen. Das Hebelmodell kann ebenfalls zur Erklärung der durch einen Prelimitingschritt zu beobachtenden 'kosmetischen' Verbesserungen herangezogen werden, welcher durch die Rückkehr zu einer flußbasierten Repräsentation auch im Kontext von Finiten Elementen eingesetzt werden kann.

Die vorgestellte FEM-FCT Formulierung ist für beliebige Zeitdiskretisierungen anwendbar, wobei unser Schwerpunkt auf den impliziten Verfahren liegt, die aufgrund ihrer uneingeschränkten Stabilität/Positivität den Einsatz von großen Zeitschrittweiten erlauben. Die auftretenden Nichtlinearitäten werden mit Hilfe einer iterativen Fixpunkt-Defektkorrektur behandelt, wobei der diskrete Transportoperator niedriger Ordnung als Vorkonditionierer eingesetzt wird. Dabei werden wir den Basis Limiter als zeitschrittabhängig entlarven, so daß das Potential der uneingeschränkten Stabilität impliziter Methoden nicht vollständig ausgenutzt werden kann. Um diese Einschränkung zu überwinden, werden wir eine iterative Limitingstrategie vorstellen, welche die bereits akzeptierte Antidiffusion in jedem Schritt der Defektkorrektur bei der Berechnung der positivitätserhaltenden Zwischenlösung berücksichtigt und Zalesaks Limiter auf die Flußdifferenz zwischen aktueller und bereits akzeptierter Antidiffusion anwendet.

Im **zweiten Kapitel** werden wir die entwickelten Werkzeuge und Methoden anhand einer Reihe von skalaren Testproblemen analysieren und ihre Performanz sowohl für lineare als auch für nichtlineare Testfälle demonstrieren. Neben transienten Problemen in ein und zwei Dimensionen werden wir stationäre Transportprobleme untersuchen, für die die Vorteile der iterativen FEM-FCT Formulierung deutlich erkennbar werden.

Im zweiten Teil dieser Arbeit möchten wir die für skalare Transportprobleme entwickelten Lösungsmethoden auf Systeme von hyperbolischen Gleichungen am Beispiel der kompressiblen Eulergleichungen verallgemeinern. Dazu werden im **dritten Kapitel** die Erhaltungsgleichungen der Kontinuumsmechanik vorgestellt und die Konstitutivgesetze der Gasdynamik in der für das weitere Vorgehen ausreichenden Detailtiefe betrachtet. Anschließend wird die für numerische Verfahren übliche quasi-lineare Form der Eulergleichungen hergeleitet. Den Abschluß dieses einführenden Kapitels bildet der Übergang zu einer nichtkonservativen Formulierung in den primitiven Variablen, welche auf den für die Randbehandlung wichtigen Begriff der Riemanninvarianten führen.

Das **vierte Kapitel** ist vollständig der Erweiterung von FEM-FCT auf hyperbolische Systeme gewidmet. Zu Beginn stellen wir eine kantenbasierte Aufbauroutine der globalen Jacobimatrizen vor, welche auf dem Konzept von lokalen Roe Matrizen für jede 'numerische' Kante basiert. Diese kantenweise Assemblierung führt zu einer Minimierung des Rechenaufwandes und bringt im weiteren Vorgehen erhebliche Einsparungen des Speicherplatzes mit sich.

Wie im skalaren Fall ist die im FCT-Algorithmus verwendete Diskretisierung niedriger Ordnung eine wichtige 'Zutat', so daß der Verallgemeinerung von *discrete Upwinding* auf hyperbolische Systeme eine große Bedeutung zukommt. Zunächst werden wir den klassischen Ansatz von Godunov vorstellen, der auf das exakte Lösen von lokalen Riemann Problemen hinausläuft, um anschließend die Idee hinter Roes approximativem Riemann Löser zu erläutern. Für den Einsatz innerhalb eines FCT-Algorithmus wird sich die Verwendung von skalarer Dissipation, deren Wert proportional zum Spektralradius der Roematrix gewählt wird, als 'beste' Wahl herausstellen, da sie effizient zu implementieren ist und für die Erhaltung der Positivität ausreicht. Die geringere Diffusivität in Roes Methode, die mit erheblichem Mehraufwand verbunden ist, wird in einigen Fällen sogar 'schädlich' für das Endresultat sein.

Anschließend werden wir eine auf Systeme verallgemeinerte Variante des FEM-FCT Algorithmus vorstellen, die auf einer iterativen Defektkorrektur zur Behandlung der Nichtlinearitäten aufbaut. Für transiente Probleme empfiehlt sich weiter ein entkoppleter Löseransatz. Dies bewerkstelligt ein blockdiagonaler Vorkonditionirer, so daß das zu lösende Problem in eine Sequenz von skalaren Teilproble-

men für die einzelnen Variablen zerfällt. Zur Konstruktion des Vorkonditionierers greifen wir auf die Diagonalblöcke des diskreten Operators niedriger Ordnung zurück, so daß die zu lösenden linearen Probleme besser konditioniert sind.

Im Anschluß daran werden wir eine synchronisierte Version des Zalesak-Limiters für Systeme vorstellen, die für einen beliebigen (nichtkonservativen) Variablensatz ihre Gültigkeit behält. Zuletzt möchten wir auf eine Technik zur Implementierung von Randbedingungen innerhalb eines Finite Elemente Codes eingehen. Wir werden ein semi-implizites Vorgehen vorschlagen, welches die Werte in den Randknoten durch Übergang zu den Riemann Invarianten berechnet und ohne eine *ad hoc* Extrapolation von Informationen aus dem Inneren des Rechengebiets auskommt. Diese Prinzip ist auch für *free slip* Randbedingungen anwendbar, so daß eine einheitliche Behandlung sämtlicher Randbedingungen möglich ist, kann dort jedoch ohne den Umweg über die Riemann Invarianten verkürzt werden.

Die numerischen Ergebnisse für ausgesuchte Benchmarkkonfigurationen im **fünften Kapitel** geben einen Überblick über die beeindruckende Leistungsfähigkeit der neuen FEM-FCT Methoden. Beginnend mit transienten Problemen wie das von Sod vorgeschlagene Shock Tube Problem werden wir das Potential von impliziten Diskretisierungen für stationäre Testfälle anhand einer Reihe von supersonischen Strömungen durch Kanäle mit keilförmigen Hindernissen demonstrieren.

Danksagung

Mein Dank geht an Stefan Turek, der mich auf den Weg der Numerik geführt und einen wesentlichen Teil meines Studiums geprägt hat. Ganz besonders möchte ich Dmitri Kuzmin danken, der mir ein Mentor im Bereich von modernen hochauflösenden Diskretisierungsverfahren war. Kuzmin und Turek haben mein heutiges Verständnis von moderner Numerik wesentlich geformt.

Ich möchte ebenfalls den Mitarbeitern des Lehrstuhls für angewandte Mathematik in Dortmund danken, die einige Anregungen für meine Arbeit geliefert haben.

Schließlich möchte ich meinen Eltern danken, die mein Studium ermöglicht und mit viel Rücksicht gefördert haben.

Dortmund, im Juli 2003

Matthias Möller

Inhaltsverzeichnis

Notation

Notation für PDE's:

Ω	Gebiet $\Omega \subset \mathbb{R}^d$ mit Ortsdimension $d = 1, 2, 3$
Γ	Rand des Gebietes Ω
u	skalare Größe $u(x, y, z, t)$
$\mathbf{v}$, U	vektorwertige Größe $\mathbf{v}(x, y, z, t)$, $U = [u_1, u_2]^T$
$d\Delta u$	Diffusionsoperator $d\frac{\partial^2 u}{\partial^2 x} + d\frac{\partial^2 u}{\partial^2 y} + d\frac{\partial^2 u}{\partial^2 z}$
∇u	Gradientoperator $(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z})$
$\nabla \cdot \mathbf{v}$	Divergenzoperator $\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$
$\mathbf{v} \cdot \nabla u$	Transportoperator $v_1\frac{\partial u}{\partial x} + v_2\frac{\partial u}{\partial y} + v_3\frac{\partial u}{\partial z}$

Notation für Diskretisierungen und Finite Elemente:

h	Gitterweitenparameter
Δt	Zeitschrittparameter
θ	Implizitheitsparameter
LW	Lax-Wendroff
CN	Crank-Nicolson
BE	Backward-Euler

Notation für Matrizen und Numerische Lineare Algebra:

M_C	konsistente Massenmatrix
M_L	'gelumpte' Massenmatrix
K	diskreter Transportoperator hoher Ordnung
L	diskreter Transportoperator niedriger Ordnung
D	diskreter Diffusionsoperator aus $D(N \times N; \mathbb{R})$

Kapitel 1

Die skalare Theorie

1.1 FEM für skalare Erhaltungsgleichungen

Wir betrachten im folgenden eine abstrakte zeitabhängige Erhaltungsgleichung für eine skalare Größe u

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f} = q \qquad \text{in } \Omega. \tag{1.1}$$

Hierbei bezeichne q einen Quellterm und $\mathbf{f}$ einen möglicherweise nichtlinear von der Lösung abhängenden Flußvektor. Üblicherweise kann letzterer in einen konvektiven und einen diffusiven Anteil aufgespalten werden, so daß sich die Form

$$\mathbf{f} = \mathbf{v}u - d\nabla u \tag{1.2}$$

ergibt. Der Term $\mathbf{v}u$ bezeichnet den konvektiven Transport, der durch das extern vorgeschriebene Geschwindigkeitsfeld $\mathbf{v}$ bewirkt wird. Der diffusive Transport, der aufgrund von Konzentrationsgefällen in der Erhaltungsgröße (z.B. Masse oder Wärme) stattfindet, wird durch den Term $-d\nabla u$ wiedergegeben. Für große Diffusionskoeffizienten d liegt ein parabolisches Problem vor, welches im Bereich von FEM numerisch weitgehend handhabbar geworden ist. Interessanter und bisher noch nicht zufriedenstellend geklärt ist die Situation für konvektionsdominante Problemstellungen. Wenn d im Vergleich zu $\mathbf{v}$ klein ist, tritt der hyperbolische

Charakter der Gleichung (1.1) zum Vorschein und macht die numerische Behandlung ausgesprochen schwierig. In beiden Fällen muß das zu lösende Problem durch kompatible Anfangs- und Randbedingungen vervollständigt werden. Zur Klassifizierung von PDEs und zur Aufstellung geeigneter Randdaten vergleiche man die einführende Literatur zur Theorie partieller Differentialgleichungen [62].

Wenn wir mit V einen geeigneten Hilbertraum bezeichnen, so lautet die variationelle Formulierung der Gleichung (1.1) wie folgt

$$\int_\Omega w \left[\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f} - q \right] \, \mathrm{d}\mathbf{x} = 0, \qquad \forall w \in V. \tag{1.3}$$

Eine Standardwahl für die Ortsdiskretisierung ist die Galerkin Methode. Dabei werden für die auftretenden Variablen endlichdimensionale Approximationsräume V_h gewählt. Am Beispiel der Crouzeix-Raviart [10] oder Rannacher-Turek [63] Elemente erkennt man, daß es in einigen Fällen durchaus sinnvoll sein kann, auf die Konformität $V_h \subseteq V$ zu verzichten. Letztere eignen sich insbesondere für die Diskretisierung von inkompressiblen Strömungen. Das Tupel $\tilde{Q}_1/Q_0$, also die Wahl von rotiert bilinearen Elementen für die Geschwindigkeit und konstanten Ansatzfunktionen für den Druck, erfüllt die Babuška-Brezzi Bedingung ohne zusätzliche Stabilisierung und wird mit Erfolg bei der numerischen Behandlung der inkompressiblen Navier-Stokes Gleichungen eingesetzt [73].

Für die gängigen Basisfunktionen gilt die Beziehung $\sum_i \varphi_i \equiv 1$. Aufsummieren aller Gleichungen und Anwenden des Satzes von Gauß liefert aus (1.3) die integrale Form der Erhaltungsgleichung

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega u \, \mathrm{d}\mathbf{x} = \int_\Omega q \, \mathrm{d}\mathbf{x} - \int_\Gamma \mathbf{f} \cdot \mathbf{n} \, \mathrm{d}s, \tag{1.4}$$

wobei $\mathbf{n}$ den äußeren Normaleneinheitsvektor bezeichnet. Aus der obigen Beziehung geht hervor, daß sich der Wert von u in Ω ausschließlich aufgrund von Flüssen durch den Rand und von internen Quell- und Senkbeiträgen ändert, so daß die Finite Elemente Galerkin Diskretisierung konservativ im Integralsinn ist. Im Vergleich zur differentiellen Formulierung läßt (1.4) auch unstetige Lösungen zu. Die Erhaltung der Masse ist eine wichtige Eigenschaft. Falls ein konsistentes numerisches Verfahren gegen eine Funktion f konvergiert, so garantiert diese Eigenschaft, daß f eine schwache Lösung ist. Für nichtlineare hyperbolische Systeme ist die Eindeutigkeit der schwachen Lösung nicht mehr gesichert, so daß auf zusätzliche Mechanismen zurückgegriffen werden muß, um die physikalisch relevante Entropielösung zu berechnen [46]. Finite Volumen und *discontinuous Galerkin* Verfahren wenden (1.4) direkt auf jedes einzelne Element an, was die lokale Massenerhaltung garantiert. Während der Einsatz von *flux correction* für unstetige Ansätze wenig Schwierigkeiten bereitet, stellt der Einbau von Limiter-Techniken in herkömmliche FEM eine große Herausforderung dar.

1.2 Standard FEM-FCT

Betrachten wir die allgemeine zeitabhängige Erhaltungsgleichung (1.1). Wir interessieren uns insbesondere für die Situation $d \ll 1$, also den konvektionsdominanten Fall. Für diesen gibt es heute noch keine völlig zufriedenstellende Antwort auf die Frage nach der numerischen Behandlung. Viele der bisherigen Verfahren leiden unter übermäßiger numerischer Diffusion und führen somit zu unbrauchbaren Resultaten. Das andere Extrem bilden lineare Verfahren höherer Ordnung, die zwar für glatte Lösungen eine hohe Ortsgenauigkeit aufweisen, jedoch zu unphysikalischen Oszillationen an Unstetigkeitsstellen des Lösungsverlaufes neigen. Insbesondere führt die Bildung von Über- und Unterschwingern zu – aus physikalischer Sicht beurteilt – unakzeptablen Ergebnissen. Spätestens seit 1959 ist mit dem Satz von Godunov bekannt, daß keine lineare Methode der Ordnung zwei und höher die Monotonie der numerischen Lösung erhält [20]. Ein Ausweg aus dieser Zwickmühle zwischen hoher Genauigkeit einerseits und Monotonieerhaltung andererseits besteht in der nichtlinearen Kombination von diffusiven Methoden erster Ordnung in Gebieten, in denen der Lösungsverlauf steile Gradienten aufweist, und Methoden höherer Ordnungen in glatten Lösungsbereichen.

Diese Idee wurde erstmals 1973 von Boris und Book [4] als das Konzept von *flux-corrected transport* (FCT) vorgestellt. Der ursprüngliche FCT Algorithmus der beiden Autoren trug den Namen SHASTA (*Sharp And Smooth Transport Algorithm*) und war eine sehr spezialisierte Implementierung eines eindimensionalen Finite Differenzen Schemas. Seine erste weitreichende Verbesserung erfuhr das neuartige FCT Konzept durch Zalesak [79]. Er stellte 1979 eine mehrdimensionale Verallgemeinerung des FCT Paradigmas vor, welche eine nahezu beliebige Kombination von Methoden hoher und niedriger Ordnung zuläßt. Zalesak präsentierte seinen Limiter für eine Finite Differenzen Diskretisierung, so daß dieser zunächst nur auf Tensorproduktgittern zum Einsatz kam.

Diese Barriere wurde erstmals von Parrott und Christie [58] durchbrochen, die eine FCT Formulierung im Kontext der Finiten Elemente veröffentlichten. Seine einstweilige Reife erfuhr FCT jedoch in der zweiten Hälfte der 1980er Jahre durch die Arbeiten von Löhner *et al.* [51],[52]. Ihnen gelang es, einen sehr effizienten expliziten FEM-FCT Code zu entwickeln und bis dahin unerreichte Simulationsergebnisse zu präsentieren. Im folgenden wollen wir näher auf den von Löhner *et al.* vorgestellten FCT Algorithmus eingehen.

Der Prozess von *flux-correction* beginnt mit der Einführung einer starken künstlichen Diffusion in die Methode hoher Ordnung. Die Erkenntnisse von Godunov zeigen schnell, daß sich dadurch die Genauigkeit der Methode unausweichlich auf die Ordnung eins reduziert. Der entscheidende Trick bei FCT besteht nun dar-

in, eine kompensierende Menge an Antidiffusion zu addieren, um so den durch die künstliche Diffusion hervorgerufenen Fehler zu reduzieren. Dies erfolgt jedoch nur in Bereichen der Lösung, die eine genügende Glattheit aufweisen und in denen daher die Taylorentwicklung mit Termen höherer Ordnung sinnvoll ist. Dieser Prozess wird als *Limiting* bezeichnet und bildet den delikatesten Teil des gesamten FCT Algorithmus.

Nach diesen Vorüberlegungen zum grundsätzlichen Vorgehen können wir das Konzept von FCT in einen numerischen Algorithmus umformen. Der von Löhner *et al.* [51], [52] vorgeschlagene sieht folgendermaßen aus:

F.1 Diskretisiere die zugrunde liegende Gleichung mit einer expliziten FEM-Methode hoher Ordnung (z.B. Taylor-Galerkin):

$$M_C \Delta u^H = b, \quad \Delta u^H = u^H - u^n.$$

F.2 Führe *mass lumping* durch und addiere ausreichende künstliche Diffusion, um eine monotonieerhaltende Methode niedriger Ordnung zu konstruieren:

$$M_L \Delta u^L = b + c_d (M_C - M_L) u^n.$$

F.3 Berechne die antidiffusiven Elementbeiträge aus der Differenz zwischen den Diskretisierungen hoher und niedriger Ordnung:

$$F_e = M_L^{-1}\big|_e (\hat{M}_L - \hat{M}_C)(c_d \hat{u}^n + \Delta \hat{u}^H).$$

F.4 Begrenze die antidiffusiven Terme so, daß in u^{n+1} keine unphysikalischen Extrema auftreten:

$$F_e^* = \alpha_e F_e, \quad 0 \leq \alpha_e \leq 1.$$

F.5 Addiere die korrigierten Terme elementweise zur Lösung niedriger Ordnung:

$$u_i^{n+1} = u_i^L + \sum_e F_{e,i}^*.$$

Entscheidend für den Erfolg von FCT ist der 5. Schritt. Falls keine Antidiffusion eingeführt wird ($\alpha_e = 0$), erhält man die Methode niedriger Ordnung und damit eine diffusive Lösung. Die Wahl von $\alpha_e = 1$ bewirkt ein Umschalten auf die Methode hoher Ordnung und möglicherweise das Entstehen von numerischen Oszillationen. Das Ziel des Limiting-Schrittes besteht also darin, den Anteil an Antidiffusion in Abhängigkeit von der Glattheit der Lösung so groß wie möglich zu wählen, ohne dadurch unphysikalische Extrema entstehen zu lassen. Im folgenden wollen wir die einzelnen Schritte des obigen Algorithmus näher betrachten.

1.2.1 Diskretisierung hoher Ordnung

Die Diskretisierung der Gleichung (1.1) in Zeit und Ort mit einer expliziten Methode hoher Ordnung kann in folgender Form geschrieben werden

$$M_C \Delta u = b. \tag{1.5}$$

Hierbei bezeichnet M_C die konsistente Massenmatrix, $\Delta u = u^{n+1} - u^n$ den Differenzvektor zwischen neuer und alter Lösung und b den Lastvektor, der sich aus konvektiven und diffusiven Beträgen ausgewertet zum vorherigen Zeitschritt zusammensetzt. In den Arbeiten von Löhner *et al.* wird als Methode hoher Ordnung eine zweischrittige Taylor-Galerkin Diskretisierung vom Lax-Wendroff Typ vorgeschlagen, wobei prinzipiell jedes explizite Verfahren eingesetzt werden kann. Die Lösung des Problems (1.5) muß offensichtlich die folgende Gleichung erfüllen

$$M_L \Delta u^H = b + (M_L - M_C)\Delta u^H. \tag{1.6}$$

Hierbei weist das hochgestellte H auf die Methode hoher Ordnung hin, und M_L bezeichnet die durch konservatives *mass lumping* [21] erzeugte Diagonalmatrix. Der zweite Term in der rechten Seite von Gleichung (1.6) repräsentiert den antidiffusiven Beitrag, der implizit in der konsistenten Massenmatrix verborgen ist.

1.2.2 Diskretisierung niedriger Ordnung

Der zweite Teil des Algorithmus wird von einer monotonieerhaltenden Methode niedriger Ordnung ausgemacht. Ein prädestinierter Kandidat für diese Aufgabe wäre das Upwind-Verfahren. Löhner *et al.* [51] schlagen das folgende Vorgehen zur Konstruktion der Methode niedriger Ordnung vor. Zunächst wird die konsistente Massenmatrix durch ihr 'gelumptes' Pendant ersetzt. Anschließend wird explizit ein konstanter Anteil an Massendiffusion addiert, um die Methode hoher Ordnung in eine von niedriger Ordnung – gekennzeichnet mit L – zu überführen

$$M_L \Delta u^L = b + c_d(M_C - M_L)u^n. \tag{1.7}$$

Die Größe der Massendiffusion wird durch den konstanten Diffusionskoeffizienten c_d definiert. Die Wahl von $c_d = 1$, wie es in [13], [67] vorgeschlagen wird, führt zu

$$M_L u^L = M_C u^n + b, \tag{1.8}$$

was als Methode hoher Ordnung mit ausschließlich auf der rechten Seite durchgeführtem *mass lumping* interpretierbar ist. Für die eindimensionale Lax-Wendroff Methode erhält man mit diesem Vorgehen ein Schema, welches für Courant Zahlen $|\nu| \leq \sqrt{2/3}$ stabil und monoton ist. Diese Bedingung ist restriktiver als

die CFL-Bedingung für die klassische Upwind-Diskretisierung. Desweiteren sind keine Aussagen über das Verhalten der Lösung für allgemeinere Fälle bekannt.

Kommen wir zur Formulierung (1.6) zurück. Schon in den Anfängen von SHASTA wurde die Monotonieerhaltung durch die Addition einer *konstanten* Diffusion erzielt. Seitdem wurde diese Technik von vielen Autoren mit Erfolg eingesetzt. Dennoch bleibt zu beobachten, daß die Wahl eines zu kleinen/großen Diffusionskoeffizienten zu Problemen führen kann. Daher sollte der Wert von c_d und der verwendete Zeitschritt Δt sorgfältig gewählt werden, um die Entstehung von unphysikalischen Oszillationen zu vermeiden. Wir werden in dieser Arbeit ein neues Konstruktionsprinzip für die Methode niedriger Ordnung vorstellen, das sich auf eine fundierte Theorie stützt und keine empirischen Richtgrößen verwendet.

1.2.3 Antidiffusive Elementbeiträge

Die Differenzbildung zwischen den Methoden hoher und niedriger Ordnung löscht den unhandlichen Term b aus und läßt die folgende Darstellung der antidiffusiven Elementbeiträge zu

$$F_e = M_L^{-1}\big|_e \, (\hat{M}_L - \hat{M}_C)(c_d \hat{u}^n + \Delta \hat{u}^H), \tag{1.9}$$

wobei zu berücksichtigen ist, daß $\Delta u^H - \Delta u^L = u^H - u^L$ gilt. In der obigen Gleichung bezeichnet $\hat{M}_L - \hat{M}_C$ den lokalen Antidiffusionsoperator, der auf die Knotenwerte des jeweiligen Elements angewendet wird. Seine Konstruktion erfolgt mit Hilfe der elementweisen Massenmatrizen. Der resultierende Ergebnisvektor entspricht in seiner Länge der Anzahl der lokalen Freiheitsgrade des Elements. Im letzten Schritt erhält man den antidiffusiven Elementbeitrag durch Division durch den Diagonaleintrag der globalen Matrix M_L.

1.2.4 Zulässigkeitsbereich der Lösung

Wie in der Arbeit von Boris und Book [4] geometrisch motiviert, ist der zulässige Bereich der endgültigen Lösung durch die lokalen Extrema der Lösung niedriger Ordnung u^L und die der Lösung des vorhergehenden Zeitschritts u^n [79] festgelegt. Löhner *et al.* schlagen den folgenden dreischrittigen Algorithmus zur Bestimmung der lokalen Schranken $u^{\max}_{\min}$ vor.

S.1 Bestimme u^* als knotenweises Maximum/Minimum aus u^L und u^n:

$$u_i^* = \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} \{u_i^L, u_i^n\}.$$

S.2 Berechne für jedes Element das Maximum/Minimum seiner u^*-Werte:

$$u_e^{**} = \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} u_i^*, \quad i \in \mathsf{N}_e.$$

S.3 Bestimme für jeden Knoten das Maximum/Minimum der u^{**}-Werte aller Elemente, zu denen der Knoten gehört:

$$u_i^{\max\limits_{\min}} = \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} u_e^{**}, \qquad e \in \mathsf{E}_i.$$

Diese Wahl der Schranken für den neuen Funktionswert u_i^{n+1} garantiert, daß dieser durch die Extrema in der Umgebung des Knotens i beschränkt bleibt. Die Einbeziehung der alten Werte zusätzlich zu denen der aktuellen Lösung wurde erstmals von Zalesak zur Vermeidung des sogenannten *'peak clipping'* vorgeschlagen, das ein Problem des ursprünglichen SHASTA-Codes war. Dieses Phänomen, daß lokale Spitzen in der Lösung übermäßig ausgeglättet werden, läßt sich wie folgt erklären. Der Funktionswert in der Spitze wird von der Methode niedriger Ordnung nur unzureichend aufgelöst. Falls die Korrekturfaktoren für die kompensierende Antidiffusion jedoch ausschließlich aus u^L und nicht aus u^n berechnet werden, so wird der *exakte* Wert in der Spitze bereits als unphysikalisches Über-/Unterschwinger interpretiert und eliminiert.

Auch wenn das Einbeziehen der alten Lösung in die Bestimmung der Korrekturfaktoren in vielen Testfällen zu Verbesserungen führen kann, lassen sich Konfigurationen insbesondere mit nicht divergenzfreien Geschwindigkeitsfeldern angeben, bei denen dies zur Entstehung von kleinen Oszillationen führt. Im Kapitel 2 werden wir eine Problemstellung untersuchen, für welche die physikalischen Extrema mit der Zeit abklingen und so die Berücksichtigung der alten Funktionswerte unerwünschte Über-/Unterschwinger in der Lösung produziert. Allgemein ist es daher ratsam, zu den Ursprüngen von Boris und Book zurückzukehren und $u^* = u^L$ zu wählen. Als wichtigster Schritt des FCT Algorithmus folgt nun die Berechnung der Korrekturfaktoren α_e. Diese bestimmen den Anteil an kompensierender Antidiffusion, die zu der hohen Genauigkeit des Verfahrens führt.

1.3 Zalesak Limiter

Die ‘richtige’ Bestimmung der Korrekturfaktoren entscheidet maßgeblich über den Erfolg eines FCT Verfahrens und soll daher im folgenden genauer analysiert werden. Durch die Wahl von $0 \leq \alpha_e \leq 1$ kann zwischen den Methoden niedriger und hoher Ordnung über eine beliebige Kombination dazwischen variiert werden. Selbstverständlich wird man versuchen, die korrigierende Antidiffusion so groß wie möglich zu wählen, ohne dadurch jedoch unphysikalische Oszillationen entstehen zu lassen. Löhner *et al.* [51], [52] setzen in ihren Arbeiten zu diesem Zweck den mehrdimensionalen Limiter von Zalesak ein, den wir an dieser Stelle genauer betrachten wollen. Da wir im Verlauf dieser Arbeit mehrmals auf diese Komponente des FCT Algorithmus zurückgreifen werden, stellen wir ihn in einer allgemeingültigen Form vor und schränken ihn nicht auf die spezielle Notation von Löhner *et al.* ein. Dazu gehen wir von einer beliebigen monotonen Zwischenlösung $\tilde{u}$ aus. Im Falle des expliziten Algorithmus von Löhner *et al.* reduziert sich diese auf $\tilde{u} = u^L$. Weiterhin wollen wir mit f_{ij} einen Fluß vom Knoten j in den Knoten i bezeichnen, der im Algorithmus von Löhner *et al.* durch den Elementbeitrag $F_{e,i}$ des Elements e zum Knoten i ersetzt werden muß.

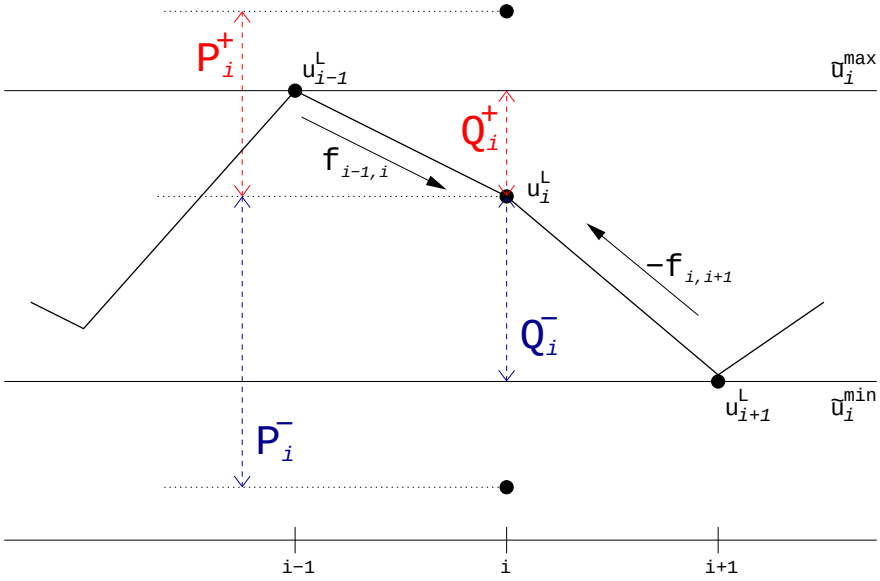


Abbildung 1.1: Zalesaks Limiter, explizit in 1D.

Die grundsätzliche Idee des Limiters von Zalesak ist für den eindimensionalen, expliziten Fall in Abbildung 1.1 dargestellt. Es bezeichne S_i die Indexmenge bestehend aus dem Knoten i und seinen Nachbarn. Unser Verständnis von Knotennachbarschaft geht hier über die graphentheoretische Definition über Kantenbeziehungen hinaus und faßt Knoten als Nachbarn auf, falls sich die Träger

ihrer Basisfunktionen überlappen. Weiter seien $\tilde{u}_i^{\max}$ und $\tilde{u}_i^{\min}$ das Maximum bzw. Minimum der Knotenwerte der den Knoten i umgebenden Elemente, d.h.

$$\tilde{u}_i^{\substack{\max \\ \min}} = \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} \tilde{u}_j, \qquad j \in \mathsf{S}_i. \tag{1.10}$$

Der Wert im Knoten i wird von allen antidiffusiven Flüssen, die von benachbarten Knoten j in den Knoten i hineinfließen, beeinflußt. Im schlimmsten Fall stimmen alle ihre Vorzeichen überein, so daß sie sich zu einem positiven oder negativen Beitrag akkumulieren und dadurch ein bestehendes Extremum weiter verstärken können. Wir wollen, der ursprünglichen Bezeichnungsweise von Zalesak folgend, die Summe aller positiven/negativen Beiträge zum Knoten i mit

$$P_i^{\pm} = \frac{1}{m_i} \sum_{j \neq i} \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} \{0, f_{ij}\} \tag{1.11}$$

bezeichnen. Für den Knoten i ist die größtmögliche Erhöhung/Verminderung durch den Abstand des Wertes der Zwischenlösung zu den lokalen Schranken

$$Q_i^{\pm} = \tilde{u}_i^{\substack{\max \\ \min}} - \tilde{u}_i \tag{1.12}$$

festgelegt. Die geometrische Interpretation der Größen $P_i^{\pm}$ und $Q_i^{\pm}$ ist in Abbildung 1.1 dargestellt. Offensichtlich kann das Zusammenwirken aller antidiffusiven Flüsse den Wert der Lösung an einem Knoten i über die zulässigen Schranken hinaustreiben und so zur Entstehung von unphysikalischen Über- und Unterschwingern führen. Der maximal zulässige Anteil an Antidiffusion kann durch

$$R_i^{\pm} = \begin{cases} \min\{1, Q_i^{\pm}/P_i^{\pm}\}, & \text{für } P_i^{\pm} \neq 0, \\ 1, & \text{für } P_i^{\pm} = 0 \end{cases} \tag{1.13}$$

beschrieben werden. Wir dürfen nicht vergessen, daß der Flußaustausch zwischen zwei Knoten auf einer 'bilateralen' Basis stattfindet. Ein positiver Fluß f_{ij} vom Knoten j in den Knoten i ist gleichzeitig ein negativer Fluß $f_{ji} = -f_{ij}$ vom Knoten i in den Knoten j und umgekehrt. Daher muß bei der Berechnung der Korrekturfaktoren α_{ij} das Vorzeichen des Flußes berücksichtigt werden und dementsprechend das Minimum über die zugehörigen Faktoren genommen werden

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\}, & \text{für } f_{ij} \geq 0, \\ \min\{R_j^+, R_i^-\}, & \text{für } f_{ij} < 0. \end{cases} \tag{1.14}$$

Eine solche Wahl von α_{ij} garantiert, daß der Wert der neuen Lösung für jeden Knoten innerhalb des zulässigen Bereichs $\tilde{u}_i^{\min} \leq \tilde{u}_i^{n+1} \leq \tilde{u}_i^{\max}$ liegt. Damit bleibt die neue Lösung positiv, solange die Zwischenlösung $\tilde{u}$ diese Eigenschaft besitzt.

Im Falle der von Löhner *et al.* benutzten antidiffusiven Elementbeiträge erhält man für den Beitrag zum Knoten i die abgewandelte Vorschrift

$$P_i^{\pm} = \sum_{e \in \mathsf{E}_i} \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} \{0, F_{e,i}\}. \tag{1.15}$$

Der durch die lokalen Schranken festgelegte Wert $Q_i^{\pm}$ wird wiederum wie in (1.12) bestimmt. Der endgültige Wert der Korrekturfaktoren ergibt sich als

$$\alpha_e = \min_{i \in \mathsf{N}_e} \begin{cases} R_i^+, & \text{für } F_{e,i} \geq 0, \\ R_i^-, & \text{für } F_{e,i} < 0, \end{cases} \tag{1.16}$$

womit gesichert ist, daß durch das Zusammenspiel aller antidiffusiven Elementbeiträge keine unphysikalischen Oszillationen hervorgerufen werden.

1.3.1 Prelimiting

Wir wollen an dieser Stelle noch auf eine Erweiterungsmöglichkeit des Limiting-Algorithmus eingehen. Insbesondere auf explizite FCT Schemata wirkt es sich günstig aus, wenn alle antidiffusiven Flüsse, die dem Gradienten von $\tilde{u}$ entgegen gerichtet sind, ausgelöscht werden

$$f_{ij} := 0 \qquad \text{für} \quad f_{ij}(\tilde{u}_i - \tilde{u}_j) \leq 0 \tag{1.17}$$

und zwar *vor* dem eigentlichen Limiting-Schritt. Kurz gesagt wird keinem antidiffusiven Fluß erlaubt, sich wie ein diffusiver zu verhalten. Dieser Trick wurde schon von Boris und Book [4] vorgeschlagen, geriet danach jedoch in Vergessenheit. Zalesak erwähnt diesen Ansatz zwar in seiner Arbeit [79], bezeichnet ihn jedoch als 'kosmetische' Spielerei und rät von seinem prinzipiellen Gebrauch ab. Seiner Auffassung nach diene der Großteil der Antidiffusion dazu, den Anstieg des Gradienten zu erhöhen, so daß die durch (1.17) bewirkten Verbesserungen nur marginal seien. Diese Einschätzung führte dazu, daß der sogenannte Prelimiting-Schritt in vielen Implementierungen fehlt. Eine Positivitätserhaltung ist damit zwar garantiert, nicht aber die Sicherstellung der Monotonie [11], [37], [38]. DeVore [11] hat ein Preprocessing der antidiffusiven Flüsse als Mittel zur 'Monotonisierung' erkannt und damit besonders für dynamische Strömungen Qualitätsgewinne erzielt.

Ein derartiges Prelimiting fehlt im FEM-FCT Algorithmus von Löhner *et al.* [51], was jedoch auch daran liegt, daß die auf Elementbeiträgen basierende Formulierung eine vorgeschaltete Modifikation der Form (1.17) in mehreren Dimensionen unmöglich macht. Erst die Rückkehr zu einer flußbasierten Formulierung macht die Anwendung dieses 'Kniffes' im Kontext der Finiten Elemente möglich. Wie sich anhand von numerischen Simulationen herausgestellt hat, ist ein solches Preprocessing der antidiffusiven Flüsse auch für implizite Schemata empfehlenswert.

1.3.2 Postlimiting

In [37] wurde bereits auf die Probleme von FEM-FCT für glatte Funktionen, deren Gradient am Rand nicht verschwindet, hingewiesen. Bei der Verwendung des Zalesak-Limiters etwa für die lineare Konvektionsgleichung mit $v = 1$ und $u(x, 0) = x$ entstehen am Rand kleine Oszillationen, die sich ins Innere des Rechengebiets ausbreiten (Abb. 1.2, links). Dieses Phänomen ist auch am Einflußrand zu beobachten, wird jedoch durch das Setzen von Dirichletwerte gemindert.

Abbildung 1.2 (Mitte) zeigt für einen kleinen Ausschnitt am Rand eine Serie von Lösungsprofilen zu den Zeitpunkten $t = t_0 + ih$, $i = 0, 1, \ldots$, durch die man die Oszillationen als regelmäßige 'Treppenstruktur' identifizieren kann, deren 'Knickstellen' gerade in den Gitterpunkten liegen.

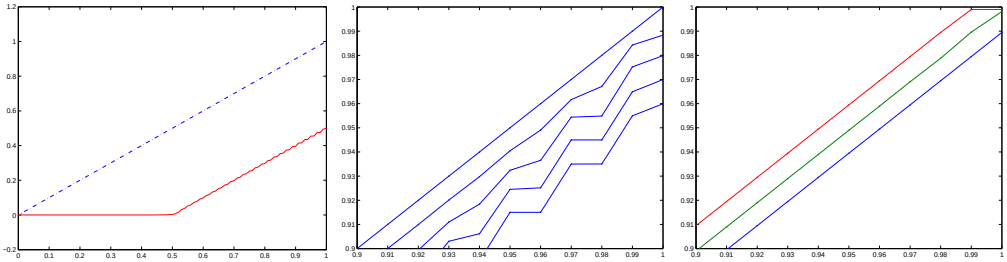


Abbildung 1.2: Pathologisches Verhalten von Zalesaks Limiter.

Das rechte Diagramm gibt die Situation nach *einem* Zeitschritt wieder, wobei die mittlere Linie die berechnete Lösung $\tilde{u}$ darstellt, die nach oben und unten durch die aus dem Zalesak-Limiter stammenden Werte $\tilde{u}^{\max}$ und $\tilde{u}^{\min}$ beschränkt ist. Der 'Knick' in der oberen Schranke wird dadurch hervorgerufen, daß der Limiter ohne das Wissen um den Funktionsverlauf über die Grenzen des Rechengebietes hinaus nicht mit Gewissheit den Funktionswert der Zwischenlösung $\tilde{u}$ am Rand als ein lokales Maximum detektieren kann und daher in diesem Knoten $u^{\max} = \tilde{u}$ setzt. Diese Fehlinterpretation eines Extremums führt dazu, daß der antidiffusive Fluß über die Kante vollständig begrenzt wird ($\alpha_{ij} = 0$) und somit die Endlösung im Randknoten den Wert der Methode niedriger Ordnung erhält.

Dem diskreten Massenerhaltungsprinzip folgend wird der im Randelement durch einen im Vergleich zur exakten Lösung zu kleinen Wert im Randknoten künstlich erzwungene Massenverlust durch eine Umverteilung der Masse ausgeglichen, indem der Funktionswert im linken Nachbarknoten angehoben wird (Abb. 1.3, links). Dadurch wird im angrenzenden Element eine erneute Massenänderung bewirkt, so daß sich das Prinzip von Anheben und Absenken von Knotenwerten sukzessive bis zum gegenüberliegenden Rand fortsetzt.

Wir führen an dieser Stelle das sogenannte *Hebelmodell* ein, das man sich als

viele an Drehpunkten befestigte und in ihrer Länge ‘variable’ Hebel vorstellen kann, die an ihren Berührungspunkten stetig miteinander verbunden sind (Abb. 1.3, rechts). Wenn man den rechten Hebel nach unten zieht, bewirkt dies eine wechselseitige Scherung der übrigen Hebel um ihre Drehpunkte.

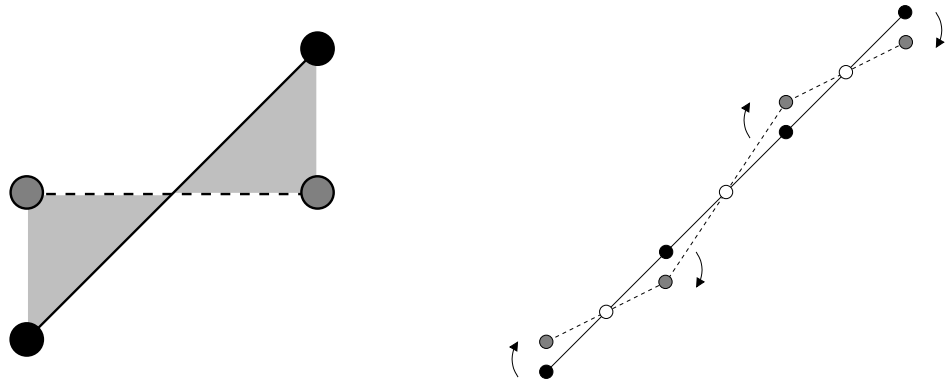


Abbildung 1.3: Massenausgleich und ‘Hebelmodell’.

Während sich der am Rand begangene Fehler im Modell sofort auf die Neigung aller ‘Hebel’ auswirkt, ist die Ausbreitung bei der numerischen Methode zeitabhängig, da Informationen in einem Zeitschritt nur über einen begrenzten Bereich transportiert werden können. Für jedes einzelne Element setzt sich der Gesamtfehler aus den von links konvektierten Fehlern aus vorhergehenden Zeitschritten und dem am rechten Rand aufgrund von falsch detektierten Extrema neu entstehenden Fehler zusammen. Diese Akkumulation von Fehlern terminiert erst, wenn die Lösung knotenweise mit den oberen bzw. unteren Schrankenwerten übereinstimmt (vgl. Abb. 1.4).

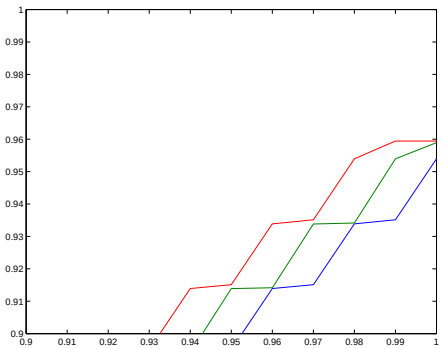


Abbildung 1.4: Lösungsverlauf am Rand.

Das oben eingeführte Hebelmodell erklärt ebenfalls die Wirkungsweise von Prelimiting für Lösungsprofile mit steilen Flanken (vgl. Abbildung 1.5). Ein diffusiver Fluß bewirkt, daß der Gradient geglättet wird, während ihn ein antidiffusiver Fluß steiler macht. Mit Hilfe des Hebelmodells läßt sich dieser Vorgang als eine

'Drehung' um den Kantenmittelpunkt interpretieren, bei der die Massenerhaltung garantiert wird. Ein diffusiver Fluß dreht die Kante gerade in die 'falsche' Richtung, so daß eine 'Ecke' im Lösungsprofil entsteht. Das vorgeschaltete Prelimiting verhindert dies, indem es antidiffusive Flüsse, die sich aufgrund des Vorzeichens wie diffusive verhalten, eliminiert.

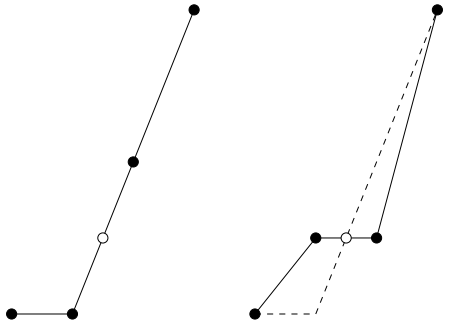


Abbildung 1.5: Prelimiting-Effekt.

Eine erste Möglichkeit, den Zalesak-Limiter zu modifizieren, besteht darin, die Lösung um einen Knoten über den Rand hinaus (linear) zu extrapolieren, so daß der Wert im Randknoten nicht mehr als Extremum interpretiert wird und alle Schranken korrekt berechnet werden. Dieses Vorgehen ist in mehreren Dimensionen aufwendig und führt zu längeren Rechenzeiten. Ein deutlich effizienteres Postlimiting erhält man durch Reinitialisierung der vom Algorithmus 'falsch' bestimmten Skalierungsfaktoren am Ein- und Ausflußrand

$$R_i^{\pm} := 1 \qquad \forall i \in \mathsf{N}_b, \tag{1.18}$$

wobei N_b gerade die Menge der entsprechenden Randknoten bezeichnet. Die damit berechneten Korrekturfaktoren α_{ij} berücksichtigen nur die für innere Knoten korrekt bestimmten Schranken. Trotz der korrekt gesetzen Dirichlet Werte ist ein solches Postprocessing, daß zwischen (1.13) und (1.14) durchzuführen ist, auch am Einflußrand ratsam. Da die Nebendiagonaleinträge des Operators hoher Ordnung für Knoten am Ausflußrand üblicherweise nichtnegativ sind [37], [41], stellt diese Modifikation keine Gefahr für die Positivitätserhaltung dar. Ferner würden etwaige Über- und Unterschwinger unverzüglich aus dem Rechengebiet konvektiert werden.

1.4 Galerkin Flusszerlegung

Ein numerisches Verfahren wird als konservativ bezeichnet, wenn sich die diskretisierten Terme in Form von Flüssen zwischen benachbarten Knoten darstellen lassen. Peraire *et al.* [60] haben als erste eine Flußzerlegung für Finite Elemente Galerkin Diskretisierungen vorgestellt, die jedoch auf Simplexelemente mit linearen Basisfunktionen und stückweise konstanten Gradienten beschränkt ist. Die dabei verwendete kantenbasierte Datenstruktur reduziert ferner die Rechenzeit und den Speicherbedarf. Für die recht mühsame Herleitung möchten wir auf die Monographien von Lyra [53] und Löhner [49] verweisen. Eine darüber hinausgehende Beschreibung von kantenbasierten Finite Elemente Methoden findet sich in dem Übersichtsartikel von Morgan und Peraire [55], der sich insbesondere mit hochauflösenden Verfahren für unstrukturierte Gitter befaßt.

Im folgenden möchten wir eine Technik zur konservativen Flußzerlegung vorstellen, die im Gegensatz zu derjenigen von Peraire *et al.* [60] auf beliebige Elementtypen und Ansatzfunktionen anwendbar ist und keiner Einschränkung bezüglich des Rechengitters unterliegt. Die Anwendung von partieller Integration in der schwachen Formulierung (1.3) zusammen mit dem Satz von Gauß liefert

$$\int_\Omega w\frac{\partial u}{\partial t}\,\mathrm{d}\mathbf{x} - \int_\Omega \nabla w\cdot\mathbf{f}\,\mathrm{d}\mathbf{x} + \int_\Gamma w\,\mathbf{f}\cdot\mathbf{n}\,\mathrm{d}s - \int_\Omega w\,q\,\mathrm{d}\mathbf{x} = 0, \qquad \forall w\in V. \quad (1.19)$$

Als gängige Praxis wird bei der Behandlung von kompressiblen Strömungen der gleiche Ansatzraum für die Flüsse und die Lösung gewählt. Dieses von Fletcher als *group finite element formulation* [15] bezeichnete Vorgehen ermöglicht einen effizienten Matrixaufbau, der zu bemerkenswerten Einsparungen in der Rechenzeit führt. Fletcher setzt diese Technik für die Diskretisierung der zweidimensionalen Burgers Gleichung ein und stellt einen Effizienzgewinn um den Faktor 2.5 im Vergleich zu konventionellen FEM fest. Dieser werde um so deutlicher, wenn man die Dimension, die Komplexität des verwendeten Elements oder die Konnektivität der Matrizen erhöhe oder nichtlineare Kopplungen hinzunehme. Desweiteren beobachtet Fletcher, daß die *group finite element formulation* für uniforme Gitter einen kleinen Genauigkeitsgewinn liefere.

Wir verwenden daher für die Approximation der Lösung, der Flüsse und der Quellterme die gleichen Ansatzfunktionen $\varphi\in V_h$

$$u=\sum_j u_j\varphi_j, \qquad \mathbf{f}=\sum_j \mathbf{f}_j\varphi_j, \qquad q=\sum_j q_j\varphi_j. \quad (1.20)$$

Einsetzen der Terme (1.20) in die schwache Formulierung (1.19) und Substitution der Gewichtsfunktionen durch ihr diskretes Pendant liefert

$$\sum_j \left[\int_\Omega \varphi_i\varphi_j \,\mathrm{d}\mathbf{x}\right](\dot{u}_j - q_j) - \sum_j \left[\int_\Omega \nabla\varphi_i\varphi_j \,\mathrm{d}\mathbf{x} - \int_\Gamma \varphi_i\varphi_j\mathbf{n}\,\mathrm{d}s\right]\cdot\mathbf{f}_j = 0. \quad (1.21)$$

Summieren wir das System von ODEs über i, so erhalten wir die integrale Darstellung der Erhaltungsgleichung, aus welcher die globale Erhaltungseigenschaft folgt. Für die Galerkin Methode ändert sich demnach der Wert der Größe u in Ω ausschließlich durch Randflüsse und aufgrund von Quell- und Senktermen im Inneren. Im folgenden werden wir für den internen Flußanteil

$$\sum_j \left[\int_\Omega \nabla\varphi_i\varphi_j \,\mathrm{d}\mathbf{x}\right]\cdot\mathbf{f}_j = \sum_j \mathbf{c}_{ji}\cdot\mathbf{f}_j, \qquad \text{mit} \quad \mathbf{c}_{ji} = \int_\Omega \nabla\varphi_i\varphi_j \,\mathrm{d}\mathbf{x} \quad (1.22)$$

eine Zerlegung in antisymmetrische numerische Flüsse herleiten. Die Vertauschung der Indizes in der Definiton der Koeffizienten $\mathbf{c}_{ji}$ weist auf die Anwendung von partieller Integration hin. Wie bereits erwähnt, gilt für die meisten Ansatzfunktionen die Beziehung $\sum_i \varphi_i \equiv 1$, so daß die Summe der Ableitungen verschwindet. Somit besitzt die Koeffizientenmatrix $\{\mathbf{c}_{ij}\}$ die Zeilensumme Null. Diese Eigenschaft ermöglicht es, die Diagonaleinträge als Summe über Nebendiagonaleinträge auszudrücken

$$\sum_j \mathbf{c}_{ij} = 0 \quad \Rightarrow \quad \mathbf{c}_{ii} = -\sum_{j\neq i} \mathbf{c}_{ij}. \quad (1.23)$$

Damit läßt sich der interne Anteil des Flußes konservativ umformen

$$\sum_j \mathbf{c}_{ji}\cdot\mathbf{f}_j = \sum_{j\neq i} g_{ij}, \qquad \text{mit} \quad g_{ij} = \mathbf{c}_{ji}\cdot\mathbf{f}_j - \mathbf{c}_{ij}\cdot\mathbf{f}_i. \quad (1.24)$$

Offensichtlich gilt für den neu eingeführten Ausdruck g_{ij}, der den Fluß vom Knoten j in den Knoten i bezeichnet, die Antisymmetrie $g_{ji} = -g_{ij}$, so daß der Knoten j den gleichen Beitrag mit entgegengesetztem Vorzeichen erhält. Wir können g_{ij} auch als 'Projektion' eines gemittelten Flußes auf die Kante zwischen den beiden beteiligten Knoten interpretieren, weshalb wir ihn als *Galerkin Fluß* vom Knoten j in den Knoten i bezeichnen. Zur Veranschaulichung betrachten wir den eindimensionalen Fall mit linearen Finiten Elementen, so daß die Gewichte durch $c_{ji} = -c_{ij} = 1/2$ gegeben sind. Der resultierende Ausdruck $g_{ij} = (f_i + f_j)/2$ demonstriert die Äquivalenz zwischen zentraler Differenzenapproximation und Galerkin Diskretisierung in 1D. Für lineare Dreiecks- bzw. Tetraederelemente korrespondieren die Galerkin Flüsse mit den 'physikalischen' Kanten der Triangulierung. Bei der Verwendung von multilinearen Elementtypen oder solchen von höherer Ordnung lassen sich die g_{ij} mit den Kanten der Konnektivitätsmatrix assoziieren und sind somit losgelöst von der Topologie des Rechengitters.

Ein vielversprechender Zugang zur Herleitung von hochauflösenden FEM besteht darin, den Galerkin Fluß (1.24) durch einen 'geeigneten' konsistenten numerischen Fluß zu ersetzen. Dieses Vorgehen wird beispielsweise in den Publikationen [53], [54] und [55] demonstriert, in denen die Autoren im wesentlichen eindimensionale Limiter in Verbindung mit der kantenbasierten Datenstruktur von Peraire *et al.* [60] auf unstrukturierten Gittern einsetzen. Desweiteren lassen sich mit Hilfe von modifizierten Galerkin Flüssen eine Vielzahl von eindimensionalen Diskretisierungstechniken für hyperbolische Erhaltungsgleichungen, wie etwa *approximate Riemann solver*, *Upwind-biasing* oder *scalar limited dissipation* Schemata in den Kontext von Finiten Elementen einbetten.

Auch wenn der vollständige Übergang zu einer kantenbasierten Formulierung vom Standpunkt der Effizienz sinnvoll ist, reicht es für das weitere Vorgehen aus, lediglich die anti-/diffusiven Flüsse in dieser Form darzustellen. Dies ermöglicht eine Integration der im folgenden vorgestellten Verfahren in existierende Codes mit einem traditionellen elementweisen Matrizenaufbau.

Wenn wir die Lösung, die Quellterme und den konvektiven Flußanteil mit Hilfe der *group finite element formulation* approximieren

$$u = \sum_j u_j \varphi_j, \qquad \mathbf{v}u = \sum_j (\mathbf{v}_j u_j)\, \varphi_j, \qquad q = \sum_j q_j \varphi_j, \tag{1.25}$$

so liefert die Standard Galerkin Diskretisierung der schwachen Formulierung (1.3)

$$\sum_j \left[\int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x} \right] (\dot{u}_j - q_j) + \sum_j \left[\int_\Omega (\varphi_i \mathbf{v}_j \cdot \nabla \varphi_j + d \nabla \varphi_i \cdot \nabla \varphi_j) \, \mathrm{d}\mathbf{x} \right] u_j = 0. \tag{1.26}$$

Hierbei gehen wir davon aus, daß kein diffusiver Fluß am Rand auftritt, so daß der Beitrag des durch partielle Integration entstehenden Oberflächenintegrals verschwindet. Das obige System von ODEs läßt sich in kompakter Matrixform als

$$M_C \frac{\mathrm{d}u}{\mathrm{d}t} = Ku + M_C q \tag{1.27}$$

schreiben, wobei $M_C = \{m_{ij}\}$ die konsistente Massenmatrix und $K = \{k_{ij}\}$ den diskreten Transportoperator bezeichnet. Die einzelnen Matrixeinträge sind durch

$$m_{ij} = \int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x}, \qquad k_{ij} = -\mathbf{v}_j \cdot \mathbf{c}_{ij} - d\, s_{ij} \tag{1.28}$$

gegeben. Die Koeffizienten $\mathbf{c}_{ij}$ und s_{ij} ergeben sich aus der Diskretisierung von Differentialoperatoren für die ersten bzw. zweiten Ableitungen

$$\mathbf{c}_{ij} = \int_\Omega \varphi_i \nabla \varphi_j \, \mathrm{d}\mathbf{x}, \qquad s_{ij} = \int_\Omega \nabla \varphi_i \cdot \nabla \varphi_j \, \mathrm{d}\mathbf{x}, \tag{1.29}$$

wobei jetzt die spezielle Indizierungsweise der Galerkin Koeffizienten gerechtfertigt wird. Wir möchten bemerken, daß die Koeffizienten m_{ij}, $\mathbf{c}_{ij}$ und s_{ij} für ein festes Rechengitter konstant bleiben, so daß sie nur einmal zu Beginn der Simulation mittels numerischer Kubatur berechnet werden müssen und abgespeichert werden können. Im weiteren Verlauf lassen sich die Nebendiagonaleinträge ohne teure numerische Integration gemäß (1.28) aufbauen. Gleichzeitig ist dieses abkürzende Vorgehen für den im folgenden vorgestellten Algorithmus nicht verpflichtend, so daß eine bestehende Aufbauroutine für den diskreten Transportoperator beibehalten werden kann.

1.5 Eigenschaften diskreter Operatoren

1.5.1 Einleitung

In diesem Abschnitt wollen wir einige mathematische Werkzeuge vorstellen, die sich bei der Herleitung von hochauflösenden FEM-Schemata als wertvoll herausstellen werden. Zum einen sollte ein numerisches Verfahren denselben physikalischen Gesetzen, etwa der Positivitätserhaltung von Dichte und Temperatur, genügen wie die kontinuierliche Modellgleichung. Desweiteren ist das Konvergenzverhalten eines numerischen Verfahrens von Interesse. Da wir unsere Theorie sowohl für lineare als auch für nichtlineare PDEs entwickeln werden, müssen wir neben der Frage, ob ein numerisches Verfahren überhaupt konvergiert, auch noch die beiden folgenden Probleme berücksichtigen:

- Die Methode kann gegen eine schwache Lösung der zugrunde liegenden Erhaltungsgleichung konvergieren, die nicht der Entropiebedingung genügt.
- Die Methode kann gegen eine Funktion konvergieren, die keine schwache Lösung der zugrunde liegenden Erhaltungsgleichung ist.

Der erste Punkt ist einfach einzusehen. Wenn mehrere schwache Lösungen existieren, müssen wir einen Mechanismus finden, der garantiert, daß das Verfahren gegen die eindeutig bestimmte physikalisch 'richtige' Entropielösung konvergiert. Aber auch der zweite Fall kann für zwei verschiedene Erhaltungsgleichungen mit äquivalenten stetigen Lösungen eintreten, wenn ihre schwachen Lösungen unterschiedlich sind. Als Beispiel betrachte man die folgenden PDEs

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}u^2\right) = 0, \qquad \frac{\partial u^2}{\partial t} + \frac{\partial}{\partial x}\left(\frac{2}{3}u^3\right) = 0, \tag{1.30}$$

welche dieselben stetigen Lösungen besitzen. Wie man anhand der Rankine-Hugoniot Bedingung verifizieren kann, ist die Schockgeschwindigkeit der beiden Probleme unterschiedlich [46].

Die obigen Forderungen an das numerische Verfahren führen direkt auf die Klasse der *monotonen* Methoden. Für zwei unterschiedliche Anfangslösungen u^0 und v^0 mit $v^0 \geq u^0$ muß gelten

$$v_i^n \geq u_i^n, \quad \forall i \qquad \Rightarrow \qquad v_i^{n+1} \geq u_i^{n+1} \quad \forall i. \tag{1.31}$$

Für eine konsistente monotone Methode wurde von Harten *et al.* [24] gezeigt, daß sie gegen die Entropielösung konvergiert. Dieses Ergebnis zusammen mit

dem Lax-Wendroff Theorem [44] garantiert, daß eine numerische Lösung, die mit einer konsistenten, monotonen und konservativen Methode berechnet wurde und die fast überall gleichmäßig gegen eine Funktion v konvergiert, die eindeutig bestimmte physikalisch 'richtige' schwache Lösung der Erhaltungsgleichung ist.

Nach dem Satz von Godunov können *lineare* monotone Methoden höchstens von erster Ordnung genau sein, so daß sie trotz ihrer oben angesprochenen 'schönen' Eigenschaften eine stark eingeschränkte Klasse bilden. Als Relaxierung bietet es sich an, von *monotonieerhaltenden* Methoden zu sprechen. Dazu genügt es, daß für monotone Anfangsdaten gilt

$$u_i^0 \geq u_j^0, \quad j \neq i \qquad \Rightarrow \qquad u_i^n \geq u_j^n, \quad j \neq i, \quad \forall n, \tag{1.32}$$

wodurch garantiert wird, daß keine Oszillationen an isolierten Unstetigkeitsstellen entstehen können. Die Aufgabe besteht nun darin, hinreichende Kriterien zu finden, aus denen die Monotonieerhaltung folgt.

1.5.2 Ortsdiskretisierung

Lax [45] hat die Beobachtung gemacht, daß für physikalisch zulässige Lösung einer skalaren Erhaltungsgleichung die *totale Variation* (TV), die durch

$$TV(u) = \int_{-\infty}^{+\infty} \left| \frac{\partial u}{\partial x} \right| \mathrm{d}x \tag{1.33}$$

definiert ist, nicht anwachsen kann. Obwohl sie ursprünglich für stetige Funktionen hergeleitet wurde, behält die Gleichung (1.33) auch für unstetige $u(x)$ zumindest im Distributionssinne ihre Gültigkeit. Nach Harten [22] wird eine numerische Methode als *total variation diminishing* (TVD) bezeichnet, wenn gilt

$$TV(u^{n+1}) \leq TV(u^n), \quad \forall n. \tag{1.34}$$

Mit Hilfe der TVD Eigenschaft hat Harten [23] eine Klasse von nicht oszillierenden TVD Schemata entwickelt, welche notwendig stabil im Sinne der totalen Variation (*total variation stability*) sind. Daß für solche Methoden die Monotonie erhalten bleibt, resultiert aus dem nachfolgenden Satz.

Satz 1.5.1. *Jedes TVD Schema ist monotonieerhaltend [46].*

Beweis. Dies folgt aus der Tatsache, daß Oszillationen zu einem Anwachsen der totalen Variation führen würden. Angenommen für die Anfangslösung gilt $u_i^0 \geq u_{i+1}^0$ für alle i und $TV(u^0) < \infty$. Dann muß $TV(u^0) = |u_{-\infty}^0 - u_{\infty}^0|$ gelten. Aufgrund der Endlichkeit des Einflußgebietes folgt weiter $u_j^n \to u_{\pm\infty}^0$ für alle

späteren Zeiten t^n. Also folgt $TV(u^n) \geq |u^0_{-\infty} - u^0_{\infty}|$. Da wir das Schema als *total variation diminishing* vorausgesetzt haben, muß die Gleichheit $TV(u^n) = |u^0_{-\infty} - u^0_{\infty}|$ gelten und somit u^n oszillationsfrei sein. □

In einer Dimension ist der TVD Begriff eng mit der von Jameson [30], [31] eingeführten *local extremum diminishing* (LED) Eigenschaft veknüpft. Das semidiskrete Problem sei in der folgenden Form darstellbar

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_j c_{ij} u_j \qquad \text{mit} \qquad \sum_j c_{ij} = 0, \tag{1.35}$$

wobei u_i die Funktionswerte in den Knoten bezeichnet und die Koeffizienten c_{ij} von der jeweiligen räumlichen Diskretisierung stammen. Im Abschnitt 1.6 werden wir demonstrieren, wie diese Darstellung mit Hilfe der auf den konvektiven Flußanteil angewendeten Flußzerlegung aus Abschnitt 1.4 erzeugt werden kann.

Da nach Voraussetzung die Koeffizientenmatrix $\{c_{ij}\}$ die Zeilensumme Null hat, läßt sich die Gleichung (1.35) auch in der folgenden Form schreiben

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \neq i} c_{ij}(u_j - u_i). \tag{1.36}$$

Wir gehen zunächst davon aus, daß für die Koeffizienten c_{ij} gilt

$$c_{ij} \geq 0, \quad j \neq i. \tag{1.37}$$

Dann läßt sich für die Darstellung (1.36) leicht die L_∞-Stabilität zeigen [30], [31].

Lemma 1.5.2. *Das semi-diskrete Problem lasse sich in der Form (1.36) darstellen. Weiter seien alle Koeffizienten c_{ij} für $j \neq i$ nichtnegativ. Dann ist die Methode stabil in der L_∞-Norm.*

Beweis. Sei o.B.d.A. u_i ein Maximum. Dann gilt $u_j - u_i \leq 0$ für beliebige u_j, und es folgt $\mathrm{d}u_i/\mathrm{d}t \leq 0$. Also kann das Maximum nicht weiter anwachsen. Analog zeigt man, daß ein Minimum nicht absinken kann. □

Im allgemeinen sind Koeffizientenmatrizen nur dünnbesetzt und besitzen die Eigenschaft, daß ein Eintrag c_{ij} nur dann von Null verschieden ist, wenn die Knoten i und j im numerischen Sinne adjazent sind, d.h. wenn sich die Träger ihrer Basisfunktionen überschneiden. Wenn wir den obigen Beweis wiederholen und uns nur auf die unmittelbaren Nachbarn eines lokalen Extremums beschränken, erhalten wir die LED Eigenschaft, mit deren Hilfe Jameson [32], [33] eine Klasse von hochauflösenden Schemata für unstrukturierte Gitter hergeleitet hat. Aus

der LED Eigenschaft folgt die Positivitätserhaltung, denn für $u_i \geq 0 \, \forall i$ folgt, daß auch das globale Minimum positiv ist, da es nicht absinken kann.

Laney und Caughey [42] haben gezeigt, daß der Wert eines Extremums in den Variationen der beidseitig angrenzenden Segmente auftritt, so daß, wenn für den linken und rechten Endpunkt des Definitionsintervals homogene Randwerte angenommen werden, für die globale Variation die Gleichung

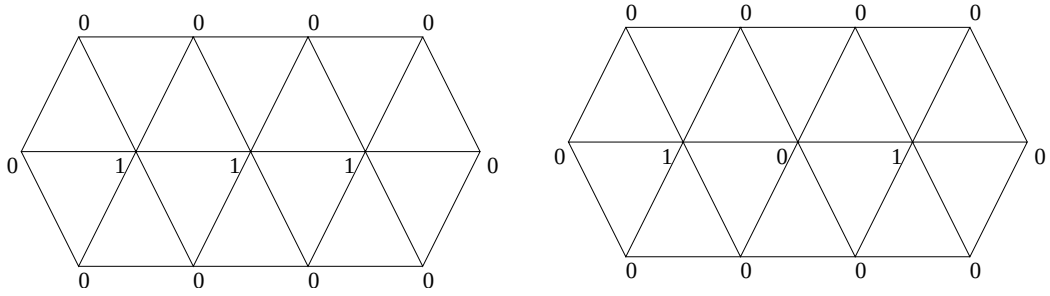
$$TV(u) = 2 \left(\sum \max u - \sum \min u \right) \tag{1.38}$$

gilt. Man sieht leicht, daß in einer Dimension aus der LED die TVD Eigenschaft folgt. Tadmor [74] hat darüberhinaus die Äquivalenz beider Begriffe in 1D gezeigt, welche sich dahingehend ausnutzen läßt, daß Konvergenzbeweise für konsistente, TV-stabile Methoden leicht auf LED Schemata übertragen werden können [22], [46]. Erfreulicherweise ermöglicht die in der physikalischen Lösung zwangsläufig vorkommende und von der numerischen explizit geforderte Eigenschaft von TVD die Konstruktion von nichtlinearen Methoden mit hoher Genauigkeitsordnung. Im Gegensatz dazu fallen sämtliche Methoden auf erste Ordnung zurück, wenn sie versuchen, andere Eigenschaften der exakten Lösung 'nachzuahmen' [46].

Der eindeutige Vorteil der LED Charakterisierung im Gegensatz zu TVD wird erst in mehreren Dimensionen ersichtlich. Als Veranschaulichung nehmen wir eine Dreieckszerlegung des Gebietes Ω und die entsprechende mehrdimensionale Definition der totalen Variation [32]

$$TV(u) = \int ||\nabla u|| \mathrm{d}s. \tag{1.39}$$

Das folgende Beispiel [32] zeigt, daß zwei einzelne Spitzen eine kleinere Variation bewirkten (links, $TV = 4{+}2\sqrt{3}$ (L_1), 6 (L_2), $2{+}2\sqrt{3}$ (L_∞)) als ein ganzer Kamm (rechts, $TV = 6{+}\sqrt{3}$ (L_1), 7 (L_2), $5{+}3\sqrt{3}$ (L_∞)). Im Gegensatz zum weitgehend eindimensionalen Charakter von TVD läßt sich das LED Konzept auf mehrdimensionale Problemstellungen und unstrukturierte Gitter verallgemeinern.



1.5.3 Zeitdiskretisierung

Erinnern wir uns daran, daß die Gleichungen (1.35) und (1.36) ein semi-diskretes Problem beschreiben. Für die Zeitdiskretisierung verwenden wir das übliche einschrittige θ-Schemata, nach dessen Anwendung die vollständig diskretisierte Gleichung die folgende Form besitzt

$$\frac{u_i^{n+1}-u_i^n}{\Delta t} = \theta \sum_{j\neq i} c_{ij}(u_j^{n+1}-u_i^{n+1}) + (1-\theta)\sum_{j\neq i} c_{ij}(u_j^n - u_i^n), \quad 0 \leq \theta \leq 1. \tag{1.40}$$

Die Wahl von θ entscheidet über die Art des Zeitschrittverfahrens. Die beiden Extremfälle $\theta = 0$ und $\theta = 1$ führen auf die bekannten *Forward* und *Backward Euler* Methoden, welche in der Zeit von erster Ordnung genau sind. Eine Methode von zweiter Genauigkeitsordnung erhält man mit der Wahl von $\theta = 0.5$, die allgemein als *Crank-Nicolson* bekannt ist. Weiterhin gilt das folgende [37]

Positivitätstheorem 1.5.3. *Jedes LED Schema, das als Zeitdiskretisierung die Backward Euler Methode einsetzt, ist uneingeschränkt positiv. Andere Zeitschrittverfahren* $(0 \leq \theta < 1)$ *erhalten die Positivität unter der CFL-artigen Bedingung*

$$1 + \Delta t(1-\theta)\min_i c_{ii} \geq 0. \tag{1.41}$$

Beweis. Zunächst beweisen wir die uneingeschränkte Positivität der Backward Euler Methode. Für diese vollimplizite Zeitdiskretisierung verschwindet der letzte Term in Gleichung (1.40). Wir nehmen an, daß die diskrete Lösung u^{n+1} an einigen Stellen negativ ist, wobei k den Knoten bezeichnet, an dem das globale Minimum angenommen wird. Die neue Lösung im Knoten k erfüllt dann zum Zeitpunkt t^{n+1} die Gleichung

$$u_k^{n+1} = u_k^n + \Delta t \sum_{j\neq k} c_{kj}(u_j^{n+1} - u_k^{n+1}). \tag{1.42}$$

Nach Induktionsannahme gilt für die alte Lösung u^n, daß sie überall nichtnegativ ist. Weiter gilt aufgrund der LED Voraussetzung, daß auch die Koeffizienten c_{kj} stets nichtnegativ sind. Da $u_k^{n+1} < 0$ angenommen wurde, folgt daraus $u_j^{n+1} - u_k^{n+1} < 0$ für ein j, was zu einem Widerspruch führt, da u_k^{n+1} als globales Minimum angenommen wurde.

Betrachten wir nun den Fall $\theta < 1$. Nach den Vorüberlegungen zum vollimpliziten Schema folgt, daß Gleichung (1.40) die Positivität erhält, solange der explizite Anteil der folgenden Ungleichung genügt

$$u_i^n + \Delta t(1-\theta)\sum_{j\neq i} c_{ij}(u_j^n - u_i^n) \geq 0, \qquad \forall i. \tag{1.43}$$

Da nach Voraussetzung $u_i^n \geq 0$ und $c_{ij} \geq 0$ gilt, reicht es aus, für den Zeitschritt

$$1 + \Delta t(1-\theta) \min_i c_{ii} \geq 0 \tag{1.44}$$

zu fordern, wobei sich die Diagonaleinträge der Koeffizientenmatrix nach (1.35) als $c_{ii} = -\sum_{j \neq i} c_{ij}$ schreiben lassen. □

Als direkte Konsequenz aus dem Positivitätstheorem läßt sich für beliebige LED Schemata eine rigorose Abschätzung des maximal zulässigen Zeitschritts bei expliziten oder semi-impliziten Methoden angeben. Bemerkenswert ist, daß die Herleitung der oberen Schranke keinerlei Wissen über die zugrunde liegende PDE oder das verwendete Gitternetz erfordert, sondern auf semi-diskreter Ebene anhand der Diagonalkoeffizienten c_{ii} erfolgt.

Wenn jedoch das diskrete Schema nicht die LED Eigenschaft besitzt, brauchen wir einen anderen Mechanismus, der uns die Positivitätserhaltung garantiert. Ein allgemeines Kriterium baut auf dem Konzept von M-Matrizen auf.

Definition 1.5.4. *Ein nichtsingulärer diskreter Operator $A \in \mathbb{R}^{N \times N}$ heißt M-Matrix, falls alle $a_{ij} \leq 0$ für beliebige $i \neq j$ und alle Einträge der Inversen A^{-1} nichtnegativ sind.*

Eine hinreichende Charakterisierung einer M-Matrix leistet das folgende

Lemma 1.5.5. *Es seien $A \in \mathbb{R}^{N \times N}$ strikt diagonaldominant, alle Diagonaleinträge positiv ($a_{ii} > 0$) und alle Nebendiagonaleinträge nichtpositiv ($a_{ij} \leq 0$ für $i \neq j$). Dann ist A eine M-Matrix.*

Beweis. Sei $D = \mathrm{diag}\{a_{ii}\}$ der Diagonalanteil von A, so daß die Aufspaltung $A = D - N$ gilt. Offensichtlich ist D nichtsingulär und $N \geq 0$. Nach dem Satz von Gerschgorin folgt, daß alle Eigenwerte von $D^{-1}N$ im Innern des Einheitskreises $\mathbb{D}$ liegen. Insgesamt folgt aus

$$A^{-1} = [I - D^{-1}N]^{-1} D^{-1} = \sum_{k=0}^{\infty} [D^{-1}N]^k D^{-1} \geq 0 \tag{1.45}$$

die Behauptung. □

Offensichtlich folgt für eine M-Matrix A, daß $Ax \geq 0$, nur wenn $x \geq 0$ gilt, woraus sich ein zweites Positivitätstheorem herleiten läßt.

Positivitätstheorem 1.5.6. *Das numerische Schema lasse sich mit abstrakten Matrizenoperatoren in der folgenden Form schreiben*

$$Au^{n+1} = Bu^n. \tag{1.46}$$

Dann ist eine hinreichende Bedingung für die Positivitätserhaltung der Methode, daß A eine M-Matrix ist und alle Einträge von B nichtnegativ sind ($B \geq 0$).

Beweis. Da A eine M-Matrix ist, ist ihre Inverse nichtnegativ ($A^{-1} \geq 0$). Es folgt $u^{n+1} = A^{-1}Bu^n \geq 0$ solange $u^n \geq 0$. □

Weiterhin ergeben sich die folgenden Eigenschaften.

Bemerkung 1.5.7. *Die Voraussetzungen des obigen Theorems sind hinreichend (aber nicht notwendig), um zu garantieren, daß die numerische Lösung das diskrete Maximumprinzip erfüllt.*

Bemerkung 1.5.8. *Die Wahl des Zeitschritts Δt beeinflußt das Vorzeichen der Matrixeinträge, so daß aus der Bedingung $B \geq 0$ eine CFL-artige obere Schranke für explizite Schemata folgt.*

Mit den Positivitätstheoremen (1.5.3) und (1.5.6) stehen zwei wichtige Bausteine für die Herleitung von monotonieerhaltenden Methoden zur Verfügung. In einer Vielzahl von Arbeiten werden Techniken zur Konstruktion von diskreten Operatoren mit den notwendigen Eigenschaften aus (1.5.3) oder (1.5.6) beschrieben. Ein Konzept beruht auf der Einführung von künstlicher Viskosität oder Upwindbiasing Techniken [32], [34], zum Teil auch mit Hilfe von *flux-splitting* Ansätzen. Letztlich können aber nach Godunov lineare Methoden höchstens von erster Ordnung genau sein, falls sie die Monotonie erhalten. In der Literatur findet sich auch eine Vielzahl von sogenannten hochauflösenden LED Schemata, die auf algebraischen *flux-limiting* oder geometrischen *slope-limiting* Techniken aufbauen. Zur ersten Klasse gehören etwa die von Sweby eingeführten Methoden [72], welche die TVD Eigenschaft garantieren. Zu der zweiten Klasse zählen prominente Vertreter wie das von van Leer [76] aus der Godunov Methode weiterentwickelte MUSCL, das von Colella und Woodward [9] vorgestellte PPM-Schema oder die von Harten und Osher [25] entwickelten ENO-Verfahren.

1.6 Discrete Upwinding

Mitentscheidend für den Erfolg des gesamten FEM-FCT Algorithmus ist die Methode niedriger Ordnung, da sich von ihr produzierte Oszillationen auf die Endlösung übertragen würden. Von einer 'perfekten' Methode niedriger Ordnung erwarten wir, daß sie gerade genug künstliche Diffusion enthält, um die Positivität zu garantieren, jedoch die Lösung nicht übermäßig stark 'verschmiert'. Für Finite Differenzen oder Finite Volumen Diskretisierungen bietet sich das klassische Upwind-Verfahren an. Im Kontext von Finiten Elementen fehlte lange Zeit ein entsprechendes Äquivalent, und viele Upwind-artige Verfahren greifen zur Diskretisierung des konvektiven Terms auf Finite Volumen zurück. Löhner *et al.* [51], [52] schlagen die Verwendung von konstanter Massendiffusion (1.7) vor, was eine effiziente Implementierung ermöglicht. Dennoch liefert dieses Vorgehen aus Sicht der Genauigkeit kein 'optimales' Verfahren. Falls der freie Parameter c_d zu groß gewählt wird, nimmt die Diffusivität der Methode zu und schränkt ihren Stabilitäts- bzw. Positivitätsbereich unnötig ein. Auf der anderen Seite bilden sich kleine unphysikalische Extrema, wenn die künstliche Diffusion zu gering ausfällt. Diese Nachteile sind auch Georghiou *et al.* [19] aufgefallen, die in ihrem 'verbesserten' FEM-FCT Algorithmus versuchen, den 'optimalen' Diffusionskoeffizienten in Abhängigkeit von der lokalen Courantzahl zu bestimmen. Das in Anlehnung an Finite Differenzen konstruierte Verfahren funktioniert nur für reguläre Gitter und kann auch dann keine Positivitätsgarantie geben. Andere Autoren [68] greifen die Idee von konstanter künstlicher Diffusion wieder auf, ohne dabei auf die Originalarbeiten von Löhner [51], [52] zu verweisen, und bestimmen den Koeffizienten c_d so, daß die M-Matrix Eigenschaft sichergestellt wird.

Wir wollen im folgenden auf eine von Kuzmin und Turek [37] vorgeschlagene Diskretisierungstechnik eingehen, die unter Erhaltung der Positivität zu einem numerischen Verfahren mit geringst möglicher Diffusivität führt, welches als *discrete Upwinding* bezeichnet wird. Dabei gehen wir von der semi-diskreten ODE-Form (1.27) aus. Im ersten Schritt wird die konsistente Massenmatrix durch die via *mass lumping* erzeugte Diagonalmatrix $M_L = \{m_i\}$ ersetzt, wodurch die implizit in $M_C = \{m_{ij}\}$ enthaltene Massendiffusion eliminiert wird. Damit erhalten wir für das semi-diskrete Problem die folgende Darstellung

$$M_L \frac{\mathrm{d}u}{\mathrm{d}t} = Ku + M_L q, \tag{1.47}$$

die sich für jede Komponente u_i in die äquivalente Form

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \neq i} k_{ij}(u_j - u_i) + \delta_i u_i + m_i q_i \tag{1.48}$$

bringen läßt. In der obigen Gleichung bezeichnet der Summenterm den inkompressiblen Anteil des diskreten Transportterms, $\delta_i u_i$ als diskretes Analogon von

$u\nabla \cdot \mathbf{v}$ den kompressiblen, welcher für divergenzfreie Geschwindigkeitsfelder verschwindet, und $m_i q_i$ den Beitrag der Quell- und Senkterme. Die einzelnen Matrixeinträge sind gemäß

$$m_i = \sum_j m_{ij}, \qquad \delta_i = \sum_j k_{ij} \tag{1.49}$$

definiert. Unser Ziel besteht darin, eine oszillationsfreie Methode niedriger Ordnung zu konstruieren. Die beiden letzten Terme in Gleichung (1.48) sind für ein physikalisches Anwachsen von lokalen Extrema aufgrund von Kompressibilität oder Quellen/Senken verantwortlich und sollten daher nicht manipuliert werden. Der Einfachheit halber nehmen wir im folgenden $\delta_i = q_i = 0$ an, so daß sich unser semi-diskretes Problem in der Form (1.36) schreiben läßt

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \neq i} c_{ij}(u_j - u_i), \qquad \text{mit} \qquad c_{ij} = \frac{k_{ij}}{m_i} \tag{1.50}$$

wobei $K = \{k_{ij}\}$ die Zeilensumme Null besitzt. Im vorherigen Abschnitt haben wir für ein semi-diskretes Problem der Form (1.50) die LED Eigenschaft nachgewiesen, falls alle c_{ij} mit $j \neq i$ nichtnegativ sind. Unser Ziel besteht darin, den diskreten Transportoperator durch Addition von künstlicher Diffusion so zu modifizieren, daß alle negativen Nebendiagonaleinträge eliminiert werden.

1.6.1 Diskrete Diffusionsoperatoren

Kuzmin und Turek [37] haben zu diesem Zweck das Konzept von *diskreten Diffusionsoperatoren* eingeführt. Wir verlangen von dem Operator $D = \{d_{ij}\}$ die beiden nachfolgenden Eigenschaften:

Symmetrie

$$d_{ij} = d_{ji} \tag{1.51}$$

und Zeilen-/Spaltensumme Null

$$\sum_j d_{ij} = \sum_i d_{ij} = 0. \tag{1.52}$$

Man rechnet leicht nach, daß die diskreten Diffusionsoperatoren $D(N \times N; \mathbb{R})$ bezüglich der Addition eine Gruppe bildet. Die Anwendung eines diskreten Diffusionsoperators D auf einen Vektor aus Knotenwerten ergibt mit Hilfe der Nullzeilensumme die Darstellung

$$(Du)_i = \sum_j d_{ij} u_j = \sum_{j \neq i} d_{ij}(u_j - u_i). \tag{1.53}$$

Wir definieren den Fluß f_{ij} vom Knoten j in den Knoten i als

$$f_{ij} = d_{ij}(u_j - u_i). \tag{1.54}$$

Damit läßt sich Gleichung (1.53) als Summe von antisymmetrischen Flüssen

$$(Du)_i = \sum_{j \neq i} f_{ij}, \qquad f_{ji} = -f_{ij} \tag{1.55}$$

umschreiben. Diffusive Terme, die aus der Anwendung eines diskreten Diffusionsoperators resultieren, lassen sich also als Summe von numerischen Flüssen ähnlich denen, die in konservativen Finiten Differenzen Schemata auftreten, darstellen und die mit den Kanten des Konnektivitätsgraphen der Finiten Elemente Matrix assoziiert werden können. Für lineare Approximationen auf Dreiecks- oder Tetraedergittern stimmen die numerischen Kanten mit den physikalischen Kanten der Finiten Elemente überein. Für multilineare Elemente oder Approximationen höherer Ordnung kommen die 'virtuellen' Kanten der Konnektivitätsmatrix hinzu, welche die Abhängigkeiten der Freiheitsgrade untereinander widerspiegeln. Einfach ausgedrückt tauscht ein Knoten mit all den umliegenden Knoten Masse aus, die mit ihm ein gemeinsames Element teilen. Aufgrund der Antisymmetrie des Flußvektors ist der Nettofluß zwischen je zwei Knoten gleich Null, so daß die Massenerhaltung garantiert ist.

Im folgenden wollen wir einige spezielle Diffusionsoperatoren untersuchen, die im Bereich von FEM Verwendung finden.

Diskreter Laplace Operator Die Einträge der Steifigkeitsmatrix

$$s_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, \mathrm{d}\mathbf{x}, \tag{1.56}$$

resultieren aus der Diskretisierung einer physikalischen Dissipation.

Stromliniendiffusions-Operator Die entlang der Stromlinie des Geschwindigkeitsfeldes hinzugefügte künstliche Diffusion des Operators

$$d^s_{ij} = \int_\Omega \mathbf{v} \cdot \nabla\varphi_i \; \mathbf{v} \cdot \nabla\varphi_j \, \mathrm{d}\mathbf{x} \tag{1.57}$$

bewirkt eine Stabilisierung des konvektiven Terms. Dieses Konzept wurde erstmals von Brooks und Hughes [5] innerhalb einer konsistenten Petrov-Galerkin Formulierung vorgeschlagen. Einen ähnlichen Ansatz, basierend auf einer charakteristischen Stromliniendiffusion, schlägt Johnson [35] in seinen Arbeiten vor.

Den gleichen Operator d^s_{ij} erhält man bei der Anwendung der *least-squares* Formulierung von Carey und Jiang [7]. Nicht zuletzt lassen sich die von Donea *et al.* [13] für Taylor-Galerkin Schemata eingeführten Terme höherer Ordnung in der zeitlichen Taylorentwicklung als Operator im Sinne von (1.57) auffassen.

Massendiffusions-Operator Durch Subtraktion der 'gelumpten' von der konsistenten Massenmatrix $(M_C - M_L)$ erhält man die Massendiffusion

$$d^m_{ij} = \int_\Omega \varphi_i(\varphi_j - \delta_{ij})\,\mathrm{d}\mathbf{x}, \tag{1.58}$$

die unter anderem in der Methode niedriger Ordnung des FEM-FCT Algorithmus von Löhner *et al.* [51], [52] zur 'Monotonisierung' der expliziten Taylor-Galerkin Diskretisierung hoher Ordnung verwendet wurde.

1.6.2 Konstruktion eines linearen LED Schemas

Im folgenden werden wir das Konzept von verallgemeinerten Diffusionsoperatoren nutzen, um den diskreten Transportoperator in (1.50) in einen Operator mit der LED Eigenschaft zu überführen. Dazu definieren wir die künstliche Dissipationsmatrix $D = \{d_{ij}\}$ als

$$d_{ij} = d_{ji} = \max\{0, -k_{ij}, -k_{ji}\}, \quad \forall\ i < j, \qquad d_{ii} = -\sum_{k \neq i} d_{ik}. \tag{1.59}$$

Man sieht leicht, daß die notwendigen Eigenschaften (1.51) und (1.52) erfüllt sind. Wenn wir D auf den Transportoperator K anwenden, so erhalten wir sein LED Pendant niedriger Ordnung $L = K + D$. Zum Aufbau des discrete Upwind-Operators bietet sich ein kantenweises Vorgehen an. Ausgehend vom Galerkin Operator $L = K$ wird für jede Kante $\vec{ij}$ die folgende Modifikation durchgeführt

$$\begin{aligned} l_{ii} &= l_{ii} - d_{ij}, & l_{ij} &= l_{ij} + d_{ij}, \\ l_{ji} &= l_{ji} + d_{ij}, & l_{jj} &= l_{jj} - d_{ij}, \end{aligned} \tag{1.60}$$

was im wesentlichen der Anwendung eines eindimensionalen Diffusionsoperators entlang der (fiktiven) Kanten zwischen adjazenten Knoten entspricht. Der globale Matrixaufbau verläuft nunmehr standardmäßig. Es bleibt anzumerken, daß die Koeffizienten k_{ij} bei ausreichender physikalischer Diffusion bereits nichtnegativ und somit die Matrizen K und L identisch sind, was zeigt, daß nur die tatsächlich notwendige künstliche Diffusion addiert wird, so daß die resultierende Methode niedriger Ordnung die am wenigsten diffusive ist.

Nach dem Positivitätstheorem (1.5.3) ist ein LED Schema für die Backward Euler Zeitdiskretisierung uneingeschränkt positiv und für andere Verfahren positivitätserhaltend, solange der Zeitschritt die folgende Bedingung erfüllt

$$\Delta t \leq \frac{1}{1-\theta} \min_i \{-m_i/l_{ii} : \ l_{ii} < 0\}. \tag{1.61}$$

Aus dieser CFL-artigen Bedingung läßt sich eine obere Schranke für den maximal zulässigen Zeitschritt Δt ablesen, die etwa zur Steuerung eines adaptiven Zeitschritt-Verfahrens dienen kann. Die obere Schranke hängt vom Grad der Implizitheit θ ab und wird von der in l_{ii} beherbergten Diffusion beeinflußt, so daß für übermäßige Diffusion nicht nur die Genauigkeit der Methode abnimmt, sondern auch unwirtschaftlich kleine Zeitschrittweiten notwendig werden. Wir wollen das obige Vorgehen an einem einfachen Beispiel verdeutlichen.

Beispiel 1: Dazu betrachten wir die eindimensionale Konvektionsgleichung

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0, \tag{1.62}$$

auf einem uniformen Gitter. Wir gehen von einer konstanten Geschwindigkeit $v > 0$ aus, was auf die folgenden Elementmatrizen führt

$$\hat{M}_L = \frac{\Delta x}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \hat{K} = \frac{v}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}. \tag{1.63}$$

Nach dem Aufbau der globalen Matrizen erhält man daraus für die inneren Knoten die zentrale Differenzen-Approximation des konvektiven Terms

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -v\ \frac{u_{i+1} - u_{i-1}}{2\,\Delta x}. \tag{1.64}$$

Die negativen Nebendiagonaleinträge von K können eliminiert werden, indem die künstliche Diffusion als $\hat{d}_{12} = v/2$ gewählt wird. Auf ein einzelnes Element eingeschränkt gilt dann für den Diffusionsoperator und den daraus entstehenden diskreten Transportoperator niedriger Ordnung

$$\hat{D} = \frac{v}{2}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad \Rightarrow \quad \hat{L} = v\begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}. \tag{1.65}$$

Für innere Knoten ergibt sich das für Finite Differenzen bekannte am wenigsten diffusive Upwind-Verfahren

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = -v\ \frac{u_i - u_{i-1}}{\Delta x}, \tag{1.66}$$

welches die Positiviät unter der CFL-artigen Bedingung

$$v\frac{\Delta t}{\Delta x} \leq \frac{1}{1-\theta} \tag{1.67}$$

erhält. Dieses Beispiel zeigt die Äquivalenz von discrete Upwinding und dem klassischen Upwind für Finite Differenzen in einer Dimension.

1.6.3 Quelltermlinearisierung

Bisher sind wir stets von $\delta_i = q_i = 0$ und der Darstellung (1.50) ausgegangen. Wir betrachten nun den Fall, daß ein beliebiger Quellterm $s(u)$ auftritt, so daß das vollständig diskretisierte Schema die folgende Gestalt besitzt

$$[M_L - \theta \Delta t L]\; u^{n+1} = [M_L + (1-\theta)\Delta t L]\; u^n + \Delta t\; s(u^{n+\theta}). \tag{1.68}$$

Um für negative Quellterme die Voraussetzungen des Positivitätstheorems 1.5.6 nicht zu verletzen, greifen wir auf den von Patankar [59] vorgeschlagenen Linearisierungsansatz zurück und passen ihn an die Positivitätskriterien an.

Der Quellterm $s(u)$ läßt sich in einen konstanten Term s_C und einen nichtlinear von der Lösung abhängenden Anteil s_V aufteilen

$$s(u) = s_C + s_V u. \tag{1.69}$$

Dabei wird von s_V verlangt, daß er nichtpositiv ($s_V \leq 0$) ist, was Patankar als *negative-slope linearization* des Quellterms bezeichnet. Diese Forderung ist nicht etwa willkürlich, sondern hat einen physikalischen Hintergrund. Angenommen s_V sei positiv und die abhängige Variable u sei gerade die Temperatur T, dann würde ein Anstieg von T zu einem Anwachsen des Quellterms führen. Wenn keine Möglichkeit zum Abbau der Temperatur vorgesehen ist, bewirkt dieser unendliche Zyklus eine 'Explosion' des Wertes von T. Aus numerischer Sicht sorgt die Forderung $s_V \leq 0$ weiterhin für die nötige Stabilität. Wir wollen zusätzlich von einem nichtnegativen $s_C \geq 0$ ausgehen. Damit läßt sich vermöge der Identität

$$s = s^+ - s^- = s^+ + \left(\frac{-s^-}{u}\right) u \tag{1.70}$$

ein einfaches Splitting der Form (1.69) angeben, bei dem (s^+) den positiven und ($-s^-$) den negativen Anteil des Quellterms bezeichnet. Offensichtlich müssen diese beiden Größen in jeder Iteration anhand des aktuellen Wertes für u neu berechnet werden. Der Notation von Patankar folgend bezeichnen wir die zur Zeit beste Approximation von u mit u^*. Damit ergibt sich für (1.69) die Setzung

$$s_C = s^+ \qquad \text{und} \qquad s_V = -s^-/u^*. \tag{1.71}$$

Übrig bleibt noch die Festlegung des Zustandes von u, um den herum der Quellterm linearisiert wird. Für $s = s(u^{n+\theta})$ ist die natürliche Wahl

$$u = \theta u^{n+1} + (1-\theta) u^n, \tag{1.72}$$

wobei dann auch der negative Anteil $(1-\theta) s_V u^n$ in der rechten Seite der Gleichung (1.68) auftritt, was zu einer restriktiveren Bedingung für den maximal zulässigen

Zeitschritt führt. Alternativ kann man den Quellterm unabhängig von θ vollimplizit um $u = u^{n+1}$ herum linearisieren. Der Vorteil dieses Ansatzes besteht darin, daß sich die Gleichung (1.68) mit Hilfe der Diagonalmatrizen

$$S^+ = \mathrm{diag}\left\{\frac{s^+}{u^n}\right\} \qquad \text{und} \qquad S^- = \mathrm{diag}\left\{\frac{s^-}{u^*}\right\} \tag{1.73}$$

in der bevorzugten Form (1.46) darstellen läßt

$$[M_L - \theta\Delta t L + \Delta t S^-]u^{n+1} = [M_L + (1-\theta)\Delta t L + \Delta t S^+]u^n. \tag{1.74}$$

Man sieht leicht ein, daß diese Wahl von $S^\pm$ die im Positivitätstheorem 1.5.6 von A und B geforderten Eigenschaften noch verstärkt. Nach Konstruktion sind nun alle Nebendiagonaleinträge von A nichtpositiv, während diejenigen von B nichtnegativ sind. Um auch die Positivität der Diagonaleinträge von B sicherzustellen, genügt es, eine obere Schranke für den Zeitschritt anzugeben, die im Vergleich zur Bedingung (1.61) noch den positiven Anteil des Quellterms berücksichtigt

$$\Delta t \le \min_i \left\{ \frac{-m_i\ u_i^n}{(1-\theta) l_{ii}\ u_i^n + s_i^+} \ :\ l_{ii} < 0 \right\}. \tag{1.75}$$

Offensichtlich stimmt diese Bedingung für $s^+ = 0$ mit (1.61) überein.

1.7 Flussbasiertes FEM-FCT

Ein weiterer wichtiger Baustein des FEM-FCT Algorithmus ist eine lineare Methode hoher Ordnung. In der Literatur findet man eine Vielzahl von Finite Elemente Schemata, die versuchen, den konvektiven Term mit Hilfe von Stromliniendiffusion zu stabilisieren [5],[7],[12]. Die Klasse der Taylor-Galerkin Methoden nutzt zur Stabilisierung Zeitableitungen höherer Ordnung aus der Taylorentwicklung. Dies führt zu verbesserten Zeitdiskretisierungen, die mit der traditionellen Galerkin Ortsdiskretisierung kombiniert werden. Der bekannteste Vertreter dieser Klasse ist das Lax-Wendroff Verfahren. Eine genaue Untersuchung der *modified equation* zeigt, daß die eingeführte Dissipation gerade ausreicht, um die immanente negative Diffusion auszugleichen, welche sonst das explizite Euler/Galerkin Verfahren für rein konvektive Probleme instabil machen würde. Eine detaillierte Untersuchung der Standard-Galerkin bzw. Lax-Wendroff Methode und anderer Taylor-Galerkin Verfahren höherer Ordnung findet man in [12],[13].

Während für vollexplizite Zeitschrittverfahren eine Stabilisierung des konvektiven Terms notwendig ist, kommen implizite Finite Elemente Verfahren, die auf einer Crank-Nicolson oder Backward Euler Zeitdiskretisierung beruhen, ohne diese aus. Lineare Diskretisierungen dieser Art sind für sich genommen nur wenig nützlich, da sie unphysikalische Oszillationen entstehen lassen. Im folgenden werden wir zwei auf einer flußbasierten Darstellung der diffusiven/antidiffusiven Terme aufbauende FCT Formulierungen vorstellen, die als lineare Methode hoher Ordnung das Lax-Wendroff Verfahren und als Methode niedriger Ordnung das im vorherigen Abschnitt vorgestellte monotonieerhaltende discrete Upwinding einsetzen. Ein nichtlinearer Limiter kombiniert beide Diskretisierungen zu einem hochauflösenden Verfahren, das frei von unphysikalischen Oszillationen ist.

1.7.1 Basis Formulierung

Wir wollen der Übersichtlichkeit halber auf den (linearisierten) Quellterm verzichten. Nach der Zeitdiskretisierung mit Hilfe des einschrittigen θ-Schemas lassen sich die Methoden hoher und niedriger Ordnung folgendermaßen kombinieren

$$Au^H = b^n + f(u^n, u^H), \tag{1.76}$$

wobei A und b^n die Matrix bzw. die rechte Seite niedriger Ordnung bezeichnet

$$A = M_L - \theta\Delta t L, \qquad b^n = [M_L + (1-\theta)\Delta t L]u^n. \tag{1.77}$$

Ferner ergibt sich der für die hohe Genauigkeit verantwortliche Antidiffusionsterm f als Differenz zwischen beiden Diskretisierungen

$$\begin{aligned} f(u^n, u^H) &= [(\underbrace{M_C - M_L}_{D^m}) - (1-\theta)\Delta t(\underbrace{L-K}_{D})]u^n + \Delta t D^s u^n \\ &- [(\underbrace{M_C - M_L}_{D^m}) + \theta\Delta t(\underbrace{L-K}_{D})]u^H. \end{aligned} \tag{1.78}$$

Hierbei bezeichnet D^s den für die Stabilisierung im Lax-Wendroff Verfahren verantwortlichen Stromliniendiffusionsoperator (1.57), während das hochgestellte H auf die Lösung der Methode hoher Ordnung hinweist. Der vollständige Verzicht auf den antidiffusiven Term $f(u^n, u^H)$ liefert die positivitätserhaltende Methode niedriger Ordnung, welche jedoch recht diffusiv ist. Wird hingegen die komplette Antidiffusion beibehalten, so ergibt sich die Methode hoher Ordnung mit ausgeprägten Oszillationen. Die Idee von FCT besteht darin, gerade soviel Antidiffusion zuzulassen, daß die überflüssige Diffusion aus der Methode niedriger Ordnung entfernt wird ohne jedoch unphysikalische Oszillationen entstehen zu lassen, wobei uns das Positivitätstheorem 1.5.6 einen großen Dienst erweisen wird.

Nach Konstruktion (vgl. Abschnitt 1.6) ist der Operator A der linken Seite eine M-Matrix. Das Ziel besteht darin, die rechte Seite in der Form

$$b = B\tilde{u} \tag{1.79}$$

mit einer nichtnegativen Matrix $B \geq 0$ und einer positivitätserhaltenden Zwischenlösung $\tilde{u}$ darzustellen, so daß aus der Positivität von u^n mit Hilfe des Theorems 1.5.6 für die Lösung zum nächsten Zeitschritt $u^{n+1} \geq 0$ folgt. Dazu betrachten wir das folgende explizite Teilproblem mit der rechten Seite aus (1.77)

$$M_L \tilde{u}^n = b^n, \tag{1.80}$$

wobei M_L für die 'gelumpte' Massenmatrix steht. Die Lösung $\tilde{u}^n$ läßt sich als Zwischenlösung der expliziten Methode niedriger Ordnung zum Zeitpunkt $t^{n+1-\theta}$ interpretieren. Im Grenzfall der vollimpliziten Backward Euler Methode stimmt $\tilde{u}^n$ mit u^n überein, und es gilt $\tilde{u}^n = u^L$ für $\theta = 0$. Entscheidend für den Erfolg von FCT ist die Positivität von $\tilde{u}^n$. Für die vollimplizite Zeitdiskretisierung ist dies stets gewährleistet, und für $0 \leq \theta < 1$ liefert das Positivitätstheorem 1.5.3 eine Bedingung für den größtmöglichen Zeitschritt, unter dem $\tilde{u}^n$ positiv ist.

Um den zulässigen Anteil an kompensierender Antidiffusion zu bestimmen, gehen wir zu einer Flußdarstellung des Terms (1.78) über. Die Matrizen D^m, D und D^s gehören zur Gruppe der diskreten Diffusionsoperatoren $D(N\times N;\, \mathbb{R})$ und besitzen daher die Zeilen- und Spaltensumme Null. Demnach läßt sich $f(u^n, u^H)$ in eine Summe von antisymmetrischen Flüssen $f_{ij} = -f_{ji}$ für $i \neq j$ zerlegen

$$\begin{aligned} f_{ij} &= (m_{ij} - (1-\theta)\Delta t d_{ij})\ (u_j^n - u_i^n) + \Delta t d_{ij}^s (u_j^n - u_i^n) \\ &- (m_{ij} + \theta\Delta t d_{ij})\ (u_j^H - u_i^H). \end{aligned} \tag{1.81}$$

Diese 'Rohflüsse' gleichen den Fehler aus, der durch *mass lumping* und discrete Upwinding entsteht. Die Kunst besteht darin, gerade soviel Antidiffusion zu addieren, daß keine unphysikalischen Über- und Unterschwinger erzeugt werden. Daraus ergibt sich der in [37] vorgestellte Basis Algorithmus

$$m_i u_i^{n+1} - \theta\Delta t \sum_j l_{ij} u_j^{n+1} = m_i \tilde{u}_i^n + \sum_{j\neq i} \alpha_{ij} f_{ij}, \qquad \alpha_{ij} = \alpha_{ij}(\tilde{u}^n, f_{ij}), \tag{1.82}$$

der mit Hilfe von symmetrischen Korrekturfaktoren $\alpha_{ji} = \alpha_{ij}$, auf die wir im Abschnitt 1.7.4 genauer eingehen werden, für eine oszillationsfreie und gleichzeitig hochgenaue Lösung sorgt. Wir erkennen bereits, daß eine 'geschickte' Bestimmung des Wertes $0 \leq \alpha_{ij} \leq 1$ ausreicht, um unter Berücksichtigung der Positivität von $\tilde{u}^n$ zu garantieren, daß alle Einträge der rechten Seite nichtnegativ sind. Zusammen mit der M-Matrixeigenschaft der linken Seite sichert das Positivitätstheorem 1.5.6, daß aus $u^n \geq 0$ auch $u^{n+1} \geq 0$ folgt.

Diese neue Familie von FEM-FCT Schemata unterscheidet sich von den bisher bekannten Ansätzen, indem sie gleichermaßen auf explizite und implizite Zeitdiskretisierungen angewendet werden kann. Weiterhin ist die vollexplizite Variante bis auf die mathematisch motivierte Konstruktion der Methode niedriger Ordnung mit dem von Löhner *et al.* [51], [52] eingeführten Standard FEM-FCT konsistent, so daß sich der neue Ansatz nahtlos in die FCT Familie einfügt. Wir wollen darauf hinweisen, daß für implizite Schemata *zwei* nichtsymmetrische lineare Gleichungssysteme pro Zeitschritt gelöst werden müssen: Eins für die Lösung hoher Ordnung, aus der die antidiffusiven Flüsse berechnet werden, und ein weiteres für die Endlösung zum nächsten Zeitpunkt. Von diesem Standpunkt aus gesehen, sind die Kosten von expliziten Verfahren pro Zeitschritt deutlich geringer. Die Effizienz von impliziten Verfahren und der Vorteil ihrer numerischen Robustheit wirkt sich erst für größere Zeitschrittweiten und damit vor allem bei der Behandlung von stationären Problemen aus.

1.7.2 Defektkorrektur für nichtlineare Probleme

Bei der Simulation von praxisrelevanten Anwendungen wird man häufig mit nichtlinearen Erhaltungsgleichungen konfrontiert. Ein typisches Beispiel ist die nichtviskose Burgers Gleichung, die einen eindimensionalen Prototyp der Euler und Navier-Stokes Gleichungen darstellt

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0. \tag{1.83}$$

Allgemein hängen für nichtlineare Problemstellungen die Matrizen $K = K(u)$ und $L = L(u)$ von der unbekannten Lösung u ab, so daß implizite Schemata zusätzliche *äußere* Iterationen erfordern. Die Linearisierung der Erhaltungsgleichung mit

Hilfe einer Zeitextrapolation kann zu Problemen führen, da ein Massenverlust nicht auszuschließen ist.

Im folgenden werden wir daher die einfachste ableitungsfreie Methode zur Behandlung von nichtlinearen Gleichungssystemen einsetzen: Fixpunkt-Defektkorrektur. Wir gehen zunächst von einem abstrakten nichtlinearen System

$$A(u)u = b \tag{1.84}$$

aus, für das ein allgemeines Fixpunkt-Defektkorrektur-Verfahren die Darstellung

$$u^{(m+1)} = u^{(m)} + [C^{(m)}]^{-1} r^{(m)}, \qquad m = 0, 1, \ldots \tag{1.85}$$

besitzt. Hierbei bezeichnet $r^{(m)}$ den Residuenvektor der m-ten Iteration und die reguläre Matrix $C \in \mathbb{R}^{N \times N}$ einen 'geeigneten' Vorkonditionierer. In der Praxis wird man (1.85) nicht direkt berechnen, sondern ein lineares Teilproblem für die Differenz zweier Iterierter lösen

$$C^{(m)} \Delta u^{(m)} = r^{(m)} \tag{1.86}$$

und das Ergebnis zur alten Iterierten hinzuaddieren

$$u^{(m+1)} = u^{(m)} + \Delta u^{(m)}, \qquad u^{(0)} = u^n. \tag{1.87}$$

Dabei genügt es, nur eine mittlere Anzahl an inneren Iterationen durchzuführen, da bereits eine Verbesserung des Residuums um 1–2 Stellen pro Iteration für eine hohe Genauigkeit der Endlösung ausreicht. Für den Vorkonditionierer empfiehlt sich der diskrete Operator niedriger Ordnung

$$C^{(m)} = M_L - \theta \Delta t L(u^{(m)}), \tag{1.88}$$

da sich dessen M-Matrixeigenschaft positiv auf die Robustheit des iterativen Lösers auswirkt. Der Residuenvektor ergibt sich als

$$r^{(m)} = b^n + f(u^n, u^{(m)}) - A(u^{(m)})u^{(m)}, \tag{1.89}$$

wobei die rechte Seite der Methode niedriger Ordnung durch

$$b^n = [M_L + (1-\theta)\Delta t L(u^n)]u^n \tag{1.90}$$

und der antidiffusive Term gemäß

$$\begin{aligned} f(u^n, u^{(m)}) &= [(M_C - M_L) - (1-\theta)\Delta t D(u^n)]u^n \\ &- [(M_C - M_L) + \theta \Delta t D(u^{(m)})]u^{(m)} \end{aligned} \tag{1.91}$$

definiert sind. Ersetzen wir letzteren im Residuenvektor (1.89) durch den *korrigierten* Fluß $f^*(u^{(m)}, u^n)$, so erhalten wir die Basis Formulierung des verallgemeinerten (expliziten und impliziten) FEM-FCT Algorithmus (vgl. Abschnitt

1.7.1). Die Flußkorrektur kann entweder nach jedem äußeren Iterationsschritt durchgeführt werden oder nur einmal, nachdem die Methode hoher Ordnung zur Konvergenz geführt wurde. Wie wir in Abschnitt 1.7.4 nachweisen werden, kann in beiden Fällen die Positivität der resultierenden Lösung bei Anwendung des Zalesak-Limiters garantiert werden.

Aber auch lineare Probleme können von dieser iterativen Behandlung mittels Defektkorrektur profitieren. Für große Zeitschritte bei impliziten Schemata degeneriert die Matrix der Methode hoher Ordnung zunehmend, so daß iterative Verfahren divergieren können. Beim Übergang zu einem Defektkorrektur-Algorithmus wird die Originalmatrix durch einen 'gutartigen' Vorkonditionierer approximiert und so die Konvergenz der Iterationen gesichert.

1.7.3 Iterative Formulierung

In den beiden vorherigen Abschnitten sind wir zunächst von der allgemeinen FCT Formulierung (1.82) ausgegangen und haben *anschließend* eine iterative Defektkorrekturschleife zur Behandlung der Nichtlinearitäten von außen 'aufgesetzt'. Flußkorrektur und iterative Defektkorrektur wurden dabei unabhängig voneinander betrachtet. Da letztere für explizite Verfahren entfällt, ist diese Separation zumindest in diesem Fall gerechtfertigt. Die resultierende FEM-FCT Formulierung ist aber auch für implizite Diskretisierungen zulässig und bildet bei Verwendung von moderaten Zeitschrittweiten ein hochauflösendes Verfahren.

Für stationäre Probleme spielt die *Zeit*genauigkeit keine Rolle, da es vielmehr auf die Minimierung des *Orts*fehlers ankommt. Da vollimplizite LED Schemata uneingeschränkt positivitätserhaltend sind, kann der als künstlicher Relaxationsparameter fungierende Zeitschritt aus Sicht der Zeitgenauigkeit beliebig groß gewählt werden. Sein Wert geht aber über die Flüsse in den Zalesak-Limiter ein, der diese in Abhängigkeit von der Größe von Δt begrenzt und so für große Zeitschritte mehr von der aus dem discrete Upwinding stammenden Diffusion beibehält, was im Endeffekt zu einer starken 'Verschmierung' der numerischen Ergebnisse führt.

Im folgenden wollen wir den in [40] eingeführten iterativen FEM-FCT Algorithmus betrachten, der die Zeitschrittabhängigkeit des Zalesak-Limiters umgeht. Da große Zeitschritte nur für implizite Verfahren zulässig sind, gehen wir von Anfang an von der Defektkorrekturschleife (1.86)–(1.87) aus

$$C^{(m)}\Delta u^{(m)} = r^{(m)} \qquad m = 0, 1, \ldots \tag{1.92}$$

$$u^{(m+1)} = u^{(m)} + \Delta u^{(m)}, \qquad u^{(0)} = u^n. \tag{1.93}$$

Im Gegensatz dazu findet man bei Schär und Smolarkiewicz [66] die etwas fragwürdige Konstruktion eines iterativen FCT Algorithmus für *explizite* Zeitdiskretisierungen, der zudem noch auf Finite Differenzen beschränkt bleibt. Als Vorkonditionierer benutzen wir den diskreten Operator niedriger Ordnung, der mit Ausnahme der Situation, daß der Diskretisierung eine lineare PDE zugrunde liegt, in jedem Iterationsschritt neu berechnet werden muß

$$C^{(m)} = M_L - \theta \Delta t L(u^{(m)}). \tag{1.94}$$

In der Basis FCT Formulierung haben wir die positivitätserhaltende Zwischenlösung $\tilde{u}^n$ nur *einmal* zu Beginn eines jeden Zeitschritts bestimmt

$$M_L \tilde{u}^n = b^n. \tag{1.95}$$

Hierbei bezeichnet b^n die rechte Seite der Methode niedriger Ordnung (1.90). Dieses Vorgehen birgt den Nachteil, daß der Limiter in jedem Iterationsschritt ganz von vorne mit der Begrenzung der kompletten Antidiffusion beginnt. Der in Abschnitt 1.7.1 vorgestellte FEM-FCT Algorithmus führt im Kontext einer iterativen Defektkorrektur auf die folgende Darstellung der rechten Seite

$$b_i^{(m+1)} = b_i^n + \sum_{j \neq i} \alpha_{ij}^{(m)} f_{ij}^{(m)}, \qquad \alpha_{ij}^{(m)} = \alpha_{ij}(\tilde{u}^n, f_{ij}^{(m)}). \tag{1.96}$$

Die Korrekturfaktoren $\alpha_{ij}^{(m)}$ hängen von der Hilfslösung $\tilde{u}^n$ und dem Zusammenspiel der antidiffusiven 'Rohflüsse' $f_{ij}^{(m)}$ ab. Den Vorteil, $\tilde{u}^n$ nur einmal zu Beginn eines jeden Zeitschritts berechnen zu müssen, erkauft man sich mit der zunehmenden Diffusivität der Endlösung für große Zeitschritte. Anhand von Gleichung (1.81) erkennt man, daß der Beitrag der Ortsdiskretisierung zum 'Rohfluß' proportional zur Größe von Δt ist. Gleichzeitig hängt der Anteil an zulässiger Antidiffusion ausschließlich von $\tilde{u}^n$ ab, so daß mit zunehmender Zeitschrittweite eine immer stärkere Begrenzung des Flusses erfolgt.

Die Hauptidee bei der iterativen Formulierung besteht darin, die bereits akzeptierte Antidiffusion bei der Berechnung der positivitätserhaltenden Zwischenlösung $\tilde{u}^{(m)}$ zu berücksichtigen. *De facto* bestimmt der Limiter nur im ersten Iterationsschritt die Korrekturfaktoren für die komplette Antidiffusion und wird anschließend auf die Flußdifferenz zwischen der vollständigen und der bereits akzeptierten angewandt, so daß mehr und mehr Antidiffusion in die Endlösung aufgenommen wird. Dazu muß die Zwischenlösung $\tilde{u}^{(m)}$ mit Hilfe des Vektors der rechten Seite $b^{(m)}$ in *jedem* Schritt der Iteration aktualisiert werden

$$M_L \tilde{u}^{(m)} = b^{(m)}, \qquad b^{(0)} = b^n. \tag{1.97}$$

Dies ist – erinnern wir uns daran, daß M_L eine Diagonalmatrix darstellt – ohne das explizite Lösen eines linearen Gleichungssystems möglich, so daß der notwendige Mehraufwand vernachlässigt werden darf. Wir folgen weiterhin der ursprünglichen Strategie des Zalesak-Limiters, bestimmen die Koeffizienten $\alpha_{ij}^{(m)}$

jedoch aufbauend auf den mit Hilfe von $\tilde{u}^{(m)}$ berechneten Schranken

$$\alpha_{ij}^{(m)} = \alpha_{ij}(\tilde{u}^{(m)}, \Delta f_{ij}^{(m)}) \tag{1.98}$$

und wenden die Korrekturfaktoren auf die Flußdifferenz zwischen den aktuellen 'Rohflüssen' und den in früheren Iterationen akzeptierten Beiträgen an

$$\Delta f_{ij}^{(m)} = f_{ij}^{(m)} - g_{ij}^{(m)}. \tag{1.99}$$

Anschließend wird die korrigierte Antidiffusion in Vorbereitung auf den nachfolgenden Iterationsschritt zur Summe ihrer Vorgänger addiert

$$g_{ij}^{(m+1)} = g_{ij}^{(m)} + \alpha_{ij}^{(m)} \Delta f_{ij}^{(m)} = \sum_{j \neq i} \alpha_{ij}^{(m)} \Delta f_{ij}^{(m)}, \qquad g_{ij}^{(0)} = 0 \tag{1.100}$$

und in den globalen Lastvektor eingefügt

$$b_i^{(m+1)} = b_i^{(m)} + \sum_{j \neq i} \alpha_{ij}^{(m)} \Delta f_{ij}^{(m)}. \tag{1.101}$$

Wie zuvor sichert der Limiter die Existenz einer nichtnegativen Matrix $B \geq 0$ mit $b^{(m+1)} = B\tilde{u}^{(m)}$, für die das inverse Diffusionsproblem (1.97) eine positivitätserhaltende Lösung $\tilde{u}^{(m+1)} = M_L^{-1} B \tilde{u}^{(m)}$ besitzt. Für die vollexplizite Zeitdiskretisierung ergibt sich der klassische FCT Algorithmus mit $u^{n+1} = \tilde{u}^{(1)}$. Im impliziten Fall wird ein sich sukzessiv vergrößernder Anteil an Antidiffusion in die provisorische Lösung $\tilde{u}^{(m)}$ eingebracht, so daß der Limiter eine geringere Korrektur der Flußdifferenz $\Delta f_{ij}^{(m)}$ durchführen muß. Dies steht im Gegensatz zu der Basis Formulierung, welche vom Limiter in jeder Iteration einen *cold start* erforderte. Für beide FEM-FCT Formulierungen läßt sich der Defektvektor für das lineare Gleichungssystem (1.92) einheitlich in folgender Weise darstellen

$$r^{(m)} = b^{(m+1)} - A(u^{(m)})u^{(m)}. \tag{1.102}$$

Offensichtlich stimmt der vorgestellte Algorithmus für $m = 0$ mit der Basis Formulierung (1.96) überein, so daß sich letztere als Spezialfall in die iterative FCT Variante einbetten läßt. Wenn wir als Konsistenzcheck die Korrekturfaktoren der m-ten Iteration als $\alpha_{ij}^{(m)} \equiv 1$ festsetzen, folgt durch sukzessives Einsetzen

$$b_i^{(m+1)} = b_i^n + \sum_{j \neq i} (g_{ij}^{(m)} + \Delta f_{ij}^{(m)}) = b_i^n + \sum_{j \neq i} f_{ij}^{(m)}, \tag{1.103}$$

so daß sich korrekterweise die Standard Galerkin Diskretisierung ergibt.

1.7.4 Limiting und Positivitätsbeweis

Entscheidend für die Monotonie- und Positivitätserhaltung der Lösung u^{n+1} ist die Bestimmung der 'optimalen' Korrekturfaktoren. Durch die Wahl von $0 \leq \alpha_{ij} \leq 1$ wird zwischen den Methoden niedriger und hoher Ordnung über eine beliebige Kombination dazwischen variiert. Im folgenden werden wir den mehrdimensionalen Limiter von Zalesak, der bereits in Abschnitt 1.3 vorgestellt wurde, auf eine andere Weise deuten, um die Positivitätserhaltung auch im impliziten Fall zu beweisen. Es bezeichne $\tilde{u}_i^{\substack{\max \\ \min}}$ das lokale Maximum/Minimum der Lösung in der Umgebung des Knotens i und seiner Nachbarn

$$\tilde{u}_i^{\substack{\max \\ \min}} = \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} \tilde{u}_j, \qquad j \in \mathsf{S}_i. \tag{1.104}$$

Hierbei steht $\tilde{u} = u^L(t^{n+1-\theta})$ für die positive Zwischenlösung des Hilfsproblems (1.95) oder (1.97). Wie in der ursprünglichen Variante von Boris und Book [4] wird die alte Lösung u^n nicht in die Berechnung der lokalen Extrema einbezogen.

Der klassischen FCT Theorie folgend werden alle antidiffusiven Flüsse, die ein bestehendes Extremum verstärken, eliminiert

$$\alpha_{ij} = 0 \qquad \text{falls} \quad \left\{ \begin{array}{lll} \tilde{u}_i = \tilde{u}_i^{\max} & \text{und} & f_{ij} > 0 \\ \tilde{u}_i = \tilde{u}_i^{\min} & \text{und} & f_{ij} < 0. \end{array} \right. \tag{1.105}$$

Falls dies für alle zum Knoten i gehörenden Flüsse der Fall ist, sind wir fertig. Ansonsten müssen die übrigen Flüsse so begrenzt werden, daß sie die Positivität der Endlösung nicht verletzen. Dazu stellen wir zunächst die rechte Seite der Gleichung (1.96) bzw. (1.101) in der folgenden Form dar

$$b_i = m_i \tilde{u}_i + \sum_{j \neq i} \alpha_{ij} f_{ij} = m_i \tilde{u}_i + c_i Q_i, \tag{1.106}$$

wobei der Koeffizient c_i als

$$c_i = \frac{1}{Q_i} \sum_{j \neq i} \alpha_{ij} f_{ij} \tag{1.107}$$

gesetzt wird. Der Multiplikator Q_i ergibt sich aus der Setzung

$$Q_i = \left\{ \begin{array}{ll} Q_i^+ = \tilde{u}_i^{\max} - \tilde{u}_i, & \\ Q_i^- = \tilde{u}_i^{\min} - \tilde{u}_i, & \text{falls} \quad \displaystyle\sum_{j \neq i} \alpha_{ij} f_{ij} \left\{ \begin{array}{l} > 0, \\ < 0, \\ = 0. \end{array} \right. \\ 1, & \end{array} \right. \tag{1.108}$$

Aufgrund von (1.105) gilt stets $Q_i \neq 0$, so daß in (1.107) keine Division durch Null stattfindet. Weiterhin ist der Koeffizient c_i stets nichtnegativ. Wir gehen davon

aus, daß das lokale Extremum $\tilde{u}_i^{\max\atop\min}$ an einem Knoten k, der adjazent zum Knoten i ist, angenommen wird. Dann weist der antidiffusive Term die LED Eigenschaft auf, so daß sich Gleichung (1.106) als

$$b_i = m_i\tilde{u}_i + c_i(\tilde{u}_k - \tilde{u}_i) = (m_i - c_i)\tilde{u}_i + c_i\tilde{u}_k \tag{1.109}$$

schreiben läßt, wobei $c_i \geq 0$ gilt. Da nach Konstruktion alle übrigen Voraussetzungen des Positivitätstheorems 1.5.3 erfüllt sind, genügt es, an die rechte Seite (1.96) bzw. (1.101) die weitere Forderung $m_i \geq c_i$ zu stellen, aus der sich eine allgemeine Regel zur Bestimmung der Korrekturfaktoren α_{ij} herleiten läßt.

Gerade diese Aufgabe erfüllt der Limiter von Zalesak [79]. Die Größen $P_i^{\pm}$ und $R_i^{\pm}$ seien gemäß (1.11) und (1.13) als

$$P_i^{\pm} = \frac{1}{m_i}\sum_{j\neq i}\left\{\begin{matrix}\max\\ \min\end{matrix}\right\}\{0, f_{ij}\}, \qquad R_i^{\pm} = \begin{cases} \min\{1, Q_i^{\pm}/P_i^{\pm}\}, & \text{für } P_i^{\pm} \neq 0,\\ 1, & \text{für } P_i^{\pm} = 0 \end{cases}$$

gewählt. Da der flußbasierte Massenaustausch auf einer bilateralen Basis stattfindet, werden weiter die Korrekturfaktoren wie in (1.14) bestimmt

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\}, & \text{für } f_{ij} \geq 0,\\ \min\{R_j^+, R_i^-\}, & \text{für } f_{ij} < 0. \end{cases}$$

Der daraus resultierende Limiter erfüllt automatisch die Bedingung (1.105), da aus $Q_i^{\pm} = 0$ sofort $R_i^{\pm} = \alpha_{ij} = 0$ folgt, so daß lokale Extrema nicht verstärkt werden. Weiter gilt nach Setzung der jeweiligen Größen die Ungleichungskette

$$\sum_{j\neq i}\alpha_{ij}f_{ij} \leq \sum_{j\neq i}\alpha_{ij}\max\{0, f_{ij}\} \leq m_iR_i^+P_i^+ \leq m_iQ_i^+ \tag{1.110}$$

und analog dazu die Abschätzung nach unten

$$\sum_{j\neq i}\alpha_{ij}f_{ij} \geq \sum_{j\neq i}\alpha_{ij}\min\{0, f_{ij}\} \geq m_iR_i^-P_i^- \geq m_iQ_i^-. \tag{1.111}$$

Zalesaks Limiter stellt also sicher, daß die Summe über die korrigierten Flüsse nach oben und nach unten durch den Abstand zum lokalen Extremum der positivitätserhaltenden Zwischenlösung $\tilde{u}$ beschränkt bleibt

$$m_iQ_i^- \leq \sum_{j\neq i}\alpha_{ij}f_{ij} \leq m_iQ_i^+ \tag{1.112}$$

und damit stets die nach dem Theorem 1.5.6 für die Positivitätserhaltung hinreichende Bedingung $m_i \geq c_i$ erfüllt wird.

1.8 Zusammenfassung des Algorithmus

Im folgenden wollen wir die wesentlichen Schritte des FEM-FCT Algorithmus zusammenfassen, wobei wir die Basis Formulierung als Spezialfall der iterativen Variante interpretieren. Wir beschreiben einen Defektkorrekturzyklus, wobei gewisse Initialisierungen nur im ersten Schritt durchgeführt werden.

In der kantenbasierten Aufbauroutine:

F.1 Berechne die Einträge k_{ij} und k_{ji} vom diskreten Transportoperator hoher Ordnung nach (1.28) oder greife auf die abgespeicherten Werte zurück.

F.2 Bestimme die Diffusionskoeffizienten d_{ij} entsprechend (1.59) und konstruiere den Operator niedriger Ordnung (1.60) und den Vorkonditionierer $C^{(m)}$.

F.3 Addiere den Kantenbeitrag zum Residuum $r^{(m)}$ definiert in (1.102) und für $m = 0$ und $\theta < 1$ ebenfalls zur rechten Seite b^n (1.90).

F.4 Berechne die antidiffusiven Flüsse f_{ij} entsprechend (1.81) und die zu begrenzende Flußdifferenz Δf_{ij} nach (1.99).

Im Flußkorrektur-Modul:

F.5 Berechne (für die nichtiterative Formulierung nur für $m = 0$) die positivitätserhaltende Zwischenlösung $\tilde{u}^{(m)}$ nach (1.97).

F.6 Bestimme mit Hilfe des Zalesak-Limiters (mit Pre- und Postlimiting) die Korrekturfaktoren $\alpha_{ij}^{(m)}$ und begrenze die antidiffusiven Rohflüsse.

F.7 Addiere die begrenzte Antidiffusion zum Residuenvektor $r^{(m)}$ in (1.102).

F.8 Aktualisiere in der iterativen Formulierung $b^{(m)}$ nach (1.101) und die akkumulierte Antidiffusion $g_{ij}^{(m)}$ entsprechend (1.100).

In der Defektkorrektur Schleife:

F.9 Löse das lineare System für $\Delta u^{(m+1)}$ (1.86) mit dem Defektvektor $r^{(m)}$.

F.10 Aktualisiere die neue Lösung $u^{(m+1)}$ entsprechend (1.87) und fahre mit der nichtlinearen Iterationsschleife fort oder gehe zum nächsten Zeitschritt.

Kapitel 2

Numerische Beispiele für skalare Problemstellungen

Im folgenden Kapitel werden wir anhand einer Reihe von unterschiedlichen Testfällen die Leistungsfähigkeit der neuen FEM-FCT Formulierungen demonstrieren und somit bereits einen Ausblick auf das Potential dieser Methodik für realistische Anwendungen geben. Zu Beginn eines jeden Abschnitts werden wir kurz auf die Herleitung der analytischen Lösung eingehen, die mit den numerischen Ergebnissen verglichen wird. Neben den hier wiedergegebenen Benchmarks liegen noch eine Vielzahl von Resultaten für andere Problemstellungen vor, auf die wir aber aus Platzgründen nicht näher eingehen können. Stattdessen möchten wir den Leser auf die Arbeiten [37] und [38] verweisen.

2.1 Eindimensionale Benchmarks

Wir beginnen die numerische Analyse von FEM-FCT mit einem repräsentativen Querschnitt durch allgemein anerkannte eindimensionale Testfälle. Aufgrund der geringen Rechenzeiten kann hier eine numerische Bestimmung der Konvergenzordnung durchgeführt werden. Weiterhin lassen sich einzelne Phänomene wie Pre- und Postlimiting in 1D leichter veranschaulichen.

2.1.1 Lineare Konvektionsgleichung

Konstante Konvektionsgeschwindigkeit Als ersten Testfall betrachten wir die lineare Konvektionsgleichung mit konstanter Geschwindigkeit $v > 0$

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0 \tag{2.1}$$

zusammen mit der Anfangsbedingung

$$u(x,0) = u^0(x). \tag{2.2}$$

Die analytische Lösung ergibt sich einfach als

$$u(x,t) = u^0(x - vt), \tag{2.3}$$

wobei diese für einmal stetig differenzierbare Anfangsdaten der klassischen Lösung der Differentialgleichung (2.1) entspricht und sonst im schwachen Sinne zu verstehen ist. Das Anfangsprofil wird mit der Geschwindigkeit v konvektiert und bleibt entlang der Charakteristiken $x - vt = x^0$ mit dem Anfangspunkt x^0 konstant. Die Charakteristiken bezeichnen die Kurven in der $x-t$–Ebene, die der ODE

$$x'(t) = v, \qquad x(0) = x^0 \tag{2.4}$$

genügen und entlang derer die materielle Ableitung der Lösung verschwindet

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u(x(t),t) &= \frac{\partial u(x(t),t)}{\partial t} + \frac{\partial u(x(t),t)}{\partial x}\, x'(t) \\ &= \frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0. \end{aligned} \tag{2.5}$$

Damit ist u Lösung der PDE (2.1) mit den Anfangsbedingungen (2.2).

'Step function' Als ersten Benchmark betrachten wir die lineare Konvektion einer unstetigen Anfangslösung mit Geschwindigkeit $v = 1$

$$u(x,0) = \begin{cases} 0, & \text{falls} \quad x \in [0, 0.2] \cup [0.6, 2] \\ 1, & \text{falls} \quad x \in (0.2, 0.6). \end{cases} \tag{2.6}$$

Für eine Ortsdiskretisierung mit 200 linearen Elementen erhalten wir bei einem Zeitschritt von $\Delta t = 10^{-3}$ die zugehörige Courantzahl

$$\nu = v\frac{\Delta t}{\Delta x} = 0.1. \tag{2.7}$$

Betrachten wir zunächst das vollexplizite FEM-FCT Schema, das auf dem von zweiter Ordnung genauen Lax-Wendroff Zeitschrittverfahren basiert. Zum Zeitpunkt $t = 1.0$ weist die Methode hoher Ordnung wie erwartet gravierende Über-

und Unterschwinger auf, während die Ergebnisse, die mit Hilfe von discrete Upwinding berechnet wurden, monoton und frei von Oszillationen, jedoch mit großer numerischer Diffusion behaftet sind. Die nichtlineare Kombination beider Methoden im FEM-FCT Algorithmus bringt eine enorme Verbesserung mit sich, die durch ein vorgeschaltetes Prelimiting noch deutlich gesteigert werden kann. Als Endresultat erhalten wir eine hochgenaue Methode, die keine unphysikalischen Oszillationen und Artefakte aufweist (vgl. Abb. 2.1).

Betrachten wir als nächstes die ebenfalls von zweiter Ordnung genaue semiimplizite Crank-Nicolson Zeitdiskretisierung. Die numerischen Ergebnisse (vgl. Abb. 2.2) weisen mit Ausnahme der weiterhin oszillierenden Methode hoher Ordnung kaum Unterschiede zu denen auf, die mit Hilfe des expliziten Schemas produziert wurden. Da implizite Verfahren ohne zusätzliche Stabilisierung auskommen, größere Zeitschritte zulassen und bessere Stabilitätseigenschaften bieten, werden wir im folgenden auf die Anwendung der vollexpliziten Methode verzichten.

Als letztes wenden wir uns dem vollimpliziten Backward-Euler Zeitschrittverfahren zu, welches nur von erster Ordnung genau ist. Dies zeigt sich darin, daß die in Abbildung 2.3 dargestellten Ergebnisse bei gleichem Zeitschritt diffusiver als CN und LW sind. Wir möchten jedoch darauf hinweisen, daß eine übermäßige Diffusivität nur bei stark zeitabhängigen Problemen hervortritt und auch dort für kleiner werdende Zeitschritte Δt verschwindet. Auf der anderen Seite bewirkt die numerische Diffusion, daß bereits die Methode hoher Ordnung kleinere und 'glattere' Oszillationen als beispielsweise die Lax-Wendroff Methode aufweist. Desweiteren genießen vollimplizite LED Verfahren den Vorteil, daß sie für beliebige Zeitschrittweiten positivitätserhaltend sind (vgl. Positivitätstheorem 1.5.3).

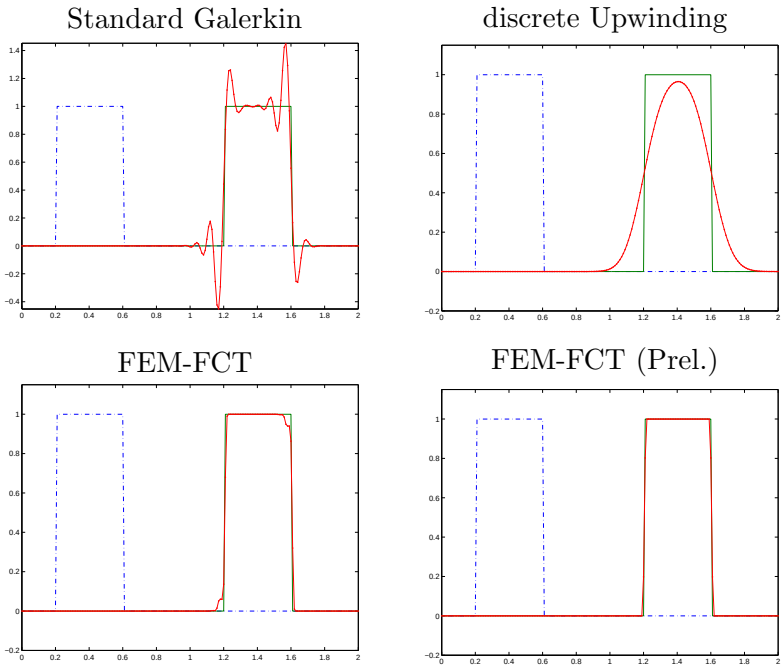


ABBILDUNG 2.1: Konvektion einer *step function*, Lax-Wendroff, $t = 1.0$.

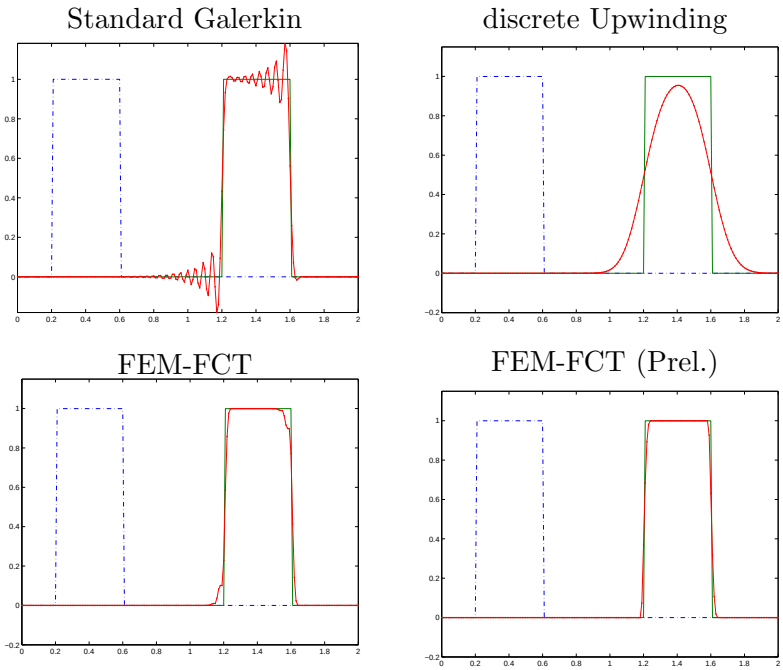


ABBILDUNG 2.2: Konvektion einer *step function*, Crank-Nicolson, $t = 1.0$.

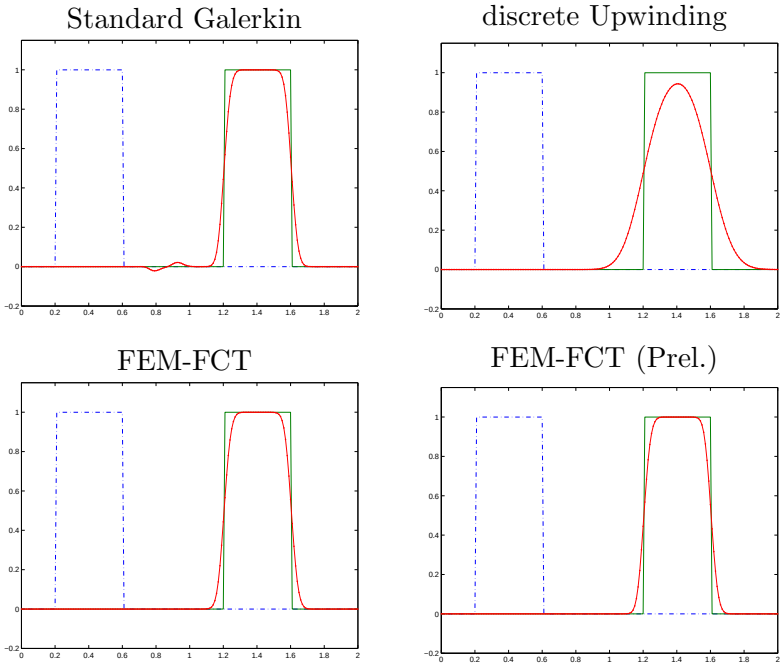


ABBILDUNG 2.3: Konvektion einer *step function*, Backward-Euler, $t = 1.0$.

Konvergenzordnung Nach dieser qualitativen Begutachtung der Ergebnisse wollen wir eine quantitative Untersuchung des Fehlers durchführen und numerisch die Konvergenzordnung des Verfahrens bestimmen. Der Gesamtfehler

$$E_h(\Delta t) = ||e_h||, \qquad e_h = u - u_h \tag{2.8}$$

setzt sich aus dem Ortsfehler E_h zur Gitterweite h sowie dem Zeitfehler $E(\Delta t)$ zusammen. Zur Bestimmung der Konvergenzordnung betrachten wir das Verhalten des Ortsfehlers E_h für $h \to 0$, welcher als Grenzwert des Gesamtfehlers $E_h(\Delta t)$ für $\Delta t \to 0$ definiert ist. Da dieser Grenzwert numerisch nicht exakt bestimmt werden kann, müssen wir die Zeitschrittweite so klein wählen, daß der Anteil des Zeitfehlers zu vernachlässigen ist. Dazu verkleinern wir bei fester Gitterweite h sukzessive den Zeitschritt, bis sich eine fest vorgegebene Anzahl an Dezimalstellen des Fehlers nicht mehr ändert. Exemplarisch ist dies für CN/discrete Upwinding in Tabelle 2.1 für das gröbste Gitter $h = 0.1$ aufgelistet.

Wenn wir eine derartige Elimination des Zeitfehlers für sukzessiv feinere Gitter durchführen, erhalten wir daraus eine sogenannte E–h–Tabelle, die den Ortsfehler als Funktion in der Gitterfeinheit darstellt.

Für hinreichend kleine Gitterweiten nehmen wir an, daß sich der Ortsfehler wie

$$E_h \approx ch^p \tag{2.9}$$

verhält. Daraus ergibt sich bei Halbierung der Gitterweite die Beziehung

$$\frac{E_h}{E_{2h}} \approx 0.5^p, \tag{2.10}$$

aus der sich approximativ die Konvergenzordnung p gemäß

$$p \approx \frac{\log(E_h/E_{2h})}{\log 0.5} \tag{2.11}$$

bestimmen läßt. Wenn wir dies für alle Paare von Ortsfehlerwerten durchführen, erhalten wir eine sogenannte p–h–Tabelle. Diese Vorgehensweise zur numerischen Bestimmung der Konvergenzordnung durch sukzessive Gitterhalbierung wurde von Sokolichin in [71] vorgeschlagen, mit deren Hilfe er eine detaillierte Analyse von eindimensionalen TVD Verfahren durchführt.

Δt	$E_h(\Delta t)$	Δt	$E_h(\Delta t)$
2e-1	0.515300515479162	1e-3	0.420004074557626
1e-1	0.483775946497917	5e-4	0.419481270968687
5e-2	0.458787580582928	2e-4	0.419165887026115
2e-2	0.437591053121022	1e-4	0.419060473438830
1e-2	0.428850897160120	5e-5	0.419007712907444
5e-3	0.424063431617647	2e-5	0.418976039368565
2e-3	0.421039191472726	1e-5	0.418965478650264

Tabelle 2.1: Gesamtfehler als Funktion von Δt für $h = 0.1$.

h	E_h	h	E_h
1.0e-1	4.1895e-1	5.0e-3	1.1279e-1
5.0e-2	3.3707e-1	2.5e-3	7.9771e-2
2.5e-2	2.5000e-1	1.0e-3	5.0458e-2
1.0e-2	1.5943e-1	5.0e-4	3.5680e-2

Tabelle 2.2: Ortsfehler als Funktion von der Gitterweite h.

h	1.0e-1	5.0e-2	2.5e-2	1.0e-2	5.0e-3	2.5e-3	1.0e-3
p	0.3137	0.4311	0.4909	0.4993	0.4997	0.4999	0.4999

Tabelle 2.3: Konvergenzordnung p für unterschiedliche Gitterweiten h.

Zur Visualisierung verwenden wir ferner p–h–Diagramme (Abb. (2.4), links) mit logarithmisch skalierter h–Achse und linear aufgetragener Konvergenzordnung p. Der numerisch bestimmte Wert von $p = 0.5$ für discrete Upwinding stimmt mit dem für klassisches Upwind für diesen Benchmark berechneten Wert [71] überein.

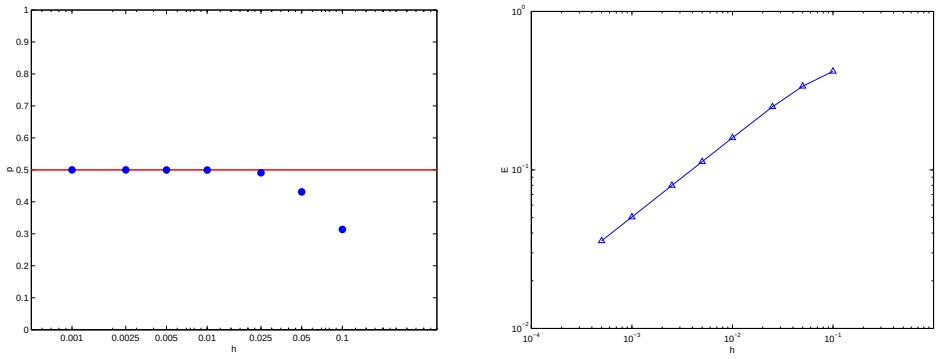


Abbildung 2.4: p–h / E–h–Diagramme für CN/discrete Upwinding.

Als weitere Darstellungsform setzen wir E–h–Diagramme ein, in denen der Ortsfehler gegen die Gitterweite aufgetragen wird, wobei beide Achsen logarithmisch skaliert sind. Bei hinreichend feiner Gitterweite folgt aus (2.8)

$$\log E_h \approx \log c + p \log h, \tag{2.12}$$

woraus sich bei doppeltlogarithmischer Skalierung ein zur Konvergenzordnung p proportionaler, linearer Verlauf des Ortsfehlers ergibt (vgl. Abb. (2.4), rechts). Wir werden auf die ausführliche Darstellung der Zwischenschritte verzichten und lediglich die Endergebnisse in Form von Diagrammen präsentieren.

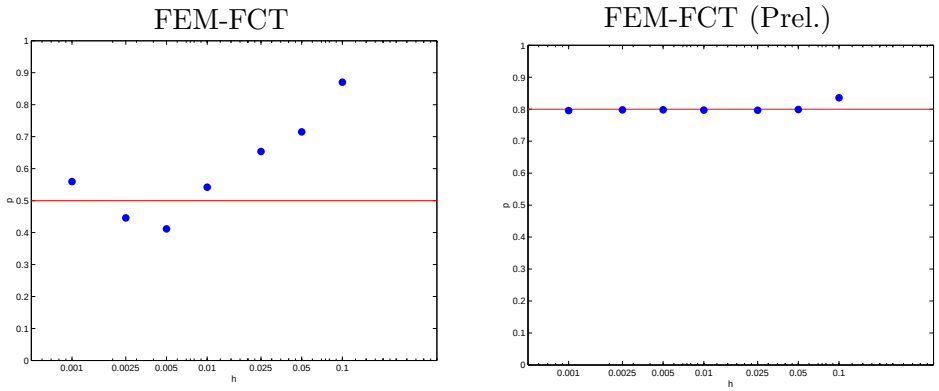


Abbildung 2.5: p–h–Diagramme.

Das in Abbildung 2.5 links dargestellte Diagramm zeigt, daß für FEM-FCT ohne Prelimiting keine vernünftige Bestimmung der Konvergenzordnung möglich ist,

da der Einfluß der 'Eckenausbildung' zu stark ist. Für FEM-FCT mit Prelimiting erhält man hingegen eine Konvergenzordnung von $p = 0.8$, welche damit über den mit TVD Verfahren erreichten Werten von typischerweise $0.6 - 0.7$ für diesen Benchmark [71] liegt. Das zugehörige E–h–Diagramm (Abb. 2.6) verdeutlicht, daß sich der Ortsfehler sowohl für discrete Upwinding als auch für das mit Prelimiting behandelte FEM-FCT linear verhält.

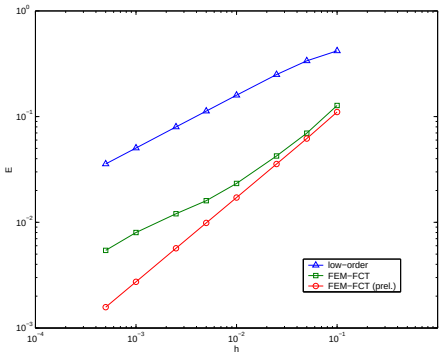


ABBILDUNG 2.6: E–h–Diagramm.

Konvektion-Reaktion Als zweiten Benchmark wollen wir eine Erweiterung des ersten Testfalls betrachten, der durch Hinzunahme eines Quellterms auf die Konvektions-Reaktionsgleichung führt

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} + ru = 0, \tag{2.13}$$

wobei $r > 0$ einen konstanten Reaktionskoeffizienten bezeichnet. Indem wir in der homogenen Gleichung (2.5) die rechte Seite durch den reaktiven Term $-ru$ ersetzen, erhalten wir die analytische Lösung

$$u(x,t) = u^0(x - vt)\ e^{-t^2/2}. \tag{2.14}$$

Die Art des hinzugekommenen Quellterms erlaubt es, ihn wie in Abschnitt 1.6.3 erläutert vollimplizit in der linken Seite zu berücksichtigen. Für die numerische Simulation wurde wie in [8] ein Reaktionskoeffizient von $r = 0.5$ und eine Geschwindigkeit von $v = 1$ gewählt. Die mit discrete Upwinding und FEM-FCT mit einem Zeitschritt von $\Delta t = 10^{-3}$ berechneten Ergebnisse sind in Abbildung 2.7 bis zur Zeit $t = 1.0$ dargestellt. Erwartungsgemäß ist CN/FEM-FCT frei von Oszillationen und liefert eine gute Übereinstimmung mit der exakten Lösung.

Dieser Testfall zeigt insbesondere, daß die Quelltermlinearisierung innerhalb von FEM-FCT durchführbar ist und daß unter Berücksichtigung der modifizierten Zeitschrittbedingung (1.75) die Positivität garantiert werden kann.

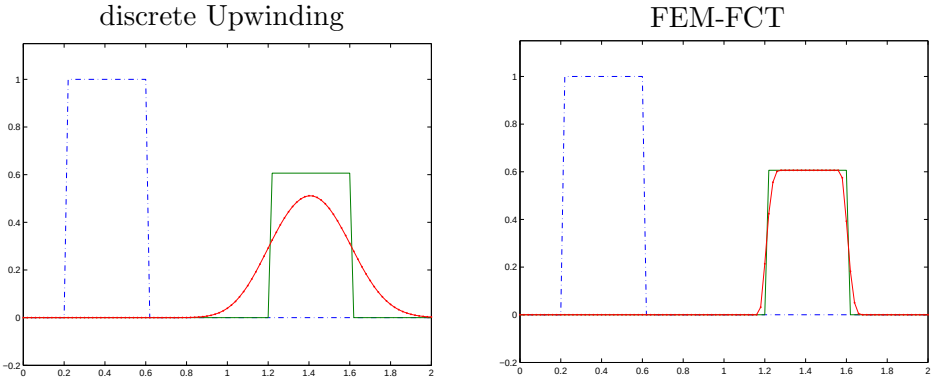


ABBILDUNG 2.7: Konvektion-Reaktion einer *step function*, CN, $t = 1.0$.

Cosinusprofil Im Abschnitt 1.3.2 sind wir auf die Problematik von entstehenden Randoszillationen durch falsch berechnete Schranken im Zalesak-Limiter eingegangen und haben einen Lösungsansatz vorgeschlagen. Diesen wollen wir anhand der Konvektion eines glatten Cosinusprofils mit einer Geschwindigkeit von $v = 1$ numerisch verifizieren. Die Anfangslösung genüge in $[0, 2]$ der Bedingung

$$u(x, 0) = \cos(\pi x) + 1. \tag{2.15}$$

Die numerischen Ergebnisse wurden auf einem Gitter mit 100 linearen Elementen zum Zeitpunkt $t = 0.5$ berechnet. Wie man anhand von Abbildung 2.8 (oben) erkennt, liefert die Standard Galerkin Methode für dieses glatte Profil ausgezeichnete Ergebnisse, während FEM-FCT ohne Postlimiting zu Oszillationen am Ausflußrand führt. Mit Postlimiting stimmen die Ergebnisse für hinreichend kleine Zeitschritte gut mit der analytischen Lösung überein (Abb. 2.8, unten).

Für dieses glatte Anfangsprofil erhält man für discrete Upwinding wie erwartet eine Konvergenzordnung von $p = 1.0$ (Abb. 2.9, oben links). Obwohl sich die numerischen Ergebnisse mit und ohne Prelimiting in der '*picture*-Norm' nicht unterscheiden, zeigt sich ein kleiner Unterschied im zugehörigen p–h–Diagramm, der auch im E–h–Diagramm ersichtlich ist. Weiterhin erkennt man deutlich das lineare Verhalten des Ortsfehlers bezogen auf die Gitterweite h.

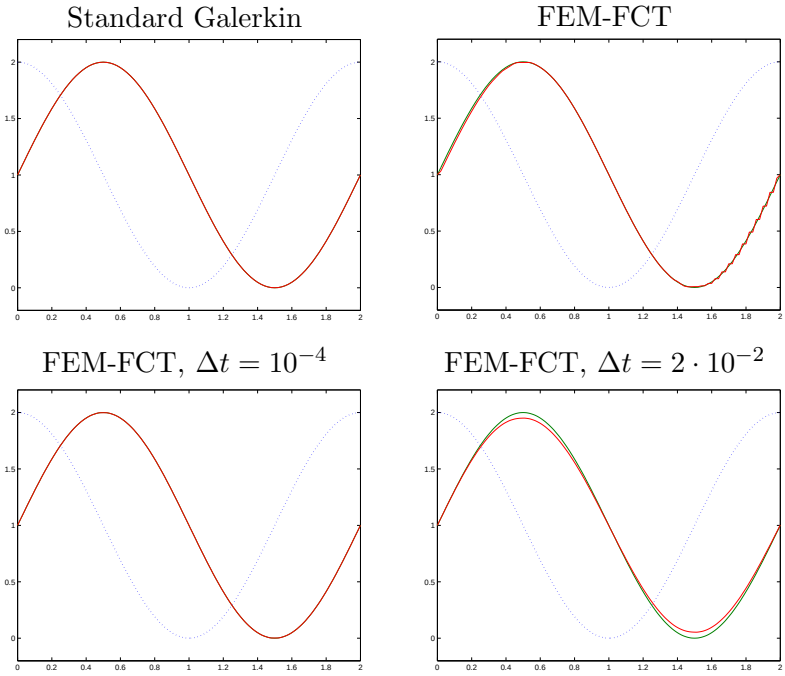


ABBILDUNG 2.8: Konvektion eines Cosinusprofils, CN und BE, $t = 0.5$.

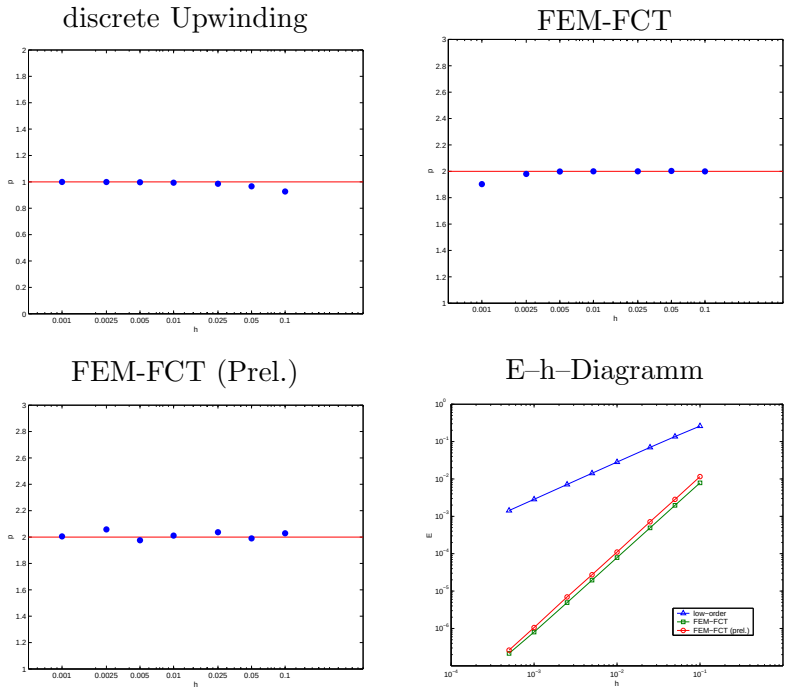


ABBILDUNG 2.9: Konvergenzordnung und Ortsfehler.

Variable Konvektionsgeschwindigkeit Zum Abschluß der Betrachtung von linearen Konvektionsgleichungen wollen wir eine im Ort variable Geschwindigkeit $v = v(x)$ zulassen. Die resultierende Konvektionsgleichung lautet dann

$$\frac{\partial u}{\partial t} + \frac{\partial (vu)}{\partial t} = 0, \tag{2.16}$$

wobei die entsprechende Anfangsbedingung durch

$$u(x, 0) = u^0(x) \tag{2.17}$$

gegeben ist. Die obige Konvektionsgleichung läßt sich auch in der Form

$$\frac{\partial u}{\partial t} + v(x)\frac{\partial u}{\partial x} = -v'(x)u \tag{2.18}$$

darstellen. Analog zum Fall einer konstanten Konvektionsgeschwindigkeit genügen die Charakteristiken der Gleichung (2.16) der Differentialgleichung

$$x'(t) = v(x), \qquad x(0) = x^0. \tag{2.19}$$

Für die Ableitung der Lösung entlang einer Charakteristik folgt daraus

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} u(x(t), t) &= \frac{\partial u(x(t), t)}{\partial t} + \frac{\partial u(x(t), t)}{\partial x} x'(t) \\ &= \frac{\partial u}{\partial t} + v(x)\frac{\partial u}{\partial x} = -v'(x)u, \end{aligned} \tag{2.20}$$

so daß sich mit Hilfe dieser ODE und der Anfangsbedingung $u(x(0), 0) = u^0(x^0)$ die eindeutig bestimmte analytische Lösung berechnen läßt

$$u(x, t) = u^0(xe^{-t})\ e^{-t}. \tag{2.21}$$

'Step function' Wir betrachten erneut das durch (2.6) festgelegte unstetige Anfangsprofil und gehen jetzt von der variablen Konvektionsgeschwindigkeit $v = x$ aus. Dieser Testfall eines nicht gleichmäßigen Geschwindigkeitsfeldes demonstriert das Verhalten der FEM-FCT Methode für die Situation eines physikalischen Anwachsens oder Abnehmens von Extrema. Dazu ist es notwendig, daß der Limiter zwischen physikalischen und unphysikalischen Extrema unterscheiden kann und nur letztere an ihrer Entstehung hindert. Die auf einem gleichmäßigen Gitter von 100 linearen Elementen berechnete Lösung ist in Abbildung 2.10 dargestellt. Sowohl Crank-Nicolson als auch der vollimplizite Euler liefern oszillationsfreie Ergebnisse, die jedoch recht diffusiv sind. Es fällt weiterhin auf, daß die linke Sprungstelle besser erfaßt wird als die rechte, was auf die mit x wachsende Courantzahl zurückzuführen ist. Wir wollen bemerken, daß die übermäßige Diffusivität nicht als Schwäche der FEM-FCT Methode anzulasten ist, sondern daß bereits die Methode hoher Ordnung recht diffusiv ist.

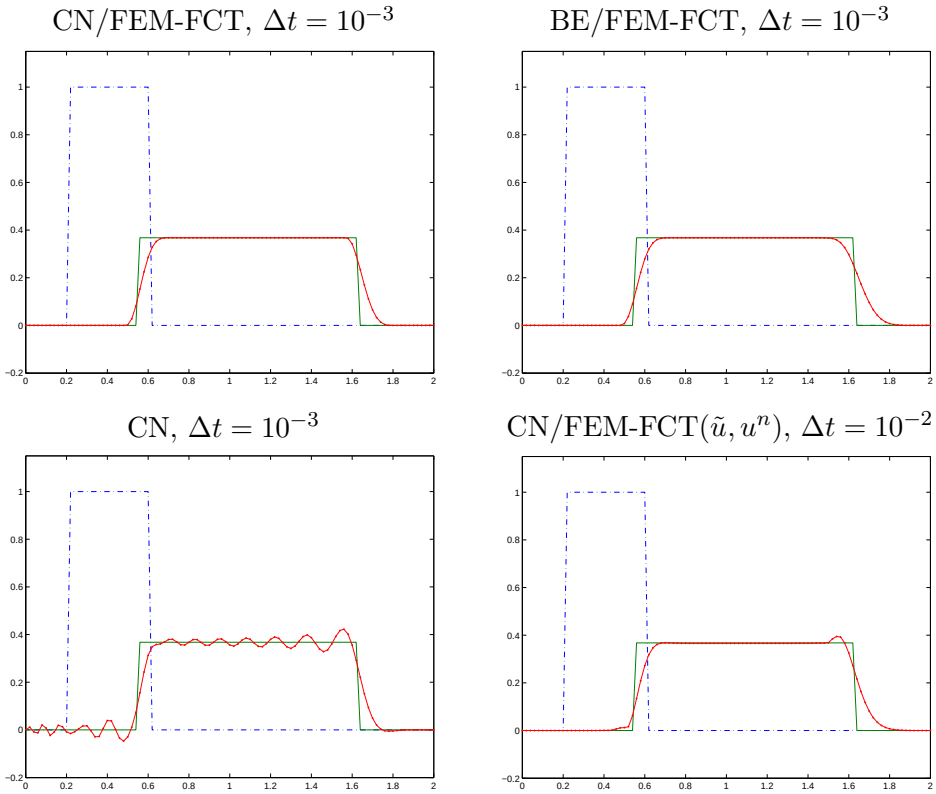


ABBILDUNG 2.10: Variables Geschwindigkeitsfeld $v = x$, $t = 1.0$.

Abbildung 2.10 unten rechts zeigt das Verhalten des Limiters, wenn bei der Berechnung der oberen und unteren Schranken sowohl die monotone Zwischenlösung $\tilde{u}$ als auch die Lösung des letzten Zeitschritts u^n berücksichtigt wird. Zur besseren Verdeutlichung wurde ein größerer Zeitschritt gewählt, so daß die Ausbildung von kleinen 'Hörnern' verstärkt wird. Da für die physikalisch richtige Lösung der Wert des lokalen Extremums an der rechten Flanke in jedem Zeitschritt abnimmt, ist eine obere Schranke, die sich als Maximum von $\tilde{u}$ und u^n ergibt, zu unscharf und läßt kleine Oszillationen entstehen. Dieses Beispiel zeigt, daß es ratsam ist, nur die monotone Zwischenlösung $\tilde{u}$ in die Berechnung der Schranken einfließen zu lassen und die Lösung des letztes Zeitschritts zu 'vergessen'.

Da es sich um ein unstetiges Anfangsprofil handelt, erwarten wir für das von erster Ordnung genaue discrete Upwinding eine Konvergenzordnung von $p = 0.5$, die auch numerisch (vgl. Abb. 2.11) nachgewiesen werden konnte. Für FEM-FCT zeigen die p–h–Diagramme einen nicht ganz zufriedenstellenden Verlauf. Im Mittel liegt die Konvergenzordnung mit und ohne Prelimiting bei $p = 0.88$, was etwas besser als im Falle von Konvektion mit konstanter Geschwindigkeit ist. Dort hat der Prelimiting-Schritt jedoch eine stärkere Auswirkung auf die Konvergenzordnung und auf deren Monotonie. Im E–h–Diagramm erkennt man für alle drei

Methoden (zumindest für hinreichend kleines h) eine nahezu lineare Fehlerreduktion. Darüber hinaus liegt für FEM-FCT mit Prelimiting der Ortsfehler um eine Größenordnung unter dem des Upwind Verfahrens.

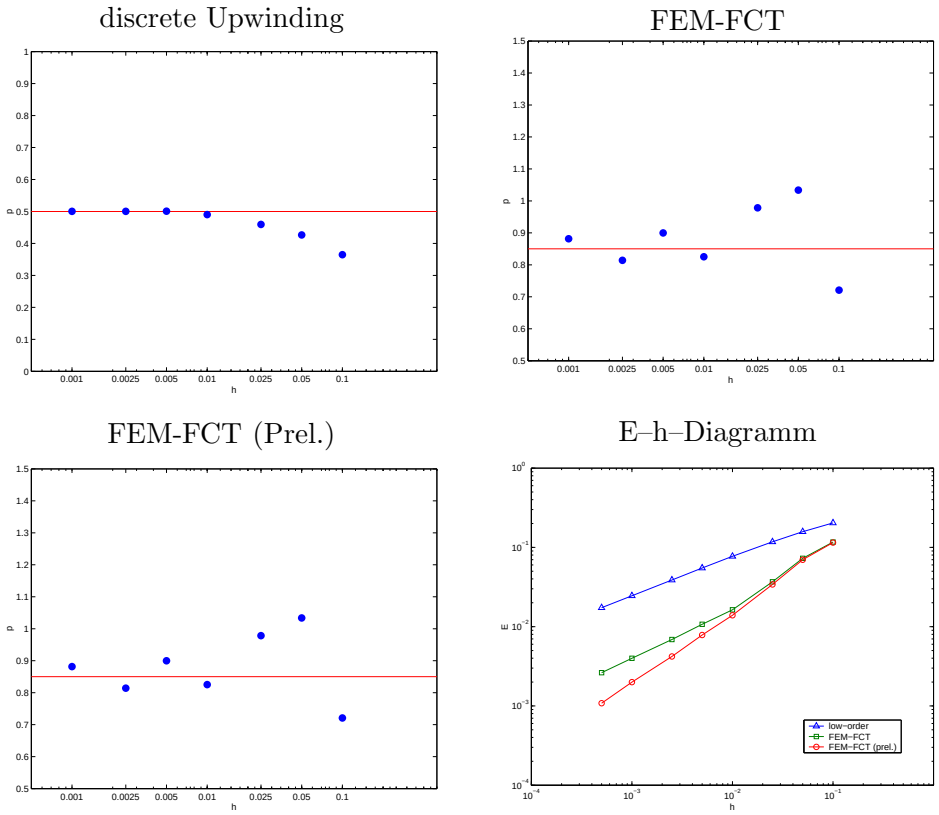


Abbildung 2.11: Konvergenzordnung und Ortsfehler.

2.1.2 Burgers Gleichung

Als zweite Klasse von Benchmarks wollen wir in Hinblick auf die kompressiblen Eulergleichungen hyperbolische Gleichungen der Form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{2.22}$$

betrachten, wobei $f(u)$ eine nichtlinear von u abhängende Funktion ist. Auch wenn diese für die Buckley-Leverett Gleichung (vgl. etwa [46]), ein einfaches Modell zur Simulation von zweiphasigen Strömungen in porösen Medien, nichtkonvex sein kann, gehen wir im folgenden von einer konvexen Funktion $f(u)$ aus, so daß $f''(u) > 0$ gilt. Die Forderung nach Konvexität entspricht bei Systemen der Annahme von 'echter Nichtlinearität' [46].

Im folgenden betrachten wir den wohl bekanntesten Spezialfall

$$f(u) = \frac{1}{2}u^2, \tag{2.23}$$

aus dem man die *nichtviskose Burgers Gleichung* erhält

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0. \tag{2.24}$$

Diese ergibt sich als Grenzwert der von Burgers [6] untersuchten Gleichung

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \epsilon\frac{\partial^2 u}{\partial x^2} \tag{2.25}$$

mit verschwindender Viskositätskonstante ϵ. Mit Hilfe der 1950 von Cole und Hopf entwickelten Transformation kann (2.24) auf die Wärmeleitungsgleichung zurückgeführt und so eine exakte Lösung berechnet werden. Alternativ kann für kleine Zeiten und glatte Anfangslösungen auch die Methode der Charakteristiken verwendet werden, solange sich die einzelnen Charakteristiken nicht überkreuzen. Es gilt wiederum die Differentialgleichung

$$x'(t) = u(x(t), t), \qquad x(0) = x^0, \tag{2.26}$$

aus der analog zum linearen Fall mit konstanter Konvektionsgeschwindigkeit folgt, daß die Lösung u entlang von Charakteristiken konstant ist

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u(x(t), x) &= \frac{\partial u(x(t), t)}{\partial t} + \frac{\partial u(x(t), t)}{\partial x}x'(t) \\ &= \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0. \end{aligned} \tag{2.27}$$

Demnach entsprechen die Charakteristiken eindeutig durch ihre Steigung (2.26) und ihren Anfangswert bestimmten Geraden. Für glatte Anfangsprofile erhält man die analytische Lösung, indem man das Gleichungssystem

$$x(t) = \xi + u(\xi, 0)t \tag{2.28}$$

$$u(x(t), t) = u(\xi, 0) \tag{2.29}$$

zuerst nach ξ und dann nach u auflöst.

'Step function' Wir betrachten das unstetige Anfangsprofil

$$u(x, 0) = \begin{cases} 0, & \text{falls} \quad x \in [0, 0.2] \cup [0.7, 1], \\ 1, & \text{falls} \quad x \in (0.2, 0.7), \end{cases} \tag{2.30}$$

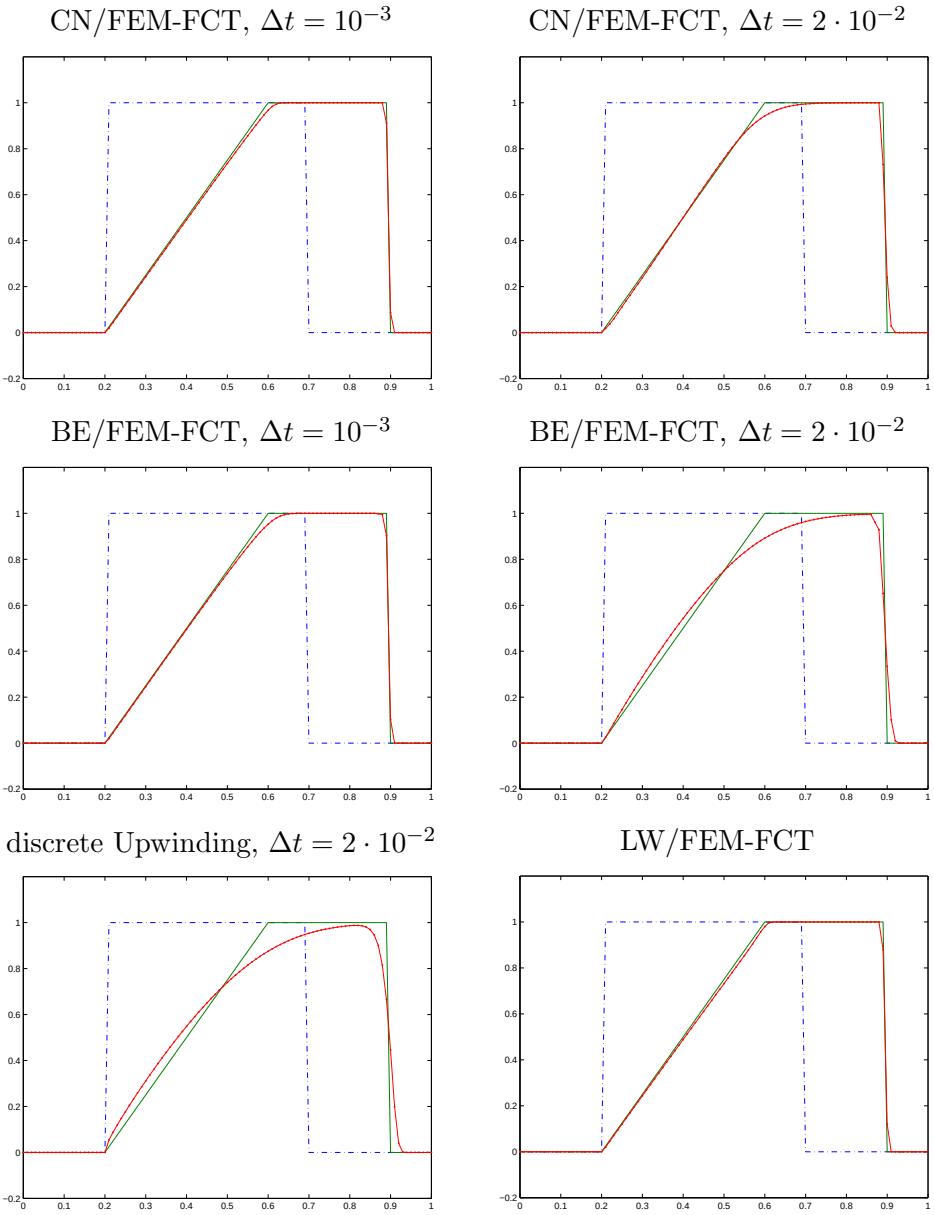


ABBILDUNG 2.12: Nichtviskose Burgers Gleichung, $t = 0.4$.

das bis zum Zeitpunkt $t = 0.4$ konvektiert wird. Dieser Testfall demonstriert das Verhalten des numerischen Verfahrens in Situationen, in denen Schocks entstehen und transportiert werden. Abbildung 2.12 zeigt die numerischen Ergebnisse, die auf einem gleichmäßigen Gitter mit 100 linearen Elementen produziert wurden.

Die oben links dargestellte Lösung verdeutlicht, daß die semi-implizite Zeitdiskretisierung mit moderaten Zeitschritten eine außerordentlich hohe Genauigkeit und gleichzeitig absolut keine Oszillationen aufweist. Selbst das vollimplizite Backward Euler Verfahren zeigt bei dieser Wahl des Zeitschritts nur wenig Diffusion.

Jeweils auf der rechten Seite ist der Lösungsverlauf für einen großen Zeitschritt von $\Delta t = 2h$ dargestellt, der die exakte Lösung trotz der erwarteten Diffusivität qualitativ gut wiedergibt. Wir möchten darauf hinweisen, daß sich der Schock mit der richtigen Geschwindigkeit bewegt, was darauf zurückzuführen ist, daß sowohl discrete Upwinding als auch FEM-FCT die globale Massenerhaltung garantieren. Selbst für das vollexplizite Lax-Wendroff Verfahren sind die Ergebnisse mit und ohne Prelimiting nahezu identisch, so daß in Abbildung 2.12 (unten rechts) nur die mit Prelimiting berechneten Ergebnisse dargestellt sind.

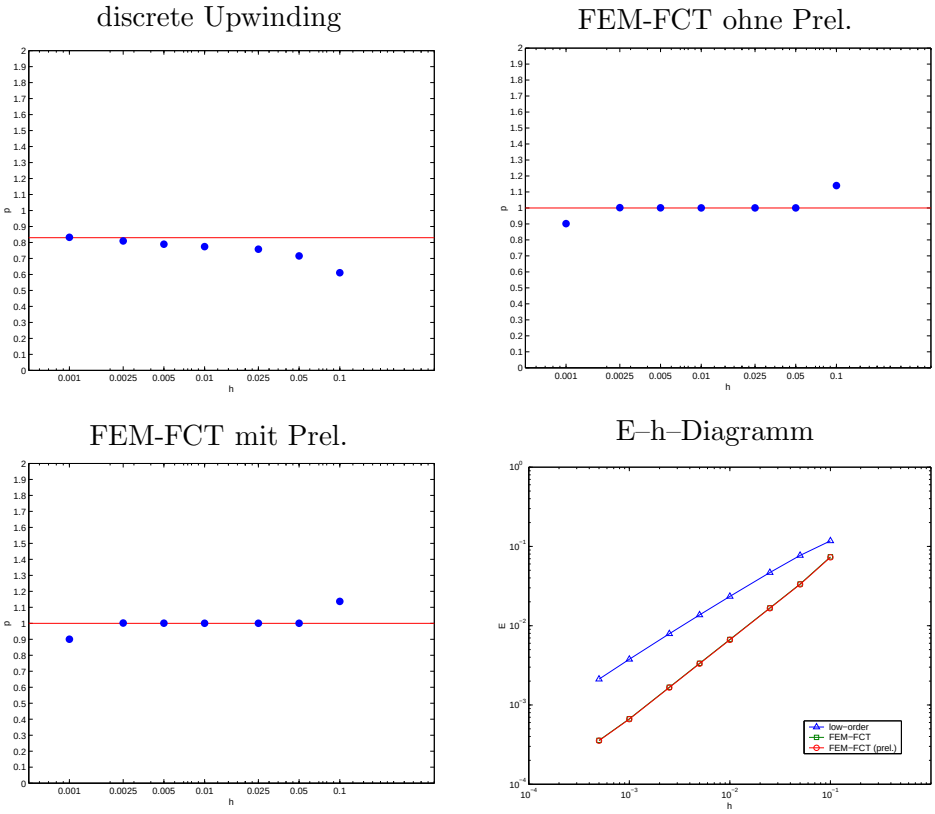


Abbildung 2.13: Konvergenzordnung und Ortsfehler.

Da das Anfangsprofil zunächst unstetig ist und erst mit wachsender Zeit t die linke 'Flanke' geglättet wird, liegt die Konvergenzordnung von discrete Upwinding mit $p = 0.83$ erwartungsgemäß zwischen der von reiner Konvektion eines unstetigen Profils ($p = 0.5$) und der eines glatten Profils ($p = 1.0$). Für FEM-FCT beträgt die Konvergenzordnung mit und ohne Prelimiting $p = 1.0$, was mit den von Sokolichin [71] für TVD Verfahren bestimmten Werten korrespondiert. In Abbildung 2.13 (unten rechts) erkennt man, daß der Ortsfehler bei FEM-FCT um eine ganze Größenordnung unter dem von discrete Upwinding liegt. Weiter findet man auch im E–h–Diagramm keinerlei Unterschiede zwischen den Ergebnissen mit und ohne Anwendung von Prelimiting.

2.1.3 Stationäre Probleme

In den obigen Testfällen haben wir festgestellt, daß das vollimplizite BE/FEM-FCT für transiente Probleme monoton und positivitätserhaltend, aber bei großen Zeitschrittweiten recht diffusiv ist. Diese Diffusion verschwindet für sukzessiv kleineres Δt, so daß der Zeitfehler eliminiert wird. Andererseits eignet sich die Backward Euler Methode hervorragend als iterativer Löser für die zeitunabhängige Konvektions-Diffusions-Gleichung der Form

$$v\frac{\partial u}{\partial x} - \epsilon\frac{\partial^2 u}{\partial x^2} = 0, \qquad u(x,t)|_\Gamma = u^0, \tag{2.31}$$

deren stationäre Lösung man als konvergierte Lösung des entsprechenden zeitabhängigen Problems erhält. Dabei können mögliche Nichtlinearitäten innerhalb derselben Iterationsschleife behandelt werden. Der Zeitschritt spielt hier die Rolle eines künstlichen Relaxationsparameters, der sich auf die Konvergenzrate, jedoch nicht auf die Qualität der konvergierten Lösung auswirkt. Daher ist es durchaus zulässig, eine *local time-stepping*-Strategie [3] anzuwenden, die in der Regel die Konvergenz von expliziten Schemata gegen die stationäre Lösung beschleunigt. Wie bereits erwähnt, hängt die Genauigkeit der stationären Lösung hauptsächlich von der Ortsdiskretisierung ab, so daß man vollimpliziten Methoden mit großen Zeitschritten und damit geringerem Rechenaufwand den Vorzug geben sollte. Gerade hier sind explizite Verfahren mit ihrer meist strengen CFL-Bedingung nur eingeschränkt konkurrenzfähig. Desweiteren wird die Qualität der numerischen Ergebnisse durch die bei der Lax-Wendroff-Methode vorkommende Stromliniendiffusion, die vom künstlichen Zeitschrittparameter abhängt, beeinflußt.

Wir werden für das elliptische Problem (2.31) die Randbedingungen $u(0,t) = 1$ und $u(1,t) = 0$ vorschreiben und mit einer einfachen Approximation für das Anfangsprofil $u(x,0) = 1 - x$ starten. Für die Konvektionsgeschwindigkeit und den Diffusionskoeffizienten setzen wir $v = 1$ und $\epsilon = 10^{-2}$, was einer Pecletzahl von $Pe = 100$ entspricht. Die Herausforderung dieses 'singulär gestörten' Problems besteht in der Auflösung der steilen Flanke am Ausfluß $x = 1$. Die Ausbildung einer Randschicht geht auf die Tatsache zurück, daß die Lösung für den Grenzfall $\epsilon = 0$ die homogene Dirichlet Randbedingung nicht länger erfüllt.

Die numerischen Ergebnisse in Abbildung 2.14 wurden auf einem gleichmäßigen, mit 10 linearen Elementen diskretisierten Rechengebiet berechnet und zeigen das Verhalten des vollimplizten Backward Euler Schemas mit und ohne FCT. Wie in Beispiel 1 gezeigt, reduziert sich die Standard Galerkin Methode auf die zentrale Differenzen Approximation, für die die numerische Lösung auf diesem groben Gitter oszilliert. Dagegen ist die FEM-FCT Lösung in den Knotenpunkten exakt, und selbst discrete Upwinding produziert (hier nicht dargestellte) hervorragende Ergebnisse. Im Gegensatz zu der von Löhner [51], [52] vorgeschlagenen Metho-

de niedriger Ordnung, die auf dem Einsatz konstanter Massendiffusion beruht, wird bei discrete Upwinding die vorhandene physikalische Dissipation berücksichtigt. Bei der Konstruktion des diskreten Diffusionsoperators D, der aus dem Operator hoher Ordnung den Operator der Methode niedriger Ordnung macht ($K+D=L$), wird die fehlende physikalische Diffusion gerade durch genug künstliche Dissipation ausgeglichen, um die Positivitätserhaltung zu garantieren. Als die am wenigsten diffusive monotonieerhaltende Methode ist discrete Upwinding demnach für $\epsilon > 0$ weniger diffusiv als ein klassisches Upwind Verfahren.

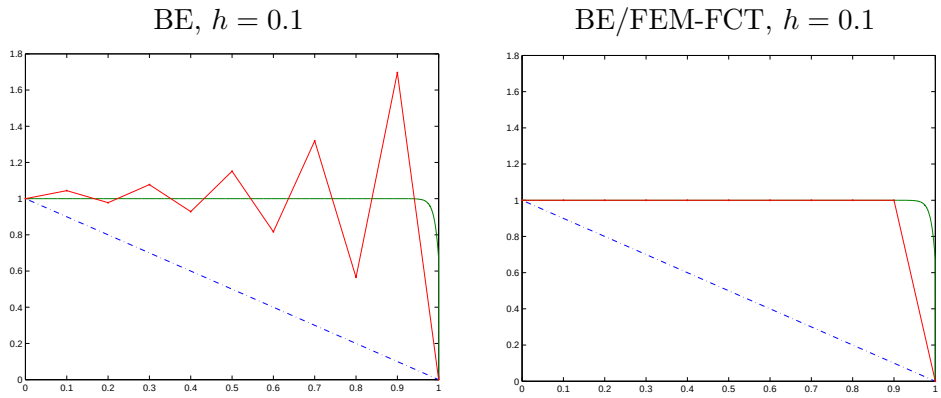


Abbildung 2.14: Stationäre Konvektion-Diffusion mit $\epsilon = 10^{-2}$.

2.2 Mehrdimensionale Benchmarks

Im folgenden Abschnitt wollen wir die neue FEM-FCT Methode auf mehrdimensionale skalare Erhaltungsgleichungen anwenden und damit zeigen, daß es sich im Gegensatz zu diversen Splitting-Verfahren um einen 'echt' mehrdimensionalen Ansatz handelt. Für die zweidimensionalen Verallgemeinerungen eines unstetigen 'Blockprofils' sowie eines glatten Cosinusprofils möchten wir aus Platzgründen auf die numerischen Beispiele in [37] verweisen. Stattdessen wollen wir neue mit FEM-FCT berechnete Simulationsergebnisse vorstellen, die einen Überblick über die Stärken der neuen Methodik geben. Aufgrund der in mehreren Dimensionen um ein Vielfaches größeren Rechenzeiten ist eine numerische Bestimmung der Konvergenzordnung mit vertretbarem Aufwand nicht durchführbar.

2.2.1 Lineare Konvektionsgleichung

Im folgenden betrachten wir die Rotation dreier Festkörper, die von LeVeque [48] als Benchmark für konvektionsdominante Probleme vorgeschlagen wurden. Das Anfangsprofil (vgl. Abb. 2.15) wird dabei gegen den Uhrzeigersinn um den Mittelpunkt des Rechengebiets $\Omega = (0,1)^2$ rotiert, wobei das Geschwindigkeitsfeld als $\mathbf{v} = (0.5 - y, x - 0.5)$ definiert ist. Diese Wahl eines variablen Geschwindigkeitsfeldes führt dazu, daß die lokale Courantzahl mit zunehmendem Abstand vom Mittelpunkt $(0.5, 0.5)$ anwächst. Am Zuflußrand werden bei nach innen gerichteter Normalengeschwindigkeit homogene Dirichlet Randwerte gesetzt.

Den ersten Körper bildet ein glatter Hügel mit dem Mittelpunkt in $(0.25, 0.5)$ und Radius $r^0 = 0.15$, der entsprechend der Vorschrift

$$u(x,y,0) = 0.25 + \cos(\pi\tilde{r}(x,y))/4, \qquad \tilde{r}(x,y) = \min\left\{r(x,y)/r^0, 1\right\} \tag{2.32}$$

erzeugt wird, wobei die Distanzfunktion $r(x,y)$ für alle drei Körper gemäß

$$r(x,y) = \sqrt{(x-x^0)^2 + (y-y^0)^2} \tag{2.33}$$

definiert ist. Desweiteren wird durch

$$u(x,y,0) = 1 - r(x,y)/r^0, \tag{2.34}$$

ein spitzer Kegel mit Mittelpunkt in $(0.5, 0.25)$ festgelegt, der von Smolarkiewicz [69] vorgeschlagen wurde. Den dritten Körper der Konfiguration bildet ein geschlitzter Zylinder mit dem Mittelpunkt in $(0.5, 0.75)$ und

$$u(x,y,0) = \begin{cases} 1, & \text{falls} \quad r(x,y) < r^0 \wedge (|x - x^0| > 0.025 \vee y > 0.85), \\ 0, & \text{sonst.} \end{cases} \tag{2.35}$$

Im Vergleich zu Zalesaks [79] klassischem Zylinder besitzt der obige 'dünnere' Wände und hat das Rotationszentrum weiter außerhalb liegen. Die Herausforderung bei diesem Benchmark besteht darin, die scharfen Kanten des Zylinders und die Spitze des Kegels präzise aufzulösen, und trotzdem die stetigen Übergänge des Hügels nicht durch übermäßige Antidiffusion 'aufzurauhen'.

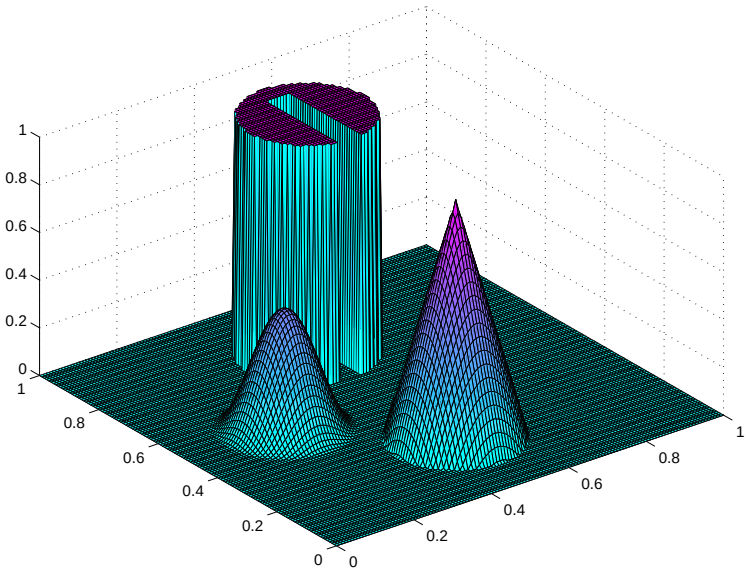


Abbildung 2.15: Anfangsprofil für die Festkörperrotation.

Zur Ortsdiskretisierung wurde ein gleichmäßiges Gitter mit 128×128 Q_1-Elementen verwendet. In den nachfolgenden Abbildungen sind die numerischen Ergebnisse der CN/FEM-FCT und BE/FEM-FCT Methoden nach einer vollen Umdrehung ($t = 2\pi$) dargestellt, so daß die exakte Lösung und das Anfangsprofil zusammenfallen. Mit beiden Zeitdiskretisierungen produziert FEM-FCT oszillationsfreie Lösungen, wobei der Einsatz von Prelimiting anzuraten ist. Ohne diesen Schritt ist die numerische Lösung durch winzige Oszillationen am Schlitz des Zylinders 'verunreinigt'. Wir haben uns in dieser Arbeit auf das konforme Q_1-Element beschränkt und möchten bemerken, daß die Lösung für P_1-Elemente keine Unterschiede aufweist [37]. Um einen fairen Vergleich zwischen dem von erster Ordnung genauen Backward Euler und dem von zweiter Ordnung genauen Crank-Nicolson Verfahren durchführen zu können, wurde der Zeitschritt $\Delta t = 10^{-3}$ gewählt. Um einen Vergleich zwischen den vorgestellten FEM-FCT Verfahren und den von LeVeque [48] eingesetzten TVD Methoden zu ermöglichen, haben wir die Ergebnisse entlang von Cutlines wiedergegeben, wobei die exakte Lösung durch Linien und die numerischen Resultate durch Punkte dargestellt werden.

Die Unstetigkeiten des Zylinders werden durch CN/FEM-FCT sehr gut aufgelöst

(vgl. Abb. 2.16). Die schmale Brücke zwischen den beiden Zylinderhälften bleibt weitgehend erhalten, und lediglich an der Spaltöffnung sind minimale Artefakte zu erkennen, wobei ansonsten kein *fill-in* im Inneren des Zylinders entsteht. *Peak clipping* bewirkt ein Absinken der Kegelspitze um weniger als 10%, wobei zu berücksichtigen ist, daß diese nur durch etwa 25 Gitterpunkte aufgelöst wird. Gleichzeitig wird der glatte Hügel nahezu exakt reproduziert.

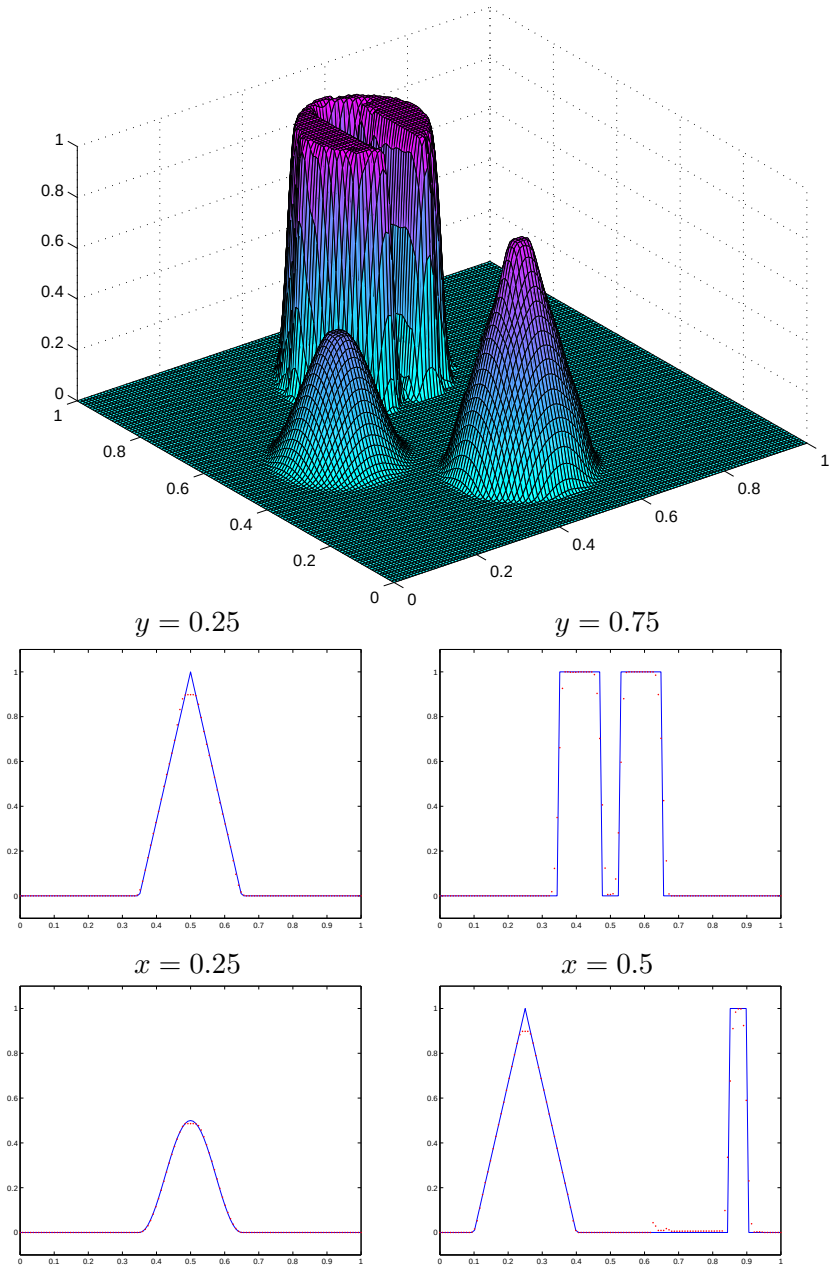


Abbildung 2.16: CN/FEM-FCT zum Zeitpunkt $t = 2\pi$.

Bei Verwendung des iterativen FEM-FCT Algorithmus werden die Kanten des Zylinders schärfer aufgelöst und die Brücke nahezu vollständig erhalten. Schließlich verschwinden die 'kosmetischen' Fehler an der Zylinderöffnung komplett. Der größere Anteil an akzeptierter Antidiffusion bewirkt eine bessere Wiedergabe von Unstetigkeiten, führt aber im Vergleich zu dem bei TVD Methoden gängigen *Superbee*-Limiter nicht dazu, daß glatte Profile steiler gemacht werden.

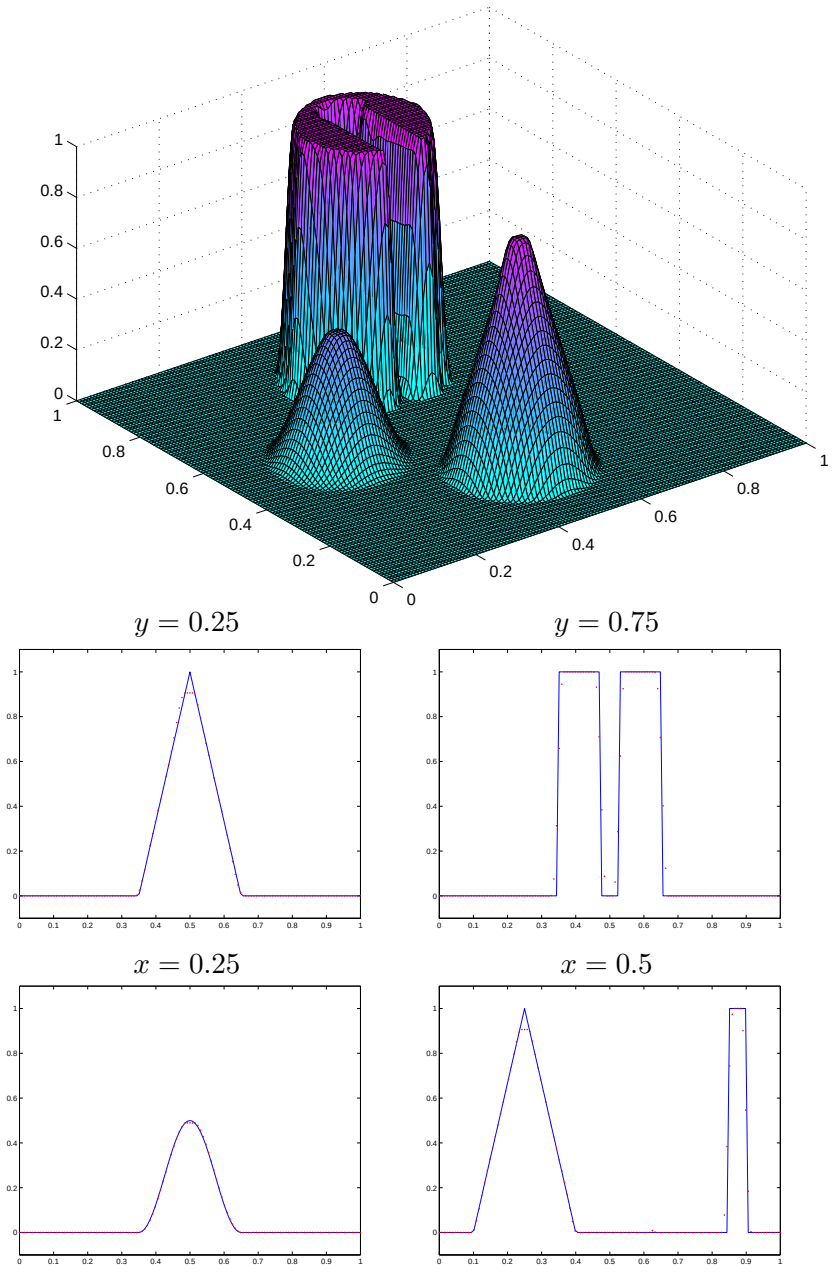


ABBILDUNG 2.17: Iteratives CN/FEM-FCT zum Zeitpunkt $t = 2\pi$.

Das vollimplizite BE/FEM-FCT führt zu einer erkennbaren ‘Abtragung’ des Zylinders (vgl. Abb. 2.18), wobei der maximale Funktionswert dennoch erhalten bleibt. Im Inneren des Schlitzes bildet sich ein *fill-in*, der mit zunehmender Courantzahl (linear) anwächst. Während der maximale Funktionswert in der Kegelspitze kaum von dem mit CN/FEM-FCT berechneten abweicht, werden alle drei Körper jeweils am Boden durch ausgeprägte numerische Diffusion geglättet.

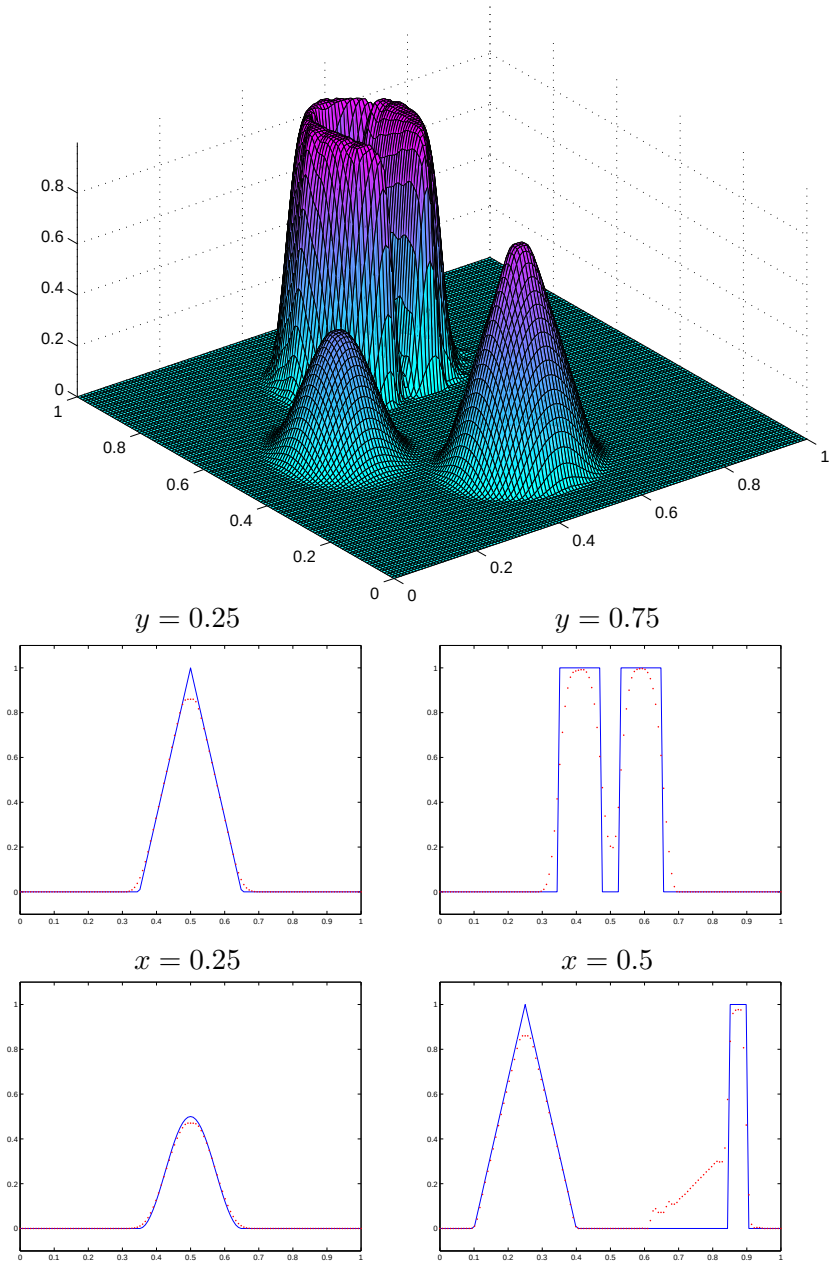


Abbildung 2.18: BE/FEM-FCT zum Zeitpunkt $t = 2\pi$.

Auch für das vollimplizite BE/FEM-FCT führt der iterative Limiter zu leicht weniger diffusiven Ergebnissen, die in Abbildung 2.19 dargestellt sind. Wir möchten jedoch bemerken, daß der Sinn der iterativen Formulierung darin besteht, die Zeitschrittabhängigkeit des Zalesak-Limiters zu umgehen, weshalb sie ihre volle Leistungsfähigkeit erst bei großen Zeitschritten ausspielen kann. Diese werden wir am Beispiel von stationären Problemen in Abschnitt 2.2.3 demonstrieren.

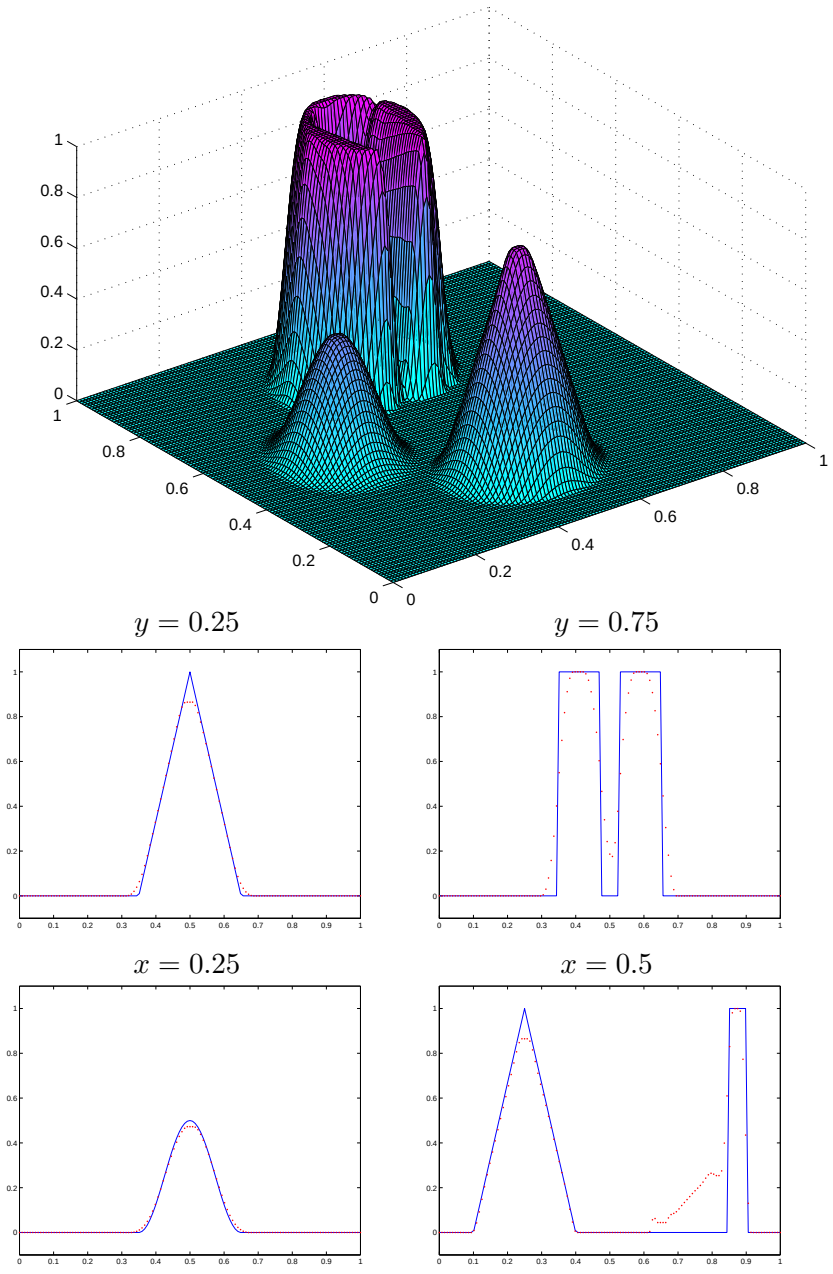


Abbildung 2.19: Iteratives BE/FEM-FCT zum Zeitpunkt $t = 2\pi$.

Um die Anwendbarkeit der neuen FEM-FCT Methode auf beliebigen Gittern zu verifizieren, haben wir den obigen Benchmark auf einem stochastisch gestörten Gitter durchgeführt, von dem ein Ausschnitt in Abbildung 2.20 dargestellt ist. Die Toleranz lag bei 25%, was bei einer Gitterweite von 1/128 einer maximalen Verschiebung jedes Gitterpunktes um etwa 0.002 von seinem Ursprung entspricht.

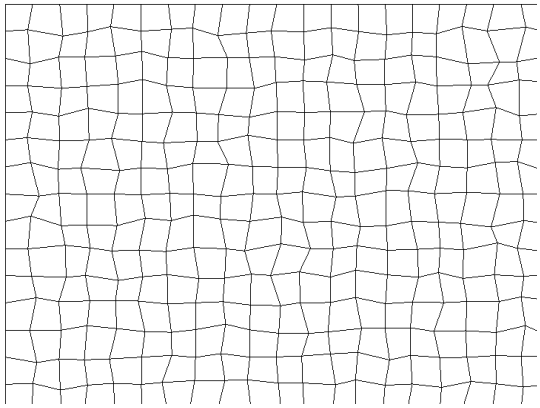


ABBILDUNG 2.20: Stochastisch gestörtes Gitter.

Da sich die numerischen Ergebnisse in der 'picture-Norm' kaum von denen in Abbildung 2.16 unterscheiden, haben wir auf ihre Darstellung verzichtet und lediglich die Lösungsprofile entlang der Cutlines wiedergegeben.

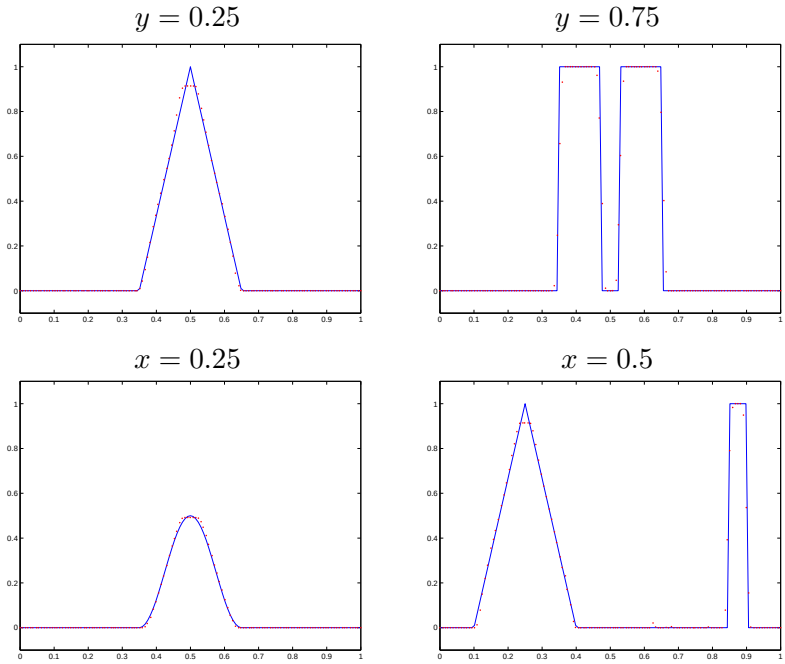


ABBILDUNG 2.21: Iteratives CN/FEM-FCT zum Zeitpunkt $t = 2\pi$.

2.2.2 Lineare Konvektion-Diffusion

Im folgenden wollen wir eine von Lapin [43] vorgeschlagene Methode zur quantitativen Untersuchung des Diffusionsverhaltens eines numerischen Verfahrens vorstellen und auf FEM-FCT anwenden. Dazu betrachten wir eine zweidimensionale Gaußsche Glockenfunktion, die durch eine Rotation um den Ursprung allmählich verschmiert wird. Das Geschwindigkeitsfeld im Rechengebiet $\Omega = (-1,1)^2$ ist als $\mathbf{v} = (-y, x)$ definiert.

In einem rotierenden Lagrangeschen Bezugssystem reduziert sich die zugrundeliegende Konvektions-Diffusionsgleichung

$$\frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u = \epsilon \Delta u \tag{2.36}$$

zu einem reinen Diffusionsproblem. Die analytische Lösung entspricht in ihrer Struktur der Gaußschen Normalverteilung

$$u(x,y,t) = \frac{1}{4\pi\epsilon t} \exp\left(\frac{-r^2}{4\epsilon t}\right), \qquad r^2 = (x-\hat{x})^2 + (y-\hat{y})^2 \tag{2.37}$$

und beschreibt die Wahrscheinlichkeitsdichte der normalverteilten Variablen x und y mit den Mittelwerten $\hat{x}$ und $\hat{y}$. Diese sind in unserem Fall zeitabhängig und entsprechen den Koordinaten das maximalen Funktionswertes

$$\hat{x}(t) = \hat{x}^0 \cos t - \hat{y}^0 \sin t, \qquad \hat{y}(t) = -\hat{x}^0 \sin t + \hat{y}^0 \cos t. \tag{2.38}$$

Der Zusammenhang zwischen Diffusionsgleichung und Normalverteilung wird offensichtlich, wenn wir die normalverteilte Häufigkeit an Kollisionen zwischen Teilchen betrachten und die dadurch induzierte Bewegung in Richtung einer geringeren Kollisionshäufigkeit als Diffusion verstehen.

Das Anfangsprofil wird durch die Dirac Deltadistribution $u(x,y,0) = \delta(\hat{x}^0, \hat{y}^0)$ festgelegt, welche in einer praktischen Implementierung nicht vorgeschrieben werden kann. Ein mögliches Vorgehen besteht darin, das Diracfunktional zu approximieren, indem die gesamte Masse in einem Punkt konzentriert wird. Für eine diskrete Funktion läßt sich das Gebietsintegral als Summe der Knotenwerte multipliziert mit den Einträgen der 'gelumpten' Massenmatrix berechnen

$$\int_\Omega u_h \, \mathrm{d}\mathbf{x} = \int_\Omega \sum_i u_i \varphi_i \, \mathrm{d}\mathbf{x} = \sum_i u_i m_i. \tag{2.39}$$

Es bezeichne P_i den zu $(\hat{x}^0, \hat{y}^0)$ nächstgelegenen Gitterpunkt. Da die Gesamtmasse der Deltadistribution gerade Eins ist, setzen wir als approximierte Anfangslösung

$$u^0(P_i) = 1/M_L(P_i), \qquad u^0(P_j) = 0 \quad \text{für} \quad j \neq i. \tag{2.40}$$

Genaugenommen weichen die Koordinaten des Maximalwertes leicht von denen in (2.38) ab und können als Schwerpunkt der numerischen Lösung

$$\hat{x}_h(t) = \int_\Omega x u_h(x,y,t)\mathrm{d}\mathbf{x}, \qquad \hat{y}_h(t) = \int_\Omega y u_h(x,y,t)\mathrm{d}\mathbf{x} \tag{2.41}$$

berechnet werden. Die Approximationsgüte kann durch die Standardabweichung

$$\sigma_h^2(t) = \int_\Omega r_h^2 u_h(x,y,t)\mathrm{d}\mathbf{x}, \qquad r_h^2 = (x-\hat{x}_h)^2 + (y-\hat{y}_h)^2 \tag{2.42}$$

gemessen werden, die sich aus der physikalischen und der numerischen Diffusion zusammensetzt und für die exakte Lösung $\sigma^2 = 4\epsilon t$ beträgt. Die Differenz aus den Varianzen von exakter und numerischer Lösung liefert ein Maß zur Quantifizierung der numerischen Diffusion [43].

Um den numerischen Fehler in den Anfangsdaten weitgehend zu minimieren, beginnen wir mit der exakten Lösung nach einer Viertelumdrehung ($t^0 = \pi/2$), so daß der Mittelpunkt des Anfangsprofils im Punkt $(-0.5, 0)$ liegt. Die exakte Lösung für $\epsilon = 10^{-3}$ nach einer kompletten Umdrehung ($t = 5\pi/2$) ist in Abbildung 2.22 (links) dargestellt. Die rechte Abbildung zeigt die mit dem semi-impliziten CN/FEM-FCT berechneten Ergebnisse bei einem Zeitschritt von $\Delta t = 10^{-3}$ auf einem gleichmäßigen Gitter mit 128×128 Q_1-Elementen. Während sich die numerischen Ergebnisse für den iterativen und den Basis Limiter in der 'picture-Norm' nicht unterscheiden, treten leichte Unterschiede bezüglich des maximalen Funktionswertes hervor. Für den leicht diffusiven Basis Limiter gilt $||u||_\infty = 10.1291$, wohingegen der etwas antidiffusive iterative Limiter einen Wert von $||u||_\infty = 10.1519$ liefert.

exakte Lösung, $||u||_\infty = 10.1356$ CN/FEM-FCT, $\Delta t = 10^{-3}$

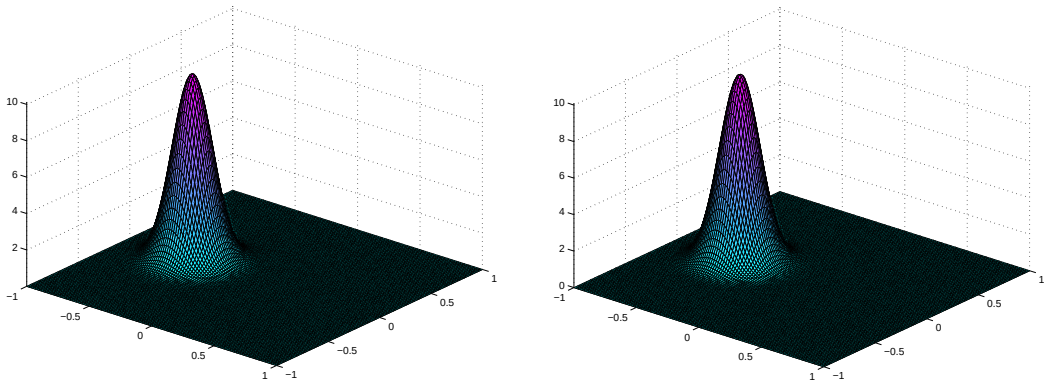


Abbildung 2.22: Konvektion-Diffusion einer Gaußschen Glockenfunktion.

Wir wollen im folgenden die Abhängigkeit der numerischen Varianz vom verwendeten Zeitschritt untersuchen. In Abbildung 2.23 betrachten wir dazu das

Verhalten des relativen Fehlers

$$\Delta\sigma_{\text{rel}} = \frac{\sigma_h^2 - \sigma^2}{\sigma^2} = \frac{\sigma_h^2}{4\epsilon t} - 1 \tag{2.43}$$

für unterschiedliche Zeitschrittweiten im Bereich von 10^{-3} bis 10^{-2}. Die Backward Euler Methode ist nur von erster Ordnung in der Zeit genau, was sich anhand einer starken Diffusivität zeigt. Diese nimmt linear mit der Größe von Δt ab, so daß die vollimplizite Diskretisierung für kleine Zeitschritte die Genauigkeit des von zweiter Ordnung genauen Crank-Nicolson Verfahrens erreicht.

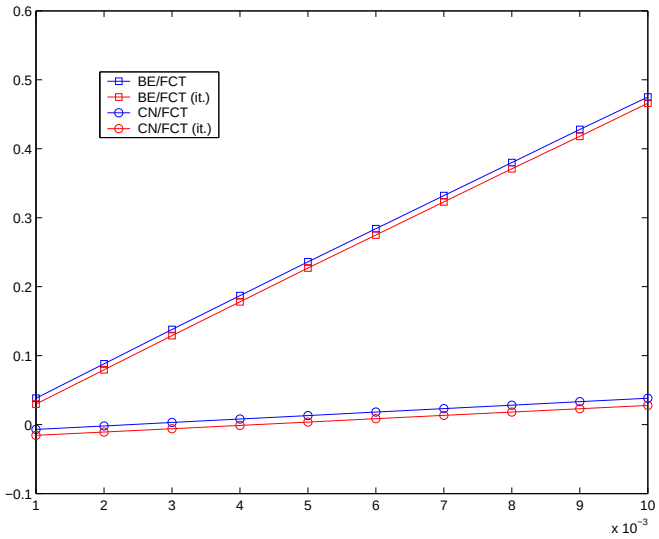


Abbildung 2.23: Relativer Fehler der Varianz als Funktion von Δt.

Man erkennt ebenfalls, daß die Basis FEM-FCT Methode für stark zeitabhängige Probleme und bei kleinen Zeitschrittweiten nur geringfügig diffusiver als die iterative FEM-FCT Formulierung ist. Dies zeigt umsomehr, daß sich die neue Formulierung ohne die Einschränkungen des zeitschrittabhängigen Zalesak-Limiters hauptsächlich für stationäre Probleme empfiehlt, während sie keine übermäßigen Vorteile für zeitabhängige Probleme und kleine Zeitschritte mit sich bringt. Als letzten Benchmark werden wir ein stationäres Problem betrachten, für das sich die iterative Formulierung als 'Wunderwaffe' herausstellen wird.

2.2.3 Stationäre Probleme

Im folgenden wollen wir die zweidimensionale Verallgemeinerung einer 'singulär gestörten' Konvektions-Diffusionsgleichung der Form

$$\mathbf{v} \cdot \nabla u - \epsilon \Delta u = 0, \qquad \text{in} \quad \Omega = (0,1)^2 \tag{2.44}$$

betrachten. Das Geschwindigkeitsfeld ist dabei durch $\mathbf{v} = (\cos 10°, \sin 10°)$ gegeben. Das Problem wird durch die Randbedingungen

$$\frac{\partial u}{\partial y}(x,1) = 0, \qquad u(1,y,t) = 0 \tag{2.45}$$

$$u(x,0,t) = 0, \qquad u(0,y,t) = \begin{cases} 1, & \text{falls} \quad y \geq 0.5, \\ 0, & \text{falls} \quad y < 0.5 \end{cases} \tag{2.46}$$

vervollständigt. Wie schon im entsprechenden eindimensionalen Testproblem erfüllt die Lösung des reduzierten Problems ($\epsilon = 0$) nicht die homogene Dirichlet Randbedingung, so daß sich bei $x = 1$ eine steile Flanke ausbildet. Als Anfangsprofil für das Pseudo-Zeitschrittverfahren wählen wir die lineare Approximation

$$u(x,y,0) = \begin{cases} 1 - x, & \text{falls} \quad y \geq 0.5 \\ 0, & \text{falls} \quad y < 0.5. \end{cases} \tag{2.47}$$

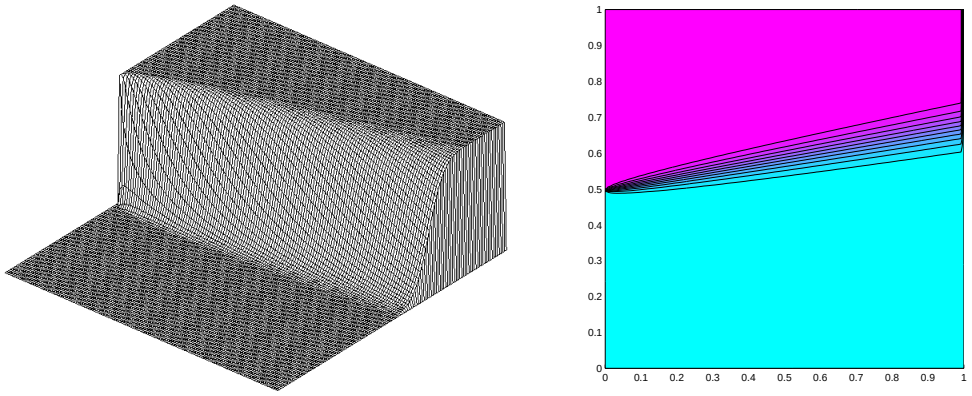
Da es sich bei dem obigen Testfall um eine stationäre Problemstellung handelt, ist es in der Praxis sinnvoll, zunächst eine konvergierte Lösung mit Hilfe der Methode niedriger Ordnung zu berechnen und anschließend die Genauigkeit durch den Einsatz von FEM-FCT zu erhöhen. Für realistische Anwendungen erweist sich dieser '*educated guess*' als sehr hilfreich und kann einen beträchtlichen Rechenzeitgewinn liefern. Desweiteren ist der Gebrauch der konsistenten Massenmatrix für stationäre Probleme nicht länger gerechtfertigt, so daß für die Methode hoher Ordnung die günstige 'gelumpte' Massenmatrix zum Einsatz kommen sollte.

Die in Abbildung 2.24 dargestellten numerischen Ergebnisse wurden auf einem gleichmäßigen Gitter von 128×128 Q_1-Elementen mit dem vollimpliziten Backward Euler Verfahren berechnet. Der Zeitschritt wurde mit $\Delta t = 0.1$ absichtlich groß gewählt, um die Unterschiede zwischen der (zeitschrittabhängigen) Basis Formulierung und dem iterativen Limiter zu verdeutlichen. Weiterhin beträgt der Wert des Diffusionskoeffizienten $\epsilon = 10^{-3}$. Die beiden oberen Diagramme zeigen die mit der (nicht iterativen) Basis Formulierung berechnete Lösung. Diese ist absolut frei von Oszillationen und liefert eine zufriedenstellende Auflösung der Randschicht. Dennoch ist der Übergang zwischen dem Bereich mit Funktionswert Eins und dem mit Funktionswert Null recht diffusiv, was auf die Zeitschrittabhängigkeit des Zalesak-Limiters zurückzuführen ist.

Abbildung 2.24 (unten) gibt die mit dem iterativen Limiter berechneten numerischen Ergebnisse wieder, in denen die Front mit einer wesentlich besseren

Auflösung reproduziert wird. Vergleichbare Ergebnisse erhält man mit der Basis Formulierung nur für deutlich kleinere Zeitschritte ($\Delta t = 10^{-3}$). Insbesondere erkennt man einen strukturellen Unterschied. Während der iterative Limiter die Front mit einer gleichmäßigen Breite erfaßt, wird das Gefälle in der Basis Formulierung stärker verschmiert.

Uniformes Gitter, Basis Limiter



Uniformes Gitter, iterativer Limiter

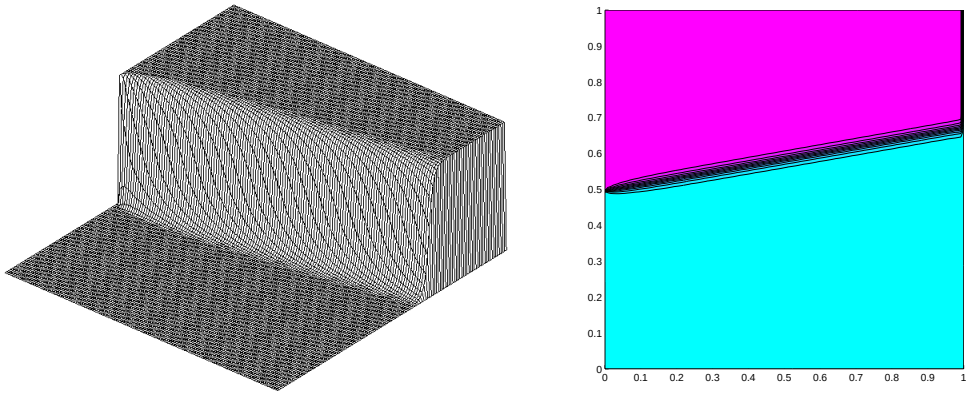


ABBILDUNG 2.24: Stationäre Konvektion-Diffusion in 2D, $\epsilon = 10^{-3}$.

Um die Anwendbarkeit der neuen FEM-FCT Verfahren auf nicht gleichförmigen Gittern zu demonstrieren, haben wir ein adaptives, dem Strömungsverlauf manuell angepaßtes Grobgitter verwendet. Nach vierfacher Verfeinerung erhält man daraus das aus 1920 Q_1-Elementen bestehende Rechengitter (vgl. Abb. 2.2.3).

Die mit dem iterativen Limiter auf diesem Gitter berechnete Lösung ist in Abbildung 2.26 wiedergegeben. Zur Visualisierung des nicht uniformen Gitters mußte auf die Software GMV zurückgegriffen werden, so daß sich die Darstellung leicht von den übrigen Plots unterscheidet. Trotz der um den Faktor 8 größeren Anzahl

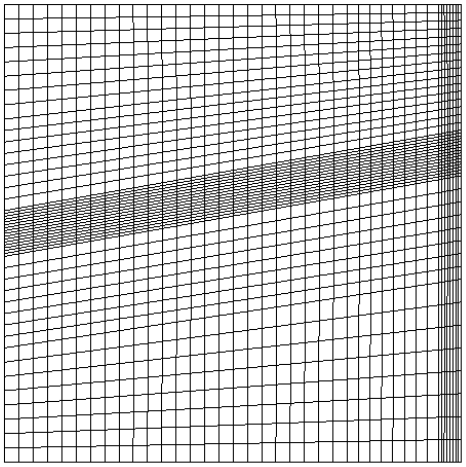


Abbildung 2.25: Adaptives Grobgitter für stationäre Konvektion-Diffusion.

an Elementen im uniformen Gitter, weisen die auf dem adaptiven Gitter produzierten Ergebnisse eine wesentlich höhere Genauigkeit auf, und das bei einem deutlich geringeren Rechenaufwand. Bei Verwendung des Basis Limiters erhält man nahezu identische Ergebnisse, so daß wir auf ihre Darstellung verzichten können.

Nicht uniformes Gitter, iterativer Limiter

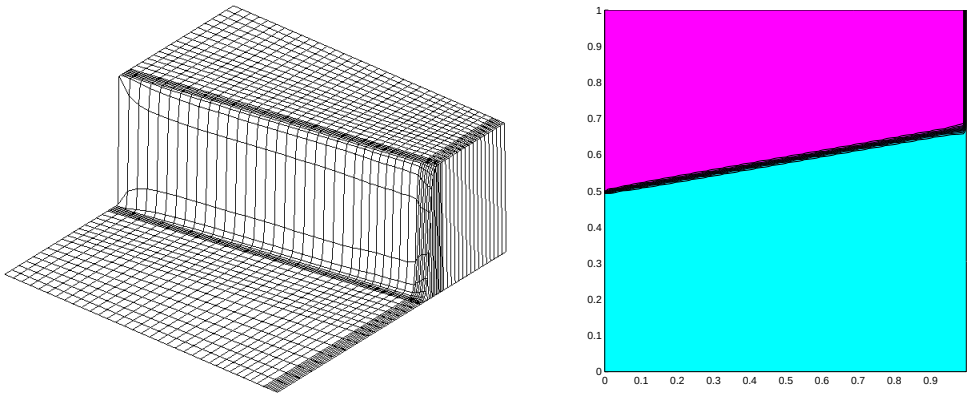


Abbildung 2.26: Stationäre Konvektion-Diffusion in 2D, $\epsilon = 10^{-3}$.

Kapitel 3

Die Grundgleichungen der Strömungsmechanik

Unter einem Fluid versteht man ein Medium, das aufgrund seiner molekularen Struktur keinerlei Widerstand gegenüber äußeren Kräften ausübt, so daß bereits kleinste Krafteinwirkungen zu Deformationen führen. Üblicherweise kann ein Fluid als ein Kontinuum angenommen werden, dessen Strömungsverhalten durch die Anwendung von Oberflächen- und Körperkräften bestimmt wird. Die Schwerkraft zählt etwa zur letztgenannten Klasse, wohingegen die bei der Umströmung einer rauhen Oberfläche entstehenden Druck- und Scherkräfte ein Beispiel für die erstgenannte sind. Während verschiedene Fluide ein weitgehend ähnliches Verhalten in Bezug auf Krafteinwirkungen aufweisen, unterscheiden sich ihre makroskopischen Eigenschaften wie Dichte und Viskosität erheblich.

Wir werden die Verallgemeinerung der skalaren FEM-FCT Methodik auf hyperbolische Gleichungssysteme am Beispiel der kompressiblen Eulergleichungen durchführen. Die Eigenschaften eines Fluides hängen stark von seiner Geschwindigkeit bzw. der Machzahl M ab, die das Verhältnis aus Strömungs- und Schallgeschwindigkeit angibt. Allgemein spricht man ab $M > 0.3$ von Kompressibilität. Häufig wirkt sich die Viskosität nur in Randnähe auf das Strömungsverhalten aus, so daß im Inneren von einem reibungsfreien Fluid ausgegangen werden kann. Die Eulergleichungen werden zur Modellierung von kompressiblen Strömungen bei hohen Machzahlen eingesetzt, für die die Auswirkungen von Körperkräften und viskosen Spannungen vernachlässigt werden können.

3.1 Die konservative Formulierung

Die Kontinuumsmechanik befaßt sich mit Deformation und Bewegung von kontinuierlichen Körpern, die wir als zusammenhängende kompakte Mengen von Partikeln $\mathcal{X}$ verstehen. Es gibt im wesentlichen zwei Möglichkeiten, eine Feldgröße u darzustellen: die Lagrangesche und die Eulersche Betrachtungsweise. Die erste Darstellungsform legt eine ausgezeichnete Referenzkonfiguration zugrunde und wird in der Regel im Zusammenhang mit Festkörperproblemen verwendet, während die zweite Betrachtungsweise in einem festen Bezugssystem gilt und für die Behandlung von Fluiden besser geeignet ist. Bei Verwendung der Lagrangeschen Darstellung spricht man von der materiellen Zeitableitung $\mathrm{d}u/\mathrm{d}t$, welche die Änderung von u aus der Sicht eines mit einem Teilchen $\mathcal{X}$ mitwandernden Beobachters beschreibt. Unter der lokalen Zeitableitung versteht man die partielle Zeitableitung in räumlicher Darstellung $\partial u/\partial t$, welche die Änderung von u für einen im Ort stationären Beobachter wiedergibt. Wenn wir die Geschwindigkeit $\mathbf{v}$ als $\mathrm{d}x/\mathrm{d}t$ definieren, so ergibt sich mit Hilfe der Kettenregel der Zusammenhang

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u, \tag{3.1}$$

wobei sich die materielle Zeitableitung aus einem lokalen und einem konvektiven Term zusammensetzt. Wir wollen ein Volumen als materiell bezeichnen, wenn es stets aus den gleichen Partikeln besteht. Betrachten wir die zeitliche Änderung unserer Feldgröße u über Ω, so liefert das Reynoldssche Transporttheorem

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega u \,\mathrm{d}\Omega = \frac{\partial}{\partial t} \int_\Omega u \,\mathrm{d}\Omega + \int_\Gamma u\mathbf{v} \cdot \mathbf{n} \,\mathrm{d}s, \tag{3.2}$$

wobei Γ die Oberfläche von Ω und $\mathbf{n}$ den äußeren Normaleneinheitsvektor auf Γ bezeichnen. Gleichung (3.2) gilt auch für nicht-materielle Volumina, wie sie in der Fluiddynamik auftreten, wenn Ω ein Kontrollvolumen bezeichnet, dessen Rand Γ sich mit der Geschwindigkeit $\mathbf{v}$ bewegt. Das Ziel besteht darin, die Bewegung eines Körpers oder Feldes unter Vorgabe von Anfangs- und Randbedingungen aus den Feldgleichungen zu berechnen. Dazu werden sowohl die Bilanzgleichungen für Masse, Impuls und Energie als auch die materialspezifischen Konstitutivgesetze benötigt. Wir werden letztere direkt mit den für die Anwendung ausreichenden Vereinfachungen herleiten. Die zeitliche Änderung einer physikalischen Feldgröße $u = \rho\varphi$ mit Dichte ρ in einem materiellen Volumen Ω setzt sich aus dem Molekulartransport und den Oberflächeneffekten sowie den Beiträgen von Quellen und Senken zusammen. Aus (3.2) erhalten wir die integrale Form der Bilanzgleichung

$$\frac{\partial}{\partial t} \int_\Omega \rho\varphi \,\mathrm{d}\Omega + \int_\Gamma \rho\varphi\mathbf{v} \cdot \mathbf{n} \,\mathrm{d}s = \int_\Gamma \sigma \cdot \mathbf{n} \,\mathrm{d}s + \int_\Omega \rho f \,\mathrm{d}\Omega, \tag{3.3}$$

wobei $\sigma \cdot \mathbf{n}$ die Normalenflußdichte der Oberflächenkräfte bezeichnet, und f für die Intensität der spezifischen Körperkräfte steht.

3.1.1 Erhaltungsprinzip der Masse

Das Massenerhaltungsprinzip besagt, daß *Masse weder erzeugt noch vernichtet* werden kann. Damit erfüllt die (Massen-)Dichte ρ die sich aus (3.3) mit $\varphi = 1$ ergebende Kontinuitätsgleichung

$$\frac{\partial}{\partial t}\int_\Omega \rho \,\mathrm{d}\Omega + \int_\Gamma \rho \mathbf{v}\cdot\mathbf{n}\,\mathrm{d}s = 0. \tag{3.4}$$

Da die obige Gleichung für beliebige Kontrollvolumina gültig ist, erhält man mit Hilfe des Gaußschen Integralsatzes unter der Annahme von hinreichnender Stetigkeit die differentielle Form

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = 0. \tag{3.5}$$

3.1.2 Erhaltungsprinzip des Impulses

Newtons zweites Gesetz besagt, daß die *Impulsänderung eines Fluids gleich der Summe aller einwirkenden Kräfte ist*. Die Anwendung von (3.3) auf die einzelnen Komponenten des Geschwindigkeitsvektors $\mathbf{v}$ führt bei Konkretisierung des Oberflächenanteils auf den Stressterm τ und Einschränkung der spezifischen Körperkräfte $\mathbf{f}$ auf die Erdbeschleunigung $\mathbf{g}$ zu

$$\frac{\partial}{\partial t}\int_\Omega \rho\mathbf{v}\,\mathrm{d}\Omega + \int_\Gamma \rho\mathbf{v}\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}S = \int_\Gamma \tau\cdot\mathbf{n}\,\mathrm{d}S + \int_\Omega \rho\mathbf{g}\,\mathrm{d}\Omega. \tag{3.6}$$

Unter der Annahme eines newtonschen Fluids ergibt sich für den Streßtensor, der die molekulare Transportrate des Impulses beschreibt, die Darstellung

$$\tau = -\left(p + \frac{2}{3}\mu\nabla\cdot\mathbf{v}\right)\mathcal{I} + \mu\mathcal{D}(\mathbf{v}), \tag{3.7}$$

wobei μ die dynamische Viskosität, $\mathcal{I}$ den Einheitstensor, p den Ruhedruck und $\mathcal{D}(\mathbf{v})$ den elastischen Deformationstensor

$$\mathcal{D}(\mathbf{v}) = \nabla\mathbf{v} + (\nabla\mathbf{v})^T \tag{3.8}$$

bezeichnen. Die Anwendung des Gaußschen Integralsatzes unter den notwendigen Stetigkeitsvoraussetzungen führt zu der Impulserhaltungsgleichung

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\otimes\mathbf{v}) = \nabla\cdot\tau + \rho\mathbf{g}. \tag{3.9}$$

Wenn wir im folgenden die Annahme eines nicht viskosen Fluids ($\mu = 0$) machen, so daß sich der Streßtensor auf $\tau = -p\mathcal{I}$ reduziert, und die externen Körperkräfte $\mathbf{g}$ vernachlässigen, erhalten wir die vereinfachte Impulsbilanz

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\otimes\mathbf{v}) + \nabla p = \mathbf{0}. \tag{3.10}$$

3.1.3 Erhaltungsprinzip der Energie

Der erste Hauptsatz der Thermodynamik besagt, daß die *Änderung der totalen Energie E eines Fluids gleich der Summe der verrichteten Arbeit und der zugeführten Wärme ist.* Dabei setzt sich die spezifische totale Energie aus der spezifischen internen und der spezifischen kinetischen Energie zusammen

$$E = e + \frac{\mathbf{v}^2}{2}. \tag{3.11}$$

Im folgenden werden wir die übrigen Terme der allgemeinen Bilanzgleichung – den Oberflächen- und den Volumenanteil – festlegen. Ersterer ist als Summe der von Scherkräften verrichteten Arbeit $\tau \cdot \mathbf{v}$ und des diffusiven Flusses $-\kappa \nabla T$ definiert. Dieser ergibt sich nach Fouriers Gesetz zur Modellierung der molekularen Wärmeausbreitung in einem ruhenden Medium, wobei T die absolute Temperatur in Kelvin und κ den Wärmeleitkoeffizienten bezeichnen. Faßt man beide Terme zusammen, so erhält man für den Oberflächenanteil

$$\sigma = \tau \cdot \mathbf{v} + \kappa \nabla T. \tag{3.12}$$

Der Volumenanteil f setzt sich aus der von der Schwerkraft $\mathbf{g}$ geleisteten Arbeit und der durch Quellen q erzeugten spezifischen Wärme zusammen

$$f = \mathbf{g} \cdot \mathbf{v} + q. \tag{3.13}$$

Beide Anteile eingesetzt in (3.3) liefern die integrale Energieerhaltungsgleichung

$$\frac{\partial}{\partial t} \int_\Omega \rho E \, \mathrm{d}\Omega + \int_\Gamma \rho E \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}s = \int_\Gamma (\tau \cdot \mathbf{v} + \kappa \nabla T) \cdot \mathbf{n} \, \mathrm{d}s + \int_\Omega \rho(\mathbf{g} \cdot \mathbf{v} + q) \, \mathrm{d}\Omega. \tag{3.14}$$

Die Anwendung des Gaußschen Integralsatzes führt auf

$$\frac{\partial(\rho E)}{\partial t} + \nabla \cdot (\rho E \mathbf{v} - \tau \cdot \mathbf{v} - \kappa \nabla T) = \rho \mathbf{g} \cdot \mathbf{v} + \rho q. \tag{3.15}$$

Wenn wir wie zuvor den Einfluß der Schwerkraft und die Wärmequellen vernachlässigen und von einem nicht viskosen Fluid ($\mu = 0$) ausgehen, so liefert dies die vereinfachte Energieerhaltungsgleichung

$$\frac{\partial(\rho E)}{\partial t} + \nabla \cdot (\rho E \mathbf{v}) + \nabla \cdot (p \mathbf{v}) = 0. \tag{3.16}$$

Mit Hilfe der folgenden Definition für die spezifische totale Enthalpie

$$H = E + \frac{p}{\rho} \tag{3.17}$$

ergibt sich aus (3.16) die gebräuchliche Form

$$\frac{\partial(\rho E)}{\partial t} + \nabla \cdot (\rho H \mathbf{v}) = 0. \tag{3.18}$$

3.1.4 Die Eulergleichungen

Die vereinfachten Erhaltungsgleichungen (3.5), (3.10) und (3.18) bilden das System der kompressiblen Eulergleichungen

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{3.19}$$

$$\frac{\partial (\rho \mathbf{v})}{\partial t} + \nabla \cdot (\rho \mathbf{v} \otimes \mathbf{v}) + \nabla p = 0, \tag{3.20}$$

$$\frac{\partial (\rho E)}{\partial t} + \nabla \cdot (\rho H \mathbf{v}) = 0. \tag{3.21}$$

Wir stellen zunächst fest, daß alle Gleichungen eine gemeinsame Struktur besitzen und sich in der folgenden Form darstellen lassen

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f} = 0. \tag{3.22}$$

Im allgemeinen beschreibt der Fluß $\mathbf{f}$ den Transport der Masse, des Impulses, der Enthalpie oder einer anderen physikalischen Größe (Konzentration, Temperatur, Energie). Für skalare Erhaltungsgleichungen der Form (3.22) haben wir die Theorie von FEM-FCT im ersten Kapitel vorgestellt und im darauffolgenden seine Leistungsfähigkeit demonstriert. Die FCT Methodik soll im vierten Kapitel auf das stark gekoppelte System der Eulergleichungen übertragen werden. Um (3.19) – (3.21) in kompakter Form darstellen zu können, wollen wir mit U den Vektor der Erhaltungsgrößen Dichte, Impuls und totale Energie bezeichnen

$$U = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho v_3 \\ \rho E \end{bmatrix} \tag{3.23}$$

und daraus den konservativen Flußvektor $\mathbf{F} = (F^1, F^2, F^3)$ aufbauen

$$F^1 = \begin{bmatrix} \rho v_1 \\ \rho v_1^2 + p \\ \rho v_1 v_2 \\ \rho v_1 v_3 \\ v_1(\rho E + p) \end{bmatrix}, F^2 = \begin{bmatrix} \rho v_2 \\ \rho v_1 v_2 \\ \rho v_2^2 + p \\ \rho v_2 v_3 \\ v_2(\rho E + p) \end{bmatrix}, F^3 = \begin{bmatrix} \rho v_3 \\ \rho v_1 v_3 \\ \rho v_2 v_3 \\ \rho v_3^2 + p \\ v_3(\rho E + p) \end{bmatrix}. \tag{3.24}$$

Wir stellen weiterhin fest, daß sich dadurch das aus den Gleichungen (3.19) – (3.21) bestehende System völlig analog zu (3.22) als

$$\frac{\partial U}{\partial t} + \nabla \cdot \mathbf{F} = \mathbf{0}, \qquad \text{mit} \qquad \nabla \cdot \mathbf{F} = \sum_{d=1}^{3} \frac{\partial F^d}{\partial x_d} \tag{3.25}$$

formulieren läßt. Bisher setzt sich (3.25) aus fünf Erhaltungsgleichungen zusammen, welche die drei unbekannten Zustandsvariablen Dichte, Druck und interne Energie (ρ, p, e) sowie die drei Geschwindigkeitskomponenten und die Temperatur ($\mathbf{v}, T$) beinhalten. Um daraus ein abgeschlossenes System zu machen, muß es durch zwei weitere sogenannte Zustandsgleichungen vervollständigt werden.

3.1.5 Thermodynamische Aspekte

Die Thermodynamik beschäftigt sich mit den mikroskopischen Eigenschaften eines Mediums im Gegensatz zu den makroskopischen Eigenschaften wie Geschwindigkeit und kinetischer Energie, die eher in der Strömungsmechanik von Interesse sind. Jedes einfache thermodynamische System besitzt höchstens zwei unabhängige Zustandsvariablen, aus denen die übrigen eindeutig bestimmt werden können.

Wir werden für den Rest dieser Arbeit die Annahme eines *perfekten Gases* machen, für das die folgende Zustandsgleichung gilt

$$p = \rho \mathcal{R} T. \tag{3.26}$$

Hierbei bezeichnet $\mathcal{R}$ die charakteristische Gaskonstante. Die zweite Zustandsgleichung erhalten wir aus der Theorie der Thermodynamik für perfekte Gase, die besagt, daß die interne Energie e ausschließlich der Temperatur T abhängt

$$e = e(T). \tag{3.27}$$

Weiter wollen wir die spezifische Enthalpie einführen

$$h = e + \frac{p}{\rho}, \tag{3.28}$$

die für ein perfektes Gas als Funktion von der Temperatur darstellbar ist

$$h = e(T) + \mathcal{R}T = h(T). \tag{3.29}$$

Damit lassen sich die spezifischen Wärmekapazitäten bei konstantem Volumen und konstantem Druck wie folgt festlegen

$$c_v = \frac{\mathrm{d}e}{\mathrm{d}T}, \qquad c_p = \frac{\mathrm{d}h}{\mathrm{d}T}, \tag{3.30}$$

für deren Differenz die Beziehung $\mathcal{R} = c_p - c_v$ gilt. Für ein *kalorisch perfektes Gas* ist $c_v = const$ und damit auch $c_p = const$, so daß sich der Isentropenexponent

$$\gamma = \frac{c_p}{c_v} \tag{3.31}$$

definieren läßt, welcher für Luft gleich $\gamma = 7/5$ ist. Bisher haben wir keinen Mechanismus eingeführt, der die Irreversibilität eines kontinuumsmechanischen Prozesses sichert. Wenn wir als Beispiel den Wärmefluß von einem wärmeren in ein kälteres Medium betrachten, so ist seine Umkehrung nicht vollkommen unmöglich, jedoch auf mikroskopischer Ebene betrachtet sehr unwahrscheinlich und daher in der Praxis nicht zu beobachten. Dieses stochastische Argument wird in der makroskopischen Theorie durch die Entropie S ersetzt. Von ihr nehmen wir axiomatisch an, daß sie (i) eine Zustandsgröße und (ii) additiv ist, sie (iii) einer Bilanzgleichung genügt und daß nach dem zweiten Hauptsatz der Thermodynamik (iv) die Entropie des Universums bei allen ablaufenden Prozessen niemals abnimmt. Im folgenden bedeutet δ eine kleine Änderung in den Zustandsvariablen

$$T\,\delta S = \delta e + p\,\delta\frac{1}{\rho} = \delta h - \frac{1}{\rho}\delta p. \tag{3.32}$$

Wir bezeichnen einen thermodynamischen Prozess als *adiabatisch*, wenn dem System keine Wärme hinzugeführt oder entzogen wird. Unter der Annahme eines perfekten Gases läßt sich das System der Eulergleichungen (3.25) durch die Hinzunahme der beiden Zustandsgleichungen (3.26) und (3.27) vervollständigen. Elimination der Temperatur liefert für den Druck die Beziehung

$$p = p(\rho, e) = (\gamma - 1)\rho e, \tag{3.33}$$

die das System der Eulergleichungen zusammen mit der Definition der (totalen) Enthalpie (3.17) bzw. (3.28) abschließt.

Eine wichtige Größe in der Fluiddynamik ist die Machzahl M, welche die Strömungsgeschwindigkeit des Mediums mit der Schallgeschwindigkeit in Beziehung setzt. Letztere ist für ein perfektes Gas durch

$$c^2 = \gamma\mathcal{R}T = \frac{\gamma p}{\rho} \tag{3.34}$$

gegeben, so daß wir für die Machzahl den folgenden Ausdruck erhalten

$$M = \frac{|\mathbf{v}|}{c}. \tag{3.35}$$

Wie bereits in der Einleitung zu diesem Kapitel bemerkt, zeigen Strömungen mit $M < 0.3$ ein inkompressibles Verhalten und sind nicht Gegenstand dieser Arbeit. Oberhalb davon unterteilt man sie in subsonische ($M < 1$), transsonische ($M \sim 1$), supersonische ($M > 1$) und hypersonische ($M > 5$) Strömungen. Bei der Behandlung von letzteren treten aufgrund der hohen Temperaturen chemische Effekte in den Vordergrund, so daß wir sie nicht weiter betrachten wollen. Für uns werden die ersten drei Strömungstypen von Interesse sein, wobei die Machzahl unter anderem für die Aufstellung von Randbedingungen benötigt wird.

3.2 Die quasi-lineare Formulierung

In diesem Abschnitt werden wir eine alternative und für die Anwendung besser geeignete Darstellung der Eulergleichungen herleiten. Die Gleichung (3.25) enthält lediglich erste Ableitungen in der Zeit und im Ort und bildet somit ein System von PDEs erster Ordnung. Es bezeichne A^d die Jacobimatrix entlang der Koordinatenrichtung x_d, die als

$$A^d = \frac{\partial F^d}{\partial U}, \qquad d = 1, 2, 3 \tag{3.36}$$

definiert ist. In drei Dimensionen haben die Jacobimatrizen die Gestalt [29]

$$A^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{\gamma-1}{2}|\mathbf{v}|^2 - v_1^2 & (3-\gamma)v_1 & (1-\gamma)v_2 & (1-\gamma)v_3 & \gamma-1 \\ -v_1v_2 & v_2 & v_1 & 0 & 0 \\ -v_1v_3 & v_3 & 0 & v_1 & 0 \\ \frac{\gamma-1}{2}v_1|\mathbf{v}|^2 - v_1H & -v_1^2(\gamma-1)+H & (1-\gamma)v_1v_2 & (1-\gamma)v_1v_3 & \gamma v_1 \end{bmatrix}$$

$$A^2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ -v_1v_2 & v_2 & v_1 & 0 & 0 \\ \frac{\gamma-1}{2}|\mathbf{v}|^2 - v_2^2 & (1-\gamma)v_1 & (3-\gamma)v_2 & (1-\gamma)v_3 & \gamma-1 \\ -v_2v_3 & 0 & v_3 & v_2 & 0 \\ \frac{\gamma-1}{2}v_2|\mathbf{v}|^2 - v_2H & (1-\gamma)v_1v_2 & -v_2^2(\gamma-1)+H & (1-\gamma)v_2v_3 & \gamma v_2 \end{bmatrix}$$

$$A^3 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ -v_1v_3 & v_3 & 0 & v_1 & 0 \\ -v_2v_3 & 0 & v_3 & v_2 & 0 \\ \frac{\gamma-1}{2}|\mathbf{v}|^2 - v_3^2 & (1-\gamma)v_1 & (1-\gamma)v_2 & (3-\gamma)v_3 & \gamma-1 \\ \frac{\gamma-1}{2}v_3|\mathbf{v}|^2 - v_3H & (1-\gamma)v_1v_3 & (1-\gamma)v_2v_3 & -v_3^2(\gamma-1)+H & \gamma v_3 \end{bmatrix}$$

Die ein- und zweidimensionalen Analoga erhält man durch Streichen der entsprechenden Zeilen und Spalten sowie Elimination der nicht auftretenden Geschwindigkeitskomponenten v_d. Mit Hilfe der Jacobimatrizen $\mathbf{A} = (A^1, A^2, A^3)$ läßt sich das System der Eulergleichungen in quasi-linearer Form wie folgt schreiben

$$\frac{\partial U}{\partial t} + \mathbf{A} \cdot \nabla U = \mathbf{0}, \qquad \text{mit} \qquad \mathbf{A} \cdot \nabla U = \sum_{d=1}^{3} A^d \frac{\partial U}{\partial x_d}. \tag{3.37}$$

Unter der Annahme eines perfekten Gases ist der Flußvektor $\mathbf{F} = (F^1, F^2, F^3)$ eine homogene Funktion ersten Grades in den konservativen Variablen [77]

$$\mathbf{F}(\mu U) = \mu \mathbf{F}(U) \qquad \forall \mu \in \mathbb{R}. \tag{3.38}$$

Indem wir in (3.38) nach μ differenzieren und dann $\mu = 1$ setzen, erhalten wir für die Komponenten des Flußvektors die für die weitere Analyse nützliche Beziehung

$$F^d = \frac{\partial F^d}{\partial U} U = A^d U, \qquad d = 1, 2, 3. \tag{3.39}$$

3.3 Die nichtkonservative Formulierung

Im folgenden wollen wir zeigen, daß es sich bei den Eulergleichungen um ein hyperbolisches System handelt, und auf die spezielle Behandlung solcher Systeme eingehen. Der Einfachheit halber beschränken wir uns dabei auf den eindimensionalen Fall und möchten für die mehrdimensionale Theorie auf [29] verweisen.

Definition 3.3.1. *[75] Ein System der Form*

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x_1} + B = 0 \tag{3.40}$$

heißt hyperbolisch in einem Punkt (x,t), falls die Matrix $A \in \mathbb{R}^{N\times N}$ genau N reelle Eigenwerte $\lambda_1,\ldots,\lambda_N \in \mathbb{R}$ hat und die entsprechenden rechten Eigenvektoren $r_1\ldots,r_N$ linear unabhängig sind. Es wird als strikt hyperbolisch bezeichnet, falls alle λ_i paarweise verschieden sind ($\lambda_i \neq \lambda_j$ für $i \neq j$).

Diese Eigenschaft werden wir ausnutzen, um eine Formulierung der Eulergleichungen in den primitiven Variablen Dichte, Geschwindigkeit und Druck

$$W = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} \rho \\ v_1 \\ p \end{bmatrix} \tag{3.41}$$

herzuleiten, die sich wie folgt in die konservativen umrechnen lassen

$$U = \begin{bmatrix} w_1 \\ w_1 w_2 \\ w_1/(\gamma-1) + w_1 w_2^2/2 \end{bmatrix}. \tag{3.42}$$

Daraus ergeben sich die Jacobimatrix $Q = \partial U/\partial W$ und ihre Inverse als

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ v_1 & \rho & 0 \\ v_1^2/2 & \rho v_1 & \frac{1}{\gamma-1} \end{bmatrix}, \qquad Q^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -v_1/p & 1/\rho & 0 \\ \frac{\gamma-1}{2}v_1^2 & (1-\gamma)v_1 & \gamma-1 \end{bmatrix}, \tag{3.43}$$

so daß man durch Multiplikation von links und rechts mit Q^{-1} und Q die Eulergleichungen in nichtkonservativer Form erhält

$$\frac{\partial W}{\partial t} + Q^{-1}AQ\frac{\partial W}{\partial x_1} = 0. \tag{3.44}$$

Die resultierende Jacobimatrix ergibt sich als

$$Q^{-1}AQ = \begin{bmatrix} v_1 & \rho & 0 \\ 0 & v_1 & 1/\rho \\ 0 & \rho c^2 & v_1 \end{bmatrix} \tag{3.45}$$

mit den Eigenwerten

$$\lambda_1 = v_1 - c, \qquad \lambda_2 = v_1, \qquad \lambda_3 = v_1 + c \tag{3.46}$$

und den zugehörigen linear unabhängigen rechten Eigenvektoren

$$R_1 = \frac{1}{2}\begin{bmatrix} -\rho/c \\ 1 \\ -\rho c \end{bmatrix}, \qquad R_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad R_3 = \frac{1}{2}\begin{bmatrix} \rho/c \\ 1 \\ \rho c \end{bmatrix}. \tag{3.47}$$

Wenn wir (3.44) von rechts und links mit der Matrix aus rechten Eigenvektoren und ihrer Inversen multiplizieren, erhalten wir das folgende Gleichungssystem

$$\frac{\partial v_1}{\partial t} + (v_1 - c)\frac{\partial v_1}{\partial t} - \frac{1}{\rho c}\frac{\partial p}{\partial t} - \frac{v_1 - c}{\rho c}\frac{\partial p}{\partial x_1} = 0, \tag{3.48}$$

$$\frac{\partial v_1}{\partial t} + v_1\frac{\partial v_1}{\partial t} - \frac{1}{c^2}\frac{\partial p}{\partial t} - \frac{v_1}{c^2}\frac{\partial p}{\partial x_1} = 0, \tag{3.49}$$

$$\frac{\partial v_1}{\partial t} + (v_1 + c)\frac{\partial v_1}{\partial t} - \frac{1}{\rho c}\frac{\partial p}{\partial t} - \frac{v_1 + c}{\rho c}\frac{\partial p}{\partial x_1} = 0. \tag{3.50}$$

Nun benötigen wir die in (3.32) eingeführte Entropie, die für ein perfektes Gas bis auf eine Konstante eindeutig durch

$$S = c_v \ln \frac{p}{\rho^\gamma} \tag{3.51}$$

bestimmt ist, so daß wir die Gleichung (3.49) nach kurzer Rechnung durch

$$\frac{\partial S}{\partial t} + v_1\frac{\partial S}{\partial x_1} = 0 \tag{3.52}$$

ersetzen können. Solange die Unbekannten differenzierbar sind, bleibt die Entropie demnach entlang von Stromlinien konstant und erhöht sich, sobald ein Fluidpartikel einen Schock durchläuft, was dem vierten Axiom, daß die Entropie in einem ablaufenden Prozeß niemals abnehmen kann, entspricht. Durch Umformung von (3.51) erhält man

$$p = \exp\left(\frac{S}{c_v}\right)\rho^\gamma, \tag{3.53}$$

so daß sich (3.48) und (3.50) für eine lokal konstante Entropie zu

$$\frac{\partial}{\partial t}(v_1 - \frac{2c}{\gamma - 1}) + (v_1 - c)\frac{\partial}{\partial x_1}(v_1 - \frac{2c}{\gamma - 1}) = 0 \tag{3.54}$$

$$\frac{\partial}{\partial t}(v_1 + \frac{2c}{\gamma - 1}) + (v_1 + c)\frac{\partial}{\partial x_1}(v_1 + \frac{2c}{\gamma - 1}) = 0 \tag{3.55}$$

umschreiben lassen. Insgesamt folgt für den eindimensionalen Fall, daß die drei Riemann Invarianten, welche durch

$$R^+ = v_1 + \frac{2c}{\gamma - 1}, \qquad R^- = v_1 - \frac{2c}{\gamma - 1} \qquad \text{bzw.} \qquad S \tag{3.56}$$

definiert sind, entlang der Charakteristiken

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = v_1 \pm c \qquad \text{bzw.} \qquad \frac{\mathrm{d}x_1}{\mathrm{d}t} = v_1 \tag{3.57}$$

konstant sind. In Abschnitt 4.4 werden wir genauer darauf eingehen, inwiefern diese Charakteristiken die Anzahl der aufzustellenden Randbedingungen bestimmen und wie eine numerische Behandlung der Ein- und Ausflußränder basierend auf den Riemann Invarianten durchgeführt werden kann. In mehreren Dimensionen wird man statt (3.56) die folgenden Riemann Variablen betrachten

$$R^+ = \mathbf{v} \cdot \mathbf{n} + \frac{2c}{\gamma - 1}, \qquad R^- = \mathbf{v} \cdot \mathbf{n} - \frac{2c}{\gamma - 1}, \qquad \text{bzw.} \qquad S, \tag{3.58}$$

bei denen die skalare Größe v_1 durch die Normalenkomponente des Geschwindigkeitsvektors ersetzt wird.

Kapitel 4

Die Eulergleichungen

Im folgenden Kapitel wollen wir die für skalare Transportgleichungen entwickelte FEM-FCT Theorie am Beispiel der kompressiblen Eulergleichungen auf hyperbolische Erhaltungssätze verallgemeinern. Zuerst werden wir einen effizienten kantenbasierten Matrixaufbau vorstellen, der sich lokaler Roe Matrizen bedient. Zur Konstruktion eines auf Systeme verallgemeinerten discrete Upwindings wird ein künstlicher Diffusionstensor zum Operator hoher Ordnung hinzuaddiert, so daß die Nebendiagonalblöcke zu positiv definiten Matrizen werden. Als Alternative zu Roes *approximate Riemann Solver* werden wir ein auf skalarer, zum Spektralradius der Roe Matrix proportionaler Dissipation aufbauendes Vorgehen vorstellen, welches in Verbindung mit einem iterativen Defektkorrekturansatz zu einer Minimierung des Rechenaufwandes führt. Dabei werden wir aus dem Operator niedriger Ordnung einen blockdiagonalen Vorkonditionierer für die iterative Defektkorrekturschleife konstruieren, die dadurch in eine Sequenz von skalaren Teilproblemen zerfällt. Auf jedes einzelne läßt sich separat die in Kapitel 1 vorgestellte skalare FEM-FCT Theorie anwenden. Zuletzt muß eine Synchronisierung der Korrekturfaktoren für alle antidiffusiven Flüsse durchgeführt werden, für die wir eine auf beliebigen Indikatorvariablen basierende Technik vorschlagen.

4.1 Galerkin-Matrixaufbau

Beginnen wir mit den in Kapitel 3 hergeleiteten Eulergleichungen der Gasdynamik ohne Quellterme, die in der gebräuchlichen Divergenzform durch

$$\frac{\partial U}{\partial t} + \nabla \cdot \mathbf{F} = \mathbf{0} \tag{4.1}$$

gegeben sind. Für die Standard Galerkin Diskretisierung liefert die *group finite element formulation* [15] zur Interpolation der Flüsse die Darstellung

$$\sum_j \left[\int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x} \right] \frac{\mathrm{d}U_j}{\mathrm{d}t} + \sum_j \left[\int_\Omega \varphi_i \nabla \varphi_j \, \mathrm{d}\mathbf{x} \right] \cdot \mathbf{F}_j = \mathbf{0}. \tag{4.2}$$

Damit nach einer impliziten Zeitdiskretisierung ein (nicht-)lineares Problem der Form $AU = B$ entsteht, muß das semi-diskrete Schema hoher Ordnung folgendermaßen umgeschrieben werden

$$M_C \frac{\mathrm{d}U}{\mathrm{d}t} = KU, \tag{4.3}$$

wobei $M_C = \{\mathrm{M}_{ij}\}$ die blockdiagonale konsistente Massenmatrix des gekoppelten Systems und K das diskrete Analogon des Operators $-\mathbf{A} \cdot \nabla$ aus der quasilinearen Formulierung (3.37) bezeichnet. Da gebräuchliche Basisfunktionen die Eigenschaft $\sum_i \varphi_i = 1$ besitzen, verschwindet die Summe ihrer Ableitungen, so daß der in (1.29) definierte Koeffizient $\mathbf{c}_{ij}$ der Beziehung $\mathbf{c}_{ii} = -\sum_{j \neq i} \mathbf{c}_{ij}$ genügt. Für den Knoten i läßt sich damit die rechte Seite von (4.3) als

$$(KU)_i = -\sum_j \mathbf{c}_{ij} \cdot \mathbf{F}_j = -\sum_{j \neq i} \mathbf{c}_{ij} \cdot (\mathbf{F}_j - \mathbf{F}_i) \tag{4.4}$$

schreiben, wobei $\mathbf{F}_i$ und $\mathbf{F}_j$ die Knotenwerte der Flüsse bezeichnen. Wenn wir gezielt den zur Kante $\vec{ij}$ gehörenden Summanden aus (4.4) herausgreifen, so können wir den Kantenbeitrag zu den beiden beteiligten Knoten wie folgt darstellen

$$\begin{array}{rcl} \mathbf{c}_{ij} \cdot (\mathbf{F}_i - \mathbf{F}_j) & \longrightarrow & (KU)_i \\ \mathbf{c}_{ji} \cdot (\mathbf{F}_j - \mathbf{F}_i) & \longrightarrow & (KU)_j. \end{array} \tag{4.5}$$

Desweiteren läßt sich auch für Systeme eine konservative Flußzerlegung angeben

$$(KU)_i = -\sum_{j \neq i} \mathrm{G}_{ij}, \qquad \text{mit} \qquad \mathrm{G}_{ij} = \mathbf{c}_{ji} \cdot \mathbf{F}_j - \mathbf{c}_{ij} \cdot \mathbf{F}_i, \tag{4.6}$$

wobei der *Galerkin Fluß* antisymmetrisch ist: $\mathrm{G}_{ji} = -\mathrm{G}_{ij}$. In einer Dimension erhält man für ein uniformes Gitter und lineare Finite Elemente $c_{ji} = -c_{ij} = 1/2$ und damit die bekannte Form $\mathrm{G}_{ij} = (F_i + F_j)/2$.

Wie wir schon in Abschnitt 3.2 gesehen haben, bildet $\mathbf{F}$ eine in den konservativen Variablen U homogene Funktion ersten Grades, so daß sich mit Hilfe von Jacobimatrizen eine quasi-lineare Darstellung (3.37) der Eulergleichungen herleiten läßt. Dies wollen wir jetzt auf die diskrete Formulierung übertragen. Durch Anwendung von partieller Integration erhalten wir die Beziehung

$$\mathbf{c}_{ij} + \mathbf{c}_{ji} = -\int_\Gamma \mathbf{n}\,\varphi_i\varphi_j\,\mathrm{d}s, \tag{4.7}$$

wobei $\mathbf{n}$ den äußeren Normaleneinheitsvektor bezeichnet. Somit läßt sich der Koeffizient $\mathbf{c}_{ij}$ in einen internen und einen Randanteil zerlegen: $c_{ij} = -(\mathbf{a}_{ij} + \mathbf{b}_{ij})$

$$\begin{aligned}
\mathbf{a}_{ij} &= -\frac{\mathbf{c}_{ij} - \mathbf{c}_{ji}}{2} = -\mathbf{c}_{ij} + \frac{1}{2}\int_\Gamma \mathbf{n}\,\varphi_i\varphi_j\,\mathrm{d}s, \\
\mathbf{b}_{ij} &= -\frac{\mathbf{c}_{ij} + \mathbf{c}_{ji}}{2} = -\frac{1}{2}\int_\Gamma \mathbf{n}\,\varphi_i\varphi_j\,\mathrm{d}s.
\end{aligned} \tag{4.8}$$

Für diese Separation gilt, daß der Koeffizient für den internen Beitrag antisymmetrisch ist ($\mathbf{a}_{ji} = -\mathbf{a}_{ij}$), während die Randkomponente ein symmetrisches Verhalten aufweist ($\mathbf{b}_{ji} = \mathbf{b}_{ij}$). Da die zu inneren Knoten gehörenden Basisfunktionen am Rand verschwinden, ergibt sich für Kanten mit mindestens einem internen Knoten die Vereinfachung $\mathbf{a}_{ij} = -\mathbf{c}_{ij}$ und $\mathbf{b}_{ij} = 0$. Für lineare oder multilineare Finite Elemente erhalten wir damit explizit den Randterm $\mathbf{b}_{ij} = -\mathbf{n}\,t_{ij}/2$, der nur dann von Null verschieden ist, wenn beide Knoten i und j am Rand liegen. Dabei bezeichnet $t_{ij} = \int_\Gamma \varphi_i\varphi_j\,\mathrm{d}s$ die Einträge der Massenmatrix für die Randtriangulierung, die sich in zwei Dimensionen etwa als $|\Gamma_{ij}|/6$ ergeben. Im folgenden werden wir sehen, welchen Vorteil diese Separation in interne und Randbeiträge bietet. Unser Ziel besteht weiterhin darin, eine quasi-lineare Formulierung der Eulergleichungen auf diskreter Ebene herzuleiten.

Anstelle der in Abschnitt 3.2 verwendeten Jacobimatrizen $\mathbf{A} = (A^1, A^2, A^3)$ führen wir die sogenannten Roe-Matrizen $\hat{\mathbf{A}}_{ij} = (\hat{A}^1_{ij}, \hat{A}^2_{ij}, \hat{A}^3_{ij})$ ein, die sich durch Auswertung von $\mathbf{A}$ in den Roe Mittelwerten [64] bezüglich der Kante $\vec{ij}$ ergeben

$$\hat{\rho}_{ij} = \sqrt{\rho_i\rho_j}, \qquad \hat{\mathbf{v}}_{ij} = \frac{\sqrt{\rho_i}\mathbf{v}_i + \sqrt{\rho_j}\mathbf{v}_j}{\sqrt{\rho_i} + \sqrt{\rho_j}}, \qquad \hat{H}_{ij} = \frac{\sqrt{\rho_i}H_i + \sqrt{\rho_j}H_j}{\sqrt{\rho_i} + \sqrt{\rho_j}}. \tag{4.9}$$

Wir werden in Abschnitt 4.2 genauer auf diese Wahl von Mittelwerten eingehen, da sie für die Konstruktion von '*hyperbolic Upwinding*' eine wichtige Rolle spielen. Für den Moment wollen wir unter den Roe Mittelwerten einfach einen speziellen Kantenmittelwert verstehen, welcher der Beziehung

$$\mathbf{F}_j - \mathbf{F}_i = \hat{\mathbf{A}}_{ij}(U_j - U_i) \tag{4.10}$$

genügt. Während im eindimensionalen Fall nur eine (eindeutige) Jacobimatrix für jede Kante existiert, benötigen wir im Mehrdimensionalen eine geeignete Linearkombination aller Roe-Matrizen für die einzelnen Koordinatenrichtungen. Zu

diesem Zweck definieren wir die sogenannte *kumulative Roe Matrix* als

$$\mathrm{R}_{ij} = -\mathbf{c}_{ij} \cdot \hat{\mathbf{A}}_{ij} = -\sum_{d=1}^{3} c_{ij}^{d} \hat{A}_{ij}^{d}. \tag{4.11}$$

Diese kann als eine Art Projektion des mehrdimensionalen Jacobi-Tensors auf die numerische Kante zwischen den Knoten i und j interpretiert werden. Eine solche Kante existiert immer dann, wenn die Träger der Basisfunktionen der beiden Knoten einen nichtleeren Schnitt haben. Offensichtlich läßt sich die Kantenliste aus dem Besetzungsmuster der Finiten Elemente Matrix ablesen.

An dieser Stelle greifen wir auf die zuvor hergeleitete Zerlegung (4.8) des Koeffizienten $\mathbf{c}_{ij}$ in einen internen und einen Randbeitrag zurück, so daß sich für die kumulative Roe Matrix die folgende Darstellung ergibt

$$\mathrm{R}_{ij} = \mathrm{A}_{ij} + \mathrm{B}_{ij} \qquad \text{mit} \qquad \begin{array}{l} \mathrm{A}_{ij} = \mathbf{a}_{ij} \cdot \hat{\mathbf{A}}_{ij}, \\ \mathrm{B}_{ij} = \mathbf{b}_{ij} \cdot \hat{\mathbf{A}}_{ij}. \end{array} \tag{4.12}$$

Da gilt $-\mathbf{c}_{ij} \cdot (\mathbf{F}_j - \mathbf{F}_i) = -\mathbf{c}_{ij} \cdot \hat{\mathbf{A}}_{ij}(U_j - U_i) = \mathrm{R}_{ij}(U_j - U_i)$, läßt sich (4.4) für jeden Knoten i auf diskreter Ebene in die äquivalente quasi-lineare Form

$$(KU)_i = \sum_{j \neq i} (\mathrm{A}_{ij} + \mathrm{B}_{ij})(U_j - U_i) \tag{4.13}$$

bringen. Wenn wir analog zu (4.5) einen einzelnen Summanden betrachten, so liefert die Kante $\vec{ij}$ den folgenden Beitrag zu den beteiligten Knoten

$$\begin{array}{lcl} (\mathrm{A}_{ij} + \mathrm{B}_{ij})(U_j - U_i) & \longrightarrow & (KU)_i \\ (\mathrm{A}_{ij} - \mathrm{B}_{ij})(U_j - U_i) & \longrightarrow & (KU)_j. \end{array} \tag{4.14}$$

Da der Randterm B_{ij} für interne Knoten verschwindet, muß für jede Kante $\vec{ij}$ nur *eine* kumulative Roe Matrix, nämlich A_{ij}, ausgewertet und ihr Beitrag entsprechend (4.14) auf die Knoten i und j verteilt werden, was einen effizienten kantenbasierten Matrixaufbau ermöglicht. An dieser Stelle möchten wir bereits bemerken, daß der Randbeitrag $\mathrm{B}_{ij}(U_j - U_i)$ aufgrund seiner Antisymmetrie innerhalb des FCT Algorithmus als (anti-)diffusiver Fluß interpretiert werden kann. Wie zuvor ist für den Matrixaufbau keine numerische Quadratur notwendig, solange das Gitter fest und damit der Koeffizient $\mathbf{c}_{ij}$ konstant bleibt und während der Initialisierungsphase einmal berechnet und abgespeichert werden kann.

Die Konnektivität der globalen Systemmatrix hängt von dem zugrunde liegenden Gitter und der verwendeten Finiten Elemente Approximation ab, so daß sich die Kantenliste beim Übergang von skalaren Problemen zu Systemen nicht

ändert. Während im skalaren Fall nur Interaktionen zwischen Basisfunktionen unterschiedlicher Knoten stattfinden, interagieren bei Systemen auch Basisfunktionen für unterschiedliche Variablen miteinander. Als natürliche Verallgemeinerung des skalaren Falles wird jeder Koeffizient des diskreten Operators durch eine Matrix ersetzt, deren Größe sich aus dem Quadrat der Variablenanzahl ergibt. Für die Kante $\vec{ij}$ erhält man so die vier 5×5 Blöcke

$$\begin{aligned} \mathrm{K}_{ii} &= -\mathrm{A}_{ij} - \mathrm{B}_{ij}, & \mathrm{K}_{ij} &= \mathrm{A}_{ij} + \mathrm{B}_{ij}, \\ \mathrm{K}_{ji} &= -\mathrm{A}_{ij} + \mathrm{B}_{ij}, & \mathrm{K}_{jj} &= \mathrm{A}_{ij} - \mathrm{B}_{ij}. \end{aligned} \tag{4.15}$$

Die Einträge K_{ij}^{kl} müssen auf ihre Positionen in den 25 Blöcken $K_{kl} \in \mathbb{R}^{N \times N}$ mit $k, l = 1, \ldots, 5$ verteilt werden, wobei N die Anzahl der Knoten pro Variable bezeichnet. Die endgültige Assemblierung der globalen Systemmatrix $\mathbf{K} \in \mathbb{R}^{5N \times 5N}$ ist schematisch in Abbildung 4.1 dargestellt.

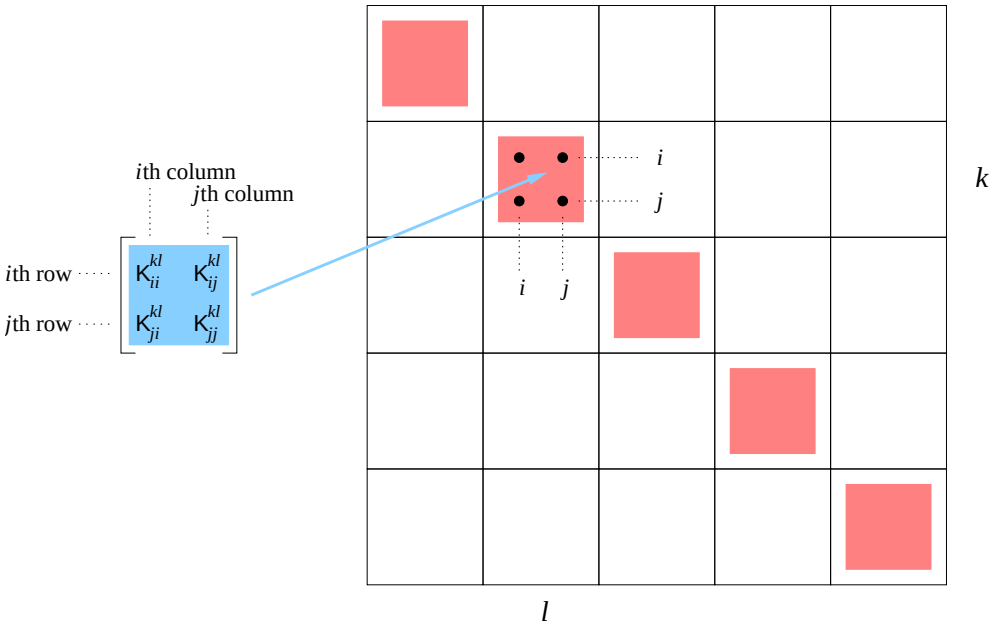


Abbildung 4.1: Globaler Matrixaufbau.

Nach der Zeitdiskretisierung läßt sich das System hoher Ordnung in abstrakter Matrizennotation wie folgt schreiben

$$A^H(U^H)U^H = B^H. \tag{4.16}$$

Dabei setzt sich die globale Systemmatrix hoher Ordnung A^H aus den 25 Matrizen jeweils von Dimension $N \times N$ zusammen

$$A_{kl}^H = M_C \delta_{kl} - \theta \Delta t K_{kl}, \tag{4.17}$$

und die Einträge des globalen Lastvektors B^H werden durch die fünf Vektoren

$$b_k^H = M_C u_k^n + (1-\theta)\Delta t \sum_l K_{kl} u_l^n \tag{4.18}$$

mit jeweils N Einträgen definiert. Da es sich bei den vollständig diskretisierten Eulergleichungen um ein nichtlineares Gleichungssystem handelt, werden wir eine Quasi-Fixpunkt-Defektkorrektur mit einem blockdiagonalen Vorkonditionierer anwenden. Wie wir im weiteren sehen werden, reicht es dafür aus, daß lediglich die Diagonalblöcke A_{kk}^H (und damit auch nur wie in Abbildung 4.1 verdeutlicht die Diagonalblöcke K_{kk}) explizit aufgebaut und gespeichert werden. Die Beiträge aller Nebendiagonalblöcke können innerhalb der kantenbasierten Aufbauroutine in den Defektvektor eingefügt werden. In zwei bzw. drei Dimensionen führt dies beim benötigten Speicherplatz für die globale Matrix A^H zu Einsparungen im Bereich von $75-80\%$. Ein weiterer Vorteil eines kantenorientierten Zuganges gegenüber einer elementbasierten Routine ergibt sich dadurch, daß die Projektionskoeffizienten $\mathbf{c}_{ij}$ typischerweise einmal zu Beginn der Simulation berechnet werden und sich dann bei festem Gitter nicht ändern, so daß die Matrizen ohne numerische Quadratur berechnet werden können.

Wir fassen den oben beschriebenen Matrixaufbau noch einmal zusammen. In *einer* Schleife über die Kantenliste der Konnektivitätsmatrix werden die beiden Komponenten A_{ij} und B_{ij} der kumulativen Jacobimatrix R_{ij} mit Hilfe der Projektionskoeffizienten $\mathbf{a}_{ij}$ und $\mathbf{b}_{ij}$ und der Roe Matrix $\hat{\mathbf{A}}_{ij}$ aufgebaut und ihre Einträge in die Diagonalblöcke K_{kk} der globalen Matrix oder in den Defekt bzw. den antidiffusiven Fluß eingefügt. Dieses Vorgehen entspricht in seiner Struktur der folgenden Aufteilung der durchzuführenden Operationen wie in [49]

- Sammeln (*gather*) von Knotenwerten für die Kante
- Berechnen des lokalen Kantenbeitrages (Datenlokalität)
- Hinzuaddieren (*scatter-add*) des Kantenbeitrages zu den Matrixeinträgen der beiden beteiligten Knoten und zum Defekt

Wir möchten zuletzt darauf hinweisen, daß die Numerierung der Knoten eines Gitters und ebenso die Reihenfolge der Kanten in der Kantenliste viele Freiheiten zuläßt. Eine 'geeignete' Traversierung der Kanten des Konnektivitätsgraphen kann zu einer Minimierung der durch die indirekte Adressierung und *Cache-Misses* verbrauchten Rechenzeit führen [49].

4.2 Künstliche Viskosität

Wie bereits im Kapitel 1 über die skalare Theorie festgestellt, hängt die Fähigkeit eines FEM-FCT Algorithmus, oszillationsfreie Lösungen zu produzieren, stark von der zugrunde liegenden Methode niedriger Ordnung ab. Zur Konstruktion eines Operators ähnlich dem in Abschnitt 1.6 werden die zuvor hergeleiteten lokalen Roe Matrizen durch die Hinzunahme von künstlicher Viskosität modifiziert. Bei der Analyse von skalaren positivitätserhaltenden Methoden haben wir festgestellt, daß der Anteil an künstlicher Dissipation gerade so gewählt werden muß, daß negative Nebendiagonaleinträge eliminiert werden. Für Gleichungssysteme bestehen die Nebendiagonaleinträge nicht länger aus Skalaren, sondern sind ihrerseits Matrizen. Als konsequente Verallgemeinerung des LED Kriteriums muß der Dissipationstensor für Systeme der Gestalt sein, daß die Nebendiagonalblöcke des resultierenden Operators *positiv definite* Matrizen sind.

Im ersten Schritt führen wir konservatives *mass lumping* durch, um die implizit in der konsistenten Massenmatrix beherbergte Antidiffusion zu eliminieren, und ersetzen die Diskretisierung hoher Ordnung (4.3) durch

$$M_L \frac{\mathrm{d}U}{\mathrm{d}t} = LU. \tag{4.19}$$

Hierbei bezeichnen M_L die diagonale gelumpte Massenmatrix und L den Jacobi-Operator niedriger Ordnung. Um all seine Nebendiagonalblöcke zu positiv definiten Matrizen umzuformen, berechnen wir für jede Kante $\vec{ij}$ den künstlichen Diffusionstensor D_{ij} und führen in Anlehnung an (4.15) einen kantenweisen Aufbau des diskreten Operators durch

$$\begin{aligned} \mathrm{L}_{ii} &= -\mathrm{A}_{ij} - \mathrm{D}_{ij}, & \mathrm{L}_{ij} &= \mathrm{A}_{ij} + \mathrm{D}_{ij}, \\ \mathrm{L}_{ji} &= -\mathrm{A}_{ij} + \mathrm{D}_{ij}, & \mathrm{L}_{jj} &= \mathrm{A}_{ij} - \mathrm{D}_{ij}. \end{aligned} \tag{4.20}$$

Dieses Vorgehen entspricht im skalaren Fall der Modifikation (1.60).

Wie schon im vorherigen Abschnitt für das System hoher Ordnung demonstriert, läßt sich das aus (4.19) nach der Zeitdiskretisierung entstehende System niedriger Ordnung in abstrakter Matrizennotation schreiben

$$A^L(U^L)U^L = B^L, \tag{4.21}$$

wobei die Blöcke der globalen Steifigkeitsmatrix A^L und des Lastvektors B^L in Analogie zu (4.17) und (4.18) durch folgende Vorschrift definiert sind

$$A^L_{kl} = M_L\delta_{kl} - \theta\Delta t L_{kl} \tag{4.22}$$

$$b^L_k = M_L u^n_k + (1-\theta)\Delta t \sum_l L_{kl} u^n_l. \tag{4.23}$$

Das Hauptziel dieses Abschnitts besteht nunmehr darin, einen 'geeigneten' Dissipationstensor D_{ij} zu konstruieren. Wir werden der Vollständigkeit halber kurz das klassische Godunov Schema vorstellen, welches als historischer Meilenstein in der Lösungstheorie der Eulergleichungen betrachtet werden kann. Im Anschluß daran werden wir den Übergang zu approximativen Riemann Lösern vollziehen, aus denen eine Klasse von zahlreichen (z.T. effizient implementierbaren) Verfahren entwickelt wurde. Insbesondere werden wir näher auf den Zugang von Roe [64] eingehen und anschließend den Vorteil von skalarer Dissipation insbesondere in Kombination mit dem FCT Paradigma herausstellen.

4.2.1 Godunov Schema

Das älteste Upwind Verfahren für die Eulergleichungen ist das konservative Godunov Schema [20], das auf der Idee basiert, lokale Riemann Probleme vorwärts in der Zeit exakt zu lösen und daraus die globale Lösung zusammenzusetzen. Im Kontext von Finiten Volumen werden die Variablen jeweils mit stückweise konstanten Funktionen approximiert. Als entsprechendes Analogon für Finite Elemente seien die Variablen in einer Umgebung der Knoten i und j konstant, so daß ein unstetiger Übergang im Mittelpunkt der Kante $\vec{ij}$ stattfindet. Riemann hat bereits 1860 für dieses nach ihm benannte Problem die exakte Lösung hergeleitet, die sich in Abhängigkeit vom Verhältnis $U_i : U_j$ als eine von zehn möglichen Kombinationen von Expansionsfächer, Kontaktunstetigkeit und Verdichtungsstoß [75] ergibt. Die Lösung des lokalen Riemann Problems sei mit

$$U_{ij}^* = RP(U_i, U_j) \tag{4.24}$$

bezeichnen. Zur Realisierung des von erster Ordnung genauen Godunov Verfahrens muß der Galerkin Fluß (4.6) durch den konsistenten numerischen Fluß [46]

$$\mathrm{G}_{ij}^* = \mathbf{F}(U_{ij}^*). \tag{4.25}$$

ersetzt werden. Eine genaue Beschreibung des Godunov Schemas und seiner Spezialisierung auf lineare hyperbolische Systeme findet sich ebenfalls in [46] und [75]. Der größte Nachteil der Godunov Methode liegt im enormen rechnerischen Aufwand, den das exakte Lösen der lokalen Riemann Probleme (4.24) erfordert. Gleichzeitig wird der Großteil der zusätzlichen Strukturinformation, die der exakte Riemann Löser liefert, im Verfahren nicht ausgenutzt, da die exakte Lösung zell- bzw. elementweise gemittelt wird. Aus diesem Grund ist es sinnvoll, einen approximativen Riemann Löser einzusetzen, der bei gleichem numerischen Fehler weniger rechenaufwendig ist. In der Literatur findet man eine Vielzahl von verschiedenen Ansätzen zur Konstruktion eines approximativen Riemann Lösers, darunter die von Roe [64], Roe und Pike [65] sowie von Osher und Solomon [61]. Allen gemeinsam ist das Prinzip, ein hyperbolisches Upwindverfahren durch Elimination von negativen Eigenwerten der Jacobimatrix zu konstruieren.

4.2.2 Roe's approximate Riemann Solver

Der in der Praxis am häufigsten verwendete approximative Riemann Löser geht auf Roe [64] zurück. Die Grundidee besteht darin, die Koeffizientenmatrix des nichtlinearen Systems von Erhaltungssätzen entlang der Kante zwischen den Knoten i und j zu linearisieren. Dabei werden von der sogenannten *Roe Matrix* $\hat{\mathbf{A}}_{ij}$ die folgenden Eigenschaften gefordert:

R.1 Hyperbolizität des linearisierten Systems.
Die Matrix $\hat{\mathbf{A}}_{ij}$ sei diagonalisierbar mit reellen Eigenwerten und einer Basis aus linear unabhängigen rechten Eigenvektoren.

R.2 Konsistenz mit der exakten Jacobimatrix.
$\hat{\mathbf{A}}_{ij} \rightarrow \nabla \cdot \mathbf{F}(\overline{U}_{ij})$ stetig, wenn $U_i \rightarrow \overline{U}_{ij}$ und $U_j \rightarrow \overline{U}_{ij}$, wobei $\overline{U}_{ij}$ die stückweise konstante Approximation der Lösung bezeichnet.

R.3 Konservativität an Unstetigkeiten.
$\hat{\mathbf{A}}_{ij}(U_j - U_i) = \mathbf{F}_j - \mathbf{F}_i$.

Offensichtlich garantiert die erste Bedingung R.1, daß für das linearisierte Problem der mathematische Charakter des nichtlinearen Systems erhalten bleibt. Die zweite Bedingung R.2 sichert ein 'sinnvolles' Verhalten der Methode für glatte Lösungen. Schließlich folgt aus R.3, daß die Lösung des approximativen Riemann Lösers mit der exakten übereinstimmt, falls U_i und U_j durch einen einfachen isolierten Verdichtungsstoß oder eine Kontaktunstetigkeit miteinander verbunden sind. Für allgemeine hyperbolische Systeme kann die Bestimmung einer Matrix mit den Eigenschaften R.1–R.3 sehr kompliziert und damit für den algorithmischen Gebrauch unattraktiv sein. Für die Eulergleichungen der Gasdynamik hingegen hat Roe in [64] eine relativ einfache Konstruktionsvorschrift basierend auf den Roe Mittelwerten (4.9) angegeben. Eine rigorose Herleitung der Roe Matrix sowie eine weiterführende Analyse der Matrizeneigenschaften R.1–R.3 findet man neben der Originalarbeit auch in [46] und [75].

Für hyperbolisches Upwinding ist es notwendig, daß sich die Matrix $\hat{\mathbf{A}}_{ij}$ diagonalisieren läßt, um ihre negativen Eigenwerte zu eliminieren. Dazu gehen wir zunächst auf die Definition 3.3.1 eines 'hyperbolischen Systems' zurück, welches stets eine diagonalisierbare Jacobimatrix

$$A = R\Lambda R^{-1} \tag{4.26}$$

besitzt. Dabei bezeichnet Λ die Matrix aus Eigenwerten von A und R die Matrix aus den dazu gehörenden rechten Eigenvektoren. Aufgrund der Hyperbolizität

existiert eine vom Koeffizientenvektor $\mathbf{a}_{ij}$ abhängige reguläre Matrix aus rechten Eigenvektoren, so daß die kumulative Roe Matrix A_{ij} diagonalisiert werden kann

$$\mathrm{A}_{ij} = R(\mathbf{a}_{ij})\Lambda(\mathbf{a}_{ij})R(\mathbf{a}_{ij})^{-1}, \tag{4.27}$$

wobei sich die Diagonalmatrix

$$\Lambda(\mathbf{a}_{ij}) = |\mathbf{a}_{ij}|\mathrm{diag}\{\lambda_1, \ldots, \lambda_5\} \tag{4.28}$$

aus den Eigenwerten von A_{ij} zusammensetzt

$$\lambda_1 = \hat{v}_{ij} - \hat{c}_{ij}, \quad \lambda_2 = \lambda_3 = \lambda_4 = \hat{v}_{ij}, \quad \lambda_5 = \hat{v}_{ij} + \hat{c}_{ij}. \tag{4.29}$$

Dabei bezeichnet der Skalierungsfaktor $|\mathbf{a}_{ij}| = \sqrt{\mathbf{a}_{ij} \cdot \mathbf{a}_{ij}}$ die euklidische Norm des Koeffizientenvektors $\mathbf{a}_{ij}$. Da die in (4.27) und (4.28) auftretenden Matrizen als kantengemittelte Größen zu verstehen sind, ist es sinnvoll, auch eine auf die Kante $\vec{ij}$ bezogene Geschwindigkeit einzuführen. Wir betrachten daher die folgende 'Projektion' der dichtegemittelten Geschwindigkeit

$$\hat{v}_{ij} = \frac{\mathbf{a}_{ij} \cdot \hat{\mathbf{v}}_{ij}}{|\mathbf{a}_{ij}|} \tag{4.30}$$

und die kantenbezogene Schallgeschwindigkeit

$$\hat{c}_{ij} = \sqrt{(\gamma - 1)\left(\hat{H}_{ij} - \frac{|\hat{\mathbf{v}}_{ij}|^2}{2}\right)}. \tag{4.31}$$

Diese beiden Größen verdeutlichen, daß die Eigenwerte (4.29) auch in mehreren Dimensionen charakteristische Geschwindigkeiten darstellen, wie sie aus der eindimensionalen Theorie für hyperbolische Erhaltungssätze bekannt sind [46].

Mit Hilfe der Darstellung (4.27) und (4.28) läßt sich der Diffusionstensor in (4.20) durch Elimination von negativen Eigenwerten aus A_{ij} entsprechend

$$\mathrm{D}_{ij} = |\mathrm{A}_{ij}| = R(\mathbf{a}_{ij})|\Lambda(\mathbf{a}_{ij})|R(\mathbf{a}_{ij})^{-1} \tag{4.32}$$

definieren, wobei sich die Matrix $|\Lambda(\mathbf{a}_{ij})|$ aus den Beträgen der Eigenwerte ergibt

$$|\Lambda(\mathbf{a}_{ij})| = |\mathbf{a}_{ij}|\mathrm{diag}\{|\lambda_1|, \ldots, |\lambda_5|\}. \tag{4.33}$$

Häufig wird diese Art von künstlicher Viskosität zur Konstruktion von *upwind-biased* Finiten Differenzen Schemata für hyperbolische Systeme eingesetzt [46], [53]. Der vorgestellte Ansatz erlaubt es ferner, Roes approximativen Riemann Löser auf Finite Elemente mit Hilfe von *flux difference splitting* zu übertragen.

Ein alternatives Vorgehen zur Durchführung von '*hyperbolic Upwinding*', besteht darin, den durch die konservative Flußzerlegung erzeugten Galerkin Fluß (4.6) durch einen konsistenten numerischen Fluß zu ersetzen

$$\mathrm{G}^*_{ij} = \mathrm{G}_{ij} + \frac{1}{2}\mathrm{D}_{ij}(U_j - U_i). \tag{4.34}$$

In der ursprünglichen Formulierung von Roe ist zu berücksichtigen, daß für *jeden* lokalen Dissipationstensor zwei Matrix-Matrix-Multiplikationen durchgeführt werden müssen. Roe und Pike [65] haben einen modifizierten Algorithmus vorgestellt, der, basierend auf analytischen Ausdrücken für die Mittelwerte, ohne das explizite Aufstellen der Matrix $|\mathrm{A}_{ij}|$ und damit das Berechnen und Invertieren der Matrix aus rechten Eigenvektoren auskommt. In [75] (vgl. 11.3, The Roe-Pike Method) findet man eine algebraische Herleitung der Bestimmungsgleichungen und eine einfach zu implementierende Darstellung des Verfahrens.

Ferner besitzt D_{ij} genau wie die Matrix A_{ij} die Dimension 5×5, so daß sich die CPU-Zeit für den kantenweisen Matrixaufbau (4.20) ungefähr verdoppelt. Auch wenn sich der approximative Riemann Löser von Roe gut als eigenständige Methode niedriger Ordnung geeignet ist, macht ihn der erforderliche Rechenaufwand in Kombination mit FEM-FCT wenig konkurrenzfähig.

4.2.3 Scalar limited dissipation

Da die Genauigkeit der Methode niedriger Ordnung bei FCT zugunsten der Effizienz durchaus etwas geringer ausfallen darf, bevorzugen wir die Anwendung von skalarer Viskosität, deren Wert dem Spektralradius der Roe Matrix A_{ij} entspricht

$$d_{ij} = |\mathbf{a}_{ij}| \max_i |\lambda_i| = |\mathbf{a}_{ij}|(|\hat{v}_{ij}| + \hat{c}_{ij}) \tag{4.35}$$

Der künstliche Dissipationstensor (4.32) reduziert sich dadurch auf eine Diagonalmatrix mit dem betragsmäßig größten Eigenwert auf der Hauptdiagonalen

$$\mathrm{D}_{ij} = d_{ij} I, \tag{4.36}$$

wobei I die Einheitsmatrix bezeichnet. Im Gegensatz zu Roes Dissipationstensor (4.32) geht das aus der Anwendung von skalarer Viskosität entstehende D_{ij} lediglich in die Berechnung der Diagonalblöcke (4.20) ein, und ist sogar für jede Komponente gleich. Damit reduziert sich der Rechenaufwand im Vergleich zu Roes Ansatz auf etwa 1/5 (in 3D), wobei das Verfahren formal der Ersetzung des Galerkin Flusses durch den konsistenten numerischen Fluß (4.34) entspricht.

Die leicht stärkere Diffusivität von *scalar limited dissipation* im Vergleich zu Roes approximativem Riemann Löser kann vernachlässigt werden, solange sie in der

nachfolgenden Flußkorrektur entfernt wird, so daß sich der erhöhte Rechenaufwand für das Verfahren von Roe nicht auszahlt. Darüberhinaus hat sich herausgestellt, daß eine leicht 'überdiffusive' Methode niedriger Ordnung in Kombination mit dem FEM-FCT Ansatz zu besseren Resultaten führen kann, was in der verbesserten Phasengenauigkeit begründet liegt [38].

4.3 FEM-FCT Algorithmus

Nach der Zeitdiskretisierung mit Hilfe des einschrittigen θ–Schemas lassen sich die beiden Methoden hoher und niedriger Ordnung in ähnlicher Weise wie in (1.85) in einer nichtlinearen Defektkorrektur kombinieren

$$U^{(m+1)} = U^{(m)} + [C^{(m)}]^{-1} R^{(m)}, \qquad m = 0, 1, \ldots. \tag{4.37}$$

Der globale Residuenvektor der Galerkin Diskretisierung ist analog zu (1.102) als

$$R^{(m)} = B^{(m+1)} - A(U^{(m)})U^{(m)} \tag{4.38}$$

definiert, wobei sich die globale rechte Seite $B^{(m+1)}$ als direkte Verallgemeinerung der skalaren Situation (1.101) ergibt. Dabei gilt für die Initialisierung

$$B^{(0)} = B^n = [M_L + (1-\theta)\Delta t L(U^n)]U^n, \qquad L = K + D \tag{4.39}$$

und für den aktuellen antidiffusiven 'Rohfluß'

$$\begin{aligned} F(U^n, U^{(m)}) &= [(M_C - M_L) - (1-\theta)\Delta t D(U^n)]U^n \\ &- [(M_C - M_L) + \theta \Delta t D(U^{(m)})]U^{(m)}. \end{aligned} \tag{4.40}$$

Die übrigen Setzungen können ohne Probleme aus der skalaren Theorie (vgl. Abschnitt 1.7.3) übernommen werden. Wie in den Abschnitten 4.1 und 4.2 erläutert, lassen sich die Vektoren B^n, $B^{(m+1)}$ und $R^{(m)}$ sowie die antidiffusiven Flüsse $F(U^n, U^{(m)})$ und der Vorkonditionierer $C^{(m)}$ in einer praktischen Implementierung in *einer* globalen kantenbasierten Aufbau- bzw. Updateroutine assemblieren. Das zu lösende lineare Gleichungssystem für das Lösungsinkrement $\Delta U^{(m+1)}$ läßt sich symbolisch in folgender Form schreiben

$$\begin{bmatrix} C_{11}^{(m)} & C_{12}^{(m)} & C_{13}^{(m)} & C_{14}^{(m)} & C_{15}^{(m)} \\ C_{21}^{(m)} & C_{22}^{(m)} & C_{23}^{(m)} & C_{24}^{(m)} & C_{25}^{(m)} \\ C_{31}^{(m)} & C_{32}^{(m)} & C_{33}^{(m)} & C_{34}^{(m)} & C_{35}^{(m)} \\ C_{41}^{(m)} & C_{42}^{(m)} & C_{43}^{(m)} & C_{44}^{(m)} & C_{45}^{(m)} \\ C_{51}^{(m)} & C_{52}^{(m)} & C_{53}^{(m)} & C_{54}^{(m)} & C_{55}^{(m)} \end{bmatrix} \begin{bmatrix} \Delta u_1^{(m+1)} \\ \Delta u_2^{(m+1)} \\ \Delta u_3^{(m+1)} \\ \Delta u_4^{(m+1)} \\ \Delta u_5^{(m+1)} \end{bmatrix} = \begin{bmatrix} r_1^{(m)} \\ r_2^{(m)} \\ r_3^{(m)} \\ r_4^{(m)} \\ r_5^{(m)} \end{bmatrix}. \tag{4.41}$$

Gerade für ein solch 'monströses' Gleichungssystem ist die 'richtige' Wahl des Vorkonditionierers von großer Bedeutung. Während er im skalaren Fall ausschließlich für die Konditionierung des linearen Systems verantwortlich ist, läßt sich (4.41) für zeitabhängige Probleme effizient mit einer entkoppelten Lösungsstrategie behandeln. Anstatt $C^{(m)}$ als einen *globalen* Vorkonditionierer eines schlecht konditionierten Gleichungssystems zu betrachten, berücksichtigen wir, daß die Komponenten des Lösungsvektors $\Delta U^{(m+1)}$ den konservativen Variablen entsprechen, die über *skalare* Erhaltungsgleichungen definiert sind.

Wir werden (4.41) mit Hilfe eines Block-Jacobi Vorkonditionierers in skalare Teilprobleme zerlegen, so daß $C_{kl}^{(m)} = 0, \quad \forall l \neq k$. Als konsequente Verallgemeinerung von (1.88) definieren wir dazu die fünf Diagonalblöcke von $C^{(m)}$ durch die entsprechenden Komponenten des Operators niedriger Ordnung

$$C_{kk}^{(m)} = M_L - \theta\Delta t L_{kk}^{(m)}, \qquad k = 1, \ldots, 5, \tag{4.42}$$

so daß das gekoppelte System der kompressiblen Eulergleichungen in eine Sequenz von skalaren Teilproblemen zerfällt

$$C_{kk}^{(m)} \Delta u_k^{(m)} = r_k^{(m)} \qquad k = 1, \ldots, 5, \tag{4.43}$$

$$u_k^{(m+1)} = u_k^{(m)} + \Delta u_k^{(m)}, \qquad u_k^{(0)} = u_k^n. \tag{4.44}$$

Die gut konditionierten linearen Gleichungssysteme für die fünf Inkrementvektoren können nacheinander oder besser noch parallel gelöst werden. Jede Komponente des Residuenvektors genügt der Vorschrift

$$r_k^{(m)} = b_k^{(m+1)} - M_L u_k^{(m)} + \theta\Delta t \sum_l L_{kl}^{(m)}\, u_l^{(m)}, \tag{4.45}$$

wobei der explizite Anteil der rechten Seite die Form

$$b_k^{(0)} = b_k^n = M_L u_k^n + (1-\theta)\Delta t \sum_l L_{kl}^n\, u_l^n \tag{4.46}$$

besitzt und der implizite entsprechend (1.96) bzw. (1.101) aktualisiert wird. Der unbegrenzte Antidiffusionsterm ergibt sich wie im skalaren Fall (1.91) gemäß

$$\begin{aligned} f(u_k^n, u_k^{(m)}) &= (M_C - M_L)u_k^n - (1-\theta)\Delta t \sum_l D_{kl}^n\, u_l^n \\ &- (M_C - M_L)u_k^{(m)} + \theta\Delta t \sum_l D_{kl}^{(m)}\, u_l^{(m)}. \end{aligned} \tag{4.47}$$

An dieser Stelle möchten wir abermals darauf hinweisen, daß der in den Abschnitten 4.1 und 4.2 vorgestellte kantenbasierte Matrixaufbau zu einer Steigerung der Effizienz führt, da keine CPU-belastenden Matrix-Vektor-Multiplikationen erforderlich sind, sondern stets lokale Kantenbeiträge in den Defektvektor bzw. in einen Diagonalblock des Vorkonditionierers eingefügt werden. Insbesondere wird der Performancezuwachs durch den Einsatz von *scalar limited dissipation* (vgl. Abschnitt 4.2.3) gegenüber dem klassischen Roe Löser aus Abschnitt 4.2.2 ersichtlich. Zum einen sind lediglich die Diagonalblöcke D_{kk} des künstlichen Diffusionstensors von Null verschieden, so daß sich der antidiffusive Rohfluß zu

$$\begin{aligned} f(u_k^n, u_k^{(m)}) &= [(M_C - M_L) - (1-\theta)\Delta t D_{kk}^n]u_k^n \\ &- [(M_C - M_L) + \theta\Delta t D_{kk}^{(m)}]u_k^{(m)} \end{aligned} \tag{4.48}$$

vereinfacht. Desweiteren ist der *skalare* Diffusionskoeffizient für alle fünf Komponenten identisch, nämlich proportional zum größten Eigenwert $|\mathbf{a}_{ij}|(|\hat{v}_{ij}| + \hat{c}_{ij})$ (4.35) der Roe Matrix, so daß ein effizienter kantenbasierter Aufbau der antidiffusiven Flüsse möglich ist.

4.3.1 Zalesak Limiter für Systeme

Während die Entwicklung auf dem Gebiet von skalaren Flux-Corrected Transport Schemata zu bemerkenswerten Fortschritten geführt und eine weitgehend vollständige Lösungstheorie hervorgebracht hat (vgl. [4], [79], [51], [52], [37] und [38]), stellen hyperbolische Gleichungssysteme weiterhin eine Herausforderung für FCT Methoden dar. Eine mögliche Vorgehensweise besteht darin, die Korrekturfaktoren α_{ij}^k für jede der fünf Komponenten mittels Operator-Splitting einzeln zu bestimmen. Die unabhängige Behandlung von physikalisch gekoppelten Variablen führt jedoch in einigen Fällen zur Ausbildung von Oszillationen [17], [18]. Dieser Nachteil hat die Entwicklung eines Limiting-Prozesses mit einem 'systemhaften Charakter' [49] vorangetrieben, bei dem alle antidiffusiven Flüsse mit einer Kombination der Korrekturfaktoren $\alpha_{ij} = \mathcal{S}(\alpha_{ij}^1, \ldots, \alpha_{ij}^5)$ *gemeinsam* begrenzt werden. Die beobachteten Verbesserungen in der numerischen Lösung lassen sich auf die Synchronisierung der Phasenfehler für die einzelnen Gleichungen zurückführen [51]. Dennoch haftet der Konstruktion solcher Limiter ein gewisses Maß an Empirismus an, so daß die verschiedenen Limiter-Varianten für unterschiedliche Anwendungen mehr oder weniger gut geeignet sind. Löhner schlägt in [49], [52] zwei Kriterien zur Konstruktion eines synchronisierten Limiters vor

S.1 Nutze die Korrekturfaktoren **einer spezifischen Indikatorvariable** (z.B. Dichte, Druck oder Entropie).

S.2 Nutze das Minimum der Korrekturfaktoren einer **Gruppe von Indikatorvariablen** (z.B. Dichte/Energie oder Dichte/Druck).

Es hat sich herausgestellt [49], daß die Kombination von Dichte/Energie insbesondere für dynamische Strömungen, die sich durch die Ausbreitung von Schocks und die Interaktion mehrerer Schockwellen auszeichnen, geeignet ist. Demgegenüber profitieren stationäre Probleme von der Kombination aus Dichte und Druck.

Wir wollen an dieser Stelle auf zwei weitere Aspekte der Limitersynchronisierung eingehen. Die für die Kombination aus Dichte und Druck notwendige Transformation von den konservativen Variablen in die primitiven Variablen läßt sich zu einem allgemeinen Konstruktionsprinzip für synchronisierte Limiter ausbauen [49]. Es bezeichne W einen beliebigen Variablensatz (z.B. primitive, charakteristische), für den die Korrekturfaktoren bestimmt werden sollen, und

$$T : \{U\} \rightarrow \{W\} \quad \text{mit} \quad W_i = T(U_i)U_i \tag{4.49}$$

die eindeutig bestimmte Transformationsabbildung von den konservativen in die nichtkonservativen Variablen. Insbesondere wird bei der Berechnung der oberen und unteren Schranken im Zalesak-Limiter die monotonieerhaltende Zwischenlösung $\tilde{U}$ verwendet, so daß wir in (4.49) zu $\tilde{W}_i = T(\tilde{U}_i)\tilde{U}_i$ übergehen und

für diese Lösung die Schranken berechnen können. Weiterhin betrachten wir für jede Kante die transformierten, nichtkonservativen Flüsse

$$\mathrm{H}_{ij} = T(\tilde{U}_i)\mathrm{F}_{ij}, \qquad \mathrm{H}_{ji} = T(\tilde{U}_j)\mathrm{F}_{ij}, \tag{4.50}$$

wobei durch die Transformation möglicherweise die Antisymmetrie verloren geht ($\mathrm{H}_{ij} \neq -\mathrm{H}_{ji}$). Dem allgemeinen Zalesak-Limiter folgend bestimmen wir für jede Komponente die Korrekturfaktoren β_{ij}^k und berechnen damit die korrigierten konservativen Flüsse mittels der folgenden Rücktransformationen

$$\begin{aligned} F_{ij}^*(U^{(m)}, U^n) &= T(\tilde{U}_i)^{-1} \left\{ \sum_{j\neq i} [\beta_{ij}^1, \ldots, \beta_{ij}^5] \mathrm{H}_{ij} \right\}, \\ F_{ji}^*(U^{(m)}, U^n) &= T(\tilde{U}_j)^{-1} \left\{ \sum_{j\neq i} [\beta_{ij}^1, \ldots, \beta_{ij}^5] \mathrm{H}_{ji} \right\}. \end{aligned} \tag{4.51}$$

Auch für diese ist die Antisymmetrieeigenschaft im allgemeinen nicht vorhanden. Die Verwendung eines synchronisierten Limiters führt jedoch dazu, daß nur noch ein gemeinsamer skalarer Korrekturfaktor $\beta_{ij} = \mathcal{S}(\beta_{ij}^1, \ldots, \beta_{ij}^5)$ für alle Flußkomponenten verwendet wird, so daß sich (4.51) stark vereinfacht

$$F_{ij}^*(U^{(m)}, U^n) = \beta_{ij}\mathrm{F}_{ij} = -F_{ji}^*(U^{(m)}, U^n) \tag{4.52}$$

und die Antisymmetrie der Flüsse erhalten bleibt. Das Vorgehen bei diesem Limiter läßt sich kurz wie folgt zusammenfassen:

T.1 Bestimme die Korrekturfaktoren für beliebige Indikatorvariablen W.

T.2 Synchronisiere diese und wende den resultierenden skalaren Korrekturfaktor β_{ij} auf die konservativen Flüsse F_{ij} an.

Desweiteren möchten wir auf die Verbesserungen hinweisen, die der iterative FCT Algorithmus für die Synchronisierung des Limiters bewirkt. Die FEM-FCT Basis Formulierung reagiert empfindlich auf die Zusammensetzung der Gruppe von Indikatorvariablen und produziert zum Teil deutlich schlechtere numerische Ergebnisse, falls die Geschwindigkeits-/Impulskomponenten mitberücksichtigt werden. Unter Verwendung von Prelimiting gestattet der iterative Limiter, das Minimum über die Faktoren *aller* Variablen als gemeinsame Korrektur zu nehmen, ohne die Qualität der numerischen Ergebnisse dadurch zu beeinträchtigen.

4.4 Implementierung von Randbedingungen

Die in Kapitel 3 vorgestellten Eulergleichungen der Gasdynamik beschreiben allgemeine reibungsfreie und adiabatische Strömungen. Erst durch das Aufstellen von speziellen Randbedingungen lassen sie sich zur Simulation eines konkreten strömungsdynamischen Problems nutzen. Aber gerade die Spezifizierung und noch mehr die numerische Behandlung von Randbedingungen stellt eine große Herausforderung dar und wird in vielen Publikationen gerne 'dem Leser überlassen'. Wir möchten an dieser Stelle nicht genauer auf die Theorie von Randbedingungen eingehen, zu der eine detaillierte Darstellung in [16], [29] zu finden ist, sondern praktische Aspekte ihrer Implementierung betrachten.

Sämtliche Randbedingungen können einer der beiden folgenden Klassen zugeordnet werden. Die *physikalischen* Randbedingungen (PBC) sind notwendig, um die Wohlgestelltheit des zu lösenden (kontinuierlichen) Problems zu garantieren, während die *numerischen* Randbedingungen (NBC) die Stabilität und das Konvergenzverhalten des (diskreten) Lösungsalgorithmus beeinflussen. Die Differenz zwischen der Anzahl der Lösungsvariablen $\mathsf{N}_{\mathsf{var}}$ und der vorgegebenen physikalischen Randbedingungen N_{p} legt fest, wie viele numerische Randbedingungen $\mathsf{N}_{\mathsf{n}} = \mathsf{N}_{\mathsf{var}} - \mathsf{N}_{\mathsf{p}}$ aufgestellt werden müssen. Dabei entspricht N_{p} am Ein- und Ausflußrand der Anzahl der in das Innere des Rechengebietes zeigenden Charakteristiken. Für N_{n} müssen dagegen die das Rechengebiet verlassenden Charakteristiken in Betracht gezogen werden. Tabelle 4.1 gibt eine Übersicht über die Anzahl der jeweiligen Randbedingungen in Abhängigkeit von der Raumdimension.

	Einflußrand						Ausflußrand					
	subsonisch			supersonisch			subsonisch			supersonisch		
	1D	2D	3D	1D	2D	3D	1D	2D	3D	1D	2D	3D
N_{p}	2	3	4	3	4	5	1	1	1	0	0	0
N_{n}	1	1	1	0	0	0	2	3	4	3	4	5

Tabelle 4.1: Physikalische und numerische Randbedingungen

Für viele Strömungskonfigurationen werden die Randbedingungen in Form von nichtkonservativen Variablen wie Entropie, Enthalpie oder Druck angegeben, was ihre direkte Implementierung als Dirichletwerte für den konservativen Lösungsvektor unmöglich macht. Als gängige Praxis hat sich die Transformation zu den charakteristischen Variablen etabliert. Dazu werden die hereinkommenden Riemann Invarianten (3.58) aus den physikalischen Randbedingungen berechnet und die herausgehenden aus den durch Extrapolation aus dem Inneren des Rechengebiets gewonnenen Werten bestimmt [36], [78]. Die Rücktransformation in die

konservativen Variablen liefert Knotenwerte, aus denen für Finite Volumen 'geeignete' Randflüsse konstruiert werden können und die sich für Finite Differenzen direkt als Dirichletwerte in den Randknoten setzen lassen. Uns ist jedoch keine Publikation bekannt, die solch ein charakteristisches Vorgehen im Kontext von (impliziten) Finite Elemente Diskretisierungen anwendet. Darüber hinaus fehlt dem auf unstrukturierten Gittern recht kostspieligen, heuristischen Ansatz, die Riemann Invarianten einfach aus dem Inneren zu extrapolieren, eine theoretische Rechtfertigung. Wir haben uns nach einigen Experimenten mit unterschiedlichen Implementierungen von Randbedingungen für ein einfaches Verfahren entschieden [40], das im folgenden vorgestellt werden soll.

Zunächst möchten wir bemerken, daß bei den von uns verwendeten Q_1-Elementen die *Knoten*werte am Rand bekannt sind, so daß keine Extrapolation von Informationen aus dem Inneren des Rechengebiets notwendig ist. In der Literatur [73] werden im wesentlichen drei Varianten zur Implementierung von Dirichlet Randbedingungen angegeben. Beim vollexpliziten Vorgehen werden sowohl die Zeilen als auch die Spalten der Systemmatrix eliminiert, die zu einem Randknoten gehören. Gleichzeitig werden die bekannten Werte multipliziert mit den entsprechenden Matrixeinträgen in den Vektor der rechten Seite eingesetzt. Zuletzt werden die zu spezifizierenden Dirichletwerte explizit in den Lösungsvektor eingebaut. Bei der zweiten, in der Praxis im häufigsten verwendeten semi-impliziten Methode werden nur die mit Randknoten korrespondierenden Zeilen der Systemmatrix durch die entsprechenden Zeilen der Identitätsmatrix ersetzt und die Dirichlet Randwerte in den Lösungsvektor bzw. in den Vektor der rechten Seite eingefügt. Zusammen mit einem iterativen Lösungsansatz garantiert diese Setzung, daß die konvergierte Lösung die vorgeschriebenen Randbedingungen erfüllt. Zuletzt sei noch das vollimplizite Vorgehen erwähnt, bei dem die Randwerte vor und nach jedem Iterationsschritt wie ein Filter in den Lösungsvektor und den Vektor der rechten Seite eingesetzt werden, die Systemmatrix jedoch unverändert bleibt.

Wir haben uns für eine semi-implizite Behandlung entschieden, da sie sich gut in das restliche iterative Vorgehen einfügt. Erinnern wir uns daran, daß wir in jedem Schritt eine Sequenz von skalaren Teilproblemen (4.43)–(4.44) für die einzelnen Komponenten $k = 1, \ldots, 5$ lösen müssen

$$u_k^{(m+1)} = u_k^{(m)} + [C_{kk}^{(m)}]^{-1} r_k^{(m)}. \tag{4.53}$$

Um die Randkomponenten des Lösungsvektors

$$U_i^{(m)} = [u_{1,i}^{(m)}, u_{2,i}^{(m)}, u_{3,i}^{(m)}, u_{4,i}^{(m)}, u_{5,i}^{(m)}]^T \tag{4.54}$$

explizit kontrollieren zu können, muß die Kopplung mit den Nachbarknoten aufgehoben werden. Dazu werden für jeden Randknoten i die zu den Komponenten $k = 1, \ldots, 5$ korrespondierenden Zeilen der fünf Diagonalblöcke des Vorkonditio-

nierers $C^{(m)} = \{c_{ij}^{kl}\}$ betrachtet und darin alle Nebendiagonaleinträge eliminiert

$$\forall i \in \mathsf{N}_b : \qquad c_{ij}^{kk} := 0 \qquad \forall j \neq i, \quad \forall k = 1, \ldots, 5. \tag{4.55}$$

Die zugehörigen Einträge des Residuums

$$R_i^{(m)} = [r_{1,i}^{(m)}, r_{2,i}^{(m)}, r_{3,i}^{(m)}, r_{4,i}^{(m)}, r_{5,i}^{(m)}]^T \tag{4.56}$$

werden durch die Diagonaleinträge des Vorkonditionierers dividiert und anschließend zum Lösungsvektor hinzuaddiert, um diesen explizit zu aktualisieren

$$\forall i \in \mathsf{N}_b : \qquad u_{k,i}^* = u_{k,i}^{(m)} + \frac{r_{k,i}^{(m)}}{c_{ii}^{kk}}, \qquad k = 1, \ldots, 5. \tag{4.57}$$

Die modifizierte Lösung U_i^* wird in die vorläufigen Riemann Invarianten W_i^* transformiert, welche die physikalischen Randbedingungen im allgemeinen nicht erfüllen. Da die Anzahl der PBC der Zahl der hereinkommenden Riemann Invarianten entspricht, können letztere analytisch berechnet und als Ersatz für die provisorischen Werte aus W_i^* genommen werden. Die mit den numerischen Randbedingungen korrespondierenden herausgehenden Riemann Invarianten bleiben von dieser Modifikation unberührt. Anschließend wird der 'korrigierte' Vektor W_i^{**} in die konservativen Variablen U_i^{**} zurücktransformiert und die Komponenten des Defektvektors auf Null gesetzt $R_i^{(m)} := 0$, bevor die skalaren Teilprobleme (4.43)–(4.44) gelöst werden. Die Sequenz von algebraischen Transformationen ist für einen Iterationsschritt in Abbildung 4.2 dargestellt.

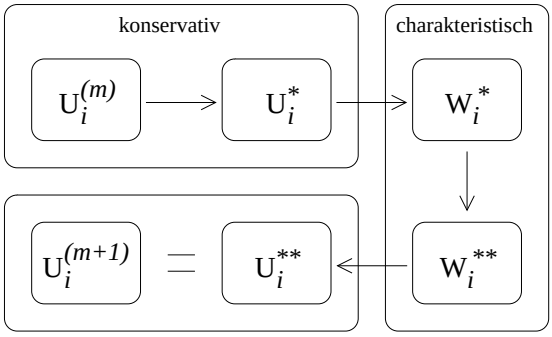


Abbildung 4.2: Randbehandlung: algebraische Transformationen.

Berücksichtigen wir, daß für einen Randknoten i lediglich die fünf Diagonaleinträge c_{ii}^{kk} des Vorkonditionierers $C^{(m)}$ von Null verschieden und die entsprechenden Einträge des Defektvektors $R_i^{(m)}$ gerade gleich Null sind, so gilt in jedem Iterationsschritt $\Delta U_i^{(m)} = 0$ und somit $U_i^{(m+1)} = U_i^{**}$. Nachdem die Defektkorrektur konvergiert hat, fungieren die 'korrigierten' Knotenwerte U_i^{**} *quasi* als Dirichletwerte für die Lösung zum Zeitschritt t^{n+1}, so daß der resultierende Vektor U^{n+1}

die PBC erfüllt. Offensichtlich ist keine *ad hoc* Extrapolation von Werten aus dem Inneren und auch keine Unterscheidung von unterschiedlichen Randtypen notwendig. Dennoch existiert für *solid walls* eine 'billigere' Implementierung.

Eine Wand zeichnet sich dadurch aus, daß sie undurchlässig ist. Diese *no-penetration* Eigenschaft wird im allgemeinen durch die Forderung nach einem verschwindenden Normalenfluß oder

$$\mathbf{v} \cdot \mathbf{n} = 0 \tag{4.58}$$

ausgedrückt. Die alternative Bezeichnung *free-slip* geht auf die äquivalente Festlegung zurück, daß die Tangentialgeschwindigkeit der Strömung entlang der Wand nicht dem Einfluß von Reibungseffekten ausgesetzt ist.

Bedingung (4.58) kann im vorgestellten Algorithmus ohne die Transformationen zu den Riemann Invarianten implementiert werden, indem die drei Impulskomponenten *nach* der Korrektur (4.57) mit Hilfe des äußeren Normalenvektors $\mathbf{n}_i = [n_{1,i}, n_{2,i}, n_{3,i}]^T$ im Knoten i auf die Tangentialebene projiziert werden

$$U_i^{**} = \begin{bmatrix} u_{1,i}^* \\ u_{2,i}^* \\ u_{3,i}^* \\ u_{4,i}^* \\ u_{5,i}^* \end{bmatrix} - \sum_{k=1}^{3} u_{k+1,i}^* n_{k,i} \begin{bmatrix} 0 \\ n_{1,i} \\ n_{2,i} \\ n_{3,i} \\ 0 \end{bmatrix}. \tag{4.59}$$

Das restliche Vorgehen bleibt unverändert, was zu einer übersichtlichen und unkomplizierten Implementierung beiträgt.

Alternativ können die 'unerwünschten' Normalengeschwindigkeiten aus den Impulskomponenten des Defektvektors eliminiert werden

$$R_i^* = \begin{bmatrix} r_{1,i}^{(m)} \\ r_{2,i}^{(m)} \\ r_{3,i}^{(m)} \\ r_{4,i}^{(m)} \\ r_{5,i}^{(m)} \end{bmatrix} - \sum_{k=1}^{3} r_{k+1,i}^{(m)} n_{k,i} \begin{bmatrix} 0 \\ n_{1,i} \\ n_{2,i} \\ n_{3,i} \\ 0 \end{bmatrix}, \tag{4.60}$$

wobei der Vektor $U_i^{(m)}$ unverändert bleibt.

Das vorgestellte Verfahren läßt sich problemlos auf die allgemeine Situation eines (im Blocksinne) 'vollbesetzen' Vorkonditionierers übertragen. Es soll gelten

$$\forall i \in \mathsf{N}_b : \qquad c_{ij}^{kl} = 0 \qquad \forall j \neq i, \quad \forall l \neq k, \tag{4.61}$$

so daß die Voraussetzungen für unseren Algorithmus weiterhin erfüllt sind.

4.5 Zusammenfassung des Algorithmus

Wie wir gesehen haben, läßt sich unsere FEM-FCT Formulierung zusammen mit Zalesaks Limiter auf Systeme hyperbolischer Erhaltungsgleichungen verallgemeinern. Wir haben an entsprechender Stelle gezeigt, welche Effizienzsteigerung durch den Einsatz von einem kantenbasierten Matrizenaufbau, einer *scalar limited dissipation*, einer Defektkorrektur mit blockdiagonalem Vorkonditionierer zur Entkopplung des Systems und zuletzt eines synchronisierten Limiters möglich ist. Viele dieser Techniken lassen sich unabhängig voneinander in einem bestehenden FEM-Code implementieren. Wir wollen die wesentlichen Schritte von FEM-FCT für hyperbolische Systeme von Erhaltungsgleichungen im folgenden zusammenfassen und dabei wie schon im skalaren Fall (vgl. Abschnitt 1.8) die Basis Formulierung als Spezialfall des iterativen FEM-FCT Algorithmus interpretieren.

In der kantenbasierten Aufbauroutine:

F.1 Berechne die Vektoren $\mathbf{c}_{ij}$ und $\mathbf{c}_{ji}$ für *alle* Komponenten des diskreten Operators hoher Ordnung wie im skalaren Fall oder greife auf die abgespeicherten Werte zurück und bilde damit die Koeffizientenvektoren des internen und des Randbeitrags $\mathbf{a}_{ij}$ und $\mathbf{b}_{ij}$ entsprechend (4.8).

F.2 Berechne die Roe Mittelwerte (4.9), werte damit die Roe Matrizen (4.12) aus und speichere die lokalen Matrizen A_{ij} und B_{ij}.

F.3 Bestimme den Diffusionskoeffizienten D_{ij} proportional zum Spektralradius der Roe Matrix entsprechend (4.36) und baue zusammen mit A_{ij} die lokalen Matrizen L_{ij} entsprechend (4.20) auf.

F.4 Konstruiere die Diagonalblöcke des Operators niedriger Ordnung (4.42) und des Vorkonditionierers $C_{kk}^{(m)}$. Addiere die Kantenbeiträge entsprechend (4.38) zum Residuum $R^{(m)}$ und für $m = 0$ und $\theta < 1$ ebenfalls zur rechten Seite B^n (4.39) hinzu.

F.5 Berechne die antidiffusiven Flüsse F_{ij} nach (4.48) und analog zum skalaren Vorgehen die zu begrenzende Flußdifferenz $\Delta\mathrm{F}_{ij}$.

Im Flußkorrektur-Modul:

F.6 Berechne komponentenweise (für nichtiteratives FCT nur für $m = 0$) die positivitätserhaltende Zwischenlösung $\tilde{U}^{(m)}$ wie im skalaren Fall.

F.7 Bestimme mit Hilfe des Zalesak-Limiters (mit Pre- und Postlimiting und ggf. Variablentransformation) die synchronisierten Korrekturfaktoren $\alpha_{ij}^{(m)}$ und begrenze die antidiffusiven Rohflüsse.

F.8 Addiere die begrenzte Antidiffusion zum Residuenvektor $R^{(m)}$.

F.9 Aktualisiere in der iterativen Formulierung die Hilfsgrößen $B^{(m)}$ und $\mathrm{G}_{ij}^{(m)}$ für jede Komponente wie im skalaren Fall.

In der FCT Defektkorrekturschleife:

F.10 Löse die Sequenz von linearen Systemen für $\Delta u_k^{(m+1)}$ (4.44) mit dem Defektvektor $R^{(m)}$.

F.11 Aktualisiere die neue Lösung $U^{(m+1)}$ und fahre mit der nichtlinearen Iterationsschleife fort oder gehe zum nächsten Zeitschritt.

Ein Vergleich des obigen Verfahrens mit dem skalaren Vorgehen (vgl. Abschnitt 1.8) zeigt, daß der Algorithmus weitgehend übernommen werden kann, wobei die skalaren Prozeduren für jede Komponente einzeln und möglicherweise auch parallel durchgeführt werden können. Lediglich zum Abschluß des Limiters müssen die berechneten Korrekturfakturen synchronisiert werden. Desweiteren besteht kein prinzipieller Unterschied zwischen der Simulation von ein- oder mehrdimensionalen Problemen. Im Idealfall genügt es daher, in einer vorausschauenden Implementierung lediglich den Schleifenindex für die skalaren Prozeduren zu erhöhen und die Routinen für die Gittergenerierung und den Aufbau der globalen Operatoren an die Raumdimension anzupassen.

Kapitel 5

Numerische Beispiele für die Eulergleichungen

Im folgenden Kapitel wollen wir die Leistungsfähigkeit der auf hyperbolische Gleichungssysteme verallgemeinerten FEM-FCT Methodik anhand von einigen Testfällen für die kompressiblen Eulergleichungen demonstrieren. Um eine Vergleichbarkeit mit anderen in der Literatur gebräuchlichen Verfahren zu ermöglichen, werden wir auf etablierte Benchmarkprobleme zurückgreifen.

In Abschnitt 5.1 beschäftigen wir uns mit zeitabhängigen Problemen und studieren dazu das Verhalten von FEM-FCT für das von Sod 1978 eingeführte Shock Tube Problem [70]. Weiter werden wir einen von Leveque [47] vorgeschlagenen Benchmark zur Analyse der Symmetrieeigenschaften eines numerischen Verfahrens in mehreren Dimensionen vorstellen. In Abschnitt 5.2 konzentrieren wir uns auf stationäre Probleme und betrachten dazu eine Reihe von *internal flow* Anwendungen. Wir werden dazu supersonische Strömungen durch Kanäle simulieren, die durch keilförmige Hindernisse abgelenkt werden, so daß *oblique shocks* und *expansion waves* entstehen.

5.1 Transiente Benchmarks

Schon anhand der Benchmarks für skalare Erhaltungsgleichungen haben wir festgestellt, daß die semi-implizite Crank-Nicolson Zeitdiskretisierung die beste Wahl für zeitabhängige Problemstellungen darstellt, da sie nicht so diffusiv wie vollimplizite Verfahren ist und ohne die für Forward Euler notwendige Stabilisierung auskommt. Bei den im folgenden diskutierten Simulationen haben wir uns eines 'Standardtricks' bedient und $\theta = 0.55$ gesetzt.

5.1.1 Shock Tube

Ein klassisches Beispiel zum Vergleich von Lösern für kompressible Strömungen ist das von Sod vorgeschlagene Shock Tube Problem [70]. Sein physikalisches Gegenstück ist eine zu Beginn mit einem ruhenden Gas gefüllte Röhre, die durch eine Membran in zwei Bereiche mit unterschiedlichen Gaskonfigurationen (im allgemeinen verschiedene Drücke und Dichten) unterteilt wird. Sobald die Membran entfernt wird, beginnt ein Massentransport entlang des Konzentrations- bzw. Druckgefälles in Richtung der geringeren Gaskonzentration bzw. Gasdruckes. Wenn wir in beiden Bereichen von einer anfänglich gleichmäßigen Gasverteilung ausgehen, so läßt sich das Strömungsverhalten nach dem Entfernen der Membran durch die eindimensionalen Eulergleichungen beschreiben.

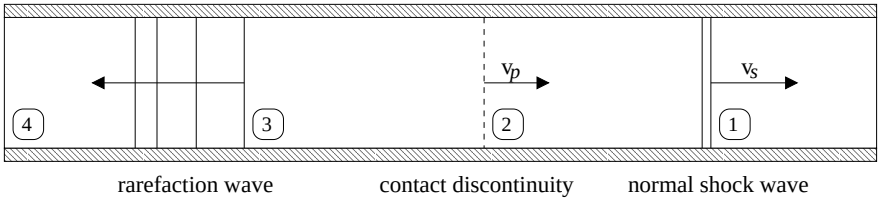


Abbildung 5.1: Shock Tube Problem. Struktur der Wellenausbreitung.

Abbildung 5.1 zeigt die Ausbildung von drei unterschiedlichen Wellen, welche das Gas in vier Abschnitte einteilen, in denen die Zustandsvariablen konstant sind. Nach dem abrupten Entfernen der Membran zum Zeitpunkt $t = 0$ bewegt sich eine Schockwelle (*normal shock*) in Richtung des geringeren Gasdrucks (1), wobei die Schockgeschwindigkeit v_s die Sprungbedingung

$$v_s \, [U] = [\mathbf{F}(U)] \tag{5.1}$$

erfüllt. Es ist bekannt, daß alle primitiven Variablen beim Passieren eines Schocks einen unstetigen Übergang erfahren. Durch den Druckunterschied setzt sich eine zweite Welle mit einer Geschwindigkeit v_p in Bewegung, so daß Masse in gleicher

Richtung vorangetrieben wird. Zu beiden Seiten (2) und (3) dieser Kontaktunstetigkeit bleiben die Geschwindigkeit und der Druck konstant, und nur die Dichte erfährt einen unstetigen Übergang. Schließlich bildet sich in entgegengesetzter Richtung ein Expansionsfächer aus, durch den alle Zustandsvariablen stetig zu ihren ursprünglichen Werten im linken Bereich (4) übergehen. Dieses Strömungsmuster bleibt so lange erhalten, bis eine der Wellen die linke oder rechte Röhrenwand erreicht und Überlagerungen aufgrund von Reflektionsgesetzen entstehen.

Für das eindimensionale Riemann Problem gelten die Anfangsbedingungen:

$$\begin{bmatrix} \rho_L \\ v_L \\ p_L \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 1.0 \end{bmatrix} \quad \text{für } x \in (0, 0.5], \qquad \begin{bmatrix} \rho_R \\ v_R \\ p_R \end{bmatrix} = \begin{bmatrix} 0.125 \\ 0.0 \\ 0.1 \end{bmatrix} \quad \text{für } x \in (0.5, 1). \tag{5.2}$$

Die hier gezeigten numerischen Ergebnisse wurden auf einem uniformen Gitter mit 100 linearen Finiten Elementen bei einem Zeitschritt von $\Delta t = 10^{-3}$ berechnet. Die exakte Lösung wurde mit Hilfe der in Referenz [1] vorgestellten Technik bestimmt und ist als gestrichelte Linie dargestellt. Abbildung 5.2 zeigt einen Vergleich zwischen zum Spektralradius der Roe Matrix proportionaler skalarer Dissipation (links) und dem approximativen Riemann Löser von Roe (rechts) zum Zeitpunkt $t = 0.231$. Beide Ansätze reproduzieren die Lösung mit vergleichbarer Güte und sind frei von Oszillationen. Die bei skalarer Viskosität erkennbare leicht stärkere 'Verschmierung' des Lösungsprofils wird durch den Einsatz des nichtlinearen Limiters ausgeglichen.

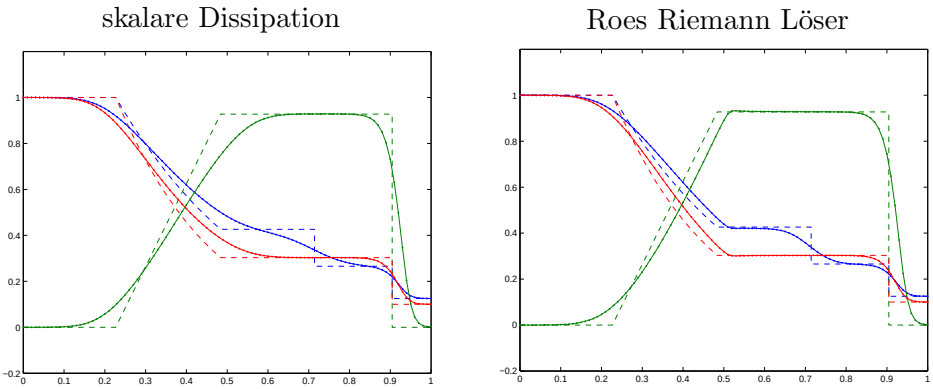


ABBILDUNG 5.2: Shock Tube Problem in 1D, 'hyperbolic Upwinding', $t = 0.231$.

Abbildung 5.3 zeigt einen Vergleich verschiedener Limitervarianten und -synchronisierungen bei Verwendung von skalarer Dissipation zur Konstruktion der Methode niedriger Ordnung. In den folgenden Simulationen wurde die Auflösung der Unstetigkeitsstellen durch die Anwendung von Prelimiting verbessert. Für die beiden oberen Simulationen wurde, den Vorschlägen von Löhner [49] folgend, das Minimum der Korrekturfaktoren für die Dichte und die Energie als gemeinsamer

Faktor gewählt. Sowohl der iterative als auch der Basis Limiter reproduzieren die exakte Lösung mit überzeugender Genauigkeit. Für die beiden unteren Abbildungen wurde zur Synchronisierung das Minimum über alle Variablen gewählt, was bei der Basis Formulierung zu diffusiveren Resultaten, insbesondere im rechten Teil des stetigen Expansionsfächers, führt. Da der iterative Limiter in jedem Iterationsschritt mehr Antidiffusion zuläßt, gleicht er die übermäßige Viskosität hinreichend aus. Wir möchten an dieser Stelle vom Einsatz einer konstanten Diffusion abraten, da die damit produzierten Ergebisse mit einer starken 'Verschmierung' behaftet sind. Gleichzeitig möchten wir bemerken, daß ein Splitting der Eulergleichungen in skalare Teilprobleme, die unabhängig voneinander begrenzt werden, die Ausbildung von kleinen Oszillationen zur Folge haben kann.

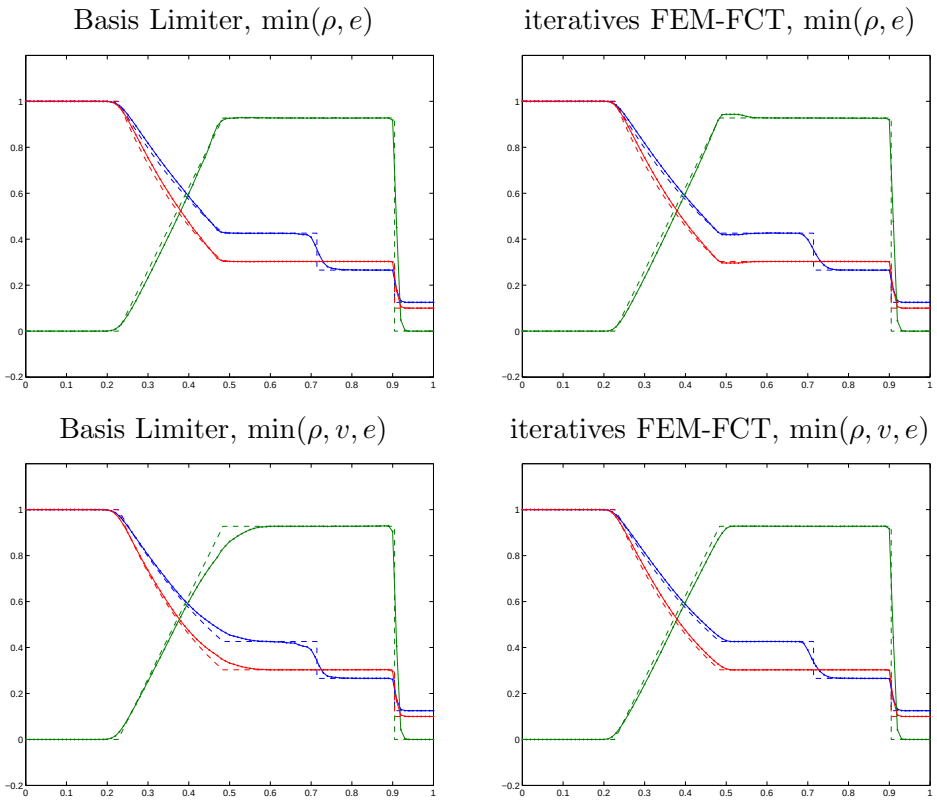


Abbildung 5.3: Shock Tube Problem in 1D, Synchronisierung, $t = 0.231$.

Im folgenden betrachten wir die zweidimensionale Variante des Shock Tube Problems, wobei die Wellenausbreitung entlang der x-Achse stattfindet. Die Anfangsbedingungen für das Riemann Problem werden analog zu (5.2) gewählt, wobei die Geschwindigkeit in 2D zu einem Vektor wird. Die numerischen Ergebnisse wurden auf einem uniformen Gitter mit 128×128 Q_1-Elementen bei einer Zeitschrittweite von $\Delta t = 10^{-3}$ bis zum Zeitpunkt $t = 0.231$ berechnet. Abbildung 5.4 zeigt die Simulationsergebnisse bei Anwendung eines über die Dichte und die Energie synchronisierten Limiters. Unten rechts ist ein Schnitt durch das Rechen-

gebiet bei $y = 0.5$ dargestellt, anhand dessen man eine gute Übereinstimmung mit den eindimensionalen Ergebnissen erkennen kann.

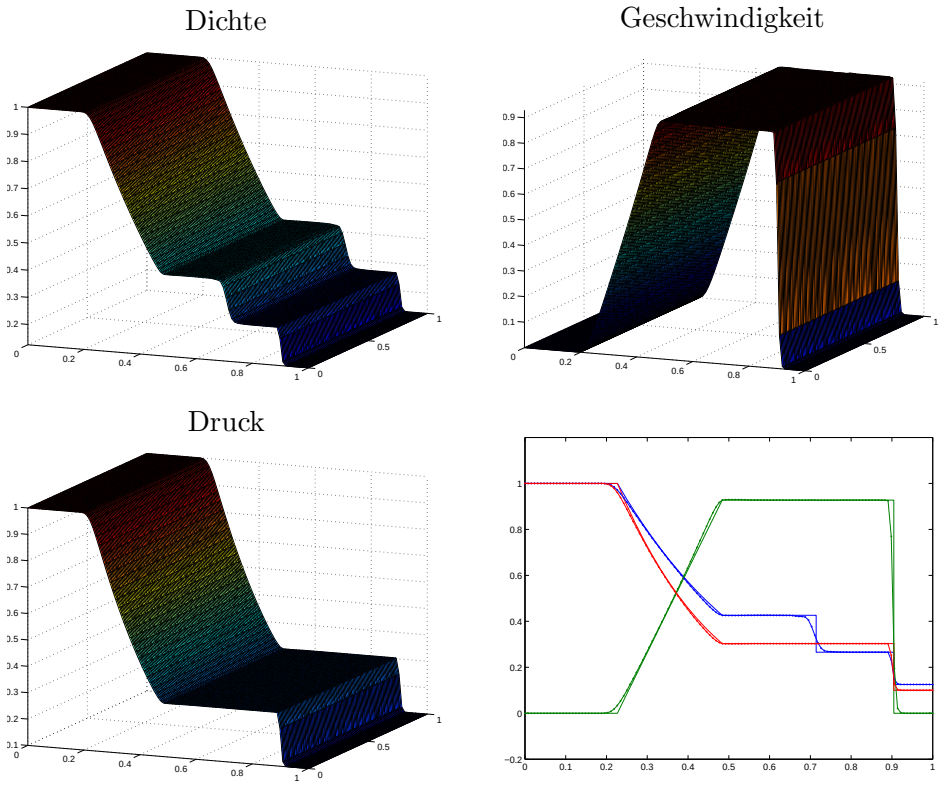


Abbildung 5.4: Shock Tube Problem in 2D. CN/FEM-FCT, $t = 0.231$.

Die Anwendung des iterativen Limiters bei gleicher Synchronisierung der Korrekturfaktoren liefert zu denen in Abbildung 5.4 identische Ergebnisse, so daß wir auf ihre Darstellung verzichtet haben. Dies überrascht nicht weiter, da der Zeitschritt aus Genauigkeitsgründen so klein gewählt werden muß, daß sich die Vorteile der iterativen Formulierung nicht bemerkbar machen.

Abbildung 5.5 zeigt die numerischen Resultate, wenn für die Synchronisierung des iterativen Limiters das Minimum aus Dichte, Energie und Impuls in x–Richtung gewählt wird, was dem quasi-eindimensionalen Charakter des Problems Rechnung trägt. Man erkennt bereits etwas 'überschüssige' Diffusion im rechten Teil des Expansionsfächers, die jedoch im Vergleich zur Basis Formulierung mit Synchronisierung über alle konservativen Variablen deutlich geringer ausfällt.

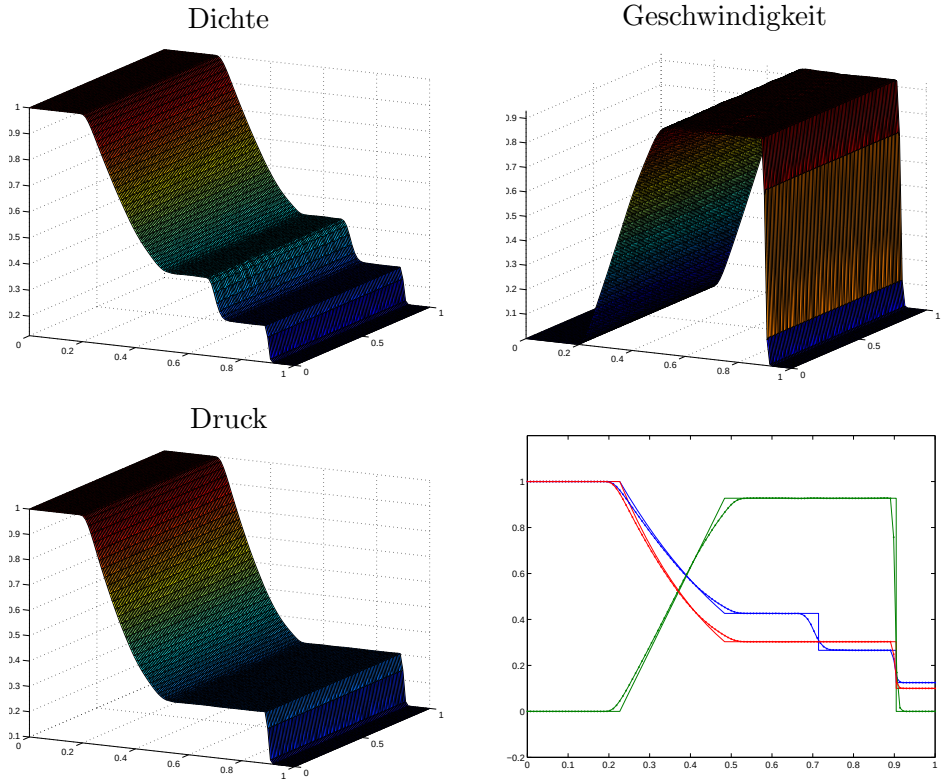


ABBILDUNG 5.5: Shock Tube Problem in 2D. CN/iteratives FEM-FCT, $t = 0.231$.

5.1.2 Radialsymmetrisches Riemann Problem

Als zweites transientes Testproblem betrachten wir den von LeVeque [47] vorgeschlagenen Benchmark, der ein numerisches Verfahren auf seine Fähigkeit hin untersucht, radiale Symmetrie zu erhalten. Zum Zeitpunkt $t = 0$ wird innerhalb einer Kreisscheibe eine höhere Dichte und ein höherer Druck vorgeschrieben als in dem sie umgebenden Gebiet. Zu Beginn befindet sich das Medium in beiden Zonen im Ruhezustand. Nachdem die 'künstliche' Membran, durch die beide Regionen voneinander getrennt werden, abrupt entfernt wird, breitet sich eine Schockwelle aufgrund des Druckunterschiedes radial-symmetrisch aus.

Die Anfangsdaten entsprechen denen von LeVeque [47]. Innerhalb des Rechengebietes $(-0.5, 0.5)^2$ sei das Anfangsprofil entsprechend

$$U(x, y, 0) = \begin{cases} U_L, & \text{für } \sqrt{x^2 + y^2} < 0.13, \\ U_R, & \text{für } \sqrt{x^2 + y^2} \geq 0.13 \end{cases} \tag{5.3}$$

gewählt, wobei für die Geschwindigkeit $\mathbf{v} = \mathbf{0}$ gilt, sowie für Dichte und Druck

$$\rho_L = 2, \quad p_L = 15, \qquad \rho_R = 1, \quad p_R = 1. \tag{5.4}$$

Die numerischen Ergebnisse wurden auf einem uniformen Gitter mit 128×128 Q_1-Elementen bei einem Zeitschritt von $\Delta t = 10^{-3}$ erzeugt und sind jeweils zum Zeitpunkt $t = 0.13$ bei einer Auflösung von 20 Konturlinien wiedergegeben.

In Abbildung 5.6 (links) ist die resultierende Dichteverteilung bei Anwendung der Basis Formulierung dargestellt, die eine exzellente Symmetrie aufweist. Auf der rechten Seite ist ein Schnitt durch das Rechengebiet entlang der x-Achse wiedergegeben, wobei die numerische Lösung durch Punkte gekennzeichnet ist. Die durchgezogene Linie stellt im Vergleich dazu die Simulationsergebnisse dar, die auf einem stark verfeinerten Gitter mit 1.048.576 Elementen erzeugt wurde, was einer Diskretisierung der x-Achse mit 1.025 Punkten entspricht. Wir möchten darauf hinweisen, daß die dargestellten numerischen Lösungen im Vergleich zu anderen veröffentlichten tatsächlich symmetrisch sind und eine gute Auflösung der beiden Unstetigkeiten nahe des Ursprungs und am Rand aufweisen.

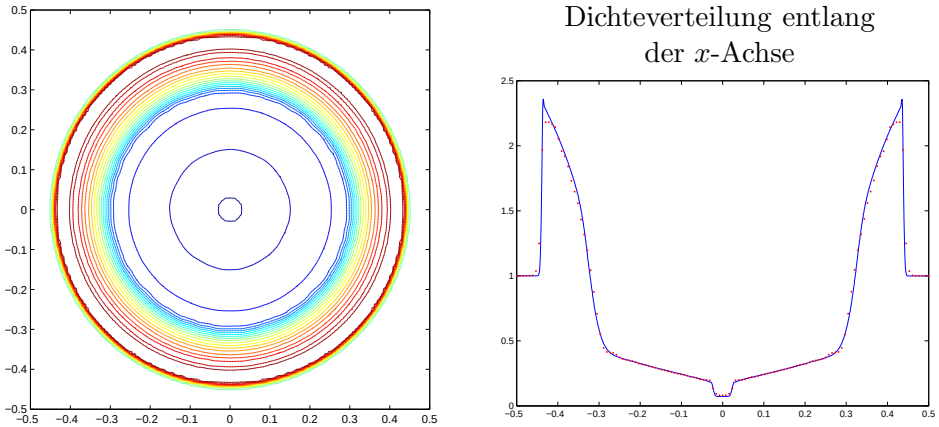


Abbildung 5.6: Radial-symmetrisches Riemann Problem, $t = 0.13$.

Abbildung 5.7 zeigt die Simulationsergebnisse, die mit Hilfe des iterativen Limiters berechnet wurden. Sowohl bei der Synchronisierung der Korrekturfaktoren über das Minimum aus Dichte und Energie als auch bei der Synchronisierung über alle Variablen bleiben die hervorragenden Symmetrieeigenschaften vollständig erhalten. Für letzteres Vorgehen erkennt man die bereits zuvor beobachtete leichte Zunahme der Diffusivität.

Abweichend vom üblichen Vorgehen, einen zweiten Schnitt entlang der Diagonalen $x = y$ zu legen [2], haben wir in Abbildung 5.8 eine Cutline um 5.37° gegen den Uhrzeigersinn gedreht (bezogen auf die x-Achse) dargestellt, um so eine Abhängigkeit der Symmetrie vom kartesischen Gitter weitgehend auszuschließen. Links ist die auf dem feinen Gitter berechnete 'Referenzlösung' dargestellt, die ebenso wie die auf dem gröberen Gitter berechnete Lösung (Punkte) eine sehr gute Übereinstimmung mit der Lösung entlang der x-Achse (Linie) besitzt.

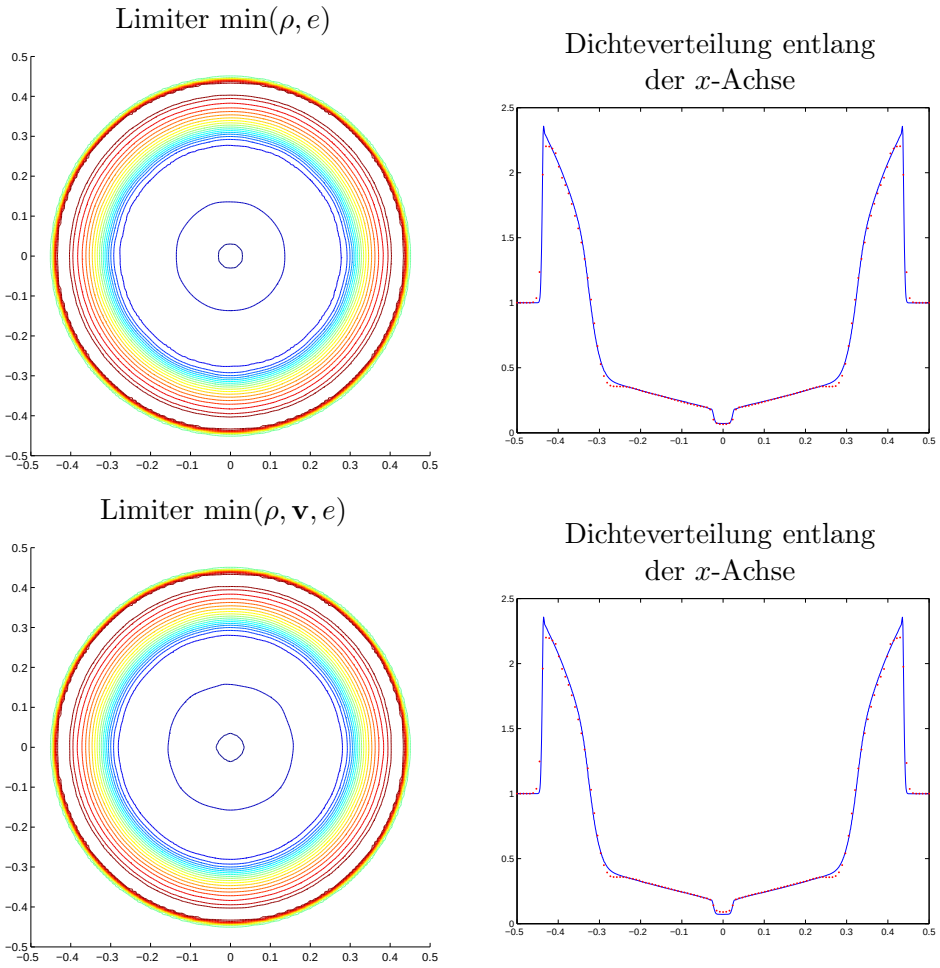


ABBILDUNG 5.7: Radial-symmetrisches Riemann Problem, $t = 0.13$.

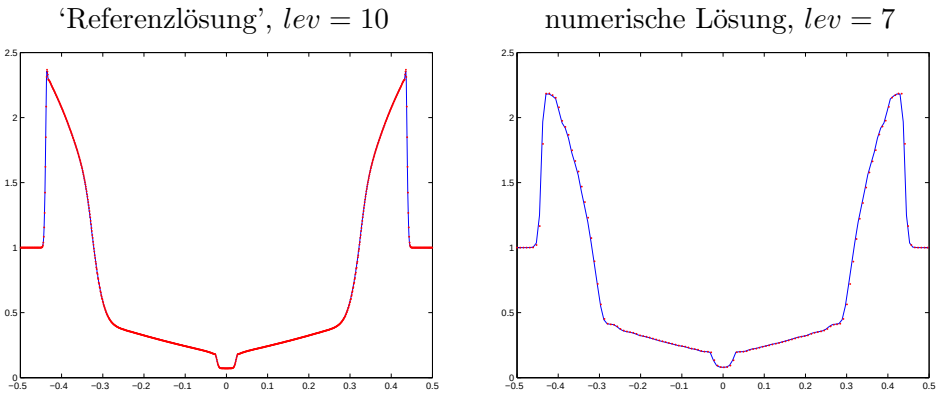


ABBILDUNG 5.8: Radial-symmetrisches Riemann Problem, $t = 0.13$.

5.2 Stationäre Benchmarks

Der zweite Teil des fünften Kapitels ist stationären Testfällen gewidmet und soll auch dafür die Leistungsfähigkeit von FEM-FCT unter Beweis stellen. Da der Zeitschritt Δt für diese Art von Problemen zu einem künstlichen Relaxationsparameter wird, kann er aus Sicht der Zeitgenauigkeit beliebig groß gewählt werden. Wir werden im folgenden die vollimplizite Backward Euler Zeitdiskretisierung ($\theta = 1$) verwenden, da sie uneingeschränkt positivitätserhaltend ist. Für große Zeitschritte werden wir Unterschiede zwischen der Basis Formulierung, die aufgrund der Zeitschrittabhängigkeit des Zalesak-Limiters zu diffusiveren Ergebnissen führt, und dem iterativen FEM-FCT Algorithmus erkennen.

5.2.1 Oblique Shocks

In Abschnitt 5.1.1 haben wir uns mit der Entstehung von Schockwellen beschäftigt, die normal zur Ausbreitungsrichtung verlaufen. In der Aerodynamik gibt es eine Vielzahl von Verdichtungsstößen, die sich in irgendeiner Form 'schräg' bezüglich der Strömungsrichtung ausbreiten. Diese sogenannten *oblique shocks* entstehen beispielsweise immer dann, wenn eine supersonische Strömung auf ein keilförmiges Hindernis wie eine Projektilspitze trifft. In diesem Abschnitt werden wir kurz auf die analytische Berechnung solcher Shocks eingehen und anschließend beide FEM-FCT Formulierungen auf den bekannten *compression corner* Benchmark anwenden. Dabei werden wir teilweise ein vollkommen unstrukturiertes Gitter verwenden, um numerisch zu demonstrieren, daß sich die neuen FEM-FCT Verfahren auf beliebigen Gittern anwenden lassen. In Abbildung 5.9 ist ein Prototyp einer *compression corner* dargestellt. Dabei bezeichnet θ den tatsächlichen Ablenkungswinkel und $\theta_{\max}$ den maximalen Ablenkungswinkel, bei dem der Schock noch 'attached' zum Hindernis ist.

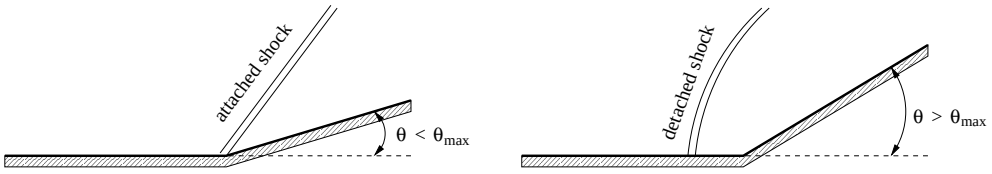


Abbildung 5.9: Schockwellen an einer *compression corner*.

Aus der allgemeinen Theorie von *oblique shocks* ist die $\theta - \beta - M$ Beziehung bekannt, die θ mit der (*upstream*) Machzahl M_1 *vor* dem Schock und dem resultierenden Verdichtungsstoßwinkel β in Relation setzt [1]

$$\tan\theta = 2\cot\beta \frac{M_1^2 \sin^2\beta - 1}{M_1^2(\gamma + \cos 2\beta) + 2}. \tag{5.5}$$

Die in Anschnitt 5.1.1 betrachteten normalen Schocks sind ein Spezialfall mit $\beta = \pi/2$, so daß die Zustandsänderungen über den Schock alleine durch M_1 bestimmt werden. Typischerweise verwendet man Gleichung (5.5), um ein sogenanntes $\theta - \beta - M$–Diagramm (etwa in [56]) zu erzeugen, in dem ausgezeichnete Niveaulinien der Machzahl in der $\theta - \beta$–Ebene dargestellt sind. Zum einen erkennt man anhand einer solchen Darstellung, daß es abhängig von M_1 einen maximalen Ablenkungswinkel $\theta_{\max}$ gibt, so daß für $\theta > \theta_{\max}$ keine Lösung der Gleichung (5.5) für einen geradlinigen, schrägen Schock existiert. In der Praxis äußert sich diese Tatsache darin, daß sich ein von der Keilspitze losgelöster gekrümmter Verdichtungsstoß ausbildet (vgl. Abbildung 5.9, rechts). Wir wollen uns hier mit der Situation $\theta < \theta_{\max}$ befassen, für die sich aus Gleichung (5.5) zwei Lösungen für einen geradlinigen, schrägen Schock ergeben: den starken und den schwachen Verdichtungsstoß. Ohne den physikalischen Hintergrund zu erläutern, gehen wir davon aus, daß sich der schwache Schock in der Praxis durchsetzt, so daß wir eine Situation wie in Abbildung 5.9 links vorfinden.

Für den Benchmark betrachten wir eine tangential zur x–Achse verlaufende supersonische Strömung mit $M = 2.5$, die auf einen Keil mit einem Ablenkungswinkel von 15° trifft. Dieser Testfall ist Bestandteil der NPARC *CFD Verification and Validation* Datenbank [57] und dort hinreichend dokumentiert. Gleichung (5.5) ergibt einen Winkel von $\beta = 36.94°$ für die exakte Lösung des schwachen Verdichtungsstoßes. Für die Berechnung der Machzahl M_2 *hinter* dem Verdichtungsstoß möchten wir auf die Literatur [1] verweisen und lediglich bemerken, daß sich für den vorliegenden Benchmark ein Wert von $M_2 = 1.87$ ergibt.

Zunächst wurden die Simulationen auf einem strukturierten Gitter mit 128×128 Q_1-Elementen bei einem Zeitschritt von $\Delta t = 10^{-2}$ durchgeführt. Ein so kleiner Zeitschritt ist trotz der uneingeschränkten Stabilität der Backward Euler Zeitdiskretisierung notwendig, um die Konvergenz der nichtlinearen Iteration zu sichern. Diese Tatsache überrascht nicht weiter, da ein entkoppelter Löser nicht robust genug ist, so daß sich für stationäre Strömungen ein vollgekoppelter Ansatz anbieten würde [28], [73], auf den wir kurz im Ausblick eingehen werden.

Abbildung 5.10 zeigt die Ergebnisse, die mit Hilfe von discrete Upwind berechnet wurden. Man erkennt, daß der qualitative Lösungsverlauf bereits von der Methode niedriger Ordnung erfaßt wird. Ferner sind die Werte von M_1 und M_2 korrekt erfaßt. Die idealerweise 'dünne' Schockwelle wird jedoch aufgrund der starken Diffusivität zu einem 'Fächer' verschmiert. Dennoch ist es, wie schon im skalaren Fall, durchaus empfehlenswert, zunächst mit der Methode niedriger Ordnung eine provisorische Lösung bis zum stationären Limes zu berechnen und ihre Ortsgenauigkeit *anschließend* durch die Hinzunahme einer der beiden FEM-FCT Verfahren zu verbessern. Bereits die Basis Formulierung liefert bei gleicher Zeitschrittweite eine deutlich genauere Auflösung der Schockwelle.

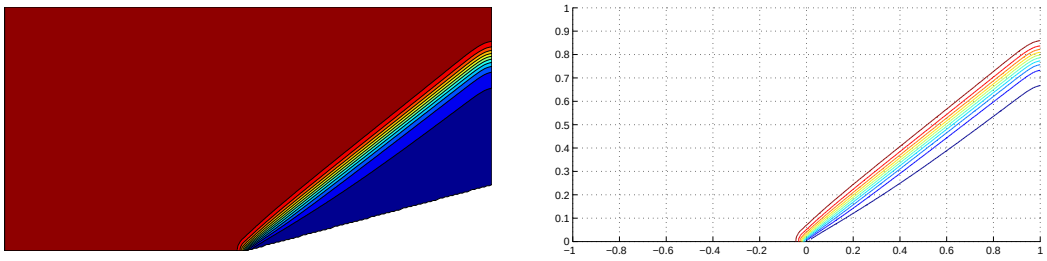


ABBILDUNG 5.10: Discrete Upwinding, $M = 2.5$, $\theta = 15°$

Wenn wir die in Abbildung 5.11 dargestellten Ergebnisse betrachten, so stellen wir fest, daß die Schockdicke über das gesamte Rechengebiet nahezu gleichmäßig einer Breite von $5 - 6$ Zellen entspricht.Wir möchten darauf hinweisen, daß die Genauigkeit, mit welcher der Schock reproduziert wird, aufgrund der Zeitschrittabhängigkeit des Zalesak-Limiters vom Wert des Zeitschritts Δt abhängt.

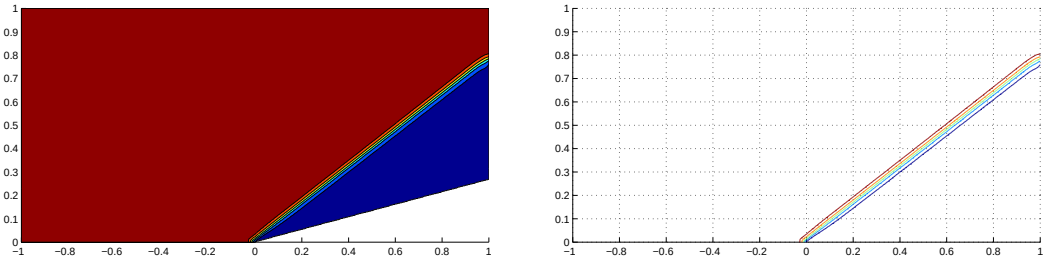


ABBILDUNG 5.11: Basis Formulierung, $M = 2.5$, $\theta = 15°$

Beim Einsatz der iterativen FEM-FCT Formulierung (vgl. Abbildung 5.12) wird dieser Nachteil überwunden, so daß wir bei ansonsten unveränderter Konfiguration eine Schockdicke von $3 - 4$ Zellen bekommen, welche *unabhängig* von der Wahl des Zeitschritts ist. Dieser Vorteil muß jedoch mit der Zunahme der nichtlinearen Iterationszahl pro Zeitschritt erkauft werden, damit die zurückgewiesene Antidiffusion effektiv 'recycled' werden kann.

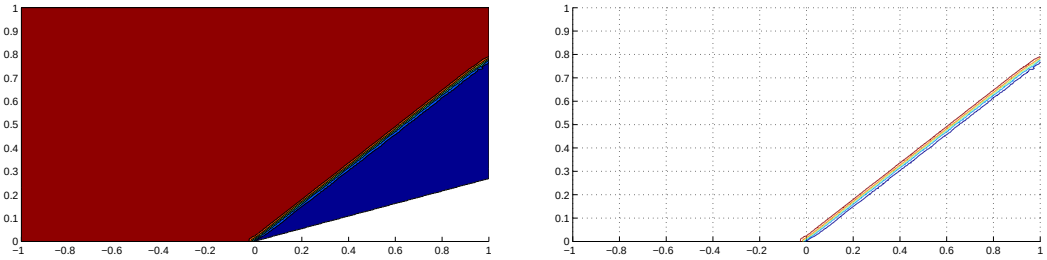


ABBILDUNG 5.12: iteratives FEM-FCT, $M = 2.5$, $\theta = 15°$

Im folgenden werden wir für diesen Benchmark ein unstrukturiertes Gitter verwenden, um auch numerisch zu belegen, daß die vorgestellten FEM-FCT Formulierungen vollkommen unabhängig von der zugrunde liegenden Gittertopologie

sind. Da für diesen Testfall die Struktur der exakten Lösung *a priori* bekannt ist, haben wir das in Abbildung 5.13 dargestellte Grobgitter mit zweimaliger Verfeinerung benutzt, was einer Zahl von 10.016 Viereckselementen entspricht.

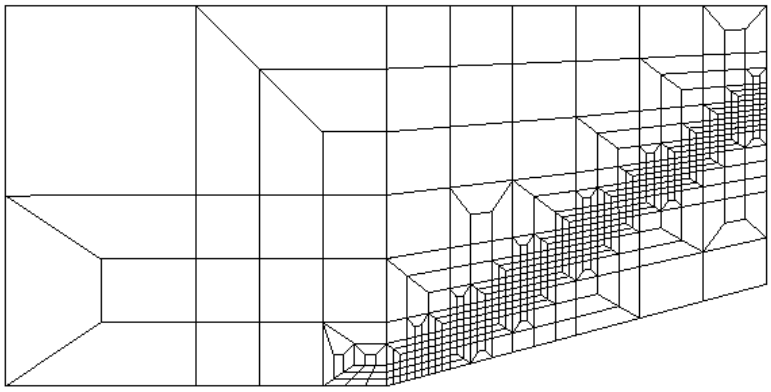


Abbildung 5.13: Adaptives Gitter

In Abbildung 5.14 sind die auf diesem Gitter mit der iterativen FEM-FCT Formulierung berechneten Simulationsergebnisse dargestellt. Trotz extremer Größenunterschiede zwischen den Elementen ist die Genauigkeit, mit welcher der Schock aufgelöst wird, überzeugend. Die Schockdicke entspricht etwa der Breite von $5-6$ Zellen, wobei zu berücksichtigen ist, daß diese im Bereich des Schocks kleinere Ausmaße als im gleichmäßigen Gitter haben. Betrachtet man den absoluten Wert der Schockdicke, so stimmt dieser für beide Gitter nahezu überein und liegt bei etwa 0.016. Die Basis Formulierung führt zu identischen Ergebnissen, so daß wir sie hier nicht präsentieren. Diese Ergebnisse belegen, daß die neuen FEM-FCT Verfahren auf beliebigen auch nicht rechteckigen Gittern verwendet werden können.

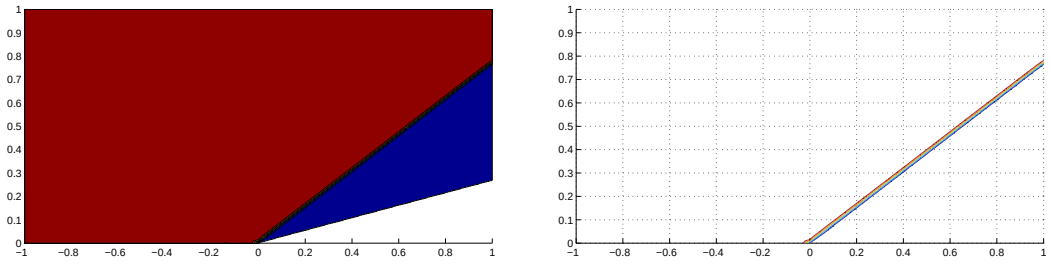


Abbildung 5.14: iteratives FEM-FCT, $M = 2.5$, $\theta = 15°$

5.2.2 Prandtl-Meyer Eckenströmung

Als letzten Testfall möchten wir die Entstehung eines Expansionsfächers bei der Umströmung eines Eckenprofils simulieren. Im Gegensatz zu den in Abschnitt 5.2.1 betrachteten *oblique shocks*, die entstehen, wenn eine supersonische Strömung 'in sich hinein gelenkt' wird, zeichnen sich *expansion waves* durch einen stetigen Übergang der Zustandsvariablen aus und entstehen, wenn eine supersonische Strömung 'von sich weg gelenkt' wird. Wir werden uns im folgenden auf die von Prandtl und Meyer um 1907 analysierten zentrierten Expansionswellen beschränken, die wie in Abbildung 5.15 dargestellt entlang einer scharfkantigen konvexen Ecke entstehen. Das stetige Expansionsgebiet ist von der vorderen und hinteren Machlinie begrenzt, deren Winkel sich aus der geometrischen Beziehung

$$\mu_1 = \sin^{-1}(1/M_1), \qquad \mu_2 = \sin^{-1}(1/M_2) \tag{5.6}$$

ergeben, wobei M_1 und M_2 die Machzahl vor bzw. hinter der Expansionswelle bezeichnet. Zur Bestimmung der Machlinienwinkel und der Änderungen der Zustandsgrößen dazwischen ist es also notwendig, den Wert von M_2 zu berechnen.

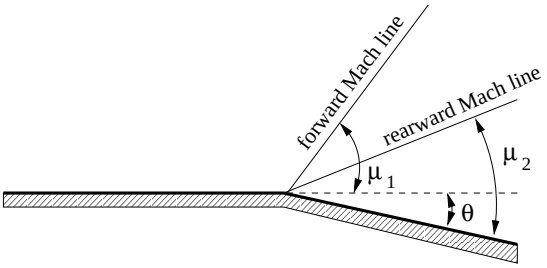


Abbildung 5.15: Schockwellen an einem keilförmigen Hindernis.

Wir möchten auf ihre Herleitung verzichten und die sogenannte Prandtl-Meyer Funktion (für ein kalorisch perfektes Gas) einführen [1]

$$\nu(M) = \sqrt{\frac{\gamma+1}{\gamma-1}}\tan^{-1}\sqrt{\frac{\gamma-1}{\gamma+1}(M^2-1)} - \tan^{-1}\sqrt{M^2-1}, \tag{5.7}$$

welche zu $\nu(M=1)=0$ normiert wird. Zusammen mit der Beziehung

$$\theta = \nu(M_2) - \nu(M_1) \tag{5.8}$$

läßt sich aus den Werten des Ablenkungswinkels θ und der Machzahl M_1 vor dem Expansionsgebiet der Wert von M_2 sowie die Änderung aller Zustandsvariablen aus den bekannten Relationen für isentropische Gase berechnen [1].

Für den folgenden Benchmark haben wir wie an Abschnitt 5.2.1 eine *upstream* Machzahl von $M_1 = 2.5$ und einen Ablenkungswinkel $\theta = 15°$ gewählt. Auch dieser Testfall ist in der Datenbank von NPARC [57] dokumentiert. Aus Gleichung (5.7) ergibt sich $\nu(M_1) = 39.12$ und zusammen mit Beziehung (5.8) $M_2 = 3.25$, was für die Winkel der Machlinien die Werte $\mu_1 = 23.58°$ bzw. $\mu_2 = 18.0°$ liefert. Die Herausforderung bei diesem Testfall besteht neben der korrekten Reproduktion der Machlinienwinkel und der Machzahl hinter dem Expansionsgebiet darin, das Zentrum aller Machwellen an der Unstetigkeitsstelle möglichst exakt wiederzugeben. Wir verwenden wieder ein gleichmäßiges Gitter mit 128×128 Q_1-Elementen und wählen einen moderaten Zeitschritt von $\Delta t = 10^{-2}$.

Abbildung 5.16 zeigt die mit Hilfe von discrete Upwind berechnete numerische Lösung sowie 20 Konturlinien. Auch wenn die maximale Machzahl mit $M_2 = 3.26$ mit der 'exakten' übereinstimmt, ist die Unstetigkeitsstelle in einem nicht zu akzeptierenden Maße verschmiert. Wie bereits zuvor angesprochen, wollen wir im folgenden aus Effizienzgründen von dieser 'konvergierten' Lösung als '*educated guess*' für den anschließenden FCT Löser ausgehen.

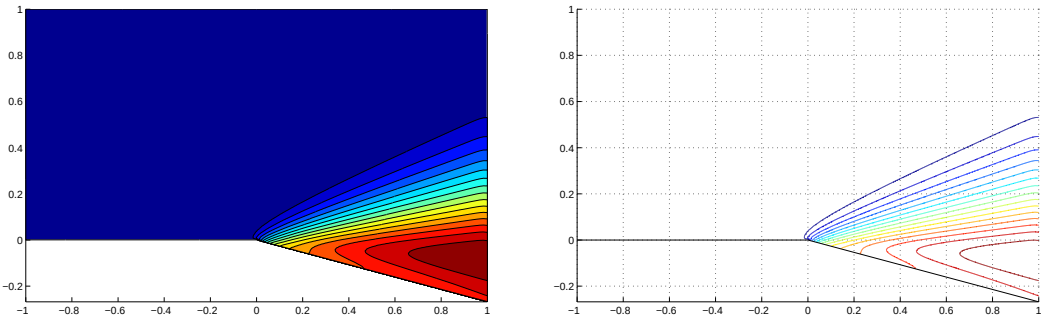


ABBILDUNG 5.16: discrete Upwind, $M = 2.5$, $\theta = 15°$

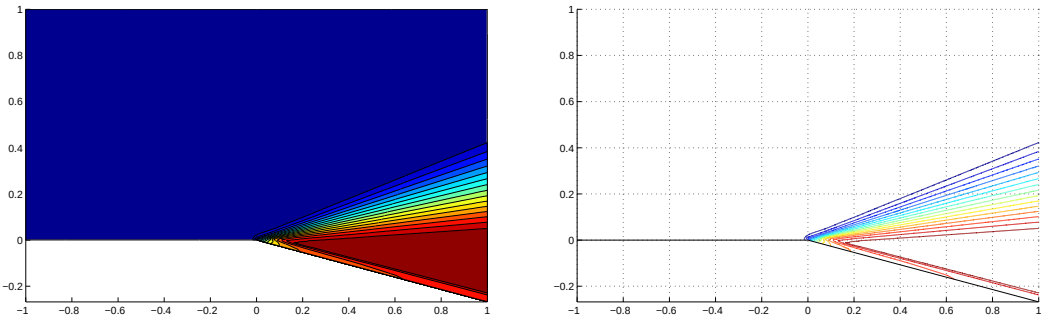


ABBILDUNG 5.17: Basis Formulierung, $M = 2.5$, $\theta = 15°$

Beide FCT Varianten reproduzieren die exakte Lösung mit einer hohen Genauigkeit. Da für stetige Lösungsprofile kaum eine Flußkorrektur notwendig ist, unterscheidet sich die Basis Formulierung kaum vom iterativen FEM-FCT Algo-

rithmus. Im Gegensatz zu den mit discrete Upwinding erzeugten Ergebnissen erkennt man deutlich die *Geradlinigkeit* der Stromlinien, die sich strahlenförmig von einem stark lokalisierten Zentrum ausbreiten.

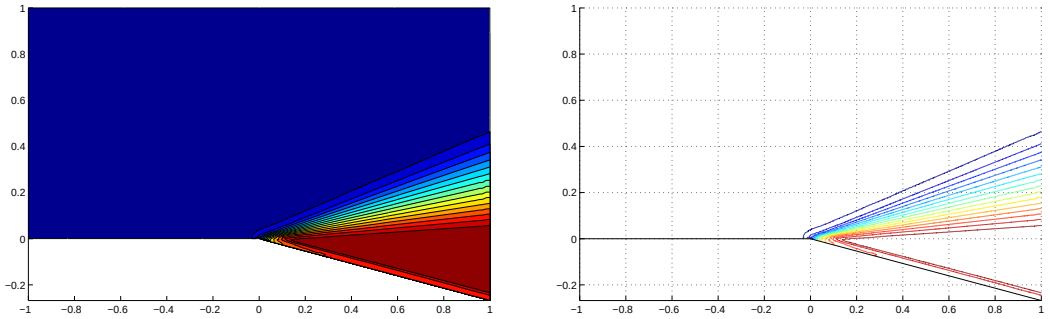


Abbildung 5.18: iteratives FEM-FCT, $M = 2.5$, $\theta = 15°$

Zusammenfassung und Ausblick

Ausgehend von der klassischen FCT Theorie haben wir in dieser Arbeit eine neue Familie von *flux-corrected transport* Schemata vorgestellt und sie für die kompressiblen Eulergleichungen verallgemeinert.

Für skalare Erhaltungsgleichungen haben wir mathematisch fundierte Kriterien zur Konstruktion von positivitätserhaltenden Diskretisierungen hergeleitet und mit ihrer Hilfe eine entsprechende Methode niedriger Ordnung entwickelt. Dazu wurden nach der Durchführung von konservativem *mass lumping* zur Entfernung der impliziten Antidiffusion aus der konsistenten Massenmatrix alle negativen Nebendiagonaleinträge des Transportoperators hoher Ordnung durch Addition von künstlicher Viskosität eliminiert, um so eine LED Diskretisierung des konvektiven Anteils zu erhalten. Dabei wurde die bereits vorhandene physikalische Diffusion berücksichtigt, so daß *discrete Upwinding* unter allen positivitätserhaltenden Verfahren das am wenigsten diffusive ist. Ein möglicherweise vorhandener Quellterm konnte mit Hilfe eines einfachen Splittings linearisiert werden, um die auf dem Konzept von M-Matrizen basierenden Positivitätskriterien zu erfüllen.

Eine konservative Galerkin Flußzerlegung wurde vorgestellt, die im Vergleich zu einer auf der kantenbasierten Datenstruktur von Peraire *et al.* [60] beruhenden Zerlegung [53] nicht auf Dreiecks- und Tetraedergitter beschränkt ist. Für das weitere Vorgehen reichte es aus, nur die antidiffusiven Terme als Flüsse zwischen im Sinne des Konnektivitätsgraphen der Finiten Elemente Matrix benachbarten Knoten darzustellen, so daß die vorgestellte FCT Methodik in bestehende FEM-Codes, die einen konventionellen elementbasierten Zugang verwenden, eingebaut werden kann. Ferner war ein vollständiger Übergang zu einer aus Effizienzgründen empfehlenswerten kantenbasierten Datenstruktur möglich. Die Antisymmetrie der Flüsse garantierte die strikte Massenerhaltung und ermöglichte die Übertragung der Theorie für 1D mit Hilfe von *flux difference splitting* Techniken.

Die Diskretisierungen hoher und niedriger Ordnung wurden in einer FEM-FCT Prozedur vereint, um eine hochgenaue Endlösung ohne unphysikalische Oszillationen zu erhalten. Auf der Basis von mathematisch fundierten Kriterien wurde

gezeigt, daß Zalesaks mehrdimensionaler Limiter gleichermaßen für explizite wie implizite Zeitdiskretisierungen garantiert positivitätserhaltend ist. Zur Korrektur der bei nicht verschwindendem Gradienten des Lösungsprofils am Rand entstehenden Oszillationen haben wir einen einfachen Postlimiting Schritt vorgestellt. Zur Erklärung des pathologischen Verhaltens des Zalesak-Limiters am Rand haben wir das sogenannte Hebelmodell eingeführt. Dieses diente ferner der Rechtfertigung eines auf Boris und Book [4] zurückgehenden Prelimiting Schrittes, der durch den Übergang zu einer flußbasierten Darstellung des antidiffusiven Terms erstmals im Kontext von Finiten Elementen anwendbar war.

Für die vollimplizite Backward Euler Methode konnte die uneingeschränkte Stabilität nachgewiesen werden. Für semi-implizite und explizite Zeitschrittverfahren ließ sich aus den Positivitätskriterien eine rigorose obere Schranke für den maximal zulässigen Zeitschritt angeben. Der Freiheit bei der Wahl des Zeitschritts im vollimpliziten Fall wurden jedoch Grenzen durch die Zeitschrittabhängigkeit des Zalesak-Limiters gesetzt, da die Basis Formulierung des FCT Algorithmus mit wachsender Größe des Zeitschritts zunehmend diffusiver wurde. Dieser Nachteil konnte durch die besonders für (semi-)implizite Diskretisierungen geeignete iterative Limitertechnik aufgehoben werden, bei der sich die in einem Iterationsschritt zurückgewiesene Antidiffusion 'recyclen' ließ , so daß sukzessive mehr Antidiffusion in die Endlösung integriert wurde.

Zur iterativen Behandlung von Nichtlinearitäten und impliziten Zeitschrittverfahren wurde eine mit dem Operator niedriger Ordnung vorkonditionierte Fixpunkt-Defektkorrektur vorgeschlagen, dessen M-Matrixeigenschaft zu einem gut konditionierten linearen Gleichungssystem für das Lösungsinkrement führte.

Den zweiten Teil dieser Arbeit machte die Verallgemeinerung der skalaren Theorie auf die kompressiblen Eulergleichungen mit Hilfe von *flux difference splitting* Techniken aus. Als direkte Verallgemeinerung des skalaren LED Konzeptes wurde die verwendete Diskretisierung niedriger Ordnung durch Addition von künstlicher Viskosität gewonnen, die alle Nebendiagonalblöcke zu positiv definiten Matrizen transformierte. Um die negativen Eigenwerte der lokalen Jacobimatrizen zu eliminieren, konnte der approximative Riemann Löser von Roe verwendet werden. Ferner haben wir aus Performancegründen den Einsatz von *scalar limited dissipation* empfohlen, die darüber hinaus in Kombination mit einem Flußkorrektur Algorithmus bessere Ergebnisse lieferte.

Zalesaks Limiter wurde auf hyperbolische Gleichungssysteme angewandt. Dazu mußte eine Synchronisierung der individuellen Korrekturfaktoren durchgeführt werden. Aufbauend auf den Vorschlägen von Löhner [52] zur Wahl der für die Synchronisierung relevanten Korrekturfaktoren haben wir eine Technik vorgestellt [49], bei der die Korrekturfaktoren nach einer Transformation für einen

beliebigen Satz an Variablen berechnet und nach der Synchronisierung auf die konservativen Flüsse angewandt wurde.

Die Implementierung von Randbedingungen für Finite Elemente im Bereich von CFD-Anwendungen wurde angesprochen. Wir haben einen einfachen semi-impliziten Zugang vorgestellt, bei dem die Randwerte über den Umweg von Riemann Invarianten in den Vektor der konservativen Variablen direkt eingesetzt wurden. Eine Extrapolation von herausgehenden Riemann Invarianten war dabei nicht erforderlich. Für *free-slip* Ränder ließ sich dieses Vorgehen noch abkürzen.

Ein kantenbasierter Matrixaufbau wurde vorgeschlagen. Dieser ließ insbesondere in Kombination mit einer iterativen Fixpunkt-Defektkorrektur und der Verwendung von skalarer Viskosität eine effiziente Implementierung zu. Ein besonders für transiente Probleme geeigneter Block-Jacobi Löseransatz wurde vorgestellt, der durch eine Vorkonditionierung mit den Diagonalblöcken des diskreten Operators niedriger Ordnung auf eine Sequenz von relativ gut konditionierten, linearen Gleichungssystemen führte.

Die numerischen Ergebnisse haben das Potential der neuen FEM-FCT Methodik bestätigt. Sowohl für skalare Benchmarks als auch für die kompressiblen Eulergleichungen produzierten die Verfahren hochgenaue und absolut oszillationsfreie Lösungen. Ferner konnten diese durch den Einsatz von Prelimiting mehr als 'kosmetisch' verbessert werden. In verschiedenen Testkonfigurationen konnte gezeigt werden, daß der zugrunde liegende Algorithmus 'echt mehrdimensional' und auf beliebige (un-)strukturierte Gitter anwendbar ist (vgl. das stochastisch gestörte Gitter aus Abb. 2.20 oder das a-priori adaptierte Gitter aus Abb. 5.13). Hierin liegt der eindeutige Vorteil der Finiten Elemente gegenüber Finiten Differenzen. In Zukunft soll die FEM-FCT Methodik – aufbauend auf den Arbeiten von Hartmann, Houston und Suli [26], [27] – um a-posteriori eine adaptive Gitteranpassung erweitert werden. Vorstellbar wäre auch, den Wert der Korrekturfaktoren als einen qualitativen Indikator für die Gitterverfeinerung zu nutzen, da der Limiter Unstetigkeiten in der Lösung ohne zusätzlichen Aufwand detektiert. Gleichzeitig soll der Einsatz von Finiten Elementen höherer Ordnung untersucht werden, wodurch die Voraussetzungen für ein auf h/p–Adaptivität basierendes Verfahren geschaffen werden.

Insbesondere zur Simulation von stationären Strömungen hat sich der iterative Limiter als 'Wunderwaffe' herausgestellt. Daß dieser noch durch die notwendige Verwendung moderater Zeitschritte ausgebremst wurde, lag an dem eingesetzten Block-Jacobi Löser. Für die folgenden Vergleichstest haben wir den stationären *compression corner* Benchmark ausgewählt und für unterschiedlichen Zeitschrittweiten zwischen $10^{-3} - 1$ jeweils 10 Zeitschritte mit discrete Upwinding simuliert. Die nachfolgenden Tabellen zeigen die Gesamtzahl an durchgeführten nichtlinea-

ren Iterationen sowie die durchschnittliche Anzahl an linearen pro nichtlinearen Iteration in Anhängigkeit vom verwendeten Gitterlevel.

			Δt			
NLEV	NVT	NEL	1.0	0.1	0.01	0.001
6	4425	4096	NaN	NaN	2/54	1/30
7	16641	16384	NaN	NaN	3/85	1/34
8	66049	65536	NaN	NaN	NaN	2/42

Tabelle 5.1: Anzahl linearer/nichtlinearer Iterationen: Block-Jacobi.

Allererste Experimente mit einem gekoppelten Ansatz lassen bereits erkennen, daß hier noch bedeutende Fortschritte zu erwarten sind. Da detaillierte Vergleichsrechnungen mit dem neuen Code noch ausstehen, möchten wir die nachfolgenden Ergebnisse mehr als Tendenz für die weitere Entwicklung verstanden wissen. Als erstes fand ein mit einem (Block-)Gauss-Seidel Vorkonditionierer ausgestattetes BiCGSTAB-Verfahren zur Lösung des linearen Gleichungssystems $C^{(m)}\Delta U^{(m+1)} = R^{(m)}$ (vgl. (4.41)) Verwendung. Erste Tests ergaben, daß dieser Übergang von einer approximativen blockdiagonalen Defektkorrektur zu einem im Blocksinne 'vollbesetzten' Vorkonditionierer innerhalb einer vollen Fixpunkt Defektkorrektur die Verwendung von großen Zeitschritten ermöglicht.

			Δt			
NLEV	NVT	NEL	1.0	0.1	0.01	0.001
6	4425	4096	23.4/60	14.2/78	3.9/48	2/30
7	16641	16384	45.7/68	28.4/91	5.7/62	2/35
8	66049	65536	100.6/81	63.5/109	9.7/75	2.1/42

Tabelle 5.2: Anzahl linearer/nichtlinearer Iterationen: BiCGSTAB.

Um die Abhängigkeit des linearen Lösers von der Gitterweite weiter zu reduzieren ($\#\text{ite}_{BiCGSTAB} \sim h^{-1}$), haben wir die Mehrgitterroutinen aus der FEAT-Bibliothek [14] eingebunden. Die folgende Tabelle verdeutlicht, daß unser Ansatz einen Schritt in die richtige Richtung für vollimplizite Verfahren mit beliebigen Zeitschrittweiten darstellt: Der Einsatz von Mehrgitterverfahren bringt eine CPU-Ersparnis um den Faktor 10 und mehr. Eine genaue Analyse von Mehrgitterverfahren für die FEM-FCT Methodik soll erfolgen und ein effizienter und robuster gekoppelter Löser entwickelt werden.

			Δt			
NLEV	NVT	NEL	1.0	0.1	0.01	0.001
6	4425	4096	4.4/60	3.2/78	1.7/48	1/30
7	16641	16384	6.4/68	4.4/91	2/62	1/35
8	66049	65536	9.9/81	4.9/109	2.2/75	1/42

Tabelle 5.3: Anzahl linearer/nichtlinearer Iterationen: Mehrgitter.

Schließlich ist die Anwendung auf realistische Problemstellungen geplant. Da die verwendeten Komponenten – kantenbasierter Matrizenaufbau, discrete Upwinding und FCT Limiter – für beliebige Dimensionen gültig sind, sollte die Erweiterung auf 3D ohne konzeptionelle Schwierigkeiten durchführbar sein. Interessant wird letztendlich ein Vergleich mit mehrdimensionalen TVD Methoden [41] sein, die sich direkt auf stationäre Strömungen anwenden lassen.

Literaturverzeichnis

[1] J. D. Anderson, Jr., *Modern Compressible Flow*, McGraw-Hill, 1990.

[2] ANumE, Web Site: `http://rubens.math.uni-magdeburg.de/~anume/`

[3] H. Blank, M. Rudgyard und A. Wathen, Stabilised finite element methods for steady incompressible flow. *Comput. Methods Appl. Mech. Engrg.* **174** (1999), no. 1–2, 91–105.

[4] J. P. Boris und D. L. Book, Flux-corrected transport. I. SHASTA, A fluid transport algorithm that works. *J. Comput. Phys.* **11** (1973) 38–69.

[5] A. N. Brooks und T. J. R. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.* **32** (1982) 199–259.

[6] J. M. Burgers, A mathematical model illustrating the theory of turbulence, *Adv. Appl. Mech.* **1** (1948) 171–199.

[7] G. F. Carey und B. N. Jiang, Least-squares finite elements for first-order hyperbolic systems. *Int. J. Numer. Meth. Fluids* **26** (1988) 81–93.

[8] G. F. Carey und R. J. MacKinnon, Positivity-preserving, flux-limited finite-difference and finite-element methods for reactive transport. *Int. J. Numer. Meth. Fluids* **41** (2003) 151–183.

[9] P. Colella und P. R. Woodward, The piecewise parabolic method (PPM) for gas dynamical simulations. *J. Comput. Phys.* **54** (1984) 174–201.

[10] M. Crouzeix und P. A. Raviart, Conforming and nonconforming finite elements for solving the stationary Stokes equations. *RAIRO, Série Rouge Anal. Num.* **7** (1973), R-3, 33–76.

[11] C. R. DeVore, An improved limiter for multidimensional flux-corrected transport. *NASA Technical Report* AD-A360122 (1998).

[12] J. Donea, L. Quartapelle und V. Selmin, An analysis of time discretization in the finite element solution of hyperbolic problems. *J. Comput. Phys.* **70** (1987) 463–499.

[13] J. Donea, V. Selmin und L. Quartapelle, Recent developments of the Taylor-Galerkin method for the numerical solution of hyperbolic problems. *Numerical Methods for Fluid Dynamics III*, Oxford (1988) 171–185.

[14] S. Turek *et al.*, *FEATFLOW: finite element software for the incompres sible Navier-Stokes equations.* User manual, University of Dortmund, 2000. Available at the URL *http://www.featflow.de*.

[15] C. A. J. Fletcher, The group finite element formulation. *Comput. Methods Appl. Mech. Engrg.* **37** (1983) 225–243.

[16] C. A. J. Fletcher, *Computational Techniques for Fluid Dynamics.* Springer, 1988.

[17] M. A. Fry und D. L. Book, Adaptation of Flux-Corrected Transport Codes for Modelling Dusty Flows. *Proc. 14th Int. Symp. on Shock Tubes and Waves*, New South Wales Univ. Press (1983)

[18] D. E. Fyfe, J. H. Gardner, M. Picone und M. A. Fry, Fast three-dimensional flux-corrected transport code for highly resolved compressible flow calculations. *Springer Lecture Notes in Physics* **218** (1985) 230–234.

[19] G. E. Georghiou, R. Morrow und A. C. Metaxas, An improved finite-element flux-corrected transport algorithm. *J. Comput. Phys.* **148** (1999) 605–620.

[20] S. K. Godunov, Finite difference method for numerical computation of discontinous solutions of the equations of fluid dynamics. *Mat. Sbornik* **47** (1959) 217–306.

[21] P. Hansbo, Aspects of conservation in finite element flow computations. *Comput. Methods Appl. Mech. Engrg.* **117** (1994) 423-437.

[22] A. Harten, High Resolution Schemes for Hyperbolic Conservation Laws. *J. Comput. Phys.* **49** (1983) 357–393.

[23] A. Harten, H. C. Yee und R. F. Warming, Implicit Total Variation Diminishing (TVD) Schemes for Steady-State Calcculations. *J. Comput. Phys.* **57** (1983) 327–360.

[24] A. Harten, J. M. Hyman und P. D. Lax, On finite-difference approximations and entropy conditions for shocks. *Comm. Pure Appl. Math.* **29** (1976) 297–322.

[25] A. Harten und S. Osher, Uniformly high-order accurate nonoscillatory schemes I. *SIAM J. Num. Anal.* **24** (1987) 279–309.

[26] R. Hartmann, *Adaptive Finite Element Methods for the Compressible Euler Equations.* PhD Thesis, University of Heidelberg, 2002.

[27] P. Houston, R. Hartmann und E. Suli, Adaptive Discontinuous Galerkin Finite Element Methods for Compressible Fluid Flows. *Numerical Methods for Fluid Dynamics VII*, (2001) 347–353.

[28] P. W. Hemker und B. Koren, Defect correction and nonlinear multigrid for steady Euler equations. In: W.G. Habashi and M.M. Hafez (ed.). *Computational fluid dynamics techniques.* London: Gordon and Breach Publishers, 1995, 699–718.

[29] C. Hirsch, *Numerical Computation of Internal and External Flows, Volume 2*, Wiley, 1988.

[30] A. Jameson, Computational algorithms for aerodynamic analysis and design. *Appl. Numer. Math.* **13** (1993) 383–422.

[31] A. Jameson, Positive schemes and shock modelling for compressible flows. *Int. J. Numer. Meth. Fluids* **20** (1995) 743–776.

[32] A. Jameson, Analysis and Design of Numerical Schemes for Gas Dynamics 1. Artificial Diffusion, Upwind Biasing, Limiters and Their Effect on Accuracy and Multigrid Convergence. *RIACS Technical Report* **94.15**, *Int. J. Comput. Fluid Dyn.* **4** (1995) 171–218

[33] A. Jameson, Analysis and Design of Numerical Schemes for Gas Dynamics 2. Artificial Diffusion and Discrete Shock Structure. *RIACS Report* **94.16**, *Int. J. Comput. Fluid Dyn.* **5** (1995) 1–38

[34] A. Jameson, Artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence in transonic and hypersonic flows. *AIAA Paper 86–0103*

[35] C. Johnson, The characteristic streamline diffusion finite element method. *Mat. Aplic. Comp.* **10** (1991), no. 3, 229–242.

[36] N. Kroll und R. K. Jain, Solution of Two-Dimensional Euler Equations – Experience with a Finite Volume Code. Forschungsbericht *DFVLR-FB* **87-41**, Institut für Entwurfsaerodynamik, Braunschweig.

[37] D. Kuzmin und S. Turek, Flux correction tools for finite elements. *J. Comput. Phys.* **175** (2002) 525–558.

[38] D. Kuzmin, M. Möller und S. Turek, Multidimensional FEM-FCT schemes for arbitrary time-stepping. Technical report No. **215**, University of Dortmund, 2002, erscheint in: *J. Comput. Phys.*

[39] D. Kuzmin, M. Möller und S. Turek, Implicit flux-corrected transport algorithm for finite element simulation of the compressible Euler equations. Technical report No. **221**, University of Dortmund, 2002, erscheint in: Proceedings of the Conference *Finite Element Methods: 50 Years of Conjugate Gradients*, University of Jyväskylä, Finland, June 11-12, 2002

[40] D. Kuzmin, M. Möller und S. Turek, High-resolution FEM-FCT schemes for multidimensional conservation laws. Technical report No. **231**, University of Dortmund, 2003, eingereicht bei: *Comput. Methods Appl. Mech. Engrg.*

[41] D. Kuzmin und S. Turek, High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter. Technical report No. **229**, University of Dortmund, 2003, eingereicht bei: *J. Comput. Phys.*

[42] B. Laney und D. A. Caughey, Extremum control II: Semi-discrete approximations to conservation laws. *AIAA paper 91-0632*, AIAA 29th Aerospace Sciences Meeting, Reno, Nevada (1991)

[43] A. Lapin, University of Stuttgart, *persönliche Mitteilung*

[44] P. D. Lax und B. Wendroff, Systems of conservation laws, *Comm. Pure Appl. Math* **13** (1960) 217–237.

[45] P. D. Lax, Systems of Conservation Laws and Mathematical Theory of Shock Waves, *SIAM Publ.* (1973)

[46] R. J. LeVeque, *Numerical Methods for Conservation Laws*. Birkhäuser, 1992.

[47] R. J. LeVeque, Simplified multi-dimensional flux limiting methods. *Numerical Methods for Fluid Dynamics* **IV** (1993) 175–190.

[48] R. J. LeVeque, High-resolution Conservative Algorithms for Advection in Incompressible Flow, *SIAM J. Num. Anal.* (1995)

[49] R. Löhner, *Applied CFD Techniques*. Wiley, 2001.

[50] R. Löhner, Web Site: `http://www.science.gmu.edu/~rlohner/`

[51] R. Löhner, K. Morgan, J. Peraire und M. Vahdati, Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. *Int. J. Numer. Meth. Fluids* **7** (1987) 1093–1109.

[52] R. Löhner, K. Morgan, M. Vahdati, J. P. Boris und D. L. Book, FEM-FCT: combining unstructured grids with high resolution. *Commun. Appl. Numer. Methods* **4** (1988) 717–729.

[53] P. R. M. Lyra, *Unstructured Grid Adaptive Algorithms for Fluid Dynamics and Heat Conduction*. PhD thesis, University of Wales, Swansea, 1994.

[54] P. R. M. Lyra, K. Morgan, J. Peraire und J. Peiro, TVD algorithms for the solution of the compressible Euler equations on unstructured meshes. *Int. J. Numer. Meth. Fluids* **19** (1994) 827–847.

[55] K. Morgan und J. Peraire, Unstructured grid finite element methods for fluid mechanics. *Reports on Progress in Physics*, **61** (1998), no. 6, 569–638.

[56] NACA Report 1135, Ames Research Staff, *Equations, Tables and Charts for Compressible Flow*, 1953.

[57] NPARC Alliance, Computational Fluid Dynamics (CFD) Verification and Validation, Web Site: `http://www.grc.nasa.gov/WWW/wind/valid/`

[58] A. K. Parrott und M. A. Christie, FCT applied to the 2-D finite element solution of tracer transport by single phase flow in a porous medium. *Proc. ICFD Conf. on Numerical Methods in Fluid Dynamics*, Oxford University Press, 1986, 609–619.

[59] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*. McGraw-Hill, New York, 1980.

[60] J. Peraire, M. Vahdati, J. Peiro und K. Morgan, The construction and behaviour of some unstructured grid algorithms for compressible flows. *Numerical Methods for Fluid Dynamics* **IV**, Oxford University Press, 1993, 221–239.

[61] S. Osher und F. Solomon, Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp.* **38** (1982) 339–374.

[62] R. Rannacher, *Numerische Methoden für Partialle Differentialgleichungen*, Vorlesungsskriptum, University of Heidelberg, 2000.

[63] R. Rannacher und S. Turek, A simple nonconforming quadrilateral Stokes element. *Numer. Meth. PDEs* **8** (1992), no. 2, 97–111.

[64] P. L. Roe, Approximate Riemann solvers, parameter vectors and difference schemes. *J. Comput. Phys.* **43** (1981) 357–372.

[65] P. L. Roe und J. Pike, Efficient Construction and Utilisation of Approximate Riemann Solutions. *Comput. Methods Appl. Sci. and Engrg.* (1984) 499–518

[66] C. Schär und P. K. Smolarkiewicz, A Synchronous and Iterative Flux-Correction Formalism for Coupled Transport Equations. *J. Comput. Phys.* **128** (1996) 101–120.

[67] V. Selmin, Finite element solution of hyperbolic equations. I. One-dimensional case. *INRIA Research Report* **655** (1987).

[68] T. W. H. Sheu und C. C. Fang, High resolution finite-element analysis of shallow water equations in two dimensions. *Comput. Methods Appl. Mech. Engrg.* **190** (2001) 2581–2601.

[69] P. K. Smolarkiewicz und W. W. Grabowski, The multidimensional positive definite advection transport algorithm: nonoscillatory option. *J. Comput. Phys.* **86** (1990) 355–375.

[70] G. Sod, A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *J. Comput. Phys.* **27** (1978) 1–31.

[71] A. Sokolichin, *Mathematische Modellbildung und numerische Simulation von Gas-Flüssigkeits-Blasenströmungen*. Habilitationsschrift, University of Stuttgart, 2002.

[72] P. K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws, *SIAM J. Num. Anal.* **21** (1984) 995–1011.

[73] S. Turek, *Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approach*, LNCSE 6, Springer, 1999.

[74] E. Tadmor, Convenient total variation diminishing conditions for nonlinear difference schemes. *SIAM J. Num. Anal.* **25** (1988) 1002–1014

[75] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics, A Practical Introduction*, Springer, 1999.

[76] B. van Leer, Towards the ultimate conservative difference scheme V. A second order sequel to Godunov's method. *J. Comput. Phys.* **32** (1979) 101–136.

[77] P. Wesseling, *Principles of Computational Fluid Dynamics*. Springer, 2001.

[78] H. C. Yee, Numerical approximation of boundary conditions with applications to inviscid gas dynamics. *NASA report* TM-81265, 1981.

[79] S. T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* **31** (1979) 335–362.