Sigrun Ortleb

Ein diskontinuierliches Galerkin-Verfahren hoher Ordnung auf Dreiecksgittern mit modaler Filterung zur Lösung hyperbolischer Erhaltungsgleichungen

kassel university press

Sigrun Ortleb

Ein diskontinuierliches Galerkin-Verfahren hoher Ordnung auf Dreiecksgittern mit modaler Filterung zur Lösung hyperbolischer Erhaltungsgleichungen

kassel university press

Die vorliegende Arbeit wurde vom Fachbereich Mathematik und Naturwissenschaften der Universität Kassel als Dissertation zur Erlangung des akademischen Grades einer Doktorin der Naturwissenschaften (Dr. rer. nat.) angenommen.

Erster Gutachter: Prof. Dr. Andreas Meister
Zweiter Gutachter: Prof. Dr. Thomas Sonar

Tag der mündlichen Prüfung 4. Oktober 2011

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar

Zugl.: Kassel, Univ., Diss. 2011
ISBN print: 978-3-86219-218-2
ISBN online: 978-3-86219-219-9
URN: http://nbn-resolving.de/urn:nbn:de:0002-32193

Printed in Germany

Vorwort

Die vorliegende Arbeit ist während meiner Beschäftigung als wissenschaftliche Bedienstete des Fachbereichs 10 Mathematik und Naturwissenschaften der Universität Kassel entstanden. Viele Menschen haben am Gelingen dieser Arbeit Anteil genommen und mich unterstützt - hierfür möchte ich mich herzlich bedanken.

Ganz besonderer Dank gebührt meinem Doktorvater, Professor Dr. Andreas Meister, für die hervorragende und warmherzige Betreuung, professionelle Beratung und vorbildliche Unterstützung. Seine vielen hilfreichen Anregungen haben mir zahlreiche wichtige Impulse für meine Arbeit gegeben, mir aber auch in hohem Maße Eigenständigkeit erlaubt.

Professor Dr. Thomas Sonar danke ich für die Übernahme des Koreferats sowie die vielen anregenden Gespräche und die stete Bereitschaft, an seinem immensen fachlichen und mathematikgeschichtlichen Wissen teilhaben zu lassen.

Der Ausgangspunkt meiner Forschung, die thematische Eingrenzung meiner Doktorarbeit auf den Bereich der diskontinuierlichen Galerkin-Verfahren für hyperbolische Erhaltungsgleichungen, ergab sich durch ein im Jahr 2006 angeschobenes gemeinsames Projekt von Professor Dr. Meister und Professor Dr. Sonar. Für die Bereitstellung dieses Themas möchte ich mich deshalb bei beiden herzlich bedanken. Gleichermaßen bedanke ich mich hiermit bei der Deutschen Forschungsgemeinschaft für die finanzielle Unterstützung dieser Forschungskooperation zwischen Kassel und Braunschweig, insbesondere im Rahmen des DFG-Projekts ME 1889/3-1, in dem ich seit August 2010 angestellt bin.

In entscheidendem Maße hat sich ebenso das positive, offene Umfeld der Mitarbeiter des Instituts für Mathematik der Universität Kassel auf die Qualität dieser Arbeit ausgewirkt, für die herzliche und konstruktive Atmosphäre möchte ich mich insbesondere bei Dipl.-Math. Bettina Messerschmidt, Dr. Philipp Birken, Dipl.-Math. Stefan Kopecz und Dr. Philipp Zardo bedanken.

Von ganzem Herzen möchte ich mich bei meiner Mutter, ihrem Lebensgefährten Peter und meinem Vater dafür bedanken, dass sie mich beständig dazu ermutigt haben, wissenschaftlich zu arbeiten.

Abschließend gilt mein innigster Dank meinem Lebensgefährten Christian, für seine fortwährende Ermutigung, die Doktorarbeit abzuschließen, für sein großes Interesse an meiner Forschungstätigkeit und für unsere gemeinsame Tochter Elenore Marie.

Inhaltsverzeichnis

1 Einleitung

Die vorliegende Arbeit beschäftigt sich mit der Anwendung diskontinuierlicher Galerkin-Verfahren zur numerischen Lösung hyperbolischer Erhaltungsgleichungen. Diese Klasse zeitabhängiger partieller Differentialgleichungen stellt besondere Anforderungen an ein zur approximativen Berechnung ihrer Lösungen ausgewähltes numerisches Verfahren, da die Theorie hyperbolischer Erhaltungsgleichungen abweichend vom klassischen Lösungsbegriff unstetige Lösungen zulässt. Hinzu kommt, dass die Eindeutigkeit dieser Lösungen im Allgemeinen nur durch das Aufstellen zusätzlicher Entropiebedingungen gewährleistet werden kann. Während numerische Verfahren erster Ordnung dazu neigen, Sprungunstetigkeiten der exakten Lösung zu verschmieren, reagieren Verfahren höherer Ordnung ohne an die spezielle Situation hyperbolischer Erhaltungsgleichungen angepasste zusätzliche Dämpfungsmechanismen mit der Ausbildung unphysikalischer Oszillationen in unmittelbarer Nähe der Unstetigkeitsstellen. Da diese Oszillationen in Kombination mit der Nichtlinearität der Gleichungen die Stabilität des numerischen Verfahrens gefährden können, werden in der Literatur verschiedene zusätzliche Dämpfungsmechanismen vorgeschlagen, die auf die konkrete gewählte Verfahrensklasse ausgerichtet sind.

Bei den in dieser Arbeit betrachteten diskontinuierlichen Galerkin-Verfahren in zwei Raumdimensionen wird die exakte Lösung im Raum durch stückweise polynomiale, im Allgemeinen unstetige Funktionen approximiert. Das Augenmerk liegt hierbei auf der Verwendung hoher Polynomgrade, so dass die Einführung zusätzlicher künstlicher Dämpfung notwendig ist. Wie bei Finite-Elemente-Methoden wird im Fall der diskontinuierlichen Galerkin-Verfahren zunächst eine Diskretisierung des Rechengebiets in Form einer Zerlegung in eine endliche Anzahl von Teilgebieten vorgenommen, die sich auf ein Referenzelement transformieren lassen. Besonders hohe Flexibilität, die der Diskretisierung komplex strukturierter Gebiete entgegenkommt, lässt sich hierbei mit Dreiecksgittern erreichen, die in dieser Arbeit ausschließlich verwendet werden sollen. Diskontinuierliche Galerkin-Verfahren zur Diskretisierung hyperbolischer Erhaltungsgleichungen wurden erst in jüngerer Zeit entwickelt und lassen sich als eine Erweiterung der relativ ausgefeilten und sehr verbreiteten Klasse der Finite-Volumen-Verfahren auffassen, die die zeitliche Entwicklung von Zellmittelwerten approximieren und im Fall höherer Ordnung punktweise Zustände innerhalb einer Zelle beispielsweise durch gewichtete Interpolationen benachbarter Zellmittelwerte berechnen. Aufgrund ihrer geschichtlichen Entwicklung in enger Verwandtschaft zu Finite-Volumen-Verfahren erben viele Varianten diskontinuierlicher Galerkin-Verfahren zur Lösung hyperbolischer Erhaltungsgleichungen die Dämpfungsmechanismen von Finite-Volumen-Verfahren – insbesondere Steigungslimiter und wesentlich nicht-oszillierende Rekonstruktionen – bei denen die Zustände innerhalb benachbarter Elemente zur Oszillationsdämpfung einbezogen werden. Im Gegensatz zur derartigen impliziten Einführung numerischer Viskosität besteht eine andere Art der Einführung zusätzlicher numerischer Dämpfung darin, explizite Diffusionsterme zur Erhaltungsgleichung hinzuzufügen. Die Diskretisierung dieser zusätzlichen Terme erfordert dann die Modifikation des Basisverfahrens sowie das Erfüllen schärferer Restriktionen an die Zeitschrittweite im Fall der Verwendung expliziter Zeitintegrationsverfahren.

In der vorliegenden Arbeit wird die Konstruktion und Untersuchung eines Dämpfungsmechanismus vorgenommen, der im Kontext der Lösung hyperbolischer Erhaltungsgleichungen mittels diskontinuierlicher Galerkin-Verfahren auf Triangulierungen neuartig ist. Die Dämpfungsstrategie sieht ebenfalls die Einführung expliziter Viskosität vor, in spe-

zieller als spektrale Viskosität bezeichneter Form. Allerdings wird hier anstelle des üblichen Laplace-Operators der zu den Ansatzfunktionen des diskontinuierlichen Galerkin-Verfahrens gehörige Sturm-Liouville-Operator eingesetzt, dessen Eigenfunktionen die Polynome der im Verfahren verwendeten Polynombasis sind. Der Vorteil einer derartigen Formulierung besteht in der Möglichkeit, den Dämpfungsterm in Form eines modalen Filters zu implementieren, welcher ausschließlich auf den Koeffizienten der stückweise polynomialen Entwicklung operiert. Dadurch benötigt die Dämpfungsstrategie vergleichsweise geringen Rechenaufwand und erfordert keine weitere Verringerung des Zeitschritts. Die Idee einer derartigen Dämpfungsstrategie hat ihren Ursprung in der Klasse der Spektralmethoden, deren Näherungslösungen in Ansatzräumen liegen, die von global auf dem gesamten Rechengebiet definierten glatten Basisfunktionen aufgespannt werden. Dementsprechend werden im Rahmen dieser Arbeit diejenigen Ideen und Herangehensweisen der Spektralmethoden, die innerhalb von diskontinuierlichen Galerkin-Verfahren mit hohem Polynomgrad Verwendung finden, zusammengestellt, übertragen – insbesondere auf das Vorliegen von Dreiecksgittern – und erweitert. Insbesondere gehen wir auf die Dämpfungseigenschaften des zur Viskositätsformulierung betrachteten Sturm-Liouville-Operators ein. In diesem Zusammenhang ergibt sich eine in der Literatur bisher nicht aufgeführte, auch für sich genommen interessante, Abschätzung für die zugehörige gewichtete Norm. Desweiteren erhalten wir ein neues Approximationsresultat für modal gefilterte Entwicklungen hinreichend glatter Funktionen in Proriol-Koornwinder-Dubiner-Polynome.

Ein Aspekt, der bei der Nutzung modaler Filter innerhalb von Spektralmethoden im Allgemeinen ebenfalls außer Acht gelassen wird, ist die adaptive Steuerung der durch die modale Filterung gegebenen numerischen Viskosität auf dem gegebenen Rechengebiet. So ist der Ordnungsverlust, der bei Verwendung globaler modaler Filterung unter Gitterverfeinerung auftritt, bisher nicht thematisiert worden, da Spektralmethoden in ihrer ursprünglichen Form keine Gebietszerlegung vorsehen. Wir werden diesbezüglich zwei verschiedene Indikatoren zur elementweisen Steuerung der modalen Filterung betrachten, die bereits im Kontext von diskontinuierlichen Galerkin-Verfahren zur Steuerung expliziter Viskosität verwendet wurden. Allerdings wurde die Nutzbarkeit dieser Indikatoren zur Steuerung der hier verwendeten modalen Filterung bisher noch nicht untersucht und ist deshalb ebenfalls Gegenstand dieser Arbeit.

Desweiteren ist bereits im Kontext von Spektralmethoden bekannt, dass trotz der Nutzung modaler Filter Oszillationen in der Nähe der Unstetigkeitsstellen der exakten Entropielösung verbleiben. Die Verwendung modaler Filter geht bei dieser Verfahrensklasse daher einher mit einer geeigneten Technik zur Nachbearbeitung der Näherungslösung zur End- beziehungsweise Visualisierungszeitpunkten. Dementsprechend soll nicht nur das Konzept modaler Filterung, sondern das Gesamtpaket bestehend aus Stabilisierungstechnik und Nachbearbeitungsprozedur für diskontinuierliche Galerkin-Verfahren nutzbar gemacht werden. Während zur Nachbearbeitung der Näherungslösung im eindimensionalen Fall häufig Reprojektionstechniken eingesetzt werden, basierend auf der Projektion der Näherungslösung in einen neuen Ansatzraum, ist die Übertragung einer derartigen Technik auf den zweidimensionalen Fall nicht trivial. In dieser Arbeit wird daher auf eine andere bereits im Bereich der Spektralmethoden verwendete Nachbearbeitungsmethode zurückgegriffen, deren Einführung in diskontinuierliche Galerkin-Verfahren bisher noch nicht in Betracht gezogen wurde. Hierbei handelt es sich um den sogenannten digitalen TV-Filter, der ursprünglich als Bildverarbeitungsmethode entworfen wurde. Im neuen Kontext der Anwendung auf Näherungslösungen diskontinuierlicher Galerkin-Verfahren

auf Dreiecksgittern wird zudem eine Modifikation der Filtervorschrift entwickelt, die erstmals die Anwendung des Filters auf DTV-Graphen ermöglicht, die eine an die Struktur der Dreiecksgitter angepasste Form besitzen. Ein Erkenntnisgewinn ergibt sich zudem durch eine Betrachtung der Fehlerverläufe fern von Stößen, die sich unter Gitterverfeinerung für die DTV-gefilterten Näherungslösungen im Fall einer hyperbolischen Erhaltungsgleichung mit unstetiger Lösung ergeben. Ein Ordnungsverlust bei globaler Anwendung des DTV-Filters zeigt hierbei, dass eine adaptive Form der Filterung notwendig ist, die in dieser Arbeit ebenfalls entwickelt und untersucht werden soll.

Die vorliegende Arbeit untergliedert sich wie folgt. Das anschließende zweite Kapitel enthält eine knappe Zusammenstellung der grundlegenden Definitionen und Eigenschaften hyperbolischer Erhaltungsgleichungen. Desweiteren werden dort diejenigen Gleichungen vorgestellt, die zur Definition von Testfällen für das entwickelte numerische Verfahren dienen. Kapitel 3 beschäftigt sich mit den Eigenschaften der zur Konstruktion des diskontinuierlichen Galerkin-Verfahrens verwendeten orthogonalen Polynombasis auf einem Dreiecksgebiet. Insbesondere wird auf das zur Polynombasis gehörige Sturm-Liouville-Problem eingegangen, welches zur Herleitung des modalen Filters verwendet wird. In Kapitel 4 wird die Konstruktion der diskontinuierlichen Galerkin-Methode einschließlich der verwendeten Zeitintegrationsverfahren beschrieben. Zudem gehen wir auf die Stabilitäts- und Konvergenzeigenschaften dieses Basisverfahrens ein. In Kapitel 5 werden modale Filter sowie die Formulierung spektraler Viskosität erklärt. Das neue Approximationsresultat für modal gefilterte Proriol-Koornwinder-Dubiner-Entwicklungen hinreichend glatter Funktionen wird in diesem Zusammenhang nachgewiesen. Ein modaler Filter für diskontinuierliche Galerkin-Verfahren wird anschließend aus einer Formulierung spektraler Viskosität auf dem Dreieck hergeleitet und es wird die adaptive Anwendung des Filters erläutert. Kapitel 6 ist den zur Nachbearbeitung verwendeten Varianten des digitalen TV-Filters gewidmet. Schließlich werden in Kapitel 7 die mit der hier konstruierten Methode erzielten numerische Ergebnisse präsentiert. Kapitel 8 liefert abschließend eine Zusammenfassung der Arbeit und einen Ausblick auf mögliche zukünftige Weiterentwicklungen.

2 Hyperbolische Erhaltungsgleichungen

Hyperbolische Erhaltungsgleichungen sind zeitabhängige Systeme partieller Differentialgleichungen erster Ordnung, die Wellenausbreitungen und Transportprozesse beschreiben. Sie modellieren die zeitliche Änderung von Dichten physikalischer Erhaltungsgrößen, wie zum Beispiel Masse, Impuls, Energie oder elektrische Ladung. Ein wichtiges Beispiel sind die Euler-Gleichungen der Gasdynamik, die unter anderem zur Simulation der Umströmung von Tragflächen eines Flugzeugs oder von Fahrzeugen unter Vernachlässigung von Reibungseffekten genutzt werden. Weitere Anwendungen ergeben sich aus der Beschreibung von Flachwasserwellen, Verkehrsflüssen, Klimamodellen oder Mehrphasenströmungen. In diesem Kapitel sollen zunächst in knapper Form die diese Gleichungen betreffenden theoretischen Grundlagen präsentiert werden, die in ausführlicher Darstellung unter anderem in den Lehrbüchern von Godlewski und Raviart [34], LeVeque [63] und Warnecke [94] zu finden sind. Anschließend werden die in dieser Arbeit zur Untersuchung des entwickelten numerischen Verfahrens verwendeten skalaren Testprobleme vorgestellt, während im letzten Abschnitt des Kapitels die Euler-Gleichungen behandelt werden.

2.1 Grundlagen

Grundsätzlich betrachtet man die Änderung einer *Erhaltungs-* oder *Zustandsgröße*

$$\mathbf{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \to S \subseteq \mathbb{R}^m$$

in Abhängigkeit von den Variablen $\mathbf{x} \in \mathbb{R}^d$ und $t \in \mathbb{R}_0^+$, die Raum und Zeit darstellen. Die Einschränkung des Wertebereichs auf den *Zustandsraum* S trägt unter anderem der Tatsache Rechnung, dass bestimmte physikalische Grössen keine negativen Werte annehmen können. Im eigentlichen Sinn sind die Variablen u_j Dichten physikalischer Erhaltungsgrössen, wobei sich der Begriff der Erhaltung darauf bezieht, dass die Gesamtmasse $\int_{\mathbb{R}^d} u_j(\mathbf{x}, t) d\mathbf{x}$ in der Zeit konstant bleibt.
Die in dieser Arbeit betrachteten *Erhaltungsgleichungen*, die auf einem solchen Erhaltungsprinzip beruhen, haben die Form

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x}, t) + \nabla_{\mathbf{x}} \cdot \mathcal{F}(\mathbf{u}(\mathbf{x}, t)) \;=\; \mathbf{0}, \qquad (\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R}^+, \tag{2.1}$$

mit der im Allgemeinen nichtlinearen *Flussfunktion* $\mathcal{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_d)^T,\ \mathbf{f}_l \in C^1(S, \mathbb{R}^m)$, und der Kurznotation $\nabla_{\mathbf{x}} \cdot \mathcal{F}(\mathbf{u}) = \sum_{l=1}^{d} \frac{\partial}{\partial x_l} \mathbf{f}_l(u)$. Bei Hinzunahme von Anfangswerten

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^d, \tag{2.2}$$

mit einer gegebenen Funktion $\mathbf{u}_0 : \mathbb{R}^d \to S$, wird das zeitliche Verhalten von $\mathbf{u}$ ausgehend von einem Anfangszustand beschrieben. Im Fall einer skalaren Zustandsgröße, d.h. für $m = 1$, spricht man von skalaren Erhaltungsgleichungen, für $m > 1$ von Systemen.
Die Erhaltung der Gesamtmassen, durch die die Bezeichnung von (2.1) als Erhaltungsgleichung motiviert ist, ergibt sich – unter der Annahme, dass $\mathbf{u}$ einen kompakten Träger besitzt – durch Integration der Gleichung über ein Lipschitz-Gebiet $\Omega^d \subset \mathbb{R}^d$ mit $\mathbf{u}|_{\mathbb{R}^d \setminus \Omega^d \times \mathbb{R}_0^+} \equiv \mathbf{0}$ und die Verwendung des Gaußschen Integralsatzes. Es gilt zunächst

$$\frac{d}{dt}\int_{\mathbb{R}^d} \mathbf{u}(\mathbf{x}, t)\, d\mathbf{x} = \frac{d}{dt}\int_{\Omega^d} \mathbf{u}(\mathbf{x}, t)\, d\mathbf{x} = -\int_{\Omega^d} \nabla_{\mathbf{x}} \cdot \mathcal{F}(\mathbf{u})\, d\mathbf{x} = -\int_{\partial\Omega^d} \mathcal{F}(\mathbf{u}) \cdot \mathbf{n}\, d\sigma,$$

mit $\mathcal{F}(\mathbf{u}) \cdot \mathbf{n} = \sum_{l=1}^{d} \mathbf{f}_l(\mathbf{u}) n_l$, wobei mit $\mathbf{n}$ das Feld äußerer Normalenvektoren an $\partial\Omega^d$ bezeichnet ist. Da $\mathcal{F}(\mathbf{u})$ auf $\partial\Omega^d$ den konstanten Wert $\mathcal{F}(\mathbf{0})$ annimmt, erhält man

$$\frac{d}{dt} \int_{\mathbb{R}^d} \mathbf{u}(\mathbf{x}, t)\, d\mathbf{x} = \mathbf{0},$$

so dass die Gesamtmassen in der Zeit konstant bleiben.
Beschränkt man sich nicht auf hyperbolische Erhaltungsgleichungen, so kann $\mathcal{F}$ auch von $\nabla_{\mathbf{x}}\mathbf{u}$ abhängig sein, Gleichungen in dieser allgemeineren Form sollen hier jedoch nicht betrachtet werden. Zur Definition der Hyperbolizität betrachtet man (2.1) in *quasilinearer Form*

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x}, t) + \sum_{l=1}^{d} \mathbf{A}_l(\mathbf{u}) \frac{\partial}{\partial x_l}\mathbf{u}(\mathbf{x}, t) \;=\; \mathbf{0}, \tag{2.3}$$

wobei $\mathbf{A}_l$ die zu $\mathbf{f}_l$ gehörige Jacobi-Matrix ist. Die Erhaltungsgleichung (2.1) wird nun als *hyperbolisch* bezeichnet, wenn für alle $\mathbf{u} \in S$ und alle Vektoren $\boldsymbol{\nu} \in \mathbb{R}^d$ die Linearkombination $\sum_{l=1}^{d} \nu_l \mathbf{A}_l(\mathbf{u})$ reell diagonalisierbar ist. Aus dieser Definition ergibt sich, dass eine skalare Gleichung der Form

$$\frac{\partial}{\partial t}u(\mathbf{x}, t) + \nabla_{\mathbf{x}} \cdot \mathcal{F}(u(\mathbf{x}, t)) \;=\; 0, \tag{2.4}$$

mit stetig differenzierbarer Flussfunktion $\mathcal{F} = (f_1, \ldots, f_d)^T$ und der quasilinearen Form

$$\frac{\partial}{\partial t}u(\mathbf{x}, t) + \mathbf{a}(u) \cdot \nabla_{\mathbf{x}}(u(\mathbf{x}, t)) \;=\; 0, \qquad \mathbf{a}(u) = (f_1'(u), \ldots, f_d'(u))^T, \tag{2.5}$$

immer hyperbolisch ist.

Hyperbolische Erhaltungsgleichungen zeichnen sich dadurch aus, dass die Existenz einer klassischen Lösung, d.h. einer Funktion $\mathbf{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \to S$ mit $\mathbf{u} \in C^1(\mathbb{R}^d \times \mathbb{R}_0^+)^m$, die die Differentialgleichung (2.1) und die Anfangsbedingung (2.2) punktweise erfüllt, selbst unter der Voraussetzung glatter Anfangsdaten $\mathbf{u}_0$ möglicherweise nur für ein endliches Zeitintervall gewährleistet ist. Das Phänomen des Zusammenbruchs der klassischen Lösung tritt bereits bei skalaren Erhaltungsgleichungen in einer Raumdimension auf, wie nachfolgend gezeigt werden soll. Gehen wir zunächst von der globalen Existenz einer klassischen Lösung einer skalaren Erhaltungsgleichung (2.4) aus, so lässt sich diese durch das Verfolgen von Charakteristiken der quasilinearen Gleichung (2.5) konstruieren.

Definition 2.1 *Sei $u : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}$ klassische Lösung der quasilinearen partiellen Differentialgleichung* (2.5)*. Eine im Raum-Zeit-Koordinatensystem durch $\{(\boldsymbol{\gamma}(t), t) \mid t \in \mathbb{R}_0^+\}$ gegebene Kurve, die die Differentialgleichung*

$$\frac{d}{dt}\boldsymbol{\gamma}(t) = \mathbf{a}(u(\boldsymbol{\gamma}(t), t))$$

erfüllt, heißt charakteristische Grundkurve *von* (2.5)*. Auf einer Lösungsfläche der Gleichung liegende Kurven $\{(\boldsymbol{\gamma}(t), t, u(\boldsymbol{\gamma}(t), t) \mid t \in \mathbb{R}_0^+\}$, die die obige Bedingung erfüllen, heißen* Charakteristiken.

Besitzt eine skalare quasilineare Gleichung (2.5), die nicht notwendigerweise eine Erhaltunggleichung sein muss, eine klassische Lösung, so kann man diese durch das Lösen des obigen Systems gewöhnlicher Differentialgleichungen und den Aufbau einer Lösungsfläche aus den Charakteristiken konstruieren. Ist die gegebenen skalare Gleichung eine Erhaltungsgleichung, so lassen sich die charakteristischen Grundkurven besonders einfach bestimmen.

Lemma 2.2 *Sei $u : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}$ eine klassische Lösung von (2.4). Dann ist u konstant entlang der charakteristischen Grundkurven, die in diesem Fall Geraden sind.*

Beweis: Sei $\boldsymbol{\gamma}$ eine charakteristische Grundkurve von (2.4). Dann gilt

$$\begin{aligned} \frac{d}{dt}u(\boldsymbol{\gamma}(t),t) &= \frac{\partial}{\partial t}u(\boldsymbol{\gamma}(t),t) + \underbrace{\frac{d}{dt}\boldsymbol{\gamma}(t)}_{=\mathbf{a}(u)} \cdot \nabla_{\mathbf{x}} u(\boldsymbol{\gamma}(t),t) \\ &= \frac{\partial}{\partial t}u(\boldsymbol{\gamma}(t),t) + \nabla_{\mathbf{x}}\mathcal{F}(u(\boldsymbol{\gamma}(t),t)) = 0. \end{aligned}$$

Somit ist u konstant entlang $t \mapsto (\boldsymbol{\gamma}(t), t)$ und man erhält $\frac{d}{dt}\boldsymbol{\gamma}(t) = \mathbf{a}(u(\boldsymbol{\gamma}(0),0))$, so dass $\boldsymbol{\gamma}$ gegeben ist durch

$$\boldsymbol{\gamma}(t) = \mathbf{a}(u_0(\mathbf{x}_0), t) + \mathbf{x}_0 \tag{2.6}$$

mit $\mathbf{x}_0 = \boldsymbol{\gamma}(0)$. □

Ist die Flussfunktion $\mathcal{F}$ einer skalaren Erhaltungsgleichung nichtlinear, so ist $\mathbf{a}(u) = \mathcal{F}'(u)$ nicht konstant, so dass die Steigung der Geraden von den Anfangsdaten u_0 abhängig ist. Schneiden sich charakteristische Grundkurven mit verschiedener Steigung, die dann verschiedene Werte von u_0 transportieren, so ist die klassische Lösung spätestens im Schnittpunkt der Geraden nicht definiert. Das Standardbeispiel für das Entstehen eines solchen Verdichtungsstoßes ist die *Burgers-Gleichung* ohne dissipative Terme, die durch

$$\frac{\partial}{\partial t}u + u \cdot \frac{\partial}{\partial x}u = 0. \tag{2.7}$$

gegeben ist.

Beispiel 2.3 Man betrachte die Burgers-Gleichung (2.7) mit den Anfangswerten

$$u_0(x) = \begin{cases} 1, & x < 0, \\ \cos(\pi x), & 0 \le x \le 1, \\ -1, & x > 1. \end{cases} \tag{2.8}$$

Die zugehörigen charakteristischen Grundkurven haben nach (2.6) die Form $\gamma(t) = u_0(x_0) \cdot t + x_0$ und schneiden sich auf der Geraden $x = \frac{1}{2}$ zu einem Zeitpunkt t^*, wie in Abbildung 2.1 dargestellt.
Für $t > t^*$ existiert somit keine klassische Lösung des gegebenen Cauchy-Problems. Im Allgemeinen lässt sich zeigen, dass klassische Lösungen der Burgers-Gleichung (2.7) mit Anfangsdaten $u_0 \in C^1(\mathbb{R})$ genau dann zusammenbrechen, wenn es eine Stelle $x \in \mathbb{R}$ mit $u_0'(x) < 0$ gibt, und dass der Zeitpunkt, zu dem die klassische Lösung aufhört zu existieren, durch $t^* = -1/\min_{x\in\mathbb{R}} u_0'(x)$ gegeben ist. In diesem Beispiel gilt daher $t^* = 1/\pi$.

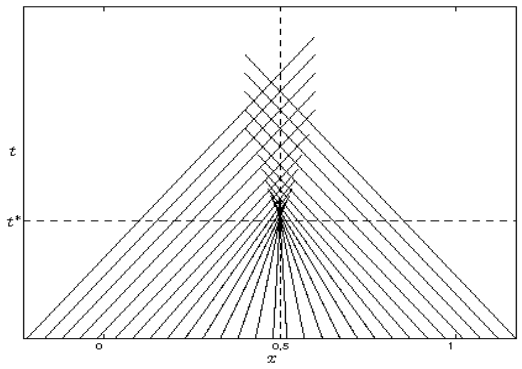


Abb. 2.1: Charakteristische Grundkurven der Burgers-Gleichung.

Mit der Nichtexistenz einer globalen klassischen Lösung begnügt man sich allerdings nicht. Die physikalischen Vorgänge, die durch hyperbolische Erhaltungsgleichungen unter Vernachlässigung dissipativer Effekte modelliert werden, können tatsächlich Zustandsänderungen auf einer sehr kleinen Skala im Vergleich zum Gesamtsystem aufweisen, so dass sich mit dem bloßen Auge kein Unterschied zu einem unstetigen Sprung erkennen lässt. Es ist somit sinnvoll, unstetige Funktionen als Lösungen einer Gleichung der Form (2.1) zuzulassen und den Lösungsbegriff zu erweitern. Dazu wird die Gleichung (2.1) mit Testfunktionen $\boldsymbol{\varphi} \in C_0^1(\mathbb{R}^d \times \mathbb{R}_0^+)^m$ multipliziert und über Raum und Zeit integriert. Ist $\mathbf{u}$ eine klassische Lösung von (2.1), die der Anfangsbedingung (2.2) genügt, so gilt unter Verwendung der Greenschen Integralformel

$$\begin{aligned} 0 &= \int_{\mathbb{R}^d\times\mathbb{R}^+} \left[\frac{\partial}{\partial t}\mathbf{u} + \sum_{l=1}^{d} \frac{\partial}{\partial x_l}\mathbf{f}_l(\mathbf{u})\right] \cdot \boldsymbol{\varphi}\, d\mathbf{x}dt \\ &= -\int_{\mathbb{R}^d\times\mathbb{R}^+} \left[\mathbf{u}\cdot\frac{\partial}{\partial t}\boldsymbol{\varphi} + \sum_{l=1}^{d} \mathbf{f}_l(\mathbf{u})\cdot\frac{\partial}{\partial x_l}\boldsymbol{\varphi}\right] d\mathbf{x}dt + \int_{\partial\Omega^{d+1}} \left(\mathbf{u}\cdot n_t + \sum_{l=1}^{d}\mathbf{f}_l(\mathbf{u})\cdot n_l\right)\cdot\boldsymbol{\varphi}\, d\sigma, \end{aligned}$$

wobei $\Omega^{d+1} \subset \mathbb{R}^d \times \mathbb{R}^+$ ein Lipschitz-Gebiet mit dem äußeren Normalenvektorfeld $\mathbf{n} = (n_1, \ldots, n_d, n_t)$ ist, dessen Abschluss $\overline{\Omega^{d+1}}$ den Träger von $\boldsymbol{\varphi}$ enthält. Da $\boldsymbol{\varphi}$ für $t > 0$ auf $\partial\Omega^{d+1}$ verschwindet und der äußere Normalenvektor in einem Punkt $(\mathbf{x}, 0) \in \partial\Omega^{d+1}$ durch $\mathbf{n} = (0, \ldots, 0, -1)$ gegeben ist, gilt mit Einbeziehen der Anfangswerte die Gleichung

$$\int_{\mathbb{R}^d\times\mathbb{R}^+} \left[\mathbf{u}\cdot\frac{\partial}{\partial t}\boldsymbol{\varphi} + \sum_{l=1}^{d}\mathbf{f}_l(\mathbf{u})\cdot\frac{\partial}{\partial x_l}\boldsymbol{\varphi}\right] d\mathbf{x}dt + \int_{\mathbb{R}^d}\mathbf{u}_0(\mathbf{x})\cdot\boldsymbol{\varphi}(\mathbf{x},0)d\mathbf{x} = 0. \tag{2.9}$$

Die Integrale in (2.9) existieren ebenso, falls nur die Voraussetzungen $\mathbf{u}_0 \in L_{loc}^\infty(\mathbb{R}^d)^m$ sowie $\mathbf{u} \in L_{loc}^\infty(\mathbb{R}^d \times \mathbb{R}^+)^m$ erfüllt sind, so dass der Begriff der schwachen Lösung wie folgt erklärt ist.

Definition 2.4 *Sei $\mathbf{u}_0 \in L_{loc}^\infty(\mathbb{R}^d)^m$. Eine Funktion $\mathbf{u} \in L_{loc}^\infty(\mathbb{R}^d \times \mathbb{R}_0^+)^m$ heißt* schwache Lösung *des Cauchy-Problems* (2.1), (2.2)*, falls $\mathbf{u}(\mathbf{x}, t) \in S$ für fast alle $(\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R}^+$ gilt und die Gleichung* (2.9) *für alle Testfunktionen $\boldsymbol{\varphi} \in C_0^1(\mathbb{R}^d \times \mathbb{R}_0^+)^m$ erfüllt ist.*

Ein wichtiges Beispiel schwacher Lösungen sind stückweise glatte Funktionen. Dabei wird eine Funktion $\mathbf{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}^m$ als stückweise glatt bezeichnet, wenn eine endliche

Menge glatter d-dimensionaler Flächen Γ in $\mathbb{R}^d \times \mathbb{R}_0^+$ existiert, außerhalb derer $\mathbf{u}$ stetig differenzierbar ist und auf denen $\mathbf{u}$ einseitige Grenzwerte $\mathbf{u}^\pm$ besitzt. Die Unstetigkeiten stückweise glatter Lösungen einer hyperbolischen Erhaltungsgleichung können nicht beliebig geartet sein, sondern sind der folgenden Restriktion unterworfen.

Satz 2.5 *Sei $\mathbf{u} : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}^m$ eine stückweise glatte Funktion. Dann ist $\mathbf{u}$ genau dann eine schwache Lösung der Erhaltungsgleichung* (2.1)*, wenn gilt*

a) $\mathbf{u}$ ist klassische Lösung von (2.1) *überall dort, wo $\mathbf{u}$ stetig differenzierbar ist,*

b) $\mathbf{u}$ erfüllt die Rankine-Hugoniot-Bedingung

$$(\mathbf{u}^+ - \mathbf{u}^-)n_t + \sum_{l=1}^{d} \left(\mathbf{f}_l(\mathbf{u}^+) - \mathbf{f}_l(\mathbf{u}^-)\right) n_l = \mathbf{0}, \tag{2.10}$$

auf d-dimensionalen Flächen Γ, auf denen $\mathbf{u}$ unstetig ist, wobei $\mathbf{n} = (n_1, \ldots, n_d, n_t)$ ein stetiges Normalenvektorfeld an Γ ist.

Beweis: Siehe [34]. □

Gilt $(n_1, \ldots, n_d) \neq \mathbf{0}$ in einem Punkt von Γ, so lässt sich dort ein Normalenvektor $\mathbf{n} = (\mathbf{v}, -s)$ an Γ wählen, mit einem durch $\|\mathbf{v}\|_2 = 1$ normierten Vektor $\mathbf{v} \in \mathbb{R}^d$ und einer Zahl $s \in \mathbb{R}$, die als Richtung und Geschwindigkeit der Bewegung der Unstetigkeit Γ interpretiert werden können.
Im Fall $d = 1$ lässt sich dies leicht einsehen, wenn die Unstetigkeitskurve parametrisiert wird durch $\Gamma = \{(x(t), t) \mid t \in \mathbb{R}_0^+\}$. Dann erhält man mit $s = x'(t)$ den Normalenvektor $\mathbf{n} = (1, -s)$ und die Rankine-Hugoniot-Bedingung nimmt die Form

$$s[\mathbf{u}] = [\mathbf{f}(\mathbf{u})]$$

an, mit der Notation $\mathbf{f} = \mathbf{f}_1$ und den Sprungklammern $[\mathbf{u}] = \mathbf{u}^+ - \mathbf{u}^-$ und $[\mathbf{f}(\mathbf{u})] = \mathbf{f}(\mathbf{u}^+) - \mathbf{f}(\mathbf{u}^-)$. Unstetigkeiten stückweise glatter Funktionen, die die Sprungbedingung (2.10) erfüllen, werden auch als *Stöße* und s als die *Stoßgeschwindigkeit* bezeichnet.

Für die Burgers-Gleichung mit den Anfangsbedingungen aus dem Beispiel 2.3 erhält man eine schwache Lösung über den Zeitpunkt t^* hinaus, indem man die klassische Lösung durch einen an der Stelle $x = \frac{1}{2}$ startenden Stoß mit der Stoßgeschwindigkeit $s = 0$ fortsetzt, da die Rankine-Hugoniot-Bedingung wegen $f(u^-) = \frac{1}{2}\cos^2\left(\pi\left(\frac{1}{2} - \epsilon\right)\right) = \frac{1}{2}\cos^2\left(\pi\left(\frac{1}{2} + \epsilon\right)\right) = f(u^+)$ und damit

$$[f(u)] = 0 = s[u]$$

in diesem Fall erfüllt ist. Die charakteristischen Grundkurven laufen von $t = 0$ aus in den Stoß hinein, wie in der Abbildung 2.2 dargestellt.
Durch die Erweiterung des Lösungsbegriffs ergibt sich allerdings das Problem, dass zu einer Erhaltungsgleichung und einem Anfangsdatensatz mehrere schwache Lösungen existieren können. Dies lässt sich leicht am Spezialfall des Riemann-Problems verdeutlichen, welches auch bei der Konstruktion numerischer Verfahren eine wichtige Rolle spielt.

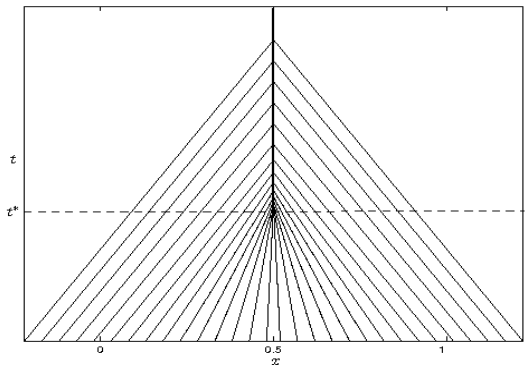


Abb. 2.2: Ausbildung eines Stoßes.

Definition 2.6 *Eine eindimensionale Erhaltungsgleichung*

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

zusammen mit stückweise konstanten Anfangswerten mit nur einer Unstetigkeitsstelle

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}^-, & x < 0, \\ \mathbf{u}^+, & x > 0, \end{cases}$$

wird als Riemann-Problem *bezeichnet.*

Beispiel 2.7 Man betrachte das durch die Burgers-Gleichung (2.7) mit den Anfangswerten

$$u(x,0) = \begin{cases} u^-, & x < 0, \\ u^+, & x > 0, \end{cases}$$

gegebene Riemann-Problem. Gilt $u^- < u^+$, so ist nach Satz 2.5 sowohl der Stoß

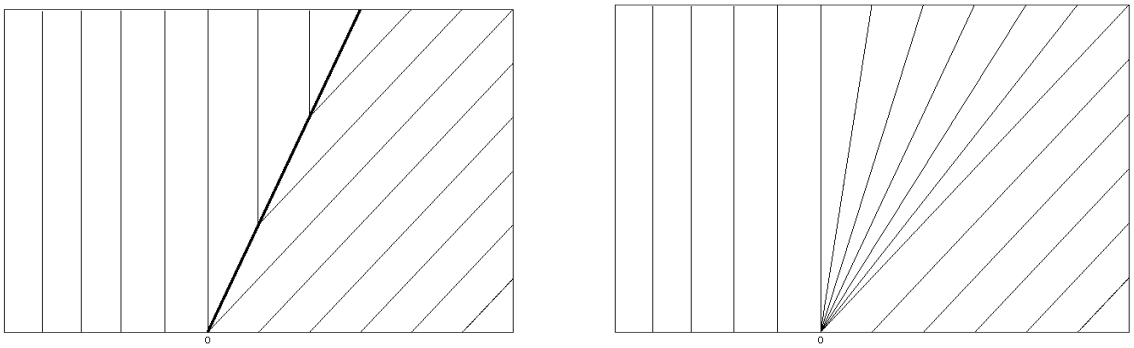
$$u(x,t) = \begin{cases} u^-, & x < st, \\ u^+, & x > st, \end{cases} \tag{2.11}$$

mit Stoßgeschwindigkeit $s = [\frac{1}{2}u^2]/[u] = \frac{u^+ + u^-}{2}$, als auch die *Verdünnungswelle*

$$u(x,t) = \begin{cases} u^-, & x < u^- t, \\ x/t & u^- t \le x \le u^+ t \\ u^+, & x > u^+ t, \end{cases} \tag{2.12}$$

eine schwache Lösung des gegebenen Cauchy-Problems. Der jeweilige Verlauf der charakteristischen Grundkurven im Fall $u^- = 0,\ u^+ = 1$ ist in der Abbildung 2.3 gezeigt. Desweiteren lassen sich noch unendlich viele weitere schwache Lösungen zu diesem Beispiel konstruieren.

Die Mehrdeutigkeit von Lösungen ist insbesondere aus zwei Gründen unerwünscht. Zum einen ist man vom mathematischen Gesichtspunkt aus generell an der Existenz einer eindeutigen Lösung für das Cauchy-Problem interessiert, zum anderen kann bei der Modellierung eines physikalischen Vorgangs durch eine hyperbolische Erhaltungsgleichung nur eine Lösung tatsächlich relevant sein.



(a) (b)

Abb. 2.3: Mehrdeutigkeit der schwachen Lösung; (a) Stoß (2.11), (b) Verdünngungswelle (2.12).

Im Fall der ersteren schwachen Lösung (2.11) zu Beispiel 2.7 laufen charakteristische Grundkurven mit fortschreitender Zeit aus dem Stoß heraus, die sich nicht auf die Grundlinie $t = 0$ zurückbeziehen lassen (siehe Abbildung 2.3a), so dass entlang dieser Kurven keine Informationen über die Anfangsdaten transportiert werden. Zudem erhielte man eine vollständig andere Lösung für leicht gestörte Anfangsdaten, zum Beispiel für einen rapiden, aber glatten Übergang von u^- zu u^+ anstelle des Sprungs. Diese schwache Lösung möchte man daher als nicht sinnvoll deklarieren.
Eine Möglichkeit der Kennzeichnung einer sinnvollen schwachen Lösung stellt die *Viskositätsmethode* dar, die dadurch motiviert ist, dass hyperbolische Erhaltungsgleichungen oft aus komplexeren Gleichungen durch Vernachlässigung dissipativer Terme hervorgehen. Dies ist zum Beispiel bei den Euler-Gleichungen der Gasdynamik der Fall, die sich unter Vernachlässigung von Viskosität und Wärmeleitung aus den kompressiblen Navier-Stokes-Gleichungen herleiten lassen. Geht man von der Existenz von Lösungen der kompressiblen Navier-Stokes-Gleichungen für eine gegebene Nullfolge von Viskositäts- und Wärmeleitungskoeffizienten aus, so erwartet man die Konvergenz in einem geeigneten Sinn gegen Lösungen der Euler-Gleichungen.
Bei der Viskositätsmethode werden nun nur solche schwachen Lösungen einer hyperbolischen Erhaltungsgleichung zugelassen, die sich aus dem Grenzfall verschwindender Dissipation eines zugehörigen dissipativen Systems ergeben. Für theoretische Zwecke wird hierbei die einfache Modifikation

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x},t) + \nabla_{\mathbf{x}} \cdot \mathcal{F}(\mathbf{u}(\mathbf{x},t)) = \epsilon \Delta \mathbf{u}, \qquad \epsilon > 0, \tag{2.13}$$

der Gleichung (2.1) betrachtet und das Verhalten der Lösungen für $\epsilon \to 0$ untersucht.
Die Viskositätsmethode stellt ebenso eine Grundlage für die Konstruktion ausreichend dissipativer numerischer Verfahren dar, die Näherungen an sinnvolle schwache Lösungen hyperbolischer Erhaltungsgleichungen liefern sollen. In diesem Kontext wird dann oft eine komplexere Form des dissipativen Terms $\epsilon\Delta\mathbf{u}$ gewählt.
Aus praktischen Gründen wünscht man sich allerdings handlichere Kriterien als die Viskositätsmethode, die sich direkt auf eine gegebene schwache Lösung anwenden lassen. Zu diesem Zweck entwickelte Bedingungen, die zusätzlich an die schwache Lösung gestellt werden, können verschiedene Formen annehmen und sind sämtlich mit dem Begriff der *Entropie-Bedingung* bezeichnet – in Analogie zur physikalischen Entropie, die den Informationsverlust in einem System beschreibt und nach dem zweiten Hauptsatz der

Thermodynamik mit fortschreitender Zeit nicht geringer werden darf. Angelehnt an diese physikalische Größe werden zu einer gegebenen hyperbolischen Erhaltungsgleichung mathematische Entropiefunktionen definiert.

Definition 2.8 *Eine* Entropiefunktion *für die Erhaltungsgleichung* (2.1)*, deren quasilineare Form durch* (2.3) *gegeben ist, ist eine glatte, konvexe Funktion* $\eta : \mathbb{R}^m \to \mathbb{R}$*, für die eine glatte Funktion* $\mathbf{q} = (q_1, \ldots, q_d)^T : \mathbb{R}^m \to \mathbb{R}^d$ *existiert mit*

$$(\nabla_{\mathbf{u}}\eta)^T \mathbf{A}_l(\mathbf{u}) = (\nabla_{\mathbf{u}} q_l)^T,\ 1 \le l \le d,$$

für alle $\mathbf{u} \in S$*. Hierbei wird* $\mathbf{q}$ *als zu* η *gehöriger* Entropiefluss *und* $(\eta, \mathbf{q})$ *als* Entropie-Entropiefluss-Paar *bezeichnet.*

Im allgemeinen Fall eines Systems von Erhaltungsgleichungen ist die Existenz einer Entropiefunktion nicht zwingend gewährleistet, jedoch ist es für alle aus der Physik hergeleiteten Beispiele möglich, eine Entropiefunktion mit physikalischer Bedeutung zu finden.
Ist zu einer Erhaltungsgleichung (2.1) ein Entropie-Entropiefluss-Paar $(\eta, \mathbf{q})$ im Sinne der Definition 2.8 gegeben, so erfüllt eine klassische Lösung $\mathbf{u}$ mit

$$\begin{aligned} 0 &= (\nabla_{\mathbf{u}}\eta)^T \left[\frac{\partial}{\partial t}\mathbf{u} + \sum_{l=1}^{d} \mathbf{A}_l(\mathbf{u}) \frac{\partial}{\partial x_l}\mathbf{u} \right] \\ &= \frac{\partial}{\partial t}\,\eta(\mathbf{u}) + \sum_{l=1}^{d} \frac{\partial}{\partial x_l}\, q_l(\mathbf{u}) \end{aligned} \tag{2.14}$$

eine zusätzliche skalare Gleichung.
Für eine unstetige schwache Lösung gilt eine schwache Form der Gleichung (2.14) nicht zwingenderweise, wie das Beispiel der Burgers-Gleichung mit dem Entropie-Entropiefluss-Paar $\left(u^2, \frac{2}{3}u^3\right)$ zeigt. Für eine stückweise glatte Lösung ist die Stoßgeschwindigkeit s einer Unstetigkeitskurve in diesem Fall nach der Rankine-Hugoniot-Bedingung durch $s = \left[\frac{1}{2}u^2\right] / [u]$ gegeben. Andererseits lässt sich für stückweise glatte Lösungen der Gleichung (2.14) analog zu Satz 2.5 die Sprungbedingung $[\eta(\mathbf{u})]\, s = \sum_{l=1}^{d} [q_l(\mathbf{u})]\, v_l$ nachweisen, mit dem Normalenvektorfeld $(\mathbf{v}, -s) = \mathbf{n}$ an die Unstetigkeitsfläche, so dass für das gegebene Entropie-Entropiefluss-Paar noch die Bedingung $[u^2] \cdot s = \left[\frac{2}{3}u^3\right]$ zu erfüllen ist. Insbesondere kann der Fall $u^- = -u^+$ daher ausgeschlossen werden. Insgesamt erhält man somit die Gleichung

$$0 = \frac{2}{3} \cdot \frac{(u^+)^3 - (u^-)^3}{(u^+)^2 - (u^-)^2} - \frac{1}{2} \cdot \frac{(u^+)^2 - (u^-)^2}{u^+ - u^-} = \frac{1}{6} \cdot \frac{(u^+ - u^-)^2}{u^+ + u^-}\,,$$

die nur für $u^+ = u^-$ erfüllt ist. Daraus ergibt sich, dass durch die zusätzliche Restriktion (2.14) keine unstetigen schwachen Lösungen der Burgers-Gleichung zugelassen werden. Hingegen zeigt sich, dass eine mit der Viskositätsmethode erhaltene schwache Lösung anstelle dessen eine entsprechende Entropieungleichung erfüllt.

Herleitung der Entropieungleichung aus der Viskositätsmethode Betrachtet man eine Lösung u^ϵ der viskosen Burgers-Gleichung

$$\frac{\partial}{\partial t}u + u \cdot \frac{\partial}{\partial x}u = \epsilon \cdot \frac{\partial^2}{\partial x^2}u, \tag{2.15}$$

so gilt für eine konvexe Funktion $\eta \in C^2(\mathbb{R})$ und einen zugehörigen Entropiefluss q wegen $\eta'' \geq 0$ die Beziehung

$$\begin{aligned}\frac{\partial}{\partial t}\eta(u^\epsilon) + \frac{\partial}{\partial x}q(u^\epsilon) &= \epsilon\eta'(u^\epsilon)\frac{\partial^2}{\partial x^2}u^\epsilon \\ &= \epsilon\frac{\partial^2}{\partial x^2}\,\eta(u^\epsilon) - \epsilon\eta''(u^\epsilon)\cdot\left(\frac{\partial}{\partial x}u^\epsilon\right)^2 \leq \epsilon\frac{\partial^2}{\partial x^2}\eta(u^\epsilon).\end{aligned}$$

Für Testfunktionen $\varphi \in C_0^\infty(\mathbb{R}\times\mathbb{R}^+)$ mit $\varphi \geq 0$ gilt daher die Ungleichung

$$\int_{\mathbb{R}\times\mathbb{R}^+}\left[\eta(u^\epsilon)\frac{\partial}{\partial t}\,\varphi + q(u^\epsilon)\frac{\partial}{\partial x}\,\varphi\right]dxdt \geq -\epsilon\int_{\mathbb{R}\times\mathbb{R}^+}\eta(u^\epsilon)\frac{\partial^2}{\partial x^2}\,\varphi\,dxdt. \tag{2.16}$$

Es sei nun eine Folge von glatten Lösungen $(u^\epsilon)_\epsilon$ der Gleichung (2.15) gegeben, die für $\epsilon \to 0$ fast überall gegen eine Funktion $u \in L^\infty(\mathbb{R}\times\mathbb{R}_0^+)$ konvergiert und für die eine von ϵ unabhängige Konstante $C > 0$ existiert mit $\|u^\epsilon\|_{L^\infty(\mathbb{R}^d\times\mathbb{R}_0^+)} \leq C$. Unter diesen Voraussetzungen erhält man durch den Übergang $\epsilon \to 0$ in der Ungleichung (2.16) und mit Hilfe des Lebesgueschen Satzes von der majorisierten Konvergenz

$$\int_{\mathbb{R}\times\mathbb{R}^+}\left[\eta(u)\frac{\partial}{\partial t}\,\varphi + q(u)\frac{\partial}{\partial x}\,\varphi\right]dxdt \geq 0.$$

d.h. die Funktion u erfüllt die Entropiebedingung

$$\frac{\partial}{\partial t}\,\eta(u) + \frac{\partial}{\partial x}\,q(u) \leq 0 \tag{2.17}$$

im Distributionensinn.
Für eine stückweise glatte schwache Lösung u, die die Ungleichung (2.17) erfüllt, lässt sich in Analogie zur Rankine-Hugoniot-Bedingung die Sprung*ungleichung*

$$[\eta(u)]s \geq [q(u)] \tag{2.18}$$

herleiten, die entlang von Unstetigkeitskurven $x(t)$ mit der Stoßgeschwindigkeit $s = x'(t)$ zu erfüllen ist. Für die Burgers-Gleichung mit dem Entropie-Entropiefluss-Paar $(\eta, q) = (u^2, \frac{2}{3}u^3)$ gilt in (2.18) aufgrund der obigen Ausführungen sogar die strikte Ungleichung

$$[\eta(u)]s > [q(u)]. \tag{2.19}$$

Aus der obigen Ungleichung ergibt sich zudem das Abfallen der Energie bei Auftreten von Stößen, wie nachfolgend erläutert.

Energiedissipation über Stöße Man betrachte das durch

$$E[u](t) = \int_{-\infty}^{\infty}\frac{1}{2}u^2(x,t)dx$$

gegebene *Energiefunktional*, angewandt auf eine stückweise glatte schwache Lösung u der Burgers-Gleichung. Wir nehmen an, dass u genau eine stetig differenzierbare Unstetigkeitskurve $x(t)$ besitzt und die Entropiebedingung (2.17) erfüllt, aus der sich die Gültigkeit der strikten Ungleichung (2.19) für das Entropie-Entropiefluss-Paar $(\eta, q) = (u^2, \frac{2}{3}u^3)$

ergibt. Unter der weiteren Annahme, dass u einen in x kompakten Träger besitzt, ist die Energie endlich. Es gilt dann

$$\begin{aligned}
\frac{dE[u]}{dt} &= \frac{d}{dt}\left(\int_{-\infty}^{x(t)} \frac{1}{2}u^2(x,t)\,dx + \int_{x(t)}^{\infty} \frac{1}{2}u^2(x,t)\,dx\right) \\
&= \frac{1}{2}(u^-)^2 \cdot x'(t) + \int_{-\infty}^{x(t)} u\frac{\partial}{\partial t}u\,dx - \frac{1}{2}(u^+)^2 \cdot x'(t) + \int_{x(t)}^{\infty} u\frac{\partial}{\partial t}u\,dx \\
&= \frac{(u^-)^2 - (u^+)^2}{2} \cdot s - \int_{-\infty}^{x(t)} u^2\frac{\partial}{\partial x}u\,dx - \int_{x(t)}^{\infty} u^2\frac{\partial}{\partial x}u\,dx \\
&= -\frac{1}{2}(\eta(u^+) - \eta(u^-)) \cdot s - \int_{-\infty}^{x(t)} \frac{1}{2}\frac{\partial}{\partial x}q(u)\,dx - \int_{x(t)}^{\infty} \frac{1}{2}\frac{\partial}{\partial x}q(u)\,dx \\
&= -\frac{1}{2}([\eta(u)] \cdot s - [q(u)]) < 0.
\end{aligned}$$

Eine analoge Aussage ergibt sich für eine stückweise glatte, 2π-periodische, schwache Lösung u der Burgers-Gleichung, wenn das Energiefunktional über eine Periodenlänge definiert wird, d.h.

$$E_{per}[u](t) = \int_{-\pi}^{\pi} \frac{1}{2}u^2(x,t)dx.$$

Für mehrdimensionale Systeme hyperbolischer Erhaltungsgleichungen ist die Entropiebedingung durch die mehrdimensionale Form der im Distributionensinn zu verstehenden Ungleichung (2.17) gegeben. In Analogie zum Begriff der schwachen Lösung wird desweiteren der Begriff der Entropielösung zur Kennzeichnung der "richtigen" schwachen Lösung eingeführt.

Definition 2.9 *Eine schwache Lösung* $\mathbf{u}$ *des Cauchy-Problems* (2.1), (2.2) *wird als* Entropielösung *bezeichnet, wenn* $\mathbf{u}$ *für alle Entropie-Entropiefluss-Paare* $(\eta, \mathbf{q})$ *von* (2.1) *und für alle Testfunktionen* $\varphi \in C_0^1(\mathbb{R}^d \times \mathbb{R}_0^+),\ \varphi \geq 0$, *die Ungleichung*

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} \left[\eta(\mathbf{u})\frac{\partial}{\partial t}\varphi + \sum_{l=1}^{d} q_l(\mathbf{u})\frac{\partial}{\partial x_l}\varphi\right] d\mathbf{x}dt + \int_{\mathbb{R}^d} \eta(\mathbf{u}_0(\mathbf{x}))\varphi(\mathbf{x},0)d\mathbf{x} \geq 0$$

erfüllt.

Im Fall skalarer Erhaltungsgleichungen lässt sich ein Existenz- und Eindeutigkeitsresultat für Entropielösungen nachweisen, während dies für allgemeine hyperbolische Systeme von Erhaltungsgleichungen nur vermutet werden kann.

Satz 2.10 *Sei* $u_0 \in L^\infty(\mathbb{R}^d)$. *Dann existiert zu der skalaren Erhaltungsgleichung* (2.4) *und der Anfangsbedingung* $u(\mathbf{x},0) = u_0$ *eine eindeutige Entropielösung.*

Beweis: Siehe [57]. □

Bei der Entwicklung numerischer Verfahren für Systeme von Erhaltungsgleichungen setzen wir hingegen grundsätzlich die Existenz einer eindeutigen Entropielösung zu einem gegebenen Testproblem voraus.

2.2 Skalare Testprobleme

Bei den nachfolgenden skalaren Testproblemen, die in Kapitel 7 zur Validierung der numerischen Methode verwendet werden, handelt es sich um zweidimensionale Erhaltungsgleichungen der Form (2.4) auf beschränkten, polygonal berandeten Gebieten $\Omega \subset \mathbb{R}^2$. In der Situation beschränkter Rechengebiete sind zusätzlich zu den Anfangsbedingungen ebenso Randbedingungen vorzugeben. Hierbei muss unterschieden werden zwischen Einströmrändern, an denen Werte festzulegen sind, und Ausströmrändern, an denen sich diese durch den Verlauf der charakteristischen Grundkurven aus den im Inneren des Gebiets vorliegenden Werten ergeben.

Testproblem 2.11 (Lineare Advektion) Das Cauchy-Problem

$$\frac{\partial}{\partial t}u + \mathbf{a} \cdot \nabla_{\mathbf{x}} u \;=\; 0, \qquad u(\mathbf{x}, 0) = u_0(\mathbf{x}), \tag{2.20}$$

mit einem konstanten Vektor $\mathbf{a} \in \mathbb{R}^2$ beschreibt den linearen Transport von Anfangsdaten $u_0(\mathbf{x})$ mit Geschwindigkeitsvektor $\mathbf{a}$ und wird als *lineare Advektionsgleichung* bezeichnet. Die zugehörigen charakteristischen Grundkurven sind nach (2.6) durch die Parallelen $\boldsymbol{\gamma}(t) = \mathbf{a}\,t + \mathbf{x}_0$ gegeben, so dass sich die Lösung der Gleichung in der Form

$$u(\mathbf{x}, t) = u_0(\mathbf{x} - \mathbf{a}t)$$

darstellen lässt. Die exakte Lösung lässt sich somit durch Verschieben der Anfangsbedingungen bestimmen.
Das numerische Verfahren wird angewendet auf die lineare Advektionsgleichung mit Geschwindigkeitsvektor $\mathbf{a} = (1, 1)^T$ im räumlichen Gebiet $\Omega = (0, 1)^2$, so dass Randbedingungen am linken und unteren Rand dieses Rechengebiets vorgegeben werden müssen. Bei Vorliegen eines Anfangsdatensatzes u_0 auf dem gesamten Gebiet $\mathbb{R}^2$ sind die vorzugebenden Randbedingungen durch die exakte Lösung der linearen Advektionsgleichung bestimmt.

Testproblem 2.12 Als zweiter Testfall auf dem Gebiet $\Omega = (0, 1)^2$ soll die Gleichung

$$\frac{\partial u}{\partial t} + u \cdot \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2} = 0\,. \tag{2.21}$$

betrachtet werden. Durch $u_0(\mathbf{x}) = 0$ für $\mathbf{x} \in \Omega$ seien Anfangsbedingungen vorgegeben. Die zu (2.21) gehörigen charakteristischen Grundkurven sind gegeben durch $\boldsymbol{\gamma}(t) = (u_0(\mathbf{x}_0), 1)^T + \mathbf{x}_0$, so dass der obere Rand $\{\mathbf{x} \in \partial\Omega \mid x_2 = 1\}$ des Rechengebiets als Ausflussrand zu behandeln ist. An den übrigen Randpunkten werden die Randwerte

$$u_0(x_1, x_2, t) = \begin{cases} 1.5 & \text{falls} \quad x_1 = 0, \\ -1 & \text{falls} \quad x_1 = 1, \\ 1.5 - 2.5x_1 & \text{falls} \quad x_2 = 0, \end{cases}$$

vorgegeben.

Für $t \to \infty$ ergibt sich eine stationäre Lösung des Problems, die sich durch das Verfolgen der charakteristischen Grundkurven exakt berechnen lässt. Die Konstruktion dieser stationären Lösung, die in Abbildung 2.4 dargestellt ist, ergibt sich aus der Betrachtung des Problems für $\frac{\partial}{\partial t}u = 0$, mit der Gleichung

$$\frac{\partial u}{\partial x_2} + u \cdot \frac{\partial u}{\partial x_1} = 0.$$

Hierbei handelt es sich um die Burgers-Gleichung (2.7), deren Lösung durch Betrachten der charakteristischen Grundkurven in der x_1-x_2-Ebene, die sich mittels der Gleichung $\frac{dx_1}{dx_2} = u\left(x_1(x_2), x_2\right)$ bestimmen lassen, konstruiert werden kann. Entlang dieser Kurven ist die Lösung konstant, so dass es sich bei den von einem Punkt $Q = (\xi, 0),\ \xi \in [0,1]$, der unteren Kante des Quadrats ausgehenden charakteristischen Grundkurven um Geraden mit der Gleichung

$$x_1 = \xi + u_0(\xi, 0)x_2 \tag{2.22}$$

handelt. Die durch den Punkt $(0,0)$ verlaufende Gerade g_1 ist gegeben durch $x_1 = \frac{3}{2}x_2$, während die durch den Punkt $(1,0)$ verlaufende Gerade g_2 die Gleichung $x_1 = 1 - x_2$ erfüllt. Das Einsetzen von $x_2 = \frac{2}{5}$ in die Gleichungen der Form (2.22) ergibt desweiteren $x_1 = \xi + \left(\frac{3}{2} - \frac{5}{2} \cdot \xi\right) \cdot \frac{2}{5} = \frac{3}{5}$, so dass sich alle durch die Punkte auf dem unteren Rand des Quadrats verlaufenden Geraden im Punkt $P = (\frac{3}{5}, \frac{2}{5})$ treffen. Die vom linken und rechten Rand des Quadrats ausgehenden charakteristischen Grundkurven verlaufen jeweils parallel zu den Geraden g_1 und g_2. Im Punkt P entsteht ein Stoß entlang der Geraden g_3, die die beiden konstanten Zustände $u^- = 1.5$ und $u^+ = -1.0$ trennt. Nach der Rankine-Hugoniot-Bedingung ist die zugehörige Stoßgeschwindigkeit durch $s = \frac{1}{2}(u^- + u^+) = \frac{1}{4}$ gegeben. Desweiteren ist der Wert der stationären Lösung entlang der Strecke $\overline{PQ}$ durch $u(\mathbf{x}) = u_0(\xi, 0) = 1.5 - 2.5 \cdot \xi$ bestimmt. Auf dem dreiecksförmigen Gebiet, welches vom unteren Rand des Quadrats und dem Punkt P aufgespannt wird, besitzt die stationäre Lösung die explizite Form

$$u(x_1, x_2) \quad = \quad 1.5 - 2.5\,\frac{x_1 - 1.5x_2}{1 - 2.5x_2}.$$

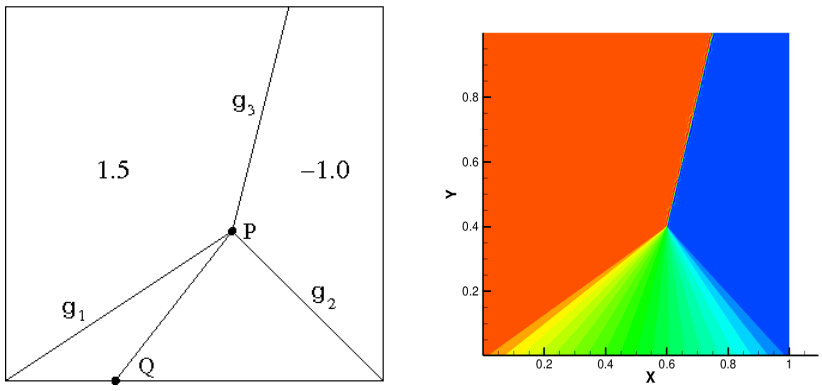


Abb. 2.4: Die stationäre Lösung der Gleichung (2.21).

Testproblem 2.13 (Burgers-Gleichung in 2D) Als weiterer Testfall soll die Gleichung

$$\frac{\partial}{\partial t}u + u \cdot \frac{\partial}{\partial x_1}u + u \cdot \frac{\partial}{\partial x_2}u = 0, \tag{2.23}$$

auf dem Gebiet $\Omega = (-1,1)^2$, mit den periodischen Randbedingungen

$$u(-1, x_2, t) = u(1, x_2, t), \quad u(x_1, -1, t) = u(x_1, 1, t),$$

betrachtet werden. Als Anfangsdatensatz wird hierbei die Funktion

$$u_0(\mathbf{x}) = \frac{1}{2} + \frac{1}{4}\sin(\pi(x+y))$$

vorgegeben.

Testproblem 2.14 Auf dem Gebiet $\Omega = (-1.8, 1.8) \times (-2.3, 1.3)$ wird die Erhaltungsgleichung

$$\frac{\partial u}{\partial t} + \frac{\partial \sin u}{\partial x_1} + \frac{\partial \cos u}{\partial x_2} = 0\,, \tag{2.24}$$

mit der Anfangsbedingung

$$u(x_1, x_2, 0) = \begin{cases} 3.5\pi & \text{für } x_1^2 + x_2^2 < 1, \\ 0.25\pi & \text{andernfalls,} \end{cases}$$

betrachtet. Bei diesem Testfall handelt es sich um eine nichtkonvexe Erhaltungsgleichung, da die Funktion

$$\mathbf{f}''(u) \cdot \mathbf{n} = a_1'(u)n_1 + a_2'(u)n_2 = -\sin(u)n_1 - \cos(u)n_2$$

für alle normierten Vektoren $\mathbf{n} \in \mathbb{R}^2$ sowohl positive als auch negative Werte annehmen kann. Derartige Gleichungen zeichnen sich dadurch aus, dass ihre Riemann-Problemen Entropielösungen mit Folgen von aneinander grenzenden Stößen und Verdünnungswellen zulassen, während die Entropielösung eines skalaren Riemann-Problems im Fall einer konvexen oder konkaven Flussfunktion entweder ein Stoß oder eine Verdünnungswelle ist. Untersuchungen von Qiu und Shu [78] sowie von Kurganov et al. [59] haben gezeigt, dass einige Verfahren höherer Ordnung für nichtkonvexe Erhaltungsgleichungen nicht die korrekte Entropielösung liefern, so dass die Untersuchung des Verhaltens des in dieser Arbeit konstruierten DG-Verfahrens auch für nichtkonvexe Gleichungen anhand der obigen Testgleichung sinnvoll ist.
Da die zur Gleichung (2.24) gehörigen charakteristischen Grundkurven außerhalb des durch $x_1^2 + x_2^2 = 1$ gegebenen Kreises die Form

$$\gamma(t) = (\cos(0.25\pi), -\sin(0.25\pi))^T \cdot t + \mathbf{x}_0 = (1/\sqrt{2}, -1/\sqrt{2})^T \cdot t + \mathbf{x}_0$$

besitzen, müssen Randbedingungen am linken und oberen Rand des Rechengebiets vorgegeben werden. Dementsprechend werden die Bedingungen $u(x_1, x_2, t) = 0.25\pi$ für $t \in \mathbb{R}^+$ und $\mathbf{x} \in \partial\Omega$ mit $x_1 = -1.8$ oder $x_2 = 1.3$ festgelegt.

2.3 Die Euler-Gleichungen der Gasdynamik

Die Euler-Gleichungen der Gasdynamik, mit denen insbesondere Strömungen kompressibler Gase beschrieben werden, sind ein besonders wichtiges Beispiel eines Systems hyperbolischer Erhaltungsgleichungen. Zum einen finden sie vielfach Anwendung in praxisrelevanten Problemstellungen, so dass mathematisch fundierte Lösungsverfahren für diese Gleichungen auch im ingenieurwissenschaftlichen Bereich auf großes Interesse stoßen. Zum anderen erlaubt die Vielzahl bereits untersuchter und gut dokumentierter Testfälle für diese Gleichungen die Untersuchung und Validierung numerischer Verfahren in Bezug auf ihr

Verhalten für Systeme von Erhaltungsgleichungen, deren Lösungen vielfältigere Strukturen aufweisen können als skalare Gleichungen. Insbesondere lassen sich die Eigenschaften eines neuartigen oder modifizierten Verfahrens bei Auftreten von Verdichtungsstößen, Verdünnungswellen und Kontaktunstetigkeiten sowie der nichtlinearen Interaktion derartiger Phänomene untersuchen und mit bestehenden Verfahren vergleichen.
Das System der Euler-Gleichungen lässt sich aus den physikalischen Prinzipien der Erhaltung von Masse, Impuls und Energie herleiten. Eine ausführliche Behandlung der entsprechenden physikalischen Grundlagen findet man in [26] und [97]. An dieser Stelle soll nur eine verkürzte Darstellung der Herleitung der Gleichungen gegeben werden. Hierzu wird die Strömung eines idealen, polytropen Gases betrachtet, welches durch die als Funktionen von Raum und Zeit gegebenen physikalischen Größen *Dichte* $\rho : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}$, *Druck* $p : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}$, *Geschwindigkeit* $\mathbf{v} : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}^d$ sowie *spezifische innere Energie* $e : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}$ beschrieben werden kann. Betrachtet man exemplarisch die physikalische Dichte des Gases, so ist eine Änderung der durch $\int_\Omega \rho(\mathbf{x}, t)\, d\mathbf{x}$ gegebenen Masse innerhalb eines als *Kontrollvolumen* bezeichneten beschränkten Gebiets Ω nur durch Ein- oder Ausströmen des Gases über den Rand $\partial\Omega$ möglich. Der entsprechende Massenfluss, der die Masse ausströmenden Gases in einer Zeiteinheit beschreibt, ist gegeben durch $\int_{\partial\Omega} \rho(\mathbf{x}, t)\mathbf{v}(\mathbf{x}, t) \cdot \mathbf{n}\, d\sigma$, wobei $\mathbf{n}$ die normierte äußere Normale an $\partial\Omega$ ist. Die zeitliche Änderung der Masse ist damit durch

$$\frac{d}{dt} \int_\Omega \rho\, d\mathbf{x} = - \int_{\partial\Omega} \rho\mathbf{v} \cdot \mathbf{n}\, d\sigma$$

beschrieben. Unter der Voraussetzung, dass ρ und $\mathbf{v}$ stetig differenzierbare Funktionen sind, kann die zeitliche Differentiation und räumliche Integration vertauscht und der Gaußsche Integralsatz angewendet werden, so dass sich die Gleichung

$$\int_\Omega \frac{\partial}{\partial t}\rho\, d\mathbf{x} = \int_\Omega \nabla_{\mathbf{x}} \cdot (\rho\mathbf{v})\, d\mathbf{x}$$

ergibt. Da die Erhaltungseigenschaft für beliebige Kontrollvolumina Ω gilt, und die Integranden als stetig vorausgesetzt wurden, erhält man die Differentialgleichung

$$\frac{\partial}{\partial t}\rho + \nabla_{\mathbf{x}} \cdot (\rho\mathbf{v}) = 0,$$

die als *Kontinuitätsgleichung* bezeichnet wird und die erste Gleichung des Systems darstellt.
Im Allgemeinen ergibt sich für jede skalare Größe $z : \mathbb{R}^d \times \mathbb{R}_0^+ \to \mathbb{R}$, die mit der Strömung transportiert wird, eine zeitliche Änderung hervorgerufen durch den Fluss $z\mathbf{v}$. Die ursprünglich von Euler aufgestellten *Impulserhaltungsgleichungen*, die die zeitliche Änderung der Komponenten ρv_i des Impulses $\rho\mathbf{v}$ beschreiben, beinhalten daher Flüsse der Form $\rho v_i\mathbf{v}$. Zusätzlich kann eine Beschleunigung der Strömung – und damit eine Änderung des Impulses – durch auf das Gas wirkende äußere Kräfte, wie beispielsweise Gravitation, sowie durch innere Kräfte, beispielsweise Druck oder durch Reibung verursachte Scherkräfte, hervorgerufen werden. Können äußere und durch Reibung verursachte innere Kräfte vernachlässigt werden, so ergeben sich weitere Impulsänderungen nur aufgrund von Druckunterschieden. Die Impulserhaltungsgleichungen sind dementsprechend durch

$$\frac{\partial}{\partial t}\rho v_i + \nabla_{\mathbf{x}} \cdot (\rho v_i\mathbf{v}) + \frac{\partial}{\partial x_i}p = 0, \qquad i = 1, \ldots, d,$$

gegeben. Aus ähnlichen Überlegungen ergibt sich die *Energieerhaltungsgleichung*

$$\frac{\partial}{\partial t}\rho E + \nabla_{\mathbf{x}} \cdot ((\rho E + p)\mathbf{v}) = 0,$$

wobei mit $E = e + \frac{\|\mathbf{v}\|^2}{2}$ die aus der spezifischen inneren und der spezifischen kinetischen Energie zusammengesetzte *spezifische Gesamtenergie* bezeichnet ist. Zusammenfassend ergibt sich im zweidimensionalen Fall $d = 2$ das System

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho E \end{pmatrix} + \frac{\partial}{\partial x_1}\begin{pmatrix} \rho v_1 \\ \rho v_1^2 + p \\ \rho v_1 v_2 \\ (\rho E + p) v_1 \end{pmatrix} + \frac{\partial}{\partial x_2}\begin{pmatrix} \rho v_2 \\ \rho v_1 v_2 \\ \rho v_2^2 + p \\ (\rho E + p) v_2 \end{pmatrix} = \mathbf{0}. \tag{2.25}$$

Zum System (2.25) gehört eine zusätzliche Gleichung, die den Druck als Funktion der Dichte und der spezifischen inneren Energie beschreibt und damit das andernfalls unterbestimmte System schließt. In einem *idealen* Gas ist der Druck bestimmt durch

$$p = R\rho T, \tag{2.26}$$

wobei T die *Temperatur* des Gases und R die sogenannte *Gaskonstante* ist. Das Gas ist *polytrop*, wenn die spezifische innere Energie durch

$$e = c_V T$$

gegeben ist, wobei c_V eine stoffabhängige Konstante ist, die als die *Wärmekapazität bei konstantem Volumen* bezeichnet wird. Aus den obigen Zusammenhängen ergibt sich die *Zustandsgleichung für ideale, polytrope Gase*,

$$p = (\gamma - 1)\rho e = (\gamma - 1)\rho\left(E - \frac{\|\mathbf{v}\|^2}{2}\right),$$

mit dem Isentropenkoeffizienten $\gamma = \left(\frac{R}{c_V} + 1\right)$. Für ein diatomisches Gas wie trockene Luft, die primär aus den Molekülen N_2 und O_2 zusammengesetzt ist, wird beispielsweise ein Wert von $\gamma = 1.4$ angenommen.

Hyperbolizität und Rotationsinvarianz der Euler-Gleichungen In den konservativen Variablen $\mathbf{u} = (\rho, \rho v_1, \rho v_2, \rho E)^T$ sind die zum System der Euler-Gleichungen gehörigen Flussfunktionen durch

$$\begin{aligned}
\mathbf{f}_1(\mathbf{u}) &= \left(u_2,\ \frac{u_2^2}{u_1} + (\gamma - 1)\left[u_4 - \frac{u_2^2 + u_3^2}{2u_1}\right],\ \frac{u_2 u_3}{u_1},\ \left[\gamma u_4 - (\gamma - 1)\frac{u_2^2 + u_3^2}{2u_1}\right]\frac{u_2}{u_1}\right)^T, \\
\mathbf{f}_2(\mathbf{u}) &= \left(u_3,\ \frac{u_2 u_3}{u_1},\ \frac{u_3^2}{u_1} + (\gamma - 1)\left[u_4 - \frac{u_2^2 + u_3^2}{2u_1}\right],\ \left[\gamma u_4 - (\gamma - 1)\frac{u_2^2 + u_3^2}{2u_1}\right]\frac{u_3}{u_1}\right)^T,
\end{aligned}$$

gegeben.

Dementsprechend errechnet man die Jacobi-Matrizen

$$\mathbf{A}_1(\mathbf{u}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{\gamma-3}{2}v_1^2 + \frac{\gamma-1}{2}v_2^2 & (3-\gamma)v_1 & (1-\gamma)v_2 & \gamma-1 \\ -v_1v_2 & v_2 & v_1 & 0 \\ (\gamma-1)v_1\|v\|^2 - \gamma v_1 E & \gamma E - \frac{\gamma-1}{2}(3v_1^2+v_2^2) & (1-\gamma)v_1v_2 & \gamma v_1 \end{pmatrix},$$

$$\mathbf{A}_2(\mathbf{u}) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ -v_1v_2 & v_2 & v_1 & 0 \\ \frac{\gamma-1}{2}v_1^2 + \frac{\gamma-3}{2}v_2^2 & (1-\gamma)v_1 & (3-\gamma)v_2 & \gamma-1 \\ (\gamma-1)v_2\|v\|^2 - \gamma v_2 E & (1-\gamma)v_1v_2 & \gamma E - \frac{\gamma-1}{2}(v_1^2+3v_2^2) & \gamma v_2 \end{pmatrix}.$$

Zum Nachweis der Hyperbolizität lässt sich nun die Eigenschaft der Rotationsinvarianz der Euler-Gleichungen ausnutzen, die besagt, dass für beliebige normierte Vektoren $\boldsymbol{\nu} \in \mathbb{R}^2,\ \|\boldsymbol{\nu}\| = 1$, mit der Definition der Drehmatrix

$$\mathbf{Q}(\boldsymbol{\nu}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \nu_1 & \nu_2 & 0 \\ 0 & -\nu_2 & \nu_1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

die Gleichung

$$\nu_1\mathbf{f}_1(\mathbf{u}) + \nu_2\mathbf{f}_2(\mathbf{u}) = \mathbf{Q}(\boldsymbol{\nu})^{-1}\mathbf{f}_1\left(\mathbf{Q}(\boldsymbol{\nu})\mathbf{u}\right) \tag{2.27}$$

erfüllt ist. Durch Differentiation bezüglich $\mathbf{u}$ erhält man daraus die Gleichung

$$\nu_1\mathbf{A}_1(\mathbf{u}) + \nu_2\mathbf{A}_2(\mathbf{u}) = \mathbf{Q}(\boldsymbol{\nu})^{-1}\mathbf{A}_1\left(\mathbf{Q}(\boldsymbol{\nu})\mathbf{u}\right)\mathbf{Q}(\boldsymbol{\nu}),$$

so dass die Betrachtung der Eigenwerte der Matrix $\mathbf{A}_1\left(\mathbf{Q}(\boldsymbol{\nu})\mathbf{u}\right)$ hinreichend ist.
Die Eigenwerte der Matrix $\mathbf{A}_1\left(\mathbf{Q}(\boldsymbol{\nu})\mathbf{u}\right)$ sind gegeben durch

$$\begin{aligned} \lambda_1 = \lambda_2 &= \boldsymbol{\nu}\cdot\mathbf{v}, \\ \lambda_3 &= \lambda_1 + c, \\ \lambda_4 &= \lambda_1 - c, \end{aligned} \tag{2.28}$$

mit der als *Schallgeschwindigkeit* bezeichneten Größe $c = \sqrt{\frac{\gamma p}{\rho}}$. Die zugehörigen Eigenvektoren der Matrix $\mathbf{A}_1\left(\mathbf{Q}(\boldsymbol{\nu})\mathbf{u}\right)$ sind

$$\begin{aligned} \mathbf{r}_1 &= \left(1,\, \boldsymbol{\nu}\cdot\mathbf{v},\, \nu_1v_2 - \nu_2v_1,\, \|\mathbf{v}\|^2/2\right)^T, \\ \mathbf{r}_2 &= \left(0,\, 0,\, -1,\, \nu_2v_1 - \nu_1v_2\right)^T, \\ \mathbf{r}_3 &= \left(1,\, \boldsymbol{\nu}\cdot\mathbf{v} + c,\, \nu_1v_2 - \nu_2v_1,\, H + c\boldsymbol{\nu}\cdot\mathbf{v}\right)^T, \\ \mathbf{r}_4 &= \left(1,\, \boldsymbol{\nu}\cdot\mathbf{v} - c,\, \nu_1v_2 - \nu_2v_1,\, H - c\boldsymbol{\nu}\cdot\mathbf{v}\right)^T, \end{aligned}$$

mit der als *Enthalpie* bezeichneten Größe $H = E + \frac{p}{\rho}$. Für physikalisch sinnvolle Werte von positiver Dichte und positivem Druck ist die gegebene Matrix daher reell diagonalisierbar, so dass es sich bei den Euler-Gleichungen um ein hyperbolisches System handelt.

Randbedingungen Neben der Betrachtung von periodischen Randbedingungen sowie numerischen Ein- und Ausströmrändern, wie im Fall einer skalaren Gleichung, beinhalten einige Testfälle der Euler-Gleichungen feste Wände. An diesen physikalischen Rändern wird die Undurchlässigkeitsbedingung $\mathbf{v} \cdot \mathbf{n} = 0$ gefordert, wobei $\mathbf{n}$ äußeres Normalenvektorfeld an den Gebietsränd $\partial\Omega$ ist. Aus dieser Bedingung folgt für die Flussfunktion der Eulergleichungen, siehe (2.25),

$$n_1 \cdot \mathbf{f}_1(\mathbf{u}) + n_2 \cdot \mathbf{f}_2(\mathbf{u}) = \begin{pmatrix} \rho \mathbf{v} \cdot \mathbf{n} \\ \rho v_1 \mathbf{v} \cdot \mathbf{n} + p n_1 \\ \rho v_2 \mathbf{v} \cdot \mathbf{n} + p n_2 \\ (\rho E + p) \mathbf{v} \cdot \mathbf{n} \end{pmatrix} = \begin{pmatrix} 0 \\ p n_1 \\ p n_2 \\ 0 \end{pmatrix} =: \mathbf{f}_W(\mathbf{u}, \mathbf{n}). \tag{2.29}$$

3 Orthogonale Polynome und Quadraturformeln auf dem Dreieck

Die Lösungen hyperbolischer Erhaltungsgleichungen sollen im Zuge des in dieser Arbeit verwendeten numerischen Verfahrens durch stückweise polynomiale Funktionen auf Dreiecksgittern approximiert werden. Der Hintergrund zur Verwendung von Dreiecksgittern ist, dass diese im Vergleich zu kartesischen Gittern mehr Flexibilität im Hinblick auf die Diskretisierung komplexer Gebiete ermöglichen. Desweiteren soll eine hohe räumliche Auflösung des Verfahrens durch Verwendung hoher Polynomgrade bei vergleichsweise groben Gittern erreicht werden. Mit der Verwendung hoher Polynomgrade nähert man sich der Klasse der spektralen Methoden, die auf der Entwicklung der exakten Lösung mittels einer Basis trigonometrischer Funktionen oder orthogonaler Polynome basieren, die auf einfachen Rechengebieten global definiert sind. Im Gegensatz zur klassische Formulierung einer Spektralmethode sehen spätere Spektral-Element-Methoden zudem auch Zerlegungen des Rechengebiets vor. Herangehensweisen der Spektralmethoden sollen im Rahmen dieser Arbeit insbesondere auf das Vorliegen von Dreieckszerlegungen übertragen werden. Die Grundlagen spektraler Methoden sowie der Spektral-Element-Methoden sind beispielsweise in den Büchern von Canuto, Hussaini, Quarteroni und Zang [11, 12], Hesthaven, Gottlieb und Gottlieb [43] sowie Karniadakis und Sherwin [52] zu finden.
Die Effizienz spektraler Methoden ist im Fall der Verwendung kartesischer Gitter insbesondere auf die Konstruktion orthogonaler Tensorproduktbasen bestehend aus Produkten eindimensionaler Basisfunktionen zurückzuführen. Um eine ähnliche Effizienzsteigerung auch für Dreiecksgebiete zu ermöglichen, konstruierte Dubiner in [30] eine orthogonale verallgemeinerte Tensorproduktbasis auf dem Dreieck, die auf die konkrete Anwendung in spektralen Verfahren ausgerichtet war. Tatsächlich ist diese Basis ein Spezialfall einer bereits von Proriol [74] beschriebenen Familie von Polynombasen, die auch im Rahmen einer Untersuchung orthogonaler Polynome in zwei Variablen auf unterschiedlichen Gebieten von Koornwinder [54] aufgegriffen wurde. Von Karniadakis und Sherwin wurden die Proriol-Koornwinder-Dubiner(PKD)-Polynome in konkreten Implementationen zur effizienten Nutzung in adaptiven Spektral-Element-Verfahren verwendet und auf den dreidimensionalen Fall erweitert, siehe [52]. Da die PKD-Polynome ein wesentlicher Bestandteil des in dieser Arbeit verwendeten numerischen Verfahrens sind, soll im Folgenden auf ihre Konstruktion und ihre Approximationseigenschaften eingegangen werden. Insbesondere sind diese Polynome Eigenfunktionen eines singulären Sturm-Liouville-Problems auf dem Dreieck, dessen zugehöriger Differentialoperator bei der Betrachtung an die Polynombasis angepasster künstlicher Viskosität in Kapitel 5 den Laplace-Operator in der Viskositätsformulierung (2.13) ersetzen wird. Ebenso werden die im Verfahren genutzten Quadraturformeln beschrieben, die bereits in [30] konstruiert wurden und in engem Zusammenhang zu den PKD-Polynomen stehen.

3.1 Konstruktion der Proriol-Koornwinder-Dubiner-Polynome

Die PKD-Polynome lassen sich mit Hilfe eines zusammenfallenden Koordinatensystems *"collapsed coordinate system"*, durch das ein Dreieck als Quadrat mit zwei identischen Ecken aufgefasst wird, aus den eindimensionalen Jacobi-Polynomen $P_n^{\alpha,\beta}$ konstruieren. Sie bilden eine hierarchische Basis bestehend aus Polynomen in den Dreieckskoordinaten, die bezüglich der Integration über die Dreiecksfläche orthogonal sind.

Zur Konstruktion betrachten wir die im Folgenden als Standarddreieck bezeichnete Menge

$$\mathbb{T} = \{(r,s) \in \mathbb{R}^2 \mid -1 \le r,s;\ r+s \le 0\},$$

sowie das Standardquadrat

$$R = \{(a,b) \in \mathbb{R}^2 \mid -1 \le a,b \le 1\}.$$

Die Transformation

$$\Psi : \begin{pmatrix} r \\ s \end{pmatrix} \mapsto \begin{pmatrix} 2\frac{1+r}{1-s} - 1 \\ s \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \tag{3.1}$$

bildet das durch $\mathbb{T}' = \mathbb{T} \setminus \{(-1,1)\}$ gegebene nach oben offene Dreieck auf das nach oben offene Quadrat $R' = R \setminus \{(a,1) \mid -1 \le a \le 1\}$ ab und kann als Auseinanderziehen des Dreiecks an der oberen Ecke veranschaulicht werden, siehe Abbildung 3.1. Das Koordi-

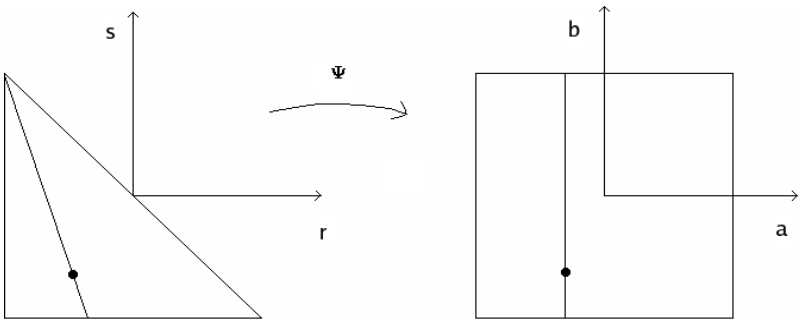


Abb. 3.1: Abbildung auf das zusammenfallende Koordinatensystem.

natensystem (a,b) wird dann als das zusammenfallende Koordinatensystem bezeichnet. Seien $p, q : [-1,1] \to \mathbb{R}$ Polynome in einer Variablen, wobei p den Grad l und q den Grad m besitze, so ist

$$\Phi_{lm} : \mathbb{T}' \to \mathbb{R}; \quad \Phi_{lm}(r,s) = p\left(2\frac{1+r}{1-s} - 1\right)\left(\frac{1-s}{2}\right)^l q(s)$$

zu einem Polynom in den Variablen r und s auf ganz $\mathbb{T}$ fortsetzbar. Zudem lässt sich Φ_{lm} in $\mathbb{T}'$ als Produkt zweier eindimensionaler Polynome

$$\Phi_{lm}(r,s) = \Phi_l^1(a(r,s)) \cdot \Phi_{lm}^2(b(r,s))$$

schreiben, mit $\Phi_l^1(a) = p(a)$ und $\Phi_{lm}^2(b) = \left(\frac{1-b}{2}\right)^l q(b)$.
Die PKD-Polynome erhält man nun durch die Wahl der Jacobi-Polynome l-ten und m-ten Grades,

$$\begin{aligned} p(a) &= P_l^{0,0}(a), \\ q(b) &= P_m^{2l+1,0}(b), \end{aligned}$$

zu den Gewichtsfunktionen $\omega(x) = 1$ bzw. $\omega(x) = (1-x)^{2l+1}$. Zur Definition und zu den Eigenschaften der Jacobi-Polynome sei an dieser Stelle auf den Anhang A.1 verwiesen. Für die PKD-Polynome ergibt sich aus dieser Konstruktion die Definition

$$\Phi_{lm}(r,s) = P_l^{0,0}\Big(2\frac{1+r}{1-s} - 1\Big)\Big(\frac{1-s}{2}\Big)^l P_m^{2l+1,0}(s), \quad l,m \in \mathbb{N}_0.$$

Diese Polynome erzeugen nun die gewünschte orthogonale verallgemeinerte Tensorproduktbasis, wie nachfolgend gezeigt wird.

Lemma 3.1 *Die Polynome Φ_{lm} sind orthogonal bezüglich des L^2-Skalarproduktes*

$$(\Phi, \tilde{\Phi})_{L^2(\mathbb{T})} = \int_{\mathbb{T}} \Phi(r,s)\tilde{\Phi}(r,s)\, drds, \qquad \Phi, \tilde{\Phi} \in L^2(\mathbb{T}),$$

und es gilt

$$\gamma_{lm} := (\Phi_{lm}, \Phi_{lm})_{L^2(\mathbb{T})} = \frac{2}{(2l+1)(l+m+1)}. \tag{3.2}$$

Zudem ist für ein fest gewähltes $N \in \mathbb{N}_0$ die Menge $\{\Phi_{lm} \mid 0 \le l+m \le N\}$ eine Basis des Raumes $\mathcal{P}^N(\mathbb{T}) = Span\{r^l s^m \,|\, 0 \le l+m \le N\}$ der Polynome auf $\mathbb{T}$ vom Grad kleiner oder gleich N.

Beweis: Für die Jacobimatrix der Rücktransformation des Quadrats auf das Dreieck gilt

$$\left|\det \begin{pmatrix} \partial r/\partial a & \partial r/\partial b \\ \partial s/\partial a & \partial s/\partial b \end{pmatrix}\right| = \frac{1-b}{2}\,.$$

Daher ergibt sich für das innere Produkt zweier PKD-Polynome

$$\begin{aligned}
(\Phi_{lm}, \Phi_{jk})_{L^2(\mathbb{T})} &= \int_{\mathbb{T}} \Phi_{lm}(r,s)\Phi_{jk}(r,s)\, drds \\
&= \int_R P_l^{0,0}(a) P_j^{0,0}(a) \Big(\frac{1-b}{2}\Big)^{l+j+1} P_m^{2l+1,0}(b) P_k^{2j+1,0}(b)\, dadb \\
&= \int_{-1}^{1} P_l^{0,0}(a) P_j^{0,0}(a)\, da \cdot \int_{-1}^{1} \Big(\frac{1-b}{2}\Big)^{l+j+1} P_m^{2l+1,0}(b) P_k^{2j+1,0}(b)\, db\,.
\end{aligned}$$

Für $l \neq j$ erhält man nun aufgrund der Orthogonalitätseigenschaft der Legendre-Polynome das Verschwinden des ersten Faktors und damit $(\Phi_{lm}, \Phi_{jk})_{L^2(\mathbb{T})} = 0$. Für $l = j$ ist der erste Faktor von Null verschieden und es gilt

$$(\Phi_{lm}, \Phi_{lk})_{L^2(\mathbb{T})} = \int_{-1}^{1} \big(P_l^{0,0}(a)\big)^2\, da \cdot \int_{-1}^{1} \Big(\frac{1-b}{2}\Big)^{2l+1} P_m^{2l+1,0}(b) P_k^{2l+1,0}(b)\, db\,.$$

Da Aufgrund der Orthogonalitätseigenschaft der Jacobi-Polynome, siehe (A.1), der zweite Faktor nur für $k = m$ von Null verschieden ist, folgt der erste Teil der Behauptung. Für das innere Produkt bei gleichen Indizes l, m errechnet man unter Verwendung von (A.3) den Wert

$$\gamma_{lm} = (\Phi_{lm}, \Phi_{lm})_{L^2(\mathbb{T})} = \frac{2}{2l+1} \cdot \frac{1}{2^{2l+1}} \cdot \frac{2^{2l+2}}{2m+2l+2} = \frac{2}{(2l+1)(l+m+1)}\,.$$

Die lineare Unabhängigkeit der PKD-Polynome folgt direkt aus deren Orthogonalität, denn betrachtet man eine verschwindende Linearkombination von Polynomen

$$\sum_{l,m} \alpha_{lm}\Phi_{lm} = 0\,,$$

so gilt für ein beliebiges Element (j,k) der Indexmenge

$$0 = \left(\sum_{l,m} \alpha_{lm}\Phi_{lm}, \Phi_{jk}\right)_{L^2(\mathbb{T})} = \sum_{l,m} \alpha_{lm} \left(\Phi_{lm}, \Phi_{jk}\right)_{L^2(\mathbb{T})} = \alpha_{jk}\|\Phi_{jk}\|_{L^2(\mathbb{T})}$$

und damit $\alpha_{jk} = 0$.
Zu zeigen bleibt nun $\Phi_{lm} \in \mathcal{P}^N(\mathbb{T})$ für alle Indizes $l, m \in \mathbb{N}_0$ mit $l + m \leq N$, dann folgt der zweite Teil der Behauptung aus der Gleichheit der Dimensionen. Schreibt man das Polynom $P_l^{0,0}$ in der Form $P_l^{0,0}(x) = \sum_{k=0}^{l} \alpha_k x^k$, so erhält man

$$\begin{aligned}
\Phi_{lm}(r,s) &= \sum_{k=0}^{l} \alpha_k \Big(2\frac{1+r}{1-s} - 1\Big)^k \Big(\frac{1-s}{2}\Big)^l P_m^{2l+1,0}(s) \\
&= \sum_{k=0}^{l} \alpha_k \left[\sum_{\nu=0}^{k} 2^\nu (-1)^{k-\nu} \binom{\nu}{k} \left(\frac{1+r}{1-s}\right)^\nu\right] \Big(\frac{1-s}{2}\Big)^l P_m^{2l+1,0}(s) \\
&= \sum_{k=0}^{l} \sum_{\nu=0}^{k} \alpha_k (-1)^{k-\nu} \binom{\nu}{k} (1+r)^\nu \Big(\frac{1-s}{2}\Big)^{l-\nu} P_m^{2l+1,0}(s)\,.
\end{aligned}$$

Die in der letzten Summe auftretenden Summanden sind jeweils Polynome vom Grad ν in r und vom Grad $l + m - \nu$ in s und sind daher in $\mathcal{P}^N(\mathbb{T})$ enthalten. Somit gilt auch $\Phi_{lm}(r,s) \in \mathcal{P}^N(\mathbb{T})$.

□

Dem obigen Beweis entnimmt man zudem die Eigenschaft $\Phi_{lm}(r,s) \in \mathcal{P}^{l+m}(\mathbb{T})$. Die PKD-Polynome bilden demnach eine *hierarchische* oder auch *modale* Basis, d.h. die Menge der den Raum $\mathcal{P}^N(\mathbb{T})$ aufspannenden PKD-Polynome ist in der dem Raum $\mathcal{P}^{N+1}(\mathbb{T})$ zugeordneten Basis enthalten.

3.2 Hochgenaue Quadraturformeln auf dem Dreieck

Das zur Konstruktion der PKD-Polynome verwendete zusammenfallende Koordinatensystem erlaubt zusätzlich die Herleitung von Quadraturformeln beliebig hoher Ordnung zur numerischen Integration über die Dreiecksfläche mit Hilfe eindimensionaler Quadraturformeln. Sollen hierzu für einen gegebenen Exaktheitsgrad $M \in \mathbb{N}_0$ alle Polynome in $\mathcal{P}^M(\mathbb{T})$ exakt integriert werden, so reduziert sich die Problemstellung aufgrund der Linearität des Integrals zur Forderung, dass dies für alle Basispolynome $\Phi_{lm},\ l + m \leq M$, gilt. Mit Hilfe der Transformation auf das Standardquadrat lässt sich das Integral über die Dreiecksfläche als Produkt zweier eindimensionaler Integrale schreiben, so dass sich die Umformung

$$\begin{aligned}
\int_{\mathbb{T}} \Phi_{lm}(r,s)\, dr ds &= \int_R P_l^{0,0}(a) \Big(\frac{1-b}{2}\Big)^{l+1} P_m^{2l+1,0}(b)\, da db \\
&= \int_{-1}^{1} P_l^{0,0}(a)\, da \cdot \int_{-1}^{1} \Big(\frac{1-b}{2}\Big)^{l+1} P_m^{2l+1,0}(b)\, db
\end{aligned}$$

ergibt. Unter den Quadraturformeln, die sich bei diesem Ansatz ergeben, haben nun diejenigen den höchsten Exaktheitsgrad, die durch Verwendung der Gauß-Jacobi-Quadratur (siehe A.1) zur numerischen Berechnung der eindimensionalen Integrale hervorgehen. Eine entsprechende Wahl der Stützstellen und Gewichte wird unter anderem in [52] vorgenommen. Die Verwendung der Gauß-Jacobi-Quadraturformeln ermöglicht es insbesondere, den durch die Transformation auf das Standardquadrat auftretenden zusätzlichen Faktor $\frac{1-b}{2}$ im zweiten Integral des letzten Terms über die Gauß-Jacobi-Regel zur Gewichtsfunktion $\omega_{10}(x) = (1-x)$ einzubeziehen. Sind nun durch die Stützstellen

$a_i,\ i = 0,\dots,p$, sowie $b_i,\ j = 0,\dots,q$, mit zugehörigen Gewichten $w_i^a,\ j = 0,\dots,Q_a$, und $w_j^b,\ j = 0,\dots,Q_b$, Quadraturformeln gegeben, die jeweils die Integrale $\int_{-1}^{1} p(a)da$ beziehungsweise $\int_{-1}^{1}(1-b)q(b)db$ für alle Polynome $p, q \in \mathcal{P}^M([-1,1])$ exakt berechnen, so erhält man mit den Stützstellen

$$(r_{ij}, s_j) = \Psi^{-1}(a_i, b_j), \quad i = 0,\dots,Q_a,\ j = 0,\dots,Q_b,$$

und den Gewichten

$$w_{ij} = w_i^a \cdot \frac{w_j^b}{2}, \quad i = 0,\dots,Q_a,\ j = 0,\dots,Q_b,$$

aufgrund der Beziehung

$$\begin{aligned} \sum_{\substack{i=0,\dots,Q_a \\ j=0,\dots,Q_b}} w_{ij}\Phi_{lm}(r_{ij}, s_j) &= \sum_{i=0}^{Q_a} w_i^a P_l^{0,0}(a_i) \cdot \sum_{j=0}^{Q_b} \frac{w_j^b}{2}\Big(\frac{1-b_j}{2}\Big)^l P_m^{2l+1,0}(b_j) \\ &= \int_{-1}^{1} P_l^{0,0}(a)\,da \cdot \int_{-1}^{1} \Big(\frac{1-b}{2}\Big)^{l+1} P_m^{2l+1,0}(b)\,db \end{aligned} \tag{3.3}$$

die Exaktheit der Integration für die Basisfunktionen $\Phi_{lm},\ l+m \le M$, und demnach für alle Polynome in $\mathcal{P}^M(\mathbb{T})$.
In Bezug auf den Typ der Quadraturformel stellt sich nun die Frage, inwiefern Randpunkte von $\mathbb{T}$ als Stützstellen berücksichtigt werden sollen. Während sich die minimale Stützstellenanzahl zu vorgegebenem Exaktheitsgrad ergibt, wenn keine Randpunkte gewählt werden, kann es zur Vereinfachung des Setzens von Randbedingungen dennoch sinnvoll sein, Randpunkte einzubeziehen. Es kann dann sowohl in a- als auch in b-Richtung jeweils eine Formel vom Lobatto-Typ gewählt werden, d.h. beide Randpunkte des Intervalls $[-1,1]$ sind in den Stützstellenmengen der eindimensionalen Quadraturformeln enthalten, so dass im Standarddreieck $\mathbb{T}$ die Stützstellen für $b = 1$ durch die Transformation Ψ^{-1} zur oberen Ecke zusammenfallen. Zur Vermeidung der Auswertung von Informationen an der oberen Ecke von $\mathbb{T}$, an der Ψ nicht definiert ist, ist jedoch insbesondere die Verwendung einer Radau-Formel in b-Richtung sinnvoll, wie auch in [52] vorgeschlagen wird. Werden die Stützstellen nur zum Zweck der Integration genutzt, ist diese Abwandlung allerdings nicht notwendig. Im in dieser Arbeit verwendeten numerischen Verfahren werden aufgrund der resultierenden geringstmöglichen Stützstellenanzahl nur innere Punkte von $\mathbb{T}$ als Stützstellen gewählt. In a-Richtung werden die klassischen Gauß-Punkte verwendet, während in b-Richtung Jacobi$(1,0)$-Punkte genutzt werden. Für den Exaktheitsgrad M sind dann Stützstellenanzahlen Q_a in a-Richtung und Q_b in b-Richtung von

$$Q_a = Q_b = \begin{cases} \frac{M}{2} + 1 & M \text{ gerade} \\ \frac{M+1}{2} & M \text{ ungerade} \end{cases} \tag{3.4}$$

notwendig. In Abbildung 3.2 sind diese sich durch den verallgemeinerten Tensorprodukt-Ansatz und die Gauß-Jacobi-Quadratur ergebenden Stützstellen für die Exaktheitsgrade $M = 0,\dots,3$ und $M = 6,7$ skizziert.
Die so konstruierten Quadraturformeln zur Integration über die Dreiecksfläche besitzen den ästhetischen Nachteil, dass sie nicht invariant unter den Symmetrien des Dreiecks

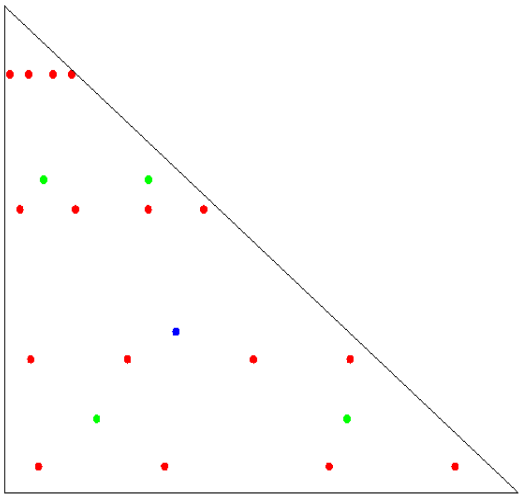


Abb. 3.2: Stützstellenwahl für die Exaktheitsgrade $M = 0, 1$ (blau), $M = 2, 3$ (grün) und $M = 6, 7$ (rot).

sind und sich an der oberen Ecke verdichten. Vom Standpunkt der Implementierung ergibt sich jedoch der Vorteil der möglichen Reduktion des Rechenaufwands durch Ausnutzung der Tensorproduktstruktur, wie nachfolgend erläutert werden soll. Zunächst sei noch bemerkt, dass Quadraturformeln auf dem Dreieck mit minimaler Stützstellenanzahl, die bezüglich der Effizienz der numerischen Integration prinzipiell ebenso von Interesse sind, zum einen nicht zu jedem Exaktheitsgrad bekannt sind und zum anderen negative Gewichte oder Stützstellen außerhalb des Integrationsgebiets enthalten können. Für einen diesbezüglichen Überblick der derzeit bekannten Quadraturformeln auf verschiedenen Integrationsgebieten sei auf die elektronische Bibliothek [24] verwiesen. In aktuellen Arbeiten zur numerischen Integration über Simplizes im Kontext von Finite-Elemente- und Spektral-Elemente-Verfahren, siehe [90, 101], werden Quadraturformeln mit möglichst geringer Stützstellenanzahl berechnet, die nur Stützstellen im Inneren des Gebiets und nur positive Gewichte besitzen. Hierbei werden zum Teil nur symmetrische Formeln betrachtet. Für hohe Polynomgrade wird die mögliche Beschleunigung der Auswertung bei Tensorprodukt-Quadraturformeln durch die etwas geringere Stützstellenanzahl der Formeln, die keine derartige Struktur aufweisen, allerdings nicht aufgewogen.

Verringerung des Rechenaufwands durch Ausnutzung der Tensorproduktstruktur Die Beschleunigung der Auswertung von Quadraturformeln mit Tensorproduktstruktur lässt sich nicht bei der Berechnung eines einzelnen Integrals erreichen. Vielmehr muss bei den in dieser Arbeit betrachteten numerischen Verfahren häufig zwischen der Repräsentation eines Polynoms auf $\mathbb{T}$ in den Koeffizienten und den Werten an den Stützstellen gewechselt werden, so dass hier eine effiziente Berechnung der jeweiligen Daten gewünscht ist. Sollen demnach für eine Funktion $u : \mathbb{T} \to \mathbb{R}$ die Koeffizienten

$$\hat{u}_{lm} = \frac{(u, \Phi_{lm})_{L^2(\mathbb{T})}}{(\Phi_{lm}, \Phi_{lm})_{L^2(\mathbb{T})}} = \frac{1}{\gamma_{lm}} (u, \Phi_{lm})_{L^2(\mathbb{T})}$$

numerisch berechnet werden, so sind zunächst $\frac{(N+1)(N+2)}{2}$ Summationen der Form

$$\gamma_{lm}\hat{u}_{lm} = \sum_{\substack{i=0,\dots,Q_a \\ j=0,\dots,Q_b}} w_{ij}\, u(r_{ij}, s_j)\Phi_{lm}(r_{ij}, s_j)$$

auszuführen, so dass sich ein Aufwand von $\mathcal{O}(N^2 Q_a Q_b)$ ergibt. Schreibt man die obige Summe um, wobei neben der Tensorprodukteigenschaft der Quadraturformel ebenso die verallgemeinerte Tensorprodukteigenschaft der PKD-Basis ausgenutzt wird, so erhält man

$$\gamma_{lm}\hat{u}_{lm} = \sum_{j=0}^{Q_b} \left[\sum_{i=0}^{Q_a} w_i^a u(r_{ij}, s_j) \overset{1}{\Phi}_l\,(a_i) \right] \frac{w_j^b}{2} \overset{2}{\Phi}_{lm}\,(b_j),$$

so dass sich eine effizientere Auswertung ergibt, wenn zunächst $(Q_b+1)(N+1)$ Summationen der Form

$$S(j,l) = \sum_{i=0}^{Q_a} w_i^a u(r_{ij}, s_j) \overset{1}{\Phi}_l\,(a_i)$$

und anschließend $\frac{(N+1)(N+2)}{2}$ Summationen der Form

$$\gamma_{lm}\hat{u}_{lm} = \sum_{j=0}^{Q_b} S(j,l) \frac{w_j^b}{2} \overset{2}{\Phi}_{lm}\,(b_j)$$

ausgeführt werden, mit dem Gesamtaufwand $\mathcal{O}(NQ_aQ_b + N^2Q_b)$.
Im Verfahren selbst ist tatsächlich die Auswertung von Summen der Form

$$\sum_{\substack{i=0,\dots,Q_a \\ j=0,\dots,Q_b}} w_{ij}\, v(r_{ij}, s_j) \frac{\partial}{\partial r}\Phi_{lm}(r_{ij}, s_j) \quad \text{sowie} \quad \sum_{\substack{i=0,\dots,Q_a \\ j=0,\dots,Q_b}} w_{ij}\, v(r_{ij}, s_j) \frac{\partial}{\partial s}\Phi_{lm}(r_{ij}, s_j)$$

notwendig. Der obige Ansatz der gestaffelten Addition lässt sich übertragen, wenn die Ableitungen der PKD-Polynome als Produkte eindimensionaler Polynome beziehungsweise als Summe von Produkten eindimensionaler Polynome dargestellt werden können. Eine derartige Darstellung ist gegeben durch

$$\frac{\partial}{\partial r}\Phi_{lm}(r_{ij}, s_j) = \begin{cases} 0, & l = 0, \\ \frac{d}{da}P_l^{0,0}(a_i) \cdot \left(\frac{1-b_j}{2}\right)^{l-1} P_m^{2l+1,0}(b_j), & l > 0, \end{cases}$$

sowie

$$\frac{\partial}{\partial s}\Phi_{lm}(r_{ij}, s_j) = \begin{cases} \frac{d}{db}P_m^{1,0}(b_j), & l = 0, \\ \frac{1}{2}\frac{d}{da}P_l^{0,0}(a_i)(a_i+1) \cdot \left(\frac{1-b_j}{2}\right)^{l-1} P_m^{2l+1,0}(b_j) & \\ \quad + P_l^{0,0}(a_i) \cdot \left[\left(\frac{1-b_j}{2}\right)^{l-1} \left(-\frac{l}{2}P_m^{2l+1,0}(b_j) + \frac{1-b_j}{2}\frac{d}{ds}P_m^{2l+1,0}(b_j) \right) \right], & l > 0. \end{cases}$$

Bei der Auswertung einer polynomialen Darstellung von u an allen Stützstellen kann analog zur Berechnung von Koeffizienten aus Werten an Stützstellen vorgegangen werden. Eine entsprechende Umformung ergibt

$$u(r_{ij}, s_j) = \sum_{0 \le l+m \le N} \hat{u}_{lm} \Phi_{lm}(r_{ij}, s_j) = \sum_{l=0}^{N} \left[\sum_{m=0}^{N-l} \hat{u}_{lm} \overset{2}{\Phi}_{lm}(b_j) \right] \overset{1}{\Phi}_l(a_i),$$

so dass, wie im obigen Fall, durch vorherige Berechnung und Abspeicherung der inneren Summen die Auswertung beschleunigt werden kann.

3.3 Sturm-Liouville-Gleichung und Approximationseigenschaften der Proriol-Koornwinder-Dubiner-Polynome

Im Kontext spektraler Verfahren liegt das Interesse an Sturm-Liouville-Problemen darin begründet, dass die Entwicklung einer glatten Funktion in Eigenfunktionen eines derartigen Problems bei Festlegung geeigneter Randbedingungen beziehungsweise im Fall singulärer Sturm-Liouville-Probleme sogenannte spektrale Genauigkeit aufweist. Dementsprechend sind spektrale Verfahren so konzipiert, dass sie eine Approximation an die Entwicklung der Lösung einer gegebenen Differentialgleichung in Eigenfunktionen eines Sturm-Liouville-Operators liefern. Für die PKD-Polynome wurde unabhängig voneinander von Warburton [93] sowie von Wingate und Taylor [98] ein zugehöriges singuläres Sturm-Liouville-Problem hergeleitet.
Die klassische Definition eines *Sturm-Liouville-Problems* ist gegeben durch eine Differentialgleichung der Form

$$-(p(x)u'(x))' + q(x)u(x) = \lambda w(x)u(x), \qquad x \in (-1, 1), \tag{3.5}$$

mit geeigneten Randbedingungen für u. An die Koeffizientenfunktionen $p, q, w : (-1, 1) \to \mathbb{R}$ werden hierbei die Bedingungen gestellt, dass p stetig differenzierbar, strikt positiv und an ± 1 stetig fortsetzbar ist, dass q und w stetig und nicht negativ sind, und dass q beschränkt und w integrierbar ist. Das Sturm-Liouville-Problem wird als *singulär* bezeichnet, wenn p an mindestens einem der Randpunkte verschwindet.
Die PKD-Polynome sind Lösungen eines auf das Standarddreieck angepassten Sturm-Liouville-Problems, wie nachfolgend gezeigt werden soll. In den Koordinaten des Referenzdreiecks $\mathbb{T}$ ist die zugehörige Gleichung gegeben durch

$$\mathcal{L}_{r,s}\, \Phi(r, s) + \lambda\, \Phi(r, s) = 0\,, \quad (r, s) \in \mathbb{T}, \tag{3.6}$$

mit dem Differentialoperator

$$\begin{aligned} \mathcal{L}_{r,s} \;=\; & \frac{\partial}{\partial r}\left((1+r)\left[(1-r)\frac{\partial}{\partial r} - (1+s)\frac{\partial}{\partial s} \right] \right) \\ & + \frac{\partial}{\partial s}\left((1+s)\left[(1-s)\frac{\partial}{\partial s} - (1+r)\frac{\partial}{\partial r} \right] \right). \end{aligned} \tag{3.7}$$

Mit der Definition der von den Dreieckskoordinaten abhängigen symmetrischen Matrix

$$\mathbf{B}(r, s) = \begin{pmatrix} 1 - r^2 & -(1+r)(1+s) \\ -(1+r)(1+s) & 1 - s^2 \end{pmatrix},$$

sowie dem Operator

$$\nabla_{r,s} = \left(\frac{\partial}{\partial r}, \frac{\partial}{\partial s}\right)^T$$

erhält man zunächst $\mathcal{L}_{r,s} = \nabla_{r,s} \cdot \mathbf{B}(r,s)\nabla_{r,s}$. Analog zur strikten Positivität der Koeffizientenfunktion p in der klassischen Sturm-Liouville-Gleichung (3.5) ist die Matrix $\mathbf{B}$ im Inneren des Referenzdreiecks $\mathbb{T}$ positiv definit, da sich $\mathbf{B}$ für $r \neq -1$ mit Hilfe des symmetrischen Gauß-Algorithmus in die Diagonalmatrix

$$\begin{pmatrix} 1-r^2 & 0 \\ 0 & -2(r+s)\frac{1+s}{1-r} \end{pmatrix}$$

überführen lässt, die positive Diagonaleinträge im Inneren von $\mathbb{T}$ besitzt.
Es gelten desweiteren die folgende Aussagen.

Satz 3.2 *1. Die Eigenwerte des Sturm-Liouville-Operators $\mathcal{L}_{r,s}$ sind gegeben durch $-\lambda_{lm} = -(l+m)(l+m+2)$ mit den PKD-Polynomen $\Phi_{lm}(r,s)$ als zugehörigen Eigenfunktionen.*

2. Das Problem (3.6), (3.7) ist singulär in dem Sinn, dass die Normalenvektoren $\mathbf{n}_i$, $i = 1,2,3$, des Standarddreiecks $\mathbb{T}$ die Gleichung

$$\mathbf{B}(r,s) \cdot \mathbf{n}_i = \mathbf{0}, \quad i = 1,2,3,$$

erfüllen.

3. Der Operator $\mathcal{L}_{r,s}$ ist selbstadjungiert, d.h. es gilt

$$(u, \mathcal{L}_{r,s} v)_{L^2(\mathbb{T})} = (\mathcal{L}_{r,s} u, v)_{L^2(\mathbb{T})}, \qquad u, v \in H^2(\mathbb{T}). \tag{3.8}$$

Beweis:

1. Wir zeigen zunächst, dass die Polynome

$$\tilde{\Phi}_{lm}(a,b) := \Phi_{lm} \circ \Psi^{-1}(a,b) = P_l^{0,0}(a)\Big(\frac{1-b}{2}\Big)^l P_m^{2l+1,0}(b)$$

die Gleichung

$$\frac{2}{1-b}\left\{\frac{\partial}{\partial a}\left[(1-a^2)\frac{\partial \tilde{\Phi}_{lm}}{\partial a}\right] + \frac{\partial}{\partial b}\left[(1-b^2)\frac{1-b}{2}\frac{\partial \tilde{\Phi}_{lm}}{\partial b}\right]\right\} = -\lambda_{lm}\tilde{\Phi}_{lm} \tag{3.9}$$

erfüllen. Für den zweiten Summanden in Gleichung (3.9) gilt

$$\begin{aligned}
&\frac{\partial}{\partial b}\left[(1-b^2)\frac{1-b}{2}\frac{\partial\tilde{\Phi}_{lm}}{\partial b}\right] \\
&= P_l^{0,0}(a)\cdot\frac{\partial}{\partial b}\left[(1-b^2)\frac{1-b}{2}\frac{\partial}{\partial b}\left[\left(\frac{1-b}{2}\right)^l P_m^{2l+1,0}(b)\right]\right] \\
&= P_l^{0,0}(a)\cdot\frac{\partial}{\partial b}\left[-l(1+b)\left(\frac{1-b}{2}\right)^{l+1}P_m^{2l+1,0}(b)\right. \\
&\qquad\qquad \left. +(1-b^2)\left(\frac{1-b}{2}\right)^{l+1}\frac{\partial}{\partial b}P_m^{2l+1,0}(b)\right] \\
&= P_l^{0,0}(a)\cdot\left(\frac{\partial}{\partial b}\left[-l(1+b)\left(\frac{1-b}{2}\right)^{l+1}\right]\cdot P_m^{2l+1,0}(b)\right. \\
&\qquad\qquad -l(1+b)\left(\frac{1-b}{2}\right)^{l+1}\frac{\partial}{\partial b}P_m^{2l+1,0}(b) \\
&\qquad\qquad \left.+\frac{\partial}{\partial b}\left[(1-b^2)\left(\frac{1-b}{2}\right)^{l+1}\frac{\partial}{\partial b}P_m^{2l+1,0}(b)\right]\right). \qquad (3.10)
\end{aligned}$$

Die Sturm-Liouville-Gleichung für die Jacobi-Polynome $P_m^{2l+1,0}$ hat die Form

$$\frac{\partial}{\partial b}\left[(1-b)^{2l+2}(1+b)\frac{\partial}{\partial b}P_m^{2l+1,0}(b)\right] = -m(m+2l+2)(1-b)^{2l+1}P_m^{2l+1,0}(b)\,,$$

siehe auch A.1. Andererseits lässt sich die linke Seite der obigen Gleichung unter Verwendung der Produktregel umformen zu

$$\begin{aligned}
\frac{\partial}{\partial b}\left[(1-b)^{2l+2}(1+b)\frac{\partial}{\partial b}P_m^{2l+1,0}(b)\right] &= (1-b)^l\cdot\frac{\partial}{\partial b}\left[(1-b)^{l+2}(1+b)\frac{\partial}{\partial b}P_m^{2l+1,0}(b)\right] \\
&\quad -l(1-b)^{2l+1}(1+b)\frac{\partial}{\partial b}P_m^{2l+1,0}(b)\,,
\end{aligned}$$

wodurch wir

$$\begin{aligned}
&\frac{\partial}{\partial b}\left[(1-b)^{l+2}(1+b)\frac{\partial}{\partial b}P_m^{2l+1,0}(b)\right]-l(1-b)^{l+1}(1+b)\frac{\partial}{\partial b}P_m^{2l+1,0}(b) \\
&\quad = -m(m+2l+2)(1-b)^{l+1}P_m^{2l+1,0}(b)
\end{aligned}$$

folgern können. Einsetzen in (3.10) ergibt nun

$$\begin{aligned}
&\frac{\partial}{\partial b}\left[(1-b^2)\frac{1-b}{2}\frac{\partial\tilde{\Phi}_{lm}}{\partial b}\right] \\
&= P_l^{0,0}(a)P_m^{2l+1,0}(b)\left(\frac{\partial}{\partial b}\left[-l(1+b)\left(\frac{1-b}{2}\right)^{l+1}\right]-\,m(m+2l+2)\left(\frac{1-b}{2}\right)^{l+1}\right).
\end{aligned}$$

Mit

$$\frac{\partial}{\partial b}\left[-l(1+b)\left(\frac{1-b}{2}\right)^{l+1}\right] = -l\cdot\left(\frac{1-b}{2}\right)^{l+1} + \frac{l}{2}(l+1)(1+b)\left(\frac{1-b}{2}\right)^{l}$$

erhält man schließlich

$$\begin{aligned}&\frac{\partial}{\partial b}\left[(1-b^2)\frac{1-b}{2}\frac{\partial\tilde{\Phi}_{lm}}{\partial b}\right]\\ &\quad= \frac{1}{2}\left(-l(1-b)+l(l+1)(1+b)-m(m+2l+2)(1-b)\right)\cdot\tilde{\Phi}_{lm}\,.\end{aligned}$$

Mit der zu den Legendre-Polynomen gehörigen Sturm-Liouville-Gleichung schreibt sich der erste Summand in (3.9) als

$$\frac{\partial}{\partial a}\left[(1-a^2)\frac{\partial\tilde{\Phi}_{lm}}{\partial a}\right] = -l(l+1)\tilde{\Phi}_{lm}\,.$$

Die Terme in l, m und b lassen sich dann zusammenfassen zu

$$\frac{\partial}{\partial a}\left[(1-a^2)\frac{\partial\tilde{\Phi}_{lm}}{\partial a}\right] + \frac{\partial}{\partial b}\left[(1-b^2)\frac{1-b}{2}\frac{\partial\tilde{\Phi}_{lm}}{\partial b}\right] = -\frac{1-b}{2}(l+m)(l+m+2)\tilde{\Phi}_{lm}.$$

Nach Multiplikation mit dem Vorfaktor $\frac{2}{1-b}$ ergibt sich dadurch die Gültigkeit der Gleichung (3.9).

Im nächsten Schritt formen wir den in (3.9) gegebenen Differentialoperator in den Variablen a und b zu dem in den Dreieckskoordinaten r und s gegebenen Operator $\mathcal{L}_{r,s}$ um. Mit der durch $a = 2\frac{1+r}{1-s} - 1,\ b = s$ gegebenen Transformation Ψ sowie

$$\begin{aligned}\frac{\partial\tilde{\Phi}_{lm}}{\partial a}\circ\Psi &= \frac{1-s}{2}\cdot\frac{\partial\Phi_{lm}}{\partial r}\,,\\ \frac{\partial\tilde{\Phi}_{lm}}{\partial b}\circ\Psi &= -\frac{1+r}{1-s}\cdot\frac{\partial\Phi_{lm}}{\partial r} + \frac{\partial\Phi_{lm}}{\partial s}\end{aligned}$$

ergibt sich

$$\begin{aligned}
\left(\tilde{\mathcal{L}}_{a,b}\tilde{\Phi}_{lm}\right)\circ\Psi &:= \frac{2}{1-b}\left\{\frac{\partial}{\partial a}\left[(1-a^2)\frac{\partial\tilde{\Phi}_{lm}}{\partial a}\right]+\frac{\partial}{\partial b}\left[(1-b^2)\frac{1-b}{2}\frac{\partial\tilde{\Phi}_{lm}}{\partial b}\right]\right\}\circ\Psi \\
&= \frac{2}{1-s}\left\{\frac{1-s}{2}\frac{\partial}{\partial r}\left[\left(1-\left(2\frac{1+r}{1-s}-1\right)^2\right)\frac{1-s}{2}\frac{\partial\Phi_{lm}}{\partial r}\right]\right. \\
&\quad \left.+\left(-\frac{1+r}{1-s}\frac{\partial}{\partial r}+\frac{\partial}{\partial s}\right)\left[(1-s^2)\frac{1-s}{2}\left(-\frac{1+r}{1-s}\frac{\partial\Phi_{lm}}{\partial r}+\frac{\partial\Phi_{lm}}{\partial s}\right)\right]\right\} \\
&= \frac{\partial}{\partial r}\left[\left(-\frac{2(1+r)^2}{1-s}+2(1+r)\right)\frac{\partial\Phi_{lm}}{\partial r}\right] \\
&\quad +\left(-\frac{2(1+r)}{(1-s)^2}\frac{\partial}{\partial r}+\frac{2}{1-s}\frac{\partial}{\partial s}\right)\left[-\frac{(1-s^2)(1+r)}{2}\frac{\partial\Phi_{lm}}{\partial r}\right. \\
&\qquad \left.+\frac{(1-s^2)(1-s)}{2}\frac{\partial\Phi_{lm}}{\partial s}\right] \\
&= \frac{\partial}{\partial r}\left[\left(-\frac{2(1+r)^2}{1-s}+2(1+r)\right)\frac{\partial\Phi_{lm}}{\partial r}\right] \\
&\quad +(1+r)\frac{\partial}{\partial r}\left[(1+r)\frac{1+s}{1-s}\frac{\partial\Phi_{lm}}{\partial r}-(1+s)\frac{\partial\Phi_{lm}}{\partial s}\right] \\
&\quad +\frac{2}{1-s}\frac{\partial}{\partial s}\left[\left(\frac{1-s^2}{2}\right)\left(-(1+r)\frac{\partial\Phi_{lm}}{\partial r}+(1-s)\frac{\partial\Phi_{lm}}{\partial s}\right)\right].
\end{aligned}$$

Durch die Verwendung der Gleichungen

$$(1+r)\frac{\partial}{\partial r}\left(\frac{1}{1+r}\cdot v\right)=\frac{\partial v}{\partial r}-\frac{v}{1+r}\quad\text{und}\quad\frac{2}{1-s}\frac{\partial}{\partial s}\left(\frac{1-s}{2}\cdot w\right)=\frac{\partial w}{\partial s}-\frac{w}{1-s}$$

für die Funktionen

$$\begin{aligned}
v &= (1+r)^2\frac{1+s}{1-s}\frac{\partial\Phi_{lm}}{\partial r}-(1+r)(1+s)\frac{\partial\Phi_{lm}}{\partial s}, \\
w &= -(1+r)(1+s)\frac{\partial\Phi_{lm}}{\partial r}+(1-s^2)\frac{\partial\Phi_{lm}}{\partial s}
\end{aligned}$$

und das Ausnutzen der Beziehung $v/(1+r)=-w/(1-s)$ ergibt sich daraus

$$\begin{aligned}
\left(\tilde{\mathcal{L}}_{a,b}\tilde{\Phi}_{lm}\right)\circ\Psi &= \frac{\partial}{\partial r}\left[\left(-\frac{2(1+r)^2}{1-s}+2(1+r)\right)\frac{\partial\Phi_{lm}}{\partial r}\right] \\
&\quad +\frac{\partial}{\partial r}\left[(1+r)^2\frac{1+s}{1-s}\frac{\partial\Phi_{lm}}{\partial r}-(1+r)(1+s)\frac{\partial\Phi_{lm}}{\partial s}\right] \\
&\quad +\frac{\partial}{\partial s}\left[-(1+s)(1+r)\frac{\partial\Phi_{lm}}{\partial r}+(1-s^2)\frac{\partial\Phi_{lm}}{\partial s}\right].
\end{aligned}$$

Aus der Gleichung

$$\begin{aligned}
&\frac{\partial}{\partial r}\left[\left(-\frac{2(1+r)^2}{1-s}+2(1+r)+(1+r)^2\frac{1+s}{1-s}\right)\frac{\partial\Phi_{lm}}{\partial r}\right] \\
&\quad = \frac{\partial}{\partial r}\left[(1+r)\frac{1-r-s+rs}{1-s}\frac{\partial\Phi_{lm}}{\partial r}\right] = \frac{\partial}{\partial r}\left[(1-r^2)\frac{\partial\Phi_{lm}}{\partial r}\right]
\end{aligned}$$

erhält man daher schließlich

$$\left(\tilde{\mathcal{L}}_{a,b}\tilde{\Phi}_{lm}\right)\circ\Psi \;=\; \mathcal{L}_{r,s}\Phi_{lm}$$

und damit die Behauptung.

2. Man zerlege den Rand des Standarddreiecks in die Vereinigung der drei Kanten $\partial\mathbb{T} = \kappa_1 \cup \kappa_2 \cup \kappa_3$ mit

$$\begin{aligned}\kappa_1 &= \{(r,-1)\in\mathbb{R}^2 \mid -1\le r\le 1\},\\ \kappa_2 &= \{(r,-r)\in\mathbb{R}^2 \mid -1\le r\le 1\},\\ \kappa_3 &= \{(-1,s)\in\mathbb{R}^2 \mid -1\le s\le 1\}.\end{aligned}$$

Die zugehörigen Normalenvektoren sind $\mathbf{n}_1 = (0,-1)^T$, $\mathbf{n}_2 = \frac{1}{\sqrt{2}}(1,1)^T$ sowie $\mathbf{n}_3 = (-1,0)^T$. Die Vektoren $\mathbf{n}_i$ liegen wegen $s=-1$ für $i=1$, $s=-r$ für $i=2$ bzw. $r=-1$ für $i=3$ jeweils offensichtlich im Kern der Matrix $\mathbf{B}(r,s)$.

3. Durch zweimalige Verwendung der Greenschen Integralformel und unter Ausnutzung der Symmetrie von $\mathbf{B} = \mathbf{B}(r,s)$ ergibt sich

$$\begin{aligned}\int_{\mathbb{T}} u\,\mathcal{L}_{r,s}v\,drds &= \int_{\mathbb{T}} u\,\nabla\cdot\mathbf{B}\nabla v\,drds = -\int_{\mathbb{T}}\nabla u\cdot\mathbf{B}\nabla v\,drds + \int_{\partial\mathbb{T}} u\,\mathbf{B}\nabla v\cdot\mathbf{n}\,d\sigma\\ &= -\int_{\mathbb{T}}\mathbf{B}^T\nabla u\cdot\nabla v\,drds + \int_{\partial\mathbb{T}} u\,\mathbf{B}\nabla v\cdot\mathbf{n}\,d\sigma\\ &= \int_{\mathbb{T}}\nabla\cdot\mathbf{B}\nabla u\,v\,drds - \int_{\partial\mathbb{T}}\mathbf{B}\nabla u\cdot\mathbf{n}\,v\,d\sigma + \int_{\partial\mathbb{T}} u\,\mathbf{B}\nabla v\cdot\mathbf{n}\,d\sigma.\end{aligned}$$

Entfallen die Randterme auf der rechten Seite der obigen Gleichung, so ist die Behauptung gezeigt. Dazu liefert die zweite Aussage des Lemmas

$$\mathbf{B}\nabla u\cdot\mathbf{n}_i = \nabla u^T\mathbf{B}\mathbf{n}_i = 0\,,$$

und damit wie gewünscht

$$\int_{\mathbb{T}} u\,\mathcal{L}_{r,s}v\,drds = \int_{\mathbb{T}}\nabla\cdot\mathbf{B}\nabla u\,v\,drds = \int_{\mathbb{T}}\mathcal{L}_{r,s}u\,v\,drds\,.$$

□

Aus der Eigenschaft der PKD-Polynome, als Eigenfunktionen des selbstadjungierten Operators $\mathcal{L}_{r,s}$ aufzutreten, mit den zugehörigen Eigenwerten $-\lambda_{lm} = \mathcal{O}((l+m)^2)$, $l+m\to\infty$, lassen sich die folgenden asymptotischen Aussagen für PKD-Entwicklungen hinreichend glatter Funktionen herleiten.

Satz 3.3 *Sei $u\in H^{2k}(\mathbb{T})$, $k\in\mathbb{N}$. Für die abgeschnittene PKD-Entwicklung von u, gegeben durch*

$$\mathcal{P}_N u = \sum_{l+m\le N}\hat{u}_{lm}\Phi_{lm}, \qquad \hat{u}_{lm} = \frac{(u,\Phi_{lm})_{L^2(\mathbb{T})}}{\gamma_{lm}},$$

mit γ_{lm} aus (3.2), gelten die Abschätzungen

$$\sqrt{\gamma_{lm}}\,|\hat{u}_{lm}| = \mathcal{O}((l+m)^{-2k}), \quad l+m \to \infty, \tag{3.11}$$
$$\|u - \mathcal{P}_N u\|_{L^2(\mathbb{T})} = \mathcal{O}(N^{-2k}), \quad N \to \infty. \tag{3.12}$$

Für $k > 1$ gilt zusätzlich die punktweise Abschätzung

$$|u(r,s) - \mathcal{P}_N u(r,s)| = \mathcal{O}(N^{-2k+2}), \qquad (r,s) \in \mathbb{T}. \tag{3.13}$$

Bemerkung 3.4 *Im Sinne der den Sobolevschen Einbettungssätzen [16, S. 114] zu entnehmenden Einbettung $H^2(\mathbb{T}) \subset C^0(\mathbb{T})$ ist in jeder Äquivalenzklasse im Sobolev-Raum $H^2(\mathbb{T})$ eine stetige Funktion in $C^0(\mathbb{T})$ enthalten. Dementsprechend bezieht sich die punktweise Eigenschaft (3.13) auf den entsprechenden stetigen Repräsentanten von u.*

Zum Beweis von (3.13) müssen zunächst die Werte der PKD-Polynome $\Phi_{lm}(r,s)$ für $l, m \to \infty$ abgeschätzt werden. Wir benötigen daher ein vorbereitendes Lemma.

Lemma 3.5 *Für $(r,s) \in \mathbb{T}\backslash\{(-1,1)\}$ und für alle $l, m \in \mathbb{N}$ gilt die Abschätzung*

$$|\Phi_{lm}(r,s)| \leq \left(\frac{2}{1-s}\right)^{3/4}. \tag{3.14}$$

An der oberen Ecke $(r,s) = (-1,1)$ des Standarddreiecks besitzen die PKD-Polynome die Werte

$$\Phi_{lm}(-1,1) = 0, \qquad l > 0, \tag{3.15}$$
$$\Phi_{0m}(-1,1) = m+1. \tag{3.16}$$

Beweis: Da die Legendre Polynome durch $\max\limits_{-1\leq x\leq 1} |P_l^{0,0}(x)| = P_l^{0,0}(1) = 1$ gleichmäßig beschränkt sind, siehe auch (A.4), ergibt sich für einen beliebigen festen Punkt $(r,s) \in \mathbb{T}$ die obere Schranke

$$\begin{aligned} |\Phi_{lm}(r,s)| &= \left| P_l^{0,0}\left(2\frac{1+r}{1-s} - 1\right)\left(\frac{1-s}{2}\right)^l P_m^{2l+1,0}(s)\right| \\ &\leq \left|\left(\frac{1-s}{2}\right)^l P_m^{2l+1,0}(s)\right|. \end{aligned} \tag{3.17}$$

Mit der aus [86, S. 164] entnommenen Abschätzung

$$\left(\frac{1-s}{2}\right)^{\frac{\alpha}{2}+\frac{1}{4}} |P_m^{\alpha,0}(s)| \leq 1, \qquad s \in [-1,1],\ \alpha \geq -\frac{1}{2},$$

erhält man daher die Ungleichung (3.14). Desweiteren liefert (3.17) für $(r,s) = (-1,1)$ die in (3.15) angegebenen Werte, während sich die Werte in (3.16) wegen $P_m^{1,0}(1) = \binom{m+1}{m} = m+1$ direkt aus der Definition von Φ_{0m} ergeben. □

Beweis von Satz 3.3:

1. Durch mehrmalige Anwendung der Gleichung (3.8) und mit $\gamma_{lm} = \|\Phi_{lm}\|^2_{L^2(\mathbb{T})}$ ergibt sich für $l+m \neq 0$ die Gleichung

$$\begin{aligned}\gamma_{lm} \cdot \hat{u}_{lm} &= (u,\, \Phi_{lm})_{L^2(\mathbb{T})} = \left(u, \frac{1}{-\lambda_{lm}} \mathcal{L}_{r,s}\, \Phi_{lm}\right)_{L^2(\mathbb{T})} = \frac{1}{-\lambda_{lm}} \left(\mathcal{L}_{r,s} u, \Phi_{lm}\right)_{L^2(\mathbb{T})} \\ &= \frac{1}{(-\lambda_{lm})^k} \left(\mathcal{L}^k_{r,s} u, \Phi_{lm}\right)_{L^2(\mathbb{T})}. \end{aligned} \tag{3.18}$$

   Mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\left|\left(\mathcal{L}^k_{r,s} u, \Phi_{lm}\right)_{L^2(\mathbb{T})}\right| \leq \|\mathcal{L}^k_{r,s} u\|_{L^2(\mathbb{T})} \|\Phi_{lm}\|_{L^2(\mathbb{T})}$$

   erhält man daher für jedes fest gewählte $k \in \mathbb{N}$ die Abschätzung (3.11) mit

$$\sqrt{\gamma_{lm}}\, |\hat{u}_{lm}| \leq \frac{1}{\lambda^k_{lm}} \left\|\mathcal{L}^k_{r,s} u\right\|_{L^2(\mathbb{T})} = \mathcal{O}\left((l+m)^{-2k}\right), \quad l+m \to \infty.$$

2. Da der Raum der polynomialen Funktionen auf $\mathbb{T}$ dicht in $L^2(\mathbb{T})$ liegt, bilden die PKD-Polynome ein vollständiges Orthogonalsystem von $L^2(\mathbb{T})$, so dass die *Parsevalsche Gleichung*

$$\|v\|^2_{L^2(\mathbb{T})} = \sum_{l,m \in \mathbb{N}_0} \frac{1}{\gamma_{lm}} \left|(v, \Phi_{lm})_{L^2(\mathbb{T})}\right|^2, \tag{3.19}$$

   für alle $v \in L^2(\mathbb{T})$ Gültigkeit hat. Für den Abschneidefehler erhält man unter Verwendung der Gleichung (3.18) daher

$$\begin{aligned}\|u - \mathcal{P}_N u\|^2_{L^2(\mathbb{T})} &= \sum_{l+m>N} \gamma_{lm} \cdot \hat{u}^2_{lm} = \sum_{l+m>N} \frac{1}{\lambda^{2k}_{lm} \cdot \gamma_{lm}} \left|\left(\mathcal{L}^k_{r,s} u, \Phi_{lm}\right)_{L^2(\mathbb{T})}\right|^2 \\ &\leq \frac{1}{\lambda^{2k}_{N+1}} \left\|\mathcal{L}^k_{r,s} u\right\|^2_{L^2(\mathbb{T})}, \end{aligned}$$

   wobei die Notation $\lambda_{N+1} = (N+1)(N+3)$ verwendet wurde und sich die letzte Ungleichung aus der erneuten Anwendung von (3.19) auf $\mathcal{L}^k_{r,s} u$ ergibt. Aus der Abschätzung

$$\left\|\mathcal{L}^k_{r,s} u\right\|^2_{L^2(\mathbb{T})} \leq \tilde{C}_k \|u\|^2_{H^{2k}(\mathbb{T})}$$

   mit einer von u unabhängigen Konstanten $\tilde{C}_k > 0$ erhält man

$$\|u - \mathcal{P}_N u\|^2_{L^2(\mathbb{T})} \leq \tilde{C}_k [(N+1)(N+3))]^{-2k} \|u\|^2_{H^{2k}(\mathbb{T})}$$

   und daher (3.12) mit

$$\|u - \mathcal{P}_N u\|_{L^2(\mathbb{T})} \leq C_k N^{-2k} \|u\|_{H^{2k}(\mathbb{T})} = \mathcal{O}(N^{-2k}), \quad N \to \infty.$$

3. Zunächst soll für eine Funktion $u \in H^4(\mathbb{T})$ die punktweise Konvergenz der PKD-Reihe gegen u, d.h. die Gültigkeit der Gleichung

$$u(r,s) \;=\; \sum_{l,m\in\mathbb{N}_0} \hat{u}_{lm}\Phi_{lm}(r,s), \qquad \forall (r,s)\in\mathbb{T}, \tag{3.20}$$

gezeigt werden. Mit $|\hat{u}_{lm}| \le \frac{1}{\sqrt{\gamma_{lm}}\lambda_{lm}^2}\left\|\mathcal{L}_{r,s}^2 u\right\|_{L^2(\mathbb{T})}$ für $l+m\neq 0$ sowie $\sqrt{\gamma_{lm}} < l+m+2$ gilt zunächst für alle $N\in\mathbb{N}$ die Abschätzung

$$\begin{aligned}\sum_{0<l+m\le N} |\hat{u}_{lm}| \;&<\; \sum_{0<l+m\le N} (l+m)^{-2}(l+m+2)^{-1}\left\|\mathcal{L}_{r,s}^2 u\right\|_{L^2(\mathbb{T})} \\ &<\; \left\|\mathcal{L}_{r,s}^2 u\right\|_{L^2(\mathbb{T})}\cdot\sum_{k\in\mathbb{N}} k^{-2}\,.\end{aligned}$$

Zudem sind die Werte der PKD-Polynome auf jeder Teilmenge der Form $\Omega_{\tilde{s}} = \{(r,s)\in\mathbb{T} \mid s\le\tilde{s}\}\subset\mathbb{T}$ nach (3.14) durch

$$|\Phi_{lm}(r,s)| \;\le\; \left(\frac{2}{1-\tilde{s}}\right)^{3/4}, \qquad \forall\, l,m\in\mathbb{N}_0,\ \forall\,(r,s)\in\Omega_{\tilde{s}}$$

beschränkt. Somit ist die PKD-Reihe auf $\Omega_{\tilde{s}}$ normal konvergent. Dies bedeutet, dass die Reihe reeller Zahlen $\sum_{l,m\in\mathbb{N}_0}\|\hat{u}_{lm}\Phi_{lm}(r,s)\|_{\Omega_{\tilde{s}}}$, mit der Definition $\|f\|_M = \sup_{x\in M}|f(x)|$ für eine reelle Funktion f auf einer Menge M, einen Grenzwert besitzt. Nach dem Weierstraßschen Majorantenkriterium, siehe beispielsweise [51, S. 255], folgt desweiteren aus der normalen Konvergenz die gleichmäßige Konvergenz der Reihe $\sum_{l,m\in\mathbb{N}_0}\hat{u}_{lm}\Phi_{lm}$ auf $\Omega_{\tilde{s}}$ und damit die Stetigkeit der Grenzfunktion auf $\Omega_{\tilde{s}}$. Aufgrund der Abschätzung (3.12) und der Stetigkeit von u muss die Summe der Reihe auf $\mathbb{T}\backslash\{(-1,1)\}$ daher mit u übereinstimmen. Da die Abschätzung (3.14) für $(r,s)=(-1,1)$ nicht anwendbar ist, muss die Gültigkeit von (3.20) an dieser Stelle gesondert überprüft werden. Hierzu konstruieren wir die stetige Funktion $w:[-1,1]\to\mathbb{R}$ mit

$$w(b) = \begin{cases} u(-1,1), & b=1,\\ \frac{1}{2}\int_{-1}^{1} u(r(a,b),b)\,da, & b\neq 1,\end{cases}$$

so dass für $b\neq 1$ der integrale Mittelwert von u entlang der Parallelen $s=b$ zur unteren Kante des Standarddreiecks $\mathbb{T}$ angenommen wird. Aufgrund der Konstruktion sind die Koeffizienten $\hat{w}_m^{1,0}$ der Entwicklung von w basierend auf den Jacobi-Polynomen $P_m^{1,0}$ gegeben durch

$$\hat{w}_m^{1,0} = \frac{1}{\gamma_{0m}}\int_{-1}^{1}(1-b)w(b)P_m^{1,0}(b)\,db = \hat{u}_{0m},$$

und es gilt

$$\int_{-1}^{1}(1-b)\left[w(b)-\sum_{m\in\mathbb{N}_0}\hat{w}_m^{1,0}P_m^{1,0}(b)\right]^2 = 0. \tag{3.21}$$

Wegen $\sum_{0<m\leq N} |\hat{u}_{0m}| < \left\|\mathcal{L}_{r,s}^2 u\right\|_{L^2(\mathbb{T})} \cdot \sum_{m\in\mathbb{N}} m^{-3}$ und der in (A.4) gegebenen oberen Schranke $|P_m^{1,0}(b)| \leq m+1$ ist die Reihe $\sum_{m\in\mathbb{N}_0} \hat{u}_{0m} P_m^{1,0}$ normal konvergent und daher auch stetig auf dem Intervall $[-1,1]$, so dass die Summe der Reihe wegen (3.21) durch die Funktion w gegeben sein muss. Insgesamt erhält man daher

$$\sum_{l,m\in\mathbb{N}_0} \hat{u}_{lm}\Phi_{lm}(-1,1) = \sum_{m\in\mathbb{N}_0} \hat{u}_{0m} P_m^{1,0}(1) = w(1) = u(-1,1).$$

Mit (3.20) und (3.18) ergibt sich nun

$$u(r,s) - \mathcal{P}_N u(r,s) \;=\; \sum_{l+m>N} \hat{u}_{lm}\Phi_{lm}(r,s) = \int_{\mathbb{T}} R_N(r,s,\tilde{r},\tilde{s}) \mathcal{L}_{\tilde{r},\tilde{s}}^k u(\tilde{r},\tilde{s})\, d\tilde{r}d\tilde{s},$$

mit

$$R_N(r,s,\tilde{r},\tilde{s}) \;=\; \sum_{l+m>N} \frac{\Phi_{lm}(\tilde{r},\tilde{s})\Phi_{lm}(r,s)}{\gamma_{lm}(-\lambda_{lm})^k}.$$

Für $s \neq 1$ kann R_N abgeschätzt werden durch

$$\begin{aligned} \|R_N(r,s,\tilde{r},\tilde{s})\|_{L^2(\mathbb{T})}^2 &= \sum_{l+m>N} \lambda_{lm}^{-2k}\gamma_{lm}^{-1}\Phi_{lm}^2(r,s) < \left(\frac{2}{1-s}\right)^{3/2} \sum_{l+m>N} (l+m)^{-4k+2} \\ &< \left(\frac{2}{1-s}\right)^{3/2} \int_N^\infty (t+1)t^{-4k+2}dt < \left(\frac{2}{1-s}\right)^{3/2} \frac{2}{N^{4k-4}}, \end{aligned}$$

wobei die in Lemma 3.5 gegebene Abschätzung (3.14) für die Werte der PKD-Polynome verwendet wurde. Unter Verwendung von (3.15) und (3.16) erhalten wir desweiteren

$$\|R_N(-1,1,\tilde{r},\tilde{s})\|_{L^2(\mathbb{T})}^2 \;=\; \sum_{m>N} [m(m+2)]^{-2k} \frac{(m+1)^3}{2} < \int_N^\infty t^{-4k+3}dt < \frac{1}{N^{4k-4}}.$$

Mit der Cauchy-Schwarzschen Ungleichung gilt daher

$$|u(r,s) - \mathcal{P}_N u(r,s)| \;\leq\; \|R_N(r,s,\tilde{r},\tilde{s})\|_{L^2(\mathbb{T})} \cdot \|\mathcal{L}_{\tilde{r},\tilde{s}}^k u(\tilde{r},\tilde{s})\|_{L^2(\mathbb{T})} = \mathcal{O}(N^{-2k+2}),$$

so dass auch die Abschätzung (3.13) gezeigt ist. □

Insbesondere zeigen die Abschätzungen aus Satz 3.3, dass die Koeffizienten $\hat{u}_{lm}$ der Entwicklung einer beliebig glatten Ausgangsfunktion $u \in C^\infty(\mathbb{T})$ in eine PKD-Reihe schneller als polynomial gegen Null konvergieren und dies auch für der Abschneidefehler $u - \mathcal{P}_N u$ sowohl in der L^2-Norm als auch im punktweisen Sinn gilt. Diese Eigenschaft wird mit dem Begriff der *exponentiellen* oder *spektralen Konvergenz* bezeichnet und begründet die sehr guten Approximationseigenschaften spektraler Verfahren für partielle Differentialgleichungen mit glatten Lösungen.

Dämpfungseigenschaften des PKD-Operators $\mathcal{L}_{r,s}$ Der Operator $\mathcal{L}_{r,s}$, der sich grundsätzlich als Dämpfungsoperator verwenden lässt, führt zu keiner wesentlich stärkeren, aber zu einer möglicherweise deutlich schwächeren Energiedissipation im Vergleich zur Verwendung des Laplace-Operators. Dies wird durch die nachfolgende Abschätzung der zugehörigen Normen ersichtlich.
Zunächst sind aufgrund der positiven Definitheit der Matrix $\mathbf{B}(r,s)$ durch

$$(\mathbf{v},\mathbf{w})_{L^2_{\mathbf{B}}(\mathbb{T})} = (\mathbf{B}\mathbf{v},\mathbf{w})_{L^2(\mathbb{T})}, \quad \|\mathbf{v}\|_{L^2_{\mathbf{B}}(\mathbb{T})} = (\mathbf{B}\mathbf{v},\mathbf{v})^{1/2}_{L^2(\mathbb{T})}, \qquad \mathbf{v},\mathbf{w} \in [L^2(\mathbb{T})]^2, \quad (3.22)$$

ein Skalarprodukt und eine Norm auf $L^2(\mathbb{T})$ definiert.

Lemma 3.6 *Es gilt*

$$\|\mathbf{v}\|_{L^2_{\mathbf{B}}(\mathbb{T})} \leq \sqrt{2}\|\mathbf{v}\|_{L^2(\mathbb{T})}$$

für alle $\mathbf{v} \in [L^2(\mathbb{T})]^2$. Desweiteren gilt für alle $\mathbf{v} \in [\mathcal{P}^N(\mathbb{T})]^2$ die Abschätzung

$$\|\mathbf{v}\|_{L^2(\mathbb{T})} \leq CN\|\mathbf{v}\|_{L^2_{\mathbf{B}}(\mathbb{T})}, \quad (3.23)$$

mit einer geeigneten Konstanten C.

Eine derartige Abschätzung ist auch für sich genommen interessant und nach Kenntnis der Autorin bisher noch nicht nachgewiesen worden. Lemma 3.6 zeigt zudem, dass Lösungen $u : \mathbb{T} \times \mathbb{R}^+ \to \mathbb{R}$ der Differentialgleichung

$$\frac{\partial}{\partial t}u(r,s,t) = \epsilon\mathcal{L}_{r,s}u(r,s,t)$$

mit $u(\cdot,t) \in \mathcal{P}^N(\mathbb{T})$ die Energiedissipation

$$\frac{1}{2}\frac{d}{dt}\|u\|^2_{L^2(\mathbb{T})} = -\epsilon\|\nabla u\|^2_{L^2_{\mathbf{B}}(\mathbb{T})} \in \left[-\epsilon\sqrt{2}\|\nabla u\|^2_{L^2(\mathbb{T})}, -\frac{\epsilon}{C^2}N^{-2}\|\nabla u\|^2_{L^2(\mathbb{T})}\right]$$

besitzen. Die rechte Intervallgrenze liefert hierbei die Möglichkeit einer deutlich geringeren Energiedissipation für hohe Polynomgrade N im Vergleich zur Energiedissipation des üblichen Laplace-Operators unter Vernachlässigung von Randtermen $\int_{\partial\mathbb{T}} u\nabla u \cdot \mathbf{n}\, d\sigma$.

Beweis von Lemma 3.6: Es gilt zunächst

$$\|\mathbf{u}\|_{L^2_{\mathbf{B}}(\mathbb{T})} \leq (1-r^2)u_1^2 + (1-s^2)u_2^2 + 2(1+s)(1+r)|u_1u_2|.$$

Auf dem Standarddreieck $\mathbb{T}$ wird der Ausdruck $(1+r)(1+s)$ offensichtlich maximal für $r = s = 0$. Somit gilt mit

$$\|\mathbf{u}\|^2_{L^2_{\mathbf{B}}(\mathbb{T})} \leq (2-r^2)u_1^2 + (2-s^2)u_2^2 \leq 2\|\mathbf{u}\|^2_{L^2(\mathbb{T})}$$

die erste Abschätzung.
Die Grundidee des Nachweises der zweiten Abschätzung ist der Herleitung eines analogen Resultats für die durch $\omega(x) = 1-x^2$ gewichtete Norm auf $L^2[-1,1]$ von Bernardi und Maday in [4] entnommen. Es wird hierbei die Existenz von Quadraturformeln auf $\mathbb{T}$ vom Exaktheitsgrad $2N+2$ ausgenutzt, deren Stützstellenmenge $X_N = \{\xi_i \mid i = 0,\ldots,Q\}$ zum

einen einen Abstand $\geq \frac{2}{CN^2}$ vom Rand des Standarddreiecks besitzt, wobei C eine geeignete Konstante ist, und deren Gewichte $\omega_i,\ i = 0, \ldots, Q$, zum anderen alle positiv sind. Aus der ersten geforderten Eigenschaft der Quadraturformeln werden wir im Folgenden die Existenz einer Konstante $C_N \leq CN^2$ herleiten, so dass

$$\mathbf{v}^T\mathbf{v} \leq C_N \mathbf{v}^T \mathbf{B}(\boldsymbol{\xi})\mathbf{v} \tag{3.24}$$

für alle $\mathbf{v} \in \mathbb{R}^2$ und alle Stützstellen $\boldsymbol{\xi} \in X_N$ gilt. Da die Quadraturformeln nach Voraussetzung einen ausreichenden Exaktheitsgrad besitzen und die Gewichte positiv sind, ist die Integration auch bezüglich der Gewichtsmatrix $\mathbf{B}$ exakt, und es gilt mit (3.24)

$$\begin{aligned} \|\mathbf{u}\|^2_{L^2(\mathbb{T})} &= \sum_{i=0}^{Q} \omega_i \mathbf{u}(\boldsymbol{\xi}_i)^T \mathbf{u}(\boldsymbol{\xi}_i) \leq C_N \sum_{i=0}^{Q} \omega_i \mathbf{u}(\boldsymbol{\xi}_i)^T \mathbf{B}(\boldsymbol{\xi}_i)\mathbf{u}(\boldsymbol{\xi}_i) \\ &\leq CN^2 \sum_{i=0}^{Q} \omega_i \mathbf{u}(\boldsymbol{\xi}_i)^T \mathbf{B}(\boldsymbol{\xi}_i)\mathbf{u}(\boldsymbol{\xi}_i) = CN^2 \|\mathbf{u}\|^2_{L^2_{\mathbf{B}}(\mathbb{T})}, \end{aligned}$$

so dass die Abschätzung (3.23) gezeigt ist. Es bleibt also, Quadraturformeln auf $\mathbb{T}$ mit den gewünschten Eigenschaften zu konstruieren, und nachzuweisen, dass eine Konstante $C_N \leq CN^2$ existiert, so dass die Matrix

$$C_N\mathbf{B}(r,s) - \mathbf{I} = \begin{pmatrix} (1-r^2)C_N - 1 & -C_N(1+r)(1+s) \\ -C_N(1+r)(1+s) & (1-s^2)C_N - 1 \end{pmatrix} \tag{3.25}$$

für alle $\boldsymbol{\xi} = (r,s) \in X_N$ positiv definit ist, so dass (3.24) Gültigkeit hat. Hierzu soll zunächst die positive Definitheit der obigen Matrizen unter Voraussetzung der Existenz entsprechender Quadraturformeln gezeigt werden und abschließend die Konstruktion der Quadraturformeln selbst vorgenommen werden.

Definitheit der Matrizen (3.25):
Durch Anwendung des symmetrischen Gauß-Algorithmus ergibt sich aus (3.25) die Matrix

$$\begin{pmatrix} (1-r^2)C_N - 1 & 0 \\ 0 & \frac{-2C_N^2(r+s)(1+r)(1+s) - C_N(2-r^2-s^2)+1}{(1-r^2)C_N - 1} \end{pmatrix},$$

so dass für jede Konstante C_N mit

$$C_N \geq \max_{(r,s)\in X_N} \left\{ \frac{1}{1-r^2}, \frac{2-r^2-s^2}{-2(r+s)(1+r)(1+s} \right\} \tag{3.26}$$

die Matrix (3.25) für alle $(r,s) \in X_N$ positiv definit ist. Zur Herleitung einer oberen Schranke für C_N betrachten wir zunächst den zweiten der zu maximierenden Terme, der sich aufspalten lässt in

$$\frac{2-r^2-s^2}{-2(r+s)(1+r)(1+s)} = \frac{1-r}{-2(r+s)(1+s)} + \frac{1-s}{-2(r+s)(1+r)}.$$

Getrenntes Differenzieren der beiden auf dem gesamten Dreiecksinneren definierten Summanden zeigt, dass diese dort keine Extremwerte annehmen. Insbesondere gilt

$$\frac{\partial}{\partial r}\left(\frac{1-r}{-2(r+s)(1+s)}\right) = \frac{\partial}{\partial s}\left(\frac{1-s}{-2(r+s)(1+r)}\right) = \frac{1}{2(r+s)^2} > 0 \quad \text{für } (r,s) \in \mathbb{T}\backslash\partial\mathbb{T}.$$

Es genügt daher, diejenigen Stützstellen in X_N mit dem geringsten Abstand ϵ zu $\partial\mathbb{T}$ zu betrachten. Ohne Einschränkung können wir $\epsilon \leq \frac{1}{2}$ annehmen. Für die Stützstellen nahe der linken Dreieckskante, mit $r = -1+\epsilon, -1+\epsilon \leq s \leq 1-2\epsilon$, gilt dann

$$\frac{1-r}{-2(r+s)(1+s)} = \frac{2-\epsilon}{2(1-\epsilon-s)(1+s)} \leq \frac{1}{2\epsilon(1-\epsilon)} \leq \frac{1}{\epsilon}$$

sowie

$$\frac{1-s}{-2(r+s)(1+r)} = \frac{1-s}{2(1-\epsilon-s)\epsilon} \leq \frac{1}{2\epsilon(1-\epsilon)} \leq \frac{1}{\epsilon}.$$

Diese obere Schranke erhält man ebenso für die Stützstellen (r,s) nahe der unteren Dreieckskante mit $s = -1+\epsilon, -1+\epsilon \leq r \leq 1-2\epsilon$, sowie für diejenigen mit $-r-s = \epsilon, -1+\epsilon \leq r \leq 1-2\epsilon$, wie sich leicht analog zur obigen Abschätzung nachrechnen lässt. Mit der Wahl von $C_N = \frac{2}{\epsilon}$ ist daher wegen $\frac{1}{1-r^2} \leq \frac{1}{\epsilon(2-\epsilon)} < \frac{1}{\epsilon}$ die Bedingung (3.26) erfüllt, und aufgrund der Stützstellenkonstruktion gibt es eine Konstante C mit $\epsilon \geq \frac{2}{CN^2}$, so dass $C_N \leq CN^2$ gilt.

Konstruktion der Quadraturformeln:
Quadraturformeln auf dem Standarddreieck, die die genannten Bedingungen erfüllen, erhält man beispielsweise, indem man $\mathbb{T}$ zunächst durch Verbinden der Eckpunkte mit dem Schnittpunkt $P = \left(-\frac{1}{1+\sqrt{2}}, -\frac{1}{1+\sqrt{2}}\right)$ der Winkelhalbierenden in drei Teildreiecke $\tau_i,\ i = 1,2,3$, zerlegt, siehe Abbildung 3.3. Der Abstand einer in einem Teildreieck τ_i liegenden Stützstelle zum Rand $\partial\mathbb{T}$ ist dann gegeben durch ihren Abstand zur gemeinsamen Kante von τ_i und $\mathbb{T}$. Auf jedes der drei Teildreiecke lässt sich durch affine Transformationen wiederum das Standarddreieck einschließlich dort definierter Quadraturformeln abbilden, so dass die obere Ecke $(-1,1) \in \mathbb{T}$ auf den gemeinsamen Eckpunkt P abgebildet wird.
Es gilt also, eine Quadraturformel auf $\mathbb{T}$ vom Exaktheitsgrad $2N+2$ und mit positiven Gewichten zu konstruieren, deren Stützstellen alle einen Abstand $\geq \tilde{C}N^{-2}$ von der unteren Kante $s = -1$ von $\mathbb{T}$ besitzen, für eine Konstante $\tilde{C} > 0$. Es bietet sich an, hierbei sowohl das zusammenfallende Koordinatensystem mit der Transformation (3.1) und eindimensionale Quadraturformeln, als auch die folgende aus [86, Th. 6.21.2] entnommene Abschätzung für die Nullstellen der Legendre-Polynome zu nutzen, da sich so die Verteilung der Stützstellen der Gauss-Quadratur auf dem Intervall $[-1,1]$ auf den vorliegenden Fall übertragen lässt.
Seien die Nullstellen von $P_N^{0,0}(x)$ in absteigender Reihenfolge durch $1 > x_1 > x_2 > \cdots > x_N > -1$ gegeben und sei $x_\nu = \cos(\theta_\nu),\ \nu = 1,\ldots,N$, mit $0 < \theta_1 < \theta_2 < \cdots < \theta_N < \pi$, so gilt

$$\frac{2\nu-1}{2N+1}\pi \leq \theta_\nu \leq \frac{2\nu}{2N+1}\pi, \quad \nu = 1,\ldots,N. \tag{3.27}$$

Bei der Wahl der eindimensionalen Gauß-Quadratur in (3.3), sowohl in a- als auch in b-Richtung, lässt sich der durch die Transformation auf das Standardquadrat hinzukommende Faktor $\frac{1-b}{2}$ nicht in die Quadraturformel einbeziehen, sondern es ist anstelle dessen in b-Richtung die Exaktheit der Quadratur für alle Polynome in $\mathcal{P}^{2N+3}(-1,1)$ erforderlich. Dieser Exaktheitsgrad ist aber erfüllt, wenn sowohl in a- als auch in b-Richtung die Nullstellen von $P_{N+2}^{0,0}(x)$ als Stützstellen gewählt werden. Für den Abstand der kleinsten

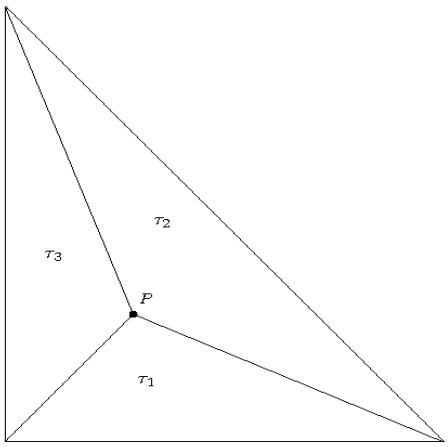


Abb. 3.3: Zerlegung des Standarddreiecks.

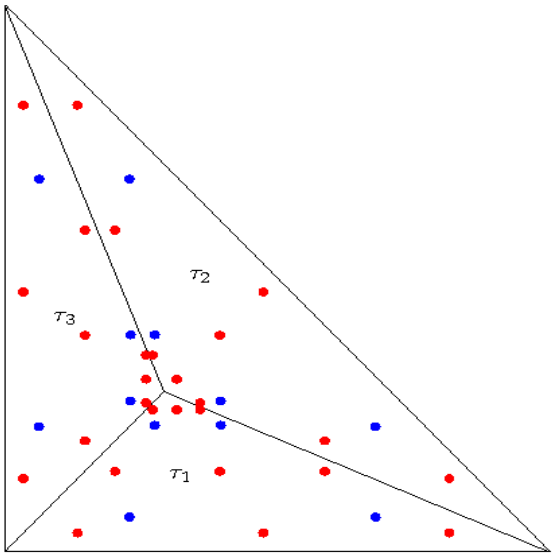


Abb. 3.4: Stützstellen mit Exaktheitsgrad $2N+2$ auf $\mathbb{T}$ für $N=0$ (blau) und $N=1$ (rot).

Stützstelle x_{N+2} vom linken Intervallrand gilt bei dieser Wahl

$$|-1-cos(\theta_{N+2})| \geq 1+\cos\left(\frac{-1}{2N+5}\pi+\pi\right) = \frac{1}{2}\left(\frac{\pi}{2N+5}\right)^2 + \mathcal{O}(N^{-4}) \geq \tilde{C}N^{-2}.$$

Da die Gewichte der Gauß-Quadratur desweiteren alle positiv sind, ist somit eine Quadraturformel auf $\mathbb{T}$ mit den gewünschten Eigenschaften erzeugt, die entsprechenden Stützstellen sind in Abbildung 3.4 dargestellt. □

4 Diskontinuierliche Galerkin-Verfahren auf Dreiecksgittern

Diskontinuierliche Galerkin(DG)-Verfahren zur numerischen Lösung hyperbolischer Erhaltungsgleichungen basieren zum einen auf einer Zerlegung des räumlichen Rechengebiets in kleinere Teilgebiete, die als *Zellen* oder *Elemente* bezeichnet werden und sich leicht auf ein Standardgebiet (wie zum Beispiel ein Rechteck oder Dreieck im zweidimensionalen Fall) transformieren lassen. Der zweite Grundstein dieser Verfahrensklasse ist die Darstellung der Näherungslösung zu gegebenem Zeitpunkt t als stückweise polynomiale Funktion im Raum, die nur an den Zellgrenzen Unstetigkeitsstellen besitzen darf. Das Herzstück des DG-Verfahrens ist dann der Entwurf einer variationellen Formulierung der Erhaltungsgleichung zur Diskretisierung der räumlichen Ableitungen. Aus dieser Formulierung ergibt sich eine gewöhnliche Differentialgleichung, die die zeitliche Entwicklung der räumlichen Freiheitsgrade der polynomialen Näherungslösung beschreibt. Diese Gleichung wird als *semidiskrete Gleichung* bezeichnet und im nächsten Diskretisierungsschritt mit Hilfe eines geeignet zu wählenden Verfahrens zur Lösung gewöhnlicher Differentialgleichungen behandelt, welches in diesem Kontext als Zeitintegrationsverfahren bezeichnet wird. Dem in dieser Arbeit konstruierten diskontinuierlichen Galerkin-Verfahren zur numerischen Lösung hyperbolischer Erhaltungsgleichungen liegen hauptsächlich die von Cockburn und Shu in einer Reihe von Veröffentlichungen [20, 19, 18, 17, 22, 23] entwickelten diskontinuierlichen Galerkin-Verfahren mit Zeitdiskretisierung durch explizite TVD-stabile Runge-Kutta-Verfahren zugrunde, die zusätzlich die Übertragung numerischer Flussfunktionen aus dem Kontext der Finite-Volumen-Verfahren beinhalten. Der zweite Ausgangspunkt ist die Verwendung orthogonaler Polynombasen und hochgenauer Quadraturformeln auf dem Dreieck nach Karniadakis und Sherwin [52], wie in Kapitel 3 beschrieben. Im Folgenden soll zunächst die durch diese Ansätze gegebene räumliche und anschließend die zeitliche Diskretisierung einer gegebenen hyperbolischen Erhaltungsgleichung erläutert werden. Abschließend werden wir auf die Stabilitäts- und Konvergenzeigenschaften des so konstruierten Basisverfahrens eingehen, welche zusätzlich anhand erster Testrechnungen veranschaulicht werden.

4.1 Räumliche Diskretisierung

Wir betrachten im Folgenden eine zweidimensionale hyperbolische Erhaltungsgleichung der Form

$$\frac{\partial}{\partial t}\,\mathbf{u}(\mathbf{x},t) + \frac{\partial}{\partial x_1}\mathbf{f}_1(\mathbf{u}(\mathbf{x},t)) + \frac{\partial}{\partial x_2}\mathbf{f}_2(\mathbf{u}(\mathbf{x},t)) = 0, \qquad (\mathbf{x},t) \in \Omega \times \mathbb{R}^+, \tag{4.1}$$

für eine vektorwertige Funktion $\mathbf{u} : \Omega \times \mathbb{R}^+ \to S \subseteq \mathbb{R}^m$. Zu (4.1) seien zudem Anfangsbedingungen $\mathbf{u}(\mathbf{x},0) = \mathbf{u}_0(\mathbf{x})$, sowie geeignete Randbedingungen gegeben.

Triangulierungen Da Zerlegungen des räumlichen Rechengebiets Ω in Dreieckselemente eine besonders hohe Flexibilität bei der Diskretisierung komplexer Strukturen aufweisen, werden in dieser Arbeit ausschließlich Dreiecksgitter betrachtet. Im Folgenden beschränken wir uns zudem auf polygonal berandete Gebiete Ω und konforme Triangulierungen.

Definition 4.1 *Eine* Triangulierung $\mathcal{T}^h$ *von* $\overline{\Omega}$ *ist eine endliche Menge von Dreiecken* $\tau_i \subset \overline{\Omega},\ i = 1, \ldots, \#\mathcal{T}^h$*, für die gilt:*

- $\overline{\Omega} = \cup_{i \in \{1,\ldots,\#\mathcal{T}^h\}} \tau_i$,
- *jedes Dreieck* $\tau_i \in \mathcal{T}^h$ *ist abgeschlossen und hat ein nichtleeres Inneres,*
- *die Dreiecke sind nicht überlappend, d.h. für* $\tau_i, \tau_j \in \mathcal{T}^h$ *mit* $i \neq j$ *gilt* $\overset{\circ}{\tau}_i \cap \overset{\circ}{\tau}_j = \emptyset$.

Eine Triangulierung $\mathcal{T}^h$ *heißt* konform*, wenn zusätzlich gilt:*

- *jede Kante eines Dreiecks* $\tau_i \in \mathcal{T}^h$ *ist entweder Teilmenge von* $\partial\Omega$ *oder Kante genau eines anderen Dreiecks* $\tau_j \in \mathcal{T}^h, j \neq i$.

Sei nun $\mathcal{T}^h$ eine konforme Triangulierung von $\overline{\Omega}$. Als zu $\mathcal{T}^h$ gehöriger Ansatzraum, in dem die "Momentaufnahme" der numerischen Approximation zu einem Zeitpunkt $t \in \mathbb{R}_0^+$ enthalten sein soll, wird der Raum

$$V^{h,N} = \{v \in L^\infty(\Omega) \mid v|_{\tau_i} \in \mathcal{P}^N(\tau_i) \quad \forall\, \tau_i \in \mathcal{T}^h\}$$

der stückweise polynomialen Funktionen vom Grad $\leq N$ gewählt.

Bemerkung 4.2 *Unter Ausnutzung geeigneter Strategien zur p-Adaption ist es auch möglich, auf verschiedenen Dreieckselementen unterschiedliche Polynomgrade zuzulassen. In dieser Arbeit gehen wir jedoch von einem über das gesamte Rechengebiet konstant gewählten Polynomgrad aus.*

Zur Herleitung der semidiskreten Gleichung wird die Gleichung (4.1) zunächst mit Testfunktionen $\mathbf{v} \in \left(V^{h,N}\right)^m$ multipliziert und anschließend über das räumliche Gebiet Ω integriert. Unter Verwendung der Greenschen Integralformel ergibt sich daraus

$$\begin{aligned} \frac{d}{dt}\int_\Omega \mathbf{u}\,\mathbf{v}\,d\mathbf{x} + \sum_{\tau_i \in \mathcal{T}^h} \Bigg(\int_{\partial\tau_i} (\mathbf{f}_1(\mathbf{u})\,n_1 + \mathbf{f}_2(\mathbf{u})\,n_2) \cdot \mathbf{v}\,d\sigma \\ - \int_{\tau_i} \mathbf{f}_1(\mathbf{u}) \cdot \frac{\partial \mathbf{v}}{\partial x_1} + \mathbf{f}_2(\mathbf{u}) \cdot \frac{\partial \mathbf{v}}{\partial x_2}\,d\mathbf{x} \Bigg) &= 0, \quad \forall \mathbf{v} \in \left(V^{h,N}\right)^m . \end{aligned} \tag{4.2}$$

Verwendung numerischer Flussfunktionen Fordert man nun die Gültigkeit der Gleichung (4.2) von einer stückweise polynomialen Funktion $\mathbf{u}_{h,N} : \Omega \times \mathbb{R}^+ \to \mathbb{R}^m$ mit $\mathbf{u}_{h,N}(\cdot, t) \in \left(V^{h,N}\right)^m$ für alle $t \in \mathbb{R}^+$, die sich im obigen Ansatzraum bewegt und die Lösung $\mathbf{u}$ in (4.2) ersetzen soll, so ergibt sich zunächst die Schwierigkeit, dass über die Elementgrenzen hinweg Unstetigkeiten zugelassen werden und die entsprechenden Flüsse im zweiten Integral nicht eindeutig definiert sind. Aus diesem Grund wird das Konzept einer numerischen Flussfunktion aus dem Kontext von Finite-Volumen-Verfahren übernommen. Dieser Funktion kommt die Aufgabe zu, durch exakte oder approximative Lösung des zugehörigen Riemann-Problems, aus den links- und rechtsseitigen Grenzwerten von $\mathbf{u}_{h,N}$ an der Zellkante einen eindeutig definierten Fluss zu bestimmen, welcher den Ausdruck $\mathbf{f}_1(\mathbf{u})n_1 + \mathbf{f}_2(\mathbf{u})n_2$ in den Integralen über Zellgrenzen ersetzt.

Definition 4.3 *Es sei* $B_1 = \{\mathbf{n} \mid \|\mathbf{n}\|_2 = 1\}$*. Eine* numerische Flussfunktion *für die Gleichung* (4.1) *ist eine Funktion* $\mathbf{H} : S \times S \times B_1 \to S$*, die folgenden Eigenschaften erfüllt:*

- $\mathbf{H}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n})$ *ist Lipschitz-stetig bezüglich der ersten beiden Argumente,* $\mathbf{u}^- \in S$ *und* $\mathbf{u}^+ \in S$*,*
- $\mathbf{H}$ *ist* konsistent *mit den Flussvektoren* $\mathbf{f}_1$ *und* $\mathbf{f}_2$*, d.h. für alle* $\mathbf{u} \in S$ *und* $\mathbf{n} \in B_1$ *gilt* $\mathbf{H}(\mathbf{u}, \mathbf{u}, \mathbf{n}) = \mathbf{f}_1(\mathbf{u})n_1 + \mathbf{f}_2(\mathbf{u})n_2$*,*
- $\mathbf{H}$ *ist* konservativ*, d.h. für alle* $\mathbf{u}^-, \mathbf{u}^+ \in S$ *und alle* $\mathbf{n} \in B_1$ *gilt die Gleichung* $\mathbf{H}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n}) = -\mathbf{H}(\mathbf{u}^+, \mathbf{u}^-, -\mathbf{n})$*.*

Um bestimmte Aussagen zu Stabilität und Konvergenzordnung der DG-Methode treffen zu können, werden meistens zusätzliche Eigenschaften der numerischen Flussfunktion gefordert. Im Fall skalarer Erhaltungsgleichungen kommt den folgenden Klassen numerischer Flussfunktionen besondere Bedeutung zu.

Definition 4.4 *Sei* $\mathbf{H}$ *eine numerische Flussfunktion für eine skalare Erhaltungsgleichung.*

1. $\mathbf{H}$ *wird als* E-Fluss *bezeichnet, falls gilt*

$$\big(\mathbf{H}(u^-, u^+, \mathbf{n}) - \mathbf{f}(u) \cdot \mathbf{n}\big)\,(u^+ - u^-) \leq 0, \quad \forall u \in [\min\{u^-, u^+\}, \max\{u^-, u^+\}].$$

2. $\mathbf{H}$ *heißt* monoton*, wenn* $\mathbf{H}(\,\cdot\,, u^+, \mathbf{n})$ *monoton steigend ist.*

3. *Eine monotone numerische Flussfunktion* $\mathbf{H}$ *heißt* Upwind-Fluss*, falls zusätzlich die Bedingungen*

$$\mathbf{H}(u^-, u^+, \mathbf{n}) = \begin{cases} \mathbf{f}(u^-) \cdot \mathbf{n}, & \text{falls } \mathbf{f}'(u) \cdot \mathbf{n} \geq 0 \text{ für alle } u \text{ zwischen } u^- \text{ und } u^+, \\ \mathbf{f}(u^+) \cdot \mathbf{n}, & \text{falls } \mathbf{f}'(u) \cdot \mathbf{n} < 0 \text{ für alle } u \text{ zwischen } u^- \text{ und } u^+, \end{cases}$$

erfüllt sind.

In den Testrechnungen für skalare Erhaltungsgleichungen mit $\mathbf{f} = (f_1, f_2)^T$ wird der *Godunov-Fluss*

$$\mathbf{H}(u^-, u^+, \mathbf{n}) = \begin{cases} \min\limits_{u^- \leq u \leq u^+} \mathbf{f}(u) \cdot \mathbf{n} & \text{falls } u^- \leq u^+, \\ \max\limits_{u^+ \leq u \leq u^-} \mathbf{f}(u) \cdot \mathbf{n} & \text{sonst.} \end{cases}$$

verwendet, der die exakte Lösung des Riemann-Problems liefert. Offensichtlich ist der Godunov-Fluss nach obiger Definition ein Upwind-Fluss. Im Fall der linearen Advektionsgleichung (2.20) erhält man bei dieser Wahl den numerischen Fluss

$$\mathbf{H}(u^-, u^+, \mathbf{n}) = \begin{cases} u^- \, \mathbf{a} \cdot \mathbf{n} & \text{falls } \mathbf{a} \cdot \mathbf{n} \geq 0, \\ u^+ \, \mathbf{a} \cdot \mathbf{n} & \text{sonst.} \end{cases}$$

Für die Euler-Gleichungen wird zunächst die Rotationsinvarianz (2.27) bezüglich der Drehmatrix $\mathbf{Q}(\mathbf{n})$ ausgenutzt, so dass der Fluss $\mathbf{f}_1(\mathbf{u})n_1 + \mathbf{f}_2(\mathbf{u})n_2$ über Zellgrenzen durch den Term

$$\mathbf{H}(\mathbf{u}^-, \mathbf{u}^+, \mathbf{n}) = \mathbf{Q}(\mathbf{n})^{-1} \mathbf{H}_{1D}(\mathbf{Q}(\mathbf{n})\mathbf{u}^-, \mathbf{Q}(\mathbf{n})\mathbf{u}^+)$$

ersetzt wird, wobei $\mathbf{H}_{1D}$ eine numerische Flussfunktion für die Euler-Gleichungen in einer Raumdimension ist. In dieser Arbeit wurde zur Definition von $\mathbf{H}_{1D}$ das Flussvektor-Splitting-Verfahren nach van Leer [61] verwendet, welches im Anhang A.2 beschrieben ist.

Mit dem Ziel der Aufnahme der numerischen Flussfunktion in die Formulierung (4.2) betrachten wir die Gleichung auf einer Zelle $\tau_i \in \mathcal{T}^h$. Zur Bezeichnung der linksseitigen Grenzwerte der Näherungslösung $\mathbf{u}_{h,N}$ an der Zellgrenze wird für deren Spur die Kurznotation $\mathbf{u}^i_{h,N} = \mathbf{u}_{h,N}|_{\partial\tau_i}$ verwendet und der Rand des Dreiecks τ_i in die Kanten Γ_{ij} mit Normalenvektoren $\mathbf{n}_{ij}$ zerlegt, $\partial\tau_i = \cup_{j=1}^3 \Gamma_{ij}$. Gibt es ein Dreieck $\tau_k \in \mathcal{T}^h$ mit $\Gamma_{ij} = \partial\tau_i \cap \partial\tau_k$, so sind die rechtsseitigen Grenzwerte $\mathbf{u}^{ij}_{h,N} : \Gamma_{ij} \times \mathbb{R}^+ \to S$ an der Zellkante Γ_{ij} durch $\mathbf{u}^{ij}_{h,N}(\mathbf{x},t) = \mathbf{u}^k_{h,N}(\mathbf{x},t)$ gegeben.
Mit Hilfe der numerischen Flussfunktion lassen sich auch bestimmte Randbedingungen in die variationelle Formulierung aufnehmen. In dem Fall, dass in Bezug auf die charakteristischen Richtungen einströmende Werte vorliegen, definiert man $\mathbf{u}^{ij}_{h,N}(\mathbf{x},t) = \mathbf{g}(\mathbf{x},t)$, mit der durch die Randbedingungen vorgegebenen Funktion $\mathbf{g}$. An diesen Rändern werden daher anstelle des Erzwingens von Dirichlet-Randbedingungen Riemann-Probleme gelöst. Im Fall des Ausströmens wird $\mathbf{u}^{ij}_{h,N} = \mathbf{u}^i_{h,N}$ gesetzt, so dass aufgrund der Konsistenz der numerischen Flussfunktion an diesen Rändern der linksseitge Fluss $f(\mathbf{u}^i_{h,N})$ verwendet wird. Liegt bei einem Testfall der Euler-Gleichungen eine feste Wand vor, so wird nicht die numerische Flussfunktion, sondern der in (2.29) definierte Fluss $\mathbf{f}_W$ verwendet. Für diese Situation definieren wir die Menge

$$W(i) = \{j \in \{1,2,3\} \mid \Gamma_{ij} \text{ unterliegt der Randbedingung für eine feste Wand}\}.$$

Zusammengefasst erhalten wir die lokale Diskretisierung

$$\begin{aligned}\frac{d}{dt}\int_{\tau_i} \mathbf{u}_{h,N}\,\tilde{\mathbf{v}}\,d\mathbf{x} \;&+\; \sum_{\substack{j=1\\ j\notin W(i)}}^{3} \int_{\Gamma_{ij}} \mathbf{H}(\mathbf{u}^i_{h,N}, \mathbf{u}^{ij}_{h,N}, \mathbf{n}_{ij})\cdot\tilde{\mathbf{v}}\,d\sigma \\ &+\; \sum_{j\in W(i)} \int_{\Gamma_{ij}} \mathbf{f}_W(\mathbf{u}^i_{h,N}, \mathbf{n}_{ij})\cdot\tilde{\mathbf{v}}\,d\sigma \\ &-\; \int_{\tau_i} \mathbf{f}_1(\mathbf{u}_{h,N})\cdot\frac{\partial\tilde{\mathbf{v}}}{\partial x_1} + \mathbf{f}_2(\mathbf{u}_{h,N})\cdot\frac{\partial\tilde{\mathbf{v}}}{\partial x_2}\,d\mathbf{x} \;=\; 0, \end{aligned} \tag{4.3}$$

für alle Dreiecke $\tau_i \in \mathcal{T}^h$ und alle vektorwertigen Testfunktionen $\tilde{\mathbf{v}} \in \left(\mathcal{P}^N(\tau_i)\right)^m$.

Verwendung der orthogonalen Polynombasis Da $\mathbf{u}_{h,N}(\cdot,t)$ polynomial auf jeder Zelle τ_i ist, kann (4.3) als System gewöhnlicher Differentialgleichungen für die zeitabhängigen Koeffizienten von $\mathbf{u}_{h,N}$ bezüglich einer geeigneten Basis von $\mathcal{P}^N(\tau_i)$ aufgefasst werden. Hierzu sei $\Lambda_i : \tau_i \to \mathbb{T},\ \mathbf{x} \mapsto \mathbf{A}_i\mathbf{x} + \mathbf{b}_i$, mit $\mathbf{A}_i \in \mathbb{R}^{2\times 2}$ und $\mathbf{b}_i \in \mathbb{R}^2$, eine orientierungserhaltende affine Transformation, die das spezifische Dreieck τ_i in das Referenzelement $\mathbb{T}$ abbildet. Ausgehend von der in Kapitel 3 beschriebenen Basis der Proriol-Koornwinder-Dubiner(PKD)-Polynome erhält man mit Hilfe der Transformation Λ_i dann eine Basis des Vektorraums $\mathcal{P}^N(\tau_i)$, die aus den Polynomen $\Phi_{lm} \circ \Lambda_i$, für $l,m \in \mathbb{N}_0$ mit $0 \le l+m \le N$, besteht. Dadurch lässt sich $\mathbf{u}_{h,N}|_{\tau_i}$ darstellen als

$$\mathbf{u}_{h,N}(\mathbf{x},t) = \sum_{l+m\le N} \hat{\mathbf{u}}^i_{lm}(t)\cdot\Phi_{lm}\circ\Lambda_i(\mathbf{x})$$

mit zeitabhängigen, vektorwertigen Koeffizienten $\hat{\mathbf{u}}^i_{lm}$. Für diese liefert die Orthogonalitätseigenschaft der PKD-Polynome die Berechnungsvorschrift

$$\hat{\mathbf{u}}^i_{lm}(t) \;=\; \frac{2}{\gamma_{lm}|\tau_i|}\cdot\int_{\tau_i}\mathbf{u}_{h,N}(\mathbf{x},t)\,\Phi_{lm}\circ\Lambda_i(\mathbf{x})\,d\mathbf{x}, \tag{4.4}$$

mit $\gamma_{lm}=\|\Phi_{lm}\|^2_{L^2(\mathbb{T})}=\frac{2}{(2l+1)(l+m+1)}$ und der Determinante $\frac{2}{|\tau_i|}$ der Jacobi-Matrix A_i von Λ_i.

Mit dem Einsatz der orthogonalen Basis als Testfunktionen erhält man daher direkt die Darstellung der Gleichung (4.3) in den Koeffizienten der PKD-Entwicklung. Es ergibt sich das System gewöhnlicher Differentialgleichungen

$$\begin{aligned}\frac{d}{dt}\hat{\mathbf{u}}^i_{lm} \;=\;& -\frac{2}{\gamma_{lm}|\tau_i|}\sum_{\substack{j=1\\ j\notin W(i)}}^{3}\int_{\Gamma_{ij}}\mathbf{H}(\mathbf{u}^i_{h,N},\mathbf{u}^{ij}_{h,N},\mathbf{n}_{ij})\cdot(\Phi_{lm}\circ\Lambda_i)\,d\sigma\\ &-\frac{2}{\gamma_{lm}|\tau_i|}\sum_{j\in W(i)}\int_{\Gamma_{ij}}\mathbf{f}_W(\mathbf{u}^i_{h,N},\mathbf{n}_{ij})\cdot(\Phi_{lm}\circ\Lambda_i)\,d\sigma\\ &+\frac{1}{\gamma_{lm}}\int_{\mathbb{T}}\mathcal{F}(\mathbf{u}_{h,N}\circ\Lambda_i^{-1})\cdot\mathbf{A}_i^T\nabla_{r,s}\Phi_{lm}\,drds,\quad 0\le l+m\le N,\end{aligned} \tag{4.5}$$

wobei die Transfomation $\nabla_{\mathbf{x}}=\mathbf{A}_i^T\nabla_{r,s}$ der räumlichen Ableitungen in das Standarddreieck sowie die Notation $\mathcal{F}=(\mathbf{f}_1,\mathbf{f}_2)^T$ verwendet wurde.

Verwendung von Quadraturformeln Zur endgültigen Definition der räumlichen Diskretisierung wird eine letzte Modifikation der variationellen Formulierung vorgenommen. Diese ergibt sich aus der Feststellung, dass die im zweiten und dritten Term von (4.3) auftretenden Integrale im Fall von nichtlinearen Funktionen $\mathbf{f}_1,\mathbf{f}_2$ und $\mathbf{H}$ zumeist nicht exakt berechnet werden können. Diese Integrale werden daher durch Quadraturformeln mit geeignetem Exaktheitsgrad approximiert. Dazu werden an den Zellkanten Gauß-Quadraturformeln verwendet, während für das Volumenintegral durch die im vorangegangenden Kapitel beschriebene singuläre Transformation des Standarddreiecks $\mathbb{T}$ in das Quadrat $[-1,1]^2$ hochgenaue Quadraturformeln auf dem Standarddreieck konstruiert werden.

Mit $\xi_\nu\in[-1,1],\ \nu=1,\dots,n_K$, seien Gaußsche Integrationspunkte bezeichnet, die zugehörigen Gewichte mit ω_ν. Die Integrationspunkte auf $\mathbb{T}$ seien $(r_\mu,s_\mu)\in\mathbb{T},\ \mu=1,\dots,n_I$, mit den zugehörigen Gewichten $\tilde{\omega}_\mu$. Desweiteren sei $\mathbf{x}_{ij}:[-1,1]\to\Gamma_{ij}$ eine affine Transformation der Gauß-Punkte von $[-1,1]$ nach Γ_{ij}. Um einen Abschneidefehler der Ordnung $N+1$ gewährleisten zu können, siehe hierzu [17], ist für die Integration über die Dreiecksfläche ein Exaktheitsgrad von $2N$ erforderlich, während auf jeder Zellkante des Randes $\partial\tau_i$ ein Exaktheitsgrad von $2N+1$ verlangt wird.

Die semidiskrete Gleichung für die Koeffizienten $\hat{\mathbf{u}}^i_{lm}$ ist dann gegeben durch

$$\begin{aligned}\frac{d}{dt}\hat{\mathbf{u}}^i_{lm} &= -\frac{1}{\gamma_{lm}|\tau_i|}\sum_{\substack{j=1\\ j\notin W(i)}}^{3}|\Gamma_{ij}|\sum_{\nu=1}^{n_K}\Big\{\omega_\nu \mathbf{H}\left(\mathbf{u}^i_{h,N}(\mathbf{x}_{ij}(\xi_\nu),t),\mathbf{u}^{ij}_{h,N}(\mathbf{x}_{ij}(\xi_\nu),t),\mathbf{n}_{ij}\right)\\ &\qquad\qquad\cdot\Phi_{lm}(\Lambda_i(\mathbf{x}_{ij}(\xi_\nu))\Big\}\\ &\quad -\frac{1}{\gamma_{lm}|\tau_i|}\sum_{j\in W(i)}|\Gamma_{ij}|\sum_{\nu=1}^{n_K}\omega_\nu \mathbf{f}_W\left(\mathbf{u}^i_{h,N}(\mathbf{x}_{ij}(\xi_\nu),t),\mathbf{n}_{ij}\right)\cdot\Phi_{lm}(\Lambda_i(\mathbf{x}_{ij}(\xi_\nu))\\ &\quad +\frac{1}{\gamma_{lm}}\sum_{\mu=1}^{n_I}\tilde{\omega}_\mu\,\mathcal{F}(\mathbf{u}_{h,N}\circ\Lambda_i^{-1}(r_\mu,s_\mu))\cdot\mathbf{A}_i^T\nabla_{r,s}\Phi_{lm}(r_\mu,s_\mu),\end{aligned}\tag{4.6}$$

für $0\le l+m\le N$ und $i=1,\dots,\#\mathcal{T}^h$, mit den Stützstellenanzahlen $n_K=N+1$ und $n_I=(N+1)^2$, vergleiche (3.4).

Durch die vorangegangenen Schritte erhalten wir somit ein System gewöhnlicher Differentialgleichungen

$$\frac{d}{dt}\mathbf{U}(t)=\mathcal{L}_{h,N}(\mathbf{U}(t),t),\tag{4.7}$$

bei dem die Funktion $\mathbf{U}$ den gesamten Satz aller Koeffizienten auf den Dreiecken der Triangulierung beinhaltet, während der Operator $\mathcal{L}_{h,N}$ die Diskretisierung der räumlichen Ableitungen darstellt, in Form der rechten Seiten von (4.6). Die Lösung dieses Systems mittels eines geeigneten Zeitintegrationsverfahrens ist Gegenstand des nächsten Abschnittes. Bezüglich der Implementation der Gleichung (4.6) sei noch angemerkt, dass die Werte $\Phi_{lm}(\Lambda_i(\mathbf{x}_{ij}(\xi_\nu)))$ und die Gradienten $\nabla_{r,s}\Phi_{lm}(r_\mu,s_\mu)$ der Basisfunktionen an den Stützstellen im Referenzdreieck, auf die im Laufe der zeitlichen Evolution mehrfach zurückgegriffen werden muss, vor dem Verfahrensablauf berechnet und abgespeichert werden. Zur einmaligen Berechnung der Werte und Ableitungen der entsprechenden Jacobi-Polynome werden die in Anhang A.1 angegebenen Rekursionsformeln verwendet.

4.2 Zeitliche Diskretisierung

Wie üblich, werden Zeitpunkte, an denen eine Approximation der Lösung des Systems (4.7) berechnet wird, mit $t^n,\ n=0,1,\dots$, die zugehörigen Zeitschritte mit $\Delta t^n=t^{n+1}-t^n$, die berechneten numerischen Lösungen mit $\mathbf{U}^n$, und die zugehörigen Werte des diskreten räumlichen Operators mit $\mathbf{L}^n=\mathcal{L}_{h,N}(\mathbf{U}^n,t^n)$ bezeichnet.
Im Allgemeinen ist die Wahl eines geeigneten Zeitintegrationsverfahrens unter anderem von der gewünschten Genauigkeit der Approximation, der vorhandenen Speicherkapazität und der Rechengeschwindigkeit abhängig. Während Mehrschrittverfahren pro Zeitschritt nur eine Auswertung von $\mathcal{L}_{h,N}$ benötigen, müssen die Lösungen zu vorherigen Zeitpunkten gespeichert werden. Mit Einschrittverfahren lässt sich an dieser Stelle Speicherplatz einsparen, allerdings sind in diesem Fall für Verfahren höherer Ordnung mehr als eine Auswertung von $\mathcal{L}_{h,N}$ pro Zeitschritt notwendig. In dieser Arbeit wurden zugunsten des geringeren Speicherplatzverbrauchs verschiedene explizite Runge-Kutta-Verfahren der

Form

$$\begin{aligned}
\mathbf{U}^{(0)} &= \mathbf{U}^n, \\
\mathbf{U}^{(k)} &= \sum_{j=0}^{k-1} \left(\alpha_{kj} \mathbf{U}^{(j)} + \beta_{kj} \Delta t^n \mathcal{L}_{h,N}(\mathbf{U}^{(j)}, t^n + \gamma_j \Delta t^n) \right), \ k = 1, \ldots, s, \\
\mathbf{U}^{n+1} &= \mathbf{U}^{(s)}
\end{aligned} \tag{4.8}$$

genutzt. Für eine derartige Kombination der DG-Raumdiskretisierung mit einem Runge-Kutta-Verfahren zur Zeitintegration verwenden wir die von Cockburn und Shu eingeführte Bezeichnung RKDG-Verfahren. Implementiert wurden im Kontext dieser Arbeit neben dem expliziten Euler-Verfahren

$$\mathbf{U}^{n+1} = \mathbf{U}^n + \Delta t^n \mathbf{L}^n$$

die bereits von Shu und Osher [84] betrachteten TVD-stabilen Runge-Kutta-Verfahren zweiter Ordnung,

$$\begin{aligned}
\mathbf{U}^{(1)} &= \mathbf{U}^n + \Delta t^n \mathbf{L}^n, \\
\mathbf{U}^{n+1} &= \frac{1}{2} \left(\mathbf{U}^n + \mathbf{U}^{(1)} + \Delta t^n \mathcal{L}_{h,N}(\mathbf{U}^{(1)}, t^n + \Delta t^n) \right),
\end{aligned} \tag{4.9}$$

und dritter Ordnung,

$$\begin{aligned}
\mathbf{U}^{(1)} &= \mathbf{U}^n + \Delta t^n \mathbf{L}^n, \\
\mathbf{U}^{(2)} &= \frac{1}{4} \left(3\mathbf{U}^n + \mathbf{U}^{(1)} + \Delta t^n \mathcal{L}_{h,N} \left(\mathbf{U}^{(1)}, t^n + \Delta t^n \right) \right) \\
\mathbf{U}^{n+1} &= \frac{1}{3} \left(\mathbf{U}^n + 2\mathbf{U}^{(2)} + 2\Delta t^n \mathcal{L}_{h,N} \left(\mathbf{U}^{(2)}, t^n + \frac{1}{2} \Delta t^n \right) \right).
\end{aligned} \tag{4.10}$$

Als TVD-Stabilität wird hierbei die Eigenschaft des Zeitintegrationsverfahrens bezeichnet, die Stabilität des expliziten Euler-Verfahrens, die bezüglich einer gegebenen Halbnorm (beispielsweise die der Totalvariation oder der Totalvariation der Zellmittelwerte) unter geeigneter Schrittweitenrestriktion vorausgesetzt wird, zu erhalten, gegebenenfalls unter einer schärferen Zeitschrittrestriktion. Dementsprechend wurden derartige Verfahren von Gottlieb, Shu und Tadmor in [40] zu SSP("strong stability preserving")-Methoden umbenannt. Die genannte Stabilitätseigenschaft ergibt sich für nichtnegative Koeffizienten α_{kj}, β_{kj} mit $\alpha_{kj} \neq 0$ für $\beta_{kj} \neq 0$ in (4.8) aus der Eigenschaft, dass sich die Zwischenwerte $\mathbf{U}^{(k)}$ wegen $\sum_{j=0}^{k-1} \alpha_{kj} = 1$ als Konvexkombination

$$\mathbf{U}^{(k)} = \sum_{j=0}^{k-1} \alpha_{kj} \left(\mathbf{U}^{(j)} + \frac{\beta_{kj}}{\alpha_{kj}} \Delta t^n \mathcal{L}_{h,N}(\mathbf{U}^{(j)}, t^n + \gamma_j \Delta t^n) \right)$$

von Euler-Schritten schreiben lassen. Gilt nun für eine beliebige Halbnorm $|\cdot|$ und einen Zeitschritt Δt_E^n des expliziten Euler-Verfahrens

$$|\mathbf{U}^n + \Delta t_E^n \mathbf{L}^n| \leq |\mathbf{U}^n|,$$

so gilt für alle $\Delta t^n \leq \min_{k,j} \frac{\alpha_{kj}}{\beta_{kj}} \Delta t_E^n$ die Stabilität $|\mathbf{U}^{n+1}| \leq |\mathbf{U}^n|$ des Runge-Kutta-Verfahrens (4.8) bezüglich derselben Halbnorm.

Kann eine derartige Stabilitätseigenschaft schon für das explizite Euler-Verfahren nicht garantiert werden, so ist die Beschränkung auf TVD-stabile Methoden nicht unbedingt notwendig. Als ein Runge-Kutta-Verfahren mit geringem Speicherbedarf wurde daher wie in [45] desweiteren die von Carpenter und Kennedy [13] entwickelte Methode vierter Ordnung,

$$\begin{aligned} &\mathbf{U}^{(0)} = \mathbf{U}^n, \\ &\left.\begin{aligned} \tilde{\mathbf{U}}^{(k)} &= a_k \tilde{\mathbf{U}}^{(k-1)} + \Delta t^n \mathcal{L}_{h,N}(\mathbf{U}^{(k-1)}, t^n + c_k \Delta t^n) \\ \mathbf{U}^{(k)} &= \mathbf{U}^{(k-1)} + b_k \tilde{\mathbf{U}}^{(k)} \end{aligned}\right\}, \; k = 1, \ldots, 5, \\ &\mathbf{U}^{n+1} = \mathbf{U}^{(5)} \end{aligned} \tag{4.11}$$

mit den Koeffizienten

$$\begin{aligned} a_1 &= 0, & b_1 &= \tfrac{1432997174477}{9575080441755} \approx 0.15, & c_1 &= 0, \\ a_2 &= -\tfrac{567301805773}{1357537059087} \approx -0.42, & b_2 &= \tfrac{5161836677717}{13612068292357} \approx 0.38, & c_2 &= \tfrac{1432997174477}{9575080441755} \approx 0.15, \\ a_3 &= -\tfrac{2404267990393}{2016746695238} \approx -1.19, & b_3 &= \tfrac{1720146321549}{2090206949498} \approx 0.82, & c_3 &= \tfrac{2526269341429}{6820363962896} \approx 0.37, \\ a_4 &= -\tfrac{3550918686646}{2091501179385} \approx -1.7, & b_4 &= \tfrac{3134564353537}{4481467310338} \approx 0.7, & c_4 &= \tfrac{2006345519317}{3224310063776} \approx 0.62, \\ a_5 &= -\tfrac{1275806237668}{842570457699} \approx -1.51, & b_5 &= \tfrac{2277821191437}{14882151754819} \approx 0.15, & c_5 &= \tfrac{2802321613138}{2924317926251} \approx 0.96, \end{aligned}$$

implementiert.

4.3 Stabilität und Konvergenz des DG-Verfahrens

In diesem Unterkapitel soll zum einen die Frage nach notwendigen Zeitschrittbedingungen für die in Abschnitt 4.2 aufgeführten expliziten Zeitintegrationsverfahren zur vollständigen Diskretisierung der semidiskreten Formulierung (4.6) beantwortet werden. Zunächst gehen wir auf die Stabilitätseigenschaften der semidiskreten DG-Formulierung im Fall einer skalaren Erhaltungsgleichung ein. Mit Hilfe einer von Neumannschen Stabilitätsanalyse in Matrixform [46] für skalare lineare Erhaltungsgleichungen in zwei Raumdimensionen mit periodischen Randbedingungen wird die lineare L^2-Stabilität der vollständigen Diskretisierung untersucht. Aus den resultierenden Zeitschrittrestriktionen im linearen Fall wird anschließend die Wahl des Zeitschritts im Fall nichtlinearer hyperbolischer Systeme von Erhaltungsgleichungen motiviert.
Zum anderen wird anhand von Testrechnungen mit der linearen Advektionsgleichung (2.20) die experimentelle Konvergenzordnung des DG-Verfahrens für lineare Erhaltungsgleichungen untersucht und den bekannten theoretischen Konvergenzresultaten gegenübergestellt. Anhand der Anwendung des DG-Verfahrens auf die nichtlinearen Testgleichungen (2.21) und (2.23) veranschaulichen wir zudem die Problematik der nicht hinreichenden numerischen Dämpfung des expliziten RKDG-Verfahrens in seiner bisher beschriebenen Form im Fall nichtlinearer hyperbolischer Erhaltungsgleichungen mit unstetiger Lösung.

4.3.1 Lineare und nichtlineare L^2-Stabilität

Nichtlineare L^2-Stabilität Für die semidiskrete DG-Formulierung (4.3) einer skalaren Erhaltungsgleichung

$$\frac{\partial}{\partial t} u(\mathbf{x}, t) + \nabla_{\mathbf{x}} \cdot \mathcal{F}(u(\mathbf{x}, t)) = 0, \quad \mathcal{F} = (f_1, \ldots, f_d)^T,$$

auf beliebigen Gittern, insbesondere auf Triangulierungen, gilt unter Voraussetzung periodischer Randbedingungen sowie der Verwendung eines E-Flusses die von Jiang und Shu in [48] bewiesene Zellentropie-Ungleichung. Auf Dreiecksgittern hat diese Ungleichung die Form

$$\frac{d}{dt}\int_{\tau_i} u_{h,N}^2(\mathbf{x},t)\,d\mathbf{x} + \sum_{j=1}^{3}\int_{\Gamma_{ij}} \tilde{q}_{ij}(u_{h,N}^i(\mathbf{x},t), u_{h,N}^{ij}(\mathbf{x},t), \mathbf{n}_{ij})\,d\sigma \leq 0, \quad i = 1,\ldots,\#\mathcal{T}^h.$$

Hierbei ist $\tilde{q}_{ij}$ eine numerische Flussfunktion, die konsistent zum Entropiefluss über die Kante Γ_{ij} ist, d.h. zu $\mathbf{q}(u)\cdot\mathbf{n}_{ij}$, mit $\mathbf{q}(u)$ definiert durch $\mathbf{q}(u) = \left(\int^u uf_1(u)du, \int^u uf_2(u)du\right)^T$. Im Fall einer eindimensionalen konvexen skalaren Erhaltungsgleichung garantiert nun die Gültigkeit der Zellentropie-Ungleichung, dass bei vorausgesetzter Konvergenz der Grenzwert der semidiskreten DG-Methode die eindeutige Entropielösung ist. Durch Summation der Zellentropie-Ungleichung über alle Dreiecke der Triangulierung erhält man aufgrund der vorausgesetzten periodischen Randbedingungen die (starke) L^2-Stabilität der semidiskreten Formulierung,

$$\frac{d}{dt}\int_{\Omega} u_{h,N}^2(\mathbf{x},t)d\mathbf{x} \leq 0,$$

d.h. es gilt

$$\|u_{h,N}(\cdot,t)\|_{L^2} \leq \|u_{h,N}(\cdot,0)\|_{L^2}, \quad \forall t > 0. \tag{4.12}$$

Diskontinuierliche Galerkin-Verfahren mit bestimmten impliziten Zeitdiskretisierungen, beispielsweise dem impliziten Euler- oder dem Crank-Nicholson-Verfahren, erfüllen eine analoge diskrete Zellentropieungleichung, aus der die L^2-Stabilität der vollständigen Diskretisierung folgt. Allerdings ist im Allgemeinen, insbesondere im Fall expliziter Zeitdiskretisierungen, nur bedingte Stabilität zu erwarten.

Lineare L^2-Stabilität Bekanntermaßen ist bei der Verwendung expliziter Verfahren zur Zeitintegration die Größe des Zeitschritts durch Stabilitätsanforderungen eingeschränkt. Zur Ermittlung einer zulässigen Zeitschrittweite Δt wird üblicherweise eine Stabilitätsanalyse für die skalare lineare Erhaltungsgleichung (2.20),

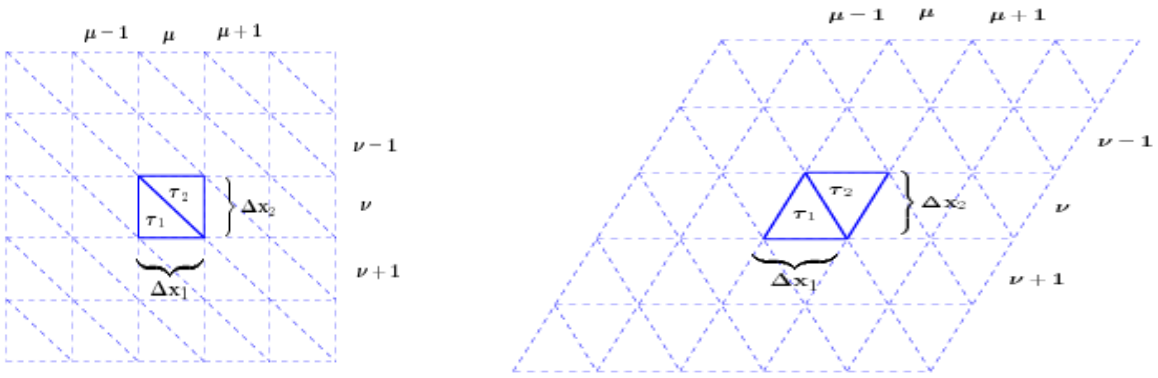
$$\frac{\partial}{\partial t}u + \mathbf{a}\cdot\nabla_{\mathbf{x}}u \;=\; 0,$$

auf Gittern mit periodischen Randbedingungen durchgeführt. Dadurch erhält man eine notwendige Bedingung der Form

$$\Delta t \leq cfl \cdot \frac{h}{\|\mathbf{a}\|_2}, \tag{4.13}$$

die nach Courant, Friedrichs und Lewy [25] als *CFL-Bedingung* bezeichnet wird. Hierbei steht der Parameter h für ein geeignet zu wählendes Längenmaß für die Größe der Elemente der räumlichen Diskretisierung und der Parameter $cfl \in (0,1]$ ist die sogenannte *CFL-Zahl*, die abhängig ist vom Polynomgrad N und der konkreten Zeitintegrationsmethode des gegebenen RKDG-Verfahrens. In einer Raumdimension sind CFL-Zahlen für verschiedene Polynomgrade und spezielle Runge-Kutta-Verfahren von Cockburn und Shu

berechnet worden, vgl. [23]. In [58] berechneten Kubatko et al. CFL-Zahlen für RKDG-Verfahren niedriger Ordnung (Polynomgrade bis $N = 3$) bei Verwendung verschiedener SSP-RK-Verfahren mit hoher (die Ordnung des Verfahrens übersteigender) Stufenanzahl. Die dort durchgeführte von Neumannsche Stabilitätsanalyse auf zwei verschiedenen strukturierten Gittern soll hier auf RKDG-Verfahren hoher Ordnung im Raum (Polynomgrade bis $N = 10$) unter Verwendung der in Abschnitt 4.2 aufgeführten expliziten RK-Verfahren übertragen werden.
Die betrachteten Gitter, dargestellt in Abbildung 4.1, werden jeweils erzeugt durch ein Grundmuster bestehend aus einem in zwei Dreiecke τ_1, τ_2 zerlegten Parallelogramm, welches in Blöcken wiederholt wird, die durch μ, ν indiziert sind. Die Elemente des Gittertyps I, der durch Zerlegen eines quadratischen Gitters entsteht, sind gleichschenklige Dreiecke mit einem rechten Winkel, während der Gittertyp II aus gleichseitigen Dreiecken besteht.



(a) Gittertyp I (b) Gittertyp II

Abb. 4.1: Zur Stabilitätsanalyse verwendete Gittertypen.

Werden die elementweisen Freiheitsgrade, d.h. die Koeffizienten $\hat{\mathbf{u}} = [\hat{u}_k]_{k=1,\ldots,\frac{(N+1)(N+2)}{2}}$, die zur besseren Lesbarkeit anstelle der Indizes l, m hier nur einen Index k tragen, auf jedem Block zusammengefasst zu

$$\mathbf{U}_{\mu,\nu} = \begin{bmatrix} \hat{\mathbf{u}}^{\mu,\nu,1} \\ \hat{\mathbf{u}}^{\mu,\nu,2} \end{bmatrix},$$

so lässt sich die semidiskrete Gleichung (4.3) in Blockform schreiben als

$$\frac{d}{dt}\mathbf{U}_{\mu,\nu} = \mathbf{S}\mathbf{U}_{\mu,\nu} - \mathbf{B}_{\pm}\mathbf{U}_{\mu\pm1,\nu} - \mathbf{C}_{\pm}\mathbf{U}_{\mu,\nu\pm1}\,. \tag{4.14}$$

Die Matrizen $\mathbf{S}, \mathbf{B}_{\pm}$ und $\mathbf{C}_{\pm}$ besitzen Blockstruktur und ihre Einträge sind insbesondere von der Strömungsrichtung abhängig. Die Matrix $\mathbf{S}$ setzt sich zusammen aus einem durch Integrale über die Dreiecksflächen gegebenen Anteil $\mathbf{S}_E$ sowie einem Anteil $\mathbf{S}_K$, der die Flüsse zum einen über Ausströmränder des Grundmusters und zum anderen über dessen innere Kante beinhaltet. Zur Konstruktion von $\mathbf{S}$ verwenden wir die folgenden Bezeichnungen:

$\lvert\tau\rvert$,	Flächeninhalt der Gitterelemente, $\lvert\tau\rvert := \lvert\tau_1\rvert = \lvert\tau_2\rvert$,
$\mathbf{A}_i$,	Jacobi-Matrix der Abbildung $\Lambda_i : \tau_i \to \mathbb{T}$ auf das Standarddreieck, $i = 1, 2$,
Φ_k^i	$:= \Phi_k \circ \Lambda_i$, auf τ_i transformierte Basisfunktionen,
$I_{\mathbf{a}}^i$	$:= \{j \in \{1,2,3\} \mid \mathbf{n}_{ij} \cdot \mathbf{a} \geq 0\,;\ \Gamma_{ij} \subset \partial(\tau_1 \cup \tau_2)\}$, Indexmenge zur Kennzeichnung der im Rand des Grundmusters liegenden Ausströmkanten Γ_{ij} von τ_i.

Die Matrix $\mathbf{S}$ ist dann gegeben durch

$$\mathbf{S} = \mathbf{S}_E - \mathbf{S}_K = \begin{pmatrix} \mathbf{S}_E^{1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_E^{2,2} \end{pmatrix} - \begin{pmatrix} \mathbf{S}_K^{1,1} & \mathbf{S}_K^{1,2} \\ \mathbf{S}_K^{2,1} & \mathbf{S}_K^{2,2} \end{pmatrix},$$

mit den Blöcken

$$\begin{aligned} \mathbf{S}_E^{i,i} &= \left[\frac{1}{\gamma_k} (\mathbf{A}_i \, \mathbf{a}) \cdot \int_{\mathbb{T}} \Phi_l \nabla_{r,s} \Phi_k \, dr ds \right]_{k,l}, \\ \mathbf{S}_K^{i,i} &= \left[\frac{2}{\gamma_k |\tau|} \, \mathbf{a} \cdot \sum_{j \in I_{\mathbf{a}}^i} \mathbf{n}_{ij} \int_{\Gamma_{ij}} \Phi_k^i \Phi_l^i \, d\sigma \right]_{k,l}, \end{aligned}$$

für $i = 1, 2$, sowie

$$\begin{aligned} \mathbf{S}_K^{i_1,i_2} &= \left[\frac{2}{\gamma_k |\tau|} \, \mathbf{a} \cdot \mathbf{n} \int_{\partial\tau_1 \cap \partial\tau_2} \Phi_k^{i_1} \Phi_l^{i_2} \, d\sigma \right]_{k,l}, \\ \mathbf{S}_K^{i_2,i_1} &= \mathbf{0}, \end{aligned}$$

wobei für die letzteren Blöcke die Indizes i_1, i_2 so gesetzt sind, dass die innere Kante $\partial\tau_1 \cap \partial\tau_2$ des Grundmusters Einströmkante für τ_{i_1} und Ausströmkante für τ_{i_2} ist und $\mathbf{n}$ den von τ_{i_1} nach τ_{i_2} gerichteten Normalenvektor an diese Kante bezeichnet. Die Matrizen $\mathbf{B}_\pm, \mathbf{C}_\pm$ beinhalten die zu den Flüssen über Einströmkanten des Grundmusters gehörigen Anteile der semidiskreten Gleichung und haben die Form

$$\mathbf{B}_+ = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{B}_K^{2,1} & \mathbf{0} \end{pmatrix}, \; \mathbf{B}_- = \begin{pmatrix} \mathbf{0} & \mathbf{B}_K^{1,2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \; \mathbf{C}_+ = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{C}_K^{2,1} & \mathbf{0} \end{pmatrix}, \; \mathbf{C}_- = \begin{pmatrix} \mathbf{0} & \mathbf{C}_K^{1,2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

mit Blöcken $\mathbf{B}_K^{i,j}, \mathbf{C}_K^{i,j}$, deren Einträge sich analog zur Berechnung der Blöcke $\mathbf{S}_K^{i,j}$ ergeben. Insbesondere gilt $\mathbf{B}_\pm \neq \mathbf{0} \Rightarrow \mathbf{B}_\mp = \mathbf{0}$ sowie $\mathbf{C}_\pm \neq \mathbf{0} \Rightarrow \mathbf{C}_\mp = \mathbf{0}$.

Die von Neumannsche Stabilitätsanalyse basiert auf der Entwicklung der Lösungen der Gleichung (4.14) in eine Fourier-Reihe. Sei dementsprechend

$$\mathbf{U}_{\mu,\nu}(t) = \mathbf{v}(t) \cdot e^{i(\mu k_1 \Delta x_1 + \nu k_2 \Delta x_2)}$$

eine periodische Lösung von (4.14) mit dem Wellenvektor $\mathbf{k} = (k_1, k_2)$ und den in Abbildung 4.1 gekennzeichneten Gitterweiten Δx_1 und Δx_2 in x_1- bzw. x_2- Richtung. Dann erfüllen die Amplituden die Differentialgleichung

$$\frac{d}{dt}\mathbf{v} = \mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha}) \cdot \mathbf{v},$$

mit der Matrix

$$\mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha}) = \mathbf{S} - \mathbf{B}_\pm e^{\pm i\alpha_1} - \mathbf{C}_\pm e^{\pm i\alpha_2},$$

wobei h ein geeignetes Längenmaß ist und $\alpha_1 = k_1 \Delta x_1, \alpha_2 = k_2 \Delta x_2$ gesetzt wird.
Die Anwendung eines Runge-Kutta-Verfahrens mit dem Zeitschritt Δt zur vollständigen Diskretisierung von (4.14) liefert die Berechnungsvorschrift

$$\mathbf{U}_{\mu,\nu}^{n+1} = P\left(\Delta t \, \mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha})\right) \mathbf{U}_{\mu,\nu}^n,$$

beziehungsweise

$$\mathbf{v}^{n+1} = P\left(\Delta t\, \mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha})\right) \mathbf{v}^n,$$

für die Amplituden. Mit P ist hierbei das charakteristische Polynom des gewählten Runge-Kutta-Verfahrens bezeichnet. Die starke Stabilität der diskreten Lösung in der diskreten L^2-Norm,

$$\|\mathbf{U}^n\|_2 \leq \|\mathbf{U}^0\|_2, \quad \forall n > 0,$$

analog zur Situation (4.12) im semidiskreten Fall, gilt genau dann, wenn die Bedingung $\|P\left(\Delta t\, \mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha})\right)\|_2 \leq 1$ erfüllt ist. Sie erfordert daher die notwendige Bedingung, dass der Zeitschritt des Verfahrens klein genug gewählt wird, so dass für jeden Eigenwert λ der Matrix $\mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha})$ der skalierte Wert $\Delta t\lambda$ im Stabilitätsgebiet $\mathcal{S} = \{z \in \mathbb{C} \,|\, |P(z)| \leq 1\}$ des gegebenen Runge-Kutta-Verfahrens liegt. Diese Bedingung ist zudem hinreichend, falls $\mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha})$ eine normale und damit unitär diagonalisierbare Matrix ist, allerdings tritt dieser Fall im Allgemeinen nur für den Polynomgrad $N = 0$ auf.
Die den in Abschnitt 4.2 aufgeführten Runge-Kutta-Verfahren der Ordnung $k = 1, \ldots, 4$ zugeordneten charakteristischen Polynome P_k sind gegeben durch

$$\begin{aligned}
P_1(z) &= 1 + z, \\
P_2(z) &= 1 + z + \frac{1}{2}z^2, \\
P_3(z) &= 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3, \\
P_4(z) &= 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4 + \frac{1}{200}z^5.
\end{aligned}$$

Die zugehörigen mit der Ordnung des RK-Verfahrens wachsenden Stabilitätsgebiete sind in Abbildung 4.2 dargestellt.
Für einen festen Gittertyp skalieren die Eigenwerte der Matrix $\mathbf{L}(h, \mathbf{a}, \boldsymbol{\alpha})$ offensichtlich linear mit der Transportgeschwindigkeit $\|\mathbf{a}\|_2$ sowie dem Kehrwert der Gitterweite h, so dass die Betrachtung normierter Vektoren $\mathbf{a} = (\cos\theta, \sin\theta)^T$ und Gitterweiten $h = 1$ ausreichend ist. Dies führt zur Definition der CFL-Zahl als

$$cfl = \max\{\Delta t \mid \forall(\theta, \boldsymbol{\alpha}) \in [0, 2\pi]^3 \; \forall \lambda \in \sigma(\mathbf{L}(1, \theta, \boldsymbol{\alpha})) : |P(\Delta t\lambda)| \leq 1\}. \tag{4.15}$$

In dieser Form ist der Wert von cfl noch abhängig vom konkreten Längenmaß h, für das in der Praxis beispielsweise die kürzeste Kante, der Inkreisradius oder die kürzeste Höhe des Dreiecks verwendet werden. Die Definition von h als die kürzeste Höhe des Dreiecks ist aus zweierlei Sicht vorteilhaft und wird deshalb auch in dieser Arbeit verwendet. Zum einen lassen sich bei dieser Wahl im Fall $N = 0$ die CFL-Zahlen auf Dreiecksgittern unabhängig vom Gittertyp zu eindimensionalen CFL-Bedingungen in Beziehung setzen, vgl. [58], zum anderen zeigt eine numerische Berechnung der CFL-Zahlen in [91] für RKDG-Verfahren mit der Zeitintegration (4.11) auf einer zweiparametrigen Schar von strukturierten Gittern, dass die Definition von h als die kleinste Höhe im Vergleich zu den genannten Alternativen die geringste Abhängigkeit der CFL-Zahl von der Verzerrung des Gitters liefert.
Für die in Abbildung 4.1 dargestellten Gittertypen wurden die CFL-Zahlen der RKDG-Verfahren für die Polynomgrade $N = 0, \ldots, 10$ und die in Abschnitt 4.2 angegebenen

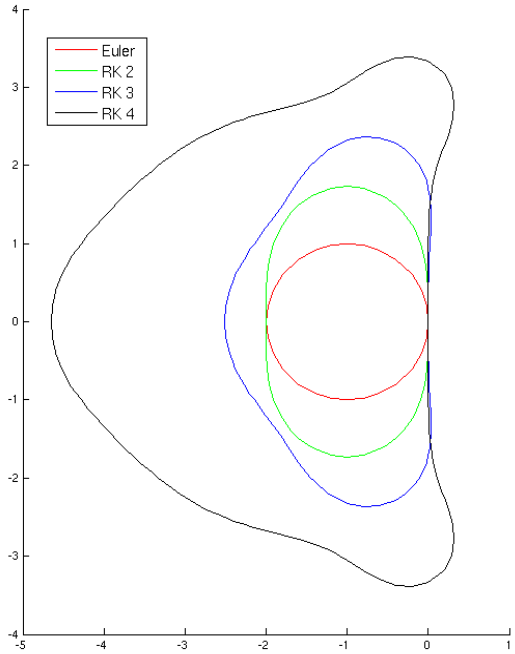


Abb. 4.2: Stabilitätsgebiete der in Abschnitt 4.2 aufgeführten RK-Verfahren.

Zeitintegrationsverfahren numerisch berechnet, die entsprechenden Werte sind Tabelle 4.1 aufgeführt. Hierbei wurden die Eigenwerte der Matrizen $\mathbf{L}(1, \theta, \boldsymbol{\alpha})$ für diskrete Werte von θ und $\boldsymbol{\alpha}$ numerisch berechnet und das Maximum analog zur Definition (4.15) über die entsprechende diskrete Menge gebildet. Ein Eintrag "–" in Tabelle 4.1 bedeutet, dass das RKDG-Verfahren im hier angenommenen Fall eines konstanten Verhältnisses des Zeitschritts Δt zur Gitterweite h für $h \to 0$ nicht stabil ist.

Gittertyp I											
N	0	1	2	3	4	5	6	7	8	9	10
Euler	0.5000	–	–	–	–	–	–	–	–	–	–
RK 2	0.5000	0.2449	–	–	–	–	–	–	–	–	–
RK 3	0.6282	0.3055	0.1732	0.1175	0.0824	0.0632	0.0488	0.0397	0.0324	0.0274	0.0232
RK 4	1.1116	0.5168	0.2981	0.2003	0.1417	0.1080	0.0837	0.0682	0.0557	0.0472	0.0400
Gittertyp II											
N	0	1	2	3	4	5	6	7	8	9	10
Euler	0.5000	–	–	–	–	–	–	–	–	–	–
RK 2	0.5000	0.2449	–	–	–	–	–	–	–	–	–
RK 3	0.6282	0.3030	0.1732	0.1173	0.0824	0.0632	0.0487	0.0397	0.0324	0.0275	0.0232
RK 4	1.1115	0.5084	0.2981	0.1993	0.1417	0.1080	0.0837	0.0682	0.0558	0.0472	0.0400

Tabelle 4.1: Numerisch ermittelte CFL-Zahlen für RKDG-Verfahren auf Dreiecksgittern.

In Bezug auf den Gittertyp unterscheiden sich die berechneten CFL-Zahlen nur geringfügig, allerdings zeigen die Berechnungen in [91] für eine größere Anzahl unterschiedlich stark verzerrter strukturierter Gitter Abweichungen der CFL-Zahlen von bis zu 12% des minimalen Wertes, zudem ergeben sich Unterschiede bei Verwendung eines Lax-Friedrich- anstelle des upwind-Flusses. Da in dieser Arbeit die Dämpfung von Oszillationen der nu-

merischen Lösung im Fall *nichtlinearer* Erhaltungsgleichungen mit möglicherweise *unstetigen* Lösungen im Vordergrund steht, muss sichergestellt werden, dass nicht die Verletzung linearer L^2-Stabilität eine grundsätzlich sinnvolle Dämpfungsstrategie als unzureichend erscheinen lässt. Aus diesem Grund wird die CFL-Zahl in den numerischen Experimenten jeweils auf höchstens 80% des kleineren der beiden in Tabelle 4.1 aufgeführten Werte gesetzt, vgl. Tabelle 4.2.

N	0	1	2	3	4	5	6	7	8	9	10
Euler	0.40	–	–	–	–	–	–	–	–	–	–
RK 2	0.40	0.19	–	–	–	–	–	–	–	–	–
RK 3	0.50	0.24	0.13	0.093	0.065	0.05	0.038	0.031	0.025	0.021	0.018
RK 4	0.88	0.40	0.23	0.159	0.113	0.086	0.066	0.054	0.044	0.037	0.032

Tabelle 4.2: In den numerischen Berechnungen verwendete CFL-Zahlen.

Im allgemeinen Fall der Diskretisierung eines nichtlinearen Systems hyperbolischer Erhaltungsgleichungen ,

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{x},t) + \nabla_{\mathbf{x}} \cdot \mathcal{F}(\mathbf{u}(\mathbf{x},t)) = \mathbf{0},$$

wird der globale Zeitschritt Δt^n ausgehend von den in Tabelle 4.2 gegebenen Werten gewählt als $\Delta t^n = \min_{i=1,\dots,\#\mathcal{T}^h} \Delta t_i^n$, mit dem lokal auf jedem Element berechneten maximal wählbaren Zeitschritt

$$\Delta t_i^n := cfl \cdot \frac{h(\tau_i)}{\max\limits_{\substack{k=1,\dots,m \\ \|\boldsymbol{\nu}\|_2=1}} |\lambda_k(\mathbf{u}^{n,i},\boldsymbol{\nu})|}\,. \tag{4.16}$$

Hierbei bezeichnet $\lambda_k(\mathbf{u}^{n,i},\boldsymbol{\nu})$ den k-ten Eigenwert der durch $\nu_1\mathbf{A}_1(\mathbf{u}^{n,i}) + \nu_2\mathbf{A}_2(\mathbf{u}^{n,i})$ gegebenen Jacobi-Matrix des Flusses $\mathcal{F}(\mathbf{u}^{n,i})\cdot\boldsymbol{\nu}$. Zur Bestimmung des Maximums in (4.16) wird die mit $\mathbf{u}^{n,i}$ bezeichnete numerische Lösung auf τ_i zum Zeitpunkt t^n ausgewertet an den $n_I + 3n_K$ Stützstellen, die zur Quadratur über die Dreiecksfläche sowie über die Dreieckskanten genutzt werden. Eine derartige Wahl des Zeitschritts in diesem allgemeinen Fall ist dadurch motivert, dass die Rolle der Transportgeschwindigkeit $\|\mathbf{a}\|_2$ in (4.13) offensichtlich von der maximalen charakteristische Ausbreitungsgeschwindigkeit im Nenner von (4.16) eingenommen wird. Im Fall der Euler-Gleichungen handelt es sich hierbei um die Größe

$$\max_{\|\boldsymbol{\nu}\|_2=1} |\mathbf{v}^{n,i}\cdot\boldsymbol{\nu} + c^{n,i}| = \max\{\|\mathbf{v}^{n,i}\|_2 + c^{n,i}\},$$

wie sich bei der Betrachtung der in (2.29) gegebenen Eigenwerte der Jacobi-Matrix ergibt.

4.3.2 Konvergenz

Die meisten Untersuchungen zur Konvergenzrate des DG-Verfahrens unter Gitterverfeinerung bei festem Polynomgrad N beziehen sich auf die semidiskrete DG-Formulierung oder die vollständige Diskretisierung durch den DG-Ansatz, d.h. der Einfluss eines gesonderten Zeitintegrationsverfahrens wird häufig nicht berücksichtigt.

Für die vollständige DG-Diskretisierung einer skalaren linearen Erhaltungsgleichung in Raum und Zeit unter Verwendung eines upwind-Flusses, beziehungsweise die DG-Diskretisierung eines stationären Problems im Raum, wie ursprünglich von Reed und Hill [79] zur Diskretisierung der Neutronentransport-Gleichung betrachtet, ergeben sich optimale Konvergenzraten von $\mathcal{O}\left(h^{N+1}\right)$ bei Verwendung kartesischer Gitter [62] oder speziell strukturierter Dreiecksgitter [80], während im allgemeinen Fall regulärer Triangulierungen (d.h. die Winkel der Dreieckselemente der Gitterfolge sind durch eine positive Konstante nach unten beschränkt) die optimale Konvergenzrate von $\mathcal{O}\left(h^{N+1/2}\right)$ nachgewiesen wurde, siehe [50, 73]. Im Fall zeitabhängiger skalarer linearer Erhaltungsgleichung wiesen Cockburn und Shu in [21] für die semidiskrete DG-Formulierung (4.3) auf regulären Triangulierungen eine L^2-Fehlerabschätzung von $\mathcal{O}\left(h^{N+1/2}\right)$ nach, unter der Voraussetzung, dass der verwendete numerische Fluss ein E-Fluss ist. Hierbei bezeichnet h den maximalen Durchmesser der Dreieckselemente des gegebenen Gitters.
Konvergenzresultate für die vollständige RKDG-Diskretisierung allgemeiner nichtlinearer Erhaltungsgleichungen wurden von Zhang und Shu in [102, 103] bewiesen. Bei Verwendung einer monotonen Flussfunktion sowie des RK-Verfahrens zweiter Ordnung (4.9) zur Zeitintegration ergibt sich danach im Fall skalarer Erhaltungsgleichungen mit glatter Lösung eine L^2-Fehlerabschätzung von $\mathcal{O}\left(h^{N+1/2} + \Delta t^2\right)$, siehe [102]. Hierbei bezeichnen h und Δt die maximale Elementlänge (betrachtet werden RKDG-Verfahren auf Viereckgittern) und die maximale Zeitschrittweite der jeweiligen konkreten Diskretisierung. Bei der Verwendung eines upwind-Flusses erhalten Zhang und Shu unter den obigen Voraussetzungen die optimale Konvergenzordnung $\mathcal{O}\left(h^{N+1} + \Delta t^2\right)$. In [103] wurden diese Konvergenzresultate für RKDG-Verfahren auf den Fall symmetrisierbarer hyperbolischer Systeme, wie beispielsweise die Euler-Gleichungen, übertragen.

Experimentelle Konvergenzordnung Die numerische Konvergenzordnung des DG-Verfahrens im Fall linearer Gleichungen mit glatter Lösung soll am Beispiel des Transports, siehe Gleichung (2.20), einer skalierten Gauß-Kurve

$$u_0(\mathbf{x}) = 0.2 \exp\left(-500 \cdot ((x_1 - 0.2)^2 + (x_2 - 0.3)^2)\right)$$

über unstrukturierte Dreiecksgitter untersucht werden. Die Randbedingungen für diesen Testfall wurden entsprechend der exakten Lösung $u(x_1, x_2, t) = u_0(x_1 - t, x_2 - t)$ gesetzt. Die Näherungen der DG-Methode wurden auf einer Hierarchie von Triangulierungen berechnet, bei der jede Gitterverfeinerung dadurch entsteht, dass die Dreieckselements des Gitters durch Verbinden ihrer drei Seitenhalbierenden jeweils in vier Teildreiecke zerlegt werden. Dies wird auch als Rot-Verfeinerung bezeichnet.
Das gröbste Gitter sowie die erste Rot-Verfeinerung sind in der Abbildung 4.3 gezeigt. Das gröbste Gitter besteht aus $K = 296$ Dreiecken während die verfeinerten Gitter $K = 1184$, $K = 4736$ bzw. $K = 18944$ Elemente besitzen.
Die Abbildung 4.4 zeigt den Anfangszustand sowie die Näherungslösung zum Zeitpunkt $T = 0.5$ für $N = 5$, berechnet auf dem gröbsten Gitter.
Zur Zeitintegration wurde das in Abschnitt 4.2 angegebene Runge-Kutta-Verfahren vierter Ordnung mit geringem Speicherverbrauch (4.11) verwendet, mit einem Zeitschritt von $\Delta t = 10^{-5}$, so dass der Fehler in der Zeit vernachlässigbar ist und somit die Konvergenzordnung der semidiskreten Formulierung 4.6 numerisch untersucht wird. In der Tabelle 4.3 sowie in der Abbildung 4.5 ist die experimentelle Konvergenzordnung für den Fehler der DG-Methode zum Ausgabezeitpunkt $T = 0.5$ in der L^2-Norm dargestellt. Während

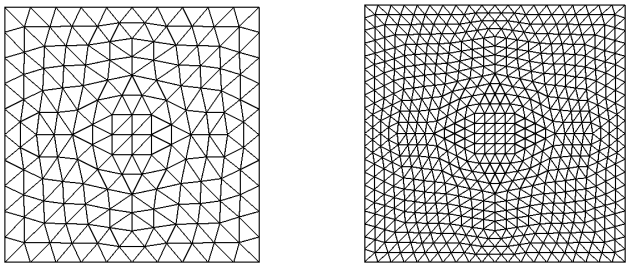


Abb. 4.3: Gröbstes Gitter mit 296 Dreiecken und erste Rot-Verfeinerung.

die Tabelle 4.3 die jeweiligen Fehler sowie die experimentelle Konvergenzordnung für die Polynomgrade $3 \leq N \leq 8$ verzeichnet, ist in der Abbildung 4.5 der L^2-Fehler in logarithmischer Skala in Abhängigkeit der Elementanzahl des jeweiligen Gitters dargestellt. Man erhält eine experimentelle Konvergenzordnung von $\mathcal{O}(h^{N+1})$ – mit Ausnahme des Falls $K = 18944$ und $N = 8$, in dem die Maschinengenauigkeit keinen geringeren Fehler ermöglicht – die besser ist, als die durch die Theorie gegebene Fehlerschätzung. Während diese höhere Konvergenzrate in der Praxis auch auf Triangulierungen vielfach beobachtet wurde, beispielsweise in [80, 45], zeigen Untersuchungen in [73], dass die theoretisch ermittelte Konvergenzrate von $\mathcal{O}(h^{N+1/2})$ strikt ist.

Desweiteren sind in den im Anhang zu findenden Tabellen A.1, A.2, A.3 und A.4 die numerischen Fehler in der L^1-, L^2- und L^∞-Norm sowie die Rechenzeiten in Sekunden verzeichnet, die sich bei Anwendung des RKDG-Verfahrens für Polynomgrade $N \leq 8$ und Zeitintegrationsverfahren verschiedener Ordnung ergeben, wenn die Zeitschrittwahl (4.16) mit den in der Tabelle 4.2 angegebenen CFL-Zahlen verwendet wird. Bei dieser Zeitschrittwahl ergeben die numerischen Berechnungen keine Stabilitätsprobleme. Für höhere Polynomgrade und feine Gitter führt die niedrigere Ordnung der Zeitintegration erwartungsgemäß zu einer Verschlechterung der experimentellen Konvergenzordnung, so dass für hohe Genauigkeitsanforderungen eine Anpassung der Ordnung der Zeitintegration in Erwägung gezogen werden sollte. Für $N > 3$ liefert die Verwendung der RK-Zeitintegration vierter Ordnung sowohl im Hinblick auf die Fehlerentwicklung als auch auf die Laufzeiten bessere Ergebnisse als das RK-Verfahren dritter Ordnung, so dass wir im Wesentlichen diese Zeitdiskretisierung verwenden werden.

4.3.3 Nichtlineare Erhaltungsgleichungen mit unstetigen Lösungen

Sowohl die Zellentropie-Ungleichung als auch die daraus folgende L^2-Stabilität (4.12) der semidiskreten DG-Formulierung sind insbesondere auch dann gültig, wenn die Lösung der Erhaltungsgleichung unstetig ist. Allerdings stellen diese Resultate im Fall unstetiger Lösungen keine ausreichende Kontrolle von Oszillationen der RKDG-Approximation im Bereich der Unstetigkeitsstellen der exakten Lösung dar, wie schon die Anwendung des Verfahrens auf die nichtlinearen Modellgleichungen (2.21) und (2.23) zeigt. Zur Zeitintegration wurde bei den folgenden Berechnungen das Runge-Kutta-Verfahren (4.11) vierter Ordnung verwendet. Analog zur Bedingung (4.16) wurde der Zeitschritt gewählt als

$$\Delta t^n := cfl \cdot \min_{\tau_i \in \mathcal{T}^h} \frac{h(\tau_i)}{\max\limits_{\mathbf{x} \in X_i} \|\nabla_u \mathbf{f}(u(\mathbf{x}, t^n))\|_2}\,, \tag{4.17}$$

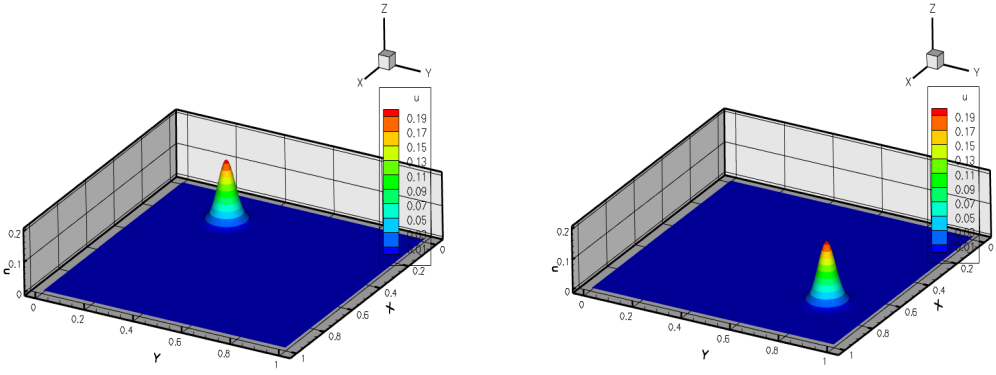


Abb. 4.4: Lineare Advektion: Anfangsbedingung und Näherung zum Zeitpunkt $T = 0.5$ für $N = 5$.

K	N	L^2-Fehler	EOC	N	L^2-Fehler	EOC
296	3	1.171084e-03		4	3.521460e-04	
1184		6.818670e-05	4.102210e+00		7.670696e-06	5.520672e+00
4736		2.774005e-06	4.619448e+00		2.138201e-07	5.164888e+00
18944		1.681435e-07	4.044205e+00		6.559287e-09	5.026715e+00
296	5	7.996430e-05		6	2.044012e-05	
1184		1.059437e-06	6.237986e+00		1.311833e-07	7.283676e+00
4736		1.540299e-08	6.103944e+00		1.041515e-09	6.976757e+00
18944		2.389907e-10	6.010112e+00		8.228855e-12	6.983776e+00
296	7	4.142200e-06		8	1.007003e-06	
1184		1.855943e-08	7.802101e+00		2.157085e-09	8.866769e+00
4736		7.270799e-11	7.995823e+00		4.558819e-12	8.886207e+00
18944		3.075400e-13	7.885196e+00		1.279585e-13	5.154912e+00

Tabelle 4.3: Experimentelle Konvergenzordnung (EOC) für die lineare Transportgleichung.

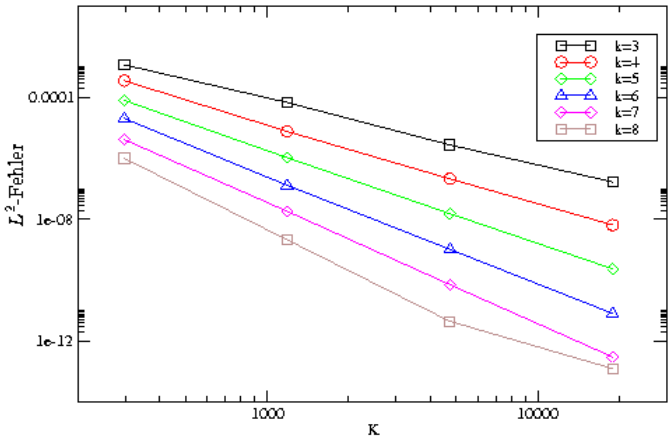
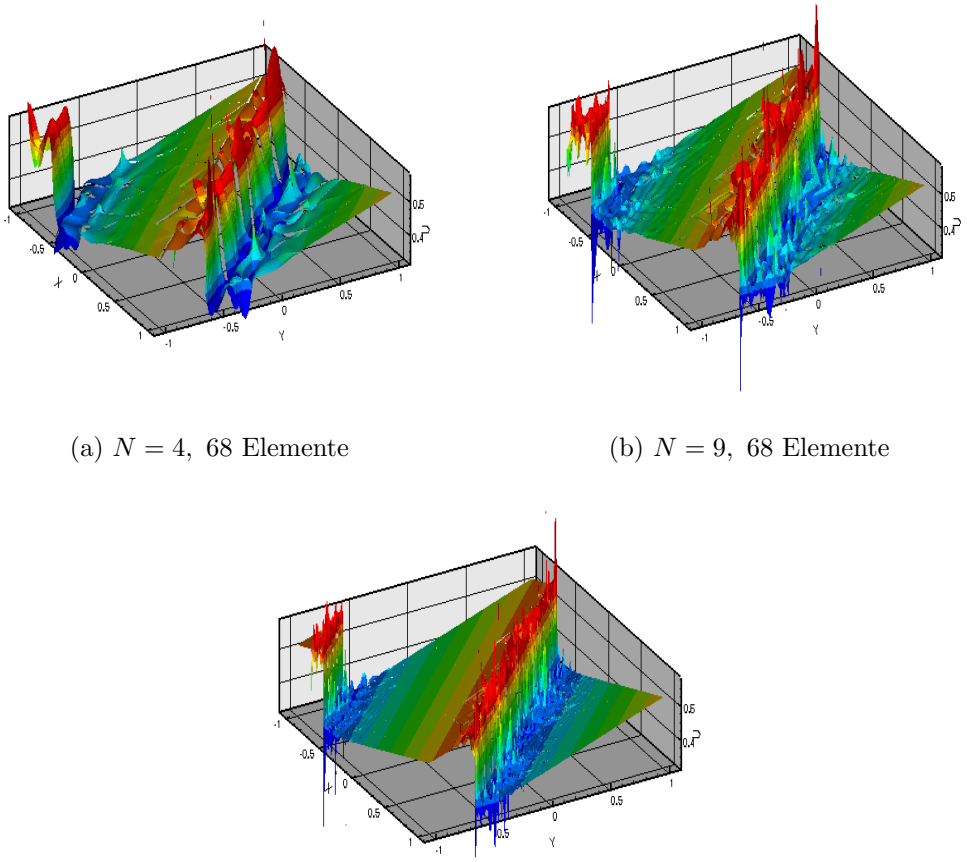


Abb. 4.5: L^2-Fehler unter Gitterverfeinerung. K ist die Elementanzahl.



(a) $N = 4,\ 68$ Elemente

(b) $N = 9,\ 68$ Elemente

(c) $N = 9,\ 272$ Elemente

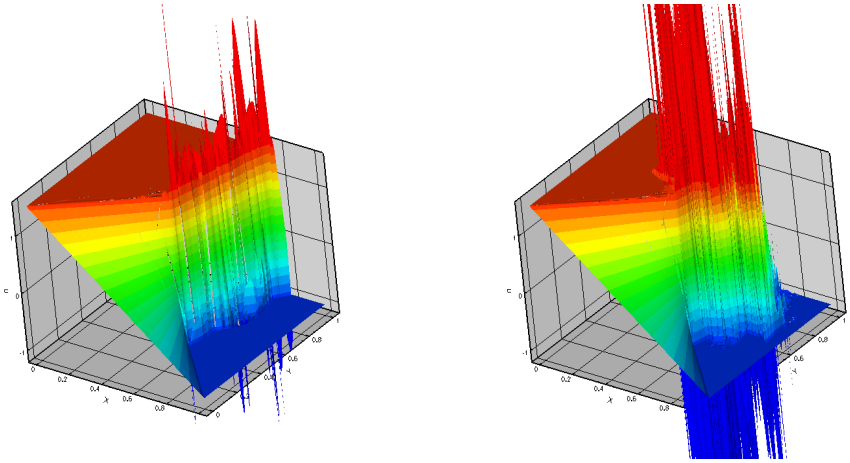
Abb. 4.6: Näherungslösungen des RKDG-Verfahrens zur Gleichung (2.23) zum Zeitpunkt $T = 1.4$.

mit $\|\nabla_u \mathbf{f}(u)\|_2 = \sqrt{2}|u|$ für die Gleichung (2.23) und $\|\nabla_u \mathbf{f}(u)\|_2 = \sqrt{u^2+1}$ für die Gleichung (2.21), mit dem numerisch ermittelten Parameter cfl aus der Tabelle 4.2 und der Bestimmung der maximalen Ausbreitungsgeschwindigkeit über die Menge X_i der im DG-Verfahren verwendeten Stützstellen auf τ_i.

In Abbildung 4.6 sind die Näherungslösungen zum Zeitpunkt $T = 1.4$ des auf die Burgers-Gleichung (2.23) angewendeten RKDG-Verfahrens dargestellt. Offensichtlich weisen die Näherungen starke Oszillationen nahe des Stoßes auf.

Die approximative Berechnung der stationären Lösung der Modellgleichung (2.21) unter Verwendung hoher Polynomgrade N stagniert sogar. Während sich für den Polynomgrad $N = 5$ und das in Abbildung 4.3 gezeigte grobe Gitter noch eine – sehr stark oszillierende – Näherungslösung berechnen lässt, siehe Abbildung 4.7a, liefert eine näherungsweise Berechnung der räumlichen L^∞- und L^2-Normen der Näherungslösungen für Polynomgrade $N \geq 8$ mit der Zeit sehr stark steigende Werte, die dazu führen, dass der Zeitschritt entsprechend der Bedingung (4.17) verschwindend gering wird. Zur Veranschaulichung ist die ermittelte stark oszillative Näherungslösung für $N = 8$ zum Zeitpunkt $T = 0.9$ in Abbildung 4.7b dargestellt.

Man könnte vermuten, dass die Verwendung kleinerer Zeitschritte während der gesamten



(a) $N = 5,\ T = 2$ (b) $N = 8,\ T = 0.9$

Abb. 4.7: Näherungslösungen des RKDG-Verfahrens zur Gleichung (2.21).

N	θ	t_{stag}	L^2-Norm	L^∞-Norm
8	1	0.95	1.5e+09	1.5e+11
	0.5	0.85	6.0e+08	4.9e+10
	0.1	0.99	1.4e+08	1.1e+10
	0.01	0.91	1.9e+07	1.4e+09
9	1	0.76	2.0e+09	9.2e+10
	0.5	0.82	3.9e+08	4.1e+10
	0.1	0.86	2.9e+08	1.4e+10
	0.01	0.84	1.1e+07	1.2e+09
10	1	0.56	1.8e+09	1.0e+11
	0.5	0.56	3.7e+08	6.4e+10
	0.1	0.56	2.1e+08	1.1e+10
	0.01	0.56	1.3e+07	9.6e+08

Tabelle 4.4: Stagnationszeiten t_{stag} sowie L^2- und L^∞-Normen der Näherungslösungen zur Gleichung (2.21).

Berechnung Abhilfe schafft. Ersetzen wir in (4.17) den Parameter cfl durch $cfl^* = \theta \cdot cfl$, unter Verwendung eines Sicherheitsfaktors $\theta \leq 1$, und definieren wir die Berechnung als stagniert zum Zeitpunkt $t_{stag} := t^n$, wenn $\Delta t^n < 10^{-14}$, so liefern diverse Berechnungen jedoch die in Tabelle 4.4 angegebenen Resultate, die anzeigen, dass Konvergenz (in der Zeit) gegen eine stationäre Lösung nicht gegeben ist.

Ansätze zur Stabilisierung der DG-Methode im nichtlinearen Fall Um ein derartiges Anwachsen der numerischen Lösung im Fall nichtlinearer Erhaltungsgleichungen mit unstetiger Lösung zu verhindern, wurden in der Literatur bisher verschiedene Dämpfungsstrategien vorgeschlagen.
Cockburn und Shu verwenden einen modifizierten minmod-Limiter, der den linearen Anteil der Entwicklung in polynomiale Basisfunktionen innerhalb einer Zelle mit den Differenzen des dort gegebenen Zellmittelwertes zu den Zellmittelwerten der benachbarten Elemente vergleicht, anhand dieser Daten gegebenenfalls den Polynomgrad in der betrachteten Zelle auf $N = 1$ reduziert und zusätzlich die Steigung limitiert. Die Modifikation

des ursprünglichen minmod-Limiters von van Leer besteht in der Erweiterung um die von einem benutzerdefinierten Parameter $M > 0$ abhängige Bedingung, dass die numerische Lösung auf einer Zelle nur dann modifiziert wird, wenn der Betrag der dort ermittelten Steigung größer ist als Mh^2. Durch diese Modifikation lässt sich im Fall einer eindimensionalen skalaren Erhaltungsgleichung, diskretisiert auf Zerlegungen des Rechengebiets Ω in Intervalle $\Omega = \sum_i \Omega_i$, die Beschränktheit der Totalvariation $TV(u_{h,N}) = \sum_i |u_{h,N}^{i+1} - u_{h,N}^i|$ der numerischen Lösung des RKDG-Verfahrens nachweisen, falls eine TVD-Zeitintegration verwendet wird. Für den eindimensionalen skalaren Fall folgt daraus die Konvergenz einer Teilfolge der unter Gitterverfeinerung erzeugten Näherungslösungen gegen eine schwache Lösung der Erhaltungsgleichung. Zudem überträgt sich die TVB-Eigenschaft auf lineare Systeme, falls die Limitierung in charakteristischen Variablen ausgeführt wird. Für skalare Erhaltungsgleichungen in mehreren Raumdimensionen bleibt die Gültigkeit eines lokalen Maximumprinzips bestehen, welches jedoch keine Konvergenzaussage erlaubt.
Ein inhärentes Problem der von Cockburn und Shu vorgeschlagenen Limitierung ist die Abhängigkeit von dem schwer zu determinierenden Parameter M, der die Genauigkeit an Extremstellen der Lösung erhalten soll. Zudem gehen die in den Koeffizienten zu den Basisfunktionen höheren als ersten Grades enthaltenen Informationen bei der Limitierung verloren. In [77, 105, 75, 76] wurden daher WENO- beziehungsweise HWENO-Rekonstruktionen verwendet, die die eigentliche Limitierung der Näherungslösung ausführen, während der Limiter vom minmod-Typ von Cockburn und Shu ausschließlich als ein Indikator dient zur Ermittlung "gefährdeter" Zellen (*"troubled cells"*), die dieser Rekonstruktionsprozedur zu unterwerfen sind. Desweiteren wurden in den Arbeiten [5, 55, 100] sogenannte Momentenlimiter entworfen, die alle Koeffizienten in die Limitierung einbeziehen. Allerdings sind die beiden letztgenannten Vorgehensweisen relativ rechenintensiv.
Andere Ansätze zur Stabilisierung der DG-Methode, beispielsweise von Jaffre, Johnson und Szepessy [47], von Feistauer und Kučera [31] sowie von Persson und Peraire [72] beinhalten den Einbau von expliziten Dämpfungstermen in die semidiskrete Form oder die Erhaltungsgleichung selbst, die dann geeignet zu diskretisieren sind. Für die DG-Methode von Jaffre, Johnson und Szepessy kann mit Hilfe der Theorie maßwertiger Lösungen von DiPerna [27] im Fall skalarer Erhaltungsgleichungen (auch in mehreren Raumdimensionen) Konvergenz gegen die Entropielösung gezeigt werden.
Die vorliegende Arbeit verfolgt ebenso den Ansatz der Einführung expliziter Viskosität in die DG-Methode, hierbei in der Form sogenannter spektraler Viskosität. Eine derartige Strategie wurde von Tadmor für die Verfahrensklasse der Spektralmethoden eingeführt – ihre Anpassung und Verwendung im Kontext von DG-Verfahren auf Dreiecksgittern ist Gegenstand des nachfolgenden Kapitels.

5 Modale Filter und spektrale Viskosität für DG-Verfahren auf Dreiecksgittern

Zur Einführung expliziter Viskosität in die DG-Methode orientieren wir uns an der von Tadmor entwickelten Methode spektraler Viskosität für den Verfahrenstyp der Spektralmethoden. Der Vorteil dieses Zugangs gegenüber den am Ende des letzten Kapitels genannten Strategien liegt in der Möglichkeit der effizienten Implementierung durch modale Filter, die direkt auf den Koeffizienten der Näherungslösung wirken. In diesem Kapitel wird zunächst das Gibbsche Phänomen beschrieben, welches den Grundstein legt für die Problematik der Anwendung von Spektralverfahren auf hyperbolische Erhaltungsgleichungen mit unstetiger Lösung. Anschließend werden modale Filter sowie die Formulierung spektraler Viskosität erläutert. Kern des Kapitels ist die Herleitung eines modalen Filters für DG-Verfahren auf Dreiecksgittern zur effizienten Implementation von auf dem Standarddreieck definierter spektraler Viskosität. Abschließend wird die Notwendigkeit der adaptiven Verwendung des modalen Filters zur Steuerung der Viskosität des DG-Verfahrens nachgewiesen und es werden mögliche Stoßindikatoren vorgestellt.

5.1 Das Gibbs-Phänomen

Da die Verfahrensklassen der spektralen Verfahren, der Spektral-Element-Verfahren, sowie der an diese Methoden angelehnten diskontinuierlichen Galerkin-Verfahren hoher Ordnung auf der Entwicklung der Lösung einer Differentialgleichung mittels orthogonaler Basisfunktionen basieren, sind zunächst die Eigenschaften derartiger Reihenentwicklungen für das Verhalten der Verfahren von Bedeutung. Während die Projektion einer glatten Funktion in einen derartigen Ansatzraum eine sehr gute punktweise Näherung an die gegebene Funktion darstellt, wie in Satz 3.3 für die Entwicklung in PKD-Polynome gezeigt, besitzt die Entwicklung einer nur stückweise stetigen Funktionen einige ungünstige Eigenschaften, die die Verwendung spektraler Verfahren im Kontext hyperbolischer Erhaltungsgleichungen erschweren. Hierzu wird häufig auf das sogenannte *Gibbssche Phänomen* verwiesen, welches bei der Entwicklung einer unstetigen oder nichtperiodischen Funktion in eine Fourier-Reihe auftritt. Betrachtet man hierzu beispielsweise die Sprungfunktion

$$u(x) = \begin{cases} -1, & -\pi < x < 0, \\ 1, & 0 < x < \pi, \end{cases}$$

die die Fourier-Reihe $\frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin[(2n-1)x]}{2n-1}$ besitzt, so veranschaulichen die in Abbildung 5.1 dargestellten Partialsummen der Länge $N = 3, 8, 20$, dass in der Nähe der Unstetigkeitstelle $x = 0$ der Funktion u sowie an den Intervallrändern keine gleichmäßige Konvergenz der Fourier-Reihe zu erwarten ist. Durch eine genaue Untersuchung wie in [106, S. 61] lässt sich nachweisen, dass die jeweiligen maximalen Über- und Unterschwinger der Partialsummen für $N \to \infty$ gegen den Wert $\pm\frac{2}{\pi} \int_0^{\pi} \frac{\sin t}{t} dt \approx \pm 1.18$ konvergieren.
Während das Gibbs-Phänomen im ursprünglichen Sinn genau dieses Fehlen gleichmäßiger Konvergenz bezeichnet, umfasst der Begriff aus der Sicht der spektralen Verfahren – unter anderem nach dem Verständnis von Gottlieb und Shu [39] – ebenso den globalen Genauigkeitsverlust der punktweisen Approximation durch eine abgeschnittene Fourier-Entwicklung. So konvergiert der punktweise Fehler für $x \in (-\pi, 0) \cup (0, \pi)$ im obigen Beispiel nur mit einer Rate von $\mathcal{O}(1/N)$ gegen Null.

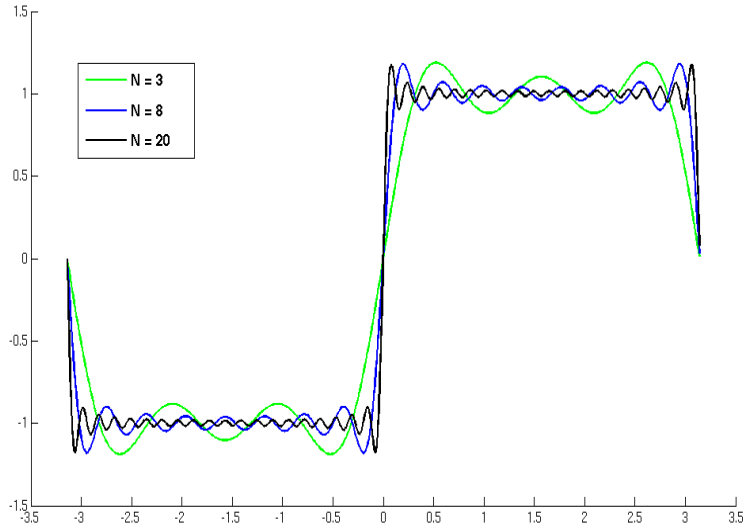


Abb. 5.1: Zur Funktion u gehörige Fourier-Partialsummen für $N = 3, 8, 20$.

Nachteilig für die Nutzung spektraler Verfahren insbesondere für nichtlineare hyperbolische Erhaltungsgleichungen, deren Lösungen Unstetigkeiten entwickeln können, wirken sich zudem die über das gesamte Intervall ausgebreiteten Oszillationen aus. Diese Oszillationen werden in diesem Kontext auch als *Gibbssche Oszillationen* bezeichnet.
Obwohl sich der Begriff des Gibbs-Phänomens genau genommen auf den Spezialfall von Fourier-Entwicklungen bezieht, lassen sich die fehlende gleichmäßige Konvergenz, der globale Genauigkeitsverlust sowie das Auftreten von Oszillationen ebenfalls unter anderem bei Chebyshev- und Legendre-Entwicklungen unstetiger Funktionen beobachten, wie in [38] beschrieben wird.

5.2 Modale Filter

Eine effiziente Methode zur Reduktion der Gibbsschen Oszillationen ist eine direkte Modifikation der Koeffizienten der Reihenentwicklung mittels eines *modalen Filters*. Die Beschreibung derartiger Filter ist in der Literatur zu spektralen Verfahren verbreitet und beispielsweise in den Büchern [43, 11, 12, 52] zu finden.

Definition 5.1 *Sei $p \geq 1$ eine natürliche Zahl und sei $\sigma \in C^{p-1}[0,1]$ mit Werten $\sigma(\eta) \in [0,1]$. Wir bezeichnen σ als einen* Filter der Ordnung p*, falls die Eigenschaften*

$$\sigma(0) = 1, \tag{5.1}$$
$$\sigma^{(l)}(0) = 0, \quad 1 \leq l \leq p-1, \tag{5.2}$$

erfüllt sind.

Mit Hilfe des Filters werden die in spektralen Verfahren auftretenden Reihenentwicklungen modifiziert. Man betrachte dazu eine Funktion u_N der Form

$$u_N = \sum_{k \in I_N} \hat{u}_k \Phi_k, \quad N \geq 1,$$

mit Basisfunktionen Φ_k, zugehörigen Koeffizienten $\hat{u}_k$ und einer von der Wahl der Basisfunktionen abhängigen Indexmenge I_N, die zum Beispiel durch $I_N = \{k \in \mathbb{Z} \mid |k| \leq N\}$ für Fourier-Entwicklungen, durch $I_N = \{k \in \mathbb{N}_0 \mid k \leq N\}$ für Chebychev- oder Legendre-Entwicklungen, sowie durch $I_N = \{(l,m) \in \mathbb{N}_0^2 \mid l+m \leq N\}$ für PKD-Entwicklungen gegeben ist. Von einem modalen Filter σ spricht man, wenn σ direkt auf den Koeffizienten der Entwicklung wirkt, so dass die gefilterte Funktion die Form

$$u_N^\sigma = \sum_{k \in I_N} \sigma\left(\frac{|k|}{N}\right) \hat{u}_k \Phi_k$$

annimmt, mit der Vereinbarung $|(l,m)| = l+m$. Die modale Filterung modifiziert demnach die Koeffizienten unterschiedlich stark, abhängig davon, ob es sich bei den zugehörigen Termen um hoch- oder niedrigfrequente Anteile der Funktion u_N handelt.
Soll durch die Anwendung eines modalen Filters neben der Reduktion hochfrequenter Oszillationen ebenso der Genauigkeitsverlust im Fall der Entwicklung unstetiger Funktionen behoben und die Konvergenzrate fern der Unstetigkeitsstellen verbessert werden, so wird von einem Filter σ der Ordnung p die zusätzliche Bedingung

$$\sigma^{(l)}(1) \;=\; 0, \quad 0 \leq l \leq p-1\,, \tag{5.3}$$

verlangt. Diese Bedingung beschreibt, wie stark die hochfrequenten Anteile der Entwicklung "weggefiltert" werden. Handelt es sich bei der Funktion u_N um eine abgeschnittene Reihenentwicklung, so sorgt (5.3) dafür, dass bei der gefilterten Funktion u_N^σ der Übergang zum Weglassen der Terme für $|k| > N$ umso glatter ist, je höher die Filterordnung p gewählt wurde. Die Bedeutung der Zusatzbedingung (5.3), die von manchen Autoren auch in die Definition eines Filters p-ter Ordung aufgenommen wird, wurde zuerst von Vandeven erkannt, der in [92] den Einfluss modaler Filterung auf Fourier-Entwicklungen glatter und stückweise stetiger Funktionen analytisch untersuchte.
Für eine 2π-periodische Funktion $u : \mathbb{R} \to \mathbb{R}$ mit der abgeschnittenen Fourier-Entwicklung

$$u_N(x) = \sum_{|k| \leq N} \hat{u}_k e^{ikx}, \qquad \hat{u}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x) e^{-ikx} dx,$$

und deren gefilterte Partialsumme

$$u_N^\sigma(x) = \sum_{|k| \leq N} \sigma\left(\frac{|k|}{N}\right) \hat{u}_k e^{ikx}\,.$$

zeigte Vandeven die folgenden Fehlerschranken.

Satz 5.2 *Sei σ ein Filter der Ordnung p und sei $u : \mathbb{R} \to \mathbb{R}$ eine 2π-periodische, stückweise glatte Funktion. Dann gelten die folgenden Abschätzungen:*
1) Sei $u \in C^p(\mathbb{R})$. Dann gilt

$$|u(x) - u_N^\sigma(x)| \leq C_1 \cdot \frac{1}{N^{p-1/2}}\,, \quad x \in \mathbb{R}, \tag{5.4}$$

mit einer von x und N unabhängigen Konstanten C_1.
2) Der Filter σ erfülle die zusätzliche Bedingung (5.3)*. Falls u eine oder mehrere Sprungstellen besitzt und an der Stelle x stetig ist, so gilt*

$$|u(x) - u_N^\sigma(x)| \leq C_2 \cdot \frac{1}{[d(x)]^{p-1} N^{p-1}}, \tag{5.5}$$

wobei mit $d(x)$ der Abstand von x zur nächstgelegenen Sprungstelle bezeichnet und C_2 eine von x und N unabhängige Konstante ist.

Bemerkung 5.3 *Die Konstanten C_1 und C_2 in Satz 5.2 sind abhängig von der gegebenen Funktion u sowie der Wahl des Filters σ.*

Die theoretische Untersuchung des asymptotischen Verhaltens allgemeiner modal gefilterter Reihendarstellungen ist noch unvollständig. Zunächst gelten die in Satz 5.2 aufgeführten Approximationseigenschaften ebenso für Chebyshev-Reihen, da – wie auch in [92] angemerkt wurde – die Chebychev-Koeffizienten einer Funktion $v : [-1, 1] \to \mathbb{R}$ proportional zu den Fourier-Koeffizienten von $u(\theta) = v(\cos\theta)$ sind. Konkret erhält man für die gewichtete Integration von v gegen das k-te Chebyshev-Polynom $T_k(x) = \cos(k \cdot \arccos x)$ durch die Substitution $x = \cos\theta$ die Beziehung

$$\int_{-1}^{1} v(x)T_k(x)(1-x^2)^{-1/2}dx = \frac{1}{2}\int_{-\pi}^{\pi} u(\theta)\cos k\theta\, d\theta.$$

Da u eine gerade Funktion ist, lassen sich aufgrund der Orthogonalitätseigenschaft der Chebyshev-Polynome,

$$\int_{-1}^{1} T_n(x)T_m(x)(1-x^2)^{-1/2}dx = \begin{cases} 0 & \text{für} \quad n \neq m, \\ \pi & \text{für} \quad n = m = 0, \\ \pi/2 & \text{für} \quad n = m \neq 0, \end{cases}$$

die Chebyschev-Koeffizienten $\hat{v}_k^C$ von v aus den Fourier-Koeffizienten $\hat{u}_k^F$ berechnen durch

$$\hat{v}_0^C = \hat{u}_0^F, \qquad \hat{v}_k^C = 2\hat{u}_k^F = 2\hat{u}_{-k}^F,\ k \geq 1.$$

Für die gefilterte Chebychev-Entwicklung von v an der Stelle $x = \cos\theta$ gilt daher

$$v_N^{C,\sigma}(\cos\theta) = \sum_{0\leq k\leq N} \sigma\left(\frac{k}{N}\right)\hat{v}_k^C \cos k\theta = \sum_{|k|\leq N} \sigma\left(\frac{|k|}{N}\right)\hat{u}_k^F e^{ik\theta} = u_N^{F,\sigma}(\theta),$$

so dass man für den punktweisen Fehler der gefilterten Chebychev-Entwicklung die Darstellung

$$v_N^{C,\sigma}(\cos\theta) - v(\cos\theta) = u_N^{F,\sigma}(\theta) - u(\theta),$$

erhält und die Aussagen des Satzes 5.2 ebenso für die Entwicklung einer Funktion in eine Chebychev-Reihe gültig sind. Desweiteren untersuchten Hesthaven und Kirby [44] das asymptotisches Verhalten modal gefilterter Legendre-Entwicklungen mit dem Resultat einer obere Schranke von $\mathcal{O}(N^{1-p})$ für hinreichend glatte Funktionen, so dass die Aussage (5.4) ebenso auf den Fall von Legendre-Reihen übertragen wurde. Allerdings ist der Nachweis einer Abschätzung der Form (5.5) bei Vorliegen einer unstetigen Funktion bisher nur für den speziellen Fall erbracht, dass sich genau eine Sprungstelle in der Mitte des Intervalls befindet und x einer der Randpunkte ist, siehe [44]. Resultate für nur stückweise stetige Funktionen liegen auch außerhalb der Reichweite dieser Arbeit. Wir beschränken uns hingegen auf die Übertragung der Resultate von Vandeven sowie Hesthaven und Kirby auf PKD-Entwicklungen hinreichend glatter Funktionen.

Für gefilterte PKD-Partialsummen

$$\begin{aligned} u_N^\sigma(r,s) &= \sum_{l+m\le N} \sigma\left(\frac{l+m}{N}\right)\hat{u}_{lm}\Phi_{lm}(r,s), \quad N\ge 1, \\ \hat{u}_{lm} &= \frac{1}{\gamma_{lm}}\int_{\mathbb{T}} u(r,s)\Phi_{lm}(r,s)\,drds, \end{aligned}$$

hinreichend glatter Funktionen $u : \mathbb{T} \to \mathbb{R}$ lässt sich durch eine geeignete Modifikation der Beweise in [92, 44] die folgende Abschätzung nachweisen, siehe auch [69].

Satz 5.4 *Sei $u \in H^{2p}(\mathbb{T})$, $p > 1$, und sei σ ein Filter der Ordnung $2p-1$, mit der zusätzlichen Eigenschaft $\sigma \in C^{2p-1}[0,\epsilon)$ in einem Teilintervall $[0,\epsilon) \subset [0,1]$, $\epsilon > 0$. Dann gilt*

$$\begin{aligned} |u(r,s) - u_N^\sigma(r,s)| &< C_1 \cdot \frac{1}{(1-s)^{3/4}N^{2p-2}}, \quad (r,s)\in\mathbb{T}\backslash\{(-1,1)\}, \\ |u(-1,1) - u_N^\sigma(-1,1)| &< C_2 \cdot \frac{1}{N^{2p-2}}. \end{aligned}$$

mit Konstanten C_1 und C_2.

Beweis: Da u hinreichend glatt ist, erhält man unter Verwendung des in Kapitel 3 beschriebenen Sturm-Liouville-Operators $\mathcal{L}_{r,s}$ mit (3.18) die Gleichung

$$\hat{u}_{lm} = \frac{1}{\gamma_{lm}}\int_{\mathbb{T}} u(\tilde{r},\tilde{s})\Phi_{lm}(\tilde{r},\tilde{s})\,d\tilde{r}d\tilde{s} = \frac{1}{\gamma_{lm}(-\lambda_{lm})^p}\int_{\mathbb{T}} \mathcal{L}^p_{\tilde{r},\tilde{s}}u(\tilde{r},\tilde{s})\Phi_{lm}(\tilde{r},\tilde{s})\,d\tilde{r}d\tilde{s}$$

für $l+m>0$, und mit

$$u(r,s) = \sum_{l,m\in N_0}\hat{u}_{lm}\Phi_{lm}(r,s),$$

siehe auch Gleichung (3.20), ergibt sich

$$\begin{aligned} |u(r,s) - u_N^\sigma(r,s)| &= \left|\sum_{l+m\le N}\left(1-\sigma\left(\frac{l+m}{N}\right)\right)\hat{u}_{lm}\Phi_{lm}(r,s) + \sum_{l+m>N}\hat{u}_{lm}\Phi_{lm}(r,s)\right| \\ &= \left|\int_{\mathbb{T}}\left[Q_N(r,s,\tilde{r},\tilde{s}) + R_N(r,s,\tilde{r},\tilde{s})\right]\mathcal{L}^p_{\tilde{r},\tilde{s}}u(\tilde{r},\tilde{s})\,d\tilde{r}d\tilde{s}\right|, \end{aligned} \tag{5.6}$$

wobei $Q_N(r,s,\tilde{r},\tilde{s})$ und $R_N(r,s,\tilde{r},\tilde{s})$ durch

$$\begin{aligned} Q_N(r,s,\tilde{r},\tilde{s}) &= \sum_{1\le l+m\le N}\left(1-\sigma\left(\frac{l+m}{N}\right)\right)\frac{\Phi_{lm}(\tilde{r},\tilde{s})\Phi_{lm}(r,s)}{\gamma_{lm}(-\lambda_{lm})^p}, \\ R_N(r,s,\tilde{r},\tilde{s}) &= \sum_{l+m>N}\frac{\Phi_{lm}(\tilde{r},\tilde{s})\Phi_{lm}(r,s)}{\gamma_{lm}(-\lambda_{lm})^p} \end{aligned}$$

gegeben sind. Aus dem Beweis zur in Satz 3.3 gegebenen Abschätzung (3.13) entnimmt man

$$\|R_N(r,s,\tilde{r},\tilde{s})\|^2_{L^2(\mathbb{T})} < \left(\frac{2}{1-s}\right)^{3/2}\frac{2}{N^{4p-4}}, \qquad \text{für } s\ne 1,$$

sowie

$$\|R_N(-1,1,\tilde{r},\tilde{s})\|^2_{L^2(\mathbb{T})} \;<\; \frac{1}{N^{4p-4}}\,.$$

Analog zum Beweis der Abschätzung (3.13) erhält man mit Hilfe von Lemma 3.5 eine obere Schranke für die L^2-Norm von Q_N.
Für $s \neq 1$ ergibt sich unter Verwendung der oberen Schranke (3.14) für die Werte der Jacobi-Polynome und mit der Ungleichung $\lambda_{lm}^{-2p}\gamma_{lm}^{-1} < (l+m)^{-4p+2}$ die Abschätzung

$$\begin{aligned}\|Q_N(r,s,\tilde{r},\tilde{s})\|^2_{L^2(\mathbb{T})} &= \sum_{1\le l+m\le N}\left[1-\sigma\left(\frac{l+m}{N}\right)\right]^2\lambda_{lm}^{-2p}\gamma_{lm}^{-1}\Phi^2_{lm}(r,s)\\ &< \left(\frac{2}{1-s}\right)^{3/2}2N^{-4p+4}\left(\frac{1}{N}\sum_{1\le k\le N}\left[1-\sigma\left(\frac{k}{N}\right)\right]^2\left(\frac{k}{N}\right)^{-4p+3}\right).\end{aligned}$$

Der letzte Faktor ist eine Riemannsche Summe, die für $N\to\infty$ dem Integral

$$\int_0^1 (1-\sigma(\eta))^2\eta^{-4p+3}d\eta \tag{5.7}$$

entspricht. Eine Unbeschränktheit des Integrals kann sich nur aufgrund eines unbeschränkten Verhaltens des Integranden an der Stelle $\eta = 0$ ergeben. Da jedoch nach Voraussetzung ein reelles $\epsilon > 0$ mit $\sigma \in C^{2p}[0,\epsilon)$ existiert, lässt sich σ um $\eta = 0$ in eine Taylor-Reihe entwickeln. Man erhält

$$\sigma(\eta) = \sum_{j=0}^{2p-1}\frac{1}{j!}\sigma^{(j)}(0)\eta^j + o(\eta^{2p-1}) = 1 + \frac{1}{(2p-1)!}\sigma^{(2p-1)}(0)\eta^{2p-1} + o(\eta^{2p-1}),$$

für $\eta\in[0,\epsilon)$, wobei die Eigenschaften $\sigma(0)=1$ und $\sigma^{(j)}(0)=0,\ j=1,\dots,2p-2$, eines Filters der Ordnung $2p-1$ ausgenutzt wurden. Daher ist das Integral (5.7) beschränkt und man erhält

$$\|Q_N(r,s,\tilde{r},\tilde{s})\|^2_{L^2(\mathbb{T})} \;<\; \left(\frac{2}{1-s}\right)^{3/2}2N^{-4p+4}\cdot C\,,$$

mit einer von u abhängigen Konstanten C.
Für $(r,s)=(-1,1)$ ergibt sich mit (3.15) und (3.16) analog

$$\|Q_N(-1,1,\tilde{r},\tilde{s})\|^2_{L^2(\mathbb{T})} \;<\; N^{-4p+4}\left(\frac{1}{N}\sum_{1\le m\le N}\left[1-\sigma\left(\frac{m}{N}\right)\right]^2\left(\frac{m}{N}\right)^{-4p+3}\right).$$

Dieser Ausdruck kann wieder durch $C\cdot N^{-4p+4}$, mit einer geeigneten Konstanten C, abgeschätzt werden. Mit den obigen Abschätzungen für R_N und Q_N sowie der Cauchy-Schwarzschen Ungleichung angewendet auf (5.6) ist die Behauptung gezeigt. □

Beispiel 5.5 Beispiele von Filtern σ wachsender Ordnung sind die folgenden Funktionen.

$\sigma_1(\eta)$	$= 1-\eta$	Fejér-Filter (Cesàro-Mittel)
$\sigma_2(\eta)$	$= \frac{\sin(\pi\eta)}{\pi\eta}$	Lanczos-Filter
$\sigma_3(\eta)$	$= \frac{1}{2}(1+cos(\pi\eta))$	Raised-cosine-Filter
$\sigma_{4,p,\alpha}(\eta)$	$= \exp(-\alpha\eta^p)$	Exponentieller Filter p-ter Ordnung
$\sigma_{5,p}(\eta)$	$= 1-\frac{(2p-1)!}{(p-1)!}\int_0^\eta [t(1-t)]^{p-1}dt$	Vandeven-Filter p-ter Ordnung

In Abbildung 5.2 sind die Filter $\sigma_i,\ i=1,\ldots,3$, sowie $\sigma_{4,p,\alpha}(\eta)$ mit $p=8$ und $\alpha=35$ dargestellt. Je höher die Ordnung des Filters ist, desto weniger modifiziert dieser die zu den niedrigfrequenten Basisfunktionen gehörigen Koeffizienten und desto stärker ist der Einfluss auf die hochfrequenten Anteile.

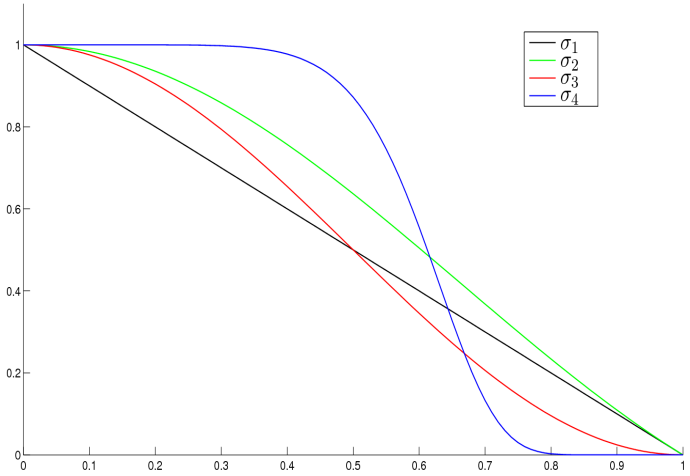


Abb. 5.2: Filterfunktionen $\sigma_i,\ i=1,\ldots,4$.

Bei dem Fejér-Filter handelt es sich nach Definition 5.1 um einen Filter erster Ordnung. Dieser Filter hat die aus theoretischer Sicht interessante Eigenschaft, dass die entsprechend gefilterten Fourier-Entwicklungen stückweise stetiger Funktion keine neuen Maxima oder Minima aufweisen, siehe [106, S. 89], und daher keine Über- und Unterschwinger wie in Abbildung 5.1 besitzen. Aus praktischer Sicht ist der Fejér-Filter für spektrale Verfahren jedoch nicht geeignet, da die Genauigkeit der Approximation nicht verbessert und an Unstetigkeitsstellen zu stark geglättet wird.
Der Lanczos-Filter ist stetig differenzierbar und erfüllt an $\eta=0$ die Bedingung $\sigma_2'(0)=0$. Dementsprechend handelt es sich beim Lanczos-Filter wie auch beim Raised-cosine-Filter um Filter zweiter Ordnung nach Definition 5.1, wobei der Raised-cosine-Filter die zusätzliche Bedingung (5.3) für $p=2$ erfüllt.
Der Parameter p des exponentiellen Filters entspricht dessen Ordnung, wie sich leicht errechnen lässt. Die Zusatzbedingung (5.3) ist jedoch formal noch nicht einmal für $l=0$ erfüllt. Deshalb wird der Parameter α oft so groß gewählt, dass zumindest der Funktionswert $\sigma_4(1)$ unterhalb der Rechengenauigkeit des spezifischen Computers liegt, beispiels-

weise wählt man $\alpha \approx -\ln(10^{-16})$, wie in Abbildung 5.2. Da die modale Filterung als Bestandteil spektraler Verfahren vorwiegend den Zweck hat, die Gibbschen Oszillationen zu reduzieren, und das Wiedergewinnen spektraler Genauigkeit *während der zeitlichen Evolution* üblicherweise nicht angestrebt wird, kann man mit der Wahl kleinerer Werte α auch von einer numerischen Gültigkeit von Bedingungen der Form (5.3) abweichen. Die Koeffizienten mit höheren Indizes werden in diesem Fall weniger stark modifiziert. Die abgeschnittene Form des exponentiellen Filters

$$\tilde{\sigma}_{4,p,\alpha,\eta}(\eta) = \begin{cases} 1 & 0 \leq \eta \leq \eta_c, \\ \exp\left(-\alpha\left(\frac{\eta-\eta_c}{1-\eta_c}\right)^p\right) & \eta_c \leq \eta \leq 1, \end{cases}$$

mit dem zusätzlichen Parameter η_c, findet ebenfalls Verwendung. In diesem Fall werden die niedrig frequenten Anteile in Form aller Koeffizienten mit Indizes k, die Bedingung $|k| \leq \eta_c N$ erfüllen, beibehalten.
Der Vandeven-Filter $\sigma_{5,p}$ ist so konstruiert, dass er für alle $p \geq 1$ neben der Definition eines Filters der Ordnung p auch die zusätzliche Bedingung (5.3) erfüllt. Für $p = 1$ entspricht dieser Filter dem Fejér-Filter σ_1. Fern der Unstetigkeitsstellen kann die exponentielle Genauigkeit der gefilterten Entwicklung, d.h. eine schnellere als polynomiale Konvergenzrate, durch die Wahl der Filterordnung p beispielsweise des Vandeven-Filters als wachsende Funktion von N erreicht werden, man vergleiche hierzu auch mit den Abschätzungen in Satz 5.2.
Alle der oben angegebenen Filter erfüllen zudem die für den Satz 5.4 relevante zusätzliche Glattheitseigenschaft $\sigma \in C^p[0,1]$, wobei p die jeweilige Ordnung des Filters ist.

Modale Filter zur Stabilisierung spektraler Verfahren Neben der Möglichkeit, durch die Verwendung modaler Filter punkweise hochgenaue Approximationen an stückweise stetige Funktionen zu rekonstruieren – zumindest fern der Unstetigkeitsstellen – wurde diese Art der Reduktion hoher Frequenzen auch zur Stabilisierung spektraler Methoden verwendet. Die grundlegende Idee ist hierbei, dass die Kontrolle der Oszillationen, die im Fall einer nichtlinearen Erhaltungsgleichung durch ihre Interaktion mit dem glatten Bereich der Lösung noch verstärkt werden können, notwendig für die Stabilität des spektralen Verfahrens ist. Jedoch ist nicht unbedingt die vollständige Beseitigung der Oszillationen erforderlich.
Die Anwendung des exponentiellen Filters auf die Näherungslösung (nach jedem Zeitschritt oder in bestimmten Zeitintervallen) zur Stabilisierung spektraler und Spektral-Element-Verfahren findet sich unter anderem bei Gottlieb, Lustman und Orszag [37] und Don [28] für die Euler-Gleichungen sowie bei Don, Gottlieb und Jung [29] für die kompressiblen Navier-Stokes-Gleichungen gekoppelt mit chemischen Reaktionen. Modale Filter werden ebenso innerhalb sogenannter Large-Eddy-Simulationen eingesetzt, die auf dem Konzept der Trennung von räumlichen Skalen basieren und mit deren Hilfe turbulente Strömungen modelliert werden können, siehe [64, 6, 83]. In [89] wird eine derartige Filterung zudem innerhalb einer Spektral-Element-Methode zur Berechnung glatter Lösungen der Flachwassergleichungen vorgeschlagen, mit dem Ziel der Stabilisierung der nichtlinearen Terme. Während die beschriebenen modalen Filter für klassische Spektral-Element-Methoden, die eine kontinuierliche Approximation auf dem gesamten Rechengebiet verlangen, zur Gewährleistung der Stetigkeit an Zellgrenzen modifiziert werden müssen, siehe [7], ist dies insbesondere für die in dieser Arbeit betrachteten diskontinuierlichen Galerkin-Verfahren nicht notwendig. Die Möglichkeit der einfachen Implementation

künstlicher Diffusion durch Verwendung modaler Filterroutinen bleibt dadurch für diesen Verfahrenstyp erhalten. Die Entwicklung einer Dämpfungsstrategie basierend auf der Anwendung modaler Filter scheint daher ebenso im Kontext von RKDG-Verfahren auf Dreiecksgittern ein vielversprechender Ansatz zu sein.
Eine Schwierigkeit bei der Nutzung modaler Filter zur Stabilisierung des numerischen Verfahrens stellt die Auswahl einer geeigneten Filterfunktion beziehungsweise deren Parameter dar. Während die notwendige Wahl eines *problemabhängigen* Filterparameters vermutlich akzeptiert werden kann, betrifft dies zunächst auch Veränderungen der Diskretisierungsparameter N und h unter Beibehalten der zu lösenden Gleichung. Einen möglichen Zugang zur Steuerung der Filterfunktion in Abhängigkeit von der Diskretisierung liefert die Methode spektraler Viskosität, welche im skalaren Fall Konvergenzaussagen erlaubt und desweiteren einer Idee von Gottlieb folgend oft in Form eines exponentiellen Filters implementiert wird. Ein solcher Zusammenhang zwischen modalen Filtern und der Methode spektraler Viskosität ist im Fall stückweiser Entwicklungen in PKD-Polynome auf Dreiecksgittern allerdings bisher noch nicht aufgestellt worden, so dass wir diesen im Folgenden erarbeiten werden.

5.3 Spektrale Viskosität

In [87] zeigte Tadmor, dass die naive Implementation eines spektralen Verfahrens zur Lösung nichtlinearer hyperbolischer Erhaltungsgleichungen die mit der Gleichung einhergehende Entropiebedingung verletzen kann. Die Konvergenz der spektralen Methode gegen die eindeutige Entropielösung kann in einem solchen Fall nicht garantiert werden und diesbezügliche numerische Experimente lieferten entsprechend schlechte Ergebnisse. Eine typische Herangehensweise zur Gewährleistung von Entropiebedingungen besteht in der Einführung eines dissipativen Mechanismus in Form von künstlicher Viskosität im Geiste der in Kapitel 2 beschriebenen Viskositätsmethode, d.h. die gegebenen Erhaltungleichung wird durch einen zusätzlichen dissipativen Term modifiziert und die entsprechend eingeführte Viskosität verschwindet mit steigender Anzahl von Freiheitsgraden, die innerhalb des numerischen Verfahrens berücksichtigt werden.
Tadmor entwickelte hierzu die “Spectral-Vanishing-Viscosity”- sowie die “Super-Viscosity”-Methode. Beide Methoden wurden später unter dem gemeinsamen Namen *Methode spektraler Viskosität* zusammengefasst. Die Grundidee dieser Methode ist die Einführung künstlicher Viskosität zur Kontrolle der Entropiedissipation in einer an die spezielle Situation spektraler Verfahren angepassten Form, die berücksichtigt, dass die hohe Genauigkeit spektraler Verfahren gerade durch deren geringe Dämpfung ermöglicht wird. Daher ist der Viskositätsterm so gewählt, dass dessen Einfluss mit steigenden Frequenzen der Lösung wächst und die niedrigen Frequenzen nur wenig oder überhaupt nicht modifiziert werden. Die ursprüngliche Fassung der SV-Methode wurde von Tadmor für das Fourier-Spektralverfahren entwickelt, zur numerischen Lösung skalarer periodischer Erhaltungsgleichungen der Form

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}f(u(x,t)) = 0\,, \qquad (x,t) \in [-\pi,\pi] \times [0,T]\,, \tag{5.8}$$

mit konvexer Flussfunktion f sowie mit 2π-periodischen Anfangsbedingungen

$$u(\cdot,0) = u_0 : [-\pi,\pi] \to \mathbb{R}$$

und periodischen Randbedingungen. Die Fourier-Projektion

$$\mathcal{P}_N u(x,t) \;=\; \sum_{|k|\le N} \hat{u}(k,t)e^{ikx}, \qquad \hat{u}(k,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} u(x,t)e^{-ikx}dx,$$

der Lösung $u(x,t)$ von (5.8) wird hierbei durch ein trigonometrisches Polynom N-ten Grades der Form

$$u_N(x,t) = \sum_{|k|\le N} \hat{u}_k(t)e^{ikx}$$

approximiert. Ausgehend von der Projektion der Anfangsbedingung

$$u_N(x,0) = \mathcal{P}_N u_0(x)$$

ist die zeitliche Entwicklung der mit dem klassischen Fourier-Spektralverfahren berechneten Näherungslösung durch die semidiskrete Gleichung

$$\frac{\partial}{\partial t}u_N(x,t) + \frac{\partial}{\partial x}\mathcal{P}_N f(u_N(x,t)) \;=\; 0 \tag{5.9}$$

gegeben.

Mangelnde Entropiedissipation des Spektralverfahrens Die fehlende Entropiedissipation der obigen Formulierung wird nun ersichtlich, wenn (5.9) mit u_N multipliziert und über das Intervall $[-\pi,\pi]$ integriert wird. Für das in Kapitel 2 definierte Energiefunktional $E_{per}[u_N](t) = \frac{1}{2}\int_{-\pi}^{\pi} u_N^2(x,t)dx$ ergibt sich daraus

$$\frac{d}{dt}E_{per}[u_N](t) \;=\; -\int_{-\pi}^{\pi} u_N \cdot \frac{\partial}{\partial x} f(u_N(x,t))\,dx = \int_{-\pi}^{\pi} \frac{\partial}{\partial x} q(u_N(x,t))\,dx, \tag{5.10}$$

für eine Funktion $q(u)$ mit $q'(u) = u\cdot f'(u)$, die nach Definition 2.8 ein zur Entropiefunktion $\eta(u) = \frac{1}{2}u^2$ gehöriger Entropiefluss für die skalare Gleichung (5.8) ist. Aufgrund der periodischen Randbedingungen folgt aus (5.10)

$$\frac{d}{dt}E_{per}[u_N](t) \;=\; [q(u_N(\cdot,t))]_{-\pi}^{\pi} = 0,$$

so dass die Energie der mit dem Fourier-Spektralverfahren erhaltenen numerischen Lösung zeitlich konstant ist. Konvergiert die Folge der Näherungen u_N für $N\to\infty$ nun gegen eine periodische Funktion $u : \mathbb{R}\times\mathbb{R}_0^+ \to \mathbb{R}$ in dem Sinn, dass für alle $t>0$ die starke L^2-Konvergenz,

$$\int_{-\pi}^{\pi} [u_N(x,t) - u(x,t)]^2 dx \to 0,$$

gilt, so ergibt sich für alle $t>0$

$$E_{per}[u](t) \;=\; \frac{1}{2}\int_{-\pi}^{\pi} u^2(x,t)\,dx = \lim_{N\to\infty}\frac{1}{2}\int_{-\pi}^{\pi} u_N^2(x,0)\,dx = \int_{-\pi}^{\pi}\frac{1}{2}u_0^2(x)\,dx,$$

so dass auch u in der Zeit konstante Energie besitzt. (Ein entsprechender Erhalt der Energie in der Zeit wurde in [87] auch für die Annahme der schwachen Konvergenz von u_N gegen eine schwache Lösung u der Gleichung (5.8) gezeigt.) Die Entropiebedingung erzwingt jedoch im Allgemeinen das Abfallen des Energiegehalts über Unstetigkeitskurven stückweise glatter Entropielösungen, wie in Kapitel 2 für das Beispiel der Burgers-Gleichung ausgeführt, so dass das klassische Fourier-Spektralverfahren (5.9) in einem solchen Fall nicht gegen die Entropielösung von (5.8) konvergieren kann.

Die Methode spektraler Viskosität Zur Kontrolle der Entropiedissipation ist der Methode spektraler Viskosität daher anstelle der Gleichung (5.9) die durch einen Viskositätsterm erweiterte Gleichung

$$\frac{\partial}{\partial t}u_N(x,t) + \frac{\partial}{\partial x}\mathcal{P}_N f(u_N(x,t)) = \epsilon_N(-1)^{p+1}\frac{\partial^p}{\partial x^p}\Big[Q_N\frac{\partial^p u_N(x,t)}{\partial x^p}\Big] \tag{5.11}$$

zugrunde gelegt, wobei der Operator Q_N definiert ist durch

$$Q_N v \;=\; \sum_{|k|\le N}\hat{Q}_k\hat{v}(k,t)e^{ikx}, \qquad v=\sum_{|k|\le N}\hat{v}(k,t)e^{ikx}, \tag{5.12}$$

mit nicht negativen Koeffizienten $\hat{Q}_k \le 1$. Der Parameter $p \in \mathbb{N}$ steuert die Ordnung des Viskositätsterms und kann als wachsende Funktion von N gewählt werden. Für $p > 1$ wurde der Begriff der Super-Viskosität eingeführt. Um niedrige Frequenzen viskositätsfrei zu belassen, wird $\hat{Q}_k = 0$ für $|k| \le m$ gewählt, für einen von N abhängigen Parameter $m < N$. Die Bandbreite wählbarer Werte für die Parameter m und ϵ_N sowie für die Koeffizienten $\hat{Q}_k$ ergibt sich aus dem folgenden Satz.

Satz 5.6 *Es sei angenommen, dass die durch die Methode spektraler Viskosität* (5.11), (5.12) *definierten Näherungslösungen u_N gleichmäßig beschränkt sind, d.h. es gelte*

$$\max_{0\le t\le T}\|u_N(\cdot,t)\|_{L^\infty[-1,1]} \le A_\infty,$$

mit einer von N unabhängigen Konstanten A_∞. Für die Viskositätsstärke ϵ_N in (5.11) *gelte hierbei*

$$\epsilon_N = \frac{C_p}{N^{2p-1}}\,, \tag{5.13}$$

wobei die von der Ordnung p des Viskositätsterms abhängige Konstante C_p durch

$$C_p \sim \sum_{k=1}^{p}|f|_{C^k}\|u_N\|_{L^\infty}^{k-1} \tag{5.14}$$

gegeben sei, mit $|f|_{C^k} = \|\partial_u^k f(u)\|_{L^\infty}$. Die Glättungsfaktoren $\hat{Q}_k$ seien zudem entsprechend den Bedingungen

$$\begin{aligned}\hat{Q}_k &= 0, \qquad |k| \le m_N,\\ 1-\left(\frac{m_N}{|k|}\right)^{\frac{2p-1}{\theta}} &\le \hat{Q}_k \le 1, \qquad |k| > m_N,\end{aligned}$$

gewählt, mit der oberen Schranke

$$m_N \le N^\theta \text{ mit } \theta < \frac{2p-1}{2p}$$

für den Abschneideparameter m_N.
Unter den obigen Annahmen konvergieren die Näherungslösungen u_N dann stark, d.h. in den Funktionenräumen $L^q([-1,1]\times[0,T])$, für alle $q < \infty$, gegen die eindeutige Entropielösung der konvexen Erhaltungsgleichung (5.8).

Beweis: Siehe [88]. □

In nachfolgenden Arbeiten wurden desweiteren SV-Methoden mit entsprechenden Konvergenzeigenschaften für mehrdimensionale periodische Erhaltungsgleichungen [15] sowie für nichtperiodische Erhaltungsgleichungen in einer Raumdimension,

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}f(u(x,t)) = 0, \qquad (x,t) \in [-1,1] \times [0,T],$$

entwickelt. Beispielsweise verwendeten Maday, Ould Kaber und Tadmor in [67] eine pseudospektrale Legendre-SV-Methode mit der Variationsformulierung

$$\left(\frac{\partial}{\partial t}u_N + \frac{\partial}{\partial x}I_N f(u_N), \Phi\right)_N = -\epsilon_N \left(Q * \frac{\partial}{\partial x}u_N, \frac{\partial}{\partial x}\Phi\right)_N + (B(u_N), \Phi)_N\,, \ \forall \Phi \in \mathcal{P}^N[-1,1].$$

Unter $(\cdot,\cdot)_N$ ist hierbei die Integration über $[-1,1]$ mittels geeigneter Quadraturformeln zu verstehen, I_N bezeichnet den zugehörigen Interpolationsoperator und $B(u_N)$ stellt die Einhaltung von Randbedingungen sicher. Der Operator Q wirkt in diesem Fall direkt auf den Koeffizienten der Legendre-Reihe, d.h.

$$Q_N * v\ (x,t) = \sum_{k=0}^{N} \hat{Q}_k(t)\hat{v}_k(t)P_k^{0,0}(x)\,, \qquad \text{für } v(x,t) = \sum_{k=0}^{N} \hat{v}_k(t)P_k^{0,0}(x)\,.$$

Während die Formulierung der SV-Methode im periodischen Fall eine effiziente Implementation durch modale Filter erlaubt, wie nachfolgend erläutert werden soll, muss der Viskositätsterm bei der Methode von Maday, Ould Kaber und Tadmor für nichtperiodische Gleichungen direkt implementiert werden. In [65] und [66] entwickelte Ma Chebychev-Legendre-SV-Methoden mit einem für Chebyshev-Entwicklungen angepassten Differentialoperator, auf die er die theoretischen Resultate von Tadmor übertragen konnte. Die SV-Methode wurde von Kirby und Sherwin [53] ebenso innerhalb eines Spektral-Element-Verfahrens auf unstrukturierten Gittern zur Lösung der inkompressiblen Navier-Stokes-Gleichungen genutzt, allerdings ohne Anpassung des Viskositätsterms an die Entwicklung in PKD-Polynome.

Zusammenhang zwischen spektraler Viskosität und modaler Filterung Bei geeigneter Wahl des Viskositätsterms in Abhängigkeit von den Ansatzfunktionen kann die SV-Methode effizient als Filter hoher Ordnung implementiert werden. Die Idee der Äquivalenz von Filterung und spektraler Viskosität sowie die dazu notwendige Modifizierung des Viskositätsterms in Abhängigkeit von den gewählten Basisfunktionen findet sich beispielsweise in den Arbeiten [35, 36, 66, 65, 29]. Die verwendeten Differentialoperatoren in der Formulierung mittels künstlicher Viskosität wie in (5.11) sind hierbei abhängig von dem spezifischen Sturm-Liouville-Problem, welches die Ansatzfunktionen erfüllen. Nachstehend aufgeführt sind die in der Literatur angegebenen Viskositätsterme zu verschiedenen Reihenentwicklungen, die bei der Verwendung spektraler Methoden auftreten.

Verwendete Viskositätsterme in Abhängigkeit der Basisfunktionen

Fourier-Reihen:	$\epsilon_N(-1)^{p+1}\frac{\partial^{2p}}{\partial x^{2p}}$
Chebychev-Reihen:	$\epsilon_N(-1)^{p+1}\left(\sqrt{1-x^2}\frac{\partial}{\partial x}\right)^{2p}$
Legendre-Reihen:	$\epsilon_N(-1)^{p+1}\left[\frac{\partial}{\partial x}(1-x^2)\frac{\partial}{\partial x}\right]^{p}$

Der Nachweis eines Konvergenzresultats analog zu Satz 5.6 ist hierbei bisher nur für den Fall von Chebychev-Reihen vorgenommen worden.
Der Zusammenhang zwischen der SV-Methode und einem direkt auf den Koeffizienten der Entwicklung wirkenden modalen Filter soll in diesem Abschnitt exemplarisch für den Fall der Fourier-Spektralmethode hergestellt werden und überträgt sich auf den Fall anderer Reihenentwicklungen bei Wahl des zur Basis gehörigen Viskositätsterms.
Die Gleichung (5.11) wird hierzu im Zeitintervall $[t^n, t^{n+1}]$ mit Hilfe eines Splittingverfahrens in Transport- und Dämpfungsanteil zerlegt. Dadurch erhält man zunächst die Gleichungen

$$\frac{\partial}{\partial t} w_N = \epsilon_N (-1)^{p+1} \frac{\partial^p}{\partial x^p} \Big[Q_N \frac{\partial^p}{\partial x^p} w_N \Big], \qquad w_N(x, t^n) = u_N(x, t^n), \tag{5.15}$$

$$\frac{\partial}{\partial t} u_N + \frac{\partial}{\partial x} \mathcal{P}_N f(u_N) = 0, \qquad u_N(x, t^n) = w_N(x, t_{n+1}). \tag{5.16}$$

Somit wird vor der Anwendung des klassischen Spektralverfahrens entsprechend der Gleichung (5.16) ein Dämpfungsschritt durchgeführt. Die zugehörige Gleichung (5.15) ist zwar von parabolischem Typ, kann allerdings effizient mit Hilfe eines modalen Filters implementiert werden. Mit

$$w_N(x,t) = \sum_{|k| \le N} \hat{w}_k(t) e^{ikx}$$

und der Darstellung des Viskositätsterms als

$$\epsilon_N (-1)^{p+1} \frac{\partial^p}{\partial x^p} \Big[Q_N \frac{\partial^p w_N(x,t)}{\partial x^p} \Big] = -\epsilon_N \sum_{m < |k| \le N} k^{2p} \hat{Q}_k \hat{w}_k(t) e^{ikx}$$

erhält man N entkoppelte gewöhnlichen Differentialgleichungen für die Fourier-Koeffizienten $\hat{w}_k(t)$:

$$\begin{aligned} \frac{d}{dt}\hat{w}_k(t) &= -\epsilon_N k^{2p} \hat{Q}_k \hat{w}_k(t), \qquad m \le |k| \le N, \\ \frac{d}{dt}\hat{w}_k(t) &= 0, \qquad |k| < m\,, \end{aligned}$$

die leicht analytisch gelöst werden können. (An dieser Stelle ergibt sich ein analoges Resultat für andere Basisfunktionen, wenn der zugehörige Viskositätsterm gewählt wird.) Mit $\Delta t = t^{n+1} - t^n$ und $\epsilon_N = C_p / N^{2p-1}$ ergibt sich

$$\hat{w}_k(t^{n+1}) = \begin{cases} \exp\left(-C_p \Delta t N \hat{Q}_k \left(\frac{k}{N}\right)^{2p} \right) \hat{w}_k(t^n), & m \le |k| \le N, \\ \hat{w}_k(t^n), & |k| < m\,. \end{cases}$$

In [36] wurde aufgrund der Beobachtungen in numerischen Experimenten vorgeschlagen, Viskositätsterme höherer Ordnung, d.h. $p > 1$, zu verwenden und von der Definition eines viskositätsfreien Spektrums abzusehen, d.h. $m_N = 0$ zu wählen, so dass sich eine Modifikation aller Koeffizienten der Entwicklung mit Ausnahme des Zellmittelwertes ergibt. Desweiteren werden die Viskositätskoeffizienten durch $\hat{Q}_k = 1,\ |k| \le N$ festgelegt. Diese spezielle Wahl der Parameter wurde auch bei der Konstruktion der Chebyshev-Legendre-Spektralmethode von Ma in [66] verwendet. Die Formulierung spektraler Viskosität hat dann die einfacherer Form

$$\frac{\partial}{\partial t} u_N + \frac{\partial}{\partial x} \mathcal{P}_N f(u_N) = \epsilon_N (-1)^{p+1} \frac{\partial^{2p}}{\partial x^{2p}} u_N. \tag{5.17}$$

In diesem Fall ist der Dämpfungsschritt durch die Gleichung

$$\frac{d}{dt}\hat{w}_k = -\epsilon_N k^{2p}\hat{w}_k, \quad 0 \leq |k| \leq N,$$

gegeben, und mit der Definition $\epsilon_N = C_p/N^{2p-1}$ in (5.13) ergibt sich die Modifikation der Fourier-Koeffizienten als

$$\hat{w}_k^\sigma = \exp\left(-C_p N\Delta t \left(\frac{k}{N}\right)^{2p}\right)\hat{w}_k.$$

Somit erhält man die modal gefilterte Näherung

$$w_N^\sigma = \sum_{|k|\leq N} \sigma\left(\frac{k}{N}\right)\hat{w}_k e^{ikx},$$

mit dem durch $\sigma(\eta) = \exp\left(-C_p N\Delta t\eta^{2p}\right)$ gegebenem exponentiellen Filter der Ordnung $2p$ und der Filterstärke $\alpha = C_p N\Delta t$.
Diese modale Filterung wird vor jedem Schritt des Spektralverfahrens auf die zum entsprechenden Zeitpunkt t^n vorliegende Näherung $u_N(x,t^n) = w_N(x,t^n)$ angewendet.
Bei expliziten Verfahren zur Zeitintegration führt die entsprechende Zeitschrittrestriktion im Fall der Fourier-Spektralmethode zu einem konstanten Wert von $N\Delta t$, siehe [36]. Bei der auf Fourier-Reihen basierten Methode ist es daher möglich eine konstante Filterstärke α für alle Werte von N zu wählen. Bei den Weiterentwicklungen der SV-Methode für Legendre- [66] und Chebyshev-Legendre-Spektralverfahren [67] ergibt sich hingegen eine Zeitschrittrestriktion der Form $\Delta t = \mathcal{O}(N^{-2})$, siehe [43]. Daher ist die Filterstärke bei Verwendung dieser Ansatzfunktionen in Abhängigkeit vom Polynomgrad N zu wählen, mit $\alpha_N = \mathcal{O}(1/N)$.

5.4 Ein modaler Filter für die Proriol-Koornwinder-Dubiner-Basis

Der im vorherigen Abschnitt hergeleitete Zusammenhang zwischen spektraler Viskosität und modaler Filterung soll nun auf den Fall stückweiser PKD-Entwicklungen auf Dreiecksgittern übertragen werden. Die Grundidee, den zur Basis der PKD-Polynome gehörigen Sturm-Liouville-Operator innerhalb der Viskositätsformulierung zu verwenden, wurde bereits in der dieser Schrift vorangegangenen Arbeit [68] sowie in [69] beschrieben, mit ersten numerischen Ergebnissen für skalare Erhaltungsgleichungen.
Wie bisher erfolgt die Herleitung des modalen Filters in der folgenden Darstellung nur für skalare Gleichungen. Im Fall von Systemen wird die Viskosität komponentenweise eingeführt, so dass der Filter auf die konservativen Variablen angewendet wird.

5.4.1 Herleitung aus einer Formulierung spektraler Viskosität auf dem Dreieck

Gegeben sei eine zweidimensionale, skalare Erhaltungsgleichung der Form

$$\frac{\partial}{\partial t}u(\mathbf{x},t) + \nabla_{\mathbf{x}}\cdot\mathbf{f}(u(\mathbf{x},t)) = 0, \quad (\mathbf{x},t)\in\Omega\times\mathbb{R}_+. \tag{5.18}$$

Um aus der Formulierung spektraler Viskosität für (5.18) einen modalen Filter basierend auf der PKD-Basis zu erhalten, wird das zu dieser Basis gehörige Sturm-Liouville-Problem betrachtet. Zudem müssen die Bedingungen an die Viskositätsstärke ϵ_N in (5.13) auf die elementweise Diskretisierung des DG-Verfahrens übertragen werden.
Im Kontext der DG-Diskretisierung werden auf den einzelnen Elementen zunächst Gleichungen der Form

$$\left(\frac{\partial}{\partial t}u_{h,N} + \nabla_{\mathbf{x}} \cdot \tilde{\mathcal{P}}_N \mathbf{f}(u_{h,N}), v\right)_{N_I} = (\mathcal{B}[u_{h,N}], v)_{N_K}, \qquad \forall\, v \in \mathcal{P}^N(\tau_i),$$

gelöst. Hierbei bezeichnen die Notationen $(\cdot,\cdot)_{N_I}$ und $(\cdot,\cdot)_{N_K}$ die Verwendung der Quadraturformeln, die der Volumen- bzw. der Randintegration entsprechen. Desweiteren bezeichnet $\tilde{\mathcal{P}}_N\mathbf{f}$ die Projektion der Flussfunktion in den Ansatzraum mit der Eigenschaft $(\tilde{\mathcal{P}}_N\mathbf{f}, v)_{N_I} = (\mathbf{f}, v)_{N_I}$, für alle $v \in \mathcal{P}^N(\tau_i)$. Der Operator $\mathcal{B}$ fasst die durch die räumliche Diskretisierung entstehenden Randterme zusammen, zur Differenz der projizierten und der numerischen Flussfunktion,

$$\mathcal{B}[u_{h,N}] = \tilde{\mathcal{P}}_N\mathbf{f}(u_{h,N}) \cdot \mathbf{n} - \mathbf{H}(u^-_{h,N}, u^+_{h,N}, \mathbf{n}).$$

Mit der Transformationsabbildung $\Lambda_i(\mathbf{x}) = \mathbf{A}_i\mathbf{x} + \mathbf{b}_i$, die τ_i auf das Standardreieck abbildet, sowie der Anwendung der Kettenregel $\nabla_{\mathbf{x}} = \mathbf{A}_i^T\nabla_{r,s}$ ergibt sich für das Element τ_i die Darstellung der DG-Diskretisierung auf $\mathbb{T}$ durch eine Gleichung der Form

$$\begin{aligned}&\left(\frac{\partial}{\partial t}u_{h,N}(\Lambda_i^{-1}(r,s),t) + \nabla_{r,s} \cdot \tilde{\mathcal{P}}_N\mathbf{A}_i\mathbf{f}(u_{h,N}(\Lambda_i^{-1}(r,s),t)), \tilde{v}\right)_{N_I}\\ &\quad = \quad \left(\mathcal{B}[u_{h,N}](\Lambda_i^{-1}(r,s),t), \tilde{v}\right)_{N_K},\end{aligned} \tag{5.19}$$

die für alle $\tilde{v} \in \mathcal{P}^N(\mathbb{T})$ gefordert wird. Eine Formulierung spektraler Viskosität analog zur Fourier-Spektralmethode (5.17) unter Verwendung des zur Basis der PKD-Polynome assoziierten Sturm-Liouville-Operators, siehe hierzu die Definition (3.7), erhält man dann durch

$$\begin{aligned}&\left(\frac{\partial}{\partial t}u_{h,N}(\Lambda_i^{-1}(r,s),t) + \nabla_{r,s} \cdot \tilde{\mathcal{P}}_N\mathbf{A}_i\mathbf{f}(u_{h,N}(\Lambda_i^{-1}(r,s),t)), \tilde{v}\right)_{N_I}\\ &= \left(\mathcal{B}[u_{h,N}](\Lambda_i^{-1}(r,s),t), \tilde{v}\right)_{N_K} + \epsilon_N^i(-1)^{p+1}\left((\mathcal{L}_{r,s})^p u_{h,N}(\Lambda_i^{-1}(r,s),t), \tilde{v}\right)_{N_K},\end{aligned} \tag{5.20}$$

für alle $\tilde{v} \in \mathcal{P}^N(\mathbb{T})$, mit einer noch zu bestimmenden Viskositätsstärke ϵ_N^i.
Aufgrund der Bedingung (5.14) an die Konstante C_p ist die Viskositätsstärke ϵ_N in (5.11) abhängig von den Ableitungen des Flussvektors der gegebenen Erhaltungsgleichung. In der Gleichung (5.19) ist diese Flussfunktion durch $\mathbf{A}_i\mathbf{f}$ gegeben. Unter Vernachlässigung der Abhängigkeit von u_N ergibt sich für ein beliebiges $d \in \mathbb{R}^+$ nach (5.14) die Eigenschaft $C_p(d\cdot f) = d\cdot C_p(f)$. Aufgrund dieser Überlegung ist es in Anbetracht der Gleichung (5.19) sinnvoll, ϵ_N^i proportional zur Operatornorm $\|\mathbf{A}_i\|_\infty$ der Matrix $\mathbf{A}_i$ zu wählen. Da $\|\mathbf{A}_i\|_\infty$ im Fall ähnlicher Dreiecke τ_i zudem proportional zum Kehrwert $1/h_i$ des zugehörigen Längenmaßes ist, ergibt sich die Definition von ϵ_N^i als

$$\epsilon_N^i = \frac{C_p}{h_i(N+1)^{2p-1}}, \tag{5.21}$$

so dass ϵ_N^i damit im Wesentlichen analog zu (5.13) gewählt wird, bis auf das Einbeziehen des Längenmaßes h_i.

Wie im Fall der Fourier-Spektralmethode, wird die Gleichung (5.20) nun durch Operatorsplitting in Transport- und Dämpfungsteil zerlegt. Desweiteren lässt sich analog zur Gleichung (5.15) auch in diesem Fall der Dämpfungsteil als modaler Filter ausdrücken. Auf dem Dreieck $\tau_i \in \mathcal{T}^h$ betrachten wir hierzu eine Funktion $w^i_{h,N} : \mathbb{T} \times [t^n, t^{n+1}] \to \mathbb{R}$ mit der Eigenschaft $w^i_{h,N}(\cdot, t) \in \mathcal{P}_N(\mathbb{T})$ für alle $t \in [t^n, t^{n+1}]$, die die Anfangsbedingung

$$w^i_{h,N}(r, s, t^n) = u_{h,N}\left(\Lambda_i^{-1}(r, s), t^n\right)$$

erfüllt. Diese Anfangsbedingung entspricht hierbei einer aus einem Schritt des DG-Verfahrens resultierenden Näherung auf dem Dreieck τ_i, die mit Gibbschen Oszillationen behaftet ist.
Der zur Gleichung (5.20) zugehörige Filterschritt ist dann gegeben durch die Differentialgleichung

$$\frac{\partial}{\partial t} w^i_{h,N} = \epsilon^i_N (-1)^{p+1} (\mathcal{L}_{r,s})^p w^i_{h,N}, \tag{5.22}$$

mit der elementweisen Viskositätsstärke ϵ^i_N nach (5.21). Die Anwendung des Operators $(\mathcal{L}_{r,s})^p$ auf die PKD-Entwicklung von $w^i_{h,N}$ ergibt wegen der in Satz 3.2 beschriebenen Eigenschaft der PKD-Polynome als Eigenfunktionen den Ausdruck

$$(\mathcal{L}_{r,s})^p \, w^i_{h,N} = (-1)^p \sum_{l+m \le N} \lambda^p_{lm} \hat{w}^i_{lm} \Phi_{lm} \, .$$

Daher erhält man mit Hilfe der Entwicklung von $w^i_{h,N}$ in eine PKD-Reihe unter Ausnutzung von (5.22) und (5.21) die Gleichung

$$\frac{d}{dt} \hat{w}^i_{lm} = -\frac{C_p}{h_i (N+1)^{2p-1}} (l+m)^p (l+m+2)^p \hat{w}^i_{lm} \approx -\frac{C_p (N+1)}{h_i} \left(\frac{l+m}{N+1} \right)^{2p} \hat{w}^i_{lm}$$

für die PKD-Koeffizienten $\hat{w}^i_{lm},\ 0 \le l+m \le N$. Dementsprechend wird wie im Fall der Fourier-Spektralmethode ein exponentieller Filter

$$\sigma(\eta) = \exp(-\alpha \eta^{2p}), \qquad \eta = \frac{l+m}{N+1}, \tag{5.23}$$

der Ordnung $2p$, mit durch

$$\alpha_i = \frac{C_p (N+1) \Delta t}{h_i} \tag{5.24}$$

gegebener elementweiser Filterstärke, auf die PKD-Entwicklung angewendet.
Im Fall der Nutzung expliziter Zeitintegrationsverfahren zur Lösung der semidiskreten Gleichung (5.19) ergibt sich eine Zeitschrittrestriktion von $\Delta t = \mathcal{O}(h/N^2)$, wie sich beispielsweise der Untersuchung der Eigenwerte des advektiven Operators in [52] entnehmen lässt. Ein entsprechend gewählter Zeitschritt führt daher zur Invarianz des durch (5.23), (5.24) gegebenen modalen Filters $\sigma(\eta)$ bei Gitterverfeinerung sowie zur Bedingung $\alpha = \mathcal{O}(1/N)$ an die Filterstärke.

5.4.2 Adaptive Filterung

Ordnungsverlust durch globale Filterung Aufgrund der in jedem Zeitschritt durchgeführten Dämpfung führt die globale Anwendung des exponentiellen Filters (5.23) auf das Eingangsbeispiel der linearen Transportgleichung zu einer drastischen Verschlechterung der Genauigkeit des DG-Verfahrens, wie in in der Tabelle 5.1 ersichtlich ist. Dort sind die L^2-Fehler der Näherungslösungen für $3 \leq N \leq 6$ verzeichnet, die im Vergleich zu den in der Tabelle 4.3 gegebenen Resultaten deutlich höher sind, sowie die errechnete numerische Konvergenzordung. Wie im Fall des unmodifizierten DG-Verfahrens wurde hierbei mit einer globalen Zeitschrittweite von $\Delta t = 10^{-5}$ gerechnet. Die Filterparameter des modalen Filters (5.23), (5.24) wurden in diesem Beispiel durch $p = 3$ sowie $C_p = 0.1$ festgelegt.

K	N	L^2-Fehler	EOC	N	L^2-Fehler	EOC
296		3.634776e-03			2.526065e-03	
1184	3	1.140990e-03	1.671580e+00	4	5.657878e-04	2.158559e+00
4736		2.736219e-04	2.060031e+00		1.147468e-04	2.301807e+00
296		1.737470e-03			1.181871e-03	
1184	5	2.901764e-04	2.581986e+00	6	1.586382e-04	2.897261e+00
4736		5.767916e-05	2.330808e+00		3.253800e-05	2.285543e+00

Tabelle 5.1: Wirkung der globalen Anwendung des modalen Filters ($p = 3,\ C_p = 0.1$) auf die Entwicklung des L^2-Fehlers.

Die Daten der Tabelle 5.1 zeigen, dass die experimentelle Konvergenzordnung in diesem Fall offensichtlich nicht mit der zu erwartenden räumlichen Ordnung des DG-Verfahrens übereinstimmt. Dies steht nicht im Widerspruch zu den approximationstheoretischen Resultaten für modale Filter in Abschnitt 5.2. Während die modale Filterung der Reihenentwicklungen spektraler Methoden nach Satz 5.2 und Satz 5.4 hochgenaue Approximationen in Bezug auf steigende Polynomgrade liefern kann, lassen sich anhand dieser Resultate keine Aussagen treffen bezüglich der Konvergenz elementweise gefilterter Entwicklungen auf Gebietszerlegungen unter Gitterverfeinerung bei festgehaltenem Polynomgrad N. Tatsächlich lässt sich nachweisen, dass die Filterung diesbezüglich nur eine Nährung von erster Ordnung im Raum zulässt, falls schon die zum linearen Anteil von $u_{h,N}$ gehörigen Koeffizienten, mit Indizes $l + m = 1$, modifiziert werden, d.h. falls der gewählte Filter σ die Eigenschaft $\sigma(1/N) < 1$ besitzt. Da insbesondere für den aus der Formulierung spektraler Viskosität hergeleiteten exponentiellen Filter die Eigenschaft

$$\sigma(\eta) < 1 \text{ für alle } \eta > 0$$

gilt, modifiziert dieser bis auf den Zellmittelwert alle Koeffizienten der Entwicklung, unabhängig von der Anzahl der Terme. Dementsprechend treffen die nachfolgenden Betrachtungen unabhängig vom gewählten Polynomgrad auch auf diesen Filter zu.

Zur Abschätzung der Approximationsfehler elementweise gefilterter PKD-Entwicklungen betrachte man eine Folge regulärer, konformer Triangulierungen $\mathcal{T}^h$ des Gebiets $\Omega = [-1, 1]^2$, sowie die durch

$$u_{h,N} : \Omega \to \mathbb{R}, \qquad u_{h,N}|_{\tau_i}(\mathbf{x}) = \sum_{l+m \leq N} \hat{u}^i_{lm} \Phi_{lm}(\Lambda_i(\mathbf{x})), \quad \tau_i \in \mathcal{T}^h,$$

gegebene stückweise polynomielle Approximation N-ten Grades einer glatten Funktion $u : \Omega \to \mathbb{R}$. Desweiteren sei die elementweise gefilterte Entwicklung gegeben durch

$$u^\sigma_{h,N} : \Omega \to \mathbb{R}, \qquad u^\sigma_{h,N}|_{\tau_i}(\mathbf{x}) = \sum_{l+m\leq N} \sigma\left(\frac{l+m}{N}\right) \hat{u}^i_{lm} \Phi_{lm}(\Lambda_i(\mathbf{x})).$$

Für den Fehler in der L^2-Norm gilt nun die Abschätzung

$$\|u - u^\sigma_{h,N}\|_{L^2(\Omega)} \geq \|u_{h,N} - u^\sigma_{h,N}\|_{L^2(\Omega)} - \|u_{h,N} - u\|_{L^2(\Omega)}.$$

Für hinreichend glatte Funktionen u besitzt der Interpolationsfehler nach [16, Theorem 3.1.5] die obere Schranke $\|u_{h,N} - u\|_{L^2(\Omega)} = \mathcal{O}(h^{N+1})$ für $h \to 0$, so dass der räumliche Fehler bezüglich Gitterverfeinerung dominiert wird durch den Fehler $\|u_{h,N} - u^\sigma_{h,N}\|_{L^2(\Omega)}$, der sich durch die Filterung ergibt. Für diesen gilt

$$\|u_{h,N} - u^\sigma_{h,N}\|^2_{L^2(\Omega)} = \sum_{\tau_i\in\mathcal{T}^h} \|u_{h,N} - u^\sigma_{h,N}\|^2_{L^2(\tau_i)} \geq \sum_{\tau_i\in\mathcal{T}^h} \frac{|\tau_i|}{2} \sum_{l+m=1} \left(1 - \sigma\left(\frac{1}{N}\right)\right)^2 \gamma_{lm} \left(\hat{u}^i_{lm}\right)^2.$$

Nun kann aber für die Koeffizienten $\hat{u}^i_{lm},\ l+m=1$, keine bessere Schranke als $\mathcal{O}(h)$ erwartet werden, mit der entsprechenden Implikation, dass die Abschätzung der Form $\|u_{h,N} - u^\sigma_{h,N}\|_{L^2(\Omega)} = \mathcal{O}(h)$ strikt ist. Hierzu betrachten wir die lineare Funktion $u(\mathbf{x}) = \mathbf{a}\cdot\mathbf{x} + c$ sowie die Folge von Dreieckselementen

$$\tau_{i_n} = \{\mathbf{x} \in \mathbb{R}^2 \mid -\frac{1}{n} \leq x_1, x_2;\ x_1 + x_2 \leq 0\}.$$

Die zugehörigen Folgen von Dubiner-Koeffizienten $\{\hat{u}^{i_n}_{lm}\}_{n\in\mathbb{N}}$ für $l+m=1$ besitzen die Folgenglieder

$$\hat{u}^{i_n}_{lm} = \frac{1}{n}\cdot\frac{1}{\gamma_{lm}} \int_{\mathbb{T}} \mathbf{a}\cdot\begin{pmatrix} r \\ s \end{pmatrix} \Phi_{lm}(r,s)drds = h_n \cdot C,$$

mit dem zu τ_{i_n} gehörige Längenmaß h_n und einer von h_n unabhängigen Konstanten C, wie sich unter Anwendung der Gleichung (4.4) leicht nachrechnen lässt.

Indikatoren zur elementweisen adaptiven Filterung Eine Möglichkeit, den hier beobachteten Ordnungsverlust zu vermeiden, ist die adaptive Verwendung des modalen Filters nur in den Gitterzellen, in denen Oszillationen auftreten und das Vorliegen einer Unstetigkeit vermutet wird. Von Persson und Peraire wird in [72] für die Einführung eines klassischen Viskositätsterm zweiter Ordnung in das DG-Verfahren eine derartige Strategie verwendet. Die Autoren konstruieren einen Indikator basierend auf dem Abklingverhalten der PKD-Koeffizienten, der diejenigen Zellen kennzeichnet, auf denen zusätzliche numerische Viskosität eingeführt werden soll. Bezüglich des Rechenaufwandes ist ein derartiger Indikator im Rahmen der hier verwendeten modalen Filterung ideal, so dass dessen Anwendung in diesem neuen Kontext untersucht werden soll. Zunächst wählen wir die Filterstärke $\alpha_{N,i}$ auf einem Dreieck τ_i als

$$\alpha_{N,i} = \begin{cases} 0 & \text{falls } \omega_i < (N+1)^{-4}, \\ C_p\cdot(N+1)^{-1} & \text{sonst,} \end{cases} \tag{5.25}$$

wobei der Indikator ω_i durch

$$\omega_i = \sum_{l+m=N} \gamma_{lm}(\hat{u}^i_{lm})^2 \cdot \left(\sum_{l+m<N} \gamma_{lm}(\hat{u}^i_{lm})^2 + \epsilon \right)^{-1}, \tag{5.26}$$

gegeben ist, mit einem zusätzlichen Parameter $\epsilon > 0$ zur Vermeidung der Division durch Null. Dieser Indikator wurde auch bei den Berechnungen der in [68] dargestellten Ergebnisse verwendet. Heuristisch lässt sich der Indikator durch die Abschätzung

$$\sum_{l+m=N} \gamma_{lm}(\hat{u}^i_{lm})^2 \leq N^{-2p}(N+2)^{-2p} \|\mathcal{L}^p_{r,s}(u \circ \Lambda_i^{-1})\|^2_{L^2(\mathbb{T})} = \mathcal{O}(N^{-4p}),$$

motivieren, die sich für hinreichend glatte Funktionen aus der Gleichung (3.18) im Beweis von Satz 3.3 herleiten lässt. Dementsprechend verschwindet die Filterstärke auf den Bereichen des Rechengitters, auf denen die Lösung gut dargestellt wird. Da sich eine entsprechend gute Auflösung für glatte Funktionen unter anderem durch eine Gitterverfeinerung ergibt, darf man erwarten, dass die Konvergenzordnung des DG-Verfahrens aufgrund der verschwindenden Filterstärke erhalten bleibt. Bei der Berechnung von Näherungen an die Transportgleichung für $3 \leq N \leq 6$ ergibt sich die in Tabelle 5.2 dargestellte Fehlerentwicklung. Hierbei wurden wie im vorherigen Fall die Filterparameter durch $p = 3,\ C_p = 0.1$ belegt. Dementsprechend deuten auch die numerischen Experimente für die elementweise adaptive Einführung künstlicher Viskosität mittels (5.25), (5.26) an, dass die Ordnung des Verfahrens erhalten bleibt.

K	N	L^2-Fehler	EOC	N	L^2-Fehler	EOC
296		1.189032e-03			3.486356e-04	
1184	3	6.794500e-05	4.129276e+00	4	7.649236e-06	5.510260e+00
4736		2.774004e-06	4.614326e+00		2.138201e-07	5.160846e+00
296		7.988103e-05			2.031294e-05	
1184	5	1.050113e-06	6.249236e+00	6	1.305682e-07	7.281452e+00
4736		1.540299e-08	6.091190e+00		1.041515e-09	6.969976e+00

Tabelle 5.2: Wirkung der adaptiven Anwendung des modalen Filters ($p = 3,\ C_p = 0.1$) mittels (5.25), (5.26) auf die experimentelle Konvergenzordnung.

Die endgültige Skalierung des Indikators von Persson und Peraire erzielt eine glatte, monoton steigende Indikatorfunktion $s : \mathbb{R}^+_0 \to [0, 1]$. Barter und Darmofal [3] nutzen in ihrer Definition künstlicher Viskösität zwei verschiedene Stoßindikatoren, einen koeffizientenbasierten Indikator

$$s_{res}(\omega_i) = \min\{\, 1000(5N^4 + 1)\omega_i \,, 1 \,\} \tag{5.27}$$

ähnlich dem Indikator von Persson und Peraire, sowie einen Sprungindikator, abhängig von den Sprüngen über Elementgrenzen der DG-Diskretisierung. Letzterer Indikator basiert auf den Fehlerschätzungen von Krivodonova et al. [56], die zeigten, dass für derartige Sprünge von Komponenten der Näherungslösung oder daraus berechneter Größen $g(\mathbf{u}_{h,N})$, mit einer glatten Funktion $g : \mathbb{R}^m \to \mathbb{R}$, die Abschätzungen

$$\frac{1}{|\tau_i|} \int_{\tau_i} |g(\mathbf{u}^-_{h,N}) - g(\mathbf{u}^+_{h,N})|\, d\sigma = \begin{cases} \mathcal{O}(h^{N+1}), & \text{in Bereichen in denen die Lösung glatt ist,} \\ \mathcal{O}(1), & \text{nahe Unstetigkeitsstellen,} \end{cases}$$

gültig sind. Dementsprechend ist der Sprungindikator in Abhängigkeit von

$$\tilde{\omega}_i = \frac{1}{|\partial\tau_i|}\int_{\partial\tau_i} \frac{|g(\mathbf{u}_{h,N}^-) - g(\mathbf{u}_{h,N}^+)|}{|\{g(\mathbf{u}_{h,N})\}| + \epsilon} d\sigma,$$

mit $\{g(\mathbf{u}_{h,N})\} = \frac{1}{2}\left(g(\mathbf{u}_{h,N}^-) + g(\mathbf{u}_{h,N}^+)\right)$, gewählt als

$$s_J(\tilde{\omega}_i) = \min\{\, 5000\tilde{\omega}_i\,, 1\,\}. \tag{5.28}$$

Die beiden von Barter und Darmofal verwendeten Indikatoren (5.27), (5.28) werden in den in der vorliegenden Arbeit durchgeführten Experimenten zur Steuerung der Filterstärke mittels

$$\alpha_{N,i} = \begin{cases} s_{res/J} \cdot C_p \frac{N\Delta t}{h_i} & \text{falls } s_{res/J} > 0.01 \\ 0 & \text{andernfalls} \end{cases} \tag{5.29}$$

implementiert, wobei der Parameter C_p neben der Filterordnung in Abhängigkeit vom konkreten Testfall zu wählen ist.
Für $p = 3$ und $C_p = 0.1$ ergeben sich im Fall des betrachteten Beispiels zur linearen Advektionsgleichung die in Tabelle 5.3 verzeichneten L^2-Fehlerentwicklungen für die beiden Indikatoren s_{res} und s_J.

N	K	s_{res}		s_J	
		L^∞-Fehler	EOC	L^∞-Fehler	EOC
3	296	1.225255e-03		1.225255e-03	
	1184	8.169819e-05	3.906634e+00	8.172099e-05	3.906232e+00
	4736	6.189373e-06	3.722439e+00	7.059044e-06	3.533162e+00
4	296	3.931542e-04		3.931542e-04	
	1184	1.626865e-05	4.594929e+00	1.643935e-05	4.579870e+00
	4736	3.494515e-07	5.540859e+00	1.574537e-06	3.384154e+00
5	296	1.136967e-04		1.136967e-04	
	1184	2.528406e-06	5.490818e+00	5.984817e-06	4.247739e+00
	4736	1.602622e-08	7.301650e+00	3.938578e-07	3.925560e+00
6	296	4.407593e-05		4.415189e-05	
	1184	1.702150e-07	8.016489e+00	1.430511e-06	4.947873e+00
	4736	1.041928e-09	7.351959e+00	3.618938e-08	5.304820e+00

Tabelle 5.3: Wirkung der adaptiven Anwendung des modalen Filters ($p = 3,\ C_p = 0.1$) mittels (5.29) auf die experimentelle Konvergenzordnung.

Diese ersten Ergebnisse deuten an, dass die Wahl des Indikators s_{res} die Ordnung des Verfahrens im Wesentlichen erhält, während dies für den Indikator s_J zumindest auf den betrachteten Gittern nicht zuzutreffen scheint. Beide Indikatoren liefern jedoch für den betrachteten Testfall deutlich bessere Ergebnisse als die globale Filterung. Werden die Filterparameter mit den Werten $p = 3$ sowie $C_p = 0.01$ belegt, so ergeben sich die in Tabelle 5.4 dargestellten Ergebnisse für die adaptive Anwendung des modalen Filters nach (5.29) mittels der Indikatoren s_{res} und s_J sowie für die globale Anwendung des Filters. Bei dieser Parameterwahl entspricht die experimentelle Konvergenzordnung für beide Indikatoren derjenigen des ungefilterten DG-Verfahrens. Die zum Vergleich angegebenen Resultate der globalen Filterung zeigen hierbei erwartungsgemäß auch in diesem Fall eine deutliche Beeinträchtigung der Genauigkeit des Verfahrens auf.

N	K	s_{res}		s_J		globale Filterung	
		L^∞-Fehler	EOC	L^∞-Fehler	EOC	L^∞-Fehler	EOC
	296	1.175861e-03		1.175861e-03		1.701045e-03	
3	1184	6.914189e-05	4.09	6.914301e-05	4.09	2.286228e-04	2.90
	4736	2.912998e-06	4.57	2.969415e-06	4.54	4.289916e-05	2.41
	296	3.552213e-04		3.552213e-04		8.203638e-04	
4	1184	8.214109e-06	5.43	8.222967e-06	5.43	9.042798e-05	3.18
	4736	2.157931e-07	5.25	2.538417e-07	5.02	1.947955e-05	2.21
	296	8.248182e-05		8.248182e-05		4.255808e-04	
5	1184	1.080937e-06	6.25	1.256704e-06	6.03	4.455866e-05	3.26
	4736	1.540725e-08	6.13	1.664081e-08	6.24	1.031478e-05	2.11
	296	2.167114e-05		2.167448e-05		2.422031e-04	
6	1184	1.310204e-07	7.37	1.715729e-07	6.98	2.433940e-05	3.31
	4736	1.041520e-09	6.97	1.045177e-09	7.36	5.829489e-06	2.06

Tabelle 5.4: Wirkung der adaptiven sowie globalen Anwendung des modalen Filters für die Parameterwahl $p = 3,\ C_p = 0.01$ auf die experimentelle Konvergenzordnung.

6 Nachbearbeitung oszillationsbehafteter Näherungslösungen

Im Fall unstetiger exakter Lösungen führt das Vorhandensein Gibbsscher Oszillationen in den Näherungslösungen spektraler Verfahren dazu, dass die punktweisen Werte der berechneten Approximation keine hohe Genauigkeit aufweisen. Ist die Ermittlung vertrauenswürdiger punktweiser Werte notwendig – zum Beispiel zur Visualisierung nach dem Evolutionsprozess oder zu Zwischenzeitpunkten – so wird die Näherungslösung üblicherweise nachbearbeitet. Die Anwendung einer Nachbearbeitungsprozedur auf die Näherungslösungen spektraler Verfahren beruht hierbei auf der Vermutung, dass in der durch ein spektrales Verfahren gegebenen Näherungslösung selbst im Fall einer unstetigen exakten Lösung hochgenaue Informationen enthalten sind, die eine Rekonstruktion punktweiser Werte mit hoher Genauigkeit ermöglichen.
Im Fall linearer Gleichungen kann diese Vermutung bewiesen werden. Ein entsprechendes Resultat von Abarbanel, Gottlieb und Tadmor [1] besagt, dass für die durch eine Fourier-Spektralmethode ermittelte numerische Lösung u_N einer linearen periodischen Erhaltungsgleichung mit exakter Lösung u eine schwache Fehlerabschätzung der Form

$$|(u - u_N, \phi)_{L^2[0,2\pi]}| \leq CN^{-s}\|\phi\|_{H^s[0,2\pi]}\,, \quad \forall \phi \in H^s[0, 2\pi],$$

gilt. Die Fourier-Koeffizienten der numerischen Lösung approximieren demzufolge diejenigen der exakten Lösung mit spektraler Genauigkeit, so dass eine hochgenaue Approximation an die *Projektion* der exakten Lösung in den jeweiligen Ansatzraum gegeben ist. Aufgrund des in Abschnitt 5.1 beschriebenen Gibbs-Phänomens besitzt diese Projektion zunächst sehr schlechte punktweise Approximationseigenschaften, jedoch sind bereits sehr fortgeschrittene Techniken entwickelt worden, die unter Voraussetzung der Kenntnis beziehungsweise der hochgenauen Approximation der ersten N Fourier-Koeffizienten der exakten Lösung die Rekonstruktion punktweiser Werte bis an die Unstetigkeitsstellen heran mit spektraler Genauigkeit ermöglichen. Eine derartige Technik ist beispielsweise die von Gottlieb und Shu entwickelte Gegenbauer-Rekonstruktion [39], die auf der stückweisen Projektion der abgeschnittenen Fourier-Reihe in eine abgeschnittene Entwicklung in Gegenbauer-Polynome basiert.
Ob es für nichtlineare Erhaltungsgleichung mit unstetiger Lösung im Allgemeinen möglich ist, hochgenaue Informationen aus einer stabilen spektralen Approximation zu extrahieren, wie es bereits in [60] von Lax vermutet wurde, ist jedoch noch nicht geklärt. Ebenso ist es nicht klar, in welcher Form diese Informationen in den Näherungslösungen repräsentiert sein könnten. Insbesondere zeigen Shu und Wong in [85], dass sich das Enthaltensein hochgenauer Informationen in einer durch spektrale Viskosität stabilisierten Näherungslösung der Burgers-Gleichung nicht in einer hochgenauen Approximation der Fourier-Koeffizienten durch die spektralen Methode manifestieren muss. In den dort durchgeführten numerischen Experimenten weisen die punktweisen Fehler der mit Hilfe der Gegenbauer-Methode rekonstruierten Daten deutlich bessere Größenordnungen auf als die Fehler in den numerisch ermittelten Fourier-Koeffizienten. Die Anwendung einer Reprojektionstechnik ist in diesem Fall jedoch nicht aus der Theorie heraus begründbar, da diese auf der Kenntnis der exakten Koeffizienten beruht.
Da die Reprojektion (beispielsweise mittels Gegenbauer-Polynomen) zudem auf Teilbereichen erfolgen muss, in denen die Funktion analytisch ist, müssen diese Bereiche in einem vorhergehenden Schritt mittels Kantenerkennungsverfahren identifiziert werden, ein

Beispiel stellt das Verfahren von Gelb und Tadmor [33] dar. Neben der Rundungsfehler-Sensibilität für große Längen N der Partialsumme des Spektralverfahrens und der Notwendigkeit, für die Anwendung im konkreten Fall je zwei Parameter der Methode für jeden glatten Teilbereich zu wählen, vgl. hierzu [32], liegt die Schwierigkeit der Gegenbauer-Rekonstruktion daher in der Übertragbarkeit auf den allgemeinen zweidimensionalen Fall. In dieser Situation sind Unstetigkeits*kurven* zu erkennen und die Gebiete, in denen die neue Entwicklung zu konstruieren ist, können sich deutlich von einfachen Gebieten wie beispielsweise geradlinig berandeten Polygonen unterscheiden. Aufgrund dieser Problematik soll im folgenden anstelle einer Reprojektionstechnik ein häufig verwendetes Verfahren der Bildverarbeitung auf seine Nutzbarkeit als Nachbearbeitungstechnik für die Näherungslösungen der diskontinuierlichen Galerkin-Methode mit modaler Filterung untersucht werden.

6.1 Algorithmen der Bildverarbeitung zur Oszillationsfilterung

Das Anwendungsgebiet der Bildverarbeitung beschäftigt sich unter anderem mit der Bildwiederherstellung, d.h. mit der approximativen Rekonstruktion von aus einer realen Situation gewonnenen Bildern, deren Qualität durch ihre Erstellung, Übertragung und Aufzeichnung beeinträchtigt wurde. Die Verminderung der Bildqualität kann zum einen deterministische Gründe besitzen, die aus der Art der Bilderfassung hervorgehen (beispielsweise Unschärfe durch falsches Einstellen der Linse oder Bewegung des Objekts) und ist zum anderen auf zufällige Störungen durch stochastisches Rauschen, welches jeder Signalübertragungsmethode inherent ist, zurückzuführen. Für eine fundierte mathematische Beschreibung der grundlegenden Modelle und Methoden der Bildverarbeitung sei auf das Buch von Aubert und Kornprobst [2] verwiesen.

Die Verwendung von Methoden der Bildwiederherstellung, entweder als Teilkomponente innerhalb eines numerischen Verfahrens zur Lösung zweidimensionaler hyperbolischer Erhaltungsgleichungen oder auch als Nachbearbeitungsprozedur angewandt auf die Näherungslösung zu Visualisierungszeitpunkten, beruht auf der grundsätzlichen Idee, die numerische Approximation zu einem gegebenen Zeitpunkt t als ein verrauschtes (oder im Fall hoher Viskosität des numerischen Verfahrens als ein an Stößen unscharfes) Bild aufzufassen. Aufgabe der gewählten Bildverarbeitungsmethode ist es dann, die exakte Lösung zum Zeitpunkt t, die das Originalbild repräsentiert, aus den gegebenen Daten so gut wie möglich wiederherzustellen. Die Ziele der Bildwiederherstellung kommen hierbei auch im Kontext der numerischen Lösung hyperbolischer Erhaltunggleichungen gelegen: Die Störungen durch Rauschen oder Unschärfe sollen beseitigt werden, während die relevanten Informationen des Bildes, insbesondere die Kanten, erhalten bleiben. Während viele Ansätze der Bildwiederherstellung partielle Differentialgleichungen zur Grundlage haben und dadurch von der Ausgefeiltheit numerischer Methoden zu deren Lösung profitieren, findet sich in umgekehrter Richtung die Anwendung von Ideen und Methoden der Bildverarbeitung innerhalb numerischer Verfahren zur Lösung hyperbolischer Erhaltungsgleichungen erst in jüngerer Literatur. Die meisten Arbeiten betreffen hierbei die Anwendung der Bildverarbeitungsmethode zur expliziten Steuerung der Viskosität des numerischen Verfahrens. So wurde in [41, 42] zur Stabilisierung eines oszillierenden Lax-Wendroff-Verfahrens anisotrope, insbesondere normal zu Stößen verschwindende, Diffusion eingeführt, auf der Grundlage einer von Weickert [96] zur Bildverarbeitung entwickelten

Diskretisierung der anisotropen Diffusionsgleichung

$$\frac{\partial u}{\partial t} = \nabla \cdot (D(J_\rho(\nabla u_\sigma))\nabla u) \ . \tag{6.1}$$

Hierbei bezeichnet u_σ die Faltung von u mit dem Gaußschen Kern $k_\sigma(\mathbf{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\mathbf{x}|^2}{2\sigma^2}\right)$, zur Vermeidung einer fälschlichen Interpretation von zufälligem Rauschens als Kante des Bildes. Der Strukturtensor $J_\rho = k_\rho * \nabla u_\sigma \nabla u_\sigma^T$ sorgt für das Einbeziehen lokaler Informationen zur Steuerung des nichtlinearen Diffusionstensors D mit Werten in $\mathbb{R}^{2\times 2}$. Die Matrix $\nabla u_\sigma \nabla u_\sigma^T$ besizt eine Basis aus orthonormalen Eigenvektoren, mit einem Eigenvektor zum Eigenwert $|\nabla u_\sigma|^2$, der parallel zu ∇u_σ verläuft und einem Eigenvektor zum Eigenwert 0 senkrecht zu ∇u_σ. Der Diffusionstensor D wird nun so gewählt, dass er dieselben Eigenvektoren besitzt und die Wahl der zugehörigen Eigenwerte zur Reduktion der Diffusion parallel zu ∇u_σ für große Werte von $|\nabla u_\sigma|^2$ führt, so dass die Diffusion normal zu Kanten des Bildes verschwindet. Die Differentialgleichung (6.1) ist eine Erweiterung des Modells inhomogener Diffusion

$$\frac{\partial u}{\partial t} = \nabla \cdot \big(c(|\nabla u|^2)\nabla u\big) \tag{6.2}$$

von Perona und Malik, in dem die Diffusion durch eine glatte, monoton fallende Funktion $c : [0, \infty) \to [0, \infty)$ gesteuert wird und welches neben der Reduktion der Viskosität in Bereichen, die mit großer Wahrscheinlichkeit Kanten sind, auch zu rückwärtsgerichter Diffusion, also Verschärfung von Kanten, führen kann. Inhomogene Diffusion der Form (6.2) wurde ebenso von Wei [95] zur Stabilisierung von Finite-Differenzen-Verfahren zur Lösung hyperbolischer Erhaltungsgleichungen verwendet, allerdings mit $c(s) = \ln(s + 1)$, so dass die Diffusion nahe Stößen explizit erhöht wurde. Rückwärtsgerichtete Diffusion wurde von Breuß et al. [8] zur Verbesserung der Auflösung eines upwind-Verfahrens an Stößen eingeführt. Desweiteren untersuchte Bürgel [9] eine große Bandbreite nichtlinearer Filteralgorithmen, die sowohl aus Modellen partieller Differentialgleichungen wie in den obigen Beispielen, als auch gemäß Variationsansätzen oder aus vollständig diskreten Überlegungen wie der Idee der Restflächenverteilung hergeleitet werden können.

Im Gegensatz zu den bisher zitierten Arbeiten wird die Steuerung der Viskosität des hier entworfenen diskontinuierlichen Galerkin-Verfahrens von dem eingeführten spektralen Filter übernommen. Der Bildverarbeitungsmethode wird nun die Rolle einer Nachbearbeitungstechnik zugewiesen. Einen derartigen Ansatz findet man in der Arbeit von Sarra [81], in der der digitale TV-Filter (DTV-Filter) von Chan, Osher und Shen [14], der auch von Bürgel, Grahs und Sonar in [10] verwendet wurde und aus einem diskreten Optimierungsproblem hergeleitet werden kann, die Aufgabe der Nachbearbeitung von Näherungslösungen einer Chebyshev-Pseudospektralmethode mit spektraler Filterung übernimmt.

Aufgrund der einfachen Anwendbarkeit des DTV-Filters und der sehr guten Ergebnisse in Kombination mit numerischen Verfahren zur Lösung hyperbolischer Erhaltungsgleichungen soll diese digitale Bildverarbeitungsmethode im folgenden Abschnitt genauer beschrieben und auf seine Nutzbarkeit als Nachberarbeitungsverfahren für die oszillativen Näherungslösungen der DG-Methode mit spektraler Filterung untersucht werden. Erste Ergebnisse des DTV-Filters für skalare Erhaltungsgleichungen sind zudem in den Arbeiten [68, 69] beschrieben.

6.2 Der digitale TV-Filter

6.2.1 Grundlegende Eigenschaften

Der DTV-Filter von Chan, Osher und Shen ist auf ungerichteten Graphen $[V, E]$ mit endlicher Knotenmenge V und endlicher Kantenmenge E beschrieben. Für einen Knoten $\alpha \in V$ bezeichne $N_\alpha \subseteq V$ die Menge der *Nachbarknoten* bzw. die *Nachbarschaft* von $\alpha \in V$, d.h. die Menge derjenigen Knoten $\beta \in V$, die mit α durch eine Kante verbunden sind. Mit dem Vektor $\mathbf{u}^0 \in \mathbb{R}^{\#V}$ mit Komponenten $u^0_\alpha,\ \alpha \in V$ sei ein auf der Struktur des Graphen vorliegendes Ausgangssignal gegeben. Das Ausgangssignal $\mathbf{u}^0$ ist in dem hier interessierenden Kontext gegeben durch die nodalen Werte $u_h(\cdot, T)$ einer Näherungslösung des DG-Verfahrens mit modaler Filterung. Der Wert $T \in \mathbb{R}_+$ entspricht hierbei einem Ausgabezeitpunkt, für den eine Visualisierung und damit ein weitgehendes Entfernen restlicher Oszillationen erwünscht ist.
Der DTV-Filter ist nun durch eine iterative Prozedur $\mathbf{u}^k \to \mathbf{u}^{k+1}$ mit

$$u_\alpha^{k+1} = \sum_{\beta \in N_\alpha} h_{\alpha\beta}(\mathbf{u}^k) u_\beta^k + h_{\alpha\alpha}(\mathbf{u}^k) u_\alpha^0, \quad k = 0, 1, \ldots, \tag{6.3}$$

gegeben, wobei die Filterkoeffizienten in Abhängigkeit von den Daten $\mathbf{u} \in \mathbb{R}^{\#V}$ definiert sind durch

$$h_{\alpha\beta}(\mathbf{u}) = \frac{\omega_{\alpha\beta}(\mathbf{u})}{\lambda + \sum_{\gamma \in N_\alpha} \omega_{\alpha\gamma}(\mathbf{u})}, \qquad h_{\alpha\alpha}(\mathbf{u}) = \frac{\lambda}{\lambda + \sum_{\gamma \in N_\alpha} \omega_{\alpha\gamma}(\mathbf{u})}, \tag{6.4}$$

mit nichtnegativen, datenabhängigen Gewichten $\omega_{\alpha\beta}(\mathbf{u})$ und einem vom Benutzer einzustellenden Anpassungsparameter $\lambda \geq 0$. Die Gewichte sind gegeben durch

$$\omega_{\alpha\beta}(\mathbf{u}) = \frac{1}{|\nabla_\alpha \mathbf{u}|_\epsilon} + \frac{1}{|\nabla_\beta \mathbf{u}|_\epsilon},$$

wobei die Größe

$$|\nabla_\alpha \mathbf{u}|_\epsilon = \left[\sum_{\beta \in N_\alpha} (u_\beta - u_\alpha)^2 + \epsilon \right]^{1/2} \tag{6.5}$$

als die *regularisierte diskrete Lokalvariation* bezeichnet wird, in deren Definition der Regularisierungsparameter $\epsilon > 0$ zur Vermeidung der Division durch Null eingefügt ist. Die Filterkoeffizienten $h_{\alpha\beta}$, die den Einfluss benachbarter Werte auf die neue Iterierte steuern, nehmen demnach genau dann hohe Werte an, wenn im betrachteten Bereich die Lokalvariation gering ist, so dass hochfrequente Oszillationen mit vergleichsweise geringer Amplitude geglättet werden können. Stöße, über die hinweg die Lokalvariation hoch ist, führen hingegen zu dort vorliegenden geringen Werten der Filterkoeffizienten $h_{\alpha\beta}$, so dass der Einfluss der ursprünglichen Daten $\mathbf{u}^0$ überwiegt und der Stoß erhalten bleibt.
Der DTV-Filter wurde auf der Grundlage eines diskreten Variationsproblem konstruiert. So wurde in [14] gezeigt, dass sich die obige Wahl der Gewichte aus dem Variationsproblem

$$\min_{\mathbf{u} \in \mathbb{R}^{\#V}} \left\{ \sum_{\alpha \in V} |\nabla_\alpha \mathbf{u}|_\epsilon + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{u}^0\|_2^2 \right\}$$

ergibt. Die Wahl des Parameters λ entscheidet somit, wie sehr das Ausgangssignal $\mathbf{u}^0$ geglättet werden darf, um die Summe der Lokalvariationen zu vermindern und hat damit nicht unerheblichen Einfluss auf die Entscheidung, wann Feinstrukturen von $\mathbf{u}^0$ als wichtige Informationen und wann als überflüssige Oszillationen gelten.
Ein Konvergenznachweis für die Iterationsprozedur des DTV-Filters ist nach [14] noch nicht erbracht, jedoch deuten die numerischen Experimente in [14, 10, 9, 81] sowie eigene Berechnungen an, dass die iterative Anwendung von (6.3) für $\lambda > 0$ zu einem stationären Bild führt.

Eine Nachbearbeitungstechnik, die auf die Näherungslösung eines konservativen Verfahrens zur Lösung von Erhaltungsgleichungen angewendet wird, sollte nach Möglichkeit die Erhaltungseigenschaften der numerischen Methode nicht verletzen. Bei ausreichender Iterationszahl und unter Voraussetzung der Konvergenz ist der DTV-Filter in dieser Hinsicht zur Nachbearbeitung geeignet, da er, wie nachfolgend gezeigt wird, im Limit konservativ ist.

Lemma 6.1 *Ist das durch (6.3), (6.4) gegebene iterative Verfahren konvergent und gilt $\lambda > 0$, so erfüllt der Grenzwert $\mathbf{u} = \lim_{k\to\infty} \mathbf{u}^k$ die globale Erhaltungseigenschaft*

$$\sum_{\alpha\in V} u_\alpha = \sum_{\alpha\in V} u_\alpha^0.$$

Beweis: Da $\mathbf{u}$ ein Fixpunkt von (6.3) ist, liefert eine Multiplikation mit $\lambda + \sum_{\gamma\in N_\alpha} \omega_{\alpha\gamma}(\mathbf{u})$ die Gleichung

$$\left(\lambda + \sum_{\gamma\in N_\alpha} \omega_{\alpha\gamma}(\mathbf{u})\right) u_\alpha = \sum_{\beta\in N_\alpha} \omega_{\alpha\beta}(\mathbf{u}) u_\beta + \lambda u_\alpha^0, \quad \forall \alpha \in V,$$

die umgestellt werden kann zu

$$\sum_{\beta\in N_\alpha} \omega_{\alpha\beta}(\mathbf{u})(u_\beta - u_\alpha) + \lambda(u_\alpha^0 - u_\alpha) = 0, \quad \forall \alpha \in V.$$

Summiert man über alle Knoten in V, so liefert die Symmetrieeigenschaft $\omega_{\alpha\beta}(\mathbf{u}) = \omega_{\beta\alpha}(\mathbf{u})$ der Gewichte das Verschwinden der ersten Summe der obigen Gleichung. Man erhält somit

$$\lambda \cdot \sum_{\alpha\in V} (u_\alpha^0 - u_\alpha) = 0.$$

Da $\lambda \neq 0$ vorausgesetzt wurde, ist die Behauptung hiermit gezeigt. □

6.2.2 Modifikation zur Verwendung auf Dreiecksgittern

Es ist grundsätzlich möglich, den digitalen TV-Filter auf ein Ausgangssignal anzuwenden, welches durch die Auswertung der Näherungslösung des DG-Verfahrens mit modaler Filterung an Gitterpunkten eines kartesischen Gitters erzeugt wird. Obwohl eine solche

Vorgehensweise für verschieden feine Gitter prinzipiell realisierbar ist und optisch ozillationsfreie Ergebnisse liefern kann, ist das Erzeugen eines derartigen Ausgangssignals mit einigem Aufwand verbunden. Zum einen muss für jeden kartesischen Gitterpunkt ermittelt werden, in welchem Dreieck der Triangulierung er sich befindet, bevor das dort vorliegende Polynom an der entsprechenden Stelle ausgewertet werden kann, und zum anderen ist unklar, wie das Ausgangssignal an einem auf einer Dreieckskante liegenden Gitterpunkt zu ermitteln ist.
Mit dem Ziel einer einfachen Berechnung des Ausgangssignals $\mathbf{u}^0$ aus einer auf einem Dreiecksgitter gegebenen Näherungslösung soll die Anwendung des DTV-Filters auf Graphen ermöglicht werden, deren Knoten V die Schwerpunkte von Teildreiecken sind, die durch Zerlegen jedes Dreieckselements entstehen. Wie in der Abbildung 6.1 skizziert, werden hierzu eine feste Anzahl äquidistanter Punkte auf jede Dreieckskante gesetzt und die Punkte auf zwei verschieden Kanten eines Dreiecks durch zur dritten Kante parallele Linien verbunden. Zwei Knoten im DTV-Graphen, d.h. die Schwerpunkte zweier Teildreiecke der Zerlegung eines Dreiecksgitters, sind genau dann durch eine Kante im DTV-Graph verbunden, wenn sie eine gemeinsame Kante in der durch die Zerlegung gegebenen Triangulierung besitzen.

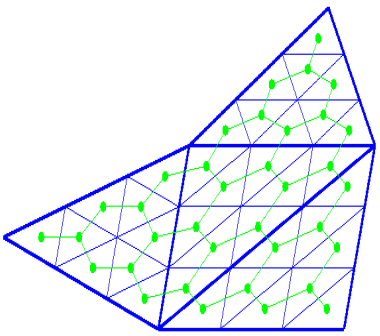
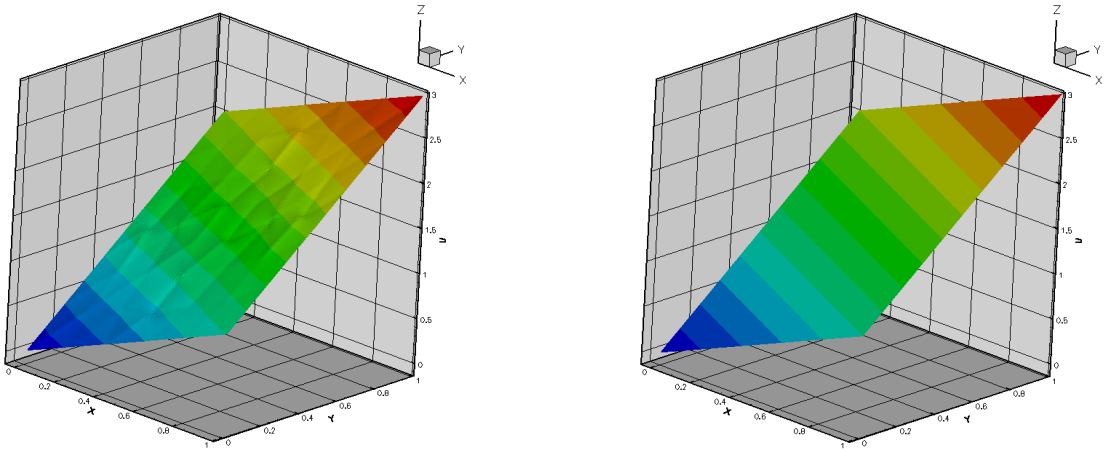


Abb. 6.1: DTV-Graph (grün) basierend auf der Zerlegung eines Dreiecksgitters (dicke blaue Kanten).

Anwendung des unmodifizierten DTV-Filters Auf Graphen, die wie oben beschrieben an die Struktur eines gegebenen Dreiecksgitters angepasst sind, lassen sich Knotenwerte leicht durch Auswertung der Näherungslösung an auf jedem Dreieck gleich verteilten Punkten berechnen. Eine direkte Anwendung des in Abschnitt 6.2.1 beschriebenen DTV-Filters auf derartig konstruierten Graphen liefert jedoch unbefriedigende Ergebnisse. Zur Veranschaulichung des Problems betrachten wir das Ausgangssignal $\mathbf{u}^0$, mit $u^0_\alpha = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \mathbf{x}_\alpha$, welches durch die Auswertung der linearen Funktion $f(\mathbf{x}) = x_1 + 2x_2$ an den mit $\mathbf{x}_\alpha \in \mathbb{R}^2$ bezeichneten Schwerpunkten der Teildreiecke eines zerlegten Dreiecksgitters gegeben ist. Die naive Anwendung der Iterationsvorschrift (6.3) des DTV-Filters auf einem Graphen, der durch die Zerlegung einer 74-elementigen Triangulierung des Gebiets $\Omega = [0,1]^2$ in je 81 Teildreiecke gegeben ist, liefert nach 100 Iterationen mit $\lambda = 2$ das in Abbildung 6.2a gezeigte gefilterte Signal. Es zeigt eine der Gitterstruktur der zugrunde liegenden Triangulierung folgende Musterung anstelle der aus Sicht der Filterung von Oszil-

lationen bzw. Rauschen sinnvolleren Wiedergabe der ursprünglichen linearen Verteilung. Dem gegenübergestellt ist das gefilterte Signal in Abbildung 6.2b, das bei der Anwendung des DTV-Filters auf 100×100 Gitterpunkten eines kartesischen Gitters erzielt wurde, bei gleicher Iterationszahl und Definition des Parameters λ. In diesem Beispiel wurden auf beiden Graphen die Werte an Randkonten sowie an Knoten, die mit einem Randknoten verbunden sind, festgehalten. Als Randknoten werden in Dreiecksgitter-basierten Graphen hierbei die Schwerpunkte von Teildreiecken bezeichnet, die mindestens eine auf dem Gebieltsrand liegende Kante besitzen – in kartesischen Graphen umfassen sie die auf dem Gebietsrand liegenen Knoten. Somit sind die Randknoten genau die Knoten, deren Nachbarschaft weniger als 3 (auf Dreiecksgitter-basierten Graphen) beziehungsweise 4 (auf kartesischen Graphen) Knoten enthält. Die veränderte Nachbarschaftsstruktur wirkt sich insbesondere auf den Wert der in (6.5) definierten Lokalvariation und die daraus berechneten Gewichte an Randknoten sowie an Knoten, die mit mindestens einem Randknoten verbunden sind aus – ein derartiger Einfluss sollte im obigen Beispiel unterbunden werden.



(a) Dreiecksgitter-basierter DTV-Graph (b) Kartesischer DTV-Graph

Abb. 6.2: Gefilterte Signale – DTV-Filter nach ursprünglicher Definition.

Modifikation der Iterationsvorschrift Um derartige Artifakte des DTV-Filters auf Graphen, denen eine Triangulierung zugrunde liegt, zu verhindern, wird im Folgenden eine Modifikation der Iterationsvorschrift (6.3) vorgenommen, die das Vorliegen einer linearen Verteilung erkennt. Dem zugrunde liegt die Erkenntnis, dass das Fehlverhalten des DTV-Filters in seiner ursprünglichen Definition bei Verwendung von Dreiecksgitter-basierten Graphen darauf zurückzuführen ist, dass Kantenlängen und Winkel der Teildreiecke der zerlegten Triangulierung sich genau über Elementgrenzen hinweg ändern können.
Zur Definition einer modifizierten Iterationsvorschrift verwenden wir zunächst die nachfolgenden Definitionen zur Unterscheidung zweier Knoten- beziehungsweise Kantenmengen. Die Abbildung $I : V \to \{1, \dots, \mathcal{T}^h\}$, die jedem Knoten des DTV-Graphs das eindeutig definierte zugehörige Element der Triangulierung zuordnet, ist durch

$$I(\alpha) = i \Leftrightarrow \mathbf{x}_\alpha \in \tau_i$$

bestimmt. Zu einem Knoten $\alpha \in V$ bezeichnen wir die Menge der Nachbarknoten im

selben Dreieck $\tau_{I(\alpha)}$ mit

$$N_\alpha^1 = \{\beta \in N_\alpha \,|\, \mathbf{x}_\beta \in \tau_{I(\alpha)}\}$$

und die Menge der Nachbarknoten in zu $\tau_{I(\alpha)}$ benachbarten Dreiecken mit $N_\alpha^2 = N_\alpha \backslash N_\alpha^1$. Die gesamte Knotenmenge V zerlegt sich dann in die Knotenmenge $V^1 = \{\alpha \in V \,|\, N_\alpha^2 = \emptyset\}$, bestehend aus Knoten, deren Nachbarschaft ganz in einem Dreieck der Triangulierung enthalten ist, und das Komplement $V^2 = V \backslash V^1$.
Der modifizierte DTV-Filter unterscheidet nun die Mengen N_α^1 und N_α^2 durch eine modifizierte Iterationsvorschrift der Form

$$u_\alpha^{k+1} = \sum_{\beta \in N_\alpha^1} \tilde{h}_{\alpha\beta}(\mathbf{u}^k) u_\beta^k + \sum_{\beta \in N_\alpha^2} \tilde{h}_{\alpha\beta}(\mathbf{u}^k) \tilde{u}_\beta^k + \tilde{h}_{\alpha\alpha}(\mathbf{u}^k) u_\alpha^0, \quad k = 0, 1, \ldots .$$

In die Iterationsvorschrift gehen modifizierte Filterkoeffizienten $\tilde{h}_{\alpha\beta}$ ein, sowie modifizierte Werten $\tilde{u}_\beta^k$ an Knoten $\beta \in N_\alpha^2$, die durch lineare Interpolation geeigneter Werten an Knoten der Menge $N_\beta \cup \{\beta\}$ gewonnen werden. Ziel der Interpolation ist das Erkennen einer möglicherweise repräsentierten linearen Verteilung in den auf dem DTV-Graphen gegebenen Daten. Hierzu wird für jeden Knoten $\beta \in N_\alpha^2$ ein Pseudoknoten konstruiert, dessen Koordinaten $\tilde{\mathbf{x}}_\beta$ gegeben sind durch die Fortsetzung der Gitterstruktur auf $\tau_{I(\alpha)}$ über die jeweilige Kante hinaus und die Berechnung der entsprechenden Schwerpunkte, wie in Abbildung 6.3 angedeutet. In der dort abgebildeten Situation werden beispielsweise zur Ermittlung der neuen Iterierten u_α^{k+1} an $\mathbf{x}_\alpha$ neben dem Ausgangssignal die Werte an den den rot markierten Knoten verwendet. Da durch die Fortsetzung der Gitterstruktur

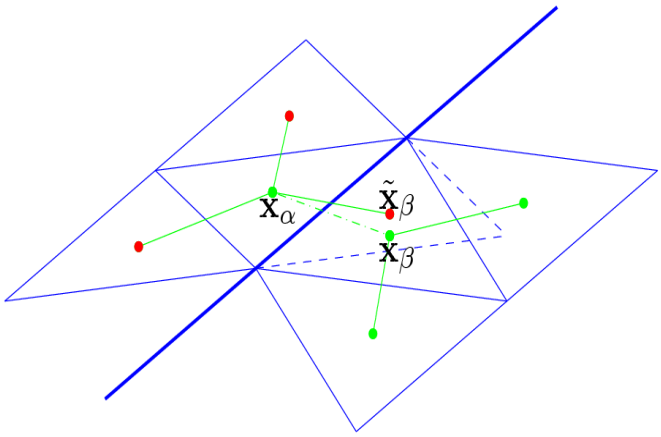


Abb. 6.3: Fortsetzung des DTV-Gitters über eine Kante zwischen zwei Dreiecken der Triangulierung.

jeweils Parallelogramme erzeugt werden, lassen sich die Koordinaten $\tilde{\mathbf{x}}_\beta$ explizit berechnen durch

$$\tilde{\mathbf{x}}_\beta = \frac{2}{3}\mathbf{p}_1^\alpha + \frac{2}{3}\mathbf{p}_2^\alpha - \frac{2}{3}\mathbf{p}_3^\alpha,$$

wobei mit $\mathbf{p}_j^\alpha$ die Koordinaten der Eckpunkte des zu α gehörigen (in $\tau_{I(\alpha)}$ liegenden) Teildreiecks bezeichnet sind, die so angeordnet werden, dass $\mathbf{p}_1^\alpha,\ \mathbf{p}_2^\alpha \in \tau_{I(\beta)}$ die Knoten

auf der Elementgrenze $\partial\tau_{I(\alpha)} \cap \partial\tau_{I(\beta)}$ sind. Unter der Annahme, dass bereits geeignete Differenzenvektoren $\mathbf{d}_\beta(\mathbf{u})$ ermittelt wurden, ist der modifizierte Wert $\tilde{u}_\beta$ definiert als

$$\tilde{u}_\beta = u_\beta + \mathbf{d}_\beta(\mathbf{u}) \cdot (\tilde{\mathbf{x}}_\beta - \mathbf{x}_\beta).$$

Der Einfluss einer veränderten Gitterstruktur über Elementgrenzen hinweg wird durch eine derartige Konstruktion behoben. Zusätzlich ist es dazu ebenso erforderlich, die Lokalvariation an Pseudoknoten zu ermitteln, die Lokalvariation an zu Pseudoknoten benachbarten Knoten geeignet zu modifizieren und die Filterkoeffizienten anzupassen.
Die modifizierte Lokalvariation an den Knoten $\alpha \in V$ wird definiert als

$$|\tilde{\nabla}_\alpha u|_\epsilon = \begin{cases} |\nabla_\alpha u|_\epsilon, & \text{falls } \alpha \in V^1, \\ \sqrt{\epsilon + \sum_{\beta\in N_\alpha^1}(u_\beta - u_\alpha)^2 + \sum_{\beta\in N_\alpha^2}(\tilde{u}_\beta - u_\alpha)^2}, & \text{falls } \alpha \in V^2. \end{cases}$$

Falls $\mathbf{u}$ eine lineare Verteilung auf dem DTV-Graphen repräsentiert, ist die Lokalvariation nach dieser Definition somit konstant für alle Knoten innerhalb eines Dreiecks, die weder Randknoten sind, noch zu Randknoten benachbart. Desweiteren definieren wir für jeden Knoten $\beta \in N_\alpha^2$, der kein Randknoten ist, durch

$$|\tilde{\nabla}_\beta^\alpha u|_\epsilon = \sqrt{\epsilon + \sum_{j=1}^{3} \left(\mathbf{d}_\beta(\mathbf{u}) \cdot (\mathbf{y}_{\alpha,j} - \mathbf{x}_\alpha)\right)^2}, \quad \mathbf{y}_{\alpha,j} = \frac{2}{3}\sum_{\substack{k=1\\k\neq j}}^{3} \mathbf{p}_k^\alpha - \frac{1}{3}\mathbf{p}_j^\alpha,$$

die Lokalvariation an den durch die Koordinaten $\tilde{\mathbf{x}}_\beta$ gegebenen Pseudoknoten. Diese wird somit gewonnen durch Übertragen der an β berechneten linearen Verteilung bzw. dem berechneten Differenzenvektor auf die Nachbarschaftsstruktur in $\tau_{I(\alpha)}$ und ist dadurch unabhängig von der Form der Dreiecke $\tau_{I(\beta)},\ \beta \in N_\alpha^2$. Um den Einfluss der veränderten Nachbarschaftsstruktur an Randknoten einzubeziehen, wird für einen Randknoten $\beta \in N_\alpha^2$ die Lokalvariation am zugehörigen Pseudoknoten definiert durch

$$|\tilde{\nabla}_\beta^\alpha u|_\epsilon = \sqrt{\epsilon + \sum_{j=1}^{2} \left(\mathbf{d}_\beta(\mathbf{u}) \cdot (\mathbf{y}_{\alpha,j} - \mathbf{x}_\alpha)\right)^2}, \quad \mathbf{y}_{\alpha,j} = \frac{2}{3}\sum_{\substack{k=1\\k\neq j}}^{3} \mathbf{p}_k^\alpha - \frac{1}{3}\mathbf{p}_j^\alpha,$$

wobei $\mathbf{p}_j^\alpha$ in diesem Fall so angeordnet sind, dass die Punkte $\mathbf{p}_1^\alpha,\ \mathbf{p}_2^\alpha \in \partial\Omega$ auf dem Rand des Rechengebiets liegen. (In der obigen Berechnungsvorschrift wurde die Annahme verwendet, dass jedes Dreieck der zugrunde liegenden Triangulierung in mehr als ein Teildreieck zerlegt wird. Dann gilt für die Anzahl von Nachbarknoten $|N_\beta| \geq 2$ für alle $\beta \in V^2$.)
Die modifizierten Filterkoeffizienten werden schließlich analog zur ursprünglichen Definition gesetzt als

$$\tilde{h}_{\alpha\beta}(\mathbf{u}) = \frac{\tilde{\omega}_{\alpha\beta}(\mathbf{u})}{\lambda + \sum_{\gamma\in N_\alpha} \tilde{\omega}_{\alpha\gamma}(\mathbf{u})}, \qquad \tilde{h}_{\alpha\alpha}(\mathbf{u}) = \frac{\lambda}{\lambda + \sum_{\gamma\in N_\alpha} \tilde{\omega}_{\alpha\gamma}(\mathbf{u})}, \tag{6.6}$$

mit

$$\tilde{\omega}_{\alpha\beta}(\mathbf{u}) = \begin{cases} (|\tilde{\nabla}_\alpha \mathbf{u}|_\epsilon)^{-1} + (|\tilde{\nabla}_\beta \mathbf{u}|_\epsilon)^{-1}, & \text{falls } \beta \in N_\alpha^1, \\ (|\tilde{\nabla}_\alpha \mathbf{u}|_\epsilon)^{-1} + (|\tilde{\nabla}_\beta^\alpha \mathbf{u}|_\epsilon)^{-1}, & \text{falls } \beta \in N_\alpha^2. \end{cases}$$

Es bleibt die Berechnung eines Differenzvektors $\mathbf{d}_\beta(\mathbf{u})$ an Knoten $\beta \in V^2$, der durch lineare Interpolation der Werte an drei Knoten in der Menge $N_\beta \cup \{\beta\}$ gegeben ist. Gilt $|N_\beta| = 3$, so werden die Werte an den Knoten $\gamma_1, \gamma_2, \gamma_3 \in N_\beta$ verwendet, um eine lineare Verteilung zu berechnen. Im Fall $|N_\beta| = 2$, verwenden wir die Knoten in der Menge $N_\beta \cup \{\beta\}$. In expliziter Form ergibt sich

$$\mathbf{d}_\beta(\mathbf{u}) = \begin{pmatrix} (\mathbf{x}_{\gamma_2} - \mathbf{x}_{\gamma_1})^T \\ (\mathbf{x}_{\gamma_3} - \mathbf{x}_{\gamma_1})^T \end{pmatrix}^{-1} \begin{pmatrix} u_{\gamma_2} - u_{\gamma_1} \\ u_{\gamma_3} - u_{\gamma_1} \end{pmatrix}.$$

Werden 100 Iterationen mit dem derartig modifizierten DTV-Filter, mit $\lambda = 2$, für das obige Ausgangssignal ausgeführt, wobei die Werte an Randknoten sowie an Knoten, die mit mindestens einem Randknoten verbunden sind, festgehalten werden, so ergibt sich dass in Abbildung 6.4 gezeigte gefilterte Signal – ohne durch die Struktur des Graphen induzierte Artifakte.

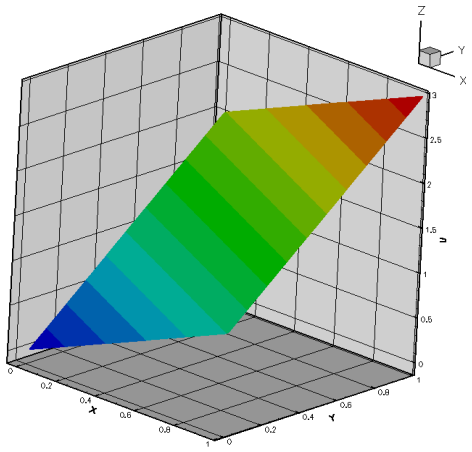


Abb. 6.4: Gefiltertes Signal – Modifizierter DTV-Filter.

Die Beobachtung, dass eine Iteration des modifizierten DTV-Filters lineare Funktionen erhält (an Knoten die weder selbst Randknoten, noch durch eine Kante mit einem Randknoten verbunden sind), kann auch theoretisch nachgewiesen werden.

Lemma 6.2 *Das Ausgangssignal* $\mathbf{u}^0$ *auf einem Dreiecksgitter-basierten DTV-Graphen repräsentiere eine lineare Verteilung, d.h. es gelte*

$$u^0_\alpha = \mathbf{d} \cdot \mathbf{x}_\alpha + c, \quad \forall \alpha \in V,$$

mit $\mathbf{d} \in \mathbb{R}^2, c \in \mathbb{R}$. *Dann gilt für jeden Knoten* $\alpha \in V$, *der weder Randknoten noch mit einem Randknoten verbunden ist, nach Ausführung einer Iteration des modifizierten DTV-Filters*

$$u^1_\alpha = u^0_\alpha.$$

Beweis: Sei $\alpha \in V$ ein Knoten mit $|N_\alpha| = 3$, d.h. kein Randknoten. Aufgrund der von der Knotenlage abhängigen Definition der Lokalvariation sind zunächst die Gewichte $\tilde{\omega}_{\alpha\beta}$ des modifizierten DTV-Filters für die gegebenen Daten $\mathbf{u}^0$ an allen von α ausgehenden

Kanten konstant, d.h. es gilt $\tilde{\omega}_{\alpha\beta} = \tilde{\omega}_\alpha$ für alle $\beta \in N_\alpha$. Hierzu genügt es zu zeigen, dass die entsprechenden Lokalvariationen konstant sind, d.h. dass $|\tilde{\nabla}_{\beta_1} u|_\epsilon = |\tilde{\nabla}_{\beta_2} u|_\epsilon$ für alle $\beta_1, \beta_2 \in N_\alpha^1$ sowie $|\tilde{\nabla}_{\beta_1} u|_\epsilon = |\tilde{\nabla}^\alpha_{\beta_2} u|_\epsilon$ für alle $\beta_1 \in N_\alpha^1,\ \beta_2 \in N_\alpha^2$ gilt. Aufgrund der vorausgesetzten linearen Verteilung gilt $\mathbf{d}_\beta(\mathbf{u}^0) = \mathbf{d}$ für alle $\beta \in N_\alpha^2$, so dass offensichtlich

$$|\tilde{\nabla}^\alpha_\beta u|_\epsilon = \sqrt{\epsilon + \sum_{j=1}^{3} \left(\mathbf{d} \cdot (\mathbf{y}_{\alpha,j} - \mathbf{x}_\alpha)\right)^2}$$

für alle $\beta \in N_\alpha^2$ konstant ist. Aufgrund der gleichmäßigen Gitterstruktur auf jedem Dreieck gilt desweiteren für einen beliebigen Knoten $\beta \in N_\alpha^1$ die Identität der Mengen

$$\begin{aligned} \{\mathbf{x}_\gamma - \mathbf{x}_\beta \,|\, \gamma \in N_\alpha^1\} \cup \{\tilde{\mathbf{x}}_\gamma - \mathbf{x}_\beta \,|\, \gamma \in N_\alpha^2\} &= \{\mathbf{x}_\alpha - \mathbf{x}_\gamma \,|\, \gamma \in N_\alpha^1\} \cup \{\mathbf{x}_\alpha - \tilde{\mathbf{x}}_\gamma \,|\, \gamma \in N_\alpha^2\} \\ &= \{\mathbf{x}_\alpha - \mathbf{y}_{\alpha,j} \,|\, j = 1,2,3\}, \end{aligned} \tag{6.7}$$

so dass

$$|\tilde{\nabla}_\beta u|_\epsilon = \sqrt{\epsilon + \sum_{\gamma \in N_\beta^1} (\mathbf{d} \cdot (\mathbf{x}_\gamma - \mathbf{x}_\beta))^2 + \sum_{\gamma \in N_\beta^2} (\mathbf{d} \cdot (\tilde{\mathbf{x}}_\gamma - \mathbf{x}_\beta))^2}$$

für alle $\beta \in N_\alpha^1$ konstant ist und der Wert mit der Lokalvariation an den zu α benachbarten Pseudoknoten übereinstimmt. Aus

$$\sum_{j=1}^{3} \mathbf{y}_{\alpha,j} = \sum_{j=1}^{3} \frac{2}{3} \sum_{\substack{k=1 \\ k \neq j}}^{3} \mathbf{p}_k^\alpha - \sum_{j=1}^{3} \frac{1}{3} \mathbf{p}_j^\alpha = 3\mathbf{x}_\alpha,$$

und der Identität (6.7) folgt desweiteren die Beziehung

$$\sum_{\beta \in N_\alpha^1} \mathbf{x}_\beta + \sum_{\beta \in N_\alpha^2} \tilde{\mathbf{x}}_\beta = 3\mathbf{x}_\alpha.$$

Eine Iteration des modifizierten DTV-Filters liefert nun

$$\begin{aligned} u_\alpha^1 &= \sum_{\beta \in N_\alpha^1} \frac{\tilde{\omega}_\alpha}{\lambda + 3\tilde{\omega}_\alpha} u_\beta^0 + \sum_{\beta \in N_\alpha^2} \frac{\tilde{\omega}_\alpha}{\lambda + 3\tilde{\omega}_\alpha} \tilde{u}_\beta^0 + \frac{\lambda}{\lambda + 3\tilde{\omega}_\alpha} u_\alpha^0 \\ &= \frac{\tilde{\omega}_\alpha}{\lambda + 3\tilde{\omega}_\alpha} \left(\sum_{\beta \in N_\alpha^1} u_\alpha^0 + \mathbf{d} \cdot (\mathbf{x}_\beta - \mathbf{x}_\alpha) + \sum_{\beta \in N_\alpha^2} u_\alpha^0 + \mathbf{d} \cdot (\tilde{\mathbf{x}}_\beta - \mathbf{x}_\alpha) \right) + \frac{\lambda}{\lambda + 3\tilde{\omega}_\alpha} u_\alpha^0 \\ &= \frac{\tilde{\omega}_\alpha}{\lambda + 3\tilde{\omega}_\alpha} \mathbf{d} \cdot \left(\sum_{\beta \in N_\alpha^1} \mathbf{x}_\beta + \sum_{\beta \in N_\alpha^2} \tilde{\mathbf{x}}_\beta - 3\mathbf{x}_\alpha \right) + u_\alpha^0 \\ &= u_\alpha^0, \end{aligned}$$

so dass der Wert am Knoten α unverändert bleibt. □

Bemerkung 6.3 *Der Nachteil der in diesem Abschnitt beschriebenen Modifikation des DTV-Filters zur Verwendung auf Dreiecksgitter-basierten Graphen ist allerdings, dass die in Lemma 6.1 gegebene Erhaltungseigenschaft nicht mehr nachweisbar ist. Dies gilt ebenso für das Festhalten von Randdaten.*

6.2.3 Adaptive Anwendung

Die Anwendung des DTV-Filters mit globalem Parameter λ kann zu einem Ordnungsverlust der nachbearbeiteten Näherungslösung in glatten Bereichen der exakten Lösung führen, wie sich in den numerischen Experimenten zeigen wird. Dementsprechend ist es sinnvoll, einen räumlich adaptiven Parameter λ zuzulassen, oder eine direkte Spezifikation von Bereichen, in denen der DTV-Filter angewendet werden soll, vorzunehmen.
In [10] wird demnach in Abhängigkeit der berechneten lokalen Varianz in der Umgebung eines Knotens ein über das räumliche Gebiet variierender Parameter λ bestimmt. Wird die lokale Varianz über einen Stoß hinweg berechnet, schlägt diese Methode fehl, so dass die Verwendung eines Kantendetektors vorgeschlagen wird. In [81] werden verschiedene Modifikationen des DTV-Filters untersucht, darunter ebenfalls die Berechnung eines adaptiven Parameters λ, sowohl mit als auch ohne Kantendetektor. Bessere Filtereigenschaften wurden hierbei allerdings im Fall hybrider Methoden aus spektraler und TV-Filterung erreicht, bei denen entweder vor der Anwendung des DTV-Filters ein schwacher modaler Filter eingesetzt wird, oder das mit dem DTV-Filter nachbearbeitete Signal nur in Teilbereichen des Rechengebiets akzeptiert wird. Derartige Teilbereiche werden entweder über den Absolutwert der Differenz benachbarter Knotenwerte in Abhängigkeit einer vorgegebenen Toleranz bestimmt, wie in [81], oder in Abhängigkeit der normalisierten Lokalvarianz, $S_\alpha = |\nabla_\alpha \mathbf{u}^0|_\epsilon / \max_{\gamma \in V} |\nabla_\gamma \mathbf{u}^0|_\epsilon$, für die eine Toleranzschwelle vorzugeben ist, siehe [82].
Die Einschränkung der Wirkung des DTV-Filters ausschließlich auf einen Teilgraphen soll auch in dieser Arbeit untersucht werden. Im Kontext der Nachbearbeitung von Näherungslösungen des DG-Verfahrens mit modaler Filterung sollte eine nochmalige Anwendung des modalen Filters auf dem übrigen Gebiet im Sinne einer hybriden Methode jedoch nicht notwendig sein.
Die hier vorgeschlagene Modifikation des DTV-Filters ist daher wie folgt beschrieben. Mit Hilfe eines geeigneten Indikators wird zunächst eine Teilmenge $\mathcal{T}^{h,DTV} \subset \mathcal{T}^h$ von Dreiecken der Triangulierung markiert. Die Berechnung neuer Iterierten u_α^{k+1} des DTV-Filters wird dann ausschließlich an den Knoten $\alpha \in V$ mit $\tau_{I(\alpha)} \in \mathcal{T}^{h,DTV}$, die im markierten Teilbereich liegen, ausgeführt. Um die Sprünge der Näherungslösung des DG-Verfahrens an Elementgrenzen zwischen den markierten und den nicht markierten Dreiecken der Triangulierung nicht zu vergrößern, wird die Lokalvariation an den betroffenen Knoten $\alpha \in V$, für die ein Knoten $\beta \in N_\alpha$ mit $\tau_{I(\beta)} \notin \mathcal{T}^{h,DTV}$ existiert, wie zuvor in Abhängigkeit von den im Nachbardreieck liegenen Knotenwerten berechnet. Der Knotenwert u_β wird in diesem Fall hingegen festgehalten.
Als DTV-Indikator zur Markierung der Dreiecke, auf denen der DTV-Filter ausgeführt werden soll, wird in dieser Arbeit der bereits zur modalen Filterung verwendete koeffizientenbasierten Indikator ω_i, siehe (5.26), eingesetzt. Ein Dreieck τ_i wird hierbei genau dann markiert, wenn $\omega_i > (N+1)^4$ gilt.

7 Numerische Experimente

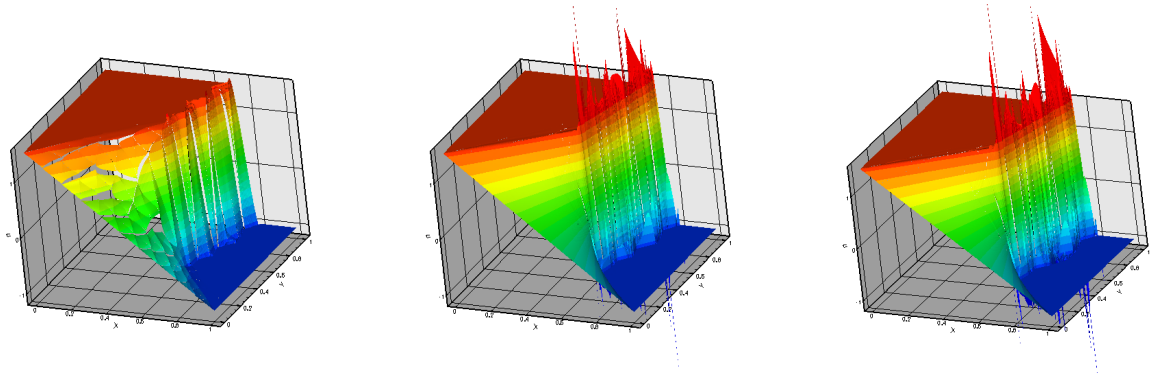
7.1 Skalare Testfälle

Kopplung der Burgers- und Advektionsgleichung Der Einfluss modaler Filterung auf die Stabilität des RKDG-Verfahrens soll zunächst anhand der nichtlinearen skalaren Testgleichung (2.21) untersucht werden, wobei die in Kapitel 5 beschriebenen Stoßindikatoren Verwendung finden. Zur Zeitintegration wird hierbei das Runge-Kutta-Verfahren vierter Ordnung (4.11) verwendet.
Zu erwarten ist, dass die Anwendung eines starken Filters, d.h. eines Filters mit niedriger Filterordnung $2p$ und hoher Filterstärke C_p zu einer stärkeren Reduktion der Oszillationen führt als die Anwendung eines schwächeren Filters. Die Abbildung 7.1, in der die Näherungslösungen zum Zeitpunkt $T = 1.5$ für $N = 5$ und $K = 296$ dargestellt sind, veranschaulicht diese Abhängigkeit von den Filterparametern. Es ist hierbei jeweils nur die Näherungslösung dargestellt, die mit dem Indikator s_{res} berechnet wurde, da sich in diesem Fall optisch keine Unterschiede zur Verwendung des Indikators s_J ergaben. In Abbildung 7.2 sind zudem die Näherungslösungen bei festgehaltenen Filterparametern $p = 2,\ C_p = 1$ für verschiedene Polynomgrade und Dreiecksgitter dargestellt. Die Oszillationen weisen hierbei eine ähnliche Stärke auf, was die in Kapitel 5 hergeleitete Wahl des Filters in Abhängigkeit der Gitterweite und des Polynomgrads unterstützt.
In den Abbildungen 7.3, 7.4 und 7.5 sind die mit den unterschiedlichen vorgestellten Varianten des DTV-Filters nachbearbeiteten Näherungslösungen zum Zeitpunkt $T = 1.5$ für verschiedene Werte der Diskretisierungs- und Filterparameter N, K, p, C_p dargestellt. Der verwendete kartesische DTV-Graph bestand aus 100×100 Punkten, während die Dreiecksgitter-basierten DTV-Graphen mit Hilfe der Zerlegung jedes Dreiecks in 100 Teildreiecke für $K = 296$ bzw. 25 Teildreiecke für $K = 1184$ konstruiert wurden. Der Anpassungsparameter wurde für den kartesischen Graphen mit $\lambda = 2$ und für die Dreiecksgitter-basierten Graphen mit $\lambda = 4$ so gewählt, dass sich im Fall der feinsten Diskretisierung für $N = 10,\ K = 1184$ ähnliche Resultate ergeben. Bei allen Varianten des DTV-Filters wurden jeweils 100 Iteration ausgeführt. Die im Fall der letzteren beiden Parametereinstellungen auftretenden Oszillationen nahe des Stoßes wurden durch den DTV-Indikator erkannt und durch die Anwendung des DTV-Filters geglättet, wie in den Abbildungen 7.4 und 7.5 ersichtlich ist. Die Ergebnisse zur globalen beziehungsweise zur adaptiven DTV-Filterung auf den Dreiecksgitter-basierten Graphen weisen hierbei kaum Unterschiede auf. Im Fall $p = 3$ (Abbildung 7.4) verbleiben jedoch bei Verwendung des DTV-Filters auf dem Dreiecksgitter-basierten Graphen stärkere Restoszillationen in der gefilterten Näherungslösung als bei der Anwendung auf dem kartesischen Graphen. Bei Verringerung des Parameters λ verschwinden auch diese Restoszillationen. Im Fall der nicht oszillierenden Näherungslösung für $N = 5$ und die Filterparameter $p = 1,\ C_p = 5$ (Abbildung 7.3) wird durch die adaptive DTV-Filterung die geringste Änderung des Ausgangssignals vorgenommen. Die Qualität der Ergebnisse der DTV-Filterung ist insgesamt stark abhängig von den gewählten Parametern, zu denen neben dem Anpassungsparameter λ auch die Feinheit des DTV-Graphen zu rechnen ist, sowie die Anzahl der Iterationen der Filtervorschrift, falls der DTV-Filter nicht bis zu Ausbildung des stationären Signals iteriert wird. Da eine aus theoretischen Überlegungen motivierte optimale Parameterwahl nicht greifbar ist, beschränken wir uns bei der Auswahl der Ergebnisse der DTV-Filterung, die für folgenden Testfällen gezeigt werden, auf konkurrenzfähige Resultate bei geschickter Parameterwahl.

In den Tabellen 7.1 und 7.2 sind die Fehlerverläufe der Näherungslösungen des RKDG-Verfahrens mit modaler Filterung sowie deren Nachbearbeitungen für verschiedene Varianten der Filter verzeichnet. Zur Nachbearbeitung wurde der modifizierte DTV-Filter auf Dreiecksgitter-basierten Graphen angewendet, wobei die DTV-Graphen jeweils mit Hilfe der Zerlegung jedes Dreiecks in 25 Teildreiecke konstruiert wurden. Die L^∞-Fehler der Näherungslösungen wurden numerisch berechnet, mit Hilfe der Auswertung der Fehler an den jeweiligen Knoten des DTV-Graphen im Teilgebiet

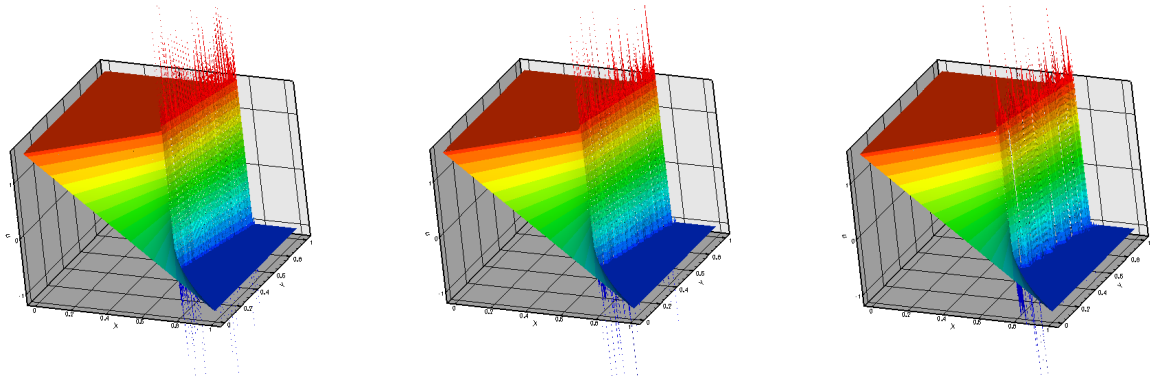
$$\Omega' = \{\mathbf{x} \in [0,1]^2 \,|\, x_1 > 1.5x_2 + 0.1;\ x_1 < 1 - x_2 - 0.1\}.$$

Auf dem Teilgebiet Ω' ist die exakte Lösung glatt aber nicht konstant, demensprechend lässt sich anhand dieser Auswertung die Ordnung des Verfahrens bei Auftreten von Stößen im Bereich fern der Unstetigkeitsstellen untersuchen. Den Resultaten lässt sich ein Ordnungsverlust bei globaler Anwendung des DTV-Filters entnehmen, der bei der adaptiven Verwendung auf Teilgraphen nicht auftritt. Daher sollte grundsätzlich eine adaptive Verwendung des DTV-Filters angestrebt werden.



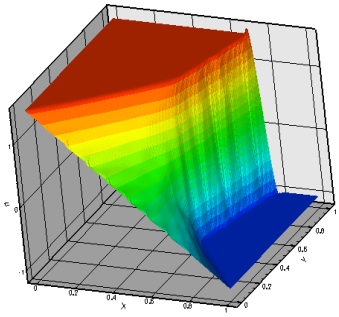
(a) $p = 1,\ C_p = 5$ (b) $p = 2,\ C_p = 1$ (c) $p = 3,\ C_p = 0.5$

Abb. 7.1: Näherungslösungen zur Gleichung (2.21) für $N = 5,\ K = 296$.

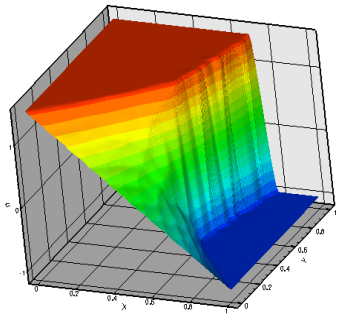


(a) $N = 5,\ K = 18944$ (b) $N = 7,\ K = 4736$ (c) $N = 10,\ K = 1184$

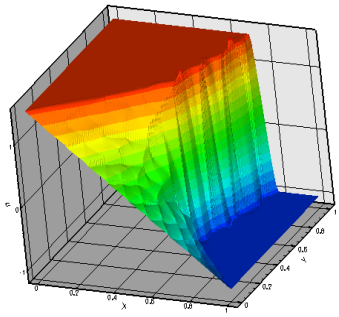
Abb. 7.2: Näherungslösungen zur Gleichung (2.21) für die Filterparameter $p = 2,\ C_p = 1$.



(a) kartesischer DTV-Graph, globale Filterung

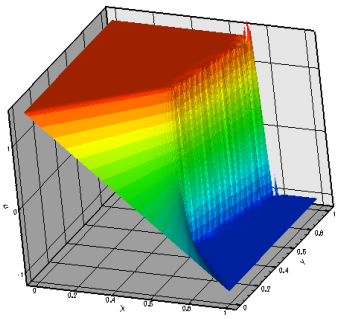


(b) Dreiecksgitter-basierter DTV-Graph, globale Filterung

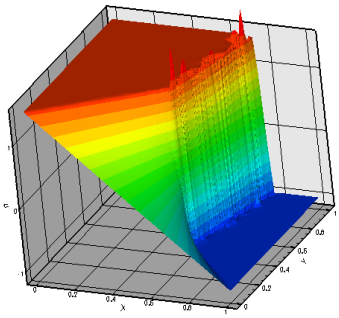


(c) Dreiecksgitter-basierter DTV-Graph, adaptive Filterung

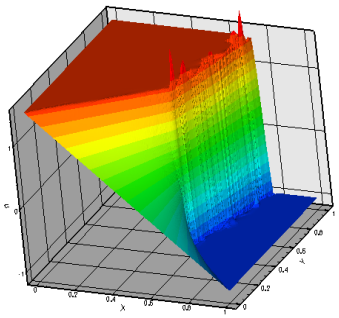
Abb. 7.3: DTV-gefilterte Näherungslösungen für $N = 5,\ K = 296,\ p = 1,\ C_p = 5.$



(a) kartesischer DTV-Graph, globale Filterung

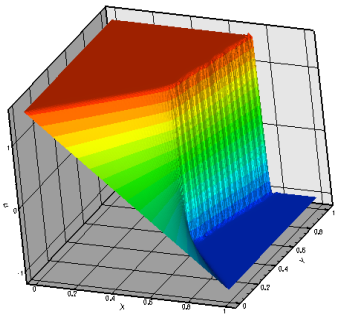


(b) Dreiecksgitter-basierter DTV-Graph, globale Filterung

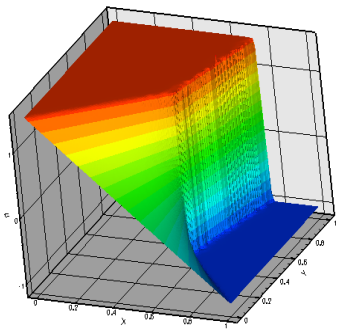


(c) Dreiecksgitter-basierter DTV-Graph, adaptive Filterung

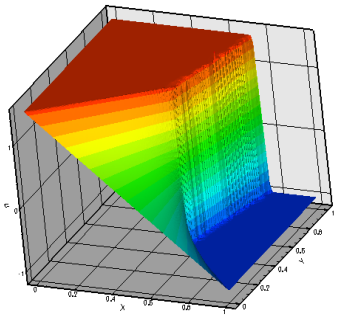
Abb. 7.4: DTV-gefilterte Näherungslösungen für $N = 5,\ K = 296,\ p = 3,\ C_p = 0.5.$



(a) kartesischer DTV-Graph, globale Filterung



(b) Dreiecksgitter-basierter DTV-Graph, globale Filterung



(c) Dreiecksgitter-basierter DTV-Graph, adaptive Filterung

Abb. 7.5: DTV-gefilterte Näherungslösungen für $N = 10,\ K = 1184,\ p = 2, C_p = 1.$

N	K	s_{res}		s_J		L^∞-Fehler nach DTV-Filterung	
		L^∞-Fehler	EOC	L^∞-Fehler	EOC	global	adaptiv
4	296	3.11e-03		5.09e-03		1.25e-02	3.11e-03
	1184	8.58e-05	5.2	2.57e-03	1.0	3.91e-03	8.58e-05
	4736	4.39e-07	7.6	2.64e-03	-0.04	1.38e-03	4.39e-07
	18944	1.42e-08	5.0	1.41e-03	0.9	4.79e-04	1.42e-08
5	296	1.00e-03		2.25e-03		1.21e-02	1.00e-03
	1184	2.10e-05	5.6	1.71e-03	0.4	3.91e-03	2.10e-05
	4736	8.25e-07	4.7	1.45e-03	0.2	1.38e-03	8.25e-07
	18944	4.47e-10	10.8	7.83e-04	0.9	4.79e-04	4.47e-10
6	296	6.68e-04		1.54e-03		1.20e-02	6.68e-04
	1184	3.39e-05	4.3	8.34e-04	0.9	3.90e-03	3.39e-05
	4736	1.26e-07	8.1	5.61e-04	0.6	1.38e-03	1.26e-07
	18944	1.24e-11	13.3	4.20e-04	0.4	4.79e-04	1.24e-11
7	296	3.52e-04		5.25e-04		1.19e-02	3.52e-04
	1184	2.57e-05	3.8	4.76e-04	0.14	3.90e-03	2.57e-05
	4736	4.00e-07	6.0	4.03e-04	0.24	1.38e-03	4.00e-07
	18944	2.12e-11	14.2	1.70e-04	1.25	4.79e-04	2.12e-11

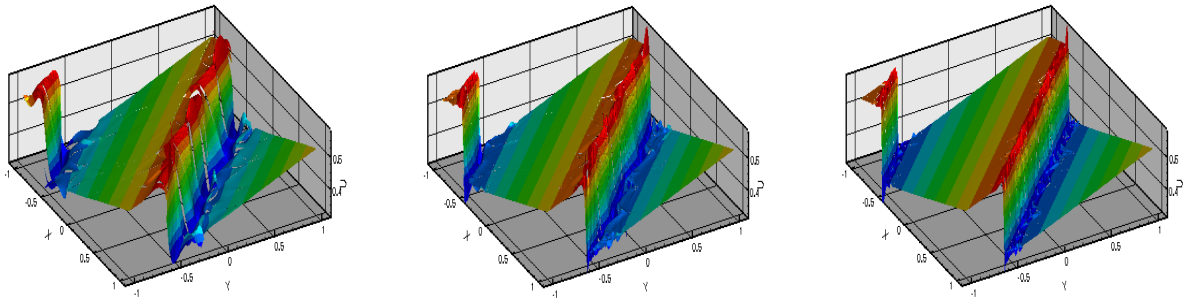
Tabelle 7.1: Fehlerverlauf für die verschiedenen Varianten der Filter am Beispiel der Modellgleichung (2.21). Modale Filterung mit den Parametern $p = 2$, $C_p = 1$. Anwendung des DTV-Filters auf die Näherungslösungen zum Indikator s_{res}.

N	K	s_{res}		s_J		L^∞-Fehler nach DTV-Filterung	
		L^∞-Fehler	EOC	L^∞-Fehler	EOC	global	adaptiv
4	296	4.19e-03		4.18e-03		1.13e-02	4.18e-03
	1184	6.14e-05	6.1	6.23e-05	6.1	3.89e-03	6.23e-05
	4736	4.19e-07	7.2	3.31e-06	4.2	1.38e-03	3.31e-06
	18944	1.42e-08	4.9	1.42e-08	7.9	4.79e-04	1.42e-08
5	296	1.58e-03		1.61e-03		1.17e-02	1.61e-03
	1184	4.73e-05	5.1	4.09e-05	5.3	3.91e-03	4.09e-05
	4736	1.00e-06	5.6	1.57e-06	4.7	1.38e-03	1.57e-06
	18944	4.47e-10	11.1	1.26e-06	0.3	4.79e-04	1.26e-06
6	296	1.54e-03		1.54e-03		1.17e-02	1.54e-03
	1184	2.58e-05	5.9	2.16e-05	6.2	3.90e-03	2.16e-05
	4736	1.79e-07	7.2	1.29e-07	7.4	1.38e-03	1.29e-07
	18944	1.24e-11	13.8	1.24e-11	13.3	4.79e-04	1.24e-11
7	296	7.70e-04		7.67e-04		1.17e-02	7.67e-04
	1184	4.19e-05	4.2	2.71e-05	4.8	3.90e-03	2.71e-05
	4736	5.57e-07	6.2	3.94e-07	6.1	1.38e-03	3.94e-07
	18944	2.34e-11	14.5	1.57e-11	14.6	4.79e-04	1.57e-11

Tabelle 7.2: Fehlerverlauf für die verschiedenen Varianten der Filter am Beispiel der Modellgleichung (2.21). Modale Filterung mit den Parametern $p = 3$, $C_p = 0.5$. Anwendung des DTV-Filters auf die Näherungslösungen zum Indikator s_J.

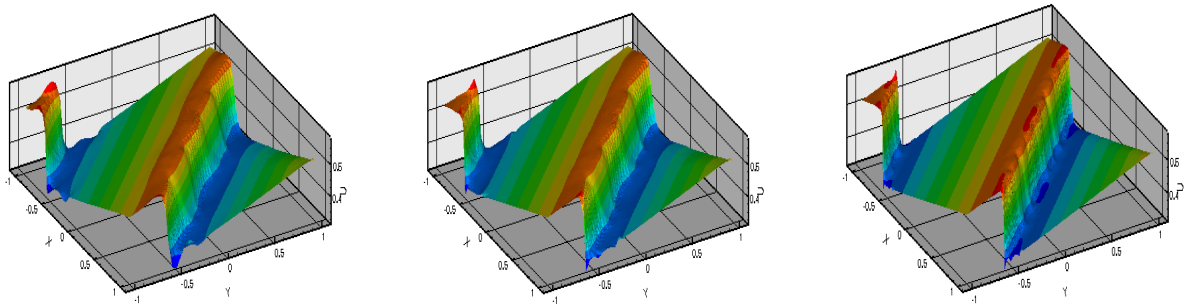
In Bezug auf die Verwendung der Indikatoren s_{res} und s_J zur adaptiven modalen Filterung wird deutlich, dass die Näherungslösungen zu s_J für die Wahl der Filterparameter $p = 2,\ C_p = 1$ deutlich schlechtere Fehlerverläufe besitzen, als die Näherungslösungen zu s_{res}. Für $p = 3,\ C_p = 0.5$ fallen die Fehler der Näherungslösungen zum Indikator s_J hingegen deutlich schneller ab, als bei der vorherigen Parameterwahl. Die Verwendung des Indikators s_{res} führt für beide Varianten der Filterparameter zu sehr schnell fallenden Fehlern.

Burgersgleichung in 2D Die Anwendung des modalen Filters auf die in (2.23) gegebene Burgers-Gleichung in zwei Raumdimensionen führt zu einer deutlichen Verringerung der Oszillationen der Näherungslösung. Hierzu vergleiche man die in Abbildung 7.6 dargestellten Näherungslösungen des DG-Verfahrens mit modaler Filterung (für die Parameter $p = 2,\ C_p = 2$ und den Indikator s_{res}) zum Zeitpunkt $T = 1.4$ mit den Näherungslösungen des DG-Verfahrens ohne Filterung in Abbildung 4.6. In Abbildung 7.7 sind die entsprechenden nachbearbeiteten Lösungen dargestellt, wobei in diesem Fall DTV-Graphen, die durch Zerlegen jedes Dreiecks in 81 Teildreiecke entstehen, verwendet und 100 Iterationen mit dem modifizierten DTV-Filter ($\lambda = 1$) ausgeführt wurden.



(a) $N = 4,\ K = 68$ (b) $N = 9,\ K = 68$ (c) $N = 9,\ K = 272$

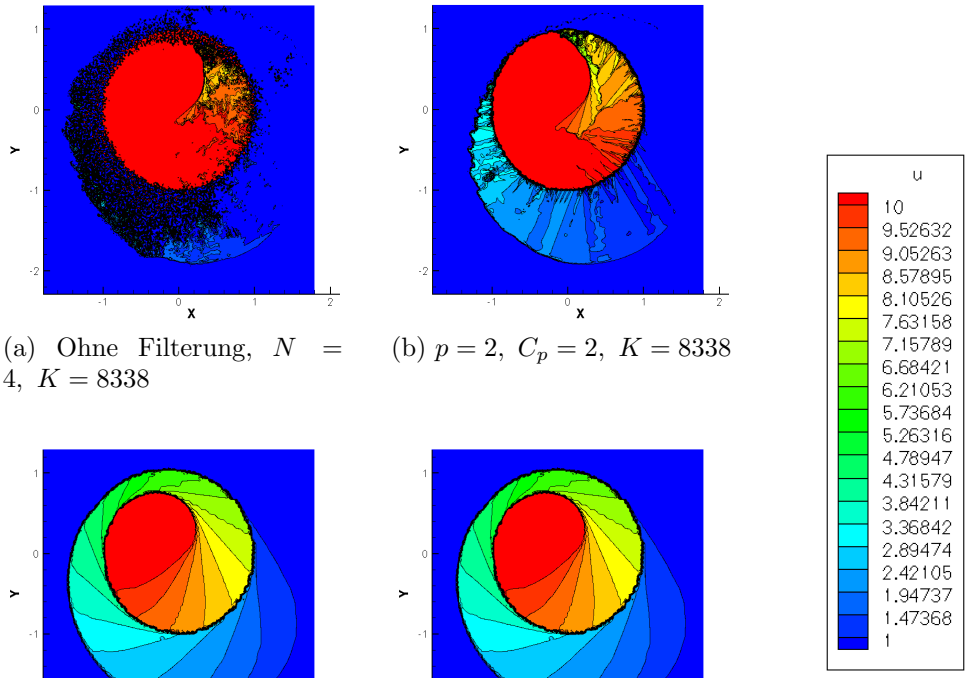
Abb. 7.6: Näherungslösungen des RKDG-Verfahrens mit modaler Filterung zur Gleichung (2.23) zum Zeitpunkt $T = 1.4$.



(a) $N = 4,\ K = 68$ (b) $N = 9,\ K = 68$ (c) $N = 9,\ K = 272$

Abb. 7.7: Nachbearbeitete Näherungen zur Gleichung (2.23) zum Zeitpunkt $T = 1.4$.

Nichtkonvexer Testfall Im Fall der nichtkonvexen Testgleichung (2.24) beobachten wir zunächst, dass das DG-Verfahren ohne Filterung nicht unbedingt die korrekte Entropielösung liefert, wie ein Vergleich der Näherungslösung für $N = 4$ (Abbildung 7.8a) auf einem Gitter mit 8338 Dreiecken mit den in [59] angegebenen Resultaten zeigt. Eine modale Filterung mit den Parametern $p = 2$, $C_p = 2$ und dem Stoßindikator s_{res} liefert die für dieses Testbeispiel noch unbefriedigende Näherungslösung nach Abbildung 7.8b, während die Wahl der Parameter $p = 1$, $C_p = 5$ (Abbildung 7.8c) ein mit der Näherungslösung des in [59] vorgeschlagenen Verfahrens vergleichbares Ergebnis liefert. Die Abbildung 7.8d zeigt die Näherungslösungen ebenfalls zum Polynomgrad $N = 4$ auf der ersten Rot-Verfeinerung des voherigen Gitters, welche $K = 23130$ Elemente enthält, so dass die letztere Parameterwahl auch unter Gitterverfeinerung gute Ergebnisse liefert. In Abbildung 7.9 sind für die Näherungslösung zu $p = 1$, $C_p = 5$ (Abbildung 7.8c) die Resultate der Nachbearbeitung durch verschiedene Varianten des DTV-Filters dargestellt. Die Abbildung 7.9a zeigt das Ergebnis der DTV-Filterung auf einem 200×200 Punkte enthaltenden kartesischen Graphen, während in der Abbildung 7.9b das Ergebnis der modifizierten DTV-Filterung auf einem durch Zerlegung jedes Dreiecks in 25 Teildreiecke entstandenen Graphen dargestellt ist. Es wurden jeweils 100 Iterationen mit $\lambda = 2$ ausgeführt. Die globale Anwendung des DTV-Filters in seiner ursprünglichen Form unter Verwendung eines kartesischen Graphen sowie die des modifizierten DTV-Filters auf einem Dreiecksgitter-basierten Graphen führt für dieses Beispiel zu fast identischen Ergebnissen. Im Fall der adaptiven Anwendung des DTV-Filters auf dem Dreiecksgitterbasierten Graphen (Abbildung 7.9c) verbleiben hingegen schwache Oszillationen fern der spiralförmigen Unstetigkeitskurve in der numerischen Lösung. Eine genauere Betrachtung des Indikators zur Markierung des nachzubearbeitenden Teilbereiches zeigt hierbei, dass in diesem Fall nur Dreiecke in unmittelbarer Nähe der Unstetigkeitskurve markiert wurden und der DTV-Filter somit fern der Sprungstellen nicht angewendet wurde.



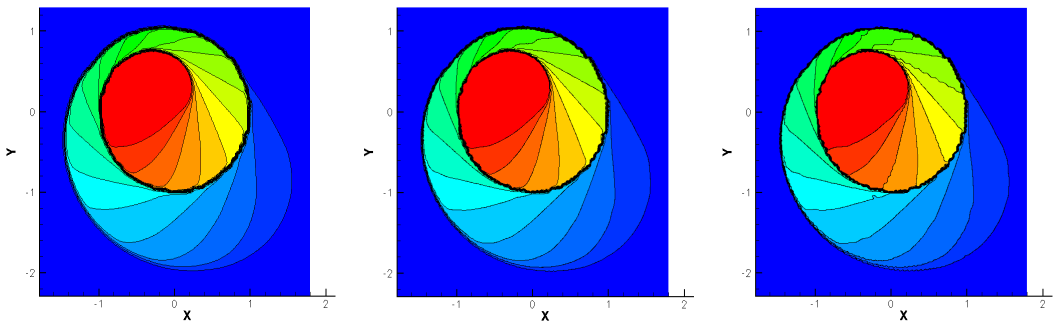
(a) Ohne Filterung, $N = 4$, $K = 8338$

(b) $p = 2$, $C_p = 2$, $K = 8338$

(c) $p = 1$, $C_p = 5$, $K = 8338$

(d) $p = 1$, $C_p = 5$, $K = 23130$

Abb. 7.8: Näherungslösungen des RKDG-Verfahrens mit modaler Filterung zur Gleichung (2.24) zum Zeitpunkt $T = 1$, Darstellung mit 20 Isolinen zu äquidistanten Werten von 1 bis 10.



(a) Kartesischer DTV-Graph

(b) Dreiecksgitter-basierter DTV-Graph, globale Filterung

(c) Dreiecksgitter-basierter DTV-Graph, adaptive Filterung

Abb. 7.9: DTV-gefilterte Näherungslösung für $p = 1$, $C_p = 5$.

7.2 Testfälle auf Grundlage der Euler-Gleichungen

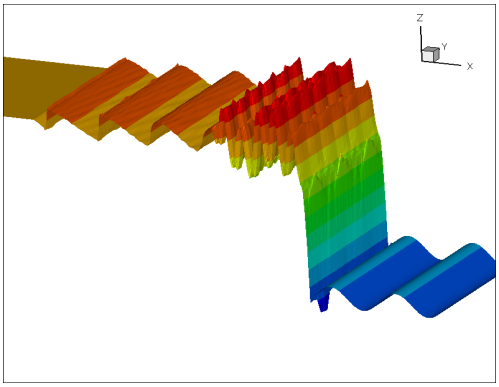
Nachfolgend werden Näherungslösungen des RKDG-Verfahrens mit modaler Filterung sowie deren Nachbearbeitungen für verschiedene klassische Testfälle auf Grundlage der Euler-Gleichungen der Gasdynamik berechnet. Als Zeitintegration wurde ausschließlich das Runge-Kutta-Verfahren vierter Ordnung (4.11) verwendet. Die Stoßindikatoren zur adaptiven Steuerung der spektralen Viskosität des DG-Verfahrens basieren zudem in allen Berechnungen auf der Komponente der Dichte. Bis auf den Fall der Interaktion eines Stoßes mit einem sich bewegenden Wirbel war es bei diesen Testfällen notwendig, negative Werte von Dichte oder Druck an den auf Dreieckskanten liegenden Stützstellen zu verhindern, unter anderem aus dem Grund, dass numerische Flussfunktionen wie das verwendete Flussvektor-Splitting-Verfahren nach van Leer, siehe Abschnitt A.2, die Berechnung der Schallgeschwindigkeit $c = \sqrt{\frac{\gamma p}{\rho}}$ erfordern. Wenn notwendig, wird bei Berechnung der Näherungslösungen daher ein schwacher modaler Filter zusätzlich iterativ angewendet, bis Druck und Dichte an allen Stützstellen auf Dreieckskanten positiv sind. Einige Ergebnisse zu den Testfällen der Interaktion eines Stoßes mit einer Dichtewelle und der Stoß-Wirbel-Interaktion wurden bereits in [70] und [71] gezeigt.

Interaktion eines Stoßes mit einer Dichtewelle

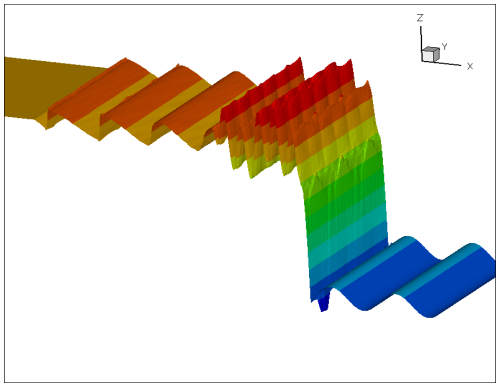
Wir betrachten zunächst den von Shu und Osher [84] verwendeten Testfall der Interaktion eines Stoßes mit einer Dichtewelle. Während dieser Testfall ursprünglich als eindimensionales Riemann-Problem definert wurde, wird es hier zu einem Testfall in zwei Raumdimensionen auf dem Gebiet $\Omega = [-5, 5] \times [0, 0.5]$ erweitert. Die Anfangsbedingungen sind dementsprechend gegeben durch

$$(\rho, v_1, v_2, p) = \begin{cases} (3.857143, 2.629369, 0, 10.333333) & \text{falls } x < -4, \\ (1 + 0.2 \cdot \sin(5x), 0, 0, 1) & \text{falls } x \geq -4. \end{cases}$$

Die Abbildung 7.10 zeigt die durch das DG-Verfahren mit modaler Filterung ($p = 2$, $C_p = 2$) unter Verwendung der Stoßindikatoren s_{res} und s_J berechnete Dichteverteilung für den Polynomgrad $N = 5$ zum Ausgabezeitpunkt $T = 1.8$. Das verwendete unstrukturierte Rechengitter mit 1250 Dreiecken ist in Abbildung 7.12 dargestellt. Die Verwendung des koeffizientenbasierten Indikators s_{res} führt hierbei zur Ausbildung kleinerer Wellen auch im glatten Bereich der Lösung, die bei der Verwendung von s_J nur in schwacher Form auftreten. Die mit dem DTV-Filter auf kartesischen Graphen nachbearbeiteten Lösungen sind in Abbildung 7.11 dargestellt. Hierbei wurden 100 Iterationen mit $\lambda = 10$ auf 500×25 kartesischen Gitterpunkten durchgeführt.
Die Abbildung 7.13 zeigt zudem eindimensionale Schnitte durch $y = 0.33$ für die Polynomgrade $N = 2, 3, 6$ und den Indikator s_{res} sowohl vor als auch nach Anwendung des DTV-Filters. Zur Erzeugung der Referenzlösung wurde der eindimensionale Shu-Osher-Testfall mit Hilfe eines Finite-Volumen-Verfahrens zweiter Ordnung mit TVD-Rekonstruktion auf einem Rechengitter mit 30000 Zellen approximiert.
Durch die zusätzliche iterative Anwendung des modalen Filters, die ausschließlich dazu dient, positive Werte der physikalischen Größen Dichte und Druck an den Zellkanten zu erzwingen, ist die innerhalb des Evolutionsprozesses eingeführte künstliche Diffusion verhältnismäßig hoch. Daher produziert das DG-Verfahren mit modaler Filterung nur geringe Überschwinger, die durch die Nachbearbeitung mit dem global angewendeten DTV-Filter auf dem gegebenen kartesischen Graphen vollständig entfernt werden. Die globale

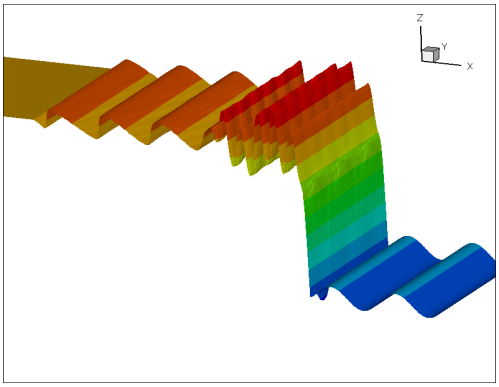


(a) Indikator s_{res}

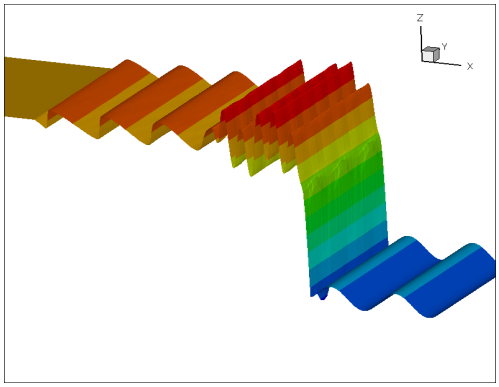


(b) Indikator s_J

Abb. 7.10: Lösung des DG-Verfahrens mit adaptiver modaler Filterung, $N = 5,\ T = 1.8$.



(a) Indikator s_{res}



(b) Indikator s_J

Abb. 7.11: Mit dem DTV-Filter nachbearbeitete Lösung.

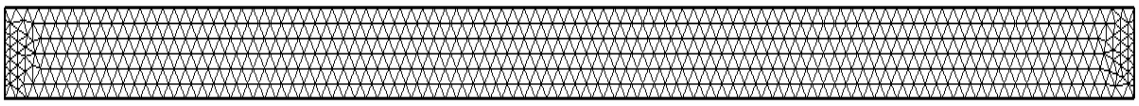
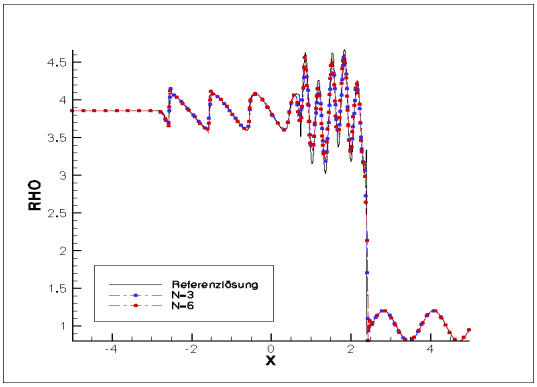
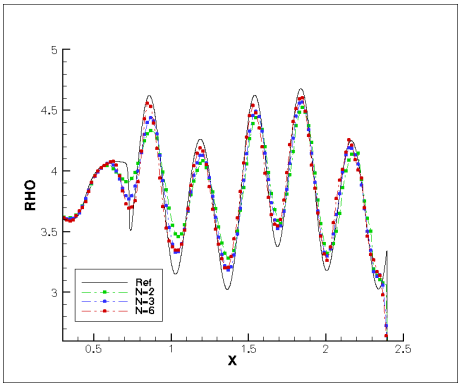


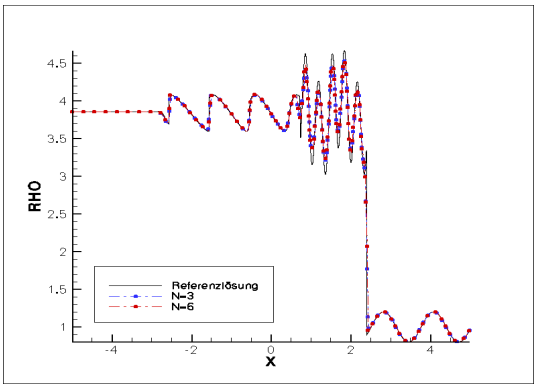
Abb. 7.12: Rechengitter für den Shu-Osher-Testfall.



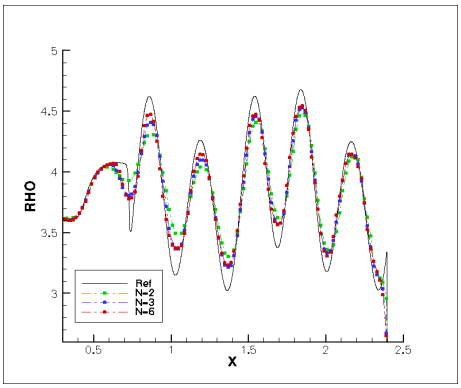
(a) Vor der DTV-Filterung, $N = 3, 6$



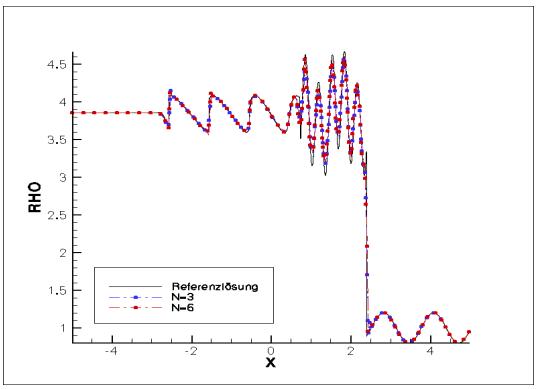
(b) Nahaufnahme, $N = 2, 3, 6$



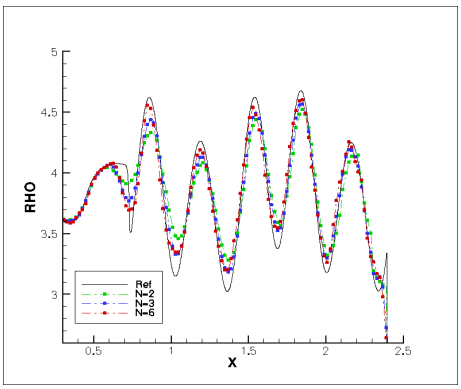
(c) Nach globaler DTV-Filterung, $N = 3, 6$



(d) Nahaufnahme, $N = 2, 3, 6$



(e) Nach adaptiver DTV-Filterung, $N = 3, 6$



(f) Nahaufnahme, $N = 2, 3, 6$

Abb. 7.13: Lösungen zur Stoß-Dichtewelle-Interaktion: Zweidimensionale Schnitte durch $y = 0.33$.

Anwendung des DTV-Filters führt allerdings auch zu einer Verringerung der Amplituden im glatten Bereich hochfrequenter Dichtefluktuationen vor dem Stoß nahe $x = 2.5$. Mit der adaptiven Anwendung des DTV-Filters lässt sich dieser Genauigkeitsverlust vermeiden, allerdings werden die leichten Überschwinger in der Nähe von $x = -2.5$ nicht beseitigt. Eine genauerer Untersuchung des Indikators für diesen Testfall zeigt, dass für $N = 5, 6, 7$ nur Dreiecke in direkter Nähe des Sprungs bei $x = 2.5$ markiert werden, so dass die numerische Lösung im restlichen Bereich unverändert bleibt. Für $N = 2, 3$ werden durch den gewählten Indikator zudem gar keine Dreiecke markiert.

Stoß-Wirbel-Interaktion

Dieser Testfall ist entnommen aus [49] und beschreibt die Interaktion eines stationären Stoßes mit einem sich bewegenden isentropen Wirbel. Das Rechengebiet ist festgelegt durch $\Omega = [0, 2] \times [0, 1]$ und die Konstruktion der Anfangsbedingungen wird wie folgt vorgenommen. Ein stationärer Stoß wird senkrecht zur x-Achse an der Stelle $x = 0.5$ positioniert. Der linke Zustand ist zunächst in primitiven Variablen durch $\mathbf{u}^-_{prim} = (\rho, v_1, v_2, p) = (1, 1.1\sqrt{\gamma}, 0, 1)$ gegeben, während sich der rechte Zustand $\mathbf{u}^+_{prim} = (\hat{\rho}, \hat{v}_1, \hat{v}_2, \hat{p})$ aus den Rankine-Hugoniot-Bedingungen berechnet. Man erhält

$$\hat{v}_1 = \frac{\gamma b - \sqrt{\gamma^2 b^2 - 2a(\gamma^2 - 1)d}}{a(\gamma + 1)},$$

mit $a = v = 1.1\sqrt{\gamma}$, $b = 1.21\gamma + 1$ und $d = \left(0.6655 + \frac{1.1}{\gamma - 1}\right)\gamma^{3/2}$ sowie

$$\begin{aligned} \hat{v}_2 &= 0, \\ \hat{\rho} &= \frac{a}{\hat{v}_1}, \\ \hat{p} &= b - a\hat{v}_1. \end{aligned}$$

Die linksseitig des Stoßes gegebene Grundströmung wird desweiteren durch eine Störung $(\delta\rho, \delta v_1, \delta v_2, \delta p)$ überlagert, d.h. man setzt $\mathbf{u}^-_{prim} = (1 + \delta\rho, 1.1\sqrt{\gamma} + \delta v_1, \delta v_2, 1 + \delta p)$. Die Störung, in Form eines in $(x_c, y_c) = (0.25, 0.5)$ zentrierten isentropen Wirbels, ist beschrieben durch die Differenzen der Geschwindigkeit, der Entropie $S = \frac{p}{\rho^\gamma}$ und der Temperatur, die bei Verwendung der Gaskonstanten $R = 1$ in der Gleichung für ideale Gase (2.26) durch $T = \frac{p}{\rho}$ definiert ist. Diese Differenzen sind gegeben durch

$$\begin{aligned} \delta v_1 &= \epsilon r_c^{-1} e^{\alpha(1-(r/r_c)^2)}(y - y_c), \\ \delta v_2 &= -\epsilon r_c^{-1} e^{\alpha(1-(r/r_c)^2)}(x - x_c), \\ \delta S &= 0, \\ \delta T &= -(\gamma - 1)\epsilon^2 e^{2\alpha(1-(r/r_c)^2)}/(4\alpha\gamma), \end{aligned}$$

mit $r = \sqrt{(x - x_c)^2 + (y - y_c)^2}$, sowie den Parametern $r_c = 0.05, \epsilon = 0.3$ und $\alpha = 0.204$. Die Störungen der Dichte und des Drucks ergeben sich aus den obigen Gleichungen für δS und δT zu

$$\begin{aligned} \delta\rho &= (1 + \delta T)^{\frac{1}{1-\gamma}} - 1, \\ \delta p &= (1 + \delta T)^{\frac{\gamma}{1-\gamma}} - 1. \end{aligned}$$

Die Randbedingungen für diesen Testfall sind durch zwei reflektierende Wände am oberen und unteren Rand des Rechengebiets, sowie durch Einströmbedingungen am linken und Ausströmbedingungen am unteren Rand gegeben.
Die Abbildung 7.15 zeigt die Druckverteilungen zu den Ausgabezeiten $T = 0.05, 0.2, 0.35$, die mit dem DG-Verfahren mit modaler Filterung unter Verwendung der Filterparameter $p = 2, C_p = 2$ und des Stoßindikators s_J für die Polynomgrade $N = 3, 5, 7$ berechnet wurden. In Abbildung 7.16 sind die entsprechenden Näherungen zu den späteren Zeitpunkten $T = 0.6, 0.8$ dargestellt. (Wie beim vorherigen Testfall weist die Verwendung des Indikators s_{res} leichte Wellen im glatten Bereich der Lösung auf, diese sind in der zweidimensionalen Darstellung allerdings nicht zu erkennen.) Das zugehörige Rechengitter besteht aus 2122 Dreiecken, wobei eine höhere Auflösung nahe der Stoßposition verwendet wurde, vergleiche Abbildung 7.14.

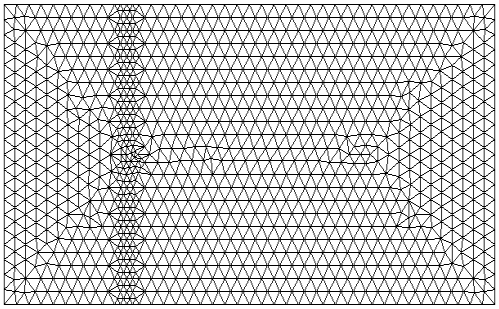


Abb. 7.14: Rechengitter für den Testfall der Stoß-Wirbel Interaktion.

Für höhere Polynomgrade $N = 5, 7$ sind leichte Verbesserungen bezogen auf die Auflösung des Stoßes und des Wirbels zu erkennen, insbesondere zu den späteren Zeitpunkten $T = 0.6, 0.8$. In allen durchgeführten Experimenten erwies sich die modale Filterung als ausreichend zur Kontrolle der Oszillationen. In Abbildung 7.17 sind für $N = 3, 7$ die Ergebnisse der Nachbearbeitung durch den DTV-Filter auf einem kartesischen Graphen mit 400×200 Knoten (100 Iterationen, $\lambda = 10$) dargestellt. Zudem zeigt die Abbildung 7.18 die Näherungslösungen in dreidimensionaler Darstellung vor und nach der Nachbearbeitungsprozedur für $N = 5$ und $T = 0.35$.
In Abbildung 7.19 sind desweiteren die elementweise konstanten Werte der Indikatoren s_{res} und s_J für $N = 5$ zu den Zeitpunkten $T = 0.2, 0.35, 0.6$ dargestellt. Durch die hier gewählte exponentielle Skalierung der Werte fällt auf, dass die Verwendung des Indikators s_J zur Anwendung des modalen Filters in einem größeren Bereich des Rechengebiets führt, wenn auch mit geringerer Filterstärke fern von Stößen. Im Fall des Indikators s_{res} ist die modale Filterung hingegen stärker auf die Bereiche in unmittelbarer Nähe von Stößen eingeschränkt.

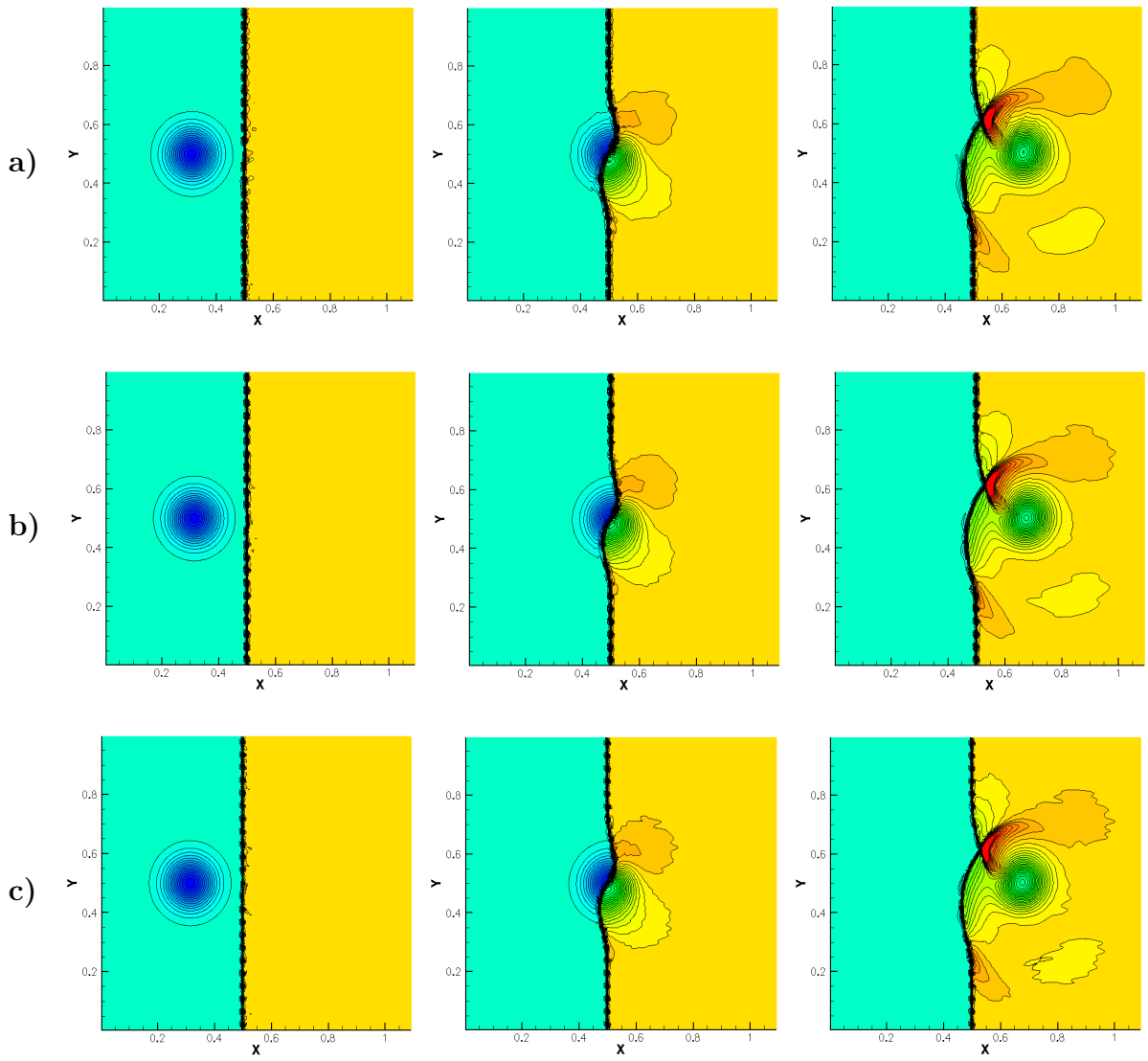


Abb. 7.15: Lösungen des RKDG-Verfahrens mit adaptiver modaler Filterung zu den Ausgabezeiten $T = 0.05, 0.2, 0.35$: Druckverteilung, 46 Isolinen zu äquidistanten Werten von 0.85 bis 1.35 für a) $N = 3$, b) $N = 5$, c) $N = 7$.

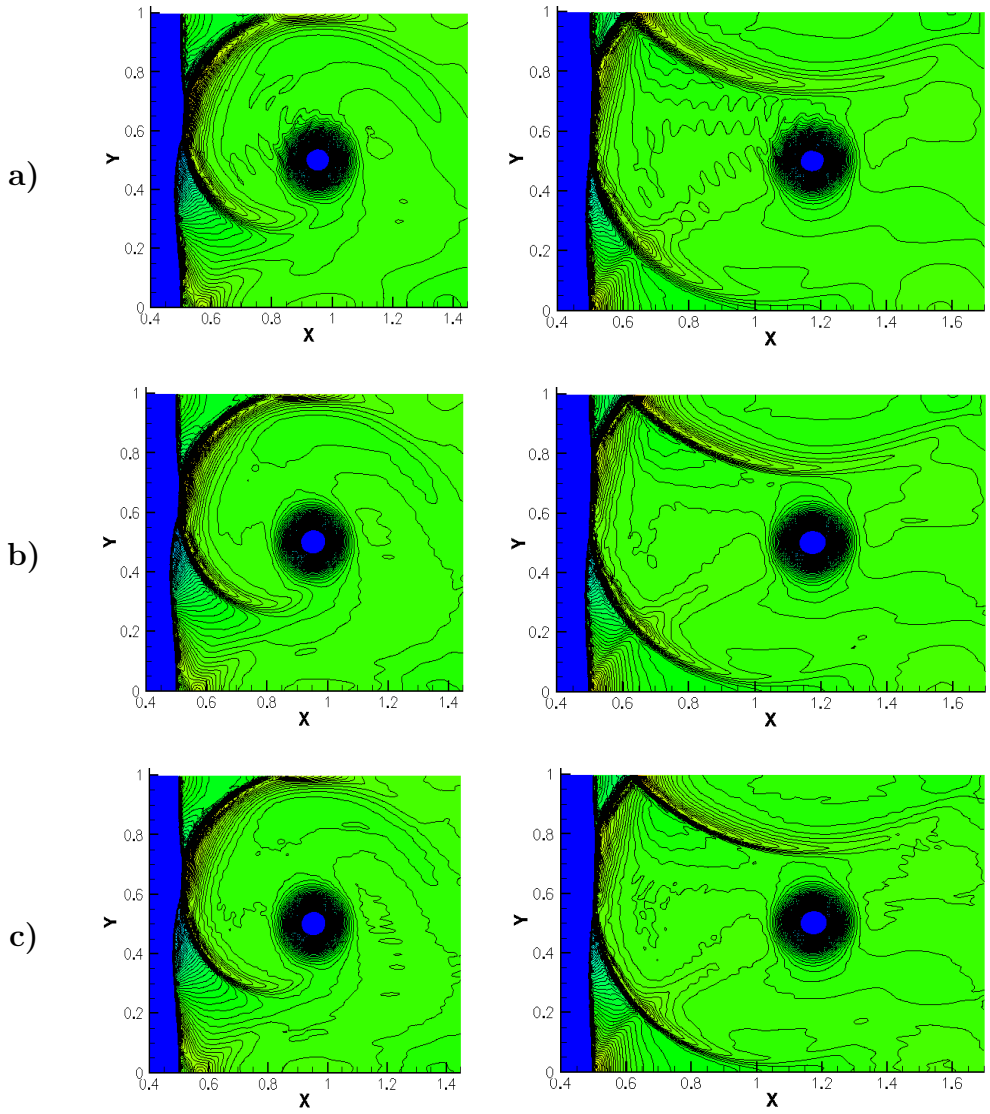


Abb. 7.16: Lösungen des DG-Verfahrens mit modaler Filterung zu den Ausgabezeiten $T = 0.6, 0.8$: Druckverteilung, 90 Isolinen zu äquidistanten Werten von 1.09 bis 1.37 für a) $N = 3$, b) $N = 5$, c) $N = 7$.

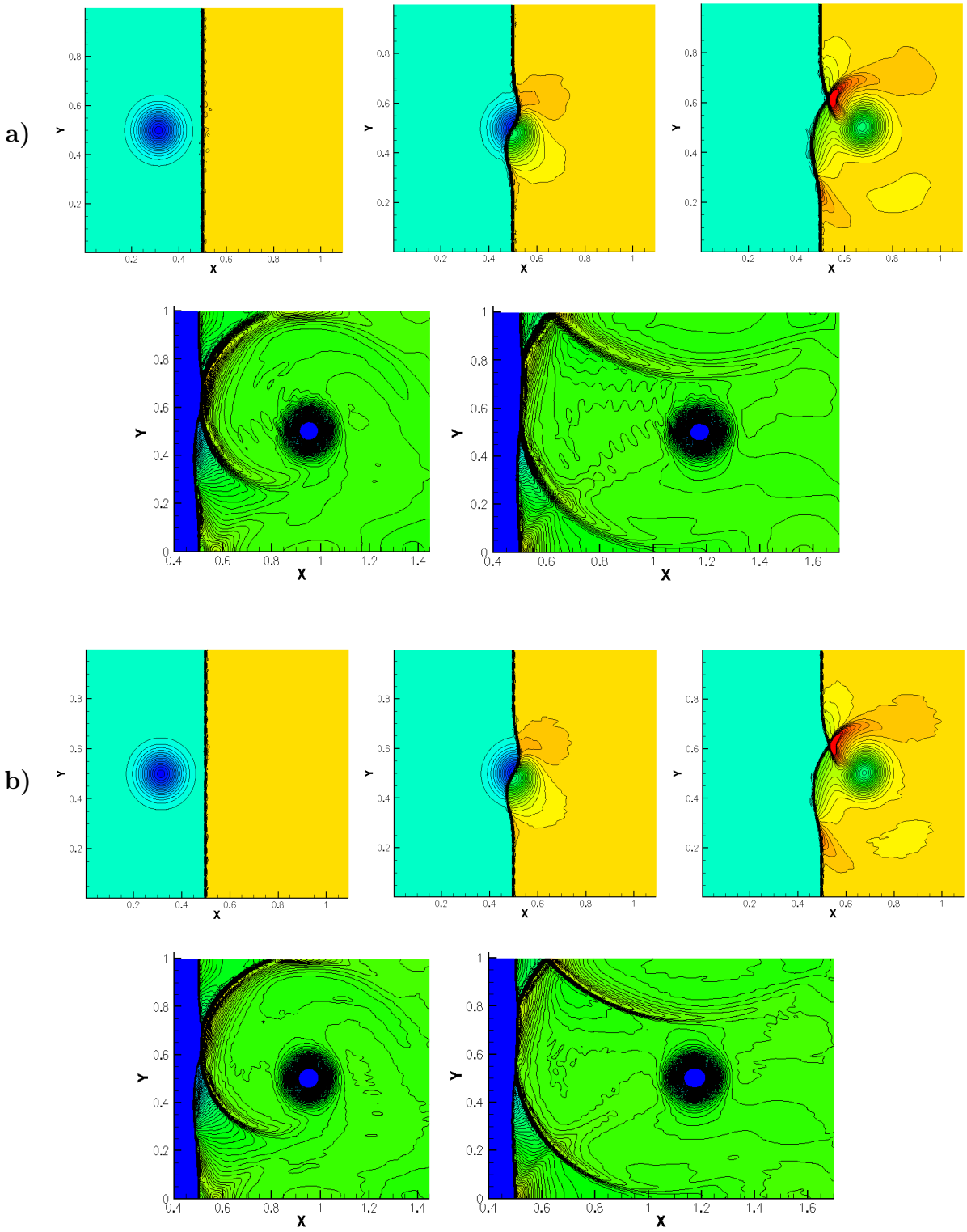
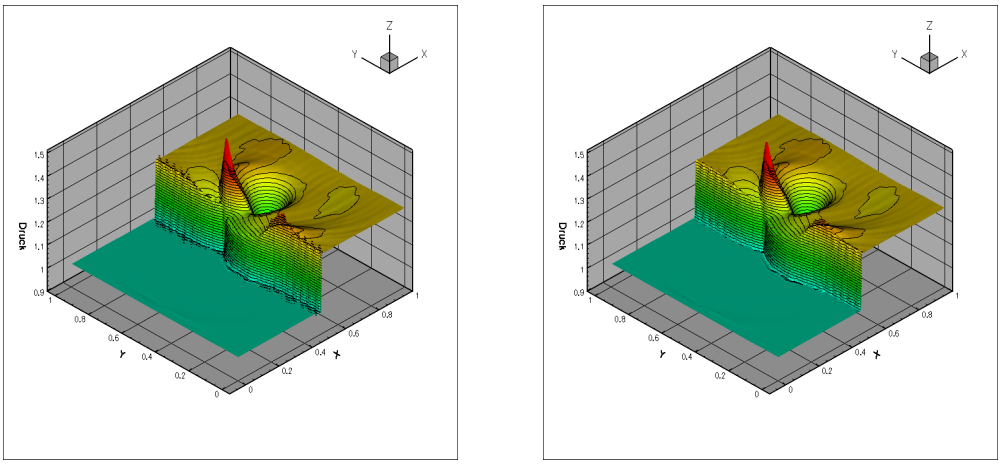


Abb. 7.17: Nachbearbeitete Lösungen für a) $N = 3$ und b) $N = 7$. Druckverteilung, 46 Isolinen zu äquidistanten Werten von 0.85 bis 1.35 für $T = 0.05, 0.2, 0.35$; 90 Isolinen zu äquidistanten Werten von 1.09 bis 1.37 für $T = 0.6, 0.8$.



(a) Vor Anwendung des DTV-Filters (b) Nach Anwendung des DTV-Filters

Abb. 7.18: Auswirkung der Nachbearbeitungsprozedur für $N = 5,\ T = 0.35$.

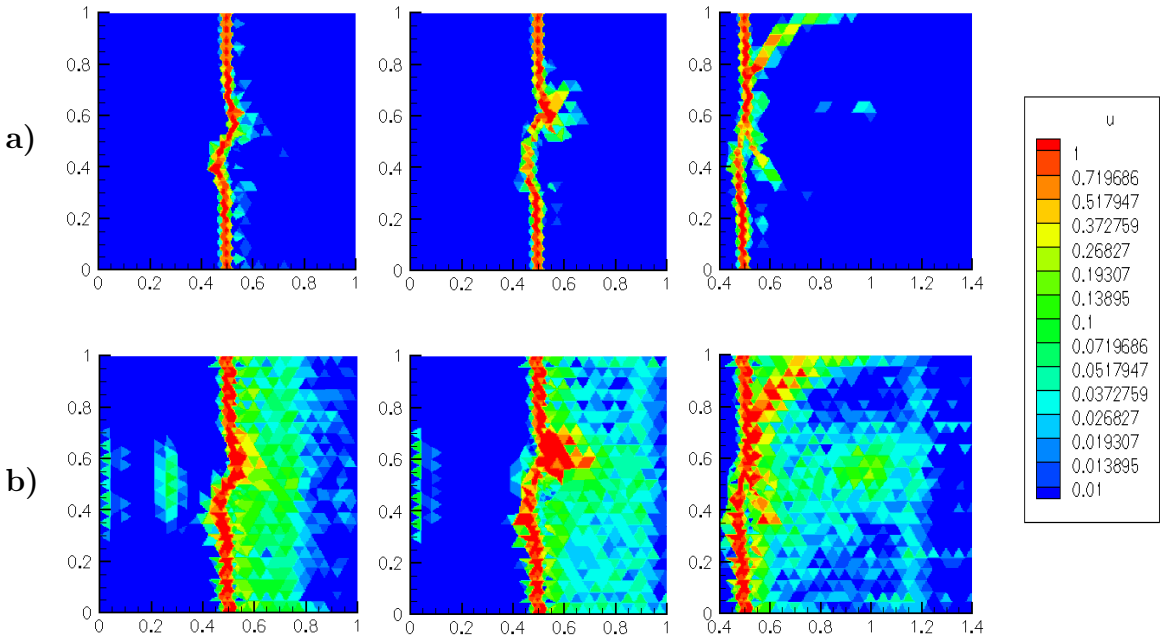


Abb. 7.19: Werte der Indikatoren a) s_{res} und b) s_J zu den Zeitpunkten $T = 0.2, 0.35, 0.6$ für $N = 5$, exponentielle Skala.

Doppel-Mach-Reflektion (Double Mach Reflection)

Dieser aus [99] entnommene Testfall beschreibt das Auftreffen eines sich horizontal mit Stoßgeschwindigkeit 10 bewegenden Stoßes auf einen Keil, durch das ein komplexes Strömungsverhalten hervorgerufen wird. Zur leichteren Darstellung des reflektierenden Keils wird der Testfall üblicherweise so implementiert, dass der Stoß schräg verläuft, während der Keil eine reflektierende feste Wand am unteren Rand des Rechengebiets bildet. In der Abbildung 7.20 ist das hier verwendete Rechengebiet $\Omega = [0, 3.5] \times [0, 1]$ einschließlich der zugehörigen Randbedingungen sowie der initialen Stoßlage dargestellt. Die Anfangsbedingungen sind gegeben durch die Strömungsdaten links und rechts des eingangs vorgegebenen Stoßes,

$$\mathbf{u}_{prim}(\mathbf{x}, 0) = \begin{cases} (\rho^-, \mathbf{v}^-, p^-) = (8\,, 8.25 \cdot \cos 30°,\ -8.25 \cdot \sin 30°\,, 116.5), & \text{für } x < \frac{1}{6} + \cdot\frac{y}{\sqrt{3}}, \\ (\rho^+, \mathbf{v}^+, p^+) = (1.4, 0, 0, 1), & \text{andernfalls,} \end{cases}$$

während die Randbedingung am oberen Einströmränd des Rechengebiets der Bewegung des Stoßes folgt, d.h. für alle $t > 0$ und alle Punkte auf dem Einströmrand $\mathbf{x} \in \partial\Omega_{in}$ setzen wir

$$\mathbf{u}_{prim}(\mathbf{x}, t) = \begin{cases} (\rho^-, \mathbf{v}^-, p^-), & \text{für } x < \frac{1}{6} + y \cdot \frac{1+20t}{\sqrt{3}}, \\ (\rho^+, \mathbf{v}^+, p^+), & \text{andernfalls.} \end{cases}$$

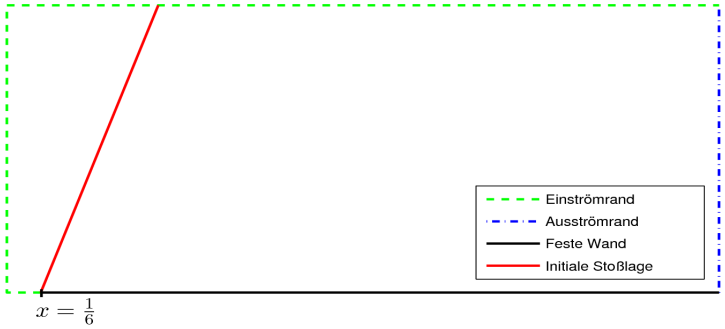


Abb. 7.20: Rechengebiet, Randbedingungen und anfängliche Stoßlage für die Doppel-Mach-Reflektion.

In Abbildung 7.21 sind die Näherungslösungen des DG-Verfahrens mit modaler Filterung zum Zeitpunkt $T = 0.2$ für $N = 2, 4$ auf einem Dreiecksgitter mit $K = 7338$ Elementen dargestellt. Die Parameter des modalen Filters wurden in diesem Beispiel mit $p = 2,\ C_p = 2$ belegt und es wurden beide Stoßindikatoren $s_{res},\ s_J$ zur Steuerung der Adaptivität verwendet, wobei aufgrund kaum erkennbarer optischer Unterschiede nur die Näherungslösungen für den Indikator s_J dargestellt sind. Die Abbildungen 7.22 und 7.23 zeigen die mit verschiedenen Varianten des modifizierten DTV-Filters auf Dreiecksgitter-basierten Graphen nachbearbeiteten Lösungen. Zur Konstruktion der DTV-Graphen wurde hierbei jedes Dreieck in 9 Teildreiecke zerlegt. Jeweils wurden 100 Iterationen des Filters für $\lambda = 1$ ausgeführt. Die Abbildung 7.22 zeigt die Resultate der adaptiven Anwendung des DTV-Filters, während in Abbildung 7.23 globale DTV-Filterung eingesetzt wurde. Die Abbildungen 7.24, 7.25 und 7.25 zeigen die entsprechenden Näherungen und ihre Nachbearbeitungen für ein feineres Gitter mit $K = 29312$ Elementen, unter Verwendung

der gleichen Parameter für modale Filterung und DTV-Nachbearbeitung wie im Fall des vorhergehenden Gitters. Sowohl die globale als auch die adaptive Anwendung des modifizierten DTV-Filters auf Dreiecksgitter-basierten Graphen liefern hierbei akzeptable Ergebnisse. Die globale DTV-Filterung führt bei diesem Testfall zumindest für die Diskretisierungsparameter $N = 2,\ K = 7338, 29312$ und $N = 4,\ K = 7338$ sogar zu besseren Resultaten.

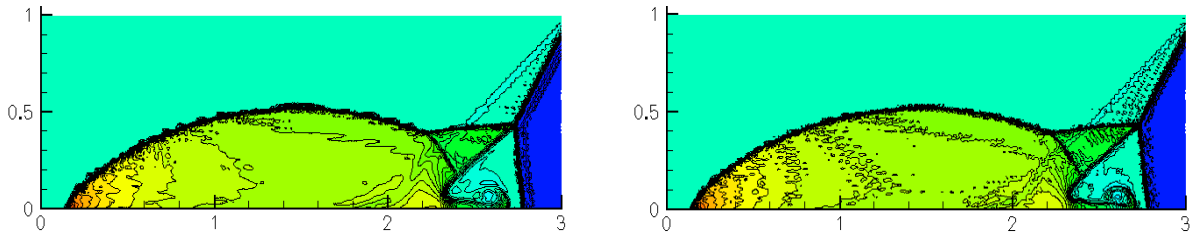


Abb. 7.21: Lösungen des DG-Verfahrens mit modaler Filterung für $N = 2$ (links) und $N = 4$ (rechts) auf einem Gitter mit $K = 7338$ Elementen. Darstellung der Dichte, 35 Isolinen zu äquidistanten Werten von 1.3965 bis 22.682.

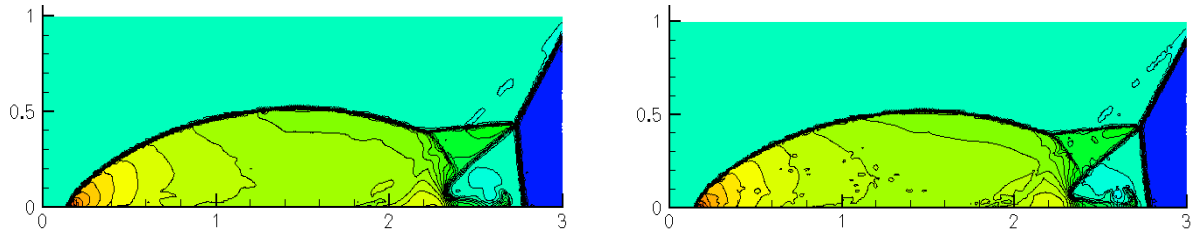


Abb. 7.22: Nachbearbeitete Lösungen, adaptive Anwendung des DTV-Filters auf Dreiecksgitter-basiertem Graphen.

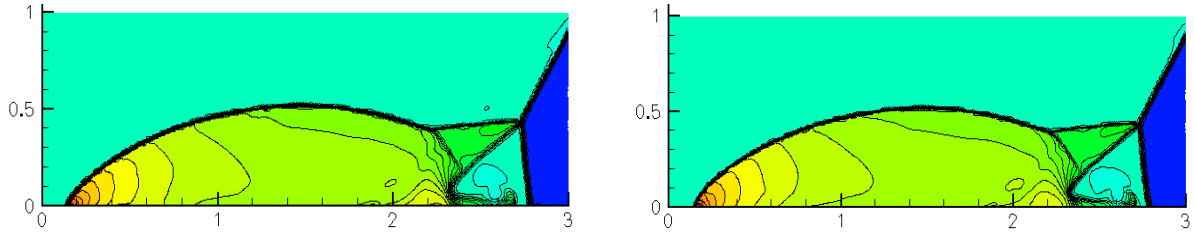


Abb. 7.23: Nachbearbeitete Lösungen, globale Anwendung des DTV-Filters auf Dreiecksgitter-basiertem Graphen.

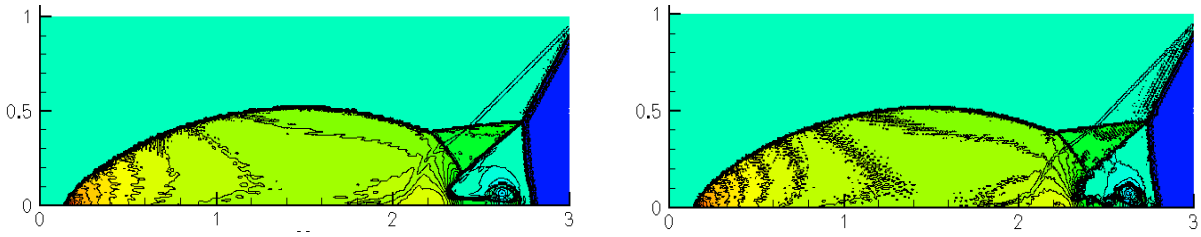


Abb. 7.24: Lösungen des DG-Verfahrens mit modaler Filterung für $N = 2$ (links) und $N = 4$ (rechts) auf einem Gitter mit $K = 29312$ Elementen. Darstellung der Dichte, 35 Isolinen zu äquidistanten Werten von 1.3965 bis 22.682.

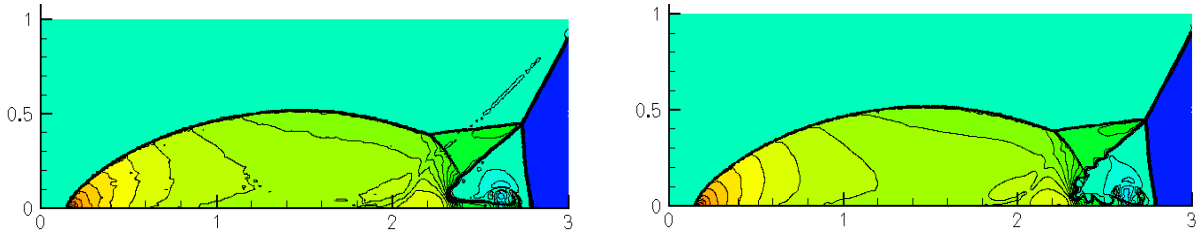


Abb. 7.25: Nachbearbeitete Lösungen, adaptive Anwendung des DTV-Filters auf Dreiecksgitter-basiertem Graphen.

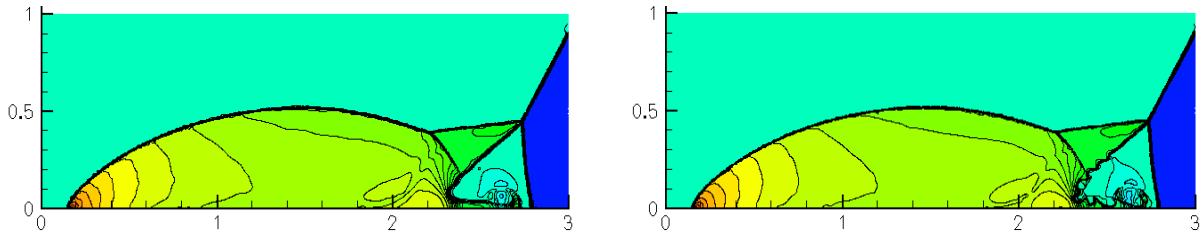


Abb. 7.26: Nachbearbeitete Lösungen, globale Anwendung des DTV-Filters auf Dreiecksgitter-basiertem Graphen.

Strömung über eine vorwärtsgerichtete Stufe (Forward Facing Step)

Dieser Testfall wird ebenfalls in [99] beschrieben und beinhaltet die Strömung eines Gases mit Isentropenkoeffizienten $\gamma = 1.4$ in einem Windkanal, der eine vorwärtsgerichtete Stufe enthält. Das im Kanal befindliche Gas besitzt anfangs konstante Werte der Dichte $\rho = 1.4$, des Druck $p = 1$ sowie der Geschwindigkeit $v_x = 3,\ v_y = 0$, und Gas mit diesen Werten strömt kontinuierlich in den Windkanal ein. Das Rechengebiet für den dieser Vorstellung entsprechenden zweidimensionalen Testfall sowie die zugehörigen Randbedingungen sind in der Abbildung 7.27 dargestellt.

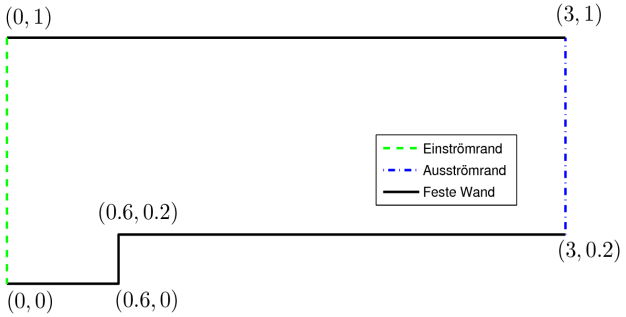


Abb. 7.27: Rechengebiet und Randbedingungen für Strömung über eine vorwärtsgerichtete Stufe.

Die Abbildung 7.28 zeigt die Näherungslösungen des DG-Verfahrens mit modaler Filterung zum Zeitpunkt $T = 4$ unter Verwendung der Parameter $p = 2,\ C_p = 2$ sowie des Stoßindikators s_J für $N = 5$ auf zwei verschiedenen Dreiecksgittern mit $K = 1624$ bzw. $K = 6496$ Elementen. Das feinere der beiden Gitter ist hierbei aus einer Rot-Verfeinerung des gröberen entstanden. Abbildung 7.29 zeigt die global bzw. adaptiv DTV-gefilterte Näherungslösung nach 100 Iterationen für das feinere der beiden Gitter, wobei ein Dreiecksgitter-basierter DTV-Graph verwendet und $\lambda = 2$ gesetzt wurde. Wie im Beispiel der Doppel-Mach-Reflektion wurde zur Konstruktion des DTV-Graphen jedes Dreieck der Triangulierung in 9 Teildreiecke zerlegt. Die sich ausbildenden feineren Strukturen und der beginnende Stoß im oberen Bereich des Rechengebiets wurden hierbei zusammen mit den Oszillationen entfernt. Unter Beibehaltung schwächerer Oszillationen lässt sich dieser Strukturverlust vermeiden, wenn zusätzlich der Parameter λ adaptiv gewählt wird. Eine diesbezügliche weitere Variation des DTV-Filters erhält man beispielsweise durch die Markierung aller Elemente $\tau_i \in \mathcal{T}^h$ mit $s_{res}(\tau_i) > 0.01$ und die Wahl des in die Berechnung der Filterkoeffizienten (6.6) eingehenden Anpassungsparameters durch $\lambda = \lambda_i = \frac{\lambda_{ref}}{s_{res}}$ mit $\lambda_{ref} = 2$. Die Ergebnisse dieser neuen adaptiven Filtervariante sind in Abbildung 7.30 dargestellt.

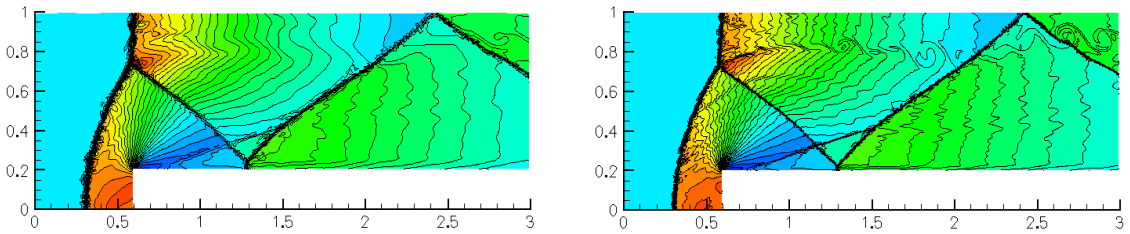


Abb. 7.28: Lösungen des DG-Verfahrens mit modaler Filterung für $N = 5$ und $K = 1624$ (links) sowie $K = 6496$ (rechts). Darstellung der Dichte, 30 Isolinien zu äquidistanten Werten von 0.090338 bis 6.2365.

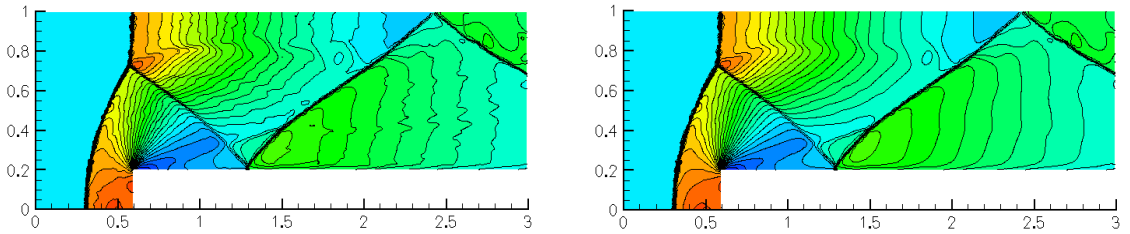


Abb. 7.29: DTV-gefilterte Näherungslösung auf Dreiecksgitter-basiertem DTV-Graphen für $N = 5$ und $K = 6496$. Adaptive DTV-Filterung (links) sowie globale DTV-Filterung (rechts).

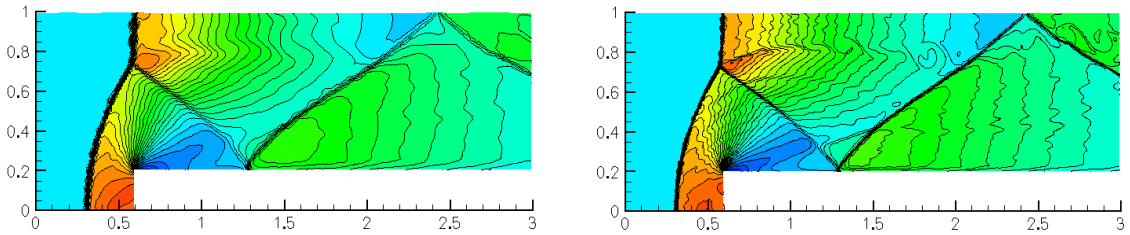


Abb. 7.30: DTV-gefilterte Näherungslösung auf Dreiecksgitter-basierten DTV-Graphen für $N = 5$ und $K = 1624$ (links) sowie $K = 6496$ (rechts). DTV-Filterung auf den durch s_{res} markierten Teilgraphen unter Verwendung eines adaptiven Anpassungsparameters λ.

8 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde ein diskontinuierliches Galerkin-Verfahren auf Dreiecksgittern mit neuartiger Stabilisierung durch modale Filterung zur Lösung hyperbolischer Erhaltungsgleichungen konstruiert. Zur Berechnung und Darstellung der stückweise polynomialen Näherunglösung auf Triangulierungen wurde die orthogonale Polynombasis der Proriol-Koornwinder-Dubiner-Polynome auf einem Referenzdreieck verwendet. Die Konstruktion des zur Stabilisierung der DG-Methode verwendeten modalen Filters erfolgte über eine neuartige Formulierung spektraler Viskosität auf dem Referenzdreieck, mit Hilfe des zur Basis der PKD-Polynome gehörigen Sturm-Liouville-Operators. Aus dieser Formulierung ergab sich ein Erkenntnisgewinn in Form von Bedingungen an die Wahl des Filterparameters C_p bei Gitterverfeinerung oder Erhöhung des Polynomgrades. Die Theorie spektraler Verfahren ergänzende neue Resultate wurden durch den Nachweis der Approximationseigenschaften gefilterter Entwicklungen in diese Basis sowie durch eine bisher noch nicht aufgestellte Abschätzung für die zur PKD-Basis gehörige gewichtete Norm erbracht. In Bezug auf Gitterverfeinerungen konnten wir nachweisen, dass die globale Anwendung des modalen Filters ein Gesamtverfahren höchstens erster Ordnung liefert. Aus diesem Grund wurde die adaptive Anwendung des modalen Filters abhängig von zwei in diesem Kontext noch nicht betrachteten Stoßindikatoren untersucht.

Die in dieser Arbeit gezeigten numerischen Resultate zu einer großen Bandbreite von Testfällen lassen darauf schließen, dass das entworfene DG-Verfahren mit adaptiver modaler Filterung unter Voraussetzung der geeigneten Wahl der Filterparameter für beide Stoßindikatoren stabile Approximationen liefert. Die neue Dämpfungsstrategie basierend auf adaptiver modaler Filterung ist daher eine ernst zu nehmende Alternative zu bisher hauptsächlich verwendeten Limitern oder aufwändigen Rekonstruktionsprozeduren. Die Abhängigkeit der Ergebnisse von der Wahl der Filterparameter ist hierbei ein Problem, welches generell im Fall explizit eingeführter Viskosität auftritt und auch im Fall des modifizierten minmod-Limiters von Cockburn und Shu bekannt ist. Optisch lieferten die untersuchten Stoßindikatoren in den betrachteten Fällen nahezu identische Ergebnisse. Allerdings wiesen die Näherungslösungen zum Indikator s_{res} im Fall des aus Burgers- und Advektionsgleichung zusammengesetzten Testfalls (2.21) bessere Fehlerverläufe fern des Stoßes der exakten Lösung auf, als die Näherungslösungen zum Indikator s_J. Die Anwendung des Verfahrens auf die Stoß-Wirbel-Interaktion von Shu und Osher zeigte zudem, dass der Indikator s_J für diesen Testfall eine deutlich größere Anzahl von Elementen markiert als der Indikator s_{res}.

Ein Konvergenznachweis für die der modalen Filterung zugrunde liegende elementweise Formulierung spektraler Viskosität auf Dreiecksgittern lag außerhalb der Zielsetzung dieser Arbeit. Grundsätzlich ist eine diesbezügliche theoretische Untersuchung jedoch dringend erforderlich. Wünschenswert wären hierbei insbesondere auch theoretische Aussagen über die verwendeten Indikatoren zur adaptiven Steuerung des modalen Filters. Die Verwendung anderer Indikatoren sollte hierbei ebenso in Betracht gezogen werden. Durch darauf aufbauende weitere Anpassungen der Filteralgorithmen an die Struktur der numerischen Lösung, die dann auf Grundlage theoretisch fundierter Stoßindikatoren ermittelt wird, wäre die Entwicklung intelligenter Black-Box-Mechanismen auf Basis der in dieser Arbeit untersuchten adaptiven modalen Filter zu Stabilisierung hochauflösender DG-Verfahren denkbar.

Im Fall hoher Polynomgrade sind bei den hier verwendeten expliziten Zeitintegrations-

verfahren aus Stabilitätsgründen sehr kleine Zeitschritte festzulegen. Ein langfristiges Ziel des dieser Arbeit übergeordneten Projekts ist daher die Verwendung impliziter Zeitintegrationsverfahren zur numerischen Lösung der semidiskreten Gleichung (4.6), die im Allgemeinen größere Zeitschritte zulassen und daher zu einem Effizienzgewinn führen können. Zu untersuchen ist hierbei, ob auch in dieser Situation die modale Filterung zur Stabilisierung der DG-Methode verwendet werden kann.

Desweiteren wird im Fall der Euler-Gleichungen die Verwendung sogenannter positivitätserhaltender Limiter nach Zhang und Shu, siehe [104], anstelle der rechenzeitintensiveren iterativen Nutzung des modalen Filters in Betracht gezogen. Die Eigenschaft der Erhaltung der Positivität von Dichte und Druck ist allerdings für diese Limiter nur nachweisbar im Fall von TVD-Zeitintegrationen sowie unter zusätzlicher Restriktion des Zeitschritts.

Zur Nachbearbeitung in der Näherungslösung verbleibender Oszillationen wurde in der vorliegenden Arbeit der in der Bildverarbeitung entwickelte digitale TV-Filter im neuen Kontext von DG-Verfahren auf Dreiecksgittern eingesetzt. Neue Erkenntnisse ergaben sich im Zusammenhang mit der Anwendung des DTV-Filters auf Dreiecksgitter-basierten Graphen, die eine einfachere Auswertung der Näherugslösungen des DG-Verfahrens an den Knoten ermöglichen. In diesem Fall führte die Anwendung des DTV-Filters auf eine lineare Ausgangsfunktion zu einer den Kanten der Triangulierung folgenden unerwünschten Musterung, so dass für diesen Gittertyp eine Modifikation der Iterationsvorschrift des Filters vorgeschlagen wurde.

Den präsentierten Ergebnissen nach zu urteilen, stellt sowohl der DTV-Filter auf kartesischen Graphen als auch dessen Modifikation auf Dreiecksgitter-basierten Graphen eine vielversprechende und richtungsweisende Alternative zu Reprojektionstechniken dar. Erstmalig wurden zudem die Fehlerverläufe fern von Stößen betrachtet, die sich bei Anwendung des DTV-Filters auf Näherungslösungen des DG-Verfahrens unter Gitterverfeinerung im Fall einer hyperbolischen Erhaltungsgleichung mit unstetiger Lösung ergeben. Bei globaler Anwendung des DTV-Filters zeigte sich hierbei ein Ordnungsverlust, so dass eine erste Variante der adaptiven DTV-Filterung im Kontext von DG-Verfahren auf Dreiecksgittern entwickelt wurde. In den meisten Fällen ergaben sich durch die adaptive DTV-Filterung ebenso gute oder bessere Ergebnisse als mit der globalen DTV-Filterung. Bessere Resultate des globalen DTV-Filters wie beispielsweise im Fall der Doppel-Mach-Reflektion sprechen eher dafür, dass die adaptive Anwendung des DTV-Filters noch zu verbessern ist. Ein erster diesbezüglicher Versuch im Fall der Strömung über eine vorwärtsgerichtete Stufe, mit adaptiver Wahl des Anpassungsparameters, zeigt, dass ein solcher Spielraum zur Verbesserung der Methode noch vorhanden ist. Andere Indikatoren zur Markierung der Elemente, auf denen der DTV-Filter anzuwenden ist, könnten ebenfalls zu besseren Ergebnissen führen. Neben der Weiterentwicklung des adaptiven DTV-Filters sollten zudem zukünftig auch andere Bildverarbeitungsmethoden zur Nachbearbeitung der Näherungslösungen des DG-Verfahrens mit modaler Filterung in Betracht gezogen werden. Denkbar wären hierbei Ansätze, die auf partiellen Differentialgleichungen beruhen, wie beispielsweise die aus dem klassischen Modell inhomogener Diffusion von Perona und Malik hervorgegangene anisotrope Diffusionsgleichung (6.1) von Weickert. Die direktere Kontrolle der Totalvariation der Näherungslösung durch die in dieser Arbeit vorgeschlagene Verwendung von Methoden der Bildverarbeitung im Vergleich zur Anwendung von Reprojektionstechniken sollte hierbei entscheidende Impulse zur Weiterentwicklung von Nachbearbeitungstechniken im Kontext komplexer praxisrelevanter Problemstellungen liefern.

Wie bereits geschildert, basiert die Anwendung von Nachbearbeitungsprozeduren im Kontext spektraler Verfahren auf der Vermutung, dass hochgenaue Informationen in den Näherungslösungen des spektralen Verfahrens enthalten sind, die eine Rekonstruktion punktweiser Werte mit hoher Genauigkeit ermöglichen. Ein Nachweis dieser Vermutung, der bestenfalls auch andeutet, in welcher Form diese Informationen in der spektralen Approximation enthalten sind, wäre wünschenswert, aber scheint mittelfristig nicht erreichbar zu sein. Bei einem Fortschritt in dieser Richtung könnten allerdings Nachbearbeitungsmethoden auf Basis entsprechender theoretischer Resultate entwickelt werden, die auch nachweisbar die Genauigkeit des Verfahrens an Stößen erhöhen.

Zusammenfassend ergibt sich, dass durch die in dieser Arbeit verfolgte erstmalige Übertragung des innerhalb der Spektralmethoden konstruierten Gesamtpakets bestehend aus der Stabilisierung durch modale Filter sowie der Nachbearbeitung mittels einer TV-Minimierungsprozedur in den Kontext von diskontinuierlichen Galerkin-Verfahren auf Dreiecksgittern ein grundlegendes Werkzeug zur hochauflösenden Simulation komplexer Strömungsphänomene geschaffen wurde. Die einzelnen Aspekte dieses Grundkonzepts bieten hierbei noch Raum für gezielte Untersuchungen in nachfolgenden Arbeiten.

A Anhang

A.1 Die Jacobi-Polynome und die Gauß-Jacobi-Quadratur

Die im folgenden aufgeführten Definitionen und Eigenschaften der Jacobi-Polynome sowie der Gauß-Jacobi-Quadratur sind im wesentlichen dem Anhang des Buches von Karniadakis und Sherwin [52] entnommen. Ein wichtiges Referenzwerk zur Theorie orthogonaler Polynome ist zudem das Buch von Szegö [86].
In den nachfolgenden Abschnitten sei $I = [-1, 1]$, $n \in \mathbb{N}_0$, sowie $\mathcal{P}^n(I)$ die Menge der Polynome $p : I \to \mathbb{R}$ vom Grad $\leq n$.

Definition und Eigenschaften der Jacobi-Polynome

Für $\alpha, \beta \in \mathbb{R}$ mit $\alpha, \beta > -1$ sind die *Jacobi-Polynome* $P_n^{\alpha,\beta} : I \to \mathbb{R}$ durch die nachstehenden Eigenschaften eindeutig definiert:

- $P_n^{\alpha,\beta} \in \mathcal{P}^n(I)$ für alle $n \in \mathbb{N}_0$,
- die Orthogonalitätseigenschaft

$$\int_{-1}^{1} \omega_{\alpha\beta}(x) P_l^{\alpha,\beta}(x) P_m^{\alpha,\beta}(x)\, dx \;=\; 0 \quad \text{für} \quad l \neq m, \tag{A.1}$$

  mit der Gewichtsfunktion $\omega_{\alpha\beta}(x) = (1-x)^\alpha (1+x)^\beta$,
- $P_n^{\alpha,\beta}(1) = \binom{n+\alpha}{n}$.

Für die Werte $\alpha = \beta = 0$ erhält man den Spezialfall der *Legendre-Polynome*.

Die Jacobi-Polynome $P_n^{\alpha,\beta}$ bilden ein vollständiges Orthogonalsystem des Hilbertraumes $L^2_{\omega_{\alpha\beta}}[-1, 1]$ der bezüglich des Gewichts $\omega_{\alpha\beta}$ quadratintegrierbaren Funktionen auf dem Intervall $[-1, 1]$ mit dem Skalarprodukt

$$(u, v)_{L^2_{\omega_{\alpha\beta}}} = \int_{-1}^{1} \omega_{\alpha\beta}(x) u(x) v(x)\, dx.$$

Mit der Definition der Gammafunktion

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \quad \text{für } z \in \mathbb{C} \text{ mit } \operatorname{Re} z > 0, \tag{A.2}$$

die für $n \in \mathbb{N}$ die Gleichung $\Gamma(n+1) = n!$ erfüllt, gilt desweiteren

$$\int_{-1}^{1} \omega_{\alpha\beta}(x) P_l^{\alpha,\beta}(x) P_l^{\alpha,\beta}(x)\, dx \;=\; \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\Gamma(n+\alpha+\beta+1)}\,. \tag{A.3}$$

Bezüglich der Werte der Jacobi-Polynome für $q = \max(\alpha, \beta) \geq -\frac{1}{2}$ gilt die Abschätzung

$$\max_{-1 \leq x \leq 1} |P^{\alpha,\beta}(x)| \;=\; \binom{n+q}{n}. \tag{A.4}$$

Zudem sind die Jacobi-Polynome die Lösungen des Sturm-Liouville-Problems

$$\begin{aligned} \frac{d}{dx}\left[(1-x)^{1+\alpha}(1+x)^{1+\beta}\frac{du(x)}{dx}\right] &= -\lambda_n (1-x)^\alpha (1+x)^\beta u(x), \\ \lambda_n &= n(n+\alpha+\beta+1). \end{aligned}$$

Rekursionsformeln Da keine einfachen expliziten Berechnungsvorschriften für die Werte der Jacobi-Polynome und deren Ableitungen bekannt sind, werden häufig die nachfolgenden Rekursionsformeln zur Berechnung verwendet.

$$\begin{aligned} P_0^{\alpha,\beta}(x) &= 1, \\ P_1^{\alpha,\beta}(x) &= \frac{1}{2}[\alpha - \beta + (\alpha + \beta + 2)x], \\ a_n^1 P_{n+1}^{\alpha,\beta}(x) &= (a_n^2 + a_n^3 x) P_n^{\alpha,\beta}(x) - a_n^4 P_{n-1}^{\alpha,\beta}(x), \end{aligned}$$

mit den Koeffizienten $a_n^1, \ldots, a_n^4 \in \mathbb{R}$ gegeben durch die Gleichungen

$$\begin{aligned} a_n^1 &= 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta), \\ a_n^2 &= (2n+\alpha+\beta+1)(\alpha^2-\beta^2), \\ a_n^3 &= (2n+\alpha+\beta)(2n+\alpha+\beta+1)(2n+\alpha+\beta+2), \\ a_n^4 &= 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2), \end{aligned}$$

sowie

$$b_n^1(x)\frac{d}{dx}P_n^{\alpha,\beta}(x) = b_n^2(x)P_n^{\alpha,\beta}(x) + b_n^3(x)P_{n-1}^{\alpha,\beta}(x),$$

mit

$$\begin{aligned} b_n^1(x) &= (2n+\alpha+\beta)(1-x^2), \\ b_n^2(x) &= n[\alpha - \beta - (2n+\alpha+\beta)x], \\ b_n^3(x) &= 2(n+\alpha)(n+\beta). \end{aligned}$$

Die Gauß-Jacobi-Quadratur

Die Gauß-Jacobi-Quadraturformeln stellen eine Verallgemeinerung der Gauß-Quadratur dar. Sie dienen der numerischen Integration einer Funktion $u : I \to \mathbb{R}$ bezüglich der zu den Jacobi-Polynomen $P_n^{\alpha,\beta}$, $\alpha, \beta > -1$, gehörigen jeweiligen Gewichtsfunktionen $\omega_{\alpha\beta}(x) = (1-x)^\alpha(1+x)^\beta$. Man unterscheidet die klassische Gauß-Jacobi-Quadratur, bei der wie bei der Gauß-Quadratur nur Stützstellen im Inneren des Intervalls I gewählt werden, von den Formeln vom Radau- und Lobatto-Typ, die einen beziehungsweise beide Randpunkte des Intervalls einbeziehen.
Für fest gewählte Parameter α, β und eine vorgegebene Stützstellenanzahl Q gilt es, Stützstellen $x_i,\ i = 0, \ldots, Q-1$, und zugehörige Gewichte $w_i,\ i = 0, \ldots, Q-1$, zu bestimmen, so dass die zugehörige Quadraturformel eine möglichst hohe Ordnung hat. Für den Exaktheitsgrad der Gauß-Jacobi-Quadratur gilt die Gleichung

$$\int_{-1}^{1}(1-x)^\alpha(1+x)^\beta u(x)dx = \sum_{i=0}^{Q} w_i u(x_i) + R(u),$$

mit $R(u) = 0$, falls $u \in \mathcal{P}^{2Q-\delta}(I)$. Der Parameter δ ist hierbei abhängig vom Typ der Quadraturformel, d.h. von den Einschränkungen an die Stützstellenwahl:

$$\delta = \begin{cases} 1 & \text{klassische Gauß-Jacobi-Quadratur,} \\ 2 & \text{Gauß-Radau-Jacobi-Formeln,} \\ 3 & \text{Gauß-Lobatto-Jacobi-Formeln.} \end{cases}$$

Die konkrete Berechnung der Stützstellen und Gewichte ergibt sich aus den folgenden aus [52] entnommenen Formeln. Mit $x_{i,Q}^{\alpha,\beta} \in (-1,1),\ i = 0,\ldots,Q-1$, seien hierbei die Nullstellen des Jacobipolynoms $P_Q^{\alpha,\beta}$ und mit Γ die in (A.2) definierte Gammafunktion bezeichnet.

Gauß-Jacobi-Formeln

$$\begin{aligned} x_i &= x_{i,Q}^{\alpha,\beta},\ i = 0,\ldots,Q-1, \\ w_i &= \frac{2^{\alpha+\beta+1}\Gamma(\alpha+Q+1)\Gamma(\beta+Q+1)}{Q!\,\Gamma(\alpha+\beta+Q+1)(1-x_i^2)} \left[\frac{d}{dx}P_Q^{\alpha,\beta}(x_i)\right]^{-2},\ i = 0,\ldots,Q-1. \end{aligned}$$

Gauß-Radau-Jacobi-Formeln

$$\begin{aligned} x_i &= \begin{cases} -1, & i = 0 \\ x_{i-1,Q-1}^{\alpha,\beta+1}, & i = 1,\ldots,Q-1, \end{cases} \\ w_i &= \begin{cases} (\beta+1)B_{0,Q-1}^{\alpha,\beta}, & i = 0 \\ B_{i,Q-1}^{\alpha,\beta}, & i = 1,\ldots,Q-1, \end{cases} \end{aligned}$$

mit

$$B_{i,Q-1}^{\alpha,\beta} = \frac{2^{\alpha+\beta}\Gamma(\alpha+Q)\Gamma(\beta+Q)(1-x_i)}{(Q-1)!\,(\beta+Q)\Gamma(\alpha+\beta+Q+1)\left[P_{Q-1}^{\alpha,\beta}(x_i)\right]^2},\ i = 0,\ldots,Q-1.$$

Gauß-Lobatto-Jacobi-Formeln

$$\begin{aligned} x_i &= \begin{cases} -1, & i = 0 \\ x_{i-1,Q-2}^{\alpha+1,\beta+1}, & i = 1,\ldots,Q-2, \\ 1, & i = Q-1 \end{cases} \\ w_i &= \begin{cases} (\beta+1)C_{0,Q-2}^{\alpha,\beta}, & i = 0 \\ C_{i,Q-2}^{\alpha,\beta}, & i = 1,\ldots,Q-2, \\ (\alpha+1)C_{Q-1,Q-2}^{\alpha,\beta}, & i = Q-1, \end{cases} \end{aligned}$$

mit

$$C_{i,Q-2}^{\alpha,\beta} = \frac{2^{\alpha+\beta+1}\Gamma(\alpha+Q)\Gamma(\beta+Q)}{(Q-1)(Q-1)!\,\Gamma(\alpha+\beta+Q+1)\left[P_{Q-1}^{\alpha,\beta}(x_i)\right]^2},\ i = 0,\ldots,Q-1.$$

A.2 Das Flussvektor-Splitting-Verfahren von van Leer

Bei der von van Leer in [61] beschriebenen Definition einer numerischen Flussfunktion für die eindimensionalen Euler-Gleichungen in Form eines Flussvektor-Splitting-Verfahrens werden zunächst die primitiven Variablen $(\rho^{\pm}, v_1^{\pm}, v_2^{\pm}, p^{\pm})$ aus den in $\mathbf{u}^{\pm}$ gegebenen konservativen Größen berechnet. Der numerische Fluss $\mathbf{H}_{1D}(\mathbf{u}^-, \mathbf{u}^+)$ wird in Abhängigkeit von $M^{\pm} = v_1^{\pm}/c^{\pm}$ definiert als

$$\mathbf{H}_{1D}(\mathbf{u}^-, \mathbf{u}^+) = \mathbf{H}^-(\mathbf{u}^-) + \mathbf{H}^+(\mathbf{u}^+),$$

mit

$$\mathbf{H}^-(\mathbf{u}) = \begin{cases} \mathbf{0} & M \geq 1, \\ \mathbf{f}^-(\mathbf{u}) & -1 < M < 1, \\ \mathbf{f}_1(\mathbf{u}) & M \leq -1, \end{cases}$$

und

$$\mathbf{H}^+(\mathbf{u}) = \begin{cases} \mathbf{f}_1(\mathbf{u}) & M \geq 1, \\ \mathbf{f}^+(\mathbf{u}) & -1 < M < 1, \\ \mathbf{0} & M \leq -1, \end{cases}$$

mit den Zwischenwerten

$$\begin{aligned}
\mathbf{f}^-(\mathbf{u}) &= \left(-\frac{\rho c}{4}(1-M)^2,\, f_1^-(\mathbf{u})\frac{C^-(\mathbf{u})}{\gamma},\, f_1^-(\mathbf{u})v_2,\, f_2^-(\mathbf{u})C^-(\mathbf{u})\frac{\gamma}{2(\gamma^2-1)} + \frac{1}{2}f_3^-(\mathbf{u})v_2 \right), \\
\mathbf{f}^+(\mathbf{u}) &= \left(\frac{\rho c}{4}(1+M)^2,\, f_1^+(\mathbf{u})\frac{C^+(\mathbf{u})}{\gamma},\, f_1^+(\mathbf{u})v_2,\, f_2^+(\mathbf{u})C^+(\mathbf{u})\frac{\gamma}{2(\gamma^2-1)} + \frac{1}{2}f_3^+(\mathbf{u})v_2 \right), \\
C^-(\mathbf{u}) &= (\gamma-1)v_1 - 2c, \\
C^+(\mathbf{u}) &= (\gamma-1)v_1 + 2c.
\end{aligned}$$

A.3 Fehlerentwicklung und Laufzeiten der RKDG-Verfahren: lineare Advektion

Dargestellt sind die L^1-, L^2- und L^∞-Fehler sowie die Rechenzeiten in Sekunden der Näherungslösungen verschiedener RKDG-Verfahren zum Testfall der lineare Advektionsgleichung, siehe Unterkapitel 4.3.2.

N	K	L^1-Fehler	EOC	L^2-Fehler	EOC	L^∞-Fehler	EOC	Laufzeit
0	296	1.753e-03		9.935e-03		1.859e-01		0.01
	1184	1.550e-03	0.18	9.100e-03	0.13	1.730e-01	0.10	0.10
	4736	1.251e-03	0.31	7.840e-03	0.22	1.531e-01	0.18	1.28
	18944	9.149e-04	0.45	6.219e-03	0.33	1.280e-01	0.26	17.52

Tabelle A.1: DG-Verfahren mit Zeitintegration durch das explizite Eulerverfahren.

N	K	L^1-Fehler	EOC	L^2-Fehler	EOC	L^∞-Fehler	EOC	Laufzeit
0	296	1.785e-03		1.010e-02		1.878e-01		0.03
	1184	1.610e-03	0.15	9.348e-03	0.11	1.762e-01	0.09	0.23
	4736	1.340e-03	0.26	8.220e-03	0.19	1.585e-01	0.15	2.69
	18944	1.015e-03	0.40	6.717e-03	0.29	1.353e-01	0.23	32.82
1	296	1.167e-03		6.608e-03		1.233e-01		0.14
	1184	5.446e-04	1.10	3.601e-03	0.88	7.602e-02	6.97	1.25
	4736	1.444e-04	1.92	1.125e-03	1.68	2.976e-02	1.35	13.72
	18944	2.484e-05	2.54	2.101e-04	2.42	6.748e-03	2.14	130.92

Tabelle A.2: DG-Verfahren mit Zeitintegration durch das RK-Verfahren (4.9) zweiter Ordnung.

N	K	L^1-Fehler	EOC	L^2-Fehler	EOC	L^∞-Fehler	EOC	Laufzeit
0	296	1.784e-03		1.010e-02		1.878e-01		0.03
	1184	1.610e-03	0.15	9.347e-03	0.11	1.762e-01	0.09	0.27
	4736	1.340e-03	0.26	8.219e-03	0.19	1.585e-01	0.15	3.23
	18944	1.015e-03	0.40	6.716e-03	0.29	1.354e-01	0.23	39.33
1	296	1.159e-03		6.571e-03		1.230e-01		0.17
	1184	5.392e-04	1.10	3.573e-03	0.88	7.574e-02	0.70	1.50
	4736	1.426e-04	1.92	1.112e-03	1.68	2.957e-02	1.46	16.46
	18944	2.449e-05	2.54	2.071e-04	2.43	6.661e-03	2.15	155.41
2	296	5.734e-04		3.375e-03		6.739e-02		0.71
	1184	8.261e-05	2.80	6.127e-04	2.46	1.537e-02	2.13	6.41
	4736	5.813e-06	3.82	4.706e-05	3.70	2.509e-03	2.61	61.33
	18944	4.965e-07	3.55	4.184e-06	3.49	3.849e-04	2.70	526.17
3	296	1.915e-04		1.175e-03		2.156e-02		2.12
	1184	8.483e-06	4.50	6.890e-05	4.09	4.029e-03	2.42	18.80
	4736	3.286e-07	4.69	2.816e-06	4.61	3.530e-04	3.51	166.14
	18944	1.921e-08	4.10	1.707e-07	4.04	2.609e-05	3.76	1366.66
4	296	5.322e-05		3.537e-04		1.030e-02		5.89
	1184	9.413e-07	5.82	7.804e-06	5.50	5.909e-04	4.12	51.01
	4736	2.590e-08	5.18	2.276e-07	5.10	3.210e-05	4.20	422.91
	18944	1.357e-09	4.25	1.140e-08	4.32	1.220e-06	4.72	3479.95
5	296	1.152e-05		8.097e-05		2.333e-03		16.29
	1184	1.242e-07	6.54	1.100e-06	6.20	1.569e-04	3.89	137.83
	4736	4.543e-09	4.77	3.743e-08	4.88	2.327e-06	6.08	1126.79
	18944	5.220e-10	3.12	4.259e-09	3.14	1.002e-07	4.54	9143.75
6	296	2.757e-06		2.060e-05		1.560e-03		30.96
	1184	2.119e-08	7.02	1.786e-07	6.85	1.281e-05	6.93	270.97
	4736	1.833e-09	3.53	1.497e-08	3.58	3.629e-07	5.14	2203.52
	18944	2.286e-10	3.00	1.866e-09	3.00	3.910e-08	3.21	19093.74
7	296	5.752e-07		4.258e-06		2.334e-04		52.22
	1184	8.268e-09	6.12	6.742e-08	5.98	2.572e-06	6.50	451.52
	4736	9.922e-10	3.06	8.101e-09	3.06	1.700e-07	3.92	3706.39
	18944	1.241e-10	3.00	1.013e-09	3.00	2.118e-08	3.01	29548.64
8	296	1.329e-07		1.074e-06		7.369e-05		91.33
	1184	4.175e-09	4.99	3.401e-08	4.98	7.562e-07	6.61	777.48
	4736	5.204e-10	3.00	4.249e-09	3.00	8.885e-08	3.09	6232.71
	18944	6.510e-11	3.00	5.313e-10	3.00	1.110e-08	3.00	49531.36

Tabelle A.3: DG-Verfahren mit Zeitintegration durch das RK-Verfahren (4.10) dritter Ordnung.

N	K	L^1-Fehler	EOC	L^2-Fehler	EOC	L^∞-Fehler	EOC	Laufzeit
0	296	1.784e-03		1.010e-02		1.879e-01		0.04
	1184	1.610e-03	0.15	9.347e-03	0.11	1.762e-01	0.09	0.29
	4736	1.340e-03	0.26	8.219e-03	0.19	1.585e-01	0.15	2.99
	18944	1.015e-03	0.40	6.716e-03	0.29	1.354e-01	0.23	41.86
1	296	1.159e-03		6.568e-03		1.300e-01		0.19
	1184	5.387e-04	1.11	3.569e-03	0.88	7.570e-02	0.70	1.70
	4736	1.424e-04	1.92	1.110e-03	1.68	2.954e-02	1.36	17.13
	18944	2.445e-05	2.54	2.067e-04	2.43	6.653e-03	2.15	167.54
2	296	5.728e-04		3.371e-03		6.730e-02		0.74
	1184	8.233e-05	2.80	6.102e-04	2.47	1.531e-02	2.14	6.55
	4736	5.773e-06	3.83	4.669e-05	3.71	2.520e-03	2.60	60.60
	18944	4.939e-07	3.55	4.164e-06	3.49	3.857e-04	2.71	520.63
3	296	1.910e-04		1.170e-03		2.144e-02		2.20
	1184	8.413e-06	4.51	6.798e-05	4.11	4.043e-03	2.41	19.42
	4736	3.239e-07	4.70	2.774e-06	4.62	3.488e-04	3.54	167.90
	18944	1.892e-08	4.10	1.681e-07	4.04	2.557e-05	3.77	1375.72
4	296	5.307e-05		3.513e-04		1.034e-02		5.81
	1184	9.310e-07	5.83	7.648e-06	5.52	5.828e-04	4.15	50.57
	4736	2.385e-08	5.29	2.138e-07	5.16	3.083e-05	4.24	419.90
	18944	6.993e-10	5.10	6.559e-09	5.03	1.185e-06	4.70	3400.64
5	296	1.145e-05		7.990e-05		2.286e-03		15.98
	1184	1.203e-07	6.57	1.050e-06	6.25	1.550e-04	3.88	139.28
	4736	1.647e-09	6.19	1.541e-08	6.09	2.858e-06	5.76	1128.48
	18944	2.518e-11	6.03	2.412e-10	6.00	5.341e-08	5.74	8962.99
6	296	2.739e-06		2.030e-05		1.556e-03		30.98
	1184	1.495e-08	7.52	1.306e-07	7.28	1.453e-05	6.74	267.54
	4736	1.097e-10	7.09	1.060e-09	6.94	1.910e-07	6.25	2171.56
	18944	1.703e-12	6.01	1.450e-11	6.19	1.507e-09	6.99	17102.34
7	296	5.631e-07		4.128e-06		2.300e-04		51.79
	1184	1.933e-09	8.19	1.846e-08	7.80	2.614e-06	6.46	448.65
	4736	1.314e-11	7.20	1.126e-10	7.36	1.313e-08	7.64	3588.52
	18944	6.493e-13	4.34	5.317e-12	4.40	1.227e-10	6.74	28284.07
8	296	1.245e-07		1.003e-06		7.289e-05		86.10
	1184	2.504e-10	8.96	2.224e-09	8.82	2.119e-07	8.43	738.72
	4736	4.600e-12	5.77	3.772e-11	5.88	1.099e-09	7.59	5944.69
	18944	2.858e-13	4.01	2.345e-12	4.01	4.496e-11	4.61	47153.97

Tabelle A.4: DG-Verfahren mit Zeitintegration durch das RK-Verfahren (4.11) vierter Ordnung.

Literaturverzeichnis

[1] Abarbanel, S. ; Gottlieb, D. ; Tadmor, E.: Spectral methods for discontinuous problems. In: N.W. Morton, M.J. B. (Hrsg.): *Numerical Methods for Fluid Dynamics II.* Oxford University Press, 1986, S. 128–153

[2] Aubert, G. ; Kornprobst, P.: *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations.* 2nd ed. Springer, 2006

[3] Barter, G.E. ; Darmofal, D.L.: Shock Capturing with Higher-Order, PDE-Based Artificial Viscosity. In: *Proceedings of the 18th AIAA Computational Fluid Dynamics Conference*, 2007. – AIAA-2007-3823

[4] Bernardi, C. ; Maday, Y.: Polynomial interpolation results in Sobolev spaces. In: *J. Comput. Appl. Math.* 43 (1992), S. 53–80

[5] Biswas, R. ; Devine, K. D. ; Flaherty, J. E.: Parallel, Adaptive Finite Element Methods for Conservation Laws. In: *Appl. Numer. Math.* 14 (1994), S. 255–283

[6] Blackburn, H. M. ; Schmidt, S.: Spectral element filtering techniques for large eddy simulation with dynamic estimation. In: *J. Comput. Phys.* 186 (2003), S. 610–629

[7] Boyd, J. P.: Two comments on Filtering (Artificial Viscosity) for Chebychev and Legendre Spectral and Spectral Element Methods: Preserving Boundary Conditions and Interpretation of the Filter as a Diffusion. In: *J. Comput. Phys.* (1998), S. 283–288

[8] Breuss, M. ; Brox, T. ; Sonar, T. ; Weickert, J.: Stabilised Nonlinear Inverse Diffusion for Approximating Hyperbolic PDEs. In: *Proc. Scale Space*, 2005, S. 536–547

[9] Bürgel, A.: *Nichtlineare, diskrete Filteralgorithmen zur numerischen Lösung hyperbolischer Erhaltungsgleichungen*, Braunschweig, Diss., 2005

[10] Bürgel, A. ; Grahs, T. ; Sonar, T.: From continuous recovery to discrete filtering in numerical approximations of conservation laws. In: *Appl. Numer. Math.* 42 (2002), S. 47–50

[11] Canuto, C. ; Hussaini, M. Y. ; Quarteroni, A. ; Zang, T.A.: *Spectral Methods. Fundamentals in Single Domains.* Berlin : Springer, 2006

[12] Canuto, C. ; Hussaini, M. Y. ; Quarteroni, A. ; Zang, T.A.: *Spectral Methods. Evolution to Complex Domains and Applications to Fluid Dynamics.* New York : Springer, 2007

[13] Carpenter, M.H. ; Kennedy, C.A.: Fourth-order 2N-storage Runge-Kutta schemes. 1994. – Forschungsbericht. – NASA Report TM 109112

[14] Chan, T. F. ; Osher, S. ; Shen, J.: The digital TV filter and nonlinear denoising. In: *IEEE Trans. Image Process.* 10 (2001), S. 231–241

[15] CHEN, G.-Q. ; DU, Q. ; TADMOR, E.: Spectral Viscosity Approximations to Multidimensional Scalar Conservation Laws. In: *Math. Comput.* 61 (1993), Nr. 204, S. 629–643

[16] CIARLET, P. G.: *The Finite Element Method for Elliptic Problems*. Bd. 40. SIAM Classics in Applied Mathematics, 2002

[17] COCKBURN, B. ; HOU, S. ; SHU, C.-W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. In: *Math. Comput.* 54 (1990), S. 545–581

[18] COCKBURN, B. ; LIN, S. Y. ; SHU, C.-W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. In: *J. Comput. Phys.* 84 (1989), S. 90–113

[19] COCKBURN, B. ; SHU, C.-W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II: General framework. In: *Math. Comput.* 52 (1989), S. 411–435

[20] COCKBURN, B. ; SHU, C.-W.: The Runge-Kutta local projection P^1-discontinuous Galerkin method for scalar conservation laws. In: *RAIRO Modél. Math. Anal. Numér.* 25 (1991), S. 337–361

[21] COCKBURN, B. ; SHU, C.-W.: The local discontinuous Galerkin method for time-dependent convection-difussion systems. In: *SIAM J. Numer. Anal.* 35 (1998), Nr. 6, S. 2440–2463

[22] COCKBURN, B. ; SHU, C.-W.: The Runge-Kutta discontinuous Galerkin finite element method for conservation laws V: Multidimensional systems. In: *J. Comput. Phys.* 141 (1998), S. 199–224

[23] COCKBURN, B. ; SHU, C.-W.: Runge-Kutta Discontinuous Galerkin Methods for Convection-Dominated Problems. In: *J. Sci. Comp.* 16 (2001), S. 173–261

[24] COOLS, R.: An encyclopaedia of cubature formulas. In: *J. Complexity* 19 (2003), S. 445–453

[25] COURANT, R. ; FRIEDRICHS, K. ; LEWY, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. In: *Mathematische Annalen* 100 (1928), S. 32–74

[26] COURANT, R. ; FRIEDRICHS, K. O.: *Supersonic Flow and Shock Waves*. Springer, 1948

[27] DIPERNA, R.J.: Measure-Valued Solutions to Conservation Laws. In: *Arch. Rational Mech. Anal.* 88 (1985), S. 223–270

[28] DON, W. S.: Numerical Study of Pseudospectral Methods in Shock Wave Applications. In: *J. Comput. Phys.* 110 (1994), S. 103–111

[29] DON, W. S. ; GOTTLIEB, D. ; JUNG, J. H.: A multidomain spectral method for supersonic reactive flows. In: *J. Comput. Phys.* 192 (2003), S. 325–354

[30] Dubiner, M.: Spectral Methods on Triangles and other Domains. In: *J. of Scientific Computing* 6 (1991), Nr. 4, S. 345–390

[31] Feistauer, M. ; Kučera, V.: On a robust discontinuous Galerkin technique for the solution of compressible flow. In: *J. Comput. Phys.* 224 (2007), Nr. 1, S. 208–221

[32] Gelb, A.: Parameter Optimization and Reduction of Round Off Error for the Gegenbauer Reconstruction Method. In: *J. Sci. Comput.* 20 (2004), S. 433–459

[33] Gelb, A. ; Tadmor, E.: Enhanced spectral viscosity approximations for conservation laws. In: *Appl. Numer. Math.* 33 (2000), S. 3–21

[34] Godlewski, E. ; Raviart, P.-A.: *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Springer, 1996 (Applied Mathematical Science 118)

[35] Gottlieb, D. ; Gottlieb, S.: Spectral methods for compressible reactive flows. In: *C. R. Mecanique* 333 (2005), S. 3–16

[36] Gottlieb, D. ; Hesthaven, J. S.: Spectral methods for hyperbolic problems. In: *J. Comput. Appl. Math.* 128 (2001), S. 83–131

[37] Gottlieb, D. ; Lustman, L. ; Orszag, S. A.: Spectral Calculations of one-dimensional inviscid compressible flows. In: *SIAM J. Sci. Stat. Comput.* 2 (1981), Nr. 3, S. 296–310

[38] Gottlieb, D. ; Orszag, S. A.: *Numerical Analysis of Spectral Methods: Theory and Application*. SIAM, 1977 (Regional Conference Series in Applied Mathematics)

[39] Gottlieb, D. ; Shu, C.-W.: On the Gibbs phenomenon and its resolution. In: *SIAM Rev.* 39 (1997), S. 644–668

[40] Gottlieb, S. ; Shu, C.-W. ; Tadmor, E.: Strong Stability-Preserving High-Order Time Discretization Methods. In: *SIAM Review* 43 (2001), Nr. 1, S. 89112

[41] Grahs, T. ; Meister, A. ; Sonar, T.: Image processing for numerical approximations of conservation laws: nonlinear anisotropic artificial dissipation. In: *SIAM J. Sci. Comput.* 23 (2002), Nr. 5, S. 1439–1455

[42] Grahs, T. ; Sonar, T.: Entropy-controlled artificial anisotropic diffusion for the numerical solution of conservation laws based on algorithms from image processing. In: *J. Visual Commun. Image Repr.* 13 (2002), S. 176–184

[43] Hesthaven, J. S. ; Gottlieb, S. ; Gottlieb, D.: *Spectral Methods for Time-Dependent Problems*. Cambridge University Press, 2007

[44] Hesthaven, J. S. ; Kirby, R. M.: Filtering in Legendre Spectral Methods. In: *Math. Comput.* 77 (2008), Nr. 263, S. 1425–1452

[45] Hesthaven, J. S. ; Warburton, T.: *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis and Applications*. Springer, 2008

[46] Hirsch, C.: *Numerical Computation of Internal and External Flows Vol I: Fundamentals of Numerical Discretisation*. Wiley, 1988

[47] JAFFRE, J. ; JOHNSON, C. ; SZPESSY, A.: Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws. In: *Mathematical Models and Methods in Applied Sciences* 5 (1995), S. 367–386

[48] JIANG, G.-S. ; SHU, C.-W.: On cell entropy inequality for discontinuous Galerkin methods. 1993. – Forschungsbericht. – ICASE Report No. 93-37

[49] JIANG, G.-S. ; SHU, C.-W.: Efficient Implementation of Weighted ENO Schemes. In: *J. Comput. Phys.* 126 (1996), S. 202–228

[50] JOHNSON, C. ; PITKÄRANTA, J.: An Analysis of the Discontinuous Galerkin Method for a Scalar Hyperbolic Equation. In: *Math. Comput.* 46 (1986), Nr. 173, S. 1–26

[51] KABALLO, W.: *Einführung in die Analysis I.* 2nd ed. Spektrum Akademischer Verlag, 2000

[52] KARNIADAKIS, G. E. ; SHERWIN, S.: *Spectral/hp Element Methods for Computational Fluid Dynamics.* 2nd ed. Oxford University Press, 2005

[53] KIRBY, R. ; SHERWIN, S.: Stabilization of Spectral/hp Element Methods Through Spectral Vanishing Viscosity: Application to Fluid Mechanics Modelling. In: *Comput. Methods Appl. Mech. Engrg.* 195 (2006), S. 3128–3144

[54] KOORNWINDER, T.: Two-variable analogues of the classical orthogonal polynomials. In: ASKEY, R. (Hrsg.): *Theory and Applications of Special Functions.* San Diego : Academic Press, 1975

[55] KRIVODONOVA, L.: Limiters for high-order discontinuous Galerkin methods. In: *J. Comput. Phys.* 226 (2007), S. 879–896

[56] KRIVODONOVA, L. ; XIN, J. ; REMACLE, J.-F. ; CHEVAUGEON, N. ; FLAHERTY, J. E.: Shock detection and limiting with discontinuous Galerkin methods for hyperbolic conservation laws. In: *Appl. Numer. Math.* 48 (2004), S. 323–338

[57] KRUŽKOV, S. N.: First order quasilinear equations in several independent variables. In: *Math. USSR Sbornik* 10 (1970), S. 217–243

[58] KUBATKO, E. J. ; DAWSON, C. ; WESTERINK, J.J.: Time step restrictions for Runge-Kutta discontinuous Galerkin methods on triangular grids. In: *J. Comput. Phys.* 227 (2008), S. 9697–9710

[59] KURGANOV, A. ; PETROVA, G. ; POPOV, B: Adaptive Semidiscrete Central-Upwind Schemes for Nonconvex Hyperbolic Conservation Laws. In: *SIAM J. Sci. Comput.* 29 (2007), S. 2381–2401

[60] LAX, P. D.: Accuracy and Resolution in the Computation of Solutions of Linear and Nonlinear Equations. In: *Recent Advances in Numerical Analysis*, Academic Press, 1978, S. 107–117

[61] LEER, B. van: Flux-vector splitting for the Euler equations. In: *Lecture Notes in Physics* Bd. 170, 1982, S. 507–512. – Proceedings of the Eighth International Conference on Numerical Methods in Fluid Dynamics

[62] Lesaint, P. ; Raviart, P. A.: On a finite element method for solving the neutron transport equation. In: Boor, C. de (Hrsg.): *Mathematical Aspects of Finite Elements in Partial Differential Equations*. Academic Press, 1974, S. 89–145

[63] LeVeque, R. J.: *Numerical Methods for Conservation Laws*. Birkhäuser Verlag, 1990 (Lectures in Mathematics)

[64] Levin, J. G. ; Iskandarani, M. ; Haidvogel, D. B.: A Spectral Filtering Procedure for Eddy-Resolving Simulations with a Spectral Element Ocean Model. In: *J. Comput. Phys.* 137 (1997), S. 130–154

[65] Ma, H.: Chebyshev-Legendre spectral viscosity method for nonlinear conservation laws. In: *SIAM J. Numer. Anal.* 35 (1998), Nr. 3, S. 869–892

[66] Ma, H.: Chebyshev-Legendre super spectral viscosity method for nonlinear conservation laws. In: *SIAM J. Numer. Anal.* 35 (1998), Nr. 3, S. 893–908

[67] Maday, Y. ; Ould Kaber, S. M. ; Tadmor, E.: Legendre pseudospectral viscosity method for nonlinear conservation laws. In: *SIAM J. Numer. Anal.* 30 (1993), Nr. 2, S. 321–342

[68] Meister, A. ; Ortleb, S. ; Sonar, Th.: On Spectral Filtering for Discontinuous Galerkin Methods on Unstructured Triangular Grids. (2009). – Preprint: Mathematische Schriften Kassel

[69] Meister, A. ; Ortleb, S. ; Sonar, Th.: Application of Spectral Filtering to Discontinuous Galerkin Methods on Triangulations. (2011). – Zur Veröffentlichung in NMPDE angenommen

[70] Ortleb, S. ; Meister, A. ; Sonar, Th.: Adaptive Spectral Filtering and Digital Total Variation Postprocessing for the DG Method on Triangular Grids: Application to the Euler Equations. In: *Spectral and High Order Methods for Partial Differential Equations - Selected papers from the ICOSAHOM '09 conference, June 22-26, Trondheim, Norway, Lecture Notes in Computational Science and Engineering, Vol. 76*, 2010

[71] Ortleb, S. ; Meister, A. ; Sonar, Th.: Adaptive Spectral Filtering and DTV Postprocessing for the DG Method on Triangular Grids. In: *Proceedings of the Conference on Topical Problems of Fluid Mechanics 2011, Praha*, 2011

[72] Persson, P.-O. ; Peraire, J.: Sub-Cell Shock Capturing for Discontinuous Galerkin Methods. In: *Proceedings of the 44th AIAA Aerospace Sciences Meeting and Exhibit*, 2006. – AIAA-2006-112

[73] Peterson, T. E.: A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. In: *SIAM J. Numer. Anal.* 28 (1991), S. 133–140

[74] Proriol, J.: Sur une famille de polynomes à deux variables orthogonaux dans un triangle. In: *C. R. Acad. Sci. Paris* 245 (1957), S. 2459–2461

[75] QIU, J. ; SHU, C.-W.: Hermite WENO schemes and their application as limiters for Runge Kutta discontinuous Galerkin method: one dimensional case. In: *J. Comput. Phys.* 193 (2004), S. 115–135

[76] QIU, J. ; SHU, C.-W.: Hermite WENO schemes and their application as limiters for Runge Kutta discontinuous Galerkin method: two dimensional case. In: *Computers and Fluids* 34 (2005), S. 642–663

[77] QIU, J. ; SHU, C.-W.: Runge-Kutta discontinuous Galerkin method using WENO limiters. In: *SIAM J. Sci. Comput.* 26 (2005), Nr. 3, S. 907–929

[78] QIU, J. ; SHU, C.-W.: Convergence of High Order Finite Volume Weighted Essentially Nonoscillatory Scheme and Discontinuous Galerkin Method for Nonconvex Conservation Laws. In: *SIAM J. Sci. Comput.* 31 (2008), Nr. 1, S. 584–607

[79] REED, W. H. ; HILL, T. R.: Triangular mesh methods for the neutron transport equation / Los Alamos Scientific Laboratory. 1973. – Forschungsbericht. – LA-UR-73-479

[80] RICHTER, G. R.: An Optimal-Order Error Estimate for the Discontinuous Galerkin Method. In: *Math. Comput.* 50 (1988), Nr. 181, S. 75–88

[81] SARRA, S. A.: Digital total variation filtering as postprocessing for Chebyshev pseudospectral methods for conservation laws. In: *Numerical Algorithms* 41 (2006), S. 17–33

[82] SARRA, S. A.: Edge Detection Free Postprocessing for Pseudospectral Approximations. In: *J. Sci. Comput.* 41 (2009), S. 49–61

[83] SENGUPTA, K. ; MASHAYEK, F. ; JACOBS, G. B.: Large-eddy simulation using a discontinuous Galerkin spectral element method. In: *Proceedings of the 45th AIAA Aerospace Sciences Meeting and Exhibit*, 2007. – AIAA-2007-402

[84] SHU, C.-W. ; OSHER, S.: Efficient Implementation of essentially non-oscillatory shock capturing schemes II. In: *J. Comput. Phys.* 83 (1989), S. 32–78

[85] SHU, C.-W. ; WONG, P. S.: A Note on the Accuracy of Spectral Method Applied to Nonlinear Conservation Laws. In: *J. Sci. Comput.* 10 (1995), Nr. 3, S. 357–369

[86] SZEGÖ, G: *Orthogonal Polynomials*. Bd. 23. AMS Colloquium Publications, 1939

[87] TADMOR, E.: Convergence of Spectral Methods for Nonlinear Conservation Laws. In: *SIAM J. Numer. Anal.* 26 (1989), Nr. 1, S. 30–44

[88] TADMOR, E.: Super Viscosity and Spectral Approximations of Nonlinear Conservation Laws. In: *Numerical Methods for Fluid Dynamics* 4 (1993), S. 69–82

[89] TAYLOR, M. ; TRIBBIA, J. ; ISKANDARANI, M.: The Spectral Element Method for the Shallow Water Equations on the Sphere. In: *J. Comput. Phys.* 130 (1997), S. 92–108

[90] Taylor, M. A. ; Wingate, B.A. ; Bos, L.P.: A cardinal function algorithm for computing multivariate quadrature points. In: *SIAM J. Numer. Anal.* 45 (2007), Nr. 1, S. 193–205

[91] Toulorge, T. ; Desmet, W.: Time stepping and linear stability of Runge-Kutta discontinuous Galerkin methods on triangular grids. In: *Proceedings of the 5th European Conference on Computational Fluid Dynamics (ECCOMAS CFD 2010)*, 2010

[92] Vandeven, H.: Family of spectral filters for discontinuous problems. In: *J. Sci. Comput.* 6 (1991), S. 159–192

[93] Warburton, T. C.: *Spectral/hp methods on polymorphic multi-domains: algorithms and applications*. Providence, RI, Brown University, Diss., 1998

[94] Warnecke, G.: *Analytische Methoden in der Theorie der Erhaltunggleichungen*. Teubner, 1999

[95] Wei, G. W.: Shock capturing by anisotropic diffusion oscillation reduction. In: *Comp. Phys. Commun.* 144 (2002), S. 317–342

[96] Weickert, J.: *Anisotropic Diffusion in Image Processing*. Teubner, 1998

[97] Whitham, G.: *Linear and Nonlinear Waves*. Wiley-Interscience, 1974

[98] Wingate, B. A. ; Taylor, M. A.: The natural function space for triangular and tetrahedral spectral elements / Los Alamos National Laboratory. 1998. – Forschungsbericht. – LA-UR-98-1711

[99] Woodward, P. ; Colella, P. R.: The Numerical Simulation of Two-Dimensional Fluid Flow with Strong Shocks. In: *J. Comput. Phys.* 54 (1984), S. 115–173

[100] Yang, M. ; Wang, Z.J.: A Parameter-Free Generalized Moment Limiter for High-Order Methods on Unstructured Grids. In: *Proceedings of the 47th AIAA Aerospace Sciences Meeting and Exhibit*, 2009. – AIAA-2009-605

[101] Zhang, L. ; Cui, T. ; Liu, H.: A set of symmetric quadrature rules on triangles and tetrahedra. In: *J. Comput. Math.* 27 (2009), Nr. 1, S. 89–96

[102] Zhang, Q. ; Shu, C.-W.: Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws. In: *SIAM J. Numer. Anal.* 42 (2004), Nr. 2, S. 641–666

[103] Zhang, Q. ; Shu, C.-W.: Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws. In: *SIAM J. Numer. Anal.* 44 (2006), Nr. 4, S. 1703–1720

[104] Zhang, X. ; Shu, C.-W.: On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. In: *J. Comput. Phys.* 229 (2010), S. 8918–8934

[105] ZHU, J. ; QIU, J. ; SHU, C.-W. ; DUMBSER, M.: Runge-Kutta discontinuous Galerkin method using WENO limiters II: Unstructured meshes. In: *J. Comput. Phys.* 227 (2008), S. 4330–4353

[106] ZYGMUND, A.: *Trigonometric series.* 2nd rev. ed. Cambridge Univ. Press, 1968

ISBN 978-3-86219-218-2